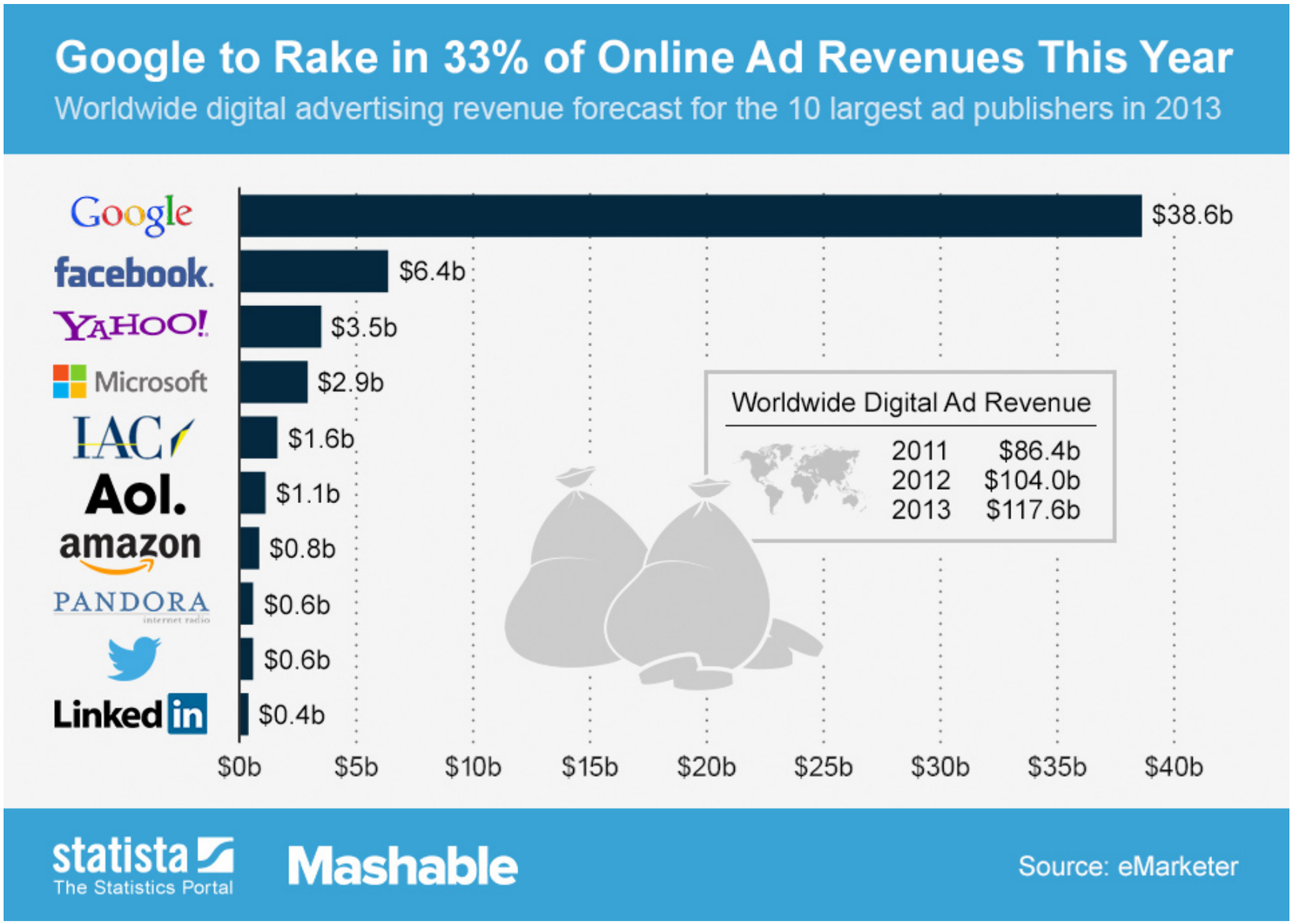
Lecture 2: Introduction

COSC 480 Data Science, Spring 2017

Michael Hay

Let's talk about $$$



Ad revenue is ≈ 89% of Google's revenue (as of 2016)

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Data and online content

Raghu Ramakrishnan, Yahoo! Research
NSF Workshop on Social Networks and Mobility in the Cloud 2012

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

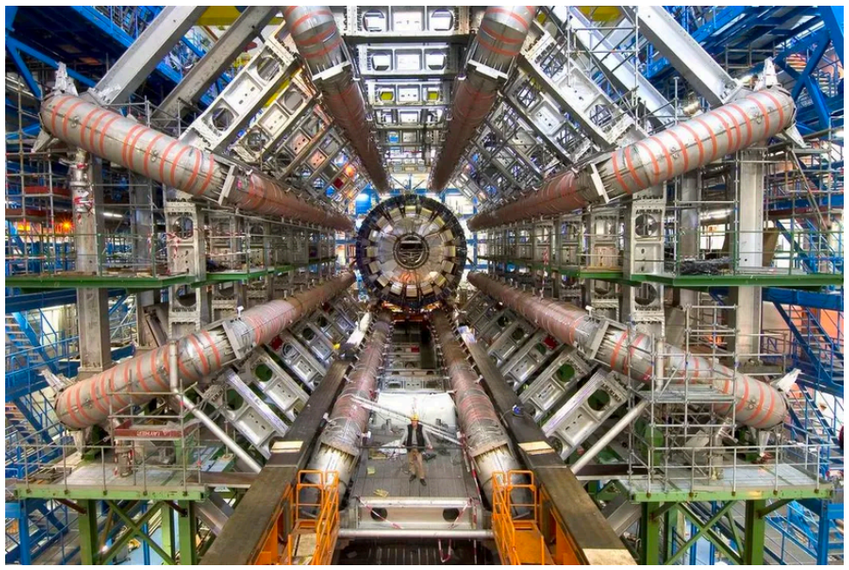
Data and commerce

Target analyzed purchases to identify women in early-stage pregnancy. Exposed teen daughter's pregnancy.

http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Data and science

- The world's largest particle collider at CERN—where the Higgs boson was confirmed—generates 30 PB of data per year



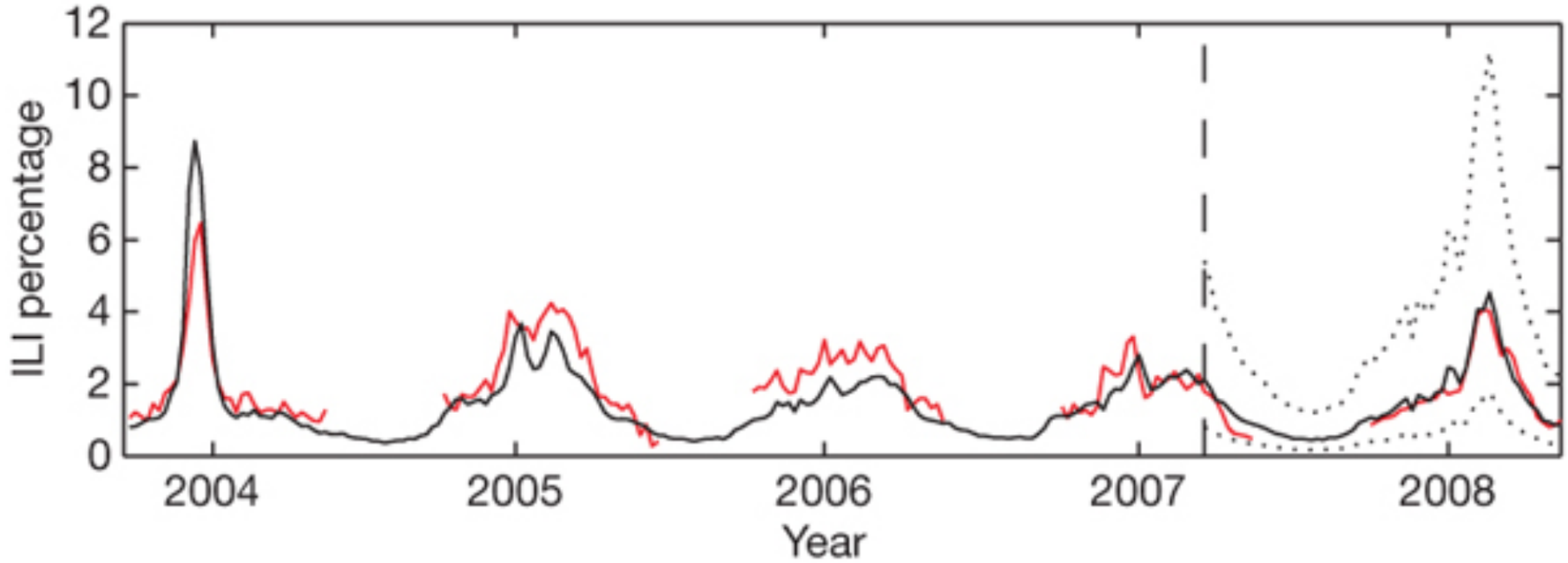
http://www.theverge.com/2016/4/25/11501078/cern-300-tb-lhc-data-open-access

http://home.web.cern.ch/about/computing

- CERN's data center has 11,000 servers with 100,000 cores... yet it still can't crunch all data!

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Data and health



Red: official numbers from Center for Disease Control and Prevention; weekly

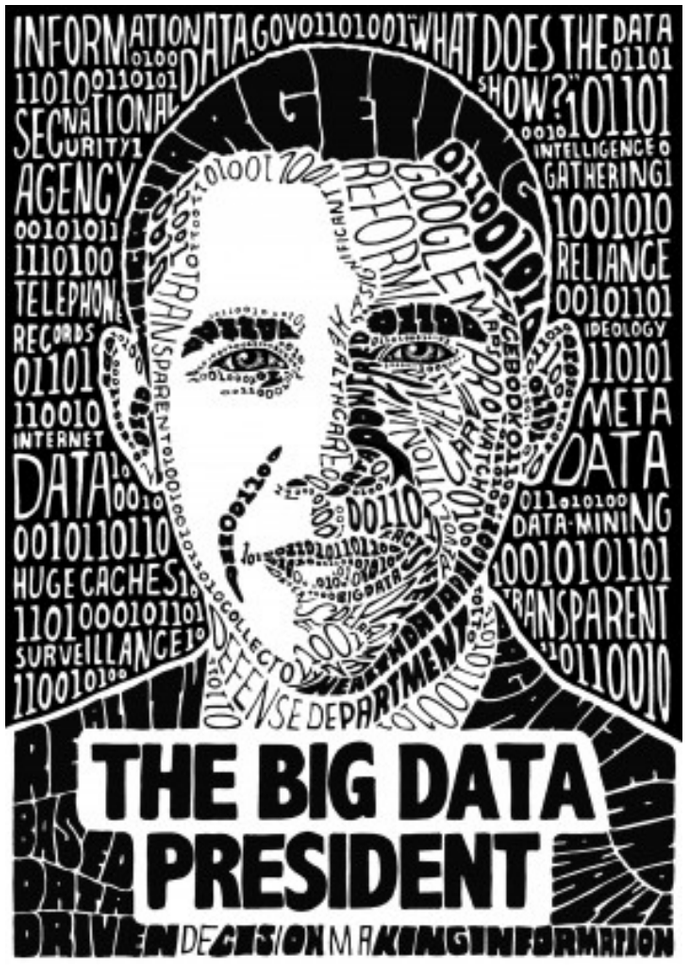
Black: based on Google search logs; daily (potentially instantaneously)

Detecting influenza epidemics using search engine query data

http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html

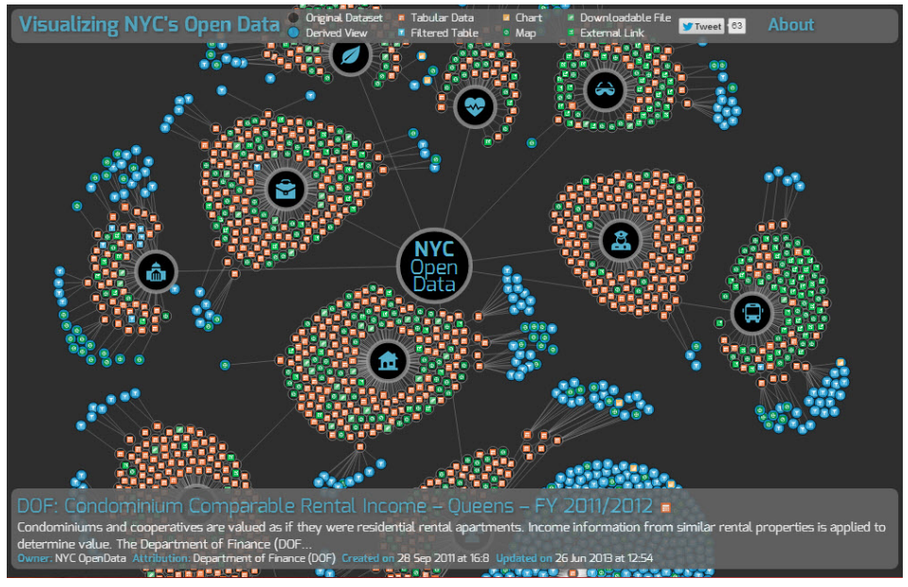
Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Data and government



http://www.washingtonpost.com/opinions/**obama-the-big-data-president**/2013/06/14/1d71fe2e-d391-11e2-b05f-3ea3f0e7bb5a_story.html

http://www.whitehouse.gov/blog/**Democratizing-Data**



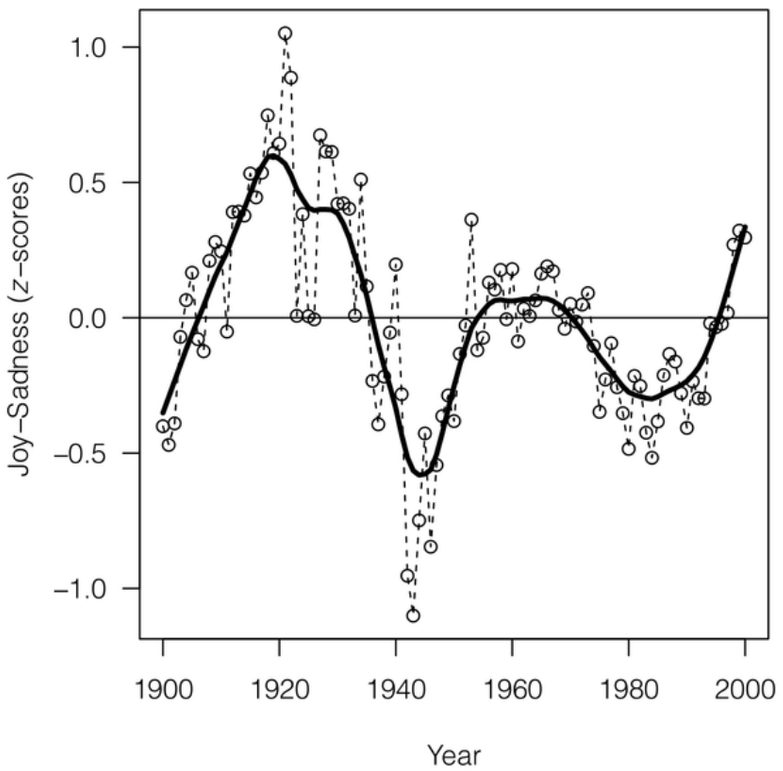
https://nycopendata.socrata.com/

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Data and culture

Quantitative Analysis of Culture Using Millions of Digitized Books

http://science.sciencemag.org/content/331/6014/176



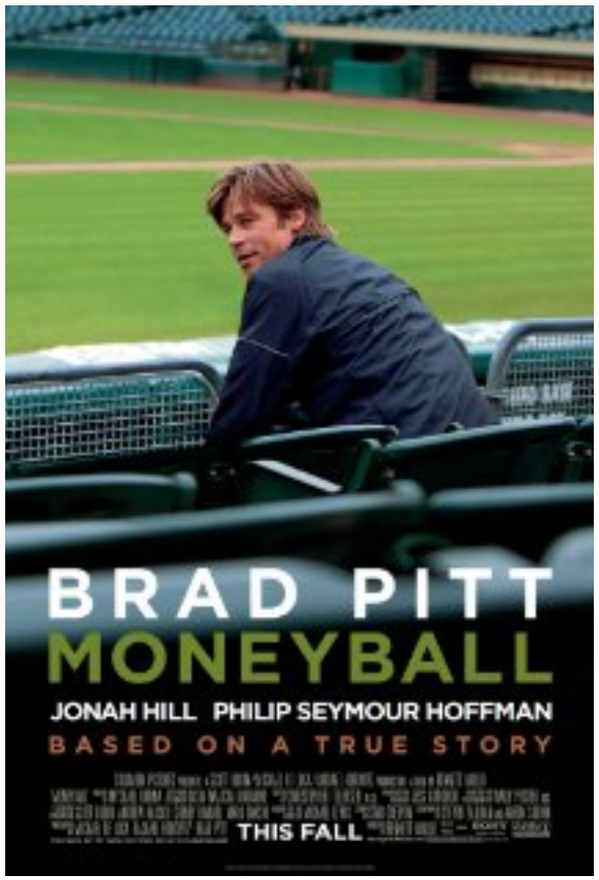
- Frequencies of emotion words in English-language books in Google's database

http://blogs.plos.org/everyone/2013/03/20/**what-are-you-in-the-mood-for-emotional-trends-in-20th-century-books**/

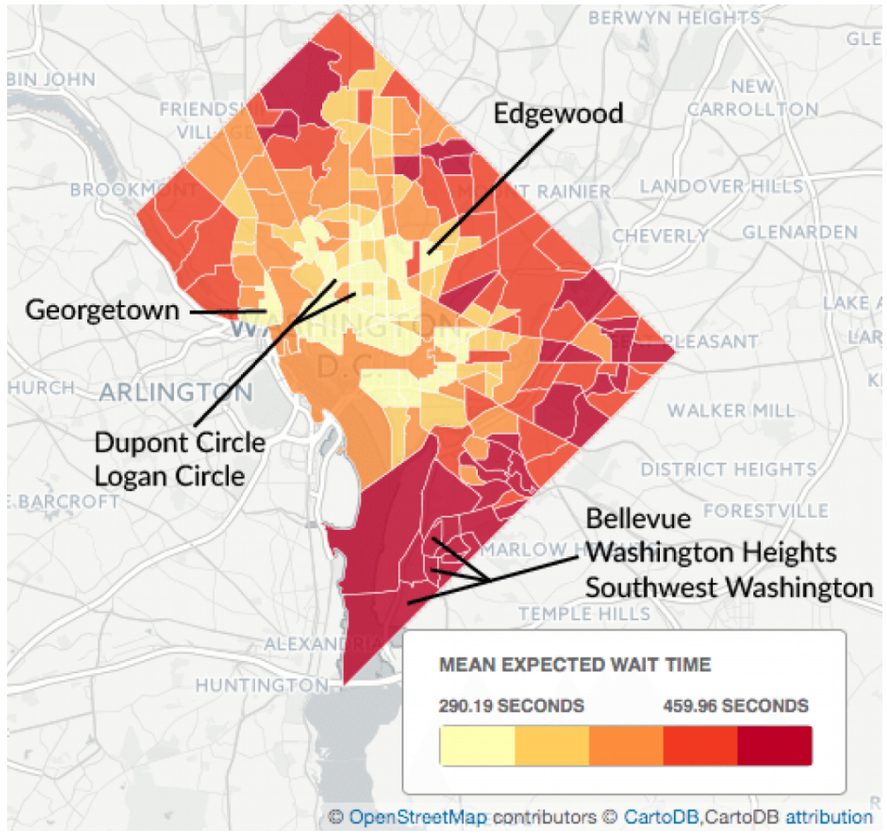
Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Data and ____ ⬅ your favorite subject

Sports



Journalism



https://www.washingtonpost.com/news/wonk/wp/2016/03/10/**uber-seems-to-offer-better-service-in-areas-with-more-white-people**-that-raises-some-tough-questions/

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Hal Varian, Chief Encomist at Google

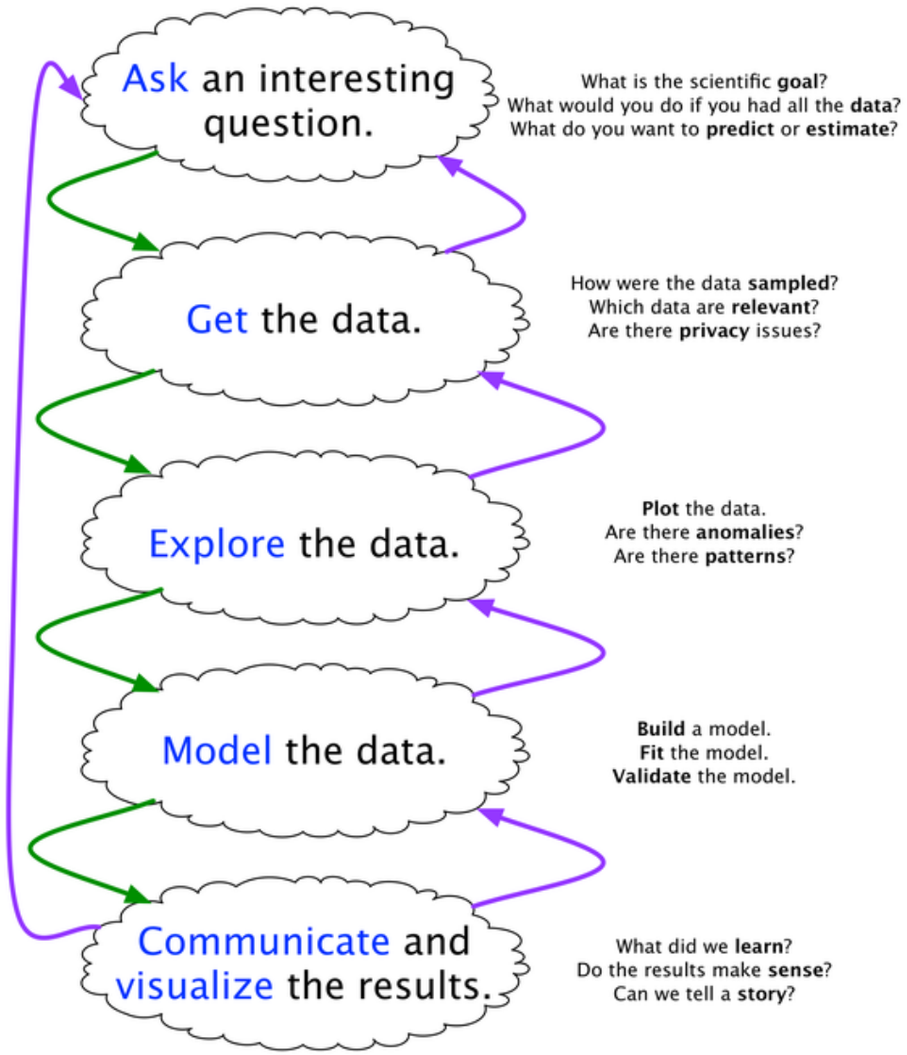
- "I keep saying ***the sexy job in the next ten years will be statisticians***.
  People think I'm joking,
  but who would've guessed
  that computer engineers would've been the sexy job of the 1990s? The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades…" (2009)

http://www.mckinsey.com/industries/high-tech/our-insights/hal-varian-on-how-the-web-challenges-managers

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

How to extract value from data?

- **Wrangle**
  - Get data you want into form you need for analysis
- **Analyze**
  - Explore, query, compute statistics, fit models
- **Communicate**
  - Use effective visualization, tell story, empower others



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Discussion of Google Flu Trends

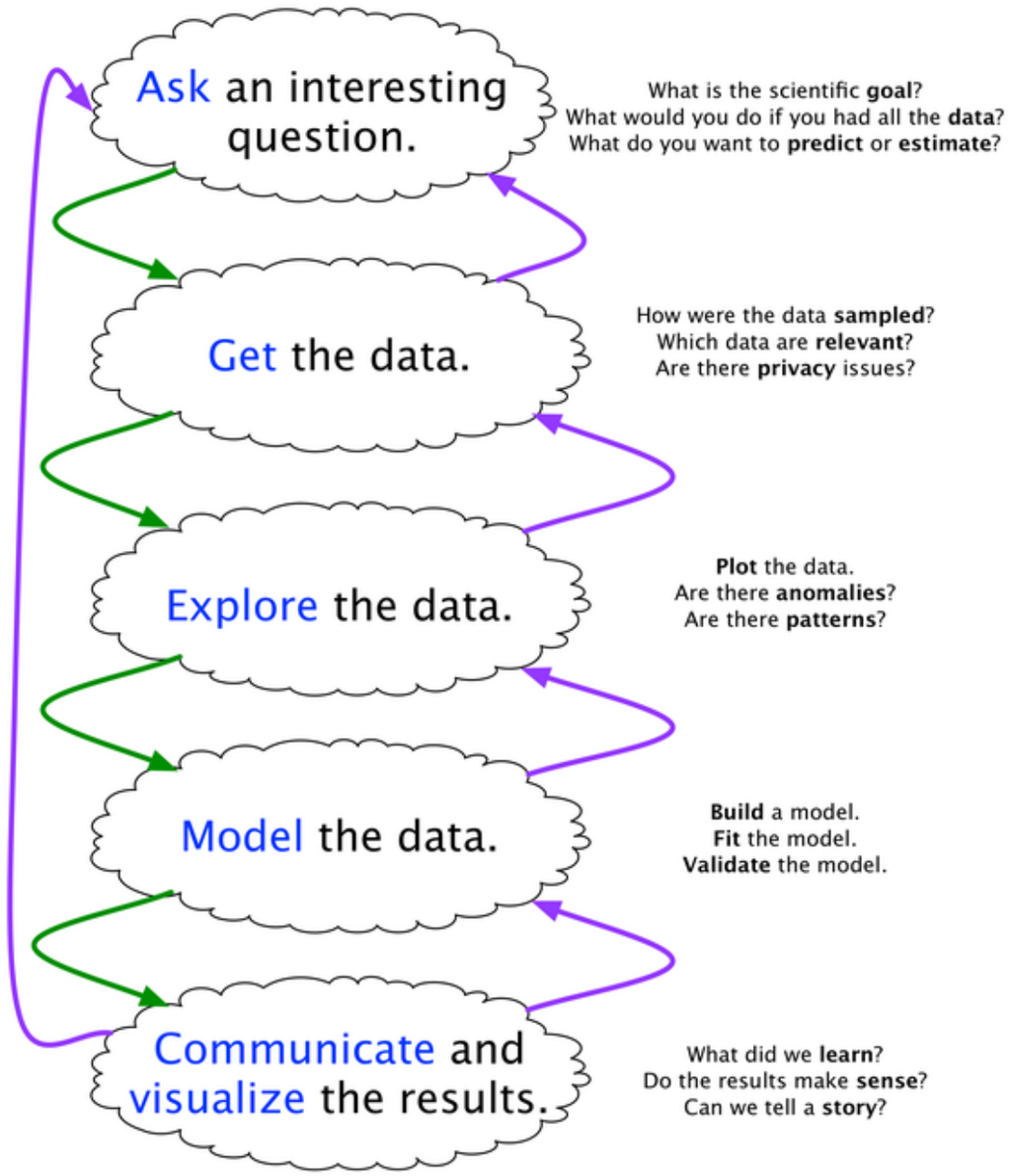
1. Align study with steps shown on right.

   A. Any steps missing?

   B. What is model?

2. Is GFT effective? How do you know?

3. Potential limitations?

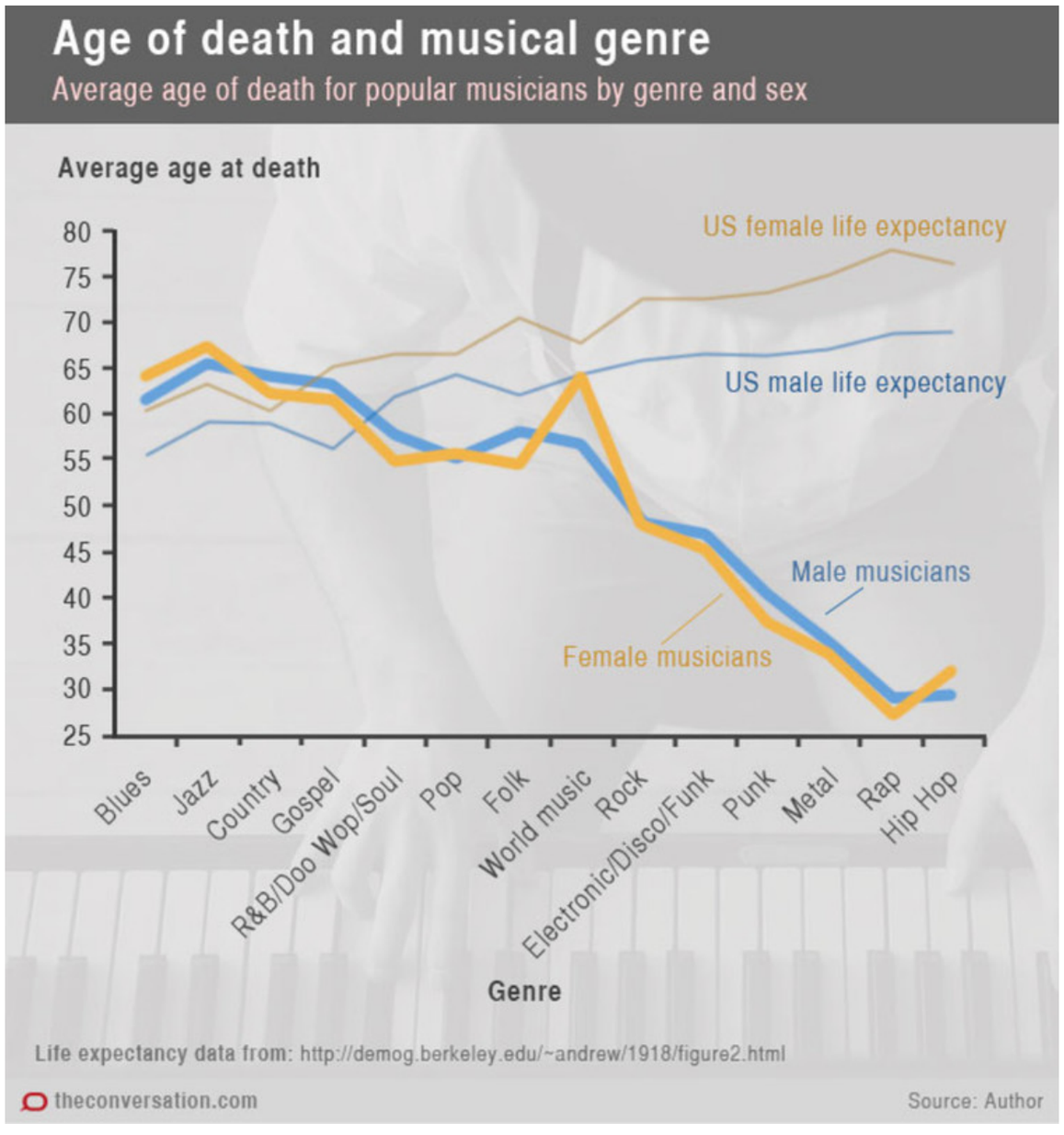


Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

http://www.realtyninja.com/blog/dark-side-of-real-estate/

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Easy to get it wrong…



Data censoring
(and other statistical traps)

http://callingbullshit.org/case_studies/case_study_musician_mortality.html

Easy to get it wrong…

BIG DATA

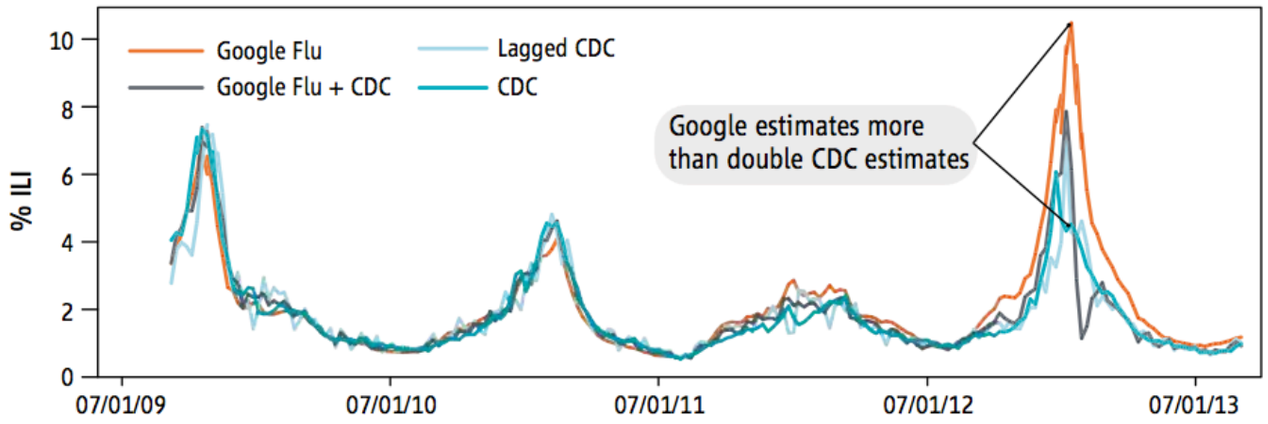
The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

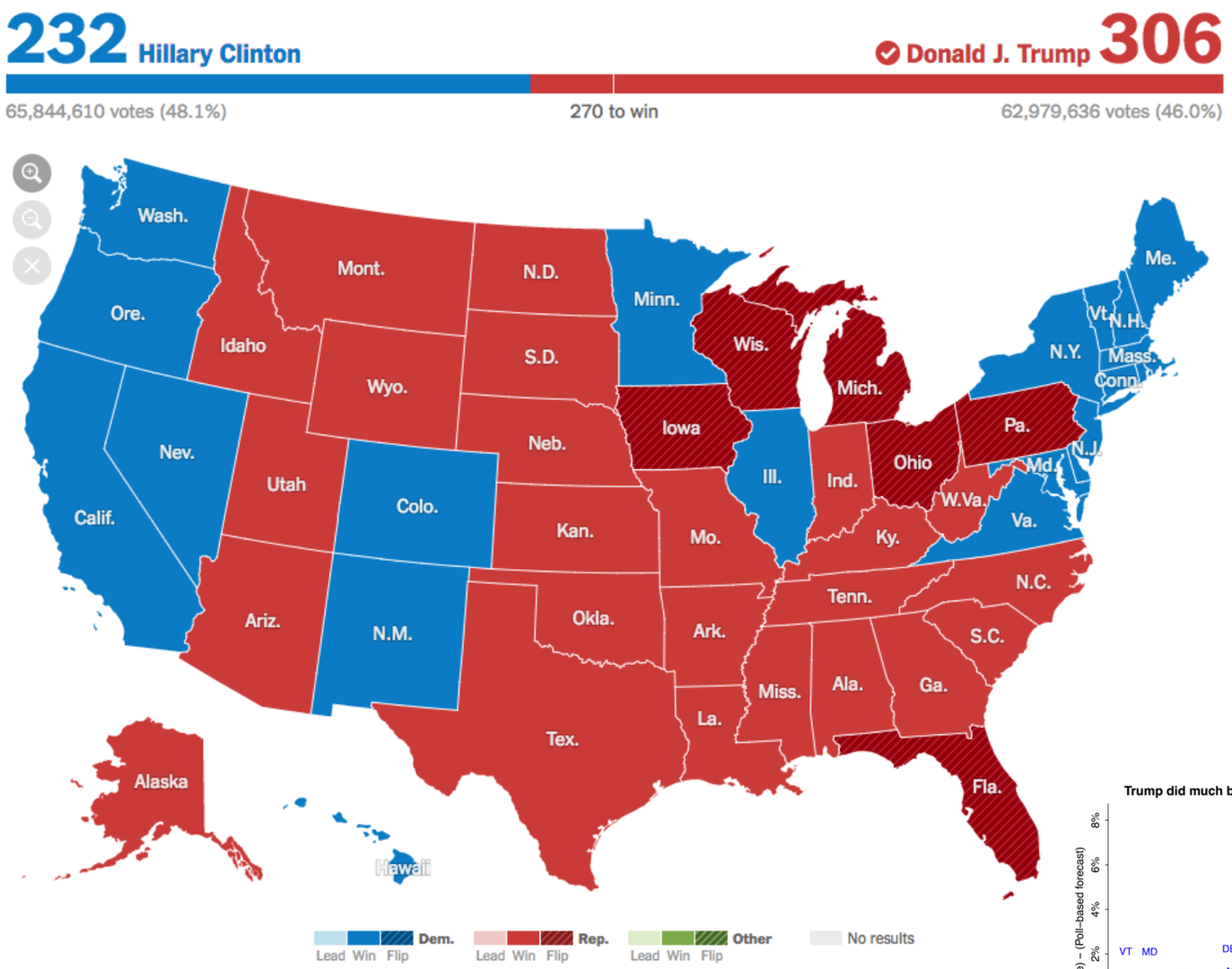
The problems we identify are not limited to GFT. Research on whether search or social media can



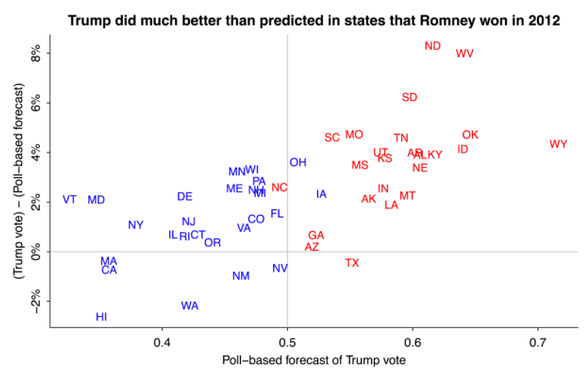
http://science.sciencemag.org/content/343/6176/1203

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Easy to get it wrong…



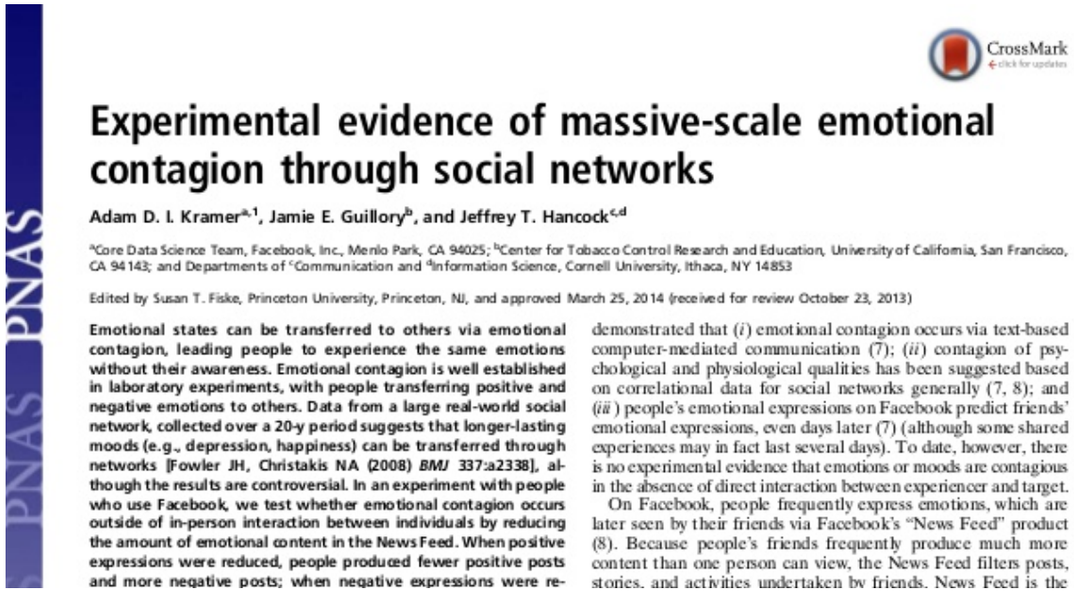
http://www.nytimes.com/elections/results/president



http://andrewgelman.com/2016/11/09/explanations-shocking-2-shift/

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Easy to abuse…. ethics



CrossMark

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b], and Jeffrey T. Hancock[c,d]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; [b]Center for Tobacco Control Research and Education, University of California, San Francisco, CA 94143; and Departments of [c]Communication and [d]Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. Emotional contagion is well established in laboratory experiments, with people transferring positive and negative emotions to others. Data from a large real-world social network, collected over a 20-y period suggests that longer-lasting moods (e.g., depression, happiness) can be transferred through networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a2338], although the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs outside of in-person interaction between individuals by reducing the amount of emotional content in the News Feed. When positive expressions were reduced, people produced fewer positive posts and more negative posts; when negative expressions were reduced, the opposite pattern occurred. These results indicate that emotions expressed by others on Facebook influence our own emotions, constituting experimental evidence for massive-scale contagion via social networks. This work also suggests that, in contrast to prevailing assumptions, in-person interaction and nonverbal cues are not strictly necessary for emotional contagion, and that the observation of others' positive experiences constitutes a positive experience for people.

demonstrated that (*i*) emotional contagion occurs via text-based computer-mediated communication (7); (*ii*) contagion of psychological and physiological qualities has been suggested based on correlational data for social networks generally (7, 8); and (*iii*) people's emotional expressions on Facebook predict friends' emotional expressions, even days later (7) (although some shared experiences may in fact last several days). To date, however, there is no experimental evidence that emotions or moods are contagious in the absence of direct interaction between experiencer and target.

On Facebook, people frequently express emotions, which are later seen by their friends via Facebook's "News Feed" product (8). Because people's friends frequently produce much more content than one person can view, the News Feed filters posts, stories, and activities undertaken by friends. News Feed is the

http://www.pnas.org/content/111/24/8788.full

Facebook emotion study breached ethical guidelines, researchers say

Lack of 'informed consent' means that Facebook experiment on nearly 700,000 news feeds broke rules on tests on human subjects, say scientists

https://www.theguardian.com/technology/2014/jun/30/facebook-emotion-study-breached-ethical-guidelines-researchers-say

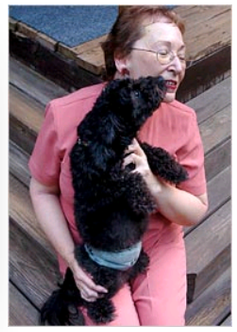
Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Easy to abuse… privacy

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-
THIS



NETFLIX

Why 'Anonymous' Data Sometimes Isn't

By Bruce Schneier 12.13.07

Last year, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using.

The Scientist » The Nutshell

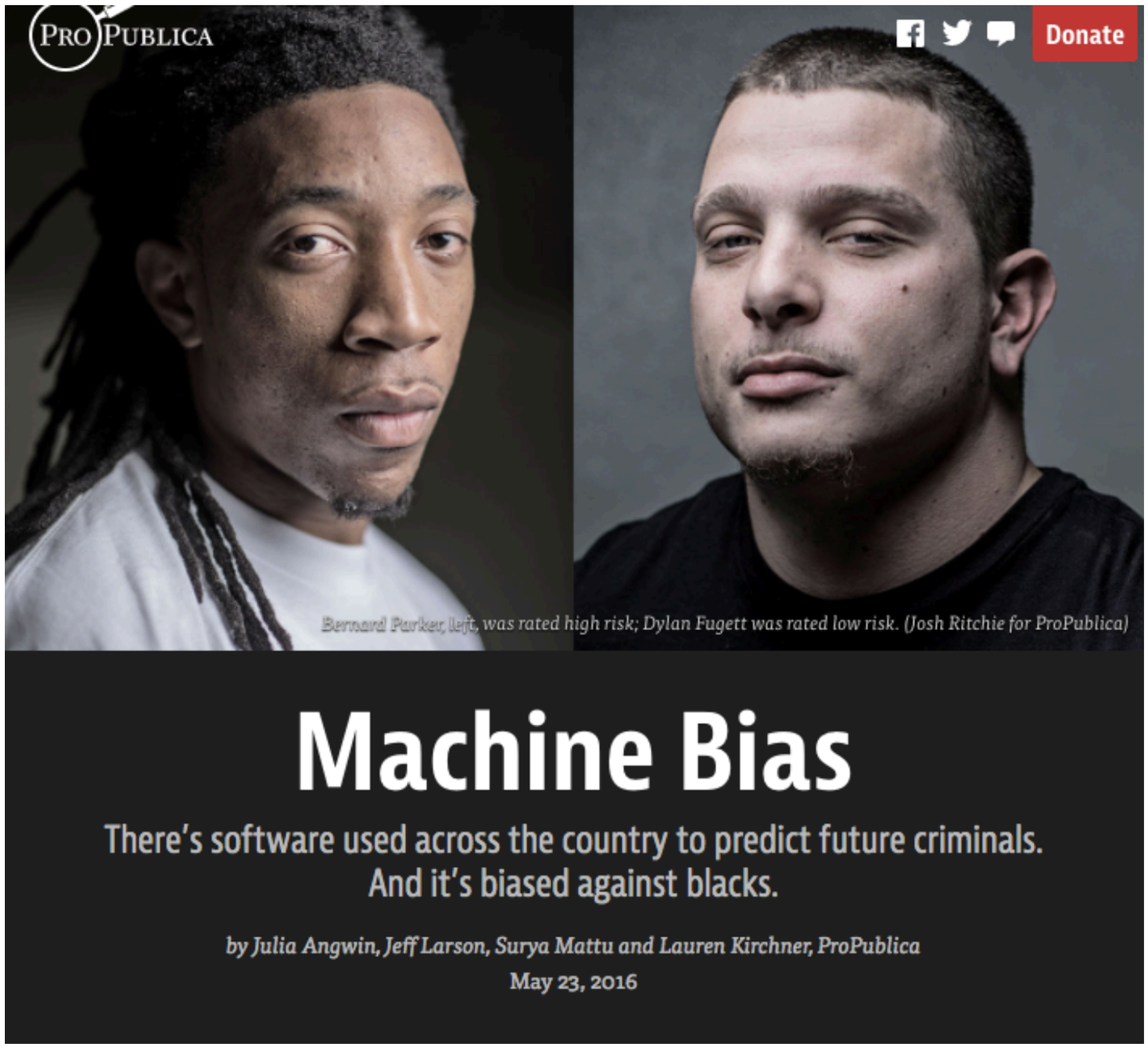
"Anonymous" Genomes Identified

The names and addresses of people participating in the Personal Genome Project can be easily tracked down despite such data being left off their online profiles.

By Dan Cossins | May 3, 2013

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Easy to misuse…. bias



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

39% of experts agree…

Thanks to many changes, including the building of "the Internet of Things," human and machine analysis of ***Big Data will cause more problems than it solves*** *by 2020. The existence of huge data sets for analysis will* ***engender false confidence in our predictive powers*** *and will lead many to make* ***significant and hurtful mistakes****. Moreover, analysis of Big Data will be* ***misused by powerful people and institutions with selfish agendas*** *who manipulate findings to make the case for what they want. And the advent of Big Data has a harmful impact because it* ***serves the majority (at times inaccurately) while diminishing the minority*** *and ignoring important outliers. Overall, the rise of Big Data is a big negative for society in nearly all respects.*

2012 Pew Research Center Report

http://pewinternet.org/Reports/2012/Future-of-Big-Data/Overview.aspx

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

But it’s here, now!

- Learn to…
  - Take advantage of it
  - Help yourself and other avoid being taken advantage of

http://rosemarynonnyknight.com/use-the-force-your-name-here/

Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

Course topics

- **Data processing**: acquisition, cleaning, processing, manipulation

- **Computational statistics**: stats, probability, inference

- **Machine learning**: foundational ideas, various methods

- **Visualization**: tools, graphics grammar, principles

- **Additional topics**: different kinds of data (text, networks, streaming); ethical: privacy, fairness

python tricks

python

- Two big (and maybe new?) ideas
  - Processing collection of data items *lazily* using iterators/generators
  - Functional programming ideas: higher-order functions, functions as “first class objects”, partial function evaluation

list comprehensions

```
In [1]: odd_numbers = [x for x in range(20) if x % 2 != 0]
```

```
In [2]: odd_numbers
Out[2]: [1, 3, 5, 7, 9, 11, 13, 15, 17, 19]
```

iterators

```
In [1]: xs = [1, 2, 3]

In [2]: it = iter(xs)

In [3]: next(it)
Out[3]: 1
```

get an iterator

take its values with `next`

```
In [4]: next(it)
Out[4]: 2

In [5]: next(it)
Out[5]: 3

In [6]: next(it)
---------------------------------------------------------------------------
StopIteration                             Traceback (most recent call last)
<ipython-input-121-5c05586d40e8> in <module>()
----> 1 next(iter)

StopIteration:
```

get a `StopIteration` exception when no values left

Credit: slide adapted from talk by Joel Grus, https://github.com/joelgrus/stupid-itertools-tricks-pydata

iterators

- serving serving up values one-at-a-time with next means you can generate them on-demand
- (laziness)
- allows us to create lazy infinite sequences

Credit: slide adapted from talk by Joel Grus, https://github.com/joelgrus/stupid-itertools-tricks-pydata

generators

```
def lazy_integers(n=0):
    while True:
        yield n
        n += 1
```

function with `yield` creates a generator

```
xs = lazy_integers()
```

infinite sequence!

```
[next(xs) for _ in range(10)]
# [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]

# maintains state
[next(xs) for _ in range(10)]
# [10, 11, 12, 13, 14, 15, 16, 17, 18, 19]
```

Credit: slide adapted from talk by Joel Grus, https://github.com/joelgrus/stupid-itertools-tricks-pydata

generator comprehensions

```
# computes nothing until next or for
squares = (x**2 for x in lazy_integers())
doubles = (2*x for x in lazy_integers())
```

```
next(squares) # 0
next(squares) # 1
next(squares) # 4
next(squares) # 9

# don't do this!!!:
bad_squares = [x**2 for x in lazy_integers()]
```

Credit: slide adapted from talk by Joel Grus, https://github.com/joelgrus/stupid-itertools-tricks-pydata

higher order functions

```
In [1]: xs = [1, 2, 3, 4]

In [2]: def double(x): return x*x

In [3]: map(double, xs)
Out[3]: [1, 4, 9, 16]
```

from functools import partial

partial function application ("currying")

```
from operator import add
```

```
def add1(x): return add(1, x)
```

could be written as

```
add1 = partial(add, 1)
```

Credit: slide adapted from talk by Joel Grus, https://github.com/joelgrus/stupid-itertools-tricks-pydata