



Statistics and Data Science

What is Statistics?

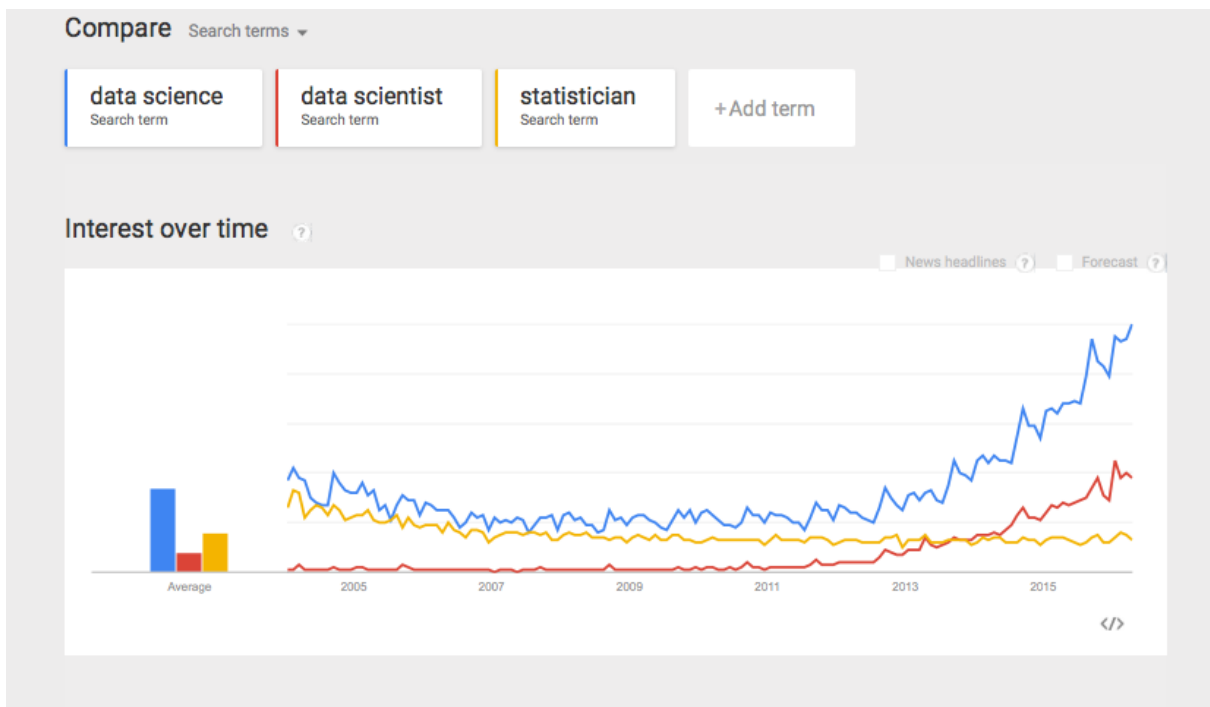
Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.

It is much more than: compilation of baseball scores, life and death tables, etc!

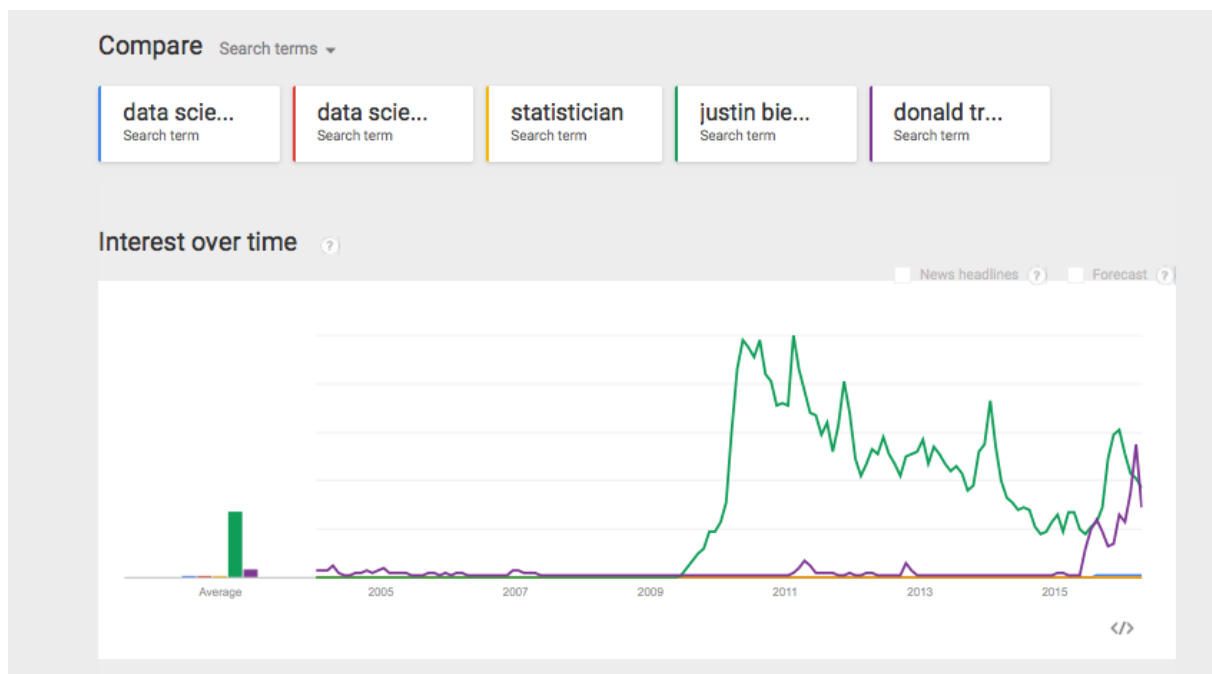
Some more definitions

- **Machine learning** constructs *algorithms that can learn from data*, especially for *prediction*
- **Statistical learning** is branch of Statistics that was born in response to Machine learning, emphasizing *statistical models and assessment of uncertainty*
- **Data Science** is the *extraction of knowledge from data*, using ideas from mathematics, statistics, machine learning, computer programming, data engineering ...

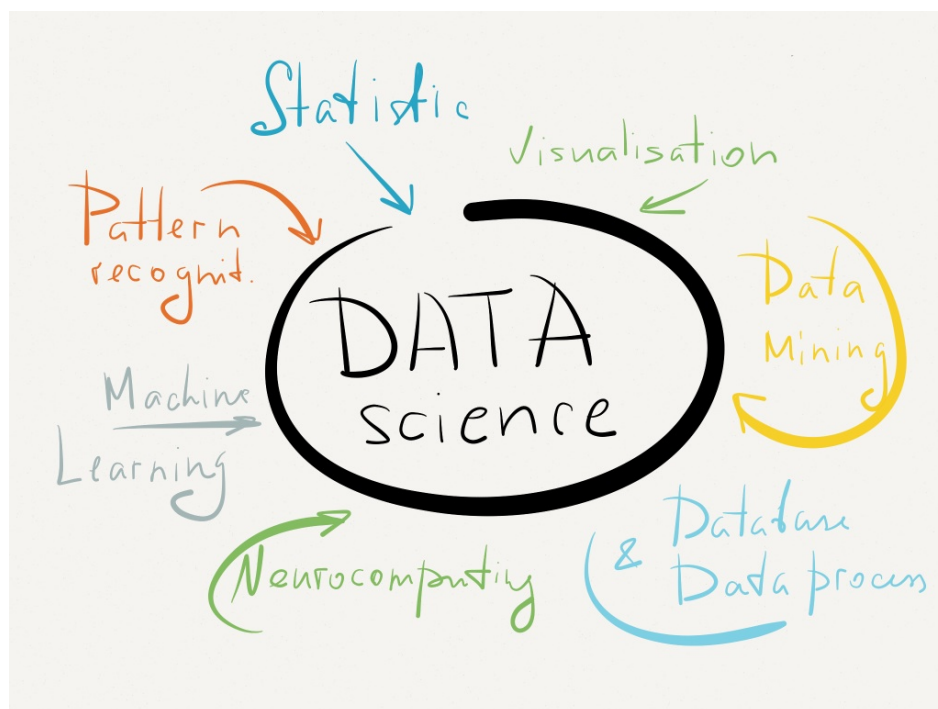
Google trends



Keeping things in perspective



One view of their relationship



20th Century Innovation

Engineering and Computer Science played key role in:

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

Statistics also played a key role

helping to answer questions like:

- Does fertilizer increase crop yields?
- Does Streptomycin cure Tuberculosis?
- Does smoking cause lung-cancer?:
- Who will win the election?

... to answer these questions

Answers:

- Does fertilizer increase crop yields? *Collect and analyze agricultural experimental data*
- Does Streptomycin cure Tuberculosis? *Analyze randomized trials data*
- Does smoking cause lung-cancer? *Collect and analyze observational studies data*
- Who will win the election? *Collect data and try to extropolate*

Answering these was the job of: boring old statisticians

Big data explosion

21st Century



In God we trust. All others bring data.

attributed to W. Deming

Emerging themes

1. The availability of massive amounts of **good data** and **fast computation** changes everything.
Examples: IBM's Watson for Playing Jeopardy, speech recognition, Google translate
2. New data is **complex and unstructured**.

Emerging themes

1. The availability of massive amounts of **good data** and **fast computation** changes everything.

Examples: IBM's Watson for Playing Jeopardy, speech recognition, Google translate

2. New data is **complex and unstructured**.

Structured data: a flat file with a fixed number of measurements. Eg. response of a patient to a drug, and 10 measurements— age, sex (not gender), lab measurements.

Unstructured data: doctor's notes, Twitter feeds, broker reports

The availability of massive amounts of *good* data
changes everything

The Unreasonable Effectiveness of Mathematics in the Natural Sciences

Eugene Wigner, 1960. “Isn’t it remarkable how well we can predict physical phenomena using simple equations like $F = ma$ or $PV = nRT$?”?

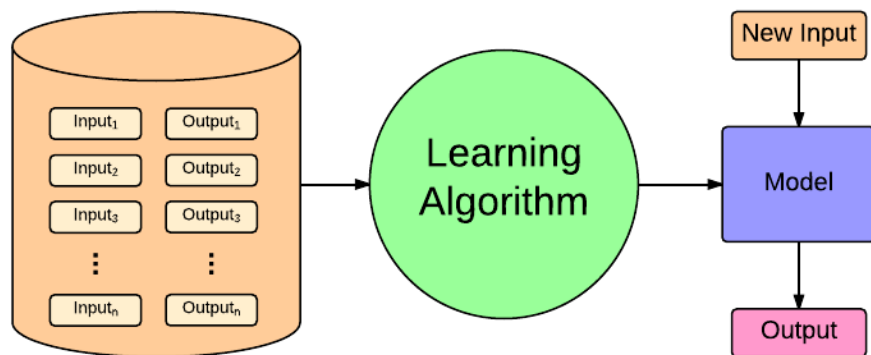
The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, & Fernando Pereira, 2009. “Isn’t it remarkable how well we can predict human phenomena using simple statistical models

21st century examples

- Use *Classification techniques* to classify which accounts are the most likely to upgrade their service contract (this helps the salesforce to know which leads / accounts to focus on to sell more)...
- Use *Regression techniques* to improve how sales people pitch prospective clients and what features of the company's services they should highlight...
- *Regression and classification* are examples of *Supervised Learning* techniques (see next slide)
- Use *Optimization techniques* to maximize the number of views of company promotional material a prospective customer sees for a given dollar amount of promotional spend

The Supervising Learning Paradigm



Training Data

Fitting

Prediction

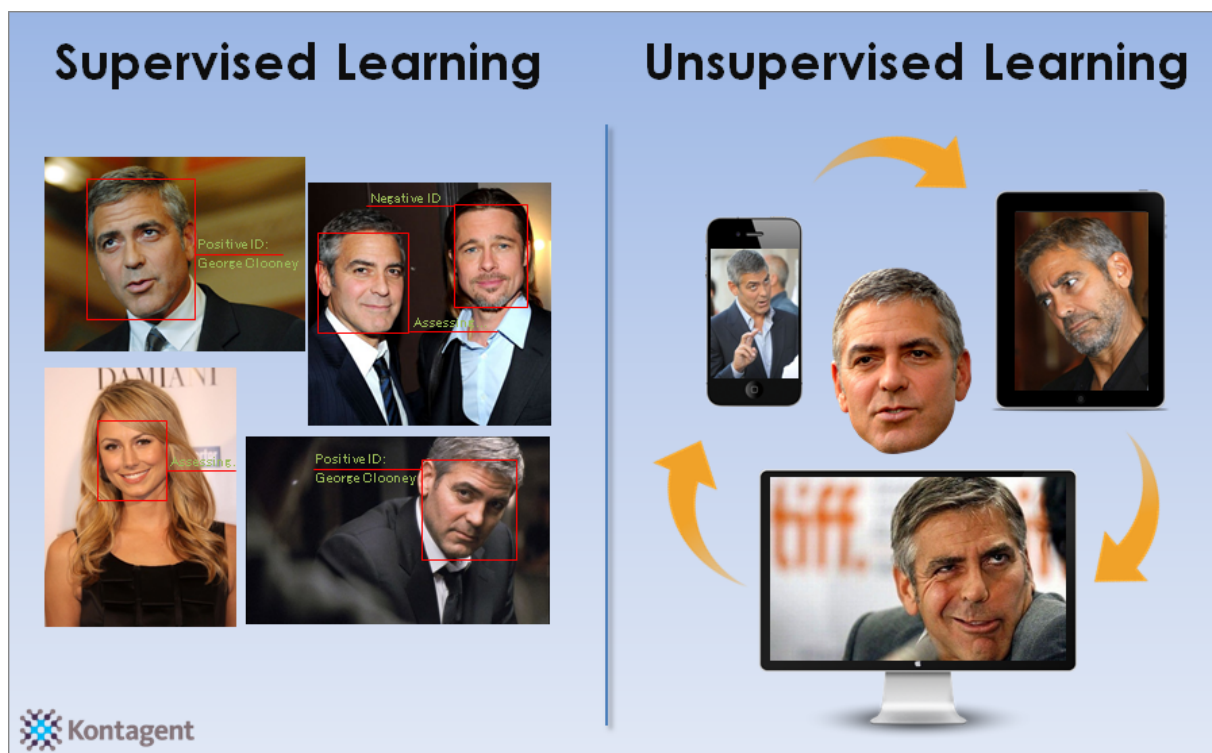
Traditional statistics: domain experts work for 10 years to learn good features; they bring the statistician a small clean dataset

Today's approach: we start with a large dataset with many features, and use a machine learning algorithm to find the good ones. **A huge change.**

Supervised vs Unsupervised learning

Supervised: Both inputs (features) and outputs (labels) in training set

Unsupervised: No output values available, just inputs.



The sexiest job?

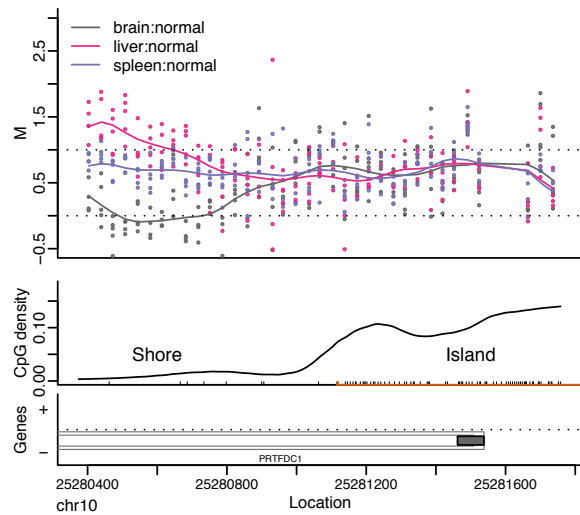
“I keep saying the sexy job in the next ten years will be statisticians. People think I’m joking, but who would’ve guessed that computer engineers would’ve been the sexy job of the 1990s”

- Hal Varian, Google’s Chief Economist

21st century version



Modern high-throughput technology



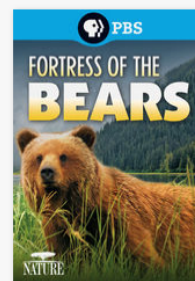
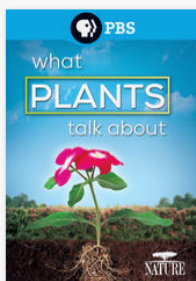
Statistics has helped bring data into focus



Business and marketing: a new era

The Netflix Recommender

Awesome, glad you enjoyed it! Try these next...



How often do you watch PBS?

This will help improve the suggestions you get overall.



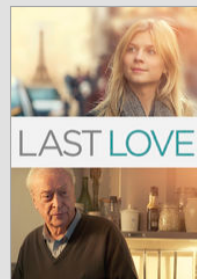
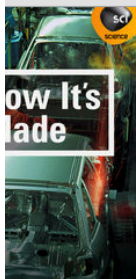
Never



Sometimes



Often



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](#) | [Compu](#)


Bits

Business • Innovation • Technology • Society

September 21, 2009, 10:15 AM

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

By STEVE LOHR



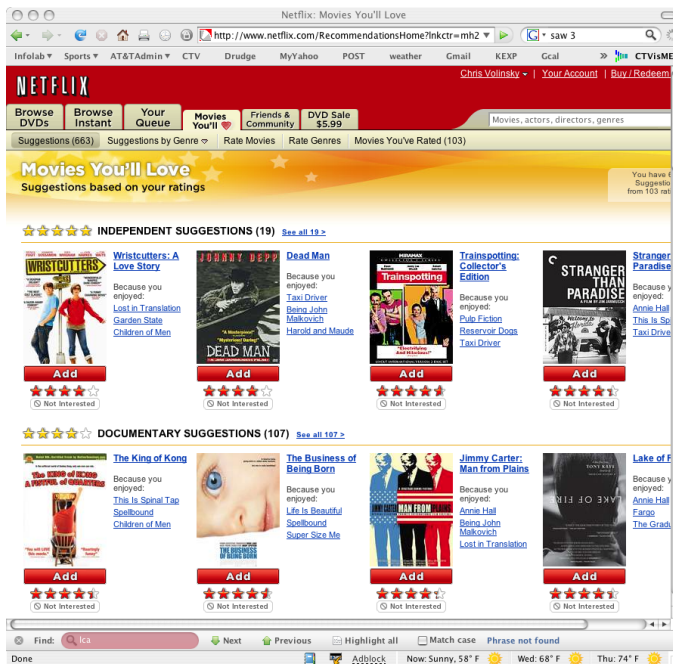
Jason Kempin/Getty Images

Netflix prize winners, from left: Yehuda Koren, Martin Chabbert, Martin Plotte, Michael Jahner, 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

- 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	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

- 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

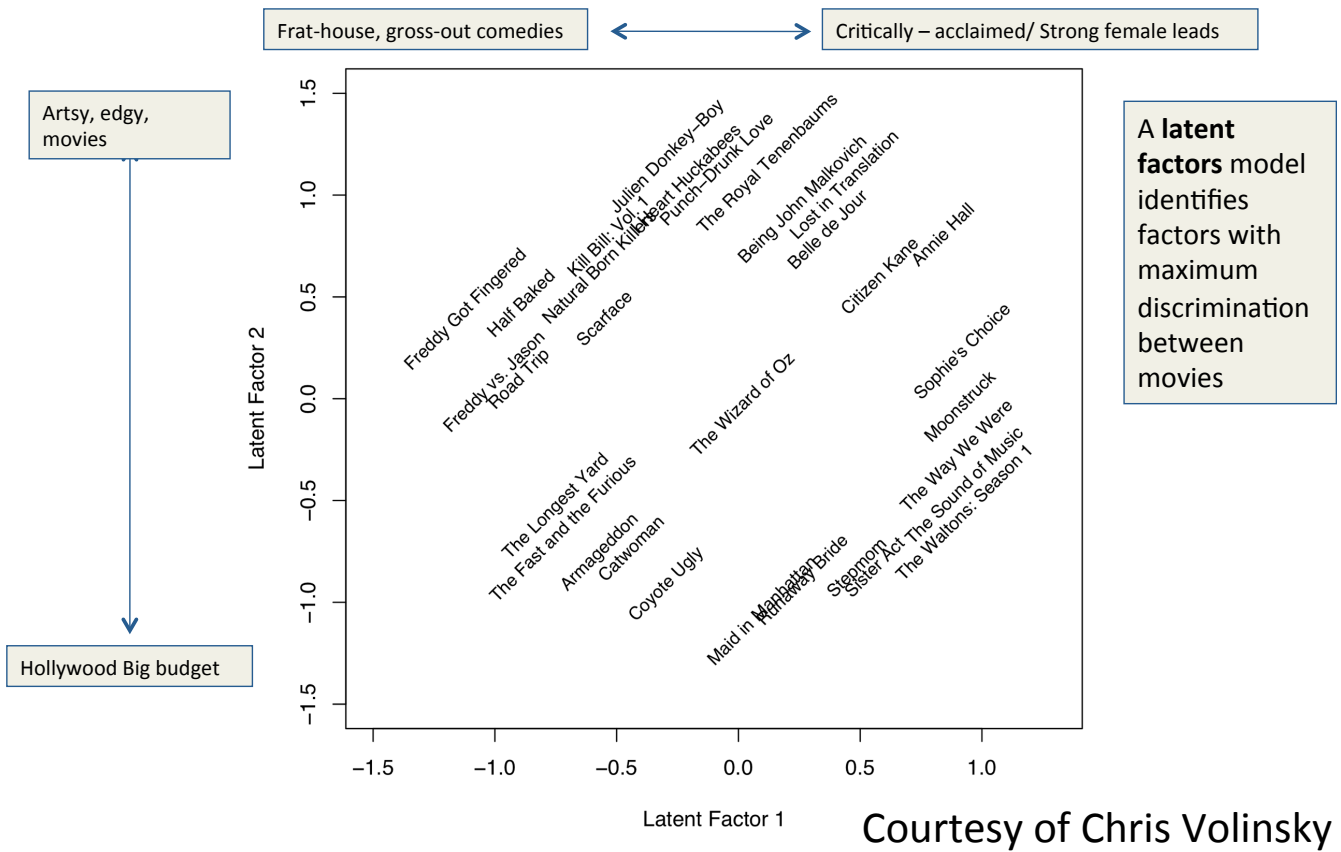
- Competition

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

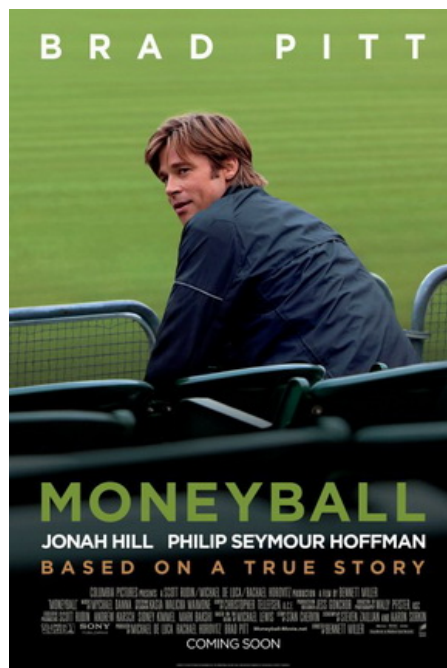
user	movie	score	date
1	21	1	2002-01-03
user	movie	score	date
1	212	?	2003-01-03
1	1123	?	2002-05-04
2	25	?	2002-07-05
2	8773	?	2002-09-05
2	98	?	2004-05-03
3	16	?	2003-10-10
4	2450	?	2004-10-11
5	2032	?	2004-10-11
5	9098	?	2004-10-11
5	11012	?	2004-12-12
6	664	?	2005-01-02
6	1526	?	2005-01-31

Courtesy of Chris Volinsky

Latent Factors Model



Data science in sports and politics



The Data Scientist

Actual

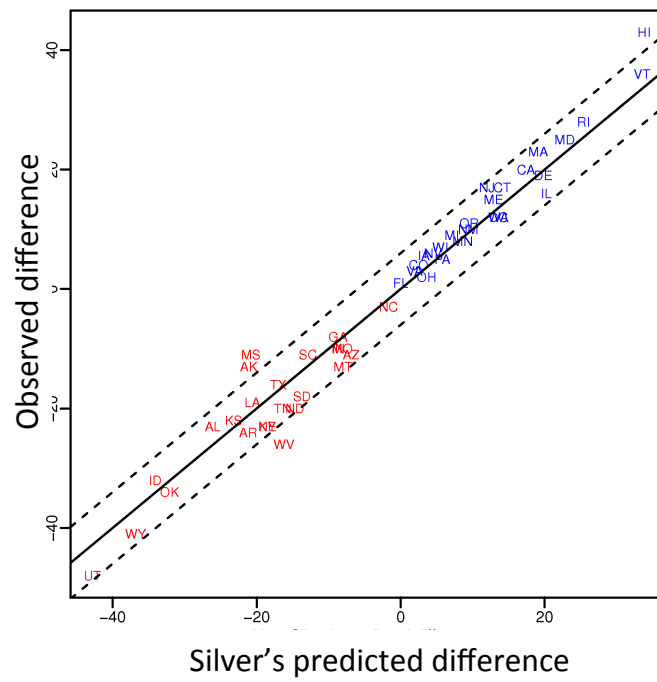


Hollywood

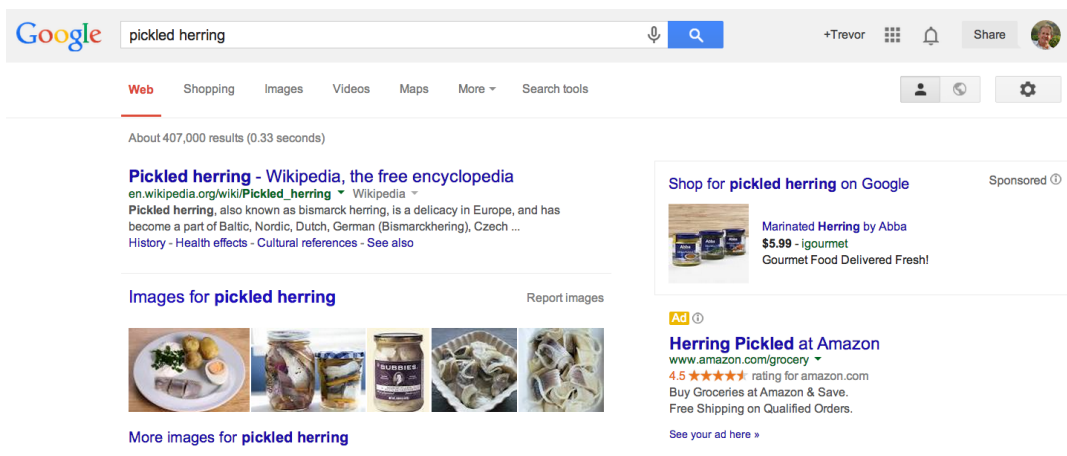


Nate Silver and 538

2012 results



How Google has changed advertising



Click-through rate. Based on the search term, knowledge of this user (IP Address), and the Webpage about to be served, what is the probability that each of the 30 candidate ads in an ad campaign would be clicked if placed in the right-hand panel?

Supervised learning with billions of training observations. Each ad exchange does this, then bids on their top candidates, and if they win, serve the ad — all within 10ms!

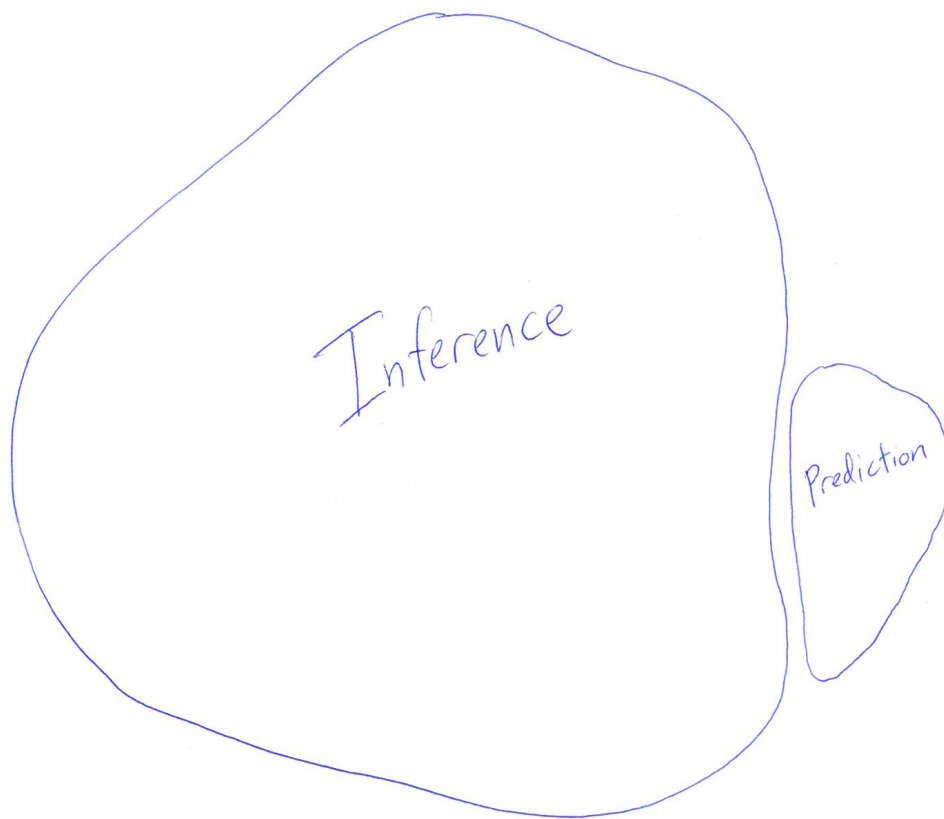
More examples of current Data Science applications

- **Adverse drug interactions.** US FDA (Food and Drug Administration) requires physicians to send in adverse drug reports, along with other patient information, including disease status and outcomes. Massive and messy data. Using natural language processing, researchers can find drug interactions associated with good and bad outcomes.
- **Social networks.** Based on who my friends are on Facebook or LinkedIn, make recommendations for who else I should invite. Predict There are more than a billion Facebook members, and two orders of magnitude more connections. Knowledge about friends informs our knowledge



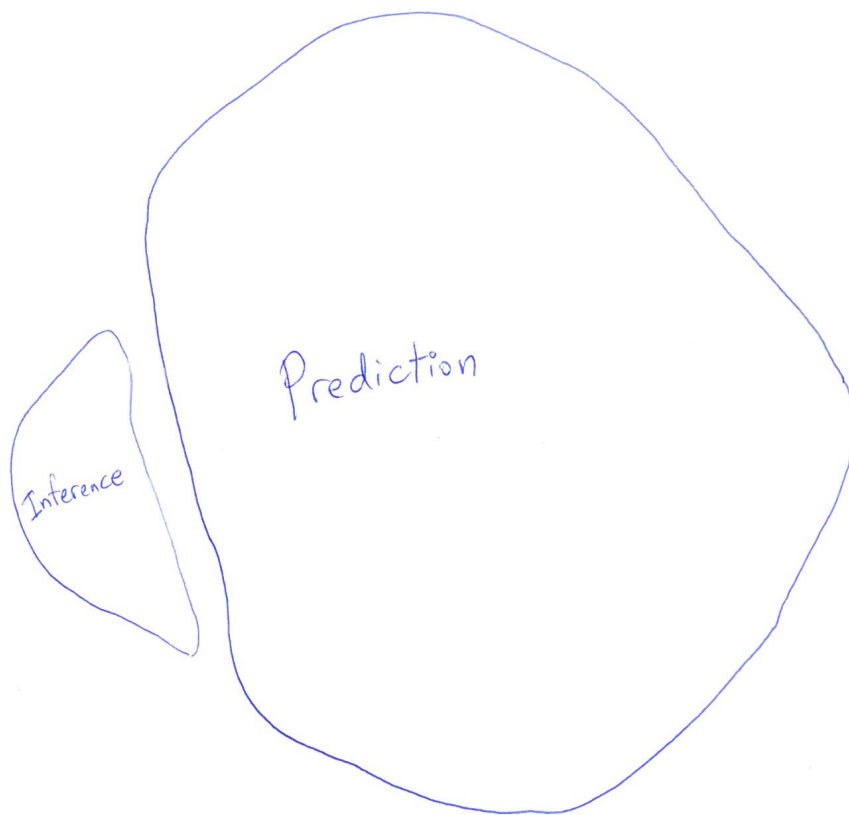
Why Statistical thinking is still important

Statistics versus Machine Learning



How statisticians see the world?

Statistics versus Machine Learning



How machine learners see the world?

Why statistical inference is important

- In many situations we care about the identity of the features— e.g. biomarker studies: Which genes are related to cancer?
- There is a crisis in reproducibility in Science: Ioannidis (2005) “Why Most Published Research Findings Are False”

