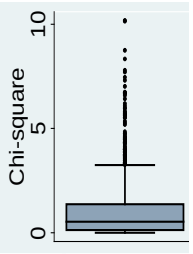
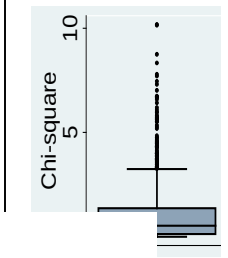
Big Data


Chi-square
0
5
10

WWW.ENTERTAINMENTWALLPAPER.COM
CAPTAIN AMERICA
THE WINTER SOLDIER

Big Data



Jasper Sitwell: Zola's algorithm is a program...for choosing Insight's targets.

Steve Rogers: What targets?

Jasper Sitwell: You! The TV anchor in Cairo, the Undersecretary of Defense, a high school valedictorian in Iowa city. Bruce Banner, Stephen Strange, anyone who's a threat to HYDRA. Now, or in the future.

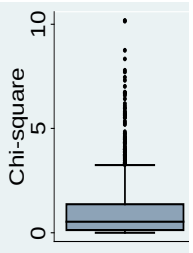
Steve Rogers The future? How can it know?

Jasper Sitwell: How could it not? The 21st century is a digital book. Zola taught HYDRA how to read it. Your bank records, medical histories, voting patterns, e-mails, phone calls, the damn SAT scores. Zola's algorithm evaluates peoples' past to predict their future.

Steve Rogers What then?

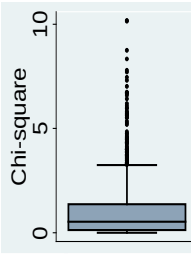
Jasper Sitwell: Then the Insight heli-carriers scratch people off the list.

Big Data



Nick Fury: We've been data-mining HYDRA's files. Looks like a lot of rats didn't go down with the ship.

That was just for fun, *but*



What 7 Creepy Patents Reveal About Facebook

The New York Times

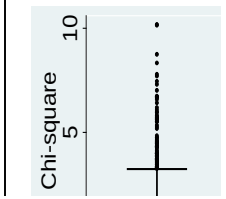
Listening to your environment

By Sahil Chi

Illustrations b

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*

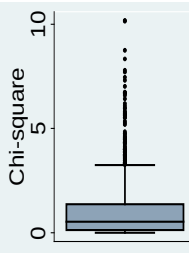


IN BRIEF

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

04 | SIGNIFICANCE | June 2018

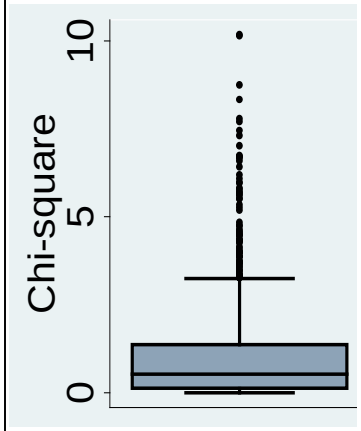
Thought #1



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

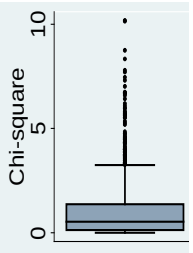
Biostat 202: Opportunities and Challenges of "Big Data"

Charles E. McCulloch,
Division of Biostatistics,
Department of Epidemiology and Biostatistics



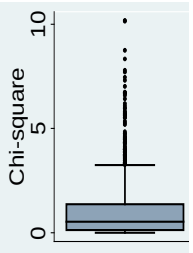
Biostat 202

Course outline



- Introduction
- Getting data (large databases, electronic health records)
- Managing data
- Digital Health
- Describing data with tables and figures
- Algorithms (predictive, clustering)
- Causal inference and big data
- Case study

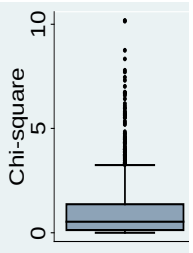
Course Grading



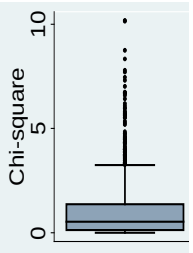
- Homework (40%)
- Lab assignment (0/1/2 grading – 15%)
  - (0=not handed in on time, 1=substantial elements missing, 2=essentially complete)
- Project
  - 3 “components” (0/1/2 grading – 15%)
  - Final project write-up (30%)

See the CLE website for details on the class due dates and project guidelines. More coming on the course project datasets.

TICR Professional Conduct Statement

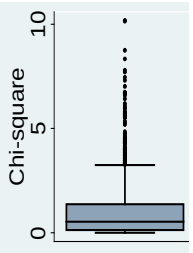


- I will maintain the highest standards of academic honesty
- I will neither give nor receive aid in examinations or assignments *unless such cooperation is expressly permitted by the instructor* (Our policy: mutual collaboration on labs and assignments is fine. The project should be your own work)
- I will conduct research in an unbiased manner, report results truthfully, and credit ideas developed and work done by others
- I will not use answer keys from prior years
- I will write answers in my own words, and, when collaboration is permitted, acknowledge collaborators when answers are jointly formulated.



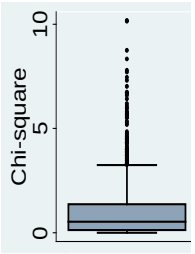
Course organization and dates

- Lectures
  - Mondays and Thursdays 1-2:30pm (8/2-9/10)
- Labs
  - Thursdays 2:45-4:15pm (8/2-9/6)
  - Interactive and hands-on learning session
  - Lab assignments due by midnight on Sunday
- Homeworks due Thursdays starting 8/9.
- Project components due Mondays from 8/20.
- Final project due Tuesday 9/11.



Course software:

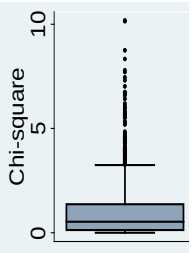
- IBM SPSS Modeler
  - Free license for a year through UCSF
  - Regular price of $4,500 per person per year
- A key aspect of the labs will be hands-on use of the software
- Allows access to many of the features of big data analytics without the need to learn programming
  - Allows more focus on concepts rather than technical programming details



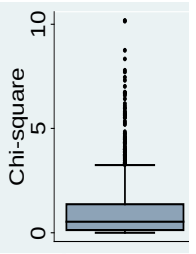
Optional reading (posted on CLE):

- *Data Mining with SPSS Modeler. Theory, Exercises and Solutions.* Free as an ebook from within the UCSF network via http://link.springer.com/book/10.1007%2F978-3-319-28709-6.
- *An Introduction to Statistical Learning*. eBook free via www-bcf.usc.edu/~gareth/ISL/. Or within UCSF network from Springer directly.

Outline for today



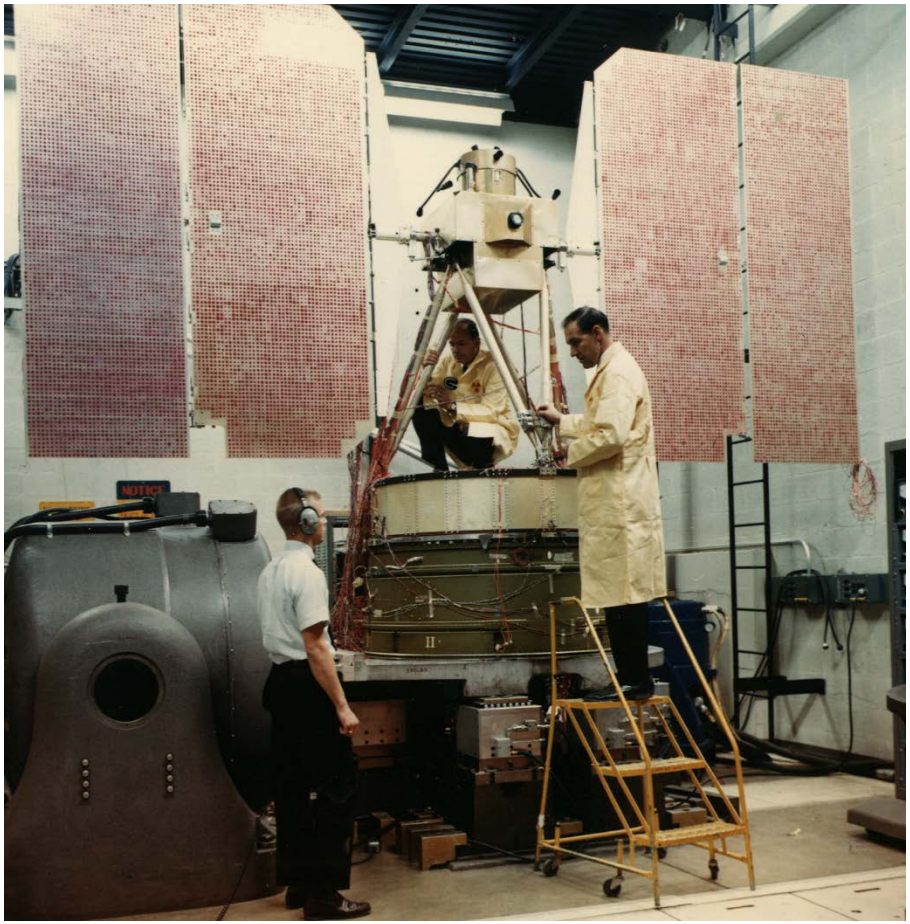
- What is big data? And since computers are so fast these days, is it really any different than what we have been doing all along?
- Data mining.
- Causality versus prediction.
- Overfitting and spurious associations.

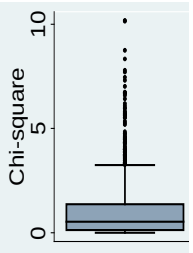


Is Big Data new?

My father actually *was* a rocket scientist.

Here is one of his satellites being tested: Nimbus weather satellite.





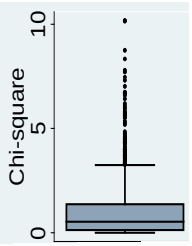
THE NIMBUS 5 USER'S GUIDE

SECTION 3

THE TEMPERATURE-HUMIDITY INFRARED RADIOMETER (THIR) EXPERIMENT

By
Andrew W. McCulloch
National Aeronautics and Space Administration
Goddard Space Flight Center

Nimbus Users Guide

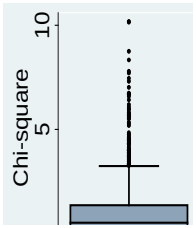


3.4.2 Digital Data

Quantitative data results when the original analog signals are digitized with full fidelity and the digital data are processed by an IBM 360 computer where calibration and geographic referencing is applied automatically.

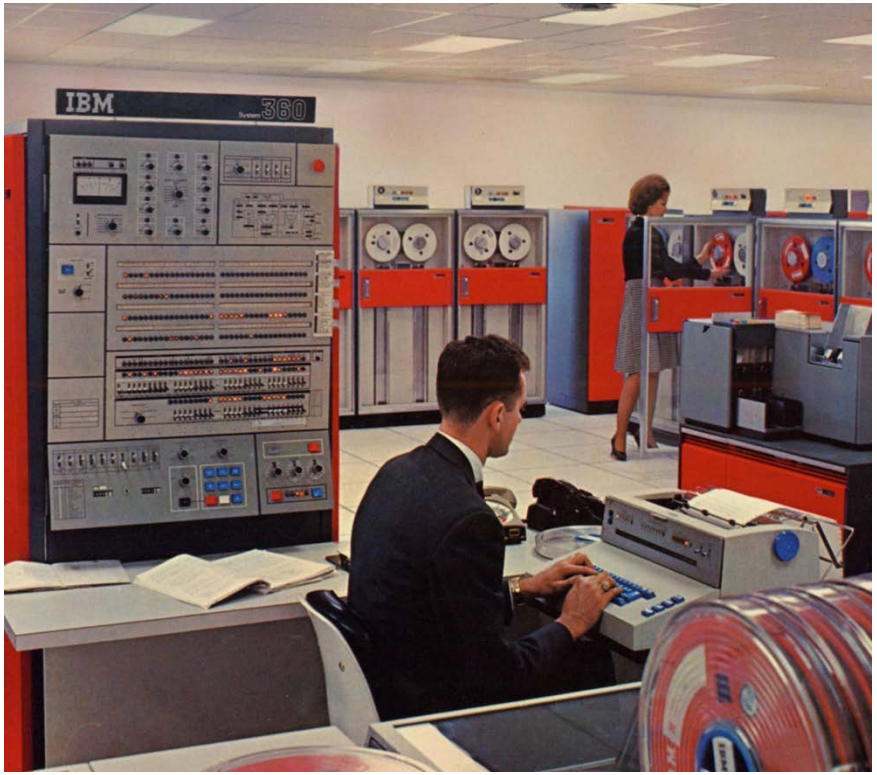
3.4.2.1 Availability of Processed Digital THIR Data

Due to the large volume of data and the long computer running time required for processing it into NMRT's, Nimbus IV THIR digital data are not routinely reduced to final NMRT Format. Only those data which are specifically requested by the user will be processed. Requests should be made through NSSDC. It is anticipated that requested THIR-NMR tapes will begin to be available through NSSDC six months after launch. The user is urged to make full use of the film strips which are abundantly available in nearly real-time from the NSSDC.



Is Big Data new?

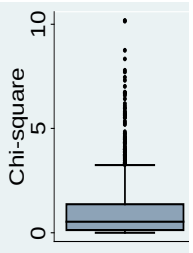
So even in my father's time, he had "big" data. And he needed the latest and greatest computing equipment (an IBM 360) to process it.



The weather files are archived and anyone can download them from the Nat'l Snow and Ice Data Center.

They are about 1 Mb per day.

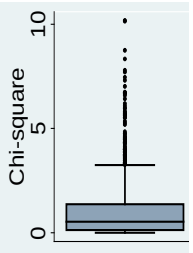
Big Data – what is it?



Every era has had big data.

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

Big Data in the 1990s



BIOLOGY OF REPRODUCTION 46, 1158–1164 (1992)

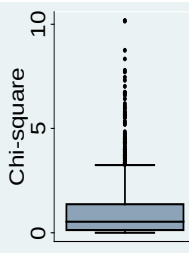
Different Patterns of Myometrial Activity and 24-H Rhythms in Myometrial Contractility in the Gravid Baboon during the Second Half of Pregnancy[1]

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

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

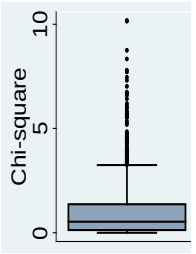
Big Data in the 1990s



- 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.

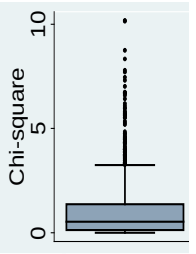
In this example, 4.8 billion data points from only 13 animals.

So the usual name of “data mining” and not “information mining” is appropriate.

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.

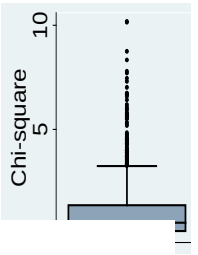
— Clifford Stoll, *High-Tech Heretic: Reflections of a Computer Contrarian* (2000), pp. 185-186.

What is “Big Data” today?



1. Electronic health records.
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., 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?

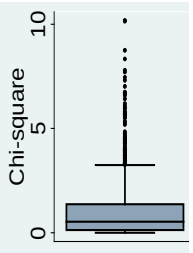


Is “big data” really any different than what we have been doing for decades other than a steady increase in volume?

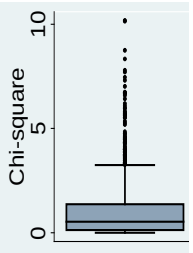
Yes

- 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 course datasets). In a cohort of individuals at risk for knee osteoarthritis, can we predict who will go on to develop painful knee arthritis? Predictors include many 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) and blood sample (numerous bone turnover markers). Full study has 1000s of predictors.



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:

2^{30} = 1.1×10^7 seconds

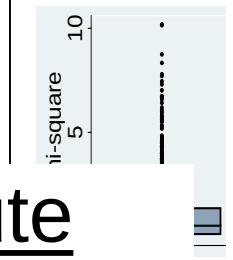
= 18 million minutes

= 300,000 hours

= 12,500 days

= 34 years to check them all

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

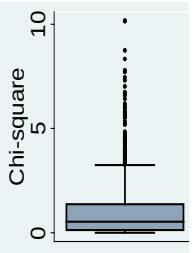
About BD2K

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?

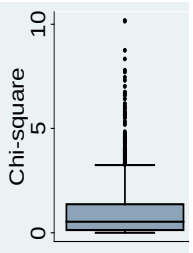
Not just a fad



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.

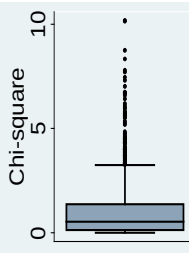
Example: OA Biomarkers Project



Predict participants who go on lose full thickness of their cartilage in the medial compartment of the knee (FT_cart_loss). We'll consider just 13 predictors, so there are 2^{13} = 8,192 possible models (assessing all 8,192 would take you 3.5 workweeks if you could do 1 per minute).

We will consider three methods for predicting the outcome, stepwise logistic regression, all subsets regression and a "tree" based method. But first some graphics.

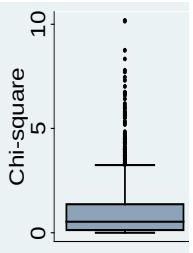
Example: Predicting progression of osteoarthritis



SPSS Modeler demonstration

- Read in a dataset
- Simple descriptive statistics
- Simple graphical description

Stata Illustration: Predicting progression of osteoarthritis



Demonstration:

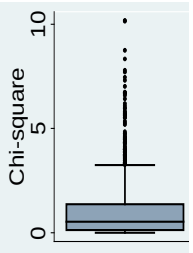
Stepwise -

sw, pe(0.05) pr(0.1): logistic FT_cart_loss_medial BMI … min_medial_JSW

All possible subsets -

gvselect <term> BMI … min_medial_JSW, nmodels(2): logistic FT_cart_loss_medial <term>

Stata Illustration: Predicting progression of osteoarthritis

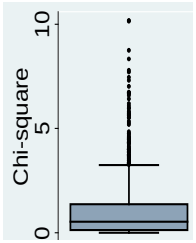


I also added in extra predictors to total 30 of them. Recall that a human would take 34 years to review all the models at 1 per second.

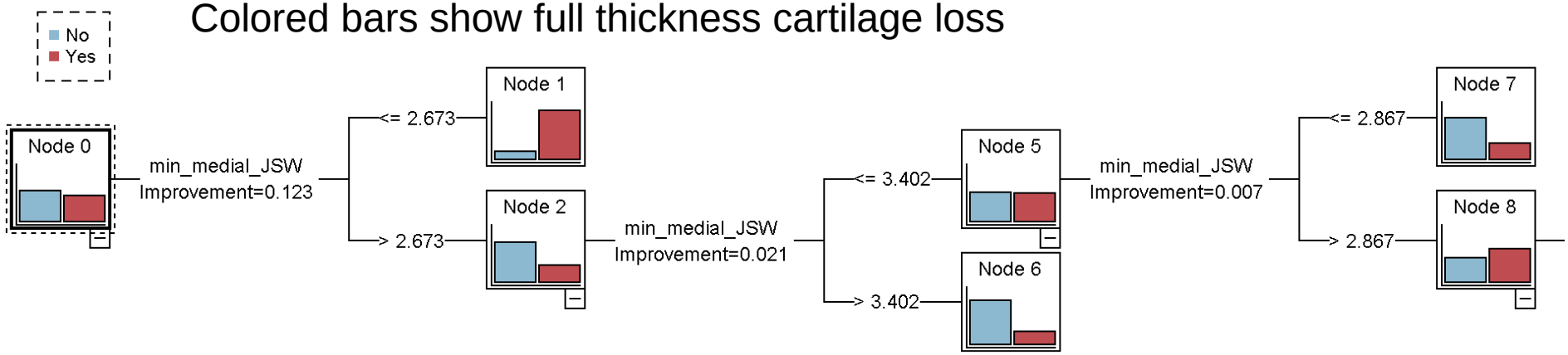
Stata took <8 hours on my laptop.

And that could easily be dropped by an order of magnitude by using more computing cores.

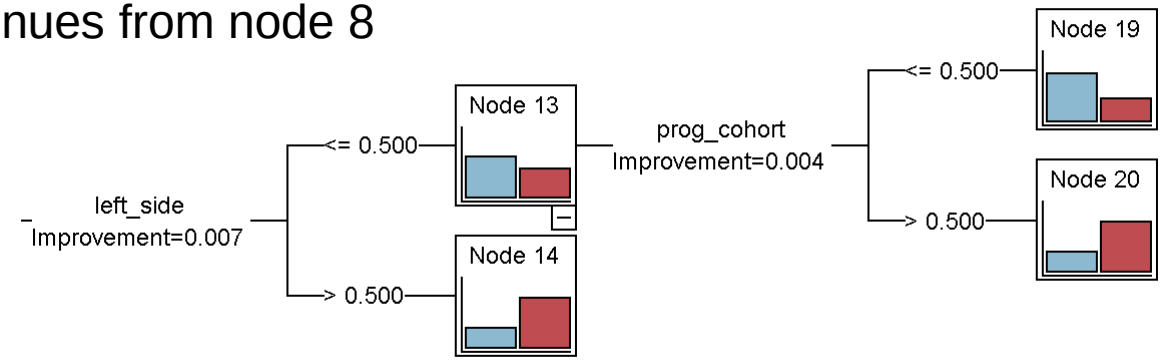
Example: Predicting progression of osteoarthritis: tree model



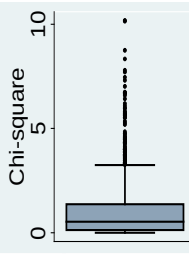
Colored bars show full thickness cartilage loss



classification tree continues from node 8

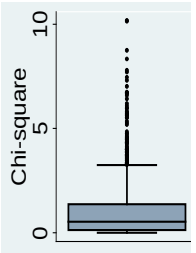


Example: Variables selected



Variable	Stepwise	All subsets	Tree
Female	✓	✓	
Kellgren Lawrence	✓	✓	
Left side	✓	✓	✓
Progression cohort	✓	✓	✓
Min Medial JSW	✓		✓✓✓
Medial JS grade		✓	
Knee pain			

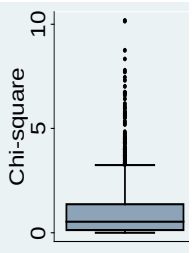
Not just a fad



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

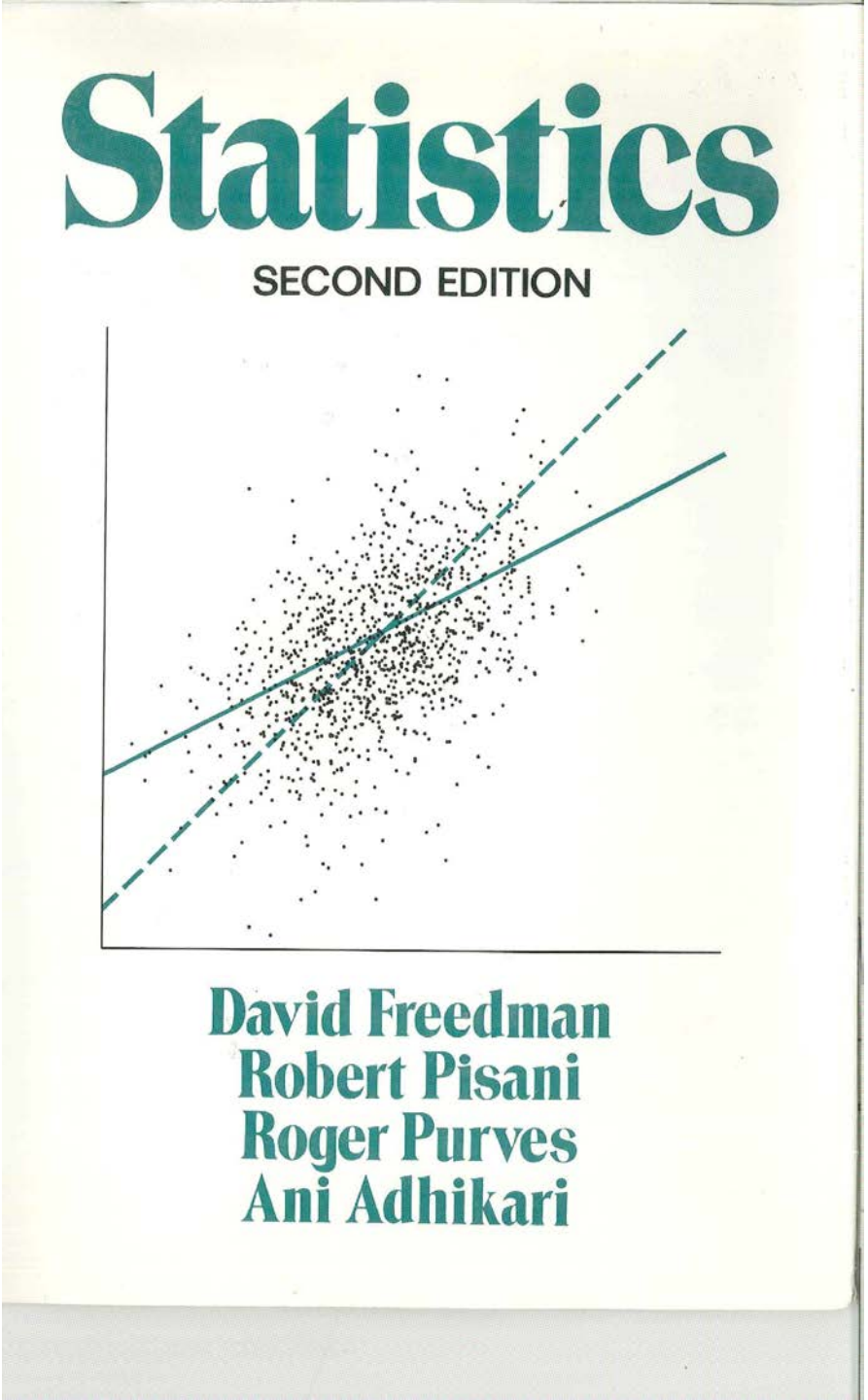
For the most up-to-date levels of evidence, see www.cebm.net/?o=1025

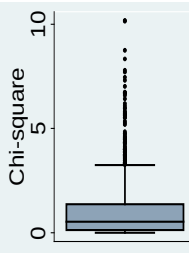
Therapy/Prevention/Etiology/Harm:

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"

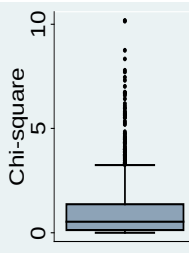
Osteoarthritis Initiative

Electronic health record based study


Statistics
SECOND EDITION
David Freedman
Robert Pisani
Roger Purves
Ani Adhikari


Chi-square
0
5
10

1991



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.

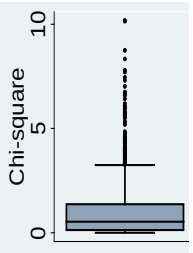
	Roosevelt's percentage
The *Digest* prediction of the election result	43
Gallup's prediction of the election result	56
The election result	62

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*.

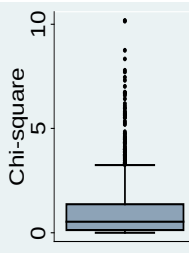
When a selection procedure is biased, taking a large sample does not help. This just repeats the basic mistake on a larger scale.



The Literary Digest Poll

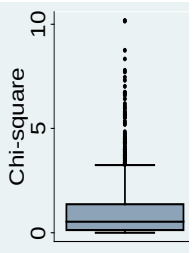
After its prediction of the election of 1936, Literary Digest folded in 1938.

Thought #3



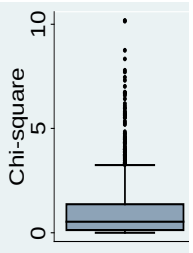
- Having big data doesn’t help if was selected in a biased way.

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 1000s of 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.

Key aspects: 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]

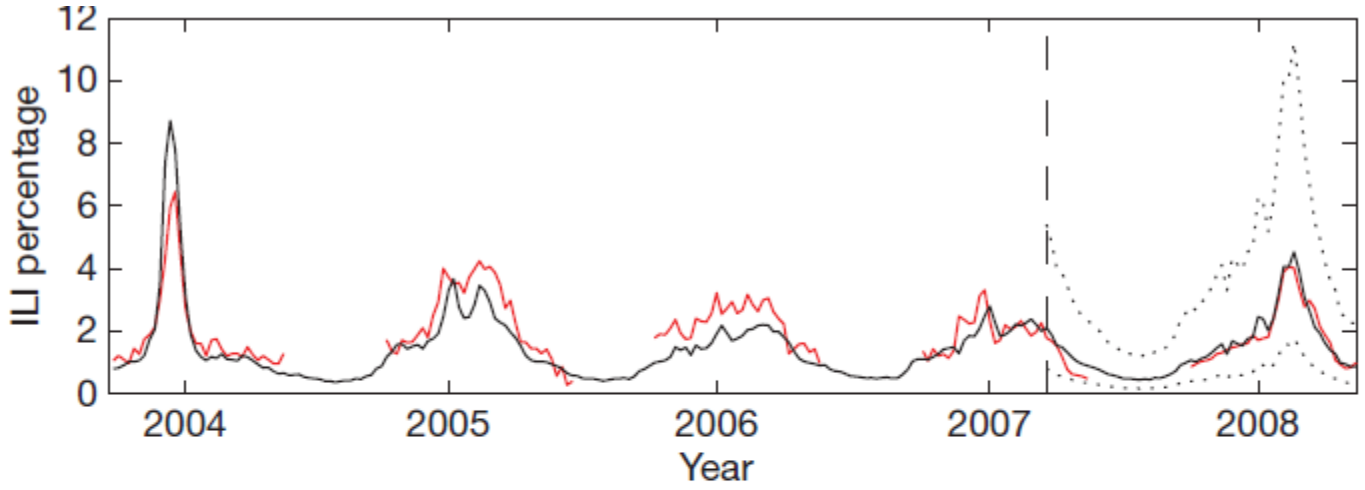
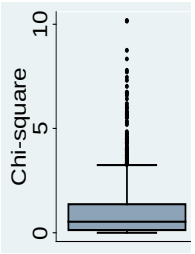


Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation 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]

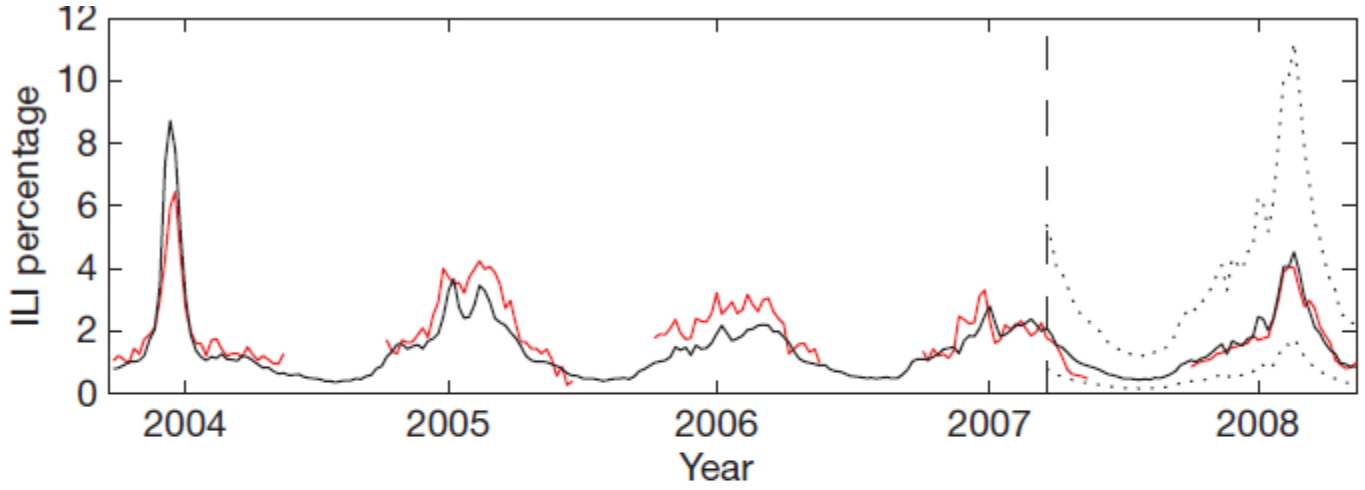
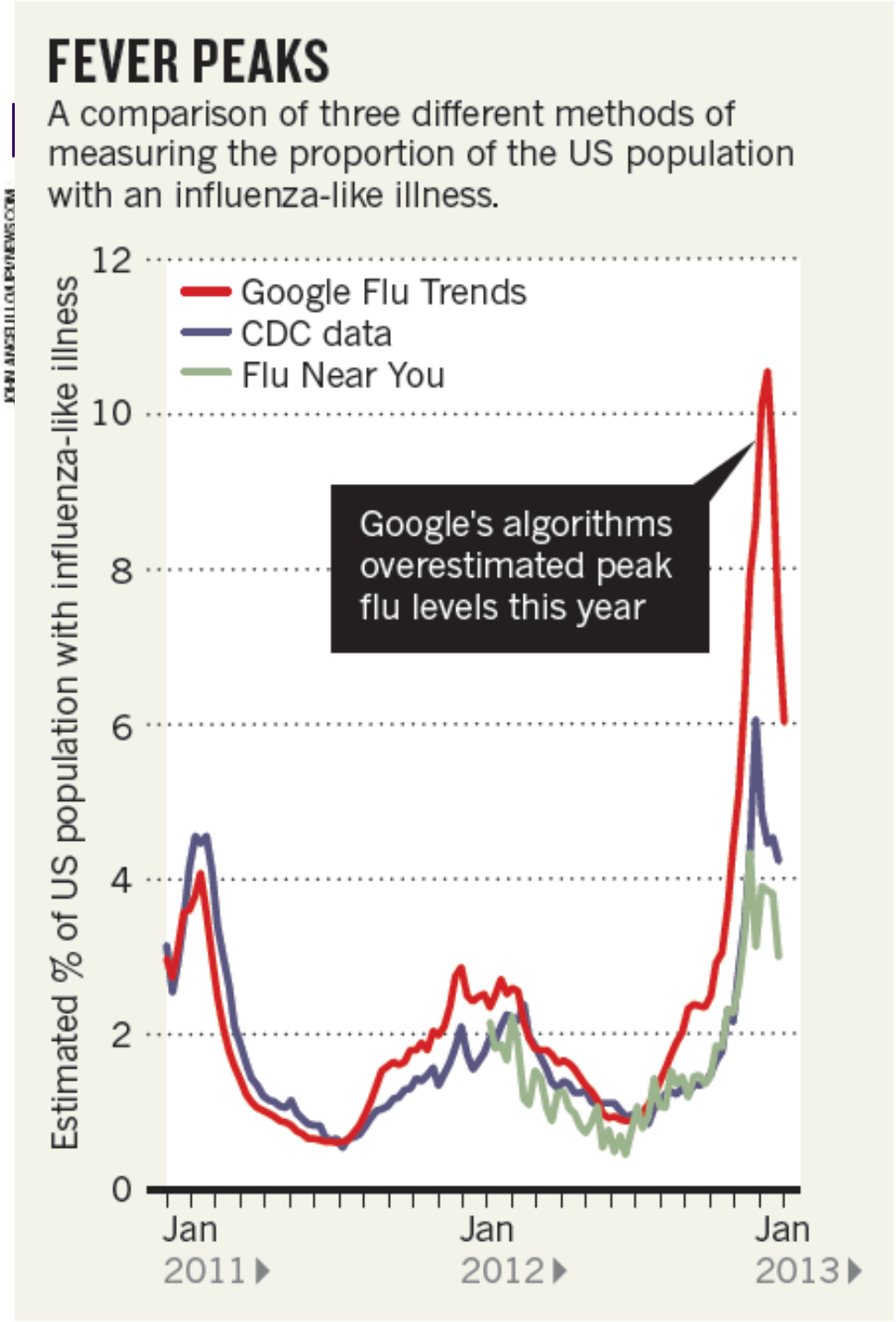
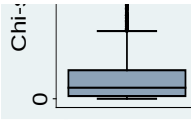


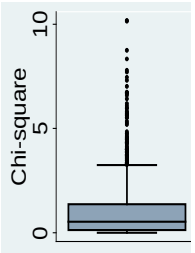
Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation 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





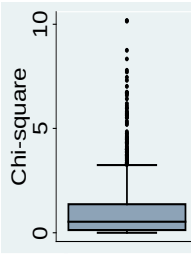
Why did it break?



- 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

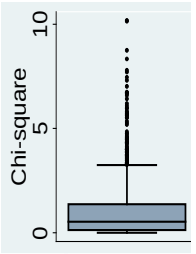
Thought #4



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

Until it isn’t.

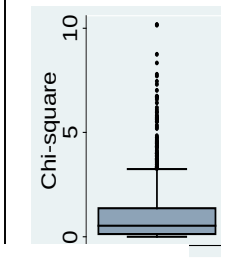
Correlation versus causation



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.

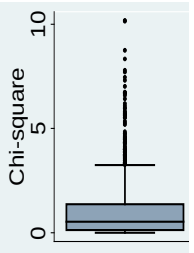
Correlation versus causation



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

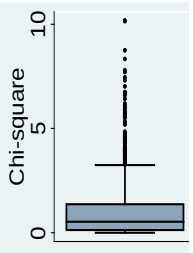


And it continues:

“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?

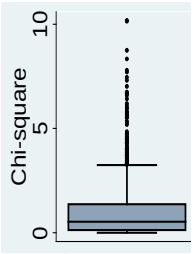
Epidemiologists have long studied the pitfalls of using observational studies to infer causation.



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.

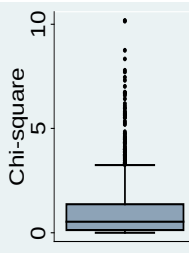
- For example, observe the same person with and without a history of knee injury to understand its impact on later osteoarthritis.
- Or recording the outcomes in a catchment area of an trauma center with and without its closure.

Thought #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.



Spurious correlations

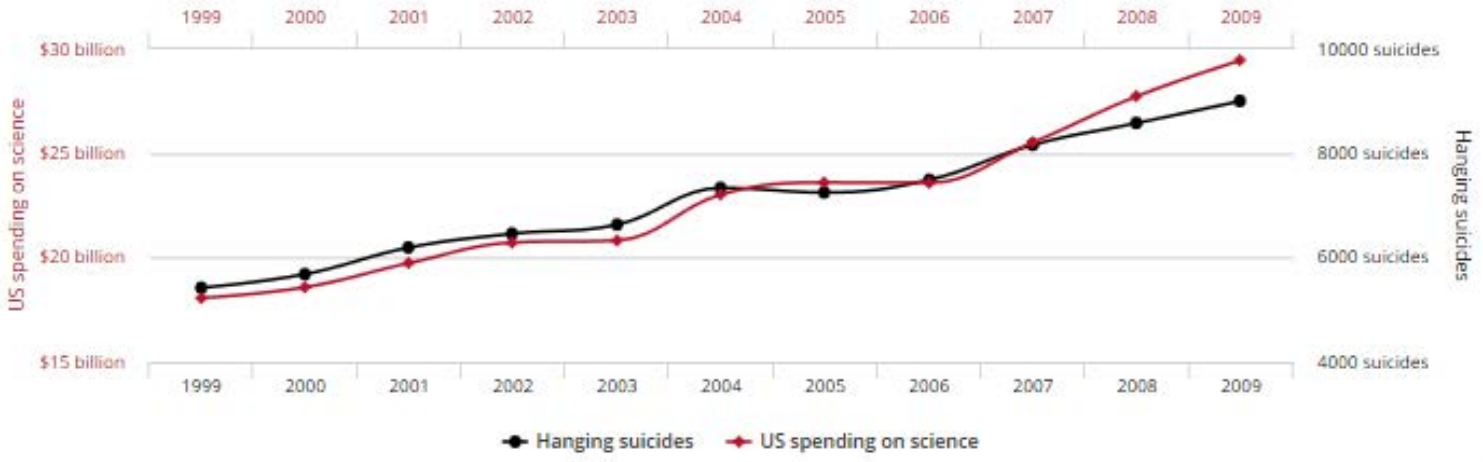
Now a ridiculous book!

- Spurious charts
- Fascinating factoids
- Commentary in the footnotes

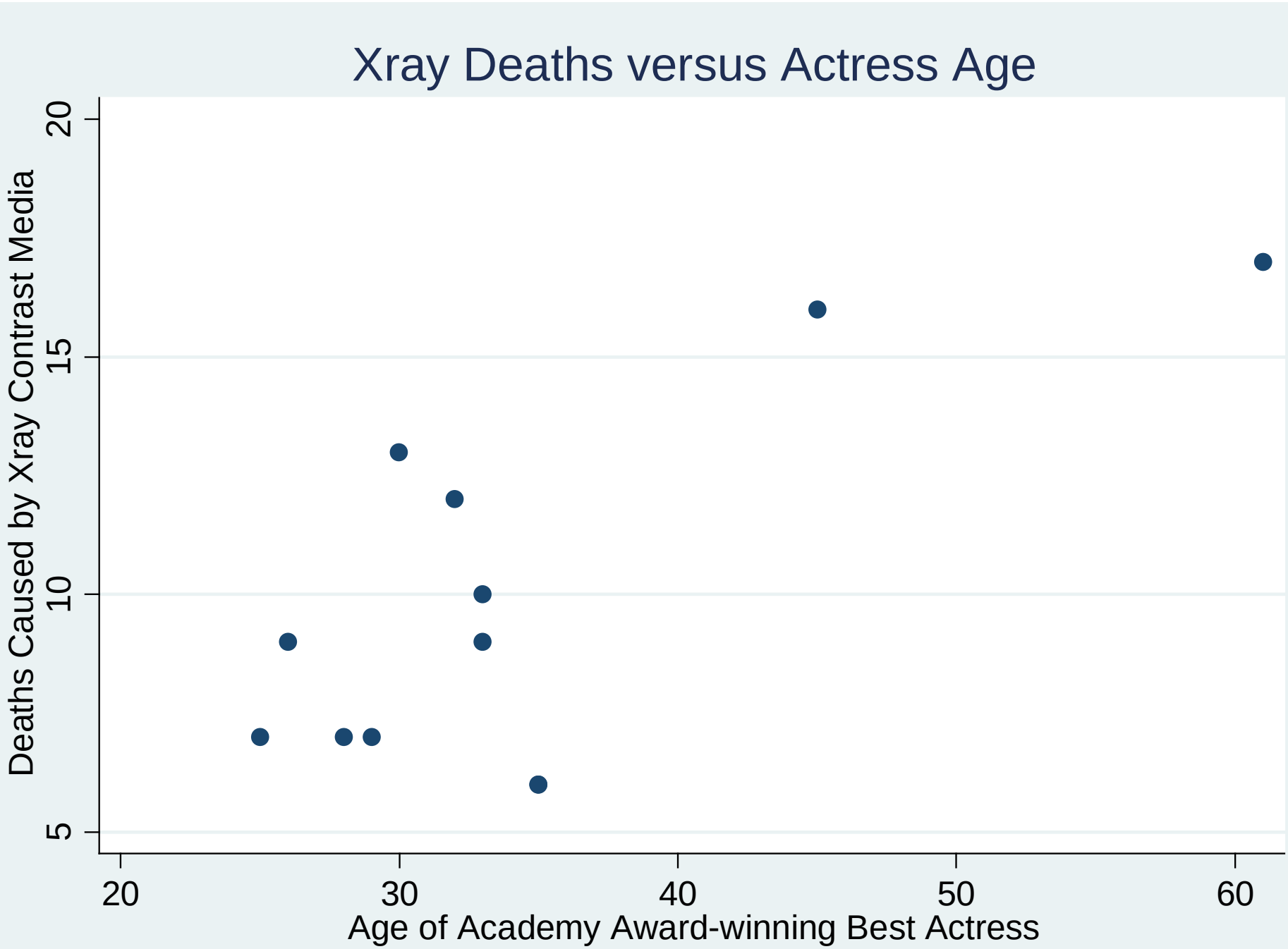
Amazon | Barnes & Noble | Indie Bound

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

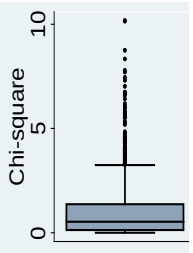
Correlation: 99.79% (r=0.99789126)



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


Xray Deaths versus Actress Age
Deaths Caused by Xray Contrast Media
5
10
15
20
20
30
40
50
60
Age of Academy Award-winning Best Actress

From the Spurious Correlation website

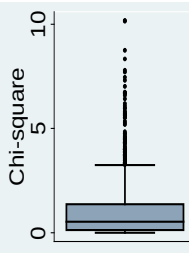


Correlation of deaths caused by X-ray contrast media and the age of the Academy Award winning best actress is 0.74 (p=0.0076).

Stepwise regression shows a further association with annual precipitation in Nebraska (p=0.003) while age of actress remains statistically significant (p=0.032).

Fortunately, number of lawyers in Ohio is not independently associated (p>0.05).

Need to beware of overfitting



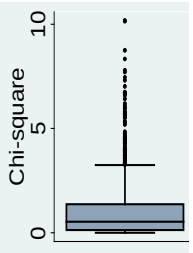
1. Stepwise with unrelated variables demonstration.
2. Revisited the OAI example and added in 3 spurious (randomly generated) variables. Were these selected by the automated methods?

   Stepwise: none

   All subsets: spurious1 and spurious3

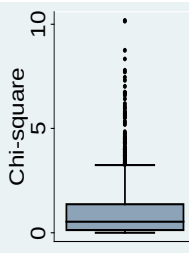
   Tree: spurious1

Thought #6



- Modern machine learning methods are just as good or even better than the more usual statistical methods at finding spurious correlations and make it harder to conduct reproducible research.
- Usual solution: validation data set or subsets.

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



"Early Supported Discharge for Improving Functional Outcomes After Stroke"

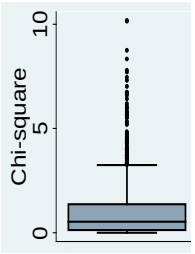
PCORI funded project to compare a more comprehensive package of post-stroke services with usual care.

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. Machine learning methods are good at finding spurious associations and hard to use reproducibly. Need to wary of over-fitting.
7. Randomization can be used with big data.