STAT121 / AC209 / E-109
CS109 Data Science

Rafael A. Irizarry
rafa@jimmy.harvard.edu

Verena Kaynig-Fittkau
vkaynig@seas.harvard.edu

Today

- What is Data Science?
- Why learn Data Science?
- How do we learn Data Science?
- Who is helping you learn Data Science?

20th Century Innovation

Engineering and Computer Science played key role

- Cars
- Airplanes
- Power grid
- Television
- Air conditioning and central heating
- Nuclear power
- Digital computers
- The internet

For more:
http://camdp.com/blogs/21st-century-problems

But how about these 20th Century questions?

- Does fertilizer increase crop yields?
- Does Streptomycin cure Tuberculosis?
- Does smoking cause lung-cancer?

What is the difference?

What is the difference?

- Deterministic versus random
- Deductive versus empirical
- Solutions deduced mostly from theory versus solutions deduced from mostly from **data**

Data

- Does fertilizer increase crop yields? Answer: Collect and analyze agricultural experimental **data**
- Does Streptomycin cure Tuberculosis? Collect and analyze randomized trials **data**
- Does smoking cause lung-cancer? Collect and analyze observational studies **data**

Analyzing these was the job of: boring ol' **statisticians**

21st Century

21st century

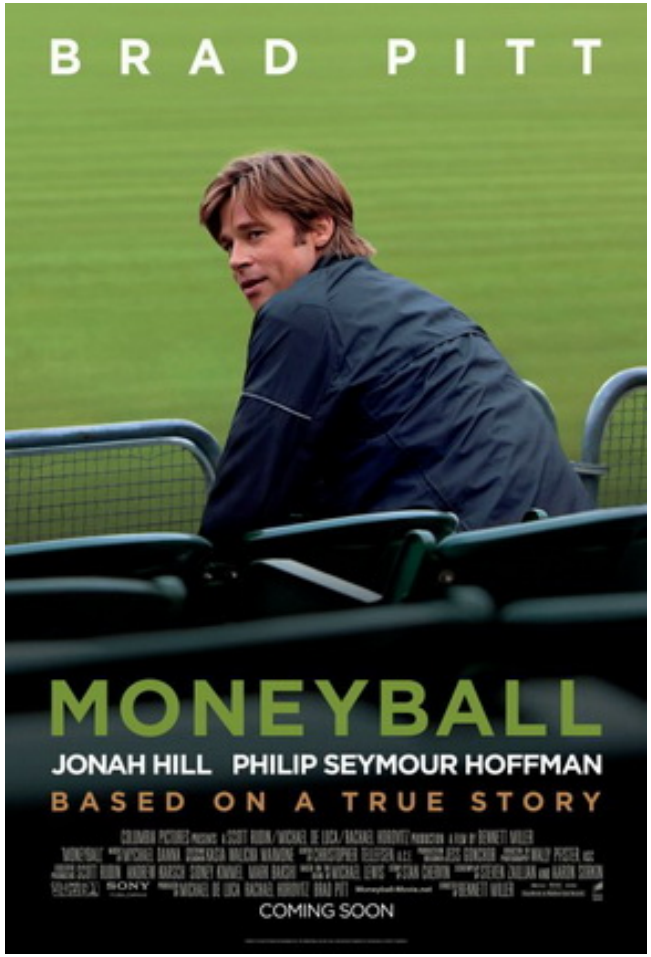
“I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?”
- Hal Varian, Google’s Chief Economist

Hal Varian Explains...

"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's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data."

– Hal Varian

Data Science Success Stories

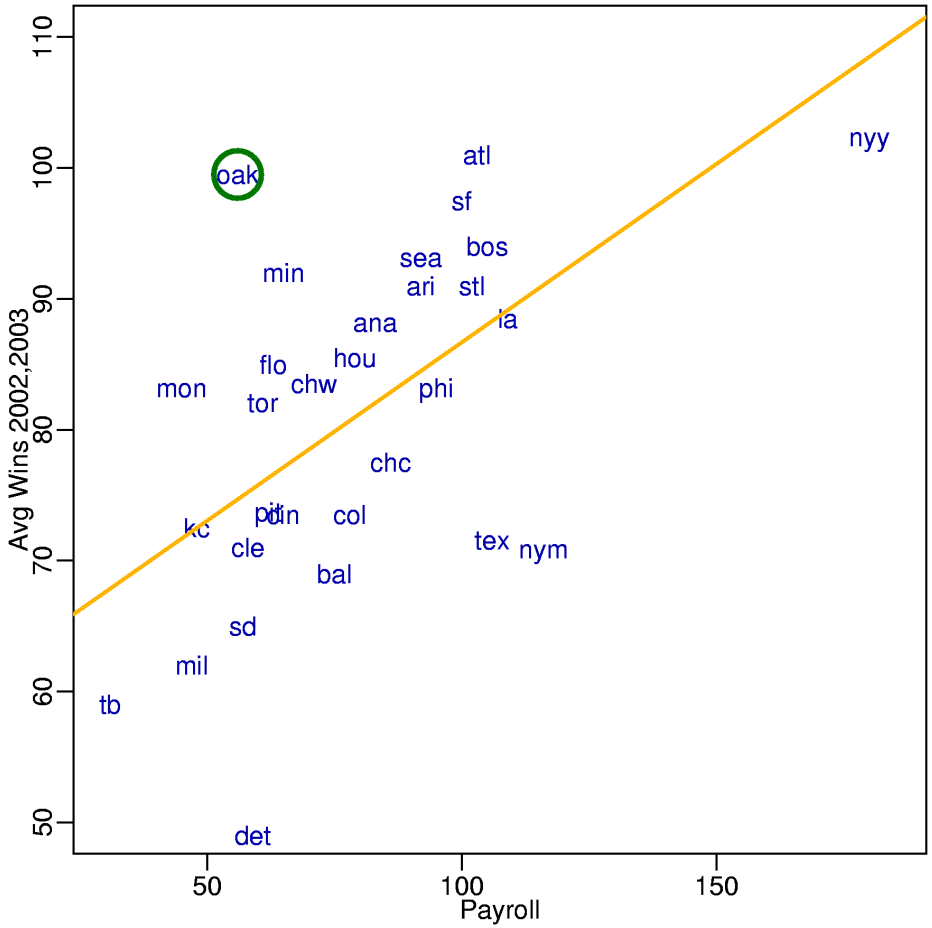

BRAD PITT
MONEYBALL
JONAH HILL PHILIP SEYMOUR HOFFMAN
BASED ON A TRUE STORY
COMING SOON

The Data Scientist

Actual

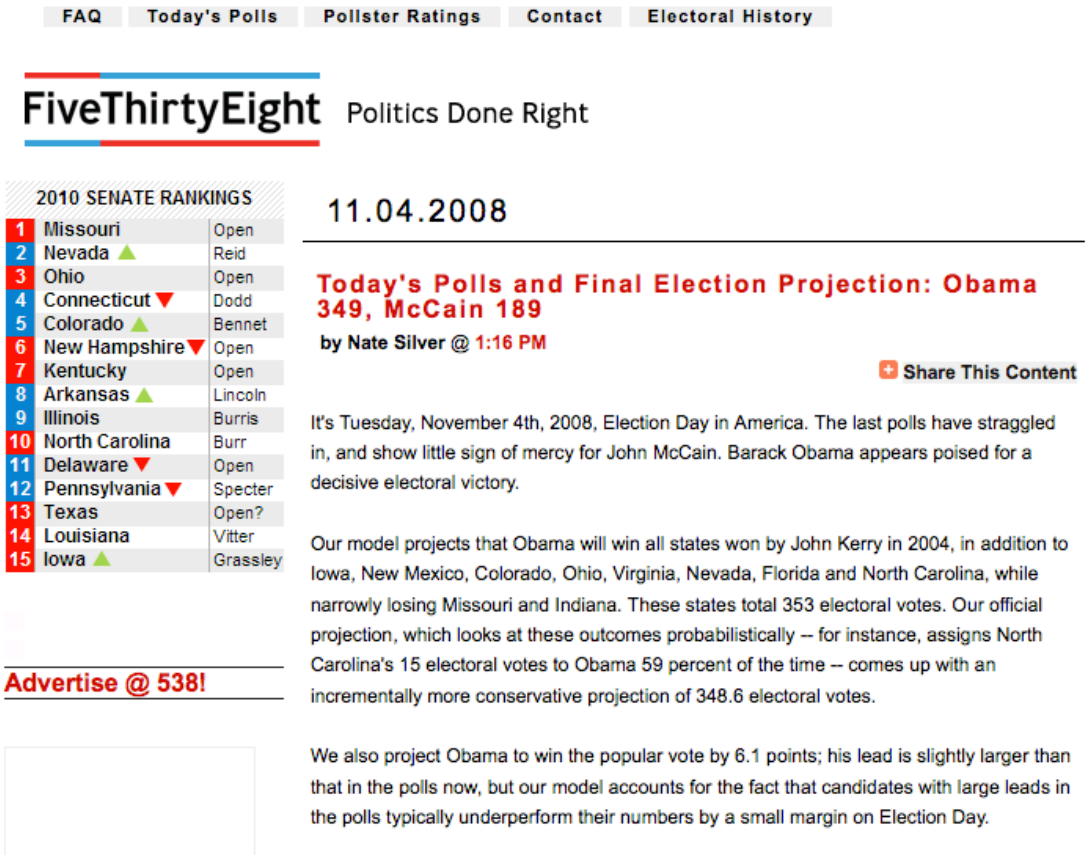
Hollywood

Money Ball



Starting around 2001, the Oakland A's picked players that scouts thought were no good but data said otherwise

“Nate Silver won the election” – Harvard Business Review



FAQ | Today's Polls | Pollster Ratings | Contact | Electoral History

FiveThirtyEight Politics Done Right

2010 SENATE RANKINGS

1	Missouri	Open
2	Nevada ▲	Reid
3	Ohio	Open
4	Connecticut ▼	Dodd
5	Colorado ▲	Bennet
6	New Hampshire ▼	Open
7	Kentucky	Open
8	Arkansas ▲	Lincoln
9	Illinois	Burris
10	North Carolina	Burr
11	Delaware ▼	Open
12	Pennsylvania ▼	Specter
13	Texas	Open?
14	Louisiana	Vitter
15	Iowa ▲	Grassley

Advertise @ 538!

11.04.2008

Today's Polls and Final Election Projection: Obama 349, McCain 189

by Nate Silver @ 1:16 PM

Share This Content

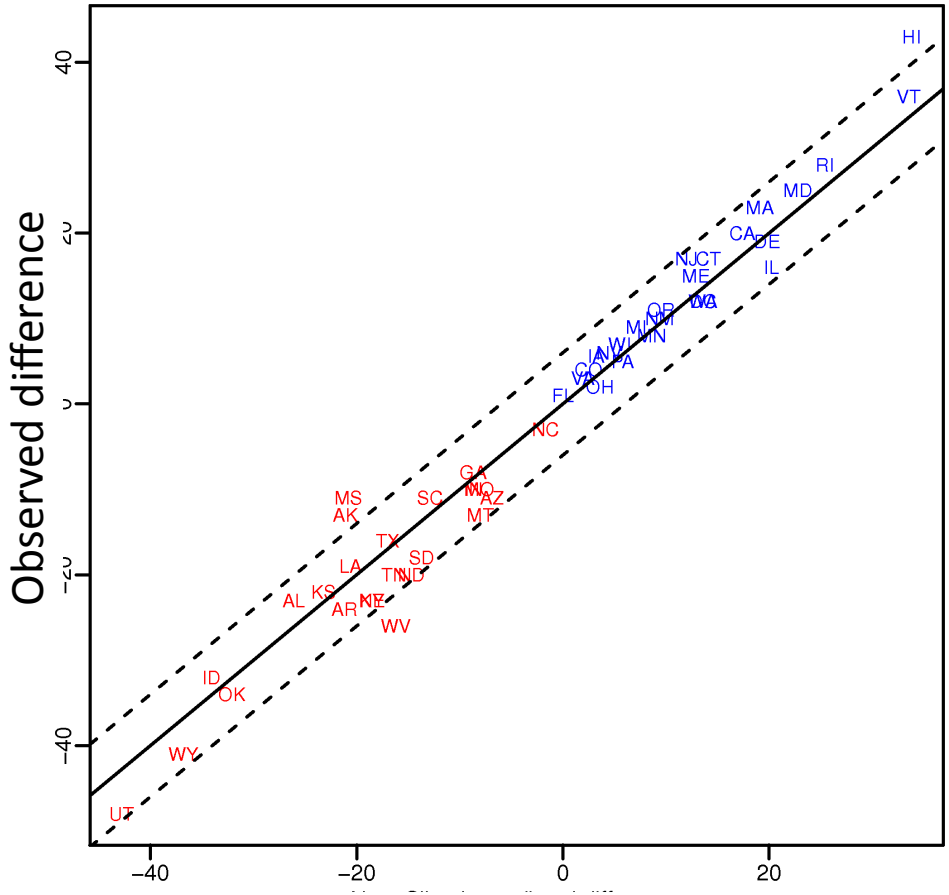
It's Tuesday, November 4th, 2008, Election Day in America. The last polls have straggled in, and show little sign of mercy for John McCain. Barack Obama appears poised for a decisive electoral victory.

Our model projects that Obama will win all states won by John Kerry in 2004, in addition to Iowa, New Mexico, Colorado, Ohio, Virginia, Nevada, Florida and North Carolina, while narrowly losing Missouri and Indiana. These states total 353 electoral votes. Our official projection, which looks at these outcomes probabilistically -- for instance, assigns North Carolina's 15 electoral votes to Obama 59 percent of the time -- comes up with an incrementally more conservative projection of 348.6 electoral votes.

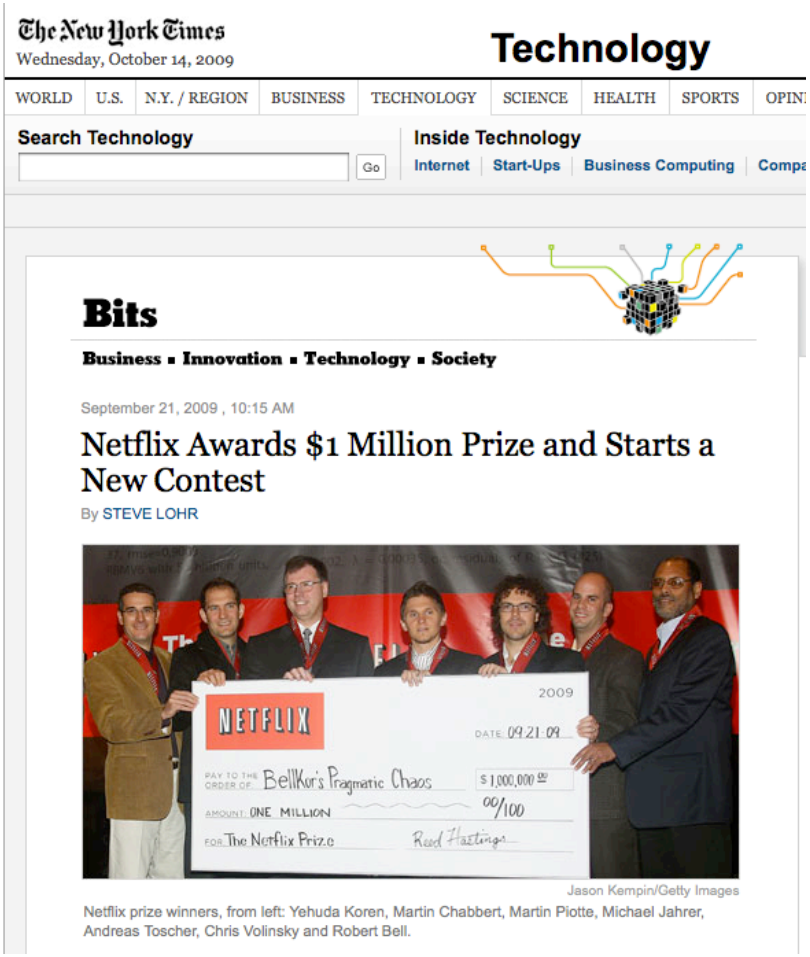
We also project Obama to win the popular vote by 6.1 points; his lead is slightly larger than that in the polls now, but our model accounts for the fact that candidates with large leads in the polls typically underperform their numbers by a small margin on Election Day.

Prediction: 349 to 189, 6.1% difference.
Actual: 365 to 173, 7.2% difference

2012 results



Netflix Challenge



The New York Times
Wednesday, October 14, 2009

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPIN

Search Technology

Go

Inside Technology

Internet | Start-Ups | Business Computing | Compa

Bits

Business ▪ Innovation ▪ Technology ▪ Society

September 21, 2009 , 10:15 AM

Netflix Awards $1 Million Prize and Starts a New Contest

By STEVE LOHR

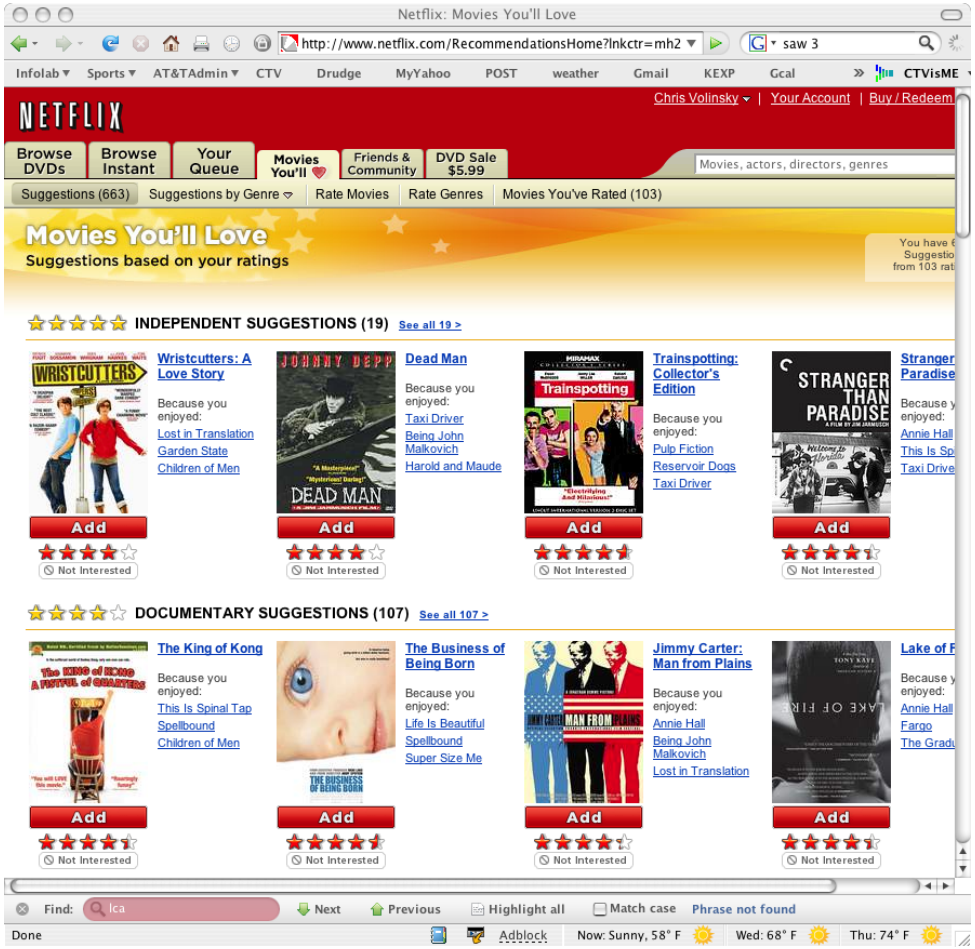
Jason Kempin/Getty Images

Netflix prize winners, from left: Yehuda Koren, Martin Chabbert, Martin Piotte, Michael Jahrer, Andreas Toscher, Chris Volinsky and Robert Bell.

In Sept 2009 a team lead by Chris Volinsky from Statistics Research AT&T Research was announced as winner!

Netflix

- A US-based DVD rental-by mail company
- >10M customers, 100K titles, ships 1.9M DVDs per day



Good recommendations = happy customers

Courtesy of Chris Volinsky

Netflix Prize

- October, 2006:
  - Offers **$1,000,000** for an improved recommender algorithm

- Training data
  - 100 million ratings
  - 480,000 users
  - 17,770 movies
  - 6 years of data: 2000-2005

user	movie	score	date
1	21	1	2002-01-03
1	213	5	2002-04-04
2	345	4	2002-05-05
2	123	4	2002-05-05
2	768	3	2003-05-03
3	76	5	2003-10-10
4	45	4	2004-10-11
5	568	1	2004-10-11
5	342	2	2004-10-11
5	234	2	2004-12-12
6	76	5	2005-01-02
6	56	4	2005-01-31

- Test data
  - Last few ratings of each user (2.8 million)
  - Evaluation via RMSE: root mean squared error
  - Netflix Cinematch RMSE: 0.9514

- Competition
  - **$1 million** grand prize for **10% improvement**
  - If 10% not met, $50,000 annual "Progress Prize" for best improvement

Courtesy of Chris Volinsky

Netflix Prize

- October, 2006:
  - Offers **$1,000,000** for an improved recommender algorithm

- Training data
  - 100 million ratings
  - 480,000 users
  - 17,770 movies
  - 6 years of data: 2000-2005

- Test data
  - Last few ratings of each user (2.8 million)
  - Evaluation via RMSE: root mean squared error
  - Netflix Cinematch RMSE: 0.9514

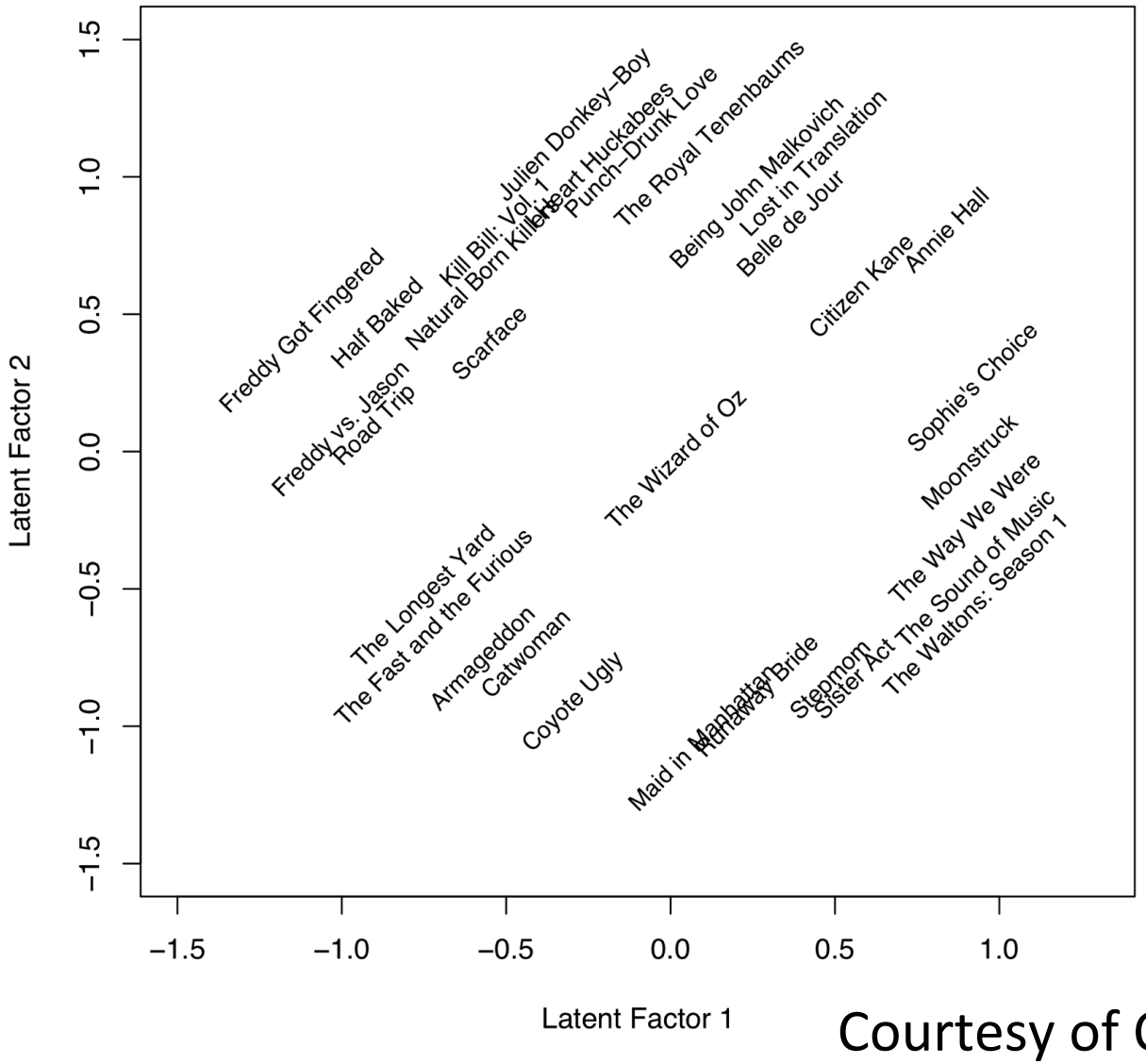
user	movie	score	date
1	21	1	2002-01-03

user	movie	score	date
1	212	?	2003-01-03
1	1123	?	2002-05-04
2	25	?	2002-07-05
2	8773	?	2002-09-05
2	98	?	2004-05-03
3	16	?	2003-10-10
4	2450	?	2004-10-11
5	2032	?	2004-10-11
5	9098	?	2004-10-11
5	11012	?	2004-12-12
6	664	?	2005-01-02
6	1526	?	2005-01-31

- Competition
  - **$1 million** grand prize for **10% improvement**
  - If 10% not met, $50,000 annual “Progress Prize” for best improvement

Courtesy of Chris Volinsky

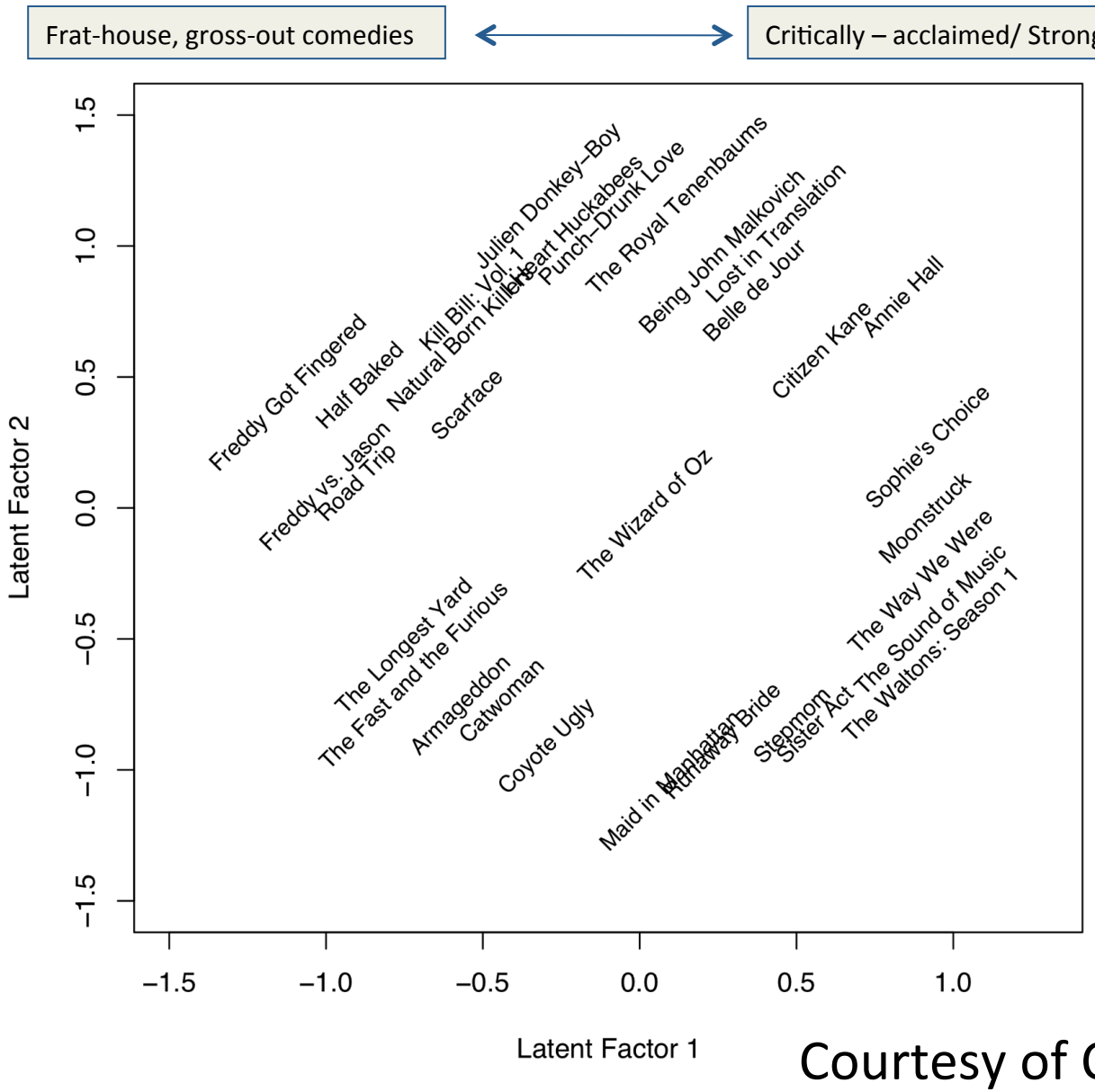
Latent Factors Model



A **latent factors** model identifies factors with maximum discrimination between movies

Courtesy of Chris Volinsky

Latent Factors Model



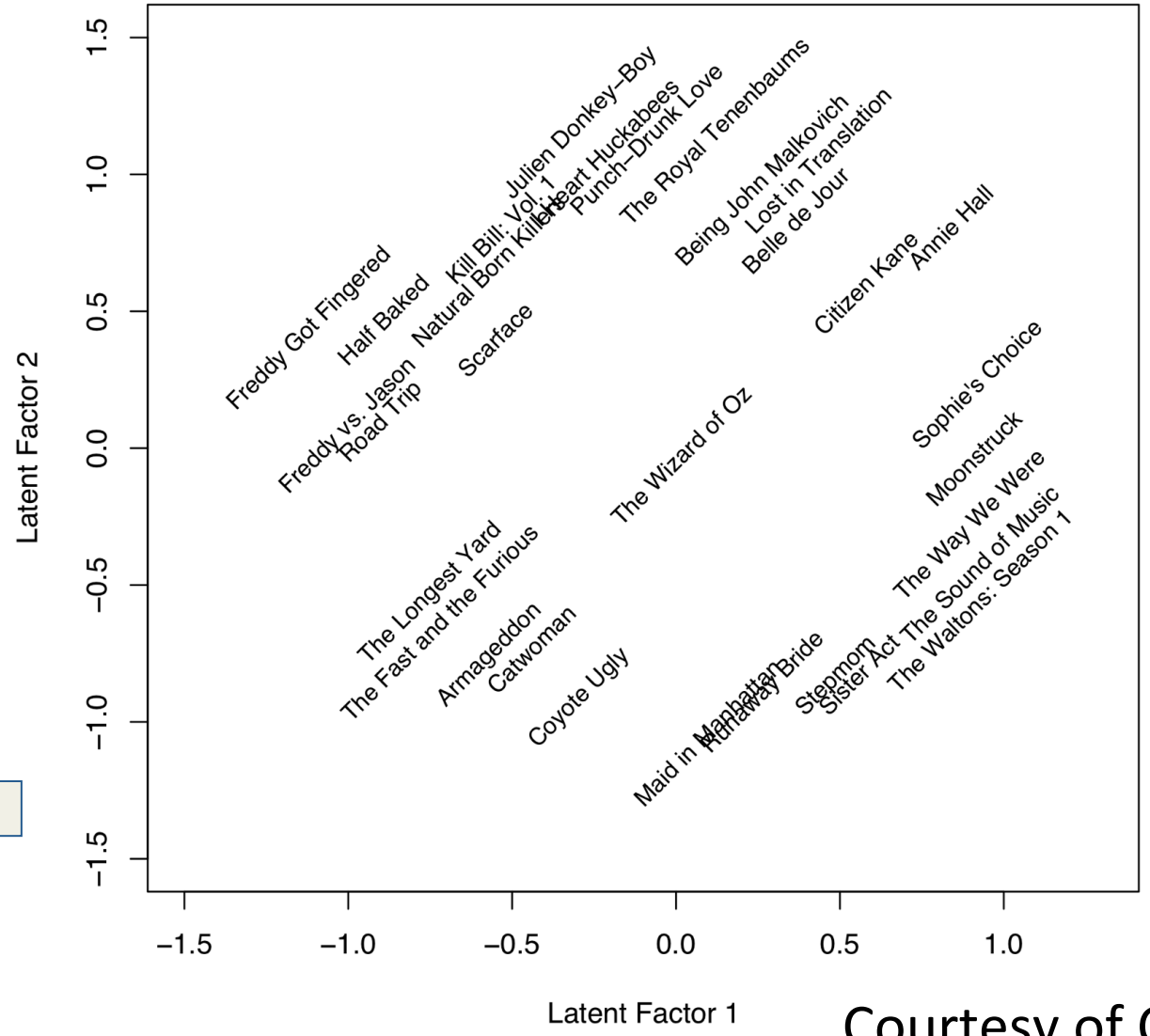
A **latent factors** model identifies factors with maximum discrimination between movies

Courtesy of Chris Volinsky

Latent Factors Model

Frat-house, gross-out comedies ⟷ Critically – acclaimed/ Strong female leads

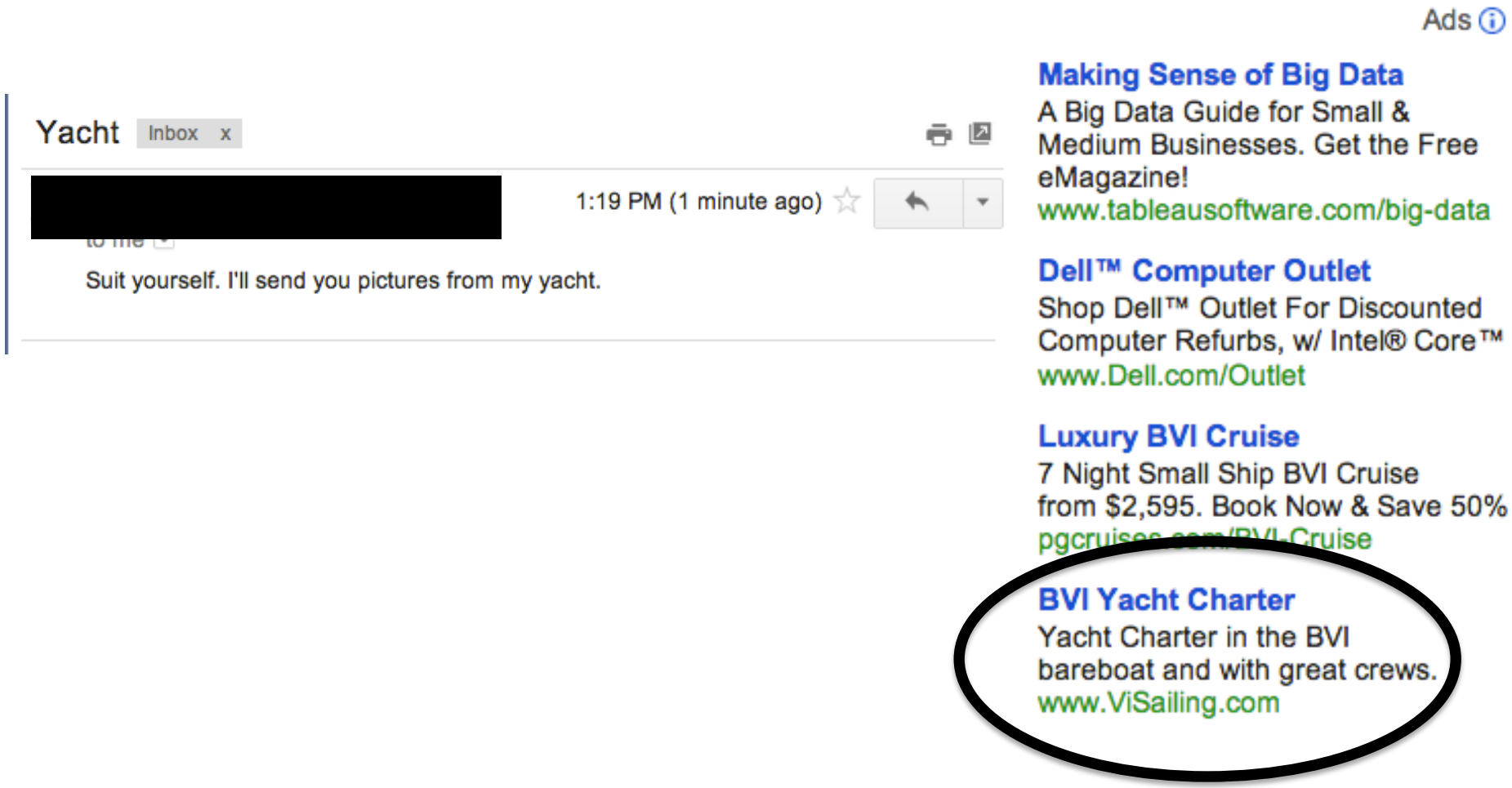
Artsy, edgy, movies ↓ Hollywood Big budget



A **latent factors** model identifies factors with maximum discrimination between movies

Courtesy of Chris Volinsky

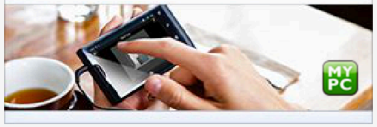
Ad-targeting



Sox Look to Take Series Against Rays In Labor Day Matinee

A win today would give Boston its third win out of four in this wraparound, four-game set against AL East rival Tampa Bay.

Game 137: Red Sox at Rays Live Updates |
Buchholz Brilliant in Sox 3-0 Win Over the Rays |
The 'Lespedes' Trade May End Up Being Billy Beane's Biggest Blunder



SCOREBOARD

Mon, Sep 1
BostonBos
Tampa BayTB
Preview | Box | Gameview

AL East

Team	W	L	Pct	GB
Baltimore	79	56	.585	-
NY Yankees	70	65	.519	9
Toronto	69	67	.507	10.5
Tampa Bay	66	71	.482	14
Boston	60	76	.441	19.5

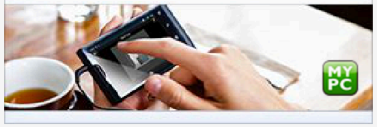
Full Standings

Leaders

Sox Look to Take Series Against Rays In Labor Day Matinee

A win today would give Boston its third win out of four in this wraparound, four-game set against AL East rival Tampa Bay.

Game 137: Red Sox at Rays Live Updates |
Buchholz Brilliant in Sox 3-0 Win Over the Rays |
The 'Lespedes' Trade May End Up Being Billy Beane's Biggest Blunder



SCOREBOARD

Mon, Sep 1
BostonBos
Tampa BayTB
Preview | Box | Gameview

AL East

Team	W	L	Pct	GB
Baltimore	79	56	.585	-
NY Yankees	70	65	.519	9
Toronto	69	67	.507	10.5
Tampa Bay	66	71	.482	14
Boston	60	76	.441	19.5

Full Standings

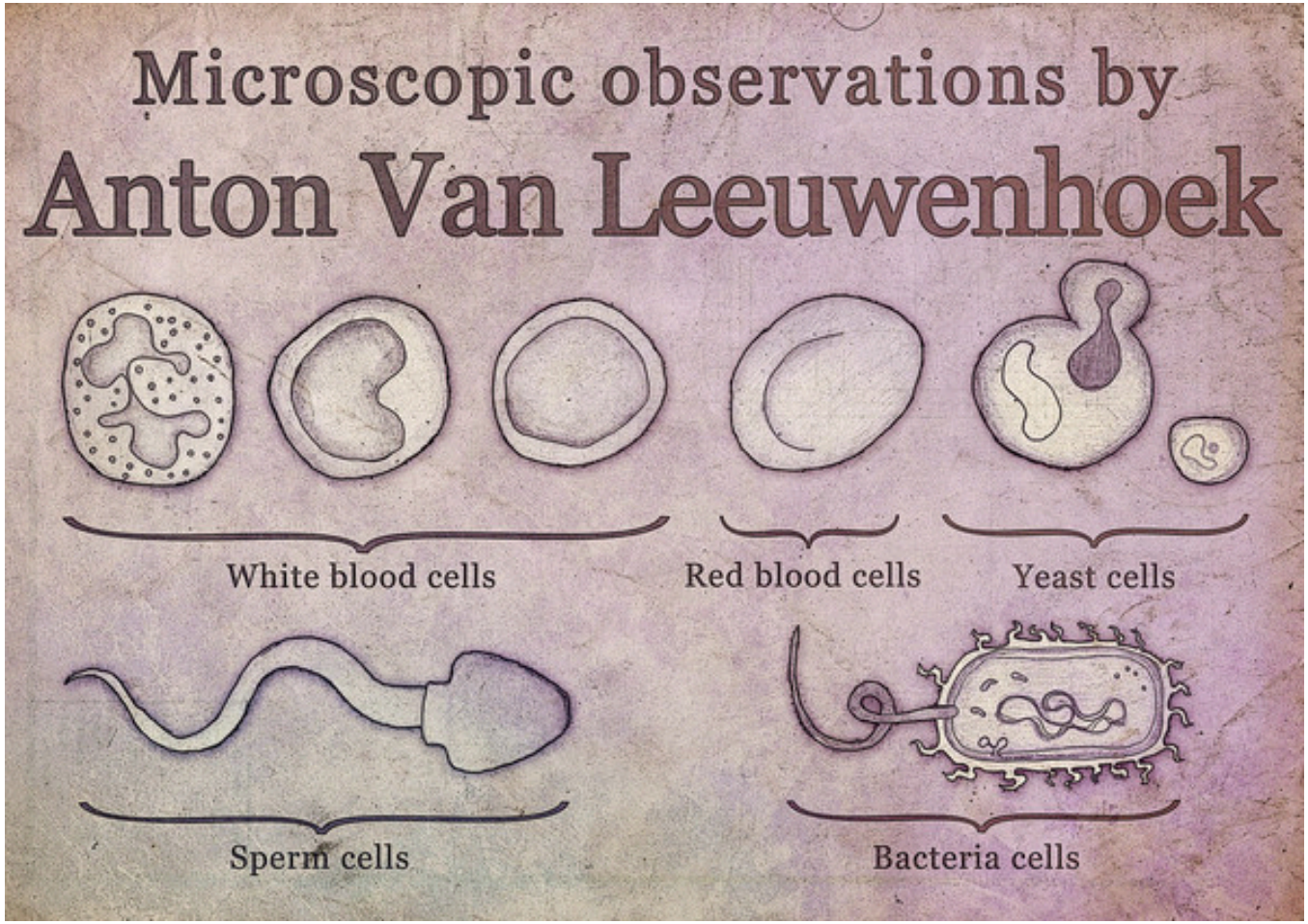
Leaders

Biology

Anton Van Leeuwenhoek (1623-1723)

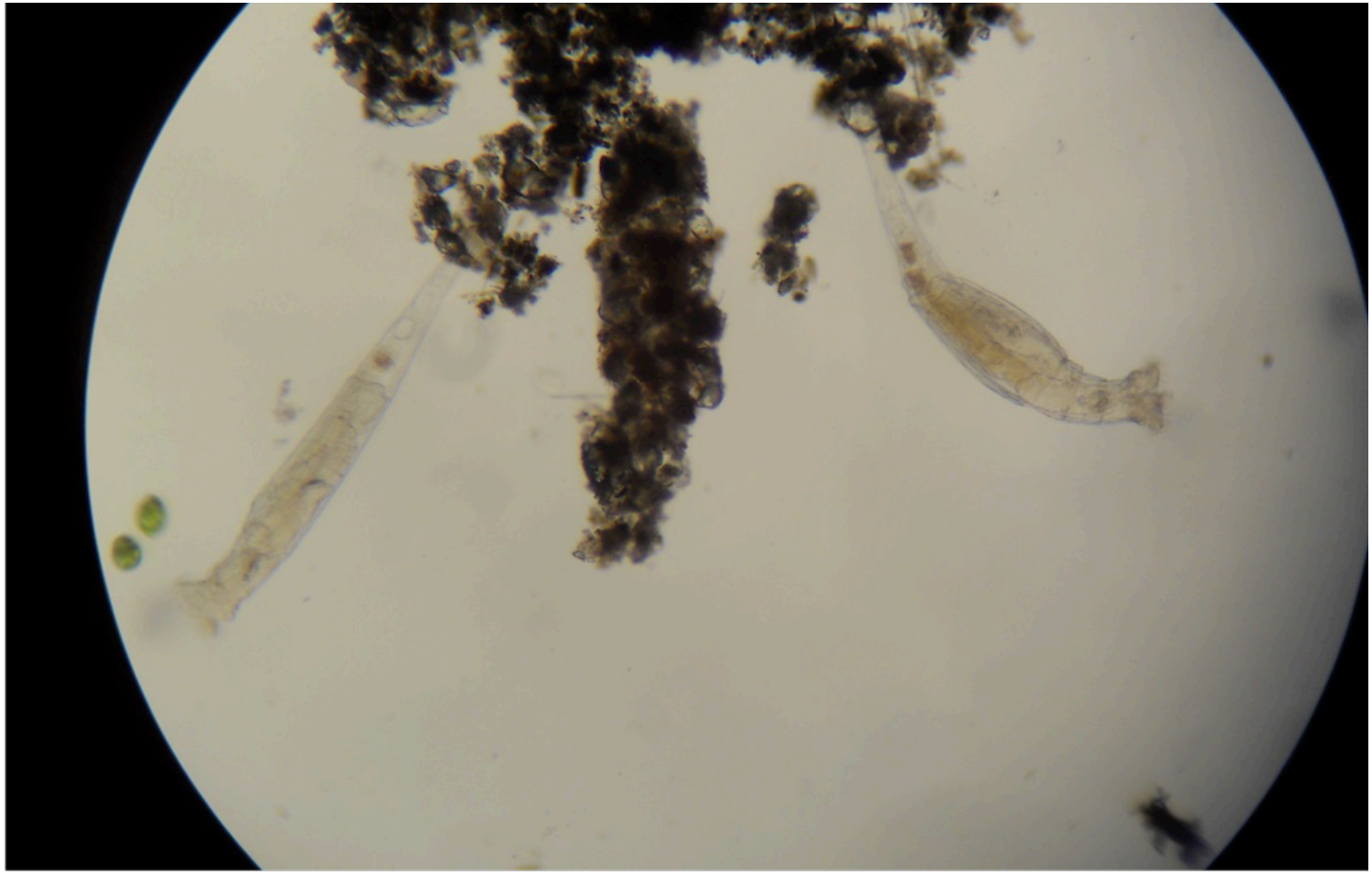
The “father of microbiology”

Some of his discoveries



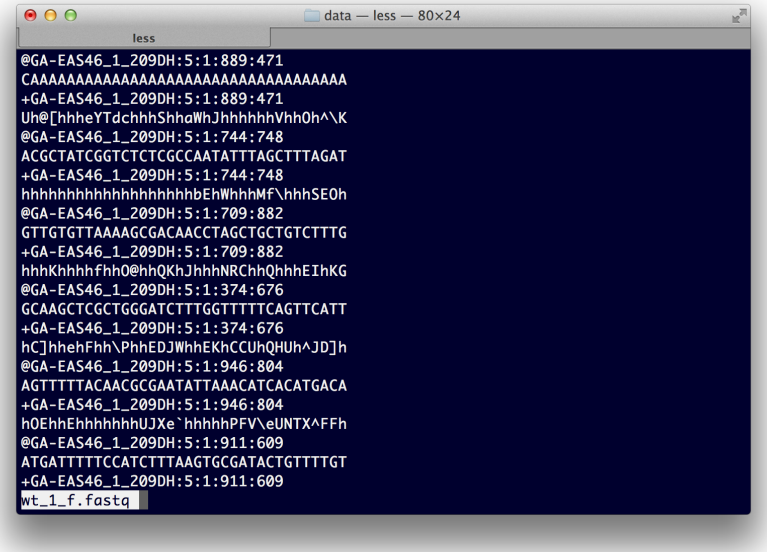
By improving the microscope

he saw what others could not



21st century version

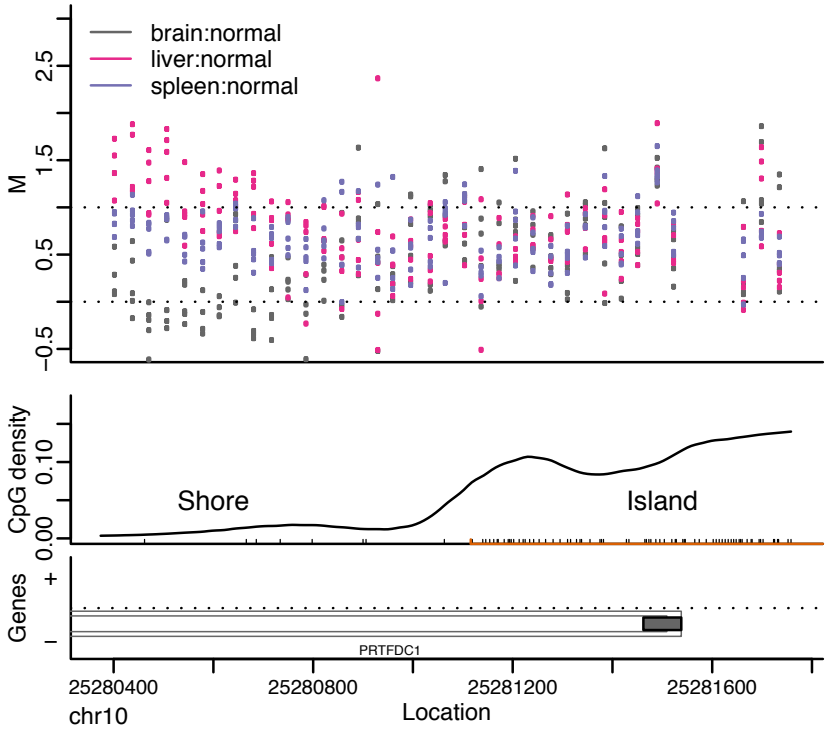
Modern high-throughput technology



Produces complex data, not images

21st century version

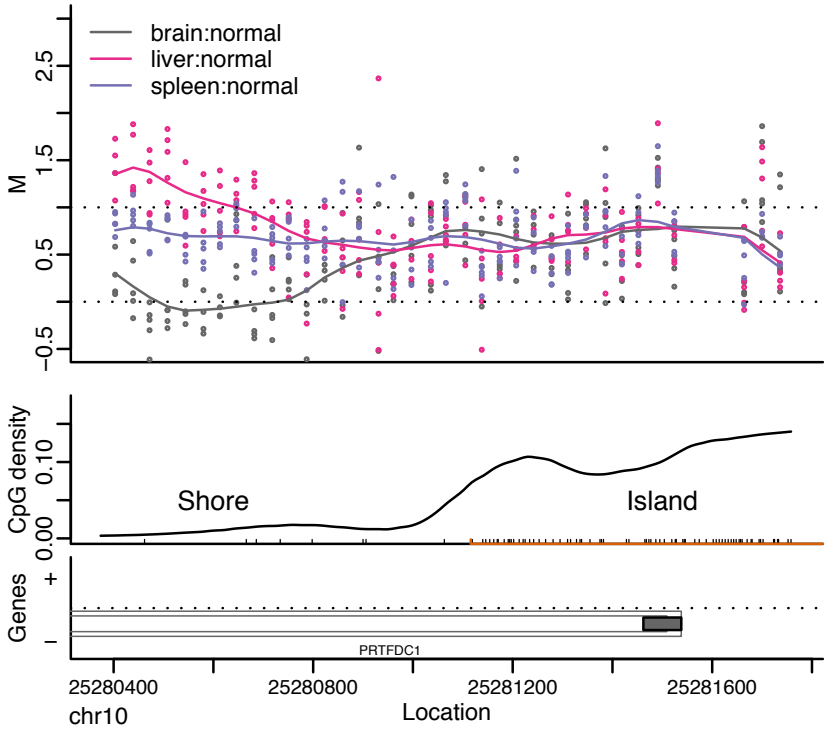
Modern high-throughput technology



My work has helped bring data into focus

21st century version

Modern high-throughput technology



My work has helped bring data into focus

Many other examples

- Spellcheckers
- Speech recognition
- Language translators
- Digitizing books
- Social sciences
- Medical diagnostics
- Personalized medicine
- Basic Biology

Last year’s projects

Some examples

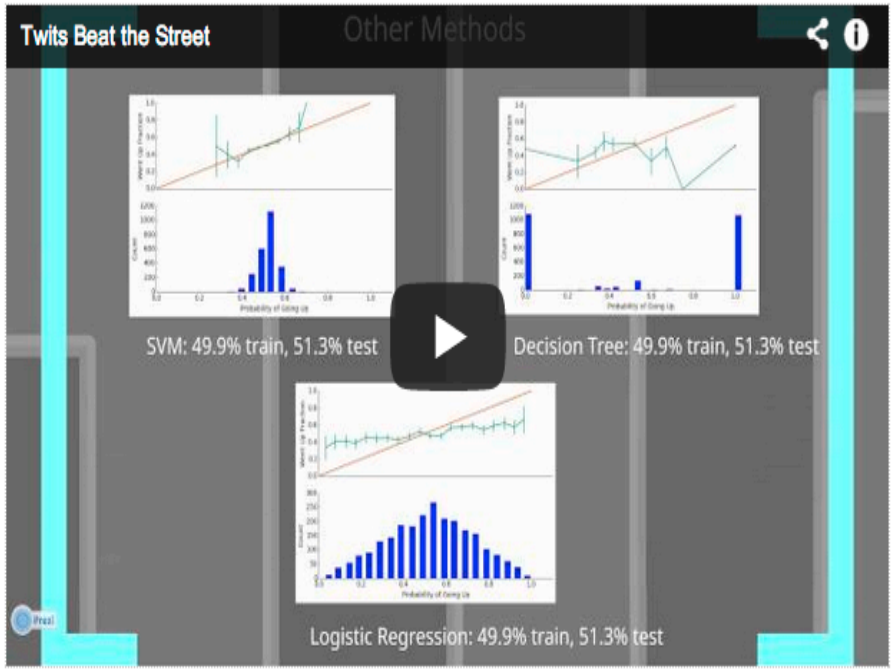
The Evolution of the American Presidency

Kathy Lin, Renzo Lucioni, Matthew Moellman and Sherrie Wang

Video

Twits Beat the Street

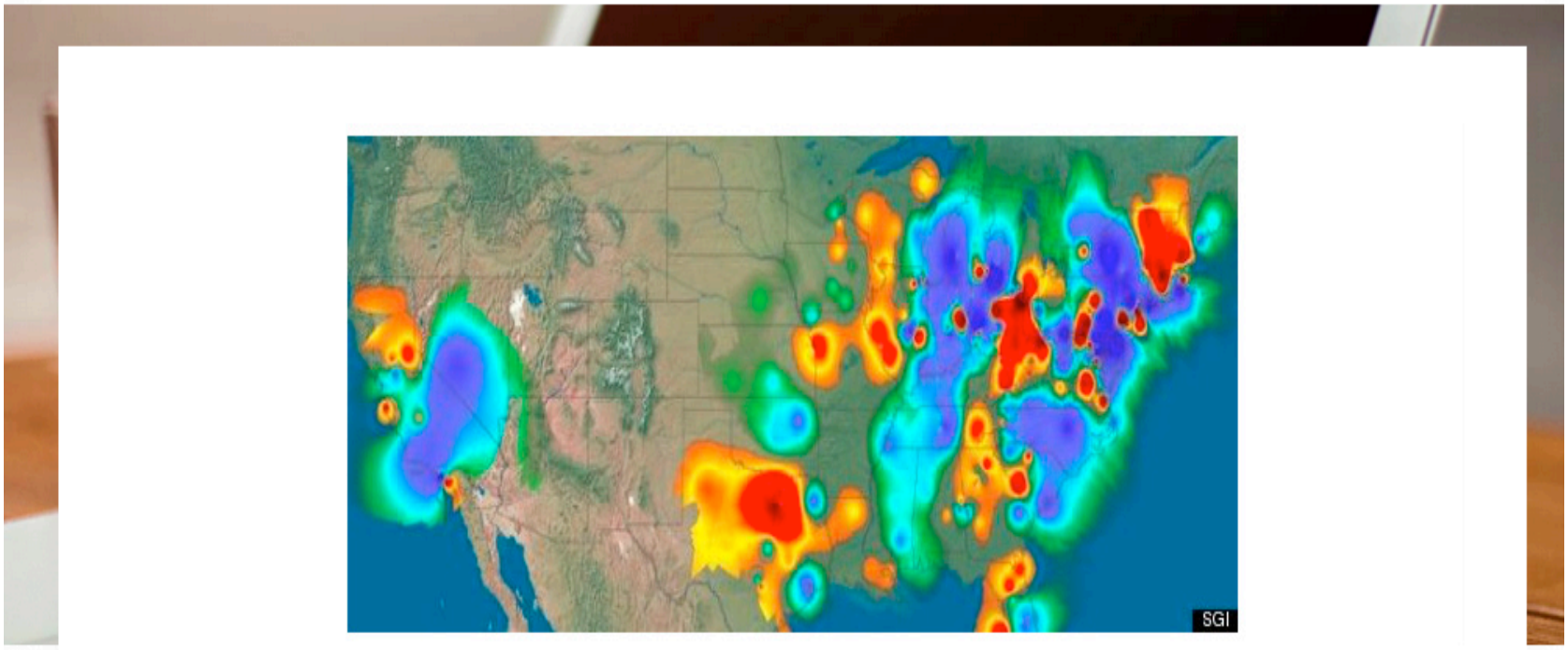
Home | Dataset | Text Classifiers | Market Predictor | Performance | About | Download the Notebook



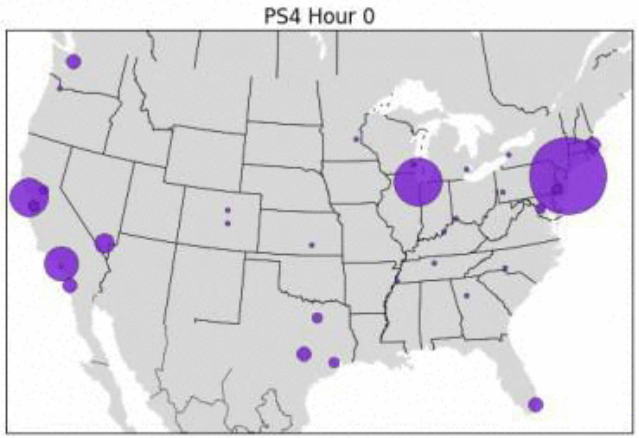
Video

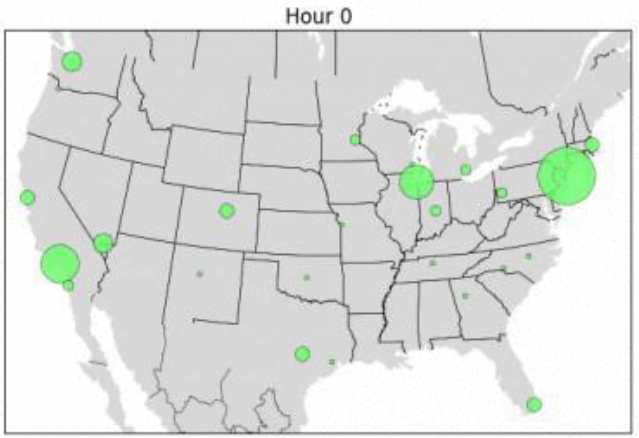
Tweets for Competitive Product Analysis

HOME ABOUT CONTACT **LOCATION ANALYSIS** MORE...



Video





PREDICTING CONCENTRATIONS

Sam Finegold | Theresa Gebert | Emi Nietfield | Jane Thomas

Video

A 3-pronged approach to the analysis of music.

video

http://cs109.joeong.com/
https://www.youtube.com/watch?v=azD8i4nJgQc

Video

How do we do Data Science?

Skills we will learn

- **Science**: determining what questions can be answered with data and what are the best datasets for answering them
- **Computer programming**: using computers to analyze data
- **Data wrangling**: getting data into analyzable form on our computers
- **Statistics**: separating signal from noise
- **Machine learning**: making predictions from data
- **Communication:** sharing findings through visualization, stories and interpretable summaries

Specific concepts and principles

- **Science**: gain experience asking questions.
- **Computer programming**: python, GitHub, cloud computing (on Amazon)
- **Data wrangling**: python libraries for reading data tables and scrapping web pages.
- **Statistics**: exploratory data analysis, inference, estimation, conditional probabilities, regression, modeling, Bayesian statistics, and more.
- **Machine learning**: support vector machines, k-nearest neighbors, regression trees, random forests, boosting,
- **Communication:** python graphing packages and in-class practice

Class will be centered around data

Homework and lectures will be based on:

- Baseball data
- World income and life expectancy
- Gene expression data
- Election data (predict 2014 midterm competition)
- Data for building predictors

Choose your own for final project

Examples of freely available data:

- Genomics
- Astronomy
- Financial
- Social media: Twitter, Reddit, Stack Overflow
- Fitbit (get your own)
- Baseball and other sports
- Movie ratings
- Baby names

To name a few…

Let's collect data for Thursday

http://bit.ly/1plnpos

Class Details

http://cs109.org

cs109.github.io/2014/

CS109 Data Science

HARVARD
School of Engineering
and Applied Sciences

Home Piazza Syllabus Schedule Homework Readings Projects Resources

Learning from data in order to gain useful predictions and insights. This course introduces methods for five key facets of an investigation: data wrangling, cleaning, and sampling to get a suitable data set; data management to be able to access big data quickly and reliably; exploratory data analysis to generate hypotheses and intuition; prediction based on statistical methods such as regression and classification; and communication of results through visualization, stories, and interpretable summaries.

We will be using Python for all programming assignments and projects. All lectures will be posted here and should be available 24 hours after meeting time.

Rafael Irizarry

- Math undergrad University of Puerto R
- PhD in Statistics UC Berkeley
- 15 years at Hopkins
- At Harvard since 2013
- DFCI is my group's home
- 5 postdocs, 2 grad students
- Taught EdX on Genomics
- Webpage: http://rafalab.org
- Blog: simplystatistics.org
- Twitter: @rafalab
- Email: rafa@jimmy.harvard.edu

Verena Kaynig-Fittkau

- Computer Science Diploma from University of Hamburg, Germany
- PhD in Machine Learning from ETH Zurich, Switzerland
- Postdoc at Harvard with Hanspeter Pfister
- Now lecturer at IACS
- Python Convert
- Working on Connectomics:
  - Image Processing
  - Deep Learning
  - Big Data
- NW 235.10
- Email: vkaynig@seas.harvard.edu

Important details

- Class web site: http://cs109.org
- Syllabus: http://cs109.github.io/2014/pages/syllabus.html
- Head TF: Stephanie Hicks shicks@jimmy.harvard.edu

Staff

- Stephanie Hicks (Head TF)
- Marc Streit (Guest lecturer)
- Mingxiang Teng
- Michael Packer
- Marcus Way
- Michael Lackner
- Amy Mir
- Tarik Adnan Moon
- Olivia Angiuli
- Yang Li
- Huihui Fan
- Antonia Oprescu
- Claudio Rosenberg
- Tudor Giurgica-Tiron
- Zhijie Zhou
- Nural Zaman
- Brian Feeny
- Joy Ming
- Rick Lee
- Felix Gonda
- Korey Tucker
- Lane Erickson
- Diana Miao
- Logan Kerr
- Stephen Klosterman

Grades

- **Homework**: 65%, assessed on your individual submission
- **Final Project Part I**: 10%, assessed on your individual submission
- **Final Project Part II**: 25%, assessed on meeting the project criteria

You are welcome to discuss the course's ideas, material, and homework with others in order to better understand it, but **the work you turn in must be your own**

Homework

- Real-World focus
- Scrape and wrangle messy data
- Explore data
- Apply statistical analysis
- Visualize and communicate results

Final Project

- Teams of up to 4 students
- Pick a project of your choosing
- **Part I**: describe question and plane for answering
- Process books, web sites, screencasts
- IPython (exceptions possible)
- **Part II:** present results and conclusions
- Best project prizes!

Details for Homework 0

Questions