INTRODUCTION TO DATA SCIENCE

DAN POTTER

CARSTEN BINNIG

ELI UPFAL

TIM KRASKA

PREREQUISITES

Required

- **Standard CS Intro Sequence CSCI 0160, 0180 or 0190**

Recommended Computer Science Courses

- **Introduction to Software Engineering CSCI 0320**
- **Introduction to Computer Systems CSCI 0330**
- **Creating Modern Web Applications CSCI 1320**

Recommended Mathematics Courses

- **Statistics APMA 1650 or CSCI 1450**
- **Linear Algebra MATH 0520, MATH 0540, CSCI 0530**

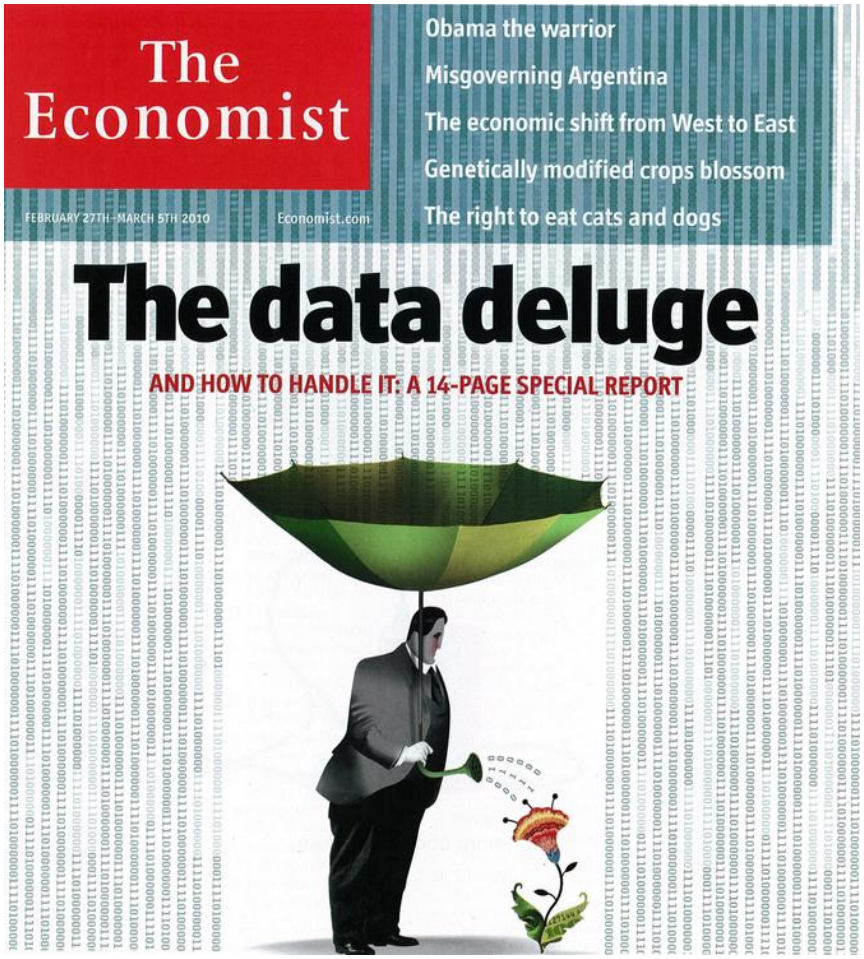
ACKNOWLEDGMENT

Introduction to Data Science **was originally developed by Prof. Tim Kraska.**

The course this year relies heavily on content he and his TAs developed last year and in prior offerings of the course.

If I have seen further, it is by standing on the shoulders of giants.

- Isaac Newton, 1676


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
FEBRUARY 27TH–MARCH 5TH 2010
Economist.com
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT


The World's Cheapest Car | 23 Hot Summer Gadgets
Get Ready for the Google Phone
WIRED
The End of Science
The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data. Welcome to the Petabyte Age.


Google

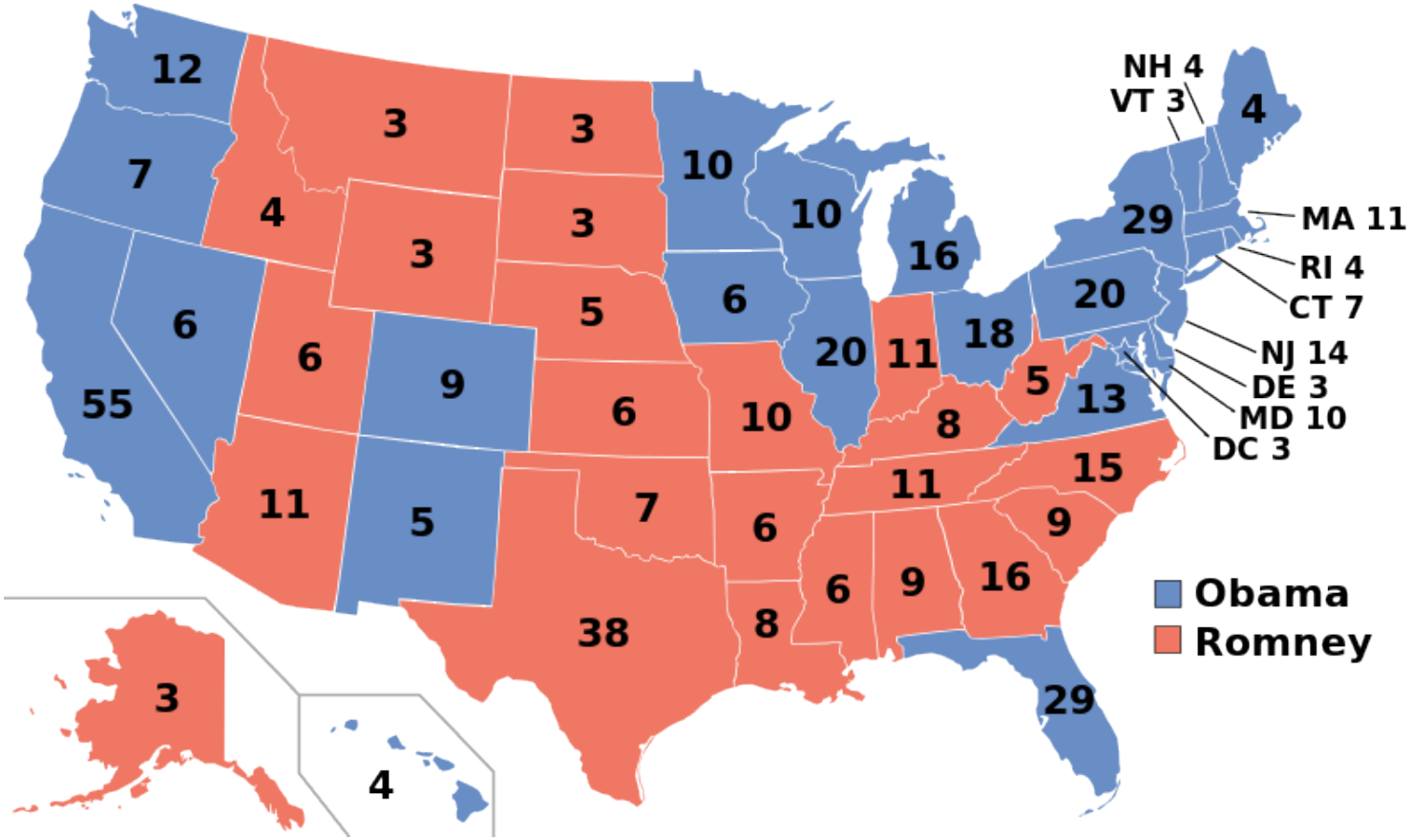
The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, Google


NETFLIX
MENU
tv

NATE SILVER

"Silver, who made his name by using cold hard math to call 49 out of 50 states in the 2008 general election and all 50 in 2012"



http://commons.wikimedia.org/wiki/File:ElectoralCollege2012.svg
(public domain)

Nate Silver

source: randy stewart

"The intuition behind this ought to be very simple: Mr. Obama is maintaining leads in the polls in Ohio and other states that are sufficient for him to win 270 electoral votes."

Nate Silver, Oct. 26, 2012

fivethirtyeight.com

"…the argument we're making is exceedingly simple. Here it is: Obama's ahead in Ohio."

Nate Silver, Nov. 2, 2012

fivethirtyeight.com

"The bar set by the competition was invitingly low. Someone could look like a genius simply by doing some fairly basic research into what really has predictive power in a political campaign."

Nate Silver, Nov. 10, 2012

DailyBeast

RELATED: OBAMA CAMPAIGN'S DATA-DRIVEN GROUND GAME

"In the 21st century, **the candidate with [the] best data**, merged with the best messages dictated by that data, **wins**."

Andrew Rasiej, Personal Democracy Forum

"…the **biggest win came from good old SQL** on a Vertica data warehouse and from providing access to data to dozens of analytics staffers who could follow their own curiosity and distill and analyze data as they needed."

Dan Woods
Jan 13 2013, CITO Research

"The decision was made to have **Hadoop** do the aggregate generations and anything not real-time, but then have Vertica to answer sort of 'speed-of-thought' queries about all the data."

Josh Hendler, CTO of H & K Strategies

Also a good read: http://fivethirtyeight.com/features/a-history-of-data-in-american-politics-part-2-obama-2008-to-the-present/

ELECTION 2016

"Donald Trump Is The Nickelback Of GOP Candidates"

"[d]isliked by most, super popular with a few"

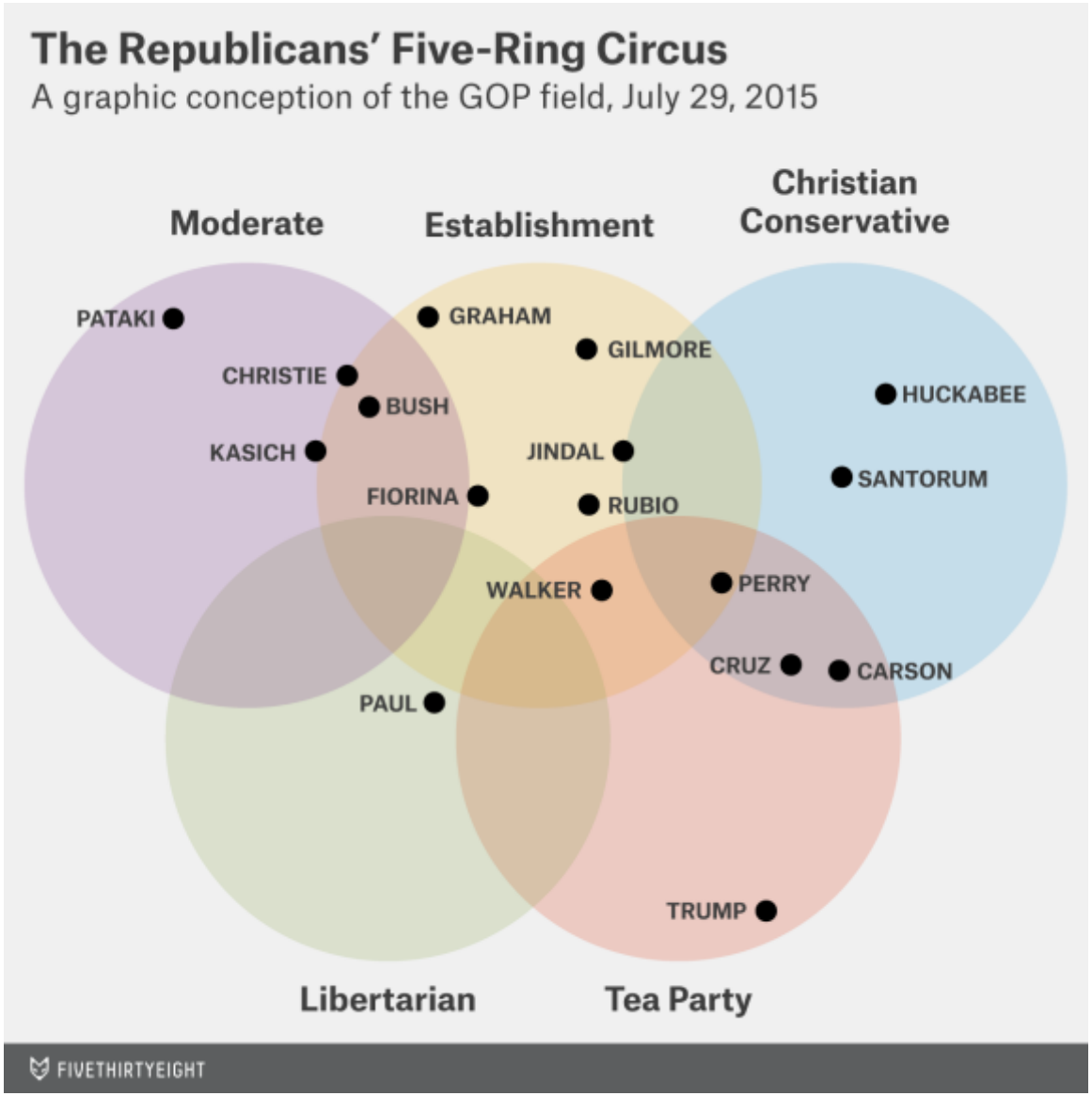
Trump Is The 13th Most Popular GOP Candidate

Average of national, Iowa and New Hampshire polls since July 18

	CANDIDATE	FAVORABLE	UNFAVORABLE	NET FAVORABLE	FIRST CHOICE
1	Walker	56%	13%	+43%	14%
2	Rubio	56%	16%	+39%	6%
3	Carson	50%	15%	+35%	7%
4	Jindal	45%	18%	+27%	2%
5	Fiorina	44%	17%	+27%	2%
6	Cruz	49%	23%	+27%	5%
7	Huckabee	52%	27%	+26%	5%
8	Perry	50%	25%	+25%	2%
9	Santorum	44%	28%	+16%	1%
10	Bush	50%	34%	+16%	12%
11	Paul	44%	30%	+14%	5%
12	Kasich	31%	17%	+14%	4%
13	Trump	47%	43%	+4%	20%
14	Christie	35%	47%	-12%	3%

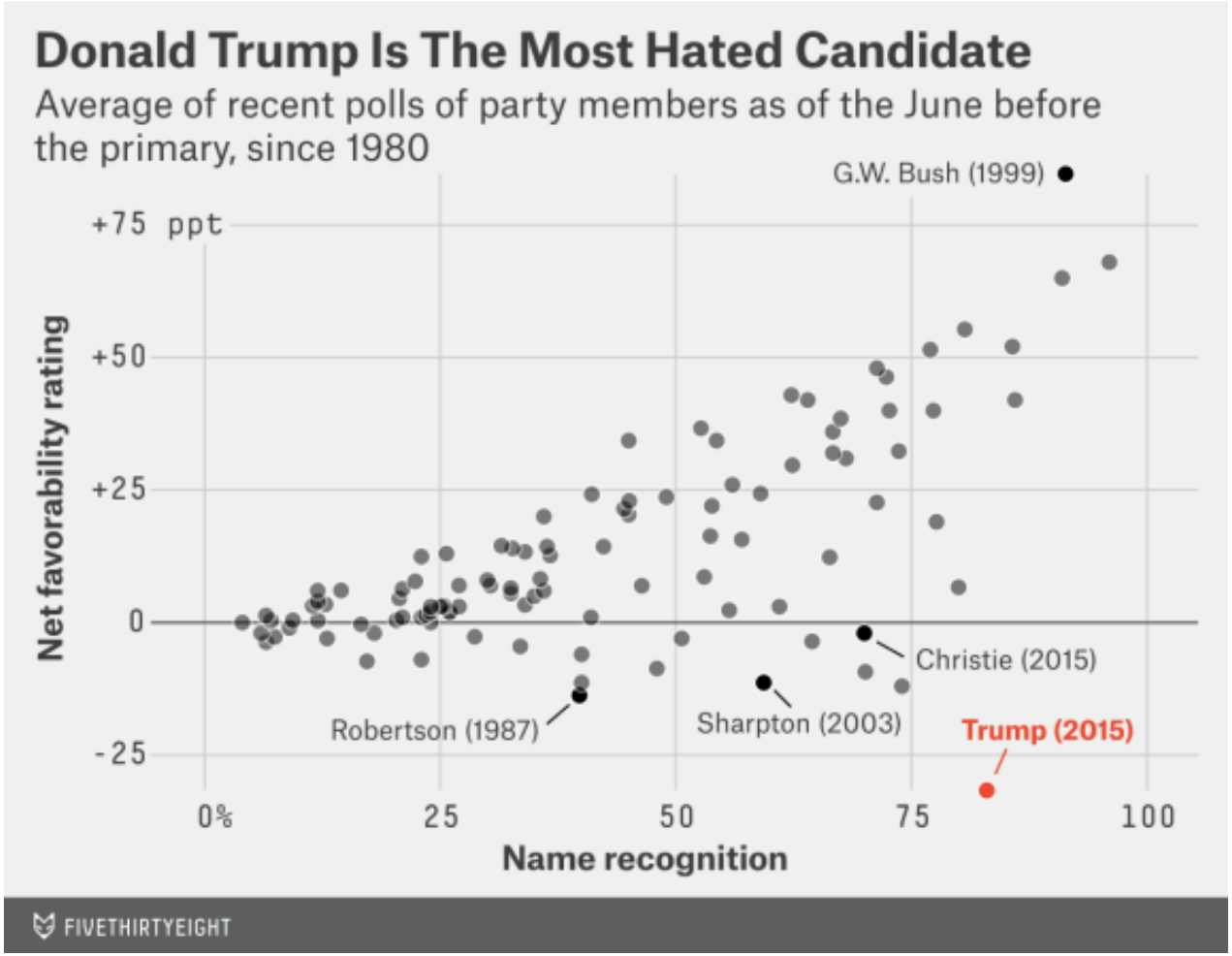
http://fivethirtyeight.com/datalab/donald-trump-is-the-nickelback-of-gop-candidates/

"HE'S HIGHLY DIFFERENTIATED FROM THE REST OF THE REPUBLICAN PACK."



ELECTION 2016

“Why Donald Trump Isn’t A Real Candidate, In One Chart” (June 2015)



http://fivethirtyeight.com/datalab/why-donald-trump-isnt-a-real-candidate-in-one-chart/

How Nate Silver Missed Donald Trump

The election guru said Trump had no shot. Where did he go wrong?

2.2k 318 980

By *Leon Neyfakh*

Polls whiz kid Nate Silver and presidential candidate Donald Trump.

Photo illustration by *Slate*. Images by Slaven Vlasic/Getty Images and Ethan Miller/Getty Images.

"If Silver's system depends largely on interpreting poll numbers, how reliable can that system be if the pre–Iowa and New Hampshire polls are basically worthless? **Garbage in, garbage out.**"

http://www.slate.com/articles/news_and_politics/politics/2016/01/nate_silver_said_donald_trump_had_no_shot_where_did_he_go_wrong.2.html

OCT 31, 2016 AT 8:40 PM

The Odds Of An Electoral College-Popular Vote Split Are Increasing

By Nate Silver

Filed under 2016 Election

Donald Trump at Macomb Community College on Monday in Warren, Michigan. CARLOS OSORIO / AP

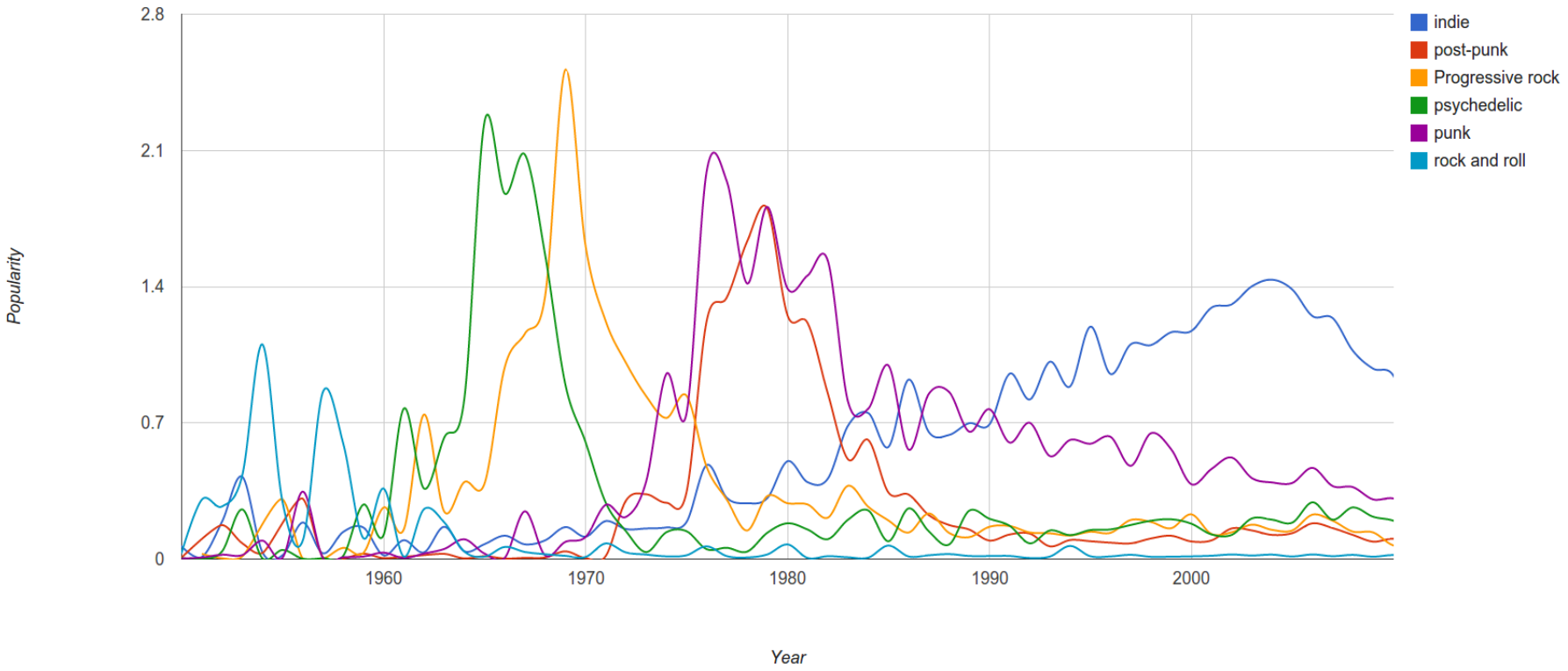
https://fivethirtyeight.com/features/the-odds-of-an-electoral-college-popular-vote-split-are-increasing/

http://fivethirtyeight.com/features/why-fivethirtyeight-gave-trump-a-better-chance-than-almost-anyone-else/

Why FiveThirtyEight Gave Trump A Better Chance Than Almost Anyone Else

Late in the cycle careful analysis based on state by state poll analysis showed that there could be a problem. Silver was virtually the only pollster to make note of this.

LAST.FM



"Since we have a massive amount of user tag data available we can easily correlate tags and years and measure "popularity" of a genre by counting the number of artists formed in a specific year."

Janni Kovacs, Last.FM

EXPRESSION OF EMOTIONS OVER THE 20TH CENTURY

1) Convert all the digitized books in the 20th century into n-grams
(Thanks, Google!)
(http://books.google.com/ngrams/)

A 1-gram: "yesterday"
A 5-gram: "analysis is often described as"

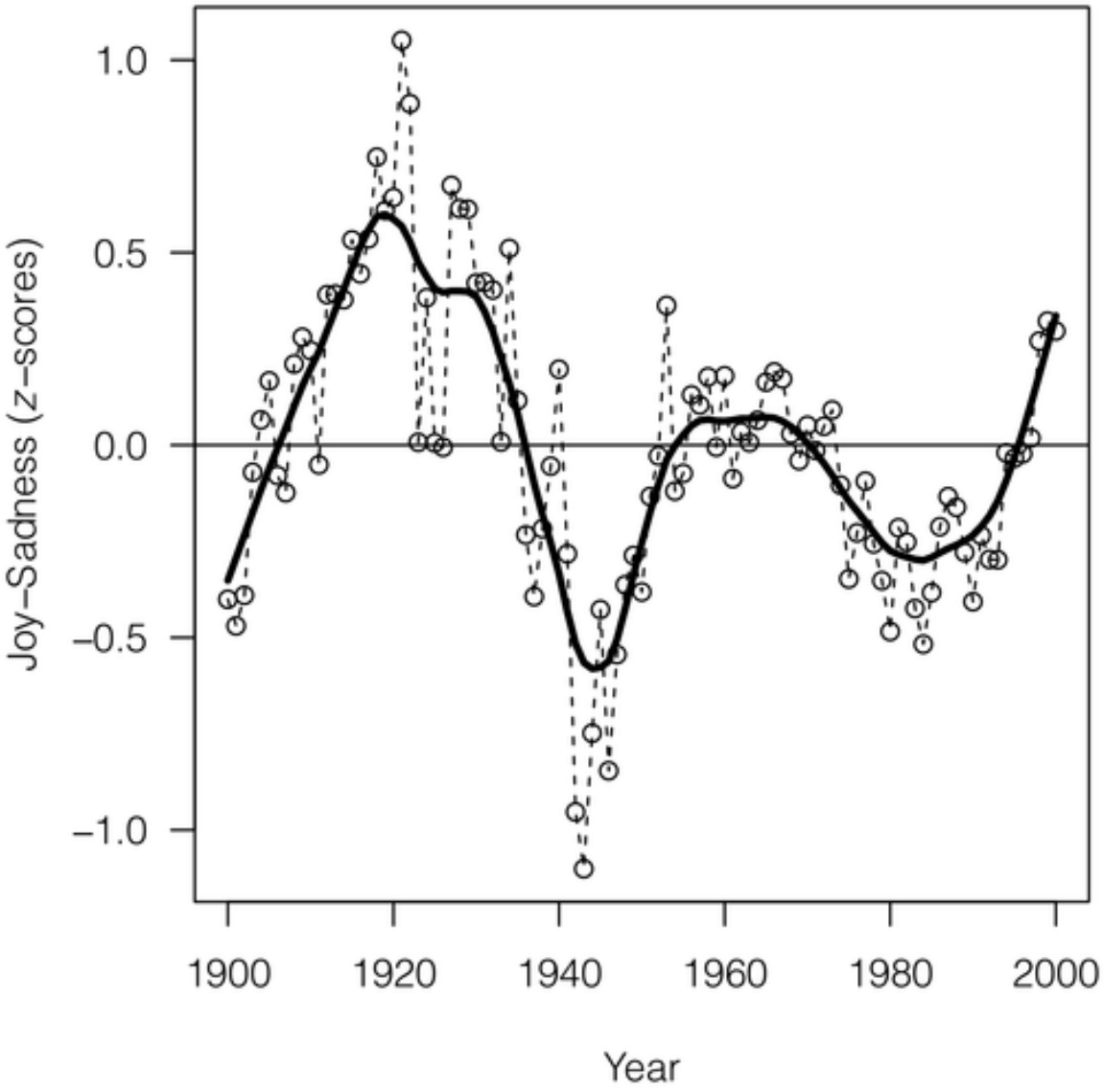
2) Label each 1-gram (word) with a mood score.
(Thanks, WordNet Affect)

3) Count the occurences of each mood word

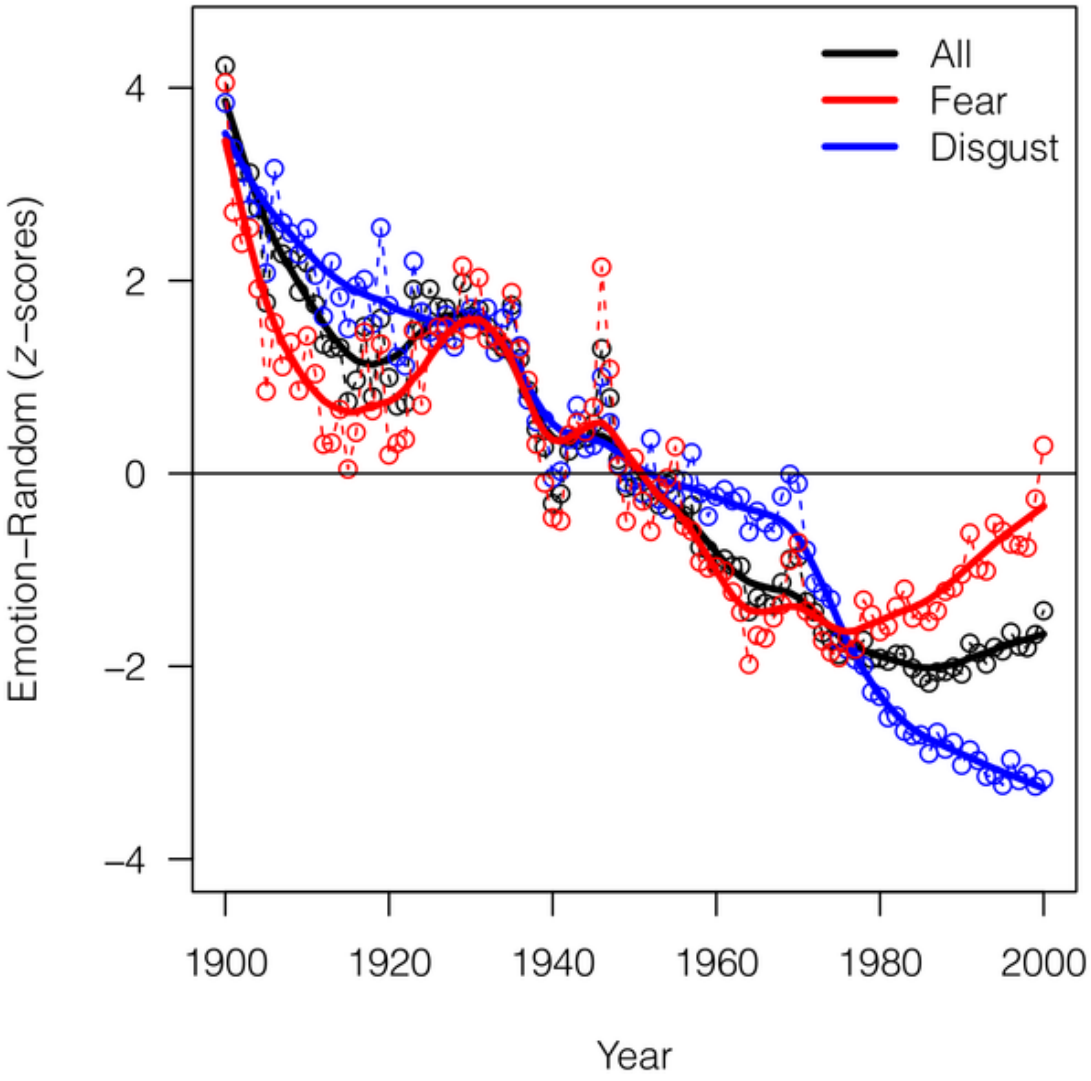
$$\mathcal{M}_Y = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{C_{\text{the}}},$$

$$\mathcal{M}z_Y = \frac{\mathcal{M}_Y - \mu_{\mathcal{M}}}{\sigma_{\mathcal{M}}},$$

Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030



Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030



Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030

PAPERS CITED BY THEM

…

2. Michel J-P, Shen YK, Aiden AP, Veres A, Gray MK, et al. (2011) ***Quantitative analysis of culture using millions of digitized books***. Science 331: 176–182. doi: 10.1126/science.1199644. Find this article online

3. Lieberman E, Michel J-P, Jackson J, Tang T, Nowak MA (2007) ***Quantifying the evolutionary dynamics of language***. Nature 449: 713–716. doi: 10.1038/nature06137. Find this article online

4. Pagel M, Atkinson QD, Meade A (2007) ***Frequency of word-use predicts rates of lexical evolution throughout Indo-European history***. Nature 449: 717–720. doi: 10.1038/nature06176. Find this article online

…

6. DeWall CN, Pond RS Jr, Campbell WK, Twenge JM (2011) ***Tuning in to Psychological Change: Linguistic Markers of Psychological Traits and Emotions Over Time in Popular U.S. Song Lyrics***. Psychology of Aesthetics, Creativity and the Arts 5: 200–207. doi: 10.1037/a0023195. Find this article online

…

Flavor network and the principles of food pairing

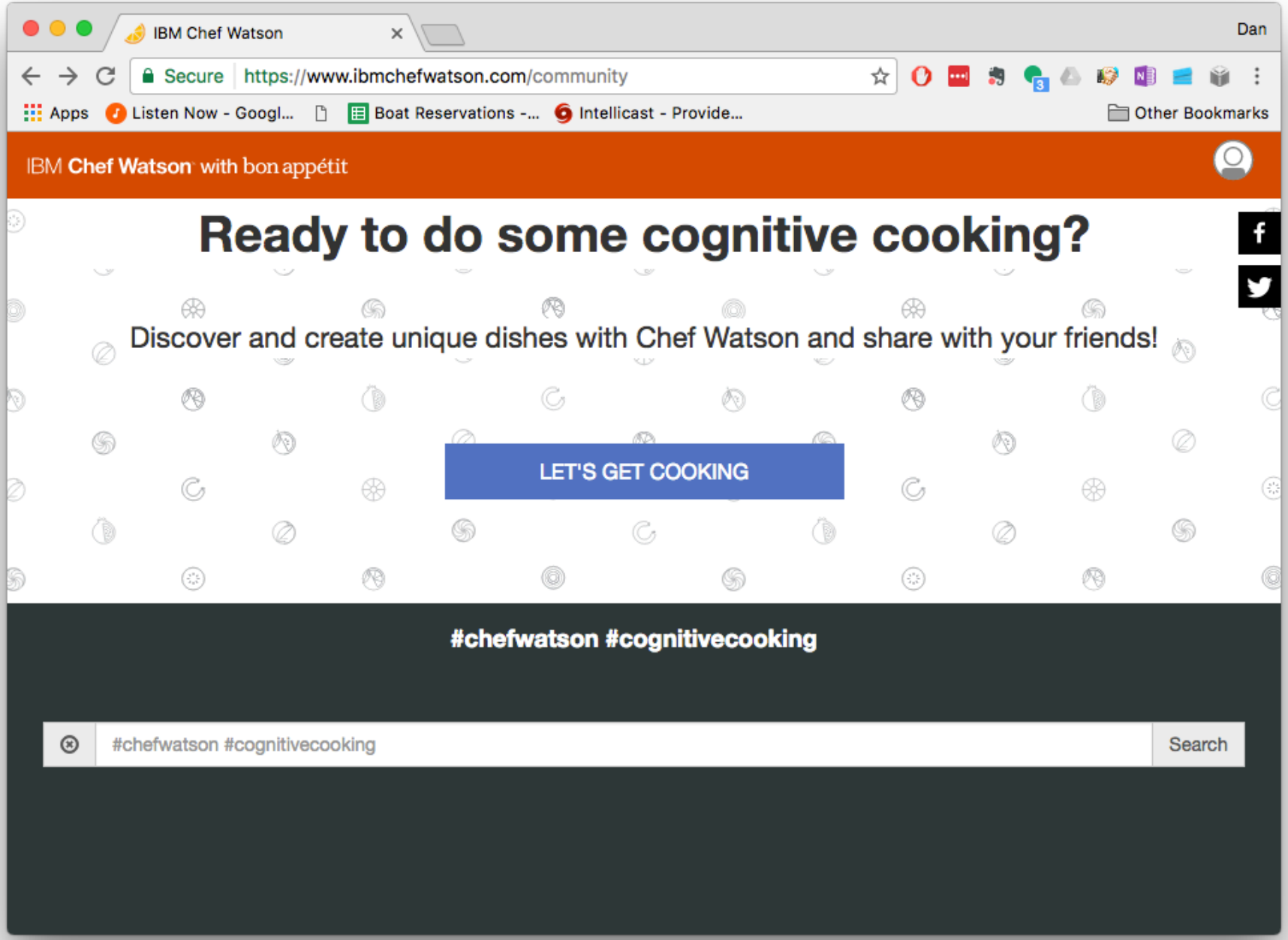
Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow & Albert-László Barabási

Affiliations | **Contributions** | **Corresponding authors**

Scientific Reports **1**, Article number: 196 | doi:10.1038/srep00196

Received 18 October 2011 | Accepted 24 November 2011 | Published 15 December 2011

Idea: Analyze the co-occurrence graph of ingredients in recipes to analyze the underlying principles of food pairing.


IBM Chef Watson
Dan
Secure
https://www.ibmchefwatson.com/community
Apps
Listen Now - Googl...
Boat Reservations -...
Intellicast - Provide...
Other Bookmarks
IBM Chef Watson with bon appétit
Ready to do some cognitive cooking?
Discover and create unique dishes with Chef Watson and share with your friends!
LET'S GET COOKING
#chefwatson #cognitivecooking
#chefwatson #cognitivecooking
Search

TRENDS + NEWS

We Spent a Year Cooking With the World's Smartest Computer—and Now You Can, Too

ONLINE EXAMPLES

Cooking with Watson

https://www.ibmchefwatson.com/community

Google Flu Datahttps://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac

3rd Party Google Flu Data in D3

http://stat4701-edav-d3.github.io/viz/cities/cities.html

Global Burden of Disease in D3

http://www.healthdata.org/gbd/data-visualizations

CS1951-A PROJECTS

Alexander Bertsch, HTA

ANATOMY OF A PROJECT

DOLLA DOLLA BILLS Y'ALL

CSCI1951A: Data Science Spring 2016 Final Project

Adam Hoff, Angelia Wang, Chris Grimm, Athyuttam Eleti

Explored ~~Wu Tang discography~~ potential arbitrage between Amazon and eBay

ANATOMY OF A PROJECT

Explored a hypothesis: Amazon and eBay aren't optimal markets and thus have arbitrage. Can we identify and predict arbitrage possibilities and item prices?

Used services like CamelCamelCamel to get Amazon and eBay time series price data

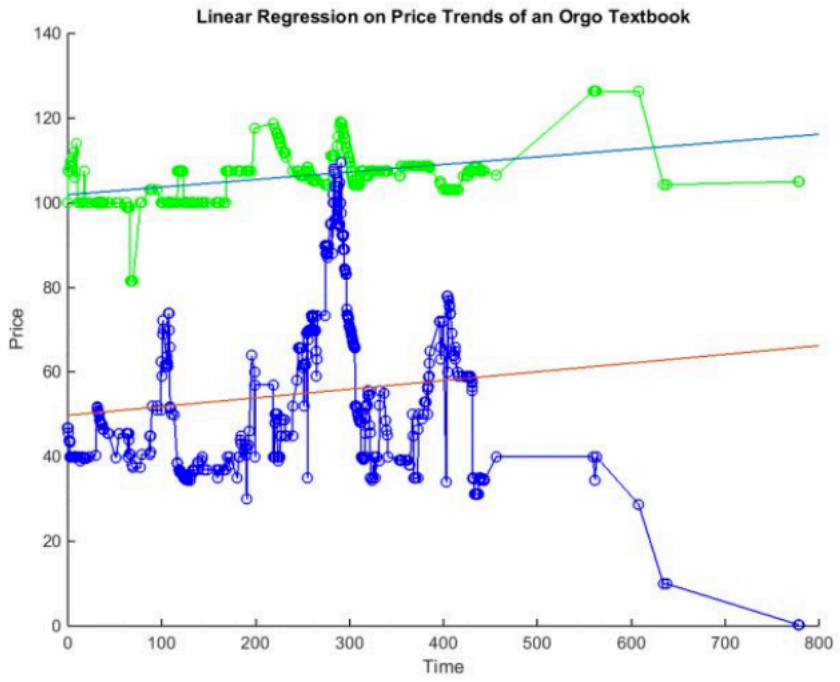
Examined both Auctions and Buy It Now on eBay

Had to build integration system that paired Amazon and eBay items

amazon.com ebay

ANATOMY OF A PROJECT

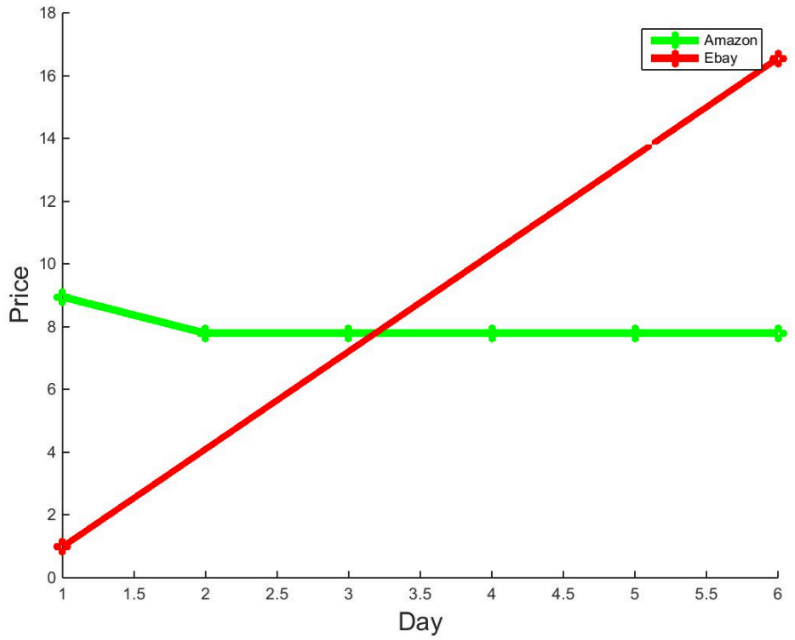
Used a linear regression to predict prices and model price differences



ANATOMY OF A PROJECT

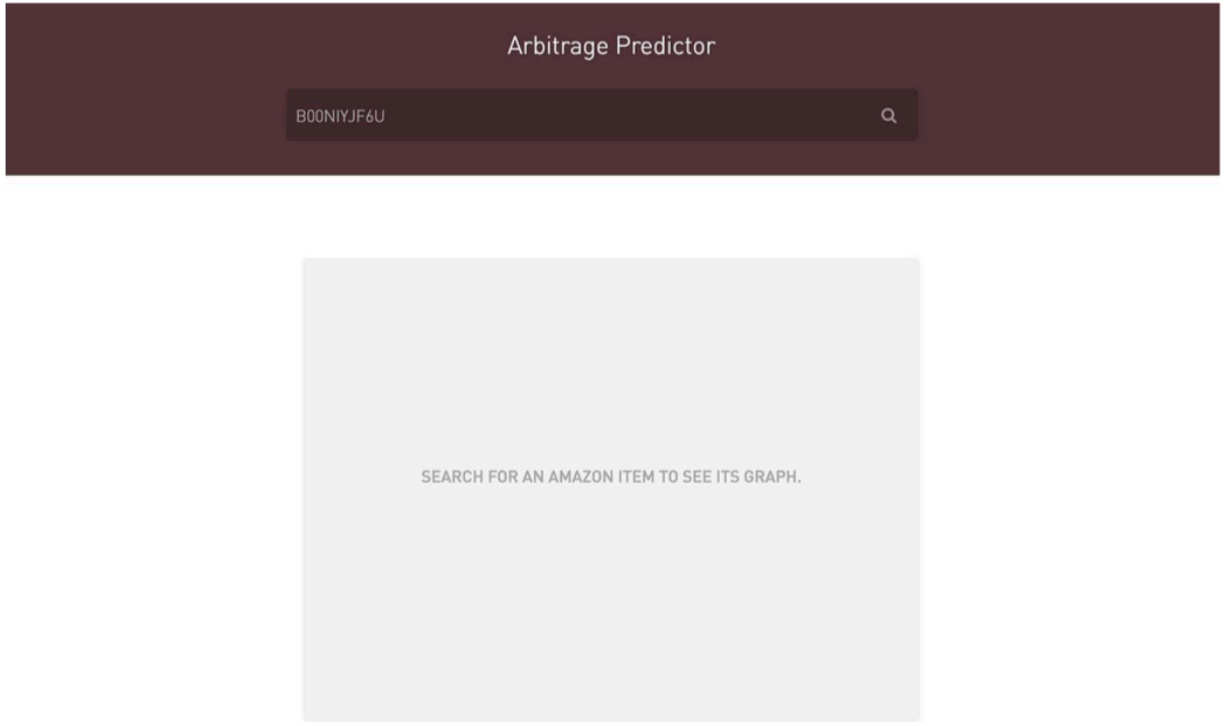
Built an arbitrage predictor with staggering 90% accuracy

Could predict future opportunities from past data



ANATOMY OF A PROJECT

As part of the capstone, built a system to enter Amazon items and predict future prices



PROJECT

- **There is no final exam!**
- **Instead, you'll work in groups of four to build a project**
- **Your goal: dive deep into a dataset**
- **Topic of your choice!**
- **In two parts:**
  - Pre-project
  - Final project

PRE-PROJECT

- **Goal: get ready to explore a dataset**
- **Begins with a pre-proposal, due mid February**
- **You'll be paired with a mentor TA**
- **Midterm report will show your progress so far**
  - Should have access to data and be partially cleaned
- **Get started early! (Can't stress this enough)**

FINAL PROJECT

- **Built on top of your pre-project**
- **Goal: demonstrate that you master the data pipeline from cleaning, model building, to presenting the result by taking a data set and deriving some *interesting* insight**
- **Weekly updates in the form of blog posts about your progress, insights, tools, etc. (We grade these.)**

CAPSTONE

For a capstone: build an dynamic data pipeline, preferably as a desktop or web app

More details to be released soon!

SOME RECURRING THEMES

simple methods

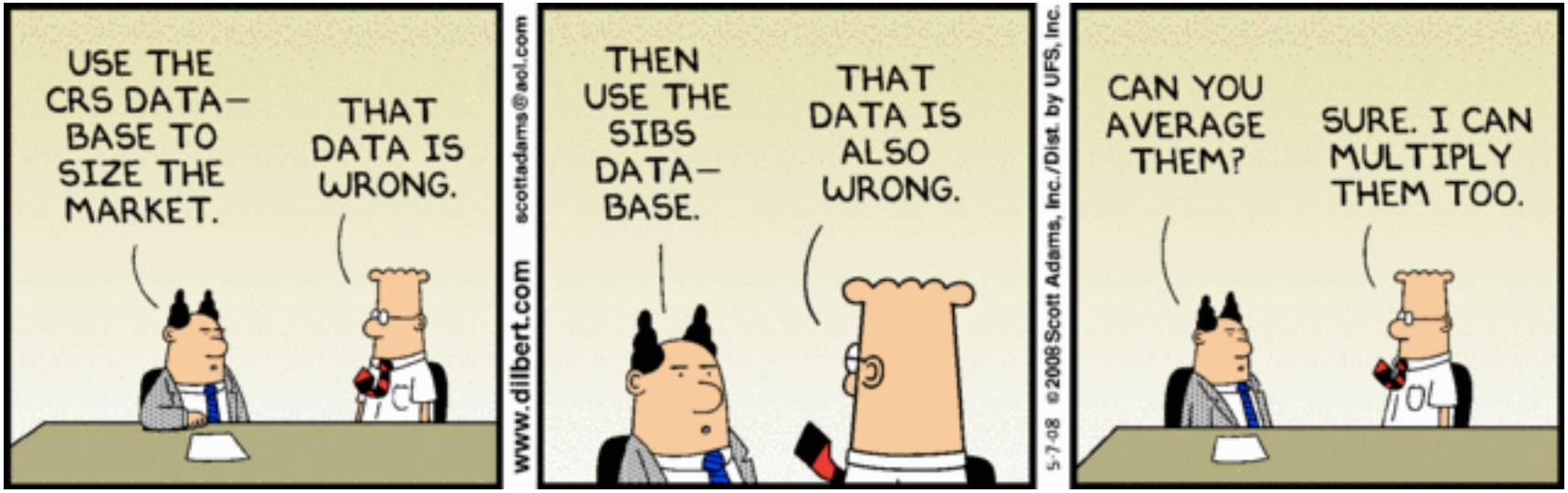
repurposing data

communication matters

bad data / bad assumptions

Other themes

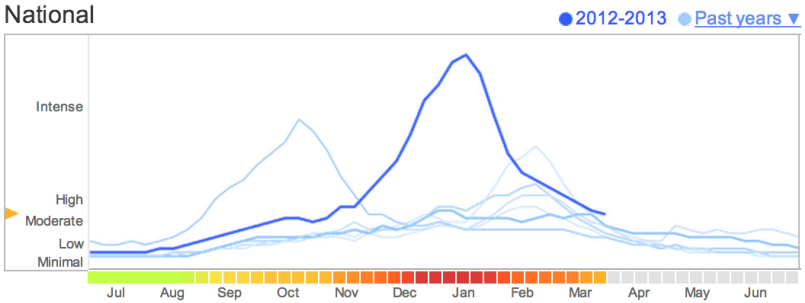
- "Data products" – not just answers
- "Speed of thought" analysis


USE THE CRS DATA-BASE TO SIZE THE MARKET.
THAT DATA IS WRONG.
www.dilbert.com scottadams@aol.com
THEN USE THE SIBS DATA-BASE.
THAT DATA IS ALSO WRONG.
5-7-08 ©2008 Scott Adams, Inc./Dist. by UFS, Inc.
CAN YOU AVERAGE THEM?
SURE. I CAN MULTIPLY THEM TOO.

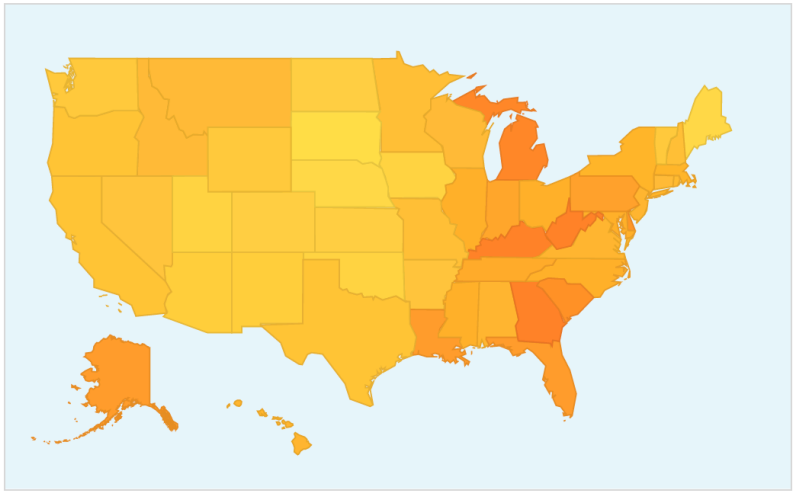
Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

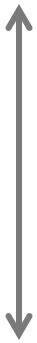
National



States | Cities (Experimental)



Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through March 30, 2013.



flu risk

"Scientific hindsight shows that Google Flu Trends far overstated this year's flu season…."

"Lots of media attention to this year's flu season skewed Google's search engine traffic."

David Wagner, Atlantic Wire, Feb 13 2013

source: http://www.google.org/flutrends/us/#US

A CLASS EXPERIMENT

Hypothesis:

Parents believe more often that their sons are gifted than their daughters

Parents care more about the appearance of their daughters than sons

RESULTS

Is my son overweight?

For every **10** U.S. Google queries about boys being overweight . . .

Is my daughter overweight?

. . . there are **17** about girls.

(In reality, boys are about 9 percent more likely to be overweight than girls.)

Is my daughter gifted?

For every **10** U.S. Google queries about girls being gifted . . .

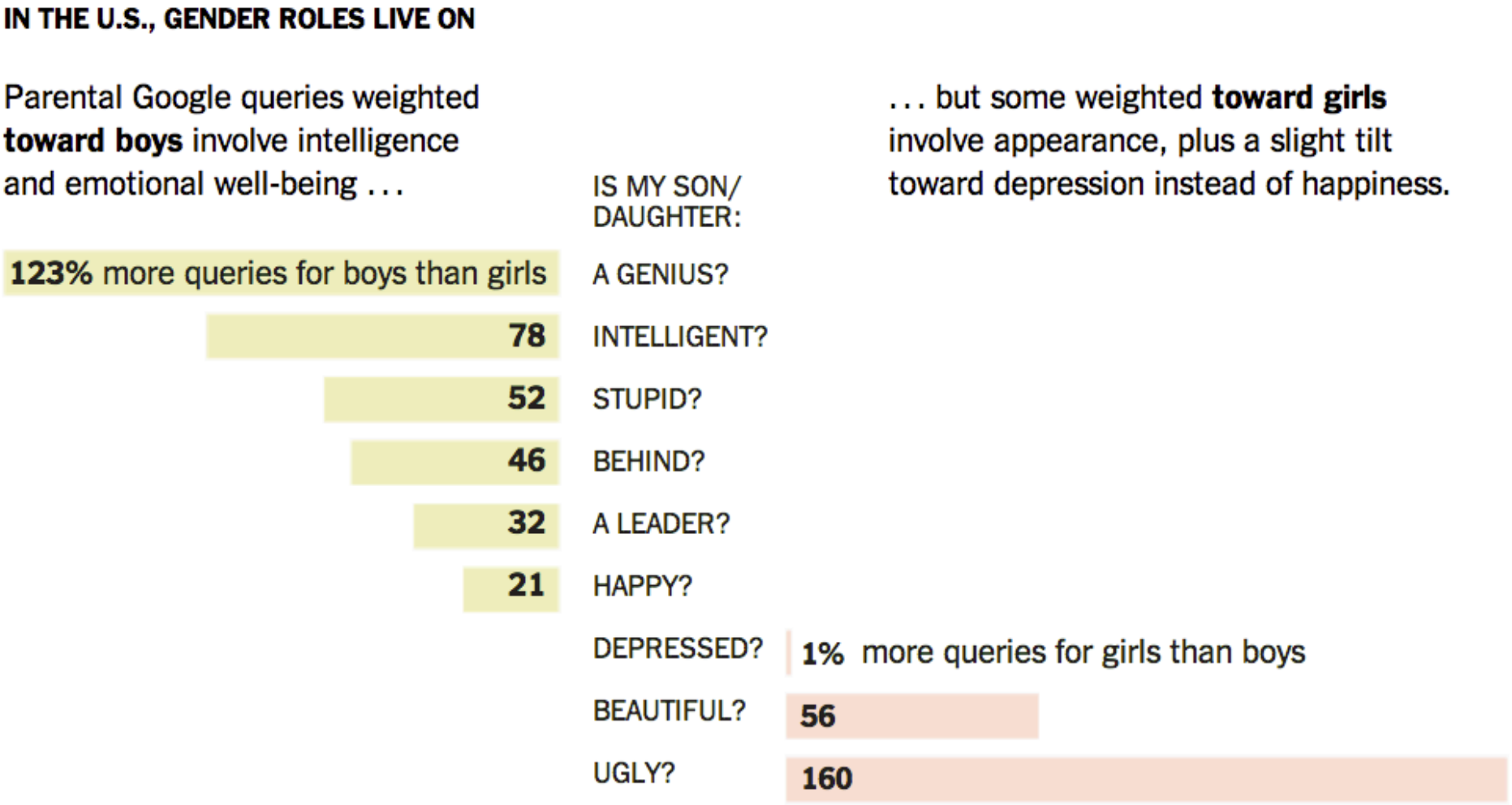
Is my son gifted?

. . . there are **25** about boys.

(In reality, girls are about 11 percent more likely to be in a gifted program.)

http://www.nytimes.com/2014/01/19/opinion/sunday/google-tell-me-is-my-son-a-genius.html?smid=fb-share&_r=1

RESULTS CTD



http://www.nytimes.com/2014/01/19/opinion/sunday/google-tell-me-is-my-son-a-genius.html?smid=fb-share&_r=1

NOT ENOUGH DATA TO DRAW CONCLUSIONS

WHAT IS DATA SCIENCE?

"Data Scientists:
The Definition of Sexy"

Forbes, 2012

"Data Scientist: The Sexiest
Job of the 21st Century"

Harvard Business Review, 2012

WHAT IS DATA SCIENCE?

Fortune

- "Hot New Gig in Tech"

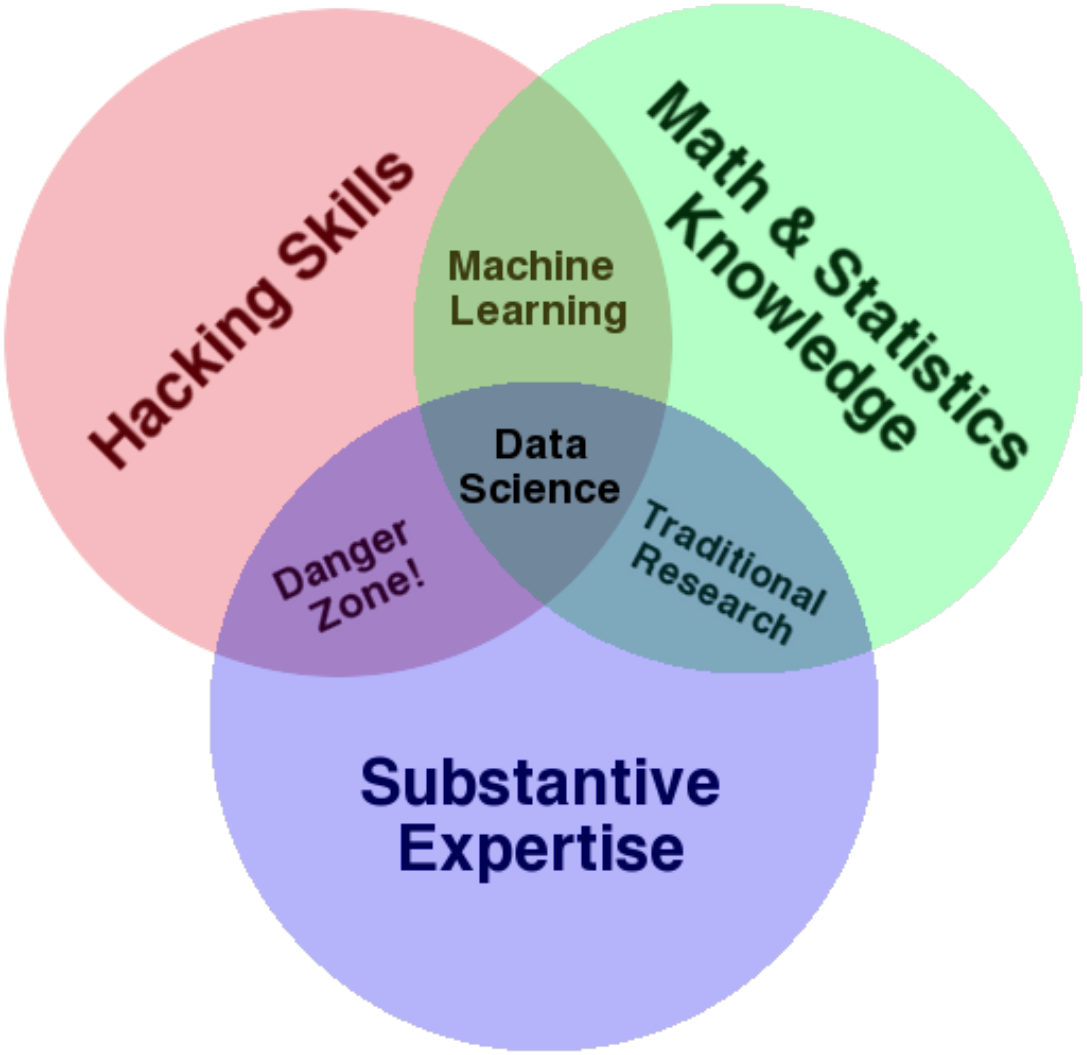
Hal Varian, Google's Chief Economist, NYT, 2009:

- "The next sexy job"
- "The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill."

Mike Driscoll, CEO of metamarkets:

- "Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics."
- "Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools & materials, coupled with a theoretical understanding of what's possible."

DREW CONWAY'S DATA SCIENCE VENN DIAGRAM



WHAT DO DATA SCIENTISTS DO?

"They need to find nuggets of truth in data and then explain it to the business leaders"

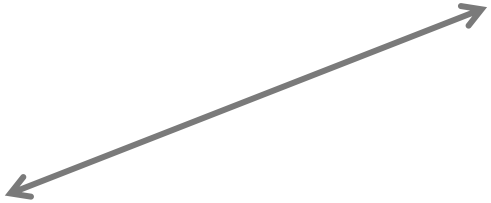
-- Richard Snee, EMC

Data scientists "tend to be "hard scientists", particularly physicists, rather than computer science majors. Physicists have a strong mathematical background, computing skills, and come from a discipline in which survival depends on getting the most from the data. They have to think about the big picture, the big problem."

-- DJ Patil, Chief Scientist at LinkedIn

MIKE DRISCOLL'S THREE SEXY SKILLS OF DATA GEEKS

"data wrangling"
"data jujitsu"
"data munging"



Data Wrangling

- parsing, scraping, and formatting data

Statistics

- traditional analysis

Visualization

- graphs, tools, etc.

"Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management and preservation of large collections of information."

An Introduction to Data Science
Jeffrey Stanton
Syracuse University School of Information Studies

"A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product."

-- Hilary Mason, chief scientist at bit.ly

DOING DATA SCIENCE (PETER HUBER)

1. **Inspection**
2. **Error checking**
3. **Modification**
4. **Comparison**
5. **Modeling and model fitting**
6. **Simulation**
7. **What-if analyses**
8. **Interpretation**
9. **Presentation of conclusions**

[http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf]

DOING DATA SCIENCE (BEN FRY)

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

[http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf]

DOING DATA SCIENCE (COLIN MALLOWS)

1. **Identify data to collect and its relevance to your problem**
2. **Statistical specification of the problem**
3. **Method selection**
4. **Analysis of method**
5. **Interpret results for non-statisticians**

[http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf]

A PRACTICAL DEFINITION

Data Science is about the whole processing pipeline to extract information out of data

Data Scientist understand and care about the whole data pipeline

A data pipeline consists of 3 steps:

1) **Preparing to run a model**

   Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping

2) **Running the model**

3) **Communicating the results**

DATA SCIENCE IS ABOUT *DATA PRODUCTS*

- "Data-driven apps"
  - Spellchecker
  - Machine Translator
- Interactive visualizations
  - Google flu application
  - Global Burden of Disease
- Online Databases
  - Enterprise data warehouse
  - Sloan Digital Sky Survey

Data science is about building data products, not just answering questions

Data products empower others to use the data.

May help communicate your results (e.g., Nate Silver's maps)

May empower others to do their own analysis (e.g., Global Burden of Disease)

(Mike Loukides)

DISTINGUISHING DATA SCIENCE FROM…

Business Intelligence

Statistics

Data(base) Management

Visualization

Machine Learning

HUGE NUMBER OF RELEVANT AREAS

Stochastic/Statistics

Machine Learning

Databases

Distributed Systems

Networking

Cloud Computing

Natural Language Processing

Visualization

…

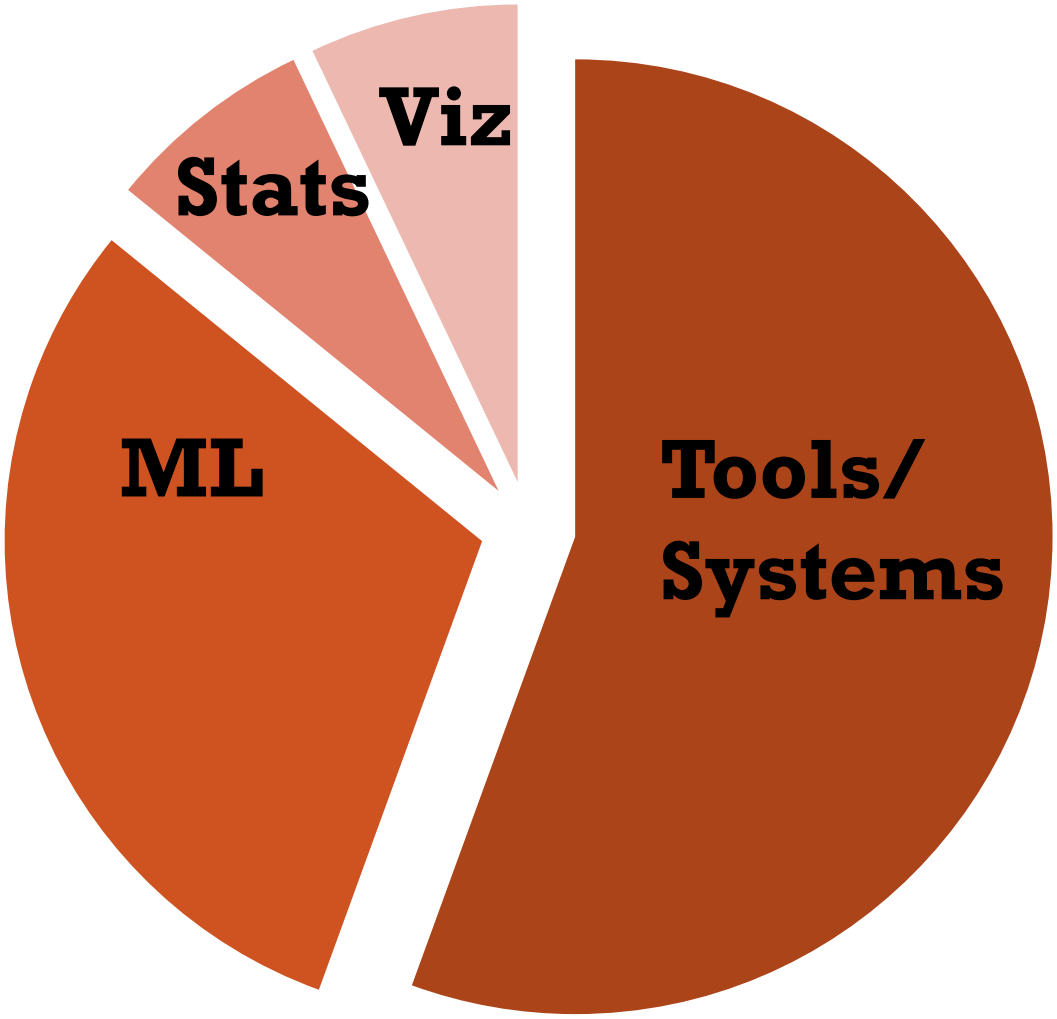
"I worry that the Data Scientist role is like the mythical "webmaster" of the 90s: master of all trades."

-- Aaron Kimball, CTO Wibidata

CS1951A

WHAT DOES IT MEAN FOR THIS COURSE

THIS COURSE



TOPICS

Databases

- Relational Model
- SQL
- Column Stores
- Compression Techniques

Modeling Tools (R/Matlab/Python)

Data Integration

- Data Wrangling
- Overview & Quality Control
- Entity Resolution

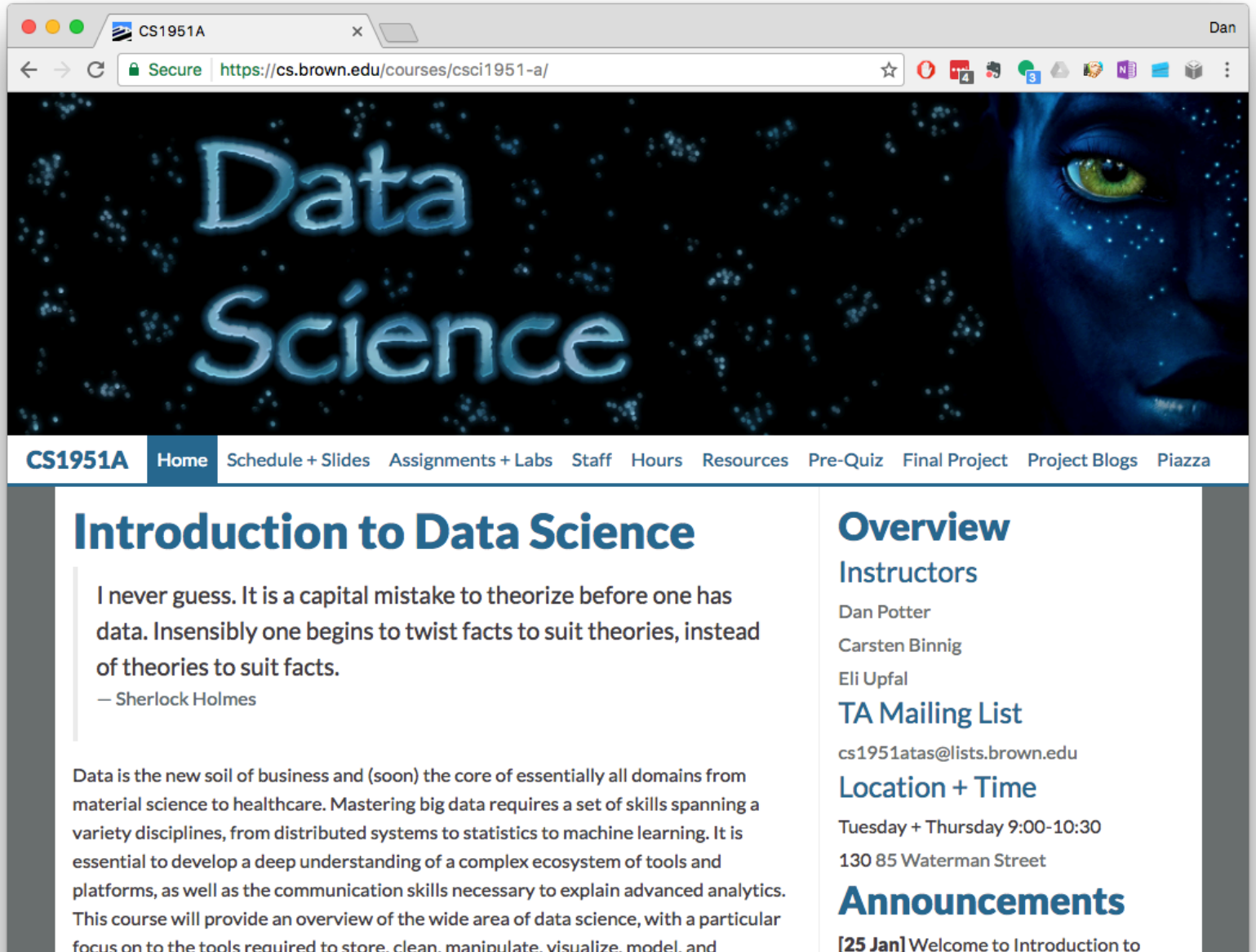
Large-Scale Analytic Frameworks

- Map/Reduce
- Languages for Hadoop
- Spark
- Scope & Reef
- NoSQL

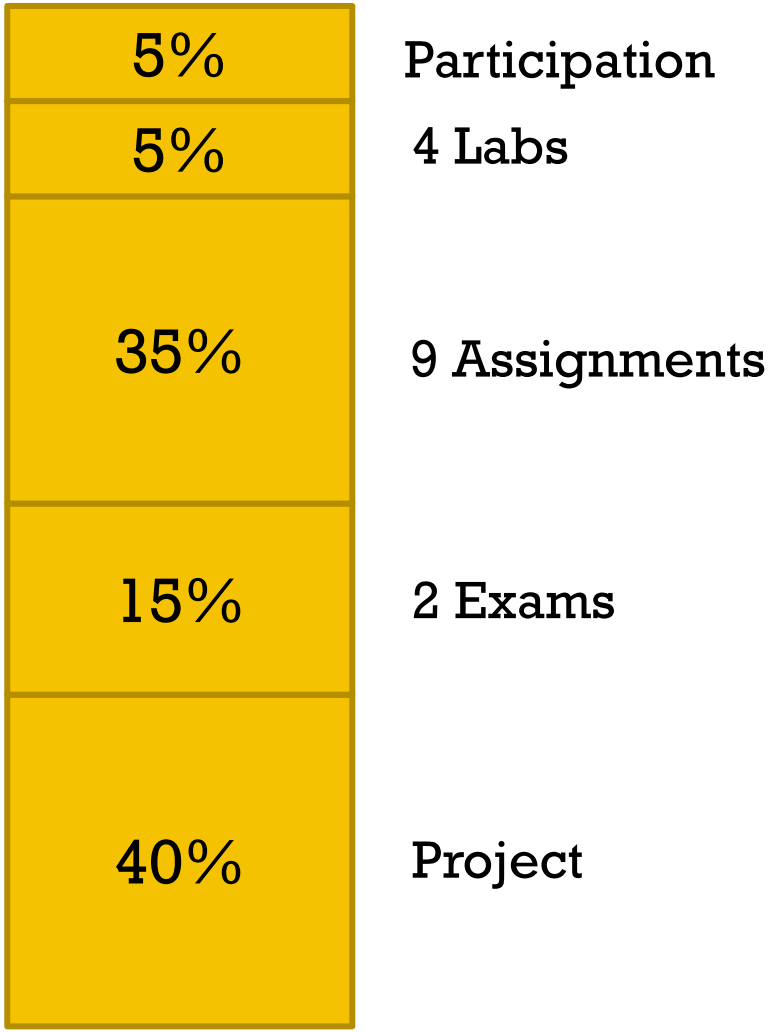
- **Statistics & Visualization**
  - Statistics Basics
  - Testing
  - Visualization Basics
  - Lying with statistics
- **Data Mining / Machine Learning**
  - Clustering (K-MEANS)
  - Classification (Regression, kNN)
  - Frequent Itemsets (Apriori)
  - Decision Tree (C4.5)
  - Graph Mining (PageRank)
  - Naive Bayes
  - Boosting
  - PCA
- **Search Engines**
  - Inverted Indexes
  - Crawling Infrastructure

COURSE WEB-PAGE

http://cs.brown.edu/courses/csci1951-a



DELIVERABLES / GRADING GUIDELINE



PRE-QUIZ (NO CREDIT)

https://cs.brown.edu/courses/csci1951-a/prequiz/

Part 1: Python
Part 2: Statistics

This is a self-assessment for your benefit

THE "LITTLE" EXAMS (15%)

2 of them:

3/7/2017

(data management, text-processing, visualization)

4/27/2017

(statistics, machine-learning)

LABS (5%)

Topics:

- Potpourri
  - Python
  - Alphabet Soup
  - Bash
- MapReduce
- D3
- ML Training

All labs are located in the Sunlab (CIT 143)

4 groups with roughly 30 students each. We'll post sign-up instructions on Piazza.

Lab Policy:

- **Attending the actual lab is optional, however turning in the lab assignment is not.** NOTE: If you haven't attended your lab, the TAs will not answer questions over email or during TA hours, except in extenuating circumstances.
- Students can attend any lab, but if the lab is full the priority is given to the students of the group (i.e., you might have to leave)

ASSIGNMENTS (35%)

- **Data Modeling**
- **SQL + Relational Algebra**
- **Data Integration + Data Warehousing**
- **MapReduce**
- **Visualization using D3**
- **Statistics**
- **Machine Learning**
- **Data Mining**
- **Regression**

LATE DAY POLICY

3 late days on any assignments, excluding the final project and labs.

Once these three late days have been exhausted, **you will receive NO credit for your late hand-in.**

piazza

Sign-up page:

piazza.com/brown/spring2017/cs1951a

Always, always search piazza first. Somebody might have already asked the same question.

If not, ask the question first on piazza publicly unless it concerns your project or reveals too much of a solution.

TA HOURS

All hours will be held in the Moonlab, CIT 277.

Monday, Tuesday 6—8:

Wednesday 6—10

Thursday 7—11

Policy: You should always ask questions on piazza first (even before going to TA hours), unless it is specific to your project

INSTRUCTOR HOURS

Please visit us for questions about our respective lectures. See Dan for special situations/circumstances. Questions about assignments labs and your project should **always first go to piazza and afterwards to the TAs** during the office hours. Only if you are still not happy with the answer, come to us.

PROJECT

There is no final exam, but essentially 2 projects!!!

Groups of 4 students (match-making on piazza)

Pre-Project (2/8 – 3/21) (10%)

- Goal: Get you started to explore a data set
- Mid-term

Final-Project (4/2 – 5/12) (30%)

- Can but does not have to build on the pre-project
- Goal: Demonstrate that you master the data pipeline from cleaning, model building, to presenting the result by taking a data set and deriving some *interesting* insight.
- You will write three blog posts (e.g., on wordpress) about your progress, interesting insights, tools you learned, etc. The blog posts will be also graded

Topic: Up to you

i>clicker®

- You can get one from the Friedman Center in the Sciences Library.
- Don't forgot to register your clicker in **Canvas** (do not use the iclicker web-page)
- Clicker participation counts 4% towards the class, but answers are not graded for correctness
- If you are asked to pay a fee, you did something wrong!!!

NEXT STEPS

Do the pre-quiz and decide yourself if have the right background

Read/Watch:

- Alon Halevy, Peter Norvig, and Fernando Pereira: The Unreasonable Effectiveness of Data
- /talks/view/lang/en//id/788
- http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html
- http://www.ted.com/talks/gary_flake_is_pivot_a_turning_point_for_web_exploration.html