Data Science for Agency Initiatives 2015

Story
Slides
Slide 1 Data Science for Agency Initiatives 2015
Slide 2 Government Leadership in the Data Age
Slide 3 Data Science for Agency Initiatives 2015: MindTouch Knowledge Base
Slide 4 CFPB Consumer Complaint Database
Slide 5 FCC Data
Slide 6 FCC Datasets Download
Slide 7 Data Science for Agency Initiatives 2015: Spreadsheet FCC Data Knowledge Base
Slide 8 GAO Government Data Sharing Community of Practice
Slide 9 GAO Government Data Sharing Community of Practice: MindTouch Knowledge Base
Slide 10 CMS.gov Data Navigator: Start
Slide 11 CMS.gov Data Navigator: Search
Slide 12 Data Science for Agency Initiatives 2015: Spreadsheet CMS Data Knowledge Base
Slide 13 Data Science for Agency Initiatives 2015: CFPB Consumer Complaints-Spotfire Cover Page
Slide 14 Data Science for Agency Initiatives 2015: CFPB Consumer Complaints-Spotfire Counts by State
Slide 15 Data Science for Agency Initiatives 2015: FCC Data-Spotfire Cover Page
Slide 16 Data Science for Agency Initiatives 2015: FCC Data-Spotfire Analytics
Slide 17 Data Science for Agency Initiatives 2015: CMS Data-Spotfire Cover Page
Slide 18 Data Science for Agency Initiatives 2015: CMS Data-Spotfire Analytics
Slide 19 Data Science for Agency Initiatives 2015: CMS Data-Spotfire Visualizations
Slide 20 Conclusions and Recommendations

Spotfire Dashboard
CFPB Consumer Complaints
FCC Data
CMS Data Navigator

Research Notes

Government Leadership in the Data Age

Agenda
Overview

http://semanticommunity.info/Data_Science/Data_Science_for_Agency_Initiatives_2015
Updated: Wed, 23 Sep 2015 08:12:44 GMT
Powered by MindTouch™
Speakers

Linda F. Powell, Chief Data Officer, Consumer Financial Protection Bureau
Tony Summerlin, Special Advisor to the CIO, FCC
Dr. Joah Iannotta, Assistant Director, Government Accountability Office
Niall Brennan, Chief Data Officer, Centers for Medicare & Medicaid Services
Chris Wilkinson, Senior Director, Market Development, immixGroup

Additional Information

Consumer Complaint Database

Get the Data

What happens when I submit a complaint?
What information do you publish?

Data by product

All
Bank account or service
Credit card
Credit reporting
Debt collection
Money transfers
Mortgage
Other financial service
Payday loan
Prepaid card
Student loan
Vehicle or other consumer loan

Download, sort, and visualize

Example visualizations

Complaints by channel
Credit card complaint responses by issue
Complaints by product

WHAT IS THIS DATA?

Field reference
More information

How we use complaint data
Reports to Congress
Snapshots of complaints

API & Documentation

Consumer Complaint Database API
Publication criteria
FCC Data

Data Innovation Initiative
Information & Data Officers
Zero-Based Data Review
Spectrum Inventory
Search FCC Databases
Download FCC Datasets

2,948 Listed Test Firms
Accredited Test Firms
Automated Reporting Management Information System
Cable Operations and Licensing System
Cable Communities Registered with the FCC
Consolidated Public Database System
Equipment Authorization Grantee Registrations
FM Service Contour Data Points
International Bureau Application Filing & Reporting System
License View Database
Raw HTML files hosting data for ECFS
Section 43.61 International Traffic Data
TCB Designating Authorities (TDA)
Telecommunication Certification Bodies (TCB)
Test Firm Accrediting Bodies (TFAB)
Universal Licensing System

Database Documentation
Database Downloads

800 MHz Vacated Spectrum
Aircraft - 47 CFR Part 87
Amateur Radio Service
Antenna Structure Registration
Airport Facilities - 47 CFR Part 17
Assignments and Transfers
BRS & EBS (Formerly known as MDS/ITFS)
Cellular - 47 CFR Part 22
Commercial Radio and Restricted Radiotelephone - FRC
General Mobile Radio Service (GMRS)
Land Mobile - Private
Land Mobile - Commercial
Land Mobile - Broadcast Auxiliary
Maritime Coast & Aviation Ground
MDS/ITFS: (Now known as BRS & EBS)
Market Based Services
Microwave - 47 CFR Parts 74 and 101, and 3650 - 3700 MHz
Ownership
Paging - 47 CFR Part 22
Ship Radio Service - 47 CFR Part 80
Spectrum Leasing
Unlicensed Wireless Microphone Registration

Develop on FCC APIs
Subscribe to FCC RSS Feeds
FOIA and Ex Parte

GAO Government Data Sharing Community of Practice

Overview
Contacts
Events

Timeline: Explore Recent Community of Practice Events

MINUTES: CoP Meeting Minutes - 1/26/15
  Background on the GAO's Government Data Sharing Community of Practice
  Data Sharing in Disaster Response and Recovery
  Panel 1: Data Sharing in Disaster Response and Recovery
  Panel 2: Disaster Response: Creating a More Agile and Responsive Federal Work force
  Slides
  Slides

MINUTES: CoP Meeting Minutes - 4/28/2014
MINUTES: Technological Challenges to Data Sharing, Meeting Minutes - 2/12/2014
MINUTES: CoP Meeting Minutes - 11/20/2013

REPORT CG FORUMS & ROUNDTABLES: Data Analytics for Oversight & Law Enforcement

April 22nd Meeting Agenda

How CMS is Using Big Data to Spur Healthcare Transformation
  Slide 1 How CMS is Using Big Data to Spur Healthcare Transformation 1
  Slide 2 How CMS is Using Big Data to Spur Healthcare Transformation 2
  Slide 3 Introduction
  Slide 4 CMS Data and Delivery System Transformation
  Slide 5 Creating Information Products
  Slide 6 Geographic Variation and Chronic Condition Data
  Slide 7 Geographic Variation Dashboard
  Slide 8 Chronic Conditions: State-Level Dashboard
Welcome to the CMS Data Navigator

About

FAQs

What is the CMS Data Navigator?
What are Publically Available Data Files – for download?
What are Publically Available Data Files – for purchase?
What are Restricted Use Data Files?
How does the CMS Data Navigator search logic work?

Start

Data Glossary, Navigator, and Catalog

Medicare Geographic Variation

Geographic Data

Drawing Causal Inference from Big Data

Overview

Speakers’ Bio Sketches

Edoardo Airoldi

Optimal design of experiments in the presence of network interference

Susan Athey

Leon Bottou

Causal Reasoning and Learning Systems

Peter Buhlmann

Susan Dumais

Dean Eckles

Identifying peer effects in social networks with peer encouragement designs

James Fowler
A Follow-up to a 61 Million Person Experiment in Social Influence and Political Mobilization

Michael Hawrylycz

Project MindScope: From Big Data to Behavior in the Functioning Cortex

David Heckerman

Jennifer Hill

Michael Jordan

On Computational Thinking, Inferential Thinking and "Big Data"

Steven Levitt

Thinking Differently about Big Data

David Madigan

Honest Inference From Observational Database Studies

Judea Pearl

Taming the challenge of extrapolation: From multiple experiments and observations to valid causal conclusions

Thomas Richardson

Non-parametric Causal Inference

James M Robins

Personalized Medicine, Optimal Treatment Strategies, and First Do No Harm: Time Varying Treatments and Big Data

Bernhard Schölkopf

Toward Causal Machine Learning

Jasjeet Sekhon

Combining Experiments with Big Data to Estimate Treatment Effects

Cosma Shalizi

Richard Shiffrin

Introduction to the Sackler Colloquium, Drawing Causal Inference from Big Data

John Stamatoyannopoulos

Hal Varian

Bin Yu

Lasso adjustments of treatment effect estimates in randomized experiments

Agenda

Thursday, March 26

Annual Sackler Lecture

Friday, March 27

NEXT
Data Science for Agency Initiatives 2015

On May 12th, I attended: Government Leadership in the Data Age, which featured four excellent government data age leadership speakers. I research their data sets for data science products for a future meetup as follows:

- Linda F. Powell, Chief Data Officer, Consumer Financial Protection Bureau
  - Consumer Complaint Database (62 MB CSV File and See Data Dictionary: Field reference)
- Tony Summerlin, Special Advisor to the CIO, FCC
  - FCC Data (It says: Over 150 data sets are available for download today, but I only find 16 links)
- Dr. Joah Iannotta, Assistant Director, Government Accountability Office
  - Government Data Sharing Community of Practice (Report and Minute PDF files that need to be researched for datasets)
- Niall Brennan, Chief Data Officer, Centers for Medicare & Medicaid Services
  - How CMS is Using Big Data to Spur Healthcare Transformation (Slides for Big Data Healthcare Analytics Forum 2014)

The panelists and moderator were excellent and the organizers and sponsors are to be complimented on an excellent event and venue!

My question to all the panel members was given all of this getting the data out, how much actual looking at the data is being done. I used my experience of hospital readmission data as an example of the value of looking at the data to understand was really going on and what to look for in the data and how to model it. The GAO panelist had an excellent response that they request data and look for outliers and why they occur and what it means to saving taxpayers money.

I started to understand the CFPB Consumer Complaint Database by importing the 62 MB CVS file into Spotfire and doing Exploratory Data Analysis as shown in the multiple, dynamically-linked adjacent visualizations below in screen capture slides and interactive dashboard forms. The data dictionary is provided in a table (Field reference) that should be imported into Spotfire as well. I will do this once I have completed the Knowledge Base for other datasets and its index in a Spreadsheet as part of the Data Mining process to produce a data product.

MORE TO FOLLOW AS I PREPARE AND ANALYZE THE DATA AND MEETUP

The purpose of the recent meeting held March 26-27, 2015 at the National Academy of Sciences on Drawing Causal Inference from Big Data was the following:

Although we are producing and storing ever greater amounts of data, we have just begun to figure out ways to analyze and understand what the data show. The problem is not restricted to science: Business, government, entertainment, social media, security agencies, social networks face the same challenges.

The two main challenges, both unprecedented in scope, are two sides of the same coin:

First, how does one find the important patterns of data?
This subject requires a Sackler Colloquium of its own. Suppose a moderately large data base has a terabyte of data ($10^{12}$). This data might perhaps contain a thousand ($10^3$) measurable factors. The number of correlations of those factors in all combinations would be on the order of $2^{10^3}$, or about a 300 digit number. The search problem for patterns is enormous.

Second, having found a pattern, how can we explain its causes?

This is the focus of the present Sackler Colloquium. If in a terabyte data base we notice factor A is correlated with factor B, there might be a direct causal connection between the two, but there might be something like $2^{300}$ other potential causal loops to be considered. Things could be even more daunting: To infer probabilities of causes could require consideration all distributions of probabilities assigned to the $2^{300}$ possibilities. Such numbers are both fanciful and absurd, but are sufficient to show that inferring causality in Big Data requires new techniques. These are under development, and we will hear some of the promising approaches in the next two days.

Whatever is developed, I am sure computational algorithms will never be sufficient to answer either question. The numbers go well beyond any conceivable rote computational approach. To me this highlights the importance of models and theories. Models and theories have always been important in science, but in Big Data they will be critically needed to guide the search for patterns and guide the search for causal accounts.

---

**Slides**

**Slide 1 Data Science for Agency Initiatives 2015**

[http://semanticommunity.info](http://semanticommunity.info)
Slide 2 Government Leadership in the Data Age

Government Leadership in the Data Age

• Linda F. Powell, Chief Data Officer, Consumer Financial Protection Bureau:
  – Consumer Complaint Database
• Tony Summerlin, Special Advisor to the CIO, FCC:
  – FCC Data (1507)
• Dr. Joah Iannotta, Assistant Director, Government Accountability:
  – GAO Government Data Sharing Community of Practice
• Niall Brennan, Chief Data Officer, Centers for Medicare & Medicaid Services:
  – How CMS is Using Big Data to Spur Healthcare Transformation (Readmissions)

Slide 3 Data Science for Agency Initiatives 2015: MindTouch Knowledge Base

Data Science for Agency Initiatives 2015
Slide 4 CFPB Consumer Complaint Database

http://www.consumerfinance.gov/complaintdatabase/

Slide 5 FCC Data

https://www.fcc.gov/data
Slide 6 FCC Datasets Download

https://www.fcc.gov/data/download-fcc-datasets

Slide 7 Data Science for Agency Initiatives 2015: Spreadsheet FCC Data Knowledge Base

http://semanticommunity.info/%40api/...?origin=mt-web
Slide 8 GAO Government Data Sharing Community of Practice

http://www.gao.gov/aac/gds_community...ctice/overview

Slide 9 GAO Government Data Sharing Community of Practice: MindTouch Knowledge Base

GAO Government Data Sharing Community of Practice
Slide 10 CMS.gov Data Navigator: Start
https://dnav.cms.gov/

Slide 11 CMS.gov Data Navigator: Search
CMS.gov Data Navigator: Search


Slide 12 Data Science for Agency Initiatives 2015: Spreadsheet CMS Data Knowledge Base

http://semanticommunity.info/%40api/...?origin=mt-web

Data Science for Agency Initiatives 2015: Spreadsheet CMS Data Knowledge Base

http://semanticommunity.info/Data_Science/Data_Science_for_Agency_Initiatives_2015

Updated: Wed, 23 Sep 2015 08:12:44 GMT

Powered by mindtouch
Slide 14 Data Science for Agency Initiatives 2015: CFPB Consumer Complaints-Spotfire Counts by State

Slide 15 Data Science for Agency Initiatives 2015: FCC Data-Spotfire Cover Page
Slide 16 Data Science for Agency Initiatives 2015: FCC Data-Spotfire Analytics

Web Player

Slide 17 Data Science for Agency Initiatives 2015: CMS Data-Spotfire Cover Page

Web Player
Slide 18 Data Science for Agency Initiatives 2015: CMS Data-Spotfire Analytics

Web Player

Slide 19 Data Science for Agency Initiatives 2015: CMS Data-Spotfire Visualizations

Web Player
Slide 20 Conclusions and Recommendations

Drawing Causal Inference from Big Data

Conclusions and Recommendations

• The Government Leadership in the Data Age Event, featured four excellent government data age leadership speakers, and prompted me to do data science products on their datasets for a future meetup.
• My question to all the panel members was given all of this getting the data out, how much actual looking at the data is being done?
  – The GAO answer was they are looking at the data outliers to save taxpayers money.
• I found a limited amount of looking at the data was being done based on very limited published data science products, except for CMS.
• This conclusion is supported by a recent National Academy of Sciences Meeting on Drawing Causal Inference from Big Data, March 26-27, 2015, which concluded: Although we are producing and storing ever greater amounts of data, we have just begun to figure out ways to analyze and understand what the data show.

Spotfire Dashboard

CFPB Consumer Complaints

For Internet Explorer Users and Those Wanting Full Screen Display Use: Web Player Get Spotfire for iPad App

Media, iframe, embed and object tags are not supported inside of a PDF.
Research Notes

Big Data
'Irish cynics' talk big data's good, bad and ugly

Source: http://fcw.com/articles/2015/05/12/d...-bad-ugly.aspx

Zach Noble
Federal Computer Week
May 12, 2015

Big data opportunities abound, sure, but if you were looking for a bubbly pep talk about innovation and data, you were bound to be a little jarred by the cold slap of pragmatism presented at FedInsider’s “Government Leadership in the Data Age” panel May 12.

The discussion was bookended by a pair of “Irish cynics” (self-labeled by Tony Summerlin) who kept the discussion firmly grounded in reality: Niall Brennan, chief data officer and director of analytics and enterprise data for the Centers for Medicare and Medicaid Services, and Summerlin, special advisor to the CIO at the Federal Communications Commission.

When it comes to big data, “[There are] way too many buzzwords, way too many white elephants, way too much trust placed in magical, out-of-the-box solutions,” said Brennan. “Big data holds incredible promise but it needs to be linked to the fundamentals.”

Those fundamentals include subject matter experts who can cast an educated eye on the data while providing a problem-solving approach and a focus on tangible outcomes rather than meaningless process, Brennan said.
Summerlin echoed Brennan’s advice and urged clear thinking as agencies jump into data collection and analytics.

“Why are you collecting this in the first place?” Summerlin said was a crucial — but oft-neglected — question.

Too often, Summerlin said, government agencies take a more-is-more approach, collecting superfluous data, performing unnecessary, basic analysis and slapping it all together in ways that aren’t useful.

“We think that just putting all this [data and analysis] together somehow creates value for somebody,” he said. “I’m not so sure.”

Another issue the pair tackled: the protection of personal privacy within massive caches of data.

“I think the horse is out of the barn,” Summerlin said of personal privacy on the Internet. “I can find out anything on anyone anywhere at any time.”

He applauded the federal agencies that try to scrub personally identifying information from data before using it for analysis, but he noted the odd position of the government as a whole.

“Part of the government is trying to ensure you have no anonymity whatsoever,” he said, referring to intelligence agencies’ surveillance programs, “while the other part is trying to protect [personal information]. You can’t have it both ways.”

Several panelists also noted the difficulties in sharing data between government agencies, and the struggle to ensure that data is kept in clean, machine-ready formats.

The panel did serve up a few success stories with the skepticism.

The Government Accountability Office’s Joah Iannotta noted how a big data mashup between FEMA insurance applications and federally backed mortgage info had enabled her agency to identify likely fraud in the wake of Hurricane Sandy. And Consumer Financial Protection Bureau Chief Data Officer Linda Powell shared how CFPB data analysts discovered medical debt — much of it undeserved or even unbeknownst to consumers who thought insurance had taken care of things — was tanking peoples’ credit scores. Credit reporting agencies now give much less weight to medical debts in calculating credit scores, thanks to the CFPB pointing out the messy insurance situation.

All four panelists agreed that big data holds immense opportunity — much of it waiting for creative minds to find it.

“Err on the side of openness,” Brennan advised, noting the value of publishing data sets because some data that might appear to be “crap” could yield valuable insights if the right person is able to use it.

---

**Government Leadership in the Data Age**


Date: Tuesday, May 12, 2015
Cost: Complimentary
Credits: 2 CPEs & Training Certificate from GWU CEPL
Agenda

7:30 - 8:30am - Registration, Breakfast, and Networking
8:30 - 10:00am - Interactive Panel Discussion with Audience Participation

Overview

The declining cost of data collection, storage, and processing coupled with ever-growing sources of digital data are spurring incremental increases in the amount of information available to government, law enforcement, intelligence, health agencies and other agencies. Big data is transforming the way we govern and live, and the Administration is taking steps to minimize potential risks such as loss of privacy, discrimination, and threats to consumer rights.

The White House has charged government agencies to leverage data to enable transparency, provide security, and foster innovation for the benefit of the American people.

This briefing will look at steps agencies are taking to push the frontiers of data science and assume a leadership role in the Data Age.

Speakers

Linda F. Powell, Chief Data Officer, Consumer Financial Protection Bureau

Tony Summerlin, Special Advisor to the CIO, FCC
Dr. Joah Iannotta, Assistant Director, Government Accountability Office

We'll explore

- Agency initiatives that are harnessing data to maximize the impact of citizen services.
- Methods to insure that proper transparency and oversight are in place for data practices.
- Policy initiatives to maintain data privacy for the American people.
- The impact of the Internet of Things on the Big Data revolution.
- Sources of digital data and their use in driving innovation.

Complimentary Registration

Please register using the form above, or call (202) 237-0300.

Participants can earn 2 CPE credits in Business Management and Organization.

Niall Brennan, Chief Data Officer, Centers for Medicare & Medicaid Services

Chris Wilkinson, Senior Director, Market Development, immixGroup
To receive CPE credit you must arrive on time and sign in on the CPE sheet (located at the registration desk). Certificates will be e-mailed to registrants. In accordance with the standards of the National Registry of CPE Sponsors, 50 minutes equals 1 CPE.

Additional Information

Prerequisites: Previous experience or education in federal IT procurement.

Advance Preparation: Basic experience in federal IT and cyber recommended.

Program Level: Intermediate

Delivery Method: Group Live Training

This event is presented by FedInsider and immixGroup.

Have questions about Government Leadership in the Data Age? Contact FedInsider, GWU CEPL, immixGroup, Cloudera, HGST, HP, Red Hat, and MySQL

Consumer Complaint Database

Source: http://www.consumerfinance.gov/complaintdatabase/

My Note: Link from Hone Page: http://www.consumerfinance.gov/

We share complaints about financial products and services to improve the financial marketplace.

My Note: See 62 MB CSV File and See Data Dictionary: Field reference

Get the Data

Use our API or download the full dataset. View all data

My Notes:

Dataset and views
The Consumer Complaint Database consists of a single Dataset -- the equivalent of a single database table or Excel worksheet. It also provides a number of persistent Views of the dataset, providing filtered slices of the data for targeted analysis.

http://catalog.data.gov/dataset
http://catalog.data.gov/dataset/cons...laint-database

https://data.consumerfinance.gov/api...sType=DOWNLOAD
What happens when I submit a complaint?

When you submit a complaint, we forward your complaint to the company and work to get a response about your issue.

Submit a complaint Learn more

What information do you publish?

We publish information about the subject and date of the complaint and the company’s response. We do not share any personal information with the public. Learn more

Data by product

All

Bank account or service

Credit card

Credit reporting

Debt collection

Money transfers

Mortgage

Other financial service

Payday loan

Prepaid card

Student loan
**Vehicle or other consumer loan**

We don’t verify all the facts alleged in these complaints but we take steps to confirm a commercial relationship between the consumer and company. Complaints are listed here after the company responds or after they have had the complaint for 15 calendar days, whichever comes first. We remove complaints if they don’t meet all of the publication criteria. Data is refreshed nightly.

**Download, sort, and visualize**

We’re using a tool called Socrata to make it easier to view and organize the data into subsets and visualizations. Additionally, Socrata provides a RESTful API for programmers and researchers.

Visualization tutorials from data.gov  Technical documentation  My Note: See excerpts below.

**Example visualizations**

**Complaints by channel**

![Complaints by channel graph](image)

**Credit card complaint responses by issue**

![Credit card complaint responses by issue](image)
Complaints by product

WHAT IS THIS DATA?

Source: http://www.consumerfinance.gov/compl...documentation/

The Consumer Complaint Database contains complaints we've received about financial products and services. It has complaints about:

- Bank accounts or services
- Consumer loans
- Credit cards
- Credit reporting
- Debt collection
- Money transfers
- Mortgages
- Private student loans
- Payday loans
- Prepaid cards
- Other consumer loan (such as pawn and title loans)
- Other financial service (such as credit repair and debt settlement)

We list complaints in the database when the company responds to the complaints, which confirms a relationship with the consumer, or after the company has had the complaint for 15 calendar days, whichever comes first. Complaints can be removed if they do not meet all of the publication criteria defined in our policy statement. In addition, we exclude any complaints that:

- Are missing critical information such as company or product,
- Have been referred to other agencies,
- Are duplicative, and/or
- Identify the incorrect company.
Data is refreshed nightly to include new complaints and update existing ones.

Field reference

The following fields are currently included in the database.

<table>
<thead>
<tr>
<th>FIELD NAME</th>
<th>DESCRIPTION</th>
<th>DATA TYPE</th>
<th>NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complaint ID</td>
<td>The unique identification number for a complaint</td>
<td>number</td>
<td></td>
</tr>
<tr>
<td>Product</td>
<td>The type of product the consumer identified in the complaint</td>
<td>plain text</td>
<td>This field is a categorical variable.</td>
</tr>
<tr>
<td>Sub-product</td>
<td>The type of sub-product the consumer identified in the complaint</td>
<td>plain text</td>
<td>This field is a categorical variable. Not all Products have Sub-products.</td>
</tr>
<tr>
<td>Issue</td>
<td>The issue the consumer identified in the complaint</td>
<td>plain text</td>
<td>This field is a categorical variable. Possible values are dependent on Product. On December 18, 2013 the issues for student loan complaints were revised. “Repaying your loan” and “Problems when you are unable to pay” were removed and “Can't pay my loan” and “Dealing with my lender or servicer” were added. Complaints received beginning on that date reflect this change. Complaints received before that date remain unchanged.</td>
</tr>
<tr>
<td>Sub-issue</td>
<td>The sub-issue the consumer identified in the complaint</td>
<td>plain text</td>
<td>This field is a categorical variable. Possible values are dependent on product and issue. Not all Issues have corresponding Sub-issues. On December 18, 2013, sub-issues were added for student loan complaints. Previously, sub-issues were not used for this type of complaint. Complaints received beginning on that date reflect this change. Complaints received before that date remain unchanged.</td>
</tr>
<tr>
<td>ZIP code</td>
<td>The consumer’s reported mailing ZIP code for the complaint</td>
<td>plain text</td>
<td>Includes only the first five digits and is blank for complaints submitted with non-numeric values. Excludes ZIP codes for areas with populations of 20,000 or fewer persons. In some instances, the consumer-submitted ZIP code and State fields may not match.</td>
</tr>
<tr>
<td>Submitted via</td>
<td>How the complaint was submitted</td>
<td>plain text</td>
<td>This field is a categorical variable.</td>
</tr>
<tr>
<td>FIELD NAME</td>
<td>DESCRIPTION</td>
<td>DATA TYPE</td>
<td>NOTES</td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>-----------</td>
<td>-------</td>
</tr>
<tr>
<td>submitted to CFPB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>State</td>
<td>The consumer's reported mailing state for the complaint</td>
<td>plain text</td>
<td>This field is a categorical variable.</td>
</tr>
<tr>
<td>Date received</td>
<td>The date the CFPB received the complaint</td>
<td>date &amp; time</td>
<td></td>
</tr>
<tr>
<td>Date sent to company</td>
<td>The date the CFPB sent the complaint to the company</td>
<td>date &amp; time</td>
<td></td>
</tr>
<tr>
<td>Company</td>
<td>The complaint is about this company</td>
<td>plain text</td>
<td>This field is a categorical variable.</td>
</tr>
<tr>
<td>Company response</td>
<td>This is how the company responded to the complaint</td>
<td>plain text</td>
<td>This field is a categorical variable.</td>
</tr>
<tr>
<td>Timely response?</td>
<td>Whether the company gave a timely response</td>
<td>plain text</td>
<td>yes/no</td>
</tr>
<tr>
<td>Consumer disputed?</td>
<td>Whether the consumer disputed the company's response</td>
<td>plain text</td>
<td>yes/no</td>
</tr>
</tbody>
</table>

**More information**

Socrata provides a number of resources for users of SODA APIs like the Consumer Complaint Database API. Code samples and other examples are available at [http://github.com/socrata](http://github.com/socrata)
How we use complaint data

Complaints help us understand the financial marketplace and protect consumers. Learn more

Reports to Congress

Consumer Response annual report: March 2015

Spring 2014 semi-annual report

Consumer Response annual report: March 2014

March 2013 semi-annual report

Consumer Response annual report: March 2013

Consumer Response annual report: March 2012

January 2012 semi-annual report

Consumer Response interim report on CFPB's credit card complaint data: November 2011

Snapshots of complaints

Complaints by the numbers: March 2015

A snapshot of complaints received: July 2014

A snapshot of complaints received: July 2013

A snapshot of complaints received: March 2013

A snapshot of complaints received: October 2012

A snapshot of complaints received: June 2012

API & Documentation

Consumer Complaint Database API

Our API is a suite of RESTful commands you can use to read and filter data from the database using scripts. It is built on the Socrata Open Data API (SODA), and users of other APIs built on this technology will find it familiar.

View documentation
Publication criteria

Certain complaint data, like personal information, is not included in this database. Complaints must meet all of the publication criteria in our policy statement.

View policy statement

View narrative data policy statement

View narrative scrubbing standard

FCC Data

Source: https://www.fcc.gov/data

Data underpins every activity at the Federal Communications Commission. By better involving data in open and transparent rulemaking, the FCC can better serve the public while enabling public innovation. The FCC has long published relevant data, though the process of improving its quality, openness, accessibility and utility warrants continuous progress.

Data Innovation Initiative

The Data Innovation Initiative is a cross-agency effort to modernize and streamline how the FCC collects, uses, and disseminates data. Its goals are improving the agency's fact-based, data-driven decision-making while lowering barriers and burdens to filing necessary information and sharing data more effectively to spur innovation, new competition, and markets.

Information & Data Officers

The FCC has Information and Data Officers in each Bureau and Office to ensure better use of data.

Zero-Based Data Review

Reviewing the FCC data collections is a core piece of the Data Innovation Initiative. We have started the process of reviewing all agency data collections, starting with the Wireline, Wireless, and Media Bureaus and a public notice for comments.

Spectrum Inventory

The FCC has assembled and put online a comprehensive, searchable baseline inventory of spectrum and holders of commercial spectrum usage rights.
Search FCC Databases

Explore granular search interfaces into more than 40 specialized FCC databases such as radio call signs and equipment authorization.

Download FCC Datasets

Over 150 data sets are available for download today.

2.948 Listed Test Firms

2.948 Listed Test Firms.XML

Accredited Test Firms

Accredited Test Firms.XML

Automated Reporting Management Information System

Welcome to the FCC's Electronic ARMIS Filing System - Data Retrieval Module--your portal to the ARMIS database containing financial and operational data of the nation's largest local exchange carriers that file this data in compliance with Part 43 of the Commission's Rules.

- ASCII Files -- This link provides users with the ability to download ASCII (comma-delimited) files, beginning with reporting year 1996.

EAFS ASCII File Download

- 

FCC Report 43-01, the Annual Summary Report

FCC Report 43-02, the USOA Report (2007 and prior years)

FCC Report 43-03, the Joint Cost Report (2007 and prior years)

FCC Report 43-04, the Access Report (2007 and prior years)

FCC Report 43-05, the Service Quality Report (2009 and prior years)

FCC Report 43-06, the Customer Satisfaction Report (2009 and prior years)

FCC Report 43-07, the Infrastructure Report (2007 and prior years)

FCC Report 43-08, the Operating Data Report
Cable Operations and Licensing System

My Note: Similar to International Bureau Application Filing & Reporting System

The following is an index to radio assignment information extracted from the various licensing systems at the Commission and made available to the public via the Commissions Web site. Although all these licensing systems contain similar administrative and technical information about licensed and applied for facilities, the structure of the data bases varies according information required to license the radio facility or system.

Notice about data usage:

The material in these data and text files are provided as-is. The FCC disclaims all warranties with regard to the contents of these files, including their fitness. In no event shall the FCC be liable for any special, indirect, or consequential damages whatsoever resulting from loss or use, data or profits, whether in connection with the use or performance of the contents of these files, action of contract, negligence, or other action arising out of, or in connection with the use of the contents of these files.

Question regarding the content accuracy of these data bases should be directed to the bureau administering the licensing system.

Cable Communities Registered with the FCC

The page you're looking for is currently unavailable to view.

Consolidated Public Database System

This single zip file contains all the public CDBS database files (except for the very large file ownership_other_int_xml_data, DTV White Space files, and the 2009 Biennial Ownership Data files):

- FTP: all-cdbs-files.zip
- HTTP: all-cdbs-files.zip (0.00 MB)

Equipment Authorization Grantee Registrations

Equipment Authorization Grantee Registrations.XML

http://semanticommunity.info/Data_Science/Data_Science_for_Agency_Initiatives_2015
Updated: Wed, 23 Sep 2015 08:12:44 GMT
Powered by Mindtouch™
FM Service Contour Data Points

In response to many requests for radio service area data, the Audio Division of the Media Bureau, Federal Communications Commission (FCC), releases predicted service contour data points on a regular basis. Generally, updated files are posted to the FCC's website by approximately 10:00 AM each day, Eastern time. These zipped files are available at the following locations (include capital letters and underscores as shown):


The data is contained in one large file:

- FM_service_contour_current.zip

The file contains data for active records (licenses, construction permits, applications) as shown in the FM Query.

Using the FCC’s F(50,50) propagation curves, distances to the FM service contours are generated from the effective radiated power in a given direction, and the radial antenna height above the average elevation of that same radial. See FM Propagation Curves and Antenna Height Above Average Terrain (HAAT) Calculations for additional information on this process.

USERS TAKE NOTE: The files are LARGE! We recommend using a broadband connection, making sure your system has enough disk space to handle these files. Current data files, unzipped, may be on the order of 200 MB in size.

International Bureau Application Filing & Reporting System

My Note: TXT and ZIP files

The following is an index to radio assignment information extracted from the various licensing systems at the Commission and made available to the public via the Commission's Web site. Although all these licensing systems contain similar administrative and technical information about licensed and applied for facilities, the structure of the data bases varies according information required to license the radio facility or system.

Notice about data usage:

The material in these data and text files are provided as-is. The FCC disclaims all warranties with regard to the contents of these files, including their fitness. In no event shall the FCC be liable for any special, indirect, or consequential damages whatsoever resulting from loss or use, data or profits, whether in connection with the use or performance of the contents of these files, action of contract, negligence, or other action arising out of, or in connection with the use of the contents of these files.

Question regarding the content accuracy of these data bases should be directed to the bureau administering the licensing system.

License View Database

http://data.fcc.gov/download/license...csv-format.zip (About 1000 MB)
Raw HTML files hosting data for ECFS

NOT FOUND

Section 43.61 International Traffic Data

Click on the file name to download the report or data.

File names ending with a .PDF extension can be viewed using Adobe Acrobat Reader. File names ending with a .ZIP extension contain word processing and spreadsheet files.

TCB Designating Authorities (TDA)

TCB Designating Authorities (TDA).XML

Telecommunication Certification Bodies (TCB)

Telecommunication Certification Bodies (TCB).XML

Test Firm Accrediting Bodies (TFAB)

Test Firm Accrediting Bodies (TFAB).XML

Universal Licensing System

The Universal Licensing System (ULS) database downloads for specific wireless radio services are available as zip files and are updated weekly. To stay abreast of the daily changes to the databases, you may also download daily transaction files.

NOTE: The 800 MHz Vacated Spectrum Files are not updated on a daily or weekly basis. Look for Public Notices regarding file availability updates.

Some of the database files are very large.

Universal Licensing System (ULS) also provides daily transactions files to these databases.

If you have any comments or questions about this information, please contact Technical Support.

Database Documentation

My Note: Lots of Items
<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Created</th>
</tr>
</thead>
<tbody>
<tr>
<td>800 MHz Vacated Spectrum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Site Based</td>
<td>23.11 MB</td>
<td>4/15/2015</td>
</tr>
<tr>
<td>Market Based</td>
<td>1.15 MB</td>
<td>4/15/2015</td>
</tr>
<tr>
<td>Aircraft - 47 CFR Part 87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applications</td>
<td>14.49 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td>Licenses</td>
<td>10.12 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td>Amateur Radio Service</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applications</td>
<td>198.13 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td>Licenses</td>
<td>120.03 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td>Antenna Structure Registration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Category</td>
<td>Size</td>
<td>Date</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td>Applications</td>
<td>149.17 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td>Registrations</td>
<td>31.83 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td>FAA Determinations</td>
<td>30.75 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td><strong>Airport Facilities - 47 CFR Part 17</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Runways</td>
<td>3 MB</td>
<td>4/30/2015</td>
</tr>
<tr>
<td>Airports</td>
<td>1.4 MB</td>
<td>4/30/2015</td>
</tr>
<tr>
<td>Assignments and Transfers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applications</td>
<td>21.22 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td><strong>BRS &amp; EBS (Formerly known as MDS/ITFS)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applications</td>
<td>3.59 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td>Licenses</td>
<td>2.36 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td>Service</td>
<td>Applications</td>
<td>Date</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>--------------</td>
<td>---------</td>
</tr>
<tr>
<td>Commercial Radio and Restricted Radiotelephone - FRC</td>
<td>35.63 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td>General Mobile Radio Service (GMRS)</td>
<td>24.88 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td>Land Mobile - Private</td>
<td>423.33 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td>Service</td>
<td>Applications Size</td>
<td>Date</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Land Mobile - Commercial</td>
<td>218.67 MB</td>
<td>5/10/15</td>
</tr>
<tr>
<td>Licenses</td>
<td>83.7 MB</td>
<td>5/10/15</td>
</tr>
<tr>
<td>Land Mobile - Broadcast Auxiliary</td>
<td>2.76 MB</td>
<td>5/10/15</td>
</tr>
<tr>
<td>Licenses</td>
<td>5.77 MB</td>
<td>5/10/15</td>
</tr>
<tr>
<td>Maritime Coast &amp; Aviation Ground</td>
<td>10.88 MB</td>
<td>5/10/15</td>
</tr>
<tr>
<td>Licenses</td>
<td>7.76 MB</td>
<td>5/10/15</td>
</tr>
</tbody>
</table>

MDS/ITFS: (Now known as BRS & EBS)

See BRS & EBS
## Market Based Services

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Applications</strong></td>
<td>16.2 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td><strong>Licenses</strong></td>
<td>12.18 MB</td>
<td>5/10/2015</td>
</tr>
</tbody>
</table>

Microwave - 47 CFR Parts 74 and 101, and 3650-3700 MHz

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Applications</strong></td>
<td>187.17 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td><strong>Licenses</strong></td>
<td>101.16 MB</td>
<td>5/10/2015</td>
</tr>
</tbody>
</table>

Ownership

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Filings</strong></td>
<td>3.36 MB</td>
<td>5/10/2015</td>
</tr>
</tbody>
</table>

Paging - 47 CFR Part 22

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Applications</strong></td>
<td>4.74 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td><strong>Licenses</strong></td>
<td>4.4 MB</td>
<td>5/10/2015</td>
</tr>
<tr>
<td>Regional Service</td>
<td>Applications</td>
<td>License</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------</td>
<td>---------</td>
</tr>
</tbody>
</table>

**Spectrum Leasing**

<table>
<thead>
<tr>
<th>Form</th>
<th>Application</th>
<th>Size</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form 603-T Application</td>
<td>0.15 MB</td>
<td>5/10/2015</td>
<td></td>
</tr>
<tr>
<td>Form 608 Application</td>
<td>4.88 MB</td>
<td>5/10/2015</td>
<td></td>
</tr>
</tbody>
</table>

**Unlicensed Wireless Microphone Registration**

<table>
<thead>
<tr>
<th>Form</th>
<th>Application</th>
<th>Size</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form 612 Application</td>
<td>0.01 MB</td>
<td>5/10/2015</td>
<td></td>
</tr>
</tbody>
</table>

---

### Develop on FCC APIs

The FCC actively promotes the innovative application of agency data in the public and private sectors. FCC.gov/Developers connects citizen developers with the tools they need to unlock government data.

### Subscribe to FCC RSS Feeds

Subscribe to exactly what you need to know in real time through an FCC RSS feed from a variety of categories from notices to blog posts.
FOIA and Ex Parte

Each year the FCC responds to over 600 Freedom of Information Act requests. We are committed to making information on Ex Parte (‘by one side’) presentations in accordance with the law and our policy.

GAO Government Data Sharing Community of Practice

Source: http://www.gao.gov/aac/gds_community...ctice/overview

Federal government agencies face challenges in sharing information, using data analytics, and leveraging resources that could assist them in their missions. This community brings together officials from all levels of government to overcome these challenges.

Overview

In January 2013, GAO co-hosted a forum alongside the Council of the Inspectors General on Integrity and Efficiency (CIGIE) and the Recovery Accountability and Transparency Board to explore using data analytics—which involve a variety of techniques to analyze and interpret data—to help identify fraud, waste, and abuse in government. Forum participants included representatives from federal, state, and local government agencies as well as the private sector. A summary of the key themes from the forum is published here.

Through facilitated discussion, forum participants identified a variety of challenges that hinder their abilities to share and use data. For example, throughout the forum, participants cited the need for greater coordination and incentives for federal, state, and local government agencies to share information among themselves and with each other. To address these and other issues related to coordination and data sharing, GAO formed the Government Data Sharing Community of Practice.


My Note: I have repurposed this PDF to MindTouch: Data Analytics For Oversight and Law Enforcement and spoken with Stephen Lord at Predictive Analytics World Government 2013

Contacts

Stephen Lord
Director, Forensic Audits and Investigative Service
GovernmentDataShare@gao.gov
202-512-4379

Events

On April 22, as part of GAO’s ongoing Government Data Sharing Community of Practice, GAO will collaborate with the Data Transparency Coalition to host a discussion on the cultural changes required for a successful transformation to open data across the government.
Open data is the official policy of the U.S. government. A 2013 executive order required federal agencies to publish their information as standardized, searchable data, and the 2014 Digital Accountability and Transparency Act (DATA Act) mandated similar standards for the federal government's spending information. Even with these changes, the transformation to open data is not an easy one. Where it has been successfully achieved, the transformation required cultural change, specifically a shift from documents-based to data-centric thinking.

The next Government Data Sharing Community of Practice will focus on cultural change success stories from across the government, as well as areas where change has not happened yet, but might soon.

CPEs for this event are available for government employees for whom this information is relevant to their work.

• **Registration is open to the public**: Please note that while information posted earlier indicated the event would be free, a tiered registration fee has been implemented to cover administrative costs. Registration for government employees continues to be free, while the fee is $20 for non-profits and students and $45 for private industry.

• **Meeting agenda**

The Government Data Sharing Community of Practice features a series of discussions on challenges and opportunities related to sharing data in government. The timeline below provides more information on upcoming events and notes from previous events.

E-mail updates notify you about upcoming events or when notes from previous events are posted to the website. E-mail updates will be sent about once a month. Enter your e-mail address below to subscribe to receive e-mail updates.

**E-mail address (Required) My Note: I registered.**

**Timeline: Explore Recent Community of Practice Events**

**MINUTES**: [CoP Meeting Minutes - 1/26/15](http://www.gao.gov/aac/gds_community...ctice/overview)

PDF

Note: These minutes summarize the discussion that took place at the Government Data Sharing Community of Practice meeting. The summary does not necessarily represent the views of GAO or the organizations that the discussion participants represent.

Government Data Sharing Community of Practice

Panel Discussion on Data Sharing in Disaster Response and Recovery

Meeting Minutes

January 26, 2015, 9:30am-12:30pm

The MITRE Corporation, 7525 Colshire Drive, McLean, VA 22102

[http://www.gao.gov/aac/gds_community...ctice/overview](http://www.gao.gov/aac/gds_community...ctice/overview)
Background on the GAO’s Government Data Sharing Community of Practice

Federal government agencies face challenges in sharing information, using data analytics, and leveraging resources that could assist them in their programmatic and oversight missions. GAO formed its Government Data Sharing Community of Practice to foster an ongoing dialogue about strategies used to overcome challenges that federal, state, and local government agencies face in trying to share data to fulfill their missions. The Community of Practice was an outcome of a forum GAO hosted with the Council of the Inspectors General on Integrity and Efficiency (CIGIE) and the Recovery Accountability and Transparency Board to explore opportunities to use data analytics to identify and prevent fraud, waste, and abuse in federal government programs. GAO’s Government Data Sharing Community of Practice is open to all stakeholders, including those from both the public and private sector.

Data Sharing in Disaster Response and Recovery

1 A summary of the key themes from the forum is published at http://www.gao.gov/products/GAO-13-680SP. Minutes from previous sessions of GAO’s Government Data Sharing Community of Practice are available at http://www.gao.gov/aac/gds_community...e/overview#t=1

In order to respond effectively to a disaster and provide basic-needs assistance to affected individuals, federal, state and local government entities need to be able to share data and leverage a skilled workforce. In this session of the Government Data Sharing Community of Practice, representatives from the Department of Defense, Department of Housing and Urban Development, the White House Innovation for Disaster Response and Recovery Initiative, GAO, and MITRE discussed how federal, state and local government entities agencies have shared data to facilitate previous emergency management and recovery efforts and ongoing challenges. Representatives from the Center for Organizational Excellence, GAO, and the Office of Personnel Management discussed how data about employees’ skills may be generated and applied by federal agencies to deploy a skilled emergency management workforce.

Panel 1: Data Sharing in Disaster Response and Recovery

Moderator

Joah Iannotta, Assistant Director, Forensic Audits, U.S. Government Accountability Office

Panelists

Denice Ross, Presidential Innovation Fellow, U.S. Department of Energy; and support for the White House Innovation for Disaster Response and Recovery Initiative

Sara Meyers, Chief of Staff for the Federal Housing Administration, U.S. Department of Housing and Urban Development

Elmer Roman, Oversight Executive, U.S. Department of Defense

Don McGarry, MITRE

Presentations

http://semanticommunity.info/Data_Science/Data_Science_for_Agency_Initiatives_2015
Updated: Wed, 23 Sep 2015 08:12:44 GMT
Powered by mindtouch
Joah Iannotta, Assistant Director, Forensic Audits, U.S. Government Accountability Office

In her opening remarks Dr. Iannotta highlighted some of GAO’s recent work to identify fraud, waste, and abuse in the Federal Emergency Management Agency’s (FEMA) response to Hurricane Sandy (GAO-15-15). In particular, she discussed how the audit team brought together disparate data sets shared from FEMA and several federal housing authorities to identify risk factors that indicate potential fraud, waste, and abuse. Such risk factors include failure to self-report accurate information about home insurance.

Denice Ross, Presidential Innovation Fellow, U.S. Department of Energy; and support for the White House Innovation for Disaster Response and Recovery Initiative

Ms. Ross discussed how federal, state, and local entities need comprehensive and accurate data to form effective disaster responses. However, as Ms. Ross described, sharing government data across federal, state, and local entities presents significant challenges. Ms. Ross discussed two examples of challenges and successes with sharing data in post-disaster environments: Open data in New Orleans (data.nola.gov); and the federal disaster.data.gov website.

Open data in New Orleans (http://Data.nola.gov)

Federal, state and local entities did not have access to the data needed to shape an effective response to Hurricane Katrina. Most publicly federal and state datasets, such as the census, were “outdated” after the Hurricane because of physical destruction and displaced populations. While the City of New Orleans had valuable data about infrastructure (such as the map of property boundaries), city officials were reluctant to share the data with others over concerns of privacy and potential legal ramifications. The lack of data sharing hampered federal, state, and local responses and recovery efforts. Based on these experiences in responding to and recovering from Hurricane Katrina and lessons learned, the City of New Orleans created http://data.nola.gov, a catalogue of public datasets, including city administrative data, property boundary maps, planning and zoning data, data on healthcare facilities, and census data. The http://data.nola.gov facilitates data sharing among all public and private stakeholders. Part of what swayed the City to move to an open data solution was the fact that so many stakeholders were asking for the same information and that citizens had started to track this information and share it themselves.

White House’s http://disaster.data.gov website

Federal, state, and local agencies also encountered challenges in coordinate response and recovery efforts after Hurricane Sandy. Federal agencies, lead by the White House Office of Science and Technology Policy, created a catalogue of federal datasets that may be used in disaster response and recovery by federal, state, local, and private entities. The site was a part of the larger data.gov site that was launched in 2009 as part of the government-wide open data initiative. http://Disaster.data.gov currently includes 168 data sets and applications and tools for analyzing and using data to respond to disasters. All of the data sets available at disaster.data.gov are machine readable and accessible to government entities and the general public. Disaster.data.gov provides high quality datasets that facilitate disaster response activities performed by public and private entities.

Sara Meyers, Chief of Staff for the Federal Housing Administration, U.S. Department of Housing and Urban Development
Ms. Meyers described her experience with data sharing as head of the Project Management Office (PMO) within the interagency Hurricane Sandy Rebuilding Task Force. The Task Force was established by executive order to help coordinate the federal response to Hurricane Sandy and made 69 recommendations regarding recovery efforts. The PMO was established to monitor the $50 billion in supplemental disaster assistance spent by 19 federal agencies and oversee the implementation of the Task Force’s recommendation. Although the PMO had specific oversight responsibilities, the supplemental Bill did not grant the Task Force the authority to collect data from federal agencies receiving funding through the bill. Ms. Meyers discussed the challenges in overseeing large programs across multiple federal agencies without a legislative mandate.

Ms. Meyers discussed how the PMO developed working relationships with the federal agencies in order to identify relevant data and to establish ad hoc data sharing policies. The PMO established a working group that included program managers and oversight professionals from the federal agencies that received supplemental funding to support Hurricane Sandy recovery efforts. In coordination with federal agencies, the PMO developed standards for collecting program-level spending and performance data. After extensive conversations with agency officials, the PMO was able to obtain county-level data for certain programs. Ms. Meyers noted that finding individuals within each agency who understood data, its applicability to measuring performance, and the benefits to the agencies of sharing data for this purpose was key to forging data sharing arrangements. Additionally, the PMO coordinated with each agency to develop performance metrics that reflect the program’s objectives and recovery activities. As Ms. Myers described, the success of the data sharing efforts was driven by a shared understanding of the value of collecting and analyzing data across agencies and programs in order to understand how money is being spent and coordinate efforts to ensure that needs are being met and minimize duplication. Ms Myers concluded by emphasizing the importance that stakeholders have a shared understanding of the importance of sharing data to improve program outcomes.

The lack of a technology system for collecting and analyzing budget and performance data also presented challenges. The PMO relied on basic spreadsheets to organize and analyze data provided by the 19 agencies about 60 programs. Although the PMO requested agencies provide data in specific standardized formats, agencies often were unable to produce standardized data. The PMO provided technical assistance to agencies in order to get the requested data in a usable format. As Ms. Meyers discussed, agencies need to be flexible when developing and implementing data sharing activities.

Ms. Meyers noted that the Hurricane Sandy Rebuilding Task Force made several recommendations about data sharing. Two key data sharing recommendations included making aggregated non-personal identifying information (non-PII) data publicly available and reviewing the System of Record Notices (SORNs) for efficiencies.

The PMO partnered with the Recovery Board to post the data it collected to a publicly available website, www.recovery.gov. The published data revealed how agencies were spending money and highlighted the outcomes each agency’s recovery efforts. The data revealed what populations were being served, and how many people / businesses have received assistance, among other things.

Elmer Roman, Oversight Executive, U.S. Department of Defense (DOD)

DOD’s efforts to promote interoperability began in the 1980’s with the development of “joint interoperability” capabilities among the services. Mr. Roman described the importance of interoperability for leveraging resources, unifying actions and avoiding duplication and discussed the challenges in establishing interoperability with other US
agencies and international entities in disaster response and recovery activities. DOD has tremendous datasharing capabilities and is working with its partners to implement the data sharing capabilities developed by DOD.

Mr. Roman discussed how data-sharing technologies can improve the efficiency of disaster recovery efforts by leveraging resources and avoiding duplicative activities. He described three data-sharing technologies that the Department of Defense (DOD) has developed to facilitate interoperability and improve the efficiency of recovery efforts among DOD, and other entities:

UNITY tool is a data-sharing visualization tool for planning and coordinating disaster response missions. Unity includes mapping tools that can be used to depict the location of response and recovery activities. DOD first used UNITY to share data and coordinate with the State Department and U.S. Agency for International Development (USAID).

All Partners Across Network (APAN) is the unclassified information sharing network. It enables coordination and knowledge sharing for all stakeholders, including non-DOD entities. After the Haiti earthquake, DOD used APAN to coordinate the response and recovery efforts with other government agencies and nongovernmental organizations.

Geospatial Tool for Security, Humanitarian Assistance and Partnership Engagement (GeoSHAPE) integrates data from multiple sources (including unclassified geo-spatial and sensor data) and displays it in a dynamic Internet-based map to provide situational awareness, monitor the progress of recovery efforts, and inform decisionmaking. It is based on an open-source platform that is jointly managed by the University of Hawaii and the Office of the Secretary of Defense. DOD has used GeoShape to coordinate with the United Nations, and American Red Cross after Typhoon Haiyan struck the Philippines.

DOD is continuing to collaborate with other US government agencies, academic researchers, and non-government entities to facilitate the use of these data sharing tools in disaster response and recovery efforts. To ensure that these data sharing tools are accessible, remain up-to-date, and available as an open resource to disaster responders, Mr. Roman noted that the federal government should take a leadership role in owning and providing these technologies.

Don McGarry, MITRE

Mr. McGarry stated that data sharing enables effective disaster response and recovery efforts that involve federal, state, and local entities. Data sharing tools need to be interoperable with the technologies used by first responders. Mr. McGarry described how effective data-sharing technologies should be developed collaboratively among three groups: (1) first responders who understand the working environment; (2) technology experts who understand the potential technologies; and (3) policy experts who understand data applications and the challenges of sharing data. Without input from all three groups, a data sharing technology may not address the most critical emergency response needs. Data-sharing technologies that are developed in a coordinated manner will be useful to local entities and enable coordinated regional and national responses.

Mr. McGarry discussed the need to develop data sharing technologies that local jurisdictions can use in "routine" emergencies (i.e. traffic incidents) and can be integrated with existing technologies. Technologies should be based on basic, standardized architectures and data formats. Data standards will enable dynamic real-time data sharing. Access to timely data enables effective disaster responses. Mr. McGarry described MITRE's work with the Department of Homeland Security's (DHS) Office for Interoperability and Compatibility within DHS' First Responders Group. Their
collaboration has focused on developing standards and open source tools for first responder agencies across federal, state and local government.

**Panel 2: Disaster Response: Creating a More Agile and Responsive Federal Work force**

**Moderator**

Robert Goldenkoff, Director, Strategic Issues, and Rebecca Shea, Assistant Director, Applied Research and Methods, U.S. Government Accountability Office

**Panelists**

Lyn McGee, Vice President, National Security Solutions, Walter Sechriest, Solutions Director, and Chuck Simpson, Principal Consultant, Center for Organizational Excellence (COE)

Melissa Kline Lee, Program Manager for GovConnect, U.S. Office of Personnel Management

Rebecca Shea, Assistant Director, Applied Research and Methods, U.S. Government Accountability Office

**Slides**

Slide 1

*Human Capital and Data Issues in Disaster Recovery*

Slide 2
Speakers:
• Center for Organizational Excellence (COE) executives:
  – Lyn McGee, Vice President, National Security Solutions
  – Walter Schriest, Solutions Director
  – Chuck Simpson, Principal Consultant
• U.S. Office of Personnel Management (OPM)
  – Melissa Kline Lee, Program Manager for GovConnect
Moderators:
• U.S. GAO
  – Robert Goldenkoff, Director, Strategic Issues
  – Rebecca Shea, Assistant Director, Applied Research and Methods

Slide 3

**What does human capital have to do with disaster recovery?**

GAO-14-723T

• Strategic human capital management plays a critical role in maximizing the government’s performance and assuring its accountability to Congress and to the nation as a whole.
• Addressing challenges in areas such as disaster response, homeland security, economic stability, and numerous other complex and evolving issues requires a skilled federal workforce able to work seamlessly with other agencies, levels of government, and across sectors.

Slide 4

**So we’re all set, right?**

• Strategic human capital management has been a GAO high-risk area since 2001.
• The federal workforce is facing fundamental capacity challenges because of budget constraints, high rates of retirement eligibility, increasing attrition, downsizing, and skill gaps.
• The gap between demands, system capacity, and speed of response is exacerbated during a disaster. Agencies, especially those that are not in the business of being first responders, have difficulty handling fast-paced challenges.

Slide 5
People + Agile Workforce

- The challenges communities face in the wake of a disaster are horizontal in nature—they cut across different agencies and levels of government. However, our capacity to respond is organized vertically—often in stove-piped bureaucracies, departments, agencies, and governmental units.
- The government response could be more nimble. Agencies need to work as an organic, horizontal team to rapidly integrate diverse expertise and capabilities. The key challenge is how to reconfigure work groups and assemble teams regardless of agency affiliation to meet the needs of a community.

People + Skills & Competencies

- Disaster response is not just about the usual suspects e.g. FEMA, DHS. Longer term recovery could involve such agencies as IRS, SBA, Medicaid, etc.
- But it is more than just technical skills and expertise. Competencies need to be considered as well.

People + Culture

- There has been some success transferring individuals during times of crisis. For example, in response to the 2010 Deepwater Horizon Spill in the Gulf of Mexico, several agencies pitched in with the clean-up effort.
- Success in a more networked environment requires a whole new way of thinking about how to align different organizations, government, and nongovernment entities.
- Raises several culture based questions...
What does human capital data have to do with disaster recovery?

- A key first step is visibility over the workforce—knowing what skills and competencies employees have and where they are located.

So what is the problem?

- Most HR systems do not identify employee skills in any detail below occupational series, title and grade. As a result, leveraging employees’ skills is difficult.
- Maintaining a valid, up to date inventory of skills and competencies is difficult.
- Knowledge about existing human capital data is stove piped; skills inventory more so.

What is the Role of DUG in all this?

- Get rid of those stovepipes. Bring federal agencies, NGOs, research organizations, academic community together.
- Through the COP we share knowledge and research:
  - raise awareness of human capital data sources
  - identify gaps in research/knowledge to suggest new areas of evaluation
  - identify data quality/reliability issues and suggest improvements
  - enhance access to data
  - share analytical techniques, statistical innovation, and ideas on "big data" analytics

Slide 9

Slide 10

Slide 11
Some questions for consideration

- What would a skills/competencies database look like?
- Does a data system (or combination of systems) already exist that would serve? (As is? With modifications?)
- Who should manage it?
- How should the data be collected and kept up-to-date?
- How should the information be validated (i.e., ensure the level of skill is accurately recorded)?
- Should it be combined with other data sources e.g. geospatial and census data?
- What are the access, privacy and “big brother” implications of such a database and what protections would be needed?
- What are the cultural implications of an “agile workforce”

Lyn McGee, Vice President, National Security Solutions, Walter Sechriest, Solutions Director, and Chuck Simpson, Principal Consultant, Center for Organizational Excellence (COE)

Ms. McGee discussed COE’s experience in building and operating the federal government’s Enterprise Human Resource Integration (EHRI) database and how it could be used to create a flexible federal workforce that could facilitate disaster responses. EHRI is a fully integrated human resources information system that includes standardized personnel information and secured information access and sharing. Currently, EHRI does not contain standardized, detailed information about employee skills. This sort of data is maintained by individual agencies and departments. According to Ms. McGee, EHRI system developers could create a system that would facilitate the collection of standardized data on employee skills and allow managers to identify federal employees with the needed expertise to support disaster response and recovery activities. A database of employee skills across the federal government could reduce barriers and stovepipes and facilitate the development of a nimble federal workforce. Federal agencies could leverage a nimble workforce to deploy federal employees with the necessary skills in disaster recovery programs. As Ms. McGee noted, federal agencies have stovepiped missions and may resist efforts to establish a nimble workforce and share employees as sharing employees could detract from an agency’s primary mission. Established an overarching disaster recovery mission for all agencies may reduce such barriers.

Ms. McGee also discussed how data standards are dynamic and often change over time. Changing data standards may create data errors. For example, if there is a standard job code, in a specific region, then the code will be invalid if the
job series is changed or there is a departmental reorganization. Nonetheless, such errors could be identified and fixed through routine data verification procedures.

Ms. McGee stated that agencies / programs must develop a capacity to use data in an emergency situation. Emergency response exercises should include data verification and analysis exercises. It is important to routinely verify the accuracy of the data to ensure that the data can be used in the event of an emergency. Furthermore, Ms. McGee indicated that an overarching disaster response mission across the federal government could facilitate data sharing and the exchange of other resources, including human capital.
What Do We Know?

- Enterprise Human Resource Initiative (EHRI)
  - Integrated HR information for 2.1M Federal employees dating back to 1972.
  - EHRI supports HR, payroll, and training data submissions from 70+ agencies and 33 providers every two weeks
  - EHRI = DW, MDR, RDR
  - Current, centralized, and integrated workforce data.
  - Data feeds send EHRI data to both FedScope.gov and data.gov, saving approximately $5.5M/year in FOIA.
  - OPM meets all federal security mandates with zero data breaches of any kind since project inception.
What Do We Need?

- Feds who say their agency suffers from a gap in data skills: 96%
- Feds who say their managers prioritize use of data in decision making: 66%
- Feds who say their agency provides adequate training in data analytics: 22%
- Feds who say their agency’s biggest skills gap is turning data insights into action: 45%

How Do We Get There?

- An on-line “opt-in” environment
  - Self-certify
  - Increased agility
- Deployment Roadmaps
  - Training
  - Certifications
  - Preparation
- Data accuracy
  - Database accuracy is driven by use
  - How can we increase data utility between disasters?

How Do We Get There: Policy

- Interagency cooperation challenges:
  - OAP: Other Agency’s Problem: agencies view emergency response as the responsibility of designated agencies such as FEMA, DOD, DOS and USAID.
  - A deployed employee is paid for by employee agency, not responsible agency
  - Demand is not predictable
- We must break the code:
  - Many federal employees already have the security clearances that are required to effectively perform in interagency operations
  - Federal employees have often been subject to medical screening and immunizations.
  - Federal employees are able to perform inherently governmental work
  - Federal civilians have the advantage of knowing and using government regulations.
  - Federal ethics rules, accountability concepts, information systems and generally accepted behavioral norms.
Melissa Kline Lee, Program Manager for GovConnect, U.S. Office of Personnel Management

Ms. Lee described GovConnect, an information technology platform for “people sharing” programs across the federal program that could facilitate disaster response programs. GovConnect is currently under development, but has been piloted at several agencies in workforce agility experiments. She presented preliminary results from three workforce agility pilot programs:

GovProject allows managers to initiate special projects. Employees with appropriate skills and/or interest volunteer to contribute to the project. Work on special projects is on a part-time basis. As people apply to contribute to special projects, agencies collect data on employees’ skills and interests. Managers may use this information in strategic planning and workforce development.

GovStart allows employees (not managers) to initiate “passion projects.” Employees across the agency are allowed to participate in these projects on a part-time basis. GovStart may improve employee morale by empowering employees to initiate projects that improve agency programs and stimulate the development of innovative programs that support the agency’s mission.

GovCloud allows employees hired by one agency to be detailed out to another agency for specific projects. GovCloud is in the earliest stage of development and a pilot program is being planned.
These programs could be used to connect federal employees to special projects to support disaster response efforts.

MINUTES: CoP Meeting Minutes - 4/28/2014
PDF

MINUTES: Technological Challenges to Data Sharing, Meeting Minutes - 2/12/2014
PDF

MINUTES: CoP Meeting Minutes - 11/20/2013
PDF

REPORT CG FORUMS & ROUNDTABLES: Data Analytics for Oversight & Law Enforcement

April 22nd Meeting Agenda

Source: http://www.datacoalition.org/changing-for-open-data/

Changing the Culture for Open Data

On April 22, the Data Transparency Coalition and the U.S. Government Accountability Office (GAO) Government Data Sharing Community of Practice co-hosted panel discussions focusing on the cultural change necessary for government agencies to embrace open data. A short recap of the event can be found here.

Open data is the official policy of the U.S. government. President Obama issued his Open Data Policy on May 9, 2013, requiring executive agencies to publish their information as standardized, searchable data. On May 9, 2014, President Obama signed the unanimously-passed Digital Accountability and Transparency Act (DATA Act), which legally mandates this transformation for the federal government’s spending information. Open data requires agencies to adopt consistent data fields and formats for information that they collect, share, and disseminate – and then make that information fully available for everyone to use.

Even with a Presidential executive order covering all government information and a law mandating open data within spending, the transformation to open data is not an easy one. Wherever it has been successfully achieved, the transformation requires cultural change: a shift from documents-based to data-centric thinking.
Success stories come from wildly different areas of government and involve diverse types of information – from campsite availability at the Forest Service to geospatial information at the Department of the Interior. Yet these cultural changes share certain characteristics in common: all of them have needed strong championship from senior leaders, all of them have relied more on relationship-building than mandatory fiat, and all of them involved some internal benefit to be derived from the shift from documents to data.

This event will explore past success stories of open-data cultural change from across the government – and then dig into some key areas of government information where cultural change hasn’t happened yet, but might soon.

**Agenda**

1:30 pm — Welcome
Joah Iannotta, GAO
Hudson Hollister, Data Transparency Coalition
Ken Melero, Socrata

1:45 pm — Panel One: Stories of Cultural Change

Moderator: Joah Iannotta, GAO
Panelist: Rick DeLappe, Recreation One-Stop Program Manager, National Park Service, USDA
Panelist: Carrie Hug, Director of Accountability, Recovery Accountability and Transparency Board
Panelist: Camille Calimlim Touton, Counselor to the Assistant Secretary for Water and Science, Department of the Interior

3:00 pm — Break

3:30 pm — Panel Two: Future Challenges for Cultural Change

Moderator: Hudson Hollister, Data Transparency Coalition
Panelist: Raphael Majma, Innovation Specialist, 18F, GSA
Panelist: Joel Gurin, president, Center for Open Data Enterprise
Panelist: Jerry Johnston, PhD, Geospatial Information Officer, Department of the Interior

---

**How CMS is Using Big Data to Spur Healthcare Transformation**

Source: [http://www.bigdatahitforum.com/session/transformation](http://www.bigdatahitforum.com/session/transformation)

General Session

10:45 - 11 AM, Thursday, November 20, 2014

SPEAKER
As the largest single payer for health care in the United States, the Centers for Medicare and Medicaid Services (CMS) generates enormous amounts of data. Historically, CMS has faced technological challenges in storing, analyzing, and disseminating this information because of its volume and privacy concerns. However, rapid progress in the fields of data architecture, storage, and analysis—the big-data revolution—over the past several years has given CMS the capabilities to use data in new and innovative ways.

In this session, Niall Brennan, acting director of the Offices of Enterprise Management at CMS, will describe the different types of CMS data being used both internally and externally, and highlight a selection of innovative ways in which big-data techniques are being used to generate actionable information from CMS data more effectively. These include the use of real-time analytics for program monitoring and detecting fraud and abuse and the increased provision of data to providers, researchers, beneficiaries, and other stakeholders.

In the long term, CMS aims to ensure that all participants in the health care system have the right data at the right time and in the format that helps them make the best care delivery decisions. Big-data tools and techniques give CMS an ever-increasing ability to accomplish these goals.

Source: http://www.bigdatahitforum.com/big-d...session-slides
Slide 2 How CMS is Using Big Data to Spur Healthcare Transformation 2

How CMS is Using Big Data to Spur Healthcare Transformation

Niall Brennan
@N_Brennan

Slide 3 Introduction

Introduction

- CMS is the largest single payer for health care services in the US
- 2.5 billion claims submitted annually
- Significant additional data sources on the way
  - Meaningful use
  - Provider quality
  - Medicare Advantage encounter data
  - Health Insurance Marketplace/Medicaid expansion data
- Receive billions of other “non-claim” data points
- Trusted to protect beneficiary privacy
Slide 4 CMS Data and Delivery System Transformation

CMS Data and Delivery System Transformation

- The big data revolution has given CMS the capabilities to use and share data in new and innovative ways
- CMS collects a wealth of data that is critical to decision making for the agency and other stakeholders in the health care system
- To promote delivery system transformation, CMS is:
  - Employing advanced analytics to create actionable information products, accelerate transparency, inform policy decisions and evaluate programs
  - Routinely and safely sharing data with numerous stakeholders to drive health care quality and efficiency improvements and lower health care costs

Slide 5 Creating Information Products

https://dnav.cms.gov/

Creating Information Products

- CMS has made a commitment to increasing data transparency through making more machine-readable, de-identified data available publicly
- CMS Data Navigator (https://dnav.cms.gov/) makes it easy to find CMS data and information products on our website

Slide 6 Geographic Variation and Chronic Condition Data


http://semanticommunity.info/Data_Science/Data_Science_for_Agency_Initiatives_2015
Updated: Wed, 23 Sep 2015 08:12:44 GMT
Geographic Variation and Chronic Condition Data

- Datasets with aggregated indicators at the state, HRR and county level
- Based on 100% Medicare claims data for beneficiaries enrolled in FFS for 2007-2012
### Slide 10 Recent High Profile PUF Releases

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Key Variable(s)</th>
<th>Years (release date)</th>
<th># of Providers</th>
<th># of Records</th>
<th># of Page Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital Inpatient</td>
<td>DRG</td>
<td>FY2011 (May '13)</td>
<td>3,000+ Hospitals</td>
<td>Over 155,000</td>
<td>350,000</td>
</tr>
<tr>
<td>Hospital Inpatient</td>
<td></td>
<td>FY2012 (June '14)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital Inpatient</td>
<td>APC</td>
<td>CY2011 (June '13)</td>
<td>3,000+ Hospitals</td>
<td>Over 40,000</td>
<td>96,000</td>
</tr>
<tr>
<td>Hospital Outpatient</td>
<td></td>
<td>CY2012 (June '14)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physicians &amp; Other</td>
<td>HCPCS, Place of Service</td>
<td>CY2012 (April '14)</td>
<td>880,000+ NPIs</td>
<td>Over 9M</td>
<td>400,000</td>
</tr>
<tr>
<td>Suppliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Extensive coverage in national television and print media (all major TV outlets, New York Times, Wall Street Journal, Washington Post, hundreds of regional newspapers and TV stations)

### Slide 11 Significant Use of Data by External Stakeholders

- Access to granular data promotes development of external tools
- Data used in several "hackathons" including at Health Datapalooza in 2014
Slide 12 User Friendly Interfaces

User Friendly Interfaces

- Search for a provider by name, address, or National Provider Identifier (NPI)
- Tool returns information about the services the provider furnished to Medicare beneficiaries

Slide 13 Data Dissemination Activity

Data Dissemination Activity

- CMS is routinely and safely sharing data to support the transformation of the delivery system
  - Research
  - Provider Performance Measurement – Qualified Entities (QEs)
  - Care Coordination – Accountable Care Organizations (ACOs) and States
- CMS has also allowed beneficiaries full and open access to their Medicare claims data through the Blue Button Initiative
Slide 14 Research Data Dissemination

Research Data Dissemination

- The Chronic Condition Warehouse (CCW) is CMS’ research data warehouse designed to support external researchers and internal CMS research and analytic functions.
- Unique beneficiary ID allows user to link data across all CCW data, including:
  - Medicare beneficiary demographics and enrollment (1999-current)
  - Medicare fee-for-service (FFS) claims (1999-current)
  - Medicare Part D event data (2006-current)
  - Medicaid eligibility and claims (1999-2009)
  - Medicare-Medicaid linked files (2006-2008)
  - Assessment data (instrument inception-current)
- New data access method: Virtual Research Data Center (RDC)

Slide 15 CMS VRDC Benefits

CMS VRDC Benefits

ACCESS
- Researchers use own laptop to securely access data remotely.
- Increases efficiency of data sharing and reduces infrastructure costs for data users.

SECURITY
- No shipping of data on external media.
- Users only see data files with the data they need to conduct their project.
- CMS can track and monitor use of the data.

PRIVACY
- Users may only remove aggregated output files; no granular identifiable output may be taken.
- Out of the VRDC.
- CMS encrypts all beneficiary identifiers.

DATA & ANALYSIS
- Users can perform their own analyses and data manipulation in the virtual environment.
- Secure File Transfer System allows users to upload their own data and download output files efficiently and securely.
Slide 16 Data Sharing for Performance Measurement

Data Sharing for Performance Measurement

- Qualified Entity (QE) Program authorizes the release of Medicare claims data to QEs for performance evaluation of providers
- QE Program changes the performance measurement landscape by facilitating the creation of actionable performance reports that cover all/most of a providers’ practice

Slide 17 Certified QEs

Certified QEs

- Data held by the 12 regional QEs includes approximately 40M covered lives (Medicare FFS data represents 25% of the covered lives)
- HCCI is the first national QE, with 92.5M covered lives across all 50 states and DC (Medicare FFS data represents just over 50% of HCCI’s data)
Slide 18 Data Sharing for Care Coordination

- CMS is sending near real-time monthly data to facilitate care coordination to:
  - Accountable Care Organizations (ACOs) for patients assigned to the ACO
  - State Medicaid programs for Medicare-Medicaid enrollees
- Data is essential for analyzing the full continuum of care for beneficiaries
- Feeds include beneficiaries' entire claims history, including all service types, procedures and supplies.
- Opportunity for private sector to help transform the data to clinical information

Slide 19 Blue Button

- VA, DoD and CMS effort to give patients access to their own data (FEHB plans beginning to also offer blue button)
- 1M+ CMS beneficiaries have downloaded their data to date
- Beneficiaries can download up to 3 years of the hospital, physician and prescription drug claims
- Private sector applications can ingest, optimize and visualize data
Welcome to the CMS Data Navigator

Source: https://dnav.cms.gov/

About

The CMS Data Navigator application is an easy-to-use, menu-driven search tool that makes the data and information resources of the Centers for Medicare and Medicaid Services (CMS) more easily available. Use the Data Navigator to find data and information products for specific CMS programs, such as Medicare and Medicaid, or on specific health care topics or settings-of-care. Navigator displays search results by data type making it easier to locate specific types of...
information (e.g., data files, publications, statistical reports, etc.). The Data Navigator development team welcomes your feedback. Write to us at DataNavigator@cms.hhs.gov.

FAQs

What is the CMS Data Navigator?

The CMS Data Navigator is an easy to use, menu driven search tool designed to guide users to CMS data and information that is already available on the web. The Data Navigator does not house or maintain data but connects users by organizing data into categories, such as program, topic, or setting of care.

What are Publically Available Data Files – for download?

Non-identifiable data that are within the public domain and are available for instant download by any user.

What are Publically Available Data Files – for purchase?

CMS Data files that have been edited and stripped of all information that could be used to identify individuals. In general the data files contain aggregate level information on Medicare beneficiary or provider utilization. These data files are available for purchase using https://www.pay.gov/.

What are Restricted Use Data Files?

Identifiable data sets to be used only for reasons compatible with the purpose for which the data are collected. Data are subject to the Privacy Act, privacy release approvals, and the availability of computing resources and require a CMS data use agreement.

How does the CMS Data Navigator search logic work?

Multiple selections within a category are treated with OR logic, thereby expanding a search. Multiple selections in separate categories are treated with AND logic, thereby refining a search. For example, if a user selects both "Medicaid" and "Medicare" keywords under the "Program" category and both "Enrollment" and "Expenditures" keywords under the "Topic" category, the search results will contain data sources tagged with the keywords Medicaid or Medicare and Enrollment or Expenditures. If your search returns a small number of results, try removing some of your search selections to widen your search.

Start

Structure your search by expanding the appropriate content labels on the left and selecting the key words that best describe the data you need. The more key words you select, the narrower your search results.

For more information about a keyword, hover over the keyword with your mouse to see the pop-up tooltip which will display the glossary definition.

For help structuring your search, or to view the Data Glossary or Frequently Asked Questions, click on Help.
Federal policymakers and health researchers have long recognized that the amount and quality of the health care services that Medicare beneficiaries receive vary substantially across different regions of the United States. The Office of Information Products and Data Analytics (OIPDA) at CMS has made several resources available to researchers, policymakers, and other users who are interested in learning more about geographic variation in Medicare.

- **Public Use File**: The Geographic Variation Public Use File is a series of downloadable tables and reports that contain demographic, spending, utilization, and quality indicators for the Medicare fee-for-service population.

- **Dashboard**: The Geographic Variation Dashboards present Medicare fee-for-service per-capita spending at the state and county level in an interactive format. We calculated the spending figures in these dashboards using standardized dollars that remove the effects of the geographic adjustments that Medicare makes for many of its payment rates.

- **Site Visit Report**: The Site Visits Summary Report details findings from site visits to 12 HRRs with varying utilization and quality patterns. CMS also conducted site visits to two HRRs that focused on the Program of All-Inclusive Care for the Elderly (PACE); these findings are detailed in the PACE Site Visit Summary Report.

Geographic Data

The Centers for Medicare & Medicaid Services (CMS) has developed data that enables researchers and policymakers to examine geographic variation in the prevalence of chronic conditions and multiple chronic conditions as well as utilization and Medicare spending for beneficiaries with multiple chronic conditions. The data are aggregated to three geographic areas: (1) the 50 U.S. states and Washington, DC, (2) hospital referral regions (HRR), and (3) U.S. counties and are available for the years 2007-2011.

The data are available as excel files with “Reports” that allow users to compare a specific geographic area to national Medicare estimates. Report 1 presents the prevalence of 15 common chronic conditions among Medicare beneficiaries and Report 2 presents the prevalence, utilization and Medicare spending for Medicare beneficiaries with multiple chronic conditions. In addition, the excel files include a brief “Overview” section describing the data source, the sample population and the methodology for calculating these indicators.
Drawing Causal Inference from Big Data

Source: http://www.nasonline.org/programs/sa.../Big-data.html

This meeting was held March 26-27, 2015 at the National Academy of Sciences 2101 Constitution Ave. NW in Washington, D.C.
Organized by Richard M. Shiffrin (Indiana University), Susan Dumais (Microsoft Corporation), Mike Hawrylycz (Allen Institute), Jennifer Hill (New York University), Michael Jordan (University of California, Berkeley), Bernhard Schölkopf (Max Planck Institute) and Jasjeet Sekhon (University of California, Berkeley)
Co-sponsored by the National Science Foundation.

Overview

This colloquium was motivated by the exponentially growing amount of information collected about complex systems, colloquially referred to as "Big Data". It was aimed at methods to draw causal inference from these large data sets, most of which are not derived from carefully controlled experiments. Although correlations among observations are vast in number and often easy to obtain, causality is much harder to assess and establish, partly because causality is a vague and poorly specified construct for complex systems. Speakers discussed both the conceptual framework required to establish causal inference and designs and computational methods that can allow causality to be inferred. The program illustrates state-of-the-art methods with approaches derived from such fields as statistics, graph theory, machine learning, philosophy, and computer science, and the talks will cover such domains as social networks, medicine, health, economics, business, internet data and usage, search engines, and genetics. The presentations also addressed the possibility of testing causality in large data settings, and will raise certain basic questions: Will access to massive data be a key to understanding the fundamental questions of basic and applied science? Or does the vast increase in data confound analysis, produce computational bottlenecks, and decrease the ability to draw valid causal inferences?

Videos of the talks are available on the Sackler YouTube Channel. More videos will be added as they are approved by the speakers.

Speakers' Bio Sketches

Source: http://www.nasonline.org/programs/sa...-big-data.html
Edoardo Airoldi

Edoardo Airoldi is well known for his research that explores modeling, inferential, and other methodological issues that often arise in applied problems where network data (i.e., measurements on pairs of units, or tuples more generally) need to be considered, and standard statistical theory and methods are no longer adequate to support the goals of the analysis. More broadly, his research encompasses statistical methodology and theory with application to molecular biology and computational social science. His areas of technical interest include approximation theorems, inequalities, convex and combinatorial optimization, and geometry.

Dr. Airoldi is an Associate Professor of Statistics and Director of Graduate Studies at Harvard University. He received a Ph.D. from Carnegie Mellon University in 2007, working at the intersection of statistical machine learning and computational social science with Stephen Fienberg and Kathleen Carley. His PhD thesis explored modeling approaches and inference strategies for analyzing social and biological networks. He was a postdoctoral fellow in the Lewis-Sigler Institute for Integrative Genomics and the Department of Computer Science at Princeton University working with Olga Troyanskaya and David Botstein.

Optimal design of experiments in the presence of network interference

Edoardo M. Airoldi

Associate Professor of Statistics

Director of Graduate Studies

Harvard University

Causal inference research in statistics has been largely concerned with estimating the effect of treatment (e.g., personalized tutoring) on outcomes (e.g., test scores) under the assumption of "lack of interference"; that is, the assumption that the outcome of an individual does not depend on the treatment assigned to others. Moreover, whenever its relevance is acknowledged (e.g., study groups), interference is typically dealt with as an uninteresting source of variation in the data. In many applications, however, from quantifying the influence of peers on learning, to understanding the role of social relations on the success of health intervention in rural villages in Africa and South America, from word-of-mouth advertising and viral marketing, to political campaigns aimed at increasing volunteering, donations, and supporters turnout on election day, and from estimating the effects of accidents on congestion in transportation networks, to assessing how social structure affects labor mark dynamics, the lack of interference assumption is not tenable. Not only interference is present in these situations, and is an important aspect of the problem that cannot be abstracted away, but we are often interested in estimating the casual effect of such interference. In this talk, we present statistical methodology for working with interference. We will review challenges in defining the inferential target, review assumptions that facilitate estimation, and state alternative problem formulations for situations when the
available network data is not well aligned with the target notion of interference. We will then introduce a two-stage strategy to define an optimal set of randomizations for estimating interference on large social and information networks.

**Susan Athey**

Susan Athey is regarded for her research in the areas of industrial organization, microeconomic theory, and applied econometrics. Her current research focuses on the design of auction-based marketplaces and the economics of the internet, primarily on online advertising and the economics of the news media. She has also studied dynamic mechanisms and games with incomplete information, comparative statics under uncertainty, and econometric methods for analyzing auction models.

Dr. Athey is a Professor of Economics at Stanford University and an NAS Member. She is also a Senior Fellow at the Stanford Institute for Economics Policy Research. She received her bachelor’s degree from Duke University and her PhD from Stanford, and she holds an honorary doctorate from Duke University. She previously taught at the economics departments at MIT, Stanford and Harvard. At the age of 36, Professor Athey received the John Bates Clark Medal. The Clark Medal was awarded by the American Economic Association every other year to “that American economist under the age of forty who is adjudged to have made the most significant contribution to economic thought and knowledge.”

**Leon Bottou**

Leon Bottou is best known for his work in machine learning and data compression. His work presents stochastic gradient descent as a fundamental learning algorithm. He is also one of the main creators of the DjVu image compression technology (together with Yann LeCun and Patrick Haffner), and the maintainer of DjVuLibre, the open source implementation of DjVu. He is the original developer of the Lush programming language.

Dr. Bottou is currently a researcher with Facebook AI Research. He obtained the Diplôme d'Ingénieur from École Polytechnique, and a PhD from Université Paris-Sud. In 1995, he returned to Bell Laboratories, where he developed a number of new machine learning methods, such as Graph Transformer Networks and applied them to handwriting recognition and OCR. In 2010 he joined the Microsoft adCenter in Redmond, WA, and in 2012 became a Principal Researcher at Microsoft Research in New York City. He is associate editor of the Journal of Machine Learning Research, the IEEE Transactions on Pattern Analysis and Machine Intelligence, and Pattern Recognition Letters.

**Causal Reasoning and Learning Systems**

Léon Bottou, Facebook AI Research
Taking the real world example of ad placement on web search result pages, this talk (1) provides a real world example demonstrating the value of causal inference for large-scale interactive machine learning systems, (2) describes a collection of practical causal inference techniques applicable to a variety of interactive machine learning problems, (3) elucidates the relation between the exploration-exploitation dilemma and the counterfactual confidence intervals, and (4) proposes a combination of causal inference techniques and dynamical system analysis techniques to clarify the connection between auction theory and machine learning.

**Peter Buhlmann**

Peter Buhlmann is a Professor in the Department of Mathematics at ETH Zurich and the Chair of the Mathematics department. His research interests include statistics, machine learning, computational biology (ranging from methodology and mathematical theory to interdisciplinary research in biology and bio-medicine).

Dr. Buhlmann received his PhD from ETH Zurich. He is a Group Leader of the Competence Center for the Systems of Physiology and Metabolic Diseases and a Member of the German-Swiss Research Group FOR916: Statistical Regularization and Qualitative Constraints. He was Co-Editor of the Annals of Statistics from 2010-2012. In 2014 he was awarded the honor of Distinguished Lecturer at the Chinese Academy of Sciences and recognized as a Highly Cited Researcher in Mathematics by Thomson Reuters. He was also the 2013 winner of the Winton Research Prize in London.

**Susan Dumais**

Susan Dumais is a Distinguished Scientist and Deputy Managing Director at Microsoft Corporation. She is best known for her work in algorithms and interfaces for improved information retrieval, as well as general issues in human-computer interaction. Susan has been at Microsoft Research since July 1997. In 2014 she was honored with the ACM-W Athena Lecture Award and the Tony Kent Strix Award.

Dr. Dumais is a member of the National Academy of Engineering. Her current research focuses on gaze-enhanced interaction, the temporal dynamics of information systems, user modeling and personalization, novel interfaces for interactive retrieval, and search evaluation. Previous research studied a variety of information access and management challenges, including personal information management, desktop search, question answering, text categorization, collaborative filtering, interfaces for improving search and navigation, and user/task modeling. She has worked closely with several Microsoft groups (Bing, Windows Desktop Search, SharePoint Portal Server, and Office Online Help) on search-related innovations. Prior to Microsoft, she co-developed a statistical method for concept-based retrieval known as Latent Semantic Indexing.
Dean Eckles

Dean Eckles is a social scientist, statistician, and member of the Data Science team at Facebook. He studies how interactive technologies affect human behavior by mediating, amplifying, and directing social influence — and the statistical methods to study these processes. His current work uses large field experiments and observational studies. His research appears in peer reviewed proceedings and journals in computer science, marketing, and statistics. Dean holds degrees from Stanford University in philosophy (BA), cognitive science (BS, MS), and statistics (MS), and communication (PhD).

Dr. Eckles completed his PhD in Clifford Nass’s CHIMe Lab at Stanford University. He was previously a member of the research staff at Nokia Research Center, Palo Alto. Before joining Nokia, he worked with BJ Fogg on research in mobile persuasive technologies in the Stanford Persuasive Technology Lab and worked at Yahoo! Research Berkeley, designing and studying mobile photo sharing apps and services.

Identifying peer effects in social networks with peer encouragement designs

Authors: Dean Eckles (presenter), René Kizilcec, Eytan Bakshy

Abstract

Interactions among humans enable the spread of information, preferences, behavior, and disease. Despite large-scale measurement of human behaviors, credible identification and estimation of peer influence effects remains difficult. After reviewing other methods for identifying peer effects, we present research designs that enable point identification of peer effects for relevant populations. In these peer encouragement designs, vertices are randomly assigned to conditions that affect adoption of a target behavior; the experimenter then observes this behavior and that of their peers (i.e., how this "spills over" to their peers). We present an example of a large peer encouragement design that identifies the effects of receiving feedback from peers in social media. We relate these designs to prior experiments in groups and to the literature on instrumental variables estimation with heterogeneous treatment effects.

James Fowler

James Fowler earned his PhD from Harvard in 2003 and is currently a Professor at the University of California, San Diego. His work lies at the intersection of the natural and social sciences, with a focus on social networks, behavior, evolution, politics, genetics, and big data.

Dr. Fowler was recently named a Fellow of the John Simon Guggenheim Foundation, one of Foreign Policy’s Top 100 Global Thinkers, TechCrunch’s Top 20 Most Innovative People, Politico’s 50 Key Thinkers, Doers, and Dreamers, and
Most Original Thinker of the year by The McLaughlin Group. He has also appeared on The Colbert Report. His research has been featured in numerous best-of lists including New York Times Magazine's Year in Ideas, Time's Year in Medicine, Discover Magazine's Year in Science, and Harvard Business Review's Breakthrough Business Ideas. Together with Nicholas Christakis, James wrote a book on social networks for a general audience called Connected.

A Follow-up to a 61 Million Person Experiment in Social Influence and Political Mobilization

James H. Fowler
Professor
University of California, San Diego

http://fowler.ucsd.edu

A previous large scale experiment showed that a single message posted on social media could directly influence real world voting behavior, and that the indirect effect of the message on friends accounted for most of its total effect on increased voter turnout. Here, we analyze a follow-up experiment conducted in the 2012 US Presidential Election. The results show that messaging had both direct and indirect effects on voting behavior in that election as well, suggesting that social media can be an effective tool for mobilizing political participation in high stakes elections.

Michael Hawrylycz

Mike Hawrylycz joined the Allen Institute in 2003. He is responsible for the direction of the data analysis and annotation effort. Hawrylycz has worked in a variety of applied mathematics and computer science areas, addressing challenges in consumer and investment finance, electrical engineering and image processing, and computational biology and genomics.

Dr. Hawrylycz received his Ph.D. in applied mathematics at the Massachusetts Institute of Technology. He subsequently was a post-doctoral researcher in the Computer Research and Applications Group at the Los Alamos National Laboratory. He received his Masters in Mathematics at Wesleyan University. He is a member of the Society for Industrial and Applied Mathematics as well as the American Statistical Association and the Society for Neuroscience. He has served as a Review Editor for Frontiers in Neurogenomics and a Reviewer for the Journal of Neuroscience, Nature Biotechnology and Physiological Genomics.

Project MindScope: From Big Data to Behavior in the Functioning Cortex

Mike Hawrylycz, Allen Institute for Brain Science

As the most complex piece of matter in the known universe, the brain gives rise to behavior, mind, and consciousness. With roughly 86 billion neurons, each coupled to as many as 10,000 others, unraveling the brain's function and what
causes us to make decisions is a daunting task. The Allen Institute for Brain Science is pursuing an ambitious project, called Project MindsScope, to further our understanding of the brain using a tripartite approach based on Components, Computation, and Cognition. From the Components viewpoint, we aim to identify the characteristics of single neurons, from physiology to morphology and genetic profile, as well as the connectivity that defines regions and local circuits. In Computation the goal is to understand the properties of neural response, i.e. the “neural code”, in sensory areas such as during behavior. Finally, in Cognition the aim is to put all of the pieces together in order to understand aspects of causal behavior, specifically the task of object recognition and the role of attention.

David Heckerman

David Heckerman is currently a Senior Director with Microsoft Corporation. In his early work, he demonstrated the importance of probability theory in Artificial Intelligence, and developed methods to learn graphical models from data, including methods for causal discovery. More recently, he is developing machine learning and statistical approaches for biological and medical applications, including HIV vaccine design and genomics (see http://github.com/microsoftgenomics). At Microsoft, David have developed numerous applications including data-mining tools in SQL Server and Commerce Server, the junk-mail filters in Outlook, Exchange, and Hotmail, handwriting recognition in the Tablet PC, text mining software in Sharepoint Portal Server, troubleshooters in Windows, and the Answer Wizard in Office.

Dr. Heckerman began his education with the intent of becoming a physicist, but his interests eventually led him into the medical sciences. While working on his MD at Stanford, he began looking at the problems of Artificial Intelligence. For his PhD work, he submitted an impressive construct he called the “probabilistic expert system” which led to Microsoft hiring him to build such systems for non-medical applications.

Jennifer Hill

Jennifer Hill is an Associate Professor of Social Sciences at New York University. She works on development of methods that help us to answer the causal question that are so vital to policy research and scientific development. In particular she focuses on situations in which it is difficult or impossible to perform traditional randomized experiments, or when even seemingly pristine study designs are complicated by missing data or hierarchically structured data.

Most recently Dr. Hill has been pursuing two major strands of research. The first focuses on Bayesian nonparametric methods that allow for flexible estimation of causal models without the need for methods such as propensity score matching. The second line of work pursues strategies for exploring the impact of violations of typical assumptions in this work that require that all confounders have been measured. Hill earned her PhD in Statistics at Harvard University in 2000 and completed a post-doctoral fellowship in Child and Family Policy at Columbia University’s School of Social Work in 2002.
Michael Jordan

Michael I. Jordan is the Pehong Chen Distinguished Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California, Berkeley. He received his Masters in Mathematics from Arizona State University, and earned his PhD in Cognitive Science in 1985 from the University of California, San Diego. He was a professor at MIT from 1988 to 1998.

Dr. Jordan’s research interests bridge the computational, statistical, cognitive and biological sciences, and have focused in recent years on Bayesian nonparametric analysis, probabilistic graphical models, spectral methods, kernel machines and applications to problems in distributed computing systems, natural language processing, signal processing and statistical genetics. Prof. Jordan is a member of the National Academy of Sciences, a member of the National Academy of Engineering and a member of the American Academy of Arts and Sciences. He is a Fellow of the American Association for the Advancement of Science. In 2015, he received the David E. Rumelhart Prize.

On Computational Thinking, Inferential Thinking and "Big Data"

Michael I. Jordan

University of California, Berkeley

The rapid growth in the size and scope of datasets in science and technology has created a need for novel foundational perspectives on data analysis that blend the inferential and computational sciences. That classical perspectives from these fields are not adequate to address emerging problems in "Big Data" is apparent from their sharply divergent nature at an elementary level— in computer science, the growth of the number of data points is a source of "complexity" that must be tamed via algorithms or hardware, whereas in statistics, the growth of the number of data points is a source of "simplicity" in that inferences are generally stronger and asymptotic results can be invoked. I aim to set the stage for the discussion of causality and Big Data by surveying recent progress at the computation/statistics interface, including fundamental tradeoffs between inferential quality, communication, runtime and privacy constraints, and mechanisms for implementing these tradeoffs, such as algorithmic weakening, subsampling and concurrency control.

Steven Levitt

Steve Levitt is the William B. Ogden Distinguished Service Professor of Economics at the University of Chicago, where he directs the Becker Center on Chicago Price Theory. Levitt received his BA from Harvard University in 1989 and his PhD from MIT in 1994. He has taught at Chicago since 1997.
In 2004, Dr. Levitt was awarded the John Bates Clark Medal, awarded to the most influential economist under the age of 40. In 2006, he was named one of Time magazine's "100 People Who Shape Our World." Steve co-authored *Freakonomics*, which spent over 2 years on the New York Times Best Seller list and has sold more than 4 million copies worldwide. *SuperFreakonomics*, released in 2009, includes brand new research on topics from terrorism to prostitution to global warming. Steve is also the co-author of the popular *Freakonomics Blog*.

**Thinking Differently about Big Data**

Steven Levitt, University of Chicago

Few phrases are touted more often in the business media than "Big Data." Big data is seemingly the answer to every company's future. But is the promise real? And when will the benefits come, if ever? University of Chicago economist Steven Levitt, winner of the prestigious John Bates Clark Medal and coauthor of Freakonomics, offers his unique and unexpected take on this issue. His conclusions may surprise you.

**David Madigan**

David Madigan received a bachelor's degree in Mathematical Sciences and a Ph.D. in Statistics, both from Trinity College Dublin. He has previously worked for AT&T Inc., Soliloquy Inc., the University of Washington, Rutgers University, and SkillSoft, Inc. He has over 100 publications in such areas as Bayesian statistics, text mining, Monte Carlo methods, pharmacovigilance and probabilistic graphical models. He is an elected Fellow of the American Statistical Association and of the Institute of Mathematical Statistics.

Dr. Madigan recently completed a term as Editor-in-Chief of *Statistical Science*. For the past three years, David has worked as a principal investigator on the OMOP research program, making significant contributions to the project's methodological work including the development, implementation, and analysis of a variety of statistical methods applied to various observational databases. His expertise is in the application of statistical methods to large-scale data problems. David is interested in large-scale predictive modeling and statistical analysis of healthcare data.

**Honest Inference From Observational Database Studies**

David Madigan, Columbia University

Observational healthcare data, such as administrative claims and electronic health records, play an increasingly prominent role in healthcare. Pharmacoepidemiologic studies in particular routinely estimate temporal associations between medical product exposure and subsequent health outcomes of interest and such studies influence prescribing patterns and healthcare policy more generally. Some authors have questioned the reliability and accuracy of such studies, but few previous efforts have attempted to measure their performance. We have conducted a series of experiments to empirically measure the performance of various observational study designs with regard to predictive accuracy for discriminating between true drug effects and negative controls. I describe this work, explore opportunities to...
expand the use of observational data to further our understanding of medical products, and highlight areas for future research and development.

**Judea Pearl**

Judea Pearl is a graduate of the Technion-Israel Institute of Technology. He came to the United States for postgraduate work in 1960, and the following year he received a master’s degree in electrical engineering from Newark College of Engineering, now New Jersey Institute of Technology. In 1965, he simultaneously received a master’s degree in physics from Rutgers University and a PhD from the Brooklyn Polytechnic Institute, now Polytechnic Institute of New York University.

Dr. Pearl joined the faculty of UCLA in 1969, where he is currently a professor of computer science and statistics and director of the Cognitive Systems Laboratory. He is known internationally for his contributions to artificial intelligence, human reasoning, and philosophy of science. He is the author of more than 350 scientific papers and three landmark books in his fields of interest. A member of the National Academy of Sciences, the National Academy of Engineering and a founding Fellow of the American Association for Artificial Intelligence, Pearl is the recipient of numerous scientific prizes, including the Turing Award and IEEE Intelligent Systems’ AI's Hall of Fame.

**Taming the challenge of extrapolation: From multiple experiments and observations to valid causal conclusions**

Judea Pearl and Elias Bareinboim, UCLA

One distinct feature of big data applications, setting them apart from traditional causal analyses, is that data are pooled together from multiple sources. Each represents a different population, studied under different conditions, some experimental and some observational, and each contaminated by different confounding factors and different sample selection mechanisms. All of these studies need be pooled together to answer causal questions in yet a new environment, unmatched by any of those studied.

The problem, as described above, appears utterly hopeless, almost as hopeless as confronting physics problems in the 16th century, before the advent of algebra.

My talk will demonstrate that the solutions to the two types of problems, data pooling on the one hand and algebraic equations on the other, are strikingly similar; they can be unraveled by symbolic operations and managed at ease by anyone who masters the two basic tools of causal inference: graphical models and the logic of causation.

The net result is the emergence of algorithms that decide what piece of information must be extracted from each of the available data sources and how to combine those pieces together so as to provide valid answers to a variety of causal questions: What if? How? and Why?
To gain an appreciation for the task, the audience may wish to consider three toy problems described in Figure 3 of this introductory paper http://ftp.cs.ucla.edu/pub/stat_ser/r400.pdf. All three problems involve the same task: We gather causal information in one environment and we wish to generalize it to another, which differs from the first in a set Z of characteristics. How do we adjust the available information so as to account for differences between the two populations?

It turns out, as readers will immediately discover, that there is no universal adjustment formula for this task; the adjustment should vary from case to case depending on the location of the set Z in the causal scheme of things. Traditional schemes like post-stratification or re-weighting would not work, except in special cases. Fortunately, however, once the set Z is allocated in the causal graph, the proper adjustment can be generated automatically using symbolic operations, in much the same way that we solve algebraic equations.

The talk will describe how we go from two to multiple environments, how we characterize the idiosyncratic features of each population or environment, and how we can generalize from unrepresentative samples to population level effects.

Related papers can be viewed here:

Tutorials

Extrapolation problem

**Thomas Richardson**

Thomas Richardson is Professor and Chair of the Department of Statistics. He is also an Adjunct Professor in the Departments of Economics and Electrical Engineering and a member of the eScience Steering Committee. He received his BA from the University of Oxford and his MS and PhD from Carnegie Mellon University. He is a Fellow of the Center for Advanced Studies in the Behavioral Sciences at Stanford University. His research interests include Graphical Models and Causality.

Dr. Richardson is known for his research in Graphical models, algorithmic model selection, Bayesian inference, causal models, applications in economics. He was a Visiting Senior Research Fellow at Jesus College in Oxford and was awarded a fellowship at the Institute for Advanced Studies at the University of Bologna.

http://semanticommunity.info/Data_Science/Data_Science_for_Agency_Initiatives_2015
Updated: Wed, 23 Sep 2015 08:12:44 GMT
Powered by mindtouch
Non-parametric Causal Inference

Thomas Richardson, University of Washington

There are three main frameworks for causal inference: counterfactuals (aka potential outcomes), nonparametric structural equation models (NPSEMs) and graphs (aka path diagrams). These approaches are similar and, in certain specific respects, equivalent. However, there are important conceptual differences and each formulation has its own strengths and weaknesses. These divergences are of relevance both in theory and when the approaches are applied in practice. This talk will introduce the different frameworks, and describe, through examples, both the commonalities and dissimilarities.

James M Robins

James M. Robins is an epidemiologist and biostatistician best known for advancing methods for drawing causal inferences from complex observational studies and randomized trials, particularly those in which the treatment varies with time. He is the 2013 recipient of the Nathan Mantel Award for lifetime achievement in statistics and epidemiology.

Dr. Robins graduated in medicine from Washington University in 1976. He is currently Mitchell L. and Robin LaFoley Dong Professor of Epidemiology at Harvard School of Public Health. He has published over 100 papers in academic journals and is an ISI highly cited researcher. In his original paper on causal inference, Robins described two new methods for confounding bias, which can be applied in the generalized setting of time-dependent exposures: The G-formula and G-Estimation of Structural Nested Models. He introduced a third class of models, Marginal Structural Models, in which the parameters are estimated using inverse probability of treatment weights. He has also contributed significantly to the theory of dynamic treatment regimes, which are of high significance in comparative effectiveness research and personalized medicine.

Personalized Medicine, Optimal Treatment Strategies, and First Do No Harm: Time Varying Treatments and Big Data

James M. Robins, Harvard School of Public Health

Enormous computerized data bases are available that contain medical information on millions of Americans. The hope is that these data bases can be mined to discover new causes of disease, to determine the efficacy and safety of numerous existing medical treatments and procedures, and, to discover individualized treatment strategies that dominate the current standard of care. In this talk I will consider the degree to which this hope is realistic., I will first discuss why big data may not be good data,, and why, even with good data, many standard analytic methods that are not explicitly causal methods may fail to produce reliable conclusions. I will then consider how one might validate the success of the enterprise of mining big data to improve health by comparing the discoveries made by mining of observational data bases with the results of randomized clinical trials.
Bernhard Schölkopf

Bernhard Schölkopf studied in Tübingen and London physics, mathematics and philosophy. He received his doctorate from the Technical University of Berlin in computer science. He was Director and Scientific Member at the Max Planck Institute for Biological Cybernetics in Tübingen. Since May 2011 he has been Director of the Max Planck Institute for Intelligent Systems in Tübingen and Stuttgart. He taught at the Humboldt University of Berlin and the Technical University of Berlin and in Tübingen, since 2002 he is honorary professor at the Technical University of Berlin. He has received numerous outstanding prizes, most recently the Royal Society Milner Award in 2014 and the 2011 Max Planck Research Award.

Dr. Scholkopf is one of the leading international experts in this field. With his research team he developed new learning method that can detect patterns in observed data. More recently, he has dealt with the problem of causal data analysis and found an interesting link between causality and description complexity.

Toward Causal Machine Learning

Bernhard Schölkopf

Director of the Max Planck Institute for Intelligent Systems

Tübingen Germany

In machine learning, we use data to automatically find dependences in the world, with the goal of predicting future observations. Most machine learning methods build on statistics, but one can also try to go beyond this, assaying causal structures underlying statistical dependences. Can such causal knowledge help prediction in machine learning tasks? We argue that this is indeed the case, due to the fact that causal models are more robust to changes that occur in real world datasets. We discuss the implications of causal models for machine learning tasks such as domain adaptation, transfer learning, and semi-supervised learning. We present an application to the removal of systematic errors for the purpose of exoplanet detection.

Machine learning and "big data" analysis currently mainly focuses on well-studied statistical methods. Some of the causal problems are conceptually harder, however, the causal point of view can provide additional insights that have substantial potential for data analysis.

Jasjeet Sekhon
Jasjeet S. Sekhon is a Robson Professor of Political Science and Statistics at University of California, Berkeley. His current research focuses on methods for causal inference in observational and experimental studies and evaluating social science, public health and medical interventions. Professor Sekhon has done research on elections, voting behavior and public opinion in the United States, multivariate matching methods for causal inference, machine learning algorithms for irregular optimization problems, robust estimators with bounded influence functions, health economic cost effectiveness analysis, and the philosophy and history of inference and statistics in the social sciences.

Dr. Sekhon studied at the University of British Columbia. He earned his MA and PhD at Cornell University. In 2012, he won the Society for Political Methodology Software Award and The Warren Miller Prize. He is a BIDS Co-PI for Moore/Sloan Data Science Environment.

**Combining Experiments with Big Data to Estimate Treatment Effects**

Jasjeet Sekhon, University of California, Berkeley

The rise of massive datasets that provide fine-grained information about human beings and their behavior provides unprecedented opportunities for evaluating the effectiveness of treatments across the human sciences. New methodological challenges have to be overcome to make the most of these opportunities. Among them are algorithms that do not scale, questions about how to combine experimental studies with massive observational data, and how to estimate heterogeneous treatment effects for subgroups. A problem that connects these challenges is the explosion of false positives that has come along with massive data. Progress on these challenges will require combining insights from diverse fields, particularly statistics and computer science. The core of this talk explores one issue of experimental design.

Our experiments are growing in size. Massive experiments often revolutionize their respective fields as they tend to be more representative of the population of interest and important effects, that normally are undetectable, are possible to estimate. Parallel to this development, several strides have recently been made in study of how to best design small-scale experiments. The recent literature reminds us that randomization alone does not solve all problems. In particular, most of the desirable properties that randomization brings are only guaranteed on average. In any specific sample and treatment assignment, there may be remarkable imbalances on prognostically important covariates and inferences can thereby be severely flawed. Many experiments will, when viewed unconditionally, have estimators with unnecessarily high variance; and when viewed conditionally on the observed imbalances, the estimators will be biased.

In simple small-scale experiments the standard method to avoid these problems is to block the sample before randomization. In its most stylized description, blocking is when the researcher divides the experimental sample into groups, or blocks, based on covariates and assigns treatment in fixed proportions within the groups but independently between them. If the blocks are formed so to make the units they contain as similar as possible this procedure will ensure that the treatment groups are balanced and thereby greatly improve the quality of any inferences.

Another, possibly more important, reason to block has become evident with the rise of large experiments: blocking can greatly improve inferences to other populations—be it subgroups in the experimental sample or reweighted subgroups so as to estimate treatment effects for other populations. As treatment assignment is independent between blocks, each group of units can be seen as an experiment in its own right. If the researcher is interested in the treatment effect in some subpopulation, she can simply extract the corresponding blocks and treat them as a separate experiments. Unlike
most other methods for analyzing fine-grained effects, this technique guarantees unbiased estimates of the treatment effect in the subpopulations. As the blocks are predefined, it is much harder to comb through the data to find seemingly significant results. Blocking thereby improves transparency and controls the rate of false positives by design.

However, although the basic idea of blocking goes back to the infancy statistics (Fisher 1926), current blocking methods are restricted to special cases, run in exponential time, or are heuristic, providing an unsatisfactory solution in many common cases. The lack of progress in the area is explained by the fact that blocking problems are isomorphic to partitioning problems in graph theory. Net of a few special cases, these problems are known to be NP-hard. In fact, the first algorithm with proven optimality for the blocking problem with covariates of high dimensionality was introduced as late as 2004 (Greevy 2004). That algorithm exploited one of the few special cases where the underlying partitioning problem is not NP-hard. While this made the algorithm possible, it restricts the blocks to contain exactly two units and therefore severely limits its applicability.

We introduce a new blocking algorithm that can be used for multiple treatments and blocks of different sizes. We describe a constant-factor approximation algorithm with polynomial time complexity. We investigate the optimization problem where, given a minimum required group size and a distance metric, one wants to find a set of groups—a blocking—so that the maximum distance between any two units within a group is minimized. Finding this blocking is an NP-hard problem. Our algorithm produces a blocking where the maximum distance is guaranteed to be at most four times the optimal value and does so in polynomial time. Unlike previous algorithms, our algorithm works for an arbitrary group size facilitating complex experiments with several treatment arms. Simulation studies indicate that the algorithm produces solutions with a maximum distance well below the theoretically guaranteed bound, often close to the optimal solution. The algorithm can successfully be used in huge experiments; millions of observations can be blocked using a desktop computer.

**Cosma Shalizi**

Cosma Shalizi is an Associate Professor of Statistics at Carnegie Mellon University. He is best known for his research in statistical inference for complex systems; nonparametric prediction for stochastic processes; causal inference; large deviations and ergodic theory; networks and information flow in neuroscience, economics and social sciences; heavy-tailed distributions; self-organization.

Dr. Shalizi earned his PhD in theoretical physics from the University of Wisconsin-Madison. As a post-doc, he moved from the mathematics of optimal prediction to devising algorithms to estimate such predictors from finite data, and applying those algorithms to concrete problems. On the algorithmic side, he devised an algorithm, CSSR, which exploits the formal properties of the optimal predictive states to efficiently reconstruct them from discrete sequence data. He also developed a reconstruction algorithm for spatio-temporal random fields. His most recent work falls into the areas of heavy tails, learning theory for time series, Bayesian consistency, neuroscience, network analysis and causal inference, with some overlap between these.
Richard Shiffrin

Richard M Shiffrin heads the Memory and Perception Laboratory in the Department of Psychological and Brain Sciences at Indiana University—the MAPLAB website gives information about present and past lab members and projects. He is a Distinguished Professor and Luther Dana Waterman Professor and has additional appointments in Cognitive Science (which he founded in 1988) and Statistics.

Dr. Shiffrin is a member of the National Academies of Sciences. His research interests are quite broad, more or less covering the fields of Cognitive Science and Psychology. Generally speaking the research involves empirical studies and quantitative and computational modeling of the results. Current projects are generally tailored toward the interests of the graduate students and postdoctoral researchers in the lab, and the need to carry out research funded by external grants (presently from NSF and AFOSR).

Introduction to the Sackler Colloquium, Drawing Causal Inference from Big Data

Richard Shiffrin, Indiana University

Data is the basis for scientific progress, and causality is the primary way humans come to understand what the data imply.

A sea change in the way this process operates has taken place in recent years: We have developed the ability to produce, measure, collect, and store amounts of data far beyond anything imagined previously.

Table

<table>
<thead>
<tr>
<th>Unit</th>
<th>Symbol</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Byte</td>
<td>8 bits</td>
<td>10**3</td>
<td>1,000</td>
</tr>
<tr>
<td>Kilobyte</td>
<td>10**3</td>
<td>1,000</td>
<td>(short story)</td>
</tr>
<tr>
<td>Megabyte</td>
<td>10**6</td>
<td>1,000,000</td>
<td>(a novel)</td>
</tr>
<tr>
<td>Gigabyte</td>
<td>10**9</td>
<td>1,000,000,000</td>
<td>(movie)</td>
</tr>
<tr>
<td>Terabyte</td>
<td>10**12</td>
<td>1,000,000,000,000</td>
<td>(x-rays in one hospital)</td>
</tr>
<tr>
<td>Petabyte</td>
<td>10**15</td>
<td>1,000,000,000,000,000</td>
<td>(US research libraries)</td>
</tr>
<tr>
<td>Exabyte</td>
<td>10**18</td>
<td>1,000,000,000,000,000,000</td>
<td>(SKA telescope/day)</td>
</tr>
<tr>
<td>Zettabyte</td>
<td>10**21</td>
<td>1,000,000,000,000,000,000,000</td>
<td>(all words ever spoken)</td>
</tr>
</tbody>
</table>

http://semanticommunity.info/Data_Science/Data_Science_for_Agency_Initiatives_2015
Updated: Wed, 23 Sep 2015 08:12:44 GMT
Powered by mindtouch
We have foreseen the need to have terms for $10^{27}$, $10^{30}$, and $10^{33}$ (so far). Information is doubling every year or two (and the rate might be accelerating).

The problem is not just collecting data from more sources, but also because we collect more precise measurements from a single source--one of my colleagues recently joined a project in which they expect to deal with 1000 terabytes of information from one-thousandth part of a single mouse brain.

Although we are producing and storing ever greater amounts of data, we have just begun to figure out ways to analyze and understand what the data show. The problem is not restricted to science: Business, government, entertainment, social media, security agencies, social networks face the same challenges.

The two main challenges, both unprecedented in scope, are two sides of the same coin:

First, how does one find the important patterns of data?

This subject requires a Sackler Colloquium of its own. Suppose a moderately large data base has a terabyte of data ($10^{12}$). This data might perhaps contain a thousand ($10^{3}$) measurable factors. The number of correlations of those factors in all combinations would be on the order of $2^{10^3}$, or about a 300 digit number. The search problem for patterns is enormous.

Second, having found a pattern, how can we explain its causes?

This is the focus of the present Sackler Colloquium. If in a terabyte data base we notice factor A is correlated with factor B, there might be a direct causal connection between the two, but there might be something like $2^{300}$ other potential causal loops to be considered. Things could be even more daunting: To infer probabilities of causes could require consideration all distributions of probabilities assigned to the $2^{300}$ possibilities. Such numbers are both fanciful and absurd, but are sufficient to show that inferring causality in Big Data requires new techniques. These are under development, and we will hear some of the promising approaches in the next two days.

Whatever is developed, I am sure computational algorithms will never be sufficient to answer either question. The numbers go well beyond any conceivable rote computational approach. To me this highlights the importance of models and theories. Models and theories have always been important in science. but in Big Data they will be critically needed to guide the search for patterns and guide the search for causal accounts.

This mention of models leads me to finish with a somewhat high level perspective: In many ways, drawing causal inference from Big Data is **Science Writ Small**: In science we find patterns of data in the real world and in experiments, use models and theories to try to explain and understand them, and use the models to guide further search for patterns and to design new experiments, and then develop new and better tuned theories. These models are in many or most cases designed to provide causal accounts, but it is recognized that the models are approximations, that there are an infinite number of alternative accounts, and that for this and other reasons we prefer models that balance a good fit to the data against model complexity. As the era of Big Data continues, I would expect one important approach to causal
explanations to follow the line that science has already developed, a line that often involves experimental tests and ‘interventions’.

However, the Big Data era also introduces the need to establish causality when experimental tests and interventions are difficult, impossible, or too far removed from the real world complexities that govern the data. If we collect a petabyte of data about the past 24 hours of world weather, and wish to understand the causes of some important weather pattern, intervention would not be feasible, and laboratory tests might not be highly relevant. In cases like this we would have to try to establish causality from data internal to the data base itself. At the least this would require both assumptions concerning the meaning of causality in large complex recurrent systems (a subject still under debate), and new statistical and computational techniques. I believe all of us are excited to hear promising approaches along these lines in the next two days.

John Stamatoyannopoulos

John Stamatoyannopoulos, M.D., is an Assistant Professor of Genome Sciences and Medicine at the University of Washington School of Medicine. He graduated from Stanford University in 1990 with degrees in Biology, Symbolic Systems, and Classics, and received an M.D. in 1995 from the University of Washington. He completed residency in Internal Medicine at Brigham and Women's Hospital, Harvard Medical School, and was a fellow in Oncology and Hematology at Dana Farber Cancer Institute and the Massachusetts General Hospital. He was awarded a Howard Hughes Medical Institute Physician-postdoctoral fellowship at Dana Farber. Dr. Stamatoyannopoulos then served as Chief Scientific Officer of biotechnology company Regulome Corp., and subsequently joined the Departments of Genome Sciences and Medicine (Oncology) at the University of Washington in 2005. He was elected to American Society for Clinical Investigation in 2009 and is a member of the Editorial Board of Genome Research.

Dr. Stamatoyannopoulos' lab focuses on understanding the large-scale cis-regulatory circuitry of the human genome, and the functional consequences of non-coding genetic variation. He is PI of the UW ENCODE Project, and Director of the Northwest Reference Epigenome Mapping Center.

Hal Varian

Hal R. Varian is the Chief Economist at Google. He is also an emeritus professor at the University of California, Berkeley in three departments: business, economics, and information management. He received his SB degree from MIT in 1969 and his MA in mathematics and Ph.D. in economics from UC Berkeley in 1973. He has also taught at MIT, Stanford, Oxford, Michigan and other universities around the world.

Dr. Varian is a fellow of the Guggenheim Foundation, the Econometric Society, and the American Academy of Arts and Sciences. Professor Varian has published numerous papers in economic theory, industrial organization, financial
Bin Yu is a Professor of Statistics, and Electrical Engineering & Computer Science, at University of California Berkeley. She is Chancellor's Professor in Statistics from 2006 to 2011. Her current research interests include statistical inference, machine learning, information theory (Minimum Description Length Principle), as well as data modeling in areas such as remote sensing, internet tomography, sensor networks, neuroscience, bioinformatics, and finance.

Dr. Yu received her B.S. degree in Mathematics from Peking University in 1984, M.S. and Ph.D. degrees in Statistics from the University of California at Berkeley in 1987 and 1990 respectively. Her doctoral research was on empirical processes for dependent data and Minimum Description Length (MDL) Principle. She is an elected Fellow of IEEE, The Institute of Mathematical Statistics (IMS) and ASA (American Statistical Association). She was a Miller Research Professor, Miller Institute, UC Berkeley. She is an Associate Editor for The Annals of Statistics, Journal of American Statistical Association (JASA), and Statistica Sinica, an Action Editor for the Journal of Machine Learning Research (JMLR). She served on the Board of IEEE Information Theory Society (two terms) and on the Council of IMS (one term).

Lasso adjustments of treatment effect estimates in randomized experiments

Bin Yu
Departments of Statistics and EECS, UC Berkeley
http://statistics.berkeley.edu/~binyu
binyu@berkeley.edu

In randomized experiments, linear regression is often used to adjust for imbalances in covariates between treatment groups, yielding an estimate of the average treatment effect with lower asymptotic variance than the unadjusted estimator. If there is large number of covariates, many of which are irrelevant to the potential outcomes, the Lasso can be used to both select relevant covariates and perform the adjustment. We study the resulting estimator under the Neyman model for randomization, and show that it is more efficient than the unadjusted estimator, and that it is possible to give a conservative estimate of the asymptotic variance. Simulations show that Lasso-based adjustment can be advantageous even when $p < n$. Moreover, when the covariates selected by the Lasso indicate the presence of heterogenous treatment effects, our method can yield conditional treatment effect estimates for subpopulations.

(This talk is based on joint work with Adam Bloniarz, Hanzhong Liu, Jas Sekhon and Cunhui Zhang.)
Thursday, March 26

Introduction, Richard Shiffrin, Indiana University, The Big Data Sea Change This Wiki

Michael Jordan, University of California, Berkeley, On Computational Thinking, Inferential Thinking and Big Data This Wiki

Judea Pearl, University of California, Los Angeles, Taming the Challenge of Extrapolation: From Multiple Experiments and Observations to Valid Causal Conclusions This Wiki YouTube Video Slides: 17

Thomas Richardson, University of Washington, Non-parametric Causal Inference This Wiki YouTube Video

David Heckerman, Microsoft Corporation, Causal Inference in the Presence of Hidden Confounders in Genomics This Wiki Slides: 35

Jasjeet Sekhon, University of California, Berkeley, Combining Experiments with Big Data to Estimate Treatment Effects This Wiki

Bin Yu, University of California, Berkeley, Lasso Adjustments of Treatment Effect Estimates in Randomized Experiments This Wiki

Bernhard Schölkopf, Max Planck Institute, Toward Causal Machine Learning This Wiki YouTube Video

John Stamatyannopoulos, University of Washington, Decoding the Human Genome: From Sequence to Knowledge This Wiki

Edo Airoldi, Harvard University, Optimal Design of Causal Experiments in the Presence of Social Interference This Wiki

Peter Buhlmann, ETH Zurich, Causal Inference Based on Invariance: Exploiting the Power of Heterogeneous Data This Wiki Slides: 34

Annual Sackler Lecture

Introduction by Ralph J. Cicerone, President, National Academy of Sciences

Sackler Lecture presented by Steven Levitt, The University of Chicago, Thinking Differently About Big Data This Wiki

Friday, March 27

Michael Hawrylycz, Allen Institute, Project MindScope: From Big Data to Behavior in the Functioning Cortex This Wiki

David Madigan, Columbia University, Honest Inference From Observational Database Studies This Wiki
Susan Athey, Stanford University, *Estimating Heterogeneous Treatment Effects Using Machine Learning in Observational Studies* This Wiki Slides: 31

Leon Bottou, Facebook AI Research, *Causal Reasoning and Learning Systems* This Wiki Slides: 92

Dean Eckles, Facebook, *Identifying Peer Effects in Social Networks* This Wiki

Hal Varian, Google, Inc., *Causal Inference, Econometrics, and Big Data* This Wiki Slides: 20

James Robins, Harvard University, *Personalized Medicine, Optimal Treatment Strategies, and First Do No Harm: Time Varying Treatments and Big Data* This Wiki

James Fowler, University of California, San Diego, *An 85 Million Person Follow-up to a 61 Million Person Experiment in Social Influence and Political Mobilization* This Wiki

General Discussion, Jennifer Hill, New York University

NEXT