Data Science Central

9 "must read" articles

My book is finally available!!

From Data Science Central to Data Science Results

Registered meteorites that has impacted on Earth visualized

Comment by Ramon Martinez on February 21, 2013 at 1:45pm
Comment by Vincent Granville on February 20, 2013 at 7:06pm
Comment by Vincent Granville on February 20, 2013 at 7:01pm

Five Announcements Email

1. New Hadoop Community
2. Data Science Book
3. Most recent data science positions
4. Three reasons to attend the big Data Science Summit Next Week, in the Bay Area
5. New Data Science Central Acquisition

Developing Analytic Talent

Chapter 1: What is Data Science?

Fake Data Science

Fake data science: Two Examples
The Face of the New University

Thirteen Problems

DUI Arrests Decreases After State Monopoly on Liquor Sales Ends
Data Science Defeats Intuition
Data Glitch Turns Data into Gibberish
Regression in Unusual Spaces
Analytics versus Seduction to Boost Sales
About Hidden Data
High Crime Rates Caused by Gasoline Lead, Really?
Boeing’s Dreamliner Problems
Seven Tricky Sentences for NLP
Data Scientists Dictate What We Eat
Increasing Amazon.com Sales with Better Relevancy
Detecting Fake Profiles or Likes on Facebook
Analytics for Restaurants

History and Milestones
Statistics Will Experience a Renaissance
A Few Random Thoughts

Modern Trends
Data Scientist Versus Data Architect

Summary
Chapter 2: Big Data is Different

Two Big Data Issues
The Curse of Big Data
When Data Flows Faster Than it Can Be Processed
Examples of Big Data Techniques
Excel for Big Data
Clustering and Taxonomy Creation for Massive Data Sets
Source Code for Keyword Correlations API
Big Data Problem That Epitomizes The Challenges of Data Science
What Map Reduce Can’t Do

Data Science: The End of Statistics?
Eight Worst Statistical Techniques
Marrying Computer Science, Statistics And Domain Expertize

The Big Data Ecosystem

Summary
Chapter 3: Becoming a Data Scientist

Types of Data Scientists
A Domain Expert, Analyst and Management Consultant
Horizontal Versus Vertical Data Scientist
Types of Data Scientists
Example of Amateur Data Science
Example of Extreme Data Science
Data Science Demographics

Training
University Programs
Certifications And Other Training

Data Science Apprenticeship
Online Training: The Basics
Special Tutorials
Data Sets
Projects
Source Code

The Independent Consultant
Finding Clients
Managing Your Finances
Salary Surveys
Sample Proposals
CRM Tools

The Entrepreneur
Our Story: Data Science Publisher
Startup Ideas For Data Scientists

Summary
Chapter 4: Data Science Craftsmanship - Part I

The Data Scientist
Data Scientist Versus Data Engineer
Data Scientist Versus Statistician
Data Scientist Versus Business Analyst

New Types of Metrics
Metrics To Optimize Digital Marketing Campaigns
Metrics For Fraud Detection

Choosing an Analytic Tool
Questions to Ask When Choosing Analytic Software
Questions to Ask When Considering Visualization Tools
Questions to Ask Regarding Real-Time Products

Programming Languages For Data science

Visualization
Producing Data Videos With R
More Sophisticated Videos

Statistical Modeling Without Models

http://semanticommunity.info/Data_Science/Data_Science_Central
Updated: Sat, 19 Sep 2015 08:46:35 GMT
Powered by mindtouch
Perl Code To Produce Data Sets
R Code To Produce Data Sets

New Types of Infographics
  Venn Diagrams
  Adding Dimensions To a Chart

Three Classes of Metrics: Centrality, Volatility, Bumpiness
  How Can Bumpiness Be Defined?
  About The Excel Spreadsheet
  Uses of the Bumpiness Coefficient

Statistical Clustering For Big Data
New Correlation and R Squared For Big Data
  A New Family of Rank Correlations
  Asymptotic Distribution, Normalization

Computer Science
  Computing q(n)
  A Theoretical Solution

Structuredness Coefficient
Identifying The Number of Clusters

Internet Topology Mapping
11 Features Any Database, SQL or NoSQL, Should Have

Additional Topics
  Belly Dancing Mathematics
  Securing Communications: Data Encoding
  10 Unusual Ways Analytics Are Used To Make Our Lives Better

Summary
Chapter 5: Data Science Craftsmanship - Part II

Data Dictionary
  What is a Data Dictionary
  How To Build a Dictionary

Hidden Decision Trees
  Implementation
  Example: Scoring Internet Traffic
  Conclusions

Model-Free Confidence Intervals
  Methodology
  The AnalyticBridge First Theorem
  Application
  Source Code
Random Numbers
Four Ways To Solve a Problem
Causation Versus Correlation
   So How Do We Detect Causes?
Lifecycle of Data Science Projects
   Predictive Modeling Mistakes
Logistic-Related Regressions
   Interactions Between Variables
   First Order Approximation
   Second Order Approximation
   Regression With Excel
Experimental Design
   Interesting Metrics
   Segmenting The Patient Population
   Customized Treatments
Analytics as a Service And API’s
   Example of Implementation
Miscellaneous Topics
   Preserving Scores When Datasets Change
   Optimizing Web Crawlers
   Hash Joins
   Simple Source Code To Simulate Clusters
Summary
Chapter 6: Data Science Applications - Case Studies
Stock Market
   Pattern To Boost Return By 500 Percent
   Optimizing Statistical Trading Strategies
   Stock Trading API: Statistical Model
   Stock Trading API: Implementation
   Stock Market Simulations
   Some Mathematics
   New Trends
Encryption
   Data Science Application: Steganography
   Solid Email Encryption
   Captcha Hack
Fraud Detection
   Click Fraud
Continuous Click Scores Versus Binary Fraud / Non Fraud
Mathematical Model, Bench-marking
Bias Due To Bogus Conversions
A Few Misconceptions
Statistical Challenges
Click Scoring to Optimize Keywords Bids
Automated, Fast Feature Selection With Combinatorial Optimization
Predictive Power of a Feature, Cross-Validation
Association Rules To Detect Collusion and Botnets
Extreme Value Theory For Pattern Detection

Digital Analytics
Online Advertising: Formula For Reach And Frequency
Email Marketing: Boosting Performance by 300 Percent
Optimizing Keyword Advertising Campaigns in 7 Days
Automated News Feed Optimization
Competitive Intelligence with Bit.ly
Measuring Return on Twitter Hashtags
Improving Google Searches with Three Fixes
Improving Relevancy Algorithms
Ad Rotation Problem

Miscellaneous
Better Sales Forecasts with Simpler Models
Better Detection of Healthcare Fraud
Attribution Modeling
Forecasting Meteorite Hits
Data Collection at Trailhead Parking
Other Application of Data Science

Summary
Chapter 7: Launching Your New Data Science Career
90 Job Interview Questions
About Your Experience
Technical Questions
General Questions
Little Data Science Projects
Testing Your Visual Thinking
Detecting Patterns with the Naked Eye
Identifying Aberrations
Misleading Time Series and Random Walks
From Statistician to Data Scientist

Data Scientists are also Statistical Practitioners
Who Should Teach Statistics to Data Scientists?
Hiring Issues, Government Statisticians
Data Scientists Work Closely with Data Architects
Who should be Involved in Strategic Thinking?
Two Types of Statisticians
Useless Statistical Jargon and Complexity
Using Big Data Versus Sampling

Taxonomy of Data Scientists
Top Data Scientists on LinkedIn
Data Science Most Popular Skill Mixes

400 Job Titles for Data Scientists

Salary Surveys
Create Your Own Salary Survey
Salary Breakdown Per Skill And Location

Summary

Chapter 8: Data Science Resources

Professional Resources
Data Sets
Books
Conferences and Organizations
Training
175 Websites
Definitions
Vendors
Top Data Scientists

Career Resources
6000 Companies Hiring Data Scientists
Sample Data Science Job Ads
Sample Resumes

New Synthetic Variance For Hadoop and Big Data
Synthetic Metrics
Hadoop, Numerical and Statistical Stability
The Abstract Concept of Variance
A New Big Data Theorem
Transformation-Invariant Metrics
Implementation: Communication versus Computational Costs
Bayesian Models, Final Comments

Summary

Addendum

1. Nine Categories of Data Scientists
2. Practical Illustration of Map-Reduce (Hadoop-Style), on Real Data
   Building a summary table: the Map step
   Building a summary table: the Reduce step
   Improvements
   Conclusions
3. Answers to Job Interview Questions
   Questions About Your Experience
   Technical questions
   General questions
   Questions about data science projects
4. Additional Topics
   Belly dancing mathematics
   Source Code to Produce the Video
   Saving and Exporting the Video
   Securing communications: data encoding
   10 unusual ways analytics are used to make our lives better
5. Improving Visuals
   Improving Venn Diagrams
   Adding Dimensions to a Chart
6. Essential Features for any Database, SQL or NoSQL

91 job interview questions for data scientists

Comment by Vincent Granville on May 5, 2013 at 7:44pm
Comment by Vincent Granville on February 16, 2013 at 7:19am
Comment by Vincent Granville on February 16, 2013 at 6:54am
Comment by Vincent Granville on February 15, 2013 at 10:22am
Comment by Vincent Granville on February 15, 2013 at 8:24am
Comment by Vincent Granville on February 14, 2013 at 6:52am

Data Science Apprenticeship
Data Sets for Download
Articles for Curriculum
Data Science eBook by Analyticbridge - 2nd Edition
   Part I - Data Science Recipes
   Part II - Data Science Discussions
   Part III - Data Science Resources
Proposal for an Apprenticeship in Data Science

Part I: Online training
Part II: Potential projects to be completed:
Part III: Students successfully completing two projects

How to enroll?

Related articles

---

Story

9 "must read" articles

Dr. Vincent Granville's recent blog below contains four things of real interest to me:

- **Big Data – From Descriptive to Prescriptive** Graphic below that addresses the subject of interest to the ESIP Earth Science Data Analytics
- **NodeXL Graph Gallery: Graph Details** featured at our recent Federal Big Data Working Group Meetup
- **The Sexiest Job of the 21st Century is Tedious, and that Needs to Change** (*)
  - (*) I (Vincent Granville) disagree with this Harvard Business Review author. Senior data scientists work on high level data from various sources, use automated processes for EDA (exploratory analysis) and spend little to no time in tedious, routine, mundane tasks (less than 5% of my time, in my case). I also use robust techniques that work well on relatively dirty data, and ... I create and design the data myself in many cases.
  - My (Brand Niemann) experience as a senior data scientist is similar in that I find the data preparation to be a very interesting and worthwhile activity that informs my data science results and story and I actually delight in "creating and designing the data myself in many cases." It is the art in the data science work along with the resulting visualizations. I also avoid coding if at all possible by using Spotfire.
- **Practical illustration of Map-Reduce (Hadoop-style), on real data**

Blog Posted by Vincent Granville on April 5, 2014 at 7:30am

Source: [http://www.datasciencecentral.com/pr...-read-articles](http://www.datasciencecentral.com/pr...-read-articles)

My selection of articles and resources recently posted in various news outlets - mostly from specialized publishers dealing with big data, machine learning, visualization and related topics. The picture below is from the first article.
Visually communicating the value of Big Data is challenging because of the need to convey different concepts simultaneously. The most popular category by far are plot charts on an X-Y axis. These charts plot analytical complexity against some sort of business value measurement in a positive correlation that looks entertainingly similar to human evolution charts we’ve all seen, with man becoming more upright and intelligent with time.

Regardless of graphic representation, they all follow the progression from **What Happened (descriptive analytics)**, **Why Did It Happen** (correlation analytics), **What Will Happen Next** (predictive analytics), and **What Should I Do About It** (prescriptive analytics).

This chart is unique in that it goes all the way back to the beginning when data is first created and gathered in raw form. So much of the resources needed to develop prescriptive analytics takes place in the very early stages of the process, and it’s nice that this graphic gives it a mention. The overwhelming majority of data available for analysis does make it to the final predictive/prescriptive model. If each circle represented the amount of actual data at that stage, the raw data circle (and cleaned data circle) would dwarf all the others, so thank you [SAP](http://www.sap.com) for giving data its due.

Continue Blog Posted by Vincent Granville on April 5, 2014 at 7:30am

My selection

- [Big Data – From Descriptive to Prescriptive](http://www.miprofs.com/big-data-desc...to-predictive/)
- [Can big data be racist?](http://www.miprofs.com/big-data-desc...to-predictive/)
- [NodeXL Graph Gallery: Graph Details](http://www.miprofs.com/big-data-desc...to-predictive/)
- [Best Metrics For Digital Marketing: Rock Your Own And Rent Strategies](http://www.miprofs.com/big-data-desc...to-predictive/)
- [Big Data: from mining to meaning](http://www.miprofs.com/big-data-desc...to-predictive/)
- [Beautiful versus useful visualizations (in French, but interesting)](http://www.miprofs.com/big-data-desc...to-predictive/)
- [Learning and Teaching Machine Learning: A Personal Journey](http://www.miprofs.com/big-data-desc...to-predictive/)
- [Big data techniques and technologies](http://www.miprofs.com/big-data-desc...to-predictive/)
- [The Sexiest Job of the 21st Century is Tedious, and that Needs to C... (*)](http://www.miprofs.com/big-data-desc...to-predictive/)
- [From the trenches: 360-degree data science](http://www.miprofs.com/big-data-desc...to-predictive/)
(*). I disagree with this Harvard Business Review author. Senior data scientists work on high level data from various sources, use automated processes for EDA (exploratory analysis) and spend little to no time in tedious, routine, mundane tasks (less than 5% of my time, in my case). I also use robust techniques that work well on relatively dirty data, and ... I create and design the data myself in many cases.

Other internal links
- [17 short tutorials all data scientists should read (and practice)]
- [10 types of data scientists]
- [66 job interview questions for data scientists]
- [Data Science Certification]
- [Update about our Data Science Apprenticeship]
- [Our Wiley Book on Data Science]
- [Data Science Top Articles]
- [Our Data Science Weekly Newsletter]
- [Practical illustration of Map-Reduce (Hadoop-style), on real data]
- [What makes up data science?]
- [Data science webinars]
- [Data science competition]

END OF BLOG

The [Addendum to the new book Developing Analytic Talent](http://semanticommunity.info/Data_Science/Data_Science_Central) contains the following section:

Practical Illustration of Map-Reduce (Hadoop-Style), on Real Data
- [Building a summary table: the Map step]
- [Building a summary table: the Reduce step]
- [Improvements]
- [Conclusions]

that helps understand what most people think big data is about and requires.

MORE TO FOLLOW

---

**Story**

**My book is finally available!!**

The Linkedin post said:

Learn the skills needed for the most in-demand tech job
Harvard Business Review calls it the sexiest tech job of the 21st century. Data scientists are in demand, and this unique book shows you exactly what employers want and the skill set that separates the quality data scientist from other talented IT professionals. Data science involves extracting, creating, and processing data to turn it into business value. This guide discusses the essential skills, such as statistics and visualization techniques, and covers everything from analytical recipes and data science tricks to common job interview questions, sample resumes, and source code.

The applications are endless and varied: automatically detecting spam and plagiarism, optimizing bid prices in keyword advertising, identifying new molecules to fight cancer, assessing the risk of meteorite impact. Complete with case studies, this book is a must, whether you’re looking to become a data scientist or to hire one.

- Explains the finer points of data science, the required skills, and how to acquire them, including analytical recipes, standard rules, source code, and a dictionary of terms
- Shows what companies are looking for and how the growing importance of big data has increased the demand for data scientists
- Features job interview questions, sample resumes, salary surveys, and examples of job ads
- Case studies explore how data science is used on Wall Street, in botnet detection, for online advertising, and in many other business-critical situations
- Developing Analytic Talent: Becoming a Data Scientist is essential reading for those aspiring to this hot career choice and for employers seeking the best candidates.

Posted to Amazon, April 16, 2014:  [http://www.amazon.com/review/R26UNBJIS4OTC0](http://www.amazon.com/review/R26UNBJIS4OTC0)

Visionary Data Scientist Publishes Excellent Book

First I am not going to spend my time on the subject of "Write a book review, earn $250". My review is without expectation of reward, although I would be honored to have an autographed copy!

I think Dr. Granville has made an outstanding contribution to the field of data science with his web sites and books. I have been eagerly awaiting the release of this book since I first saw the table of contents on-line, could order my copy, and was very pleased that Amazon got it to me 2 days before promised. Well done Amazon! And thanks to your founder and president for buying the Washington Post.

Second, lets focus these reviews on the book itself. I skim read it and could not put it down. I prepared a tutorial using it for my data science students and our Federal Big Data Working Group Meetup. The organization and content of the book is one of the best I have seen for both those new to data science and for teaching. My only complaint is that the URLs are hard to read and I always long for more data sets and results that can be downloaded to work with to gain hands-on experience.

I look forward to Dr. Granville's answers to constructive questions about the book, blogs on additional topics, and to the next book. I think this man deserves more respect and praise for his body of work and this book than he is getting in these early reviews.

Dr. Brand Niemann
Director and Senior Data Scientist
Semantic Community
http://semanticommunity.info
Story

From Data Science Central to Data Science Results

It started with an interesting email entitled Five Announcements which lead me to explore those five links, two of which were most interesting to me:

• Data Science Summit Next Week, in the Bay Area: Interesting collection of people and agenda topics, but I did not see any government people or topics, so I will not attend.

• Dr. Granville’s upcoming Wiley Data Science Book debunks all the myths about data science. If you think data science is not new, or is not a science, explore the table of contents of this 300+ page book: Very interesting because I am preparing a course and selected a book (Doing Data Science) I thought was the best, so I want to explore the table of contents.

Dr. Granville said: I have written 6 chapters out of 8. Here’s the table of content, for the material already written. I will update it regularly. It will be completed in December, published by Wiley by next April. It will also become part of our apprenticeship.

I also found mention of:

• Data Science eBook by Analyticbridge - 2nd Edition: See below
• Proposal for an Apprenticeship in Data Science: See below
• 91 interview questions for data scientists: See below
  ◦ Also see 6 Comments by Vincent Granville below on these questions

So I looked at the apprenticeship in more detail and found links to actual data sets like: Great statistical analysis: forecasting meteorite hits that led to:

• The Registered meteorites that has impacted on Earth visualized, which contained three useful comments I decided to try to implement:
  ◦ Comment by Ramon Martinez: Yes, this is important
  ◦ Comment by Vincent Granville: I did this.
  ◦ Comment by Vincent Granville: Interesting observation!

Vincent Granville is interested to see this info visually summarized in 5 dimensions, as follows:

• 2 dimensions for the location: Mouse over to see Latitude and Longitude
• 1 dimension for the size (represented by radius): Mouse over to see mass in 5 bins
• 1 dimension for the type (represented by color): Mouse over to see type of meteorite

Click on point to see Details-on-Demand. Then Unmark Marked Rows
• 1 dimension for time: turning this static image into a video, where each second represent (say) one year: Use Filter to Right to select Year. Then Reset All Filters.

See slide of screen capture and Spotfire Web Player below.

So this is my example of going from Data Science Central to Data Science Results.

Next, I will select more data sets to add to my data science training portfolio.

Registered meteorites that has impacted on Earth visualized

Source: [http://www.analyticbridge.com/profiles/blogs/registered-meteorites-that-has-impacted-on-earth-visualized](http://www.analyticbridge.com/profiles/blogs/registered-meteorites-that-has-impacted-on-earth-visualized)

• Posted by Amy on February 20, 2013 at 11:00am
• View Blog

See niche chart using Tableau, produced by the [Meteoretical Society](http://www.analyticbridge.com/profiles/blogs/registered-meteorites-that-has-impacted-on-earth-visualized). Below is a screen shot. The original chart is interactive (you can zoom in, etc.)

Download entire data set (see [Excel](http://www.analyticbridge.com/profiles/blogs/registered-meteorites-that-has-impacted-on-earth-visualized)): it's a 7MB spreadsheet consisting of 34,513 meteorites, last updated in 2012, with the following fields:

• place
• type_of_meteorite
• mass_g
• fell_found
• year
• database
• coordinate_1
• coordinates_2
Useful to make forecasts, broken down by meteorite size.

Related Articles:

- [Debunking the story about the Russian meteor event](#)
- [Can we use data science to measure distances to stars?](#)
- [Questions about astronomy and clustering](#)
- [How to identify specific spatial patterns? (application: star distr...](#)
- [Weekly digest](#)
- [Other resources](#)
- [Data Science eBook](#)
- [Data Science Apprenticeship](#)
- [Most popular blog posts](#)

Comment by **Ramon Martinez** on February 21, 2013 at 1:45pm

Thanks for featuring the data visualization "Registered meteorites that has fallen on Earth" at analyticbridge.com

I wanted to point out that this data visualization was produced & authored by Ramon Martinez using data from The Meteoritical Society. It is not produced by The Meteoritical Society as stated in this blog post.

Comment by **Vincent Granville** on February 20, 2013 at 7:06pm

Highly populated areas seem more heavily hit, but it's because meteorites in these areas are more easily detected. As inhabited areas grow, the yearly number of meteorites will increase over time. But we need to keep in mind that this is an artifact inherent to the dataset itself. It's a data bias.

Would be interested to see this info visually summarized in 5 dimensions, as follows:

- 2 dimensions for the location
- 1 dimension for the size (represented by radius)
- 1 dimension for the type (represented by color)
- 1 dimension for time: turning this static image into a video, where each second represent (say) one year

Anyway, this data set should help answer two questions, using sound statistical analysis:
• risk of hitting a populated area
• number of undetected meteorites, with trend

Comment by Vincent Granville on February 20, 2013 at 7:01pm

I actually touched a real meteorite today (see picture below), at the Pacific Science Center in Seattle. Was very, very heavy.

Photo: Meteorite at the Pacific Science Center

Slides

Slides and Slides

Spotfire Dashboard

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

Five Announcements Email

From: Advanced Business Analytics, Data Mining and Predictive Modeling [mailto:groups-noreply@linkedin.com]
Sent: Wednesday, November 27, 2013 3:32 PM
To: Brand Niemann
Subject: Five Announcements
LinkedIn Groups

- Group: Advanced Business Analytics, Data Mining and Predictive Modeling
- Subject: Five Announcements

1. New Hadoop Community

Data Science Central has created a new community: [http://www.Hadoop360.com](http://www.Hadoop360.com). This is a new DSC channel. Other channels include Analytics, Big Data, and Visualization. We will soon add more articles, including salaries for Hadoop engineers, Hadoop books and training, list of most influential Hadoop people, success stories, jobs, events etc.

2. Data Science Book

Dr. Granville's upcoming Wiley book debunks all the myths about data science. If you think data science is not new, or is not a science, explore the table of contents of this 300+ page book, at [http://bit.ly/17WNV5i](http://bit.ly/17WNV5i).

3. Most recent data science positions

Check out [http://www.analyticaltalent.com](http://www.analyticaltalent.com) to find most recent data science, big data and analytics positions.

4. Three reasons to attend the big Data Science Summit Next Week, in the Bay Area

Who's who of data science will be speaking (MIT Media Lab, Udacity, LinkedIn). An agenda geared towards you: case studies. Opportunity to collaborate in an intimate setting with your peers. Register today at [http://bit.ly/1cwztP4](http://bit.ly/1cwztP4) and enter the code Piv-VIP for a 20% discount.

5. New Data Science Central Acquisition

Data Science Central has acquired [http://www.MessagingNews.com](http://www.MessagingNews.com), a leading source of information for messaging professionals. This is separate from our other DSC channels, although Dr. Granville, co-founder, will occasionally publish articles that are relevant to both communities, for instance: new algorithm to compress terabytes of highly redundant web or email data, Algorithm to automatically categorize and prioritize email, and Better spam, copyright infringement and plagiarism detection.

Posted By Andrei Macsin
Developing Analytic Talent


- Posted by Vincent Granville on November 23, 2013 at 12:30pm
- View Blog

I have written 6 chapters out of 8. Here's the table of content ([PDF](PDF)), for the material already written. I will update it regularly. It will be completed in December, published by Wiley by next April. It will also become part of our [apprenticeship](http://www.datasciencecentral.com/profiles/blogs/my-data-science-book).

![Book Cover](image)

My Note: I ordered the book on April 4th to read and extract excerpts in the table of contents below.

**Chapter 1: What is Data Science?**

**Fake Data Science**

**Fake data science: Two Examples**
The Face of the New University

Thirteen Problems

DUI Arrests Decreases After State Monopoly on Liquor Sales Ends

Data Science Defeats Intuition

Data Glitch Turns Data into Gibberish

Regression in Unusual Spaces

Analytics versus Seduction to Boost Sales

About Hidden Data

High Crime Rates Caused by Gasoline Lead. Really?

Boeing's Dreamliner Problems

Seven Tricky Sentences for NLP

Data Scientists Dictate What We Eat

Increasing Amazon.com Sales with Better Relevancy

Detecting Fake Profiles or Likes on Facebook

Analytics for Restaurants

History and Milestones

Statistics Will Experience a Renaissance
A Few Random Thoughts

History And Milestones

Modern Trends

Data Scientist Versus Data Architect

Summary

Chapter 2: Big Data is Different

Two Big Data Issues

The Curse of Big Data

When Data Flows Faster Than it Can Be Processed

Examples of Big Data Techniques

Excel for Big Data

Clustering and Taxonomy Creation for Massive Data Sets

Source Code for Keyword Correlations API

Big Data Problem That Epitomizes The Challenges of Data Science

What Map Reduce Can't Do

Data Science: The End of Statistics?

Eight Worst Statistical Techniques
Chapter 3: Becoming a Data Scientist

Types of Data Scientists

A Domain Expert, Analyst and Management Consultant

Horizontal Versus Vertical Data Scientist

Types of Data Scientists

Example of Amateur Data Science

Example of Extreme Data Science

Data Science Demographics

Source: [http://www.datasciencecentral.com/pr...t-demographics](http://www.datasciencecentral.com/pr...t-demographics)

Posted by Mirko Krivanek on May 22, 2013 at 7:00pm

The graphs below show demographics for two popular but totally unrelated analytic websites. The demographics are strikingly similar.

USA is still the #1 country by far. Visitors from India spend more time on the site. Although you can't tell from these charts, US visitors tend to have more senior / manager roles, while in India we have more junior, entry-level analysts, and almost no executive. A few places have experienced big growth: Ireland, Singapore, London.

Data source: Quantcast.com. Data about gender, income, education level, and race is produced using

- Demographic information about zipcodes. Zip code data is derived from your IP address.
- User information from ISPs. Several big ISPs (Internet Service Providers) sell (anonymized?) information about user accounts, broken down by IP address, to Quantcast. Then Quantcast's data science engine performs statistical inferences based on this sample ISP data.
Surveys or other sources, e.g. when you install some toolbar on your browser, allowing your web browsing activity to be monitored to produce summary statistics (this methodology is subject to big biases as people installing toolbars are very different from those who don't). This was Alexa's favorite way to produce the traffic statistics, a while back.

An index of 145 means that the website in question, for the metric in question, is 45% above average. The 100 reference (100 = average) is computed across all internet users. It is clear that these two websites attract highly educated, wealthy males, predominantly with Asian origin, living mostly in US.

**Question:** are there any major differences between the two websites? I don't see any.
Training

University Programs

Certifications And Other Training

Data Science Apprenticeship

Online Training: The Basics

Special Tutorials

Data Sets

Projects

Source Code

The Independent Consultant

Finding Clients

Managing Your Finances

Salary Surveys
Sample Proposals

CRM Tools

The Entrepreneur

Our Story: Data Science Publisher

Startup Ideas For Data Scientists

Summary

Chapter 4: Data Science Craftsmanship - Part I

The Data Scientist

Data Scientist Versus Data Engineer

Data Scientist Versus Statistician

Data Scientist Versus Business Analyst

New Types of Metrics

Metrics To Optimize Digital Marketing Campaigns

Metrics For Fraud Detection

Choosing an Analytic Tool

Questions to Ask When Choosing Analytic Software

Questions to Ask When Considering Visualization Tools
Questions to Ask Regarding Real-Time Products

Programming Languages For Data science

Visualization

Producing Data Videos With R

More Sophisticated Videos

Statistical Modeling Without Models

Perl Code To Produce Data Sets

R Code To Produce Data Sets

New Types of Infographics

Venn Diagrams

Adding Dimensions To a Chart

Three Classes of Metrics: Centrality, Volatility, Bumpiness

How Can Bumpiness Be Defined?

About The Excel Spreadsheet

Uses of the Bumpiness Coefficient

Statistical Clustering For Big Data

New Correlation and R Squared For Big Data
A New Family of Rank Correlations

Asymptotic Distribution, Normalization

Computer Science

Computing q(n)

A Theoretical Solution

Structuredness Coefficient

Identifying The Number of Clusters

Internet Topology Mapping

11 Features Any Database, SQL or NoSQL, Should Have

Additional Topics

Belly Dancing Mathematics

Securing Communications: Data Encoding

10 Unusual Ways Analytics Are Used To Make Our Lives Better

Summary

Chapter 5: Data Science Craftsmanship - Part II

Data Dictionary

What is a Data Dictionary

How To Build a Dictionary
Hidden Decision Trees

Implementation

Example: Scoring Internet Traffic

Conclusions

Model-Free Confidence Intervals

Methodology

The AnalyticBridge First Theorem

Application

Source Code

Random Numbers

Four Ways To Solve a Problem

Causation Versus Correlation

So How Do We Detect Causes?

Lifecycle of Data Science Projects

Predictive Modeling Mistakes

Logistic-Related Regressions

Interactions Between Variables
First Order Approximation

Second Order Approximation

Regression With Excel

Experimental Design

Interesting Metrics

Segmenting The Patient Population

Customized Treatments

Analytics as a Service And API’s

Example of Implementation

Miscellaneous Topics

Preserving Scores When Datasets Change

Optimizing Web Crawlers

Hash Joins

Simple Source Code To Simulate Clusters

Summary

Chapter 6: Data Science Applications - Case Studies

Stock Market
Pattern To Boost Return By 500 Percent

Optimizing Statistical Trading Strategies

Stock Trading API: Statistical Model

Stock Trading API: Implementation

Stock Market Simulations

Some Mathematics

New Trends

Encryption

Data Science Application: Steganography

Solid Email Encryption

Captcha Hack

Fraud Detection

Click Fraud

Continuous Click Scores Versus Binary Fraud / Non Fraud

Mathematical Model, Bench-marking

Bias Due To Bogus Conversions

A Few Misconceptions
Statistical Challenges

Click Scoring to Optimize Keywords Bids

Automated, Fast Feature Selection With Combinatorial Optimization

Predictive Power of a Feature, Cross-Validation

Association Rules To Detect Collusion and Botnets

Extreme Value Theory For Pattern Detection

Digital Analytics

Online Advertising: Formula For Reach And Frequency

Email Marketing: Boosting Performance by 300 Percent

Optimizing Keyword Advertising Campaigns in 7 Days

Automated News Feed Optimization

Competitive Intelligence with Bit.ly

Measuring Return on Twitter Hashtags

Improving Google Searches with Three Fixes

Improving Relevancy Algorithms

Ad Rotation Problem
Miscellaneous

Better Sales Forecasts with Simpler Models

Better Detection of Healthcare Fraud

Attribution Modeling

Forecasting Meteorite Hits

Data Collection at Trailhead Parking

Other Application of Data Science

Summary

Chapter 7: Launching Your New Data Science Career

90 Job Interview Questions

About Your Experience

Technical Questions

General Questions

Little Data Science Projects

Testing Your Visual Thinking

Detecting Patterns with the Naked Eye

Identifying Aberrations
Misleading Time Series and Random Walks

From Statistician to Data Scientist

Data Scientists are also Statistical Practitioners

Who Should Teach Statistics to Data Scientists?

Hiring Issues, Government Statisticians

Data Scientists Work Closely with Data Architects

Who should be Involved in Strategic Thinking?

Two Types of Statisticians

Useless Statistical Jargon and Complexity

Using Big Data Versus Sampling

Taxonomy of Data Scientists

Top Data Scientists on LinkedIn

Data Science Most Popular Skill Mixes

400 Job Titles for Data Scientists

Salary Surveys

Create Your Own Salary Survey

Salary Breakdown Per Skill And Location
Chapter 8: Data Science Resources

Professional Resources

Data Sets

Books

Conferences and Organizations

Training

175 Websites

Definitions

Vendors

Top Data Scientists

Career Resources

6000 Companies Hiring Data Scientists

Sample Data Science Job Ads

Sample Resumes

New Synthetic Variance For Hadoop and Big Data

Synthetic Metrics
1. Nine Categories of Data Scientists

Just like there are a few categories of statisticians (biostatisticians, statisticians, econometricians, operations research specialists, actuaries) or business analysts (marketing-oriented, product-oriented, finance-oriented, etc.) we have different categories of data scientists. First, many data scientists have a job title different from data scientist, mine for instance is co-founder. Check the "related articles" section below to discover 400 potential job titles for data scientists.

Categories of data scientists

- Those strong in statistics: they sometimes develop new statistical theories for big data that even traditional statisticians are not aware of. They are expert in statistical modeling, experimental design, sampling, clustering, data reduction, confidence intervals, testing, modeling, predictive modeling and other related techniques.
- Those strong in mathematics: NSA (national security agency) or defense/military people working on big data, astronomers, and operations research people doing analytic business optimization (inventory management and forecasting, pricing optimization, supply chain, quality control, yield optimization) as they collect, analyze and extract value out of data.
- Those strong in data engineering, Hadoop, database/memory/file systems optimization and architecture, API's, Analytics as a Service, optimization of data flows, data plumbing.
- Those strong in machine learning / computer science (algorithms, computational complexity)
- Those strong in business, ROI optimization, decision sciences, involved in some of the tasks traditionally performed by business analysts in bigger companies (dashboards design, metric mix selection and metric definitions, ROI optimization, high-level database design)
• Those strong in production code development, **software engineering** (they know a few programming languages)
• Those strong in visualization
• Those strong in GIS, **spatial data**, data modeled by graphs, graph databases
• Those strong in a few of the above. After 20 years of experience across many industries, big and small companies (and lots of training), I'm strong both in stats, machine learning, business, mathematics and more than just familiar with visualization and data engineering. This could happen to you as well over time, as you build experience. I mention this because so many people still think that it is not possible to develop a strong knowledge base across multiple domains that are traditionally perceived as separated (the silo mentality). Indeed, that's the very reason why data science was created.

Most of them are familiar or expert in big data.

There are other ways to categorize data scientists, see for instance our article on Taxonomy of data scientists. A different categorization would be creative versus mundane. The "creative" category has a better future, as mundane can be outsourced (anything published in textbooks or on the web can be automated or outsourced - job security is based on how much you know that no one else know or can easily learn). Along the same lines, we have science users (those using science, that is, practitioners; often they do not have a PhD), innovators (those creating new science, called researchers), and hybrids. Most data scientists, like geologists helping predict earthquakes, or chemists designing new molecules for big pharma, are scientists, and they belong to the user category.

**Implications for other IT professionals**

You (engineer, business analyst) probably do already a bit of data science work, and know already some of the stuff that some data scientists do. It might be easier than you think to become a data scientist. Check out our book (listed below in "related articles"), to find out what you already know, what you need to learn, to broaden your career prospects.

Are data scientists a threat to your job/career? Again, check our book (listed below) to find out what data scientists do, if the risk for you is serious (you = the business analyst, data engineer or statistician; risk = being replaced by a data scientist who does everything) and find out how to mitigate the risk (learn some of the data scientist skills from our book, if you perceive data scientists as competitors).

**2. Practical Illustration of Map-Reduce (Hadoop-Style), on Real Data**

Here I will discuss a general framework to process web traffic data. The concept of Map-Reduce will be naturally introduced. Let's say you want to design a system to score Internet clicks, to measure the chance for a click to convert, or the chance to be fraudulent or un-billable. The data comes from a publisher or ad network; it could be Google. Conversion data is limited and poor (some conversions are tracked, some are not; some conversions are soft, just a click-out, and conversion rate is above 10%; some conversions are hard, for instance a credit card purchase, and conversion rate is below 1%). Here, for now, we just ignore the conversion data and focus on the low hanging fruits: click data. Other valuable data is impression data (for instance a click not associated with an impression is very suspicious). But impression data is huge, 20 times bigger than click data. We ignore impression data here.
Here, we work with complete click data collected over a 7-day time period. Let's assume that we have 50 million clicks in the data set. Working with a sample is risky, because much of the fraud is spread across a large number of affiliates, and involve clusters (small and large) of affiliates, and tons of IP addresses but few clicks per IP per day (low frequency).

The data set (ideally, a tab-separated text file, as CSV files can cause field misalignment here due to text values containing field separators) contains 60 fields: keyword (user query or advertiser keyword blended together, argh...), referral (actual referral domain or ad exchange domain, blended together, argh...), user agent (UA, a long string; UA is also known as browser, but it can be a bot), affiliate ID, partner ID (a partner has multiple affiliates), IP address, time, city and a bunch of other parameters.

- The first step is to extract the relevant fields for this quick analysis (a few days of work). Based on domain expertise, we retained the following fields:
  - IP address
  - Day
  - UA (user agent) ID - so we created a look-up table for UA's
  - Partner ID
  - Affiliate ID

These 5 metrics are the base metrics to create the following summary table. Each (IP, Day, UA ID, Partner ID, Affiliate ID) represents our atomic (most granular) data bucket.

**Building a summary table: the Map step**

The summary table will be built as a text file (just like in Hadoop), the data key (for joins or groupings) being (IP, Day, UA ID, Partner ID, Affiliate ID). For each atomic bucket (IP, Day, UA ID, Partner ID, Affiliate ID) we also compute:

- number of clicks
- number of unique UA's
- list of UA
The list of UA's, for a specific bucket, looks like ~6723|9~45|1~784|2, meaning that in the bucket in question, there are three browsers (with ID 6723, 45 and 784), 12 clicks (9 + 1 + 2), and that (for instance) browser 6723 generated 9 clicks.

In Perl, these computations are easily performed, as you sequentially browse the data. The following updates the click count:

```perl
$hash_clicks{"IP\tDay\tUA_ID\tPartner_ID\tAffiliate_ID"};
```

Updating the list of UA's associated with a bucket is a bit less easy, but still almost trivial.

The problem is that at some point, the hash table becomes too big and will slow down your Perl script to a crawl. The solution is to split the big data in smaller data sets (called subsets), and perform this operation separately on each subset. This is called the Map step, in Map-Reduce. You need to decide which fields to use for the mapping. Here, IP address is a good choice because it is very granular (good for load balance), and the most important metric. We can split the IP address field in 20 ranges based on the first byte of the IP address. This will result in 20 subsets. The splitting in 20 subsets is easily done by browsing sequentially the big data set with a Perl script, looking at the IP field, and throwing each observation in the right subset based on the IP address.

**Building a summary table: the Reduce step**

Now, after producing the 20 summary tables (one for each subset), we need to merge them together. We can't simply use hash table here, because they will grow too large and it won't work - the reason why we used the Map step in the first place.

**Here's the work around:**

Sort each of the 20 subsets by IP address. Merge the sorted subsets to produce a big summary table T. Merging sorted data is very easy and efficient: loop over the 20 sorted subsets with an inner loop over the observations in each sorted subset; keep 20 pointers, one per sorted subset, to keep track of where you are in your browsing, for each subset, at any given iteration.

Now you have a big summary table T, with multiple occurrences of the same atomic bucket, for many atomic buckets. Multiple occurrences of a same atomic bucket must be aggregated. To do so, browse sequentially table T (stored as text file). You are going to use hash tables, but small ones this time. Let's say that you are in the middle of a block of data corresponding to a same IP address, say 231.54.86.109 (remember, T is ordered by IP address). Use

```perl
$hash_clicks_small{"Day\tUA_ID\tPartner_ID\tAffiliate_ID"};
```

to update (that is, aggregate click count) corresponding to atomic bucket (231.54.86.109, Day, UA ID, Partner ID, Affiliate ID). Note one big difference between $hash_clicks and $hash_clicks_small: IP address is not part of the key in the latter one, resulting in hash tables millions of time smaller. When you hit a new IP address when browsing T, just save the stats stored in $hash_small and satellites small hash tables for the previous IP address, free the memory used by these hash tables, and re-use them for the next IP address found in table T, until you arrived at the end of table T.
Now you have the summary table you wanted to build, let's call it S. The initial data set had 50 million clicks and dozens of fields, some occupying a lot of space. The summary table is much more manageable and compact, although still far too large to fit in Excel.

Creating rules

The rule set for fraud detection will be created based only on data found in the final summary table S (and additional high-level summary tables derived from S alone). An example of rule is "IP address is active 3+ days over the last 7 days". Computing the number of clicks and analyzing this aggregated click bucket, is straightforward, using table S. Indeed, the table S can be seen as a "cube" (from a database point of view), and the rules that you create simply narrow down on some of the dimensions of this cube. In many ways, creating a rule set consists in building less granular summary tables, on top of S, and testing.

Improvements

IP addresses can be mapped to an IP category, and IP category should become a fundamental metric in your rule system. You can compute summary statistics by IP category. See details in my article Internet topology mapping. Finally, automated nslookups should be performed on thousands of test IP addresses (both bad and good, both large and small in volume).

Likewise, UA’s (user agents) can be categorized, a nice taxonomy problem by itself. At the very least, use three UA categories: mobile, (nice) crawler that identifies itself as a crawler, and other. The use of UA list such as ~6723|9~45|1~784|2 (see above) for each atomic bucket is to identify schemes based on multiple UA’s per IP, as well as the type of IP proxy (good or bad) we are dealing with.

Historical note: Interestingly, the first time I was introduced to a Map-Reduce framework was when I worked at Visa in 2002, processing rather large files (credit card transactions). These files contained 50 million observations. SAS could not sort them, it would make SAS crashes because of the many and large temporary files SAS creates to do big sort. Essentially it would fill the hard disk. Remember, this was 2002 and it was an earlier version of SAS, I think version 6. Version 8 and above are far superior. Anyway, to solve this sort issue - an O(n log n) problem in terms of computational complexity - we used the "split / sort subsets / merge and aggregate" approach described in my article.

Conclusions

I showed you how to extract/summarize data from large log files, using Map-Reduce, and then creating an hierarchical data base with multiple, hierarchical levels of summarization, starting with a granular summary table S containing all the information needed at the atomic level (atomic data buckets), all the way up to high-level summaries corresponding to rules. In the process, only text files are used. You can call this an NoSQL Hierarchical Database (NHD). The granular table S (and the way it is built) is similar to the Hadoop architecture.
3. Answers to Job Interview Questions

These are mostly open-ended questions that prospective employers may ask to assess the technical horizontal knowledge of a senior candidate for a high-level position, such as a director, as well as to junior candidates. Here we provide answers to some of the most challenging questions listed in the book.

Questions About Your Experience

My Note: This was not in the outline

These are all open-ended questions. The questions are listed in the book. We haven’t provided answers.

Technical questions

• What are lift, KPI, robustness, model fitting, design of experiments, and the 80/20 rule?

  **Answers:** *KPI* stands for Key Performance Indicator, or metric, sometimes called feature. A *robust* model is one that is not sensitive to changes in the data. *Design of experiments or experimental design* is the initial process used (before data is collected) to split your data, sample and set up a data set for statistical analysis, for instance in A/B testing frameworks or clinical trials. The 80/20 rules means that 80 percent of your income (or results) comes from 20 percent of your clients (or efforts).

• What are collaborative filtering, n-grams, Map Reduce, and cosine distance?

  **Answers:** *Collaborative filtering/recommendation engines* are a set of techniques where recommendations are based on what your friends like, used in social context networks. *N-grams* are token permutations associated with a keyword, for instance “car insurance California,” “California car insurance,” “insurance California car,” and so on. *Map Reduce* is a framework to process large data sets, splitting them into subsets, processing each subset on a different server, and then blending results obtained on each. *Cosine distance* measures how close two sentences are to each other, by counting the number of terms that their share. It does not take into account synonyms, so it is not a precise metric in NLP (natural language processing) contexts. All of these problems are typically solved with Hadoop/Map-Reduce, and the reader should check the index to find and read our various discussions on Hadoop and Map-Reduce in this book.

• What is probabilistic merging (aka fuzzy merging)? Is it easier to handle with SQL or other languages? Which languages would you choose for semi-structured text data reconciliation?

  **Answers:** You do a join on two tables A and B, but the keys (for joining) are not compatible. For instance, on A the key is first name/lastname in some char set; in B another char set is used. Data is sometimes missing in A or B. Usually, scripting languages (Python and Perl) are better than SQL to handle this.

• Toad or Brio or any other similar clients are quite inefficient to query Oracle databases. Why? What would you do to increase speed by a factor 10 and be able to handle far bigger outputs?

• What are hash table collisions? How are they avoided? How frequently do they happen?

• How can you make sure a Map Reduce application has good load balance? What is load balance?

• Is it better to have 100 small hash tables or one big hash table in memory, in terms of access speed (assuming both fit within RAM)? What do you think about in-database analytics?
• Why is Naive Bayes so bad? How would you improve a spam detection algorithm that uses Naïve Bayes?

**Answers:** It assumes features are not correlated and thus penalize heavily observations that trigger many features in the context of fraud detection. Solution: use a solution such as hidden decision trees (described in this book) or decorrelate your features.

• What is star schema? What are lookup tables?

**Answers:** The star schema is a traditional database schema with a central (fact) table (the “observations”, with database “keys” for joining with satellite tables, and with several fields encoded as ID’s). Satellite tables map ID’s to physical name or description and can be be “joined” to the central, fact table using the ID fields; these tables are known as lookup tables, and are particularly useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve multiple layers of summarization (summary tables, from granular to less granular) to retrieve information faster.

• Can you perform logistic regression with Excel and if so, how, and would the result be good? Answers: Yes; by using `linest` on log-transformed data.

• Define quality assurance, six sigma, and design of experiments. Give examples of good and bad designs of experiments.

• What are the drawbacks of general linear model? Are you familiar with alternatives (Lasso, ridge regression, and boosted trees)?

• Do you think 50 small decision trees are better than 1 large one? Why?

**Answer:** Yes, 50 create a more robust model (less subject to over-fitting) and easier to interpret.

• Give examples of data that does not have a Gaussian distribution or log-normal. Give examples of data that has a chaotic distribution.

**Answer:** Salary distribution is a typical examples. Multimodal data (modeled by a mixture distribution) is another example.

• Why is mean square error a bad measure of model performance? What would you suggest instead?

**Answer:** It puts too much emphasis on large deviations. Instead, use the new correlation coefficient described in chapter 4.

• What is a cron job?

**Answer:** A scheduled task that runs in the background on a server.

• What is an efficiency curve? What are its drawbacks, and how can they be overcome?

• What is a recommendation engine? How does it work?

• What is an exact test? How and when can simulations help when you do not use an exact test?

• What is the computational complexity of a good, fast clustering algorithm? What is a good clustering algorithm? How do you determine the number of clusters? How would you perform clustering on 1 million unique keywords, assuming you have 10 million data points—each one consisting of two keywords, and a metric measuring how similar these two keywords are? How would you create this 10 million data points table?

**Answers:** Found in Chapter 2, “Why Big Data Is Different.”

• Should removing stop words be Step 1 rather than Step 3, in the search engine algorithm described in the section “Big Data Problem That Epitomizes the Challenges of Data Science?”
Answer: Have you thought about the fact that *mine* and *yours* could also be stop words? So in a bad implementation, *data mining* would become *data mine* after stemming and then *data*. In practice, you remove stop words before stemming. So Step 3 should indeed become step 1.

**General questions**

- Should click data be handled in real time? Why? In which contexts?

  Answer: In many cases, real-time data processing is not necessary and even undesirable as end-of-day algorithms are usually much more powerful. However, real-time processing is critical for fraud detection, as it can deplete advertiser daily budgets in a few minutes.

- What is better: good data or good models? And how do you define good? Is there a universal good model? Are there any models that are definitely not so good?

  Answer: There is no universal model that is perfect in all situations. Generally speaking, good data is better than good models, as model performance can easily been increased by model testing and cross-validation. Models that are “too good to be true” usually are - they might perform very well on test data (data used to build the model) because of over-fitting, and very badly outside the test data.

- How do you handle missing data? What *imputation* techniques do you recommend?

- Compare SAS, R, Python, and Perl

  Answer: Python is the new scripting language on the block with many libraries constantly added (Pandas for data science), Perl is the old one (still great and the best - very flexible - if you don’t work in a team and don’t have to share your code). R has great visualization capabilities; also R libraries are available to handle data that is too big to fit in memory. SAS is used in a lot of traditional statistical applications; it has a lot of built-in functions, and can handle large data sets pretty past; the most recent versions of SAS have neat programming features and structures such as fast sort and hash tables. Sometimes you must use SAS because that’s what your client is using.

- What is the curse of big data? (See section “The curse of big data” in chapter 2).

- What are examples in which Map Reduce does not work? What are examples in which it works well? What is the security issues involved with the cloud? What do you think of EMC’s solution offering a hybrid approach—to both an internal and external cloud—to mitigate the risks and offer other advantages (which ones)? The reader will find more information about key players in this market, in the section “The Big Data Ecosystem” in chapter 2, and especially in the online references (links) provided in this section.

- What is sensitivity analysis? Is it better to have low sensitivity (that is, great robustness) and low predictive power, or vice versa? How to perform good cross-validation? What do you think about the idea of injecting noise in your data set to test the sensitivity of your models?

- Compare logistic regression with decision trees and neural networks. How have these technologies been vastly improved over the last 15 years?


- What is root cause analysis? How can you identify a cause versus a correlation? Give examples.

  Answer: in Chapter 5, “Data Science craftsmanship: Part II.”
• How would you define and measure the predictive power of a metric?
  
  \textbf{Answer:} in Chapter 6.

• Is it better to have too many false positives, or too many false negatives?

• Have you ever thought about creating a start-up? Around which idea / concept?

• Do you think that a typed login/password will disappear? How could they be replaced?

• What do you think makes a good data scientist?

• Do you think data science is an art or a science?

• Give a few examples of best practices in data science.

• What could make a chart misleading, or difficult to read or interpret? What features should a useful chart have?

• Do you know a few rules of thumb used in statistical or computer science? Or in business analytics?

• What are your top five predictions for the next 20 years?

• How do you immediately know when statistics published in an article (for example, newspaper) are either wrong or presented to support the author's point of view, rather than correct, comprehensive factual information on a specific subject? For instance, what do you think about the official monthly unemployment statistics regularly discussed in the press? What could make them more accurate?

• In your opinion, what is data science? Machine learning? Data mining?

\textbf{Questions about data science projects}

• How can you detect individual paid accounts shared by multiple users? This is a big problem for publishers, digital newspapers, software developers offering API access, the music / movie industry (file sharing issues) and organizations offering monthly paid access (flat fee not based on the numbers of accesses or downloads) to single users, to view or download content, as it results in revenue losses.

• How can you optimize a web crawler to run much faster, extract better information, and better summarize data to produce cleaner databases?

  \textbf{Answer:} Putting a time-out threshold to less than 2 seconds, extracting no more than the first 20 KB of each pages, and not revisiting pages already crawled (that is, avoid recurrence in your algorithm) are good starting points.

• How would you come up with a solution to identify plagiarism?

• You are about to send 1 million e-mails (marketing campaign). How do you optimize delivery? How do you optimize response? Can you optimize both separately?

  \textbf{Answer:} Not really.

• How would you turn unstructured data into structured data? Is it really necessary? Is it OK to store data as flat text files rather than in a SQL-powered RDBMS?

• How would you build nonparametric confidence intervals, for scores?

• How can you detect the best rule set for a fraud detection scoring technology? How can you deal with rule redundancy, rule discovery, and the combinatorial nature of the problem (for finding optimum rule set—the one with best predictive power)? Can an approximate solution to the rule set problem be OK? How would you find an OK approximate solution? How would you decide it is good enough and stop looking for a better one?

• How can you create keyword taxonomies?

• What is a botnet? How can it be detected?
• How does Zillow's algorithm work to estimate the value of any home in the United States?

• How can you detect bogus reviews or bogus Facebook accounts used for bad purposes?

• How would you create a new anonymous digital currency? Please focus on the aspects of security - how do you protect this currency against Internet pirates? Also, how do you make it easy for stores to accept it? For instance, each transaction has a unique ID used only once and expiring after a few days; and the space for unique ID's is very large to prevent hackers from creating valid ID's just by chance.

• You design a robust nonparametric statistic (metric) to replace correlation or R square that is independent of sample size, always between –1 and +1 and based on rank statistics. How do you normalize for sample size? Write an algorithm that computes all permutations of n elements. How do you sample permutations (that is, generate tons of random permutations) when n is large, to estimate the asymptotic distribution for your newly created metric? You may use this asymptotic distribution for normalizing your metric. Do you think that an exact theoretical distribution might exist, and therefore, you should find it and use it rather than wasting your time trying to estimate the asymptotic distribution using simulations?

• Here’s a more difficult, technical question related to previous one. There is an obvious one-to-one correspondence between permutations of n elements and integers between 1 and factorial n. Design an algorithm that encodes an integer less than factorial n as a permutation of n elements. What would be the reverse algorithm used to decode a permutation and transform it back into a number? Hint: An intermediate step is to use the factorial number system representation of an integer. You can check this reference online to answer the question. Even better, browse the web to find the full answer to the question. (This will test the candidate's ability to quickly search online and find a solution to a problem without spending hours reinventing the wheel.)

• How many "useful" votes will a Yelp review receive?

  Answer: Eliminate bogus accounts or competitor reviews. (How to detect them: Use taxonomy to classify users, and location. Two Italian restaurants in same ZZIP code could badmouth each other and write great comments for themselves.) Detect fake likes: Some companies (for example, FanMeNow.com) will charge you to produce fake accounts and fake likes. Eliminate prolific users who like everything; those who hate everything. Have a blacklist of keywords to filter fake reviews. See if an IP address or IP block of a reviewer is in a blacklist such as Stop Forum Spam. Create a honeypot to catch fraudsters. Also watch out for disgruntled employees badmouthing their former employer. Watch out for two or three similar comments posted the same day by three users regarding a company that receives few reviews. Is it a new company? Add more weight to trusted users. (Create a category of trusted users.) Flag all reviews that are identical (or nearly identical) and come from the same IP address or the same user. Create a metric to measure the distance between two pieces of text (reviews). Create a review or reviewer taxonomy. Use hidden decision trees to rate or score review and reviewers.


• This is a question about experimental design (and a bit of computer science) with Legos. Say you purchase two sets of Legos (one set to build a car A and a second set to build another car B). Now assume that the overlap between the two sets is substantial. There are three different ways that you can build the two cars. The first step consists in sorting the pieces (Legos) by color and maybe also by size. The three ways to proceed are

1. Sequentially: Build one car at a time. This is the traditional approach.

2. Semi-parallel system: Sort all the pieces from both sets simultaneously so that the pieces will be blended. Some in the red pile will belong to car A; some will belong to car B. Then build the two cars sequentially, following the instructions in the accompanying leaflets.
3. In parallel: Sort all the pieces from both sets simultaneously, and build the two cars simultaneously, progressing simultaneously with the two sets of instructions.

Which is the most efficient way to proceed?

**Answer:** The least efficient is sequential. Why? If you are a good at multitasking, the full parallel approach is best. Why? Note that with the semi-parallel approach, the first car that you build will take less time than the second car (due to easier way to find the pieces that you need because of higher redundancy) and less time than needed if you used the sequential approach (for the same reason). To test these assumptions and help you become familiar with the concept of distributed architecture, you can have your kid build two Lego cars, A and B, in parallel and then two other cars, C and D, sequentially. If the overlap between A and B (the proportion of Lego pieces that are identical in both A and B) is small, then the sequential approach will work best. Another concept that can be introduced is that building an 80-piece car takes more than twice as much time as building a 40-piece car. Why? (The same also applies to puzzles).

- How can you design an algorithm to estimate the number of LinkedIn connections your colleagues have? The number is readily available on LinkedIn for your first-degree connections with up to 500 LinkedIn connections. However, if the number is above 500, it just says 500+. The number of shared connections between you and any of your LinkedIn connections is always available.

  **Answer:** A detailed algorithm can be found in the book, in the section “job interview questions”.

- How would proceed to reverse engineer the popular BMP image format and create your own images byte by byte, with a programming language?

  **Answer:** The easiest BMP format is the 24-bit format, where 1 pixel is stored on 3 bits, the RGB (red/green/blue) components of the pixel, which determines its pixel. After producing a few sample images, it becomes obvious that the format starts with a 52-bit header. Starts with a blank image. Modify the size. This will help you identify where and how the size is encoded in the 52-bit header. On a test image (for example, 40 * 60 pixels), change the color of one pixel. See what happens to the file: Three bytes have a new value, corresponding to the red, green, and blue components of the color attached to your modified pixel. Now you can understand the format and play with it any way you want with low-level access.

4. Additional Topics

These topics are part of the Data Science Apprenticeship, but not included in our Wiley book

**Belly dancing mathematics**

This is a follow up to our article, “Structured coefficient,” published in chapter 4 in our book. I went one step further here, to create a new data video (see previous sections in this chapter about producing data videos with R), where rotating points (by accident) simulate sensual waves, not unlike belly dancing. You'll understand what I mean [when you see the video](http://semanticommunity.info/Data_Science/Data_Science_Central).
Here I provide the source code and methodology to create this video. Click here to view the video, including a link to the YouTube version.

Algorithm to Create the Data

The video features a rectangle with 20 x 20 moving points, initially evenly spaced, each point rotating around an invisible circle. Following is the algorithm source code. The parameter \( c \) governs the speed of the rotations and \( r \) the radius of the underlying circles.

```perl
open(OUT,">movie.txt");

print OUT "iter	x	y	z\n";

for ($iter=0; $iter<1000; $iter++) {
    for ($k=0; $k<20; $k++) {
        for ($l=0; $l<20; $l++) {
            $r=0.50;
            $c=(($k+$l)%81)/200;
            $color=2.4*$c;
            $pi=3.14159265358979323846264;
            $x=$k+$r*sin($c*$pi*$iter);
            $y=$l+$r*cos($c*$pi*$iter);
```
print OUT "$iter\t\$x\t\$y\t\$color\n";
}
}
}
close(OUT);

Source Code to Produce the Video

Here's the R source code, as discussed earlier in the visualization and subsequent section. The input data is the file movie.txt produced in the previous step. Note that $iter represents the frame number in the video. Here I used 1,000 frames and the video lasts about 60 seconds.

vv<-read.table("c:/vincentg/movie.txt",header=TRUE);
iter<-vv$iter;

for (n in 0:999) {
  x<-vv$x[iter == n];
  y<-vv$y[iter == n];
  z<-vv$z[iter == n];
  plot(x,y,xlim=c(3,17),ylim=c(3,17),pch=20,cex=1,col=rgb(0,0,z),
       xlab="",ylab="",axes=TRUE);
  Sys.sleep(0.01); # sleep 0.05 second between each iteration
}

Saving and Exporting the Video

I used a screen-cast software (Active Presenter, open source) to do the job. It's the same as screen shot software or print screen, except that instead of saving a static image, it saves streaming data.

Securing communications: data encoding

My Note: I did not find this in the Word Document
10 unusual ways analytics are used to make our lives better

Here are a few other applications of data science:

- **Automated patient diagnostic and customized treatment**: Instead of going to the doctor, you fill an online (decision tree type of) questionnaire. At the end of the online medical exam, customized drugs are prescribed, manufactured and delivered to you: For instance, male drug abusers receive a different mix than healthy females, for the same condition. In short, doctors and pharmacists are replaced by robots and algorithms.

- **Movie recommendations for family members**: Allows parents to select age, gender, language, and viewing history to optimize recommendations for individual family members.

- **Scoring technology**: Computation of your FICO score and algorithms to determine the chance for an Internet click (in a rare bucket of clicks lacking historical data) to convert or for a patient to recover—based on analytical scores with confidence intervals. Issues: score standardization, score blending, and score consistency over time and across clients.

- Detection of fake book reviews (Amazon.com) and fake restaurant reviews (Zagat).

- Detection of plagiarism, spam, as well as people sharing paid accounts and pirated movies with friends and colleagues.

- Automated sentencing for common crimes, using a crime score based on simple factors. This eliminates some lawyer and court costs.

- **Testing athletes for fraud**: To boost performance, some athletes now keep blood samples from themselves in the freezer, and inject it back into their system when participating in a competition, to boost red cells count and performance while making detection impossible.

- Detecting election fraud.

- **Semi-automated car driving**: Software to guess user behavior, avoiding collisions by having automated pilot to bypass human driving as needed. And also providing directions when you are driving in a place where GPS does not work (due to inability to connect with navigation satellites, typically in remote areas).

- When you obtain a mortgage, immediately sense if the calculated monthly payment makes sense. It happened to me: they mixed up 30 and 15 years (that was their explanation for the huge error in their favor): I believe it was fraud, and if you are not analytical enough to catch these "errors" right away, you will be financially abused. In this example, you need to mentally compute a mortgage monthly payment given the length and interest rate. Similar arguments apply to all financial products that you purchase.

5. Improving Visuals

Visualization is not just about images, charts, and videos. Other important types of visuals are schemas or diagrams, more generally referred to as infographics, which can be produced with tools such as Visio. But an important part of any visualization is creating dashboards that are easy to read and efficient in terms of communicating insights that lead to decisions. Open source tools such as Birt can be used for this purpose.

To demonstrate a couple of ways to make your visualizations user-friendly, let’s consider improving diagrams (using a Venn diagram as our example) and adding several new dimensions to a chart.

The three Vs of big data (volume, velocity, and variety) or the three skills that make a data scientist (hacking, statistics, and domain expertise) are typically visualized using a Venn diagram, representing all eight of the potential combinations through set intersections. In the case of big data, visualization, veracity, and value are more important than volume, velocity, variety, but that's another issue. Except that one of the Vs is visualization and all these Venn diagrams are visually wrong: The color at the intersection of two sets should be the blending of both colors of the parent sets, for easy
interpretation and easy generalization to four or more sets. For instance, if I have three sets A, B, and C painted respectively in red, green, and blue, the intersection of A and B should be yellow, and the intersection of the three should be white, as shown in the figure below.

![Venn Diagram](https://semanticscommunity.info/Data_Science/Data_Science_Central)

**Improving Venn Diagrams**

If you want to represent three sets, you need to choose three base colors. The colors for the intersections of the sets will be automatically computed using the color addition rule. It makes sense to use red, green, and blue as the base colors for two reasons:

- It maximizes the minimum distance between any two of the eight colors in the Venn diagram, making interpretation quick and easy. (I assume background color is black.)
- It is easy for the eye to reconstruct the proportion of red, green, blue in any color, making interpretation easier. Note that choosing red, green, and yellow as the three base colors would be bad because the intersection of red and green is yellow, which is the color of the third set.

Actually, you don’t even need to use Venn diagrams when using this 3-color scheme. Instead you can use eight non-overlapping rectangles, with the size of each rectangle representing the number of observations in each set/subset.

When you have four sets, and assuming the intensity for each red/green/blue component is a number between 0 and 1 (as in the rgb function in the R language), a good set of base colors is \{(0.5,0,0), (0,0.5,0), (0,0,0.5), (0.5,0.5,0.5)\}. This set corresponds to dark red, dark green, dark blue, and gray.

If you have 5 or more sets, it is better to use a table rather than a diagram, although you can find interesting but very intricate (and difficult to read) Venn diagrams on Google.

If you are not familiar with how colors blend, try this: Use your favorite graphics editor to create a rectangle filled in with yellow. Next to this rectangle, create another rectangle filled with pixels that alternate between red and green. This latter rectangle will appear yellow to your eyes, although maybe not the exact same yellow as in the first rectangle. However, if you fine-tune the brightness of your screen, you might get the two rectangles to display the exact same yellow.

This brings up an interesting fact: The eye can easily distinguish between two almost identical colors in two adjacent rectangles. But it cannot distinguish more pronounced differences if the change is not abrupt, but rather occurs via a
Gradient of colors. Great visualization should exploit features that the eye and brain can process easily and avoid signals that are too subtle for an average eye/brain to quickly detect.

**Exercise:** Can you look at the colors of all objects in the room and easily detect the red/green/blue components? It's a skill that can easily be learned. Should decision makers (who spend time understanding visualizations produced by their data science team) learn this skill? It could improve decision making. Also being able to quickly interpret maps with color gradients (in particular the famous rainbow gradient) is a great skill to have for data insight discovery.

### Adding Dimensions to a Chart

Typically, chart colors are represented by three dimensions: intensity in the red, green, and blue channels. You might be tempted to add a metallic aspect and fluorescence to your chart to provide two extra dimensions. But in reality that will make your chart more complex without adding value. Plus, it's difficult to render a metallic color on a computer screen.

For time series, producing a video rather than a chart automatically and easily adds the time dimension (as in the previous shooting stars video). You could also add two extra dimensions with sound: volume (for example, to represent the number of dots at a given time in a given frame in the video) and frequency (for example, to represent entropy). But these are summary statistics attached to each frame, and it's probably better to represent them by moving bars in the video, rather than using sound.

You could have a video where, each time you move the cursor, a different sound (attached to each pixel) is produced, but it's getting too complicated and the best solution in this case is to have two videos or two images showing two different sets of metrics, rather than trying to stuff all the dimensions in just one document.

For most people, the brain has a hard time quickly processing more than four dimensions at once, and this should be kept in mind when producing visualizations. Beyond five dimensions, any additional dimension probably makes your visual less and less useful for value extraction.

### 6. Essential Features for any Database, SQL or NoSQL

Are you interested in improving database technology? Here are suggestions for better products. This is at the core of data science research.

- Offering text standardization function to help with data cleaning, reducing volume of text information, and merging data from different sources or having different character sets.
- Ability to categorize information (text in particular, and tagged data more generally), using built-in or ad hoc taxonomies (with a customized number of categories and subcategories), together with clustering algorithms. A data record can be assigned to multiple categories.
- Ability to efficiently store images, books, and records with high variance in length, possibly though an auxiliary file management system and data compression algorithms, accessible from the database.
- Ability to process data remotely on your local machine, or externally, especially computer-intensive computations performed on a small number of records. Also, optimizing use of cache systems for faster retrieval.
- Offering SQL to NoSQL translation and SQL code beautifier.
- Offering great visuals (integration with Tableau?) and useful, efficient dashboards (integration with Birt?).

---

http://semanticommunity.info/Data_Science/Data_Science_Central

Updated: Sat, 19 Sep 2015 08:46:35 GMT

Powered by mindtouch
• API and Web/browser access: Database calls can be made with HTTPS requests, with parameters (argument and value) embedded in the query string attached to the URL. Also, allowing recovery of extracted/processed data on a server, via HTTPS calls, automatically. This is a solution to allow for machine-to-machine communications.

• DBA tools available to sophisticated users, such as fine-tuning query optimization algorithms, allowing hash joins, switching from row to column database when needed (to optimize data retrieval in specific cases).

• Real-time transaction processing and built-in functions such as computations of “time elapsed since last 5 transactions” at various levels of aggregation.

• Ability to automatically erase old records and keep only historical summaries (aggregates) for data older than, for example, 12 months.

• Security (TBD)

Note that some database systems are different from traditional DB architecture. For instance, I created a web app to find keywords related to keywords specified by a user. This system (it's more like a student project than a real app; though you can test it here) has the following features:

• It's based on a file management system (no actual database).

• It is a table with several million entries, each entry being a keyword and a related keyword, plus metrics that measure the quality of the match (how strong the relationship between the two keywords is), as well as frequencies attached to these two keywords, and when they are jointly found. The function to measure the quality of the match can be customized by the user.

• The table is split into 27 x 27 small text files. For instance, the file cu_keywords.txt contains all entries for keywords starting with letter cu. (It's a bit more complicated than that, but this shows you how I replicated the indexing capabilities of modern databases.)

• It's running on a shared server; at it peaks hundreds of users were using it, and the time to get an answer (retrieve keyword related data) for each query was less than 0.5 second—most of that time spent on transferring data over the Internet, with little time used to extract the data on the server.

It offers API and web access (and indeed, no other types of access).

91 job interview questions for data scientists


• Posted by Vincent Granville on February 13, 2013 at 8:00pm

• View Blog

We are now at 86 questions. These are mostly open-ended questions, to assess the technical horizontal knowledge of a senior candidate for a rather high level position, e.g. director.

1. What is the biggest data set that you processed, and how did you process it, what were the results?

2. Tell me two success stories about your analytic or computer science projects? How was lift (or success) measured?

3. What is: lift, KPI, robustness, model fitting, design of experiments, 80/20 rule?

4. What is: collaborative filtering, n-grams, map reduce, cosine distance?

5. How to optimize a web crawler to run much faster, extract better information, and better summarize data to produce cleaner databases?
6. How would you come up with a solution to identify plagiarism?
7. How to detect individual paid accounts shared by multiple users?
8. Should click data be handled in real time? Why? In which contexts?
9. What is better: good data or good models? And how do you define "good"? Is there a universal good model? Are there any models that are definitely not so good?
10. What is probabilistic merging (AKA fuzzy merging)? Is it easier to handle with SQL or other languages? Which languages would you choose for semi-structured text data reconciliation?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is your favorite programming language / vendor? Why?
13. Tell me 3 things positive and 3 things negative about your favorite statistical software.
15. What is the curse of big data?
16. Have you been involved in database design and data modeling?
17. Have you been involved in dashboard creation and metric selection? What do you think about Birt?
18. What features of Teradata do you like?
19. You are about to send one million email (marketing campaign). How do you optimize delivery? How do you optimize response? Can you optimize both separately? (answer: not really)
20. Toad or Brio or any other similar clients are quite inefficient to query Oracle databases. Why? How would you do to increase speed by a factor 10, and be able to handle far bigger outputs?
21. How would you turn unstructured data into structured data? Is it really necessary? Is it OK to store data as flat text files rather than in an SQL-powered RDBMS?
22. What are hash table collisions? How is it avoided? How frequently does it happen?
23. How to make sure a mapreduce application has good load balance? What is load balance?
24. Examples where mapreduce does not work? Examples where it works very well? What are the security issues involved with the cloud? What do you think of EMC's solution offering an hybrid approach - both internal and external cloud - to mitigate the risks and offer other advantages (which ones)?
25. Is it better to have 100 small hash tables or one big hash table, in memory, in terms of access speed (assuming both fit within RAM)? What do you think about in-database analytics?
26. Why is naive Bayes so bad? How would you improve a spam detection algorithm that uses naive Bayes?
27. Have you been working with white lists? Positive rules? (In the context of fraud or spam detection)
28. What is star schema? Lookup tables?
29. Can you perform logistic regression with Excel? (yes) How? (use linest on log-transformed data)? Would the result be good? (Excel has numerical issues, but it's very interactive)
30. Have you optimized code or algorithms for speed: in SQL, Perl, C++, Python etc. How, and by how much?
31. Is it better to spend 5 days developing a 90% accurate solution, or 10 days for 100% accuracy? Depends on the context?
32. Define: quality assurance, six sigma, design of experiments. Give examples of good and bad designs of experiments.
33. What are the drawbacks of general linear model? Are you familiar with alternatives (Lasso, ridge regression, boosted trees)?
34. Do you think 50 small decision trees are better than a large one? Why?
35. Is actuarial science not a branch of statistics (survival analysis)? If not, how so?

36. Give examples of data that does not have a Gaussian distribution, nor log-normal. Give examples of data that has a very chaotic distribution?

37. Why is mean square error a bad measure of model performance? What would you suggest instead?

38. How can you prove that one improvement you've brought to an algorithm is really an improvement over not doing anything? Are you familiar with A/B testing?

39. What is sensitivity analysis? Is it better to have low sensitivity (that is, great robustness) and low predictive power, or the other way around? How to perform good cross-validation? What do you think about the idea of injecting noise in your data set to test the sensitivity of your models?

40. Compare logistic regression w. decision trees, neural networks. How have these technologies been vastly improved over the last 15 years?

41. Do you know / used data reduction techniques other than PCA? What do you think of step-wise regression? What kind of step-wise techniques are you familiar with? When is full data better than reduced data or sample?

42. How would you build non parametric confidence intervals, e.g. for scores? (see the AnalyticBridge theorem)

43. Are you familiar either with extreme value theory, monte carlo simulations or mathematical statistics (or anything else) to correctly estimate the chance of a very rare event?

44. What is root cause analysis? How to identify a cause vs. a correlation? Give examples.

45. How would you define and measure the predictive power of a metric?

46. How to detect the best rule set for a fraud detection scoring technology? How do you deal with rule redundancy, rule discovery, and the combinatorial nature of the problem (for finding optimum rule set - the one with best predictive power)? Can an approximate solution to the rule set problem be OK? How would you find an OK approximate solution? How would you decide it is good enough and stop looking for a better one?

47. How to create a keyword taxonomy?

48. What is a Botnet? How can it be detected?

49. Any experience with using API's? Programming API's? Google or Amazon API's? AaaS (Analytics as a service)?

50. When is it better to write your own code than using a data science software package?

51. Which tools do you use for visualization? What do you think of Tableau? R? SAS? (for graphs). How to efficiently represent 5 dimension in a chart (or in a video)?

52. What is POC (proof of concept)?

53. What types of clients have you been working with: internal, external, sales / finance / marketing / IT people? Consulting experience? Dealing with vendors, including vendor selection and testing?

54. Are you familiar with software life cycle? With IT project life cycle - from gathering requests to maintenance?

55. What is a cron job?

56. Are you a lone coder? A production guy (developer)? Or a designer (architect)?

57. Is it better to have too many false positives, or too many false negatives?

58. Are you familiar with pricing optimization, price elasticity, inventory management, competitive intelligence? Give examples.

59. How does Zillow's algorithm work? (to estimate the value of any home in US)

60. How to detect bogus reviews, or bogus Facebook accounts used for bad purposes?

61. How would you create a new anonymous digital currency?

62. Have you ever thought about creating a startup? Around which idea / concept?

63. Do you think that typed login / password will disappear? How could they be replaced?
64. Have you used time series models? Cross-correlations with time lags? Correlograms? Spectral analysis? Signal processing and filtering techniques? In which context?

65. Which data scientists do you admire most? Which startups?

66. How did you become interested in data science?

67. What is an efficiency curve? What are its drawbacks, and how can they be overcome?

68. What is a recommendation engine? How does it work?

69. What is an exact test? How and when can simulations help us when we do not use an exact test?

70. What do you think makes a good data scientist?

71. Do you think data science is an art or a science?

72. What is the computational complexity of a good, fast clustering algorithm? What is a good clustering algorithm? How do you determine the number of clusters? How would you perform clustering on one million unique keywords, assuming you have 10 million data points - each one consisting of two keywords, and a metric measuring how similar these two keywords are? How would you create this 10 million data points table in the first place?

73. Give a few examples of "best practices" in data science.

74. What could make a chart misleading, difficult to read or interpret? What features should a useful chart have?

75. Do you know a few "rules of thumb" used in statistical or computer science? Or in business analytics?

76. What are your top 5 predictions for the next 20 years?

77. How do you immediately know when statistics published in an article (e.g. newspaper) are either wrong or presented to support the author's point of view, rather than correct, comprehensive factual information on a specific subject? For instance, what do you think about the official monthly unemployment statistics regularly discussed in the press? What could make them more accurate?

78. Testing your analytic intuition: look at these three charts. Two of them exhibit patterns. Which ones? Do you know that these charts are called scatter-plots? Are there other ways to visually represent this type of data?

79. You design a robust non-parametric statistic (metric) to replace correlation or R square, that (1) is independent of sample size, (2) always between -1 and +1, and (3) based on rank statistics. How do you normalize for sample size? Write an algorithm that computes all permutations of n elements. How do you sample permutations (that is, generate tons of random permutations) when n is large, to estimate the asymptotic distribution for your newly created metric? You may use this asymptotic distribution for normalizing your metric. Do you think that an exact theoretical distribution might exist, and therefore, we should find it, and use it rather than wasting our time trying to estimate the asymptotic distribution using simulations?

80. More difficult, technical question related to previous one. There is an obvious one-to-one correspondence between permutations of n elements and integers between 1 and n! Design an algorithm that encodes an integer less than n! as a permutation of n elements. What would be the reverse algorithm, used to decode a permutation and transform it back into a number? Hint: An intermediate step is to use the factorial number system representation of an integer. Feel free to check this reference online to answer the question. Even better, feel free to browse the web to find the full answer to the question (this will test the candidate's ability to quickly search online and find a solution to a problem without spending hours reinventing the wheel).

81. How many "useful" votes will a Yelp review receive? My answer: Eliminate bogus accounts (read this article), or competitor reviews (how to detect them: use taxonomy to classify users, and location - two Italian restaurants in same Zip code could badmouth each other and write great comments for themselves). Detect fake likes: some companies (e.g. FanMeNow.com) will charge you to produce fake accounts and fake likes. Eliminate prolific users who like everything, those who hate everything. Have a blacklist of keywords to filter fake reviews. See if IP address or IP block of reviewer is in a blacklist such as "Stop Forum Spam". Create honeypot to catch fraudsters. Also watch out for disgruntled employees badmouthing their former employer. Watch out for 2 or 3 similar comments posted the same day by 3 users regarding a company that receives very few reviews. Is it a brand new company? Add more weight to trusted users (create a category of trusted users). Flag all reviews that are identical
(or nearly identical) and come from same IP address or same user. Create a metric to measure distance between two pieces of text (reviews). Create a review or reviewer taxonomy. Use hidden decision trees to rate or score review and reviewers.

82. What did you do today? Or what did you do this week / last week?
83. What/when is the latest data mining book / article you read? What/when is the latest data mining conference / webinar / class / workshop / training you attended? What/when is the most recent programming skill that you acquired?
84. What are your favorite data science websites? Who do you admire most in the data science community, and why? Which company do you admire most?
85. What/when/where is the last data science blog post you wrote?
86. In your opinion, what is data science? Machine learning? Data mining?
87. Who are the best people you recruited and where are they today?
88. Can you estimate and forecast sales for any book, based on Amazon public data? Hint: read this article.
89. What's wrong with this picture?
90. Should removing stop words be Step 1 rather than Step 3, in the search engine algorithm described here? Answer: Have you thought about the fact that mine and yours could also be stop words? So in a bad implementation, data mining would become data mine after stemming, then data. In practice, you remove stop words before stemming. So Step 3 should indeed become step 1.
91. Experimental design and a bit of computer science with Lego's

Comment by Vincent Granville on May 5, 2013 at 7:44pm

Someone wrote:

Looks like hiring managers expect data scientists to have expertise in machine learning, statistics, business intelligence, database design, data munging, data visualization and programming. Are not these requirements too excessive?

My answer:

I think being familiar with all these domains (add computer science, map reduce) is necessary, as well as expertise in some of these domains. Mastering two programming languages (Java, Python) is a must, as well as familiarity with R and SQL. Visualization is easy to acquire.

Knowing how to quickly and independently find, learn (or if necessary, invent) and assess usefulness of the techniques needed to handle the problems, is critical, and MORE important than "knowing" the techniques in the first place. A good amount of experience with some techniques is necessary.

But you don't need to be an expert in everything. For instance, about 90% of what I learned in statistics courses, I've never had to use it to solve business problems. So why learn it in the first place? Also, machine learning (in my opinion) is a subset of statistics focusing on clustering, pattern recognition and association rules.

The mistake that many hiring managers do is looking for someone who is an expert in everything.
Comment by Vincent Granville on February 16, 2013 at 7:19am

Here’s a potential answer to question #10 (probabilistic merging). Feel free to add your answers to any of these questions.

**Answer to question #10:**

Not sure if the problem of fuzzy merging can be addressed within the framework of traditional databases. Say you have a table A with 10,000 users (key is user ID), a table B with 50,000 users (key is user ID). You could created a user mapping table C with three fields:

1. userID (= key),
2. Alternate_UserID (this field would also be a user ID) and
3. Probability (probability that userID = Alternate_UserID).

This table would be populated after some machine learning algorithm had been applied to tables A and B to identify similar users and the probability they match. Make sure that you only include (in table C) records where probability is above (say) 0.25, otherwise you risk exploding your database.

Comment by Vincent Granville on February 16, 2013 at 6:54am

Also, my 70 questions focus mostly on the tech aspects of being a data scientist. And these are high level questions, aimed at senior professionals (I think there is no such thing as a junior data scientist - they would be called data analyst, software engineer, statistician or computer scientist instead). I did not include questions about soft skills - that would be another set of 70 questions.

I will add a new one: do you think data science is an art or a science? The answer, as always, is "both". Then you can dig deeper and ask whether you are more of an artist than a scientist. My answer would be: it's more craftsmanship than art, but in my case, being a designer/architect, it's a tiny bit closer to art than to science. Certainly a blend of both.

And when bringing the issue of art vs. Science, I would also add that I like to build solutions that are elegant in the way they contribute to ROI / lift, but not in the way they contribute to statistical theory and the beauty of science. I like a dirty, ugly, imperfect solution better than a "great model" if it is more scalable, simple, efficient, easy to implement and robust.

Comment by Vincent Granville on February 15, 2013 at 10:22am

A data scientist is a bit of everything (statistician, software engineer, business analyst, computer scientist, six sigma, consultant, communicator), but most importantly she is a senior analytic practitioner

- with a very good sense for business data and business optimization at large.
• knowledge of big data - both drawbacks and potential (and able to leverage its potential)
• who enjoys swimming in unstructured data, fuzzy non-SQL "joins"
• who knows the limitation of old statistics (regression etc.) yet knows how to correctly do sampling, cross-validation, Monte Carlo simulations, design of experiments, assessing lift, identify good metrics
• who knows the limitations of MapReduce, and how they can be overcome
• who can design and develop robust, simple, efficient, reliable, scalable, useful predictive algorithms - whether or not based on statistical theory

A data scientist may not know much (but at least a little) about linear regression, statistical distributions, the complexity of the quicksort (sorting) algorithm or the limit theorems. Her knowledge of SQL can be a bit elementary, although she can run a big SQL query 10 times faster than business analysts who use tools such as Toad or Brio. Her strengths, skills and knowledge are briefly outlined above.

Comment by Vincent Granville on February 15, 2013 at 8:24am

Interviewers would pick a small subset, there's not enough time in a one-day interview to ask all these questions. Also, several of these questions are about relevant projects (e.g. questions #1 and #2). Of course, these are not yes/no question, and one would expect to spend 10-15 minutes and go in some depths answering these questions. Not being able to answer one question in no big deal - this set has 70 questions and the interviewer can easily pick another one. Indeed, this is the purpose of my list.

Comment by Vincent Granville on February 14, 2013 at 6:52am

Here's a comment from one of our readers:

Some suggestions for structure you may want to apply to your own list:

* Tools (#13 14 18 etc)
* Algorithms (#26 33 etc)
* Statistics (#35 36 37 etc)
* Techniques (#3 4 10 etc)
* Data Structures (#21 22 25 etc)
* Experience (#1 2 etc)
* Business language (52 54)
* Domain-specific (#5 6 7 8 10 19 20 21 24b 27 46 55 59 and probably others)
* Plain weirdness (#5 48 59 61 63)

It is probably worth thinking about the areas that are important to you and manage a list based on those. I don't think Vincent expects us to just use the list except for inspiration.
My favourites from the list (for senior people) are #2 9 (my answer: "valuable actions are best") and 62. Which ones are your favourites?

By Allan Engelhardt

Data Science Apprenticeship

Source: http://www.datasciencecentral.com/group/data-science-apprenticeship

• Created by Vincent Granville
• View Groups

INFORMATION

Data Science Apprenticeship by leading data scientist and entrepreneur Dr. Vincent Granville. Six-month program, with real life projects. No admission exam, state-of-the art curriculum, comprehensive program delivered entirely online, suitable for part-time learners. Training based on our Data Science eBook (2nd Edition soon to be published) and our training booklet. Perfect for self-learners. Click here for more information, and join this group to receive updates about our apprenticeship (schedule, material to get started, etc.)

If you have already earned a data science certificate or diploma, but was not requested to develop and use your own API in batch mode, and harvest/work on a data set with at least 50 million observations in a distributed environment, then what you learned is fake data science. It's time to learn the real stuff that will land you a real job!

Click here for a general overview of our apprenticeship. We have published the data and source code for our big data keyword correlation API. Read the material and download the three files (and post your comments if you have questions, I'll reply ASAP): it will teach you how API's work, and how to write your first API from scratch!
Our next API example will come with the source code of a web crawler, and will illustrate how to detect copyright infringement or how to detect the original, first version of an article published in multiple news outlets (doing a better job than Google).

All the training material will be offered for free to everyone. We have not yet put everything into a nice booklet, but some of the content is already available:

---

### Data Sets for Download

A few data sets available for download, from the following articles:

- [Source code for our Big Data keyword correlation API](#)
- [Great statistical analysis: forecasting meteorite hits](#) My Note: This data set interested me.
- Fast clustering algorithms for massive datasets
- 53.5 billion clicks dataset available for benchmarking and testing
- Over 5,000,000 financial, economic and social datasets*
- New pattern to predict stock prices, multiplies return by factor 5*
- 3.5 billion web pages*
- Another large data set - 250 million data points - available for do...* 

---

### Articles for Curriculum

The following articles will be included in our curriculum, so you can start reading them now

- [Fast clustering algorithms for massive datasets](#) *
- [Fast Combinatorial Feature Selection with New Definition of Predict...](#)
- [Internet Topology Mapping](#) *
- [Correlation and R-Squared for Big Data](#) *
- [From chaos to clusters - statistical modeling without models](#) *
- [Example of architecture for AaaS (Analytics as a Service)](#)
- Other source code posted on Analyticbridge*
- Simple solutions to make videos with R*
- What Map Reduce can't do
- The curse of big data
- Correlation vs. causation
- Three classes of metrics: centrality, volatility, and bumpiness
- Data Science dictionary
- Data Science tools
- A new random number generator
- 66 job interview questions for data scientists

---

http://semanticommunity.info/Data_Science/Data_Science_Central
Updated: Sat, 19 Sep 2015 08:46:35 GMT
Powered by mindtouch
List of potential projects for students:

• Click here for some suggestions

Starred items (*) are recent additions.

We are still in the process of writing our small booklet to teach you all the fundamentals (computer science, statistics, business analytics, Python, Map Reduce, big data etc.) in 20 pages. Also, we will publish our second edition of our Data Science eBook (required reading for the apprenticeship) this month. It will contain about twice as many articles as the current version. All these articles have been written, I just have to put them together in a nice PDF document. Click here to see a preview of the Data Science eBook - 2nd Edition.

---

Data Science eBook by Analyticbridge - 2nd Edition


• Posted by Vincent Granville on May 19, 2013 at 5:00pm
• View Blog

This 2nd edition has more than 200 pages of pure data science, far more than the first edition. This new version of our very popular book will soon be available for download: we will make an announcement when it is officially published. My Note: I know I looked at the first edition.

New contributions:

• Articles 27 to 45 in Part I
• Articles 26 to 52 in Part II
• Articles 8-23 in Part III
Follow the links below, and enjoy the reading!

Introduction

Part I - Data Science Recipes

1. New random number generator: simple, strong and fast
2. Lifetime value of an e-mail blast: much longer than you think
3. Two great ideas to create a much better search engine
4. Identifying the number of clusters: finally a solution
5. Online advertising: a solution to optimize ad relevancy
6. Example of architecture for AaaS (Analytics as a Service)
7. Why and how to build a data dictionary for big data sets
8. Hidden decision trees: a modern scoring methodology
9. Scorecards: Logistic, Ridge and Logic Regression
10. Iterative Algorithm for Linear Regression
11. Approximate Solutions to Linear Regression Problems
12. Theorems for Traders
13. Preserving metric and score consistency over time and across clients
14. Advertising: reach and frequency mathematical formulas
15. Real Life Example of Text Mining to Detect Fraudulent Buyers
16. Discount optimization problem in retail analytics
17. Sales forecasts: how to improve accuracy while simplifying models?
18. How could Amazon increase sales by redefining relevancy?
19. How to build simple, accurate, data-driven, model-free confidence i...
20. Comprehensive list of Excel errors, inaccuracies and use of non-sta...
21. 10+ Great Metrics and Strategies for Email Campaign Optimization
22. 10+ Great Metrics and Strategies for Fraud Detection
23. Case Study: Four different ways to solve a data science problem
24. Case Study: Email marketing - analytic tips to boost performance b...
25. Optimize keyword campaigns on Google in 7 days: an 11-step procedure
26. How do you estimate the proportion of bogus accounts on Facebook?
27. Stat models to solve astronomical mysteries - application to busine...
28. How to detect a pattern? Problem and solution
29. From chaos to clusters - statistical modeling without models
30. Simple solutions to make videos with R
31. Three classes of metrics: centrality, volatility, and bumpiness
32. How to optimize email campaigns? Part I
33. Simple steps to increase speed of web crawling by a factor 80,000
34. How are database joins optimized? How can you do better to handle b...
35. Correlation vs. causation
36. Seven tricky sentences for NLP and text mining algorithms
37. Great statistical analysis: forecasting meteorite hits
38. Fast clustering algorithms for massive datasets
39. How are hotel room rates determined
40. The curse of big data
41. Simple technique to improve poor predictive models
42. Simple source code to simulate nice cluster structures
43. Source code for our Big Data keyword correlation API
44. Correlation vs. causation
45. Shootings stars: producing videos about data

Part II - Data Science Discussions
1. Statisticians Have Large Role to Play in Web Analytics (AMSTAT inte...
2. Future of Web Analytics: Interview with Dr. Vincent Granville
3. Connecting with the Social Analytics Experts
4. Interesting note and questions on mathematical patents
5. Big data versus smart data: who will win?
6. Creativity vs. Analytics: Are These Two Skills Incompatible?
7. Barriers to hiring analytic people
8. Salary report for selected analytical job titles
9. Are we detailed-oriented or do we think "big picture", or both?
10. Why you should stay away from the stock market
11. Gartner Executive Programs' Worldwide Survey of More Than 2,300 CIOs
12. 4.4 Million New IT Jobs Globally to Support Big Data By 2015
13. One Third of Organizations Plan to Use Cloud Offerings to Augment BI Capabilities
14. Twenty Questions about Big Data and Data Sciences
15. Interview with Drew Rockwell, CEO of Lavastorm
16. Can we use data science to measure distances to stars?
17. Eighteen questions about real time analytics
18. Can any data structure be represented by one-dimensional arrays?
19. Data visualization: example of a great, interactive chart
20. Data science jobs not requiring human interactions
21. Featured Data Scientist: Vincent Granville, Analytic Entrepreneur
22. Healthcare fraud detection still uses cave-man data mining techniques
23. Why are spam detection algorithms so terrible?
24. What is a Data Scientist?
25. Twenty seven types of data scientists: where do you fit?
26. Seven tricky sentences for NLP and text mining algorithms
27. How maths should be taught in high school
28. An alternative to FICO scores?
30. Vertical vs. Horizontal Data Scientists
31. 14 questions about data visualization tools
32. New, fast Excel to process billions of rows via the cloud
33. When data flows faster than it can be processed
34. Car accident statistics by profession
35. Extreme Data Science
36. Six keywords characterizing milestones in the history of analytic e...
37. A new idea for an analytic business startup
38. Four innovative ideas to optimize business processes
39. Vincent's answers to data science questions - Part 2
40. How do data scientists rank?
41. History, Evolution and Classification of Programming Languages
42. Automated news feed optimization
43. Shopper Alert: Price May Drop for You Alone
44. Why are clinical trials failing?
46. The 8 worst predictive modeling techniques
47. What are the most difficult things to predict?
48. Fake Data Science
49. Analytics{Benzene} => {big Pharma, Nanotechnologies}
50. What MapReduce can't do
51. Big Data startup to fix healthcare
52. How to reverse-engineer Google?

Part III - Data Science Resources

1. Vincent’s list
2. History of 24 analytic companies over the last 30 years
3. Fifteen great data science articles from influential news outlets
4. List of publicly traded analytic companies
5. Thirty unusual applications of data sciences, analytics and big data
6. 50 unusual ways analytics are used to make our lives better
7. Berkeley course on Data Science
8. Over 5,000,000 financial, economic and social datasets
Proposal for an Apprenticeship in Data Science


- Posted by Vincent Granville on October 25, 2012 at 4:00pm
- View Blog

For motivated students who can learn on their own, here’s an option that I would like to offer: the possibility to become an expert data scientist in less than six months, for a cost well below $10,000, and with guaranteed job opportunities.

The program would be open to everyone without screening, but the degree and the guaranteed jobs would be offered only to students with a successful completion of selected projects. If you don’t succeed, you don’t pay.

The program would contain three parts:

Part I: Online training

20-pages booklet containing all the info you need to jump-start your data science career, written in simple English:

- how to download Python, Perl, Java, R, get sample programs, get started with writing an efficient web crawler, get started with Linux, Cygwin, Excel (including logistic regression)
- Hadoop, MapReduce, NoSQL their limitations and more modern technologies
- how to find data sets or download very large data sets for free on the web
- how to analyze data: from understanding business requirements to maintaining an automated (machine talking to machine) web / database application in production mode - a 12 steps process
- how to develop your first “Analytics as a Service” application and scale it
• big data algorithms, and how to make them more efficient and more robust (application in computational optimization: how to efficiently test trillions of trillions of multivariate vectors to design good scores)

• basics about statistics, monte-carlo, cross-validation, robustness, sampling, design of experiments

• tons of startup ideas for analytic people

• reference data science book available for free (click here to see 2nd Edition)

• basics of Perl, Python, real time analytics, distributed architecture and general programming practices

• data visualization, dashboards and how to communicate like a management consultant

• tips for future consultants

• tips for future entrepreneurs

• rules of thumb, best practices, craftsmanship secrets, and why is data science an art?

• additional online resources

• lift and other metrics to measure success, metrics selection, use of external data, make data silos communicate via fuzzy merging and statistical techniques

---

Part II: Potential projects to be completed:

• hacking and reverse-engineering projects, for instance a captcha attack

• web crawling projects: how many Facebook accounts are duplicate or dead? Or categorize Tweets

• taxonomy creation or improving an existing taxonomy

• optimal pricing for bid keywords on Google

• create a web app that provide (in real time) better-than-average trading signals

• find low-frequency and Botnet fraud cases in a sea of data

• internship in computational marketing with a data science start-up

• automated plagiarism detection

• use web crawlers, assess whether Google Search favors (1) its products over competitors [is this an unfair business practice?], (2) local over non-local results and (3) returns different results to web robots and humans. Identify other bias and patterns in Google search results.

---

Part III: Students successfully completing two projects

• would be featured in the largest data science community

• would receive help finding a job or advice about jump-starting their own company

• would get endorsement from a leading data scientist

• may be hired by sponsor companies funding this project

---

How to enroll?

If interested, join our Data Science Apprenticeship group to receive updates about our program and schedule, and to receive an invitation to participate as well as free training material, when the program is open.
Related articles

- 66 job interview questions for data scientists
- Fake data science
- The curse of big data
- What Map Reduce can’t do
- 53.5 billion clicks dataset available for benchmarking and testing
- Eight worst predictive modeling techniques
- Another example of misuse of statistical science
- The curse of dimensionality (it got worse with big data)
- Data Science eBook
- Data Science Apprenticeship
- Debunking lack of analytic talent
- Causation vs. Correlation
- AnalyticTalent.com
- Data Science dictionary
- How and why to build a data dictionary
- Data Science tools
- A new random number generator
- Modern books on multiple programming languages
- Assessing efficiency of approximate vs. exact algorithms (coming soon)
- Statistical comic strip
- Most popular blog posts
- $33,000 to get an outdated Applied Maths degree, offering no job pr...
- Free courses from top universities | Coursera.com
- Has a degree in mathematics become worthless? | IEEE