Data Science for the Big Data Review

Story
New Commerce Department report explores huge benefits, low cost of government data
Fostering Innovation, Creating Jobs, Driving Better Decisions: The Value of Government Data

The Value of Government Data
Foreword
Acknowledgments
Executive Summary
Government data potentially guides trillions of dollars of investments each year
The cost of Government data is small relative to its potential benefits
Government data is commercially valuable

I. Introduction

II. Government Data: Statistics for Better Decisions
What is Government Data?
Box 1: Data versus Statistics versus Information
Box 2: BrightScope Creates Value-Added Products Using Data Embedded in Public Documents

Spending on Government Statistics
Table 1: Principal Statistical Agencies of the Federal Government
Figure 1: Federal Government Spending on Principal Statistical Agencies
Figure 2: Federal Budget Authority for Selected Functions, 2004-2013 Average
Table 2: Gross Domestic Investment

Government Data Enables Better Decision Making
Data-Driven Decisions: Improving Downtown Life
Box 3: Government Data from the National Oceanic and Atmospheric Administration Parallels with the Principal Federal Statistical Agencies
Data-Driven Decisions: Screening for Cancer
Box 4: The Principal Federal Economic Indicators
Data-Driven Decisions: Transportation Planning

III. The Role of Government in Providing Data
Data Driven Decisions: Customizing Retail Product Selections
Addressing Market Failures to Provide More Information
Data-Driven Decisions: Helping Businesses Export

The Advantages of Government Production of Statistics and Other Data

Data-Driven Decisions: Choosing a Career

1. Comprehensive
   Data-Driven Decisions: Locating Hospitals

2. Consistent
   Data-Driven Decisions: Re-aligning a School District

3. Confidential
   Data-Driven Decisions: Crime Mapping

4. Credible
   Data-Driven Decisions: Setting Contract Prices

5. Relevant
   Data-Driven Decisions: Serving the Church Community

Box 5: Measuring E-Commerce: Government Data Sets the Standard After Building on Initial Private Sector Efforts

6. Accessible

The Complementary Relationship between Government and Private Sector Data

IV. The Commercial Value of Government Data

Defining Government Data-Intensive Business Activities

Figure 3: Government Data Value Chain

Estimating the Size of the Government Data-Intensive Sector

Data-Driven Decisions: Stabilizing Rents

Data-Driven Decisions: Locating a Business Establishment

Data-Driven Decisions: Investing in Real Estate

Acxiom

Bloomberg, L.P

Environmental Systems Research Institute (ESRI)

Gallup

GeoLytics, Inc.

GreatSchools

Haver Analytics

IHS Inc.

MapQuest, Inc.

Nielsen

Thomson Reuters

Trulia

Zillow

1. Lower-Bound Estimate of Size of the GDIS Based on Size of Selected Firms Active in the GDIS

https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review

Updated: Thu, 25 Jul 2019 08:57:25 GMT

Powered by mindtouch
Table 3: Revenues and Employment of Firms in the Government Data-Intensive Sector

2. Upper-Bound Estimate of the Size of the GDIS Based on the Size of Government-Data-Intensive Industries

Table 4: Revenues of Government Data-Intensive Sector (GDIS) by Industry

V. Conclusion

Footnotes

F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
F11
F12
F13
F14
F15
F16
F17
F18
F19
F20
F21
F22
F23

References

Appendix I: Definitions of Terms Used

Administrative Data
Analysts
Benchmarkers
Data Brokers
Government Data
Government data-intensive sector (GDIS)
North American Industry Classification System (NAICS)
The Study

President Obama Discusses U.S. Intelligence Programs at the Department of Justice

Big Data Report Fact Sheet
Big Data Report Blog Post

Read the study from the President’s Council of Advisors on Science and Technology

PCAST Releases Report on Big Data and Privacy

See responses to the Request for Information

Share Your Input

Big Data: Seizing Opportunities, Preserving Values

I. Big Data and the Individual

What is Big Data?

What is a Zettabyte?
The “Internet of Things”
The Scope of This Review

What is Different about Big Data?
The "3 Vs": Volume, Variety and Velocity
What are the sources of big data?
New Opportunities, New Challenges

The power and opportunity of big data applications
Finding the needle in the haystack
The benefits and consequences of perfect personalization
De-identification and re-identification
The persistence of data

Affirming our Values

II. The Obama Administration’s Approach to Open Data and Privacy

Open Data in the Obama Administration

Open Data Initiatives
My Data Initiatives
Big Data Initiative: “Data to Knowledge to Action”

U.S. Privacy Law and International Privacy Frameworks

The Fair Information Practice Principles
Sector-Specific Privacy Laws in the United States
Consumer Privacy Bill of Rights
Promoting Global Interoperability

Conclusion

III. Public Sector Management of Data

Big Data and Health Care Delivery

Learning about Learning: Big Data and Education
Protecting Children’s Privacy in the Era of Big Data

Big Data at the Department of Homeland Security
A Model for Managing Data

Upholding our Privacy Values in Law Enforcement
New Tools and New Challenges

Implications of Big Data Technology for Privacy Law
Access to Data Held by Third Parties
Data and Metadata
Government Use of Commercial Data Services
Insider Threat and Continuous Evaluation
Conclusion

IV. Private Sector Management of Data

Big Data Benefits for Enterprise and Consumer
The Advertising-Supported Ecosystem
The Consumer and the Challenge of Transparency
The Challenge with Do Not Track

The Data Services Sector
Unregulated Data Broker Services
What is a Credit Reporting Agency?
Algorithms, Alternative Scoring and the Specter of Discrimination
What is an algorithm?
Conclusion

V. Toward a Policy Framework for Big Data

Big Data and the Citizen
Big Data and the Consumer
Big Data and Discrimination
Big Data and Privacy

Federal Research in Privacy-Enhancing Technologies
Anticipating the Big Data Revolution’s Next Chapter

VI. Conclusion and Recommendations

Policy Recommendations

1. Preserving Privacy Values
2. Responsible Educational Innovation in the Digital Age
3. Big Data and Discrimination
4. Law Enforcement and Security
5. Data as a Public Resource

References

1.
Appendix

A. Methodology
B. Stakeholder Meetings
C. Academic Symposia
   - Big Data and Privacy Workshop: Advancing the State of the Art in Technology and Practice
   - The Social, Cultural, & Ethical Dimensions of 'Big Data'
   - Big Data: Values and Governance
D. PCAST Report
E. Public Request for Information
F. White House Big Data Survey
   - Figure Concern With Data Practices
   - Figure Percent Who Do Not Trust Each Entity At All

Fact Sheet: PCAST Report on Big Data and Privacy: A Technological Perspective

Big Data and Privacy: A Technology Perspective

The President's Council of Advisors on Science and Technology

PCAST Big Data and Privacy Working Group
   - Working Group Co-Chairs
   - Working Group Members
   - Working Group Staff

Letter

Executive Summary

1. Introduction
   1.1 Context and outline of this report
   1.2 Technology has long driven the meaning of privacy
   1.3 What is different today?
   1.4 Values, harms, and rights

2. Examples and Scenarios
   2.1 Things happening today or very soon
   2.2 Scenarios of the near future in healthcare and education
      2.2.1 Healthcare: personalized medicine
2.2.2 Healthcare: detection of symptoms by mobile devices
2.2.3 Education

2.3 Challenges to the home’s special status
2.4 Tradeoffs among privacy, security, and convenience

3. Collection, Analytics, and Supporting Infrastructure
3.1 Electronic sources of personal data
   3.1.1 “Born digital” data
   3.1.2 Data from sensors
3.2 Big data analytics
   3.2.1 Data mining
   3.2.2 Data fusion and information integration
   3.2.3 Image and speech recognition
   3.2.4 Social-network analysis
3.3 The infrastructure behind big data
   3.3.1 Data centers
   3.3.2 The cloud

4. Technologies and Strategies for Privacy Protection
4.1 The relationship between cybersecurity and privacy
4.2 Cryptography and encryption
   4.2.1 Well Established encryption technology
   4.2.2 Encryption frontiers
4.3 Notice and consent
4.4 Other strategies and techniques
   4.4.1 Anonymization or de-identification
   4.4.2 Deletion and non-retention
4.5 Robust technologies going forward
   4.5.1 A Successor to Notice and Consent
   4.5.2 Context and Use
   4.5.3 Enforcement and deterrence
   4.5.4 Operationalizing the Consumer Privacy Bill of Rights

5. PCAST Perspectives and Conclusions
5.1 Technical feasibility of policy interventions
5.2 Recommendations
5.3 MISSING
5.4 Final Remarks

References
1
2
Appendix A. Additional Experts Providing Input

Special Acknowledgment

NSF Technology Team Leaders
Additional NSF Contributors

Story

Commerce Gets Chief Data Officer and Releases Report on Value of Statistical Data

My suggestions at the big data workshop hosted by the White House Office of Science and Technology Policy and the Georgetown University McCourt School of Public Policy’s Massive Data Institute to Mark Doms Undersecretary for Economic Affairs, Department of Commerce, was they get a chief data officer and stress the value of the government’s statistical data - they did that! Good people listen to good advice.
New Commerce Department report explores huge benefits, low cost of government data

Source: http://www.commerce.gov/blog/2014/07...overnment-data

Submitted on July 14, 2014 - 5:59 pm

Fostering Innovation, Creating Jobs, Driving Better Decisions: The Value of Government Data

Cross post by Mark Doms, Under Secretary for Economic Affairs

Today we are pleased to roll out an important new Commerce Department report on government data. "Fostering Innovation, Creating Jobs, Driving Better Decisions: The Value of Government Data," arrives as our society increasingly focuses on how the intelligent use of data can make our businesses more competitive, our governments smarter, and our citizens better informed.

And when it comes to data, as the Under Secretary for Economic Affairs, I have a special appreciation for the Commerce Department’s two preeminent statistical agencies, the Census Bureau and the Bureau of Economic Analysis. These agencies inform us on how our $17 trillion economy is evolving and how our population (318 million and counting) is changing, data critical to our country. Although “Big Data” is all the rage these days, the government has been in this business for a long time: the first Decennial Census was in 1790, gathering information on close to four million people, a huge dataset for its day, and not too shabby by today’s standards as well.

Just how valuable is the data we provide? Our report seeks to answer this question by exploring the range of federal statistics and how they are applied in decision-making. Examples of our data include gross domestic product, employment, consumer prices, corporate profits, retail sales, agricultural supply and demand, population, international trade and much more.

Clearly, as shown in the report, the value of this information to our society far exceeds its cost – and not just because the price tag is shockingly low: three cents, per person, per day. Federal statistics guide trillions of dollars in annual investments at an average annual cost of $3.7 billion: just 0.02 percent of our $17 trillion dollar economy covers the massive amount of data collection, processing and dissemination. With a statistical system that is comprehensive, consistent, confidential, relevant and accessible, the federal government is uniquely positioned to provide a wide range of statistics that complement the vast and growing sources of private sector data.
Our federally collected information is frequently “invisible,” because attribution is not required. But it flows daily into myriad commercial products and services. Today’s report identifies the industries that intensively use our data and provides a rough estimate of the size of this sector. The lower-bound estimate suggests government statistics help private firms generate revenues of at least $24 billion annually – more than six times what we spend for the data. The upper-bound estimate suggests annual revenues of $221 billion!

This report takes a first crack at putting an actual dollars and cents value to government data. We’ve learned a lot from this initial study, and look forward to honing in even further on that figure in our next report.

Fostering Innovation, Creating Jobs, Driving Better Decisions: The Value of Government Data

Source: http://www.esa.doc.gov/Reports/fost...overnment-data

Submitted on July 11, 2014 - 1:17pm

Executive Summary

Everyone is talking about the importance of data to our society as data improves all of our decisions: those we make as individuals, as businesses, as governments.

Government has been in the data business for quite some time, going back to the first Decennial Census in 1790. Since then, the U.S. Government has played a key role in providing valuable data to our country.

Just how valuable is the data the Federal Government provides? That question can’t be answered precisely, but there are many reasons to believe that the value of the data to our society far exceeds its cost. As a first step in ascertaining better estimates of just how valuable Government data is, this report focuses on Federal statistical data, data that informs us about our huge, complex, and dynamic economy and data that tells us about our ever changing population. Examples of Government statistical data include gross domestic product, employment and unemployment, consumer prices, retail sales, housing vacancies, residential construction, agricultural supply and demand, corporate profits, and international trade; there are many more.

This report finds:

Government data potentially guides trillions of dollars of investments each year.

Government data helps governments to better target scarce resources, businesses stay competitive, and individuals stay informed about the communities in which they live. As the real world examples in this report demonstrate, individuals, businesses, other organizations, and governments use Government data to help make better informed decisions that are better, faster, and more plentiful because of the ready availability and high quality of Government data.

The cost of Government data is small relative to its potential benefits.
Since 2004, the Federal Government’s principal statistical agencies have spent an average of $3.7 billion annually on data collection, processing, and dissemination. This expenditure amounts to about three cents, per person, per day, and is only 0.02 percent of our roughly $17 trillion dollar economy.

**Government data is uniquely comprehensive, consistent, confidential, credible, relevant, and accessible.**

Acting alone, the private sector would likely provide only some of the types of data produced by the Government. The Federal Government is uniquely positioned to provide comprehensive, consistent, credible, relevant and accessible data, all while protecting confidentiality. Government data is used directly to support decision making, indirectly through commercially available value-added products, and as a benchmark and standard for private data products. Thus, the Government and the private sector complement each other.

**Government data is commercially valuable.**

Government data is a key input to a wide variety of commercial products and services in the economy, although many of these uses may not be apparent because attribution to the Government is not required. This report identifies industries that use Government data intensively and provides a rough estimate of the size of this sector. The lower-bound estimate, based on a very short and incomplete list of firms that rely heavily on Government data, suggests that Government data helps private firms generate revenues of at least $24 billion annually–far in excess of spending on Government statistical data. The upper-bound estimate suggests that this sector generates annual revenues of $221 billion. These crude estimates provide rough order-of-magnitude estimates of the range of the sector’s size and illustrate the importance of Government data as an input into commercial products and services.

The value of Government data is the extent to which the decisions based on Government data are better than the decisions that would have been made without Government data. Because such counter-factual evidence is not widely available, this report makes the case that the value of Government data far outweighs its cost by outlining the economic arguments for Government provision of data, by estimating the size of the industry sector that intensively uses Government data, and by illustrating the wide range of ways in which Government data is used.

---

**The Value of Government Data**

Source:

application/pdf icon

fosteringinnovationcreatingjobsdrivingbetterdecisionsthevalueofgovernmentdata714.pdf (PDF)
Foreword

In March, The Department of Commerce unveiled the America in Focus for Business: Strategic Plan, Fiscal Years 2014-2018. One of the plan’s five priority areas is to redefine how we manage, optimize, and enable public access to our treasure trove of data. Specifically, the plan pledges to improve government, business, and community decisions and knowledge by transforming Department of Commerce data capabilities and supporting a data-enabled economy.

The Department’s efforts are part of a broader Open Data Initiative, which has embraced innovation as a means of further unlocking the value of Government data. The initiative focuses on making raw data available to the public while rigorously protecting privacy; putting data into forms that are more accessible and usable; and making entrepreneurs and innovators aware of Government data and how to access it.

Here at the Department of Commerce, our efforts will focus on three objectives. First, every person in America should have easy access to reliable information about their communities, their climate, and how they are changing. Second, every business should have easy access to reliable information on its market, potential markets, scientific information, and changing economic conditions. Further, new data-driven businesses should be able to easily pull our data, combine it with other information, and make new products to compete in the private marketplace. Third, and finally, every government should have easy access to the information it needs to better serve its communities and to assess the efficiency of its programs.

The following report, Rethinking Innovation: Creating Jobs, Driving Better Decisions: The Value of Government Data, provides a helpful lay of the land in assessing where we are in our efforts. Focusing in particular on the value of statistical data produced by Federal agencies, this report explains what Government brings to the table in providing data and describes the wide range of sectors that are supported by Government data. Finally, it highlights the importance of collaboration between Government and the private sector by examining the use of the private sector that intensively uses Government data.

As the Government continues to push forward in its efforts to unlock data, this data-driven sector of the economy will continue to grow, guiding decisions across the country, helping make government smarter, businesses more competitive, and citizens more informed.

Penney Pritzker
U.S. Secretary of Commerce
Acknowledgments

This report was prepared by Regina Powers and David Beede in the Office of the Chief Economist, Economics and Statistics Administration.

The authors would like to thank the following persons who provided comments, suggestions, and other contributions to this report. Of course, any errors in the report are solely the authors’ responsibility.

**Economics and Statistics Administration:**
- Mark Doms, Under Secretary of Commerce for Economic Affairs
- Kenneth Arnold, Deputy Under Secretary of Commerce for Economic Affairs
- Sue Helper, the Chief Economist of the Department of Commerce
- Austin Durrer, Chief of Staff
- Robert Rubinovitz, the Deputy Chief Economist of the Department of Commerce
- Sabrina Montes, Section Chief in the Office of the Chief Economist
- Richard Lattimer and Kenneth White of the Office of Administration
- Lucas Hitt, Acting Director for Communications (on detail from the Bureau of Economic Analysis), and Jane Callen of the Communications

**Office of Management and Budget:** Paul Bugg, Melissa Chiu, and Brian Harris-Kojetin

**Bureau of Labor Statistics:** Erica L. Groshen, Commissioner; John Eltinge, Associate Commissioner for Survey Methods Research; John W. Ruser, Associate Commissioner for Productivity and Technology; David Friedman, Assistant Commissioner for Industrial Prices and Price Indexes of the Bureau of Labor Statistics; Michael D. Levi, Associate Commissioner, Office of Publications and Special Studies; and Kathy Frederickson-Mele, Deputy Associate Commissioner, Office of Publications and Special Studies

**Bureau of Economic Analysis:** Brian C. Moyer, Deputy Director and Acting Director; Dennis J. Fixler, Chief Statistician; Kathleen James, Chief Administrative Officer; Brent R. Moulton, Associate Director for National Economics; Sarahelen Thompson, Associate Director for International Economics; Joel D. Platt, Associated Director for Regional Economics; and Carol Moylan, Associate Director for Industry Economics

**Census Bureau, Economic Programs Division:** Shirin A. Ahmed, Assistant Director for Economic Programs, and Lisa Blumerman, William Davie, Carrie Dennis, Charles Funk, Mendel Gayle, Andrew Hait, Sam Jones, Steve Owens, Matt Przybocki, and Deb Stempowski

**Census Bureau, Demographic Division:** Enrique Lamas, Associate Director for Demographic Programs, Eloise Parker, Assistant Director for Demographic Programs, and David Johnson (now Chief Economist, Bureau of Economic Analysis), Ruth Ann Killion, Timothy P. Olsen, Barry Sessamen, and Tori Velkoff
Executive Summary

Everyone is talking about the importance of data to our society as data improves all of our decisions: those we make as individuals, as businesses, as governments.

Government has been in the data business for quite some time, going back to the first Decennial Census in 1790. Since then, the U.S. Government has played a key role in providing valuable data to our country.

Just how valuable is the data the Federal Government provides? That question can’t be answered precisely, but there are many reasons to believe that the value of the data to our society far exceeds its cost. As a first step in ascertaining better estimates of just how valuable Government data is, this report focuses on Federal statistical data, data that informs us about our huge, complex, and dynamic economy and data that tells us about our ever changing population. Examples of Government statistical data include gross domestic product, employment and unemployment, consumer prices, retail sales, housing vacancies, residential construction, agricultural supply and demand, corporate profits, and international trade; there are many more.

This report finds:

Government data potentially guides trillions of dollars of investments each year

Government data helps governments to better target scarce resources, businesses stay competitive, and individuals stay informed about the communities in which they live. As the real world examples in this report demonstrate, individuals, businesses, other organizations, and governments use Government data to help make better informed decisions that are better, faster, and more plentiful because of the ready availability and high quality of Government data.

The cost of Government data is small relative to its potential benefits

Since 2004, the Federal Government’s principal statistical agencies have spent an average of $3.7 billion annually on data collection, processing, and dissemination. This expenditure amounts to about three cents, per person, per day, and is only 0.02 percent of our roughly $17 trillion dollar economy. Government data is uniquely comprehensive, consistent, confidential, credible, relevant, and accessible. Acting alone, the private sector would likely provide only some of the types of data produced by the Government. The Federal Government is uniquely positioned to provide comprehensive, consistent, credible, relevant and accessible data, all while protecting confidentiality. Government data is used directly to support decision making, indirectly through commercially available value-added products, and as a benchmark and standard for private data products. Thus, the Government and the private sector complement each other.

Government data is commercially valuable

Government data is a key input to a wide variety of commercial products and services in the economy, although many of these uses may not be apparent because attribution to the Government is not required. This report identifies industries...
that use Government data intensively and provides a rough estimate of the size of this sector. The lower-bound estimate, based on a very short and incomplete list of firms that rely heavily on Government data, suggests that Government data helps private firms generate revenues of at least $24 billion annually—far in excess of spending on Government statistical data. The upper-bound estimate suggests that this sector generates annual revenues of $221 billion. These crude estimates provide rough order-of-magnitude estimates of the range of the sector's size and illustrate the importance of Government data as an input into commercial products and services.

The value of Government data is the extent to which the decisions based on Government data are better than the decisions that would have been made without Government data. Because such counter-factual evidence is not widely available, this report makes the case that the value of Government data far outweighs its cost by outlining the economic arguments for Government provision of data, by estimating the size of the industry sector that intensively uses Government data, and by illustrating the wide range of ways in which Government data is used.

I. Introduction

Modern advances in collecting, processing, disseminating, and preserving information have resulted in the proliferation of data from a wide variety of sources. The increase in the amount and variety of data that is available has highlighted long-standing questions about the role of the Federal Government in collecting and disseminating data. Why does the Government, in certain cases, mandate the collection of personal and private data from individuals, households, and businesses? Are collecting and disseminating data appropriate activities for Government or better left to the private sector? What happens to the data after the Government collects it? Do the benefits of the data outweigh the costs?

It is fairly straightforward to empirically estimate the costs to the Government of collecting, processing, and disseminating data. Measuring the benefits of Government data—the value created when people, businesses, and institutions use the data to make better decisions—is more challenging. Unlike goods and services sold on the open market, Government data—like most services provided by the Federal Government — is freely available to the public. Because Government data is not traded in the marketplace, it has no price that can be used to estimate its value.

Generally, people are unaware of the full range of both public and private sector products and services for which Government data is a vital ingredient. One reason that many uses of Government data are not visible is that there are no copyrights for Federal Government data and anyone can use the data without attribution. The combination of free dissemination of Government data and its use without required attribution has made it difficult to identity all users and uses of the data, further complicating efforts to estimate the value of the data. Because we cannot observe how decisions would be made in the absence of Government data, it is difficult to estimate the incremental value of the improvement in decision making that is made possible with Government data. Despite the measurement challenges, it is worthwhile to understand the value of Government data, particularly given the rapidly changing landscape of data production and dissemination.

The remainder of this report is organized as follows:

Section II defines Government data and discusses the costs of producing Federal statistics, as well as how those statistics are used for decision making that affects people throughout the Nation. In addition, sidebars throughout the report provide examples of data-driven decisions made by individuals, businesses, other organizations, and governments informed by Government data.
Section III discusses the rationale for Government provision of data and, focusing on the Federal statistical system, describes why the Government is involved in providing data and outlines the advantages Government often has over the private sector in collecting and disseminating statistical data. It also discusses the important role of the private sector in complementing and adding value to Government data to provide specific information needs in the market.

Section IV further explores the commercial value of Government data, illustrating the extent of the use of Government data by the private sector. It discusses estimates of the size of the Government data-intensive sector (GDIS)—the private industries that rely on Government data as a critical input to the goods and services they produce.

Section V summarizes the limited evidence about the value of Government data and concludes that the commercial value of Government data, as approximated by the revenues of data-intensive firms, far exceeds the costs of Government data. Moreover, because of its high quality and public good characteristics, the total value of Government data is greater than just its commercial value.

II. Government Data: Statistics for Better Decisions

What is Government Data?

The term “Government data” includes diverse categories of Government information products, such as the following:

- **Information published by Government agencies for statistical purposes.** Such information is compiled from raw data collected directly from individuals, households, businesses, other organizations, and governments. The information may also be indirectly compiled from Government administrative data, private sector data, and other sources. Statistical agencies summarize the data in tabulations to describe specific characteristics of groups of persons, businesses, or other entities. The agencies make such information available for use by other Government agencies and the general public to inform decision making, in program analysis and evaluation, and in research.

- **Information in Government administrative records on individuals and businesses.** Examples include publicly available data about private sector company-sponsored retirement plans reported to the Labor Department on Form 5500; Securities and Exchange Commission filings; Government contracts data; real estate appraisal records; and voter registration rolls. Historically, a Federal or State Freedom of Information Act request has often been required to obtain access to this category of information, but recent “open data” efforts have aimed to make such information more easily accessible. Other data is strictly confidential and therefore not publicly released, such as tax return data collected by the Internal Revenue Service or Social Security Administration data.

- **Physical measurements about natural phenomena recorded by Government agencies.** Examples include weather and climate data such as local temperature readings collected by the National Weather Service of the National Oceanic and Atmospheric Administration, Global Positioning System data, and measurements of the characteristics of elements and molecules collected and published by the National Institute for Standards and Technology Physical Measurement Laboratory.

**Box 1: Data versus Statistics versus Information**

The terms “data,” “statistics,” and “information” each have distinctly different formal definitions. Data is the least abstract concept – it consists of a list of discernible variations in the characteristics of measurable states-of-the-world (Boisot and Canals, 2004). For example, a datum may be the fact that person A lives in New York City, is married, and earns $73,000 per year; while another datum may be the fact that person B lives in Omaha, is unmarried, and earns $42,000 a year. Statistics are measures of some attribute of a set of data, such as the average earnings per year of unmarried persons living in New York City. Information is a set of noteworthy regularities embedded in the
In this report, we focus on a subset of the first category: Government data published by the Principal Federal Statistical Agencies within the Executive Branch of the Federal Government. These agencies’ core missions are to collect, compile, process, analyze, and disseminate information for statistical purposes. Table 1 lists these agencies and their average annual budgets for 2004-2013. It is notable that these budgets do not account for all spending on Federal statistics. In fact, funding for the statistical programs of the Principal Statistical Agencies (excluding the Decennial Census, which is associated with a surge in expenditures every decade) accounts for about one-third of all Government funding for statistics (Office of Management and Budget, 2013a, p. 4). The remaining two-thirds share of the funding for statistical purposes is devoted to programs conducted within various agencies to help them carry out their missions. For example, the Environmental Protection Agency collects data to enforce environmental regulations; the Occupational Safety and Health Administration collects data to enforce occupational health and safety regulations; and the Employee Benefits Security Administration collects data to enforce regulations of employee retirement benefits (Office of Management and Budget, 2013a, p. 5).

**Box 2: BrightScope Creates Value-Added Products Using Data Embedded in Public Documents**

BrightScope is a financial information company that provides ratings and investment analysis of retirement plans for participants and sponsors, asset managers, and financial advisors. The company maintains a comprehensive database of information on the retirement plan market and produces quantitative ratings of over 45,000 401(k) and 403(b) plans. These ratings are based on more than 200 unique data inputs per plan. BrightScope also publishes distribution rankings for funds in the retirement marketplace, allowing comparisons of the performance of asset managers. The company markets a suite of data analytics software products to Fortune 1000 companies, asset managers, broker-dealers, financial advisors, and other market participants. Definitions, criteria, and methodologies are available for free at the company’s website. BrightScope also encourages and enables plan sponsors and participants to upload current data about their retirement plans. ([http://www.brightscope.com/about/](http://www.brightscope.com/about/))

Brightscope relies upon key information from the Department of Labor’s Form 5500 Series. Employee benefit plans use the Form 5500 Series to satisfy annual reporting requirements under Title I and Title IV of the Employee Retirement Income Security Act of 1974 (ERISA) and under the Internal Revenue Code. The Form 5500 is an important compliance, research, and disclosure tool for the Department of Labor, a disclosure document for plan participants and beneficiaries, and a source of information and data for use by other Federal agencies, Congress, and the private sector in assessing employee benefit, tax, and economic trends and policies. The Form 5500 Series is part of ERISA’s overall reporting and disclosure framework, which is intended to assure that employee benefit plans are operated and managed in accordance with certain prescribed standards and that participants and
beneficiaries, as well as regulators, are provided or have access to sufficient information to protect the rights and benefits of participants and beneficiaries under employee benefit plans.  

(\url{http://www.dol.gov/ebsa/5500main.html})

### Spending on Government Statistics

On average between 2004 and 2013, the United States spent $3.7 billion annually, adjusted for inflation, on data collection and dissemination by the Principal Statistical Agencies (Table 1).  

The Decennial Census is associated with a surge in expense. Excluding the Decennial Census, the average is about $2.3 billion. (See Figure 1.)

#### Table 1: Principal Statistical Agencies of the Federal Government

<table>
<thead>
<tr>
<th>Agency</th>
<th>Parent Department or Agency</th>
<th>Average Annual Budget Authority 2004-2013 (millions of 2013 dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bureau of Economic Analysis</td>
<td>Commerce</td>
<td>$90</td>
</tr>
<tr>
<td>Bureau of Justice Statistics</td>
<td>Justice</td>
<td>55</td>
</tr>
<tr>
<td>Bureau of Labor Statistics</td>
<td>Labor</td>
<td>620</td>
</tr>
<tr>
<td>Bureau of Transportation Statistics</td>
<td>Transportation</td>
<td>29</td>
</tr>
<tr>
<td>Bureau of the Census</td>
<td>vCommerce</td>
<td>2,029</td>
</tr>
<tr>
<td>Economic Research Service</td>
<td>Agriculture</td>
<td>84</td>
</tr>
<tr>
<td>Energy Information Administration</td>
<td>Energy</td>
<td>104</td>
</tr>
<tr>
<td>National Agricultural Statistics Service</td>
<td>Agriculture</td>
<td>162</td>
</tr>
<tr>
<td>National Center for Education Statistics</td>
<td>Education</td>
<td>244</td>
</tr>
<tr>
<td>National Center for Health Statistics</td>
<td>Health and Human Services</td>
<td>135</td>
</tr>
<tr>
<td>National Center for Science and Engineering Statistics*</td>
<td>National Science Foundation</td>
<td>35</td>
</tr>
<tr>
<td>Office of Research and Statistics</td>
<td>Social Security Administration</td>
<td>28</td>
</tr>
<tr>
<td>Statistics of Income</td>
<td>Division Treasury</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>$3,657</td>
<td></td>
</tr>
</tbody>
</table>
Note: Budget information compiled from Statistical Programs of the U.S. Government and actual agency budgets. Budget amounts include spending on the Decennial Census and are converted to 2013 dollars using the Federal Government Consumption Expenditures price index.

* Formerly the Science Resources Statistics Division

**Figure 1: Federal Government Spending on Principal Statistical Agencies**

![Figure 1: Federal Government Spending on Principal Statistical Agencies](image)

Sources: Budget information compiled from Analytical Perspectives, President's Budget; Statistical Programs of the U.S. Government Supplement to President's Budget; actual agency budgets; Principles and Practices for a Federal Statistical Agency.

Note: Budget amounts converted to real 2013 dollars using Government Consumption Expenditures deflator.

Figure 2 puts the $3.7 billion average annual spending by the Principal Federal Statistical Agencies in perspective by comparing it to spending on selected other functions of the Federal Government. This chart illustrates that spending on the Principal Statistical Agencies is relatively small compared to spending on many other important government functions. Another way to put this spending in perspective is to realize that $3.7 billion is roughly 0.022 percent of 2013 GDP of nearl
Government data also could be considered a form of investment, even though they are not currently treated as investments in the National Income and Product Accounts. Employment, gross domestic product, and other Federal Government statistics are information assets built up from cumulative spending on data collection, processing and analysis, much like knowledge accumulated from years of spending on research and development. Therefore, yet another way to put expenditures on Government data in perspective is to note that the $3.7 billion average annual expenditure on (or investment in) the Federal Principal Statistical Agencies is approximately one-tenth of one percent of total annual economy-wide investments, that is, 0.11 percent of the $3.276 trillion average annual total gross domestic investment for 2013 (see Table 2).

**Table 2: Gross Domestic Investment**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic business</td>
<td>$2,108</td>
</tr>
<tr>
<td>Households and institutions</td>
<td>562</td>
</tr>
<tr>
<td>Federal Government</td>
<td>275</td>
</tr>
<tr>
<td>State and local governments</td>
<td>332</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>$3,276</strong></td>
</tr>
</tbody>
</table>

Source: Bureau of Economic Analysis, 2014.
Government Data Enables Better Decision Making

Information is valuable to the extent it affects action (Hirshleifer, 1972, p. 564). Information produced by the Federal Government affects action by aiding people in their decisions. Government data thus increases in value when it improves people’s decisions. This report illustrates how Government data informs investment decisions by people, businesses, other organizations, and governments at every level, providing evidence that, absent Government data, the loss of information would degrade the quality of decision making. In particular, given the pervasive use of Government statistics to inform investment decisions throughout the economy, the absence of such data would likely result in an annual reduction in the return on aggregate investment far greater than the $3.7 billion average annual spending associated with producing the data.

Government data provides important fact-based evidence to support a wide range of decisions in the public and private sectors. The data used for these decisions may come directly from a Government agency through its public dissemination activities, or it may come indirectly from third party intermediaries. In any event, one important concept in thinking about the value of Government data lies in the extent to which these evidence-based decisions are better than the “best guess” decisions that would have been made without the data. Assessing value in such a manner would require a body of counterfactual evidence that does not currently exist. In the absence of such counterfactual evidence, we focus instead on identifying and describing the wide range of decisions supported by Federal Government data.

Data-Driven Decisions: Improving Downtown Life

The Center City District (CCD) is a Business Improvement District in downtown Philadelphia. This organization was formed to keep Philadelphia’s downtown clean, safe, beautiful, and fun. CCD makes physical improvements to the downtown, such as installing and maintaining lighting, signs, banners, trees and landscape elements (Center City District, 2014). This organization used data from the Census Bureau’s Longitudinal Employer-Household Dynamics (LEHD) program on where people live and work to better define its downtown, thereby improving the focus of its $20 million operating budget. The LEHD program combines data from a wide range of sources, including Unemployment Insurance earnings data from States, data from the Quarterly Census of Employment and Wages, and data from the Decennial and Economic Census surveys, to create statistics on employment, earnings, and job flows at detailed levels of geography and industry and for different demographic groups. The LEHD also generates partially synthetic data on workers’ residential patterns (Gilchrest, 2013).

The pool of users of Government data for local economic development and other local public policy making is huge; U.S. Census data shows that there are 1,249 principal cities in metropolitan and micropolitan statistical areas in the United States; 729 incorporated places with a population of 50,000 or more; and over 3,140 counties. In addition, there are over 40,000 members of the American Planning Association (American Planning Association, 2014). In fact, non-profit groups repackage and tailor Government data specifically to analyze local issues. For example, the California Business Roundtable has set up a nonprofit organization called the Center for Jobs and the Economy that will provide a free online and interactive repository of Federal data from the Census Bureau and the BLS, as well as California state agencies such as the Employment Development Department (Walter, 2013).
As noted above, the National Oceanic and Atmospheric Administration (NOAA) of the U.S. Department of Commerce collects weather and climate data. Research examining that value suggests that the return to society on investment from Government meteorological data is large. For example, one study found that the overwhelming majority of survey respondents reported using weather forecasts and did so at an average rate of 3.8 per day (Lazo, et al., 2009).* This means that, based on Census population estimates, U.S. adults obtained about 301 billion forecasts in the year the study was conducted. The researchers also found a median valuation of weather forecasts per household of $286 per year, which suggests that the aggregate annual valuation of weather forecasts was about $31.5 billion. (Lazo, et al., 2009) The sum of all Federal spending on meteorological operations and research was $3.4 billion in the same year, and the private sector spent an additional $1.7 billion on weather forecasting, for a total of private and public spending of about $5.1 billion. In other words, the valuation people placed on the weather forecasts they consumed was 6.2 times as high as total expenditure on producing forecasts.

NOAA data is re-packaged and analyzed (that is, made suitable and available for mass consumption by private sector businesses) to produce 15 million weather products (Konkel, 2013). Often the entities that use data produced by NOAA’s National Weather Service (NWS), such as The Weather Channel, AccuWeather, and numerous other private sector forecasters, do so without attributing NOAA as the source of the data. As a consequence, many end-users may not realize the value that NOAA’s data provides. Private sector firms continually develop new ways of creating value for companies using data from NOAA and other Government agencies. For example, the Climate Corporation sells weather insurance to farmers and is also a high-tech agricultural consultant. The company processes 50 terabytes of new and historical data each day. One of its principal data sources is the National Weather Service’s Nexrad (Next Generation Radar), a network of 159 Doppler radar stations that scans weather data in two million locations. Another source of data is 60 years of crop-yield statistics from the Department of Agriculture. Together, these data sources enable the company to generate moisture, precipitation, and soil condition maps down to the level of a farmer’s field. Monsanto bought the company in November 2013 for nearly one billion dollars (Spector 2013).

The private sector adds value to only two of the 19 terabytes of new weather and climate data that the agency produces every day; the rest currently is not accessible to the public. NOAA is working to make the data more accessible. NOAA is developing public-private partnerships in which private companies host NOAA data on the cloud and make it readily available to the public for advanced analytics without having to download all the data (Konkel 2013).

*Incidentally, the researchers used the Census Bureau’s American Community Survey as a benchmark to determine that the respondents were generally representative of the U.S. population – an example of using Government data for benchmarking purposes to help validate statistical results from other surveys.
Data-Driven Decisions: Screening for Cancer

The Centers for Disease Control and Prevention (CDC) spent $150 million in 2013 on cancer screening services and outreach, with the potential for saving lives through early detection in tens of thousands of women. CDC’s National Breast and Cervical Cancer Early Detection Program, as well as State-based programs, need to determine where to target outreach and screening services based on the percentage of women eligible for early detection services. In order to achieve this objective, CDC helps sponsor the Census Bureau’s Small Area Health Insurance Estimates (SAHIE). These are model-based estimates that combine American Community Survey data along with data from the Census Bureau’s demographic population estimates, Census 2010, County Business Patterns, the Internal Revenue Service’s Federal tax information, the Food and Nutrition Service’s Nutrition Assistance Program, and the Centers for Medicare and Medicaid Services’ (CMS) Children’s Health Insurance Program and Medicaid. In addition to informing decisions on where to target cancer screening efforts, SAHIE provides yearly estimates of health insurance coverage for every county in the United States. SAHIE’s annual operating cost of $2 million thus helps CDC and other agencies to make better decisions on spending much larger sums of money on preventative health services than would be the case absent such data.

From the beginning of our Nation’s history, there was recognition of the need for data in order to make informed decisions. Collecting data was an essential activity envisioned by the country’s founding fathers, and the taking of a census at regular intervals for the apportionment of Representatives in Congress was written into the U.S. Constitution in 1787. The first Decennial Census was taken in 1790; its purpose was to ensure a fair and accurate system of representation and taxation.

Over time, as the country’s economy developed, the need for accurate information grew beyond a simple population count. Collection of additional information built on the foundation set by the early censuses. In the early 1800s, Congress directed that a census of manufactures be conducted in order to provide a “general view of the manufactures of the United States.” This census was followed by a census of agriculture and mineral industries (Duncan and Shelton, 1978). By the late 1800s, the Federal Government was publishing statistics on foreign trade, government expenditures and the internal commerce in the country. As the U.S. economy and society grew and changed in the 20th Century, the need for statistical information increased still further.

Today, data from Federal statistical agencies informs a multitude of decisions that help government improve its operations. One important function of Government data is to help direct the funds to programs and areas where they are most needed, such as the National School Lunch program. For example, in 2012, 28 Federal Government programs used regional income and product estimates from the Bureau of Economic Analysis to distribute $320 billion; 27 Federal programs used Local Area Unemployment Statistics from the Bureau of Labor Statistics to distribute $115 billion, and the Census Bureau’s American Community Survey was used to help determine how over $400 billion in Federal and State funds were distributed.

Government data also helps Government agencies to ensure that the benefits of their regulations justify their costs. For example, in 2013, Federal agencies evaluated the impact of over 100 "economically significant" regulations, each with a potential impact of at least $100 million or otherwise materially affecting the economy (Office of Information and Regulatory Affairs 2014). Without statistical and administrative data on the number of firms affected and the revenues of those firms, for example, estimates of the benefits and costs of these regulations would be more uncertain. As a result, there would be a greater risk that costly regulations would be enacted, while beneficial regulations would not.
Government statistics, most notably the Principal Federal Economic Indicators, are carefully watched by decision makers across the globe in order to track the macro-economy. These decision makers include Government officials at the highest levels, who use the data to gauge the overall economic health of the Nation, prepare the Federal budget, and project tax revenue; the data also informs monetary and trade policies.

Box 4: The Principal Federal Economic Indicators

The Principal Federal Economic Indicators are statistical series designated by the White House Office of Management and Budget that provide timely measures of economic activity. These indicators are published by several statistical agencies and are listed below. Because these series can have important impacts on market decisions and government policies, they are evaluated every three years, including an analysis of their accuracy and performance and an assessment of the effects of revisions.

**DEPARTMENT OF AGRICULTURE, WORLD AGRICULTURAL OUTLOOK BOARD**
- World Agricultural Supply and Demand Estimates

**DEPARTMENT OF AGRICULTURE, NATIONAL AGRICULTURAL STATISTICS SERVICES**
- Agricultural Prices
- Crop Production
- Grain Stocks
- Cattle on Feed
- Hogs and Pigs
- Plantings

**DEPARTMENT OF AGRICULTURE, FOREIGN AGRICULTURAL SERVICE**
- World Agricultural Production

**DEPARTMENT OF COMMERCE, BUREAU OF THE CENSUS**
- Construction Put in Place
- New Residential Construction
- New Residential Sales
- Monthly Wholesale Trade
- Advance Monthly Sales for Retail and Food Services
- U.S. International Trade in Goods and Services
- Manufacturing and Trade: Inventories and Sales
- Manufacturers’ Shipments, Inventories, and Orders
- Advance Report on Durable Goods -- Manufacturers’ Shipments, Inventories, and Orders
- Quarterly Financial Report -- Manufacturing, Mining, and Wholesale Trade
- Quarterly Financial Report -- Retail Trade
In 2013, the value of construction put in place for highways and streets averaged $81 billion – and the decisions leading to that construction can be informed by Census and other Government data (U.S. Census Bureau, 2013b). For example, the San Diego Association of Governments (SANDAG) adopted its 2050 Regional Transportation Plan (RTP) in 2011, with planning recommendations based on models that used data from the American Community Survey, such as population, housing, household structure, active-duty military personnel and their dependents, and school enrollment, to forecast neighborhood populations and to predict future travel patterns (National Research Council, 2013, p. 24-28). Additional data was drawn from Census Transportation Planning Products (CTPP), a Census Bureau program widely used by transportation planners. Many of the capital projects outlined in the transportation plan are currently in development.
It is noteworthy that decision makers often rely on data originating from a variety of Government agencies, and even many statistical agencies rely on the data products of other agencies. For example, the Bureau of Economic Analysis (BEA) estimates of national accounts, including gross domestic product, depend critically on data from the Economic Census produced by the Census Bureau. Similarly, the Census Bureau’s model-based Small Area Income and Poverty Estimates (SAIPE) rely on data from various Census programs, as well as data from the Internal Revenue Service’s Federal tax information, the Social Security Administration’s Supplemental Security Income, the BEA’s Regional Economic Income Statistics, and the Food and Nutrition Service’s Supplemental Nutrition Assistance Program. The SAIPE program serves as the only source for yearly estimates of school-aged children in poverty for all States, counties, and school districts.  

There are many other examples of cross-agency use of data. The Administrative Office of the U.S. Courts uses Decennial Census data on voting age, sex, race, and Hispanic origin to assist district courts in determining whether their jury pools comply with the Jury Selection and Service Act. The Social Security Administration relies on data from the Consumer Price Index published by the Bureau of Labor Statistics to adjust benefit payments for inflation.

The Census Bureau’s Equal Employment Opportunity Tabulation is the primary benchmark for comparing the race, ethnicity, and sex composition of an organization’s workforce within a specified geography and job category. The Equal Employment Opportunity Commission, the Department of Justice, the Department of Labor, and the Office of Personnel Management use the data to monitor workforce decisions of private sector businesses, Federal contractors, Federal Government agencies, and state and local government agencies.

The government decisions supported by various sources of Government data result in widely diffuse benefits that are difficult to quantify but that nonetheless illustrate the value of the data. Throughout this report, several sidebars provide additional examples of decisions made by individuals, businesses, other organizations, and governments that are informed by Government data. These examples further highlight the value of Government data.

III. The Role of Government in Providing Data

Making decisions under uncertainty is a fact of life. For example, individuals and households must continually decide on their levels of consumption (including leisure) and savings. Periodically they must make major life decisions such as choosing a career, when and where to search for a job, getting married, starting a family, making major purchases such as homes and cars, investing in education and training, and planning for retirement. Businesses have to make investments and operational decisions about such questions as what to produce, how much to produce, where to produce it, how to compensate employees, and how to protect themselves from price risks. Governments have to decide on the types and quantities of services to provide and where to provide them. Common to all these problems is a need for information to guide the decision maker in choosing the best possible action from among the many possible actions. From an economist’s perspective the best possible action is the one that provides the highest overall benefits relative to its costs.

Better information allows people, businesses, other organizations, and governments to make better decisions. For example, detailed data about a neighborhood’s characteristics (such as average residence square footage and average number of household members) can help a retailer better understand its consumers and therefore make better decisions.
about the types of merchandise to carry for sale in that neighborhood. If such data were not available, the retailer would have to either have to find a way to collect the data – a costly investment – or make do with inferior data that might lead to poorer merchandising decisions. In this section, we discuss how Government provision of data provides decision makers with more and better information than would otherwise be the case.

Data Driven Decisions: Customizing Retail Product Selections

Retail companies go to great efforts to understand their markets and tailor the products they offer. Data about local communities often weighs heavily in analyses of communities.

For example, Target Corporation, with more than 1,700 stores in 49 states and employing more than 300,000 persons, uses American Community Survey (ACS) data to tailor its merchandise offerings for customers in its neighborhoods. Compared to suburban residents, persons living in urban areas live in smaller spaces, have fewer members in their households, and are more likely to use public transportation. Based on this information, Target stocks its urban stores with smaller furniture, more folding chairs, and smaller package sizes than its suburban stores. Moreover, Target uses ACS data to identify which urban neighborhoods have attracted young professionals, so it carries apparel in more youthful fashions in those stores than in its stores in areas with older population, such as some of its suburban stores (U.S. Census Bureau, 2012).

Addressing Market Failures to Provide More Information

Government data is an information asset built up from cumulative spending on data collection, processing and analysis, with characteristics that distinguish it from products that are more efficiently produced by markets. Information products—including Government data—have special characteristics that raise the likelihood that the market will fail to produce information at a socially optimal level. Below we discuss these characteristics in more detail.

Products that can be produced efficiently by the market share five features.  

1. They are “rivalrous,” that is, only one economic agent can consume a given unit of such products. (If Jack eats a slice of bread, Jane cannot eat his slice—she has to pay to get her own slice of bread if she wants one.)

2. They are easily excludable, such that the producer has property rights to compel a consumer to pay for the product. (Thus, Jack cannot legally eat a slice of bread without paying baker Tom.)

3. They have characteristics and benefits that are “open” in the sense that they are easy to recognize or understand; generally, consumers know what they want and whether it is worth the price at which it is being sold. (Jack and Jill understand what they are buying when purchasing a loaf of bread.)

4. They generally have constant returns to scale or possibly decreasing returns to scale, at least above some level of output. (It takes twice the amount of flour, yeast, water, and other ingredients to make a slice of bread for Jack and for Jill than just for Jack alone.)
5. The production and consumption of such products generally do not impose gains or losses on others; that is, they generate no externalities. (All the costs of baking the bread, and revenues from selling it, accrue to baker Tom; and the expenditure to buy the bread, and the utility gained by eating it, are Jack’s alone.)

In contrast, information generally does not share these five characteristics.

1. The consumption of information is “non-rivalrous.” One person’s use of the information does not diminish another’s use of it.

2. Although information can be made to be excludable (for example, by encrypting it and obtaining and enforcing intellectual property rights to it), it is often costly, or against public policy, to do so. These two characteristics of information—non-rivalry and costly excludability—suggest that information has characteristics of a “public good.” Standard economic theory shows that markets generally provide public goods at less than the socially optimal level; as a result, government intervention could help the economy achieve this socially optimal level.

3. Information, as distinguished from goods that can be efficiently produced by markets, often is opaque; that is, its quality can be difficult to ascertain before acquiring the data. Information is often the product of complex data collection and processing, making it difficult and costly for consumers to assess its quality and reliability. This opacity can lead to lower than optimal production and consumption of information.

4. Information likely has increasing returns to scale at all relevant levels of production, because it has high fixed costs of production (costly collection and processing) and negligible variable costs of production (low costs of dissemination). Moreover, information production can have economies of scope (that is, the average cost of production declines as additional product types are produced), because some of the infrastructure of the production process for one type of information product may also be used to produce another information product. Such increasing returns to scale and scope give rise to natural monopoly conditions, with the result that the supply of information, if left to a monopoly market supplier, would be less than the socially optimal level.

5. The production and consumption of information can generate external benefits. For example, as more people, businesses, and governments use information to make better decisions, the information becomes more valuable. The more that information is used, the more likely it is that new and innovative ways of using the data that improve decision making would be added to the store of general knowledge about information and its beneficial uses. Moreover, users of information may find problems with the information and provide useful feedback to the information producer, or may suggest additional types of information that may enhance the quality and utility of existing data products. In addition, as noted below, setting standards for data, such as definitions of key concepts, makes all data that employs such standards potentially more valuable. In this way, there are network externalities in the production and use of information. Given these externalities, private consumers and providers do not consider the full set of benefits when deciding how much information to exchange. This means that less than the socially optimal amount of information would be produced.

Together, these five characteristics of information—non-rivalry, costly excludability, opacity, economies of scale and scope, and the presence of positive externalities—imply the market would likely produce less than the socially optimal level of information. As a result, there is a role for government intervention to help the economy achieve a more optimal level of information.
Data-Driven Decisions: Helping Businesses Export

In 2007, a coffee producer in Texas received customized market research from the U.S. Commercial Service office in San Antonio, which helped the company break into markets in Mexico, Italy, Iraq, Australia, and Japan. Market research produced by the Commercial Service includes Government data from Trade Stats Express (http://tse.export.gov/TSE/TSEhome.aspx), and even more detailed data is available at the Census Bureau’s USA Trade Online (https://usatrade.census.gov/).

The share of the company’s annual income from exports increased from 2% in 2005 to 60% in 2009, and the company added jobs as a consequence. Based on the company’s success and knowledge gained from exporting, the owner started an export management company to help other companies (Export.gov, 2014).

The Advantages of Government Production of Statistics and Other Data

The previous section laid out the reasons that government intervention may lead to a more socially optimal level of information. This section focuses on statistical information in particular and explains the advantages of having the Government produce certain types of statistics and data. The Federal Government is uniquely positioned to provide such data because it provides a system of collection and dissemination that is comprehensive, consistent, confidential, credible, relevant, and accessible.

Federal Government data includes information collected by Federal agencies from individuals, households, businesses, and institutions such as schools, hospitals, and State, local, and tribal governments to serve a wide variety of purposes. The output of Federal statistical programs is one subset of Government data. The statistical data is meant to illuminate public and private decisions on a range of topics, including the economy, the population, the environment, agriculture, crime, education, energy, health, science, and transportation (OMB, 2013a). For example, the Principal Federal Economic Indicators can move markets and shape policy. Government officials at the highest levels use them to gauge the overall economic health of the Nation, prepare the Federal budget, and project tax revenue. They inform fiscal, monetary and trade policies, exchange rates, and social welfare policies, as well as investment and financial decisions of businesses and individuals. Given their importance, there is a compelling need for timely, credible information that is accurate, representative, and comparable.

Data-Driven Decisions: Choosing a Career

About 3.4 million students were expected to graduate from high school in 2013 (National Center for Education Statistics 2013). In addition, nearly 900,000 associate’s degrees and 1.7 million bachelor’s degrees were expected to be conferred in 2013 (National Center for Education Statistics 2014). These graduates, as well as adults considering a new job or a career change, need to make informed decisions about what career to pursue.

For more than 50 years, the Bureau of Labor Statistics’ Occupational Outlook Handbook (OOH) has been a nationally recognized source of career information. It describes a variety of occupations, including what workers do...
on the job, working conditions, the training and education needed, earnings, and projected job growth. The handbook is available at http://www.bls.gov/ooh/home.htm.


This information helps our youth throughout the country. For example, a college website advising prospective students on how to get ready for college suggests that 11th graders use the OOH to explore careers and their earning potential (Spoon River College, 2014). A high school honors student trying to decide on a career path used the OOH to help him decide to major in mechanical engineering. After graduation, he obtained a job at a large tech company (Loftus, 2008).

1. Comprehensive

The mission of the Federal statistical agencies is to collect and disseminate nationally representative data that covers the entire population and the whole economy. To ensure accurate representativeness, these agencies collect data from across the U.S. population. Typically this includes population segments—such as persons from sparsely populated regions, or new small businesses—about which private sector entities have little interest. Any single business has little incentive to undertake the significant expense of collecting data from hard to reach populations for their business needs. However, the Government—because of its broader mission—has developed a robust infrastructure to provide information about all people and businesses throughout the Nation.

• The Economic Census provides comprehensive information about the Nation’s commercial activity. Taken at 5-year intervals, it provides comparable data across economic sectors, using consistent time periods, concepts, definitions, classifications and reporting units. Larger firms report information directly, while administrative records are used to provide basic information for very small firms, reducing the need to burden them with questionnaires. In 2012, over 4 million businesses were contacted for information. The Economic Census provides information on over 29 million establishments, large and small, and together with the Census of Governments and the Department of Agriculture’s Census of Agriculture covers more than 98 percent of all economic activity.

• The Population Census, conducted every 10 years, provides an official count of the entire population. Extensive efforts are directed to evaluating the coverage of the Census.

• The American Community Survey (ACS), conducted every year, uses statistical samples of the population to provide more current and detailed information about local communities. In 2012, the ACS covered about 93 percent of the U.S. population. 12

• Other programs, such as the Quarterly Census on Employment and Wages and the Census of Fatal Occupational Injuries, both programs of the Bureau of Labor Statistics, collect and disseminate data that comprehensively covers the Nation.

When the Government collects data on a sample of persons or businesses (in order to avoid higher-cost collections involving the entire population) rigorous statistical methods are used to choose the sample that is representative of the
relevant population as a whole, and the reliability of the estimates is specified through measures of sampling variability (Office of Management and Budget, 2006).

Private data collection—while perhaps produced more quickly—is often narrowly tailored to the specific needs of a business or industry. For example a private business might find it cost-effective to target specific “market segments” in its data collection efforts, such as persons with landline telephone access, or households above a certain income level. These choices make sense for that business’s specific needs, but the resulting data may be of limited use in determining whether the evidence has broader applicability.

The Government is also uniquely positioned to use administrative data to ensure coverage of the entire population while minimizing burden to respondents. Some data is already collected on nearly all employers and their employees for other purposes such as tax filings and unemployment insurance records. Data sharing arrangements enable Federal statistical agencies to use this information to construct survey frames and samples and to avoid redundancies in data collection, thereby minimizing respondent burden. Such arrangements are strictly governed by law and policy to protect the confidentiality of respondents, as discussed in the subsection on confidentiality below.

**Data-Driven Decisions: Locating Hospitals**

In 2013, $41 billion was spent in the U.S. on health care construction (U.S. Census Bureau, 2013b), and Government data can be used to help improve the decisions regarding the location of this construction. For example, a multi-organizational committee led by Michigan’s Department of Community Health was set up in 2004 to revise its procedures for approving community hospitals. To help in this effort, it used data from the 2000 Decennial Census to identify areas with inadequate access to existing hospitals (Messina, et al., 2006).

### 2. Consistent

Federal statistical agencies collect and disseminate data at regular intervals, and they work to ensure the data is consistent across data collection programs, enabling meaningful comparison and analysis. Government data collection is often broadly focused with a long time horizon, rather than targeted at specific products or specific interests in a short-time horizon based on market demand. Government data collection efforts also focus on consistent definitions and measurements that make it easier to compare data across data sets. As a result, Government data collection facilitates comparisons with fewer discontinuities.

The long-term focus of Government statistical agencies facilitates the comparability of the data over time, allowing for the assessment of changes and trends. Although methodological changes and other factors can limit the strict comparability of some Government data series, Government responsibility for data collection and dissemination better ensures that the information collection activities will be ongoing and thus that the data will be consistent and comparable.

Government data collection efforts also facilitate comparisons of data across data sets. The Office of Management and Budget sets Government-wide definitions and classifications standards for concepts such as “race, “ethnicity,”
"Metropolitan Statistical Areas," "industry," and "occupation." Without generally accepted definitions and standards, these concepts could be defined and measured in many ways, making it difficult to compare data from different sources.

In addition, the private sector might not find it profitable to provide or retain certain data for which there is not current adequate demand. To the extent that some of the value of data is in future demands that are not fully anticipated, as is often the case for basic scientific research, systematic collection (and archiving) of the data by the Government helps ensure that this future value of data is reserved. For example, data from the Department of Agriculture about U.S. horse exports to Mexico was relatively unimportant for many years, but after U.S. horse-slaughtering policies changed in 2007, the time series of the data became important in understanding the effects of those policies (C-Fare, 2013, p. 7).

Data-Driven Decisions: Re-aligning a School District

There were 13,629 public school districts in the United States in the 2009-2010 academic year (National Center for Education Statistics, 2013). The value of public educational construction put in place in 2013 was $63 billion (U.S. Census Bureau, 2013b). Government data can help improve decisions about public school district re-alignments and school construction. For example, the Fontana Unified School District in California experienced rapid population growth that required new schools to be built. In addition, changes in the distribution of population throughout the district created a need to review feeder patterns and re-align boundaries for all school levels. A consulting firm helped the district create new boundaries using analytic tools that rely on a wide range of Government data (DecisionInsite, 2013a), including data from the Census Bureau, BEA, the BLS, and other Government agencies (DecisionInsite, 2013b).

3. Confidential

Government statistical agencies recognize that the data they collect can be sensitive, and it is a top priority to maintain the confidentiality of the information. Many respondents would not provide accurate information if they believed that it would be made public or used to their detriment. Confidentiality is also the law: for example, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) ensures that individuals and organizations who supply information for statistical purposes under a pledge of confidentiality will have their identities protected and that the information will be used for statistical purposes only and may not be used for any enforcement or other regulatory purposes. Violators are subject to fines and incarceration.

Private-sector data collectors provide no similar guarantees; for example, a company that provides an Internet-based “do-it-yourself” survey application states that the anonymity of survey respondents’ information depends on the survey creator. "It is up to each survey creator to decide if they want to collect responses anonymously, or to capture respondents’ personal information." Similarly, social networking platforms are free to change privacy policies governing the use of data.

The Bureau of Economic Analysis (BEA) collection of data on services trade and on financial and operational information about investments abroad is a useful example of the value of the Government’s confidentiality protections. Because private firms have an incentive to safeguard information about internal operations and business plans, publicly reported information about their investment abroad and cross-border trade in services may be very limited or imprecise.

https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review
Updated: Thu, 25 Jul 2019 08:57:25 GMT
Powered by mindtouch
This affects the usefulness of that information. BEA collects robust trade and financial data from private firms, yet publishes statistics based on the data without revealing anything that might violate individual firms’ confidentiality. This increases the odds that firms fully and accurately report the information.

Data-Driven Decisions: Crime Mapping

Over 10 million crimes were committed in the United States in 2012, with property crimes representing the vast majority, about 9 million (Federal Bureau of Investigation, 2013). Crime mapping has emerged as a critical tool for law enforcement, pointing to relationships between various types of crime incidents and the characteristics of the neighborhood where the incidents occurred, such as population density, poverty, and housing vacancy. This helps to identify “hot spots” and focus law enforcement efforts where they are most needed. Crime mapping often relies on Census Bureau’s demographic and housing data, which is available at the neighborhood level (Reamer, 2006).

For example, the Illinois State Police department has used Census data to develop a list of risk factors to help identify methamphetamine hot spots. Of the eight risk factors that were ultimately developed, six of them came directly from Census data (Police Foundation, 2000).

4. Credible

Trust in the quality of Government statistics by data users is paramount if those statistics are to be valued. Statistical agencies foster credibility and trust in the veracity of their products by maintaining transparency and impartiality. When data is not transparent— for example if sources and methodology are unclear or undisclosed—questions arise about the quality and accuracy of the data.

Government agencies strive to be transparent by providing full documentation of the survey forms used to collect the data and by publishing the statistical methodologies used to compile the data. Changes to statistical questionnaires require public notice and comment, which allows the public to provide the Government with important information that should be considered in the collection and dissemination of data. Agencies hold data user conferences and host online training in the use of their data in order to foster a better understanding of the strengths and limits of their data. By maintaining this high level of transparency, statistical agencies help the public understand where the statistics come from and what they represent, increasing the level of trust in the reliability of the data.

Credibility is also fostered by impartiality. The Office of Management and Budget (OMB) coordinates the Federal statistical system (44 U.S.C. 3504(e)) and issues policy directives that, among other things, aim to insulate agencies from pressures to manipulate data for political purposes or other special interest objectives. For example, OMB’s Statistical Policy Directives Number 3 and 4 distinguish between policy-neutral data releases and interpretation of data by policy officials. 17
Data-Driven Decisions: Setting Contract Prices

The Producer Price Index (PPI) program at the Bureau of Labor Statistics (BLS) publishes monthly data on prices received by domestic producers of goods and services, allowing for the measurement of price changes from the seller’s perspective. Businesses can use the data to adjust prices in long-term sales and purchase contracts, allowing the parties to cope with changes in prices without otherwise revising their long-term business arrangements. These escalation clauses allow the parties to allocate risks in a mutually beneficial way over longer time periods, thereby reducing contract negotiation costs. The data can have a large financial impact. According to a survey of Producer Price Index data users, the average contract that was escalated was valued at $115 million (Bureau of Labor Statistics, 2013).

5. Relevant

To produce relevant data, statistical agencies must balance the need to provide comprehensive data consistently over time against the need to be flexible in response to changes in the economy, society, and environment, as well as to changes in information needs. The rise of new technologies, industries, occupations, and household arrangements, shifts in policy priorities, and other societal changes may require changes in the kinds of data statistical agencies collect, the ways they collect them, the products they produce with them, and the vehicles they use to disseminate those products. Because of the strictures involved in changing Federal statistical agencies’ data collection methods, private sector organizations, including academic and non-profit researchers, arguably have greater flexibility in collecting and classifying new kinds of data. To ensure they are collecting relevant data, statistical agencies frequently consult with private-sector entities in the research and development of new data products.

Statistical agencies respond to changes in data needs by collecting new data or re-classifying existing data only after careful research and consideration. For example, U.S. statistical agencies work with agencies from other countries to develop new data categories for the National Income and Product Accounts and to conform to international definitions, methodologies, and standards (United Nations Statistics Division, 2014). Proposed changes in data collection and methodologies are carefully researched and reviewed by statistical agency staff, are vetted with statistical agency advisory committees and other statistical agencies that use the data, are posted in the Federal Register for comment by the public, and are approved by the Office of Management and Budget.

Collection of data about household arrangements is a prime example of how statistical agencies alter data collection in response to societal changes. The classification of such household relationship data has evolved since 1880, when the Decennial Census questionnaire asked about the relationship of each person in the household to the householder (Simmons and O’Connell, 2003). In 1990, the category “unmarried partner” was added to the Census to reflect the growing diversity of household arrangements and an increasing tendency for couples to live together before, or instead of, marrying. This category was added to other surveys over the next several years. Given continued changes in household and family relationships, there may be continued need to update or revise categories used in these surveys. OMB established an Interagency Work Group on Measuring Relationships in Federal Household Surveys to research the measurement of marital and household relationships and to make recommendations to statistical agencies for the development of questions that more accurately capture data on marriage and family arrangements (OMB, 2013, p. 63).
Another way statistical agencies can improve the relevance of their data is by re-organizing or reclassifying their data in ways that shed light on different sectors of the economy that were previously not clear. The Bureau of Economic Analysis (BEA) has developed “satellite accounts” in conjunction with its National Income and Product Accounts (NIPA) program as a means of responding to changing information needs. To provide better information about particular sectors of the economy, such as Trade and Tourism, BEA rearranges data elements underlying the official NIPA (Bureau of Economic Analysis, 2014a). In addition, BEA has also developed satellite accounts based on concepts that differ from the official NIPA concepts and serve as a framework for estimating vital sectors of the economy that were otherwise unmeasured (United Nations Statistics Division, 2009, p. 523-543). For example, until recently, the NIPAs treated spending on research and development (R&D) as an expense, but BEA developed a satellite account that treated R&D as a capital investment, which is more consistent with the economic concept of research and development; in 2013 the R&D satellite account became a part of the official NIPAs (Bureau of Economic Analysis, 2013). BEA has also been developing a satellite account for healthcare spending, which will reclassify health care spending as expenditures on treatment of various disease groups. The goal is to develop statistics that will provide perspective on overall health care spending as well as presenting a more detailed view of household consumption of care for diseases over time, thus helping researchers, businesses, and policy makers better assess the returns to society on medical-care spending. (Bureau of Economic Analysis, 2014b).

Data-Driven Decisions: Serving the Church Community

Houses of worship of various religions and denominations use information systems based on Census Bureau data to plan expansion initiatives, identify neighborhoods where a new location might be successful, and provide services to local communities based on who lives there. For example, a house of worship observing an increase in young families and a rising birthrate in the local area might decide to upgrade its nursery facilities and develop plans for summer camp and daycare activities (Leavell Center for Evangelism and Church Health, 2014). Often, houses of worship access this information through a third-party intermediary; for example, MissionInsite is an information and query tool that provides demographic information specifically tailored to churches (MissionInsite, 2014).

Box 5: Measuring E-Commerce: Government Data Sets the Standard After Building on Initial Private Sector Efforts

Until the late 1990s, only private sector organizations published estimates of the volume of sales transactions completed over a computer network, known as e-commerce. However, these private estimates ranged widely. For example, in 1998, private sector estimates of business-to-business e-commerce sales ranged from $12 billion to $92 billion, and estimates of business-to-consumer estimates ranged from $4 billion to $27 billion (Fraumeni 2001, p. 319).

The large differences in estimates likely reflected lack of uniformity in methodologies, coverage, and definitions, but such background information about the estimates was difficult to obtain or, if available, consisted of opaque descriptions on “sizing the Internet market” (Fraumeni 2001, p. 318).
During the summer of 1999, the Census Bureau developed detailed concepts and definitions of e-commerce in collaboration with other Government agencies and private sector experts (Parker and Grove 2000). The Bureau distinguished among three main components of e-business:

- e-business processes such as buying, selling, inventory and production management, logistics, and communication and support services.
- e-business infrastructure, including computers and other hardware; telecommunications and network channels; and system and applications software.
- e-commerce, which consists of transactions completed over a computerized network involving the transfer of ownership or right to use of goods and services where the key determinant of when a transaction is “completed” is electronic agreement over the network and not payment.

The Census Bureau added e-commerce sales questions to its 1998 annual survey covering retail trade. The (revised) estimate for e-commerce retail sales in 1998 is about $5 billion (Census Bureau 2013a)—a little above the lowest private sector estimate for that year.

6. Accessible

Public access to Government data maximizes the potential beneficial uses of the data. In particular, providing access to data free of charge over the Internet helps ensure that data is used for more applications with positive benefits (net of other costs) and can be used for a variety of uses for which the benefits are uncertain or diffuse. Equity in access to Government data ensures that all users have an equal opportunity to use the information.

Government data is becoming increasingly “open.”¹⁸ Government data tabulations used to be disseminated exclusively in printed documents and made available for purchase or in libraries by the Government Printing Office; since the 1990s most Government statistical reports can be downloaded at no charge, while the data tabulations contained in those reports (and additional data) can be downloaded in electronic format. More recently, Government data has increasingly become available via Application Programming Interfaces (APIs) that allow programmers to access Government data directly and feed the data into applications. President Obama signed an Executive Order on May 9, 2013, which promotes open data as the default mode for Federal Government information, and Office of Management and Budget issued a memorandum to establish a framework for institutionalizing the requirements of the Executive Order. Making machine-readable data available through APIs is one way statistical agencies are making large amounts of data available on a timely basis. Other efforts are underway to increase the amount of Government data that is open, including efforts at the National Oceanic and Atmospheric Administration, the National Aeronautics and Space Administration, the Department of the Interior, the Internal Revenue Service, and other agencies. (U.S. Open Data Action Plan, 2014).

The Government does not necessarily have to make data more accessible on its own; the Government also supports private sector entities through grants, contracts, and public-private partnerships to make data more accessible and easier to use for the research community and the public at large. For example, the Minnesota Population Center of the University of Minnesota, under a grant from the National Institute of Child Health and Human Development, collects and
preserves Current Population Survey microdata and documentation, harmonizes the data for compatibility over time, and disseminates the data at no charge (Minnesota Population Center, 2014). The Patent and Trademark Office, in conjunction with Google, made patent awards and applications more accessible and easier to search (McKinsey Global Institute, 2013, p. 12).

The Complementary Relationship between Government and Private Sector Data

Government data is a vital source of information, but not the only important source. Government data and private data are complementary sources of information for decision making. In fact, even as private data production, aggregation, and dissemination have proliferated, much private data crucially depends on public data. For example, many private entities that create statistics, economic indicators, and other data products rely on Government data as a reference and standard for adjusting, weighting or testing the validity of their products. These entities include private sector entities that provide macroeconomic indicators that complement or attempt to forecast Government economic indicators; pollsters; and market research and consulting firms that collect and disseminate public opinion and marketing data.

For example, the Billion Prices Project is a private (academic) initiative to collect price data daily from hundreds of online retailers around the world in order to track inflation and prices. The data are commercially available through a private company, PriceStats, which uses the Consumer Price Index (CPI) as a benchmark to gauge the performance of the index. However, the PriceStats data is not intended to replace Federal Government data on prices as measured by the CPI, but instead provides an independent and complementary gauge of price changes (Lowrey, 2010).

In other instances, the role of the Government is to develop a "simple, reliable, and publicly accessible infrastructure that ‘exposes’ the underlying data" (Robinson, et al., 2009). The Government does so in a manner that is comprehensive, consistent, confidential, credible, relevant, and accessible, while the private sector is in the best position to develop custom-tailored analyses of Government data for specific users and specific purposes. That is, Government data can be used by the private sector as an input into more specialized products that the private sector is most suited to supply. Private sector entities, which are in a more direct relationship with customers, may be in a better position to perceive new or changing demands in the market. Private sector entities are freer to combine Government data from multiple Federal, State, local, overseas, and multinational agencies, as well as the private sector, and may be less restricted in presenting the data. The private sector has the flexibility to charge fees or sell advertising for data dissemination and therefore has a strong incentive to develop innovative ways to deliver Government data to the public. Moreover, the private sector is free to analyze and interpret Government data in any way it sees fit, subject to the discipline of the marketplace and peer-review. The flexibilities enjoyed by the private sector result in the ability to bring valuable products and services to the market.

IV. The Commercial Value of Government Data

A primary goal of this report is to provide an initial framework for understanding the value of Government data. Government data is an input into the products and services of many types of businesses, so an important component of the value of Government data lies in its use as an input into these private sector products and services. Because Government data is disseminated freely and with no requirements for attribution, it is difficult to identify these uses of Government data and to estimate their monetary value. Unlike goods and services that are produced and sold in the marketplace, there is no market price for Government data, so it is not possible to use its price to reflect its value as a
commercial input. Instead, this section examines the revenues attributed to firms that use Government data to illustrate the importance of Government data in the economy.

As a first step toward assessing the commercial value of Government data, we focus on defining the ecosystem of firms that rely heavily on Government data as a key input to their production processes. This section: (a) develops a definition of Government data-intensive business activities; (b) identifies some of the firms that engage in Government data-intensive business activities and estimates their size in terms of annual revenues; and (c) outlines a methodology for defining the Government data-intensive sector in terms of those industries that most closely resemble Government data-intensive activities. The goal is to develop credible, albeit rough, estimates of the size of the sector.

**Defining Government Data-Intensive Business Activities**

This section provides a first step toward assessing the value of Government data as a commercial input by looking at the ecosystem of firms that rely heavily on Government data in their production processes. These firms constitute the Government data-intensive sector (GDIS). Government data-intensive businesses provide products and services to households, businesses, other organizations, and governments.

Figure 3 lays out where GDIS activities fall in a data value chain, with Government data flowing from Government data agencies (at the bottom of the diagram) to the ultimate users of the data (at the top of the diagram). While data can flow directly from the Government agencies to these users, data can also flow through the GDIS, which uses the data as an input for products and services purchased by the ultimate end users. (Note that in Figure 3 there are data flow arrows pointing in both directions, as Government both provides public data to users and collects raw data from them.) Below, GDIS activities are described in order from those furthest upstream in the data value chain to those furthest downstream.

**Figure 3: Government Data Value Chain**

- **Value-Added Repackagers** aggregate Government data from many different Federal Government agencies, private-sector firms, other governments, and international organizations within a single application. The application adds value to the data by providing easy-to-use data interfaces and possibly the capability to create charts, tables,
and maps, and by allowing users to switch periodicity easily (for example, switch between monthly, quarterly, and annual series). For-profit examples of repackagers include Geolytics, Haver Analytics, and Socrata; non-profit examples include the Federal Reserve Economic Data (FRED) at the Federal Reserve Bank of St. Louis.

- **BENCHMARKERS** are firms that create statistics or other data products using non-Government data, but use Government data as a reference and standard with which to adjust, weight, or test the validity of their products. These firms include private providers of macroeconomic indicators that complement or attempt to forecast or substitute for Government economic indicators. Certain pollsters and market research firms that collect public opinion and marketing data and disseminate such information to private companies, marketing consulting firms, and news organizations also benchmark their survey results using Government data. Examples include Nielsen and Gallup.

- **ANALYSTS** are firms that use Government data (either directly or indirectly through repackagers) to create products for other firms or to generate research based on the data. For example, analysts create market segmentation products, in which households are classified into numerous categories sharing similar tastes, incomes, household configurations, and interests. Market segmentation is a tool for more accurately and efficiently targeting resources devoted to media advertising buys, pricing strategies, and establishment locations. In addition, social science research organizations perform work under contracts from governments, foundations, and businesses and are often very intensive users of Government data. Yet another example is news organizations that report on the principal Government economic indicators and analyze their impact on the economy or that analyze other Government data, including highly localized demographic and economic information. Examples of news organizations that report Government data include CNBC, The New York Times Company, and the News Corporation.

- **DATA BROKERS** are firms that compile Government data (and sometimes proprietary private sector data) from around the world and add value by aggregating and integrating the data (McKinsey Global Institute, 2013, p. 6). These firms sell access to the resulting repackaged data and often provide analysis of the data and even news reporting on the data. Examples include Acxiom, Bloomberg, ESRI, Experian, IHS Global Insight, and Thomson Reuters.

**Estimating the Size of the Government Data-Intensive Sector**

As previously noted, estimating the value of Government data to GDIS firms is challenging. An ideal estimate of the commercial value of Government data to GDIS firms might be an estimate of how much lower the market value of GDIS firms would be absent Government data. However, it is not possible to get a counterfactual estimate of what the value of GDIS firms would be if there were no Government data available to them. Furthermore, many of the large GDIS firms are privately held, so there is no available information about their current market valuation. Given these constraints, this report measures the size of the GDIS sector as a proxy for the commercial value of Government data. This subsection outlines a methodology for developing credible, albeit rough, estimates of the size and scope of the GDIS and reports such estimates. First, we identify specific companies that use Government data as an input into the products and services they sell. We describe the business activities of these companies and classify each company in one or more industry based on their business activities. The lower-bound estimate of the size of the GDIS is based on the revenues of the identified firms. The upper-bound estimate of the size of the GDIS is based on the revenues attributable to the industries identified by the assigned NAICS codes.

**Data-Driven Decisions: Stabilizing Rents**

There are over one million rent-stabilized apartments in New York City, regulated under the Local Emergency Housing Rent Control Act of 1962, the Local Rent Stabilization Law of 1969, and the Emergency Tenant Protection Act of 1974. These State and city laws require that data be collected every three years to determine whether rent controls should continue. The purpose of the New York City Housing and Vacancy Survey (NYCHVS) is to provide such data. The data includes the rental vacancy rate, the supply of housing accommodations, and the condition of
such accommodations. Since 1965, the City has retained the U.S. Census Bureau to design and carry out the NYCHVS in order to fulfill the data collection requirement. The latest NYCHVS survey in 2014 is estimated to cost approximately $7.8 million, which is a small investment relative to the billions of dollars of foregone market rents at stake.

Data-Driven Decisions: Locating a Business Establishment

According to the Census Bureau’s Statistics of U.S. Businesses, about 650,000 new establishments opened in 2009-2010, employing 4.9 million workers. Businesses use Government data to make informed decisions about where best to locate new establishments. For example, PETCO, a large pet supplies retailer, decided to implement a geographic information systems (GIS) site selection tool. The goal was to mitigate the risk associated with opening new locations and to maximize the return on invested capital in new stores. The tool, ESRI Business Analyst, which uses demographic data from the 2010 Decennial Census and the Census Bureau’s American Community Survey, also helped the company analyze geographic decisions related to merchandise planning and advertising. The company found that using the site selection tool reduced its real estate risk (Pillar, 2010).

Data-Driven Decisions: Investing in Real Estate

Existing-home sales in 2013 were 5.1 million on a seasonally adjusted annual rate according to the National Association of Realtors, and real estate investors accounted for roughly 20 percent of home purchases near the end of 2013, according to the Campbell/Inside Mortgage Finance survey of real estate conditions (Bharatwaj, 2013). Both investors and individual home buyers can use Census data to help make better residential real estate investment decisions. For example, a real estate investor used Government data from the Census Bureau’s “State & County QuickFacts” website (http://quickfacts.census.gov/qfd/index.html) to help to determine whether houses were under- or overpriced in a given county. By his rule of thumb, houses were underpriced in that county, so he bought houses that he later sold at a higher price. He recommends that prospective real estate investors consider Census data such as household income, population growth, unemployment rate, job growth, housing inventory, rental rates, and median home price in making investment decisions (Pierce, 2009).

To begin, a partial list of companies known to use Government data to some degree was compiled based on consultations with statistical agency staff and lists of competitors cited in the 10-K filings or annual reports of publicly-traded companies. The firms highlighted in this section are selected for illustrative purposes, to provide a general sense of the business of firms operating in the GDIS, and no endorsement of their products and services is intended.

Below we briefly describe the business activities of these firms, along with the industry or industries in which each firm operates. We classify each firm in one or more industries based on the description of the business activities. Appendix II lists and describes the industries (as defined by their six-digit North American Industrial Classification System (NAICS) codes).
Acxiom

Acxiom, headquartered in Little Rock, Arkansas, is a large data broker. One of its services is a market segmentation tool called PersonicX. Clients use PersonicX to refine products to meet the needs and desires of particular groups of potential customers; to formulate marketing strategies; and to improve targeting of advertising (Acxiom, 2014). PersonicX classifies households into 70 groups with similar “needs, motivations, likes and dislikes and behaviors” within 21 “life stage groups” (households that have attained similar life milestones, such as getting married, having a child, or buying a home).

PersonicX draws on its analysis of Acxiom’s database of demographic, contact, and purchasing behavior information for 125 million U.S. households. It gathers data from “various public information sources: driver’s license files, birth records, self-reported surveys, county tax assessors, telephone books and other public sources of data” such as Census data (Acxiom, 2010; Ayres, 2007, pp. 145-6).

**NAICS Industries: Marketing Consulting Services (NAICS 541613)**

Bloomberg, L.P.

Bloomberg, L.P. is a data broker that is the world’s biggest provider of financial information. The company’s main product (accounting for 85 percent of the company’s revenue) is its data terminal, which the company leases to its customers. The terminal provides 15,000 functions over the Internet, including Government data series as well as securities market data and other news plus analytical and computational capabilities (Elkind 2013). Another product published on the Internet, Bloomberg Government, provides many Federal Government documents and data (for example, the Federal Register, reports published by the Government Accountability Office, and data on Federal Government contracts and grants) and adds value by providing sophisticated search tools. Bloomberg Government also provides detailed profiles of Congressional districts using Census and other data.

**NAICS Industries: News Syndicates (NAICS 519110); Internet Publishing and Broadcasting and Web Search (NAICS 519130)**

Environmental Systems Research Institute (ESRI)

Environmental Systems Research Institute (ESRI) provides software tools and data that enable customers to produce maps by combining geographic information and also to help solve problems that have a spatial element to them. (For example, we describe in a sidebar elsewhere in this report an example of a business using ESRI’s Business Analyst tool to decide where to locate.) The base maps that ESRI makes available to map-makers using ESRI’s flagship software product, ArcGIS, were developed using geographic data from numerous sources, including Federal Government agencies. For example, ESRI’s World Topographic Map draws on geographic data from the U.S. Geological Survey, the Environmental Protection Agency, and the U.S. National Park Service (Environmental Systems Research Institute, 2014a). ESRI also packages, adds value, and sells economic and demographic data by geographic location from the 2010 Decennial Census and the American Community Survey that can be combined with base maps to generate maps that reflect the geographic distribution of the data (Environmental Systems Research Institute, 2014b). The same Census data is used by ESRI in the development of its Tapestry market segmentation tool (Environmental Systems Research Institute, 2014c); Tapestry categories are matched by ESRI to data from the Consumer Expenditure Survey.
conducted by the Bureau of Labor Statistics to analyze consumption patterns by product class and market segment (Environmental Systems Research Institute, 2014d).

**NAICS Industries: Internet Publishing and Broadcasting and Web Search Portals (519130); Marketing Consulting Services; Marketing Research and Public Opinion Polling (541910); and Administrative Management and General Management Consulting Services (541611)**

**Gallup**

Gallup collects data on the economic status, attitudes, wellbeing, and behaviors of individuals in more than 160 countries (Gallup, 2013a; Gallup, 2013b). In the United States, Gallup uses random-digit-dialing of landline and mobile phone numbers to generate samples of individuals to survey. After it collects and processes its survey data, Gallup assigns weights to each respondent so that the demographic profile of the weighted sample matches the demographic profile of the entire adult population. The weights are based on Census Bureau data from the American Community Survey, the Current Population Survey, and the Decennial Census on gender, race, age, educational attainment, and region. (Gallup 2010; Gallup, 2013c). Gallup publishes many of its findings on its website, and many news organizations also report its findings. Gallup Analytics makes detailed poll findings available to subscribers for an annual fee; for an additional fee subscribers can gain access to the respondent data (Gallup, 2013b).

**NAICS Industries: Marketing Research and Public Opinion Polling (NAICS 541910)**

**GeoLytics, Inc.**

GeoLytics, Inc. repackages Decennial Census, American Community Survey data from the Census Bureau and Uniform Crime Reporting Program Data from the U.S. Dept. of Justice, Federal Bureau of Investigation and adds value by providing software to make the data easier to access, analyze, and map (GeoLytics 2014a and 2014b). It produces products such as Business Demographics 2009 “for making sound business decisions” that combines Decennial Census population data with data from the Economic Census and County and Zip Code Business Patterns data as well as data from the Consumer Expenditure Survey of the Bureau of Labor Statistics and includes analytic and mapping software and GeoLytics population projections (GeoLytics 2014c). GeoLytics also sells market research data, including a market segmentation product, based on Census demographic data (GeoLytics 2014d).

**NAICS Industries: Internet Publishing and Broadcasting and Web Search Portals (519130); Marketing Consulting Services; Marketing Research and Public Opinion Polling (541910); and Administrative Management and General Management Consulting Services (541611)**

**GreatSchools**

GreatSchools is a national non-profit organization with offices in San Francisco, Milwaukee, Washington D.C., and Indianapolis. GreatSchools ratings are based on data drawn from State Departments of Education on the most recent available standardized test results for schools in each State. For each grade/subject combination, GreatSchools determines ratings based on which decile the school's test scores are in the statewide distribution of test results. For each grade the subject-level ratings are averaged to arrive at an overall rating for that grade. Similarly, schoolwide and
district-wide ratings are weighted (by the appropriate school population) averages of gradelevel and school-level ratings, respectively.

**NAICS Industries: Internet Publishing and Broadcasting and Web Search (NAICS 519130)**

**Haver Analytics**

Haver Analytics updates and manages historical time series data, providing up-to-the-minute time series data on over 200 economic and financial databases to customers in the global strategy and research community. The company maintains offices in New York, London, and Singapore and serves clients in financial services, government, academia and various industry groups from consulting to manufacturing. Data offerings range from daily market data to annual economic statistics, with extensive forecast data covering the world’s economies. Source data includes over 1,200 government and private sources. Haver databases are available through specialized software that allows for database integration and analysis, and Haver also offers web delivery of data. Telephone client support is staffed by database managers who can help clients locate and understand the data available to meet their information needs (Haver, 2014).

**NAICS Industries: Administrative Management and General Management Consulting Services (NAICS 541611)**

**IHS Inc.**

IHS Inc. is a publicly traded business information services company based in Englewood, Colorado, with experts in energy, economics, geopolitical risk, sustainability and supply chain management. The company was originally founded in 1959 as a provider of product catalog databases on microfilm for aerospace engineers, and it now employs over 8,000 people in more than 31 countries. IHS provides business clients with global market and economic information, including technical information, tools and operational and advisory services. IHS Data & Analytics services provide guidance in the areas of economics and finance, risk managements, country risk and credit risk, investment and portfolio analysis and management, corporate strategy, marketing and market intelligence, business development, and insurance underwriting. Customers range from governments and large multinational corporations to small companies and technical professionals (Ycharts, 2014).

**NAICS Industries: Marketing Consulting Services (NAICS 541613); Administrative Management and General Management Consulting Services (541611)**

**MapQuest, Inc.**

MapQuest, Inc. offers online, mobile, business and developer solutions that help people discover and explore where they would like to go, how to get there and what to do along the way and at the destination. MapQuest's suite of mobile services extends the company’s popular place search, mapping and directions services beyond the desktop to cell phones and other wireless devices. MapQuest Mobile offers free solutions for the mobile web as well as free voice-guided navigation applications to meet the needs of MapQuest users on-the-go. Recently, to provide “a more robust local discovery experience,” MapQuest has included additional information from a variety of sources, including publicly available Government data. For example, it includes airport status information by the Federal Aviation Administration; Mapquest reports that airports are the top queried search (AOL, 2014)
**NAICS Industries: Internet Publishing and Broadcasting and Web Search (NAICS 519130)**

**Nielsen**

Nielsen is a global marketing and advertising research company headquartered in Lower Manhattan, New York City. It offers critical media and marketing information, analytics, and industry expertise about what consumers buy and what consumers watch. Customers include media, entertainment, and consumer goods companies. Nielsen is active in over 100 countries, and employs approximately 32,000 people worldwide. Total revenues amounted to $5.1 billion in 2010. (Ycharts, 2014).

The company relies on Government data to adjust the results of its information collection. For example, Nielsen calculates the number of homes with televisions and the number of television viewers in the United States. To do so, Nielsen uses proprietary data from its People Meter panel, along with U.S. Census Bureau data and auxiliary sources such as state governments and the U.S. postal service. Nielsen applies TV penetration rates to convert the total household and population estimates to TV households and persons therein (Nielsen, 2013).

**NAICS Industries: Marketing Research and Public Opinion Polling (NAICS 541910)**

**Thomson Reuters**

Thomson Reuters provides “intelligent information” for businesses and professionals by combining industry expertise to deliver critical information to decision makers in the financial and risk, legal, tax and accounting, intellectual property and science and media markets. The company’s Financial & Risk Division provides solutions to the global financial community, delivering news, information and analytics to facilitate transactions relationship among trading, investing, financial and corporate professionals.

**NAICS Industries: News Syndicates (NAICS 519110), Administrative Management and General Management Consulting Services (NAICS 541611)**

**Trulia**

Trulia is an on-line search engine with a database of 3.5 million homes (Trulia, 2014a). It delivers information to homebuyers, sellers, owners, and renters about properties (including mapping homes for sale, rent, and recently sold for a chosen locality), agents and neighborhoods around the country (Trulia, 2014b). Neighborhood information includes data on schools (from GreatSchools – see description above). Crime hotspots are mapped using data provided by SpotCrime.com and CrimeReports.com. Trulia reports community statistics based on 2000 Decennial Census data for a locality with comparison data for its county and State. These statistics include median household income and the distributions of home ages, household income, and commuting times. The Trulia website also provides numerous links to other sources of data on a particular community, including Bureau of Economic Analysis data on countylevel per capita personal income and its components and Census Zip Code Business Patterns by NAICS industry code (Trulia, 2014c).

**NAICS Industries: Other Activities Related to Real Estate (NAICS 531390)**
Zillow

Zillow is an online real estate listing service headquartered in Seattle with offices in New York City, San Francisco, Chicago, Irvine, CA and Lincoln, NB.

Zillow’s website draws on a continually updated database of 110 million U.S. homes currently for sale or rent or not on the market and maps them for a chosen locality. The data on listed homes is drawn from public data on physical attributes, tax assessments, and prior and current transactions, which can also be updated by home sellers (Zillow, 2014a). Zillow also offers “Zestimates” for home values and rents using a proprietary automated valuation model based on Zillow’s database (Zillow, 2014b. Zillow also provides local and (for comparison) national demographics data from the 2000 Decennial Census, including data on relationship status, homes with children, age distribution, median household income, percent population that consists of single males and single females, median age, average household size, and average commute time compared with national numbers. In addition, using segmentation methods based on 2000 Decennial Census data, Zillow reports on a locality’s “main types of people.” Zillow also maps school catchment areas and reports ratings of local public schools using proprietary data from GreatSchools (Zillow, 2014d); see description of GreatSchools above.

NAICS Industries: Other Activities Related to Real Estate (NAICS 531390)

1. Lower-Bound Estimate of Size of the GDIS Based on Size of Selected Firms Active in the GDIS

Table 3 lists the companies identified as relatively intensive users of Government data, along with their revenues. We use the revenues of these firms as the basis of the lower bound estimate of GDIS revenues. The sources of revenues estimates for these firms were Gale Business Insights and OneSource by InfoGroup. Although this list is far from complete, it is useful to the extent that it identifies revenues attributable to some businesses that create and sell products using Government data, and thus it represents a lower-bound estimate of the value of Government data as a commercial input.

It is important to note that, even if Table 3 were a complete list of GDIS firms, their total revenues can only be a rough estimate of the value of Government data. The wide variation in the intensity with which companies use Government data to produce other goods and services presents another difficult challenge to estimating the value of the Government data they use as an input. For a handful of value-added repackagers, the bulk of their activities may be so closely associated with the use of Government data that it is almost incontrovertible that most of their revenues are a reflection of the value of Government data. For other firms, the picture is less clear—Government data is an input into their production processes (either directly or indirectly via re-packagers and brokers), but the extent to which that is the case is unclear. Moreover, there are government-supported and non-profit entities, such as the National Bureau of Economic Research and the Minnesota Population Center, that provide similar services for free. Nonetheless, the revenues of firms that are relatively intensive users of Government data reflect, at least to some extent, the commercial value of the Government data they use. While future work may provide more refined estimates, these initial estimates at least illustrate the important point that the Nation’s investment in Government data is small relative to the revenues of the firms that rely on the data to produce goods and services for economic investors.
Table 3: Revenues and Employment of Firms in the Government Data-Intensive Sector

<table>
<thead>
<tr>
<th>Company</th>
<th>Listed/Unlisted/ Subsidiary</th>
<th>Revenues</th>
<th>Employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acxiom</td>
<td>L</td>
<td>$954,464,000</td>
<td>6,300</td>
</tr>
<tr>
<td>Bloomberg</td>
<td>U</td>
<td>7,560,000,000</td>
<td>15,000</td>
</tr>
<tr>
<td>ESRI</td>
<td>U</td>
<td>1,030,470,000</td>
<td>2,900</td>
</tr>
<tr>
<td>Gallup</td>
<td>U</td>
<td>121,185,000</td>
<td>2,700</td>
</tr>
<tr>
<td>Geolytics</td>
<td>U</td>
<td>4,400,000</td>
<td>10</td>
</tr>
<tr>
<td>GreatSchools</td>
<td>U</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Haver Analytics</td>
<td>U</td>
<td>11,000,000</td>
<td>51</td>
</tr>
<tr>
<td>IHS Global</td>
<td>L</td>
<td>1,530,000,000</td>
<td>6,000</td>
</tr>
<tr>
<td>MapQuest</td>
<td>S</td>
<td>45,210,000</td>
<td>126</td>
</tr>
<tr>
<td>Nielsen, Co.</td>
<td>L</td>
<td>5,407,000,000</td>
<td>40,000</td>
</tr>
<tr>
<td>Thomson Reuters Financial &amp; Risk Division</td>
<td>L (Canada)</td>
<td>7,193,000,000</td>
<td>20,700</td>
</tr>
<tr>
<td>Trulia</td>
<td>L</td>
<td>68,085,000</td>
<td>519</td>
</tr>
<tr>
<td>Zillow</td>
<td>L</td>
<td>203,400,000</td>
<td>560</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>$24,128,214,000</td>
<td>94,866</td>
</tr>
</tbody>
</table>

Notes: Data are for 2012 except for Bloomberg (2011) and Geolytics (2010).

Sources:
- Gale Business Insights: ESRI, Bloomberg, Geolytics
- OneSource (Infogroup): Haver, Mapquest, Gallup, IHS, Trulia, Zillow.
- Annual Reports: Acxiom, Nielsen, Thompson Reuters.
2. Upper-Bound Estimate of the Size of the GDIS Based on the Size of Government-Data-Intensive Industries

Using the industry classification of each company in Table 3, we identify industries to include in the GDIS sector. Table 4 lists those industries (by 6-digit NAICS code) and their size, based on Census Bureau estimates of the revenues of employer firms in these industries.

Table 4: Revenues of Government Data-Intensive Sector (GDIS) by Industry

<table>
<thead>
<tr>
<th>NAICS Code</th>
<th>Industry Name</th>
<th>2012 Revenues ($billions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>519110</td>
<td>News Syndicates</td>
<td>$2.3</td>
</tr>
<tr>
<td>519130</td>
<td>Internet Publishing and Broadcasting and Web Search</td>
<td>57.5</td>
</tr>
<tr>
<td>531390</td>
<td>Other Activities Related to Real Estate</td>
<td>12.0</td>
</tr>
<tr>
<td>54161</td>
<td>Management Consulting Services</td>
<td>131.7</td>
</tr>
<tr>
<td>541910</td>
<td>Marketing Research and Public Opinion Polling</td>
<td>17.3</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>$220.8</td>
</tr>
</tbody>
</table>

Source of revenues data: U.S. Census Bureau. 2013d. Revenues are for employer firms only.

It is important to note that there is no bright line that separates the NAICS codes that compose the GDIS from the rest of the economy. An alternative approach to defining the GDIS has been taken by the Open Data 500 project, which is “the first comprehensive study of U.S. companies that use open Government data to generate new business and develop new products and services” and conducted by the GovLab at New York University. Open Data 500 has compiled a list of 500 U.S. firms and organizations that use Government data based on the Open Data 500’s outreach, expert advice, and research (Open Data 500, 2014). In many cases, the 17 company categories used by Open Data 500 also do not neatly map to a well-defined list of NAICS codes. Moreover, in both our GDIS classification system and that of Open Data 500, it is difficult to determine precisely the intensity with which certain firms or industries use Government data. Undoubtedly there are some firms in the GDIS that do not use Government data, while there are others that do use Government data at least to some degree but are classified in industries that are excluded from our definition of the GDIS. Other attempts to estimate the size of other data-intensive sectors such as the economic impact of the “geo-services sector” (Oxera 2013) or the size of the “data-driven marketing economy” (Deighton and Johnson 2013) also encountered challenges in parsing these sectors from the rest of the economy.

Nevertheless, even given the caveats to the estimates in Tables 3 and 4, they provide a rough order of magnitude of the size of the Government data-intensive sector. The lower-bound estimate, based on a very short and incomplete list of firms that rely on Government data, suggests that Government data helps private sector firms generate revenues of at
least $24 billion annually. The upper-bound estimate suggests that GDIS industries generate revenues of $221 billion. While there is clearly a large range of the estimate of the commercial value of Government data as measured in terms of the GDIS, the important conclusion is that Government data supports an important sector of the economy that itself supports myriad business investment decisions across the globe.

V. Conclusion

Historically, data collection and dissemination have been the purview of Government. Information, including statistics about the Nation’s economy and people, has been considered a public good. Therefore, providing such information was seen as an appropriate function of Government, helping to ensure that the optimal quantity and quality of data would be produced.

Focusing on Federal statistical data in particular, this report outlines evidence to suggest that Government data provides important information that could not or would not be produced otherwise. Given the market failures that are likely to occur in the private production of information, Government data production results in more information than would otherwise be available. Moreover, the high quality of Federal statistical data—as summarized by its comprehensiveness, consistency, confidentiality, credibility, relevance, and accessibility—results in better information than would otherwise be available. This information supports decision makers throughout the Nation, including individuals, businesses, other institutions, and governments at every level.

Government data and private data are complementary. The high quality of Government data enhances the value of private sector data by providing the benchmark that lends context and validity to private sector data. In addition, the private sector uses Government data to create new products and services in a wide range of industries. While its commercial value represents only one aspect of the value of Government data, it nonetheless provides a sense of the importance of Government data in the economy. This report estimates the commercial value of Government data using a rough estimate of the size of the Government data-intensive sector (GDIS)—the private industries that rely on Government data as a critical input to the goods and services they produce. It highlights specific examples of how firms in GDIS industries create value using Government data, including Federal statistics.

The value of the Nation’s investment in Government data lies in the usefulness of the data in improving the countless decisions it supports, either directly or indirectly through the products and services of the private industries that rely on Government data. Estimating this value would require counter-factual evidence about the decisions that would be made without Government data. Such evidence is currently not available but represents a promising area of research.

Footnotes

F1

Administrative data is data collected and used by program, administrative, and regulatory agencies and not originally intended to be used for statistical purposes. See Appendix I for definitions of other terms used in this report.
Another component of the cost of statistical data is the opportunity cost of the time spent by persons filling out survey forms. We do not address that cost in this paper, although we note that The Paperwork Reduction Act (44 U.S.C. §§ 3501-3521) is intended to minimize the burden on respondents to Federal surveys in order to “ensure the greatest possible public benefit from and maximize the utility of information created, collected, maintained, used, shared and disseminated by or for the Federal Government.”

In contrast, expenditures on research and development are treated as investments rather than expenditures in the National Income and Product Accounts.

Article I, Section 2, Clause 3 of the U.S. Constitution states, “Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers . . . . The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct....” This Clause was amended in 1868 by Section 2 of the 14th Amendment to provide that apportionment shall be based on “counting the whole number of persons in each State, excluding Indians not taxed.”

The SAIPE program, with a cost of under $4 million, results in data used to allocate $15.4 billion (in 2013) to school districts and States, assisting more than 16 million school-age children in poverty.


Because information is technically excludable, it may be more accurately characterized as a “club good” rather than a pure “public good.” Nevertheless, when exclusion costs are so high that it is not feasible to exclude non-paying consumers, as may be the case with “headline” data and even more detailed data distributed in electronic format, or when exclusion is contrary to public policy, then technical excludability is less relevant, and the case for public provision of the good is stronger (Cornes and Sandler, 1996, p. 400).
For example, the used car market is arguably less efficient than the new car market because of information asymmetries that make it difficult for a prospective purchaser to ascertain the quality of a used car until it is actually purchased (Akerlof 1970).

Interspersed throughout this report are numerous examples of publicly available descriptions of ways that Government data can be used to improve decision making. These examples, and many more not cited in this report, add to the general store of knowledge about how to use Government data in innovative ways to improve decision making.

For example, Government data agencies “have reported that intensified ties with re-users [of data] may lead to improved data quality and process efficiency since any deficiencies in the data are promptly flagged up and reported back to the” agency. “Hence, when the interest in data quality is shared, quality control is partly outsourced” (POPSIS 2011 p. 6). In addition, there are organizations such as advisory committees established by statistical agencies to obtain advice about better ways to use and improve their data. For example, the Census Bureau and BEA each have advisory committees. The Council of Professional Associations on Federal Statistics (COPAFS) “represents over 300,000 individual researchers, educators, public health professionals, civic groups, and businesses that rely on the quality and accessibility of statistics that can only be effectively collected by the Federal Government” (http://www.copafs.com). The Federal Committee on Statistical Methodology (FCSM) “is sponsored by the Office of Management and Budget to improve the quality of Federal statistics and the efficiency and effectiveness of statistical practice among Federal agencies” (http://www.fcsm.gov). The Association of Public Data Users (APDU) “is a national network that links users, producers, and disseminators of Government statistical data. APDU members share a vital concern about the collection, dissemination, preservation, and interpretation of public data. As an association, APDU is committed to helping data users identify public data that meet their needs, utilizing data in cost-effective and appropriate ways, establishing effective two-way communication between data producers and users, and bringing the perspectives and concerns of public data users to bear on important issues of Government information and statistical policy. APDU provides a venue to share news, raise concerns, and advocate on behalf of the interests of public data users” (www.apdu.com). Finally, vast numbers of researchers publish peer-reviewed studies using Government data that both lead to improvements in the quality and types of data collected as well as the development of new, useful techniques for using the data.

Complete history of ACS coverage rates is available at: https://www.census.gov/acs/www/methodology/index.php#note2.

For examples of privately held data that is no longer available for future use, and the problems this poses for ongoing research, see http://blogs.smithsonianmag.com/science/ow-be-missing/.
44 U.S.C. § 3501 note (2002). See, titles 13 and 22 for confidentiality provisions specific to the Census Bureau and the Bureau of Economic Analysis, respectively.

Section 513 of CIPSEA states that disclosure of confidential information is a class E felony punishable by imprisonment up to five years and fines up to $250,000, or both.

Additional explanation of the private survey company’s privacy and security policy is available at http://help.surveymonkey.com/article...ous-and-secure?


As defined by the McKinsey Global Institute, fully open data is accessible to everyone; are machine readable (in formats that make it easy for computers to retrieve and process data); are offered at zero cost; and have no limits on reuse and redistribution (McKinsey Global Institute 2013, p. 3).

Ultimately the goal is to define the GDIS as a list of industries within the North American Industry Classification System (NAICS). See Appendix II for definitions of NAICS industry codes for Government data-intensive industries.

We also considered the NAICS industry codes that these two databases reported for each company, but we did not find them to be accurate based on the companies’ own descriptions of their activities on their websites.

NAICS code 54161 includes these industries (for which we don’t have separate revenue estimates):

- 541611 Administrative Management and General Management Consulting Services
- 541612 Human Resources Consulting Services
- 541613 Marketing Consulting Services
- 541614 Process, Physical Distribution, and Logistics Consulting Services
- 541618 Other Management Consulting Services
Note that there are no official NAICS codes for GDIS or any set of other data-driven industries for that matter. For example, a Government Accountability Office report on “information resellers,” which are “companies with a primary line of business of collecting, aggregating, and selling personal information to third parties” notes that there is no well-defined NAICS code for such companies (U.S. Government Accountability Office, 2013 pp. 4-5).

The categories of companies used by Open Data 500 are:

- Aerospace and Defense
- Business & Legal Services
- Data/Technology
- Education
- Energy
- Environment and Weather
- Finance & Investment
- Food & Agriculture
- Geospatial/Mapping
- Governance
- Healthcare
- Housing/Real Estate
- Insurance
- Lifestyle & Consumer
- Research & Consulting
- Scientific Research
- Transportation

References


America Online. “MapQuest Overhauls Local Offering with Multiple Data-service Partnerships and Integrations,” April 9, 2014 (http://corp.aol.com/2014/04/09/mapquest...a-service-part/ accessed on June 2, 2014).


Center City District. 2014. “About Us” (http://www.centercityphila.org/about/CCD.php accessed February 27, 2014).


Gallup. 2013c. Telephone Conversation with Stephanie M. Kafka, Consultant and Survey Methodologist (December 6).


Oxera Consulting, Ltd. 2013. What Is the Economic Impact of Geo Services? (Funded by Google.)


U.S. Census Bureau, 2013b. “Table 4: December 2013 Construction at $930.5 Billion Annual Rate”. (http://www.census.gov/construction/c...f/pr201312.pdf)


Appendix I: Definitions of Terms Used

Administrative Data
Information collected and used by program, administrative, and regulatory agencies and not originally intended to be used for statistical purposes.

Analysts
GDIS firms that analyze Government data directly or indirectly, using third-party data products to create products for other firms.

Benchmarkers
GDIS firms that create statistics or other data products using non-Government data, but use Government data as a reference and standard with which to adjust or weight or test the validity of their products.

Data Brokers
GDIS firms that compile Government data (and sometimes proprietary private sector data) from around the world and add value by aggregating, integrating, and facilitating access to and analysis of the data. These firms sell access to the resulting repackaged data and often provide analysis of the data and even news reporting on the data. In other words, these firms may be vertically integrated GDIS firms.

Government Data
A diverse set of Federal Government information and data products, including

- Information published by Government agencies for statistical purposes. Such information is compiled from raw data collected directly from individuals, households, businesses, other organizations, and governments. The information may also be indirectly compiled from Government administrative sources, private sector data sources, and other sources. The data may be summarized in tabulations by statistical agencies to describe specific characteristics of groups of persons, businesses, or other entities. The agencies make such information available for use by other Government agencies and the general public to inform decision making, in program analysis and evaluation, and research.

- Information in Government administrative records on individuals and businesses. Examples include publicly available data about private sector company-sponsored retirement plans reported to the Labor Department on Form 5500; Securities and Exchange Commission filings; Government contracts data; real estate appraisal records; and voter registration rolls. Historically, a Federal Freedom of Information Act request has often been required to make much of this category of information public, but recent “open data” efforts have aimed to make such information more easily accessible. However, much data remains strictly confidential, such as tax return data collected by the Internal Revenue Service or Social Security Administration data.

- Physical measurements by Government agencies of natural phenomena. Examples include weather and climate data such as local temperature readings collected by the National Weather Service of the National Oceanic and Atmospheric Administration Global Positioning System data; and measurements of the characteristics of elements...
and molecules collected and published by the National Institute for Standards and Technology Physical Measurement Laboratory.

**Government data-intensive sector (GDIS)**

The ecosystem of firms that rely heavily on Government data as a key input to their production processes for providing services to households, governments, and firms inside and outside of the GDIS. These firms consist of firms in NAICS industry codes defined in Appendix II.

**North American Industry Classification System (NAICS)**

The standard used by Federal statistical agencies in classifying business establishments for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy. Each industry is assigned a six-digit numeric code; industry sectors at varying levels of aggregation consist of industries sharing the same first two or more digits in their NAICS codes (http://www.census.gov/eos/www/naics/).

**Principal Federal Statistical Agencies**

Thirteen agencies whose main activities are the collection, compilation, processing or analysis of information for statistical purposes. For most, funding appears as a separate line item in the President’s Budget.

**Public Good**

A good or service that is non-rival and non-excludable.

**Value-Added Repackagers**

GDIS firms that aggregate Government data from many different U.S. Government agencies, private-sector firms, other governments, and international organizations within a single application. The application adds value to the data by providing easy-to-use interfaces with the data and possibly the capability for users to create charts and tables and switch periodicity easily (for example, switch between monthly, quarterly, and annual series). Furthermore, there are synergies in using data from multiple sources, which is another way these companies add value. These companies often splice together time series for which the original source had a break in methodology, thus creating long time series that have value even if they do not meet rigorous statistical agency standards.

### Appendix II: Definitions of Government Data-Intensive Industries

<table>
<thead>
<tr>
<th>NAICS Industry Code</th>
<th>Name</th>
<th>Industry Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>519110</td>
<td>News Syndicates</td>
<td>This industry comprises establishments primarily engaged in supplying information, such as news reports, articles, pictures, and features, to the news media.</td>
</tr>
<tr>
<td>519130</td>
<td>Internet Publishing</td>
<td>This industry comprises establishments primarily engaged in</td>
</tr>
</tbody>
</table>
and Broadcasting 1) publishing and/or broadcasting content on the Internet and Web Search exclusively or 2) operating Web sites that use a search engine to generate and maintain extensive databases of Internet addresses and content in an easily searchable format (and known as Web search portals). The publishing and broadcasting establishments in this industry do not provide traditional (non-Internet) versions of the content that they publish or broadcast. They provide textual, audio, and/or video content of general or specific interest on the Internet exclusively. Establishments known as Web search portals often provide additional Internet services, such as e-mail, connections to other web sites, auctions, news, and other limited content, and serve as a home base for Internet users.

<table>
<thead>
<tr>
<th>Code</th>
<th>Industry Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>531390</td>
<td>Other Activities</td>
</tr>
<tr>
<td>541611</td>
<td>Administrative</td>
</tr>
</tbody>
</table>

This industry comprises establishments primarily engaged in related services (except lessors of Real Estate estate, offices of real estate agents and brokers, real estate property managers, and offices of real estate appraisers).

This industry comprises establishments primarily engaged in Management and providing operating advice and assistance to businesses and other General Management organizations on administrative management issues, such as Consulting Services financial planning and budgeting, equity and asset management, records management, office planning, strategic and organizational planning, site selection, new business startup, and business process improvement. This industry also includes establishments of general management consultants that provide a full range of services, such
<p>| 541612  | Human Resources | This industry comprises establishments primarily engaged in Consulting Services providing advice and assistance to businesses and other organizations in one or more of the following areas: (1) human resource and personnel policies, practices, and procedures; (2) employee benefits planning, communication, and administration; (3) compensation systems planning; and (4) wage and salary administration. |
| 541613  | Marketing Consulting | This industry comprises establishments primarily engaged in Services providing operating advice and assistance to businesses and other organizations on marketing issues, such as developing marketing objectives and policies, sales forecasting, new product developing and pricing, licensing and franchise planning, and marketing planning and strategy. |
| 541614  | Process, Physical | This industry comprises establishments primarily engaged in Distribution, and providing operating advice and assistance to businesses and other Logistics Consulting organizations in areas, such as: (1) manufacturing operations Services improvement; (2) productivity improvement; (3) production planning and control; (4) quality assurance and quality control; (5) inventory management; (6) distribution networks; (7) warehouse use, operations, and utilization; (8) transportation and shipment of goods and materials; and (9) materials management and handling. |</p>
<table>
<thead>
<tr>
<th>Code</th>
<th>Industry</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>541618</td>
<td>Other Management</td>
<td>This industry comprises establishments primarily engaged in Consulting Services providing management consulting services (except administrative and general management consulting; human resources consulting; marketing consulting; or process, physical distribution, and logistics consulting). Establishments providing telecommunications or utilities management consulting services are included in this industry.</td>
</tr>
<tr>
<td>541690</td>
<td>Other Scientific and Technical Consulting</td>
<td>This industry comprises establishments primarily engaged in providing advice and assistance to businesses and other Services organizations on scientific and technical issues (except environmental).</td>
</tr>
<tr>
<td>541910</td>
<td>Marketing Research and Public Opinion Polling</td>
<td>This industry comprises establishments primarily engaged in systematically gathering, recording, tabulating, and presenting marketing and public opinion data.</td>
</tr>
</tbody>
</table>
Data Science for the Big Data Review

Recommendation 1: Policy attention should focus more on the actual uses of big data and less on its collection and analysis.

This is "big data" (all of your content) integrated to be "massive data" that can be reused for visualization and analytics.

Special Acknowledgment: PCAST is especially grateful for the rapid and comprehensive assistance provided by an ad hoc group of staff at the National Science Foundation (NSF), Computer and Information Science and Engineering Directorate. This team was led by Fen Zhao and Emily Grumbling, who were enlisted by Suzanne Iacono. Drs. Zhao and Grumbling worked tirelessly to review the technical literature, elicit perspectives and feedback from a range of NSF colleagues, and iterate on descriptions of numerous technologies relevant to big data and privacy and how those technologies were evolving.

MORE TO FOLLOW

Seed Grants, Postdoctoral Fellowships, and MDI Seminar Series: http://mspp.georgetown.edu/massive-data-institute/
MDI Seminar Series: [http://mspp.georgetown.edu/events/MDI_Seminars/](http://mspp.georgetown.edu/events/MDI_Seminars/)

The 2013-2014 seminar series focused on the applications of "novel and high dimensional data" in the social sciences.

See This Wiki: [http://mspp.georgetown.edu/events/bi...eral-agencies/](http://mspp.georgetown.edu/events/bi...eral-agencies/)

My Recommendation to Mark Doms Undersecretary for Economic Affairs, Department of Commerce, at the big data workshop hosted by White House Office of Science and Technology Policy and the Georgetown University McCourt School of Public Policy’s Massive Data Institute, was to have a chief data officer and stress the importance of the government’s statistical data.

The Commerce Secretary announced that at the recent ESRI Conference!

---

**Slides**

---

**Spotfire Dashboard**

---

**Research Notes**

Georgetown University McCourt School of Public Policy’s Massive Data Institute My Note: I Saw the Announcement About This Earlier and Was Interested

**Big data is saving taxpayer dollars.** The Centers for Medicare and Medicaid Services have begun using predictive analytics—a big data technique—to flag likely instances of reimbursement fraud before claims are paid. The Fraud Prevention System helps identify the highest-risk health care providers for waste, fraud, and abuse in real time and has already stopped, prevented, or identified $115 million in fraudulent payments. (My Note: Look at This)

Over the past several days, severe storms have battered Arkansas, Oklahoma, Mississippi and other states. Dozens of people have been killed and entire neighborhoods turned to rubble and debris as tornadoes have touched down across the region. Natural disasters like these present a host of challenges for first responders. How many people are affected, injured, or dead? Where can they find food, shelter, and medical attention? What critical infrastructure might have been damaged?

Drawing on open government data sources, including Census demographics and NOAA weather data, along with their own demographic databases, Esri, a geospatial technology company, has created a [real-time map](https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review) (My Note: Look at This) showing where the twisters have been spotted and how the storm systems are moving. They have also used these data to show how many people live in the affected area, and [summarize potential impacts](https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review) (My Note: Look at This) from the storms. It's a powerful tool for emergency services and communities. And it's driven by big data technology. My Note: Is This True?

See 5 PDF files attached below. My Note: Why These PDF Files for Big Data?
As part of the McCourt School of Public Policy’s founding in October of 2013, we are building a Massive Data Institute devoted to the study of high-dimensional data for public policy. The Institute will use Big Data sets to increase understanding of society and human behavior and thus improve public policy decision-making.

Incredible advances in public policy may be possible by bringing together the scientists who can analyze these data and the policy practitioners in the government and nonprofit sectors that can design and implement policies. While we work to build and staff our Institute, we are pleased to announce the following initiatives:

**Seed Grants**

The Massive Data Institute is pleased to offer seed grants to aid Georgetown faculty and graduate students conducting research at the intersection of big data and public policy. Two types of grants will be available:

- **Mini travel grants:** Up to 5 grants of up to $2,000 will be awarded to faculty or graduate students who wish to attend conferences or establish collaborations with external faculty on projects using big data.
- **Research seed grants:** 4 grants of up to $40,000 will be awarded to early-stage projects with high potential to generate external funding. These projects must make use of large data sets and involve collaboration between data science and public policy. Special consideration will be given to those proposals involving faculty collaborations between computer scientists, mathematicians, statisticians, and bioinformatics faculty, on one hand, and social science and humanities faculty, on the other. Allowable expenses are up to $10,000 in summer salary support per PI as well as funds for a research assistant, and data acquisition or travel expenses.

**To Apply:**

To apply for mini grants, applicants must submit an explanation of the conference they wish to attend and the program’s relevance to their future research, as well as a proposed budget. Those who plan to use the funds for research design or data collection, must submit an explanation of the planned project, a project trajectory, and a budget. The conference or research must occur during the 2014-15 Academic Year and a summary report of no more than two pages must be submitted to the Massive Data Institute by May 31, 2015. Applications must be submitted by July 15, 2014 and applicants will be notified by August 15, 2014.

Applicants for research seed grants must provide a project description of no more than 5 pages which includes: an abstract, project goals and objectives, research methods, potential future external funding sources, and a proposed budget of no more than $40,000. Grants will be awarded for the 2014-15 academic year. At the end of the year, a progress report of no more than three pages must be submitted to the Massive Data Institute. Applications must be submitted by July 15, 2014 and applicants will be notified by August 15, 2014.

Applications should be submitted via email to: lam265@georgetown.edu with the name of the grant in the subject line.
Postdoctoral Fellowships

The McCourt School of Public Policy seeks two (2) postdoctoral fellows with research interests at the intersection of public policy and big data to join the new Massive Data Institute. Fellows may hold a Ph.D. in any discipline but they must conduct policy-relevant research that involves large data sets.

Fellows will be expected to actively engage in research leading to publications in top journals. They will also be expected to be active participants in the Massive Data Seminar Series and to give at least one seminar presentation based on their work while a fellow. Fellows will also teach one course on a big data topic during the spring semester and be expected to contribute to the intellectual life of the school by interacting with students and faculty. Fellows will work on faculty-led teams addressing key public policy and analytic issues involving high-dimensional data.

To Apply:

Please submit the following via Academic Jobs Online, position # 4009:

- CV
- One illustrative publication
- A 1-2 page description of current research and work that will be carried out during the fellowship
- Two confidential letters of recommendation

Applications must be received by June 1, 2014. Please contact Laura Masters, lam265@georgetown.edu with any questions.

MDI Seminar Series

The McCourt School of Public Policy at Georgetown University is pleased to present a Massive Data Institute seminar series. The goal of the series is to establish a cross-disciplinary research institute to further the innovative uses of data for policy and social science. The 2013-2014 seminar series focused on the applications of "novel and high dimensional data" in the social sciences.

Our 2013-2014 series has come to a close. Please visit the MDI Seminar page in the fall to view 2014-2015 events. My Note: See Below

MCCOURT SCHOOL OF PUBLIC POLICY

Old North, Suite 100

37th and O Streets, N.W., Washington D.C. 20057

Phone: (202) 687-5932

Fax: (202) 687-5544
Postdoctoral Fellow in Big Data and Public Policy

The McCourt School of Public Policy seeks two (2) postdoctoral fellows with research interests at the intersection of public policy and big data to join the new Massive Data Institute. Fellows may hold a Ph.D. in any discipline but they must conduct policy-relevant research that involves large data sets.

Fellows will be expected to actively engage in research leading to publications in top journals. They will also be expected to be active participants in the Massive Data Seminar Series and to give at least one seminar presentation based on their work while a fellow. Fellows will also teach one course on a big data topic during the spring semester and be expected to contribute to the intellectual life of the school by interacting with students and faculty. Fellows will work on faculty-led teams addressing key public policy and analytic issues involving high-dimensional data.

Fellowships will be for one academic year from August 1, 2014 to May 30, 2015. Fellows will receive a stipend of $60,000, plus health insurance coverage, and will be provided with office space.

To apply please submit the following:
- CV
- A 1-2 page description of current research and work that will be carried out during the fellowship
- One illustrative publication
- Two confidential letters of recommendation (via this site or by email to Laura Masters at lam265@georgetown.edu)

Applications must be received by June 1, 2014.

Please contact Laura Masters, lam265@georgetown.edu with any questions.

Georgetown University is an Equal Opportunity, Affirmative Action employer fully dedicated to achieving a diverse faculty and staff. All qualified candidates are encouraged to apply and will receive consideration for employment without regard to race, sex, sexual orientation, age, religion, national origin, marital status, veteran status, disability or other categories protected by law.

Massive Data Institute Seminar Series

The McCourt School of Public Policy at Georgetown University is pleased to present a Massive Data Institute seminar series. The goal of the series is to establish a cross-disciplinary research institute to further the innovative uses of data for policy and social science. Specifically, the year-long seminar series will focus on the applications of "novel and high dimensional data" in the social sciences.
2013

**Big Data, Food Purchases and the Impact of Product and Nutrient Taxes**

Matthew Harding, Department of Economics, Stanford University  
October 21, 2013  
3:30 – 5:00 PM  
Old North 205

**Google Tools for Data Analysis**

Hal Varian, Chief Economist, Google  
November 13, 2013  
3:30 – 5:00 PM  
Old North 205

2014

**Matthew Taddy, Booth School, University of Chicago**  
January 27, 2014  
3:30 – 5:00 PM  
Old North 205

**Gary King, Director of the Institute for Quantitative Social Science, Harvard University**  
February 7, 2014  
2:30 - 4:00 PM  
Old North 205

**Justin Grimmer, Department of Political Science, Stanford University**  
March 17, 2014  
3:30 – 5:00 PM  
Old North 205

**David Krakauer, Wisconsin Institute for Discovery, University of Wisconsin**  
May 5, 2014  
3:30 – 5:00 PM  
Old North 205

The Massive Data Institute Seminar Series is free and open to the public, but an RSVP is required.

Questions? Please contact Lauren Mullins at lm973@georgetown.edu.
Improving Government Performance in the Era of Big Data: Opportunities and Challenges for Federal Agencies

Source: http://mspp.georgetown.edu/events/bi...deral-agencies

A big data workshop hosted by White House Office of Science and Technology Policy and the Georgetown University McCourt School of Public Policy’s Massive Data Institute.

June 19, 2014
1:00 pm – 5:00 pm

Georgetown University
Rafik B. Hariri Building
Fisher Colloquium
37th and O Sts., NW
Washington, DC 20007

Advanced registration for this event is required.
Register here. My Note: I Did This

Invitation

PDF

Dear Guest,

The White House Office of Science and Technology Policy and the McCourt School of Public Policy’s Massive Data Institute look forward to hosting you on Thursday, June 19, at Improving Government Performance in the Era of Big Data: Opportunities and Challenges for Federal Agencies. An event agenda is attached. The event will begin promptly at 1:00 pm.

The event will be held in Fisher Colloquium, on the fourth floor of the Rafik B Hariri Building on Georgetown University’s main campus. Directions to Georgetown’s campus can be found here. Please note that parking on campus is extremely limited. Some parking is available in the Southwest Garage, which can be accessed by Georgetown’s Canal Street entrance. We recommend guests take a taxi to the main entrance of Georgetown’s campus located at 37th and O Streets, NW. Please arrive about 15 minutes early to walk to the event once you are on campus. There will be signage to guide guests to the Hariri Building, and you can find a walking map from 37th and O to the Hariri building attached.

Please note this event is at capacity and only registered guests will be admitted. If you are unable to attend, please email Lauren Mullins at lm973@georgetown.edu. The afternoon event will be webcast live at http://www.ustream.tv/GeorgetownLive.

Thank you, and we look forward to seeing you soon.
About the Workshop

On June 19, 2014, the White House Office of Science and Technology Policy and the Georgetown University McCourt School of Public Policy’s Massive Data Institute will co-host an event entitled *Improving Government Performance in the Era of Big Data: Opportunities and Challenges for Federal Agencies*. This event will convene key stakeholders and thought leaders from across government, academia, industry, and civil society at Georgetown University to share their individual views and insights.

This event is part of ongoing efforts by the Obama Administration to engage both subject matter experts and the public in a national discussion about the future of data innovation and policy. Drawing from the recent White House working group report, *"Big Data: Seizing Opportunities, Preserving Values"* (My Note: See Below), this event will focus on the opportunities and challenges posed by Federal agencies’ use of data, best practices for sharing data within and between agencies and other partners, and measures the government may use to ensure the protection of privacy and civil liberties in a big data environment. It will build on other workshops held by OSTP in collaboration with MIT (My Note: I Have This), NYU (My Note: I Looked at This), and UC Berkeley (My Note: I Looked at This). Learn more about these workshops and the big data review from the White House [here](https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review) (My Note: See Below).

This event is open to the public, and will include remarks from Counselor to the President John Podesta and U.S. Chief Technology Officer Todd Park.

About the Massive Data Institute

As part of the McCourt School of Public Policy’s founding in October of 2013, we are building a Massive Data Institute devoted to the study of high-dimensional data for public policy. The Institute will use big data sets to increase understanding of society and human behavior and thus improve public policy decision-making.

Incredible advances in public policy may be possible by bringing together the scientists who can analyze these data and the policy practitioners in the government and nonprofit sectors that can design and implement policies.

Initial Agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:30 - 1:00 p.m.</td>
<td>Check-in</td>
</tr>
<tr>
<td>1:00 - 1:10 p.m.</td>
<td>Welcome</td>
</tr>
<tr>
<td></td>
<td>Edward Montgomery, Dean, McCourt School of Public Policy</td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>1:10 - 1:30 p.m.</td>
<td>Opening Remarks</td>
</tr>
<tr>
<td>1:30 - 2:25 p.m.</td>
<td>Panel One: Open Data and Information Sharing</td>
</tr>
<tr>
<td>2:25 - 3:20 p.m.</td>
<td>Panel Two: Big Data &amp; Privacy</td>
</tr>
<tr>
<td>3:20 - 3:35 p.m.</td>
<td>Break</td>
</tr>
<tr>
<td>3:35 - 4:30 p.m.</td>
<td>Panel Three: Big Data &amp; Discrimination</td>
</tr>
<tr>
<td>4:30 - 5:00 p.m.</td>
<td>Closing Keynote</td>
</tr>
</tbody>
</table>

**Final Agenda**

PDF


1:00-1:10 Welcome Edward Montgomery, Dean, McCourt School of Public Policy, Georgetown University

1:10-1:30 Opening Remarks, Nicole Wong, Deputy Chief Technology Officer, The White House

1:30-2:25 Panel 1: Open Data and Information Sharing:

- Nick Sinai (moderator) Deputy Chief Technology Officer, The White House
- Allison Oelschlaeger Center for Strategic Planning, Centers for Medicare & Medicaid Services
- Rosye Cloud Senior Advisor for Veteran Employment, Department of Veterans Affairs
- Rajive Mathur Director, Office of Online Services, Internal Revenue Service
- Zach Goldstein Acting Chief Information Officer, National Oceanic and Atmospheric Administration
Real Progress in Meeting Agency Performance Goals

Source: http://www.whitehouse.gov/blog/2014/02/13/real-progress-meeting-agency-performance-goals

Posted by Beth Cobert on February 13, 2014 at 10:53 AM EST

Starting in 2009, the Administration established a common-sense approach to improving the performance of government. Following successful evidence-based practices used in the private and public sectors, the Administration began engaging senior Federal leaders in establishing two-year Agency Priority Goals in areas where agencies were focused on accelerated performance improvement. The Administration also established government-wide Cross-Agency Priority Goals in areas benefiting from collaboration across multiple agencies.

Today, the Administration is posting performance results on Performance.gov for the 4th Quarter of 2013, which represents the final quarterly update for the 2012-2013 Agency Priority Goals. The agency reports show significant progress across the government in delivering results and positive impact for the American people.

Below are some highlights:

- As part of the cross agency efforts to support the President’s National Export Initiative, the Department of Commerce, as Chair of the Trade Promotion Coordinating Committee (TPCC), has taken actions to help achieve a record level of exports of $2.3 trillion in 2013, which supported an additional 1.3 million U.S. jobs. For example, in 2013, the Department of Commerce’s International Trade Administration (ITA) met its goal of increasing the annual
number of new markets that current U.S. exporters enter with ITA assistance to 6,100, a 7 percent increase. The Department of Commerce and other TPCC agencies continue to advance the interests of U.S. exporters, especially small and medium sized, in markets beyond the United States.

- The Department of the Treasury estimates that it has saved the American people hundreds of millions of dollars by creating an Agency Priority Goal around increasing electronic transactions with the public to improve service, prevent fraud, and reduce costs. Included in this goal was an effort to modernize the Federal government’s payment and collection systems, which resulted in paper benefit payments dropping from 131 million in 2010 to 39 million in 2013, allowing us to get money to beneficiaries and back into the economy faster than ever. At the same time, electronic collections jumped from 85 percent of total collections in 2010 to 97 percent in 2013, reducing costs to the Federal government.

- After designating the improvement of business loan efficiency as an Agency Priority Goal, the Small Business Administration (SBA) has made considerable progress in making it more efficient for small businesses to get loans, while also reducing cost. The SBA increased the use of paperless processing in their 7(a) loan program (which provides financing for various business uses, such as working capital and real estate) from 72 percent in 2011 to 90 percent in 2013, and from 55 percent to 76 percent in their 504 loan program (which provides financing for real estate and major equipment). The adoption of electronic loan processing also contributed to a 5.6 percent increase in loan volume from 2012 to 2013, growing the number of small businesses assisted.

- The Department of State set an Agency Priority Goal of using its diplomatic mission overseas to increase the number of market-oriented economic and policy activities by 15 percent, helping to expand U.S. exports, create opportunities for U.S. businesses abroad, and increase economic growth and job creation. State uses its more than 200 diplomatic missions to promote U.S. manufactured goods and services, analyze and address foreign trade and investment barriers, and provide counseling on exports to new firms. State has exceeded its goal by 43 percent, achieving a total of 971 “success stories” – instances where an export deal is achieved, a dispute is favorably resolved, or a foreign policy is changed to help U.S. businesses expand opportunities abroad.

- After establishing an Agency Priority Goal focused on preventing Americans at-risk of foreclosure from losing their homes, the Department of Housing and Urban Development (HUD) initiated a number of measures to improve agency operations and help borrowers at the very early stages of delinquency when interventions can prevent serious delinquency. HUD increased the number of households assisted with early intervention by 31 percent between 2010 and 2013. HUD also reduced six month re-default rates from 17 percent in 2011 to 8 percent in 2013 among those who were helped by the agency’s mitigation programs.

Bringing more robust performance management practices to Federal agencies is a priority for this Administration, and we are pleased to see these efforts gaining momentum. In the coming months, the Administration will set new performance strategic plans, establish new two-year Agency Priority Goals, and select new Cross-Agency Priority Goals that span the government.

To learn more about the Administration’s performance improvement efforts, visit Performance.gov, our public portal for tracking and reporting performance progress.

Beth Cobert is the Deputy Director for Management of the Office of Management and Budget

The 90-Day Review for Big Data

Source: http://www.whitehouse.gov/issues/tec...ig-data-review
Introduction

The American people are interacting with technology more than ever — when using a cell phone, shopping online, visiting a doctor who uses electronic records, and in countless other everyday acts. These applications of technology and data are enhancing our lives in countless ways — helping us share experiences with distant loved ones, providing us with information at the click of a button, or aiding a job search. As these technologies become more sophisticated, the ways that we think about and protect individual privacy also evolve.

On January 17, President Obama spoke at the Justice Department about changes in the technology that we use for national security purposes, and what these technologies mean for our privacy broadly. He called on the administration to conduct a broad 90-day review of big data and privacy: how these technologies affect the way we live and the way we work — and how big data is being used by universities, the private sector, and the government.

- Read the Big Data Report
- Big Data Report Fact Sheet
- Read the Announcement Blog Post

Concurrent with this study, the President’s Council of Advisors for Science and Technology conducted a review of the technologies underpinning big data.

Read the study from the President’s Council of Advisors on Science and Technology

See responses to the Request for Information.

What We Heard From You

As part of the administration's 90-day review of big data and privacy, we sought out public input on these issues through a survey on WhiteHouse.gov. In it, we asked people to tell us how concerned they are with various data practices and how much they trust various institutions to keep their data safe and handle it responsibly. During the four weeks of public input through this form, responses were collected from 24,092 individuals. The White House did not include submission fields for name or contact information. Although this survey is not a statistically representative survey of public opinion, we think it provides an insightful glimpse into how people feel about the collection and use of data, and we'd like to share those insights with you here.

Respondents felt most strongly about data use and collection practices, expressing a great deal of concern about all of the practices described in the survey. Even in the area where respondents expressed the least amount of concern, collection of location data, a sizable majority (61%) were "very much concerned." And they demonstrated particularly strong feelings around ensuring that proper transparency and oversight is in place for data practices — more than 80% of respondents were very concerned with each of these areas.

Considerably more nuance was evident in respondents’ views towards different entities who collect and use data. Although majorities claimed to trust Intelligence and Law Enforcement Agencies "not at all," their views towards other government agencies (at both federal and local levels) and commercial businesses were far less negative. Further, majorities were generally trusting of how professional practices and academia handle and use their data.
varying degrees of trust expressed in the entities described here may reflect the underlying tension between the privacy concerns and the benefits of these big data practices. Aggregating health data carries the risk of exposure or misuse, but may also allow for the early detection and containment of an outbreak. Similarly, tracked location data from one's cell phone opens up some frightening possibilities, but could also help find it if it is lost or stolen, or could help rescuers locate victims of an earthquake or tornado.

Finally, we asked respondents to give us their thoughts on which technologies and data uses were most transforming their lives and if they had any general thoughts they wanted to share on the issue of big data and privacy. About two-thirds of respondents provided an answer to each of these questions, though answers to the latter were generally much more verbose — with a median length of 41 words, compared with just 9 for the first question.

The technologies that people found most transformative focused heavily around the internet, access to information, and communication through email, cell phones, and social media. Overall, respondents' general thoughts on the issue reflected the responses to the specific questions about institutions and data practices described above. These answers focused heavily on government use and collection of data, with many voicing concerns about protecting privacy, respecting personal data, and strengthening data security practices. The word clouds below depict the responses to these questions, with words and phrases weighted by the number of responses that mentioned each term.

**Which technologies or uses of data are most transforming your life?**

Are there any other thoughts on this issue that you would like to share?
Taken together, the findings from this survey indicate that respondents were most wary of how intelligence and law enforcement agencies are collecting and using data about them, particularly when they have little insight into these practices. Data privacy is an important issue for the President and his administration, and already he has taken several steps to strengthen the oversight of intelligence agencies’ data practices and increase transparency about how, when, and under what circumstances data is collected and used. These steps have included ending the bulk collection of telephone metadata by the NSA, including public interest advocates in FISA Court hearings, and limiting the scope of data that can be collected.

The 90-day review this survey was a part of is part of the foundation for future policies and actions that will help us stay at the forefront of this rapidly evolving sector.

You can read the full big data report here. My Note: See Below

Big Data Privacy Workshops

White House / UC Berkeley School of Information / Berkeley Center for Law and Technology

- PDF of Counselor John Podesta’s remarks (PDF) My Note: See Below
- Video of the remarks (via UC Berkeley)

Remarks as Delivered by Counselor John Podesta

Source: http://www.whitehouse.gov/sites/defa...big_data_1.pdf (PDF)
Good afternoon. I want to begin by thanking Deirdre for that kind introduction, and for all of her hard work putting together today’s workshop, and for her work and friendship over the years. For decades, Berkeley has been at the forefront of public debates about freedom of speech, freedom of association, and civil rights. So I think it’s very appropriate that our third workshop on big data and privacy is happening here. The discussions we’ve heard today built on and deepened some themes that have come up throughout this review of big data and privacy, particularly at the previous workshops at MIT and NYU. So I also want to thank all of today’s panelists for their thoughtfulness and their expertise. And of course I must thank you all for joining us here today.

The technologies and applications we’ve heard about today and in our previous workshops may be new, but the legal and policy questions these technologies raise are quite old. Big data requires us to ask ourselves: how do we embrace new technologies and the progress they bring to our society while at the same time protecting our fundamental freedoms and values like privacy, fairness and self-determination?

It’s been referenced that I was one of the staff authors of the Electronic Communications Privacy Act. I was working for then a young Senator, Pat Leahy, in 1984. And it might be worth just spending a second talking about how we began the journey to write the Electronic Communications Privacy Act. We actually sent a letter to the Attorney General asking whether the existing wiretap statute, which was passed in 1968—most of you who are lawyers know it as Title 3 of the Omnibus Crime Control and Safe Streets Act—applied to electronic communications. It was a simple, straightforward question: did the wiretap act apply to email and other forms of electronic communications? Months later, we heard back from his deputy, who told us that the answer to our question “was neither clear nor obvious.” So that reply didn’t inspire confidence. So for the next two years, Senator Leahy, along with his House colleagues, Don Edwards, from this area, and Bob Kastenmeier, worked to update the law for modern times. The result, as was noted, was the Electronic Communications Privacy Act, which—with some updates—and I will note with a few more needed—still governs Internet communications today.

It wasn’t that we were just moronic back in 1986. We were faced with a series of precedents about how third parties handled data that were dodgy in terms of trying to provide protection for electronic communication. So this was indeed a negotiated compromise, but I think this really was perhaps a unique law in that we took a problem before anyone knew its full dimensions and really dealt with it in a forward-looking manner. And obviously today we are engaged in the Administration to try and think about how we correct inefficiencies, particularly the one mentioned here today about the difference between 180 days and 181 days of storage, and we’re engaging again with my old boss on this topic.

A few years after that, I hosted the Washington, D.C., organizing meeting for a new group dedicated to a new civil liberties cause. I knew Lee would be here, so I somehow dug out my membership card—member number 33 of the Electronic Frontier Foundation. Lee Tien and the EFF are still fighting to find balance between civil rights, personal freedom, and technological advancement, and, of course, I’ve become “the man.”
But we all know rapid change is the one constant in technology. And in the information age, those changes come at an incredible rate. In 1998, when Google was getting its start, less than five percent of the world was online. Today, three-quarters of people in developed countries and four in 10 people worldwide have Internet access.

As Americans, our willingness to innovate, to experiment, to build new things and try new ideas has built the largest and strongest economy in the world. But from a centralized postal service all the way through to SnapChat, each technological advance has prompted serious debate about its implications for other values that we hold dear, like equality, privacy, and personal freedom.

We’re about two-thirds of the way through our study of big data and privacy. Since President Obama asked me to lead this review back in January, working with the Secretary of Commerce, the Secretary of Energy, the NEC and OSTP, we have met with privacy advocates and technology companies; with academic researchers and advertisers; with law enforcement agencies and with civil rights groups. We are receiving public input through a request for information – that comment process was due to close yesterday, but we are extending it to accept submissions through the end of this week. We also have asked for feedback from the public through a survey on the White House website—you can find that at http://whitehouse.gov/bigdata, and I encourage you to fill it out and pass it on to those in your networks.

This 90-day review is fundamentally a scoping exercise, not a policy development process. After today, I am making one exception. Listening to Ken Bamberger, we are changing our policy to ensure that Google Glasses being used in the White House Gym haven’t been hacked. Really, we hope to raise the relevant questions we must ask now and in the future. We have been casting a wide net as we ask how big data technologies have begun to transform government, commerce, and society, and as we examine the potential implications for our social and civic values.

The President’s Council of Advisors for Science and Technology is working in parallel on a report that takes a deep dive into the technologies themselves, from machine learning and algorithms to sensor technology and new models for data storage.

Both our study and PCAST’s work also aim to look ahead, as much as that’s possible, and to anticipate what today’s trends might mean for our lives, our economy, and our world in decades to come.

Without getting too far over my skis about what will be in a report that we haven’t written yet, I want to share some observations about what we’ve learned during this process.

But first, let me say a few words about the ongoing effort inside the Administration to reform how we collect information to keep our country safe. I should be clear, specific intelligence programs are not a part of this study on big data, but they are central to the debate in which we find ourselves. Last week, the President announced a plan to significantly change the way the Section 215 telephone metadata program works.

As you know, this program has been the subject of a great deal of controversy in recent months, as we as a nation debated our signals intelligence practices. Under the President’s proposal, the government will not collect or hold metadata in bulk; instead, those records will remain with telephone companies, and will be available to intelligence agencies only under court order. We’re working with Congress to put this plan into law, and are continuing a wide range of efforts to improve transparency and oversight of our intelligence programs.
The President asked us to conduct this study because he recognized that, even as the Administration takes steps to review and reform intelligence practices, the intelligence community is only part of the picture when it comes to big data. These technologies touch every sector of the economy, from government to industry to research, and will only continue to grow in importance.

I want to say up front that we recognize big data technologies can create considerable economic value and are spurring tremendous innovation. From the very beginning, the Obama Administration has worked to make more public data available through our Open Data Initiative. Entrepreneurs and developers have used these data to create countless applications and build successful businesses.

We've also known for a long time that the information age requires us to carefully consider how we protect personal data. That's why in 2012, the President released a consumer privacy blueprint and the Consumer Privacy Bill of Rights, which sets out principles that aim to protect user privacy without restricting the innovative potential of the Internet.

So in the midst of what some are calling a “big data revolution,” we are taking this opportunity to consider the landscape and to really interrogate whether our existing policies are prepared for what’s on the horizon technologically.

Throughout this process, we’ve asked three questions: First, what is new and different about big data technology and techniques, compared to traditional data analysis? Second, what questions should we be asking about big data's relationship to privacy and other rights and values? That’s the conversation we’ve been having here today. And, finally, do our existing policies and in particular the Consumer Privacy Bill of Rights provide an adequate policy framework for the era of big data, or are there gaps where new policy or new research needs to be developed?

Of course, we first had to decide on a definition of big data—something that easily could have taken up the entire 90 days of our review. So for the purposes of this study, we've been thinking of big data as data sets that are so large in volume, so diverse in variety, and changing with such velocity that traditional modes of data analysis are insufficient.

I think there are a few technological trends that bear being drawn out.

First, the declining cost of collection, storage, and processing of data, combined with new sources of data like sensors, cameras, and other observational technologies, particularly geospatial technologies, means that we live in a world of near-ubiquitous data collection. And as more companies produce web-enabled appliances, wearable technology, and advanced sensors to monitor everything from health indicators to energy use to my running speed, the “Internet of Things” will add huge amounts of new data to what is commonly collected today. Data analysis is also being conducted at a speed that is increasingly approaching real-time, with a growing potential to have an immediate effect on our environment or decisions being made about our lives.

Second, one of the most powerful things about big data analysis is what is sometimes called data fusion. By combining multiple sources and types of data, big data can lead to some remarkable insights. But it can also lead to the so-called “mosaic effect,” whereby personal information can be derived or inferred from data sets that do not include personal identifiers within.

Third, the information revealed by big data analysis isn’t necessarily perfect. Identifying a pattern doesn’t establish whether that pattern is significant. Correlation still doesn’t necessarily equal causation. So as these technologies
become more prevalent, we will need to be deliberate in developing strong, scientific standards—and strong ethical standards—for judging the results of big data analysis and predictive algorithms, as was raised on the first panel today.

Finally, there are some promising technological means to better protect privacy in a big data world. Encrypting data, perturbing data so it no longer represents real individuals, or giving users more say over how their data are used through personal profiles or controls are among the technological solutions. But none of them are perfect, and technology alone cannot protect privacy absent strong social norms and a responsive policy and legal framework.

I want to share a few examples of what all of that means in the real world.

The United States spends more on health care per capita, with worse outcomes, than any other developed country. As a consequence, health care costs are a major driver of our federal deficit. It’s urgently important that we find ways to bring down costs while also improving health outcomes. Big data is helping show us how.

One study synthesized millions of data samples from monitors in a neonatal intensive care unit to try and pinpoint which babies were likely to contract potentially fatal infections. By analyzing all of the data—not just what doctors noted on their rounds—the project was able to identify factors, like increases in temperature and heart rate, that serve as early warning signs that an infection may be taking root. These early signs of infection are not something even an experienced and attentive doctor would catch through traditional practices.

In another instance, by crunching huge amounts of data from electronic medical records, researchers determined characteristics that increase or decrease the likelihood that a patient will be readmitted to the hospital after receiving treatment. This makes it possible for doctors to proactively schedule follow-up care for the patients who are most likely to need it, and reduce costly readmissions in the process.

These are just two examples of how big data techniques are leading to astonishing advances in patient care. But data mining can also pose privacy concerns, especially in a health care context. Our medical history is among our most sensitive personal information. Latanya Sweeney, a Harvard researcher who is currently the Chief Technology Officer at the FTC, has repeatedly demonstrated that it is possible to use publicly available data to identify individual patients in supposedly de-identified medical data sets.

In other contexts, however, there are big data sets that pose no, or very remote, risks at all to personal privacy. One example is public data on climate, weather, and environmental factors. Collected by huge networks of sensors and a fleet of NOAA and NASA satellites, these data have fueled our understanding of how our climate is changing. These data get plugged into the complex computer models that predict sea-level rise, ocean acidification, storm surge risk, and other climate impacts. The economic and political import of these data is extraordinary. But by way of example, without NOAA data, we wouldn’t have the Weather Channel. Without GPS signals made available by Department of Defense satellites, our smartphones wouldn’t be able to tell us how to get from here to the Golden Gate Bridge.

At the same time, even civic-minded big data applications can have troubling implications not just for personal privacy, but for equality and nondiscrimination as well— an issue we are thinking a lot about in the course of our study. For instance, in Boston, the city released an app called Street Bump that used smartphone sensors to detect potholes and report them to the department of public works.
But what happened after Street Bump was first rolled out in Boston should give us pause. Because poor people and the elderly were less likely to carry smartphones, let alone download the app, the app wound up systematically directing city services to wealthier neighborhoods. Now, to its credit, the city of Boston figured this out and tweaked the app to account for underreporting, so that everyone would have equal access to city services. But the lesson here is that we need to pay careful attention to what unexpected outcomes the use of big data might lead to, and how to remedy any unintended discrimination or inequality that may result.

As I said earlier, this review is a scoping exercise, aimed at identifying trends and defining the questions that will inform future research and policy development. But a few big issues are coming into focus, and were reinforced throughout the day, and I’d like to touch on some of those before I conclude.

First, the notice-and-consent framework that has governed data collection for decades is coming under stress in the big data context. In public spaces, people are often unaware of the degree to which sensors, cameras, and other data collection tools are recording information about them. They may not fully understand how that information gets used and shared.

And if people aren’t reading terms of use agreements on their computers before they click “OK,” it seems unlikely that they will parse pages of legalese that may accompany their smart watch, Nike Fuel band (it’s time we let FitBit off the hook) or other wearable device.

So there is an active and ongoing debate about how we can shore up the notice and consent framework that has been the bedrock of consumer privacy for four decades, or whether we need to create other ways to ensure people’s personal information, or information collected about them, is being responsibly used in a big data world.

Next, it’s also clear that what constitutes “responsible use” will be different in different contexts. Companies and researchers working with sensitive health or financial information have different responsibilities than marketers looking to connect you with online advertisements that match what they predict you might want to buy.

In particular, as the President said in January, the United States government must hold itself to a higher standard when it comes to using data responsibly. He was speaking of how we handle data in the intelligence context, but protecting sensitive data is a challenge that many federal agencies have taken seriously for decades, including the Census Bureau, the IRS, and HHS.

Defining responsible use standards for federal agencies is essential for public trust, and will need to be balanced against the economic benefits of opening more government data.

With big data, the government can lead by example in other ways as well—for instance, by making it easier for people to access data that the government holds about them. The Department of Veterans Affairs Blue Button tool helps veterans access their medical records, manage their health care, and correct any errors in their files, all in one easily accessible place. The My Data Initiative at the Department of Education seeks to give every student access to his or her academic data so they can make more informed decisions about college choices and study needs. Efforts like these encourage greater transparency in how data are collected and used, and give people more control over their personal information.

The flow of big data is, of course, global in scale, and recognizing that, the President charged us “to consider how we can forge international norms on how to manage this data.” So to inform the process, we’ve met with a range of foreign...
counterparts, including privacy regulators from across Europe and the Americas. That's why it's important that Rainer Stentzel made the trek to be here with us today. There is tremendous interest in how big data challenges existing regulatory and ethical paradigms for privacy. There are no simple answers but it's clear that there is a whole lot riding on getting the answers at least in the right zone, including the future of a single open, interoperable, secure and reliable global Internet, and the free flow of information. Big data raises real policy questions for governments around the world. But I would note that even the European regulators we met with noted that the conversation here in the U.S. is as deep and more technologically informed than anywhere.

It's clear from the conversations amongst all the panels today, the policy implications are serious. One of the issues that has emerged powerfully today is the need to have a serious conversation about civil rights and discrimination in the big data context. Big data analysis of information voluntarily shared on social networks has showed how easy it can be to infer information about race, ethnicity, religion, gender, age, and sexual orientation, among other personal details. We have a strong legal framework in this country forbidding discrimination based on these criteria in a variety of contexts. But it's easy to imagine how big data technology, if used to cross legal lines we have been careful to set, could end up reinforcing existing inequities in housing, credit, employment, health and education. This is a place where Cynthia Dwork's question on the last panel about how to think about the relationship between utility and fairness comes into highest relief.

All this leads to a few important, technical conclusions about the world of big data in which we increasingly live. First, that the costs and physical footprint of recording data about this world is shrinking to zero, so our expectations of actions being recorded will only increase.

Second, the streams of information that might fall into the category of “surveillance” are rapidly increasing in number and diversity — and being put to some very interesting uses. Behind all the powerful functionality in, for instance, our smartphones or cars, are sensors for temperature, audio, acceleration, light, and heat that are more sensitive than anything available to consumers three decades ago.

Third, as the opportunities for big data analytics increase, so will the demand for high-performance computing capable of supporting it. The architecture for high performance computing to do computational simulation of complex systems is merging with big data analytic architectures. That will have big implications from energy to manufacturing to research and development.

And finally, it is clear that the new paradigm in big data will be defined by the move from ‘search’ to ‘prediction,’ or from a predicated search to a non-predicated search, raising a whole number of practical and ethical questions for consumers, companies, and particularly for governments.

For those of you looking for what we think the most important social and ethical question these trends raise: stay tuned. The report will come out in a matter of weeks.

So with those insights in mind, let me end on this note. Melvin Kranzberg’s six laws of technology begin by stating: Technology is neither good nor bad. Nor is it neutral. Like Kranzberg, I believe technological progress demands that we engage with the social, ethical, legal and political questions that arise.

So I want to thank you all again for your time and for your thoughtful engagement during today’s workshop and throughout our review of big data and privacy. I’m confident that this is a conversation that will continue inside the
Administration, but more importantly in every corner of America and around the world. After today, we’ll be returning to Washington to draft and review our report, and we look forward to releasing our findings in the very near term.

Thank you for your attention – I think we have some time for questions.

White House / Data & Society Research Institute / NYU Information Law Institute

- My Note: Why no PDF of John Podesta's remarks?
- Video of Counselor John Podesta’s remarks (via Data & Society Research Institute)

White House / MIT

- PDF of Counselor John Podesta's remarks (PDF) My Note: See Below
- Video of Counselor John Podesta’s remarks (via MIT)

Remarks as Delivered by Counselor John Podesta

Source: http://www.whitehouse.gov/sites/defa...a_big_data.pdf (PDF)

The White House/MIT "Big Data" Privacy Workshop March 3, 2014

Introduction

Good morning. I’m sorry to be talking to you remotely. In my world, big data squared off against big snow, and big snow won. Secretary Pritzker is travelling from NYC and I hope she’ll have better luck and be with you at lunch. And the Administration is well represented by our Deputy CTO Nicole Wong and the head of NTIA Larry Strickling. I want to start by thanking President Reif, for joining the White House in this important exploration of the technologies driving the big data revolution. And I want to thank Danny Weitzner and Elizabeth Bruce in particular, not only for putting together this outstanding event, but for their ongoing contributions to the research in this area. This workshop is the first in a series of events which the White House will be co-hosting with academic institutions across the country. So, it is a fitting time for me to provide some background about the 90-day White House study of big data and privacy, the process that we’re currently undertaking, and what we hope to accomplish in the next several weeks.

Background

As many of you will recall, on January 17, the President spoke to the American people about how to keep us safe from terrorism in a changing world, and at the same time continue to uphold America’s commitment to liberty and privacy that our values and our Constitution require. In that speech, he asked me to lead a comprehensive review of big data and privacy, recognizing that national security is not the only space where changes in technology are altering the landscape of how data is collected and used, and challenging traditional conceptions of privacy. So, one purpose of this study is to get a more holistic view of the state of the technology and the benefits and challenges that it brings. This Administration remains committed to an open, interoperable, secure and reliable internet – the fundamentals that have enabled innovation to flourish, drive markets and improve lives. We also recognize that ensuring the continued strength of
internet requires applying our timeless privacy values to these new technologies, as we have throughout our history with each new mode of communication from the mail to the telephone to the social network.

We are undergoing a revolution in the way that information about our purchases, our conversations, our social networks, our movements, and even our physical identities are collected, stored, analyzed and used. The immense volume, diversity, velocity, and potential value of data will have profound implications for privacy, the economy, and public policy. The White House working group will consider all those issues, and specifically how the present and future state of these technologies might motivate us to re-visit our policies across a range of sectors.

There is a lot of buzz these days about “Big Data” – a lot of marketing-speak and pitch materials for VC funding. For purposes of the White House study, when we talk about “big data” we’re referring to data sets that are so large, so diverse, or so complex that the conventional tools that would ordinarily be used to manage data simply don’t work. Instead, deriving value from these data sets require a series of more sophisticated techniques, such as Hadoop, NoSQL, MapReduce and machine learning. These techniques enable the discovery of insights from big data sets that were not previously possible.

There is no question that there is more data than ever before, and no sign that the trajectory is slowing its upward pace. In 2012, there were an estimated 2.4 billion global internet users. The amount of global digital information created and shared – from documents to photos to videos to tweets – grew 9x in five years to nearly 2 zettabytes in 2011 (a zettabyte is one trillion gigabytes). On Facebook, there are some 350 million photos uploaded and shared every day. On YouTube, 100 hours of video is uploaded every minute. And we are only in the very nascent stage of the “Internet of Things,” where our appliances will communicate with each other and sensors may be nearly ubiquitous.

The value that can be generated by the use of big data is not hypothetical. The availability of large data sets, and the computing power to derive value from them, is creating new business models, enabling innovations to improve efficiency and performance in a variety of public and private sector settings, and making possible valuable data-driven insights that are measurably improving outcomes in areas from education to healthcare. For example, The Cancer Genome Atlas, an NIH-funded program, is using large genomic data sets to map the genetic changes in more than 20 cancer types. Their researchers have discovered that breast and ovarian cancers have genomic similarities that may have implications for treating these diseases.

With the exponential advance of these capabilities, we must make sure that our modes of protecting privacy – whether technological, regulatory or social – also keep pace. Now, it’s certainly true that data analytics is an old science, dating to the late 1800s. In this study, we want to explore whether there is something truly new in the vast collection of data and lightning-speed analytics that are made possible by new technologies, computational strategies and cratering storage costs. My hope is that this inquiry will anticipate future technological trends to help us frame the key questions arising from the collection, availability, and use of big data — both for our government, and the nation as a whole – and develop a workplan to address them. Today’s conference – appropriately set at MIT which has been the cradle for so many game-changing technologies – is part of this 90-day endeavor, and is designed to provide a firm grounding in the current state of technologies and their likely trajectories.

The Administration’s Big Data initiatives

It is important to note that the Administration is not starting from scratch when it comes to big data or privacy.
Since the earliest days of this Administration, the Federal Government has taken unprecedented steps to make
government data more available to citizens, companies and innovators. Through the Data.gov platform which launched
in 2009, users have been able to access thousands of government datasets about a wide range of topics. The Open
Data Initiative and Executive Order that the President signed last year commits federal agencies to unlocking even more
valuable data from the vaults of government in health, energy, education, public safety, finance and global development.

The natural outgrowth of this commitment to making large data sets available for public innovation is a broad
commitment to the technologies that can harness these assets. In 2012, the Administration announced a $200 million
commitment by 6 agencies to invest in big data projects. And just last fall, we showcased 28 public-private partnerships
harnessing big data to enhance national priorities, including economic growth and job creation, education and health,
energy and sustainability, public safety and national security, and global development. Indeed, one of those projects was
launched from here as part of MIT’s Big Data Initiative. We are pleased to be collaborating with CSAIL’s Big Data
Privacy Working Group, and we look forward to hearing from some of the researchers engaged in that project later
today.

The United States can also be proud of its long history as a leader in information privacy, starting with the pioneering of
the Fair Information Practice Principles in the 1970s. Those principles -- known as the “FIPPs” -- are the underpinnings
of the Privacy Act of 1974 which articulates the rights of citizens and the obligations of government to protect personal
information. The same principles have also become the globally-recognized foundation for privacy protection, adopted
by the OECD and providing the framework for privacy regimes around the world. And President Obama, from early in his
first term, has been working to advance protections for individual privacy in this new age of information technology.

Indeed, the Administration announced a groundbreaking privacy document in 2012, with the release of its consumer
privacy blueprint, including the Consumer Privacy Bill of Rights. The blueprint refined the FIPPs to be more focused on
consumers in terms they could understand in their own lives. It also re-framed the FIPPs to better accommodate the
incredibly innovative online environment in which we all now live. While the document does not specifically use the term
“big data,” the blueprint recognized that significant data was being collected about individuals online, and that some data
would be sensitive. It also assumed that this data could deliver significant value, if properly used, to individual
consumers.

What we will be exploring in this study is whether the Consumer Privacy Bill of Rights fully addresses the changes that
today we refer to as the “big data revolution” -- recognizing that we may only be at the beginning of that revolution. What
the President wants to explore, in part, is whether our existing privacy framework can accommodate these changes, or if
there are new avenues for policy that we need to consider.

Have we fully considered the myriad ways in which this data revolution might create social value, and have we fully
contemplated the risks that it might pose to our conceptions of individual privacy, personal freedom and government
responsibility of data?

As we move from predicated analysis of data -- that is, using data to find something we already know that we’re looking
for -- to non-predicated, or pattern-based, searches -- using data to find patterns that reveal new insights -- I think we
need to be conscious of the implications for individuals.
How should we think about individuals’ sense of their identity when data reveals things they didn’t even know about themselves? In this study, we want to explore the capabilities of big data analytics, but also the social and policy implications of that capability.

Our work is still in its early stages, but already we’re learning important things about the current state of technology and its potential. For example, we recently met with some leaders in higher education to discuss the use of academic performance data to improve learning outcomes. There is some terrific research happening in this area, and it worth talking about in a bit more detail.

The Pittsburgh Science of Learning Center – an NSF-funded center that joins the disciplines of cognitive learning and computer science -- hosts the “DataShop”, the world's preeminent central repository for data on the interactions between students and educational software and a suite of tools to analyze that data. In collaboration with private sector partners, they have made largescale data sets available to develop learning models aimed at improving math, science and language curriculums for K-12 students. In one study, the researchers tested a new algebra curriculum for middle school to high school students that utilized education technologies for instruction and to measure performance.

As some of you may know, mathematics proficiency rates of students in the United States – while on the rise – are still far below what they should be and lag behind students in the top-scoring countries. While there is still more work to be done, the early results of these large-scale studies show significant gains – 8 percentile points more than usual -- an amount that is nearly double how much students learn from a typical algebra course.

Importantly, what this type of research underscores is that the use of educational technologies is improving the scope, scale and granularity of data we have about how kids learn. With that data, and applying cognitive and data science to the problem, we are better able to understand how to help our children move up the performance curve. We are gaining insights that were not previously possible, or maybe were dismissed for lack of concrete data, because of the new capability to capture student performance in detail and at scale in diverse, real-world school contexts.

Of course, there are also privacy implications to be considered when gathering and using this data. While the educators working with the students obviously knew how individual kids were doing on their tests, the researchers who developed those data-driven education tools only had de-identified data and deliberately decided not to collect the demographic data of their students. Now, given what we already know about effective education policy, demographic data might have been useful both in developing effective curriculum and addressing the needs of the individual student. But the researchers decided that collection of such information raised privacy and ethical concerns, and that they could make progress without that data.

We can see similar real-world benefits and similar privacy questions raised in a range of areas: from tracking electricity usage in a home to significantly bring down energy costs, to collecting individual location data in order to reduce traffic congestion. I believe we’ll be hearing about some of these innovative uses today, including uses in education, genomics, and transportation.

This is the power of big data analytics that could unleash real human potential, and so a goal of this study is to look at where the federal government can play a role in supporting this type of work while continuing to protect personal privacy and other values.

So that is the context of this inquiry.
The Study

Now, let me just take a few moments to explain a bit more about the review, its scope, and what you can expect over the next 90 days.

In his speech, the President asked me to lead a comprehensive review of the way that “big data” will affect the way we live and work; the relationship between government and citizens; and how public and private sectors can spur innovation and maximize the opportunities and free flow of this information while minimizing the risks to privacy. I will be joined in this effort by Secretary of Commerce Penny Pritzker (who will be your lunchtime keynote later today), Secretary of Energy Ernie Moniz, the President’s Science Advisor John Holdren, the President’s Economic Advisor Jeff Zients and other senior government officials.

This is going to be a collaborative effort with four channels of engagement.

First, the President’s Council of Advisors on Science and Technology (PCAST) is conducting a parallel study to explore in-depth the technological dimensions of the intersection of big data and privacy. Their report will feed into this broader effort and ensure a substantive grounding in the technologies at issue.

Second, our working group is consulting with a wide range of stakeholders. We have already met with privacy and civil liberties advocates, business leaders, policymakers, international partners, academics and several government agencies on the significance of and future for these technologies. In the next several weeks, we look forward to hearing from a broad range of private sector companies, particularly those who collect and use data to develop products and deliver services, whether by targeted advertising; improved medical treatment; financial services, and more.

We also will engage international audiences, including international regulators and officials, to help answer the President’s charge that we consider “whether we can forge international norms on how to manage this data; and how we can continue to promote the free flow of information in ways that are consistent with both privacy and security.”

Third, this workshop kicks off a series of events that we are co-hosting around the country to convene stakeholders to discuss these very issues and questions. The next event will be on March 17, co-hosted with the Data & Society Research Institute and NYU, and will focus on the social, cultural and ethical implications of big data. Then, on April 1, we will co-host an event with the School of Information and Berkeley Center for Law & Technology at UC Berkeley, which will focus on the legal and policy issues raised by big data.

Finally, and perhaps most importantly, we want to engage the public. This is not a discussion that should be confined to Washington or academia. This is an issue of such importance, an array of technologies already so pervasive, that it requires public participation in the conversation about how we realize the great benefits of big data while protecting individual privacy and other values. To this end, this week I will be posting a video to the White House website that describes this inquiry and asks the public “what technologies are most transformative in your life?” and “which technologies give you pause?” We have also just initiated a process to receive written comments addressing these questions in even more depth. You can find both channels for providing your input on the White House website, and we welcome your comments and ideas. All of these discussions will help to inform our study.

This study is fundamentally a scoping exercise. We are trying to get a full view of the landscape – the technologies at play, the uses by the government, industry and academia. Whether we want to examine the Administration’s consumer...
privacy blueprint—including the Consumer Privacy Bill of Rights—and how its principles can be applied in this new landscape. That may prompt us to look harder at some of our existing policies, at our research agenda, or at specific sectors where great gains could be made by the use of big data.

When we complete our work, we expect to deliver to the President a report that anticipates future technological trends and frames the key questions that the collection, availability, and use of “big data” raise – both for our government, and the nation as a whole. It will help identify technological changes to watch, whether those technological changes are addressed by the U.S.’s current policy framework and highlight where further government action, funding, research and consideration may be required.

While we don’t expect to answer all these questions, or produce a comprehensive new policy in 90 days, we expect this work to serve as the foundation for a robust and forward-looking plan of action.

This is a fascinating and complex area, so let me close by throwing out a few questions that we have been thinking about:

What is genuinely “new” about big data and what, if any, policies should be revisited because of those changes?

What business models do you think are most dependent, today, on big data? How will that change in, say, 5 years? 15?

What types of uses of big data could measurably improve social or economic outcomes or productivity with further government action, funding or research?

Can we “build in” additional privacy protection into the architecture of big data analytics and should the government and the private sector be investing more in research toward that end. For individuals, what do you think will be the most significant effects of these emerging pattern-based data mining techniques?

Thank you for your time this morning and your engagement in this national conversation. I am sorry I am not with you in person, but I’ll be watching the feed. If we can manage the technology, I think have a few more minutes to take some questions.

---

President Obama Discusses U.S. Intelligence Programs at the Department of Justice

Source: [http://www.whitehouse.gov/blog/2014/...rtment-justice](http://www.whitehouse.gov/blog/2014/...rtment-justice)

Megan Slack

Megan Slack January 17, 2014, 06:44 PM EDT
President Barack Obama delivers remarks presenting the outcome of the Administration's review of the NSA and U.S. signals intelligence programs, at the Department of Justice in Washington, D.C., Jan. 17, 2014. (Official White House Photo by Pete Souza)

Today, President Obama delivered a speech at the Department of Justice to announce the outcomes of a broad-ranging and unprecedented review of U.S. intelligence programs.

The review examined how, in light of new and changing technologies, we can use our intelligence capabilities in a way that optimally protects our national security while supporting our foreign policy, respecting privacy and civil liberties, maintaining the public trust, and reducing the risk of unauthorized disclosures.

Additionally, President Obama issued a new presidential policy directive (My Note: Look at This) for our signals intelligence activities, at home and abroad. This directive lays out new principles that govern how we conduct signals intelligence collection, and strengthens how we provide executive branch oversight of our signals intelligence activities.

Watch the speech below, or see this White House fact sheet (My Note: Look at This) to learn more.

Related Topics: Foreign Policy
FACT SHEET: Big Data and Privacy Working Group Review

Driven by the declining cost of data collection, storage, and processing; fueled by new online and real-world sources of data, including sensors, cameras, and geospatial technologies; and analyzed using a suite of creative and powerful new methods, big data is fundamentally reshaping how Americans and people around the world live, work, and communicate. It is enabling important discoveries and innovations in public safety, health care, medicine, education, energy use, agriculture, and a host of other areas. But big data technologies also raise challenging questions about how best to protect privacy and other values in a world where data collection will be increasingly ubiquitous, multidimensional, and permanent.

In January, President Obama asked his Counselor John Podesta to lead a 90-day review of big data and privacy. The review was conceived as fundamentally a scoping exercise, designed to define for the President what is new about the technologies that define the big data landscape; uncover where and how big data affects public policy and the laws and norms governing privacy; to ask how and whether big data creates new challenges for the principles animating the Consumer Privacy Bill of Rights (My Note: Look at This) embraced by the Administration in 2012; and to lay out an agenda for how government can maximize the benefits and minimize the risks of big data.

The working group—which included Commerce Secretary Pritzker, Energy Secretary Moniz, the President's Science Advisor John Holdren, the President's Economic Advisor Jeff Zients, and other Senior Administration Officials—sought public input and worked over 90 days with academic researchers and privacy advocates, regulators and the technology industry, advertisers and civil rights groups, the international community and the American public. This review was supported by a parallel effort by the President's Council of Advisors on Science and Technology (PCAST) to research the technological trends underpinning big data.

Today, Podesta and the big data working group presented their findings and recommendations to the President (My Note: See Below). The review did not set out to answer every question about big data, nor was it intended to develop a comprehensive policy approach to big data. However, by evaluating the opportunities and challenges presented by big data, the working group was able to draw important conclusions and make concrete recommendations to the President for Administration attention and policy development.

SEIZING OPPORTUNITIES

We live in a world of near-ubiquitous data collection where that data is being crunched at a speed increasingly approaching real-time. This revolution presents incredible opportunities:

- **Big data is saving lives.** Infections are dangerous—even deadly—for many babies born prematurely. By collecting and analyzing millions of data points from a neonatal intensive care unit, one study was able to identify factors, like slight changes in body temperature and heart rate, that serve as early warning signs an infection may be taking root—subtle changes that even the most experienced doctors may not have have noticed on their own.
• **Big data is making the economy work better.** Jet engines and delivery trucks now come outfitted with sensors that continuously monitor hundreds of data points and send automatic alerts when maintenance is needed. Utility companies are starting to use big data to predict periods of peak electric demand, adjusting the grid to be more efficient and potentially averting brown-outs.

• **Big data is saving taxpayer dollars.** The Centers for Medicare and Medicaid Services have begun using predictive analytics—a big data technique—to flag likely instances of reimbursement fraud before claims are paid. The Fraud Prevention System helps identify the highest-risk health care providers for waste, fraud, and abuse in real time and has already stopped, prevented, or identified $115 million in fraudulent payments. ([My Note: Look at This](#))

Big data also presents powerful opportunities in areas as diverse as medical research, agriculture, energy efficiency, global development, education, environmental monitoring, and modeling climate change impacts, among others.

**PRESERVING OUR VALUES**

The opportunities presented by big data are considerable, but big data raises serious concerns about how we protect our privacy and other values. For example:

• **Big data tools can alter the balance of power between government and citizen.** Government agencies can reap enormous benefits from using big data to improve service delivery or detect payment fraud. But government uses of big data also have the potential to chill the exercise of free speech or free association. As more data is collected, analyzed, and stored on both public and private systems, we must be vigilant in ensuring that balance is maintained between government and citizens, and revise our laws accordingly.

• **Big data tools can reveal intimate personal details.** One powerful big data technique involves merging multiple data sets, drawn from disparate sources, to reveal complex patterns. But this practice, sometimes known as “data fusion,” can also lead to the so-called “mosaic effect,” whereby personally identifiable information can be discerned even from ostensibly anonymized data. As big data becomes even more widely used in the private sector to bring a wellspring of innovations and productivity, we must ensure that effective consumer privacy protections are in place to protect individuals.

• **Big data tools could lead to discriminatory outcomes.** As more decisions about our commercial and personal lives are determined by algorithms and automated processes, we must pay careful attention that big data does not systematically disadvantage certain groups, whether inadvertently or intentionally. We must prevent new modes of discrimination that some uses of big data may enable, particularly with regard to longstanding civil rights protections in housing, employment, and credit.

**POLICY RECOMMENDATIONS**

No matter how quickly technology advances, it remains within our power to ensure that we both encourage innovation and protect our values through law, policy, and the practices we encourage in the public and private sector. To that end, the working group made six actionable policy recommendations in their report to the President:

• **Advance the Consumer Privacy Bill of Rights** ([My Note: Look at This](#)) because consumers deserve clear, understandable, reasonable standards for how their personal information is used in the big data era.

• **Pass National Data Breach Legislation** that provides for a single national data breach standard, along the lines of the Administration's 2011 [Cybersecurity legislative proposal](#) ([My Note: Look at This](#)).

• **Extend Privacy Protections to non-U.S. Persons** because privacy is a worldwide value that should be reflected in how the federal government handles personally identifiable information from non-U.S. citizens.

• **Ensure Data Collected on Students in School is used for Educational Purposes** to drive better learning outcomes while protecting students against their data being shared or used inappropriately.

https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review

Updated: Thu, 25 Jul 2019 08:57:25 GMT

Powered by [MindTouch](#)
• **Expand Technical Expertise to Stop Discrimination** because the federal government should build the technical expertise to be able to identify practices and outcomes facilitated by big data analytics that have a discriminatory impact on protected classes.

• **Amend the Electronic Communications Privacy Act** to ensure the standard of protection for online, digital content is consistent with that afforded in the physical world—including by removing archaic distinctions between email left unread or over a certain age.

---

**Big Data Report Blog Post**

Source: [http://www.whitehouse.gov/blog/2014/...g-group-review](http://www.whitehouse.gov/blog/2014/...g-group-review)

Findings of the Big Data and Privacy Working Group Review  
John Podesta  
May 01, 2014, 01:15 PM EDT

Over the past several days, severe storms have battered Arkansas, Oklahoma, Mississippi and other states. Dozens of people have been killed and entire neighborhoods turned to rubble and debris as tornadoes have touched down across the region. Natural disasters like these present a host of challenges for first responders. How many people are affected, injured, or dead? Where can they find food, shelter, and medical attention? What critical infrastructure might have been damaged?

Drawing on open government data sources, including Census demographics and NOAA weather data, along with their own demographic databases, Esri, a geospatial technology company, has created a real-time map ([My Note: Look at This](#)) showing where the twisters have been spotted and how the storm systems are moving. They have also used these data to show how many people live in the affected area, and summarize potential impacts ([My Note: Look at This](#)) from the storms. It's a powerful tool for emergency services and communities. And it's driven by big data technology.

In January, President Obama asked ([My Note: Look at This](#)) me to lead a wide-ranging review of "big data" and privacy—to explore how these technologies are changing our economy, our government, and our society, and to consider their implications for our personal privacy. Together with Secretary of Commerce Penny Pritzker, Secretary of Energy Ernest Moniz, the President's Science Advisor John Holdren, the President's Economic Advisor Jeff Zients, and other senior officials, our review sought to understand what is genuinely new and different about big data and to consider how best to encourage the potential of these technologies while minimizing risks to privacy and core American values.

Over the course of 90 days, we met with academic researchers and privacy advocates, with regulators and the technology industry, with advertisers and civil rights groups. The President's Council of Advisors for Science and Technology conducted a parallel study of the technological trends underpinning big data. The White House Office of Science and Technology Policy jointly organized three university conferences at MIT ([My Note: I Have This](#)), NYU ([My Note: I Looked at This](#)), and U.C. Berkeley ([My Note: I Looked at This](#)). We issued a formal Request for Information seeking public comment, and hosted a survey ([My Note: See Above](#)) to generate even more public input.
Today, we presented our findings to the President. We knew better than to try to answer every question about big data in three months. But we are able to draw important conclusions and make concrete recommendations for Administration attention and policy development in a few key areas.

There are a few technological trends that bear drawing out. The declining cost of collection, storage, and processing of data, combined with new sources of data like sensors, cameras, and geospatial technologies, mean that we live in a world of near-ubiquitous data collection. All this data is being crunched at a speed that is increasingly approaching real-time, meaning that big data algorithms could soon have immediate effects on decisions being made about our lives.

The big data revolution presents incredible opportunities in virtually every sector of the economy and every corner of society.

Big data is saving lives. Infections are dangerous—even deadly—for many babies born prematurely. By collecting and analyzing millions of data points from a NICU, one study was able to identify factors, like slight increases in body temperature and heart rate, that serve as early warning signs an infection may be taking root—subtle changes that even the most experienced doctors wouldn't have noticed on their own.

Big data is making the economy work better. Jet engines and delivery trucks now come outfitted with sensors that continuously monitor hundreds of data points and send automatic alerts when maintenance is needed. Utility companies are starting to use big data to predict periods of peak electric demand, adjusting the grid to be more efficient and potentially averting brown-outs.

Big data is making government work better and saving taxpayer dollars. The Centers for Medicare and Medicaid Services have begun using predictive analytics—a big data technique—to flag likely instances of reimbursement fraud before claims are paid. The Fraud Prevention System helps identify the highest-risk health care providers for waste, fraud, and abuse in real time and has already stopped, prevented, or identified $115 million in fraudulent payments.

But big data raises serious questions, too, about how we protect our privacy and other values in a world where data collection is increasingly ubiquitous and where analysis is conducted at speeds approaching real time. In particular, our review raised the question of whether the "notice and consent" framework, in which a user grants permission for a service to collect and use information about them, still allows us to meaningfully control our privacy as data about us is increasingly used and reused in ways that could not have been anticipated when it was collected.

Big data raises other concerns, as well. One significant finding of our review was the potential for big data analytics to lead to discriminatory outcomes and to circumvent longstanding civil rights protections in housing, employment, credit, and the consumer marketplace.

No matter how quickly technology advances, it remains within our power to ensure that we both encourage innovation and protect our values through law, policy, and the practices we encourage in the public and private sector. To that end, we make six actionable policy recommendations in our report to the President:

**Advance the Consumer Privacy Bill of Rights.** Consumers deserve clear, understandable, reasonable standards for how their personal information is used in the big data era. We recommend the Department of Commerce take appropriate consultative steps to seek stakeholder and public comment on what changes, if any, are needed to...
the Consumer Privacy Bill of Rights (My Note: Look at This), first proposed by the President in 2012, and to prepare draft legislative text for consideration by stakeholders and submission by the President to Congress.

**Pass National Data Breach Legislation.** Big data technologies make it possible to store significantly more data, and further derive intimate insights into a person's character, habits, preferences, and activities. That makes the potential impacts of data breaches at businesses or other organizations even more serious. A patchwork of state laws currently governs requirements for reporting data breaches. Congress should pass legislation that provides for a single national data breach standard, along the lines of the Administration's 2011 Cybersecurity legislative proposal (My Note: Look at This).

**Extend Privacy Protections to non-U.S. Persons.** Privacy is a worldwide value that should be reflected in how the federal government handles personally identifiable information about non-U.S. citizens. The Office of Management and Budget should work with departments and agencies to apply the Privacy Act of 1974 to non-U.S. persons where practicable, or to establish alternative privacy policies that apply appropriate and meaningful protections to personal information regardless of a person's nationality.

**Ensure Data Collected on Students in School is used for Educational Purposes.** Big data and other technological innovations, including new online course platforms that provide students real time feedback, promise to transform education by personalizing learning. At the same time, the federal government must ensure educational data linked to individual students gathered in school is used for educational purposes, and protect students against their data being shared or used inappropriately.

**Expand Technical Expertise to Stop Discrimination.** The detailed personal profiles held about many consumers, combined with automated, algorithm-driven decision-making, could lead—intentionally or inadvertently—to discriminatory outcomes, or what some are already calling "digital redlining." The federal government's lead civil rights and consumer protection agencies should expand their technical expertise to be able to identify practices and outcomes facilitated by big data analytics that have a discriminatory impact on protected classes, and develop a plan for investigating and resolving violations of law.

**Amend the Electronic Communications Privacy Act.** The laws that govern protections afforded to our communications were written before email, the internet, and cloud computing came into wide use. Congress should amend ECPA to ensure the standard of protection for online, digital content is consistent with that afforded in the physical world—including by removing archaic distinctions between email left unread or over a certain age.

We also identify several broader areas ripe for further study, debate, and public engagement that, collectively, we hope will spark a national conversation about how to harness big data for the public good. We conclude that we must find a way to preserve our privacy values in both the domestic and international marketplace. We urgently need to build capacity in the federal government to identify and prevent new modes of discrimination that could be enabled by big data. We must ensure that law enforcement agencies using big data technologies do so responsibly, and that our fundamental privacy rights remain protected. Finally, we recognize that data is a valuable public resource, and call for continuing the Administration's efforts to open more government data sources and make investments in research and technology.
While big data presents new challenges, it also presents immense opportunities to improve lives, the United States is perhaps better suited to lead this conversation than any other nation on earth. Our innovative spirit, technological know-how, and deep commitment to values of privacy, fairness, non-discrimination, and self-determination will help us harness the benefits of the big data revolution and encourage the free flow of information while working with our international partners to protect personal privacy. This review is but one piece of that effort, and we hope it spurs a conversation about big data across the country and around the world.

Read the Big Data Report

My Note: See Below

See the fact sheet from today's announcement.

My Note: See Above

John Podesta is a Counselor to the President.

Related Topics: Technology, Arkansas, Mississippi, Oklahoma

Read the study from the President's Council of Advisors on Science and Technology

Source: http://www.whitehouse.gov/blog/2014/...g-group-review

Findings of the Big Data and Privacy Working Group Review

John Podesta
May 01, 2014, 01:15 PM EDT

Over the past several days, severe storms have battered Arkansas, Oklahoma, Mississippi and other states. Dozens of people have been killed and entire neighborhoods turned to rubble and debris as tornadoes have touched down across the region. Natural disasters like these present a host of challenges for first responders. How many people are affected, injured, or dead? Where can they find food, shelter, and medical attention? What critical infrastructure might have been damaged?

Drawing on open government data sources, including Census demographics and NOAA weather data, along with their own demographic databases, Esri, a geospatial technology company, has created a real-time map showing where the twisters have been spotted and how the storm systems are moving. They have also used these data to show how many people live in the affected area, and summarize potential impacts from the storms. It's a powerful tool for emergency services and communities. And it's driven by big data technology.

In January, President Obama asked me to lead a wide-ranging review of "big data" and privacy—to explore how these technologies are changing our economy, our government, and our society, and to consider their implications for our personal privacy. Together with Secretary of Commerce Penny Pritzker, Secretary of Energy Ernest Moniz, the President's Science Advisor John Holdren, the President's Economic Advisor Jeff Zients, and other senior officials, our review sought to understand what is genuinely new and different about big data and to consider how best to encourage the potential of these technologies while minimizing risks to privacy and core American values.

Over the course of 90 days, we met with academic researchers and privacy advocates, with regulators and the technology industry, with advertisers and civil rights groups. The President's Council of Advisors for Science and
Technology conducted a parallel study of the technological trends underpinning big data. The White House Office of Science and Technology Policy jointly organized three university conferences at MIT (My Note: I Have This), NYU (My Note: I Looked at This), and U.C. Berkeley (My Note: I Looked at This). We issued a formal Request for Information seeking public comment, and hosted a survey to generate even more public input.

Today, we presented our findings to the President (My Note: See Below). We knew better than to try to answer every question about big data in three months. But we are able to draw important conclusions and make concrete recommendations for Administration attention and policy development in a few key areas.

There are a few technological trends that bear drawing out. The declining cost of collection, storage, and processing of data, combined with new sources of data like sensors, cameras, and geospatial technologies, mean that we live in a world of near-ubiquitous data collection. All this data is being crunched at a speed that is increasingly approaching real-time, meaning that big data algorithms could soon have immediate effects on decisions being made about our lives.

The big data revolution presents incredible opportunities in virtually every sector of the economy and every corner of society.

Big data is saving lives. Infections are dangerous—even deadly—for many babies born prematurely. By collecting and analyzing millions of data points from a NICU, one study was able to identify factors, like slight increases in body temperature and heart rate, that serve as early warning signs an infection may be taking root—subtle changes that even the most experienced doctors wouldn't have noticed on their own.

Big data is making the economy work better. Jet engines and delivery trucks now come outfitted with sensors that continuously monitor hundreds of data points and send automatic alerts when maintenance is needed. Utility companies are starting to use big data to predict periods of peak electric demand, adjusting the grid to be more efficient and potentially averting brown-outs.

Big data is making government work better and saving taxpayer dollars. The Centers for Medicare and Medicaid Services have begun using predictive analytics—a big data technique—to flag likely instances of reimbursement fraud before claims are paid. The Fraud Prevention System helps identify the highest-risk health care providers for waste, fraud, and abuse in real time and has already stopped, prevented, or identified $115 million in fraudulent payments.

But big data raises serious questions, too, about how we protect our privacy and other values in a world where data collection is increasingly ubiquitous and where analysis is conducted at speeds approaching real time. In particular, our review raised the question of whether the "notice and consent" framework, in which a user grants permission for a service to collect and use information about them, still allows us to meaningfully control our privacy as data about us is increasingly used and reused in ways that could not have been anticipated when it was collected.

Big data raises other concerns, as well. One significant finding of our review was the potential for big data analytics to lead to discriminatory outcomes and to circumvent longstanding civil rights protections in housing, employment, credit, and the consumer marketplace.

No matter how quickly technology advances, it remains within our power to ensure that we both encourage innovation and protect our values through law, policy, and the practices we encourage in the public and private sector. To that end, we make six actionable policy recommendations in our report to the President:

https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review

Updated: Thu, 25 Jul 2019 08:57:25 GMT

Powered by mindtouch
Advance the Consumer Privacy Bill of Rights. Consumers deserve clear, understandable, reasonable standards for how their personal information is used in the big data era. We recommend the Department of Commerce take appropriate consultative steps to seek stakeholder and public comment on what changes, if any, are needed to the Consumer Privacy Bill of Rights (My Note: Look at This), first proposed by the President in 2012, and to prepare draft legislative text for consideration by stakeholders and submission by the President to Congress.

Pass National Data Breach Legislation. Big data technologies make it possible to store significantly more data, and further derive intimate insights into a person’s character, habits, preferences, and activities. That makes the potential impacts of data breaches at businesses or other organizations even more serious. A patchwork of state laws currently governs requirements for reporting data breaches. Congress should pass legislation that provides for a single national data breach standard, along the lines of the Administration’s 2011 Cybersecurity legislative proposal (My Note: Look at This).

Extend Privacy Protections to non-U.S. Persons. Privacy is a worldwide value that should be reflected in how the federal government handles personally identifiable information about non-U.S. citizens. The Office of Management and Budget should work with departments and agencies to apply the Privacy Act of 1974 to non-U.S. persons where practicable, or to establish alternative privacy policies that apply appropriate and meaningful protections to personal information regardless of a person's nationality.

Ensure Data Collected on Students in School is used for Educational Purposes. Big data and other technological innovations, including new online course platforms that provide students real time feedback, promise to transform education by personalizing learning. At the same time, the federal government must ensure educational data linked to individual students gathered in school is used for educational purposes, and protect students against their data being shared or used inappropriately.

Expand Technical Expertise to Stop Discrimination. The detailed personal profiles held about many consumers, combined with automated, algorithm-driven decision-making, could lead—intentionally or inadvertently—to discriminatory outcomes, or what some are already calling "digital redlining." The federal government's lead civil rights and consumer protection agencies should expand their technical expertise to be able to identify practices and outcomes facilitated by big data analytics that have a discriminatory impact on protected classes, and develop a plan for investigating and resolving violations of law.

Amend the Electronic Communications Privacy Act. The laws that govern protections afforded to our communications were written before email, the internet, and cloud computing came into wide use. Congress should amend ECPA to ensure the standard of protection for online, digital content is consistent with that afforded in the physical world—including by removing archaic distinctions between email left unread or over a certain age.

We also identify several broader areas ripe for further study, debate, and public engagement that, collectively, we hope will spark a national conversation about how to harness big data for the public good. We conclude that we must find a way to preserve our privacy values in both the domestic and international marketplace. We urgently need to build capacity in the federal government to identify and prevent new modes of discrimination that could be enabled by big data. We must ensure that law enforcement agencies using big data technologies do so responsibly, and that our fundamental privacy rights remain protected. Finally, we recognize that data is a valuable public resource, and call for
continuing the Administration's efforts to open more government data sources and make investments in research and technology.

While big data presents new challenges, it also presents immense opportunities to improve lives, the United States is perhaps better suited to lead this conversation than any other nation on earth. Our innovative spirit, technological know-how, and deep commitment to values of privacy, fairness, non-discrimination, and self-determination will help us harness the benefits of the big data revolution and encourage the free flow of information while working with our international partners to protect personal privacy. This review is but one piece of that effort, and we hope it spurs a conversation about big data across the country and around the world.

Read the Big Data Report. My Note: See Below

See the fact sheet from today's announcement. My Note: See Above

John Podesta is a Counselor to the President.
Related Topics: Technology, Arkansas, Mississippi, Oklahoma

PCAST Releases Report on Big Data and Privacy

Source: http://www.whitehouse.gov/blog/2014/...ta-and-privacy

Posted by John P. Holdren, Susan L. Graham, and William Press on May 01, 2014 at 12:09 PM EDT

Earlier this year, President Obama asked (My Note: See Above) his counselor John Podesta to lead a comprehensive review of policy issues at the intersection of big data and privacy. As a contribution to that review, he asked his Council of Advisors on Science and Technology (PCAST) to examine current and likely future capabilities of key technologies, both those associated with the collection, analysis, and use of big data and those that can help to preserve privacy. Over the past 90 days, we have reviewed the technical literature, consulted with additional experts whose research or product-development activity focuses on the key technologies, engaged complementary perspectives from social science and the law to help put our technical insights into perspective, and deliberated over what we were learning.

Today, PCAST is releasing its analysis via a new report, Big Data: A Technological Perspective (My Note: This is a Separate Report), which details the technical aspects of big data and privacy. The ubiquity of computing and electronic communication technologies has led to the exponential growth of data from both digital and analog sources. New technical abilities to gather, analyze, disseminate, and preserve vast quantities of data raise new concerns about the nature of privacy and the means by which individual privacy might be compromised or protected.

This report begins by exploring the changing nature of privacy as computing technology has advanced and big data has come to the forefront. It proceeds by identifying the sources of these data, the utility of these data — including new data analytics enabled by data mining and data fusion — and the privacy challenges big data poses in a world where technologies for re-identification often outpace privacy-preserving de-identification capabilities, and where it is increasingly hard to identify privacy-sensitive information at the time of its collection.

The report outlines a number of recommendations, including:

• Policy attention should focus more on the actual uses of big data and less on its collection and analysis.
Policies and regulation, at all levels of government, should not embed particular technological solutions, but rather should be stated in terms of intended outcomes.

With coordination and encouragement from the White House Office of Science and Technology Policy (OSTP), the Networking and Information Technology Research and Development (NITRD) (My Note: I Have This) agencies should strengthen U.S. research in privacy-related technologies and in the relevant areas of social science that inform the successful application of those technologies.

OSTP, together with the appropriate educational institutions and professional societies, should encourage increased education and training opportunities concerning privacy protection, including career paths for professionals.

The United States should take the lead both in the international arena and at home by adopting policies that stimulate the use of practical privacy-protecting technologies that exist today. It can exhibit leadership both by its convening power (for instance, by promoting the creation and adoption of standards) and also by its own procurement practices (such as its own use of privacy-preserving cloud services).

Read the fact sheet [here](#). My Note: This is a Separate Fact Sheet (PDF) See Below

Read the full report [here](#). My Note: This is a Separate Report. See Below

Read the White House report [here](#). My Note: See Above.

Susan L. Graham and William Press are members of PCAST and co-chairs of the PCAST Big Data and Privacy Working Group.

The PCAST Big Data and Privacy Working Group also includes S. James Gates, Jr., Mark Gorenberg, John Holdren, Eric S. Lander, Craig Mundie, Maxine Savitz, and Eric Schmidt. Marjory S. Blumenthal, Executive Director of PCAST, coordinated the development and contributed to the framing of this report.

See responses to the Request for Information

Source: [http://www.whitehouse.gov/administra...shareyourinput](http://www.whitehouse.gov/administra...shareyourinput)

Share Your Input

My Note: These are all closed

**March 18, 2014 - April 14, 2014**: Request for Information: Interagency Funding for Research and Engineering Projects Conducted by Federal Researchers is now CLOSED.

**March 4, 2014 - April 4, 2014**: Request for Information soliciting public input on how ‘big data’ will affect how Americans live and work, and the implications of collecting, analyzing and using such data for privacy, the economy, and public policy is now CLOSED.

- **On March 4, 2014, OSTP issued a Request For Information** soliciting public input on how ‘big data’ will affect how Americans live and work, and the implications of collecting, analyzing and using such data for privacy, the economy, and public policy. The public comment period was open from March 4, 2014 to April 4, 2014. The comment period is now CLOSED. Comments received by OSTP during the comment period can be found [here](#). **To view a specific RFI response within the document, click the hyperlink under its number.**
February 18, 2014 - March 20, 2014: RFI comment period to submit input on ways to provide greater incentives to Federal agencies to relinquish spectrum for wireless broadband or other innovative commercial uses is now CLOSED.

- On February 18, 2014, OSTP released a Request for Information (RFI) soliciting public input on ways to provide greater incentives to Federal agencies to relinquish spectrum for wireless broadband or other innovative commercial uses. Below are the public comments received by OSTP during the comment period. The Request for Information follows last year’s Presidential Memorandum, Expanding America’s Leadership in Wireless Innovation, which aims to promote more efficient use of spectrum by Federal agencies. Among other things, the memorandum directs the Spectrum Policy Team to make recommendations to the President on approaches that could give agencies greater incentive to share or relinquish spectrum, while protecting the mission capabilities of existing and future systems that rely on spectrum use. The full Request for Information can be found here and the new report from the Science and Technology Policy Institute can be found here.

- Comments on the questions in the RFI were accepted through March 20, 2014. Received comments are presented together, in alphabetical order per the electronic filename of the submission, here.

November 12, 2013 - December 6, 2013: RFI comment period to submit input related to development of a National Plan for Civil Earth Observations is now CLOSED. Comments on the RFI were accepted through December 6, 2013 and can be found here.

---

**Big Data: Seizing Opportunities, Preserving Values**

Source: [http://www.whitehouse.gov/sites/defa...inal_print.pdf](http://www.whitehouse.gov/sites/defa...inal_print.pdf) (PDF)

### I. Big Data and the Individual

**What is Big Data?**

Since the first censuses were taken and crop yields recorded in ancient times, data collection and analysis have been essential to improving the functioning of society. Foundational work in calculus, probability theory, and statistics in the 17th and 18th centuries provided an array of new tools used by scientists to more precisely predict the movements of the sun and stars and determine population-wide rates of crime, marriage, and suicide. These tools often led to stunning advances. In the 1800s, Dr. John Snow used early modern data science to map cholera “clusters” in London. By tracing to a contaminated public well a disease that was widely thought to be caused by “miasmatic” air, Snow helped lay the foundation for the germ theory of disease. 1

Gleaning insights from data to boost economic activity also took hold in American industry. Frederick Winslow Taylor’s use of a stopwatch and a clipboard to analyze productivity at Midvale Steel Works in Pennsylvania increased output on the shop floor and fueled his belief that data science could revolutionize every aspect of life. 2 In 1911, Taylor wrote The Principles of Scientific Management to answer President Theodore Roosevelt’s call for increasing “national efficiency”:

> [T]he fundamental principles of scientific management are applicable to all kinds of human activities, from our simplest individual acts to the work of our great corporations…. [W]henever these principles are correctly applied, results must follow which are truly astounding. 3

Today, data is more deeply woven into the fabric of our lives than ever before. We aspire to use data to solve problems, improve well-being, and generate economic prosperity. The collection, storage, and analysis of data is on an upward
and seemingly unbounded trajectory, fueled by increases in processing power, the cratering costs of computation and storage, and the growing number of sensor technologies embedded in devices of all kinds. In 2011, some estimated the amount of information created and replicated would surpass 1.8 zettabytes. 4 In 2013, estimates reached 4 zettabytes of data generated worldwide. 5

What is a Zettabyte?

A zettabyte is 1,000,000,000,000,000,000,000 bytes, or units of information. Consider that a single byte equals one character of text. The 1,250 pages of Leo Tolstoy’s War and Peace would fit into a zettabyte 323 trillion times. 6 Or imagine that every person in the United States took a digital photo every second of every day for over a month. All of those photos put together would equal about one zettabyte.

More than 500 million photos are uploaded and shared every day, along with more than 200 hours of video every minute. But the volume of information that people create themselves—the full range of communications from voice calls, emails and texts to uploaded pictures, video, and music—pales in comparison to the amount of digital information created about them each day.

These trends will continue. We are only in the very nascent stage of the so-called “Inter-net of Things,” when our appliances, our vehicles and a growing set of “wearable” technologies will be able to communicate with each other. Technological advances have driven down the cost of creating, capturing, managing, and storing information to one-sixth of what it was in 2005. And since 2005, business investment in hardware, software, talent, and services has increased as much as 50 percent, to $4 trillion.

The “Internet of Things”

The “Internet of Things” is a term used to describe the ability of devices to communicate with each other using embedded sensors that are linked through wired and wireless networks. These devices could include your thermostat, your car, or a pill you swallow so the doctor can monitor the health of your digestive tract. These connected devices use the Internet to transmit, compile, and analyze data.

There are many definitions of “big data” which may differ depending on whether you are a computer scientist, a financial analyst, or an entrepreneur pitching an idea to a venture capitalist. Most definitions reflect the growing technological ability to capture, aggregate, and process an ever-greater volume, velocity, and variety of data. In other words, “data is now available faster, has greater coverage and scope, and includes new types of observations and measurements that previously were not available.” 7 More precisely, big datasets are “large, diverse, complex, longitudinal, and/or distributed datasets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future.” 8

What really matters about big data is what it does. Aside from how we define big data as a technological phenomenon, the wide variety of potential uses for big data analytics raises crucial questions about whether our legal, ethical, and social norms are sufficient to protect privacy and other values in a big data world. Unprecedented computational power and sophistication make possible unexpected discoveries, innovations, and advancements in our quality of life. But these capabilities, most of which are not visible or available to the average consumer, also create an asymmetry of power between those who hold the data and those who intentionally or inadvertently supply it.
Part of the challenge, too, lies in understanding the many different contexts in which big data comes into play. Big data may be viewed as property, as a public resource, or as an expression of individual identity. Big data applications may be the driver of America’s economic future or a threat to cherished liberties. Big data may be all of these things. For the purposes of this 90-day study, the review group does not purport to have all the answers to big data. Both the technology of big data and the industries that support it are constantly innovating and changing. Instead, the study focuses on asking the most important questions about the relationship between individuals and those who collect and use data about them.

The Scope of This Review

On January 17, in a speech at the Justice Department about reforming the United States’ signals intelligence practices, President Obama tasked his Counselor John Podesta with leading a comprehensive review of the impact big data technologies are having, and will have, on a range of economic, social, and government activities. Podesta was joined in this effort by Secretary of Commerce Penny Pritzker, Secretary of Energy Ernest Moniz, the President’s Science Advisor John Holdren, the President’s Economic Advisor Jeffrey Zients, and other senior government officials. The President’s Council of Advisors for Science & Technology conducted a parallel report to take measure of the underlying technologies. Their findings underpin many of the technological assertions in this report.

This review was conceived as fundamentally a scoping exercise. Over 90 days, the review group engaged with academic experts, industry representatives, privacy advocates, civil rights groups, law enforcement agents, and other government agencies. The White House Office of Science and Technology Policy jointly organized three university conferences, at the Massachusetts Institute of Technology, New York University, and the University of California, Berkeley. The White House Office of Science & Technology Policy also issued a “Request for Information” seeking public comment on issues of big data and privacy and received more than 70 responses. In addition, the WhiteHouse.gov platform was used to conduct an unscientific survey of public attitudes about different uses of big data and various big data technologies. A list of the working group’s activities can be found in the Appendix.

What is Different about Big Data?

This chapter begins by defining what is truly new and different about big data, drawing on the work of the President’s Council of Advisors on Science & Technology (PCAST), which has worked in parallel on a separate report, “Big Data and Privacy: A Technological Perspective.”

The “3 Vs”: Volume, Variety and Velocity

For purposes of this study, the review group focused on data that is so large in volume, so diverse in variety or moving with such velocity, that traditional modes of data capture and analysis are insufficient—characteristics colloquially referred to as the “3 Vs.” The declining cost of collection, storage, and processing of data, combined with new sources of data like sensors, cameras, geospatial and other observational technologies, means that we live in a world of near-ubiquitous data collection. The volume of data collected and processed is unprecedented. This explosion of data—from web-enabled appliances, wearable technology, and advanced sensors to monitor everything from vital signs to energy use to a jogger’s running speed—will drive demand for high-performance computing and push the capabilities of even the most sophisticated data management technologies.
There is not only more data, but it also comes from a wider variety of sources and formats. As described in the report by the President's Council of Advisors of Science & Technology, some data is “born digital,” meaning that it is created specifically for digital use by a computer or data processing system. Examples include email, web browsing, or GPS location. Other data is “born analog,” meaning that it emanates from the physical world, but increasingly can be converted into digital format. Examples of analog data include voice or visual information captured by phones, cameras or video recorders, or physical activity data, such as heart rate or perspiration monitored by wearable devices. With the rising capabilities of “data fusion,” which brings together disparate sources of data, big data can lead to some remarkable insights.

**What are the sources of big data?**

The sources and formats of data continue to grow in variety and complexity. A partial list of sources includes the public web; social media; mobile applications; federal, state and local records and databases; commercial databases that aggregate individual data from a spectrum of commercial transactions and public records; geospatial data; surveys; and traditional offline documents scanned by optical character recognition into electronic form. The advent of the more Internet-enabled devices and sensors expands the capacity to collect data from physical entities, including sensors and radio-frequency identification (RFID) chips. Personal location data can come from GPS chips, cell-tower triangulation of mobile devices, mapping of wireless networks, and in-person payments.

Furthermore, data collection and analysis is being conducted at a velocity that is increasingly approaching real time, which means there is a growing potential for big data analytics to have an immediate effect on a person’s surrounding environment or decisions being made about his or her life. Examples of high-velocity data include click-stream data that records users’ online activities as they interact with web pages, GPS data from mobile devices that tracks location in real time, and social media that is shared broadly. Customers and companies are increasingly demanding that this data be analyzed to benefit them instantly. Indeed, a mobile mapping application is essentially useless if it cannot immediately and accurately identify the phone’s location, and real-time processing is critical in the computer systems that ensure the safe operation of our cars.

**New Opportunities, New Challenges**

Big data technologies can derive value from large datasets in ways that were previously impossible—indeed, big data can generate insights that researchers didn’t even think to seek. But the technical capabilities of big data have reached a level of sophistication and pervasiveness that demands consideration of how best to balance the opportunities afforded by big data against the social and ethical questions these technologies raise.

**The power and opportunity of big data applications**

Used well, big data analysis can boost economic productivity, drive improved consumer and government services, thwart terrorists, and save lives. Examples include:

- Big data and the growing “Internet of Things” have made it possible to merge the industrial and information economies. Jet engines and delivery trucks can now be outfitted with sensors that monitor hundreds of data points and send automatic alerts when maintenance is needed. This makes repairs smoother, reducing maintenance costs and increasing safety.
The Centers for Medicare and Medicaid Services have begun using predictive analytics software to flag likely instances of reimbursement fraud before claims are paid. The Fraud Prevention System helps identify the highest risk health care providers for fraud, waste and abuse in real time, and has already stopped, pre-vented or identified $115 million in fraudulent payments—saving $3 for every $1 spent in the program’s first year. 14

During the most violent years of the war in Afghanistan, the Defense Advanced Research Projects Agency (DARPA) deployed teams of data scientists and visualizers to the battlefield. In a program called Nexus 7, these teams embedded directly with military units and used their tools to help commanders solve specific operational challenges. In one area, Nexus 7 engineers fused satellite and surveillance data to visualize how traffic flowed through road networks, making it easier to locate and destroy improvised explosive devices.

One big data study synthesized millions of data samples from monitors in a neo-natal intensive care unit to determine which newborns were likely to contract potentially fatal infections. By analyzing all of the data—not just what doctors noted on their rounds—the project was able to identify factors, like increases in temperature and heart rate, that serve as early warning signs that an infection may be taking root. These early signs of infection are not something even an experienced and attentive doctor would catch through traditional practices. 15

Big data technology also holds tremendous promise for better managing demand across electricity grids, improving energy efficiency, boosting agricultural productivity in the developing world, and projecting the spread of infectious diseases, among other applications.

Finding the needle in the haystack

Computational capabilities now make “finding a needle in a haystack” not only possible, but practical. In the past, searching large datasets required both rationally organized data and a specific research question, relying on choosing the right query to return the correct result. Big data analytics enable data scientists to amass lots of data, including unstructured data, and find anomalies or patterns. A key privacy challenge in this model of discovery is that in order to find the needle, you have to have a haystack. To obtain certain insights, you need a certain quantity of data.

For example, a genetic researcher at the Broad Institute found that having a large number of genetic datasets makes the critical difference in identifying the meaningful genetic variant for a disease. In this research, a genetic variant related to schizophrenia was not detectable when analyzed in 3,500 cases, and was only weakly identifiable using 10,000 cases, but was suddenly statistically significant with 35,000 cases. As the researcher observed, “There is an inflection point at which everything changes.” 16 The need for vast quantities of data—particularly personally sensitive data like genetic data—is a significant challenge for researchers for a variety of reasons, but notably because of privacy laws that limit access to data.

The data clusters and relationships revealed in large data sets can be unexpected but deliver incisive results. On the other hand, even with lots of data, the information revealed by big data analysis isn’t necessarily perfect. Identifying a pattern doesn’t establish whether that pattern is significant. Correlation still doesn’t equal causation. Finding a correlation with big data techniques may not be an appropriate basis for predicting outcomes or behavior, or rendering judgments on individuals. In big data, as with all data, interpretation is always important.

The benefits and consequences of perfect personalization

The fusion of many different kinds of data, processed in real time, has the power to de-liver exactly the right message, product, or service to consumers before they even ask. Small bits of data can be brought together to create a clear picture of a person to predict preferences or behaviors. These detailed personal profiles and personalized experiences
are effective in the consumer marketplace and can deliver products and offers to precise segments of the population—like a professional accountant with a passion for knitting, or a home chef with a penchant for horror films.

Unfortunately, “perfect personalization” also leaves room for subtle and not-so-subtle forms of discrimination in pricing, services, and opportunities. For example, one study found web searches involving black-identifying names (e.g., “Jermaine”) were more likely to display ads with the word “arrest” in them than searches with white-identifying names (e.g., “Geoffrey”). This research was not able to determine exactly why a racially biased result occurred, recognizing that ad display is algorithmically generated based on a number of variables and decision processes. But it’s clear that outcomes like these, by serving up different kinds of information to different groups, have the potential to cause real harm to individuals, whether they are pursuing a job, purchasing a home, or simply searching for information.

Another concern is that big data technology could assign people to ideologically or culturally segregated enclaves known as “filter bubbles” that effectively prevent them from encountering information that challenges their biases or assumptions. Extensive profiles about individuals and their preferences are being painstakingly developed by companies that acquire and process increasing amounts of data. Public awareness of the scope and scale of these activities is limited, however, and consumers have few opportunities to control the collection, use, and re-use of these data profiles.

De-identification and re-identification

As techniques like data fusion make big data analytics more powerful, the challenges to current expectations of privacy grow more serious. When data is initially linked to an individual or device, some privacy-protective technology seeks to remove this linkage, or “de-identify” personally identifiable information—but equally effective techniques exist to pull the pieces back together through “re-identification.” Similarly, integrating diverse data can lead to what some analysts call the “mosaic effect,” whereby personally identifiable information can be derived or inferred from datasets that do not even include personal identifiers, bringing into focus a picture of who an individual is and what he or she likes.

Many technologists are of the view that de-identification of data as a means of protecting individual privacy is, at best, a limited proposition. In practice, data collected and de-identified is protected in this form by companies’ commitments to not re-identify the data and by security measures put in place to ensure those protections. Encrypting data, removing unique identifiers, perturbing data so it no longer identifies individuals, or giving users more say over how their data is used through personal profiles or controls are some of the current technological solutions. But meaningful de-identification may strip the data of both its usefulness and the ability to ensure its provenance and accountability. Moreover, it is difficult to predict how technologies to re-identify seemingly anonymized data may evolve. This creates substantial uncertainty about how an individual controls his or her own information and identity, and how he or she disputes decision-making based on data derived from multiple datasets.

The persistence of data

In the past, retaining physical control over one’s personal information was often sufficient to ensure privacy. Documents could be destroyed, conversations forgotten, and records expunged. But in the digital world, information can be captured, copied, shared, and transferred at high fidelity and retained indefinitely. Volumes of data that were once unthinkably expensive to preserve are now easy and affordable to store on a chip the size of a grain of rice. As a
consequence, data, once created, is in many cases effectively permanent. Furthermore, digital data often concerns multiple people, making personal control impractical. For example, who owns a photo—the photographer, the people represented in the image, the person who first posted it, or the site to which it was posted? The spread of these new technologies are fundamentally changing the relationship between a person and the data about him or her.

Certainly data is freely shared and duplicated more than ever before. The specific responsibilities of individuals, government, corporations, and the network of friends, partners, and other third parties who may come into possession of personal data have yet to be worked out. The technological trajectory, however, is clear: more and more data will be generated about individuals and will persist under the control of others. Ensuring that data is secure is a matter of the utmost importance. For that reason, models for public-private cooperation, like the Administration’s Cybersecurity Framework, launched in February 2014, are a critical part of ensuring the security and resiliency of the critical infrastructure supporting much of the world’s data assets. 20

Affirming our Values

No matter how serious and consequential the questions posed by big data, this Administration remains committed to supporting the digital economy and the free flow of data that drives its innovation. The march of technology always raises questions about how to adapt our privacy and social values in response. The United States has met this challenge through considered debate in the public sphere, in the halls of Congress, and in the courts—and throughout its history has consistently been able to realize the rights enshrined in the Constitution, even as technology changes.

Since the earliest days of President Obama’s first term, this Administration has called on both the public and private sector to harness the power of data in ways that boost productivity, improve lives, and serve communities. That said, this study is about more than the capabilities of big data technologies. It is also about how big data may challenge fundamental American values and existing legal frameworks. This report focuses on the federal government’s role in assuring that our values endure and our laws evolve as big data technologies change the landscape for consumers and citizens.

In the last year, the public debate on privacy has largely focused on how government, particularly the intelligence community, collects, stores, and uses data. This report largely leaves issues raised by the use of big data in signals intelligence to be addressed through the policy guidance that the President announced in January. However, this report considers many of the other ways government collects and uses large datasets for the public good. Public trust is required for the proper functioning of government, and governments must be held to a higher standard for the collection and use of personal data than private actors. As President Obama has unequivocally stated, “It is not enough for leaders to say: trust us, we won’t abuse the data we collect.” 21

Recognizing that big data technologies are used far beyond the intelligence community, this report has taken a broad view of the issues implicated by big data. These new technologies do not only test individual privacy, whether defined as the right to be let alone, the right to control one’s identity, or some other variation. Some of the most profound challenges revealed during this review concern how big data analytics may lead to disparate inequitable treatment, particularly of disadvantaged groups, or create such an opaque decision-making environment that individual autonomy is lost in an impenetrable set of algorithms.

These are not unsolvable problems, but they merit deep and serious consideration. The historian Melvin Kranzberg’s First Law of Technology is important to keep in mind: “Technology is neither good nor bad; nor is it neutral.” 22
Technology can be used for the public good, but so too can it be used for individual harm. Regardless of technological advances, the American public retains the power to structure the policies and laws that govern the use of new technologies in a way that protects foundational values.

Big data is changing the world. But it is not changing Americans’ belief in the value of protecting personal privacy, of ensuring fairness, or of preventing discrimination. This report aims to encourage the use of data to advance social good, particularly where markets and existing institutions do not otherwise support such progress, while at the same time supporting frameworks, structures, and research that help protect our core values.

II. The Obama Administration's Approach to Open Data and Privacy

Throughout American history, technology and privacy laws have evolved in tandem. The United States has long been a leader in protecting individual privacy while supporting an environment of innovation and economic prosperity.

The Fourth Amendment to the Constitution protects the “right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures.” Flowing from this protection of physical spaces and tangible assets is a broader sense of respect for security and dignity that is indispensable both to personal well-being and to the functioning of democratic society. A legal framework for the protection of privacy interests has grown up in the United States that includes constitutional, federal, state, and common law elements. “Privacy” is thus not a narrow concept, but instead addresses a range of concerns reflecting different types of intrusion into a person’s sense of self, each requiring different protections.

Data collection—and the use of data to serve the public good—has an equally long history in the United States. Article I, Section 2 of the Constitution mandates a decennial Census in order to apportion the House of Representatives. In practice, the Census has never been conducted as just a simple head count, but has always been used to determine more specific demographic information for public purposes.

Since President Obama took office, the federal government has taken unprecedented steps to make more of its own data available to citizens, companies, and innovators. Since 2009, the Obama Administration has made tens of thousands of datasets public, hosting many of them on Data.gov, the central clearinghouse for U.S. government data. Treating government data as an asset and making it available, discoverable, and usable—in a word, open—strengthens democracy, drives economic opportunity, and improves citizens’ quality of life.

Deriving value from open data requires developing the tools to understand and analyze it. So the Obama Administration has also made significant investments in the basic science of data analytics, storage, encryption, cybersecurity, and computing power.

The Obama Administration has made these investments while also recognizing that the collection, use, and sharing of data pose serious challenges. Federal research dollars have supported work to address the technological and ethical issues that arise when handling large-scale data sets. Drawing on the United States’ long history of leadership on privacy issues, the Obama Administration also issued a groundbreaking consumer privacy blueprint in 2012 that included a Consumer Privacy Bill of Rights. In 2014, the President announced the Cybersecurity Framework, developed in partnership with the private sector, to strengthen the security of the nation’s critical infrastructure.
This chapter charts the intersections of these initiatives—ongoing efforts to harness data for the public good while ensuring the rights of citizens and consumers are protected.

Open Data in the Obama Administration

Open Data Initiatives

The smartphones we carry around in our pockets tell us where we are by drawing on open government data. Decades ago, the federal government first made meteorological data and the Global Positioning System freely available, enabling entrepreneurs to create a wide range of new tools and services, from weather apps to automobile navigation systems.

In the past, data collected by the government mostly stayed in the government agency that collected it. The Obama Administration has launched a series of Open Data Initiatives, each unleashing troves of valuable data that were previously hard to access, in domains including health, energy, climate, education, public safety, finance, and global development. Executive Order 13642, signed by President Obama on May 9, 2013, established an important new principle in federal stewardship of data: going forward, agencies must consider openness and machine-readability as the new defaults for government information, while appropriately safeguarding privacy, confidentiality, and security. Extending these open data efforts is also a core element of the President’s Second Term Management Agenda, and the Office of Management and Budget has directed agencies to release more of the administrative information they use to make decisions so it might be useful to others.

At Data.gov the public can find everything from data regarding complaints made to the federal Consumer Financial Protection Bureau about private student loans to 911 service area boundaries for the state of Arkansas. The idea is that anyone can use Da-ta.gov to find the open data they are looking for without having specialized knowledge of government agencies or programs within those agencies. Interested software developers can use simple tools to automatically access the datasets.

Federal agencies must also prioritize their data release efforts in part based on requests from the public. Each agency is required to solicit input through digital feedback mechanisms, like an email address or an online platform. For the first time, any advocate, entrepreneur, or researcher can connect with the federal government and suggest what data should be made available. To further improve feedback and encourage productive use of open government data, Administration officials have hosted and participated in a range of code-a-thons, brainstorming workshops (“Data Jams”), showcase events (“Datapaloozas”), and other meetings about open government data.

Pursuant to the May 2013 Executive Order, the Office of Management and Budget and the Office of Science and Technology Policy released a framework for agencies to manage information as an asset throughout its lifecycle, which includes requirements to continue to protect personal, sensitive, and confidential data. Agencies already categorize data assets into three access levels—public, restricted public, and non-public—and publish only the public catalog. To promote transparency, agencies include information in their external data inventories about technically public data assets that have not yet been posted online.
My Data Initiatives

Making public government data more open and machine-readable is only one element of the Administration’s approach to data. The Privacy Act of 1974 grants citizens certain rights of access to their personal information. That access should be easy, secure, and useful. Starting in 2010, the Obama Administration launched a series of My Data initiatives to empower Americans with secure access to their personal data and increase citizens’ access to private-sector applications and services that can be used to analyze it. The My Data initiatives include:

- **Blue Button**: The Blue Button allows consumers to securely access their health information so they can better manage their health care and finances and share their information with providers. In 2010, the U.S. Department of Veterans Affairs launched the Blue Button to give veterans the ability to download their health records. Since then, more than 5.4 million veterans have used the Blue Button tool to access their personal health information. More than 500 companies in the private sector have pledged their support to increase patient access to their health data by leveraging Blue Button, and today, more than 150 million Americans have the promise of being able to access their digital health information from health care providers, medical laboratories, retail pharmacy chains, and state immunization registries.

- **Get Transcript**: In 2014, the Internal Revenue Service made it possible for tax-payers to digitally access their last three years of tax information through a tool called Get Transcript. Individual taxpayers can use Get Transcript to download a record of past tax returns, which makes it easier to apply for mortgages, student loans, and business loans, or to prepare future tax filings.

- **Green Button**: The Administration partnered with electric utilities in 2012 to create the Green Button, which provides families and business with easy access to their energy usage information in a consumer-friendly and computer-friendly format. Today, 48 utilities and electricity suppliers serving more than 59 million homes and businesses have committed to giving their customers “Green Button” access to help them save energy. With customers in control of their energy data, they can choose which private sector tools and services can help them better manage their property’s energy efficiency.

- **MyStudentData**: The Department of Education makes it possible for students and borrowers to access and download their data from the Free Application for Federal Student Aid and their federal student loan information—including loan, grant, enrollment, and overpayment information. In both cases, the information is available via a user-friendly, machine-readable, plain-text file.

Beyond providing people with easy and secure access to their data, the My Data initiatives helps establish a strong model for personal data accessibility that the Administration hopes will become widely adopted in the private and public sectors. The ability to access one’s personal information will be increasingly important in the future, when more aspects of life will involve data transactions between individuals, companies, and institutions.

**Big Data Initiative: “Data to Knowledge to Action”**

At its core, big data is about being able to move quickly from data to knowledge to action. On March 29, 2012, six federal agencies joined forces to launch the “Big Data Research and Development Initiative,” with over $200 million in research funding to improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data.

Since the launch of this “Data to Knowledge to Action” initiative, DARPA has created an “Open Catalog” of the research publications and open source software generated by its $100 million XDATA program, an effort to process and analyze large sets of imperfect, incomplete data. The National Institutes of Health has supported a $50 million “Big Data to Knowledge” program about biomedical big data. The National Science Foundation has funded big data research projects which have reduced the cost of processing a human genome by a factor of 40. The Department of Energy...
announced a $25 million Scalable Data Management, Analysis, and Visualization Institute, which produced climate data techniques that have made seasonal hurricane predictions more than 25 per-cent more accurate. Many other research initiatives have important big data components, including the BRAIN Initiative, announced by President Obama in April 2013. As part of the Administration’s big data research initiative, the National Science Foundation has also funded specific projects examining the social, ethical, and policy aspects of big data.

U.S. Privacy Law and International Privacy Frameworks

Development of Privacy Law in the United States U.S. privacy laws have shaped and been shaped by societal changes, including the waves of technological innovation set in motion by the industrial revolution. The first portable cameras helped catalyze Samuel Warren and Louis Brandeis’s seminal 1890 article The Right to Privacy, in which they note that “[r]ecent inventions and business methods call attention to the next step which must be taken for the protection of the per-son, and for securing to the individual … the right ‘to be let alone’… numerous mechanical devices threaten to make good the prediction that what is whispered in the closet shall be proclaimed from the house-tops.” This prescient work laid the foundation for the common law of privacy in the 20th century, establishing citizens’ rights to privacy from the government and from each other.

Over the course of the last century, case law about what constitutes a “search” for purposes of the Fourth Amendment to the Constitution has developed with time and technology. In 1928, the U.S. Supreme Court held in Olmstead v. United States that placing wiretaps on a phone line located outside of a person's house did not violate the Fourth Amendment, even though the government obtained the content from discussions inside the home. But the Olmstead decision was arguably more famous for the dissent written by Justice Brandeis, who wrote that the Founders had “conferred, as against the government, the right to be let alone—the most comprehensive of rights and the right most favored by civilized men.”

The Court’s opinion in Olmstead remained the law of the land until it was overturned by the Court’s 1967 decision in Katz v. United States. In Katz, the Court held that the FBI’s placement of a recording device on the outside of a public telephone booth without a warrant qualified as a search that violated the “reasonable expectation of privacy” of the person using the booth, even though the device did not physically penetrate the booth, his person, or his property. Under Katz, an individual’s subjective expectations of privacy are protected when society regards them as reasonable.

Civil courts did not immediately acknowledge privacy as justification for one citizen to bring a lawsuit against another—what lawyers call a “cause of action.” It wasn’t until the 1934 Restatement (First) of Torts that an "unreasonable and serious" invasion of privacy was recognized as a basis to sue. Courts in most states began to recognize privacy as a cause of action, although what emerged from decisions was not a single tort, but in stead “a complex of four" potential torts:

1. Intrusion upon a person’s seclusion or solitude, or into his private affairs.
2. Public disclosure of embarrassing private facts about an individual.
3. Publicity placing one in a false light in the public eye.
4. Appropriation of one’s likeness for the advantage of another.
Some contemporary critics argue the “complex of four” does not sufficiently recognize privacy issues that arise from the extensive collection, use, and disclosure of personal information by businesses in the modern marketplace. Others suggest that automated processing should in fact ease privacy concerns because it uses computers operated under precise controls to perform tasks that used to be handled by a person.  

**The Fair Information Practice Principles**

As computing advanced and became more widely used by government and the private sector, policymakers around the world began to tackle the issue of privacy anew. In 1973, the U.S. Department of Health, Education, and Welfare issued a report entitled *Records, Computers, and the Rights of Citizens.* The report analyzed “harmful consequences that might result from automated personal data systems” and recommended certain safeguards for the use of information. Those safeguards, commonly known today as the “Fair Information Practice Principles,” or “FIPPs,” form the bedrock of modern data protection regimes.

While the principles are instantiated in law and international agreements in different ways, at their core, the FIPPs articulate basic protections for handling personal data. They provide that an individual has a right to know what data is collected about him or her and how it is used. The individual should further have a right to object to some uses and to correct inaccurate information. The organization that collects information has an obligation to ensure that the data is reliable and kept secure. These principles, in turn, served as the basis for the *Privacy Act of 1974,* which regulates the federal government’s maintenance, collection, use, and dissemination of personal information in systems of records.

By the late 1970s, several other countries had also passed national privacy laws. In 1980, the Organization for Economic Cooperation and Development (OECD) issued its “Guidelines Governing the Protection of Privacy and Transborder Flow of Personal Data.” Building on the FIPPs, the OECD guidelines have informed national privacy laws, sector-specific laws, and best practices for the past three decades. In 1981, the Council of Europe also completed work on the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108), which applied a FIPPs approach to emerging privacy concerns in Europe.

Despite some important differences, the privacy frameworks in the United States and those countries following the EU model are both based on the FIPPs. The European approach, which is based on a view that privacy is a fundamental human right, generally involves top-down regulation and the imposition of across-the-board rules restricting the use of data or requiring explicit consent for that use. The United States, in contrast, employs a sectoral approach that focuses on regulating specific risks of privacy harm in particular contexts, such as health care and credit. This places fewer broad rules on the use of data, allowing industry to be more innovative in its products and services, while also sometimes leaving unregulated potential uses of information that fall between sectors.

The FIPPs form a common thread through these sectoral laws and a variety of international agreements. They are woven into the 2004 Asia Pacific Economic Cooperation Privacy Principles, which was endorsed by APEC economies, and form the basis for the U.S.-E.U. and U.S.-Switzerland Safe Harbor Frameworks, which harness the global consensus around the FIPPs as a means to build bridges between U.S. and European law.
Sector-Specific Privacy Laws in the United States

In the United States during the 1970s and 80s, narrowly-tailored sectoral privacy laws began to supplement the tort-based body of common law. These sector-specific laws create privacy safeguards that apply only to specific types of entities and data. With a few exceptions, individual states and the federal government have predominantly enacted privacy laws on a sectoral basis. 48 The Fair Credit Reporting Act (FCRA) was originally enacted in 1970 to promote accuracy, fairness, and privacy protection with regard to the information assembled by consumer reporting agencies for use in credit and insurance reports, employee background checks, and tenant screenings. The law protects consumers by providing specific rights to access and correct their information. It requires companies that prepare consumer reports to ensure data is accurate and complete; limits when such reports may be used; and requires agencies to provide notice when an adverse action, such as the denial of credit, is taken based on the content of a report. The 1996 Health Insurance Portability and Accountability Act (HIPAA) addresses the use and disclosure of individuals’ health information by specified “covered entities” and includes standards designed to help individuals understand and control how their health information is used. 49 A key aspect of HIPAA is the principle of “minimum necessary” use and disclosure. 50 Congress and the Department of Health and Human Services have periodically updated protections for personal health data. The Children’s Online Privacy Protection Act of 1998 (COPPA) and the Federal Trade Commission’s implementing regulations require online services directed at children under the age of 13, or which collect personal data from children, to obtain verifiable parental consent to do so. In the financial sector, the Gramm-Leach-Bliley Act mandates that financial institutions respect the privacy of customers and the security and confidentiality of those customers’ nonpublic personal information. Other sectoral privacy laws safeguard individuals’ educational, communications, video rental, and genetic information. 51

Consumer Privacy Bill of Rights

In February 2012, the White House released a report titled Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy. 52 This “Privacy Blueprint” contains four key elements: a Consumer Privacy Bill of Rights based on the Fair Information Practice Principles; a call for government-convened multi-stakeholder processes to apply those principles in particular business contexts; support for effective enforcement of privacy rights, including the enactment of baseline consumer privacy legislation; and a commitment to international privacy regimes that support the flow of data across borders.

At the center of the Privacy Blueprint is the Consumer Privacy Bill of Rights, which states clear baseline protections for consumers. The rights are:

- **Individual Control**: Consumers have a right to exercise control over what personal data organizations collect from them and how they use it.
- **Transparency**: Consumers have a right to easily understandable information about privacy and security practices.
- **Respect for Context**: Consumers have a right to expect that organizations will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data.
- **Security**: Consumers have a right to secure and responsible handling of personal data.
- **Access and Accuracy**: Consumers have a right to access and correct personal data in usable formats, in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data are inaccurate.
- **Focused Collection**: Consumers have a right to reasonable limits on the personal data that companies collect and retain.
• **Accountability**: Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights.

The Consumer Privacy Bill of Rights is more focused on consumers than previous privacy frameworks, which were often couched in legal jargon. For example, it describes a right to "access and accuracy," which is more easily understood by users than previous formulations referencing "data quality and integrity." Similarly, it assures consumers that companies will respect the "context" in which data is collected and used, replacing the term "purpose specification."

The Consumer Privacy Bill of Rights also draws upon the Fair Information Practice Principles to better accommodate the online environment in which we all now live. Instead of requiring companies to adhere to a single, rigid set of requirements, the Consumer Privacy Bill of Rights establishes general principles that afford companies discretion in how they implement them. The Consumer Privacy Bill of Rights' "context" principle interacts with its other six principles, assuring consumers that their data will be collected and used in ways consistent with their expectations. At the same time, the context principle permits companies to develop new services using personal information when that use is consistent with the companies' relationship with its users and the circumstances surrounding how it collects data.

The Internet's complexity, global reach, and constant evolution require timely, scalable, and innovation-enabling policies. To answer this challenge, the Privacy Blueprint calls for all relevant stakeholders to come together to develop voluntary, enforceable codes of conduct that specify how the Consumer Privacy Bill of Rights applies in specific business contexts. The theory behind the Consumer Privacy Bill of Rights is that this combination of broad baseline principles and specific codes of conduct can protect consumers while supporting innovation.

**Promoting Global Interoperability**

The Obama Administration released the Consumer Privacy Bill of Rights as other countries and international organizations began to review their own privacy frameworks. In 2013, the OECD updated its Privacy Guidelines, which supplement the Fair Information Practice Principles with mechanisms to implement and enforce privacy protections. The APEC Cross Border Privacy Rules System, also announced in 2013, largely follows the OECD guidelines. The Council of Europe is undertaking a review of Convention 108. Building bridges among these different privacy frameworks is critical to ensuring robust international commerce.

The European Union is also in the process of reforming its data protection rules. The current E.U. Data Protection Directive only allows transfers of E.U. citizens' data to those non-E.U. countries with "adequate" privacy laws or mechanisms providing sufficient safeguards for data, such as the U.S.-E.U. Safe Harbor. In January 2014, the U.S. and E.U. began discussing how best to enhance the Safe Harbor Framework to ensure that it continues to provide strong data protection and enable trade through increased transparency, effective enforcement, and legal certainty. These negotiations continue, even as Europe—like the United States—wrestles with questions about how it will accommodate big data technologies and increased computational and storage capacities.

In March 2014, the Federal Trade Commission, together with agency officials from the European Union and Asia-Pacific Economic Cooperation economies, announced joint E.U. and APEC endorsement of a document that maps the requirements of the European and APEC privacy frameworks. The mapping project will help companies seeking certification to do business in both E.U. and APEC countries recognize overlaps and gaps between the two frameworks. Efforts like these clarify obligations for companies and help build interoperability between global privacy frameworks.
Conclusion

The most common privacy risks today still involve “small data”—the targeted compromise of, for instance, personal banking information for purposes of financial fraud. These risks do not involve especially large volumes, rapid velocities, or great varieties of information, nor do they implicate the kind of sophisticated analytics associated with big data. Protecting privacy of “small” data has been effectively addressed in the United States through the Fair Information Practice Principles, sector-specific laws, robust enforcement, and global privacy assurance mechanisms.

Privacy scholars, policymakers, and technologists are now turning to the question of how big data technology can be effectively managed under the FIPPs-based frameworks. The remainder of this report explores applications of big data in the public and private sector and then returns to consider the overall implications big data may have on current privacy frameworks.

III. Public Sector Management of Data

Government keeps the peace. It makes sure our food is safe to eat. It keeps our air and water clean. The laws and regulations it promulgates order economic and political life. Big data technology stands to improve nearly all the services the public sector delivers.

This chapter explores how big data is already helping the government carry out its obligations in health, education, homeland security, and law enforcement. It also begins to frame some of the challenges big data raises. Questions about what the government should and should not do, and how the rights of citizens should be protected in light of changing technology, are as old as the Republic itself. In framing the laws and norms of our young country, the founders took pains to demarcate private spheres shielded from inappropriate government interference. While many things about the big data world might astonish them, the founders would not be surprised to find that the Constitution and Bill of Rights are as central to the debate as Moore’s law and zettabytes.

At its core, public-sector use of big data heightens concerns about the balance of power between government and the individual. Once information about citizens is compiled for a defined purpose, the temptation to use it for other purposes can be considerable, especially in times of national emergency. One of the most shameful instances of the government misusing its own data dates to the Second World War. Census data collected under strict guarantees of confidentiality was used to identify neighborhoods where Japanese-Americans lived so they could be detained in internment camps for the duration of the war.

Because the government bears a special responsibility to protect its citizens when exercising power and authority for the public good, how big data should be put to use in the public sector, as well as what controls and limitations should apply, must be carefully considered. If unchecked, big data could be a tool that substantially expands government power over citizens. At the same time, big data can also be used to enhance accountability and to engineer systems that are inherently more respectful of privacy and civil rights.

Big Data and Health Care Delivery

Data has long been a part of health care delivery. In the past several years, legislation has created incentives for health care providers to transition to using electronic health records, vastly expanding the volume of health data available to clinicians, researchers, and patients. With the enactment of the Affordable Care Act, the model for health care
reimbursement is beginning to shift from paying for isolated and potentially uncoordinated instances of treatment—a model called “fee-for-service”—to paying on the basis of better health outcomes. Taken together, these trends are helping build a “learning” health care system where effective practices are identified from clinical data and then rapidly disseminated back to providers.

Big data can identify diet, exercise, preventive care, and other lifestyle factors that help keep people from having to seek care from a doctor. Big data analytics can also help identify clinical treatments, prescription drugs, and public health interventions that may not appear to be effective in smaller samples, across broad populations, or using traditional research methods. From a payment perspective, big data can be used to ensure professionals who treat patients have strong performance records and are reimbursed on the quality of patient outcomes rather than the quantity of care delivered.

The emerging practice of predictive medicine is the ultimate application of big data in health. This powerful technology peers deeply into a person’s health status and genetic information, allowing doctors to better predict whether individuals will develop a disease and how they might respond to specific therapies. Predictive medicine raises many complex issues. Traditionally, health data privacy policies have sought to protect the identity of individuals whose information is being shared and analyzed. But increasingly, data about groups or categories of people will be used to identify diseases prior to or very early after the onset of clinical symptoms.

But the information that stands to be discovered by predictive medicine extends beyond a single individual’s risks to include others with similar genes, potentially including the children and future descendants of those whose information is originally collected. Bio-repositories that link genomic data to health care data are on the leading edge of confronting important questions about personal privacy in the context of health research and treatment.

The privacy frameworks that currently cover information now used in health may not be well suited to address these developments or facilitate the research that drives them. Using big data to improve health requires advanced analytical models to ingest multiple kinds of lifestyle, genomic, medical, and financial data. The powerful connection between lifestyle and health outcomes means the distinction between personal data and health care data has begun to blur. These types of data are subjected to different and sometimes conflicting federal and state regulation, including the Health Insurance Portability and Accountability Act, Gramm-Leach-Bliley Act, Fair Credit Reporting Act, and Federal Trade Commission Act. The complexity of complying with numerous laws when data is combined from various sources raises the potential need to carve out special data use authorities for the health care industry if it is to realize the potential health gains and cost reductions that could come from big data analytics. At the same time, health organizations interact with many organizations that are not regulated under any of these laws. In the resulting ecosystem, personal health information of various kinds is shared with an array of firms, and even sold by state governments, in ways that might not accord with consumer expectations of the privacy of their medical data.

Though medicine is changing, information about our health remains a very private part of our lives. As big data enables ever more powerful discoveries, it will be important to re-visit how privacy is protected as information circulates among all the partners involved in care. Health care leaders have voiced the need for a broader trust framework to grant all health information, regardless of its source, some level of privacy protection. This may potentially involve crafting additional protections beyond those afforded in the Health Insurance Portability and Accountability Act and Genetic Information Non-Discrimination Act as well as streamlining data interoperability and compliance requirements. After studying health information technology, the President’s Council of Advisors on Science & Technology concluded that the
nation needs to adopt universal standards and an architecture that will facilitate controlled access to information across many different types of records. 60

Modernizing the health care data privacy framework will require careful negotiation between the many parties involved in delivering health care and insurance to Americans, but the potential economic and health benefits make it well worth the effort.

**Learning about Learning: Big Data and Education**

Education at both the K-12 and university levels is now supported inside and outside the classroom by a range of technologies that help foster and enhance the learning process. Students now access class materials, watch instructional videos, comment on class activities, collaborate with each other, complete homework, and take tests online. Technology-based educational tools and platforms offer important new capabilities for students and teachers. After only a few generations of evolution, these tools provide real-time assessment so that material can be presented based on how quickly a student learns. Education technologies can also be scaled to reach broad audiences, enable continuous improvement of course content, and increase engagement among students. 60

Beyond personalizing education, the availability of new types of data profoundly improves researchers’ ability to learn about learning. Data from a student’s experience in massive open online courses (MOOCs) or other technology-based learning platforms can be precisely tracked, opening the door to understanding how students move through a learning trajectory with greater fidelity, and at greater scale, than traditional education research is able to achieve. This includes gaining insight into student access of learning activities, measuring optimal practice periods for meeting different learning objectives, creating pathways through material for different learning approaches, and using that information to help students who are struggling in similar ways. Already, the Department of Education has studied how to harness these technologies, begun integrating the use of data from online education in the National Education Technology Plan, and laid plans for a Virtual Learning Lab to pioneer the methodological tools for this research. 62

The big data revolution in education also raises serious questions about how best to protect student privacy as technology reaches further into the classroom. While states and local communities have traditionally played the dominant role in providing education, much of the software that supports online learning tools and courses is provided by for-profit firms. This raises complicated questions about who owns the data streams coming off online education platforms and how they can be used. Applying privacy safeguards like the Family Educational Rights and Privacy Act, the Protection of Pupil Rights Amendment, or the Children’s Online Privacy Protection Act to educational records can create unique challenges.

**Protecting Children’s Privacy in the Era of Big Data**

Children today are among the first generation to grow up playing with digital devices even before they learn to read. In the United States, children and teenagers are active users of mobile apps and social media platforms. As they use these technologies, granular data about them—some of it sensitive—is stored and processed online. This data has the potential to dramatically improve learning outcomes and open new opportunities for children, but could be used to build an invasive consumer profile of them once they become adults, or otherwise pose problems later in their lives. Although youth on average are typically no less, and in many cases more, cognizant of commercial and government use of data than adults, they often face scrutiny by parents, teachers, college admissions officers, military recruiters, and case
workers. Vulnerable youth, including foster children and homeless youth, who typically have little adult guidance, are also particularly susceptible to data misuse and identity theft. Struggling to find some privacy in the face of tremendous supervision, many youth experiment with various ways to obscure the meaning of what they share except to select others, even if they are unable to limit access to the content itself. Because young people are exactly that—young—they need appropriate freedoms to explore and experiment safely and without the specter of being haunted by mistakes in the future. The Children’s Online Privacy Protection Act requires website operators and app developers to gain consent from a parent or guardian before collecting personal information from children under the age of 13. There is not yet a settled understanding of what harms, if any, are accruing to children and what additional policy frameworks may be needed to ensure that growing up with technology will be an asset rather than a liability.

Just as with health care, some of the information revealed when a user interacts with a digital education platform can be very personal, including aptitude for particular types of learning and performance relative to other students. It is even possible to discern whether students have learning disabilities or have trouble concentrating for long periods. What time of day and for how long students stay signed in to online tools reveals lifestyle habits. What should educational institutions do with this data to improve learning opportunities for students? How can students who use these platforms, especially those in K-12 education, be confident that their data is safe?

To help answer complicated questions about ownership and proper usage of data, the U.S. Department of Education released guidance for online education services in February 2014. This guidance makes clear that schools and districts can enter into agreements with third parties involving student data only so long as requirements under the Family Educational Rights and Privacy Act and Protection of Pupil Rights Amendment are met. As more online learning tools and services become available for kids, states and local governments are also watching these issues closely. Schools and districts can only share protected student information to further legitimate educational interests, and they must retain “direct control” over that information. Even with this new guidance, the question of how best to protect student privacy in a big data world must be an ongoing conversation.

The Administration is committed to vigorously pursuing these questions and will work through the Department of Education so all students can experience the benefits of big data innovations in teaching and learning while being protected from potential harms. As Secretary of Education Arne Duncan has said, “Student data must be secure, and treated as precious, no matter where it’s stored. It is not a commodity.” This means ensuring the personal information and online activity of students are protected from inappropriate uses, especially when it is gathered in an educational context.

**Big Data at the Department of Homeland Security**

Every day, two million passengers fly into, within, or over the United States. More than a million people enter the country by land. Verifying the identity of each person and determining whether he or she poses a threat falls to the Department of Homeland Security, which must process huge amounts of data in seconds to carry out its mission. The Department is not simply out to find the “needle in the haystack.” Protecting the homeland often depends on finding the most critical needles across many haystacks—a classic big data problem.

Ensuring the Department efficiently and lawfully uses the information it collects is a massive undertaking. DHS was created out of 22 separate government agencies in the wake of the 9/11 attacks. Many of the databases DHS operates today are physically disconnected, run legacy operating systems, and are unable to integrate information across...
different security classifications. The Department also carries out a diverse portfolio of missions, each governed by separate authorities in law. At all times, information must be used only for authorized purposes and in ways that protect the privacy and civil liberties afforded to U.S. citizens and foreign nationals who enter or reside in the United States. Ensuring information is properly used falls to six offices at DHS headquarters.

Beginning in 2012, representatives of the Chief Information Officer, the policy division, and the intelligence division came together with privacy, civil liberties and legal oversight officers to begin developing the first department-wide big data capability, resident in two pilot programs named Neptune and Cerberus. Neptune is designed from the ground up to be a “data lake” into which unclassified information from different sources flows. It has multiple built-in safeguards, including the ability to apply multiple data tags and fine-grained rules to determine which users can access which data for what purpose. All of the data is tagged according to a precise scheme. The rules governing usage focus on whether there is an authorized purpose, mission, or “need to know,” and whether the user has the appropriate job series and clearance to access the information. In this way, data tags can be combined with user attributes and context to govern what information is used where and by whom.

A Model for Managing Data

To build the tagging standards that govern information in its big data pilots, the Department of Homeland Security brought together the owners of the data systems, called data stewards, with representatives from privacy, civil liberties, and legal oversight offices. For each database field, the group charted its attributes and how access to the data is granted to different user communities. After developing a set of tags to encode this information, they then considered what additional rules and protections were needed to account for specific use limitations or special cases governed by law or regulation. Tagging both enables precise access control and preserves links to source data and the purpose of its original collection. The end result is a taxonomy of rules governing where information goes and tracking where it came from and under what authority. The fields in each database are grouped into three categories: core biographical data, such as name, date of birth, and citizenship status; extended biographical data, including addresses, phone number, and email; and detailed encounter data derived from electronic and in-person interactions with DHS. Encounter data is the most sensitive category. It may contain a law enforcement officer’s observations about an individual they interview as well as allegations of a risk to homeland security they may pose. These data tags then allow precise rules to be set of who can access what information for what reason. In these two pilots, the majority of rules for negotiating access are consistent across DHS’s different user communities. For example, many users will need access to the core biographic information of a particular data set to perform their missions. But some of the rules require far greater customization to account for specific use limitations.

The Neptune and Cerberus pilots also contain important controls around the types of searches that users are permitted to perform. A primary inspection agent may only need to perform a search on a specific person, because the agent is trying to confirm basic biographical information. However, an Immigration and Customs Investigator may need to perform person and characteristic searches while investigating a crime. DHS intelligence analysts may need to perform searches based on identities, characteristics, and trends when analyzing information related to a threat to homeland security. System administrators have no need to access the data contained within the system. The architecture of the database allows them to maintain the overall IT system but not to access any individual records.

The capabilities developed in these pilots are of a whole different order than the databases DHS inherited in 2002. Before these big data initiatives, it was not easy to perform searches across databases held by different components, let
alone to aggregate them. In the past, users and system administrators might have been issued a login and username and granted total access, sometimes without an audit trail monitoring their use. Now, DHS will be able to more precisely grant access according to mission needs. Most importantly, by being deliberate in tagging and organizing the data in these advanced repositories, the agency can take on new kinds of predictive and anomaly analysis while complying with the law and subjecting its activities to robust oversight.

It’s no accident that DHS was able to so carefully engineer how data is handled. DHS has both a dedicated Privacy Office and an Office for Civil Rights and Civil Liberties, each staffed with experts to help navigate this complex terrain. Each pilot is accompanied by a detailed privacy impact assessment released to the public in advance of its operation. DHS has provided public briefings on the pilots and allowed members of the public to ask questions about the initiatives. The privacy and civil liberties oversight officials not only approved the plan for the pilots, they also approve tools or widgets built in the future to increase their functionality. All of this helps drive improvements to DHS’s mission while ensuring that privacy and civil liberties concerns are considered from the start.

**Upholding our Privacy Values in Law Enforcement**

Big data can be a powerful tool for law enforcement. Recently, advanced web tools developed by DARPA’s Memex program have helped federal law enforcement make substantial progress in identifying human trafficking networks in the United States. These tools comb the “surface web” we all know, as well as “deep web” pages that are also public but not indexed by commonly used search engines. By allowing searches across a wide range of websites, the tools uncover a wealth of information that might otherwise be difficult or time-intensive to obtain. Possible trafficking rings can be identified and cross-referenced with existing law enforcement databases, helping police officers map connections between sex trafficking and other illegal activity. Already, the tools have helped detect trafficking networks originating in Asia and spreading to several U.S. cities. It’s a powerful example of how big data can help protect some of the most vulnerable people in the world.

Big data technologies provide effective tools to law enforcement and other agencies that protect our security, but they also pose difficult questions about their appropriate uses. Blending multiple data sources can create a fuller picture of a suspect’s activities around the time of a crime, but can also aid in the creation of suspect profiles that focus scrutiny on particular individuals with little or no human intervention. Pattern analysis can reveal how criminal organizations are structured or can be used to make predictions about possible future crimes. Gathering broad datasets can help catch criminals, but can also sweep up detailed personal information about people who are not subjects of an investigation. When it comes to law enforcement, we must be careful to ensure that big data technologies are used in ways that take into account the needs to protect public safety and fairly enforce the laws, as well as the civil liberties and legitimate privacy interests of citizens.

Big data will naturally—and appropriately—be used differently in national security. A powerful intelligence system that harnesses global data to identify terrorist networks, to provide warning of impending attacks, and to prevent the proliferation of weapons of mass destruction will operate under different legal authorities and oversight and have different privacy protections than a law enforcement system that helps allocate police resources to neighborhoods where higher levels of crime are predicted. Even though the applications are different, there are nevertheless important similarities in how privacy and civil rights are maintained across law enforcement and intelligence contexts. Privacy and legal officials must certify use of a system in each case, minimization rules are often employed to reduce information held, and data-tagging techniques are used to control access.
New Tools and New Challenges

The use of new technologies, especially in law enforcement, has given rise to important Constitutional jurisprudence. 71 As Justice Alito observed in a 2013 Supreme Court case concerning police placement of a GPS tracker on a suspect’s car without a court order: “[I]t is almost impossible to think of late-18th-century situations that are analogous to what took place in this case. (Is it possible to imagine a case in which a constable secreted himself somewhere in a coach and remained there for a period of time in order to monitor the movements of the coach’s owner?)” 72 Alito noted further, “Something like this might have occurred in 1791, but this would have required either a gigantic coach, a very tiny constable, or both.” 73 The “tiny constable” has enormous implications. Ubiquitous surveillance—whether by GPS tracking, closed circuit TV, or virtually undetectable sensors—will increasingly figure in litigation about reasonable expectations of privacy and the proper uses and limits of law enforcement technology.

In recent decades, the cost of surveillance and the physical size of surveillance equipment have rapidly decreased. This has made it feasible for over 70 cities in the United States to install audio sensors that can pinpoint gunfire and rapidly dispatch police to a potential crime scene. 74 Given the speed of access and decreasing cost of storage, it has likewise become practical for even local police forces to actively collect and catalog data, like license plate and vehicle information, in real-time on a city-wide scale, and to also retain it for later use. 75

The benefits of some of these technologies are tremendous. From finding missing persons to launching complex manhunts, the use of advanced surveillance technology by federal, state, and local law enforcement can mean a faster and more effective response to criminal activity. It can also increase the chances that justice is reliably served in online crime, where criminals are among the earliest adopters of new technologies and law enforcement needs to have timely access to digital evidence.

Beyond surveillance, predictive technologies offer the potential for law enforcement to be better prepared to anticipate, intervene in, or outright prevent certain crimes. Some analytics software, such as one program in use by both the Los Angeles and Memphis police departments, employs predictive analytics to identify geographically-based “hotspots.” 76 Many cities attribute meaningful declines in property crime to stepping up police patrols in “hotspot” areas.

Controversially, predictive analytics can now be applied to analyze a person’s individual propensity to criminal activity. 77 In response to an epidemic of gang-related murders, the city of Chicago conducted a pilot that shifts the focus of predictive policing from geographical factors to identity. By drawing on police and other data and applying social network analysis, the Chicago police department assembled a list of roughly 400 individuals identified by certain factors as likely to be involved in violent crime. As a result, police have a heightened awareness of particular individuals that might reflect factors beyond charges and convictions that are part of the public record. 78

Predictive analytics are also being used in other areas of criminal justice. In Philadelphia, police are using software designed to predict which parolees are more likely to commit a crime after release from prison and thus should have greater supervision. 79 The software uses about two dozen variables, including age, criminal history, and geographic location. These new techniques have come with considerable controversy about how and when they should be deployed. 80 This technology can help more precisely allocate law enforcement and other public resources, which can lead to the prevention of harmful crimes. At the same time, our Constitution and Bill of Rights grant certain rights that must not be abridged.
Police departments’ potential use of a new array of data and algorithms to try to predict criminal propensities and redirect police powers in advance of criminal activity has important consequences. It requires careful review of how we define “individualized suspicion,” which is the constitutional predicate of surveillance and search.  

The presence and persistence of authority, and the reasonable belief that one’s activities, movements, and personal affiliations are being monitored by law enforcement, can have a chilling effect on rights of free speech and association. The next section considers where changes in technology introduce tension within particular areas of the law.

**Implications of Big Data Technology for Privacy Law**

**Access to Data Held by Third Parties**

Personal documents and records have evolved from paper kept in the home, to electronic files held on the hard drive of a computer in the home, to many different kinds of computer files kept both locally and in cloud repositories accessed across multiple devices within and outside the home. As remote processing and cloud storage technologies increasingly become the norm for personal computing and records management, we must take measure of the how the law accounts for these developments.

Whether an individual reasonably expects an act to be private has framed much of our thinking about what protections are deserved. As Justice Potter Stewart in the 1967 Katz majority opinion noted: “[T]he Fourth Amendment protects people, not places. What a person knowingly exposes to the public, even in his own home or office, is not a subject of Fourth Amendment protection...But what he seeks to preserve as private, even in an area accessible to the public, may be constitutionally protected.”

Two later Supreme Court decisions further elaborated on how the Fourth Amendment applies to information that is shared with third parties. In *United States v. Miller*, in 1976, the Court found that the Fourth Amendment does not prohibit the government from obtaining “information revealed to a third-party and conveyed by him to government authorities, even if the information is revealed on the assumption that it will be used only for a limited purpose and the confidence placed in the third-party will not be betrayed.” Three years later, the Supreme Court held in *Smith v. Maryland* that the telephone numbers a person dials are not protected by a reasonable expectation of privacy because the caller voluntarily conveys dialing information to the phone company. The Court again affirmed that it had “consistently . . . held that a person has no legitimate expectation of privacy in information he voluntarily turns over to third parties.”

*Miller* and *Smith* are often cited as the Supreme Court’s foundational “third-party doctrine” cases. For decades, this doctrine has maintained that when an individual voluntarily shares information with third parties, like telephone companies, banks, or even other individuals, the government can acquire that information from the third-party absent a warrant without violating the individual’s Fourth Amendment rights. Law enforcement continues to rely on the third-party doctrine to obtain information that can be critical in criminal and national security investigations that keep the American people safe, and federal courts continue to apply the doctrine to both tangible and electronic information in a wide variety of contexts.

Against this backdrop, Congress and state legislatures have enacted statutes that provide additional safeguards for certain types of information, such as the Privacy Act of 1974 protecting personal information held by the federal government; the Electronic Communications Privacy Act of 1986 protecting (among other things) stored electronic
communications; and the Pen/Trap Act protecting (among other things) dialing information for phone calls. These legislative measures provide statutory protection in the absence of a strong Fourth Amendment right to protect records held by third parties.

In light of technological advances, especially the creation of exponentially more electronic records about personal interactions, some commentators have called for a reexamination of third-party doctrine. In 2010, the Sixth Circuit Court of Appeals in *United States v. Warshak* held that a subscriber has a reasonable expectation of privacy in his or her email communications, “analogous to a letter or a phone call” and that the government may not compel a commercial internet service provider to turn over the contents of a subscriber’s emails without first obtaining a warrant based on probable cause. In a re-cent Supreme Court case, Justice Sotomayor expressed the view in her concurring opinion that current practices around information disclosure to third parties are “ill-suited to the digital age, in which people reveal a great deal of information about themselves to third parties in the course of carrying out mundane tasks.”

Although we are not aware of any courts that have ruled that electronic content of communications can be accessed with less than a warrant, except with the consent of the user, since the Warshak case, the third-party doctrine has continued to apply to metadata of such communication and has been adapted and applied to cell-site location information and WiFi signals.

This review of big data and privacy has cast even more light on the profound issues of privacy, market confidence, and rule of law raised by the manner in which the government compels the disclosure of electronic data. We will continually need to examine our laws and policy to keep pace with technology, and should consider how the protection of content data stored remotely, for instance with a cloud provider, should relate to the protection of content data stored in a home office or on a hard drive. This is true of emails, text messages, and other communications platforms, which over the past 30 years have become an important means of private personal correspondence, and are most often stored remotely.

**Data and Metadata**

The average American transacts with businesses in one form or another multiple times a day, from purchasing goods to uploading digital photos. These interactions create records, some of which, like pharmacy purchases, contain intimate personal information. In the course of ordinary activities, users also emit lots of "digital exhaust," or trace data, that leaves behind more fragmentary bits of information, such as the geographical coordinates of a cell phone transmission or an IP address in a server log. The advent of more powerful analytics, which can discern quite a bit from even small and disconnected pieces of data, raises the possibility that data gathered and held by third parties can be amalgamated and analyzed in ways that reveal even more information about individuals. What protections this material and the information derived from it merit is now a pressing question.

An equally profound question is whether certain types of data—specifically the “metadata” or transactions records about communications and documents, versus the content of those communications and documents—should be accorded stronger privacy protections than they are currently. “Metadata” is a term describing the character of the data itself. The classic example comes from telecommunications. The phone numbers originating and terminating a call, as metadata, are considered less revealing than the conversation itself and have been accorded different privacy protections. Today, with the advent of big data, both the premise and policy may not always be so straightforward.
Experts seem divided on this issue, but those who argue that metadata today raises more sensitivities than in the past make a sufficiently compelling case to motivate review of policy on the matter. In the intelligence context, the President has already directed his Intelligence Advisory Board to consider the issue, and offer recommendations about the long-term viability of current assumptions about metadata and privacy. This review recommends that the government should broaden that examination beyond intelligence and consider the extent to which data and information should receive legal or other protections on the basis of how much it reveals about individuals.

**Government Use of Commercial Data Services**

Powerful private-sector profiling and data-mining technologies are not only used for commercial purposes. State, local, and federal agencies purchase access to many kinds of private databases for legitimate public uses, from land management to administering benefits. The sources of data that flow into these products are sometimes not publicly disclosed or may even be shielded as proprietary business information. Some legal scholars and privacy advocates have already raised concerns about the use of commercial data service products by the government, including law enforcement and intelligence agencies. 89

The Department of the Treasury has been working to implement a program to help prevent waste, fraud, and abuse in federal spending by reducing the number of payments made to the wrong person, for the wrong amount, or without the proper paperwork. To provide federal agencies with a “one-stop-shop” to check various databases and identify ineligible recipients or prevent fraud or errors, the Treasury launched a “Do Not Pay” portal. While all of the current databases available on the portal are government data-bases, Treasury anticipates that commercial databases may eventually be useful as well.

To assist the Treasury, the Office of Management and Budget issued substantial guidance to ensure that individual privacy is fully protected in the program. 90 The guidance recognized that commercial data sources “may also present new or increased privacy risks, such as databases with inaccurate or out-of-date information.” The guidelines require any commercial databases included in the Do Not Pay portal to be reviewed and approved following a 30-day period of public notice and comment. Among other requirements, the database must be relevant and necessary to the program, must be sufficiently accurate to ensure fairness to the individuals included in the database, and must not contain information that describes how any individual exercises rights guaranteed by the First Amendment, unless use of the data is expressly authorized by statute.

Given the increasing range of sensitive information available about individuals through commercial sources, this guidance is a significant step to ensure privacy protections when private-sector data is used to inform government decision-making. Similar OMB guidance should be considered for a wider range of agencies and programs, so the protections Americans have come to expect from their government exist regardless of where data originates.

**Insider Threat and Continuous Evaluation**

The 2013 shooting at the Washington Navy Yard facility by a contract employee who held a secret security clearance despite a record of arrests and troubling behavior has added urgency to ongoing efforts to more frequently evaluate employees who hold special positions of public trust. 91 It was the latest in a string of troubling breaches and acts of violence by insiders who held security clearances, including Chelsea Manning’s disclosures to WikiLeaks, the Fort Hood shooting by Major Nidal Hasan, and the most serious breach in the history of U.S. intelligence, the release of classified
National Security Agency documents by Edward Snowden. Federal government employees and contractors go through different levels of investigation, depending on the level of risk, sensitivity of their position, or their need to access sensitive facilities or systems. Currently, employees and contractors who hold “top secret” clearances are reinvestigated every five years, and those holding “secret” clearances every ten. These lengthy gaps do not allow agencies to discover new and note-worthy information about an employee in a timely manner.

Pilot programs have demonstrated the efficacy of using automated queries of appropriate official and commercial databases and social media to identify violations or irregularities, known as “derogatory information,” that may call into question a person’s suitability to continue serving in a sensitive position. The Department of Defense, for instance, recently conducted a pilot of what it calls the “Automated Continuous Evaluation System.” The pilot examined a sample of 3,370 Army service members, civilian employees, and contractor personnel, and identified that 21.7 percent of the tested population had previously unreported derogatory information that had developed since the last investigation. For 99 individuals, the pilot surfaced serious financial, domestic abuse, drug abuse, or allegations of prostitution that resulted in the revocation or suspension of their clearances. 92

The Administration recently released a review of suitability and security practices which called for expanding continuous evaluation capabilities across the federal government. 93 The Administration’s report recommends adopting practices across all agencies and security levels, although the exact extent of the information that will be used in these programs, especially social media sources, is still being determined.

These reforms will create a fundamentally different process for granting and maintaining security clearances that stands to enhance our security and safety. As the Administration works to expand the use of continuous evaluation across federal agencies, the privacy of employees and contractors will have to be carefully considered. The ability to refute or correct errant information that triggers reviews must be built into the process for appealing denials or revocations of clearance. We must ensure the big data analytics powering continuous evaluation are used in ways that protect the public as well as the civil liberties and privacy rights of those who serve on their behalf.

Conclusion

When wrestling with the vexing issues big data raises in the public sector, it can be easy to lose sight of the tremendous opportunities these technologies offer to improve public services, grow the economy, and improve the health and safety of our communities. These opportunities are real and must be kept at the center of the conversation about big data.

Big data holds enormous power to make the provision of services more efficient across the entire spectrum of government activity and to detect fraud, waste, and abuse at higher rates. Big data can also help create entirely new forms of value. New sources of precise data about weather patterns can provide meaningful scientific insights about climate change, while the ability to understand energy and natural resource use can lead to greater efficiency and reduce overall consumption. The movement, storage, and analysis of data all stands to grow more efficient and powerful. The Department of Energy, for instance, is working to develop computer memory and supercomputing frameworks that will in turn yield entire new classes of analytics tools, driving the big data revolution faster still.

There is virtually no part of government that does not stand to serve citizens better. The big data revolution will take hold across the entire government, not merely in departments and agencies that already have missions involving science and

https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review
Updated: Thu, 25 Jul 2019 08:57:25 GMT
Powered by mindtouch™
technology. Those departments and agencies that have not historically made wide use of advanced data analytics have perhaps the most significant opportunity to harness big data to benefit the citizens they serve.

The power of big data does not stop at the federal level. It will be equally transformational for states and municipalities. Cities and towns have emerged as some of the most innovative users of big data to improve service delivery. The federal agencies and pro-grams that provide grants and technical assistance to cities, towns, and counties should promote the use of these transformational municipal technologies to the greatest extent possible, replicating the successes pioneered by New York City’s Office of Data Analytics and Chicago’s Smart Data project.

Making big data work for the public good also takes people with skills that are in short supply and high demand. A recent assessment of the ability of the public and nonprofit sectors to attract and retain technical talent sounded a strong note of alarm. Though there are many young technologists who care deeply about public service and would welcome the chance to work in government, private sector opportunities are so comparatively attractive that these technologists tend to use their skills applying big data in the marketplace rather than the public sector. This means that alongside investments in technology, the federal government must create a more attractive working culture for technologists and remove hiring barriers that keep out the very experts whose creativity and technical imagination is paramount to realizing the full potential of big data in government.

IV. Private Sector Management of Data

Big data means big things all across the global economy. In the next two years, the big data technologies and services market is projected to continue its rapid ascent. This chapter considers how big data is shaping the products and services available to consumers and businesses, and highlights some of the challenges that arise when consumers have little insight into how information about them is being collected, analyzed, and used.

The Obama Administration has supported America’s leadership position in using big data to spark innovation, productivity, and value in the private sector. However, the near-continuous collection, transfer, and re-purposing of information in a big data world also raises important questions about individual control over personal data and the risks of its use to exploit vulnerable populations. While big data will be a powerful engine for economic growth and innovation, there remains the potential for a disquieting asymmetry between consumers and the companies that control information about them.

Big Data Benefits for Enterprise and Consumer

Big data is creating value for both companies and consumers. The benefits of big data can be felt across a range of sectors, in both large and small firms, as access to data and the tools for processing it are further democratized. In large enterprises, there are several drivers of investment in big data technologies: the ability to analyze operational and transactional data, to glean insights into the behavior of online customers, to bring new and exceedingly complex products to market, and to derive deeper understanding from machines and devices within organizations.

Technology companies are using big data to analyze millions of voice samples to deliver more reliable and accurate voice interfaces. Banks are using big data techniques to improve fraud detection. Health care providers are leveraging more detailed data to improve patient treatment. Big data is being used by manufacturers to improve warranty management and equipment monitoring, as well as to optimize the logistics of getting their products to market. Retailers
are harnessing a wide range of customer interactions, both online and offline, in order to provide more tailored recommendations and optimal pricing. 96

For consumers, big data is fueling an expansion of products and services that impact their daily lives. It is enabling cybersecurity experts to protect systems—from credit card readers to electricity grids—by harnessing vast amounts of network and application data and using it to identify anomalies and threats. 97 It is also enabling some of the nearly 29 percent of Americans who are "unbanked" or "underbanked" to qualify for a line of credit by using a wider range of non-traditional information—such as rent payments, utilities, mobile-phone subscriptions, insurance, child care, and tuition—to establish creditworthiness. 98

These new technologies are sensor-rich and embedded in networks. Lighting infrastructure can now detect sound, speed, temperature, and even carbon monoxide levels, and will draw data from car parks, schools, and along public streets to improve energy efficiency and public safety. Vehicles record and report a spectrum of driving and usage data that will pave the way for advanced transportation systems and improved safety. Home appliances can now tell us when to dim our lights from a thousand miles away. These are the kinds of changes that policies must accommodate. The Federal Trade Commission has already begun working to frame the policy questions raised by the Internet of Things, building on their long history of protecting consumers as new technologies come online.

The next sections discuss the online advertising and data services industries, each of which have significant histories using large datasets within long-established regulatory frameworks.

The Advertising-Supported Ecosystem

Since the earliest days of the commercial web, online advertising has been a vital driver of the growth of the Internet. One study estimated that the ad-supported Internet sustains millions of jobs in the United States and that the interactive marketing industry contributes billions to the U.S. economy each year. 99 This is a natural industry for big data to take root in and flourish. Increasingly precise data about consumers—where they are, what devices they use, and literally hundreds of categories of their interests—coupled with powerful analysis have enabled advertisers to more efficiently reach customers. Expensive television slots or full-page national magazine ads seem crude compared to the precisely segmented and instantaneously measured online ad marketplace. One study suggests that advertisers are willing to pay a premium of between 60 and 200 percent for online targeted advertising. 100

Consumers are reaping the benefits of a robust digital ecosystem that offers a broad array of free content, products, and services. The Internet also puts national or international advertising within reach of not just major companies, but mom-and-pop stores and fledgling brands. As a result, consumers are getting better, more useful ads from—and access to—a wider range of businesses, in a marketplace that is ultimately more competitive and innovative. Many different actors play a role in making this ecosystem work, including the consumer, the companies they engage with directly, and an array of other entities that provide services like analytics or security, or derive and share data. Standing between the publisher of the website a user visits and the advertiser paying for the ad displayed on the user’s page are a dizzying array of other companies. Advertising networks and ad exchanges facilitate transactions between the publishers and the advertisers. Ad content and campaigns are created and placed by agencies, optimizers, and media planners. Ad performance is measured and analyzed by yet another set of specialized companies. 101

In general, the companies with which a consumer engages directly—news websites, social media, or online or offline retailers—are called “first parties,” as they collect information directly from the consumer. But as described above, a
The Consumer and the Challenge of Transparency

For well over a decade, the online advertising industry has worked to provide consumers choice and transparency in a self-regulatory framework. Starting at the edges of the ecosystem, where the consumer can identify the website publisher and the advertiser whose ads are served, privacy policies and other forms of notice have served to inform consumers how their information is used. Under this self-regulatory regime, companies agree to a set of principles when engaged in “behavioral” or multisite advertising where they collect information about user activities over time and across different websites in order to infer user preferences. These principles include requiring notice to the user about their data collection practices; providing options for users to opt out of some forms of tracking; limiting the use of sensitive information, such as children’s information or medical or financial data; and a requirement to delete or de-identify data.

Technologies to improve transparency and privacy choices online have been slow to develop, and for many reasons have not been used widely by consumers. For example, under the self-regulatory regime adopted by advertisers and ad networks, many online behavioral ads include a standardized icon that indicates information is being collected for purposes of behavioral ad targeting, and links to a page where the consumer can opt-out of such collection. According to the online advertising industry, this icon has appeared on ads billions of times, but only a tiny fraction of users utilize this feature or understand its meaning. Advertising networks operated by some of the largest online companies have also offered users detailed dashboards for seeing the basis on which they are targeted for advertising and giving them the ability to opt out. These, too, have received little consumer attention. There are many theories about why users do not make use of these privacy features. Some assert that the privacy tools are hidden or too difficult for most users to navigate. Others argue that users have “privacy fatigue” from the barrage of privacy policies and settings they must wade through to simply use a service. It is also possible that most of the public is not very bothered by personalized ads when they enjoy a robust selection of free content, products, and services.

As we look ahead at the rising trajectory of information collection across many sources and the ability to target advertising with greater precision, the challenge to consumer transparency and meaningful choice deepens. Even employing relatively straightforward technical measures that would provide consumers with greater control over how data flows between their web browser and the servers of the webpages they visit for advertising purposes—what has become known as the “Do Not Track” browser setting—can be problematic because anti-fraud and online security activities now rely on these same data flows to track and prevent malicious activity.

The Challenge with Do Not Track

The idea behind a Do Not Track privacy setting is to provide an easy-to-use solution that empowers consumers to limit the tracking of their activities across websites. Some browsers provide a kind of Do Not Track capability by blocking
third-party cookies by default, or allowing consumers to choose to do so. Some browsers also allow consumers to send a signal instructing services not to track them. While Do Not Track technology is fairly straightforward, attempts to build consensus around the policy requirements for the websites receiving visits by users with Do Not Track technology enabled have proven far more difficult. Some websites voluntarily agreed to honor the wishes of visitors with Do Not Track indicators, but others have not, or have adopted policies that still permit partial tracking—muddling expectations for consumers and frustrating privacy advocates.

A working group of the World Wide Web Consortium, which included technologists, developers, advertising industry representatives, and privacy advocates, worked to craft a standard for implementation of the Do Not Track signal for more than three years. Recently, the working group released a final candidate for a technical Do Not Track specification, which will now go to the larger community to consider for approval.

In the meantime, the European Union amended its E-Privacy Directive in 2009 to require user consent to the use of cookies and other online tracking devices, unless they are “strictly necessary for delivery of a service requested by the user,” such as an online shopping cart. Compliance with the Directive has been uneven, although many European company websites now obtain a one-time explicit consent for the use of cookies—a solution that is widely acknowledged as clunky and which has been criticized in some circles as not providing the user the meaningful choice about privacy first envisioned by the directive.

While imperfect, these efforts reflect a growing interest in creating a technological means to allow individuals to control how commercial entities collect and use information about them.

**The Data Services Sector**

Alongside firms that focus primarily on online advertising are a related set of businesses that offer broader services drawn from information about consumers, public records, and other data sets. The “data services” sector—sometimes called “data brokers”—encompasses a class of businesses that collect data across many sources, aggregate and analyze it, and then share that information, or information derived from it. Typically, these companies have no direct relationship with the consumers whose information they collect. Instead, they offer services to other businesses or government agencies, including marketing products, verifying an individual’s identity, providing “people search” services, or detecting fraud. Some of these companies also have a specific line of business as “consumer reporting agencies,” which provide reports for purposes of credit applications, insurance, employment, or health care reports.

From a regulatory standpoint, data services fall into three broad categories:

1. Consumer reporting functions regulated under the Fair Credit Reporting Act, which generally keep the data, analysis, and reporting collected and used for these purposes in a separate system and under specific compliance rules apart from the rest of their data services operations.

2. Risk mitigation services such as identity verification, fraud detection and people-search or look up services; and

3. Marketing services to identify potential customers, enhance ad targeting information, and other advertising-related services.
The Fair Credit Reporting Act, as discussed in Chapter 2, provides affirmative rights to consumers. Consumer reporting agencies that provide reports for determining eligibility for credit, insurance, or employment, are required under the Fair Credit Reporting Act or the Equal Credit Opportunity Act to inform consumers when an adverse action, such as a denial or higher cost of credit, is taken against them based on a report. By law, consumers also have a right to know what is in their file, what their credit score is, and how to correct or delete inaccurate information. The Fair Credit Reporting Act mandates that credit reporting agencies remove negative information after certain periods, such that late payments and tax liens are deleted from a consumer’s file after seven years and bankruptcies after ten. Certain types of information—such as race, gender and religion—may not be used as factors to determine creditworthiness.

These statutory rights do not exist for risk mitigation or marketing services. As a matter of practice, data services companies may provide access and correction mechanisms to consumers for the information used in identity verification. In the context of marketing services, some companies permit consumers to opt-out of having their personal information used in marketing services.

Unregulated Data Broker Services

To assist marketers, data brokers can provide a profile of a consumer who may interact with a brand or seek services across many different channels, from online web presence to social media to mobile engagement. Data brokers aggregate purchase patterns, activities on a website, mobile, social media, ad network interactions, or direct customer support, and then further “enhance” it with information from public records or other commercially available sources. That information is used to develop a profile of a customer, whose activities or engagements can then be monitored to help the marketer pinpoint the message to send and the right moment to send it.

These profiles can be exceptionally detailed, containing upwards of thousands of pieces of data. Some large data firms have profiles on hundreds of millions of consumers. They algorithmically analyze this information to segment customers into precise categories, often with illustrative names that help their business customers identify populations for targeted advertising. Some of these categories include “Ethnic Second-City Strugglers,” “Retiring on Empty: Singles,” “Tough Start: Young Single Parents,” “Credit Crunched: City Families,” and “Rural and Barely Making It.” These products include factual information about individuals as well as “modeled” elements inferred from other data. Data brokers then sell “original lists” of consumers who fit particular criteria. They may also offer a “data append” service whereby companies can buy additional data about particular customers to help them build out more complete profiles of individuals on whom they maintain information.

What is a Credit Reporting Agency?

Since the 1950s, credit reporting companies—now known as “consumer reporting agencies”—have collected information and provided reports on individuals that are used to decide eligibility for credit, insurance or a job. In one typical scenario, a credit reporting agency collects information about an individual’s credit history, such as whether they pay their bills on time, how many and what kind of accounts they hold and for how long, whether they’ve been the subject of collection actions, and whether they have outstanding debt. The agency then uses a statistical program to compare this information to the loan repayment history of consumers with similar profiles and assigns a score that reflects the individual’s creditworthiness: how likely it is that he or she will repay a loan and make timely payments. This score
facilitates consumers’ ability to buy a home or car or otherwise engage in the economy by becoming a basis for creditors’ decisions about whether to provide credit to the consumer, and on what terms.

While this precise profiling of consumer attributes yields benefits, it also represents a powerful capacity on the part of the private sector to collect information and use that information to algorithmically profile an individual, possibly without the individual’s knowledge or consent. This application of big data technology, if used improperly, irresponsibly, or nefariously, could have significant ramifications for targeted individuals. In its 2012 Privacy Report, the Federal Trade Commission recommended that data brokers become more transparent in the services that are not already covered by the Fair Credit Report Act, and provide consumers with reasonable access to and choices about data maintained about them, in proportion to the sensitivity of data and how it is used. 109

Algorithms, Alternative Scoring and the Specter of Discrimination

The business models and big data strategies now being built around the collection and use of consumer data, particularly among the “third-party” data services companies, raise important questions about how to ensure transparency and accountability in these practices. Powerful algorithms can unlock value in the vast troves of information available to businesses, and can help empower consumers, but also raise the potential of encoding discrimination in automated decisions. Fueled by greater access to data and powerful analytics, there are now a host of products that “score” individuals beyond the scope of traditional credit scores, which are regulated by law. 110 These products attempt to statistically characterize everything from a consumer’s ability to pay to whether, on the basis of their social media posts, they are a “social influencer” or “socially influenced.”

While these scores may be generated for marketing purposes, they can also in practice be used similarly to regulated credit scores in ways that influence an individuals’ opportunities to find housing, forecast their job security, or estimate their health, outside of the protections of the Fair Credit Reporting Act or Equal Credit Opportunity Act. 111 Details on what types of data are included in these scores and the algorithms used for assigning attributes to an individual are held closely by companies and largely invisible to consumers. That means there is often no meaningful avenue for either identifying harms or holding any entity in the decision-making chain accountable.

Because of this lack of transparency and accountability, individuals have little recourse to understand or contest the information that has been gathered about them or what that data, after analysis, suggests. 112 Nor is there an industry-wide portal for consumers to communicate with data services companies, as the online advertising industry voluntarily provides and the Fair Credit Reporting Act requires for regulated entities. This can be particularly harmful to victims of identity theft who have ongoing errors or omissions impacting their scores and, as a result, their ability to engage in commerce.

What is an algorithm?

In simple terms, an algorithm is defined by a sequence of steps and instructions that can be applied to data. Algorithms generate categories for filtering information, operate on data, look for patterns and relationships, or generally assist in the analysis of information. The steps taken by an algorithm are informed by the author’s knowledge, motives, biases, and desired outcomes. The output of an algorithm may not reveal any of those elements, nor may it reveal the probability of a mistaken outcome, arbitrary choice, or the degree of uncertainty in the judgment it produces. So-called “learning algorithms” which underpin everything from recommendation engines to content filters evolve with the datasets
that run through them, assigning different weights to each variable. The final computer-generated product or decision—used for everything from predicting behavior to denying opportunity—can mask prejudices while maintaining a patina of scientific objectivity.

For all of these reasons, the civil rights community is concerned that such algorithmic decisions raise the specter of “redlining” in the digital economy—the potential to discriminate against the most vulnerable classes of our society under the guise of neutral algorithms. 113 Recently, some offline retailers were found to be using an algorithm that generated different discounts for the same product to people based on where they believed the customer was located. While it may be that the price differences were driven by the lack of competition in certain neighborhoods, in practice, people in higher-income areas received higher discounts than people in lower-income areas. 114

There are perfectly legitimate reasons to offer different prices for the same products in different places. But the ability to segment the population and to stratify consumer experiences so seamlessly as to be almost undetectable demands greater review, especially when it comes to the practice of differential pricing and other potentially discriminatory practices. It will also be important to examine how algorithmically-driven decisions might exacerbate existing socio-economic disparities beyond the pricing of goods and services, including in education and workforce settings.

Conclusion

The advertising-supported Internet creates enormous value for consumers by providing access to useful services, news, and entertainment at no financial cost. The ability to more precisely target advertisements is of enormous value to companies, which can efficiently reach audiences that are more likely to purchase their goods and services. However, private-sector uses of big data must ensure vulnerable classes are not unfairly targeted. The increasing use of algorithms to make eligibility decisions must be carefully monitored for potential discriminatory outcomes for disadvantaged groups, even absent discriminatory intent. The Federal Trade Commission should be commended for their continued engagement with industry and the public on this complex topic and should continue its plans to focus further attention on emerging practices in the data broker industry. We look forward to their forthcoming report on this important topic. Additional work should be done to identify practical ways of increasing consumer access to information about unregulated consumer scoring, with particular emphasis on the ability to correct or suppress inaccurate information. Likewise, additional research in measuring adverse outcomes due to the use of scores or algorithms is needed to understand the impacts these tools are having and will have in both the private and public sector as their use grows.

V. Toward a Policy Framework for Big Data

In what feels like the blink of an eye, the information age has fundamentally reconfigured how data affects individual lives and the broader economy. More than 6,000 data centers dot the globe. International data flows are continuous and multidirectional. To a greater degree than ever before, this data is being harnessed by businesses, governments, and entrepreneurs to improve the services they deliver and enhance how people live and work.

Big data applications create social and economic value on a scale that, collectively, is of strategic importance for the nation. Technological innovation is the animating force of the American economy. In the years to come, big data will foster significant productivity gains in industry and manufacturing, further accelerating the integration of the industrial and information economies.
Government should support the development of big data technologies with the full suite of policy instruments in its toolkit. Agencies must continue advancing the Administration’s Open Data initiative. The federal government should also invest in research and development to support big data technologies, especially as they apply to education, health care, and energy. As the preceding chapters have documented, adjusting existing policies will make possible certain new applications of big data that are clearly in the public interest, particularly in health care. The policy framework for big data will require cooperation between the public and private sectors to accelerate the revolution that is underway and identify barriers that ought to be removed for innovations driven by big data to flourish.

Like other transformative factors of production, big data generates value differently for individuals, organizations, and society. While many applications of big data are unequivocally beneficial, some of its uses impact privacy and other core values of fairness, equity, and autonomy.

Big data technologies enable data collection that is more ubiquitous, invasive, and valuable. This new cache of collected and derived data is of huge potential benefit but is also unevenly regulated. Certain private and public institutions have access to more data and more resources to compute it, potentially heightening asymmetries between institutions and individuals.

It is the responsibility of government to ensure that transformative technologies are used fairly and employed in all areas where they can achieve public good. Four areas in particular emerge as places for further policy exploration:

1. How government can harness big data for the public good while guarding against unacceptable uses against citizens;

2. The extent to which big data alters the consumer landscape in ways that implicate core values;

3. How to protect citizens from new forms of discrimination that may be enabled by big data technologies; and

4. How big data affects the core tenet of modern privacy protection, the notice and consent framework that has been in wide use since the 1970s.

**Big Data and the Citizen**

Big data will enhance how the government administers public services and enable it to create whole new kinds of value. But big data tools also unquestionably increase the potential of government power to accrue unchecked. Local police departments now have access to surveillance tools more powerful than those used by superpowers during the Cold War. The new means of surveillance that in Justice Alito’s evocative analogy deploy “tiny constables” to all areas of life, together with the ways citizens can be profiled by algorithms that redirect police powers, raise many questions about big data’s implications for First Amendment rights of free speech and free association.

Many of the laws governing law enforcement access to electronic information were passed by Congress at a time when private papers were largely stored in the home. The Stored Communications Act, which is part of the Electronic Communications Privacy Act (ECPA), articulates the rules for obtaining the content of electronic communications, including email and cloud services. ECPA was originally passed in 1986. It has served to protect the privacy of individuals’ stored communications. But with time, some of the lines drawn by the statute have become outdated and no longer reflect ways in which we use technology today. In considering how to update the Act, there are a variety of...
interests at stake, including privacy interests and the need for law enforcement and civil enforcement agencies to protect public safety and enforce criminal and civil law. Email, text messaging, and other private digital communications have become the principal means of personal correspondence and the cloud is increasingly used to store individuals’ files. They should receive commensurate protections.

Similarly, many protections afforded to metadata were calibrated for a time that predated the rise of personal computers, the Internet, mobile phones, and cloud computing. No one imagined then that the traces of digital data left today as a matter of routine can be reassembled to reveal intimate personal details. Today, most law enforcement uses of metadata are still rooted in the "small data" world, such as identifying phone numbers called by a criminal suspect. In the future, metadata that is part of the “big data” world will be increasingly relevant to investigations, raising the question of what protections it should be granted. While today, the content of communications, whether written or verbal, generally receives a high level of legal protection, the level of protection afforded to metadata is less so.

Although the use of big data technologies by the government raises profound issues of how government power should be regulated, big data technologies also hold within them solutions that can enhance accountability, privacy, and the rights of citizens. These include sophisticated methods of tagging data by the authorities under which it was collected or generated; purpose- and user-based access restrictions on this data; tracking which users access what data for what purpose; and algorithms that alert supervisors to possible abuses. All of these methods are being employed in parts of the federal government today to protect the rights of citizens and regulate how big data technologies are used, and more agencies should put them to use. Responsibly employed, big data could lead to an aggregate increase in actual protections for the civil liberties and civil rights afforded of citizens, as well as drive transformation improvements in the provision of public services.

**Big Data and the Consumer**

The technologies of collection and analysis that fuel big data are being used in every sector of society and the economy. Many of them are trained squarely on people as consumers. One of the most intensely discussed of big data analytics to date has been in the online advertising industry, where it is used to serve customized ads as people browse the web or travel around town with their mobile phone. But the information collected and the uses to which it is put are far broader and quickly changing, with data derived from the real world increasingly being combined with data drawn from online activity.

The end result is a massive increase in the amount of intimate information compiled about individuals. This information is highly valuable to businesses of all kinds. It is bought, bartered, traded, and sold. An entire industry now exists to commoditize the conclusions drawn from that data. Products sold on the market today include dozens of consumer scores on particular individuals that describe attributes, propensities, degrees of social influence over others, financial habits, household wealth, and even suitability as a tenant, job security, and frailty. While some of these scoring efforts are highly regulated, other uses of data are not.

There are enormous benefits associated with the rise of profiling and targeted advertising and the ways consumers can be tracked and offered services as they move through the online and physical world. Advertising and marketing effectively subsidize many free goods on the Internet, fueling an entire industry in software and consumer apps. As one person pointedly remarked during this review, “We don’t like putting a quarter into the machine to go do a web search.”
Data collection is also vital to securely verify identity online. The data services and financial industries have gone to extraordinary lengths to enable individuals to conduct secure transactions from computers and mobile devices. The same verification technologies that make transaction in the private sector possible also enable citizens to securely interact with the government online, opening a new universe of public services, all accessible from an arm chair.

But there are also costs to organizing the provision of commercial services in this way. Amalgamating so much information about consumers makes data breaches more consequential, highlighting the need for federal data breach legislation to replace a confusing patchwork of state standards. The sheer number of participants in this new, interconnected ecosystem of data collection, storage, aggregation, transfer, and sale can disadvantage consumers. The average consumer is unlikely to be aware of the range of data being collected or held or even to know who holds it; will have few opportunities to engage over the scope or accuracy of data being held about them; and may have limited insight into how this information feeds into algorithms that make decisions about their consumer experience or market access.

When considering what policies will allow big data to flourish in the consumer context, a crucial distinction must be drawn around the ways this collected information gets used. It is one thing for big data to segment consumers for marketing purposes, thereby providing more tailored opportunities to purchase goods and services. It is another, arguably far more serious, matter if this information comes to figure in decisions about a consumer’s eligibility for—or the conditions for the provision of—employment, housing, health care, credit, or education.

**Big Data and Discrimination**

In addition to creating tremendous social good, big data in the hands of government and the private sector can cause many kinds of harms. These harms range from tangible and material harms, such as financial loss, to less tangible harms, such as intrusion into private life and reputational damage. An important conclusion of this study is that big data technologies can cause societal harms beyond damages to privacy, such as discrimination against individuals and groups. This discrimination can be the inadvertent outcome of the way big data technologies are structured and used. It can also be the result of in-tent to prey on vulnerable classes.

An illustrative example of how one organization ensured that a big data technology did not inadvertently discriminate comes from Boston, where the city developed an experimental app in partnership with the Mayor’s Office of New Urban Mechanics. 115 Street Bump is a mobile application that uses a smartphone’s accelerometer and GPS feed to collect data about road condition, including potholes, and report them to the city’s Public Works Department. It is a marvelous example of how cities are creatively using crowdsourcing to improve service delivery. But the Street Bump team also identified a potential problem with deploying the app to the public. Because the poor and the elderly are less likely to carry smartphones or download the Street Bump app, its release could have the effect of systematically directing city services to wealthier neighborhoods populated by smartphone owners.

To its credit, the city of Boston and the StreetBump developers figured this out before launching the app. They first deployed it to city-road inspectors, who service all parts of the city equally; the public now provides additional supporting data. It took foresight to prevent an unequal outcome, and the results were worth it. The Street Bump app has to date recorded 36,992 “bumps,” helping Boston identify road castings like manholes and utility covers, not potholes, as the biggest obstacle for drivers.
More serious cases of potential discrimination occur when individuals interact with complex databases as they verify their identity. People who have multiple surnames and women who change their names when they marry typically encounter higher rates of error. This has also been true, for example, in the E-verify program, a database run jointly by the Department of Homeland Security and the Social Security Administration, which has long been a concern for civil rights advocates.

E-verify provides employers the ability to confirm the eligibility of newly hired employees to work legally in the United States. Especially given the number of queries the system processes and the volume of information it amalgamates from different sources that are themselves constantly changing, the overwhelming majority of results returned by E-verify are timely and accurate, giving employers certainty that people they hire are authorized to work in the United States. Periodic evaluations to improve the performance of E-verify have nonetheless revealed different groups receive initial verifications at different rates. A 2009 evaluation found the rate at which U.S. citizen have their authorization to work be initially erroneously unconfirmed by the system was 0.3 percent, compared to 2.1 percent for non-citizens. However, after a few days many of these workers’ status was confirmed. 116

The Department of Homeland Security and Social Security Administration have focused great attention on addressing this issue. A more recent evaluation of the program found many more people were able to verify their work status more quickly and with lower rates of error. Over five years, the rates of initial mismatch fell by 60 percent for U.S. citizens and 30 percent for non-citizens. 117 Left unresolved, technical issues like this could create higher barriers to employment or other critical needs for certain individuals and groups, making imperative the importance of accuracy, transparency, and redress in big data systems.

These two examples of inadvertent discrimination illustrate why it is important to monitor outcomes when big data technologies are applied even in instances where discriminatory intent is not present and where one might not anticipate an inequitable impact. There is, however, a whole other class that merits concern—the use of big data for deliberate discrimination.

We have taken considerable steps as a society to mandate fairness in specific domains, including employment, credit, insurance, health, housing, and education. Existing legislative and regulatory protections govern how personal data can be used in each of these contexts. Though predictive algorithms are permitted to be used in certain ways, the data that goes into them and the decisions made with their assistance are subject to some degree of transparency, correction, and means of redress. For important decisions like employment, credit, and insurance, consumers have a right to learn why a decision was made against them and what information was used to make it, and to correct the under-lying information if it is in error.

These protections exist because of the United States’ long history of discrimination. Since the early 20th century, banks and lenders have used location data to make assumptions about individuals. It was not until the Home Mortgage Disclosure Act was signed into law in 1975 that denying granting a person a loan on the basis of what neighborhood they live in rather than their personal capacity for credit became far less prevalent. “Redlining,” in which banks quite literally drew—and in cases continue to draw—boundaries around neighborhoods where they would not loan money, existed for decades as a potent tool of discrimination against African-Americans, Latinos, Asians, and Jews.
Just as neighborhoods can serve as a proxy for racial or ethnic identity, there are new worries that big data technologies could be used to “digitally redline” unwanted groups, either as customers, employees, tenants, or recipients of credit. A significant finding of this report is that big data could enable new forms of discrimination and predatory practices.

The same algorithmic and data mining technologies that enable discrimination could also help groups enforce their rights by identifying and empirically confirming instances of discrimination and characterizing the harms they caused. Civil rights groups can use the new and powerful tools of big data in service of equal treatment for the communities they represent. Whether big data will build greater equality for all Americans or exacerbate existing inequalities depends entirely on how its technologies are applied in the years to come, what kinds of protections are present in the law, and how the law is enforced.

**Big Data and Privacy**

Big data technologies, together with the sensors that ride on the “Internet of Things,” pierce many spaces that were previously private. Signals from home WiFi networks reveal how many people are in a room and where they are seated. Power consumption data collected from demand-response systems show when you move about your house. Facial recognition technologies can identify you in pictures online and as soon as you step outside. Always-on wearable technologies with voice and video interfaces and the arrival of whole classes of networked devices will only expand information collection still further. This sea of ubiquitous sensors, each of which has legitimate uses, make the notion of limiting information collection challenging, if not impossible.

This trend toward ubiquitous collection is in part driven by the nature of technology itself. Whether born analog or digital, data is being reused and combined with other data in ways never before thought possible, including for uses that go beyond the intent motivating initial collection. The potential future value of data is driving a digital land grab, shifting the priorities of organizations to collect and harness as much data as possible. Companies are now constantly looking at what kind of data they have and what data they need in order to maximize their market position. In a world where the cost of data storage has plummeted and future innovation remains unpredictable, the logic of collecting as much data as possible is strong.

Another reality of big data is that once data is collected, it can be very difficult to keep anonymous. While there are promising research efforts underway to obscure personally identifiable information within large data sets, far more advanced efforts are presently in use to re-identify seemingly “anonymous” data. Collective investment in the capability to fuse data is many times greater than investment in technologies that will enhance privacy.

Together, these trends may require us to look closely at the notice and consent framework that has been a central pillar of how privacy practices have been organized for more than four decades. In a technological context of structural over-collection, in which re-identification is becoming more powerful than de-identification, focusing on controlling the collection and retention of personal data, while important, may no longer be sufficient to protect personal privacy. In the words of the President’s Council of Advisors for Science & Technology, “The notice and consent is defeated by exactly the positive benefits that big data enables: new, non-obvious, unexpectedly powerful uses of data.”
Federal Research in Privacy-Enhancing Technologies

The research and development of privacy enhancing technologies has been a priority for the Obama Administration. Agencies across the Networking and Information Technology Research and Development (NITRD) program collectively spend over $70 million each year on privacy research. This research falls into four broad areas: support for privacy as an extension of security; research on how enterprises comply with privacy laws; privacy in health care; and basic research into technologies that enable privacy. The table below summarizes some of the research programs in progress at agencies in the NITRD. In their review of big data technologies, the President’s Council of Advisors on Science & Technology endorses strengthening U.S. research in privacy-related technologies and the social science questions surrounding their use.

<table>
<thead>
<tr>
<th>Research areas</th>
<th>Support for privacy as an extension of security</th>
<th>Research on how enterprises comply with privacy laws</th>
<th>Privacy in health care</th>
<th>Privacy research explorations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Funding est. (total $77M/year)</td>
<td>$34M/year</td>
<td>$10M/year</td>
<td>$8M/year</td>
<td>$25M/year</td>
</tr>
<tr>
<td>Sampling of key projects</td>
<td>Anonymization techniques</td>
<td>Automated privacy compliance</td>
<td>Collection and use limitation</td>
<td>Algorithmic foundations for privacy and tools</td>
</tr>
<tr>
<td></td>
<td>Confidential collaboration and communication</td>
<td>Location-privacy tools</td>
<td>Data segmentation for privacy</td>
<td>Economics of privacy</td>
</tr>
<tr>
<td></td>
<td>Homomorphic encryption</td>
<td>Protection of personally identifiable information</td>
<td>Patient consent and privacy</td>
<td>Privacy as a social-psychological construct</td>
</tr>
<tr>
<td></td>
<td>Privacy preserving data aggregation</td>
<td>Standards for legal compliance</td>
<td>Patient data quality</td>
<td>Privacy policy analysis</td>
</tr>
<tr>
<td></td>
<td>Traffic-secure routing</td>
<td>Voluntary code of conduct for smart grid</td>
<td>Preserving anonymity in health care data</td>
<td>Privacy solutions for cloud computing,</td>
</tr>
</tbody>
</table>

https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review
Updated: Thu, 25 Jul 2019 08:57:25 GMT
Powered by mindtouch™ 145
Anticipating the Big Data Revolution’s Next Chapter

For the vast majority of today’s ordinary interactions between consumers and first parties, the notice and consent framework adequately safeguards privacy protections. But as the President’s Council of Advisors on Science & Technology note, the trajectory of technology is shifting to far more collection, use and storage of data by entities that do not have a direct relationship with the consumer or individual. In instances where the notice and consent framework threatens to be overcome—such as the collection of ambient data by our household appliances—we may need to refocus our attention on the context of data use, a policy shift presently being debated by privacy scholars and technologists. The context of data use matters tremendously. Data that is socially beneficial in one scenario can cause significant harm in another. To borrow a term, data itself is “dual use.” It can be used for good or for ill.

Putting greater emphasis on a responsible use framework has many potential advantages. It shifts the responsibility from the individual, who is not well equipped to understand or contest consent notices as they are currently structured in the marketplace, to the entities that collect, maintain, and use data. Focusing on responsible use also holds data collectors and users accountable for how they manage the data and any harms it causes, rather than narrowly defining their responsibility to whether they properly obtained consent at the time of collection.

Focusing more attention on responsible use does not mean ignoring the context of collection. Part of using data responsibly could mean respecting the circumstances of its original collection. There could, in effect, be a "no surprises" rule, as articulated in the "respect for context" principle in the Consumer Privacy Bill of Rights. Data collected in a consumer context could not suddenly be used in an employment one. Technological developments support this shift toward a focus on use. Advanced data-tagging schemes can encode details about the context of collection and uses of the data already granted by the user, so that information about permissive uses travels along with the data wherever it goes. If well developed and brought widely into use, such a data-tagging scheme would not solve all the dilemmas posed by big data, but it could help address several important challenges.

Perhaps most important of all, a shift to focus on responsible uses in the big data context allows us to put our attention more squarely on the hard questions we must reckon with: how to balance the socially beneficial uses of big data with the harms to privacy and other values that can result in a world where more data is inevitably collected about more things. Should there be an agreed-upon taxonomy that distinguishes information that you do not collect or use under any circumstances, information that you can collect or use without obtaining consent, and information that you collect and use only with consent? How should this taxonomy be different for a medical researcher trying to cure cancer and a marketer targeting ads for consumer products?

As President Obama said upon the release of the Consumer Privacy Bill of Rights, “Even though we live in a world in which we share personal information more freely than in the past, we must reject the conclusion that privacy is an outmoded value.” Privacy, the President said, “has been at the heart of our democracy from its inception, and we need it now more than ever.” This is even truer in a world powered by big data.
VI. Conclusion and Recommendations

The White House review of big data and privacy, announced by President Obama on January 17, 2014, was conceived to examine the broader implications of big data technology. The President recognized the big data revolution is playing out widely across the public and private sectors and that its implications need to be considered alongside the Administration’s review of signals intelligence.

The White House big data working group set out to learn, in 90 days, how big data technologies are transforming government, commerce, and society. We wanted to understand what opportunities big data affords us, and the advances it can spur. We wanted a better grasp of what kinds of technologies already existed, and what we could anticipate coming just over the horizon. The President’s Council of Advisors for Science & Technology conducted a parallel report to take measure of the underlying technologies. Their findings underpin many of the technological assertions in this report.

Big data tools offer astonishing and powerful opportunities to unlock previously inaccessible insights from new and existing data sets. Big data can fuel developments and discoveries in health care and education, in agriculture and energy use, and in how businesses organize their supply chains and monitor their equipment. Big data holds the potential to streamline the provision of public services, increase the efficient use of taxpayer dollars at every level of government, and substantially strengthen national security. The promise of big data requires government data be viewed as a national resource and be responsibly made available to those who can derive social value from it. It also presents the opportunity to shape the next generation of computational tools and technologies that will in turn drive further innovation.

Big data also introduces many quandaries. By their very nature, many of the sensor technologies deployed on our phones and in our homes, offices, and on lampposts and rooftops across our cities are collecting more and more information. Continuing advances in analytics provide incentives to collect as much data as possible not only for today’s uses but also for potential later uses. Technologically speaking, this is driving data collection to become functionally ubiquitous and permanent, allowing the digital traces we leave behind to be collected, analyzed, and assembled to reveal a surprising number of things about ourselves and our lives. These developments challenge longstanding notions of privacy and raise questions about the “notice and consent” framework, by which a user gives initial permission for their data to be collected. But these trends need not prevent creating ways for people to participate in the treatment and management of their information. An important finding of this review is that while big data can be used for great social good, it can also be used in ways that perpetrate social harms or render outcomes that have inequitable impacts, even when discrimination is not intended. Small biases have the potential to become cumulative, affecting a wide range of outcomes for certain disadvantaged groups. Society must take steps to guard against these potential harms by ensuring power is appropriately balanced between individuals and institutions, whether between citizen and government, consumer and firm, or employee and business.

The big data revolution is in its earliest stages. We will be grappling for many years to understand the full sweep of its technologies; the ways it will empower health, education, and the economy; and, crucially, what its implications are for core American values, including privacy, fairness, non-discrimination, and self-determination.

Even at this early juncture, the authors of this report believe important conclusions are already emerging about big data that can inform how the Administration moves forward in a number of areas. In particular, there are five areas that will
each bring the American people into the national conversation about how to maximize benefits and minimize harms in a big data world:

1. **Preserving Privacy Values**: Maintaining our privacy values by protecting personal information in the marketplace, both in the United States and through interoperable global privacy frameworks;

2. **Educating Robustly and Responsibly**: Recognizing schools—particularly K-12—as an important sphere for using big data to enhance learning opportunities, while protecting personal data usage and building digital literacy and skills;

3. **Big Data and Discrimination**: Preventing new modes of discrimination that some uses of big data may enable;

4. **Law Enforcement and Security**: Ensuring big data's responsible use in law enforcement, public safety, and national security; and

5. **Data as a Public Resource**: Harnessing data as a public resource, using it to improve the delivery of public services, and investing in research and technology that will further power the big data revolution.

**Policy Recommendations**

This review also identifies six discrete policy recommendations that deserve prompt Administration attention and policy development. These are:

- **Advance the Consumer Privacy Bill of Rights**. The Department of Commerce should take appropriate consultative steps to seek stakeholder and public comment on big data developments and how they impact the Consumer Privacy Bill of Rights and then devise draft legislative text for consideration by stakeholders and submission by the President to Congress.

- **Pass National Data Breach Legislation**. Congress should pass legislation that provides for a single national data breach standard along the lines of the Administration’s May 2011 Cybersecurity legislative proposal.

- **Extend Privacy Protections to non-U.S. Persons**. The Office of Management and Budget should work with departments and agencies to apply the Privacy Act of 1974 to non-U.S. persons where practicable, or to establish alternative privacy policies that apply appropriate and meaningful protections to personal information regardless of a person’s nationality.

- **Ensure Data Collected on Students in School is Used for Educational Purposes**. The federal government must ensure that privacy regulations protect students against having their data being shared or used inappropriately, especially when the data is gathered in an educational context.

- **Expand Technical Expertise to Stop Discrimination**. The federal government’s lead civil rights and consumer protection agencies should expand their technical expertise to be able to identify practices and outcomes facilitated by big data analytics that have a discriminatory impact on protected classes, and develop a plan for investigating and resolving violations of law.

- **Amend the Electronic Communications Privacy Act**. Congress should amend ECPA to ensure the standard of protection for online, digital content is consistent with that afforded in the physical world—including by removing archaic distinctions between email left unread or over a certain age.
1. Preserving Privacy Values

Big data technologies are driving enormous innovation while raising novel privacy implications that extend far beyond the present focus on online advertising. These implications make urgent a broader national examination of the future of privacy protections, including the Administration’s Consumer Privacy Bill of Rights, released in 2012. It will be especially important to re-examine the traditional notice and consent framework that focuses on obtaining user permission prior to collecting data. While notice and consent remains fundamental in many contexts, it is now necessary to examine whether a greater focus on how data is used and reused would be a more productive basis for managing privacy rights in a big data environment. It may be that creating mechanisms for individuals to participate in the use and distribution of his or her information after it is collected is actually a better and more empowering way to allow people to access the benefits that derive from their information. Privacy protections must also evolve in a way that accommodates the social good that can come of big data use.

Advance the Consumer Privacy Bill of Rights

As President Obama made clear in February 2012, the Consumer Privacy Bill of Rights and the associated Blueprint for Consumer Privacy represent “a dynamic model of how to offer strong privacy protection and enable ongoing innovation in new information technologies.” The Consumer Privacy Bill of Rights is based on the Fair Information Practice Principles. Some privacy experts believe nuanced articulations of these principles are flexible enough to address and support new and emerging uses of data, including big data. Others, especially technologists, are less sure, as it is undeniable that big data challenges several of the key assumptions that underpin current privacy frameworks, especially around collection and use. These big data developments warrant consideration in the context of how to viably ensure privacy protection and what practical limits exist to the practice of notice and consent.

RECOMMENDATION: The Department of Commerce should promptly seek public comment on how the Consumer Privacy Bill of Rights could support the innovations of big data while at the same time responding to its risks, and how a responsible use framework, as articulated in Chapter 5, could be embraced within the framework established by the Consumer Privacy Bill of Rights. Following the comment process, the Department of Commerce should work on draft legislative text for consideration by stakeholders and for submission by the President to Congress.

Pass national data breach legislation to benefit consumers and businesses

As organizations store more information about individuals, Americans have a right to know if that information has been stolen or otherwise improperly exposed. A patchwork of 47 state laws currently governs when and how the loss of personally identifiable information must be reported.

RECOMMENDATION: Congress should pass legislation that provides for a single national data breach standard along the lines of the Administration’s May 2011 Cybersecurity legislative proposal. Such legislation should impose reasonable time periods for notification, minimize interference with law enforcement investigations, and potentially prioritize notification about large, damaging incidents over less significant incidents.

The data services industry—colloquially known as “data brokers”—should bring greater transparency to the sector
Consumers deserve more transparency about how their data is shared beyond the entities with which they do business directly, including “third-party” data collectors. This means ensuring that consumers are meaningfully aware of the spectrum of information collection and reuse as the number of firms that are involved in mediating their consumer experience or collecting information from them multiplies. The data services industry should follow the lead of the online advertising and credit industries and build a common website or online portal that lists companies, describes their data practices, and provides methods for consumers to better control how their information is collected and used or to opt-out of certain marketing uses.

Even as we focus more on data use, consumers still have a valid interest in “Do Not Track” tools that help them control when and how their data is collected.

Strengthening these tools is especially important because there is now a growing array of technologies available for recording individual actions, behavior, and location data across a range of services and devices. Public surveys indicate a clear and overwhelming demand for these tools, and the government and private sector must continue working to evolve privacy-enhancing technologies in step with improved consumer services.

The government should lead a consultative process to assess how the Health Insurance Portability and Accountability Act and other relevant federal laws and regulations can best accommodate the advances in medical science and cost reduction in health care delivery enabled by big data.

Breakthroughs in predicting, detecting, and treating disease deserve the utmost public policy attention, but are unlikely to realize their full potential without substantial improvements in the medical data privacy regime that enables researchers to combine and analyze various kinds of lifestyle and health information. Any proposed reform must also consider bringing under regulatory and legal protection the vast quantities of personal health information circulated by organizations that are not covered entities governed by the Health Insurance Portability and Accountability Act.

The United States should lead international conversations on big data that reaffirms the Administration’s commitment to interoperable global privacy frameworks.

The benefits of big data depend on the global free flow of information. The United States should engage international partners in a dialogue on the benefits and challenges of big data as they impact the legal frameworks and traditions of different nations.

Specifically, the Department of State and the Department of Commerce should actively engage with bilateral and intergovernmental partners, including the European Union, Asia Pacific Economic Cooperation (APEC), and Organization for Economic Cooperation and Development, and with other stakeholders, to take stock of how existing and pro-posed policy frameworks address big data.

The Administration should also work to strengthen the U.S.-European Union Safe Harbor Framework, encourage more countries and companies to join the APEC Cross Border Privacy Rules system, and promote collaboration on data flows between the United States, Europe and Asia through efforts to align Europe's system of Binding Corporate Rules and the APEC CBPR system.

Privacy is a worldwide value that the United States respects and which should be reflected in how it handles data regarding all persons.
For this reason the United States should extend privacy protections to non-U.S. persons.

**RECOMMENDATION**: The Office of Management and Budget should work with departments and agencies to apply the Privacy Act of 1974 to non-U.S. persons where practicable, or to establish alternative privacy policies that apply appropriate and meaningful protections to personal information regardless of a person’s nationality.

2. Responsible Educational Innovation in the Digital Age

Big data offers significant opportunities to improve learning experiences for children and young adults. Big data intersects with education in two important ways. As students begin to share information with educational institutions, they expect that they are doing so in order to develop knowledge and skills, not to have their data used to build extensive profiles about their strengths and weaknesses that could be used to their disadvantage in later years. Educational institutions are also in a unique position to help prepare children, adolescents, and adults to grapple with the world of big data.

**Ensure data protection while promoting innovation in learning**

Substantial breakthroughs stand to be made using big data to improve education as personalized learning on network-enabled devices becomes more common. Over the next five years, under the President’s ConnectED initiative, American classrooms will receive a dramatic influx of technology—with substantial potential to enhance teaching and learning, particularly for disadvantaged communities. Internet-based education tools and software enable rapid iteration and innovation in educational technologies and businesses. These technologies are already being deployed with strong privacy and safety protections for students, inside and outside of the classroom. The Family Educational Rights and Privacy Act and Children’s Online Privacy Protection Act provide a federal regulatory framework to protect the privacy of students—but FERPA was written before the Internet, and COPPA was written before smartphones, tablets, apps, the cloud, and big data. Students and their families need robust protection against current and emerging harms, but they also deserve access to the learning advancements enabled by technology that promise to empower all students to reach their full potential.

**RECOMMENDATION**: The federal government should ensure that data collected in schools is used for educational purposes and continue to support investment and innovation that raises the level of performance across our schools. To promote this innovation, it should explore how to modernize the privacy regulatory framework under the Family Educational Rights and Privacy Act and Children’s Online Privacy Protection Act and Children’s Online Privacy Protection Act to ensure two complementary goals: 1) protecting students against their data being shared or used inappropriately, especially when that data is gathered in an educational context, and 2) ensuring that innovation in educational technology, including new approaches and business models, have ample opportunity to flourish.

**Recognize digital literacy as an important 21st century skill.**

In order to ensure students, citizens, and consumers of all ages have the ability to adequately protect themselves from data use and abuse, it is important that they develop fluency in understanding the ways in which data can be collected and shared, how algorithms are employed and for what purposes, and what tools and techniques they can use to protect themselves. Although such skills will never replace regulatory protections, increased digital literacy will better prepare individuals to live in a world saturated by data. Digital literacy—understanding how personal data is collected, shared, and used—should be recognized as an essential skill in K-12 education and be integrated into the standard curriculum.
3. Big Data and Discrimination

The technologies of automated decision-making are opaque and largely inaccessible to the average person. Yet they are assuming increasing importance and being used in contexts related to individuals’ access to health, education, employment, credit, and goods and services. This combination of circumstances and technology raises difficult questions about how to ensure that discriminatory effects resulting from automated decision processes, whether intended or not, can be detected, measured, and redressed. We must begin a national conversation on big data, discrimination, and civil liberties.

The federal government must pay attention to the potential for big data technologies to facilitate discrimination inconsistent with the country’s laws and values

RECOMMENDATION: The federal government’s lead civil rights and consumer protection agencies, including the Department of Justice, the Federal Trade Commission, the Consumer Financial Protection Bureau, and the Equal Employment Opportunity Commission, should expand their technical expertise to be able to identify practices and outcomes facilitated by big data analytics that have a discriminatory impact on protected classes, and develop a plan for investigating and resolving violations of law in such cases. In assessing the potential concerns to address, the agencies may consider the classes of data, contexts of collection, and segments of the population that warrant particular attention, including for example genomic information or information about people with disabilities.

Consumers have a legitimate expectation of knowing whether the prices they are offered for goods and services are systematically different than the prices offered to others

It is implausible for consumers to be presented with the full parameters of the data and algorithms shaping their online and offline experience. Nonetheless, some transparency is appropriate when a consumer’s experience is being altered based on their personal information, particularly in situations where companies offer differential pricing to consumers in situations where they would not expect it—such as when comparing airline ticket prices on a web-based search engine or visiting the online storefront of a major retailer. The President’s Council of Economic Advisers should assess the evolving practices of differential pricing both online and offline, assess the implications for efficient operations of markets, and consider whether new practices are needed to ensure fairness for the consumer.

Data analytics can be used to shore up civil liberties

The same big data technologies that enable discrimination can also help groups enforce their rights. Applying correlative and data mining capabilities can identify and empirically confirm instances of discrimination and characterize the harms they caused. The federal government’s civil rights offices, together with the civil rights community, should employ the new and powerful tools of big data to ensure that our most vulnerable communities are treated fairly.

To build public awareness, the federal government’s consumer protection and technology agencies should convene public workshops and issue reports over the next year on the potential for discriminatory practices in light of these new technologies; differential pricing practices; and the use of proxy scoring to replicate regulated scoring practices in credit, employment, education, housing, and health care.

https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review
Updated: Thu, 25 Jul 2019 08:57:25 GMT
Powered by mindtouch™
4. Law Enforcement and Security

Big data, lawfully applied, can make our communities safer, make our nation's infrastructure more resilient, and strengthen our national security. It is crucial that the national security, homeland security, law enforcement, and intelligence communities continue to vigorously experiment with and apply lawful big data technology while adhering to full accountability, oversight, and relevant privacy requirements.

The Electronic Communications Privacy Act should be reformed

RECOMMENDATION: Congress should amend ECPA to ensure the standard of protection for online, digital content is consistent with that afforded in the physical world—including by removing archaic distinctions between email left unread or over a certain age.

The use of predictive analytics by law enforcement should continue to be subjected to careful policy review

It is essential that big data analysis conducted by law enforcement outside the context of predicated criminal investigations be deployed with appropriate protections for individual privacy and civil liberties. The presumption of innocence is the bedrock of the American criminal justice system. To prevent chilling effects to Constitutional rights of free speech and association, the public must be aware of the existence, operation, and efficacy of such programs.

Federal agencies with expertise in privacy and data practices should provide technical assistance to state, local, and other federal law enforcement agencies seeking to deploy big data techniques

Law enforcement agencies should continue to examine how federal grants involving big data surveillance technologies can foster their responsible use, as well as the potential utility of establishing a national registry of big data pilots in state and local law enforcement in order to track, identify, and promote best practices. Federal government agencies with technology leaders and experts should also report progress in developing privacy-protective technologies over the next year to help advance the development of technical skills for the advancement of the federal privacy community. Government use of lawfully-acquired commercial data should be evaluated to ensure consistency with our values. Recognizing the longstanding practice of basic commercial records searches against criminal suspects, the federal government should undertake a review of uses of commercially available data on U.S. citizens, focusing on the use of services that employ big data techniques and ensuring that they incorporate appropriate oversight and protections for privacy and civil liberties.

Federal agencies should implement best practices for institutional protocols and mechanisms that can help ensure the controlled use and secure storage of data

The Department of Homeland Security, the intelligence community, and the Department of Defense are among the leaders in developing privacy-protective technologies and policies for handling personal data. Other public sector agencies should evaluate whether any of these practices—particularly data tagging to enforce usage limitations, controlled access policies, and immutable auditing—could be integrated into their databases and data practices to provide built-in protections for privacy, civil rights, and civil liberties.

Use big data analysis and information sharing to strengthen cybersecurity
Protecting the networks that drive our economy, sustain public safety, and protect our national security has become a critical homeland security mission. The federal government’s collaboration with private sector partners to use big data in programs, pilots, and research for both cybersecurity and protecting critical infrastructure can help strengthen our resilience and cyber defenses, especially as more cyber threat data is shared. The Administration continues to support legislation that protects privacy while providing targeted liability protection for companies sharing certain threat information and appropriately defending their networks on that basis. At the same time, the Administration will continue to use executive action to increase incentives for and reduce barriers to the kind of information sharing and analytics that will help the public and private sector prevent and respond to cyber threats.

5. Data as a Public Resource

Government data is a national resource, and should be made broadly available to the public wherever possible, to advance government efficiency, ensure government accountability, and generate economic prosperity and social good—while continuing to protect personal privacy, business confidentiality, and national security. This means finding new opportunities for the government to release large data sets and ensuring all agencies make maximum use of Data.gov, a repository of federal data tools and resources. Big data can help improve the provision of public services, provide new insights to inform policymaking, and increase the efficient use of taxpayer dollars at every level of government.

**Government data should be accurate and securely stored, and to the maximum extent possible, open and accessible**

Government data—particularly statistical and census data—distinguishes itself by providing a high level of accuracy, reliability, and confidentiality. Similarly, the “My Data” initiatives that currently allow Americans easy, secure access to their own digital data in useful formats constitutes a model for personal data accessibility that should be replicated as widely as possible across the government.

**All departments and agencies should, in close coordination with their senior privacy and civil liberties officials, examine how they might best harness big data to help carry out their missions**

Departments and agencies that have not historically made wide use of advanced data analytics should make the most out of what the big data revolution means for them and the citizens they serve. They should experiment with pilot projects, develop in-house talent, and potentially expand research and development. From the earliest stages, agencies should build these projects in consultation with their privacy and civil liberties officers.

In particular, big data analytics present an important opportunity to increase value and performance for the American people in the delivery of government services. Big data also holds enormous power to detect and address waste, fraud and abuse, thereby saving taxpayer money and improving public trust. Big data can further help identify high performers across government whose practices can be replicated by similar agencies and programs and may deliver new insights into effective public-sector management.

**We should dramatically increase investment for research and development in privacy-enhancing technologies, encouraging cross-cutting research that involves not only computer science and mathematics, but also social science, communications and legal disciplines**
The Administration should lead an effort to identify areas where big data analytics can provide the greatest impact for improving the lives of Americans and encourage data scientists to develop social, ethical, and policy knowledge. To this end, the Office of Science and Technology Policy, in partnership with experts across the agencies, should work to define areas that promise significant public gains—for example, in urban informatics—and assess how to provide appropriate attention and resources.

Promising areas for basic research include data provenance, de-identification and encryption, but we also encourage focusing on lab-to-market tools that can be rapidly deployed to consumers. Because we will need a growing cadre of data and social scientists who are able to encode critical policy values into technical infrastructure, we support investment in fields such as Science and Technology Studies which emphasize teaching scientific knowledge and technology in its social and ethical context, and the teaching of module courses to data scientists and engineers to familiarize them with the broader societal implications of their work.

References


9. Harvard Professor of Science & Technology Studies Sheila Jasanoff argues that framing the policy implica-tions of big data is difficult precisely because it manifests in multiple contexts that each call up different op-erative concerns, including big data as property (who owns it); big data as common pool resources (who manages it and on what principles); and big data as identity (it is us ourselves, and thus its management raises constitutional questions about rights).


11. The distinction between data that is “born analog” and data that is “born digital” is explored at length in the PCAST report, Big Data and Privacy, p 18-22.


23. See, e.g., City of Ontario v. Quon, 560 U.S. 746, 755-56 (2010) (“The [Fourth] Amendment guarantees the privacy, dignity, and security of persons against certain arbitrary and invasive acts by officers of the Government.”); Kyllo v. United States, 533 U.S. 27, 31 (2001) (“At the very core’ of the Fourth Amendment ‘stands the right of a man to retreat into his own home and there be free from unreasonable governmental intrusion.”); Olmstead v. United States, 277 U.S. 438, 478 (1928) (Brandeis, J., dissenting) (“They [the Framers] sought to protect Americans in their beliefs, their thoughts, their emotions and their sensations. They conferred, as against the Government, the right to be let alone—the most comprehensive of rights and the right most valued by civilized men.”).
24. For example, e.g. the 1790 Census counted white men “over 16” and “under 16” separately to determine military eligibility. United States Census Bureau, “History,” https://www.census.gov/history/www/1790_1.html; Margo Anderson, The American Census: A Social History, (Yale University Press, 1988).


29. These events have helped federal agencies showcase government data resources being made freely available; collaborate with innovators about how open government data can be used to fuel new products, services, and companies; launch new challenges and incentive prizes designed to spur innovative use of data; and highlight how new uses of open government data are making a tangible impact in American lives and advancing the national interest.

30. Specifically, the Open Data Policy (OMB M-13-13) requires agencies to collect or create information in a way that supports downstream information processing and dissemination; to maintain internal and external data asset inventories; and to clarify information management responsibilities. Agencies must also use ma-chine-readable and open formats, data standards, and common core and extensible metadata.

32. In November 2013, the White House organized a “Data to Knowledge to Action” event that featured dozens of announcements of new public, private, academic and non-profit initiatives. From transforming how research universities prepare students to become data scientists to allowing more citizens and entrepreneurs to access and analyze the huge amounts of space-based data that NASA collects about the Earth, the commitments promise to spur tremendous progress. The Administration is also working to increase the number of data scientists who are actively engaged in solving hard problems in education, health care, sustainability, informed decision-making, and non-profit effectiveness.


37. Ibid at 478.

38. Katz v. United States, 389 U.S. 347, 361 (1967) (Harlan, J., concurring); see also LaFave, supra note 35 § 2.1(b) ("[L]ower courts attempting to interpret and apply Katz quickly came to rely upon the Harlan elaboration, as ultimately did a majority of the Supreme Court.").

39. Restatement (First) Torts § 867 (1939).

40. Prosser, supra note 34 at 389 (1960).

41. Ibid. See also Restatement (Second) Torts § 652A (1977) (Prosser’s privacy torts incorporated into the Restatement).
42. Ibid.


46. Ibid at 27.

47. The APEC Privacy Principles are associated with the 2004 APEC Privacy Framework and APEC Cross Border Privacy Rules system approved in 2011. See Asia-Pacific Economic Cooperation, “APEC Privacy Principles,” 2005, p. 3, http://www.apec.org/Groups/Committee...cyframewk.ashx; Consumer Data Privacy In A Net-worked World, p 49-52; http://export.gov/safeharbor for information on the U.S.-EU and U.S.-Swiss Safe Harbor Frameworks. These enforceable self-certification programs are administered by the U.S. Department of Commerce and were developed in consultation with the European Commission and the Federal Data Protection and Information Commissioner of Switzerland, respectively, to provide a streamlined means for U.S. organizations to comply with EU and Swiss data protection laws.

48. California, for example, has a right to privacy in the state Constitution. Cal. Const. art. 1 § 1.


50. This principle ensures that covered entities make reasonable efforts to use, disclose, and request only the minimum amount of protected health information needed to accomplish the intended purpose of the use, disclosure, or request. See U.S. Department of Health & Human Services, Health Information Privacy, “Min-imun Necessary Requirement,” http://www.hhs.gov/ocr/privacy/hipaa...necessary.html.

52. See Consumer Data Privacy In A Networked World, p 25.


55. See Joined Cases C-293/12 and C-594/12, Digital Rights Ireland Ltd. v. Minister for Communications, Marine and Natural Resources, et al. (Apr. 8, 2014) in which the European Court of Justice invalidated the data retention requirements applied to electronic communications on the basis that the scope of the requirements interfered in a “particularly serious manner with the fundamental rights to respect for private life and to the protection of personal data.”


59.

Latanya Sweeney, a Professor of Government and Technology in Residence at Harvard University, has studied information flows in the health care industry. A graphical map of data flows that depicts information flows outside entities regulated by HIPAA can be found at http://www.thedatamap.org.

60.

President’s Council of Advisors on Science & Technology, Realizing the Full Potential Of Health Information Technology to Improve Health Care for Americans: The Path Forward, The White House, December 2010, http://www.whitehouse.gov/sites/defa...it-report.pdf.

61.


62.


63.


64.


65.

For example, California recently passed a law prohibiting online services from gathering information about a minor’s activities for marketing purposes, or from displaying certain online advertising to minors. The law further requires online services to delete information that the minor posted on the website or service, a right for which the statute has now been dubbed “the Eraser Law.”

66.

The Department of Education is exploring data innovation and use in a wide variety of contexts, including making more educational data available through application programming interfaces. See David Soo, “How can the Department of Education Increase Innovation, Transparency and Access to Data?,” Department of Education Blog, http://www.ed.gov/blog/2014/04/how-c...ccess-to-data/.
67.

68.

69.
In the first phase, three databases, from different parts of the agency, are fed into Neptune, where the data is then tagged and sorted. From there, the Department of Homeland Security feeds this tagged data into Cerberus, which operates at the classified level. Here, DHS can compare its unclassified and classified information.

70.

71.
Most jurisprudence to date does not consider in their entirety big data technologies by the definition used in this report, but rather many of the advanced technologies, such as GPS trackers, that now play a crucial role in big data applications.

72.

73.
Ibid at n.3.

74.
Over 70 cities in the U.S. use gunshot detection technology developed and provided by SST Solutions called ShotSpotter. For more information, please visit www.shotspotter.com.

75.
76. The National Institute of Justice, the Department of Justice’s research, development, and evaluation agency, provides detailed information on the use of predictive policing at law enforcement agencies. For more information, visit www.nij.gov/topics/law-enforcement/strategies/predictive-policing.


78. The application of this particular predictive policing technology emerged out of a series of grants issued by the National Institute of Justice the Chicago Police Department, most recently involving Miles Wernick as technical investigator. For more information, see http://www.nij.gov/topics/law-enforcement/research.aspx.


81. Though some argue big data analysis is merely a new way to expand the scope of what can be considered “suspicion,” the program in question uses an algorithmic calculation heavily reliant on an individual’s associations without other criminal pretext.


86. United States v. Warshak, 631 F.3d 266 (6th Cir. 2010).

87. This assertion was not part of the Supreme Court's holding, but emphasizes the emerging discussion of third-party doctrine. United States v. Jones, 132 S.Ct. 945, 957 (2012) (Sotomayor, J., concurring).

88. The doctrine has been adapted and applied to cell-site location information multiple times, most recently by the Fifth Circuit in In re Application of the United States for Historical Cell Site Data, 724 F.3d 600 (5th Cir. 2013) (finding cell site data may be obtained without a probable cause warrant); United States v. Norris, No. 2:11-CR-00188-KJM, 2013 WL 4737197 (E.D. Cal. Sept. 3, 2013) (finding defendant who hacked a private wireless network had no reasonable expectation of privacy in his transmissions over that network). Moreover, leading commentators have argued for the continuing vitality of the third-party doctrine in the modern era, including Professor Orin Kerr in Orin S. Kerr, "The Case for the Third-Party Doctrine," 107 Michigan Law Review 561 (2009), and Orin S. Kerr, "Defending the Third-Party Doctrine: A Response to Epstein and Murphy," 24 Berkeley Technology Law Journal 1229 (2009). See also United States v. Perrine, 518 F.3d 1196, 1204 (10th Cir. 2008); United States v. Forrester, 512 F.3d 500, 510 (9th Cir. 2008).


92.
Ibid.

93.


94.


95.


96.
Ibid.

97.


98.


99.


100.


102. For information about the industry’s opt-out program, see http://www.youradchoices.com/..


108. Ibid at 22.

110.

111.

112.

113.

114.

115.
See New Urban Mechanics, http://www.newurbanmechanics.org/. All information about Street Bump comes from its former project manager James Solomon, who was interviewed by officials from the office of the White House Chief Technology Officer.

116.

117.

118.
119.

120.
Ibid at 36.

121.

122.

123.

Appendix

A. Methodology

This 90-day study was announced by President Obama in his January 17, 2014 remarks on the review of signals intelligence. He charged his Counselor John Podesta to “look how the challenges inherent in big data are being confronted by both the public and private sectors; whether we can forge international norms on how to manage this data; and how we can continue to promote the free flow of information in ways that are consistent with both privacy and security.” Podesta led a working group of senior Administration officials including Secretary of Commerce Penny Pritzker, Secretary of Energy Ernie Moniz, Director of the Office of Science and Technology Policy John Holdren, and Director of the National Economic Council Jeffrey Zients. Nicole Wong, R. David Edelman, Christopher Kirchhoff, and Kristina Costa were the principal staff authors supporting this report. To inform its deliberations, the working group initiated a broad public dialogue on the implications of technological advancements in big data.

During the course of this study, the working group met with hundreds of stakeholders from industry, academia, civil society, and the federal government through briefings at the White House. These briefings provided a chance for dialogue with key stakeholders, including privacy and civil liberties advocates; scientific and statistical agencies; international data protection authorities; the intelligence community; law enforcement officials; leading academics who study social and technical aspects of privacy and the Internet; and practitioners and executives from the health care, financial, and information services industries. A full list of briefings and participants is included in Section B of the appendix.
To further engage the public, the White House Office of Science and Technology Policy sponsored conferences at the Massachusetts Institute of Technology, New York University, and the University of California, Berkeley. Senior Administration officials, including Counselor Podesta and Secretary Pritzker, participated in these conferences, along with policy experts, academics, and representatives from business and the nonprofit community. Details of these conferences and a list of presentations is included in Section C of the appendix.

The working group also published a Federal Register notice to gather written input, and used the whitehouse.gov platform to solicit comments from the general public online. Details of these efforts are included in Sections E and F of the appendix.

B. Stakeholder Meetings

Acxiom
Adobe
Allstate
Ally Financial
Amazon
American Association of Advertising Agencies
American Association of Universities
American Civil Liberties Union
Apple
AppNexus
Archimedes Incorporated
Asian Americans Advancing Justice
Association of National Advertisers
athenahealth
Bank of America
BlueKai
Bureau of Consumer Protection
Canadian Privacy Commissioner
Capital One
Carnegie Mellon University
Cato Institute
Census Bureau
Center for Democracy & Technology
Center for Digital Democracy
Center for National Security Studies
Central Intelligence Agency
ColorOfChange
Computer Science and Artificial Intelligence Laboratory, MIT
comScore
Corelogic
Cornell University
Council of Better Business Bureaus
Data Privacy Commissioner, Mexico
Datalogix
Department of Commerce
Department of Homeland Security
Digital Advertising Alliance
Direct Marketing Association
Discover
Drug Enforcement Administration
Duke University School of Law
Data Protection Authority, Netherlands
Economics and Statistics Administration
Electronic Frontier Foundation
Electronic Privacy Information Center
Epsilon
European Union Data Protection Supervisor
European Commission: Directorate-General for Justice (Data Protection Division)
Evidera
Experian
Explorys
Facebook
Federal Bureau of Investigation
Federal Communications Commission, Bureau of Consumer Protection
Federal Trade Commission
Financial Services Roundtable
Free Press
Future of Privacy
George Washington University
Georgetown University Law Center
GNS Health care
Google
GroupM
Harvard University
Humedica
IBM Health care
IMS Health
Infogroup
Information Commissioner, United Kingdom
Interactive Advertising Bureau
International Association of Privacy Professionals
Jenner & Block LLP
Lawrence Livermore National Laboratory
LexisNexis
LinkedIn
C. Academic Symposia
Big Data and Privacy Workshop: Advancing the State of the Art in Technology and Practice

Massachusetts Institute of Technology (MIT)
Cambridge, Massachusetts
March 3, 2014

Welcome: L. Rafael Reif, President of MIT

Keynote: John Podesta, Counselor to the President

Keynote: Penny Pritzker, Secretary of Commerce

State of the Art of Privacy Protection: Cynthia Dwork, Microsoft

Panel Session 1: Big Data Opportunities and Challenges

Panel Chair: Daniela Rus, MIT
Mike Stonebraker, MIT
John Guttag, MIT
Manolis Kellis, MIT
Sam Madden, MIT
Anant Agarwal, edX

Panel Session 2: Privacy Enhancing Technologies

Panel Chair: Shafi Goldwasser
Nickolai Zeldovich, MIT
Vinod Vaikuntanathan, Assistant Professor, MIT
Salil Vadhan, Harvard University
Daniel Weitzner, MIT

Panel Session 3: Roundtable Discussion of Large-Scale Analytics Case Study

Panel Moderator: Daniel Weitzner
Chris Calabrese, American Civil Liberties Union
John DeLong, National Security Agency
Mark Gorenberg, Zetta Venture Partners
David Hoffman, Intel
Karen Kornbluh, Nielsen
Andy Palmer, KOA Lab
James Powell, Thomson Reuters
Latanya Sweeney, Harvard University
Vinod Vaikuntanathan, MIT

Concluding Statements: Maria Zuber, MIT
The Social, Cultural, & Ethical Dimensions of ‘Big Data’

The Data & Society Research Institute & New York University (NYU)
New York, New York
March 17, 2014

Introduction: danah boyd, Data & Society

Fireside Chat: John Podesta, Counselor to the President

Discussion Breakouts

Tim Hwang: On Cognitive Security
Nick Grossman: Regulation 2.0
Nuala O’Connor: The Digital Self & Technology in Daily Life
Alex Howard: Data Journalism in the Second Machine Age
Mark Latonero: Big Data and Human Trafficking
Corrine Yu: Civil Rights Principles for the Era of Big Data
Natasha Schüll: Tracking for Profit; Tracking for Protection
Kevin Bankston: The Biggest Data of All
Alessandro Acquisti: The Economics of Privacy (and Big Data)
Latanya Sweeney: Transparency Builds Trust
Deborah Estrin: You + Your Data
Clay Shirky: Analog Thumbs on Digital Scales Open Discussion
Moderators: danah boyd and Nicole Wong

Workshops

Data Supply Chains
Inferences and Connections
Predicting Human Behavior
Algorithmic Accountability
Interpretation Gone Wrong
Inequalities and Asymmetries

Public Plenary

Welcome: danah boyd, Data & Society

Video Address: John Podesta, Counselor to the President

Keynote: Nicole Wong, US Deputy Chief Technology Officer

Plenary Panel Statements
Welcome: Dean AnnaLee Saxenian, UC Berkeley School of Information
Welcome: Nicole Wong, US Deputy Chief Technology Officer

Panel Session 1: Values at stake, Values in tension: Privacy and Beyond

Moderator: Deirdre Mulligan, UC Berkeley School of Information
Amalia Deloney, Center for Media Justice
Nicole Ozer, Northern California ACLU
Fred Cate, University of Indiana
Kenneth A. Bamberger, UC Berkeley School of Law

Panel Session 2: New Opportunities and Challenges in Health and Education

Moderator: Paul Ohm, University of Colorado Law School
Barbara Koenig, University of California, San Francisco
Deven McGraw, Center for Democracy & Technology
Scott Young, Kaiser Permanente
Zachary Pardos, UC Berkeley School of Information

Panel Session 3: Algorithms: Transparency, Accountability, Values and Discretion

Moderator: Omer Tene, International Association of Privacy Professionals
Ari Gesher, Palantir
Lee Tien, Electronic Frontier Foundation
Seeta Gangadharan, New America Foundation
Thejo Kote, Automatic
James Rule, UC Berkeley

Governance Roundtable

Moderator: David Vladeck, Georgetown University Law School
Julie Brill, Federal Trade Commission
D. PCAST Report

To take measure of the shifting technological landscape, the President charged his Council of Advisors on Science & Technology (PCAST) to conduct a parallel study to assess the technological dimensions of the intersection of big data and privacy. PCAST’s statement of work reads, in part:

“PCAST will study the technological aspects of the intersection of big data with individual privacy, in relation to both the current state and possible future states of the relevant technological capabilities and associated privacy concerns. Relevant big data include data and metadata collected, or potentially collectable, from or about individuals by entities that include the government, the private sector, and other individuals. It includes both proprietary and open data, and also data about individuals collected incidentally or accidentally in the course of other activities (e.g., environmental monitoring or the “Internet of things”).

The PCAST assessment was conducted simultaneously with the 90-study on big data. PCAST shared their preliminary conclusions with the working group in order to inform its deliberations. The final PCAST report can be found at http://whitehouse.gov/bigdata and at PCAST’s own website, http://whitehouse.gov/administration/eop/ostp/pcast.

E. Public Request for Information

As part of the effort to make this review as inclusive as possible, the White House Office of Science and Technology Policy (OSTP) released a Request for Information (RFI) seeking public comment on the ways in which big data may impact privacy, the economy, and public policy. The RFI was published on March 4, 2014, and 76 comments were submitted through April 4, 2014. The comments came from nonprofits, corporations, universities, and individual citizens. The full list of respondents is included below, and the full text of all responses is publicly available at whitehouse.gov/bigdata.

The RFI posed five questions to respondents:

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

(2) What types of uses of big data could measurably improve outcomes or productivity with further government action, funding, or research? What types of uses of big data raise the most public policy concerns? Are there specific sectors or types of uses that should receive more government and/or public attention?
(3) What technological trends or key technologies will affect the collection, storage, analysis and use of big data? Are there particularly promising technologies or new practices for safeguarding privacy while enabling effective uses of big data?

(4) How should the policy frameworks or regulations for handling big data differ between the government and the private sector? Please be specific as to the type of entity and type of use (e.g., law enforcement, government services, commercial, academic research, etc.).

(5) What issues are raised by the use of big data across jurisdictions, such as the adequacy of current international laws, regulations, or norms?

The RFI can be found at: http://www.gpo.gov/fdsys/pkg/FR-2014-...2014-04660.pdf. Respondents:

Respondents:
Access
American Civil Liberties Union
Ad Self-Regulatory Council, Council of Better Business Bureaus
Annie Shebanow
The Architecture for a Digital World and Advanced Micro Devices
Association for Computing Machinery
Association of National Advertisers
Brennan Center for Justice
BSA | The Software Alliance
Center for Democracy and Technology
Center for Data Innovation
Center for Digital Democracy
Center for National Security Studies
Cloud Security Alliance
Coalition for Privacy and Free Trade
Common Sense Media
Computer and Communications Industry Association
Computing Community Consortium
Constellation Research
Consumer Action
Consumer Federation of America
Consumer Watchdog
Dell
Direct Marketing Association
Dr. Tyrone W A Grandison
Dr. A. R. Wagner
Durrell Kapan
Electronic Frontier Foundation
Electronic Privacy Information Center
Electronic Transactions Association
Entity
Federation of American Societies for Experimental Biology
Financial Services Roundtable
Food Marketing Groups
Frank Pasquale, UMD Law
Fred Cate, Microsoft, Oxford Internet Institute
Future of Privacy Forum
Georgetown University
Health care Leadership Council
IMS Health
Information Technology Industry Council
Interactive Advertising Bureau
Intrical
IT Law Group
Jackamo
James Cooper, George Mason Law
Jason Kint
Jonathan Sander, STEALTHbits
Kaliya Identity Woman
Leadership Conferences on Civil and Human Rights & Education
Making Change at Walmart
Marketing Research Association
Mary Culnan, Bentley University & Future of Privacy Forum
McKenna Long & Aldridge LLP
mediajustice.org
Microsoft
Massachusetts Institute of Technology
MITRE Corporation
Mozilla
New York University Center for Urban Science & Progress
Online Trust Alliance
Pacific Northwest National Laboratory
Peter Muhlberger
Privacy Coalition
Reed Elsevier
Sidley Austin LLP
Software & Information Industry Association
TechAmerica
TechFreedom
Technology Policy Institute
The Internet Association
U.S. Chamber of Commerce
U.S. Leadership for the Revision of the 1967 Space Treaty
F. White House Big Data Survey

Additional public input about big data and privacy issues was solicited via a short web form posted on WhiteHouse.gov and promoted via email and social media. During the four weeks the survey was open for public input, 24,092 people submitted responses. It is important to note, however, that this process was a means of gathering public input and should not be considered a statistically representative survey of attitudes about data privacy. The White House did not include submission fields for name or contact information on the survey form.

Respondents expressed a great deal of concern about big data practices. They communicated particularly strong feelings around ensuring that data practices have proper transparency and oversight—more than 80 percent of respondents were very concerned with each of these areas—but even in the area of least concern (collection of location data), 61 percent indicated that they were “very much concerned” about this practice. By contrast, considerably more nuance was evident in respondents’ views towards particular entities. Although majorities claimed to trust Intelligence and Law Enforcement Agencies “not at all,” their views towards other government agencies at both federal and local levels were far less negative. Furthermore, majorities were generally trusting of how professional practices, like law and medical offices, and academia use and handle big data.

Figure Concern With Data Practices

![Concern with data practices](https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review)
Figure Percent Who Do Not Trust Each Entity At All

![Chart showing percent who do not trust each entity at all]

Taken together, the findings from this survey indicate that respondents were most wary of how intelligence and law enforcement agencies are collecting and using data about them, particularly when they have little insight into these practices. This suggests that the Administration should work to increase the transparency about intelligence practices where possible, reassure the public that collected data is stored as securely as possible, and strengthen applicable legal structures and oversight.

For more information about the survey, visit: WhiteHouse.gov/BigData.

Fact Sheet: PCAST Report on Big Data and Privacy: A Technological Perspective

Source: [http://www.whitehouse.gov/sites/defaul..._formatted.pdf](http://www.whitehouse.gov/sites/defaul..._formatted.pdf) (PDF)

The White House
Office of Science & Technology Policy
For Immediate Release May 1, 2014

Advances in information technology have led to many new ways to collect data, analyze, and use data in ever expanding volumes. Big data holds tremendous potential to benefit society and contribute to economic growth, yet it also presents new challenges related to individual privacy. In January, President Obama asked his Council of Advisors on Science and Technology (PCAST) to analyze the technological dimension of this big data transformation and its significance for the future of privacy.

In its report, PCAST examines what distinguishes big data from data at smaller scales, how the infrastructure for handling big data is evolving through services such as cloud computing, how approaches to analyzing big data are evolving and what insights they are yielding, the opportunities and limitations technology offers in protecting privacy, and what these and other technical factors imply for public policy. It discusses a number of domains (e.g. health care, education) that accumulate big data and explores ways it can be used.
Both technology and policy play important roles in protecting privacy. PCAST concludes that technical measures alone are not sufficient for protecting privacy. In its report to the President, PCAST recommends five steps the Federal government can take around big data and privacy:

- **Recommendation 1**: Policy attention should focus more on the actual uses of big data and less on its collection and analysis.
- **Recommendation 2**: Policies and regulation at all levels of government should not embed particular technological solutions, but rather should be stated in terms of intended outcomes.
- **Recommendation 3**: With coordination and encouragement from the White House Office of Science and Technology Policy (OSTP), the agencies of the Networking and Information Technology Research and Development program should strengthen U.S. research in privacy-related technologies and in the relevant areas of social science that inform the successful application of those technologies.
- **Recommendation 4**: OSTP together with the appropriate educational institutions and professional societies should encourage increased education and training opportunities concerning privacy protection, including career paths for professionals.
- **Recommendation 5**: The United States should take the lead both in the international arena and at home by adopting policies that stimulate the use of practical privacy-protecting technologies that exist today.

Advances in technology hold the potential to both enhance and diminish personal privacy. In its report, PCAST notes that the challenge is to understand the nature of privacy in the modern world and to identify technical, educational, and policy solutions to help preserve and protect privacy and the societal benefits and economic potential around the use of big data.

###

**Big Data and Privacy: A Technology Perspective**

Source: [http://www.whitehouse.gov/sites/defa...-_may_2014.pdf](http://www.whitehouse.gov/sites/defa...-_may_2014.pdf) (PDF)

The President’s Council of Advisors on Science and Technology

PCAST Big Data and Privacy Working Group

**Working Group Co-Chairs**

**Susan L. Graham**  
Pehong Chen Distinguished Professor  
Emerita in Electrical Engineering and Computer Science  
University of California, Berkeley

**William Press**  
Raymer Professor in Computer Science and Integrative Biology  
University of Texas at Austin

https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review  
Updated: Thu, 25 Jul 2019 08:57:25 GMT  
Powered by mindtouch
Working Group Members

**S. James Gates, Jr.**
John S. Toll Professor of Physics
Director, Center for String and Particle Theory
University of Maryland, College Park

**Mark Gorenberg**
Managing Member
Zetta Venture Partners

**John P. Holdren**
Assistant to the President for Science and Technology
Director, Office of Science and Technology Policy

**Eric S. Lander**
President
Broad Institute of Harvard and MIT

**Craig Mundie**
Senior Advisor to the CEO
Microsoft Corporation

**Maxine Savitz**
Vice President
National Academy of Engineering

**Eric Schmidt**
Executive Chairman
Google, Inc.

Working Group Staff

**Marjory S. Blumenthal**
Executive Director
President’s Council of Advisors on Science and Technology

**Michael Johnson**
Assistant Director
National Security and International Affairs
Letter

EXECUTIVE OFFICE OF THE PRESIDENT
PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE AND TECHNOLOGY
WASHINGTON, D.C. 20502

President Barack Obama
The White House
Washington, DC 20502

Dear Mr. President,

We are pleased to send you this report, Big Data and Privacy: A Technological Perspective, prepared for you by the President’s Council of Advisors on Science and Technology (PCAST). It was developed to complement and inform the analysis of big-data implications for policy led by your Counselor, John Podesta, in response to your requests of January 17, 2014. PCAST examined the nature of current technologies for managing and analyzing big data and for preserving privacy, it considered how those technologies are evolving, and it explained what the technological capabilities and trends imply for the design and enforcement of public policy intended to protect privacy in big-data contexts.

Big data drives big benefits, from innovative businesses to new ways to treat diseases. The challenges to privacy arise because technologies collect so much data (e.g., from sensors in everything from phones to parking lots) and analyze them so efficiently (e.g., through data mining and other kinds of analytics) that it is possible to learn far more than most people had anticipated or can anticipate given continuing progress. These challenges are compounded by limitations on traditional technologies used to protect privacy (such as de-identification). PCAST concludes that technology alone cannot protect privacy, and policy intended to protect privacy needs to reflect what is (and is not) technologically feasible.

In light of the continuing proliferation of ways to collect and use information about people, PCAST recommends that policy focus primarily on whether specific uses of information about people affect privacy adversely. It also recommends that policy focus on outcomes, on the “what” rather than the “how,” to avoid becoming obsolete as technology advances. The policy framework should accelerate the development and commercialization of technologies that can help to contain adverse impacts on privacy, including research into new technological options. By using technology more effectively, the Nation can lead internationally in making the most of big data’s benefits while limiting the concerns it poses for privacy. Finally, PCAST calls for efforts to assure that there is enough talent available with the expertise needed to develop and use big data in a privacy-sensitive way.

PCAST is grateful for the opportunity to serve you and the country in this way and hope that you and others who read this report find our analysis useful.

Best regards,

John P. Holdren
Co-chair, PCAST
Executive Summary

The ubiquity of computing and electronic communication technologies has led to the exponential growth of data from both digital and analog sources. New capabilities to gather, analyze, disseminate, and preserve vast quantities of data raise new concerns about the nature of privacy and the means by which individual privacy might be compromised or protected.

After providing an overview of this report and its origins, Chapter 1 describes the changing nature of privacy as computing technology has advanced and big data has come to the fore. The term privacy encompasses not only the famous "right to be left alone," or keeping one’s personal matters and relationships secret, but also the ability to share information selectively but not publicly. Anonymity overlaps with privacy, but the two are not identical. Likewise, the ability to make intimate personal decisions without government interference is considered to be a privacy right, as is protection from discrimination on the basis of certain personal characteristics (such as race, gender, or genome). Privacy is not just about secrets.

Conflicts between privacy and new technology have occurred throughout American history. Concern with the rise of mass media such as newspapers in the 19th century led to legal protections against the harms or adverse consequences of "intrusion upon seclusion," public disclosure of private facts, and unauthorized use of name or likeness in commerce. Wire and radio communications led to 20th century laws against wiretapping and the interception of private communications – laws that, PCAST notes, have not always kept pace with the technological realities of today’s digital communications.

Past conflicts between privacy and new technology have generally related to what is now termed “small data,” the collection and use of data sets by private- and public-sector organizations where the data are disseminated in their original form or analyzed by conventional statistical methods. Today’s concerns about big data reflect both the substantial increases in the amount of data being collected and associated changes, both actual and potential, in how they are used.

Big data is big in two different senses. It is big in the quantity and variety of data that are available to be processed. And, it is big in the scale of analysis (termed “analytics”) that can be applied to those data, ultimately to make inferences and draw conclusions. By data mining and other kinds of analytics, nonobvious and sometimes private information can be derived from data that, at the time of their collection, seemed to raise no, or only manageable, privacy issues. Such new information, used appropriately, may often bring benefits to individuals and society – Chapter 2 of this report gives many such examples, and additional examples are scattered throughout the rest of the text. Even in principle, however, one can never know what information may later be extracted from any particular collection of big data, both because that information may result only from the combination of seemingly unrelated data sets, and because the algorithm for revealing the new information may not even have been invented at the time of collection.

The same data and analytics that provide benefits to individuals and society if used appropriately can also create potential harms – threats to individual privacy according to privacy norms both widely shared and personal. For example, large-scale analysis of research on disease, together with health data from electronic medical records and
genomic information, might lead to better and timelier treatment for individuals but also to inappropriate disqualification for insurance or jobs. GPS tracking of individuals might lead to better community-based public transportation facilities, but also to inappropriate use of the whereabouts of individuals. A list of the kinds of adverse consequences or harms from which individuals should be protected is proposed in Section 1.4. PCAST believes strongly that the positive benefits of big-data technology are (or can be) greater than any new harms.

Chapter 3 of the report describes the many new ways in which personal data are acquired, both from original sources, and through subsequent processing. Today, although they may not be aware of it, individuals constantly emit into the environment information whose use or misuse may be a source of privacy concerns. Physically, these information emanations are of two types, which can be called “born digital” and “born analog.”

When information is “born digital,” it is created, by us or by a computer surrogate, specifically for use by a computer or data processing system. When data are born digital, privacy concerns can arise from over-collection. Over-collection occurs when a program’s design intentionally, and sometimes clandestinely, collects information unrelated to its stated purpose. Over-collection can, in principle, be recognized at the time of collection.

When information is “born analog,” it arises from the characteristics of the physical world. Such information becomes accessible electronically when it impinges on a sensor such as a camera, microphone, or other engineered device. When data are born analog, they are likely to contain more information than the minimum necessary for their immediate purpose, and for valid reasons. One reason is for robustness of the desired “signal” in the presence of variable “noise.” Another is technological convergence, the increasing use of standardized components (e.g., cell-phone cameras) in new products (e.g., home alarm systems capable of responding to gesture).

Data fusion occurs when data from different sources are brought into contact and new facts emerge (see Section 3.2.2). Individually, each data source may have a specific, limited purpose. Their combination, however, may uncover new meanings. In particular, data fusion can result in the identification of individual people, the creation of profiles of an individual, and the tracking of an individual’s activities. More broadly, data analytics discovers patterns and correlations in large corpuses of data, using increasingly powerful statistical algorithms. If those data include personal data, the inferences flowing from data analytics may then be mapped back to inferences, both certain and uncertain, about individuals.

Because of data fusion, privacy concerns may not necessarily be recognizable in born-digital data when they are collected. Because of signal-processing robustness and standardization, the same is true of born-analog data – even data from a single source (e.g., a single security camera). Born-digital and born-analog data can both be combined with data fusion, and new kinds of data can be generated from data analytics. The beneficial uses of near-ubiquitous data collection are large, and they fuel an increasingly important set of economic activities. Taken together, these considerations suggest that a policy focus on limiting data collection will not be a broadly applicable or scalable strategy – nor one likely to achieve the right balance between beneficial results and unintended negative consequences (such as inhibiting economic growth).

If collection cannot, in most cases, be limited practically, then what? Chapter 4 discusses in detail a number of technologies that have been used in the past for privacy protection, and others that may, to a greater or lesser extent, serve as technology building blocks for future policies.
Some technology building blocks (for example, cybersecurity standards, technologies related to encryption, and formal systems of auditable access control) are already being utilized and need to be encouraged in the marketplace. On the other hand, some techniques for privacy protection that have seemed encouraging in the past are useful as supplementary ways to reduce privacy risk, but do not now seem sufficiently robust to be a dependable basis for privacy protection where big data is concerned. For a variety of reasons, PCAST judges anonymization, data deletion, and distinguishing data from metadata (defined below) to be in this category. The framework of notice and consent is also becoming unworkable as a useful foundation for policy.

Anonymization is increasingly easily defeated by the very techniques that are being developed for many legitimate applications of big data. In general, as the size and diversity of available data grows, the likelihood of being able to re-identify individuals (that is, re-associate their records with their names) grows substantially. While anonymization may remain somewhat useful as an added safeguard in some situations, approaches that deem it, by itself, a sufficient safeguard need updating.

While it is good business practice that data of all kinds should be deleted when they are no longer of value, economic or social value often can be obtained from applying big data techniques to masses of data that were otherwise considered to be worthless. Similarly, archival data may also be important to future historians, or for later longitudinal analysis by academic researchers and others. As described above, many sources of data contain latent information about individuals, information that can be known only if the holder expends analytic resources, or that may become knowable only in the future with the development of new data-mining algorithms. In such cases it is practically impossible for the data holder even to surface “all the data about an individual,” much less delete it on any specified schedule or in response to an individual’s request. Today, given the distributed and redundant nature of data storage, it is not even clear that data, even small data, can be destroyed with any high degree of assurance.

As data sets become more complex, so do the attached metadata. Metadata are ancillary data that describe properties of the data such as the time the data were created, the device on which they were created, or the destination of a message. Included in the data or metadata may be identifying information of many kinds. It cannot today generally be asserted that metadata raise fewer privacy concerns than data.

Notice and consent is the practice of requiring individuals to give positive consent to the personal data collection practices of each individual app, program, or web service. Only in some fantasy world do users actually read these notices and understand their implications before clicking to indicate their consent.

The conceptual problem with notice and consent is that it fundamentally places the burden of privacy protection on the individual. Notice and consent creates a non-level playing field in the implicit privacy negotiation between provider and user. The provider offers a complex, take-it-or-leave-it set of terms, while the user, in practice, can allocate only a few seconds to evaluating the offer. This is a kind of market failure.

PCAST believes that the responsibility for using personal data in accordance with the user’s preferences should rest with the provider rather than with the user. As a practical matter, in the private sector, third parties chosen by the consumer (e.g., consumer-protection organizations, or large app stores) could intermediate: A consumer might choose one of several “privacy protection profiles” offered by the intermediary, which in turn would vet apps against these profiles. By vetting apps, the intermediaries would create a marketplace for the negotiation of community standards for
privacy. The Federal government could encourage the development of standards for electronic interfaces between the intermediaries and the app developers and vendors.

After data are collected, data analytics come into play and may generate an increasing fraction of privacy issues. Analysis, per se, does not directly touch the individual (it is neither collection nor, without additional action, use) and may have no external visibility. By contrast, it is the use of a product of analysis, whether in commerce, by government, by the press, or by individuals, that can cause adverse consequences to individuals.

More broadly, PCAST believes that it is the use of data (including born-digital or born-analog data and the products of data fusion and analysis) that is the locus where consequences are produced. This locus is the technically most feasible place to protect privacy. Technologies are emerging, both in the research community and in the commercial world, to describe privacy policies, to record the origins (provenance) of data, their access, and their further use by programs, including analytics, and to determine whether those uses conform to privacy policies. Some approaches are already in practical use.

Given the statistical nature of data analytics, there is uncertainty that discovered properties of groups apply to a particular individual in the group. Making incorrect conclusions about individuals may have adverse consequences for them and may affect members of certain groups disproportionately (e.g., the poor, the elderly, or minorities). Among the technical mechanisms that can be incorporated in a use-based approach are methods for imposing standards for data accuracy and integrity and policies for incorporating usable interfaces that allow an individual to correct the record with voluntary additional information.

PCAST’s charge for this study did not ask it to recommend specific privacy policies, but rather to make a relative assessment of the technical feasibilities of different broad policy approaches. Chapter 5, accordingly, discusses the implications of current and emerging technologies for government policies for privacy protection. The use of technical measures for enforcing privacy can be stimulated by reputational pressure, but such measures are most effective when there are regulations and laws with civil or criminal penalties. Rules and regulations provide both deterrence of harmful actions and incentives to deploy privacy-protecting technologies. Privacy protection cannot be achieved by technical measures alone.

This discussion leads to five recommendations.

Recommendation 1. Policy attention should focus more on the actual uses of big data and less on its collection and analysis. By actual uses, we mean the specific events where something happens that can cause an adverse consequence or harm to an individual or class of individuals. In the context of big data, these events ("uses") are almost always actions of a computer program or app interacting either with the raw data or with the fruits of analysis of those data. In this formulation, it is not the data themselves that cause the harm, nor the program itself (absent any data), but the confluence of the two. These "use" events (in commerce, by government, or by individuals) embody the necessary specificity to be the subject of regulation. By contrast, PCAST judges that policies focused on the regulation of data collection, storage, retention, a priori limitations on applications, and analysis (absent identifiable actual uses of the data or products of analysis) are unlikely to yield effective strategies for improving privacy. Such policies would be unlikely to be scalable over time, or to be enforceable by other than severe and economically damaging measures.
Recommendation 2. Policies and regulation, at all levels of government, should not embed particular technological solutions, but rather should be stated in terms of intended outcomes.

To avoid falling behind the technology, it is essential that policy concerning privacy protection should address the purpose (the “what”) rather than prescribing the mechanism (the “how”).

Recommendation 3. With coordination and encouragement from OSTP, the NITRD agencies should strengthen U.S. research in privacy-related technologies and in the relevant areas of social science that inform the successful application of those technologies.

Some of the technology for controlling uses already exists. However, research (and funding for it) is needed in the technologies that help to protect privacy, in the social mechanisms that influence privacy-preserving behavior, and in the legal options that are robust to changes in technology and create appropriate balance among economic opportunity, national priorities, and privacy protection.

Recommendation 4. OSTP, together with the appropriate educational institutions and professional societies, should encourage increased education and training opportunities concerning privacy protection, including career paths for professionals.

Programs that provide education leading to privacy expertise (akin to what is being done for security expertise) are essential and need encouragement. One might envision careers for digital-privacy experts both on the software development side and on the technical management side.

Recommendation 5. The United States should take the lead both in the international arena and at home by adopting policies that stimulate the use of practical privacy-protecting technologies that exist today. It can exhibit leadership both by its convening power (for instance, by promoting the creation and adoption of standards) and also by its own procurement practices (such as its own use of privacy-preserving cloud services).

PCAST is not aware of more effective innovation or strategies being developed abroad; rather, some countries seem inclined to pursue what PCAST believes to be blind alleys. This circumstance offers an opportunity for U.S. technical leadership in privacy in the international arena, an opportunity that should be taken.

1. Introduction

In a widely noted speech on January 17, 2014, President Barack Obama charged his Counselor, John Podesta, with leading a comprehensive review of big data and privacy, one that would “reach out to privacy experts, technologists, and business leaders and look at how the challenges inherent in big data are being confronted by both the public and private sectors; whether we can forge international norms on how to manage this data; and how we can continue to promote the free flow of information in ways that are consistent with both privacy and security.” The President and Counselor Podesta asked the President’s Council of Advisors on Science and Technology (PCAST) to assist with the technology dimensions of the review.

For this task PCAST’s statement of work reads, in part,
PCAST will study the technological aspects of the intersection of big data with individual privacy, in relation to both the current state and possible future states of the relevant technological capabilities and associated privacy concerns.

Relevant big data include data and metadata collected, or potentially collectable, from or about individuals by entities that include the government, the private sector, and other individuals. It includes both proprietary and open data, and also data about individuals collected incidentally or accidentally in the course of other activities (e.g., environmental monitoring or the "Internet of Things").

This is a tall order, especially on the ambitious timescale requested by the President. The literature and public discussion of big data and privacy are vast, with new ideas and insights generated daily from a variety of constituencies: technologists in industry and academia, privacy and consumer advocates, legal scholars, and journalists (among others). Independently of PCAST, but informing this report, the Podesta study sponsored three public workshops at universities across the country. Limiting this report’s charge to technological, not policy, aspects of the problem narrows PCAST’s mandate somewhat, but this is a subject where technology and policy are difficult to separate. In any case, it is the nature of the subject that this report must be regarded as based on a momentary snapshot of the technology, although we believe the key conclusions and recommendations have lasting value.

1.1 Context and outline of this report

The ubiquity of computing and electronic communication technologies has led to the exponential growth of online data, from both digital and analog sources. New technological capabilities to create, analyze, and disseminate vast quantities of data raise new concerns about the nature of privacy and the means by which individual privacy might be compromised or protected.

This report discusses present and future technologies concerning this so-called “big data” as it relates to privacy concerns. It is not a complete summary of the technology concerning big data, nor a complete summary of the ways in which technology affects privacy, but focuses on the ways in which big-data and privacy interact. As an example, if Leslie confides a secret to Chris and Chris broadcasts that secret by email or texting, that might be a privacy-infringing use of information technology, but it is not a big-data issue. As another example, if oceanographic data are collected in large quantities by remote sensing, that is big data, but not, in the first instance, a privacy concern. Some data are more privacy-sensitive than others, for example, personal medical data, as distinct from personal data publicly shared by the same individual. Different technologies and policies will apply to different classes of data.

The notions of big data and the notions of individual privacy used in this report are intentionally broad and inclusive. Business consultants Gartner, Inc. define big data as “high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making,” 4 while computer scientists reviewing multiple definitions offer the more technical, “a term describing the storage and analysis of large and/or complex data sets using a series of techniques including, but not limited to, NoSQL, MapReduce, and machine learning.” 5 (See Sections 3.2.1 and 3.3.1 for discussion of these technical terms.) In a privacy context, the term "big data" typically means data about one or a group of individuals, or that might be analyzed to make inferences about individuals. It might include data or metadata collected by government, by the private sector, or by individuals. The data and metadata might be proprietary or open, they might be collected intentionally or incidentally or accidentally. They might be text, audio, video, sensor-based, or some combination. They might be data collected...
directly from some source, or data derived by some process of analysis. They might be saved for a long period of time, or they might be analyzed and discarded as they are streamed. In this report, PCAST usually does not distinguish between “data” and “information.”

The term “privacy” encompasses not only avoiding observation, or keeping one’s personal matters and relationships secret, but also the ability to share information selectively but not publicly. Anonymity overlaps with privacy, but the two are not identical. Voting is recognized as private, but not anonymous, while authorship of a political tract may be anonymous, but it is not private. Likewise, the ability to make intimate personal decisions without government interference is considered to be a privacy right, as is protection from discrimination on the basis of certain personal characteristics (such as an individual’s race, gender, or genome). So, privacy is not just about secrets.

The promise of big-data collection and analysis is that the derived data can be used for purposes that benefit both individuals and society. Threats to privacy stem from the deliberate or inadvertent disclosure of collected or derived individual data, the misuse of the data, and the fact that derived data may be inaccurate or false. The technologies that address the confluence of these issues are the subject of this report. 6

The remainder of this introductory chapter gives further context in the form of a summary of how the legal concept of privacy developed historically in the United States. Interestingly, and relevant to this report, privacy rights and the development of new technologies have long been intertwined. Today’s issues are no exception. Chapter 2 of this report is devoted to scenarios and examples, some from today, but most anticipating a near tomorrow. Yogi Berra’s much-quoted remark – “It’s tough to make predictions, especially about the future” – is germane. But it is equally true for this subject that policies based on out-of-date examples and scenarios are doomed to failure. Big-data technologies are advancing so rapidly that predictions about the future, however imperfect, must guide today’s policy development.

Chapter 3 examines the technology dimensions of the two great pillars of big data: collection and analysis. In a certain sense big data is exactly the confluence of these two: big collection meets big analysis (often termed “analytics”). The technical infrastructure of large-scale networking and computing that enables “big” is also discussed.

Chapter 4 looks at technologies and strategies for the protection of privacy. Although technology may be part of the problem, it must also be part of the solution. Many current and foreseeable technologies can enhance privacy, and there are many additional promising avenues of research.

Chapter 5, drawing on the previous chapters, contains PCAST’s perspectives and conclusions. While it is not within this report’s charge to recommend specific policies, it is clear that certain kinds of policies are technically more feasible and less likely to be rendered irrelevant or unworkable by new technologies than others. These approaches are highlighted, along with comments on the technical deficiencies of some other approaches. This chapter also contains PCAST’s recommendations in areas that lie within our charge, that is, other than policy.

### 1.2 Technology has long driven the meaning of privacy

The conflict between privacy and new technology is not new, except perhaps now in its greater scope, degree of intimacy, and pervasiveness. For more than two centuries, values and expectations relating to privacy have been continually reinterpreted and rearticulated in light of the impact of new technologies.
The nationwide postal system advocated by Benjamin Franklin and established in 1775 was a new technology designed to promote interstate commerce. But mail was routinely and opportunistically opened in transit until Congress made this action illegal in 1782. While the Constitution’s Fourth Amendment codified the heightened privacy protection afforded to people in their homes or on their persons (previously principles of British common law), it took another century of technological challenges to expand the concept of privacy rights into more abstract spaces, including the electronic. The invention of the telegraph and, later, telephone created new tensions that were slow to be resolved. A bill to protect the privacy of telegrams, introduced in Congress in 1880, was never passed. 7

It was not telecommunications, however, but the invention of the portable, consumer-operable camera (soon known as the Kodak) that gave impetus to Warren and Brandeis’s 1890 article “The Right to Privacy,” 8 then a controversial title, but now viewed as the foundational document for modern privacy law. In the article, Warren and Brandeis gave voice to the concern that “[i]nstantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that ‘what is whispered in the closet shall be proclaimed from the house-tops,’” further noting that “[f]or years there has been a feeling that the law must afford some remedy for the unauthorized circulation of portraits of private persons…” 9

Warren and Brandeis sought to articulate the right of privacy between individuals (whose foundation lies in civil tort law). Today, many states recognize a number of privacy-related harms as causes for civil or criminal legal action (further discussed in Section 1.4). 10

From Warren and Brandeis’ “right to privacy,” it took another 75 years for the Supreme Court to find, in *Griswold v. Connecticut* 11 (1965), a right to privacy in the "penumbras" and "emanations" of other constitutional protections (as Justice William O. Douglas put it, writing for the majority). 12 With a broad perspective, scholars today recognize a number of different legal meanings for “privacy.” Five of these seem particularly relevant to this PCAST report:

1. The individual’s right to keep secrets or seek seclusion (the famous “right to be left alone” of Brandeis’ 1928 dissenting opinion in *Olmstead v. United States*). 13

2. The right to anonymous expression, especially (but not only) in political speech (as in *McIntyre v. Ohio Elections Commission* 14)

3. The ability to control access by others to personal information after it leaves one’s exclusive possession (for example, as articulated in the FTC’s Fair Information Practice Principles). 15

4. The barring of some kinds of negative consequences from the use of an individual’s personal information (for example, job discrimination on the basis of personal DNA, forbidden in 2008 by the Genetic Information Nondiscrimination Act 16).

5. The right of the individual to make intimate decisions without government interference, as in the domains of health, reproduction, and sexuality (as in *Griswold*).

These are asserted, not absolute, rights. All are supported, but also circumscribed, by both statute and case law. With the exception of number 5 on the list (a right of “decisional privacy” as distinct from “informational privacy”), all are applicable in varying degrees both to citizen-government interactions and to citizen-citizen interactions. Collisions between new technologies and privacy rights have occurred in all five. A patchwork of state and federal laws have
addressed concerns in many sectors, but to date there has not been comprehensive legislation to handle these issues. Collisions between new technologies and privacy rights should be expected to continue to occur.

1.3 What is different today?

New collisions between technologies and privacy have become evident, as new technological capabilities have emerged at a rapid pace. It is no longer clear that the five privacy concerns raised above, or their current legal interpretations, are sufficient in the court of public opinion.

Much of the public’s concern is with the harm done by the use of personal data, both in isolation or in combination. Controlling access to personal data after they leave one’s exclusive possession has been seen historically as a means of controlling potential harm. But today, personal data may never be, or have been, within one’s possession – for instance they may be acquired passively from external sources such as public cameras and sensors, or without one’s knowledge from public electronic disclosures by others using social media. In addition, personal data may be derived from powerful data analyses (see Section 3.2) whose use and output is unknown to the individual. Those analyses sometimes yield valid conclusions that the individual would not want disclosed. Worse yet, the analyses can produce false positives or false negatives — information that is a consequence of the analysis but is not true or correct. Furthermore, to a much greater extent than before, the same personal data have both beneficial and harmful uses, depending on the purposes for which and the contexts in which they are used. Information supplied by the individual might be used only to derive other information such as identity or a correlation, after which it is not needed. The derived data, which were never under the individual’s control, might then be used either for good or ill.

In the current discourse, some assert that the issues concerning privacy protection are collective as well as individual, particularly in the domain of civil rights – for example, identification of certain individuals at a gathering using facial recognition from videos, and the inference that other individuals at the same gathering, also identified from videos, have similar opinions or behaviors.

Current circumstances also raise issues of how the right to privacy extends to the public square, or to quasi-private gatherings such as parties or classrooms. If the observers in these venues are not just people, but also both visible and invisible recording devices with enormous fidelity and easy paths to electronic promulgation and analysis, does that change the rules?

Also rapidly changing are the distinctions between government and the private sector as potential threats to individual privacy. Government is not just a “giant corporation.” It has a monopoly in the use of force; it has no direct competitors who seek market advantage over it and may thus motivate it to correct missteps. Governments have checks and balances, which can contribute to self-imposed limits on what they may do with people’s information. Companies decide how they will use such information in the context of such factors as competitive advantages and risks, government regulation, and perceived threats and consequences of lawsuits. It is thus appropriate that there are different sets of constraints on the public and private sectors. But government has a set of authorities – particularly in the areas of law enforcement and national security – that place it in a uniquely powerful position, and therefore the restraints placed on its collection and use of data deserve special attention. Indeed, the need for such attention is heightened because of the increasingly blurry line between public and private data.

While these differences are real, big data is to some extent a leveler of the differences between government and companies. Both governments and companies have potential access to the same sources of data and the same analytic
tools. Current rules may allow government to purchase or otherwise obtain data from the private sector that, in some cases, it could not legally collect itself, 17 or to outsource to the private sector analyses it could not itself legally perform. 18 The possibility of government exercising, without proper safeguards, its own monopoly powers and also having unfettered access to the private information marketplace is unsettling.

What kinds of actions should be forbidden both to government (Federal, state, and local, and including law enforcement) and to the private sector? What kinds should be forbidden to one but not the other? It is unclear whether current legal frameworks are sufficiently robust for today’s challenges.

1.4 Values, harms, and rights

As was seen in Sections 1.2 and 1.3, new privacy rights usually do not come into being as academic abstractions. Rather, they arise when technology encroaches on widely shared values. Where there is consensus on values, there can also be consensus on what kinds of harms to individuals may be an affront to those values. Not all such harms may be preventable or remediable by government actions, but, conversely, it is unlikely that government actions will be welcome or effective if they are not grounded to some degree in values that are widely shared.

In the realm of privacy, Warren and Brandeis in 1890 19 (see Section 1.2) began a dialogue about privacy that led to the evolution of the right in academia and the courts, later crystallized by William Prosser as four distinct harms that had come to earn legal protection. 20 A direct result is that, today, many states recognize as causes for legal action the four harms that Prosser enumerated, 21 and which have become (though varying from state to state 22) privacy “rights.” The harms are:

- **Intrusion upon seclusion.** A person who intentionally intrudes, physically or otherwise (now including electronically), upon the solitude or seclusion of another person or her private affairs or concerns, can be subject to liability for the invasion of her privacy, but only if the intrusion would be highly offensive to a reasonable person.

- **Public disclosure of private facts.** Similarly, a person can be sued for publishing private facts about another person, even if those facts are true. Private facts are those about someone’s personal life that have not previously been made public, that are not of legitimate public concern, and that would be offensive to a reasonable person.

- **“False light” or publicity.** Closely related to defamation, this harm results when false facts are widely published about an individual. In some states, false light includes untrue implications, not just untrue facts as such.

- **Misappropriation of name or likeness.** Individuals have a “right of publicity” to control the use of their name or likeness in commercial settings.

It seems likely that most Americans today continue to share the values implicit in these harms, even if the legal language (by now refined in thousands of court decisions) strikes one as archaic and quaint. However, new technological insults to privacy, actual or prospective, and a century’s evolution of social values (for example, today’s greater recognition of the rights of minorities, and of rights associated with gender), may require a longer list than sufficed in 1960.

Although PCAST’s engagement with this subject is centered on technology, not law, any report on the subject of privacy, including PCAST’s, should be grounded in the values of its day. As a starting point for discussion, albeit only a snapshot of the views of one set of technologically minded Americans, PCAST offers some possible augmentations to the established list of harms, each of which suggests a possible underlying right in the age of big data.
PCAST also believes strongly that the positive benefits of technology are (or can be) greater than any new harms. Almost every new harm is related to or “adjacent to” beneficial uses of the same technology. To emphasize this point, for each suggested new harm, we describe a related beneficial use.

- **Invasion of private communications.** Digital communications technologies make social networking possible across the boundaries of geography, and enable social and political participation on previously unimaginable scales. An individual’s right to private communication, secured for written mail and wireline telephone in part by the isolation of their delivery infrastructure, may need reaffirmation in the digital era, however, where all kinds of “bits” share the same pipelines, and the barriers to interception are often much lower. (In this context, we discuss the use and limitations of encryption in Section 4.2.)

- **Invasion of privacy in a person’s virtual home.** The Fourth Amendment gives special protection against government intrusion into the home, for example the protection of private records within the home; tort law offers protection against similar non-government intrusion. The new “virtual home” includes the Internet, cloud storage, and other services. Personal data in the cloud can be accessible and organized. Photographs and records in the cloud can be shared with family and friends, and can be passed down to future generations. The underlying social value, the “home as one’s castle,” should logically extend to one’s “castle in the cloud,” but this protection has not been preserved in the new virtual home. (We discuss this subject further in Section 2.3.)

- **Public disclosure of inferred private facts.** Powerful data analytics may infer personal facts from seemingly harmless input data. Sometimes the inferences are beneficial. At its best, targeted advertising directs consumers to products that they actually want or need. Inferences about people’s health can lead to better and timelier treatments and longer lives. But before the advent of big data, it could be assumed that there was a clear distinction between public and private information: either a fact was “out there” (and could be pointed to), or it was not. Today, analytics may discover facts that are no less private than yesterday’s purely private sphere of life. Examples include inferring sexual preference from purchasing patterns, or early Alzheimer’s disease from key-click streams. In the latter case, the private fact may not even be known to the individual in question. (Section 3.2 discusses the technology behind the data analytics that makes such inferences possible.) The public disclosure of such information (and possibly also some non-public commercial uses) seems offensive to widely shared values.

- **Tracking, stalking, and violations of locational privacy.** Today’s technologies easily determine an individual’s current or prior location. Useful location-based services include navigation, suggesting better commuter routes, finding nearby friends, avoiding natural hazards, and advertising the availability of nearby goods and services. Sighting an individual in a public place can hardly be a private fact. When big data allows such sightings, or other kinds of passive or active data collection, to be assembled into the continuous locational track of an individual’s private life, however, many Americans (including Supreme Court Justice Sotomayor, for example) perceive a potential affront to a widely accepted “reasonable expectation of privacy.”

- **Harm arising from false conclusions about individuals, based on personal profiles from big-data analytics.** The power of big data, and therefore its benefit, is often correlational. In many cases the “harms” from statistical errors are small, for example the incorrect inference of a movie preference; or the suggestion that a health issue be discussed with a physician, following from analyses that may, on average, be beneficial, even when a particular instance turns out to be a false alarm. Even when predictions are statistically valid, moreover, they may be untrue about particular individuals – and mistaken conclusions may cause harm. Society may not be willing to excuse harms caused by the uncertainties inherent in statistically valid algorithms. These harms may unfairly burden particular classes of individuals, for example, racial minorities or the elderly.

- **Foreclosure of individual autonomy or self-determination.** Data analyses about large populations can discover special cases that apply to individuals within that population. For example, by identifying differences in “learning styles,” big data may make it possible to personalize education in ways that recognize every individual’s potential and optimize that individual’s achievement. But the projection of population factors onto individuals can be misused. It is widely accepted that individuals should be able to make their own choices and pursue opportunities that are not necessarily typical, and that no one should be denied the chance to achieve more than some statistical expectation of themselves. It would offend our values if a child’s choices in video games were later used for educational tracking (for example, college admissions). Similarly offensive would be a future, akin to Philip K. Dick’s science fiction short
story adapted by Steven Spielberg in the film Minority Report, where “pre-crime” is statistically identified and punished. 25

• **Loss of anonymity and private association.** Anonymity is not acceptable as an enabler of committing fraud, or bullying, or cyber-stalking, or improper interactions with children. Apart from wrongful behavior, however, the individual’s right to choose to be anonymous is a long held American value (as, for example, the anonymous authorship of the Federalist papers). Using data to (re-) identify an individual who wishes to be anonymous (except in the case of legitimate governmental functions, such as law enforcement) is regarded as a harm. Similarly, individuals have a right of private association with groups or other individuals, and the identification of such associations may be a harm.

While in no sense is the above list intended to be complete, it does have a few intentional omissions. For example, individuals may want big data to be used “fairly,” in the sense of treating people equally, but (apart from the small number of protected classes already defined by law) it seems impossible to turn this into a right that is specific enough to be meaningful. Likewise, individuals may want the ability to know what others know about them; but that is surely not a right from the pre-digital age; and, in the current era of statistical analysis, it is not so easy to define what “know” means. This important issue is discussed in Section 3.1.2, and again taken up in chapter 5, where the attempt is to focus on actual harms done by the use of information, not by a concept as technically ambiguous as whether information is known.

# 2. Examples and Scenarios

This chapter seeks to make Chapter 1’s introductory discussion more concrete by sketching some examples and scenarios. While some of these applications of technology are in use today, others comprise PCAST’s technological prognostications about the near future, up to perhaps 10 years from today. Taken together the examples and scenarios are intended to illustrate both the enormous benefits that big data can provide and also the privacy challenges that may accompany these benefits.

In the following three sections, it will be useful to develop some scenarios more completely than others, moving from very brief examples of things happening today to more fully developed scenarios set in the future.

## 2.1 Things happening today or very soon

Here are some relevant examples:

- Pioneered more than a decade ago, devices mounted on utility poles are able to sense the radio stations being listened to by passing drivers, with the results sold to advertisers. 26
- In 2011, automatic license-plate readers were in use by three quarters of local police departments surveyed. Within 5 years, 25% of departments expect to have them installed on all patrol cars, alerting police when a vehicle associated with an outstanding warrant is in view. 27 Meanwhile, civilian uses of license-plate readers are emerging, leveraging cloud platforms and promising multiple ways of using the information collected. 28
- Experts at the Massachusetts Institute of Technology and the Cambridge Police Department have used a machine-learning algorithm to identify which burglaries likely were committed by the same offender, thus aiding police investigators. 29
- Differential pricing (offering different prices to different customers for essentially the same goods) has become familiar in domains such as airline tickets and college costs. Big data may increase the power and prevalence of this practice and may also decrease even further its transparency. 30
• The UK firm FeatureSpace offers machine-learning algorithms to the gaming industry that may detect early signs of gambling addiction or other aberrant behavior among online players. 31

• Retailers like CVS and AutoZone analyze their customers' shopping patterns to improve the layout of their stores and stock the products their customers want in a particular location. 32 By tracking cell phones, RetailNext offers bricks-and-mortar retailers the chance to recognize returning customers, just as cookies allow them to be recognized by on-line merchants. 33 Similar WiFi tracking technology could detect how many people are in a closed room (and in some cases their identities).

• The retailer Target inferred that a teenage customer was pregnant and, by mailing her coupons intended to be useful, unintentionally disclosed this fact to her father. 34

• The author of an anonymous book, magazine article, or web posting is frequently "outed" by informal crowd sourcing, fueled by the natural curiosity of many unrelated individuals. 35

• Social media and public sources of records make it easy for anyone to infer the network of friends and associates of most people who are active on the web, and many who are not. 36

• Marist College in Poughkeepsie, New York, uses predictive modeling to identify college students who are at risk of dropping out, allowing it to target additional support to those in need. 37

• The Durkheim Project, funded by the U.S. Department of Defense, analyzes social-media behavior to detect early signs of suicidal thoughts among veterans. 38

• LendUp, a California-based startup, sought to use nontraditional data sources such as social media to provide credit to underserved individuals. Because of the challenges in ensuring accuracy and fairness, however, they have been unable to proceed. 39, 40

• Insight into the spread of hospital-acquired infections has been gained through the use of large amounts of patient data together with personal information about uninfected patients and clinical staff. 41

• Individuals' heart rates can be inferred from the subtle changes in their facial coloration that occur with each beat, enabling inferences about their health and emotional state. 42

2.2 Scenarios of the near future in healthcare and education

Here are a few examples of the kinds of scenarios that can readily be constructed.

2.2.1 Healthcare: personalized medicine

Not all patients who have a particular disease are alike, nor do they respond identically to treatment. Researchers will soon be able to draw on millions of health records (including analog data such as scans in addition to digital data), vast amounts of genomic information, extensive data on successful and unsuccessful clinical trials, hospital records, and so forth. In some cases they will be able to discern that among the diverse manifestations of the disease, a subset of the patients have a collection of traits that together form a variant that responds to a particular treatment regime.

Since the result of the analysis could lead to better outcomes for particular patients, it is desirable to identify those individuals in the cohort, contact them, treat their disease in a novel way, and use their experiences in advancing the research. Their data may have been gathered only anonymously, however, or it may have been de-identified.

Solutions may be provided by specific new technologies for the protection of database privacy. These may create a protected query mechanism so individuals can find out whether they are in the cohort, or provide an alert mechanism based on the cohort characteristics so that, when a medical professional sees a patient in the cohort, a notice is generated.
2.2.2 Healthcare: detection of symptoms by mobile devices

Many baby boomers wonder how they might detect Alzheimer's disease in themselves. What would be better to observe their behavior than the mobile device that connects them to a personal assistant in the cloud (e.g., Siri or OK Google), helps them navigate, reminds them what words mean, remembers to do things, recalls conversations, measures gait, and otherwise is in a position to detect gradual declines on traditional and novel medical indicators that might be imperceptible even to their spouses?

At the same time, any leak of such information would be a damaging betrayal of trust. What are individuals' protections against such risks? Can the inferred information about individuals' health be sold, without additional consent, to third parties (e.g., pharmaceutical companies)? What if this is a stated condition of use of the app? Should information go to individuals' personal physicians with their initial consent but not a subsequent confirmation?

2.2.3 Education

Drawing on millions of logs of online courses, including both massive open on-line courses (MOOCs) and smaller classes, it will soon be possible to create and maintain longitudinal data about the abilities and learning styles of millions of students. This will include not just broad aggregate information like grades, but fine-grained profiles of how individual students respond to multiple new kinds of teaching techniques, how much help they need to master concepts at various levels of abstraction, what their attention span is in various contexts, and so forth. A MOOC platform can record how long a student watches a particular video; how often a segment is repeated, sped up, or skipped; how well a student does on a quiz; how many times he or she misses a particular problem; and how the student balances watching content to reading a text. As the ability to present different material to different students materializes in the platforms, the possibility of blind, randomized A/B testing enables the gold standard of experimental science to be implemented at large scale in these environments. 43

Similar data are also becoming available for residential classes, as learning-management systems (such as Canvas, Blackboard, or Desire2Learn) expand their roles to support innovative pedagogy. In many courses one can now get moment-by-moment tracking of the student's engagement with the course materials and correlate that engagement with the desired learning outcomes.

With this information, it will be possible not only to greatly improve education, but also to discover what skills, taught to which individuals at which points in childhood, lead to better adult performance in certain tasks, or to adult personal and economic success. While these data could revolutionize educational research, the privacy issues are complex. 44

There are many privacy challenges in this vision of the future of education. Knowledge of early performance can create implicit biases 45 that color later instruction and counseling. There is great potential for misuse, ostensibly for the social good, in the massive ability to direct students into high- or low-potential tracks. Parents and others have access to sensitive information about children, but mechanisms rarely exist to change those permissions when the child reaches majority.
2.3 Challenges to the home’s special status

The home has special significance as a sanctuary of individual privacy. The Fourth Amendment’s list, “persons, houses, papers, and effects,” puts only the physical body in the rhetorically more prominent position; and a house is often the physical container for the other three, a boundary inside of which enhanced privacy rights apply.

Existing interpretations of the Fourth Amendment are inadequate for the present world, however. We, along with the “papers and effects” contemplated by the Fourth Amendment, live increasingly in cyberspace, where the physical boundary of the home has little relevance. In 1980, a family’s financial records were paper documents, located perhaps in a desk drawer inside the house. By 2000, they were migrating to the hard drive of the home computer – but still within the house. By 2020, it is likely that most such records will be in the cloud, not just outside the house, but likely replicated in multiple legal jurisdictions – because cloud storage typically uses location diversity to achieve reliability. The picture is the same if one substitutes for financial records something like “political books we purchase,” or “love letters that we receive,” or “erotic videos that we watch.” Absent different policy, legislative, and judicial approaches, the physical sanctity of the home’s papers and effects is rapidly becoming an empty legal vessel.

The home is also the central locus of Brandeis’ “right to be left alone.” This right is also increasingly fragile, however. Increasingly, people bring sensors into their homes whose immediate purpose is to provide convenience, safety, and security. Smoke and carbon monoxide alarms are common, and often required by safety codes. Radon detectors are usual in some parts of the country. Integrated air monitors that can detect and identify many different kinds of pollutants and allergens are readily foreseeable. Refrigerators may soon be able to “sniff” for gases released from spoiled food, or, as another possible path, may be able to “read” food expiration dates from radio-frequency identification (RFID) tags in the food’s packaging. Rather than today’s annoying cacophony of beeps, tomorrow’s sensors (as some already do today) will interface to a family through integrated apps on mobile devices or display screens. The data will have been processed and interpreted. Most likely that processing will occur in the cloud. So, to deliver services the consumer wants, much data will need to have left the home.

Environmental sensors that enable new food and air safety may also be able to detect and characterize tobacco or marijuana smoke. Health care or health insurance providers may want assurance that self-declared nonsmokers are telling the truth. Might they, as a condition of lower premiums, require the homeowner’s consent for tapping into the environmental monitors’ data? If the monitor detects heroin smoking, is an insurance company obligated to report this to the police? Can the insurer cancel the homeowner’s property insurance?

To some, it seems farfetched that the typical home will foreseeably acquire cameras and microphones in every room, but that appears to be a likely trend. What can your cell phone (already equipped with front and back cameras) hear or see when it is on the nightstand next to your bed? Tablets, laptops, and many desktop computers have cameras and microphones. Motion detector technology for home intrusion alarms will likely move from ultrasound and infrared to imaging cameras – with the benefit of fewer false alarms and the ability to distinguish pets from people. Facial-recognition technology will allow further security and convenience. For the safety of the elderly, cameras and microphones will be able to detect falls or collapses, or calls for help, and be networked to summon aid.

People naturally communicate by voice and gesture. It is inevitable that people will communicate with their electronic servants in both such modes (necessitating that they have access to cameras and microphones). Companies such as PrimeSense, an Israeli firm recently bought by Apple, are developing sophisticated computer-vision software for
gesture reading, already a key feature in the consumer computer game console market (e.g., Microsoft Kinect). Consumer televisions are already among the first “appliances” to respond to gesture; already, devices such as the Nest smoke detector respond to gestures. The consumer who taps his temple to signal a spoken command to Google Glass may want to use the same gesture for the television, or for that matter for the thermostat or light switch, in any room at home. This implies omnipresent audio and video collection within the home.

All of these audio, video, and sensor data will be generated within the supposed sanctuary of the home. But they are no more likely to stay in the home than the “papers and effects” already discussed. Electronic devices in the home already invisibly communicate to the outside world via multiple separate infrastructures: The cable industry’s hardwired connection to the home provides multiple types of two-way communication, including broadband Internet. Wireline phone is still used by some home-intrusion alarms and satellite TV receivers, and as the physical layer for DSL broadband subscribers. Some home devices use the cell-phone wireless infrastructure. Many others piggyback on the home Wi-Fi network that is increasingly a necessity of modern life. Today’s smart home-entertainment system knows what a person records on a DVR, what she actually watches, and when she watches it. Like personal financial records in 2000, this information today is in part localized inside the home, on the hard drive inside the DVR. As with financial information today, however, it is on track to move into the cloud. Today, Netflix or Amazon can offer entertainment suggestions based on customers’ past key-click streams and viewing history on their platforms. Tomorrow, even better suggestions may be enabled by interpreting their minute-by-minute facial expressions as seen by the gesture-reading camera in the television.

These collections of data are benign, in the sense that they are necessary for products and services that consumers will knowingly demand. Their challenges to privacy arise both from the fact that their analog sensors necessarily collect more information than is minimally necessary for their function (see Section 3.1.2), and also because their data practically cry out for secondary uses ranging from innovative new products to marketing bonanzas to criminal exploits. As in many other kinds of big data, there is ambiguity as to data ownership, data rights, and allowed data use. Computer-vision software is likely already able to read the brand labels on products in its field of view – this is a much easier technology than facial recognition. If the camera in your television knows what brand of beer you are drinking while watching a football game, and knows whether you opened the bottle before or after the beer ad, who (if anyone) is allowed to sell this information to the beer company, or to its competitors? Is the camera allowed to read brand names when the television set is supposedly off? Can it watch for magazines or political leaflets? If the RFID tag sensor in your refrigerator usefully detects out-of-date food, can it also report your brand choices to vendors? Is this creepy and strange, or a consumer financial benefit when every supermarket can offer you relevant coupons? Or (the dilemma of differential pricing) is it any different if the data are used to offer others a better deal while you pay full price because your brand loyalty is known to be strong?

About one-third of Americans rent, rather than own, their residences. This number may increase with time as a result of long-term effects of the 2007 financial crisis, as well as aging of the U.S. population. Today and foreseeably, renters are less affluent, on average, than homeowners. The law demarcates a fine line between the property rights of landlords and the privacy rights of tenants. Landlords have the right to enter their property under various conditions, generally including where the tenant has violated health or safety codes, or to make repairs. As more data are collected within the home, the rights of tenant and landlord may need new adjustment. If environmental monitors are fixtures of the landlord’s property, does she have an unconditional right to their data? Can she sell those data? If the lease so
provides, can she evict the tenant if the monitor repeatedly detects cigarette smoke, or a camera sensor is able to
distinguish a prohibited pet?

If a third party offers facial recognition services for landlords (no doubt with all kinds of cryptographic safeguards!), can
the landlord use these data to enforce lease provisions against subletting or additional residents? Can she require such
monitoring as a condition of the lease? What if the landlord’s cameras are outside the doors, but keep track of everyone
who enters or leaves her property? How is this different from the case of a security camera across the street that is
owned by the local police?

2.4 Tradeoffs among privacy, security, and convenience

Notions of privacy change generationally. One sees today marked differences between the younger generation of
“digital natives” and their parents or grandparents. In turn, the children of today’s digital natives will likely have still
different attitudes about the flow of their personal information. Raised in a world with digital assistants who know
everything about them, and (one may hope) with wise policies in force to govern use of the data, future generations may
see little threat in scenarios that individuals today would find threatening, if not Orwellian. PCAST’s final scenario,
perhaps at the outer limit of its ability to prognosticate, is constructed to illustrate this point.

Taylor Rodriguez prepares for a short business trip. She packed a bag the night before and put it outside the front door
of her home for pickup. No worries that it will be stolen: The camera on the streetlight was watching it; and, in any case,
almost every item in it has a tiny RFID tag. Any would-be thief would be tracked and arrested within minutes. Nor is
there any need to give explicit instructions to the delivery company, because the cloud knows Taylor’s itinerary and
plans; the bag is picked up overnight and will be in Taylor’s destination hotel room by the time of her arrival.

Taylor finishes breakfast and steps out the front door. Knowing the schedule, the cloud has provided a self-driving car,
waiting at the curb. At the airport, Taylor walks directly to the gate – no need to go through any security. Nor are there
any formalities at the gate: A twenty-minute “open door” interval is provided for passengers to stroll onto the plane and
take their seats (which each sees individually highlighted in his or her wearable optical device). There are no boarding
passes and no organized lines. Why bother, when Taylor’s identity (as for everyone else who enters the airport) has
been tracked and is known absolutely? When her known information emanations (phone, RFID tags in clothes, facial
recognition, gait, emotional state) are known to the cloud, vetted, and essentially unforgeable? When, in the unlikely
event that Taylor has become deranged and dangerous, many detectable signs would already have been tracked,
detected, and acted on?

Indeed, everything that Taylor carries has been screened far more effectively than any rushed airport search today.
Friendly cameras in every LED lighting fixture in Taylor’s house have watched her dress and pack, as they do every day.
Normally these data would be used only by Taylor’s personal digital assistants, perhaps to offer reminders or fashion
advice. As a condition of using the airport transit system, however, Taylor has authorized the use of the data for
ensuring airport security and public safety.

Taylor’s world seems creepy to us. Taylor has accepted a different balance among the public goods of convenience,
privacy, and security than would most people today. Taylor acts in the unconscious belief (whether justified or not,
depending on the nature and effectiveness of policies in force) that the cloud and its robotic servants are trustworthy in
matters of personal privacy. In such a world, major improvements in the convenience and security of everyday life
become possible.
3. Collection, Analytics, and Supporting Infrastructure

Big data is big in two different senses. It is big in the quantity and variety of data that are available to be processed. And, it is big in the scale of analysis (“analytics”) that can be applied to those data, ultimately to make inferences. Both kinds of “big” depend on the existence of a massive and widely available computational infrastructure, one that is increasingly being provided by cloud services. This chapter expands on these basic concepts.

3.1 Electronic sources of personal data

Since early in the computer age, public and private entities have been assembling digital information about people. Databases of personal information were created during the days of “batch processing.” Indeed, early descriptions of database technology often talk about personnel records used for payroll applications. As computing power increased, more and more business applications moved to digital form. There now are digital telephone-call records, credit-card transaction records, bank-account records, email repositories, and so on. As interactive computing has advanced, individuals have entered more and more data about themselves, both for self-identification to an online service and for productivity tools such as financial-management systems.

These digital data are normally accompanied by “metadata” or ancillary data that explain the layout and meaning of the data they describe. Databases have schemas and email has headers, as do network packets. As data sets become more complex, so do the attached metadata. Included in the data or metadata may be identifying information such as account numbers, login names, and passwords. There is no reason to believe that metadata raise fewer privacy concerns than the data they describe.

In recent times, the kinds of electronic data available about people have increased substantially, in part because of the emergence of social media and in part because of the growth in mobile devices, surveillance devices, and a diversity of networked sensors. Today, although they may not be aware of it, individuals constantly emit into the environment information whose use or misuse may be a source of privacy concerns. Physically, these information emanations are of two types, which can be called “born digital” or “born analog.”

3.1.1 “Born digital” data

When information is “born digital,” it is created, by us or by a computer surrogate, specifically for digital use – that is, for use by a computer or data-processing system. Examples of data that are born digital include:

- email and text messaging
- input via mouse-clicks, taps, swipes, or keystrokes on a phone, tablet, computer, or video game; that is, data that people intentionally enter into a device
- GPS location data
- metadata associated with phone calls: the numbers dialed from or to, the time and duration of calls
- data associated with most commercial transactions: credit-card swipes, bar-code reads, reads of RFID tags (as used for anti-theft and inventory control)
- data associated with portal access (key card or ID badge reads) and toll-road access (remote reads of RFID tags)
- metadata that our mobile devices use to stay connected to the network, including device location and status
increasingly, data from cars, televisions, appliances: the “Internet of Things”

Consumer-tracking data provide an example of born-digital data that has become economically important. It is generally possible for companies to aggregate large amounts of data and then use those data for marketing, advertising, or many other activities. The traditional mechanism has been to use cookies, small data files that a browser can leave on a user’s computer (pioneered by Netscape two decades ago). The technique is to leave a cookie when a user first visits a site and then be able to correlate that visit with a subsequent event. This information is very valuable to retailers and forms the basis of many of the advertising businesses of the last decade. There has been a variety of proposals to regulate such tracking, and many countries require opt-in permission before this tracking is done. Cookies involve relatively simple pieces of information that proponents represent as unlikely to be abused. Although not always aware of the process, people accept such tracking in return for a free or subsidized service. At the same time, cookie-free alternatives are sometimes available. Even without cookies, so-called “fingerprinting” techniques can often identify a user’s computer or mobile device uniquely by the information that it exposes publicly, such as the size of its screen, its installed fonts, and other features. Most technologists believe that applications will move away from cookies, that cookies are too simple an idea, and that there are better analytics coming and better approaches being invented. The economic incentives for consumer tracking will remain, however, and big data will allow for more precise responses.

Tracking is also the enabling technology of some more nefarious uses. Unfortunately, many social networking apps begin by taking a person’s contact list and spamming all the recipients with advertising for the app. This technique is often abused, especially by small start-ups who may assess the value gained by reaching new customers as being greater than the value lost to their reputation for honoring privacy.

All information that is born digital shares certain characteristics. It is created in identifiable units for particular purposes. These units are in most cases “data packets” of one or another standard type. Since they are created by intent, the information that they contain is usually limited, for reasons of efficiency and good engineering design, to support the immediate purpose for which they are collected.

When data are born digital, privacy concerns can arise in two different modes, one obvious (“over-collection”), the other more recent and subtle (“data fusion”). Over-collection occurs when an engineering design intentionally, and sometimes clandestinely, collects information unrelated to its stated purpose. While your smartphone could easily photograph and transmit to a third party your facial expression as you type every keystroke of a text message, or could capture all keystrokes, thereby recording text that you had deleted, these would be inefficient and unreasonable software design choices for the default text-messaging app. In that context they would be instances of over-collection.

A recent example of over-collection was the Brightest Flashlight Free phone app, downloaded by more than 50 million users, which passed back to its vendor its location every time the flashlight was used. Not only is location information unnecessary for the illumination function of a flashlight, but it also discloses personal information that the user might wish to keep private. The Federal Trade Commission issued a complaint because the fine print on the notice-and-consent screen (see Section 4.3) had neglected to disclose that location information, whose collection was disclosed, would be sold to third parties, such as advertisers. One sees in this example the limitations of the notice-and-consent framework: A more detailed initial fine-print disclosure by Brightest Flashlight Free, which almost no one would have actually read, would likely have forestalled any FTC action without much affecting the number of downloads.

https://semanticommunity.info/Data_Science/Data_Science_for_the_Big_Data_Review
Updated: Thu, 25 Jul 2019 08:57:25 GMT
Powered by mindtouch
In contrast to over-collection, data fusion occurs when data from different sources are brought into contact and new, often unexpected, phenomena emerge (see Section 3.1). Individually, each data source may have been designed for a specific, limited purpose. But when multiple sources are processed by techniques of modern statistical data mining, pattern recognition, and the combining of records from diverse sources by virtue of common identifying data, new meanings can be found. In particular, data fusion frequently results in the identification of individual people (that is, the association of events with unique personal identities), the creation of data-rich profiles of an individual, and the tracking of an individual’s activities over days, months, or years.

By definition, the privacy challenges from data fusion do not lie in the individual data streams, each of whose collection, real-time processing, and retention may be wholly necessary and appropriate for its overt, immediate purpose. Rather, the privacy challenges are emergent properties of our increasing ability to bring into analytical juxtaposition large, diverse data sets and to process them with new kinds of mathematical algorithms.

3.1.2 Data from sensors

Turn now to the second broad class of information emanations. One can say that information is “born analog” when it arises from the characteristics of the physical world. Such information does not become accessible electronically until it impinges on a “sensor,” an engineered device that observes physical effects and converts them to digital form. The most common sensors are cameras, including video, which sense visible electromagnetic radiation; and microphones, which sense sound and vibration. There are many other kinds of sensors, however. Today, cell phones routinely contain not only cameras, microphones, and radios but also analog sensors for magnetic fields (3-D compass) and motion (acceleration). Other kinds of sensors include those for thermal infrared (IR) radiation; air quality, including the identification of chemical pollutants; barometric pressure (and altitude); low-level gamma radiation; and many other phenomena.

Examples of born-analog data providing personal information and in use today include:

- the voice and/or video content of a phone call – born analog but immediately converted to digital by the phone’s microphone and camera
- personal health data such as heartbeat, respiration, and gait, as sensed by special-purpose devices (Fitbit has been a leading provider) or cell-phone apps
- cameras/sensors in televisions and video games that interpret gestures by the user
- video from security surveillance cameras, mobile phones, or overhead drones
- imaging infrared video that can see in what people perceive as total darkness (and also see evanescent traces of past events, so-called heat scars)
- microphone networks in cities, used to detect and locate gunshots and for public safety
- cameras/microphones in classrooms and other meeting rooms
- ultrasonic motion detectors
- medical imaging, CT, and MRI scans, ultrasonic imaging
- opportunistically collected chemical or biological samples, notably trace DNA (today requiring slow, offline analysis, but foreseeably more nimble)
- synthetic aperture radar (SAR), which can image through clouds and, under some conditions, see inside of non-metallic structures
When data are born analog, they are likely to contain more information than the minimum necessary for their immediate purpose, for several valid reasons. One is that the desired information (“signal”) must be sensed in the presence of unwanted extraneous information (“noise”). The technologies typically work by sensing the environment (“signal plus noise”) with high precision, so that mathematical techniques can then be applied that will separate the two even in the worst anticipated case when the signal is smallest or the noise is largest.

Another reason is technological convergence. For example, as the cameras in cell phones become smaller and cheaper, the use of identical components in other products becomes a favored design choice, even when full images are not needed. Where a big-screen television today has separate sensors for its IR remote control, room brightness, and motion detection (a feature that turns off the picture when no one is in the room), plus a true video camera in the add-on game console, tomorrow’s model may integrate all of these functions in a single, cheap, high-resolution, IR-sensitive camera, a few millimeters in size.

In addition to the information available from digital and analog sources consciously intended to provide information about people, inadvertent disclosure abounds from the emerging “Internet of Things,” an amalgamation of sensors whose primary purpose is enhanced by “smart” network-connected computational capabilities. Examples include “smart” thermostats that detect human presence and adjust air temperatures accordingly, “smart” automobile-ignition systems, and locking systems that are biometrically triggered.

The privacy challenges of born-analog data are somewhat different from those of born-digital data. Where over-collection (as was defined above) is an irrational design choice for the principled digital designer – and therefore an identifiable red flag for privacy issues – over-collection in the analog domain can be a robust and economical design choice. A consequence is that born-analog data will often contain information that was not originally expected. Unexpected information could in many cases lead to unanticipated beneficial products and services, but it could also give opportunities for unanticipated misuse.

As a concrete example, one might consider three key parameters of video imaging: resolution (how many pixels in the image), contrast ratio (how well can the image see into dark regions), and photometric precision (how accurate is the image in brightness and color). All three parameters have improved by orders of magnitude and are likely to keep improving. Today, with special cameras, one can image a cityscape from a high rooftop and see clearly into every facing house and apartment window within several miles. 62 Or, already mentioned, the ability exists to sense remotely the pulse of an individual, giving information on health status and emotional state. 63

It is foreseeable, perhaps inevitable, that these capabilities will be present in every cell phone and securitysurveillance camera, or every wearable computer device. (Imagine the process of negotiating the price for a car, or negotiating an international trade agreement, when every participant’s Google Glass (or security camera or TV camera) is able to monitor and interpret the autonomic physiological state of every other participant, in real time.) It is unforeseeable what other unexpected information also lies in signals from the same sensors.

Once they enter the digital world, born-analog data can be fused and mined along with born-digital data. For example, facial-recognition algorithms, which might be error-prone in isolation, may yield nearly perfect identity tracking when they can be combined with born-digital data from cell phones (including unintended emanations), point-of-sale transactions,
RFID tags, and so forth; and also with other born-analog data such as vehicle tracking (e.g., from overhead drones) and automated license-plate reading. Biometric data can provide identity information that enhances the profile of an individual even more, and data on behavior (as from social networks) are being used to analyze attitudes or emotions (“sentiment analysis,” for individuals or groups 64). In short, more and more information can be captured and put in a quantified format so it can be tabulated and analyzed. 65

3.2 Big data analytics

Analytics is what makes big data come alive. Without analytics, big datasets could be stored, and they could be retrieved, wholly or selectively. But what comes out would be exactly what went in. Analytics, comprising a number of different computational technologies, is what fuels the big-data revolution. 66 Analytics is what creates the new value in big datasets, vastly more than the sum of the values of the parts. 67

3.2.1 Data mining

Data-mining, sometimes loosely equated to analytics but actually only a subset of it, refers to a computational process that discovers patterns in large data sets. It is a convergence of many fields of academic research in both applied mathematics and computer science, including statistics, databases, artificial intelligence, and machine learning. Like other technologies, advances in data mining have a research and development stage, in which new algorithms and computer programs are developed, and they have subsequent phases of commercialization and application.

Data mining algorithms can be trained to find patterns either by supervised learning, so-called because the algorithm is seeded with manually curated examples of the pattern to be recognized, or by unsupervised learning, where the algorithm tries to find related pieces of data without prior seeding. A recent success of unsupervised-learning algorithms was a program that, searching millions of images on the web, figured out on its own that “cat” was a much-posted category. 68

The desired output of data mining can take several forms, each with its own specialized algorithms. 69

- Classification algorithms attempt to assign objects or events to known categories. For example, a hospital might want to classify discharged patients as high, medium, or low risk for readmission.
- Clustering algorithms group objects or events into categories by similarity, as in the “cat” example above.
- Regression algorithms (also called numerical prediction algorithms) try to predict numerical quantities. For example, a bank may want to predict, from the details in a loan application, the probability of a default.
- Association techniques try to find relationships between items in their data set. Amazon’s suggested products and Netflix’s suggested movies are examples.
- Anomaly-detection algorithms look for untypical examples within a data set, for example, detecting fraudulent transactions on a credit-card account.
- Summarization techniques attempt to find and present salient features in data. Examples include both simple statistical summaries (e.g., average student test scores by school and teacher), and higher-level analysis (e.g., a list of key facts about an individual as gleaned from all web postings that mention her).

Data mining is sometimes confused with machine learning, the latter a broad subfield of computer science in academic and industrial research. 70 Data mining makes use of machine learning, as well as other disciplines, while machine learning has applications to fields other than data mining, for example, robotics.
There are limitations, both practical and theoretical, to what data mining can accomplish, as well as limits to how accurate it can be. It may reveal patterns and relationships, but it usually cannot tell the user the value or significance of these patterns. For example, supervised learning based on the characteristics of known terrorists might find similar persons, but they might or might not be terrorists; and it would miss different classes of terrorists who don’t fit the profile.

Data mining can identify relationships between behaviors and/or variables, but these relationships do not always indicate causality. If people who live under high-voltage power lines have higher morbidity, it might mean that power lines are a hazard to public health; or it might mean that people who live under power lines tend to be poor and have inadequate access to health care. The policy implications are quite different. While so-called confounding variables (in this example, income) can be corrected for when they are known and understood, there is no sure way to know whether all of them have been identified. Imputing true causality in big data is a research field in its infancy. 71

Many data analyses yield correlations that might or might not reflect causation. Some data analyses develop imperfect information, either because of limitations of the algorithms, or by the use of biased sampling. Indiscriminate use of these analyses may cause discrimination against individuals or a lack of fairness because of incorrect association with a particular group. 72 In using data analyses, particular care must be taken to protect the privacy of children and other protected groups.

Real-world data are incomplete and noisy. These data-quality issues lower the performance of data-mining algorithms and obscure outputs. When economics allow, careful screening and preparation of the input data can improve the quality of results, but this data preparation is often labor intensive and expensive. Users, especially in the commercial sector, must trade off cost and accuracy, sometimes with negative consequences for the individual represented in the data. Additionally, real-world data can contain extreme events or outliers. Outliers may be real events that, by chance, are over-represented in the data; or they may be the result of data-entry or data-transmission errors. In both cases they can skew the model and degrade performance. The study of outliers is an important research area of statistics.

3.2.2 Data fusion and information integration

Data fusion is the merging of multiple heterogeneous datasets into one homogeneous representation so that they can be better processed for data mining and management. Data fusion is used in a number of technical domains such as sensor networks, video/image processing, robotics and intelligent systems, and elsewhere.

Data integration is differentiated from data fusion in that integration more broadly combines data sets and retains the larger set of information. In data fusion, there is usually a reduction or replacement technique. Data fusion is facilitated by data interoperability, the ability for two systems to communicate and exchange data.

Data fusion and data integration are key techniques for business intelligence. Retailers are integrating their online, in-store, and catalog sales databases to create more complete pictures of their customers. Williams-Sonoma, for example, has integrated customer databases with information on 60 million households. Variables including household income, housing values, and number of children are tracked. It is claimed that targeted emails based on this information yield ten to 18 times the response rate of emails that are not targeted. 73 This is a simple illustration of how more information can lead to better inferences. Techniques that can help to preserve privacy are emerging. 74
There is a great amount of interest today in multi-sensor data fusion. The biggest technical challenges being tackled today, generally through development of new and better algorithms, relate to data precision/resolution, outliers and spurious data, conflicting data, modality (both heterogeneous and homogeneous data) and dimensionality, data correlation, data alignment, association within data, centralized vs. decentralized processing, operational timing, and the ability to handle dynamic vs. static phenomena. Privacy concerns may arise from sensor fidelity and precision as well as correlation of data from multiple sensors. A single sensor’s output might not be sensitive, but the combination from two or more may raise privacy concerns.

3.2.3 Image and speech recognition

Image- and speech-recognition technologies are able to extract information, in some limited cases approaching human understanding, from massive corpuses of still images, videos, and recorded or broadcast speech.

Urban-scene extraction can be accomplished using a variety of data sources from photos and videos to ground based LiDAR (a remote-sensing technique using lasers). In the government sector, city models are becoming vital for urban planning and visualization. They are equally important for a broad range of academic disciplines including history, archeology, geography, and computer-graphics research. Digital city models are also central to popular consumer mapping and visualization applications such as Google Earth and Bing Maps, as well as GPS-enabled navigation systems. Scene extraction is an example of the inadvertent capture of personal information and can be used for data fusion that reveals personal information.

Facial-recognition technologies are beginning to be practical in commercial and law-enforcement applications. They are able to acquire, normalize, and recognize moving faces in dynamic scenes. Real-time video surveillance with single-camera systems (and some with multi-camera systems, which can both recognize objects and analyze activity) has a wide variety of applications in both public and private environments, such as homeland security, crime prevention, traffic control, accident prediction and detection, and monitoring patients, the elderly, and children at home. Depending on the application, use of video surveillance is at varying levels of deployment.

Additional capabilities of image recognition include

- Video summarization and scene-change detection (that is, picking the small number of images that summarize a period of time)
- Precise geolocation in imagery from satellites or drones
- Image-based biometrics ? Human-in-the-loop surveillance systems
- Re-identification of persons and vehicles, that is, tracking the same person or vehicle as it moves from sensor to sensor
- Human-activity recognition of various kinds
- Semantic summarization (that is, converting pictures into text summaries)

Although systems are expected to become able to track objects across camera views and detect unusual activities in a large area by combining information from multiple sources, re-identification of objects remains hard to do (a challenge for inter-camera tracking), as is video surveillance in crowded environments.
Although the data they use are often captured in public areas, scene-extraction technologies like Google Street View have triggered privacy concerns. Photos captured for use in Street View may contain sensitive information about people who are unaware they are being observed and photographed. 81

Social-media data can be used as an input source for scene extraction techniques. When these data are posted, however, users are unlikely to know that their data would be used in these aggregated ways and that their social media information (although public) might appear synthesized in new forms. 82

Automated speech recognition has existed since at least the 1950s, 83 but recent developments over the last 10 years have allowed for novel new capabilities. Spoken text (e.g., news broadcasters reading part of a document) can today be recognized with accuracy higher than 95 percent using state-of-the-art techniques. Spontaneous speech is much harder to recognize accurately. In recent years there has been a dramatic increase in the corpuses of spontaneous speech data available to researchers, which has allowed for improved accuracy.

Over the next few years speech-recognition interfaces will be in many more places. For example, multiple companies are exploring speech recognition to control televisions and cars, to find a show on TV, or to schedule a DVR recording. Researchers at Nuance say they are actively planning how speech technology would have to be designed to be available on wearable computers. 84 Google has already implemented some of this basic functionality in its Google Glass product, and Microsoft’s Xbox One system already integrates machine vision and multi-microphone audio input for controlling system functions.

3.2.4 Social-network analysis

Social-network analysis refers to the extraction of information from a variety of interconnecting units under the assumption that their relationships are important and that the units do not behave autonomously. 85 Social networks often emerge in an online context. The most obvious examples are dedicated online social media platforms, such as Facebook, LinkedIn and Twitter, which provide new access to social interaction by allowing users to connect directly with each other over the Internet to communicate and share information. Offline human social networks may also leave analyzable digital traces, such as in phone-call metadata records that record which phones have exchanged calls or texts, and for how long. Analysis of social networks is increasingly enabled by the rising collection of digital data that links people together, especially when it is correlated to other data or metadata about the individual. 86 Tools for such analysis are being developed and made available, 87 motivated in part by the growing amount of social network content accessible through open applicationprogramming interfaces to online social-media platforms. This sort of analysis is an active arena for research.

Social-network analysis complements analysis of conventional databases, and some of the techniques used (e.g., clustering in association networks) can be used in either context. Social-network analysis can be more powerful because of the easy association of diverse kinds of information (i.e., considerable data fusion is possible). It lends itself to visualization of the results, which aids in interpreting the results of the analysis. It can be used to learn about people through their association with others, in a context of people’s tendency to associate with others who have some similarities to themselves. 88

Social-network analysis is yielding results that may surprise people. In particular, unique identification of an individual is easier than from database analysis alone. Moreover, it is achieved through more diverse kinds of data than many people
may understand, contributing to the erosion of anonymity. 89 The structure of an individual’s network is unique and itself serves as an identifier; co-occurrence in time and space is a significant means of identification; and, as discussed elsewhere in this report, different kinds of data can be combined to foster identification. 90

Social-network analysis is used in criminal forensic investigations to understand the links, means, and motives of those who may have committed crimes. In particular, social-network analysis has been used to better understand covert terrorist networks, whose dynamics may be different from those of overt networks. 91

In the realm of commerce, it is well-understood that what a person’s friends like or buy can influence what he or she might buy. For example, in 2010, it was reported that having one iPhone-owning friend makes a person three times more likely to own an iPhone than otherwise. A person with two iPhone-owning friends was five times more likely to have one. 92 Such correlations emerge in social-network analysis and can be used to help predict product trends, tailor marketing campaigns towards products an individual may be more likely to want, and target customers (said to have higher “network value”) with a central role (and a large amount of influence) in a social network. 93

Because disease is commonly spread via direct contact between individuals (humans or animals), understanding social networks through whatever proxies are available can suggest possible direct contacts and thereby assist in monitoring and stemming the outbreak of disease.

A recent study by researchers at Facebook analyzed the relationship between geographic location of individual users and that of their friends. From this analysis, they were able to create an algorithm to predict the location of an individual user based upon the locations of a small number of friends in their network, with higher accuracy than simply looking at the user’s IP address. 94

There are many commercial “social listening” services, such as Radian6/Salesforce Cloud, Collective Intellect, Lithium, and others, that mine data from social-networking feeds for use in business intelligence. 95 Coupled with social-network analysis, this information can be used to evaluate changing influences and the spread of trends between individuals and communities to inform marketing strategies.

### 3.3 The infrastructure behind big data

Big-data analytics requires not just algorithms and data, but also physical platforms where the data are stored and analyzed. The related security services used for personal data (see Sections 4.1 and 4.2) are also an essential component of the infrastructure. Once available only to large organizations, this class of infrastructure is now available through “the cloud” to small businesses and to individuals. To the extent that the software infrastructure is widely shared, privacy-preserving infrastructure services can also be more readily used.

#### 3.3.1 Data centers

One way to think about big-data platforms is in physical units of “data centers.” In recent years, data centers have become almost standard commodities. A typical data center is a large, warehouse-like building on a concrete slab the size of a few football fields. It is located with good access to cheap electric power and to a fiber-optic, Internet-backbone connection, usually in a rural or isolated area. The typical center consumes 20-40 megawatts of power (the equivalent of a city with 20,000-40,000 residents) and today houses some tens of thousands of servers and hard-disk drives, totaling
some tens of petabytes. Worldwide, there are roughly 6000 data centers of this scale, about half in the United States.

Data centers are the physical locus of big data in all its forms. Large data collections are often replicated in multiple data centers to improve both performance and robustness. There is a growing marketplace in selling data-center services.

Specialized software technology allows the data in multiple data centers (and spread across tens of thousands of processors and hard-disk drives) to cooperate in performing the tasks of data analytics, thereby providing both scaling and better performance. For example, MapReduce (originally a proprietary technology of Google, but now a term used generically) is a programming model for parallel operations across a practically unlimited number of processors; Hadoop is a popular open-source programming platform and program library based on the same ideas; NoSQL (the name derived from “not Structured Query Language”) is a set of database technologies that relaxes many of the restrictions of traditional, "relational" databases and allows for better scalability across the many processors in one or more data centers. Contemporary research is aimed at the next generation beyond Hadoop. One path is represented by Accumulo, initiated by the National Security Agency and transitioned to the open-source Apache community. Another is the Berkeley Data Analytics Stack, an open-source platform that outperforms Hadoop by a factor of 100 for memory-intensive data analytics and is being used by such companies as Foursquare, Conviva, Klout, Quantifind, Yahoo, and Amazon Web Services. Sometimes termed “NoHadoop” (to parallel the movement from SQL to NoSQL), technologies that fit this trend include Google’s Dremel, MPI (typically used in supercomputing), Pregel (for graphs), and Cloudscale (for realtime analytics).

3.3.2 The cloud

The “cloud” is not just the world inventory of data centers (although much of the public may think of it as such). Rather, one way of understanding the cloud is as a set of platforms and services made possible by the physical commoditization of data centers. When one says that data are “in the cloud," one refers not just to the physical hard-disk drives that exist (somewhere!) with the data, but also to the complex infrastructure of application programs, middleware, networking protocols, and (not least) business models that allow that data to be ingested, accessed, and utilized, all with costs that are competitively allocated. The commercial entities that, in aggregate, provision the cloud exist in an ecosystem that has many hierarchical levels and many different coexisting models of value added. There may be several handoffs of responsibility between the end user and the physical data center.

Today’s cloud providers offer some security benefits (and through that, privacy benefits) as compared to yesterday’s conventional corporate data centers or small-business computers. These services may include better physical protection and monitoring, as well as centralized support staffing, training, and oversight. Cloud services also pose new challenges for security, a subject of current research. Both benefits and risks come from the centralization of resources: More data are held by a given entity (albeit distributed across multiple servers or sites), and a cloud provider can perform better than separately held data centers by applying high standards to recruiting and managing people and systems.

Usage of the cloud and individual interactions with it (whether witting or not) are expected to increase dramatically in coming years. The rise of both mobile apps, reinforcing the use of cell phones and tablets as platforms, and broadly distributed sensors is associated with the growing use of cloud systems for storing, processing, and otherwise acting on information contributed by dispersed devices. Although progress in the mobile environment improves the usability of
mobile cloud applications, it may be detrimental to privacy to the extent that it more effectively hides information exchange from the user. As more core mobile functionality is transitioned to the cloud, larger amounts of information will be exchanged, and users may be surprised by the nature of the information that no longer remains localized to their cell phone. For example, cloud-based screen rendering (or “virtualized screens”) for cell phones would mean that the images shown on a cell-phone screen will actually be calculated on the cloud and transmitted to the mobile device. This means all the images on the screen of the mobile device can be accessed and manipulated from the cloud.

Cloud architectures are also being used increasingly to support big-data analytics, both by large enterprises (e.g., Google, Amazon, eBay) and by small entities or individuals who make ad hoc or routine use of public cloud platforms (e.g., Amazon Web Services, Google Cloud Platform, Microsoft Azure) in lieu of acquiring their own infrastructure. Social-media services such as Facebook and Twitter are deployed and analyzed by their providers using cloud systems. These uses represent a kind of democratization of analytics, with the potential to facilitate new businesses and more. Prospects for the future include exploration of options for federating or interconnecting cloud applications and for reducing some of the heterogeneity in application-programming interfaces for cloud applications.  

4. Technologies and Strategies for Privacy Protection

Data come into existence, are collected, and are possibly processed immediately (including adding “metadata”), possibly communicated, possibly stored (locally, remotely, or both), possibly copied, possibly analyzed, possibly communicated to users, possibly archived, possibly discarded. Technology at any of these stages can affect privacy positively or negatively.

This chapter focuses on the positive and assesses some of the key technologies that can be used in service of the protection of privacy. It seeks to clarify the important distinctions between privacy and (cyber-)security, as well as the vital, but yet limited, role that encryption technology can play. Some older techniques, such as anonymization, while valuable in the past, are seen as having only limited future potential. Newer technologies, some entering the marketplace and some requiring further research, are summarized.

4.1 The relationship between cybersecurity and privacy

Cybersecurity is a discipline, or set of technologies, that seeks to enforce policies relating to several different aspects of computer use and electronic communication.  

- identity and authentication: Are you who you say you are?
- authorization: What are you allowed to do?
- availability: Can attackers interfere with authorized functions?
- confidentiality: Can data or communications be (passively) copied by someone not authorized to do so?
- integrity: Can data or communications be (actively) changed or manipulated by someone not authorized?
- non-repudiation, auditability: Can actions (payments may provide the best example) later be shown to have occurred?

Good cybersecurity enforces policies that are precise and unambiguous. Indeed, such clarity of policy, expressible in mathematical terms, is a necessary prerequisite for the Holy Grail of cybersecurity, “provably secure” systems. At present, provable security exists only in very limited domains, for example, for certain functions on some kinds of
computer chips. It is a goal of cybersecurity research to extend the scope of provably secure systems to larger and larger domains. Meanwhile, practical cybersecurity draws on the emerging principles of such research, but it is guided even more by practical lessons learned from known failures of cybersecurity. The realistic goal is that the practice of cybersecurity should be continuously improving so as to be, in most places and at most of the time, ahead of the evolving threat.

Poor cybersecurity is clearly a threat to privacy. Privacy can be breached by failure to enforce confidentiality of data, by failure of identity and authentication processes, or by more complex scenarios such as those compromising availability.

Security and privacy share a focus on malice. The security of data can be compromised by inadvertence or accident, but it can also be compromised because some party acted knowingly to achieve the compromise – in the language of security, committed an attack. Substituting the words “breach” or “invasion” for “compromise” or “attack,” the same concepts apply to privacy. Even if there were perfect cybersecurity, however, privacy would remain at risk. Violations of privacy are possible even when there is no failure in computer security. If an authorized individual chooses to misuse (e.g., disclose) data, what is violated is privacy policy, not security policy. Or, as we have discussed (see Section 3.1.1), privacy may be violated by the fusion of data – even if performed by authorized individuals on secure computer systems.

Privacy is different from security in other respects. For one thing, it is harder to codify privacy policies precisely. Arguably this is because the presuppositions and preferences of human beings have greater diversity than the useful scope of assertions about computer security. Indeed, how to codify human privacy preferences is an important, nascent area of research.

When people provide assurance (at some level) that a computer system is secure, they are saying something about applications that are not yet invented: They are asserting that technological design features already in the machine today will prevent such application programs from violating pertinent security policies in that machine, even tomorrow. Assurances about privacy are much more precarious. Since not-yet-invented applications will have access to not-yet-imagined new sources of data, as well as to not-yet-discovered powerful algorithms, it much harder to provide, today, technological safeguards against a new route to violation of privacy tomorrow. Security deals with tomorrow’s threats against today’s platforms. That is hard enough. But privacy deals with tomorrow’s threats against tomorrow’s platforms, since those “platforms” comprise not just hardware and software, but also new kinds of data and new algorithms.

Computer scientists often work from the basis of a formal policy for security, just as engineers aim to describe something explicitly so that they can design specific ways to deal with it by purely technical means. As more computer scientists begin to think about privacy, there is increasing attention to formal articulation of privacy policy. To caricature, you have to know what you are doing to know whether what you are doing is doing the right thing. Research addressing the challenges of aligning regulations and policies with software specifications includes formal languages to express policies and system requirements; tools to reason about conflicts, inconsistencies, and ambiguities within and among policies and software specifications; methods to enable requirements engineers, business analysts, and software developers to analyze and refine policy into measurable system specifications that can be monitored over time; formalizing and enforcing privacy through auditing and accountability systems; privacy compliance in big-data systems; and formalizing and enforcing purpose restrictions.
4.2 Cryptography and encryption

Cryptography comprises a set of algorithms and system-design principles, some well-developed and others nascent, for protecting data. Cryptography is a field of knowledge whose products are encryption technology. With well-designed protocols, encryption technology is an inhibitor to compromising privacy, but it is not a “silver bullet.”

4.2.1 Well Established encryption technology

Using cryptography, readable data of any kind, termed plaintext, are transformed into what are, for all intents and purposes, incomprehensible strings of provably random bits, so-called ciphertext. Ciphertext requires no security protection of any kind. It can be stored in the cloud or sent anywhere that is convenient. It can be sent promiscuously to both the NSA and Russian FSB. If they have only ciphertext — and if it was properly generated in a precise mathematical sense — it is useless to them. They can neither read the data nor compute with it. What is needed to decrypt, to turn ciphertext back into the original plaintext, is a “key,” which is in practice a string of bits that is supposed to be known to (or computable by) only authorized users. Only with the key can encrypted data be used, i.e., their value read.

In the context of protecting privacy, it is primarily not the cryptography that is of concern. Rather, compromises of data will occur in one of two main ways:

- Data can be stolen, or mistakenly shared, before they have been encrypted or after they have been decrypted. Many attacks on supposedly encrypted data are actually attacks on machines that contain – however briefly – unencrypted plaintext. For example, in Target’s 2013 breach of one hundred million debit card number and personal-identification numbers (PINs), the PINs were present in unencrypted form only ephemerally. They were stolen nonetheless.
- Keys must be authorized, generated, distributed, and used. At every stage of a key’s life, it is potentially open to compromise or misuse that can ultimately compromise the data that the key was intended to protect. No system based on encryption is secure, of course, if persons with access to private keys can be coerced into sharing them.

Until the 1970s, keys were distributed physically, on paper or computer media, protected by registered mail, armed guards, or anything in between. The invention of “public-key cryptography” changed everything. Public-key cryptography, as the name implies, allows individuals to broadcast publicly their personal key. But this public key is only an encryption key, useful for turning plaintext into ciphertext that is meaningless to others. Its corresponding “private key,” used to transform ciphertext to plaintext, is still kept secret by the recipient. Public-key cryptography thus turns the problem of key distribution into a problem of identity determination. Alice’s messages (encrypted data transmissions) to Bob are completely protected by Bob’s public key – but only if Alice is certain that it is really Bob’s public key that she is using, and not the public key of someone merely masquerading as Bob.

Luckily, public-key cryptography also provides some techniques for helping to establish identity, namely the electronic “signing” of messages to document their authenticity. Electronic signatures, in turn, enable messages of the form “I, a person of authority known as X, certify that the following is really the public key of subordinate person Y. (Signed) X.” Messages like this are termed certificates. Certificates can be cascaded, with A certifying the identity of B, who certifies C, and so on. Certificates essentially transform the identity problem from one of validating the identity of millions of possible Y’s to validating the identity of much smaller number of top-level certificate authorities (CAs). Yet it is a matter of concern that more than 100 top-level CAs are widely recognized (e.g., accepted by most all web browsers), because there may be several intermediate steps in the hierarchy of certificates from a CA to a user, and at every step a private
key must be protected by some signer on some computer. The compromise of this private key potentially compromises the privacy of all users lower down the chain – because forged certificates of identity can now be created. Such exploits have been seen. For example, the 2011 apparent theft of a Dutch CA’s private key compromised the privacy of potentially all government records in the Netherlands. 113, 114

Many major companies have recently introduced or strengthened their use of encryption to transmit data. 115 Some are now using “(perfect) forward secrecy,” a variant of public-key cryptography that ensures that the compromise of an individual’s private key can compromise only messages that he receives subsequently, while the confidentiality of past conversations is maintained, even if their cryptotext was previously recorded by the same eavesdropper now in possession of the purloined private key. 116

4.2.2 Encryption frontiers

The technologies thus far mentioned enable the protection of data both in storage and in transit, allowing those data to be fully decrypted by users who either (i) have the right key already (as might be the case for persons storing data for their own later use), or (ii) are authorized by the data owner and have identities certified by a CA that is itself trusted by the data owner. A frontier of cryptography research, with some inventions now starting to make it into practice, is how to create different kinds of keys, ones which give only limited access of various kinds, or which allow messages to be sent to classes of individuals without knowing in advance exactly who they may be.

For example, “identity-based encryption” and “attribute-based encryption” are ways of sending a message, or protecting a file of data, for the exclusive use of “a person named Ramona Q. Doe who was born on May 23, 1980,” or for “anyone with the job title ombudsman, ombudsperson, or consumer advocate.” These techniques require a trusted third party (essentially a certificate authority), but the messages themselves do not need to pass through the hands of that third party. These tools are in early stages of adoption.

“Zero-knowledge” systems allow encrypted data to be queried for certain higher-level abstractions without revealing the low-level data. For example, a website operator could verify that a user is over age 21 without learning the user’s actual birthdate. What is remarkable is that this can be done in a way that proves mathematically that the user is not lying about his age: The operator learns with mathematical certainty that a certificate (signed by some CA of course!) attests to the user’s birthdate, without ever actually seeing that certificate. Zero-knowledge systems are just beginning to be commercialized in simple cases. They are not foreseeably extendable to complex and unstructured situations, such as what might be needed for the research mining of health-record data from non-consenting patients.

In some simpler domains, for example location privacy, practical cryptographic protection is closer to reality. The typical case might be that a group of friends want to know when they are close to one another, but without sharing their actual locations with any third party. Applications like this are, of course, much simpler if there is a trusted third party, as is de facto the case for most such commercial applications today.

Homomorphic encryption is a research area that goes beyond the mere querying of encrypted databases to actual computations (e.g., the collection of statistics) using encrypted data without ever decrypting it. These techniques are far from being practical, and they are unlikely to provide policy options on the timescale relevant to this report.
In secure multi-party computation, which is related to homomorphic encryption and is of particular interest in the financial sector, computation may be done on distributed data stores that are encrypted. Although individual data are kept private using “collusion-robust” encryption algorithms, data can be used to calculate general statistics. Parties that each know some private data use a protocol that generates useful results based on both information they know and information they do not know, without revealing to them data they do not already know.

Differential privacy, a comparatively new development related to but different from encryption, aims to maximize the accuracy of database queries or computations while minimizing the identifiability of individuals with records in the database, typically via obfuscation of query results (for example, by the addition of spurious information or “noise”). As with other obfuscation approaches, there is a tradeoff between data anonymity and the accuracy and utility of the query outputs. These ideas are far from practical application, except insofar as they may enable the risks of allowing any queries at all to be better assessed.

4.3 Notice and consent

Notice and consent is, today, the most widely used strategy for protecting consumer privacy. When the user downloads a new app to his or her mobile device, or when he or she creates an account for a web service, a notice is displayed, to which the user must positively indicate consent before using the app or service. In some fantasy world, users actually read these notices, understand their legal implications (consulting their attorneys if necessary), negotiate with other providers of similar services to get better privacy treatment, and only then click to indicate their consent. Reality is different.

Notice and consent fundamentally places the burden of privacy protection on the individual – exactly the opposite of what is usually meant by a “right.” Worse yet, if it is hidden in such a notice that the provider has the right to share personal data, the user normally does not get any notice from the next company, much less the opportunity to consent, even though use of the data may be different. Furthermore, if the provider changes its privacy notice for the worse, the user is typically not notified in a useful way.

As a useful policy tool, notice and consent is defeated by exactly the positive benefits that big data enables: new, non-obvious, unexpectedly powerful uses of data. It is simply too complicated for the individual to make fine-grained choices for every new situation or app. Nevertheless, since notice and consent is so deeply rooted in current practice, some exploration of how its usefulness might be extended seems warranted.

One way to view the problem with notice and consent is that it creates a non-level playing field in the implicit privacy negotiation between provider and user. The provider offers a complex take-it-or-leave-it set of terms, backed by a lot of legal firepower, while the user, in practice, allocates only a few seconds of mental effort to evaluating the offer, since acceptance is needed to complete the transaction that was the user’s purpose, and since the terms are typically difficult to comprehend quickly. This is a kind of market failure. In other contexts, market failures like this can be mitigated by the intervention of third parties who are able to represent significant numbers of users and negotiate on their behalf. Section 4.5.1 below suggests how such intervention might be accomplished.

4.4 Other strategies and techniques
4.4.1 Anonymization or de-identification

Long used in health-care research and other research areas involving human subjects, anonymization (also termed de-identification) applies when the data, standing alone and without an association to a specific person, do not violate privacy norms. For example, you may not mind if your medical record is used in research as long as you are identified only as Patient X and your actual name and patient identifier are stripped from that record.

Anonymization of a data record might seem easy to implement. Unfortunately, it is increasingly easy to defeat anonymization by the very techniques that are being developed for many legitimate applications of big data. In general, as the size and diversity of available data grows, the likelihood of being able to re-identify individuals (that is, re-associate their records with their names) grows substantially. 119

One compelling example comes from Sweeney, Abu, and Winn. 120 They showed in a recent paper that, by fusing public, Personal Genome Project profiles containing zip code, birthdate, and gender with public voter rolls, and mining for names hidden in attached documents, 84-97 percent of the profiles for which names were provided were correctly identified.

Anonymization remains somewhat useful as an added safeguard, but it is not robust against near-term future reidentification methods. PCAST does not see it as being a useful basis for policy. Unfortunately, anonymization is already rooted in the law, sometimes giving a false expectation of privacy where data lacking certain identifiers are deemed not to be personally identifiable information and therefore not covered by such laws as the Family Educational Rights and Privacy Act (FERPA).

4.4.2 Deletion and non-retention

It is an evident good business practice that data of all kinds should be deleted when they are no longer of value. Indeed, well-run companies often mandate the destruction of some kinds of records (both paper and electronic) after specified periods of time, often because they see little benefit in keeping the records as well as potential cost in producing them. For example, employee emails, which may be subject to legal process by (e.g.) divorce lawyers, are often seen as having negative retention value.

Counter to this practice is the new observation that big data is frequently able to find economic or social value in masses of data that were otherwise considered to be worthless. As the physical cost of retention continues to decrease exponentially with time (especially in the cloud), there will be a tendency in both government and the private sector to hold more data for longer – with obvious privacy implications. Archival data may also be important to future historians, or for later longitudinal analysis by academic researchers.

Only policy interventions will counter this trend. Government can mandate retention policies for itself. To affect the private sector, government may mandate policies where it has regulatory authorities (as for consumer protection, for example). But it can also encourage the development of stricter liability standards for companies whose data, including archived data, cause harm to individuals. A rational response by the private sector would then be to hold fewer data or to protect their use.
The above holds true for privacy-sensitive data about individuals that are held overtly – that is, the holder knows that he has the data and to whom they relate. As was discussed in Section 3.1.2, however, sources of data increasingly contain latent information about individuals, information that becomes known only if the holder expends analytic resources (beyond what may be economically feasible), or that may become knowable only in the future with the development of new data-mining algorithms. In such cases it is practically impossible for the data holder even to surface “all the data about an individual,” much less delete those data on any specified schedule.

The concepts of ephemerality (keeping data only on-the-fly or for a brief period), and transparency (enabling the individual to know what data about him or her are held) are closely related, and with the same practical limitations. While data that are only streamed, and not archived, may have lower risk of future use, there is no guarantee that a violator will play by the supposed rules, as in Target’s loss of 100 million debit card PINs, each present only ephemerally (see Section 4.2.1).

Today, given the distributed and redundant nature of data storage, it is not even clear that data can be destroyed with any useful degree of assurance. Although research on data destruction is ongoing, it is a fundamental fact that at the moment that data are displayed (in “analog”) to a user’s eyeballs or ears, they can also be copied (“re-digitized”) without any technical protections. The same holds if data are ever made available in unencrypted form to a rogue computer program, one designed to circumvent technical safeguards. Some misinformed public discussion notwithstanding, there is no such thing as automatically self-deleting data, other than in a fully controlled and rule-abiding environment.

As a current example, SnapChat provides the service of delivering ephemeral snapshots (images), visible for only a few seconds, to a designated recipient’s mobile device. SnapChat promises to delete past-date snaps from their servers, but it is only a promise. And, they are careful not to promise that the intended recipient may not contrive to make an uncontrolled and non-expiring copy. Indeed, the success of SnapChat incentivizes the development of just such copying applications. 121

From a policymaking perspective, the only viable assumption today, and for the foreseeable future, is that data, once created, are permanent. While their use may be regulated, their continued existence is best considered conservatively as unalterable fact.

4.5 Robust technologies going forward

4.5.1 A Successor to Notice and Consent

The purpose of notice and consent is that the user assents to the collection and use of personal data for a stated purpose that is acceptable to that individual. Given the large number of programs and Internet-available devices, both visible and not, that collect and use personal data, this framework is increasingly unworkable and ineffective. PCAST believes that the responsibility for using personal data in accordance with the user’s preferences should rest with the provider, possibly assisted by a mutually accepted intermediary, rather than with the user.

How might that be accomplished? Individuals might be encouraged to associate themselves with one of a standard set of privacy preference profiles (that is, settings or choices) voluntarily offered by third parties. For example, Jane might choose to associate with a profile offered by the American Civil Liberties Union that gives particular weight to individual rights, while John might associate with one offered by Consumer Reports that gives weight to economic value for the
consumer. Large app stores (such as Apple App Store, Google Play, Microsoft Store) for whom reputational value is important, or large commercial sectors such as finance, might choose to offer competing privacy-preference profiles.

In the first instance, an organization offering profiles would vet new apps as acceptable or not acceptable within each of their profiles. Basically, they would do the close reading of the provider’s notice that the user should, but does not, do. This is not as onerous as it may sound: While there are millions of apps, the most popular downloads are relatively few and are concentrated in a relatively small number of portals. The “long tail” of apps with few customers each might initially be left as “unrated.”

Simply by vetting apps, the third-party organizations would automatically create a marketplace for the negotiation of community standards for privacy. To attract market share, providers (especially smaller ones) could seek to qualify their offerings in as many privacy-preference profiles, offered by as many different third parties, as they deem feasible. The Federal government (e.g., through the National Institute of Standards and Technology) could encourage the development of standard, machine-readable interfaces for the communication of privacy implications and settings between providers and assessors.

Although human professionals could do the vetting today using policies expressed in natural language, it would be desirable in the future to automate that process. To do that, it would be necessary to have formalisms to specify privacy policies and tools to analyze software to determine conformance to those policies. But that is only part of the challenge. A greater challenge is to make sure the policy language is sufficiently expressive, the policies are sufficiently rich, and conformance tests are sufficiently powerful. Those requirements lead to a consideration of context and use.

4.5.2 Context and Use

The previous discussion, particularly that of Sections 3.1 and 3.2, illustrates PCAST’s belief that a focus on the collection, storage, and retention of electronic personal data will not provide a technologically robust foundation on which to base future policy. Among the many authors that have touched on these issues, Kagan and Abelson explain why access control does not suffice to protect privacy. Mundie gives a cogent and more complete explanation of this issue and advocates that privacy protection is better served by controlling the use of personal data, broadly construed, including metadata and data derived from analytics than by controlling collection. In a complementary vein, Nissenbaum explains that both the context of usage and the prevailing social norms contribute to acceptable use. To implement in a meaningful way the application of privacy policies to the use of personal data for a particular purpose (i.e., in context), those policies need to be associated both with data and with the code that operates on the data. For example, it must be possible to ensure that only apps with particular properties can be applied to certain data. The policies might be expressed in what computer scientists call natural language (plain English or the equivalent) and the association done by the user, or the policies might be stated formally and their association and enforcement done automatically. In either case, there must also be policies associated with the outputs of the computation, since they are data as well. The privacy policies of the output data must be computed from the policies associated with the inputs, the policies associated with the code, and the intended use of the outputs (i.e., the context). These privacy properties are a kind of metadata. To achieve a reasonable level of reliability, their implementation must be tamper-proof and “sticky” when data are copied.
There has been considerable research in areas that would contribute to such a capability, some of which is beginning to be commercialized. There is a history of using metadata (“tags” or “attributes”) in database systems to control use. While the formalization of privacy policies and their synthesis is a research topic, manual interpretation of such policies and the human determination of usage tags can be found in recent products. Identity management systems (to authenticate users and their roles, i.e., their context) are also evident both in research and in practice.

Commercial privacy systems for implementing use control exist today under the name of Trusted Data Format (TDF) implementations, developed principally for the United States intelligence community. TDF operates at the file level. The systems are primarily being implemented on a custom basis by large consulting firms, often assembled from open-source software components. Customers today are primarily government agencies, such as Federal intelligence agencies or local-government criminal intelligence units, or large commercial companies in vertically integrated industries like financial services and pharmaceutical companies looking to improve their accountability and auditing capabilities. Consulting services that have expertise in building such systems include, for example, Booz Allen, Ernst & Young, IBM, Northrop Grumman, and Lockheed; product-based companies like Palantir and new startups pioneering internal usage auditing, policy analytics, and policy reasoning engines have such expertise, as well. With sufficient market demand, more widespread market penetration could happen in the next five years. Market penetration would be further accelerated if the leading cloud-platform providers like Amazon, Google, and Microsoft implemented usage-controlled system technologies in their offerings. Wider-scale use through the government would help motivate the creation of off-the-shelf standard software.

**4.5.3 Enforcement and deterrence**

Privacy policies and the control of use in context are only effective to the extent that they are realized and enforced. Technical measures that increase the probability that a violator is caught can be effective only when there are regulations and laws with civil or criminal penalties to deter the violators. Then there is both deterrence of harmful actions and incentive to deploy privacy-protecting technologies.

It is today straightforward technically to associate metadata with data, with varying degrees of granularity ranging from an individual datum, to a record, to an entire collection. These metadata can record a wealth of auditable information, for example, provenance, detailed access and use policies, authorizations, logs of actual access and use, and destruction dates. Extending such metadata to derived or shared data (secondary use) together with privacy-aware logging can facilitate auditing. Although the state of the art is still somewhat ad hoc, and auditing is often not automated, so-called accountable systems are beginning to be deployed (Section 4.5.2). The ability to detect violations of privacy policies, particularly if the auditing is automated and continuous, can be used both to deter privacy violations and to ensure that violators are punished.

In the next five years, with regulation or market-driven encouragement, the large cloud-based infrastructure systems (e.g., Google, Amazon, Microsoft, Rackspace) could, as one example, incorporate the data-provenance and usage-compliance aspects of accountable systems into their cloud application-programming interfaces (APIs) and additionally provide APIs for policy awareness. These capabilities could then readily be included in open-source-based systems like Open Stack (associated with Rackspace) and other provider platforms. Applications intended to run on such cloud-based systems could be built with privacy concepts “baked into them,” even when they are developed by small enterprises or individual developers.
4.5.4 Operationalizing the Consumer Privacy Bill of Rights

In February 2012, the Administration issued a report setting forth a Consumer Privacy Bill of Rights (CPBR). The CPBR addresses commercial (not public sector) uses of personal data and is a strong statement of American privacy values.

For purposes of this discussion, the principles embodied in CPBR can be divided into two categories. First, there are obligations for data holders, analyzers, or commercial users. These are passive from the consumer’s standpoint – the obligations should be met whether or not the consumer knows, cares, or acts. Second, and different, there are consumer empowerments, things that the consumer should be empowered to initiate actively. It is useful here to rearrange the CPBR’s principles by category.

In the category of obligations are these elements:

- Respect for Context: Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data.
- Focused Collection: Consumers have a right to reasonable limits on the personal data that companies collect and retain.
- Security: Consumers have a right to secure and responsible handling of personal data.
- Accountability: Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights.

In the category of consumer empowerments are these elements:

- Individual Control: Consumers have a right to exercise control over what personal data companies collect from them and how they use it.
- Transparency: Consumers have a right to easily understandable and accessible information about privacy and security practices.
- Access and Accuracy: Consumers have a right to access and correct personal data in usable formats, in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data are inaccurate.

PCAST endorses as sound the principles underlying CPBR. Because of the rapidly changing technologies associated with big data, however, effective operationalization of CPBR is at risk. Up to now, debate over how to operationalize CPBR has focused on the collection contexts that motivated CPBR development. But, as discussed at multiple places in this report (e.g., Sections 3.1.2, 4.4 and 4.5.2), PCAST believes that such a focus will not provide a technologically robust foundation on which to base future policy that also applies to big data. Further, the increasing complexity of applications and uses of data undermines even a simple concept like “notice and consent.”

PCAST believes that the principles of CPBR can readily be adapted to a more robust regime based on recognizing and controlling harmful uses of the data. Some specific suggestions follow.

Turn first to the rights classified above as obligations on the data holder.

The principle of Respect for Context needs augmentation. As this report has repeatedly discussed, there are instances in which personal data are not provided by the customer. Such data may emerge as a product of analysis well after the data were collected and after they may have passed through several hands. While the intent of the right is appropriate, namely that data be used for legitimate purposes that do not produce certain adverse consequences or harms to
individuals, the CPBR’s articulation in which “consumers provide the data” is too limited. This right needs to state in some way that data about an individual – however acquired – not be used so as to cause certain adverse consequences or harms to that individual. (See Section 1.4 for a possible list of adverse consequences and harms that might be subject to some regulation.)

As initially conceived, the right to Focused Collection was to be achieved by techniques like de-identification and data deletion. As discussed in Section 4.4.1, however, de-identification (anonymization) is not a robust technology for big data in the face of data fusion; in some instances, there may be compelling reasons to retain data for beneficial purposes. This right should be about use rather than collection. It should emphasize utilizing best practices to prevent inappropriate use of data during the data’s whole life cycle, rather than depending on de-identification. It should not depend on a company’s being able itself to recognize “all” the data about a consumer that it holds, which is increasingly technically infeasible.

The principles underlying CPBR’s Security and Accountability remain valid in a use-based regime. They need to be applied throughout the value chain that includes data collection, analysis, and use.

Turn next to the rights here classified as consumer empowerments.

Where consumer empowerments have become practically impossible for the consumer to exercise meaningfully, they need to be recast as obligations of the commercial entity that actually uses the data or products of data analysis. This applies to the CPBR’s principles of Individual Control and of Transparency.

Section 4.3 explained how the non-obvious nature of big data’s products of analysis make it all but impossible for an individual to make fine-grained privacy choices for every new situation or app. For the principle of Individual Control to have meaning, PCAST believes that the burden should no longer fall on the consumer to manage privacy for each company with which the consumer interacts by a framework like “notice and consent.” Rather, each company should take responsibility for conforming its uses of personal data to a personal privacy profile designated by the consumer and made available to that company (including from a third party designated by the consumer). Section 4.5.1 proposed a mechanism for this change in responsibility.

Transparency (in the sense of disclosure of privacy practices) suffers from many of the same problems. Today, the consumer receives an unhelpful blizzard of privacy-policy notifications, many of which say, in essence, “retention, storage, and retention of data, with an emphasis on the providers can do anything we want.” As with Individual Control, the burden of conforming to a consumer’s stated personal-privacy profile should fall on the company, with notification to the consumers by a company if their profile precludes that company’s accepting their business. Since companies do not like to lose business, a positive market dynamic for competing privacy practices would thus be created.

For the right of Access and Accuracy to be meaningful, personal data must include the fruits of data analytics, not just collection. However, as this report has already explained (Section 4.4.2), it is not always possible for a company to "know what it knows" about a consumer, since that information may be unrecognized in the data; or it may become identifiable only in the future, when data sets are combined using new algorithms. When, however, the personal character of data is apparent to a company by virtue of its use of the data, its obligation to provide means for the correction of errors should be triggered. Consumers should have an expectation that companies will validate and correct data stemming from analysis and, since not all errors will be corrected, will also take steps to minimize the risk of
adverse consequences to consumers from the use of inaccurate data. Again, the primary burden must fall on the commercial user of big data and not on the consumer.

5. PCAST Perspectives and Conclusions

Breaches of privacy can cause harm to individuals and groups. It is a role of government to prevent such harm where possible, and to facilitate means of redress when the harm occurs. Technical enhancements of privacy can be effective only when accompanied by regulations or laws because, unless some penalties are enforced, there is no end to the escalation of the measures-countermeasures “game” between violators and protectors. Rules and regulations provide both deterrence of harmful actions and incentives to deploy privacy-protecting software technologies.

From everything already said, it should be obvious that new sources of big data are abundant; that they will continue to grow; and that they can bring enormous economic and social benefits. Similarly, and of comparable importance, new algorithms, software, and hardware technologies will continue to increase the power of data analytics in unexpected ways. Given these new capabilities of data aggregation and processing, there is inevitably new potential for both the unintentional leaking of both bulk and fine-grained data about individuals, and for new systematic attacks on privacy by those so minded.

Cameras, sensors, and other observational or mobile technologies raise new privacy concerns. Individuals often do not knowingly consent to providing data. These devices naturally pull in data unrelated to their primary purpose. Their data collection is often invisible. Analysis technology (such as facial, scene, speech, and voice recognition technology) is improving rapidly. Mobile devices provide location information that might not be otherwise volunteered. The combination of data from those sources can yield privacy-threatening information unbeknownst to the affected individuals.

It is also true, however, that privacy-sensitive data cannot always be reliably recognized when they are first collected, because the privacy-sensitive elements may be only latent in the data, made visible only by analytics (including those not yet invented), or by fusion with other data sources (including those not yet known). Suppressing the collection of privacy-sensitive data would thus be increasingly difficult, and it would also be increasingly counterproductive, frustrating the development of big data’s socially important and economic benefits.

Nor would it be desirable to suppress the combining of multiple sources and kinds of data: Much of the power of big data stems from this kind of data fusion. That said, it remains a matter of concern that considerable amounts of personal data may be derived from data fusion. In other words, such data can be obtained or inferred without intentional personal disclosure.

It is an unavoidable fact that particular collections of big data and particular kinds of analysis will often have both beneficial and privacy-inappropriate uses. The appropriate use of both the data and the analyses are highly contextual.

Any specific harm or adverse consequence is the result of data, or their analytical product, passing through the control of three distinguishable classes of actor in the value chain:

First, there are data collectors, who control the interfaces to individuals or to the environment. Data collectors may collect data from clearly private realms (e.g., a health questionnaire or wearable sensor), from ambiguous situations (e.g., cell-phone pictures or Google Glass videos taken at a party or cameras and microphones placed in a classroom...
for remote broadcast), or – increasing in both quantity and quality – data from the “public square,” where privacy-sensitive data may be latent and initially unrecognizable.

Second, there are data analyzers. This is where the “big” in big data becomes important. Analyzers may aggregate data from many sources, and they may share data with other analyzers. Analyzers, as distinct from collectors, create uses (“products of analysis”) by bringing together algorithms and data sets in a large-scale computational environment. Importantly, analyzers are the locus where individuals may be profiled by data fusion or statistical inference.

Third, there are users of the analyzed data – business, government, or individual. Users will generally have a commercial relationship with analyzers; they will be purchasers or licensees (etc.) of the analyzer’s products of analysis. It is the user who creates desirable economic and social outcomes. But, it is also the user who is the locus of producing actual adverse consequences or harms, when such occur.

5.1 Technical feasibility of policy interventions

Policy, as created by new legislation or within existing regulatory authorities, can, in principle, intervene at various stages in the value chain described above. Not all such interventions are equally feasible from a technical perspective, or equally desirable if the societal and economic benefits of big data are to be realized.

As indicated in Chapter 4, basing policy on the control of collection is unlikely to succeed, except in very limited circumstances where there is an explicitly private context (e.g., measurement or disclosure of health data) and the possibility of meaningful explicit or implicit notice and consent (e.g., by privacy preference profiles, see Sections 4.3 and 4.5.1), which does not exist today.

There is little technical likelihood that “a right to forget” or similar limits on retention could be meaningfully defined or enforced (see Section 4.4.2). Increasingly, it will not be technically possible to surface “all” of the data about an individual. Policy based on protection by anonymization is futile, because the feasibility of re-identification increases rapidly with the amount of additional data (see Section 4.4.1). There is little, and decreasing, meaningful distinction between data and metadata. The capabilities of data fusion, data mining, and re-identification render metadata not much less problematic than data (see Section 3.1).

Even if direct controls on collection are in most cases infeasible, however, attention to collection practices may help to reduce risk in some circumstances. Such best practices as tracking provenance, auditing access and use, and continuous monitoring and control (see Sections 4.5.2 and 4.5.3) could be driven by partnerships between government and industry (the carrot) and also by clarifying tort law and defining what might constitute negligence (the stick).

Turn next to data analyzers. One the one hand, it may be difficult to regulate them, because their actions do not directly touch the individual (it is neither collection nor use) and may have no external visibility. Mere inference about an individual, absent its publication or use, may not be a feasible target of regulation. On the other hand, an increasing fraction of privacy issues will surface only with the application of data analytics. Many privacy challenges will arise from the analysis of data collected unintentionally that was not, at the time of collection, targeted at any particular individual or even group of individuals. This is because combining data from many sources will become more and more powerful.

It might be feasible to introduce regulation at the “moment of particularization” of data about an individual, or when this is done for some minimum number of individuals concurrently. To be effective such regulation would need to be
accompanied by requirements for tracking provenance, auditing access and use, and using security measures (e.g., robust encryption infrastructure) at all stages of the evolution of data, and for providing transparency, and/or notification, at the moment of particularization.

Big data’s “products of analysis” are created by computer programs that bring together algorithms and data so as to produce something of value. It might be feasible to recognize such programs, or their products, in a legal sense and to regulate their commerce. For example, they might not be allowed to be used in commerce (sold, leased, licensed, and so on) unless they are consistent with individuals’ privacy elections or other expressions of community values (see Sections 4.3 and 4.5.1). Requirements might be imposed on conformity to appropriate standards of provenance, auditability, accuracy, and so on, in the data they use and produce; or that they meaningfully identify who (licensor vs. licensee) is responsible for correcting errors and liable for various types of harm or adverse consequence caused by the product.

It is not, however, the mere development of a product of analysis that can cause adverse consequences. Those occur only with its actual use, whether in commerce, by government, by the press, or by individuals. This seems the most technically feasible place to apply regulation going forward, focusing at the locus where harm can be produced, not far upstream from where it may barely (if at all) be identifiable.

When products of analysis produce imperfect information that may mis-classify individuals in ways that produce adverse consequences, one might require that they meet standards for data accuracy and integrity; that there are useable interfaces that allow an individual to correct the record with voluntary additional information; and that there exist streamlined options for redress, including financial redress, when adverse consequences reach a certain level.

Some harms may affect groups (e.g., the poor or minorities) rather than identifiable individuals. Mechanisms for redress in such cases need to be developed.

There is a need to clarify standards for liability in case of adverse consequences from privacy violations. Currently there is a patchwork of out-of-date state laws and legal precedents. One could encourage the drafting of technologically savvy model legislation on cyber-torts for consideration by the states.

Finally, government may be forbidden from certain classes of uses, despite their being available in the private sector.

5.2 Recommendations

PCAST’s charge for this study does not ask it to make recommendations on privacy policies, but rather to make a relative assessment of the technical feasibility of different broad policy approaches. PCAST’s overall conclusions about that question are embodied in the first two of our recommendations:

**Recommendation 1. Policy attention should focus more on the actual uses of big data and less on its collection and analysis.**

By actual uses, we mean the specific events where something happens that can cause an adverse consequence or harm to an individual or class of individuals. In the context of big data, these events (“uses”) are almost always actions of a computer program or app interacting either with the raw data or with the fruits of analysis of those data. In this formulation, it is not the data themselves that cause the harm, nor the program itself (absent any data), but the
confluence of the two. These “use events” (in commerce, by government, or by individuals) embody the necessary specificity to be the subject of regulation. Since the purpose of bringing program and data together is to accomplish some identifiable desired task, use events also capture some notion of intent, in a way that data collection by itself or program development by itself may not. The policy question of what kinds of adverse consequences or harms rise to the level of needing regulation is outside of PCAST’s charge, but an illustrative set that seem grounded in common American values was provided in Section 1.4.

PCAST judges that alternative big-data policies that focus on the regulation of data collection, storage, retention, a priori limitations on applications, and analysis (absent identifiable actual uses of big data or its products of analysis) are unlikely to yield effective strategies for improving privacy. Such policies are unlikely to be scalable over time as it becomes increasingly difficult to ascertain, about any particular data set, what personal information may be latent in it – or in its possible fusion with every other possible data set, present or future. The related issue is that policies limiting collection and retention are increasingly unlikely to be enforceable by other than severe and economically damaging measures. While there are certain definable classes of data so repugnant to society that their mere possession is criminalized, the information in big data that may raise privacy concerns is increasingly inseparable from a vast volume of the data of ordinary commerce, or government function, or collection in the public square. This dual-use character of information, too, argues for the regulation of use rather than collection.

Recommendation 2. Policies and regulation, at all levels of government, should not embed particular technological solutions, but rather should be stated in terms of intended outcomes.

To avoid falling behind the technology, it is essential that policy concerning privacy protection should address the purpose (the “what”) rather than the mechanism (the “how”). For example, regulating disclosure of health information by regulating the use of anonymization fails to capture the power of data fusion; regulating the protection of information about minors by controlling inspection of student records held by schools fails to anticipate the student information capturing by online learning technologies. Regulating control of the inappropriate disclosure of health information or student performance, no matter how the data are acquired is more robust.

PCAST further responds to its charge with the following recommendations, intended to advance the agenda of strong privacy values and the technological tools needed to support them:

Recommendation 3. With coordination and encouragement from OSTP, the NITRD agencies should strengthen U.S. research in privacy-related technologies and in the relevant areas of social science that inform the successful application of those technologies.

Some of the technology for controlling uses already exists. Research (and funding for it) is needed, however, in the technologies that help to protect privacy, in the social mechanisms that influence privacy-preserving behavior, and in the legal options that are robust to changes in technology and create appropriate balance among economic opportunity, other national priorities, and privacy protection.

Following up on recommendations from PCAST for increased privacy-related research, a 2013-2014 internal government review of privacy-focused research across Federal agencies supporting research on information technologies suggests that about $80 million supports either research with an explicit focus on enhancing privacy or research that addresses privacy protection ancillary to some other goal (typically cybersecurity). The funded
research addresses such topics as an individual’s control over his or her information, transparency, access and accuracy, and accountability. It is typically of a general nature, except for research focusing on the health domain or (relatively new) consumer energy usage. The broadest and most varied support for privacy research, in the form of grants to individuals and centers, comes from the National Science Foundation (NSF), engaging social science as well as computer science and engineering. 135, 136

Research into privacy as an extension or complement to security is supported by a variety of Department of Defense agencies (Air Force Research Laboratory, the Army's Telemedicine and Advanced Technology Research Center, Defense Advanced Research Projects Agency, National Security Agency, and Office of Naval Research) and the Intelligence Advanced Research Projects Activity (IARPA) within the Intelligence Community. IARPA, for example, has hosted the Security and Privacy Assurance Research 137 program, which has explored a variety of encryption techniques. Research at the National Institute for Standards and Technology (NIST) focuses on the development of cryptography and biometric technology to enhance privacy as well as support for federal standards and programs for identity management. 138

Looking to the future, continued investment is needed not only in privacy topics ancillary to security, but also in automating privacy protection for the broadest aspects of use of data from all sources. Relevant topics include cryptography, privacy-preserving data mining (including analysis of streaming as well as stored) data, 139 formalization of privacy policies, tools for automating conformance of software to personal privacy policy and to legal policy, methods for auditing use in context and identifying violations of policy, and research on enhancing people’s ability to make sense of the results of various big-data analyses. Development of technologies that support both quality analytics and privacy preservation on distributed data, such as secure multiparty computation, will become even more important, given the expectation that people will draw increasingly from data stored in multiple locations. The creation of tools that analyze the panoply of National, state, regional, and international rules and regulations for inconsistencies and differences will be helpful for the definition of new rules and regulations, as well as for those software developers that need to customize their services for different markets.

Recommendation 4. OSTP, together with the appropriate educational institutions and professional societies, should encourage increased education and training opportunities concerning privacy protection, including professional career paths.

Programs that provide education leading to privacy expertise (akin to what is being done for security expertise) are essential and need encouragement. One might envision careers for digital-privacy experts both on the software development side and on the technical management side. Employment opportunities should exist not only in industry (and government at all levels), where jobs focused on privacy (including but not limited to Chief Privacy Officers) have been growing, but also for consumer and citizen advocacy and support, perhaps offering “annual privacy checkups” for individuals. Just as education and training about cybersecurity has advanced over the past 20 years within the technical community, there is now opportunity to educate and train students about privacy implications and privacy enhancements, beyond the present small niche area occupied by this focus within computer science programs. 140 Privacy is also an important component of ethics education for technology professionals.

Recommendation 5. The United States should take the lead both in the international arena and at home by adopting policies that stimulate the use of practical privacy-protecting technologies that exist today. This country can exhibit leadership both by its convening power (for instance, by promoting the creation and
adoption of standards) and also by its own procurement practices (such as its own use of privacy-preserving cloud services).

Section 4.5.2 described a set of privacy-enhancing best practices that already exist today in U.S. markets. PCAST is not aware of any more effective innovation or strategies being developed abroad; rather, some countries seem inclined to pursue what PCAST believes to be blind alleys. This circumstance offers an opportunity for U.S. technical leadership in privacy in the international arena, an opportunity that should be seized. Public policy can help to nurture the budding commercial potential of privacy-enhancing technologies, both through U.S. government procurement and through the larger policy framework that motivates private-sector technology engagement.

As it does for security, cloud computing offers positive new opportunities for privacy. By requiring privacy-enhancing services from cloud-service providers contracting with the U.S. government, the government should encourage those providers to make available sophisticated privacy enhancing technologies to small businesses and their customers, beyond what the small business might be able to do on its own. 141

5.3 MISSING

5.4 Final Remarks

Privacy is an important human value. The advance of technology both threatens personal privacy and provides opportunities to enhance its protection. The challenge for the U.S. Government and the larger community, both within this country and globally, is to understand what the nature of privacy is in the modern world and to find those technological, educational, and policy avenues that will preserve and protect it.

References

1

The White House Office of Science and Technology Policy

2

NITRD refers to the Networking and Information Technology Research and Development program, whose participating Federal agencies support unclassified research in advanced information technologies such as computing, networking, and software and include both research- and mission-focused agencies such as NSF, NIH, NIST, DARPA, NOAA, DOE's Office of Science, and the DOD military-service laboratories (see http://www.nitrd.gov/SUBCOMMITTEE/nileies/index.aspx).

3


PCAST acknowledges gratefully the assistance of several contributors at the National Science Foundation, who helped to identify and distill key insights from the technical literature and research community, as well as other technical experts in academia and industry that it consulted during this project. See Appendix A.


Id. at 195.


Id. at 483-84.

McIntyre v. Ohio Elections Commission, 514 U.S. 334, 340-41 (1995). The decision reads in part, “Protections for anonymous speech are vital to democratic discourse. Allowing dissenters to shield their identities frees them to express critical minority views . . . Anonymity is a shield from the tyranny of the majority . . . . It thus exemplifies the purpose behind the Bill of Rights and of the First Amendment in particular: to protect unpopular individuals from retaliation . . . at the hand of an intolerant society.”


One Hundred Tenth Congress, “Privacy: The use of commercial information resellers by federal agencies,” Hearing before the Subcommittee on Information Policy, Census, and National Archives of the Committee on Oversight and Government Reform, House of Representatives, March 11, 2008.

For example, Experian provides much of Healthcare.gov’s identity verification component using consumer credit information not available to the government. See Consumer Reports, “Having trouble proving your identity to HealthCare.gov? Here’s how the process works,” December 18, 2013.

http://www.consumerreports.org/cro/n...ginMethod=auto


Id.
One perspective informed by new technologies and technology-mediated communication suggests that privacy is about the "continual management of boundaries between different spheres of action and degrees of disclosure within those spheres," with privacy and one’s public face being balanced in different ways at different times. See: Leysia Palen and Paul Dourish, “Unpacking ‘Privacy’ for a Networked World,” Proceedings of CHI 2003, Association for Computing Machinery, April 5-10, 2003.

“I would ask whether people reasonably expect that their movements will be recorded and aggregated in a manner that enables the Government to ascertain, more or less at will, their political and religious beliefs, sexual habits, and so on.” United States v. Jones (10-1259), Sotomayor concurrence at http://www.supremecourt.gov/opinions/11pdf/10-1259.pdf.


30


31


32


33


34


35


36

Facebook’s “The Graph API” (at https://developers.facebook.com/docs/graph-api/) describes how to write computer programs that can access the Facebook friends’ data.

37

One of four big-data applications honored by the trade journal, Computerworld, in 2013. King, Julia, "UN tackles socioeconomic crises with big data," Computerworld, June 3, 2013. http://www.computerworld.com/s/artic...with_big_data
38

39

40

41

42

43
For an overview of MOOCs and associated analytics opportunities, see PCAST’s December 2013 letter to the President. http://www.whitehouse.gov/sites/defa...t_dec-2013.pdf

44
There is also uncertainty about how to interpret applicable laws, such as the Family Educational Rights and Privacy Act (FERPA). Recent Federal guidance is intended to help clarify the situation. See: U.S. Department of Education, “Protecting Student Privacy While Using Online Educational Services: Requirements and Best Practices,” February 2014. http://ptac.ed.gov/sites/default/fil...%202014%29.pdf


See references at footnote 30.

Such databases endure and form the basis of continuing concern among privacy advocates.

Schemas are formal definitions of the configuration of a database: its tables, relations, and indices. Headers are the sometimes-invisible prefaces to email messages that contain information about the sending and destination addresses and sometimes the routing of the path between them.
In the Internet and similar networks, information is broken up into chunks called packets, which may travel independently and depend on metadata to be reassembled properly at the destination of the transmission.


DuckDuckGo is a non-tracking search engine that, while perhaps yielding fewer results than leading search engines, is used by those looking for less tracking. See: https://duckduckgo.com/


See: http://www.fitbit.com/
62

63

64

65

66

67

68

69

70

71
DARPA, for example, has a project involving machine learning and other technologies to build medical causal models from
analysis of cancer literature, leveraging the greater capacity of a computer than a person to process information from a large number of sources. See description at http://www.darpa.mil/Our_Work/I2O/Pr...Mechanism.aspx

72

“Data mining breaks the basic intuition that identity is the greatest source of potential harm because it substitutes inference for identifying information as a bridge to get at additional facts.” Barocas, Solon and Helen Nissenbaum, “Big Data’s End Run Around Anonymity and Consent,” Chapter II, in Lane, Julia, et al., Privacy, Big Data, and the Public Good, Cambridge University Press, 2014.

73


74


75


76


77


78


79

For example, Newark Airport recently installed a system of 171 LED lights (from Sensity [http://www.sensity.com/]) that contain special chips to connect to sensors and cameras over a wireless system. These systems allow for advanced automatic lighting to improve security in places like parking garages, and in doing so capture a large range of information.
This was discussed at the workshop cited in footnote 78.

Such concerns are likely to grow as commercial satellite imagery systems such as Skybox (http://skybox.com/) provide the basis for more services.


A variety of tools exist for managing, analyzing, visualizing and manipulating network (graph) datasets, such as Allegrograph, GraphVis, R, visone and Wolfram Alpha. Some, such as Cytoscape, Gephi and Netviz are open source.


89


90

Fienberg, S.E., "Is the Privacy of Network Data an Oxymoron?" Journal of Privacy and Confidentiality, 4:2, 2013.

91


92


93


94


95

A petabyte is 10^15 bytes. One petabyte could store the individual genomes of the entire U.S. population. The human brain has been estimated to have a capacity of 2.5 petabytes.


See: http://accumulo.apache.org/

See: https://amplab.cs.berkeley.edu/software/


PCAST has addressed issues in cybersecurity, both in reviewing the NITRD programs and directly in a 2013 report, Immediate Opportunities for Strengthening the Nation’s Cybersecurity. http://www.whitehouse.gov/sites/defa...y_nov-2013.pdf

There are also choices in the design and implementation of security mechanisms that affect privacy. In particular, authentication or the attempt to demonstrate identity at some level can be done with varying degrees of disclosure. See, for example: Computer Science and Telecommunications Board, Who Goes There: Authentication Through the Lens of Privacy, National Academies Press, 2003.
Such research can inform efforts to automate the checking of compliance with policies and/or associated auditing.

This future-proofing remains hard to achieve; PCAST’s cybersecurity report advocated approaches that would be more durable than the kinds of check-lists that are easily rendered obsolete. See: http://www.whitehouse.gov/sites/defa..., y nov-2013.pdf


Attacks that compromise the hardware or software that does the encrypting (for example, the promulgation of intentionally weak cryptography standards) can be considered to be a variant of attacks that reveal plaintext.


Public-key encryption originated through the secret work of British mathematicians at the U.K.’s Government Communications Headquarters (GCHQ), an organization roughly analogous to the NSA, and received broader attention through the independent work by researchers including Whitfield Diffie and Martin Hellman in the United States.
113
http://threatpost.com/final-report-d...s-103112/77170.

114
It is not publicly known whether or not the earlier 2010 compromise of servers belonging to VeriSign, a much larger CA, led to compromises of certificates or signing authorities. Bradley, Tony, “VeriSign Hacked: What We Don’t Know Might Hurt Us,” PC World, February 2, 2012.
http://www.pcworld.com/article/24924...t_hurt_us.html

115
A sample report-card: https://www.eff.org/deeplinks/2013/1...t#crypto-chart

116

117

118

119
De-identification can also be seen as a spectrum, rather than a single approach. See: “Response to Request for Information Filed by U.S. Public Policy Council of the Association for Computing Machinery,” March 2014.

120
121

122

123
Mundie, Craig, “Privacy Pragmatism: Focus on Data Use, Not Data Collection,” Foreign Affairs, March/April, 2014.

124

125

126

127
Oracle’s eXtensible Access Control Markup Language (XACML) has been used to implement attribute-based access controls for identity management systems. (Personal communication, Mark Gorenberg and Peter Guerra of Booz Allen)

128

129
See: http://www.openstack.org/
Lawyers may encourage companies to use over-inclusive language to cover the unpredictable evolution of possibilities described elsewhere in this report, even in the absence of specific plans to use specific capabilities.

Child pornography is the most universally recognized example.

NITRD refers to the Networking and Information Technology Research and Development program, whose participating Federal agencies support unclassified research in advanced information technologies such as computing, networking, and software and include both research- and mission-focused agencies such as NSF, NIH, NIST, DARPA, NOAA, DOE’s Office of Science, and the DOD military service laboratories (see http://www.nitrd.gov/SUBCOMMITTEE/ni...ies/index.aspx). There is research coordination between NITRD and Federal agencies conducting or supporting corresponding classified research.


The Secure and Trustworthy Cyberspace program is the largest funder of relevant research. See: http://www.nsf.gov/funding/pgm_summ....pims_id=504709


See: http://www.iarpa.gov/index.php/research-programs/spar
NIST is responsible for advancing the National Strategy for Trusted Identities in Cyberspace (NSTIC), which is intended to facilitate secure transactions within and across public and private sectors. See: http://www.nist.gov/nstic/


A basis can be found in the newest version of the curriculum guidance of the Association for Computing Machinery (http://www.acm.org/education/CSC2013-final-report.pdf). Given all of the pressures on curriculum, progress—as with cybersecurity—may hinge on growth in privacy-related research, business opportunities, and occupations.

A beginning can be found in the Federal Government’s FedRAMP program for certifying cloud services. Initiated to address Federal agency security concerns, FedRAMP already builds in attention to privacy in the form of a required Privacy Threshold Analysis and in some situations a Privacy Impact Analysis. The office of the U.S. Chief Information Officer provides guidance on Federal uses of information technology that addresses privacy along with security (see http://cloud.cio.gov/). It provides specific guidance on the cloud and FedRAMP (http://cloud.cio.gov/fedramp), including privacy protection (http://cloud.cio.gov/document/privacy-act-assessment).

Appendix A. Additional Experts Providing Input

Yochai Benkler Harvard

Eleanor Birrell Cornell University

Courtney Bowman Palantir

Christopher Clifton Purdue University

James Costa Sandia National Laboratory

Lorrie Faith Cranor Carnegie Mellon University

Deborah Estrin Cornell NYC

William W. (Terry) Fisher Harvard Law School

Stephanie Forrest University of New Mexico

Dan Geer In-Q-Tel
Special Acknowledgment

PCAST is especially grateful for the rapid and comprehensive assistance provided by an ad hoc group of staff at the National Science Foundation (NSF), Computer and Information Science and Engineering Directorate. This team was led by Fen Zhao and Emily Grumbling, who were enlisted by Suzanne Iacono. Drs. Zhao and Grumbling worked tirelessly to review the technical literature, elicit perspectives and feedback from a range of NSF colleagues, and iterate on descriptions of numerous technologies relevant to big data and privacy and how those technologies were evolving.

NSF Technology Team Leaders

Fen Zhao, AAAS Fellow, CISE

Emily Grumbling, AAAS Fellow, Office of Cyberinfrastructure

Additional NSF Contributors

Robert Chadduck, Program Director

Almadena Y. Chtchelkanova, Program Director

David Corman, Program Director

James Donlon, Program Director

Jeremy Epstein, Program Director

Joseph B. Lyles, Program Director

Dmitry Maslov, Program Director

Mimi McClure, Associate Program Director