

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



The image shows two media pieces from The New York Times Sunday Review:

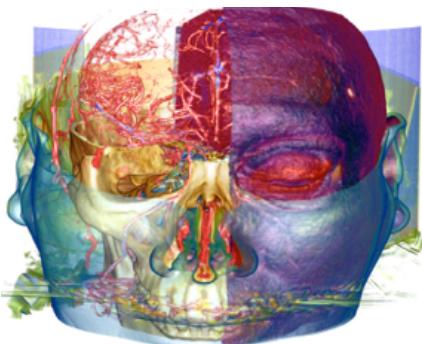
- The Age of Big Data** by Steve Lohr, published February 11, 2012. The article discusses the impact of big data on various fields.
- POPULAR SCIENCE** (Special Issue: The Future Now): The cover features a hand pointing upwards with a bright light at the tip, set against a background of binary code. Headlines include "THE CONTROL CENTERS," "OFFICER ALGORITHM," "NEW WAYS OF SEEING," and "DATA IS POWER: HOW INFORMATION IS DRIVING THE FUTURE".

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?*

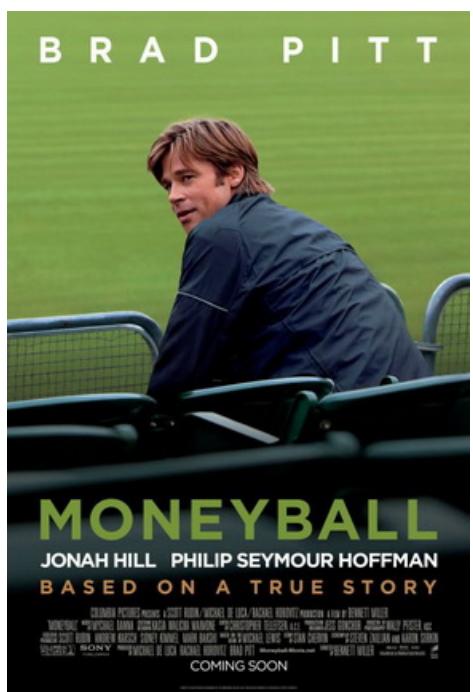
Hal Varian, Prof. Emeritus UC Berkeley
Chief Economist, Google

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



The Data Scientist

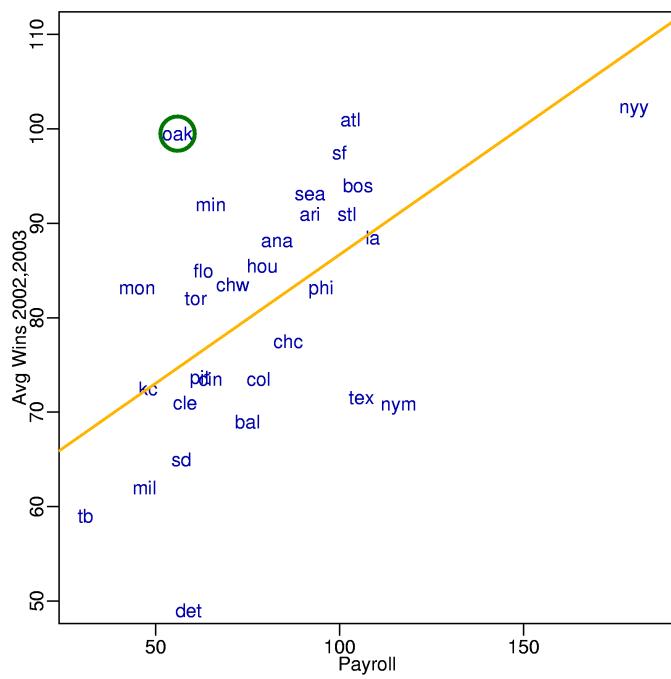
Actual



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

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.

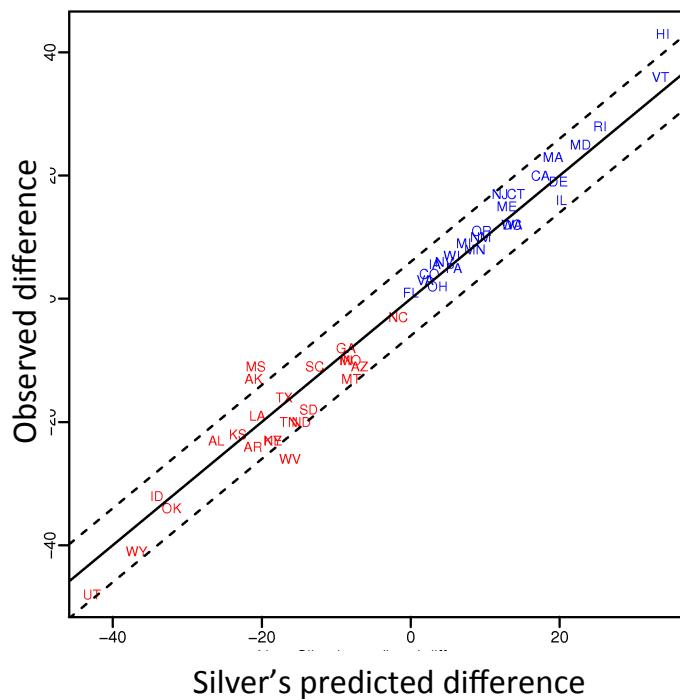
[Advertise @ 538!](#)

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 | OPINION

Search Technology Inside Technology

Internet | Start-Ups | Business Computing | Computer Software | Gadgets | Games | Home & Garden | Mobile | Software | Technology News

Bits

Business • Innovation • Technology • Society

September 21, 2009, 10:15 AM

Netflix Awards \$1 Million Prize and Starts a New Contest

By STEVE LOHR

A photograph showing seven men in suits standing behind a large ceremonial check. The check is white with a red border and features the Netflix logo at the top left. The text on the check reads: "NETFLIX", "PAY TO THE ORDER OF Bellkor's Pragmatic Chaos", "AMOUNT ONE MILLION", "00/100", "DATE 09.21.09", and "Reed Hastings".

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

Jason Kempin/Getty Images

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

The screenshot shows the Netflix Recommendations Home page. At the top, there's a navigation bar with links like 'Browse DVDs', 'Browse Instant', 'Your Queue', 'Movies You'll Love' (which is highlighted in yellow), 'Friends & Community', and 'DVD Sale \$5.99'. Below the navigation, it says 'Movies, actors, directors, genres'. Under 'Movies You'll Love', it says 'Suggestions based on your ratings' and 'You have 6 Suggs from 103 all'. It lists 'INDEPENDENT SUGGESTIONS (19)' including 'Wristcutters: A Love Story', 'Dead Man', 'Trainspotting: Collector's Edition', and 'Stranger Than Paradise'. It also lists 'DOCUMENTARY SUGGESTIONS (107)' including 'The King of Kong', 'The Business of Being Born', 'Jimmy Carter: Man from Plains', and 'Lake of Fire'. Each suggestion includes a movie poster, a brief description of why the user enjoyed it, and 'Add' and 'Not Interested' buttons.

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

- 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

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

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

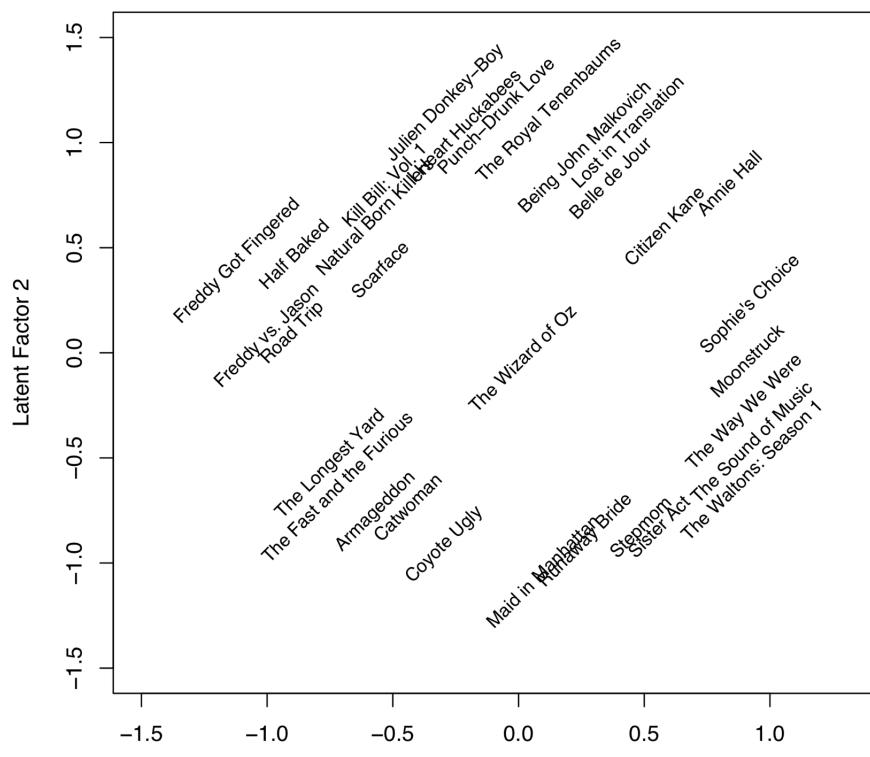
user	movie	score	date
1	21	1	2002-01-03
1	212	?	2003-01-03
1	1123	?	2002-05-04
2	25	?	2002-07-05
2	8773	?	2002-09-05
2	98	?	2004-05-03
3	16	?	2003-10-10
4	2450	?	2004-10-11
5	2032	?	2004-10-11
5	9098	?	2004-10-11
5	11012	?	2004-12-12
6	664	?	2005-01-02
6	1526	?	2005-01-31

- Competition

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

Courtesy of Chris Volinsky

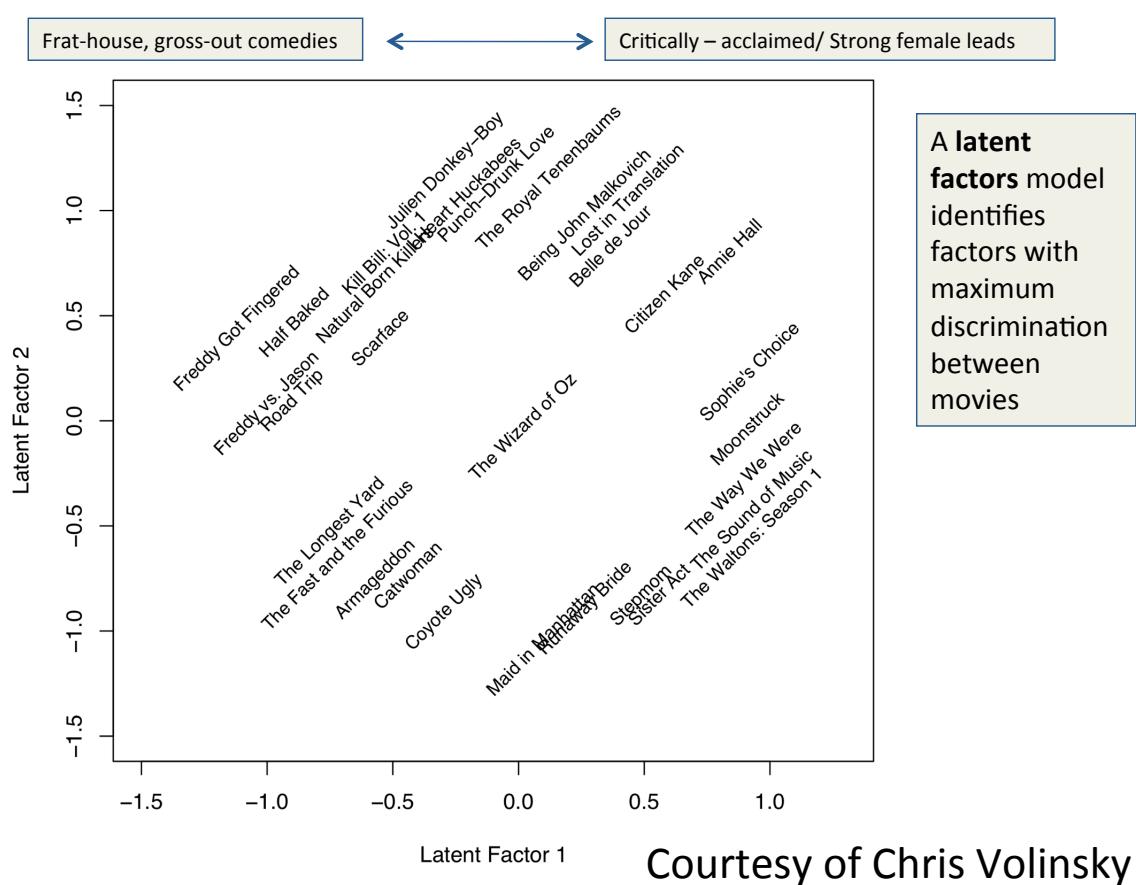
Latent Factors Model



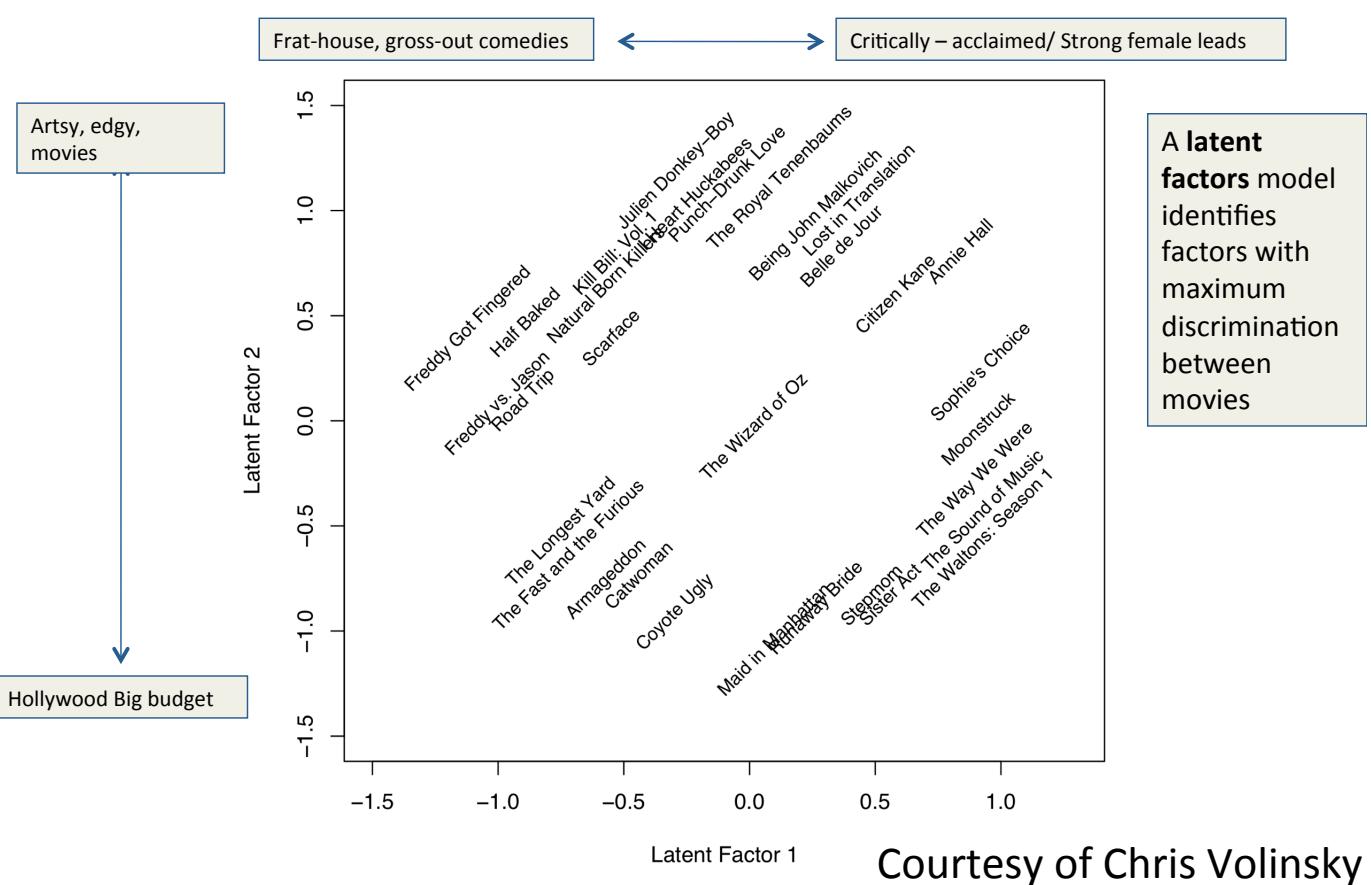
A **latent factors** model identifies factors with maximum discrimination between movies

Courtesy of Chris Volinsky

Latent Factors Model



Latent Factors Model



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.”

The screenshot shows a Mac OS X desktop with a browser window open to the Netflix Tech Blog. The title of the page is "3 Years Later...". The main content of the post discusses the evaluation of new recommendation methods offline and the decision not to implement them due to engineering costs. The post is dated Friday, April 6, 2012, and is by Xavier Amatriain and Justin Basilico. The sidebar includes links to other Netflix blogs and an RSS feed.

Friday, April 6, 2012

3 Years Later... Evaluating Some Offline Methods (Part 1)

by Xavier Amatriain and Justin Basilico (Personalization Science and Engineering)

In this post, we will compare different data mining approaches to the Netflix Prize challenge, outline the external components of our personalized service, and highlight how our task has evolved with the business. In Part 2, we will describe some of the data and models that we use, and discuss our approach to building a system that can learn from training data and predict user ratings. Every year, many teams around the world are attempting to solve this challenging problem. If you are interested in learning more about what teams have been doing, take a look at our jobs page.

In 2006 we announced the Netflix Prize, a machine learning and data mining competition for movie rating prediction. We offered \$1 million to whoever improved the accuracy of our existing system, and Cinematch 90% accuracy was the baseline and needed to improve to 97.0% accuracy on the test set. This was a hard problem. However, it was a very good one to work on because it was easier to evaluate and quantify: the root mean squared error (RMSE) of the predicted rating. The race was on to beat our RMSE of 0.9525 with the finish line of reducing it to 0.8572 or less.

After much iteration, the Netflix team found a linear blend of two algorithms that beat the baseline. They gave them this prize. And they gave us the source code. We looked at the two underlying algorithms and found that the best performance in the ensemble: Matrix Factorization (which the community generally called SVD, Singular Value Decomposition) and Restricted Boltzmann Machines (RBM). SVD by itself provided a 0.8914 RMSE, while RBM alone provided a competitive but slightly worse 0.8990 RMSE. A linear blend of these two reduced the error to 0.88. To put these algorithms to use, we had to work to overcome some limitations, for instance that they were built to handle 100 million ratings, instead of the more than 5 billion that we have, and that they were not built to adapt as members added more ratings. But once we overcame those challenges, we put the two algorithms into production, where they are still used as part of our recommendation engine.

Links

- Netflix Canada Blog
- Netflix America Latina Blog
- Netflix Brasil Blog
- Netflix DACH Blog
- Netflix EMEA Blog
- Netflix UK & Ireland Blog
- Open Positions at Netflix
- Netflix Website
- Netflix Creative Platform
- Netflix UI Engineering

RSS Feed

About the Netflix Tech Blog

This is a Netflix blog focused on technology and technology issues. We'll share our perspectives, decisions and challenges regarding the software we build and use to create the Netflix service.

Xavier Amatriain and Justin Basilico, 2012

Ad-targeting

Ads ⓘ

Yacht Inbox x

[REDACTED] 1:19 PM (1 minute ago) ⭐ ↻ ▾

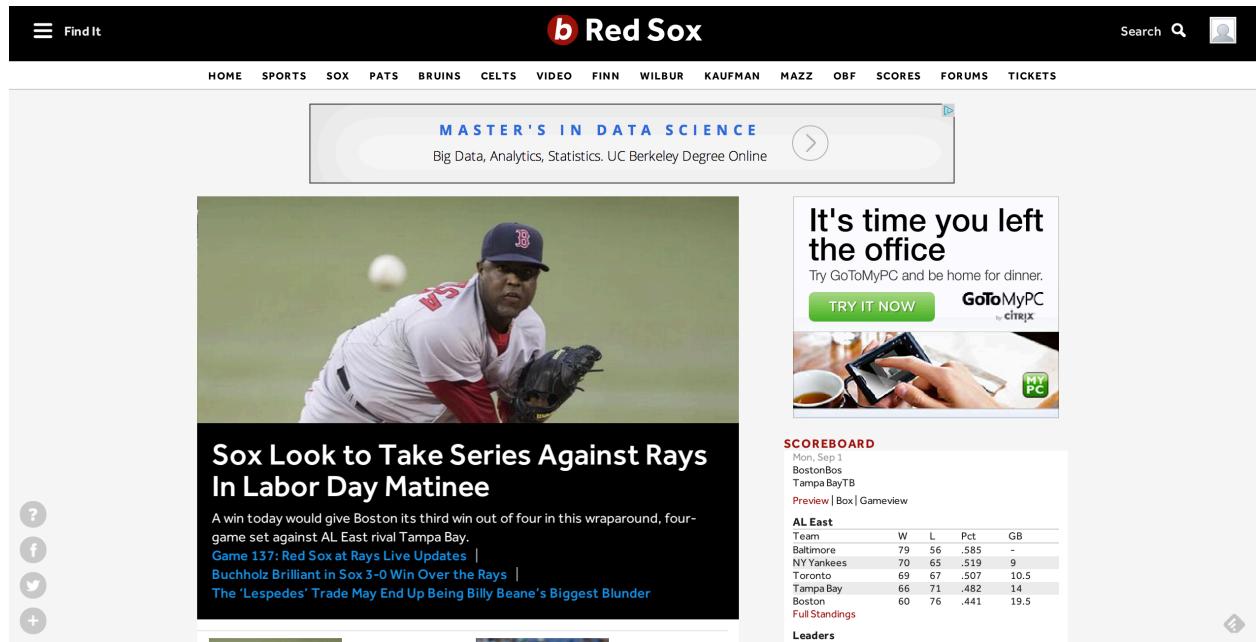
Suit yourself. I'll send you pictures from my yacht.

Making Sense of Big Data
A Big Data Guide for Small & Medium Businesses. Get the Free eMagazine!
www.tableausoftware.com/big-data

Dell™ Computer Outlet
Shop Dell™ Outlet For Discounted Computer Refurbs, w/ Intel® Core™
www.Dell.com/Outlet

Luxury BVI Cruise
7 Night Small Ship BVI Cruise
from \$2,595. Book Now & Save 50%
pgcruises.com/BVI_Cruise

BVI Yacht Charter
Yacht Charter in the BVI
bareboat and with great crews.
www.ViSailing.com



≡ Find it

b 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](#) 



SCOREBOARD
Mon, Sep 3
Boston@Tampa Bay
[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](#)



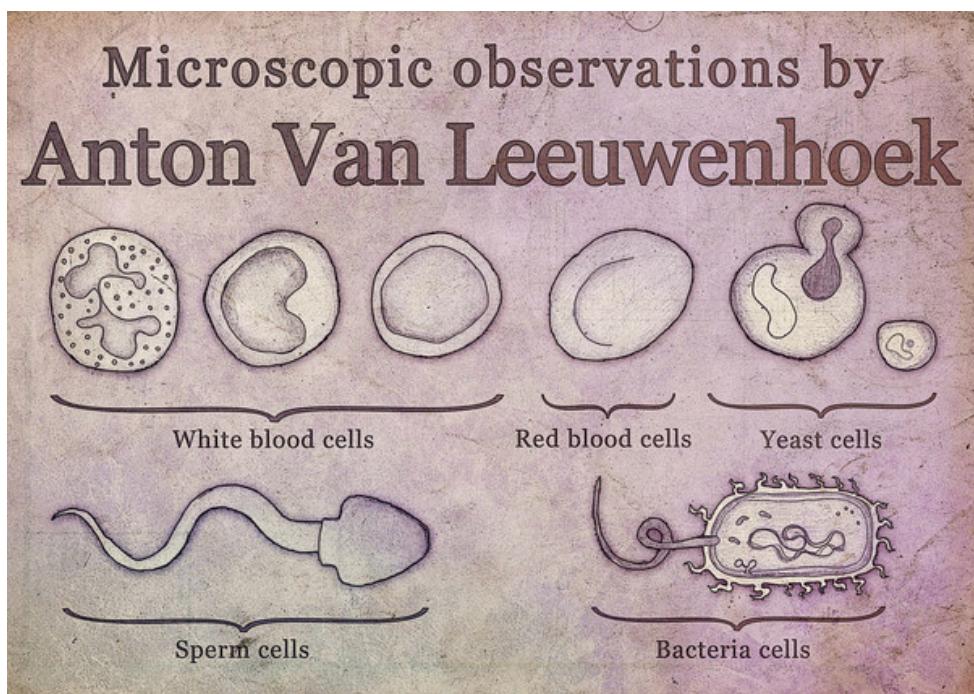
Biology

Anton Van Leeuwenhoek (1623-1723)

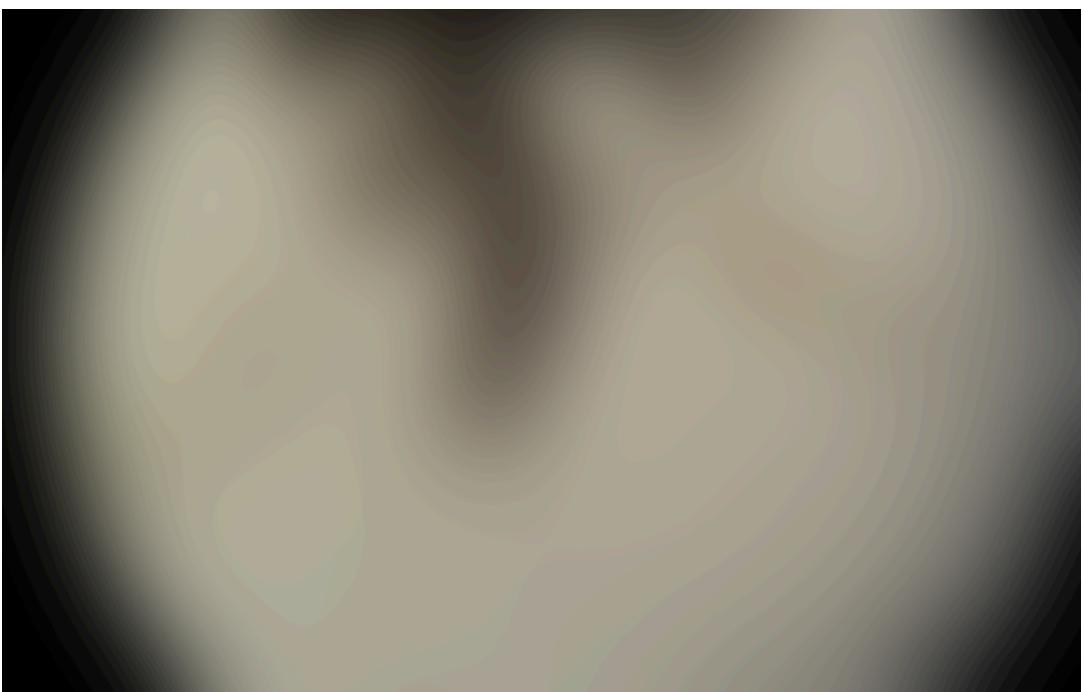


The “father of microbiology”

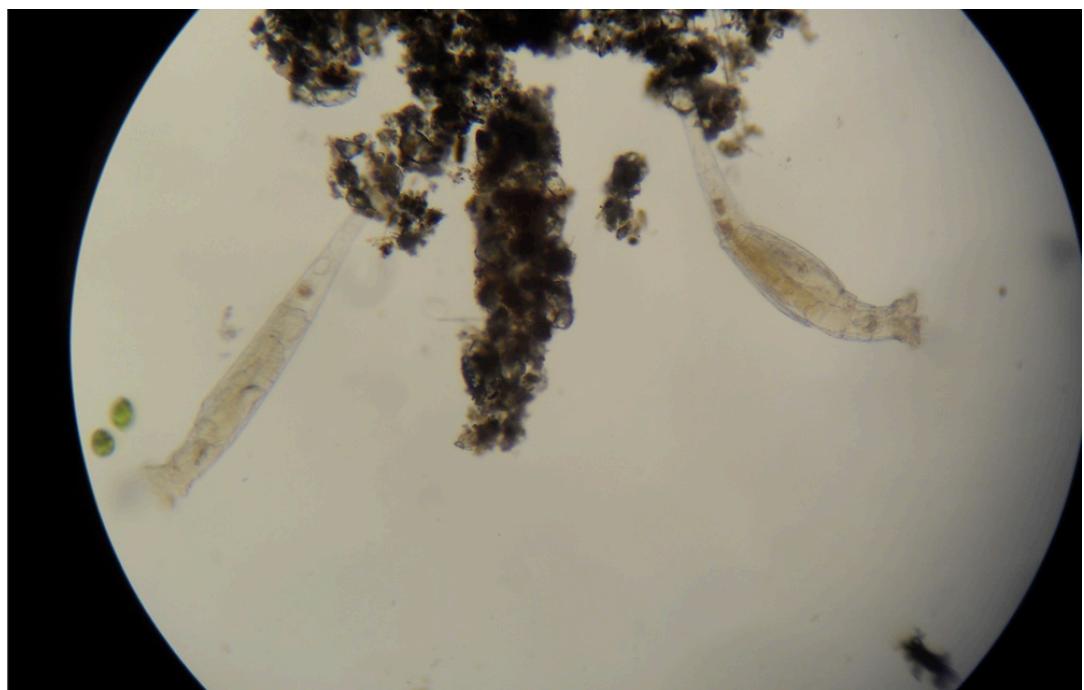
Some of his discoveries



By improving the microscope



he saw what others could not



21st century version



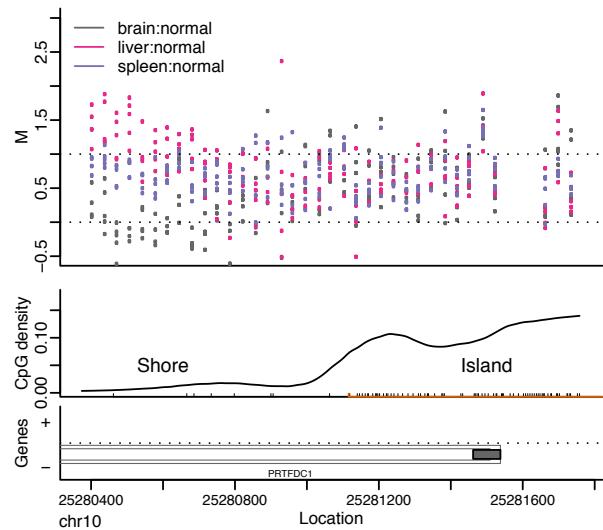
Modern high-throughput technology

Produces complex data, not images

21st century version



Modern high-throughput technology

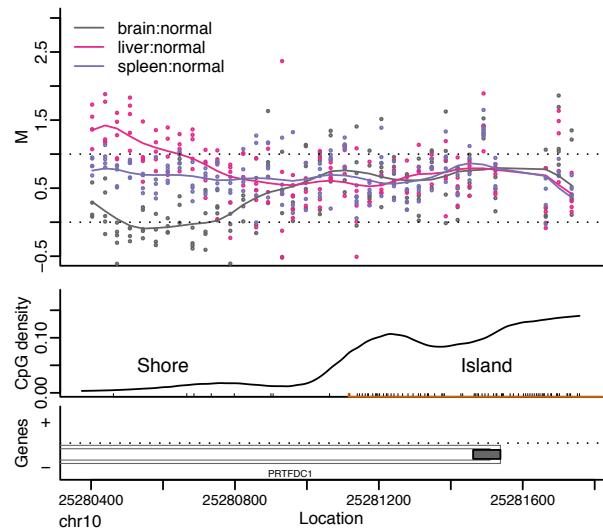


My work has helped bring data into focus

21st century version



Modern high-throughput technology

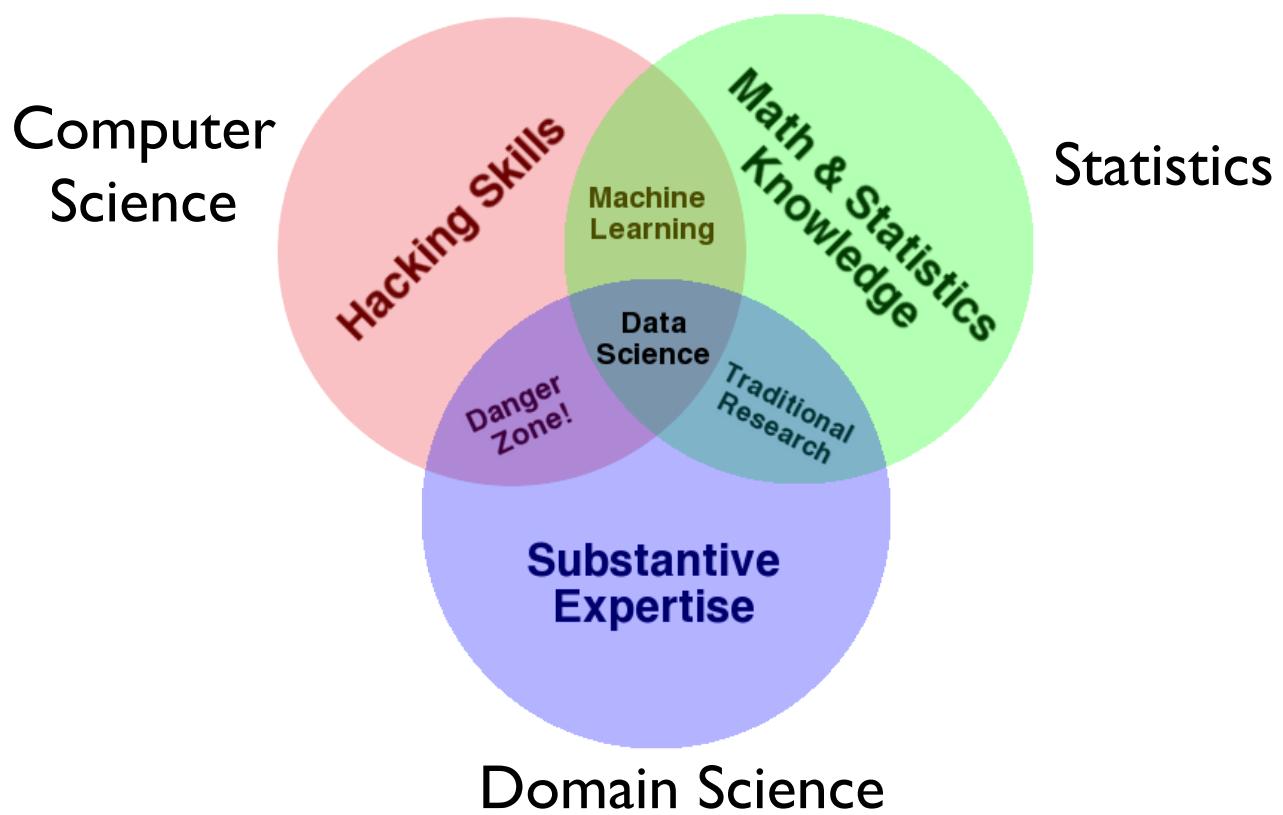


My work has helped bring data into focus

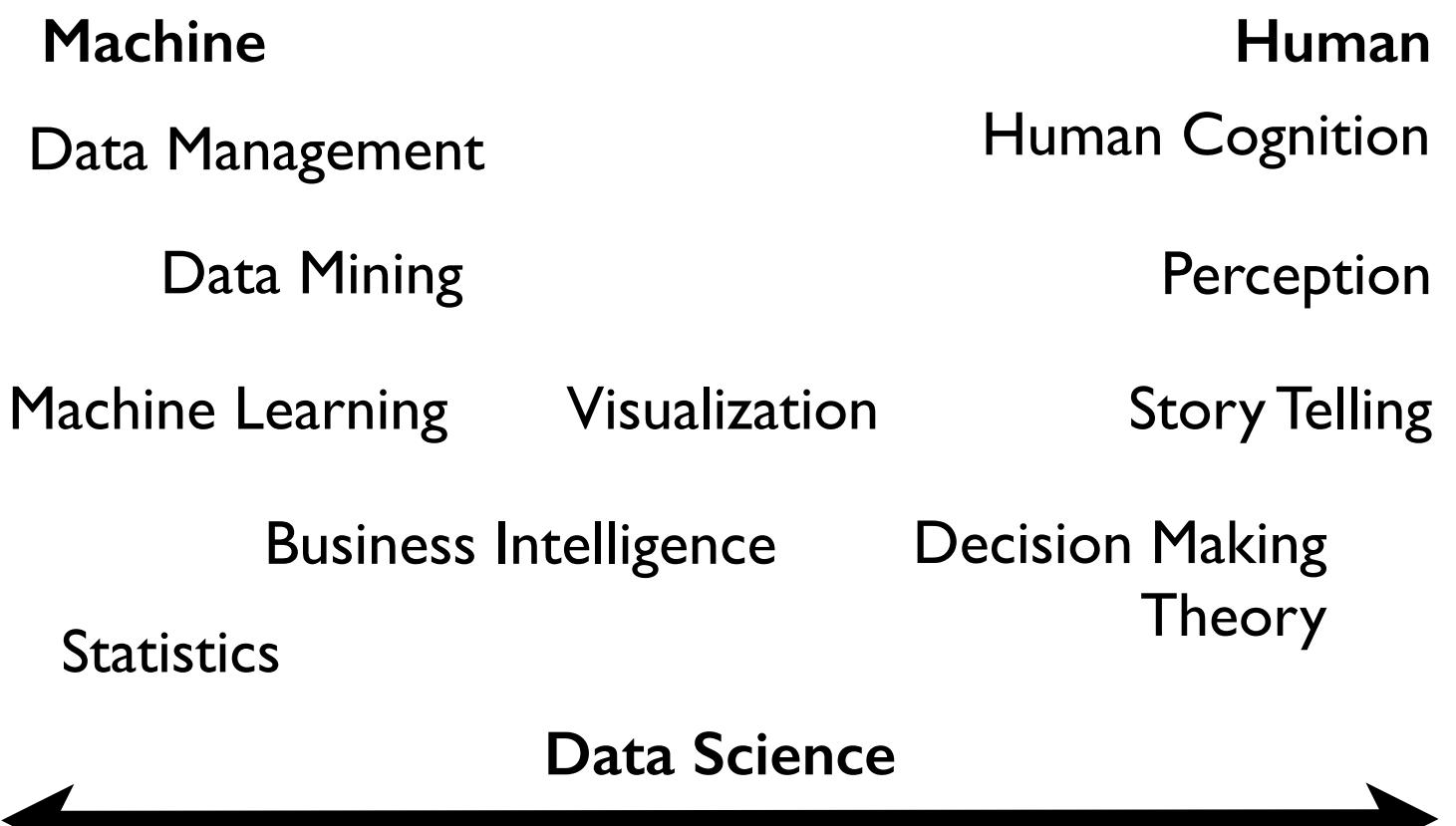
Many other examples

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

Data Science

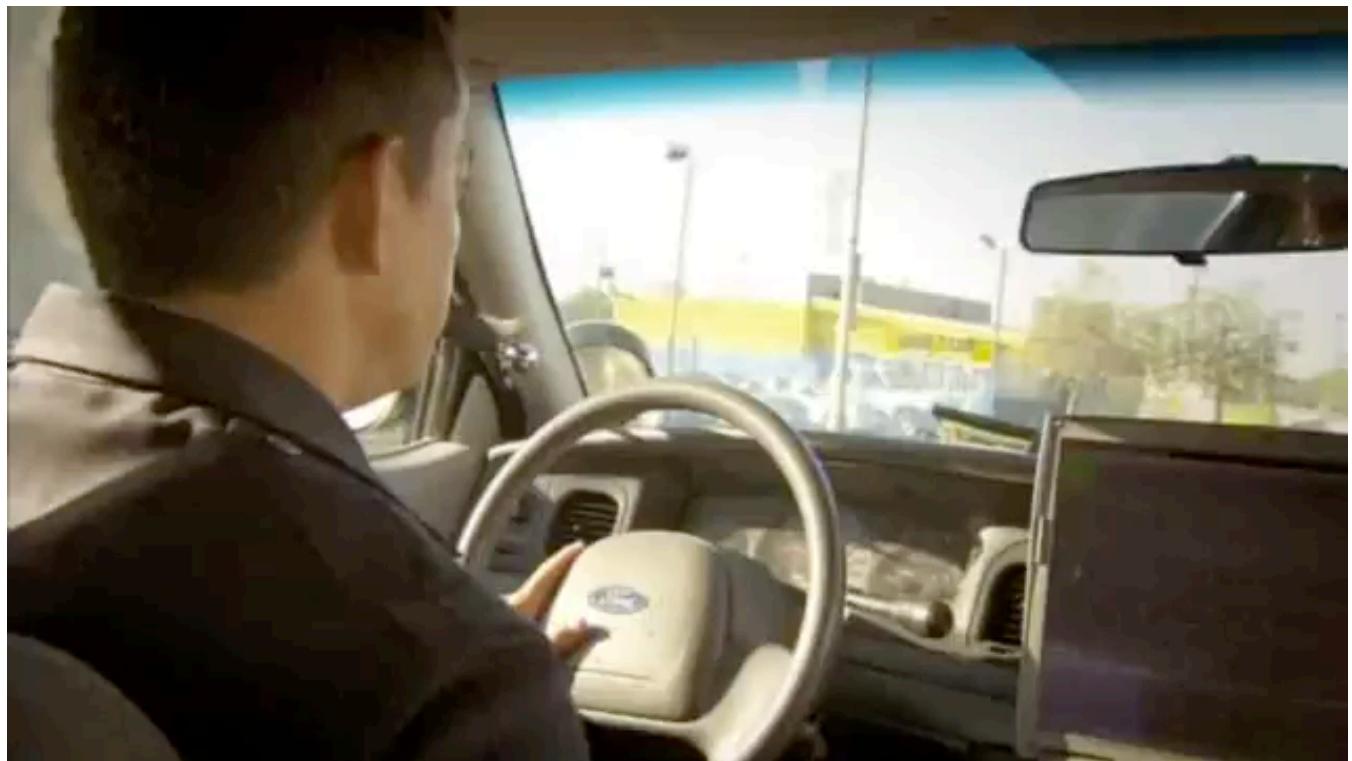


Drew Conway



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.”

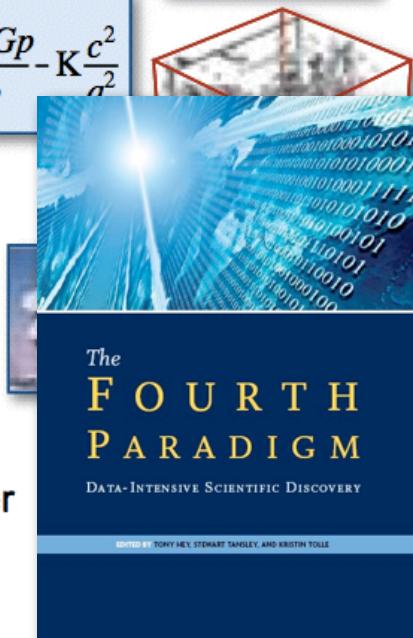
Eric Schmidt, Google (and others)

A screenshot of a Google search results page. The search query "you tube cat videos" is entered in the search bar. Below the search bar, the "Web" tab is selected, along with "Images", "Maps", "Shopping", "News", "More", and "Search tools". A red circle highlights the text "About 1,030,000,000 results (0.33 seconds)". Below this, there is an advertisement for "TheFriskies.com" encouraging users to vote for funny cat videos. The ad includes the URL "www.thefriskies.com/ContestEntry", the text "The Friskies Will Honor The Best New Cat Videos. Cast Your Vote Now!", and links to "More About The Awards" and "Visit Friskies.com".

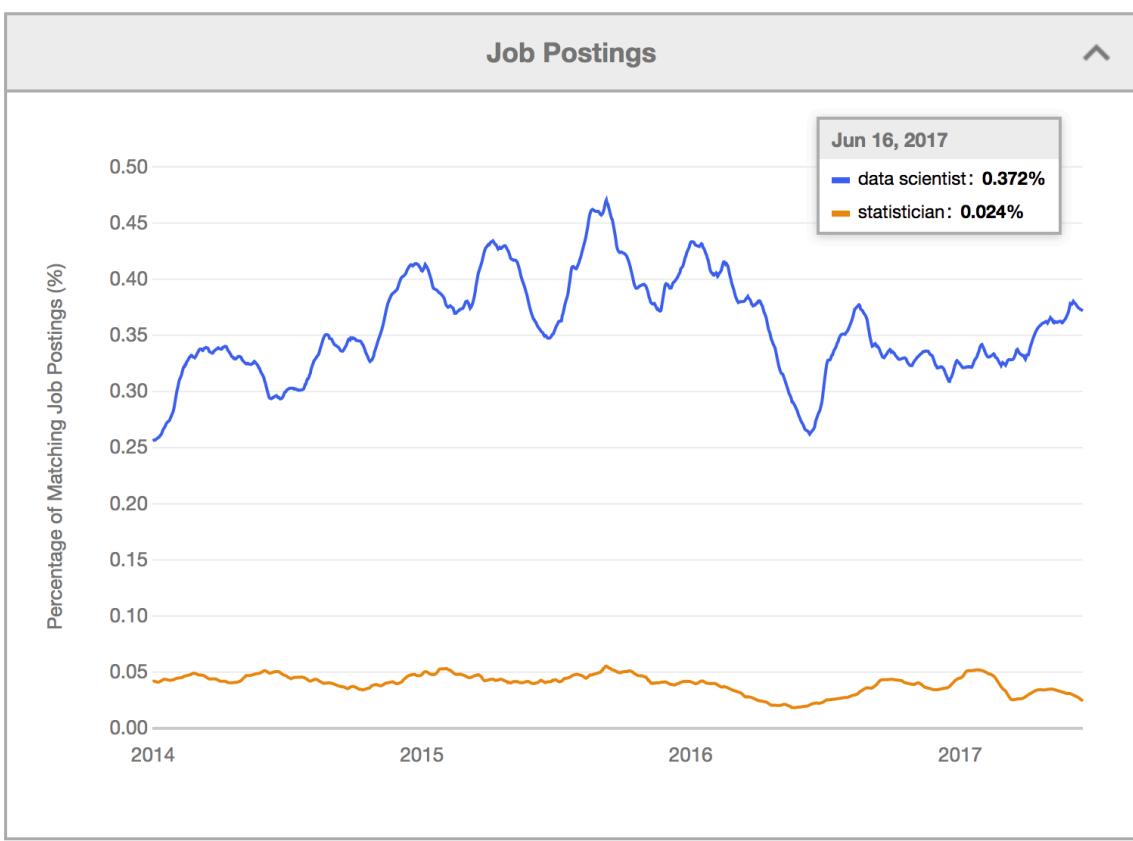
Science Paradigms

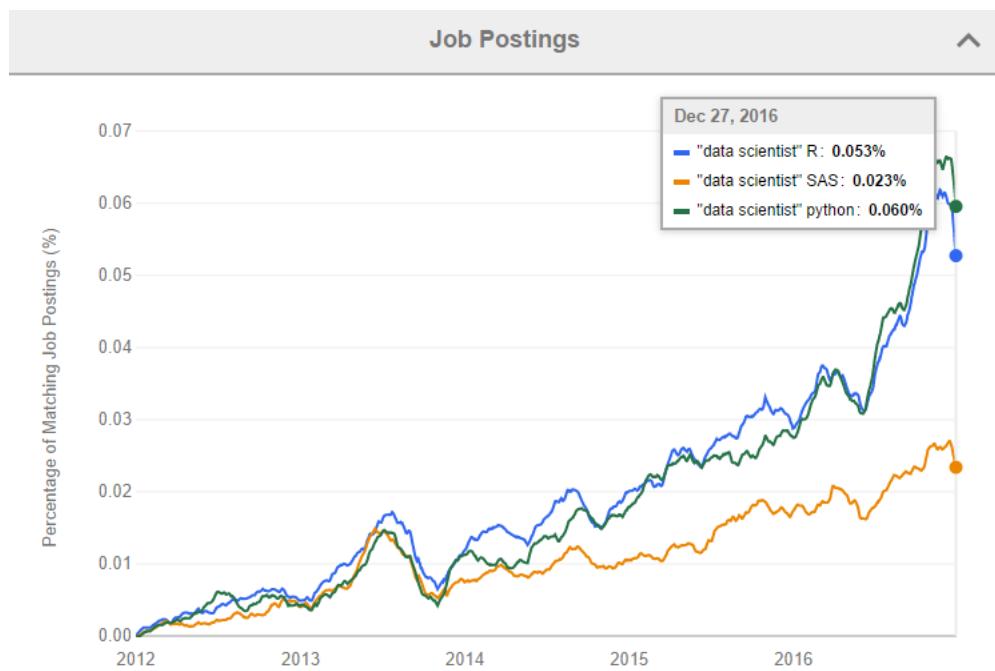
- Thousand years ago:
science was empirical
describing natural phenomena
- Last few hundred years:
theoretical branch
using models, generalizations
- Last few decades:
a computational branch
simulating complex phenomena
- Today: **data exploration (eScience)**
unify theory, experiment, and simulation
 - 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

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

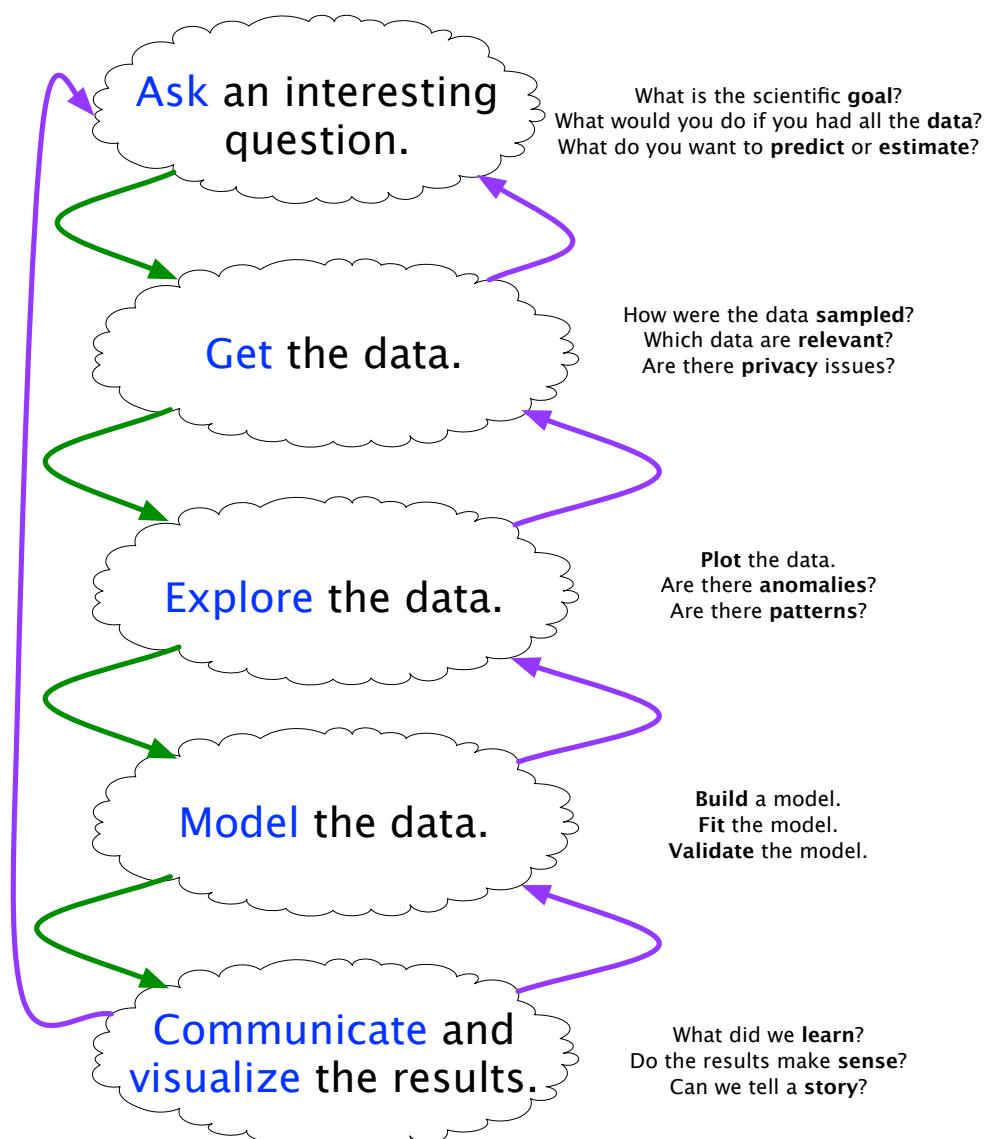


Jim Gray, Microsoft





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

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

