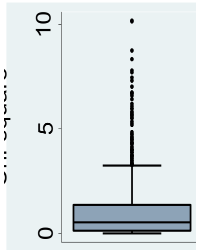
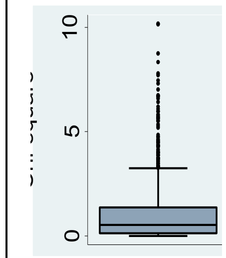
Biostat 202: Opportunities and Challenges of "Big Data"



That was just for fun, *but*



What 7 Creepy Patents Reveal About Facebook

The New York Times

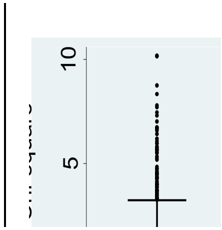
By Sahil Chi

Illustrations b

Listening to your environment

This patent application explores using your phone microphone to identify the television shows you watched and whether ads were muted. It also proposes using the electrical interference pattern created by your television power cable to guess which show is playing.

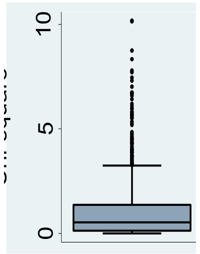
That was just for fun, *but*



IN BRIEF

What can we learn from the Facebook–Cambridge Analytica scandal?

04 | SIGNIFICANCE | June 2018

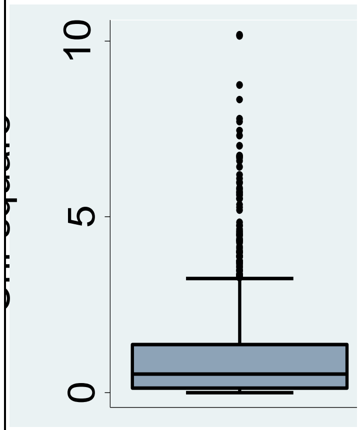


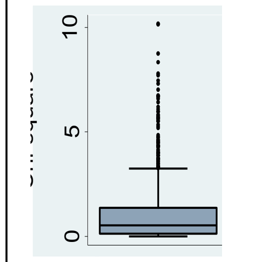
Thought #1

- Big data and data mining can be used for good or evil.

Biostat 202: Opportunities and Challenges of “Big Data”

Aaron Wolfe Scheffler,
Division of Biostatistics,
Department of Epidemiology and Biostatistics

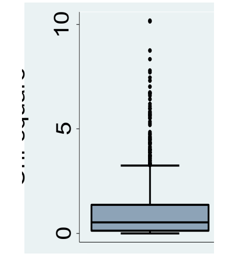




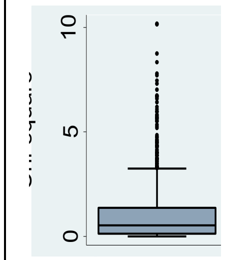
Course outline

- Introduction
- Getting data (large databases, electronic health records)
- Managing data
- Algorithms (prediction, clustering)
- Variable importance
- Causal inference and big data
- Describing data with tables and figures
- Case studies

Is this course for you?

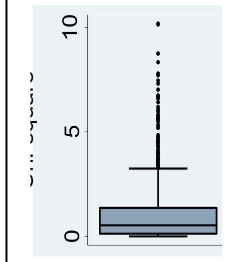


- High-level survey course
- Fast, comprehensive, but not deep
- No programming, graphical computer interface
- Not statistics!
- Goal: teach basic fluency in language and methods of machine learning/data mining in big data
- More details on the syllabus



Outline for today

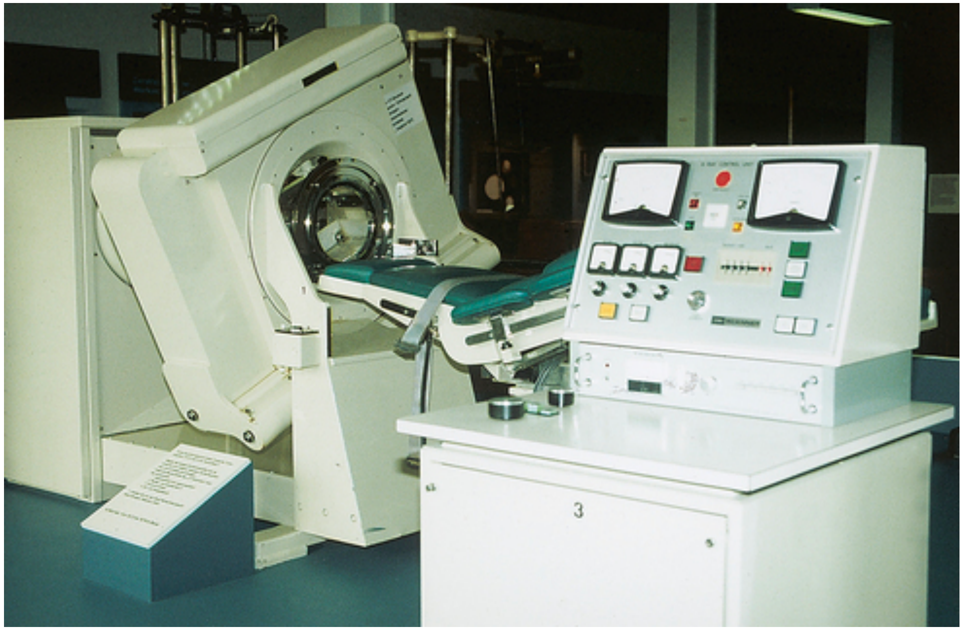
- What is big data? And since computers are so fast these days, is it really any different than what we have been doing all along?
- Data mining.
- Causality versus prediction.
- Overfitting and spurious associations.



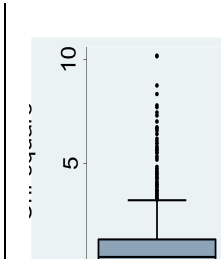
Is Big Data new?

My father was also statistician (b. 1949).

Worked with scans from early GE CT machine (~ 1984)



Images were Gb and had to be stored on main-frame in the hospital!

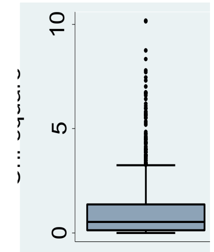


Is Big Data new?

So even in my father's time, he had "big" data.

Scan needed to be put onto 12 inch floppy disk and physically mailed to doctor's office, compressed first.

Today it goes to the cloud, straight to iPad!



Big Data in the 1990s

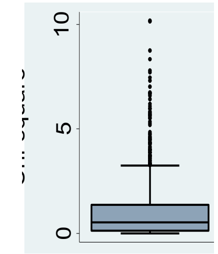
BIOLOGY OF REPRODUCTION 46, 1158–1164 (1992)

Different Patterns of Myometrial Activity and 24-H Rhythms in Myometrial Contractility in the Gravid Baboon during the Second Half of Pregnancy[1]

MARK A. MORGAN,[3] SUSAN L. SILAVIN,[3] RICHARD A. WENTWORTH,[4] JORGE P. FIGUEROA,[4] BARBERA O. M. HONNEBIER,[4] JOHN I. FISHBURNE, JR.,[3] and PETER W. NATHANIELSZ[2,4]

Department of Obstetrics and Gynecology,[3] University of Oklahoma Health Sciences Center Oklahoma City, Oklahoma 73190
Laboratory for Pregnancy and Newborn Research,[4] New York State College of Veterinary Medicine Department of Physiology, Cornell University, Ithaca, New York 14853

<u>Methods</u>: Continuous EMG …were sampled at 32 Hz … Fast Fourier Transform and power spectrum analysis were performed …to quantify myometrial activity as contractures or contractions.

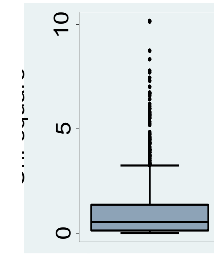


Big Data in the 1990s

- 13 baboons studied.
- At least two electrical leads per animal.
- Up to 67 days of observation.
- Record electrical activity of uterine muscles 32 times per second

About 4.8 billion data points (~4.5 GB)

Thought #2

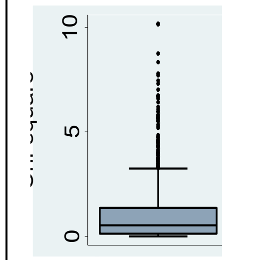


- Big data isn't the same as big information.

In this example, 4.8 billion data points from only 13 animals.

Data isn't information. ... Information, unlike data, is useful. While there’s a gulf between data and information, there’s a wide ocean between information and knowledge. What turns the gears in our brains isn't information, but ideas, inventions, and inspiration. Knowledge—not information—implies understanding. And beyond knowledge lies what we should be seeking: wisdom.

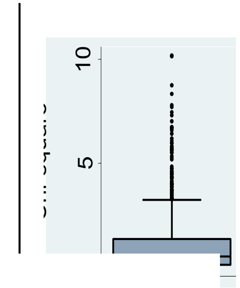
— Clifford Stoll, *High-Tech Heretic: Reflections of a Computer Contrarian* (2000), pp. 185-186.



What is “Big Data” today?

1. Electronic health records (EHR).
2. Data collected electronically at a personal level (e.g., Fitbit monitors).
3. Web based surveillance (e.g., Google Flu Trends).
4. Imaging data (e.g., CAT Scan, MRIs).
5. Digital Exhaust (noun). Unstructured information or data that is a by-product of the online activities of Internet users.
6. Omics data (e.g., genomic, proteomic, or exposome)

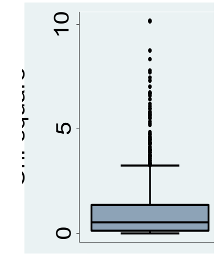
What is “Big Data” today?



Is “big data” really any different than what we have been doing for decades other than a steady increase in volume?

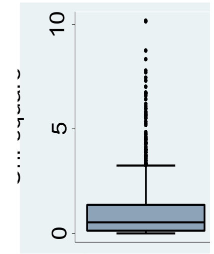
Yes

- Handling poor quality and heterogeneous data.
- Harder to practice reproducible research.
- Dealing with data “velocity.”
- Possibility of real time decision making.
- Algorithmic association detection. “Data mining” and “machine learning.”



Why data mining?

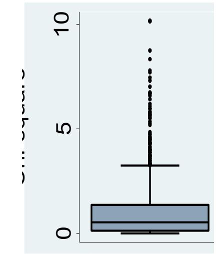
- Example: OA biomarkers project (one of the project datasets).
- In a cohort of individuals at risk for knee osteoarthritis, can we predict who will go on to develop painful knee arthritis?
- Full study has 1000s of predictors
  - measurements of the knee based on Xray (width of the space between bones at many locations)
  - MRI (degree of degeneration in various locations of the cartilage)
  - blood sample (numerous bone turnover markers)…



Why data mining?

Building a model to predict painful knee osteoarthritis with, say, 30 predictors. Since each predictor can be in or out of the model, there are 2^{30} possible models. Suppose you could review 1 per second:

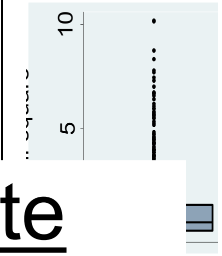
$$\begin{aligned} 2^{30} &= 1.1 \times 10^{9} \text{ seconds} \\ &= 18 \text{ million minutes} \\ &= 300{,}000 \text{ hours} \\ &= 12{,}500 \text{ days} \\ &= 34 \text{ years to check them all} \end{aligned}$$



Why data mining?

<u>Example</u>: decision tree algorithm for OA biomarkers subset

Not just a fad



<u>Patient Centered Outcomes Research Institute</u>

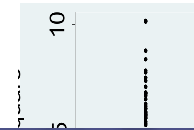
Has funded 33 health data networks (PCORnet) to the tune of $140m that link together electronic health record systems in order (in part) to conduct:

“observational studies of multiple research questions, including prevention and treatment, at low marginal cost”

by leveraging

“data from real-world clinical experiences of patients within their system”

Not just a fad



NIH Data Science at NIH

Data Science Community | BD2K | Commons | News &

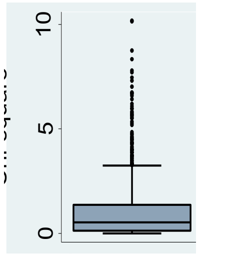
About | BD2K Organization | Funded Programs | Announcements | News | Events | Contact Us

About BD2K

Data Science Home / About BD2K

The ability to harvest the wealth of information contained in biomedical Big Data will advance our understanding of human health and disease; however, lack of appropriate tools, poor data accessibility, and insufficient training, are major impediments to rapid translational impact. To meet this challenge, the National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) initiative in 2012.

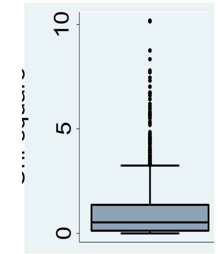
What is Biomedical Big Data?



Not just a fad

The advantages of large datasets are undeniable:

1. Ability to fit flexible models.
2. Ability to investigate heterogeneity and subgroups. "Precision medicine."
3. Easy to validate predictive rules.
4. High precision.

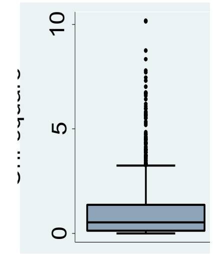


Not just a fad

The advantages of large datasets are undeniable:

1. Ability to fit flexible models.
2. Ability to investigate heterogeneity and subgroups. “Precision medicine.”
3. Easy to validate predictive rules.
4. High precision.

But there are a number of caveats.



Levels of Evidence

From the Centre for Evidence-Based Medicine, Oxford

For the most up-to-date levels of evidence, see www.cebm.net/?o=1025

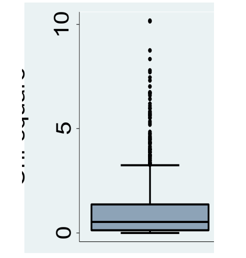
Therapy/Prevention/Etiology/Harm:

1a:	Systematic reviews (with homogeneity) of randomized controlled trials
1b:	Individual randomized controlled trials (with narrow confidence interval)
1c:	All or none randomized controlled trials
2a:	Systematic reviews (with homogeneity) of cohort studies
2b:	Individual cohort study or low quality randomized controlled trials (e.g. <80% follow-up)
2c:	"Outcomes" Research; ecological studies
3a:	Systematic review (with homogeneity) of case-control studies
3b:	Individual case-control study
4:	Case-series (and poor quality cohort and case-control studies)
5:	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"

Osteoarthritis Initiative

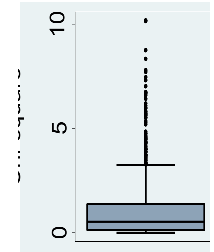
Electronic health record based study

Big Data – more is better, right?



Why does data quality vary?

In 1936, the Literary Digest conducted a presidential poll (Landon v. Roosevelt). They mailed 10 million postcards and received 2.4 million back.



The Literary Digest Poll

- Literary Digest collected 2.4 million responses and projected a solid victory for Alf Landon.
- George Gallup surveyed 50,000 individuals and predicted a solid victory for Roosevelt

Table 1. The election of 1936.

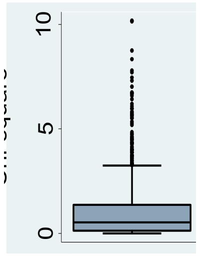
	Roosevelt's percentage
The *Digest* prediction of the election result	43
Gallup's prediction of the election result	56
The election result	62

Note: Percentages are of the major-party vote. In the election, about 2% of the ballots went to minor-party candidates.
Source: George Gallup, *The Sophisticated Poll-Watcher's Guide*, 1972.

The magnitude of the *Digest*'s error is staggering. It is the largest ever made by a major poll. Where did it come from? The number of replies was more than big enough. In fact, George Gallup was just setting up his survey organization.[4] Using his methods, he was able to predict what the *Digest* predictions were going to be, well in advance of their publication, with an error of only one percentage point. Using another sample of about 50,000 people, he correctly forecast the Roosevelt victory, although his prediction of Roosevelt's share of the vote was off by quite a bit. Gallup forecast 56% for Roosevelt; the actual percentage was

To find out where the *Digest* went wrong, you have to ask how they picked their sample. A sampling procedure should be fair, selecting people for inclusion in the sample in an impartial way, so as to get a representative cross section of the public. A systematic tendency on the part of the sampling procedure to exclude one kind of person or another from the sample is called *selection bias*.

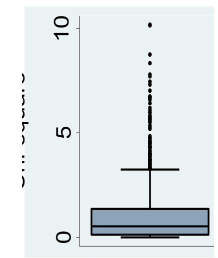
> When a selection procedure is biased, taking a large sample does not help. This just repeats the basic mistake on a larger scale.



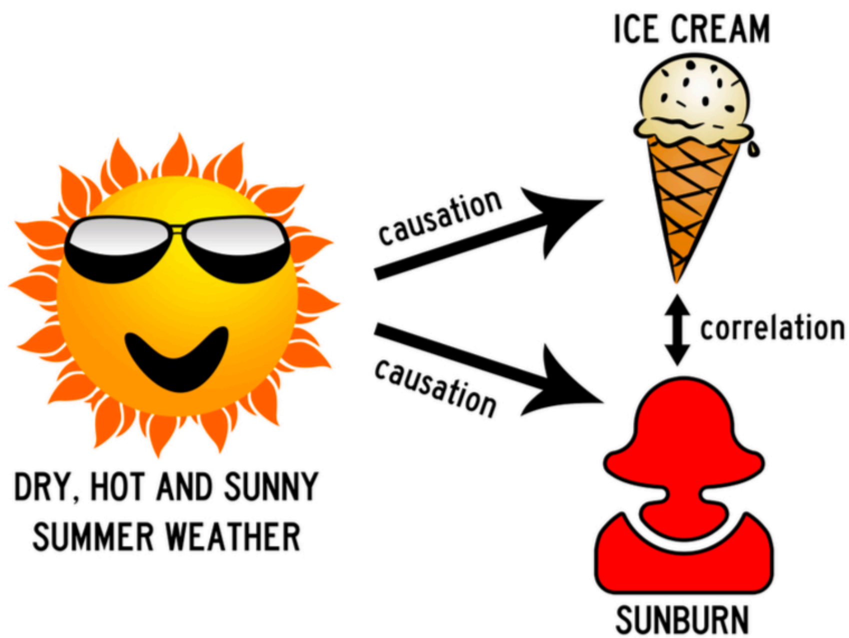
Thought #3

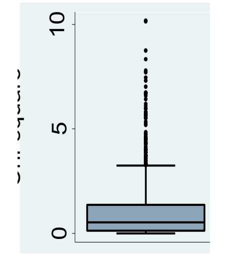
- Having big data doesn't help if it was selected in a biased way.

Prediction versus causality



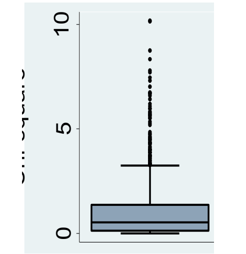
- Correlation (prediction): measure of the strength and direction of the relationship between two variables
- Causation: Will intervening upon X change occurrence of Y?





Prediction versus causality

- Everyone in the science community understands the difference between causation and correlation.
- Data-mining and machine learning algorithms are only based on correlations.
- In my work on research projects, pure prediction problems make up only a small portion.
- Usually interested in causation, i.e., understand the cause of illness or develop interventions that cause improvement.



Some bona fide prediction problems

- Early detection of cancer.
- Identifying high risk patients for closer monitoring. For example, patients likely to be readmitted to the hospital within 30 days.
- Prediction of disease epidemics.

<u>Key aspects</u>: Not concerned with causal interpretation of the effect of variables included in the model or interpreting presence of absence of the variable in the model causally. Focus on accurate prediction, not p-values and testing.

Big Data Prediction: Google Flu Trends

nature Vol 457|19 February 2009|doi:10.1038/nature07634

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

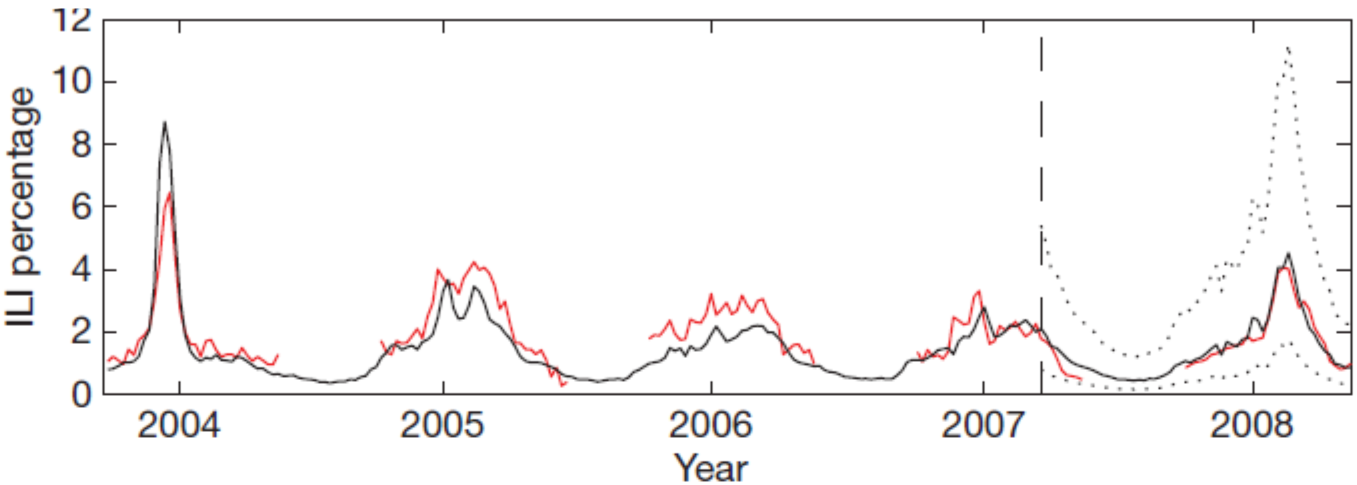
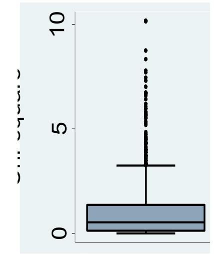


Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.



Google Flu Trends

- Used “hundreds of billions of searches” to build the model.
- Reduced to weekly counts of 50 million common searches for each state for 6 years.
- (50 million)*(50 states)*(52 weeks)*(6 years) or about 8x10^11 possible predictors.
- Built a model based on 2003 to 2007 to predict the CDC reported percentages of ILI (influenza like visits) before they were released in 2008.

Big Data Prediction: Google Flu Trends

nature Vol 457|19 February 2009|**doi:10.1038/nature07634**

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

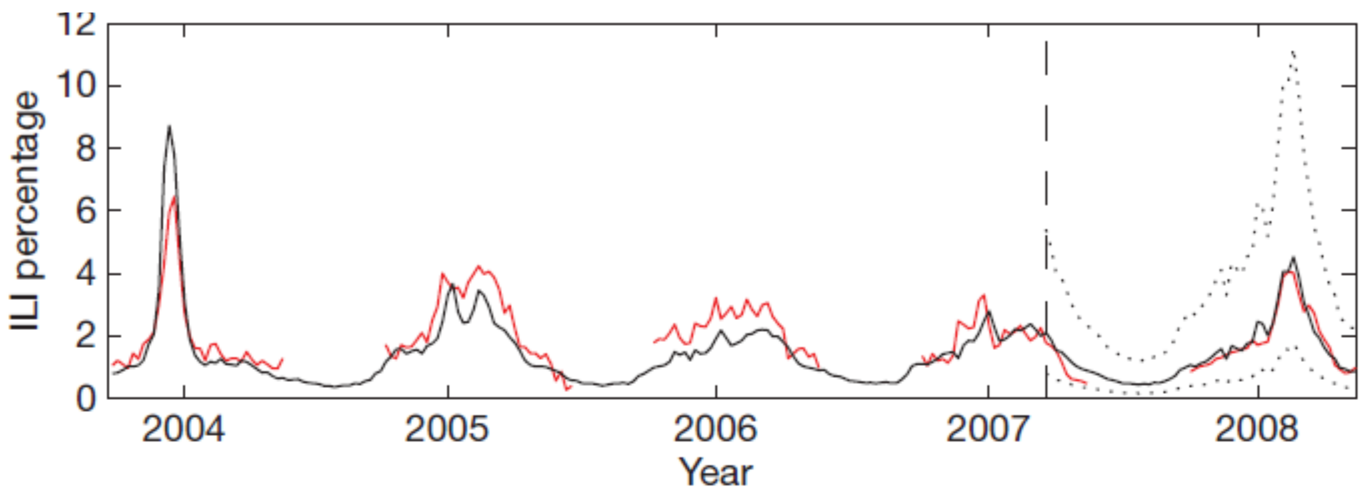
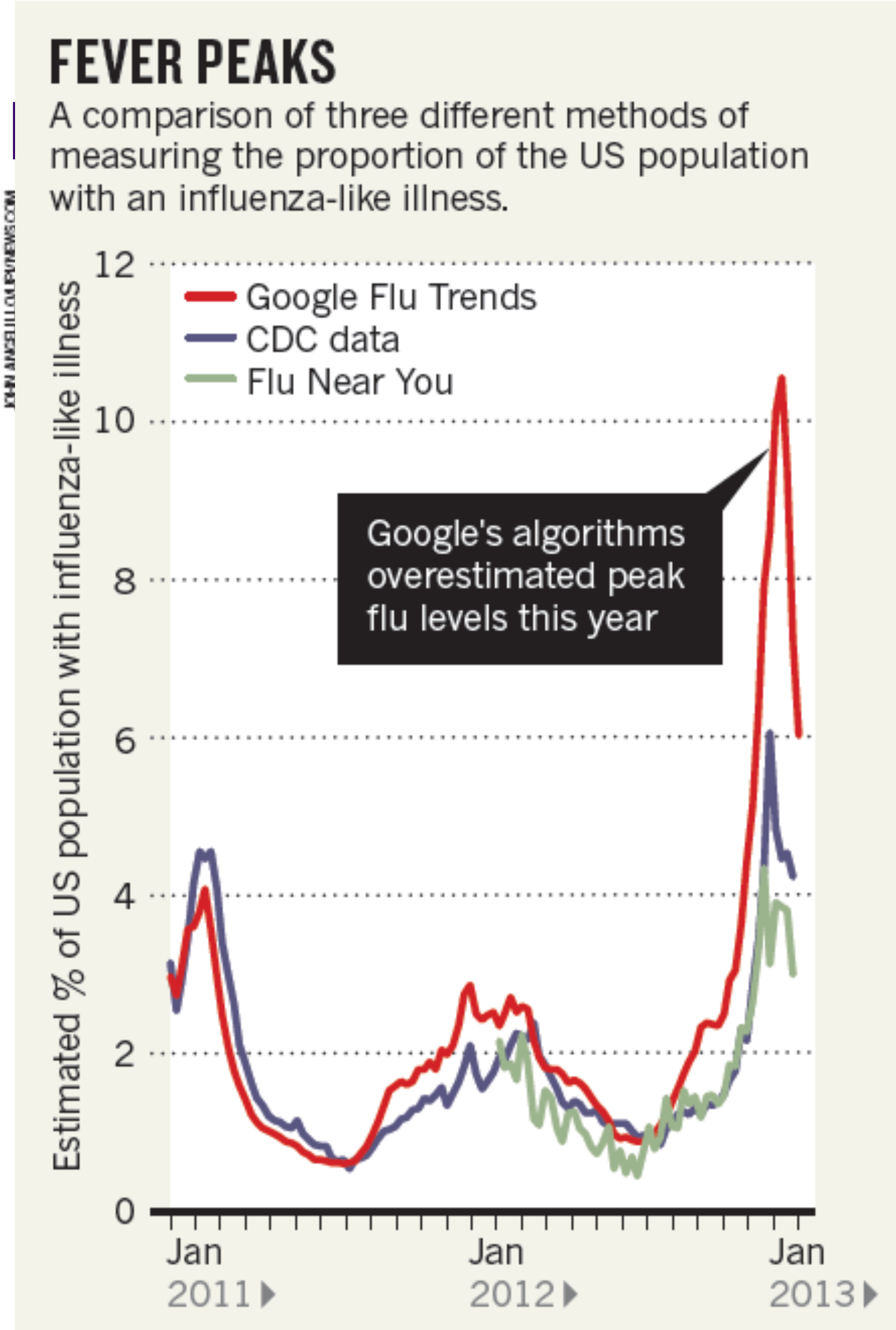
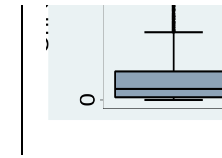


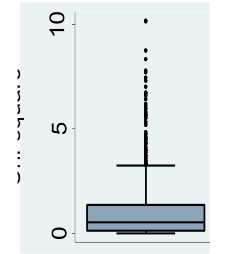
Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.

Big Data Trends



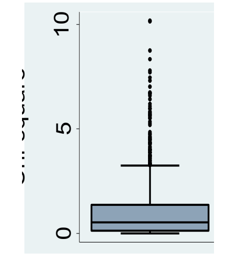


Why did it break?



- The underlying conditions under which the model was developed changed.
- The model wasn’t built with any biological insight. It was built (in a sophisticated way) on correlations.
- It was built on weekly counts of the 50 million most common search queries but only 6 years of flu data.

See Thought #2

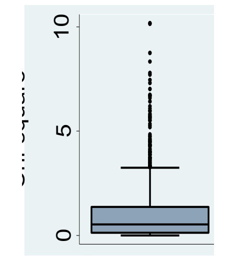


Thought #4

- Compared to causality, prediction is easy since you only have to understand associations and correlation.

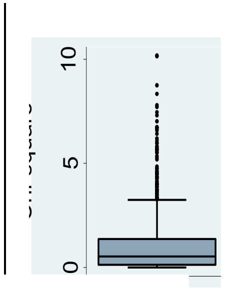
Until it isn’t.

Correlation versus causation



Many big data “successes” are solely prediction.

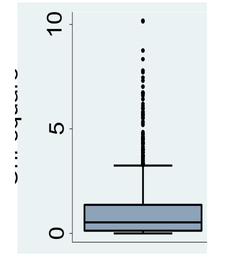
Data mining methods can help you predict 30 day readmission rates but may not tell you how to improve them.



Correlation versus causation

There is the feeling that if we just take enough measurements we will understand causal effects:

“There is now a better way. Petabytes allow us to say: ‘Correlation is enough.’ We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.” Wired Mag. 6/2008

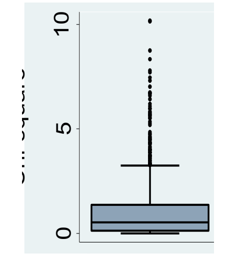


Correlation versus causation

And it continues:

"The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?

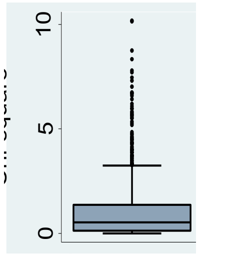


Epidemiologists have long studied the pitfalls of using observational studies to infer causation.

A fundamental idea in causal inference is to compare the same person with and without an intervention or condition. Can't usually measure the person under both conditions.

- For example, observe the same person with and without a history of knee injury to understand its impact on later osteoarthritis.

Thoughts #5

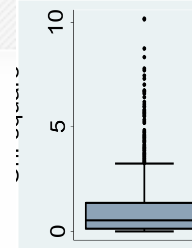


- If you are interested in causation you can't measure it all.

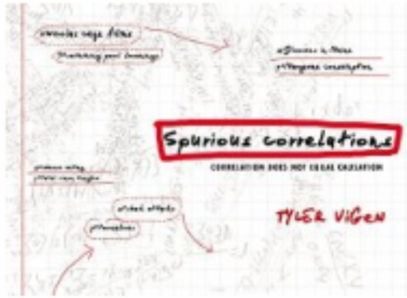
You can, at most, measure half of it all and that half is subject to selection bias.

tylervigen.com

about | twitter | email | subs



Spurious correlations



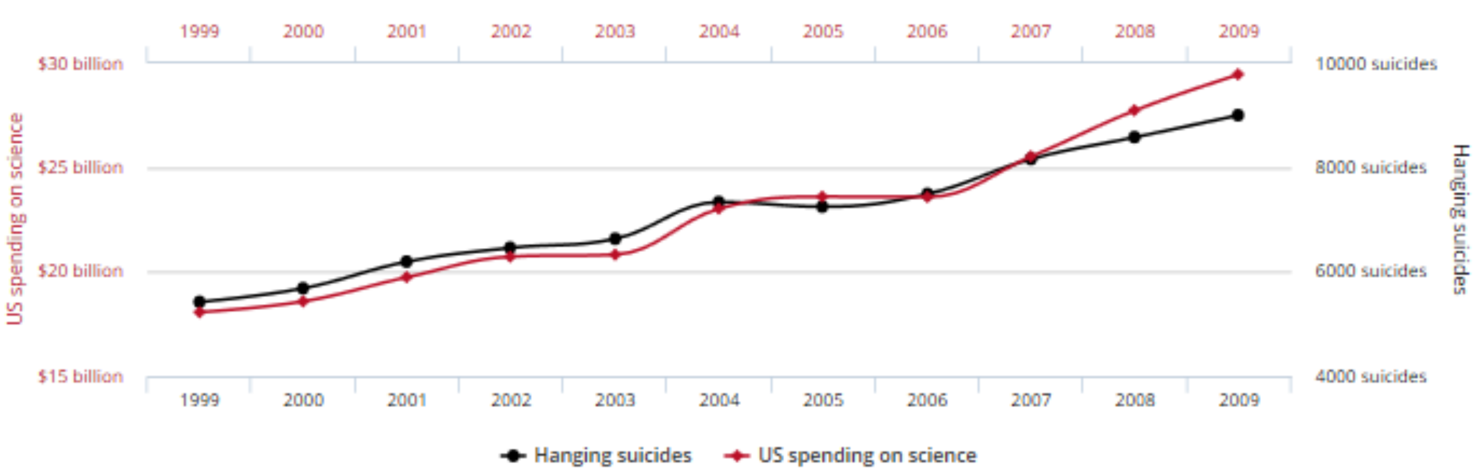
Now a ridiculous book!

- Spurious charts
- Fascinating factoids
- Commentary in the footnotes

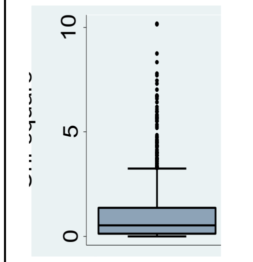
Amazon | Barnes & Noble | Indie Bound

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation

Correlation: 99.79% (r=0.99789126)



Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention



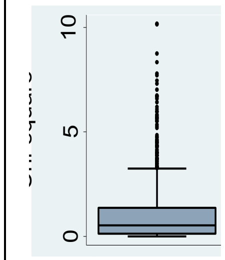
Need to beware of overfitting

1. Overfitting, trying too hard to find a model that closely matches data
2. Revisited the OA example and added in 3 spurious (randomly generated) variables, spurious1, spurious2, spurious3

    Tree: spurious1

We will work very hard to avoid overfitting!

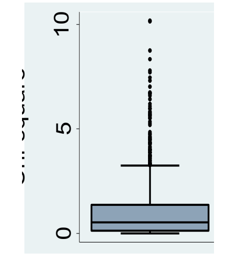
Thought #6



- Modern machine learning methods are just as good or even better than the more usual statistical methods at finding spurious correlations.
  - Usual solution: validation data set or subsets.

- Flexibility also makes it harder to conduct reproducible research.

Having our cake and eating it too: Big data and randomized trials



"Early Supported Discharge for Improving Functional Outcomes After Stroke"

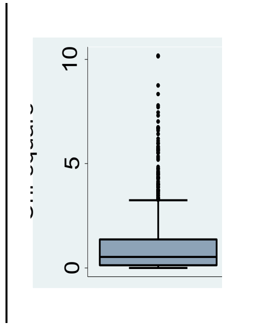
PCORI funded project to compare a more comprehensive package of post-stroke services with usual care.

Randomize most stroke patients in North Carolina by randomizing hospitals to intervention or control.

Outcomes collected electronically: death, recurrent stroke, use of transition billing codes, proportion re-hospitalized.

Thought #7

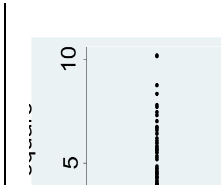
- We can have our cake and eat it too.



Thoughts on Big Data

1. Data mining can be used for good and evil.
2. Big data isn't the same as big information.
3. Modern machine learning methods can search for associations quickly and flexibly.
4. Having big data doesn't solve selection bias.
5. Distinguish prediction problems from causal investigation.
6. Data mining methods are good at finding spurious associations and hard to use reproducibly. Need to wary of overfitting.
7. Randomization can be used with big data.

Resources



First class Thursday, July 30 1:00 PM. Lab to follow.

Contact info: aaron.scheffler@ucsf.edu

Download Orange Data Mining Software: https://orange.biolab.si/download/

Course website: https://courses.ucsf.edu/course/view.php?id=7510