Data Science 101: an introduction

Data Science Team

Stanford University, Department of Statistics

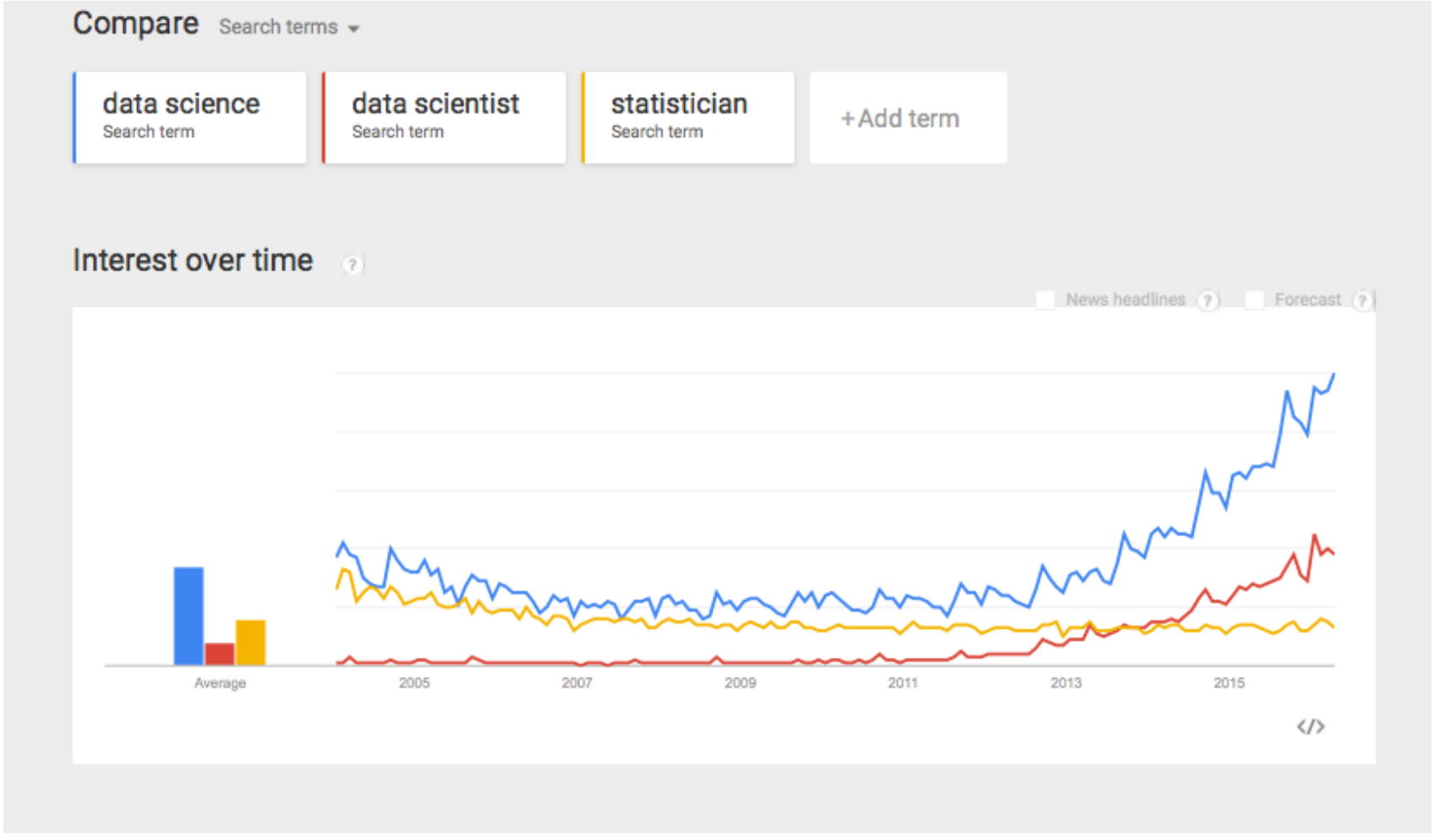
Sexy jobs

"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?"

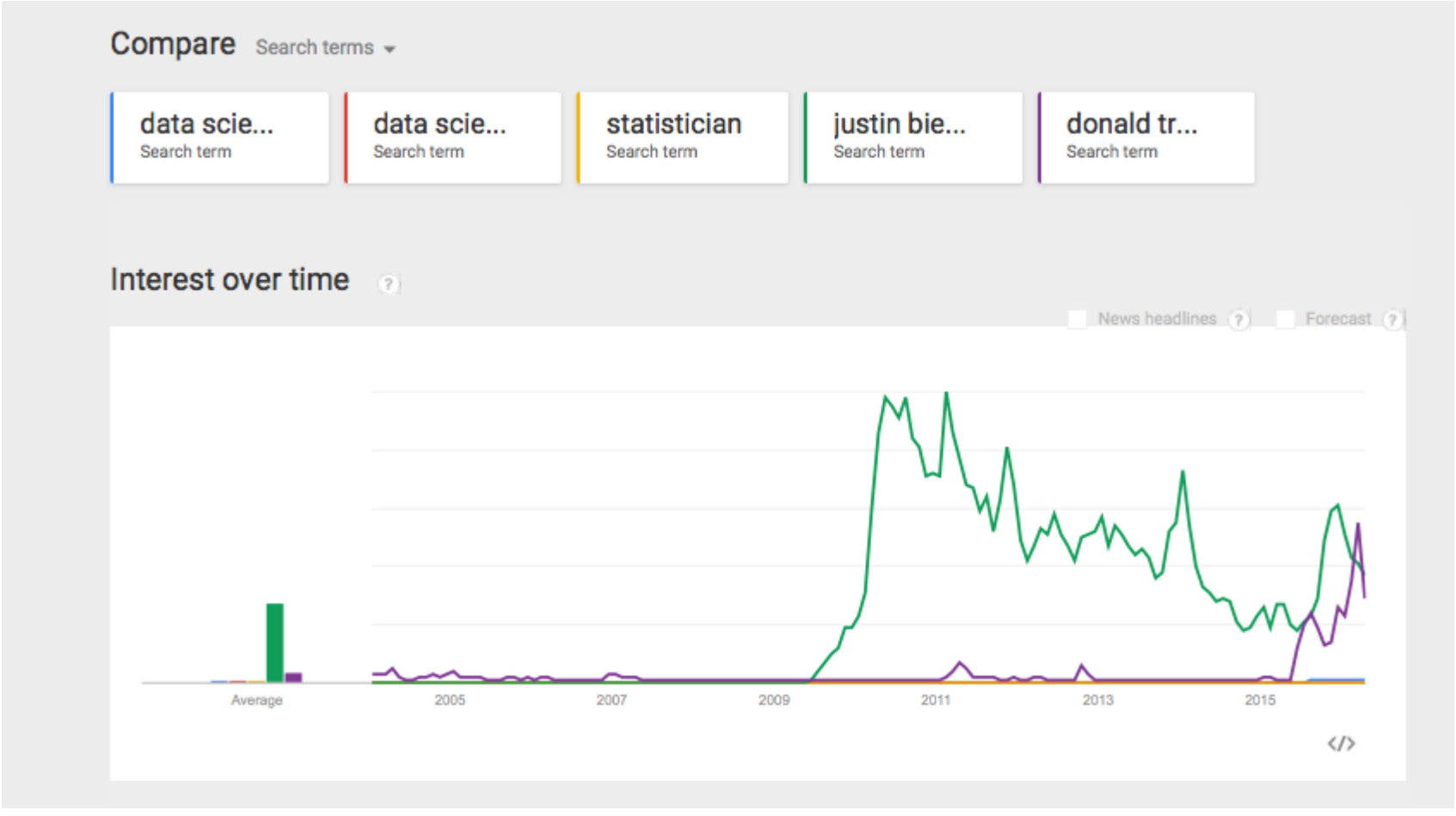
"The ability to take data, to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data."

- Hal Varian, Google's Chief Economist

Google trends



Keeping things in perspective



Data Science 101

- Not the recipe for your future start-up
- Literacy for citizenship
- Data
  - What is it?
  - Where can we find it?
  - How can we explore it?
- Science
  - What does it mean to learn from data?
  - How do we know when we are right or wrong?

There is a lot of data

Some examples of Big Data

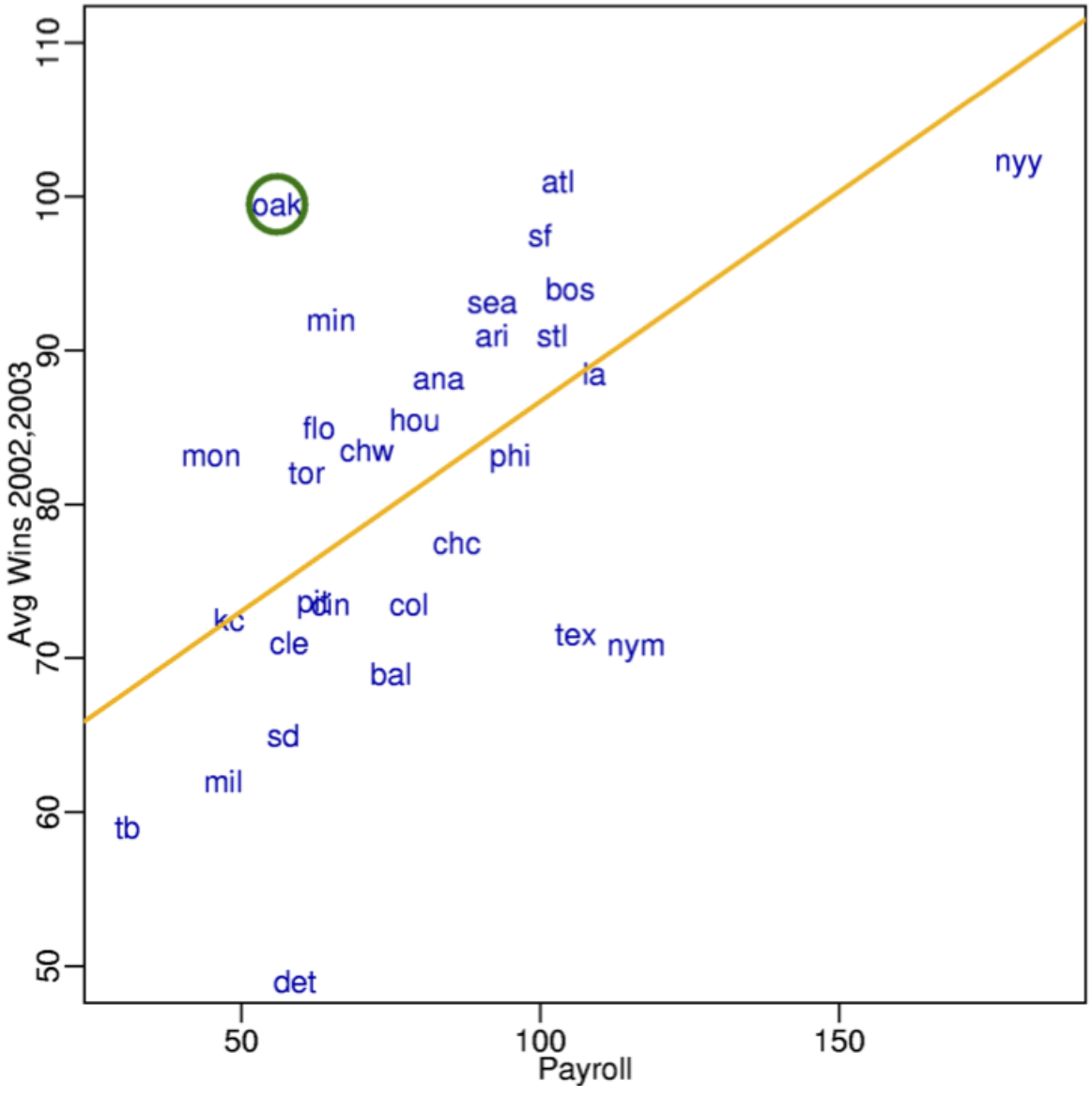
- **Genetics data**: it is easy to assess genetic variation at millions of locations in the genome, and to sequence the entire DNA of a subject; we measure expression levels of 20,000 genes in different tissues
- **Physics experiments**: the data generated in one year is 30 petabytes (petabyte of average MP3-encoded songs would require 2000 years to play)
- **Passively gathered data**
  - Products we buy
  - Topics that engage us
  - Our levels of physical activity
  - Who we talk to

Some sources of Data

- Stats for change
- Data.gov
- GEO
- dbGap
- CERN
- Private data

Data can do a lot: sports

Data can do a lot: sports



Data can do a lot: politics

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

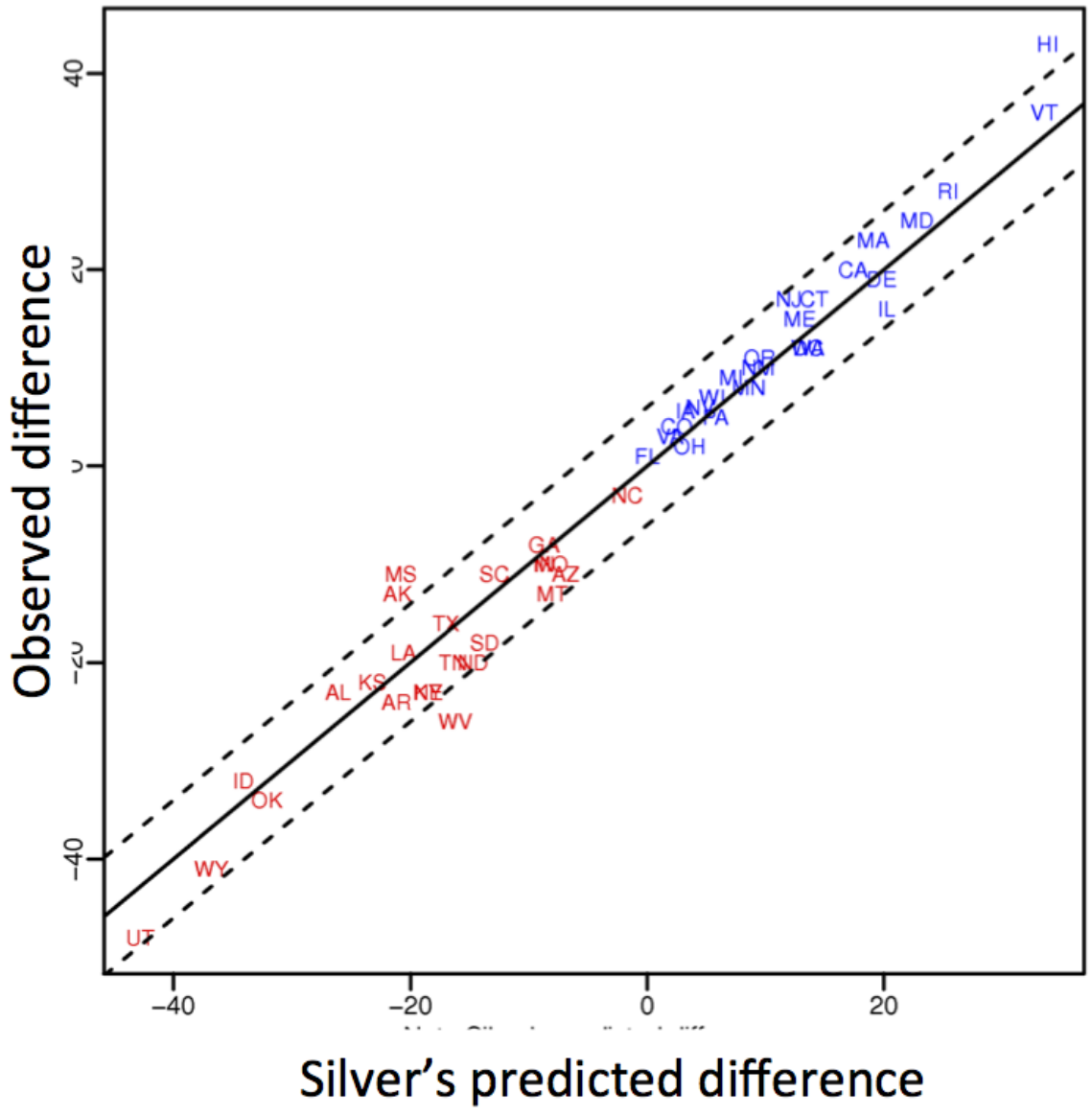
Share This Content

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.

Data can do a lot: politics



Data can do a lot: recommendations

- NETFLIX: US based streaming and DVD rental-by-mail company
- *10M customers, 10K titles, 1.9M DVDs per day*

Netflix challenge

- October 2006: Netflix offers $1M for an improved recommender algorithm.
- Training data:
  - 100M ratings
  - 480K users
  - 17,770 movies
  - 6 years of data: 2000-2005
- Test data:
  - Last few ratings of each user (2.8M)
  - Evaluation via RMSE: root mean squared error
  - Netflix Cinematch RMSE: 0.9514
- Competition:
  - $1M grand prize for 10% improvement
  - If 10% not met, $50K annual "Progress Prize" for best improvement

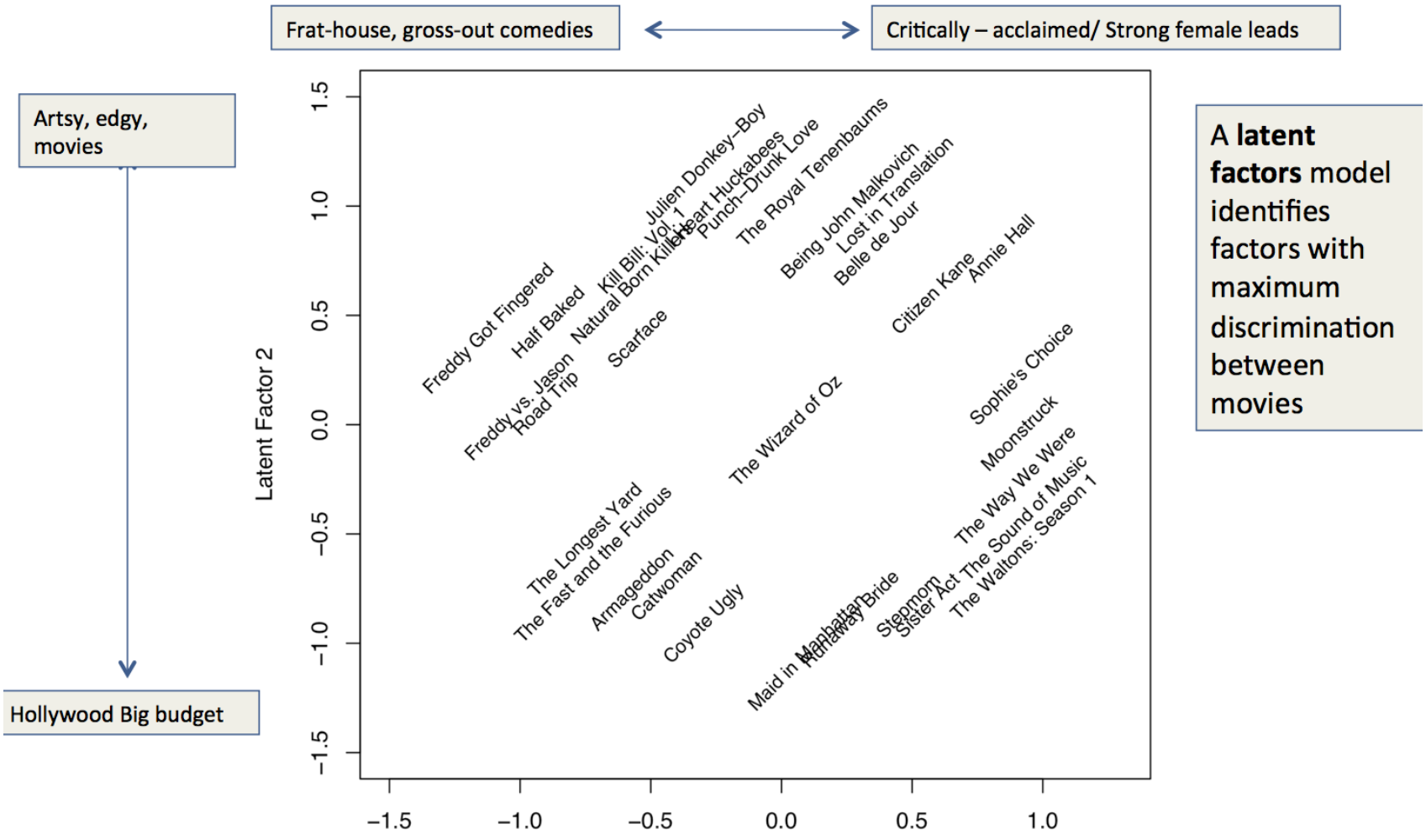
Netflix data: training

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

Netflix data: test

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

Netflix data: latent factor model



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

Netflix challenge

The New York Times
Wednesday, October 14, 2009

Technology

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.

Data can do a lot: discover bias

Stanford | News

Search Stanford news...

Home Find Stories For Journalists Contact

JUNE 15, 2016

Stanford big data study finds racial disparities in Oakland, Calif., police behavior, offers solutions

Stanford researchers analyzing thousands of data points found racial disparities in how Oakland Police Department officers treated African Americans on routine traffic and pedestrian stops. The researchers suggest 50 measures to improve police-community relations, such as better data collection, bias training and changes in cultures and systems.

BY CLIFTON B. PARKER

New Stanford research on thousands of police interactions found significant racial differences in Oakland, California, police conduct toward African Americans in traffic and pedestrian stops, while offering a big data approach to improving police-community relationships there and elsewhere.

Data can do a lot: medicine

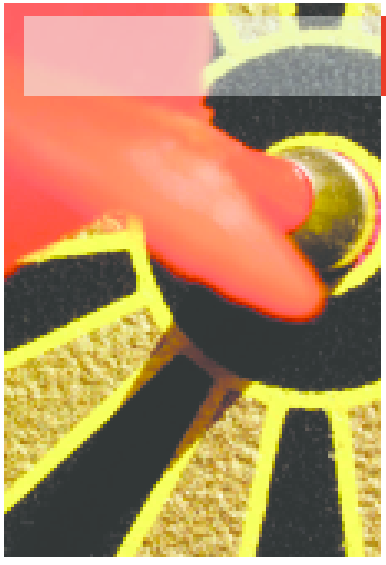
The National Institutes of Health has announced a new opportunity for organizations interested in helping engage volunteers in the *All of Us* Research Program, part of the Precision Medicine Initiative. This funding opportunity, open to national and regional organizations, as well as local community groups, will support activities to promote enrollment and retention in the *All of Us* Research Program across diverse communities.

All of Us is an ambitious effort to gather data over time from 1 million or more people living in the United States, with the ultimate goal of accelerating research and improving health. Unlike research studies that are focused on a specific disease or population, *All of Us* will serve as a national research resource to inform thousands of studies, covering a wide variety of health conditions. Researchers will use data from the program to learn more about how individual differences in lifestyle, environment and biological make-up can influence health and disease. By taking part, people will be able to learn more about their own health and contribute to an effort that will advance the health of generations to come. NIH plans to launch the program later this year.

Data can do a lot: language processing

- For example Google translate
- New York Times article.

Is data all we need?



EXPERT OPINION

Contact Editor: **Brian Brannon**, bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

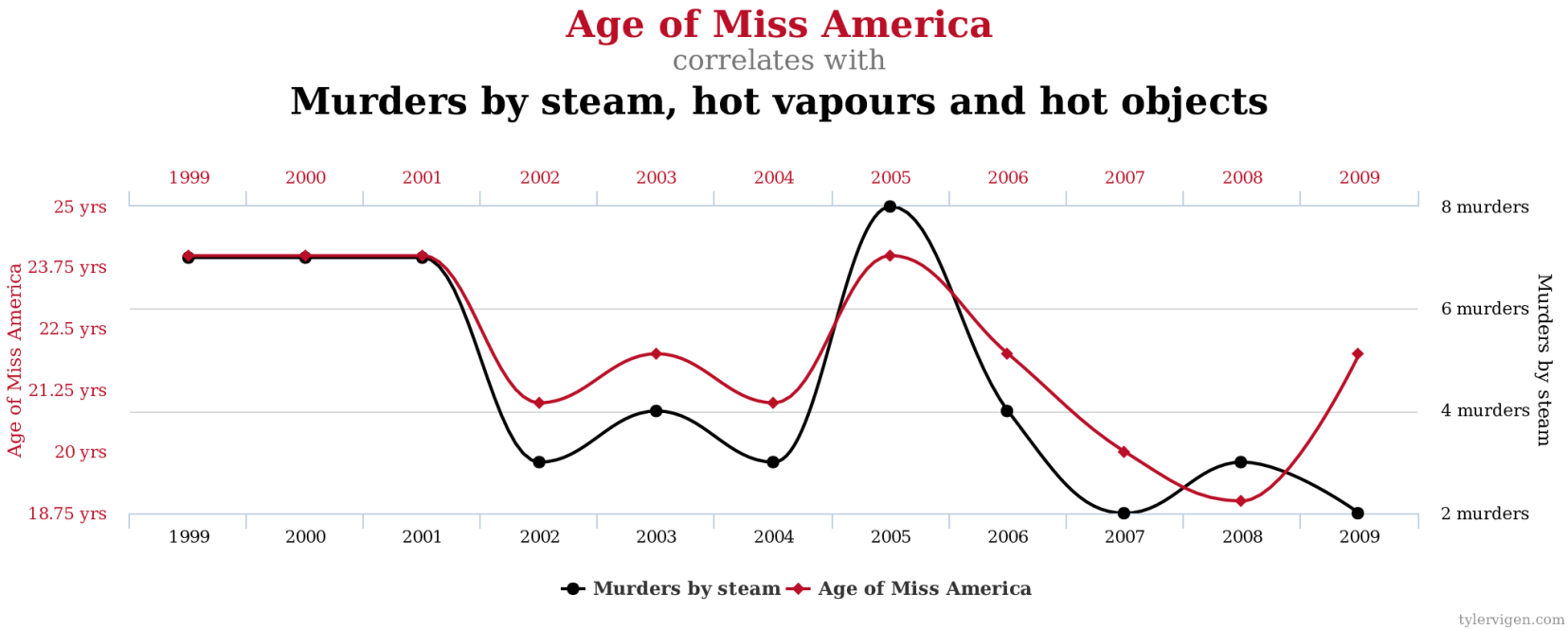
Is data all we need?

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

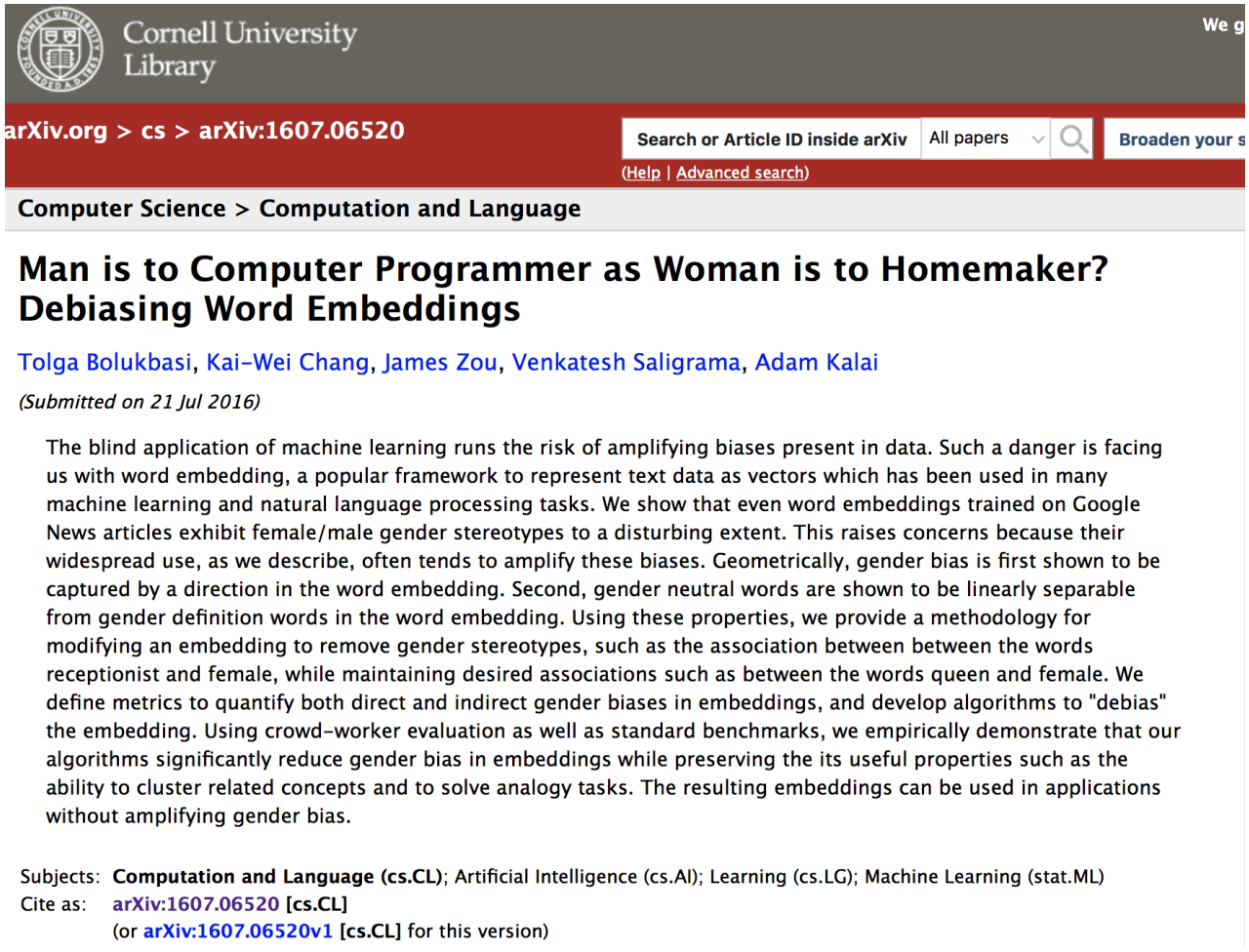
Illustration: Marian Bantjes

Is data all we need?



- It is easy to find "interesting" patterns where none exist!
- How should we judge whether a "pattern" is interesting?
- When should we worry about falsely labelling patterns "interesting"? (E.g. Google mistranslates a sentence vs. incorrect cancer diagnosis. . .)

Is data all we need?



Cornell University Library

We g

arXiv.org > cs > arXiv:1607.06520

Search or Article ID inside arXiv | All papers | Broaden your s

(Help | Advanced search)

Computer Science > Computation and Language

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

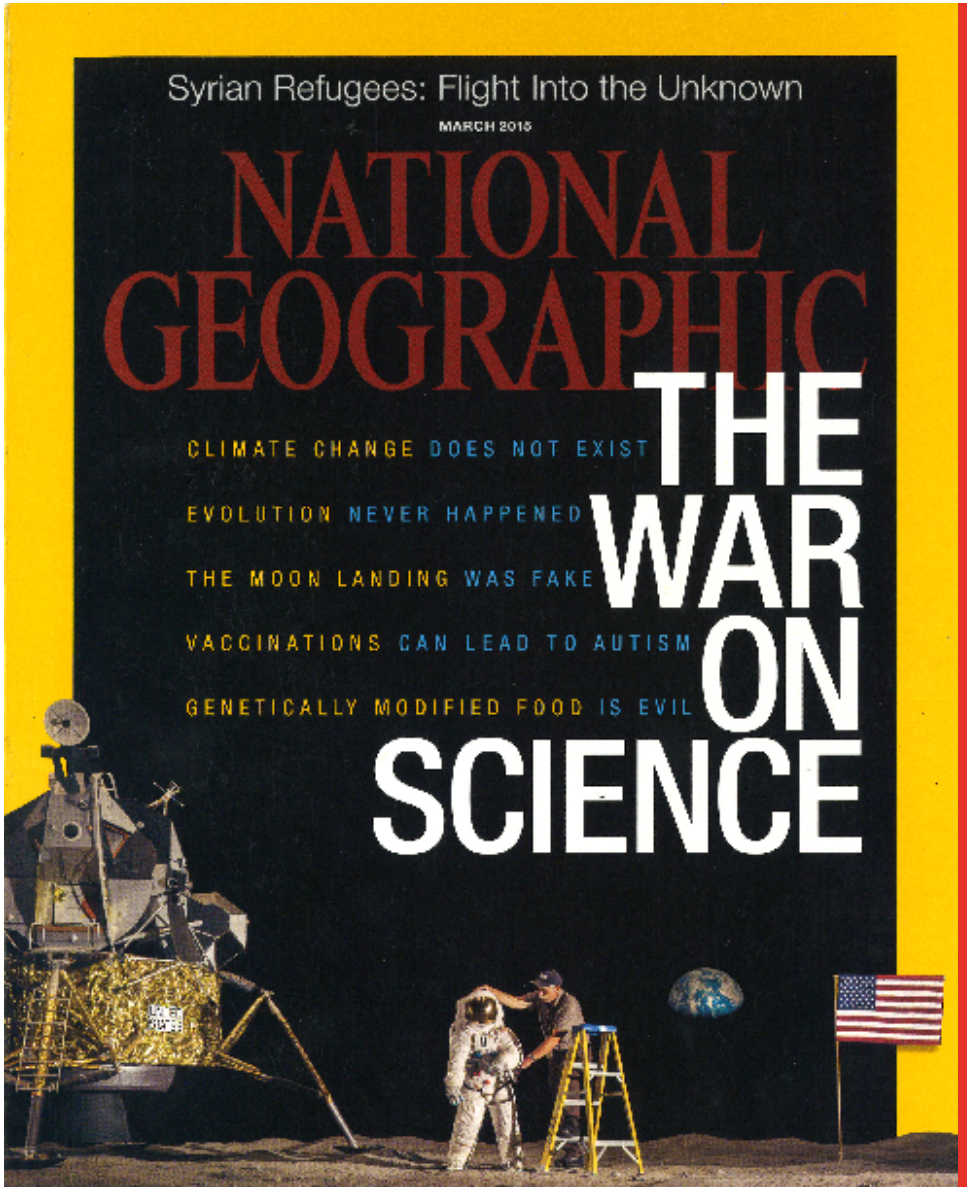
(Submitted on 21 Jul 2016)

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words receptionist and female, while maintaining desired associations such as between the words queen and female. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI); Learning (cs.LG); Machine Learning (stat.ML)
Cite as: arXiv:1607.06520 **[cs.CL]**
(or arXiv:1607.06520v1 **[cs.CL]** for this version)

Blindly used, machine learning algorithms can reinforce biases hidden in data.

Science is losing its authority?



Learn how to use data

- **Explore**: identify patterns
- **Predict**: make informed guesses
- **Infer**: quantify what you know

Data Science 101, resources

- A team of instructors, TA, and many more faculty that contribute to the course development
  - Sohom Bhattacharya (Thursday 4-6)
  - Pete Mohanty (Tuesday 2-4)
  - Chiara Sabatti (Monday 10:30-12:30)
  - Guenther Walter (Wednesday 2-4)
- Web-page: http://web.stanford.edu/class/stats101
- CANVAS site
- Piazza (linked from Canvas) for online discussion

Data Science 101: Evaluation

- Weekly homework (40% of grade)
- Participation (attendance, activities in labs, etc.) (10%)
- Midterm (30% if Midterm $>$ Final)
- In class final (30% if Final $>$ Midterm)

Data Science 101: Logistics

- Bring your laptops on Tuesdays and Thursdays
- Install `R` and `RStudio` for tomorrow
- Install extra packages this week (see course website).
- Lectures and `RNotebooks` for labs available on **stats101.stanford.edu**
- Two homeworks per module. A short one due Friday at 9:30am and a slightly longer one due the following Wednesday at 9:30am.