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?

- 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**

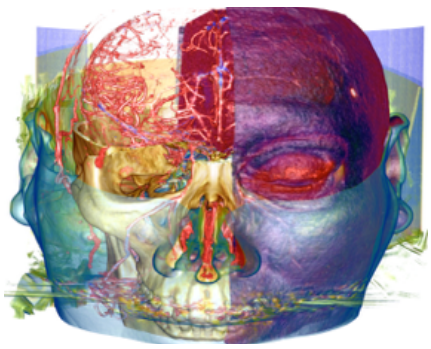
21st Century

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

Data Science

To gain insights into data through computation, statistics, and visualization



A Data Scientist Is...

"A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician."

- Josh Blumenstock

"Data Scientist = statistician + programmer + coach + storyteller + artist"

- Shlomo Aragmon

"By 2018, the US could face a shortage of up to 190,000 workers with analytical skills"

McKinsey Global Institute

"The sexy job in the next 10 years will be statisticians." *Data Scientists?*

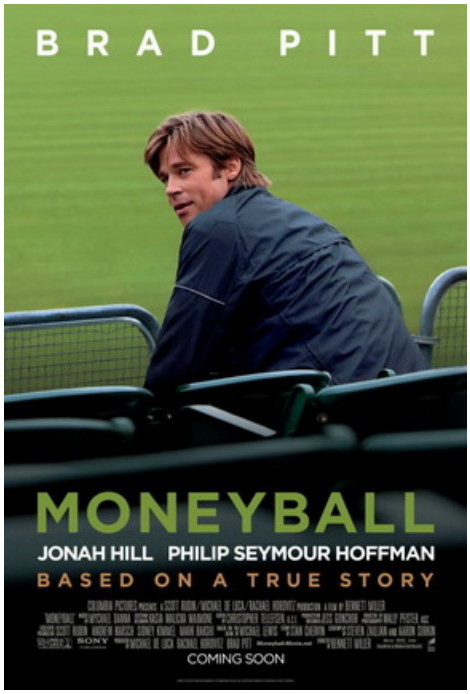
Hal Varian, Prof. Emeritus UC Berkeley
Chief Economist, Google

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.”

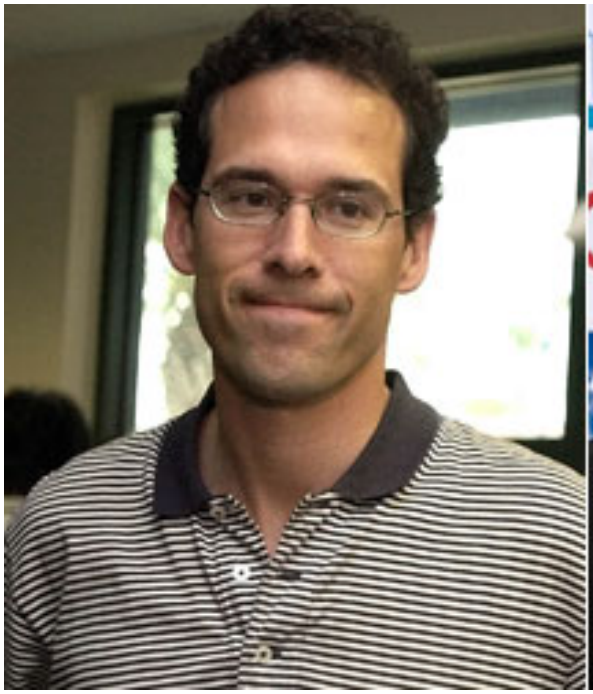
– Hal Varian

Data Science Success Stories

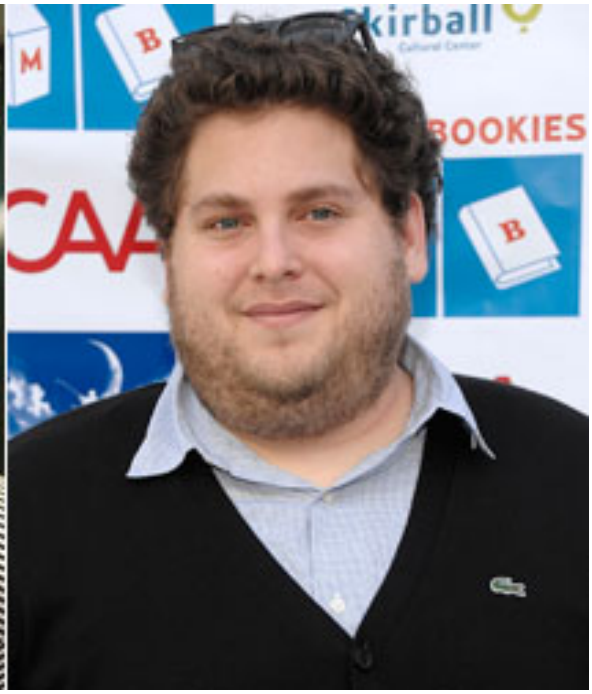

BRAD PITT
MONEYBALL
JONAH HILL PHILIP SEYMOUR HOFFMAN
BASED ON A TRUE STORY
COMING SOON

The Data Scientist

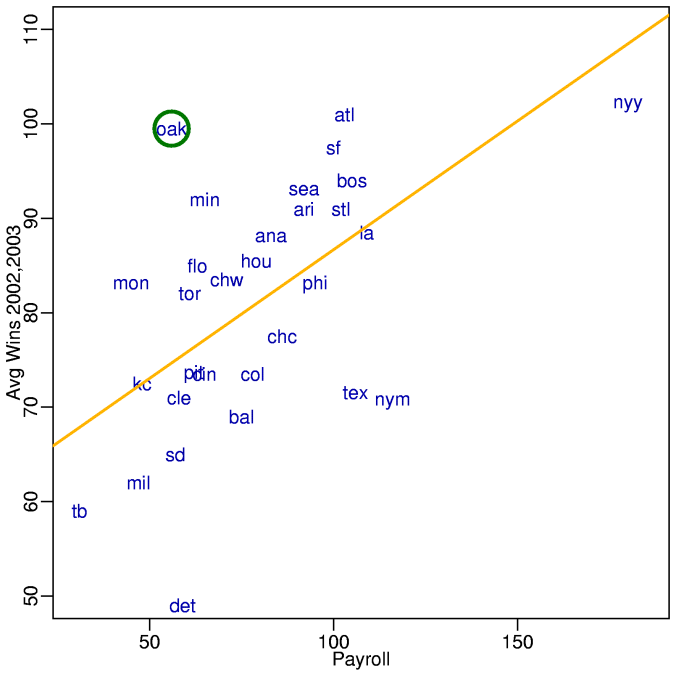
Actual



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

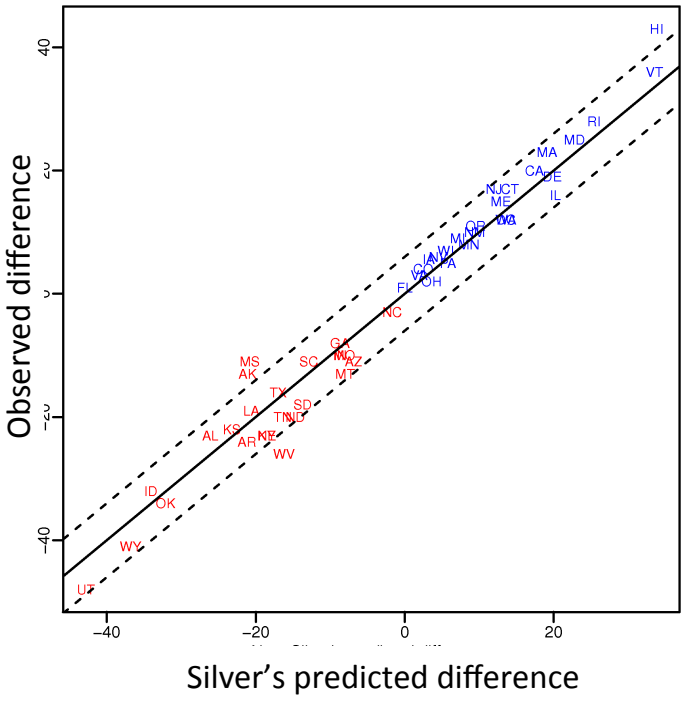
It's Tuesday, November 4th, 2008, Election Day in America. The last polls have straggled in, and show little sign of mercy for John McCain. Barack Obama appears poised for a decisive electoral victory.

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.

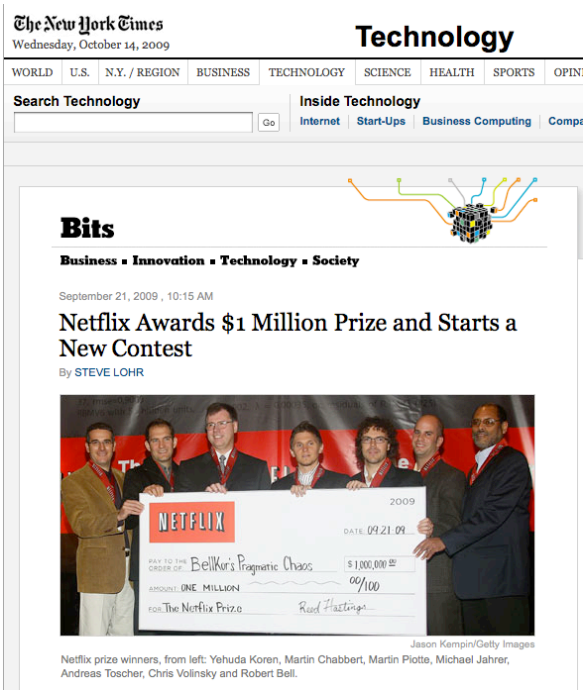
We also project Obama to win the popular vote by 6.1 points; his lead is slightly larger than that in the polls now, but our model accounts for the fact that candidates with large leads in the polls typically underperform their numbers by a small margin on Election Day.

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

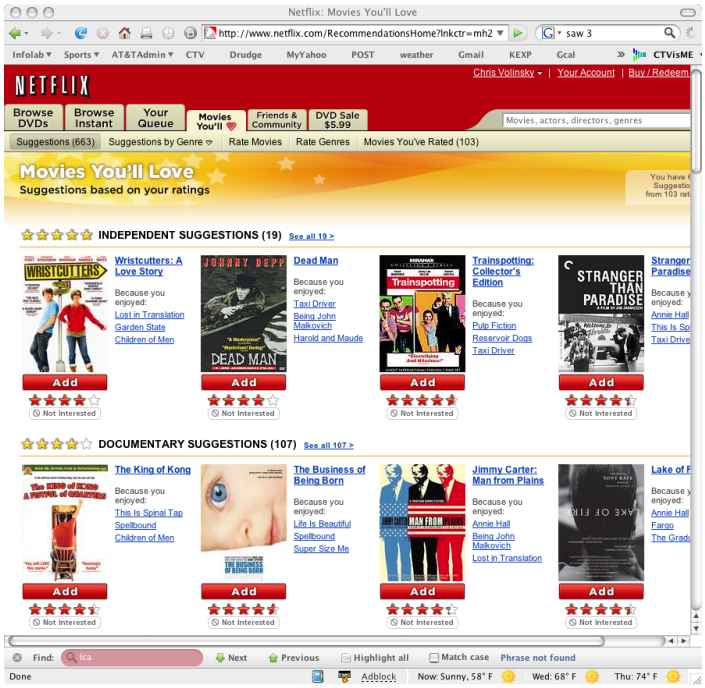
Jason Kempin/Getty Images

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

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

•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

• 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

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

• Test data

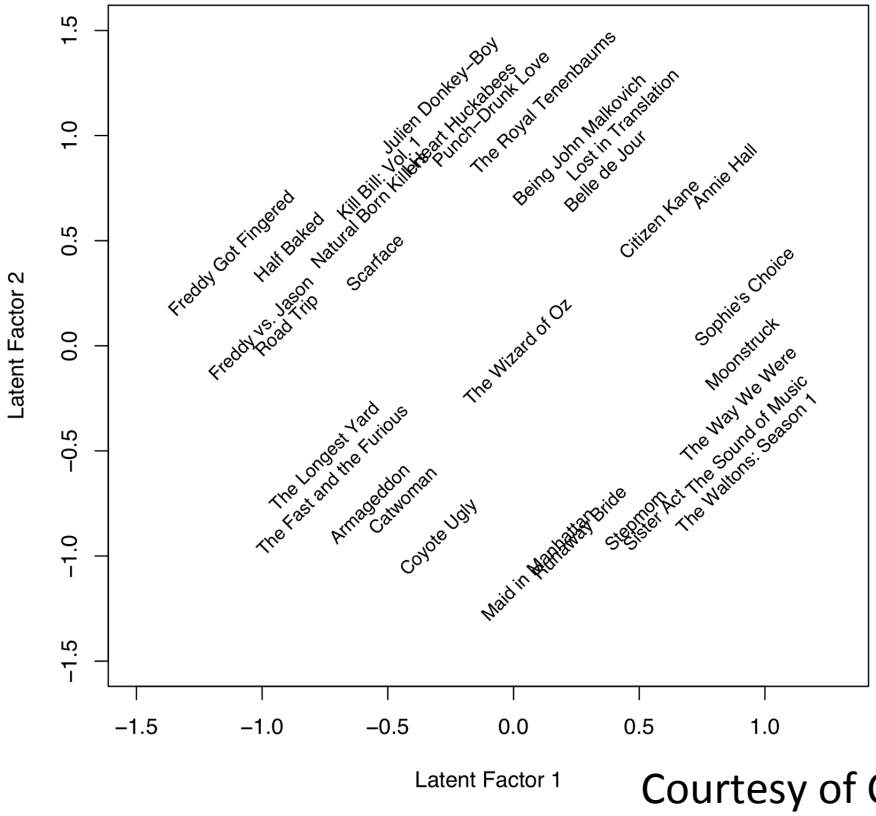
- Last few ratings of each user (2.8 million)
- Evaluation via RMSE: root mean squared error
- Netflix Cinematch RMSE: 0.9514

• Competition

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

Courtesy of Chris Volinsky

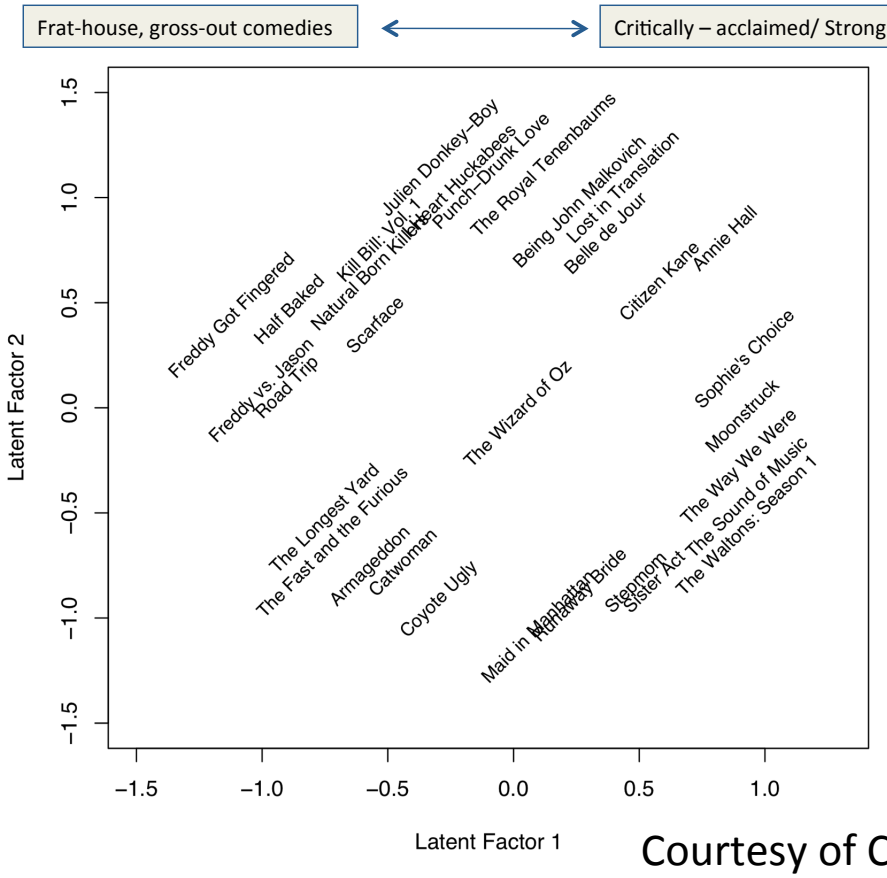
Latent Factors Model



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

Courtesy of Chris Volinsky

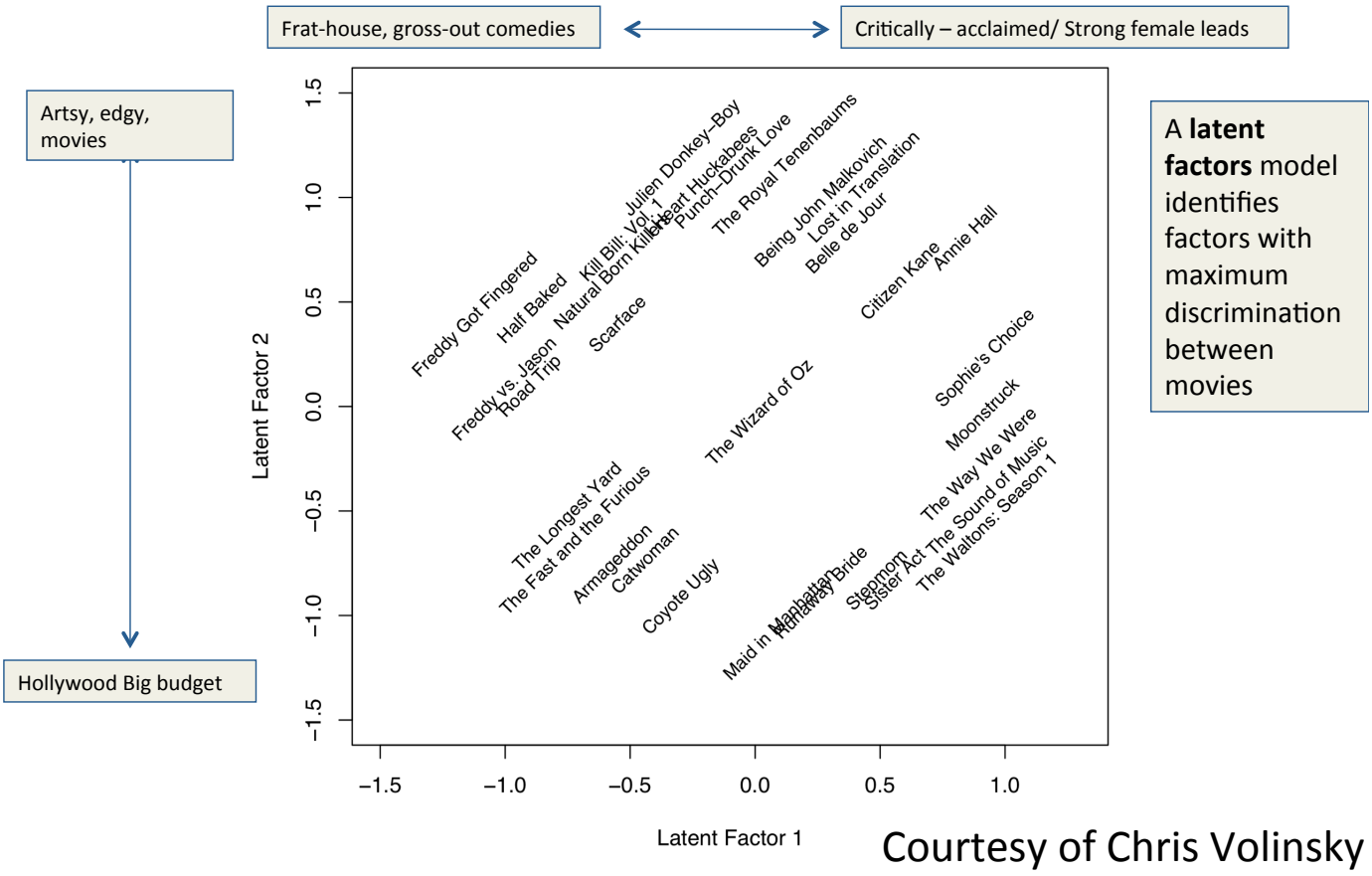
Latent Factors Model



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

Courtesy of Chris Volinsky

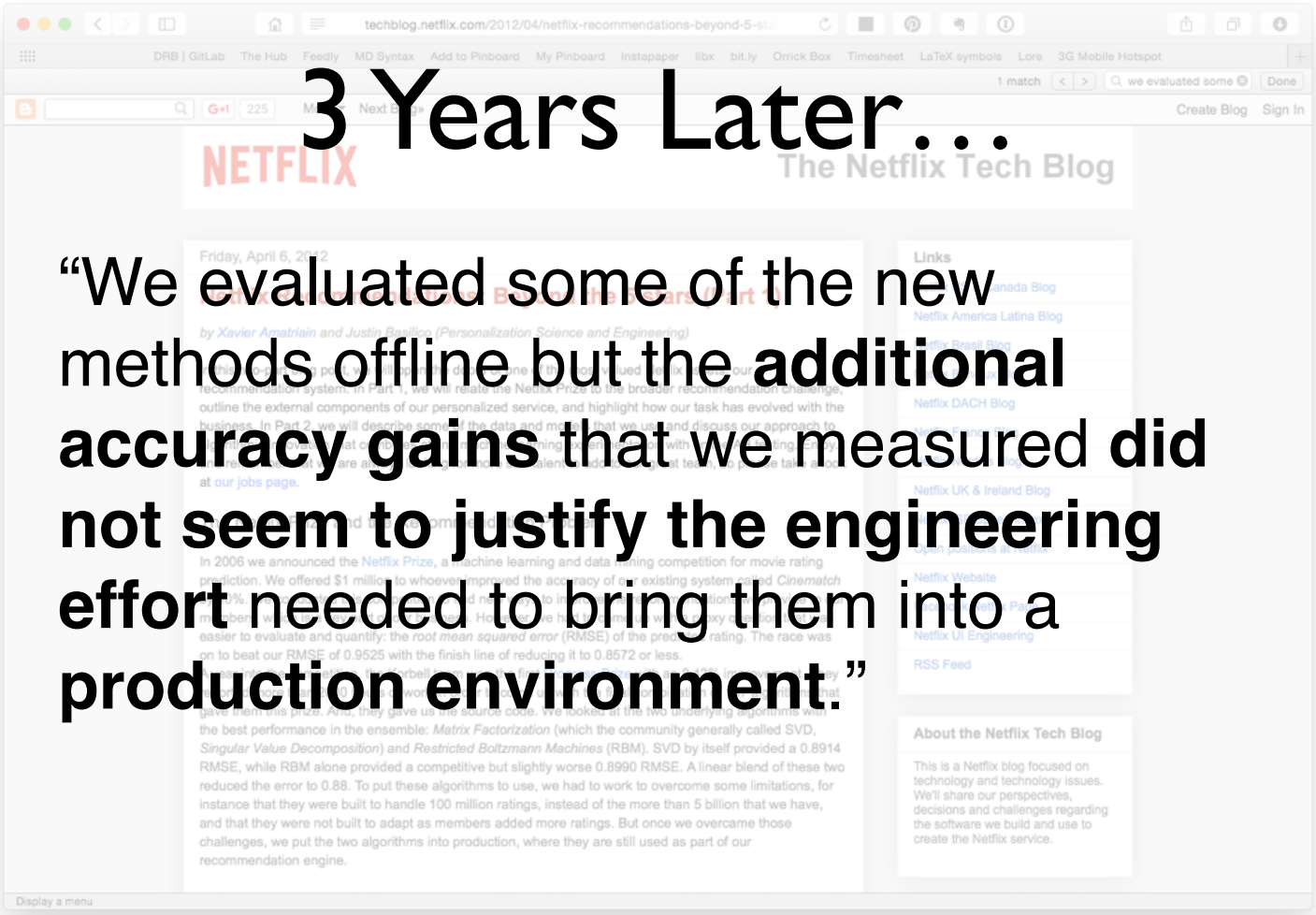
Latent Factors Model



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

Courtesy of Chris Volinsky

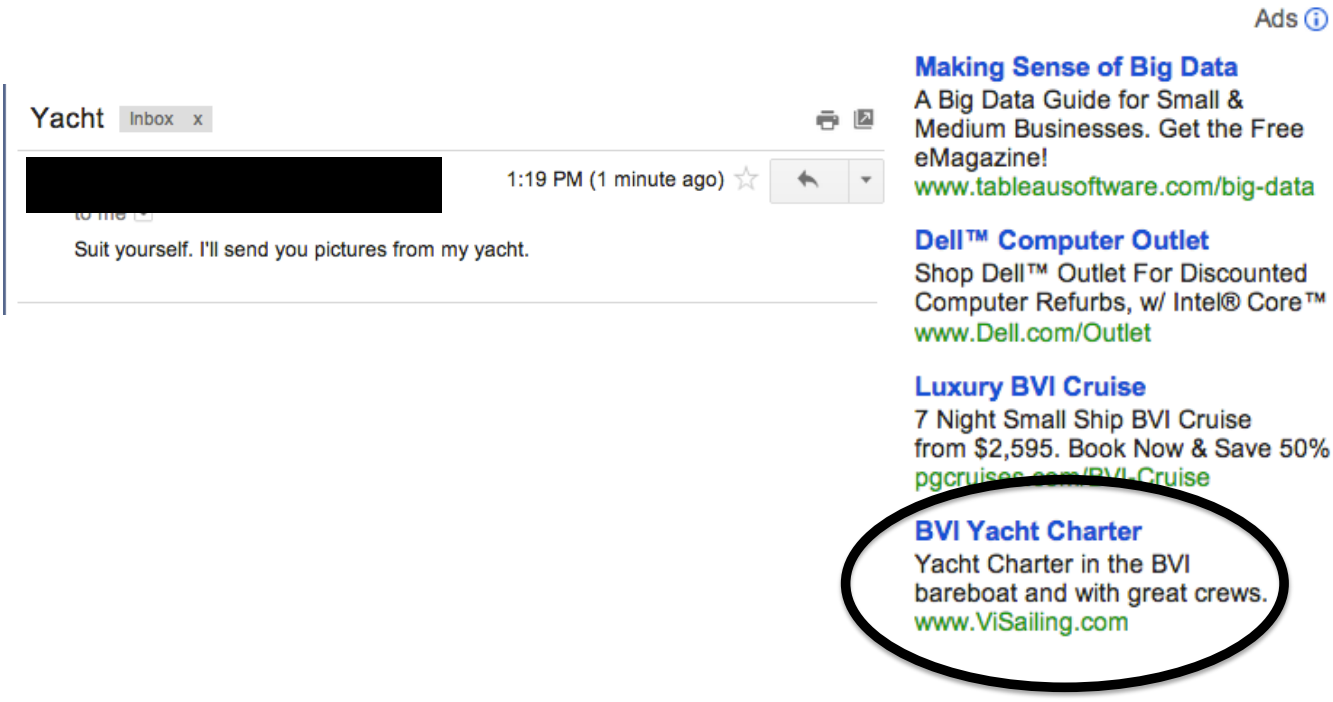
3 Years Later…



“We evaluated some of the new methods offline but the **additional accuracy gains** that we measured **did not seem to justify the engineering effort** needed to bring them into a **production environment**.”

Xavier Amatriain and Justin Basilico, 2012

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

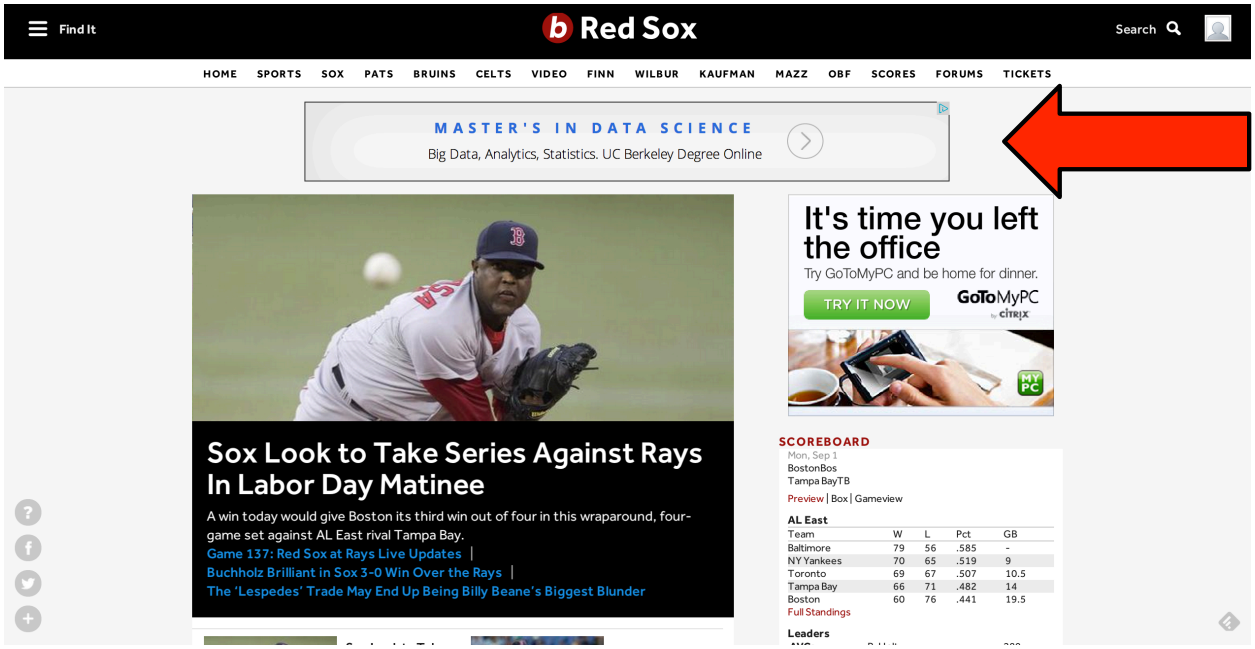
Preview | Box | Gameview

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

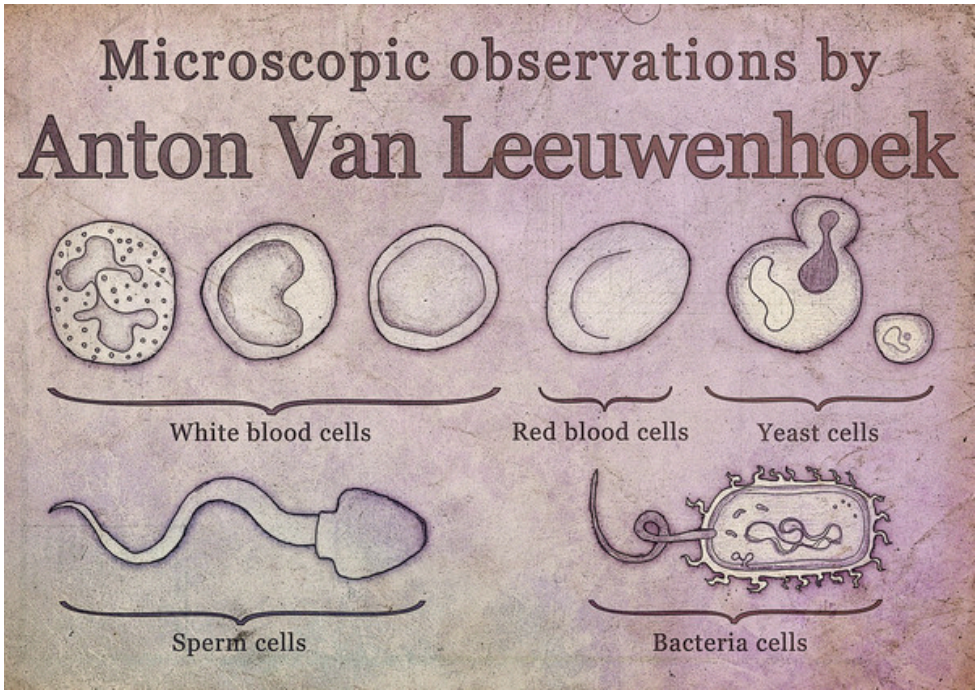

Find It
Red Sox
Search
HOME SPORTS SOX PATS BRUINS CELTS VIDEO FINN WILBUR KAUFMAN MAZZ OBF SCORES FORUMS TICKETS
MASTER'S IN DATA SCIENCE
Big Data, Analytics, Statistics. UC Berkeley Degree Online
Sox Look to Take Series Against Rays In Labor Day Matinee
A win today would give Boston its third win out 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
It's time you left the office
Try GoToMyPC and be home for dinner.
TRY IT NOW
GoToMyPC
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)

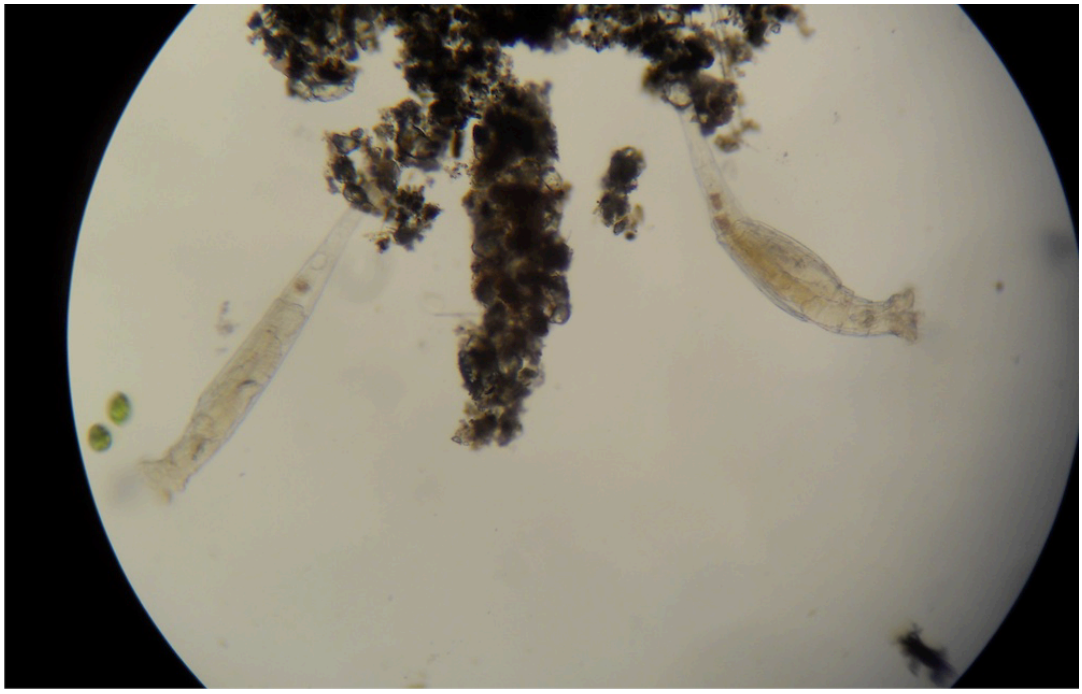
The “father of microbiology”

Some of his discoveries



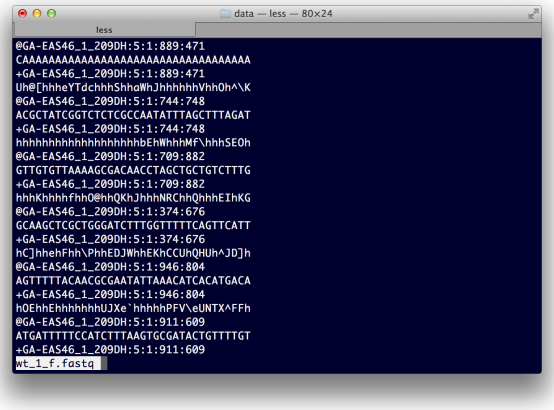
By improving the microscope

he saw what others could not



21st century version

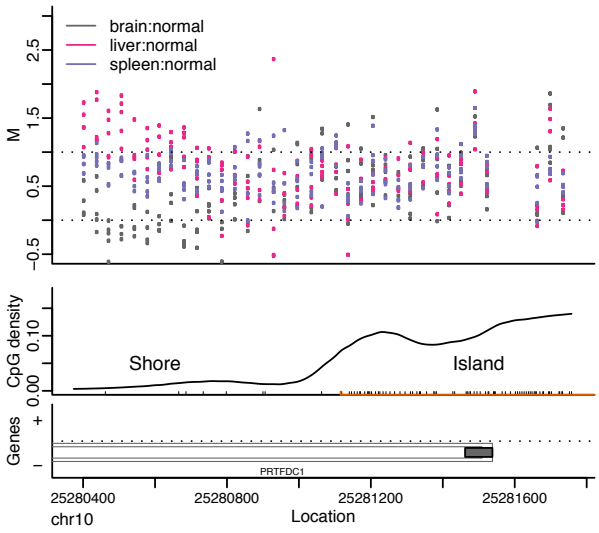
Modern high-throughput technology



Produces complex data, not images

21st century version

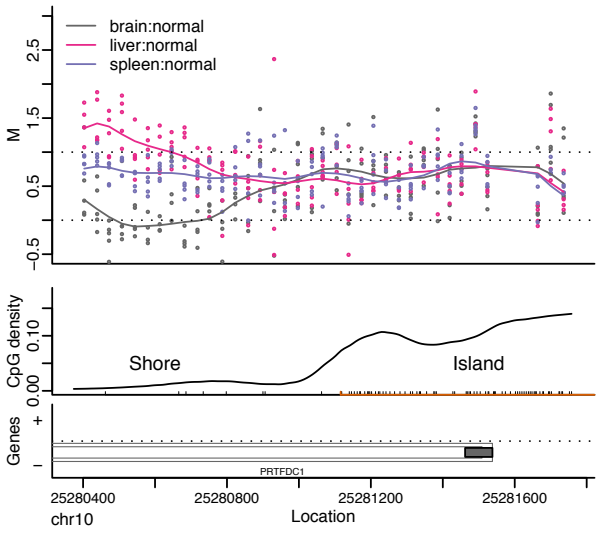
Modern high-throughput technology



My work has helped bring data into focus

21st century version

Modern high-throughput technology

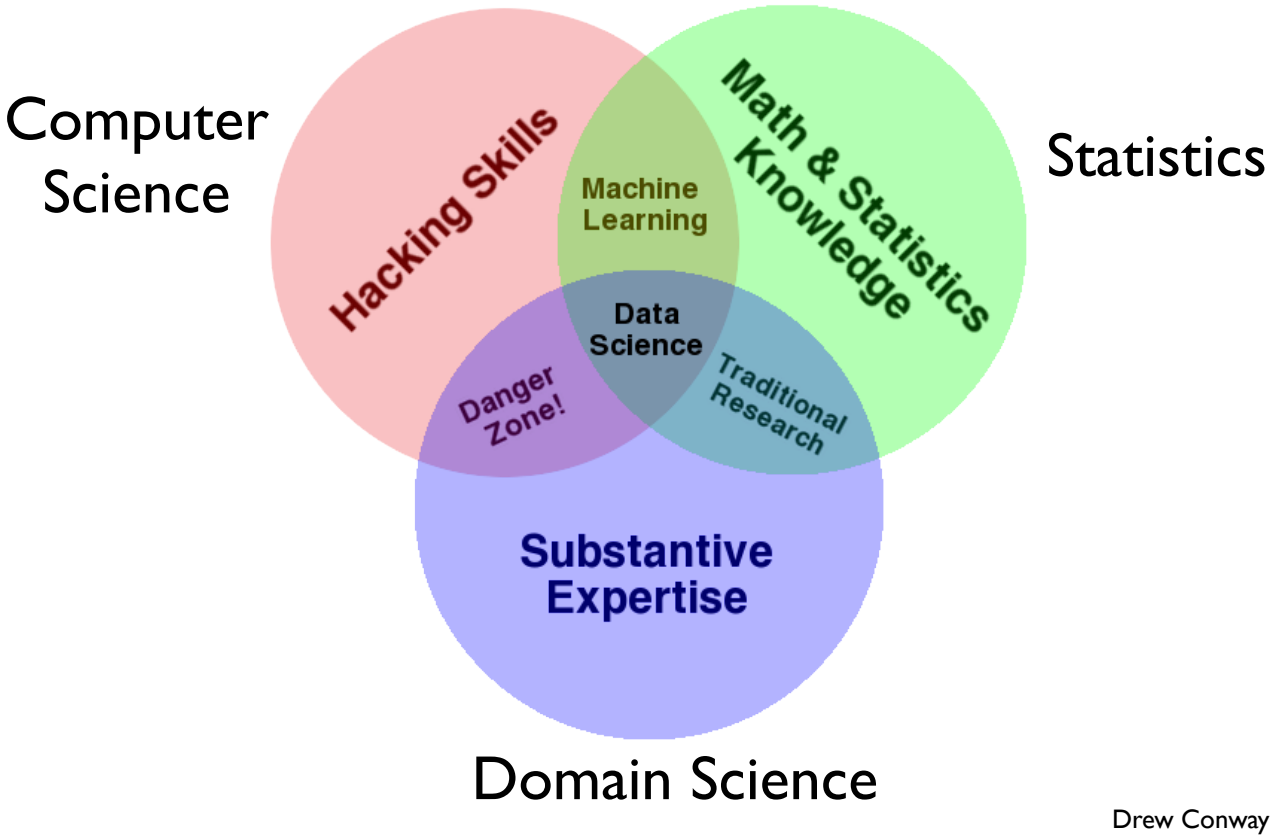


My work has helped bring data into focus

Many other examples

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

Data Science



Drew Conway

Machine **Human**

Data Management Human Cognition

Data Mining Perception

Machine Learning Visualization Story Telling

Business Intelligence Decision Making

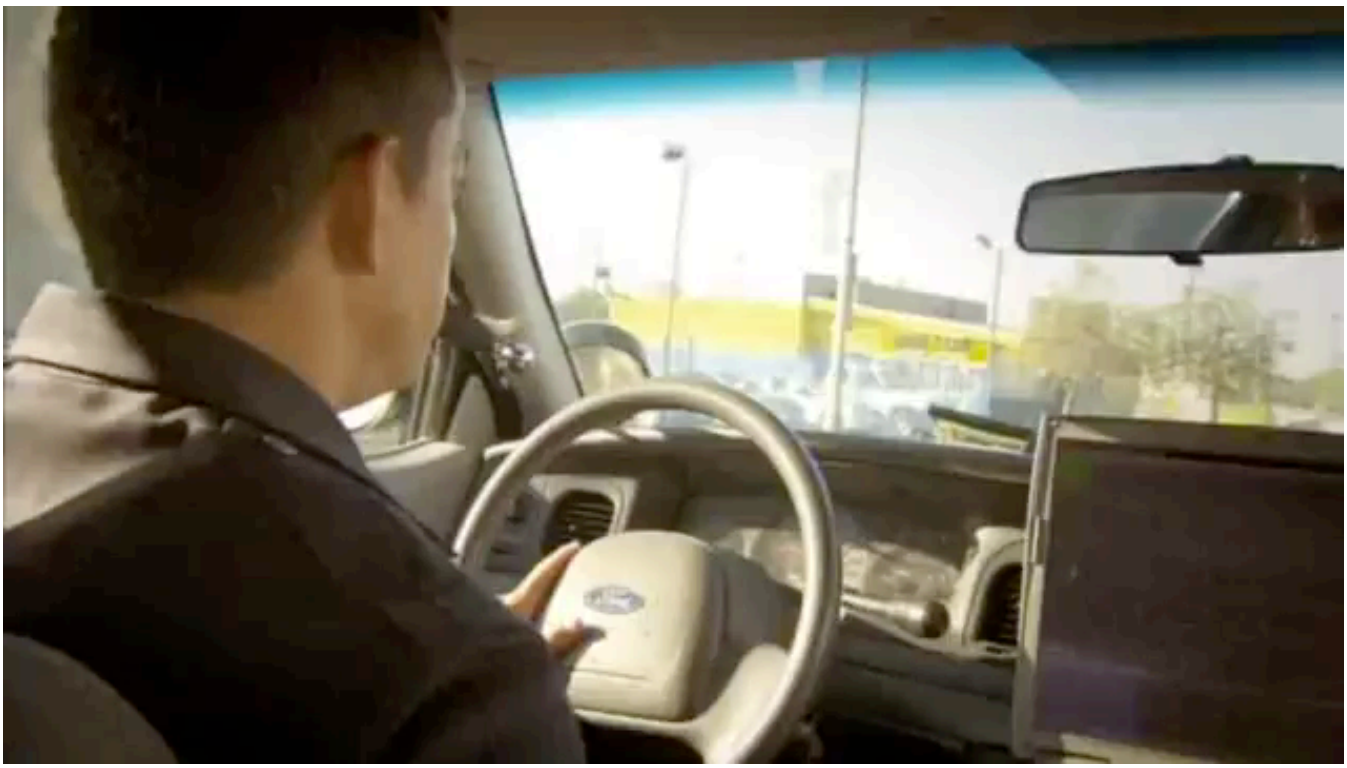
Theory

Statistics

Data Science

Inspired by Daniel Keim, "Visual Analytics: Definition, Process, and Challenges"

The Age of Big Data

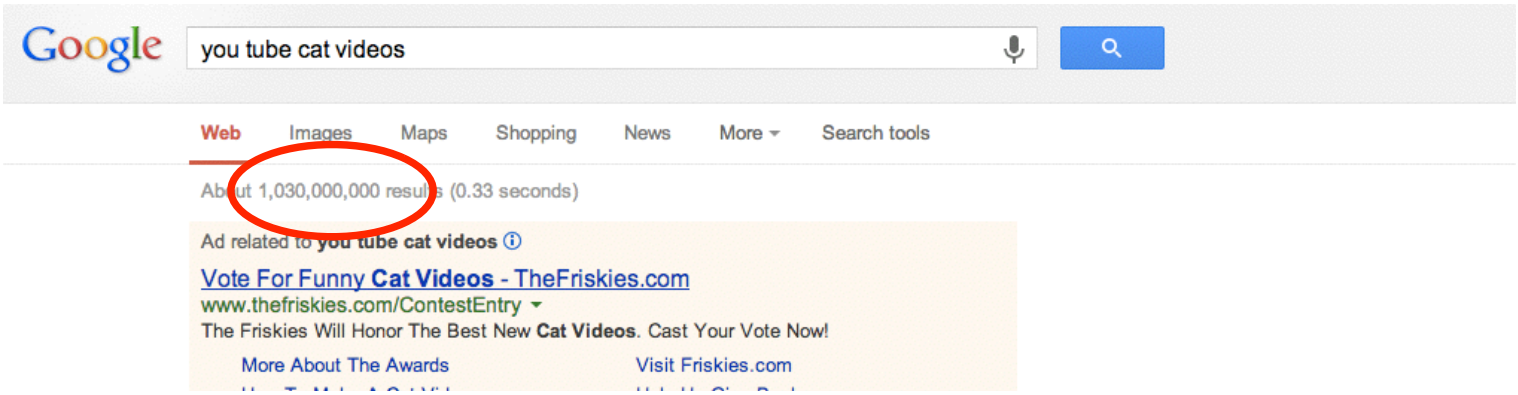


BBC, 2013

Big Data

"Between the dawn of civilization and 2003, we only created five exabytes of information; now we're creating that amount every two days."

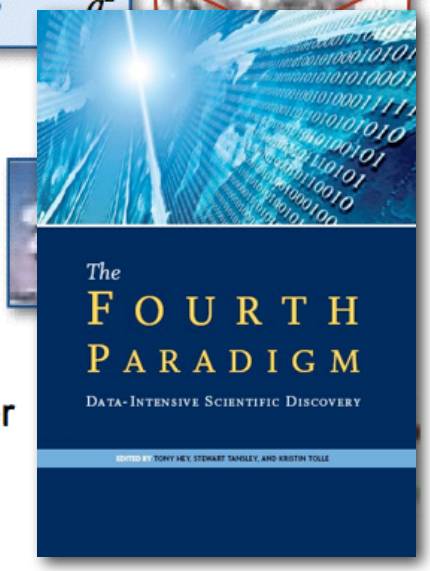
Eric Schmidt, Google (and others)



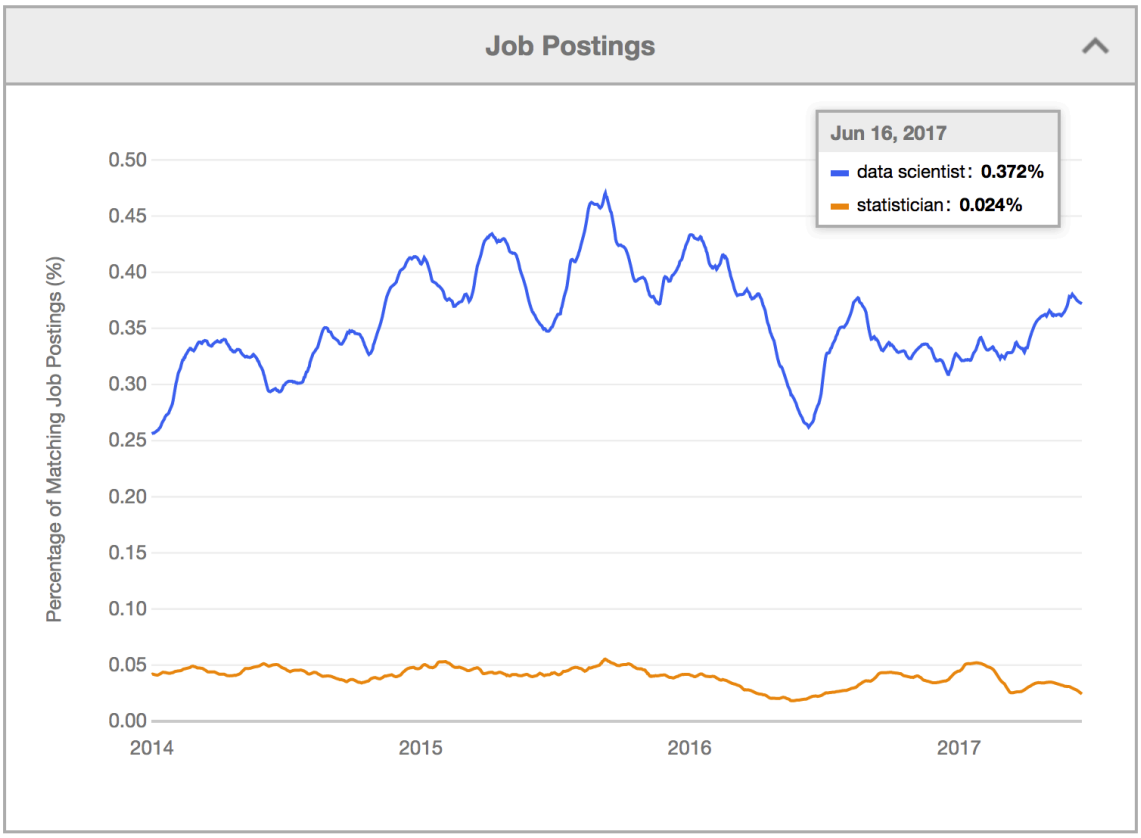
Science Paradigms

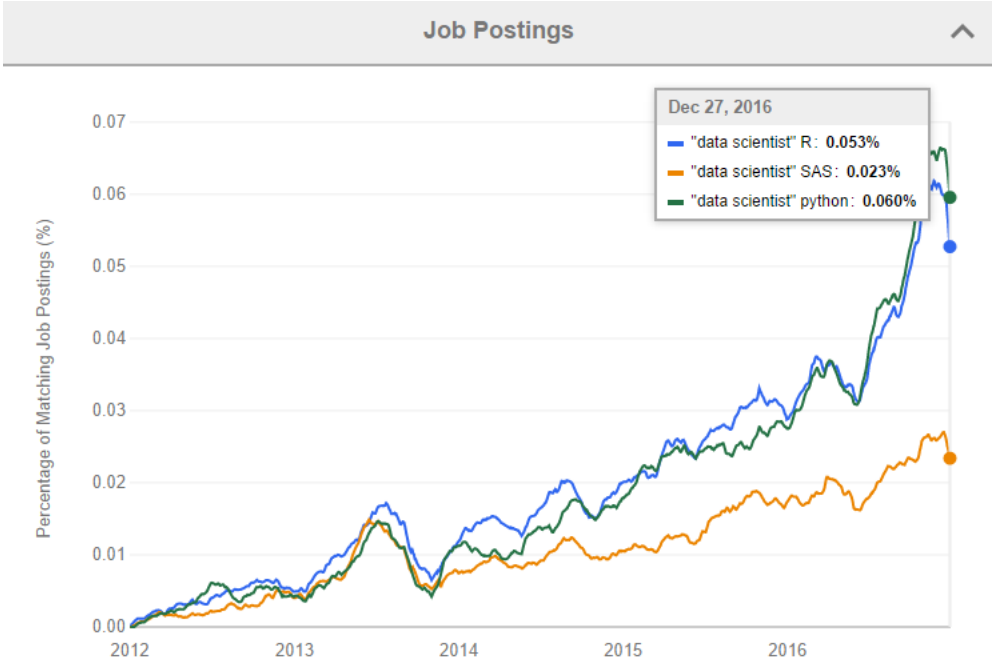
- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - \mathrm{K}\frac{c^2}{a^2}$$

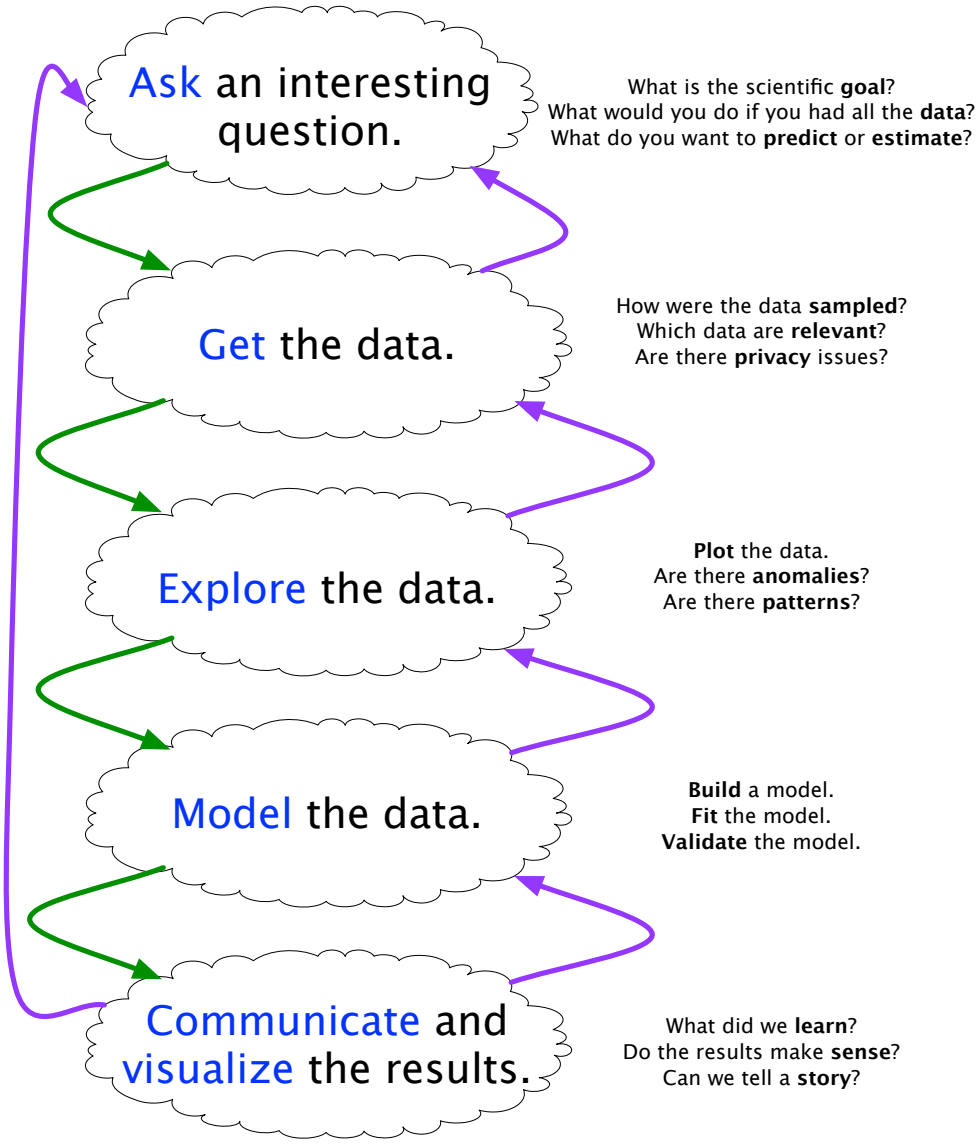


Jim Gray, Microsoft


Job Postings
Jun 16, 2017
data scientist: 0.372%
statistician: 0.024%
Percentage of Matching Job Postings (%)
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2014
2015
2016
2017


Job Postings
Dec 27, 2016
"data scientist" R: 0.053%
"data scientist" SAS: 0.023%
"data scientist" python: 0.060%
Percentage of Matching Job Postings (%)
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
2012
2013
2014
2015
2016

How do we do Data Science?


Ask an interesting question.
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?
Get the data.
How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?
Explore the data.
Plot the data.
Are there **anomalies**?
Are there **patterns**?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

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

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…

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!

Questions