Foundations of Data Science
Lecture 1, Module 1
Fall 2019

Rumi Chunara, PhD

CS6053

Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute without the instructor's permission.

Today

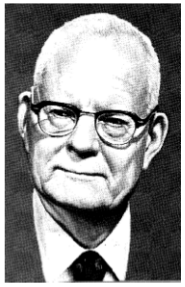
- About the Course (M0)
- What is Data Science? (M1)
- Python intro
- Statistics Review

Why We've Analyzed Data Has Had Different Focuses Over Time

1935: "The Design of Experiments"

R.A. Fisher 

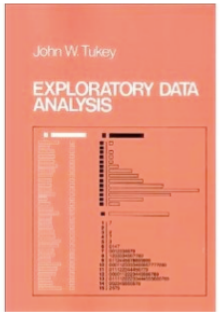
1939: "Quality Control"

W.E. Demming 

1958: "A Business Intelligence System"

Peter Luhn

1977: "Exploratory Data Analysis"



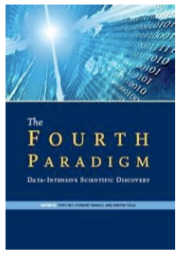
1989: "Business Intelligence"

Howard Dresner 

1997: "Machine Learning"

2010:"The Data Deluge"

2007:"The Fourth Paradigm"



1996: Google

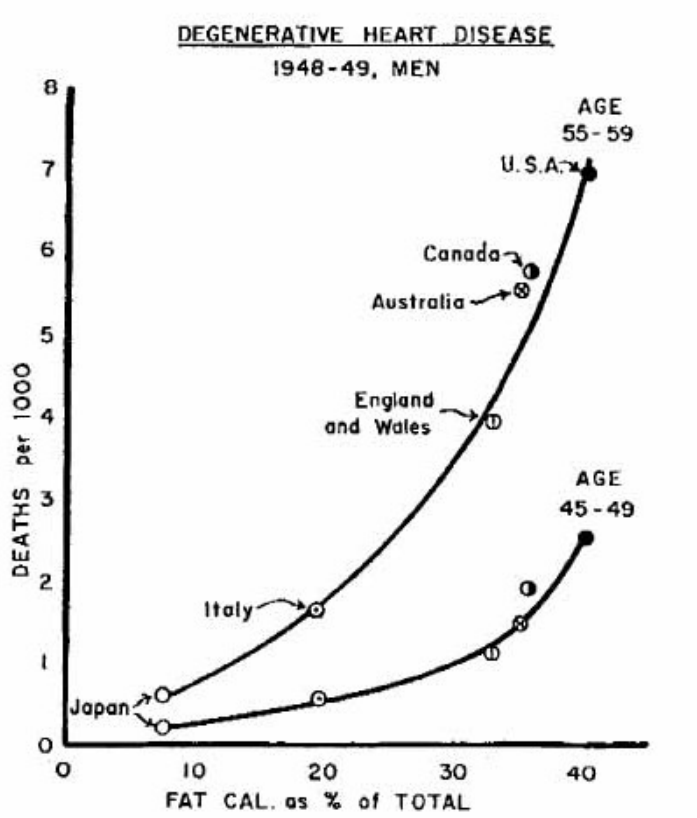
Google™

Science

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

In General: Data Helps Solve Problems

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



In General: Data Helps Solve Problems

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.

In General: Data Helps Solve Problems

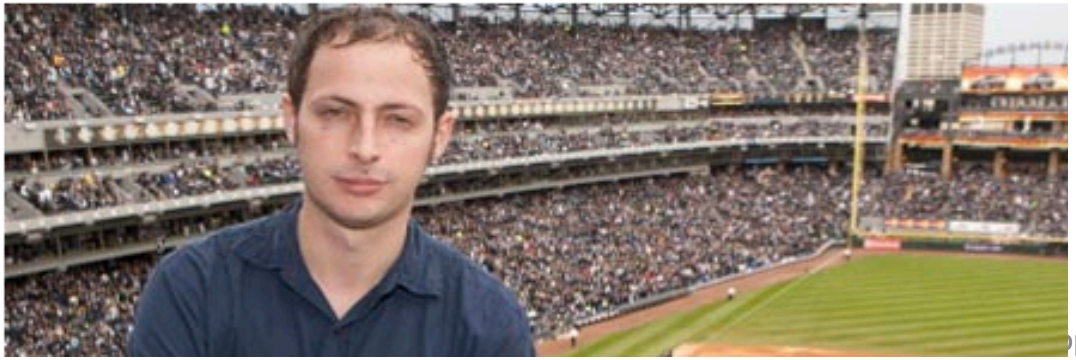
elections2012

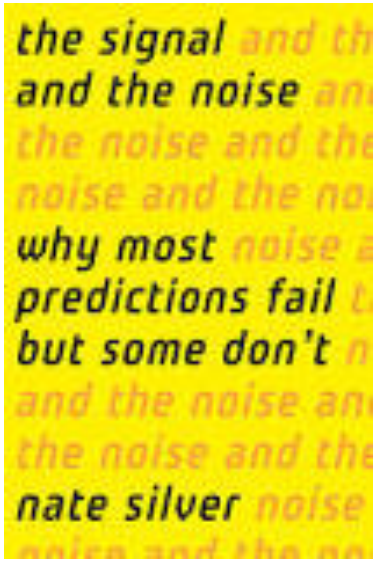
Live results | President | Senate | House | Governor | Choose your

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, Wednesday 7 November 2012 10.45 EST





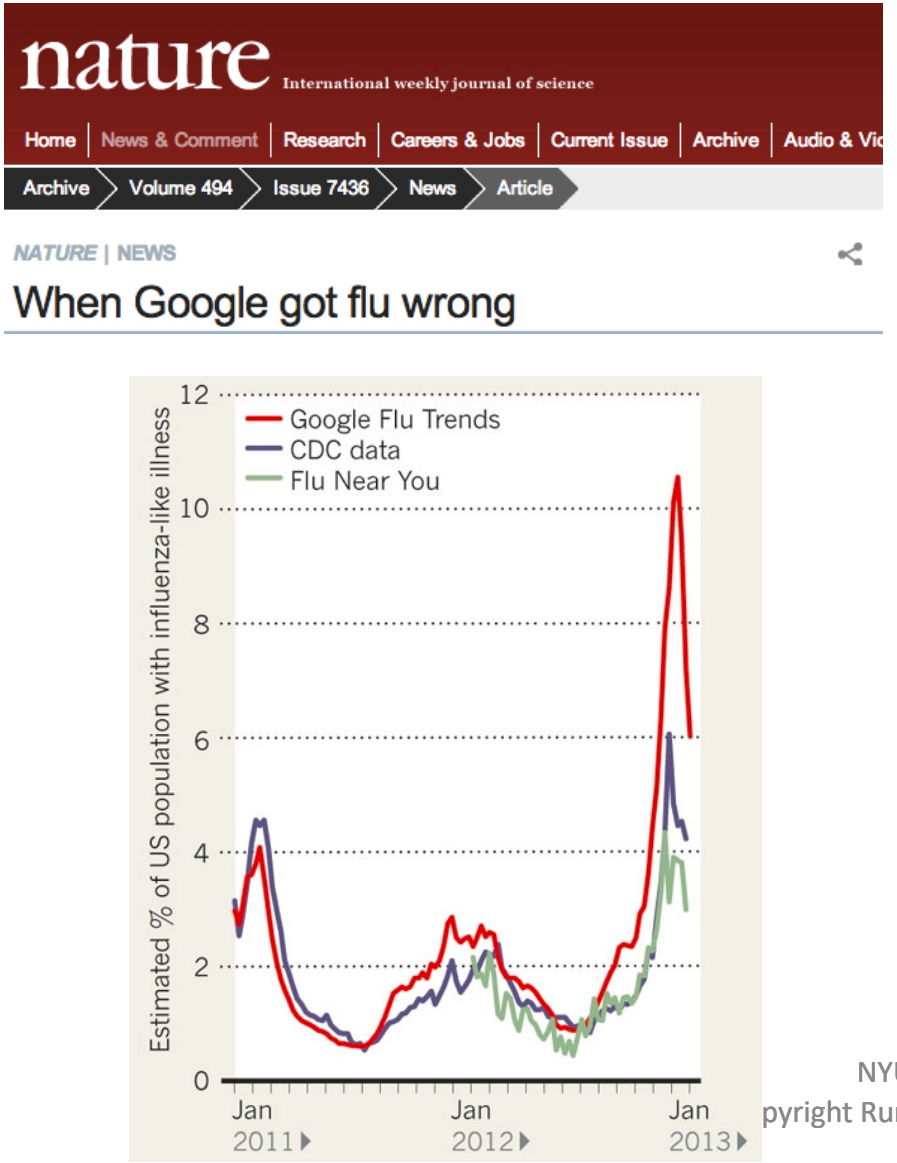
Data and Election 2012

- …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.

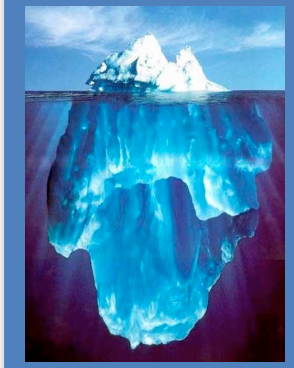
New York TImes, Wed Nov 7, 2012

More Data Brings New Challenges



Data Sources

It's All Happening On-line



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

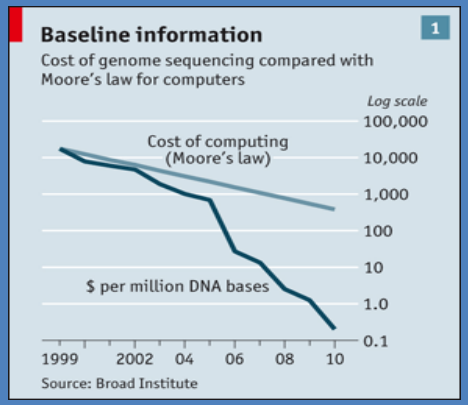
User Generated (Web & Mobile)




….
.


Internet of Things / M2M

Health/Scientific Computing

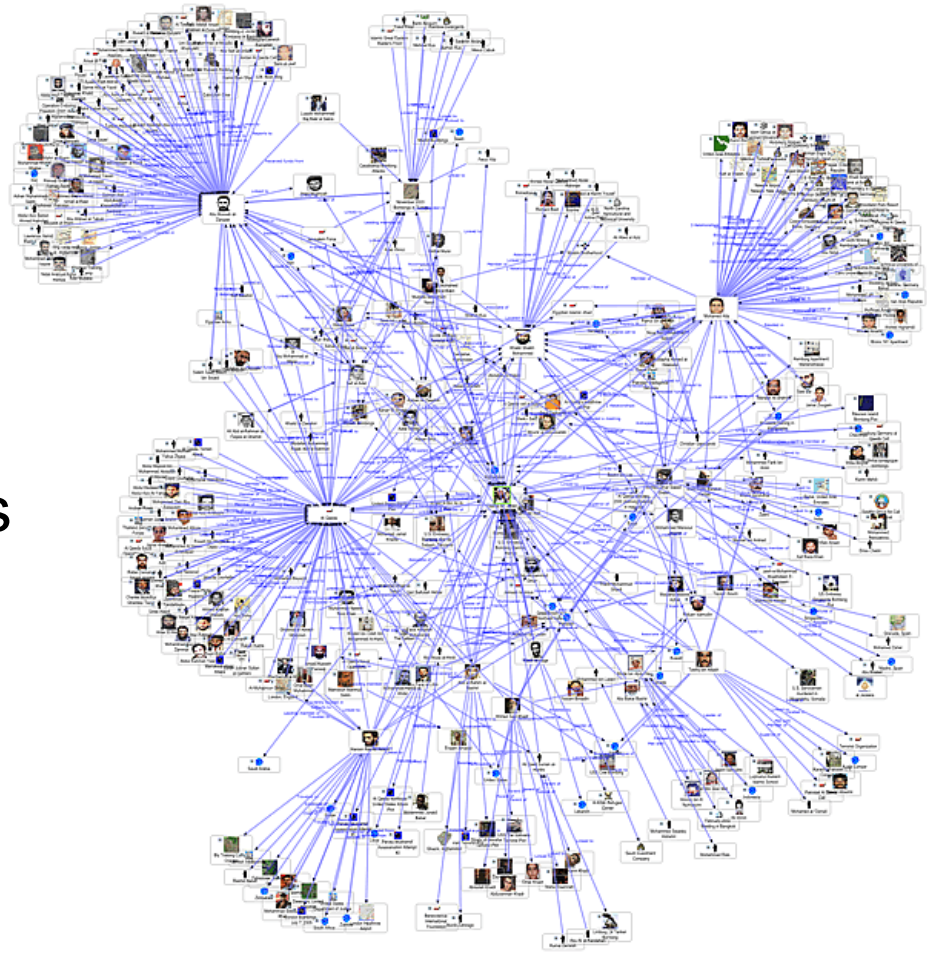


Graph Data

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)



Question Types

- **Simple (descriptive) Stats**
  - What are the genomic profiles of one group
- **Hypothesis Testing**
  - Is there a difference in movement of different people
- **Segmentation/Classification**
  - What are the common characteristics of customers
- **Prediction**
  - On what week will influenza peak in New York City?

Data Makes Everything Clearer?

Epidemiological modeling of online social network dynamics

John Cannarella[1], Joshua A. Spechler[1,*]

1 Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA

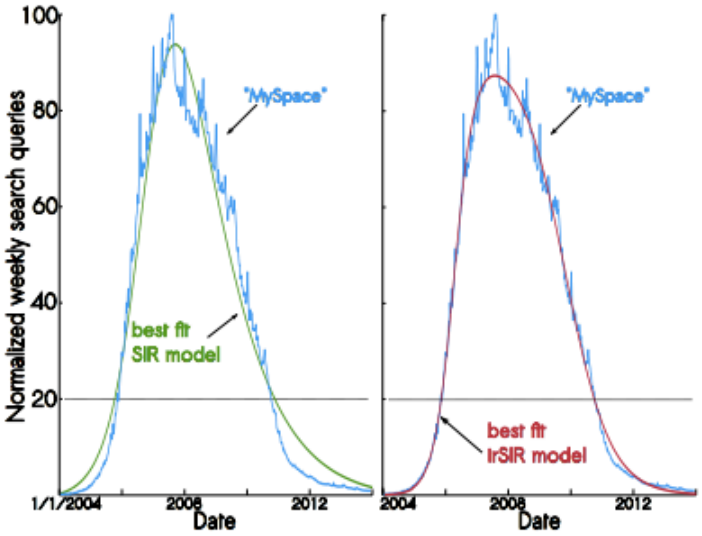
*** E-mail: Corresponding spechler@princeton.edu**

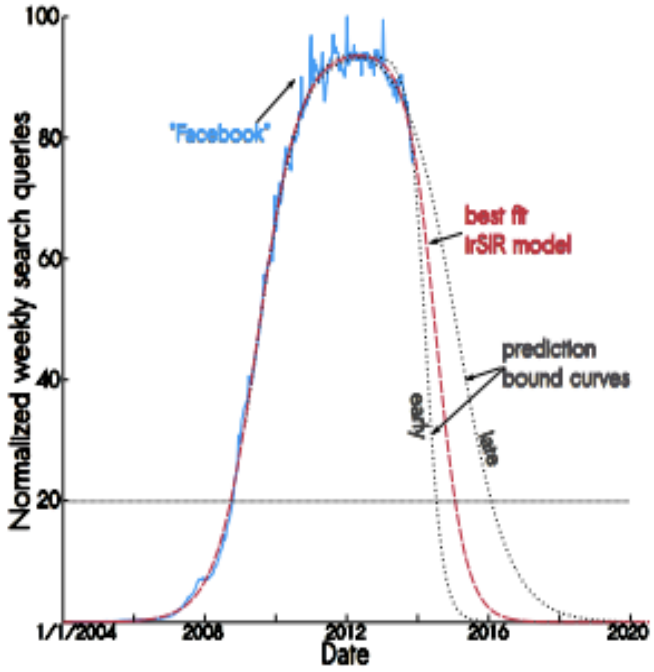
Abstract

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?

Searches for
"MySpace"



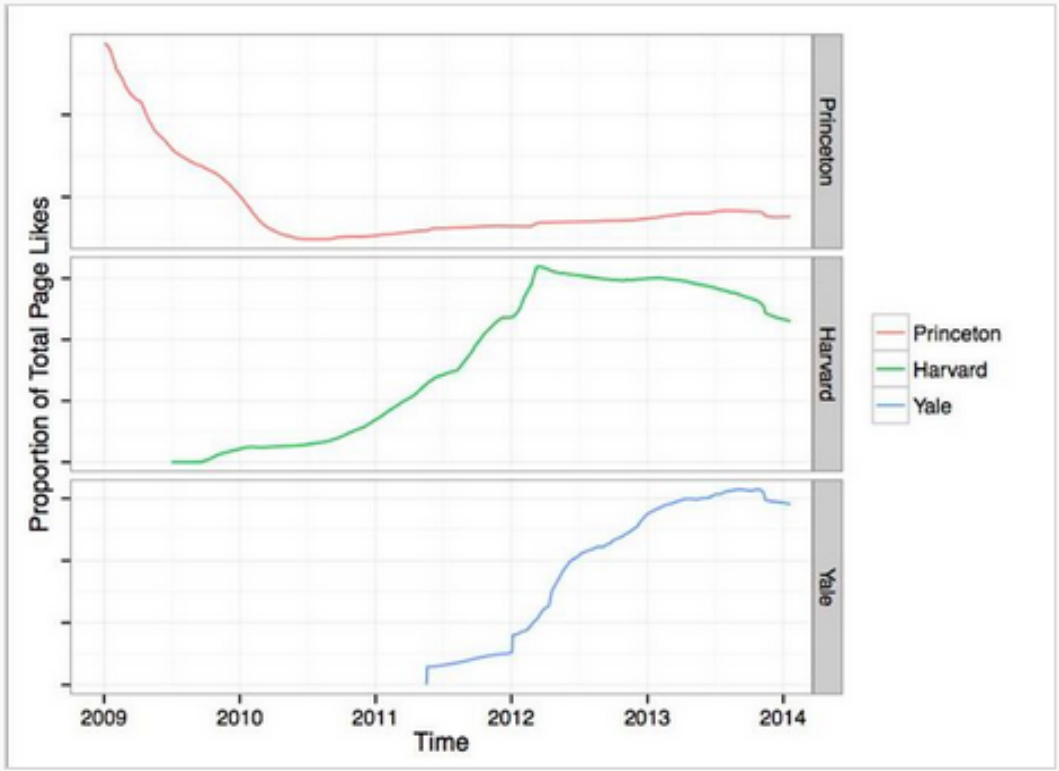


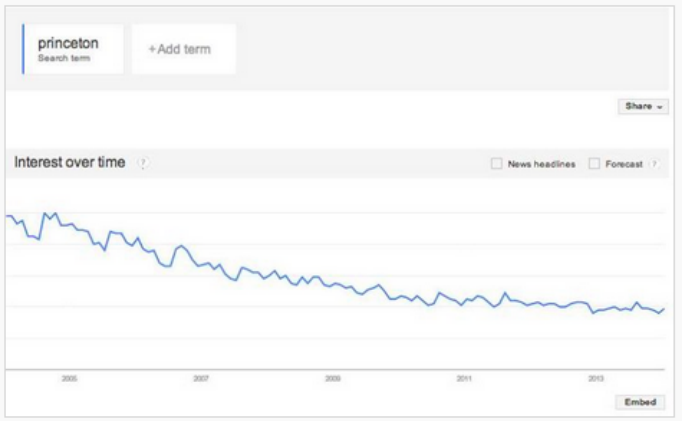
Searches for
"Facebook"

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” is so 2012

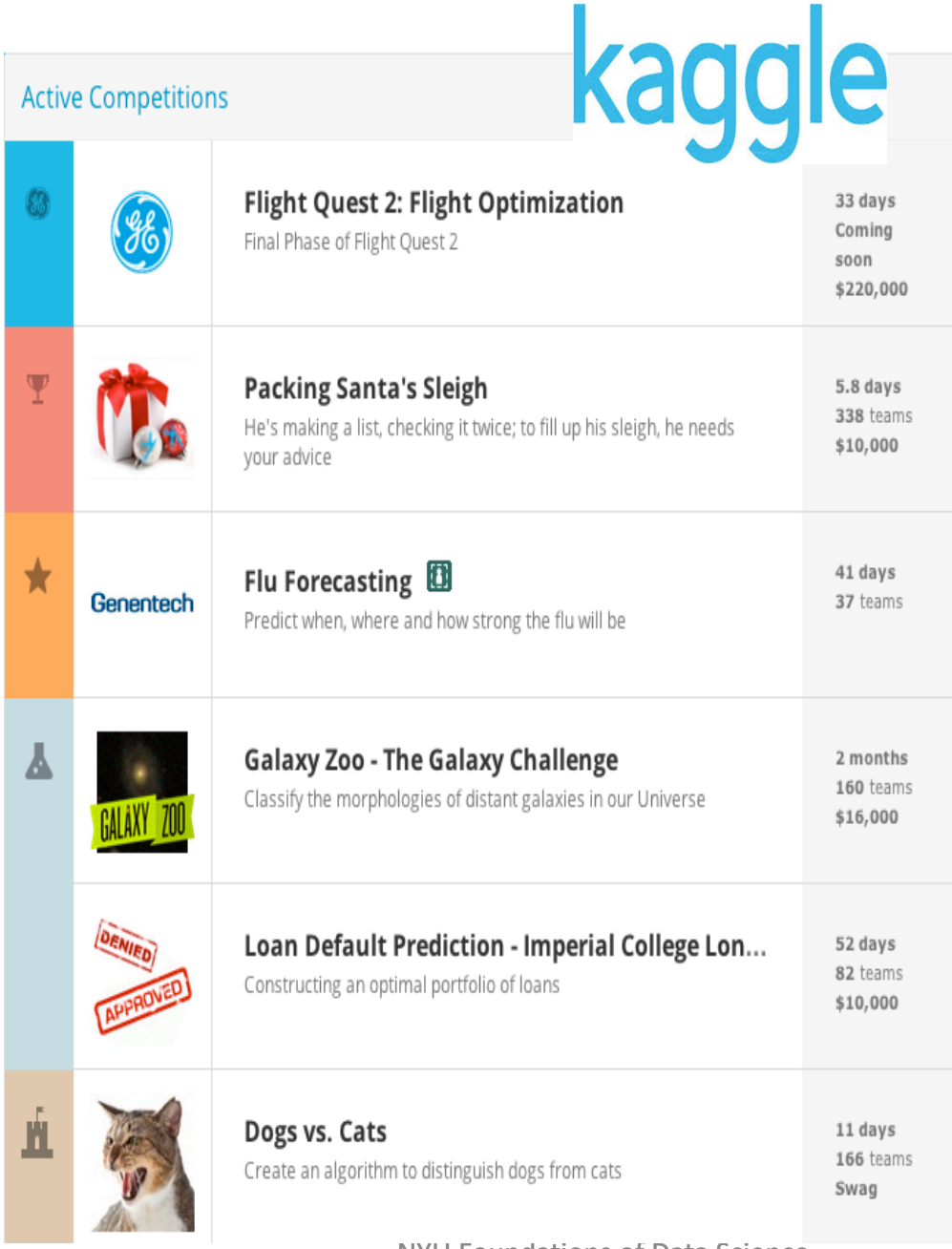
- “… the sexy job in the next 10 years will be statisticians,” Hal Varian, Google Chief Economist
- the U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. McKinsey Global Institute’s June 2011
- New Data Science institutes being created or repurposed – NYU,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

Data Science – What IS IT?

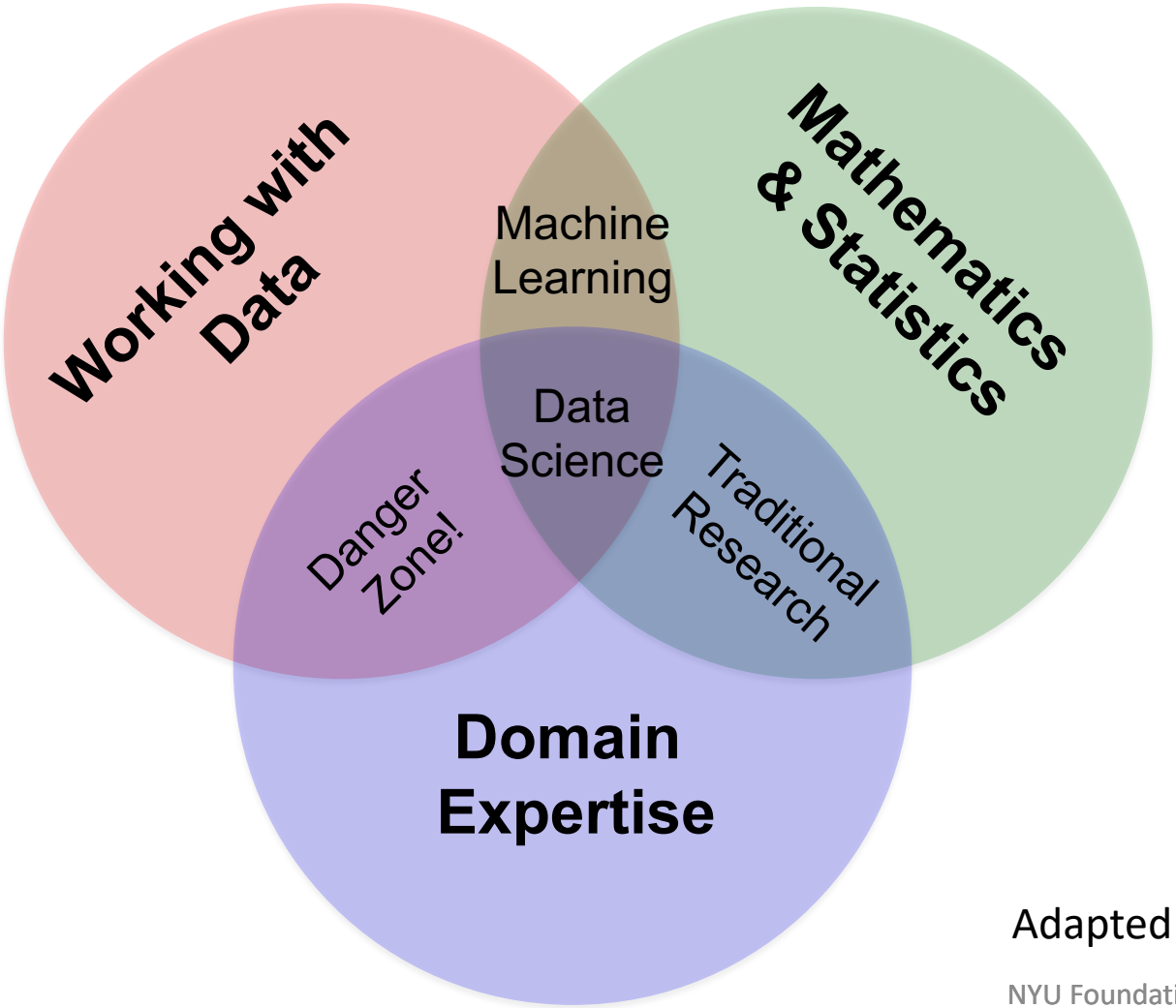
“Data Science” an Emerging Field

O’Reilly Radar report

Some recent DS Competitions



Data Science – One Definition

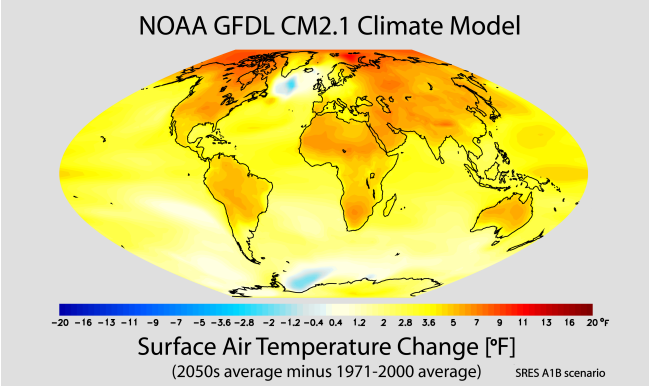


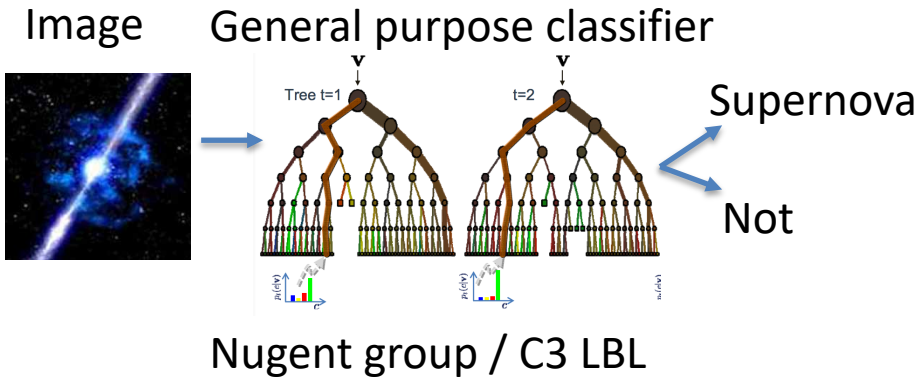
Adapted from Drew Conway

Databases vs. Data Science

	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
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,…

Scientific Computing vs. Data Science





Scientific Modeling
Physics-based models
Problem-Structured
Mostly deterministic, precise
Run on Supercomputer or High-end Computing Cluster

Data-Driven Approach
Data and inference engine replaces model
Structure not related to problem
Statistical models handle true randomness, and **unmodeled complexity**.
Run on cheaper computer Clusters (EC2)

Machine Learning vs. Data Science

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!

How to Learn Data Science?

- Masters Programs
- Get a *different* masters
- Work in Data Science

Doing Data Science

1) What Is A Data Scientist?
2) Data Science Workflow

What is A Data Scientist?

Zvi
@nivertech

Follow

"Data Scientist" is a Data Analyst who lives in California.

Reply Retweet Favorite ••• More

RETWEETS	FAVORITES
140	40

9:55 PM - 14 Mar 2012

Josh Wills
@josh_wills

Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Reply Retweet Favorite More

RETWEETS 907 FAVORITES 418

12:55 PM - 3 May 2012

Javier Nogales
@fjnogales

Follow

Data Scientist (2/2): person who is worse at statistics than any statistician and worse at software engineering than any software engineer

RETWEET 1

FAVORITES 5

9:08 AM - 27 Jan 2014

The resume: skills

Through the lens of a Data Scientist Job Description

PhD is a proxy for:
- **experience**
- **research ability**
- **technical expertise**

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a
- Comfort manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- Ability to communicate complex quantitative analysis and analytic approaches in a clear, precise, and actionable manner
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

The resume: skills

Through the lens of a Data Scientist Job Description

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a rel
- Comfort manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- Ability to communicate complex quantitative analysis and analytic approaches in a clear, precise, and actionable manner
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

You can't be a Data Scientist if you can't handle data…

The resume: skills

Through the lens of a Data Scientist Job Description

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a relevant role
- Comfort manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- Ability to communicate complex quantitative analysis and analytic approaches in a clear, precise, and actionable manner
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

This is essentially the goal of this course.

The hard skills

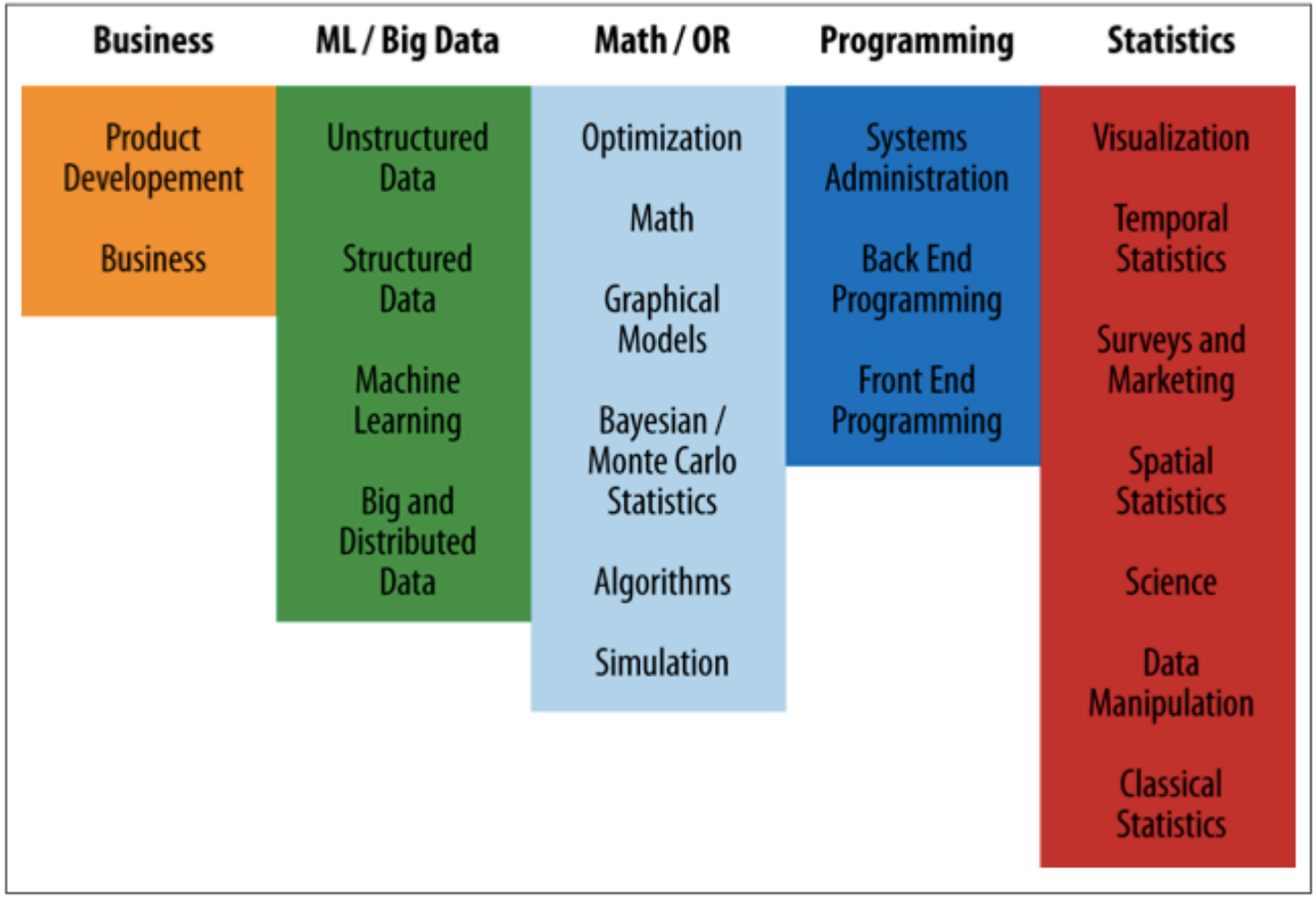
Through the lens of a Data Scientist Job Description

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a relevant role
- **Necessary: a scripting language, SQL and a scientific computing language. You will get hands-on experience with some of this in this course, and you should definitely develop these skills further outside of this course**
- approaches in a clear, precise, and actionable manner
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

Range of DS skills

They're all very similar, but some categorization still helps.



"Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways."

Data Scientist Type A (for Analysis):

- Primarily concerned with **making sense of data** or working with it in a fairly **static** way.
- Similar to a statistician, but knows all the **practical details of working with data** that aren't taught in statistics: data cleaning, dealing with large data sets, visualization, domain knowledge, etc.

Source: https://www.quora.com/What-is-data-science/answer/Michael-Hochster

"Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways."

Data Scientist Type B (for Building):

- Some statistical background, but **strong coder or software engineer**.
- Primarily concerned with **using data "in production"**: building models which interact with users (by giving recommendations, for example).

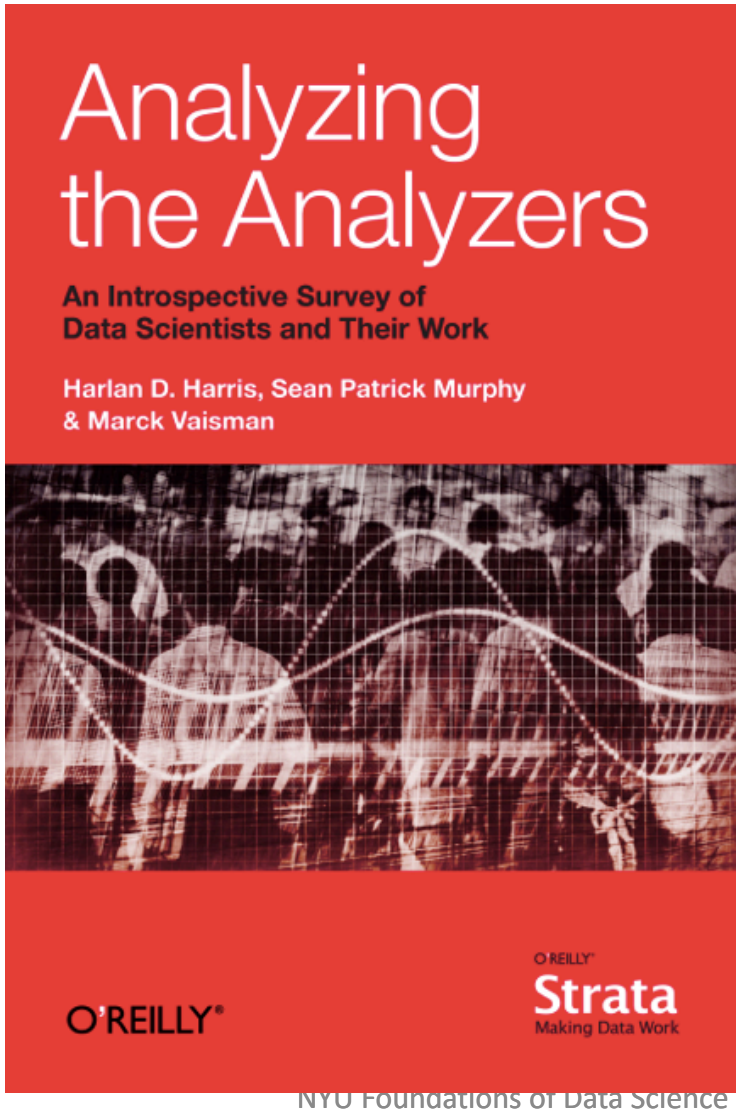
Our course is focused primarily on **Type A**.

Source: https://www.quora.com/What-is-data-science/answer/Michael-Hochster

Towards a Definition

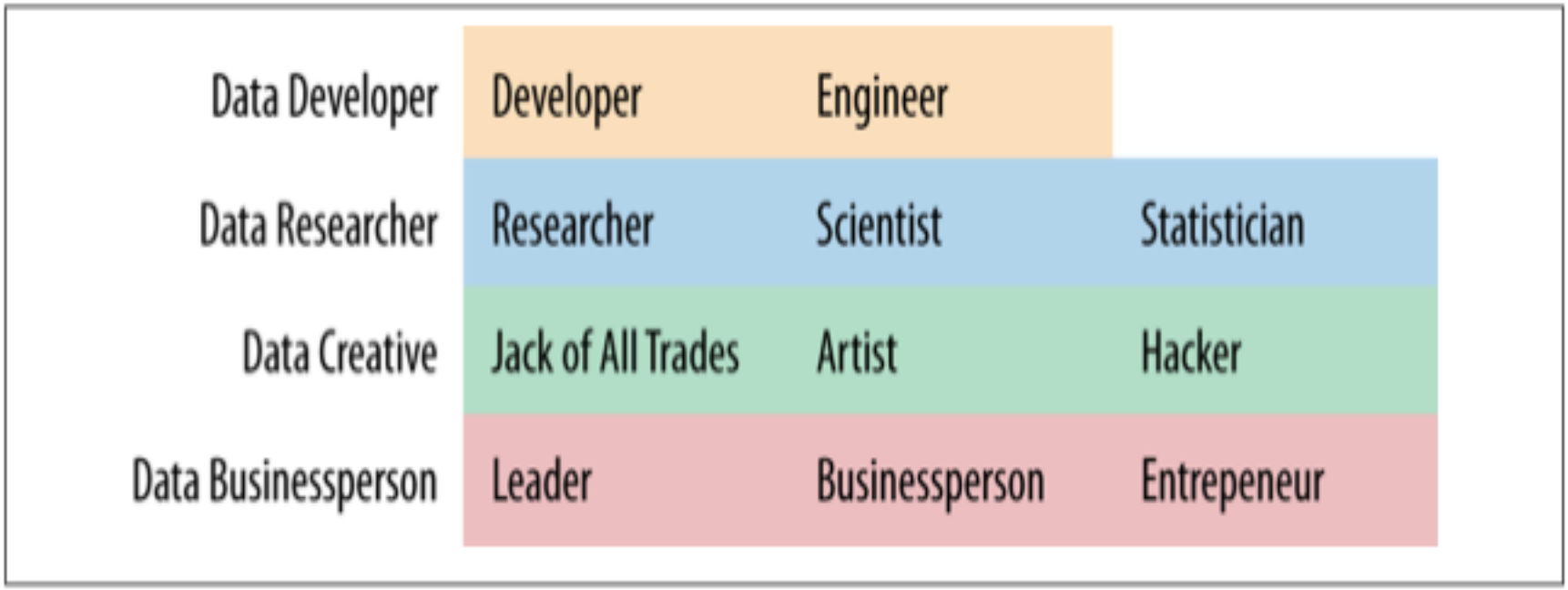
There is no 'one-size-fits-all' type of data scientist.

Luckily, people are using data science to define data science.



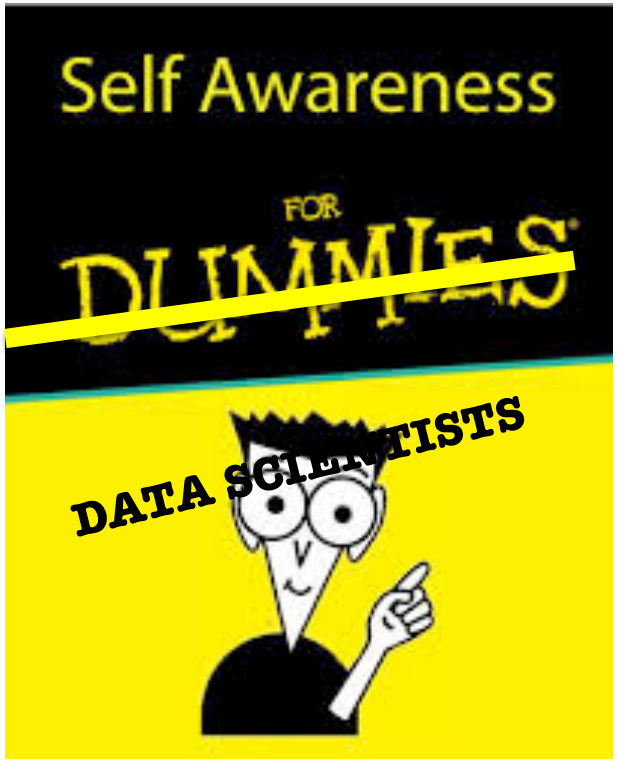
Data Roles

In Analyzing the Analyzers, the authors identified 4 types of “data scientists.”



Figuring Out Your Interests..

You don't have to fit into one bucket, but you should know where you are…



- Personal skills development
- Choosing the right job (your future boss might not know what a data scientist is, or should be)

Why Data "Science"?

We defined 4 data roles, but what is the "science" of data science?

<u>The scientific method:</u> evaluating the merit of a hypothesis with rigorous empirical testing.

I.e.,

Given raw data, constraints and a problem statement, you have an infinite set of models to choose from, with which you will use to maximize performance on some evaluation metric, that you will have to specify.

Every design choice you make can be formulated as a hypothesis, upon which you will use rigorous testing and experimentation to either validate or refute.

It's Also an "Art"

Outside of modeling competitions, seldom is a well-posed problem and clean dataset presented to you.

Putting the art into your practice means…

- Translating problems into the language of data science
- Formulating reasonable hypotheses
- Developing an intuition for good vs. bad data, good vs. bad models.
- Abstracting problems to identify similarities
- Managing the DS process from end to end

Today:

- Many mature off-the shelf tools for analysis exist
- Skills in **problem-solving**, **collaborating** and **communicating** are needed

What We'll Learn

With this course we want to emphasize the *soft* skills of data science

Art => Abstract and intuitive thinking

Science => process

We'll cover necessary Data Science tools, but with the goal of applying them towards analytic problem solving.

Data Science Workflow

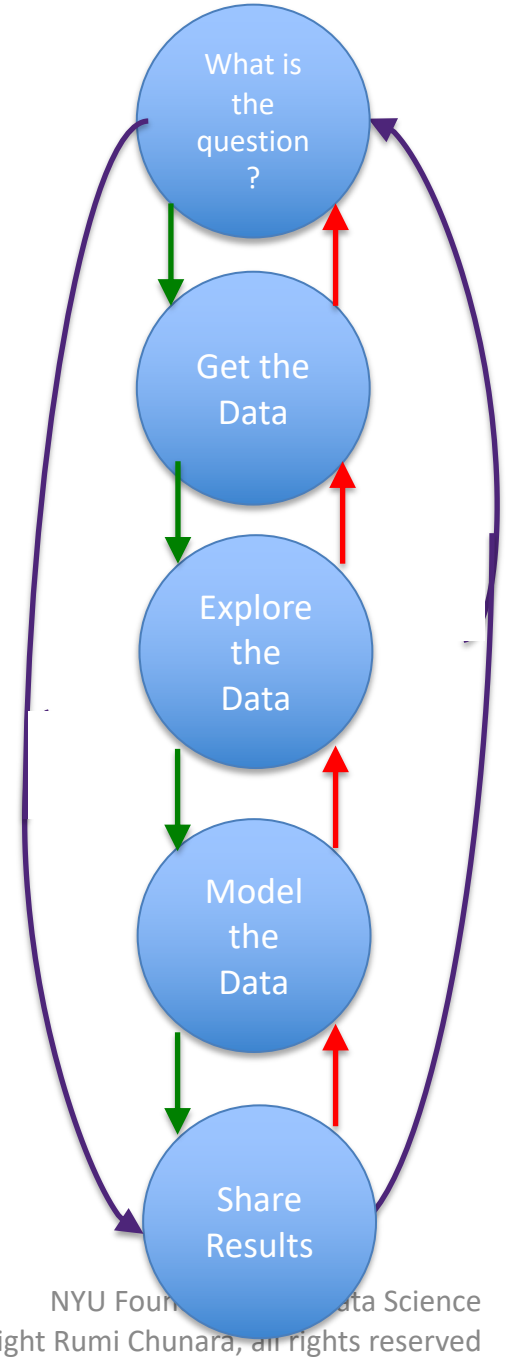
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Visualize the data
Are there **anomalies**?
Are there **patterns**?

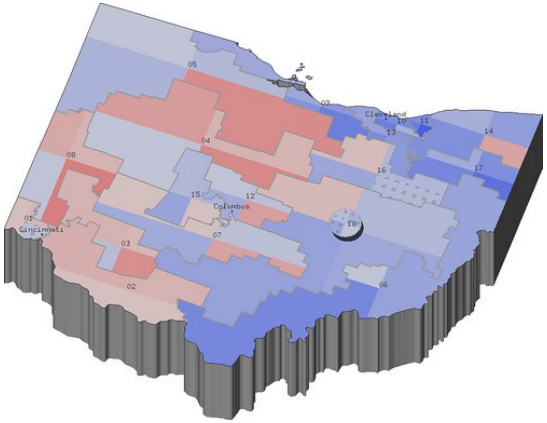
Build a model
Fit the model
Validate the model

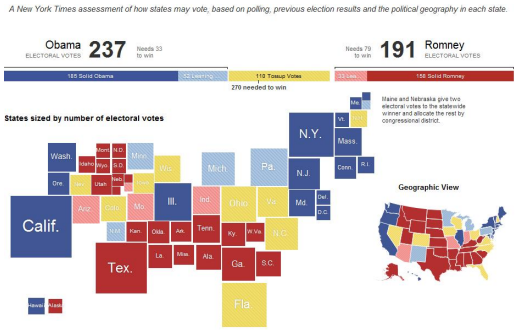
What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?



Jeff Hammerbacher's Model

1. Identify problem + Hypothesis?
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate), Mulch, Dig around
5. Build model
6. Evaluate model
7. Communicate results
8. Scale?





Example: Predicting Neonatal Infection

Problem: Children born prematurely are at high risk of developing infections, many of which are not detected until after the baby is sick

Goal: Detect subtle patterns and features in the data that predicts infection before it occurs

Data: 16 vital signs such as heart rate, respiration rate, blood pressure, etc…

Impact: Model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear

Image: http://www.babycaretips4u.com/wp-content/uploads/2014/03/premature-baby.jpg
Case Study: http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544002695

Example: Automating Insurance Claims

Problem: Processing disability claims at the Social Security Administration is a time-intensive process, with many claims taking over 2 years to adjudicate

Goal: Automate the approval of a subset of the "simplest" disability claims

Data: Free text in the claims form

Impact: Able to fully automate 20% of the simplest claims. Rating accuracy of the algorithm is higher than the average claims examiner.

Case Study: http://datamininglab.com/images/case-studies/ERI_Text_Mining_SSA_Claims_for_Disability_Approval.pdf

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 → Production transitions
- Data pipeline complexity (who do you ask?)