

# Introduction to Data Science

## Lecture 1

INSTRUCTOR

**MICHAEL CASABUENA**

# Course Description

- This course provides an introduction to the field of data science. Students will learn the fundamentals of data science, including the principles of data analysis, data visualization, and data management. Students will explore the different types of data that can be collected, the different types of data analysis techniques, and the different types of data visualization tools. Students will learn how to use these techniques to solve real-world data science problems. In addition, students will be introduced to the python programming language that is used in data science and will be able to create their own data visualizations. By the end of the course, students will be able to analyze, visualize, and manage data to answer meaningful questions.

# Course Objective

- Write programs in Python
- Collect, Manipulate, Blend Data from Different Data Sources
- Visualize Data and Perform Exploratory Data Analysis
- Understand Data Science Project Lifecycle

# Course Objective

- Understand what the goals and objectives of machine learning are
- Understand how to evaluate models generated from data.
- Apply machine learning algorithms to build real-world systems
- Develop an appreciation for what is involved in learning models from data

# Outline

- Data Science – Why all the excitement?
  - examples
- Where does data come from
- So what is Data Science
- Doing Data Science
- About the course
  - what we'll cover
  - data science first, big data later
  - requirements, workload etc.

# Data Analysis Has Been Around for a While

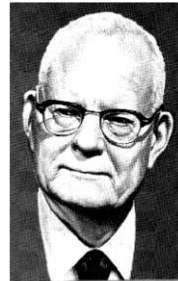
1935: “The Design of Experiments”

R.A. Fisher



W.E.  
Demming

1939: “Quality Control”



1958: “A Business Intelligence System”

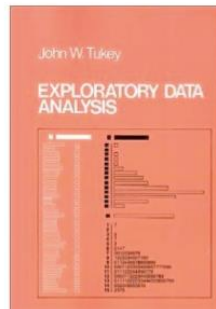


Peter Luhn

1997: “Machine Learning”



1977: “Exploratory Data Analysis”



1989: “Business Intelligence”

Howard  
Dresner

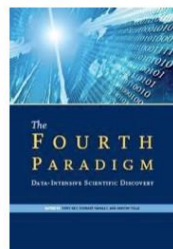


2010: “The Data Deluge”

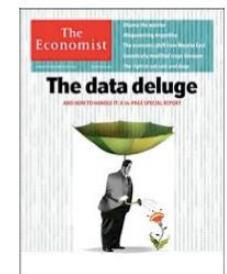
1996: Google



2007: “The Fourth Paradigm”



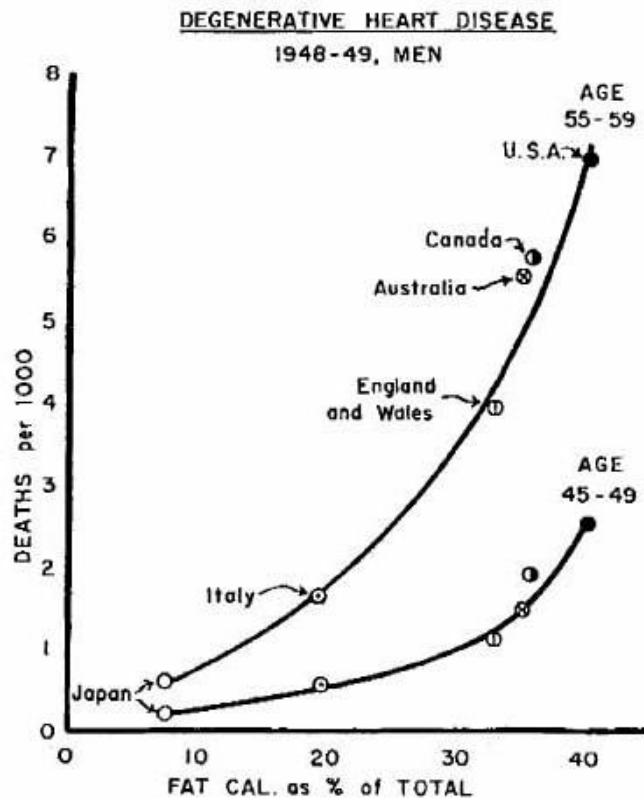
2009: “The Unreasonable Effectiveness of Data”



Abridged Version of Jeff Hammerbacher’s timeline for CS 194, 2012

# Data makes everything clearer

- Seven Countries Study (Ancel Keys, UCB 1925,28)
- 13,000 subjects total, 5-40 years follow-up.





# Data Science: Why all the Excitement?



e.g.,  
Google Flu Trends:

Detecting outbreaks  
two weeks ahead  
of CDC data

New models are estimating  
which cities are most at risk  
for spread of the Ebola virus.

# Why the all the Excitement?

**elections2012**

Live results | [President](#) | [Senate](#) | [House](#) | [Governor](#) | [Choose your](#)

## Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

**Luke Harding**

[guardian.co.uk](http://guardian.co.uk), Wednesday 7 November 2012 10.45 EST



*the signal and the  
and the noise and  
the noise and the  
noise and the no  
why most noise a  
predictions fail to  
but some don't n  
and the noise and  
the noise and the  
nate silver noise  
noise and the no*

# Data and Election 2012 (cont.)

- ...that was just one of several ways that Mr. Obama's campaign operations, some unnoticed by Mr. Romney's aides in Boston, **helped save the president's candidacy**. In Chicago, the campaign recruited a team of behavioral scientists to build an **extraordinarily sophisticated database**

...that allowed the Obama campaign not only to alter the very nature of the electorate, making it younger and less white, but also to create a portrait of shifting voter allegiances. **The power of this operation stunned Mr. Romney's aides on election night**, as they saw voters they never even knew existed turn out in places like Osceola County, Fla.

New York Times, Wed Nov 7, 2012

# A history of the (Business) Internet: 1997


BackRub Search: university

## BackRub Query Results

BackRub's Highest Ranked Sites


---

University of Illinois at Urbana-Champaign

 <http://www.uiuc.edu/>


694.687 8460 backlinks 12k - 10/25/96 - 11/1/96

Stanford University Homepage

 <http://www.stanford.edu/>

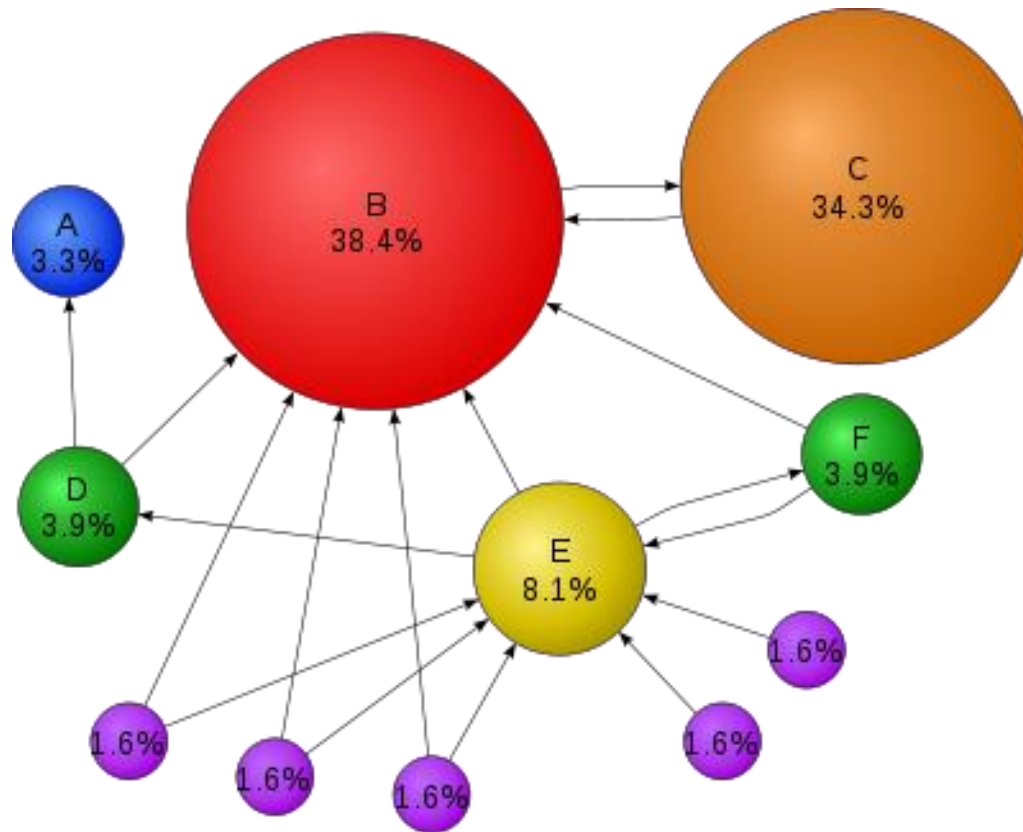
609.303 8857 backlinks 4k - none - 11/1/96

Stanford University: Portfolio Collection

 <http://www.stanford.edu/home/administration/portfolio.html>

167.919 34 backlinks

# Pagerank: The web as a behavioral dataset



# DB size = 50 billion sites



Google server farms  
2 million machines (est)

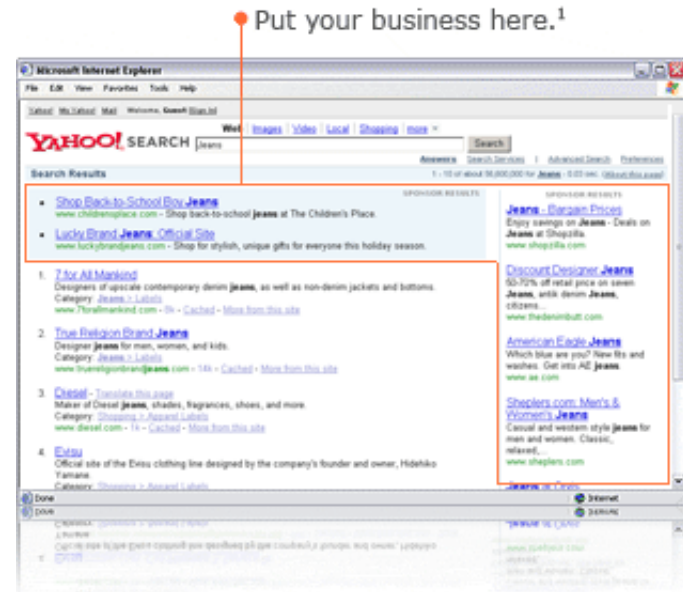
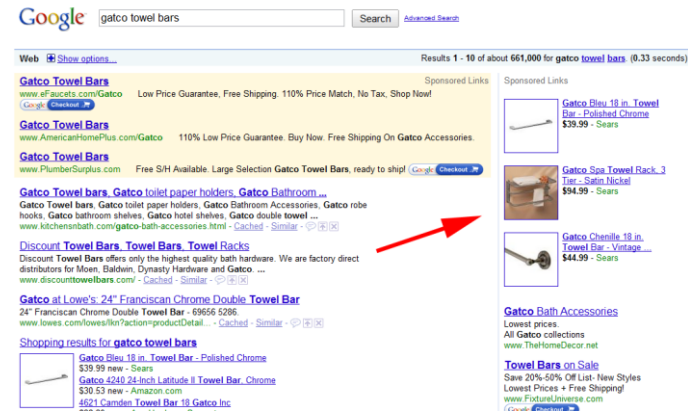




# 1998 – sponsored search



Overture



200

2

# Sponsored search

- Google revenue around \$50 bn/year from marketing, 97% of the companies revenue.
- Sponsored search uses an auction – a pure competition for marketers trying to win access to consumers.
- In other words, a competition for **models** of consumers – their likelihood of responding to the ad – and of determining the right bid for the item.
- There are around 30 billion search requests a month. Perhaps a **trillion events** of history between search providers.

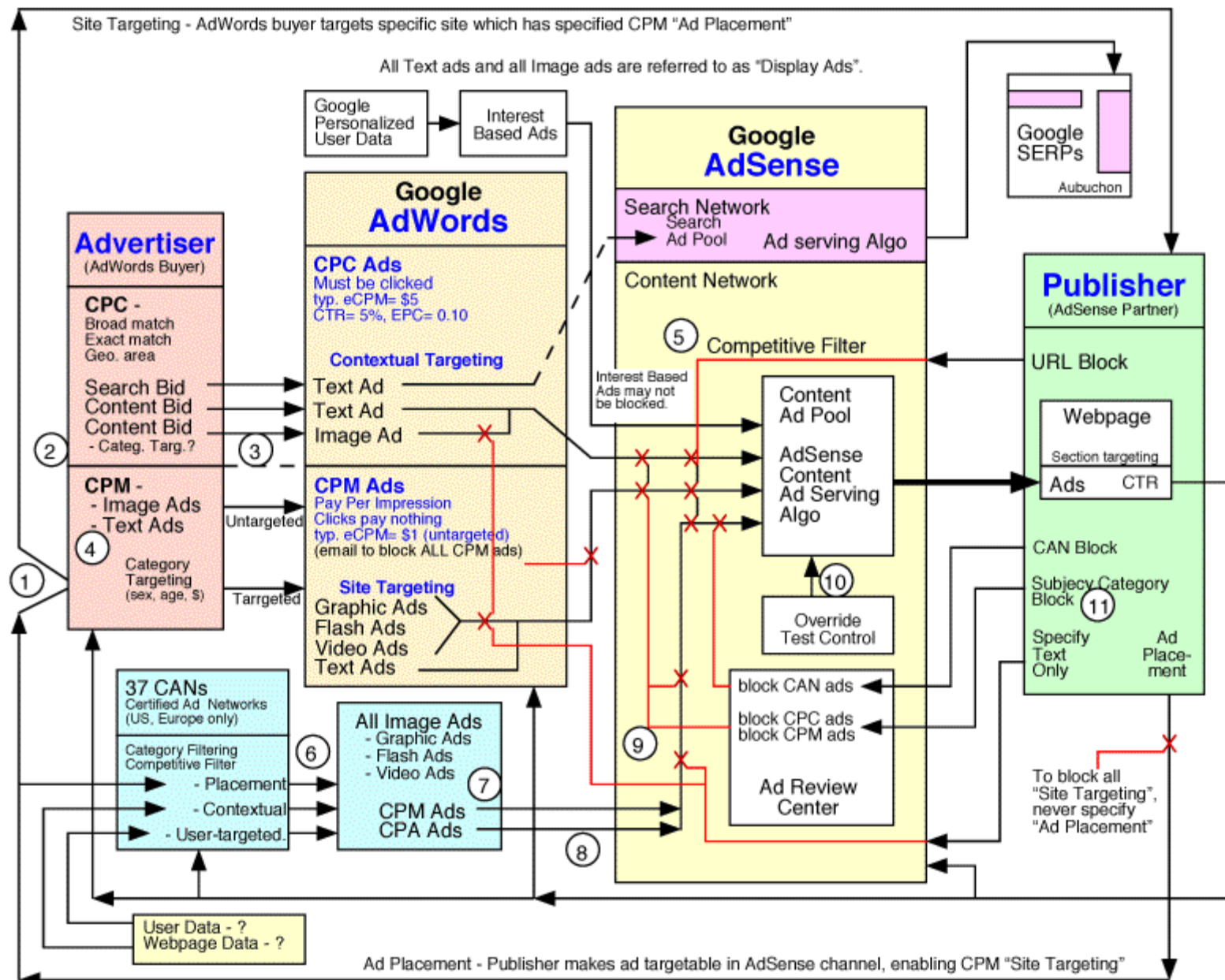


# TOP 20

## Keyword Categories

Percentages correspond to the number of keywords in the top 10,000 keywords that belong to that category.





# Data Makes Everything Clearer?

## Epidemiological modeling of online social network dynamics

John Cannarella<sup>1</sup>, Joshua A. Spechler<sup>1,\*</sup>

<sup>1</sup> Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA

\* E-mail: [Corresponding\\_spechler@princeton.edu](mailto:Corresponding_spechler@princeton.edu)

## Abstract

The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional SIR model of disease spread by incorporating infectious recovery dynamics such that contact between a recovered and infected member of the population is required for recovery. The proposed infectious recovery SIR model (irSIR model) is validated using publicly available Google search query data for “MySpace” as a case study of an OSN that has exhibited both adoption and abandonment phases. The irSIR model is then applied to search query data for “Facebook,” which is just beginning to show the onset of an abandonment phase. Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.

# Data Makes Everything Clearer

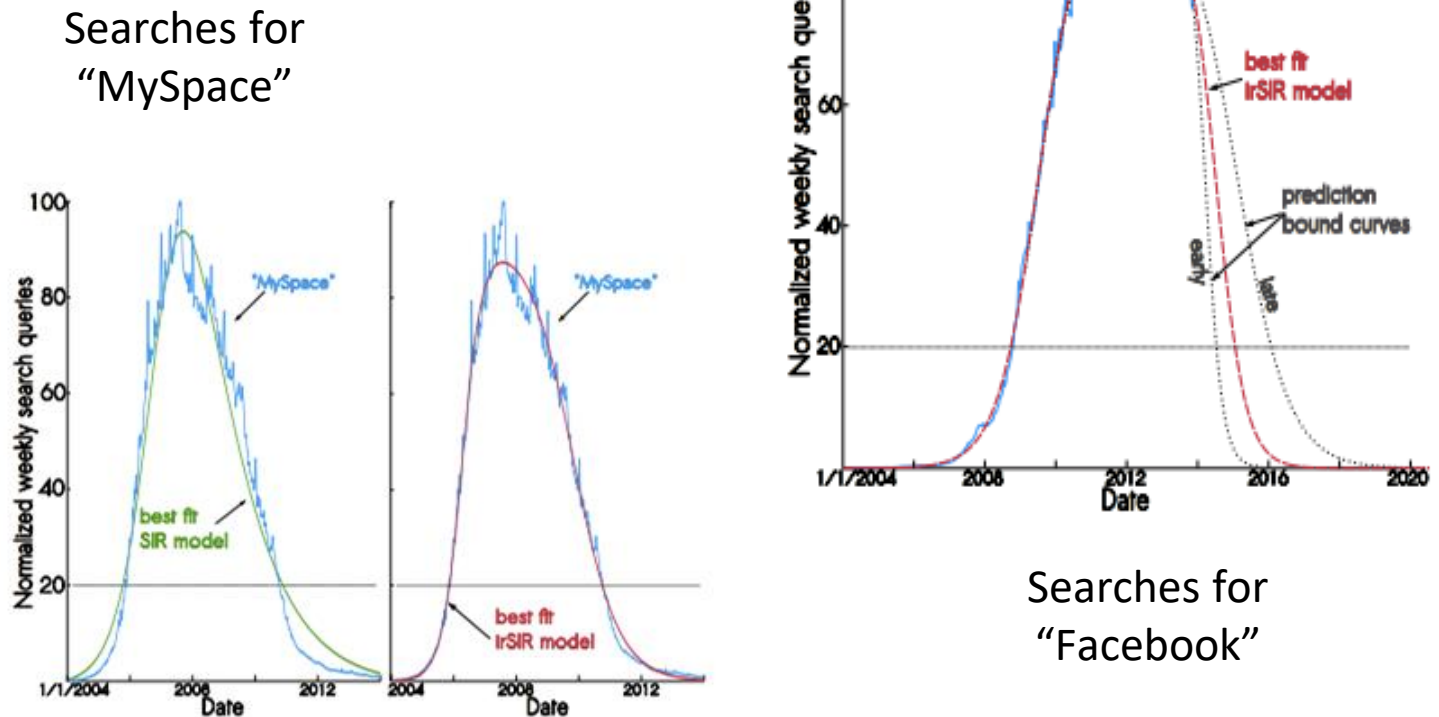
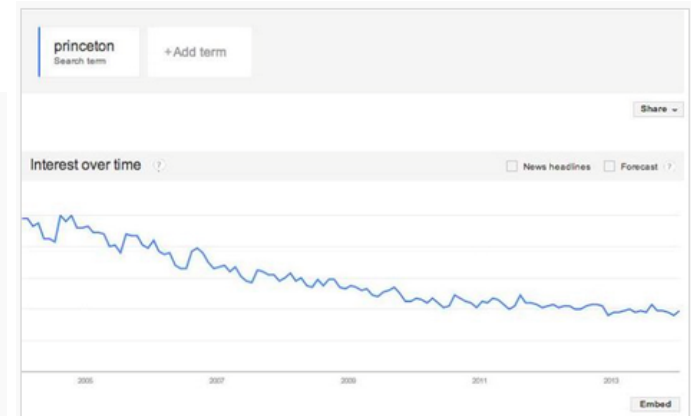
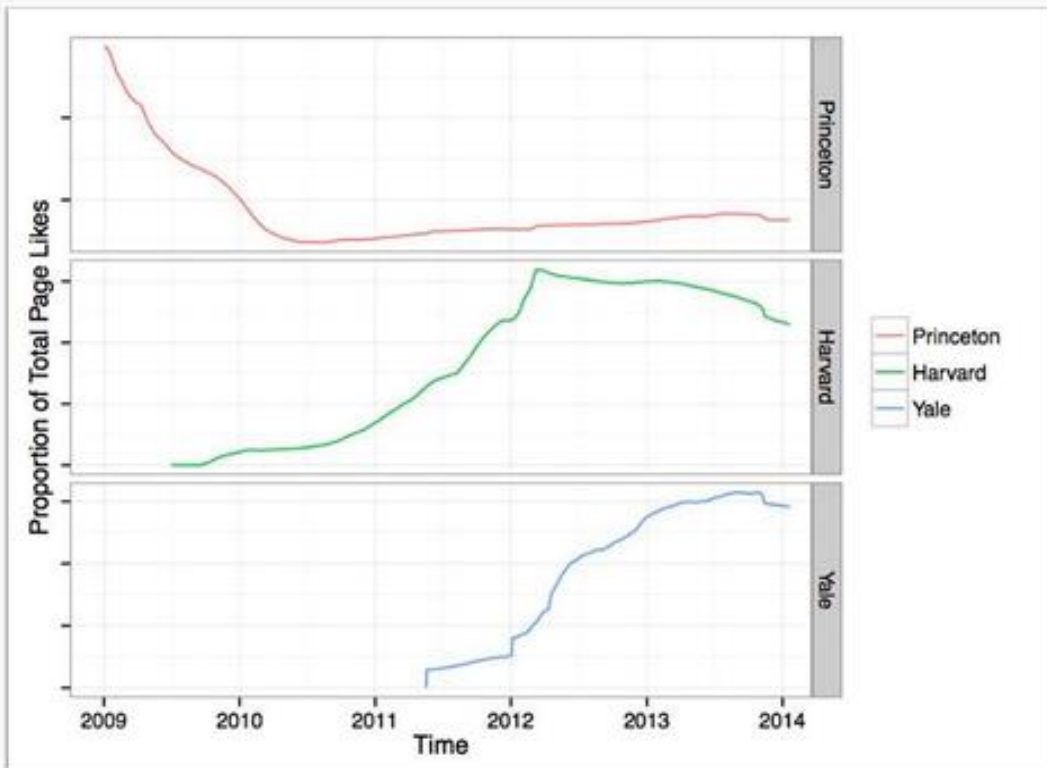


Figure 3: Data for search query "Myspace" with best fit (a) SIR and (b) IrSIR models overlaid. The search query data are normalized such that the maximum data point corresponds to a value of 100.



# Data Makes Everything Clearer

In keeping with the scientific principle “correlation equals causation,” our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely. Looking at page likes on Facebook, we find the following alarming trend:



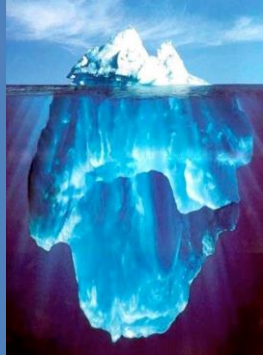
and based on Princeton search trends:

“This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,...

<http://techcrunch.com/2014/01/23/facebook-losing-users-princeton-losing-credibility/>

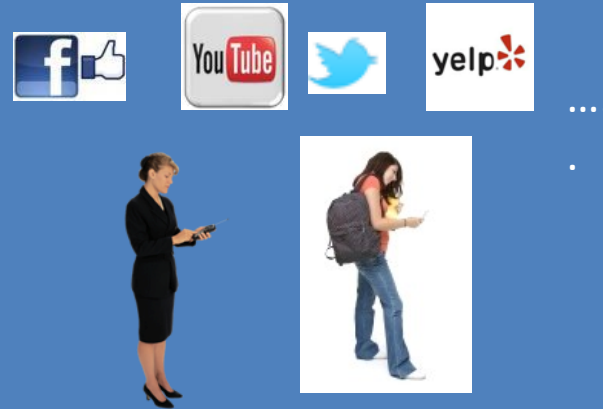
# “Big Data” Sources

## It's All Happening On-line



Every:  
Click  
Ad impression  
Billing event  
Fast Forward, pause,...  
Server request  
Transaction  
Network message  
Fault  
...

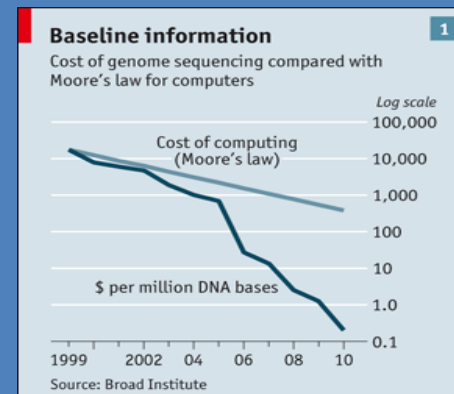
## User Generated (Web & Mobile)



## Internet of Things / M2M



## Health/Scientific Computing

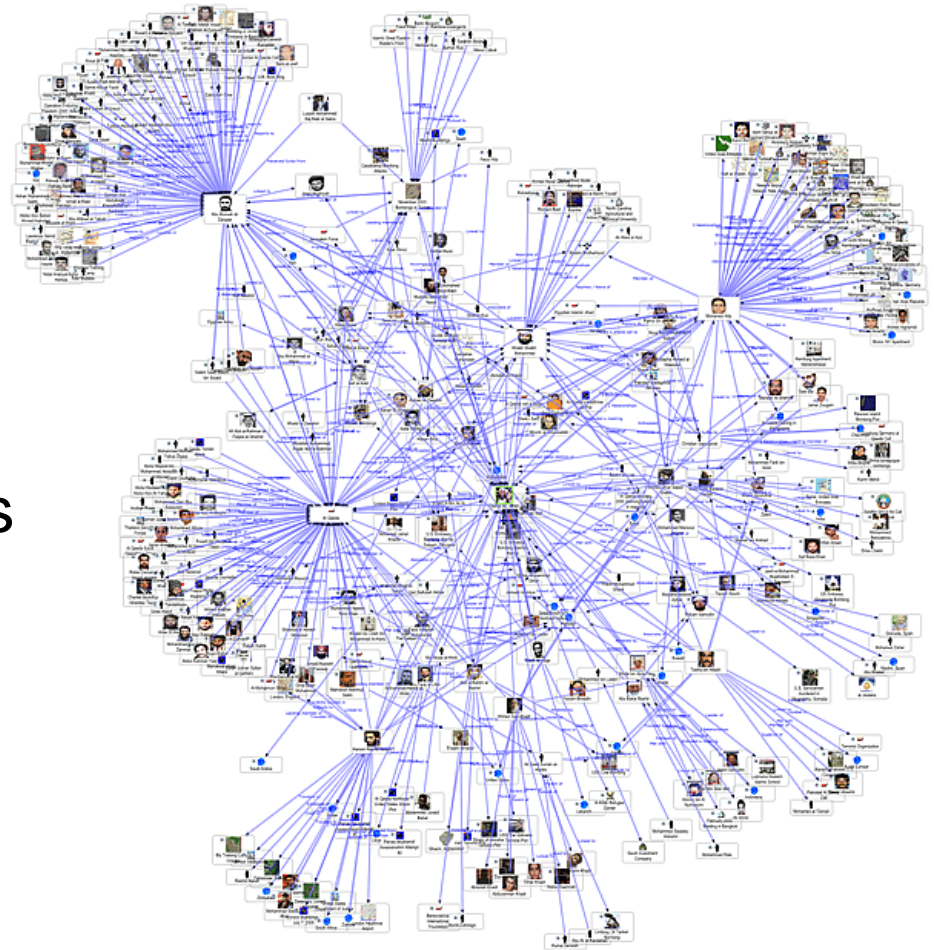


# Graph Data

Lots of interesting data has a graph structure:

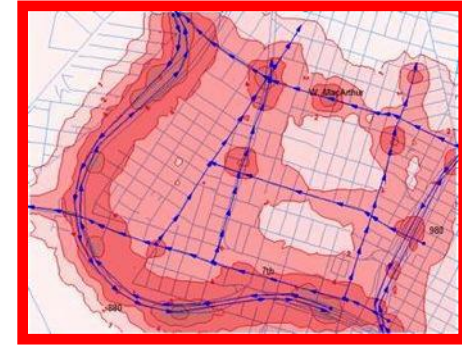
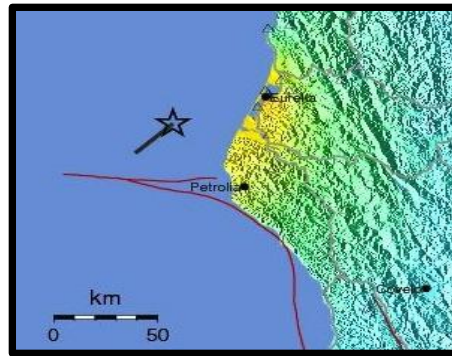
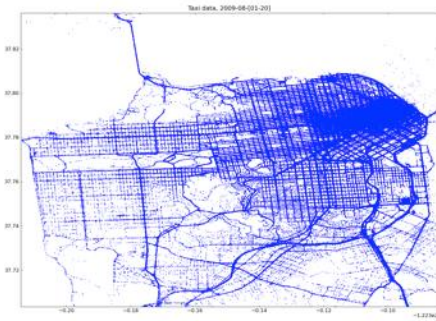
- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- ...

Some of these graphs can get quite large (e.g., Facebook\* user graph)

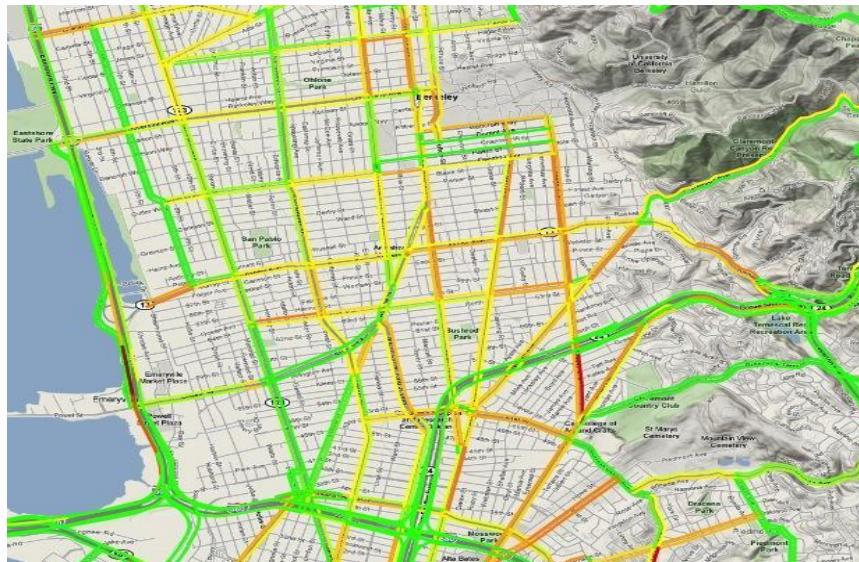




# What can you do with the data?



Crowdsourcing + physical modeling + sensing + data assimilation



From Alex Bayen, UCB

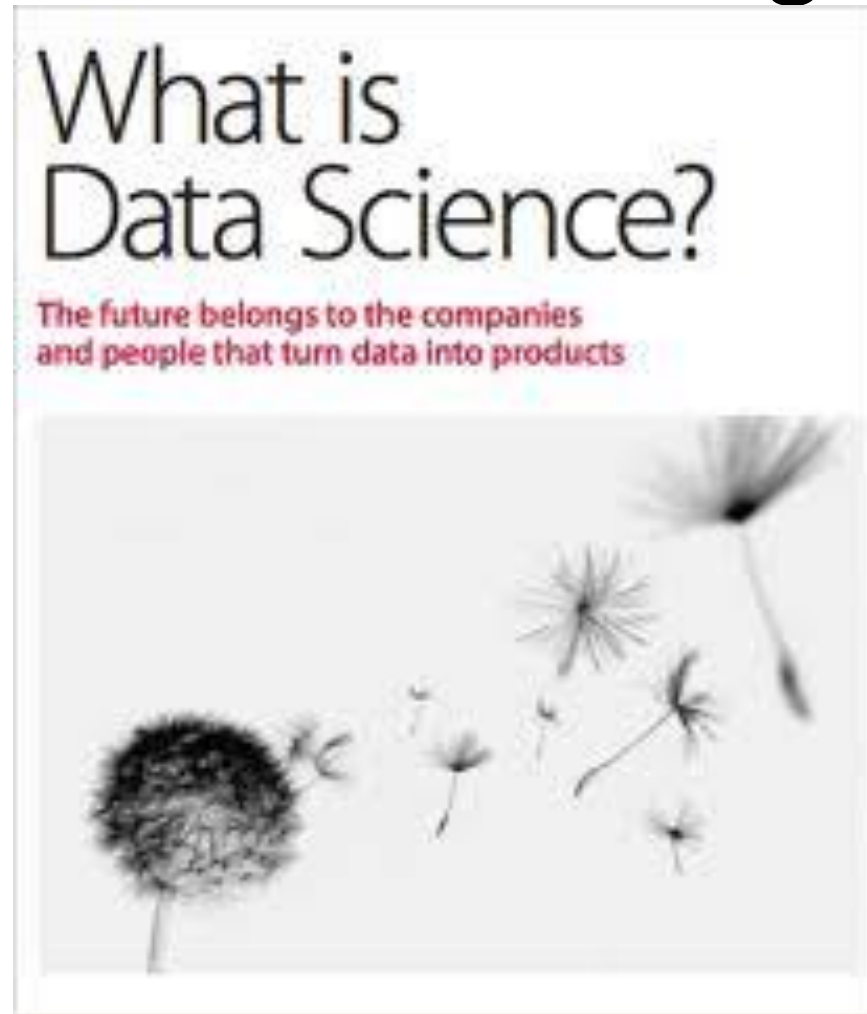


# “Big Data” is so sexy

- “... the sexy job in the next 10 years will be statisticians,” Hal Varian, Google Chief Economist
- the U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018.  
McKinsey Global Institute’s June 2011
- New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...
- New degree programs, courses, boot-camps:
  - e.g., at Berkeley: Stats, I-School, CS, Astronomy...
  - One proposal (elsewhere) for an MS in “Big Data Science”














**DATA SCIENCE – WHAT IS IT?**

# “Data Science” an Emerging Field

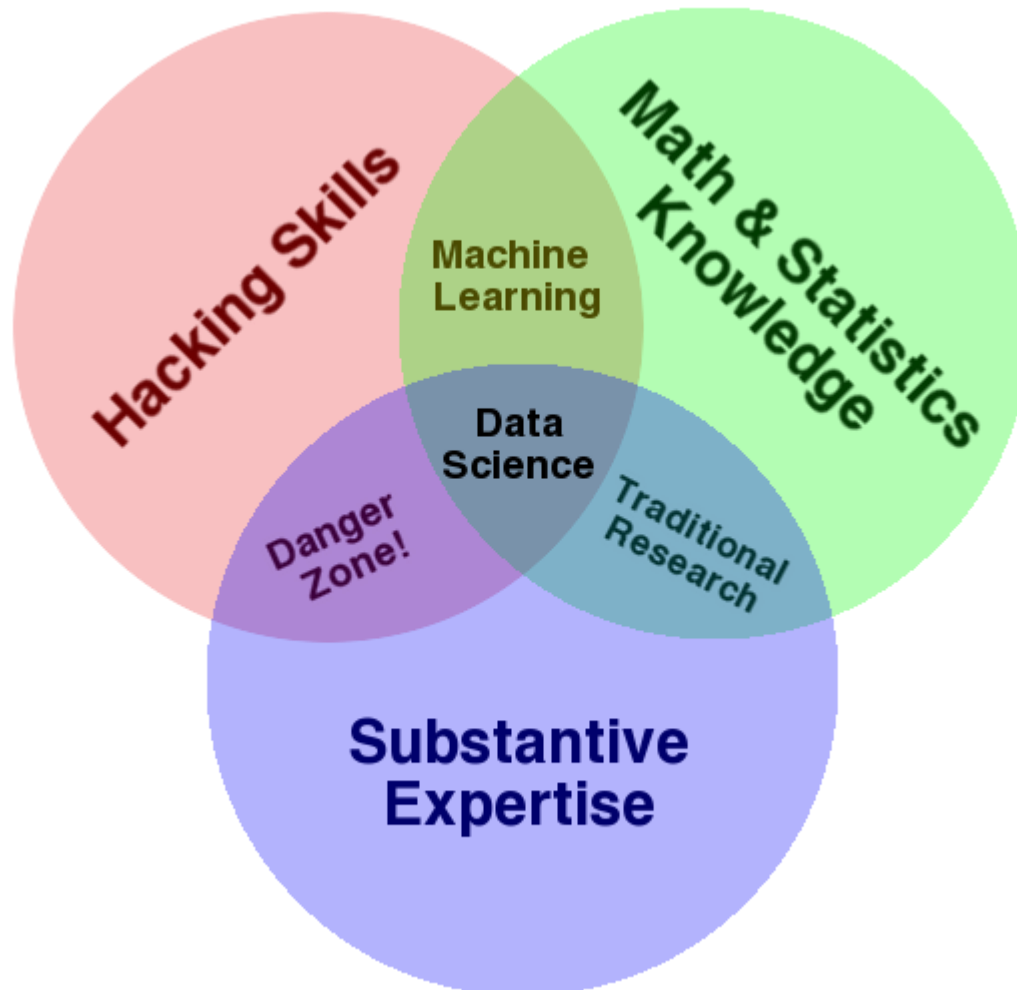


O'Reilly Radar report

# Some recent ML Competitions

Active Competitions			
		<b>Flight Quest 2: Flight Optimization</b> Final Phase of Flight Quest 2	33 days Coming soon \$220,000
		<b>Packing Santa's Sleigh</b> He's making a list, checking it twice; to fill up his sleigh, he needs your advice	5.8 days 338 teams \$10,000
		<b>Flu Forecasting</b>  Predict when, where and how strong the flu will be	41 days 37 teams
		<b>Galaxy Zoo - The Galaxy Challenge</b> Classify the morphologies of distant galaxies in our Universe	2 months 160 teams \$16,000
		<b>Loan Default Prediction - Imperial College Lon...</b> Constructing an optimal portfolio of loans	52 days 82 teams \$10,000
		<b>Dogs vs. Cats</b> Create an algorithm to distinguish dogs from cats	11 days 166 teams Swag

# Data Science – One Definition



# Contrast: Databases

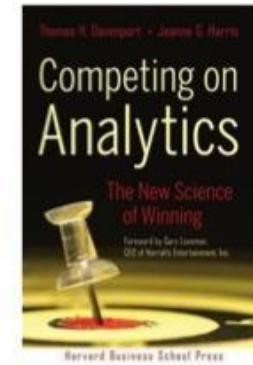
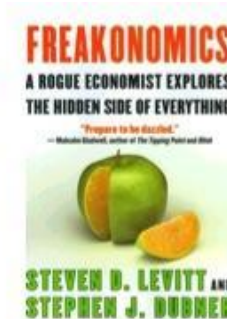
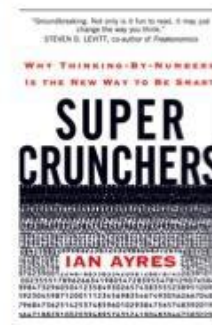
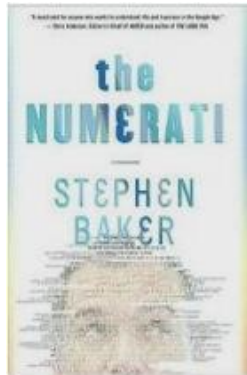
	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB, Hbase, Cassandra,...

ACID = Atomicity, Consistency, Isolation and Durability

CAP = Consistency, Availability, Partition Tolerance

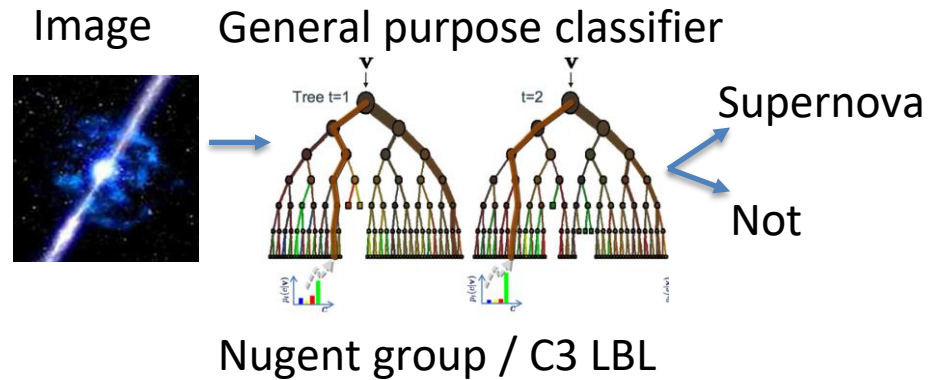
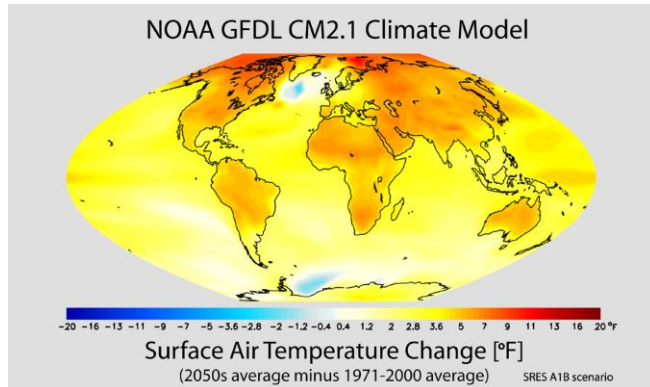
# Contrast: Databases

Databases	Data Science
Querying the past	Querying the future



**Business intelligence (BI)** is the transformation of raw data into meaningful and useful information for business analysis purposes. BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities - Wikipedia

# Contrast: Scientific Computing



## Scientific Modeling

Physics-based models

Problem-Structured

Mostly deterministic, precise

Run on Supercomputer or  
High-end Computing Cluster

## Data-Driven Approach

General inference engine replaces model

Structure not related to problem

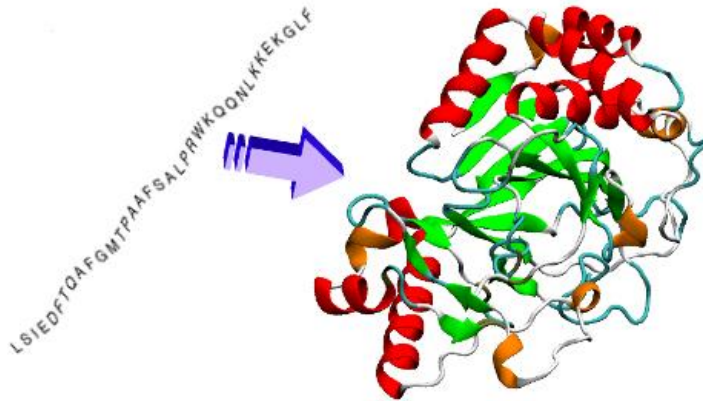
Statistical models handle true randomness,  
and **unmodeled complexity**.

Run on cheaper computer Clusters (EC2)



# Contrast: Computational Science

## CASP: A Worldwide, Biannual Protein Folding Contest



### Quark

Rich, Complex  
Energy Models

Faithful, Physical  
Simulation

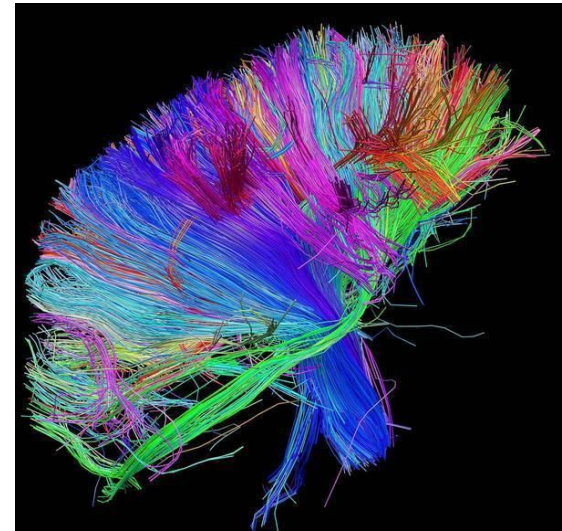
### Raptor-X

Data-intensive,  
general ML models

Feature-based inference

Conditional Neural Fields

## Brain Mapping: Allen Institute, White House, Berkeley



### Techniques (Massive ML)

Principal Component Analysis

Independent Component Analysis

Sparse Coding

Spatial (Image) Filtering

# Contrast: Machine Learning

## Machine Learning

Develop new (individual) models

Prove mathematical properties of models

Improve/validate on a few, relatively clean, small datasets

Publish a paper

## Data Science

Explore many models, build and tune hybrids

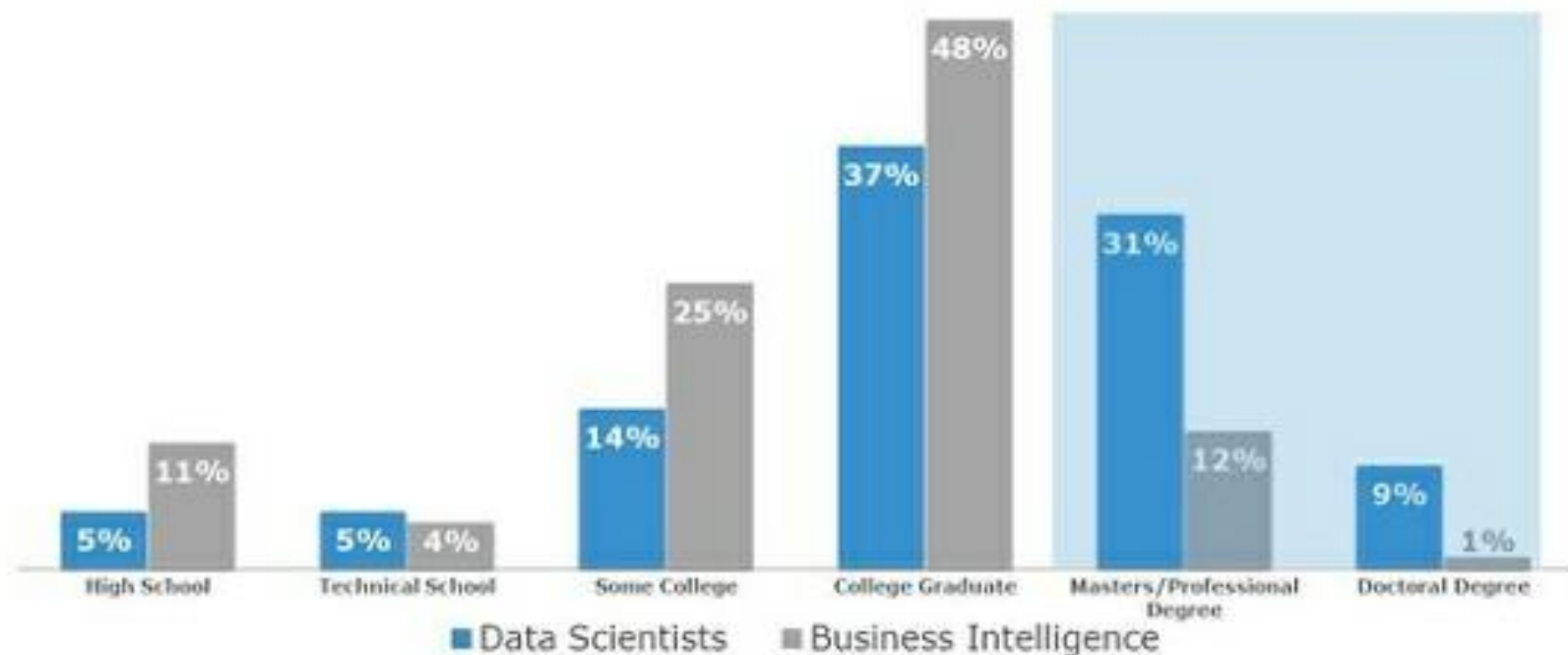
Understand empirical properties of models

Develop/use tools that can handle massive datasets

Take action!

5-min break

# Data science requires greater education



40% of data science professionals have an advanced degree – and nearly one in ten have a doctorate. In contrast, less than 1% of BI professionals have a PhD.

# Analyzing the Analysts

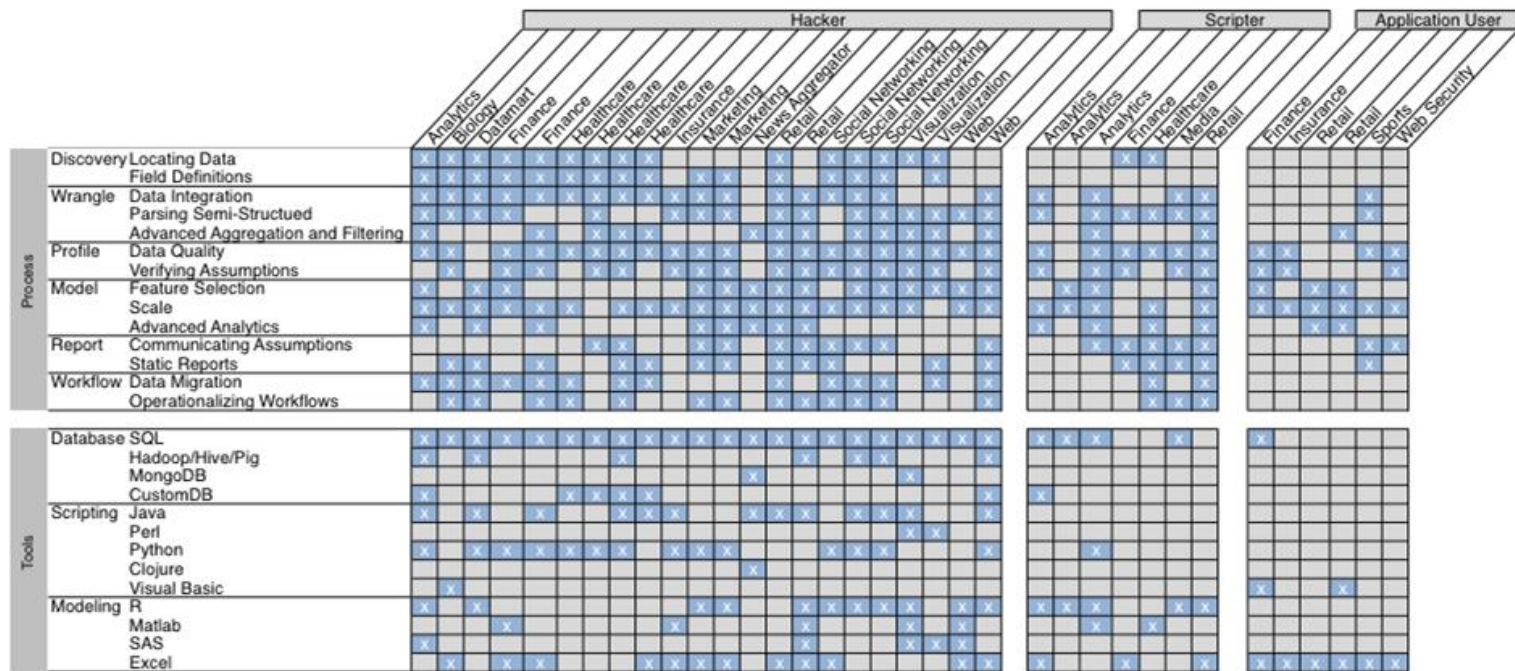


Fig. 1. Respondents, Challenges and Tools. The matrix displays interviewees (grouped by archetype and sector) and their corresponding challenges and tools. *Hackers* faced the most diverse set of challenges, corresponding to the diversity of their workflows and toolset. *Application users* and *scripters* typically relied on the IT team to perform certain tasks and therefore did not perceive them as challenges.

From Kandel, Paepcke, Hellerstein and Heer, “Enterprise Data Analysts and Visualization: An Interview Study”, IEEE VAST 2012

**DOING DATA SCIENCE**

# Ben Fry's Model

1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

# Jeff Hammerbacher's Model

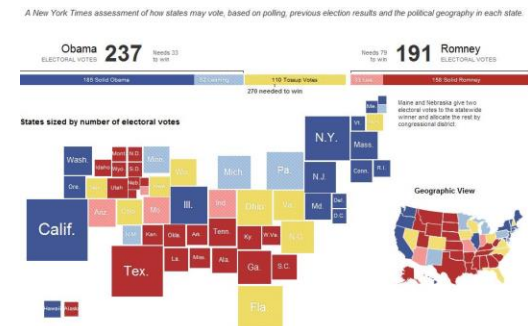
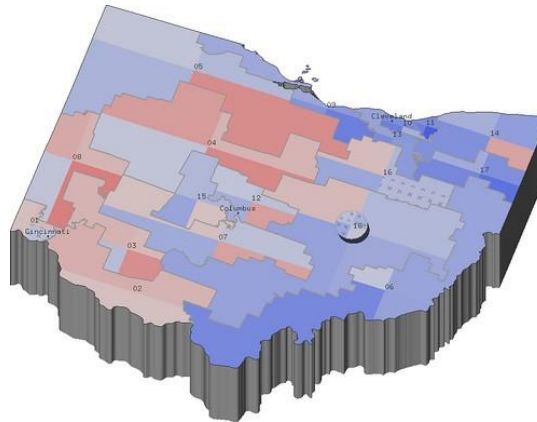
1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate)



5. Build model

6. Evaluate model

7. Communicate results





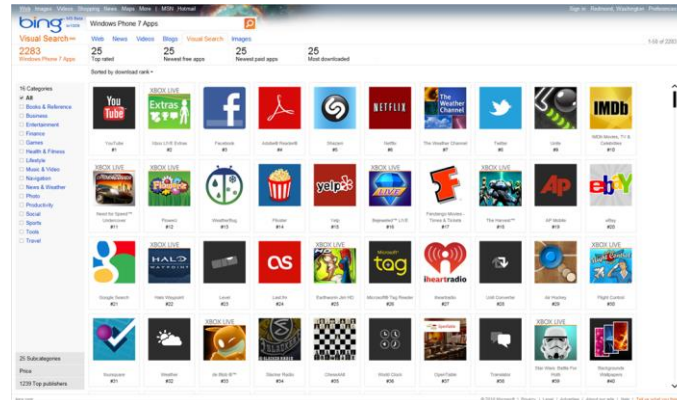
# From the Trenches

Yahoo [KDD 2009, best app. paper]

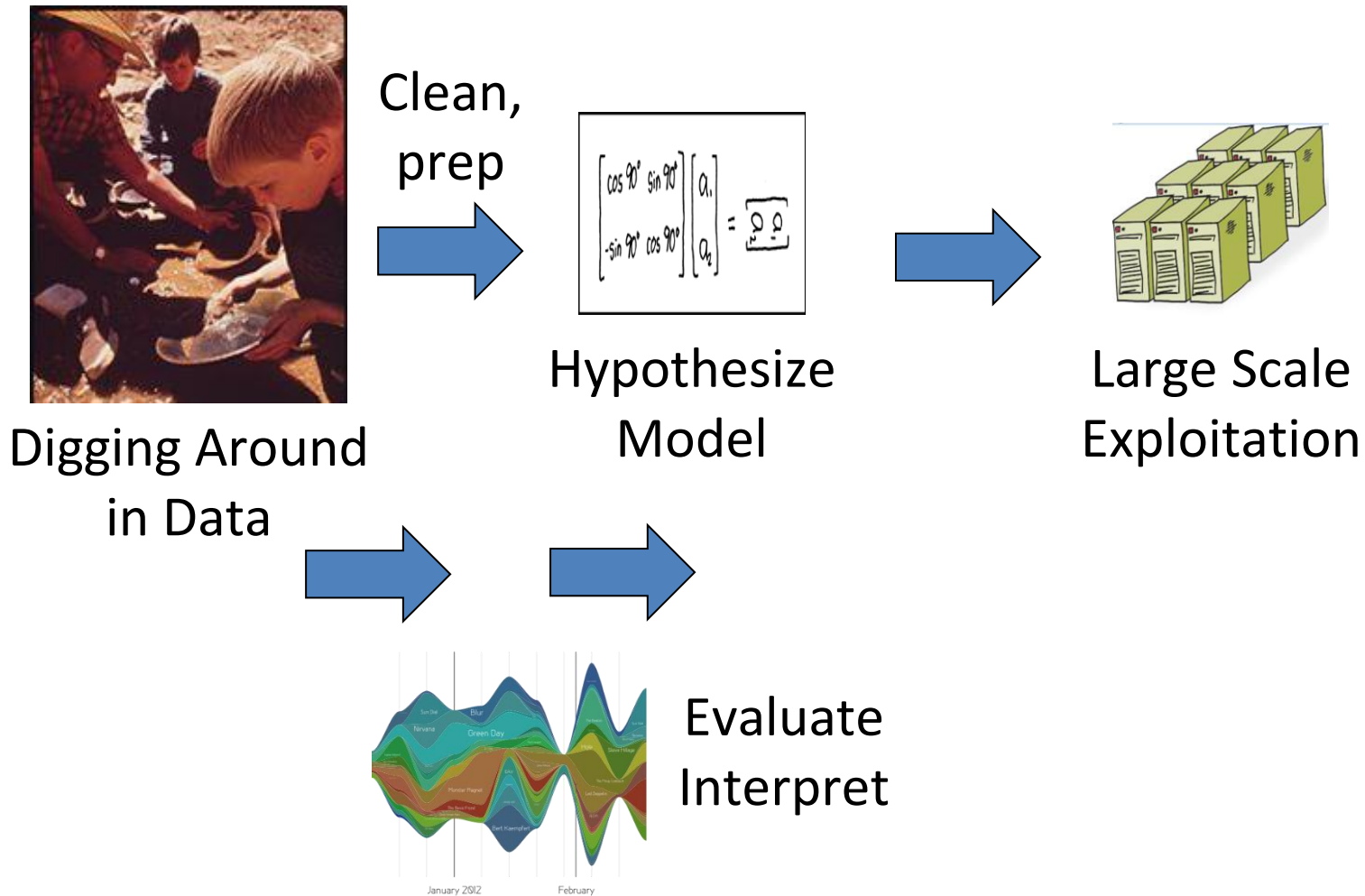
Ebay [SIGIR 2011, hon. mention]

Quantcast [2012]

Microsoft [CIKM 2014]



# Data Scientist's Practice



# What's Hard about Data Science

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype  $\rightarrow$  Production transitions
- Data pipeline complexity (who do you ask?)

# Analysis

What kinds of data will you use?

- Almost anything is OK, except other predictions.
- History: individual or pair-wise?
- Team or players?
- Numerical or text?
- What kind of model will you build?
- What assumptions are safe to make?

# **BASIC PYTHON PROGRAMMING USING REPL.IT**