Causal Inference

Lecture 1

Samantha Kleinberg
samantha.kleinberg@stevens.edu

About me: Samantha Kleinberg

- PhD in CS
- Have worked in biomathematics, bioinformatics/ computational biology, biomedical informatics
- Current research
  - Time series data, causal inference
  - Decision-making/cognition
  - Mobile health

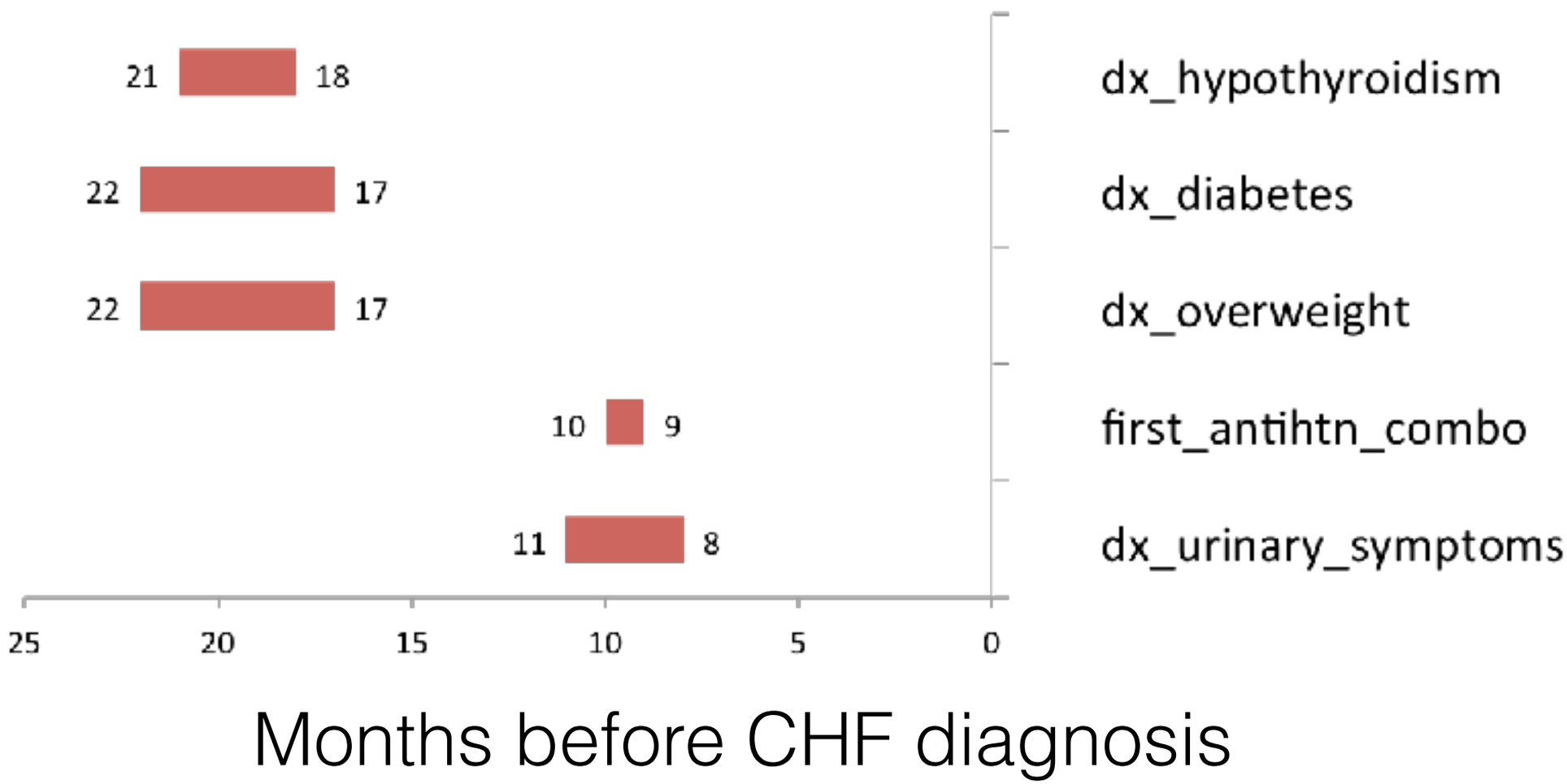
Translating data into knowledge

- Potential
  - Finding new treatments
  - Understand causes of recovery/poor outcome
- Challenges
  - Missing data
  - Data quality
  - Observation vs. underlying physiology

Successes

- Congestive heart failure
  - Early risk factors
- Diabetes
  - Link between exercise and hyperglycemia
- Stroke
  - Different mechanisms for brain swelling

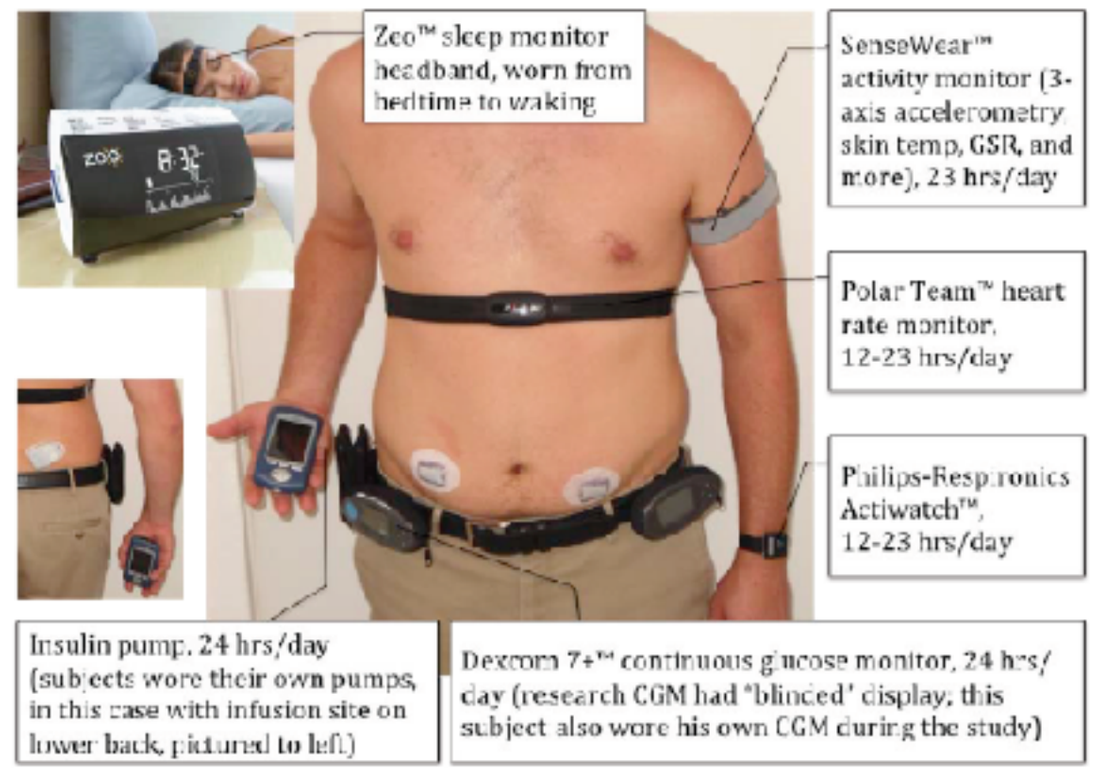
Congestive heart failure



Kleinberg S, Elhadad N (2013) Lessons Learned in Replicating Data-Driven Experiments in Multiple Medical Systems and Patient Populations. In: AMIA Annual Symposium.

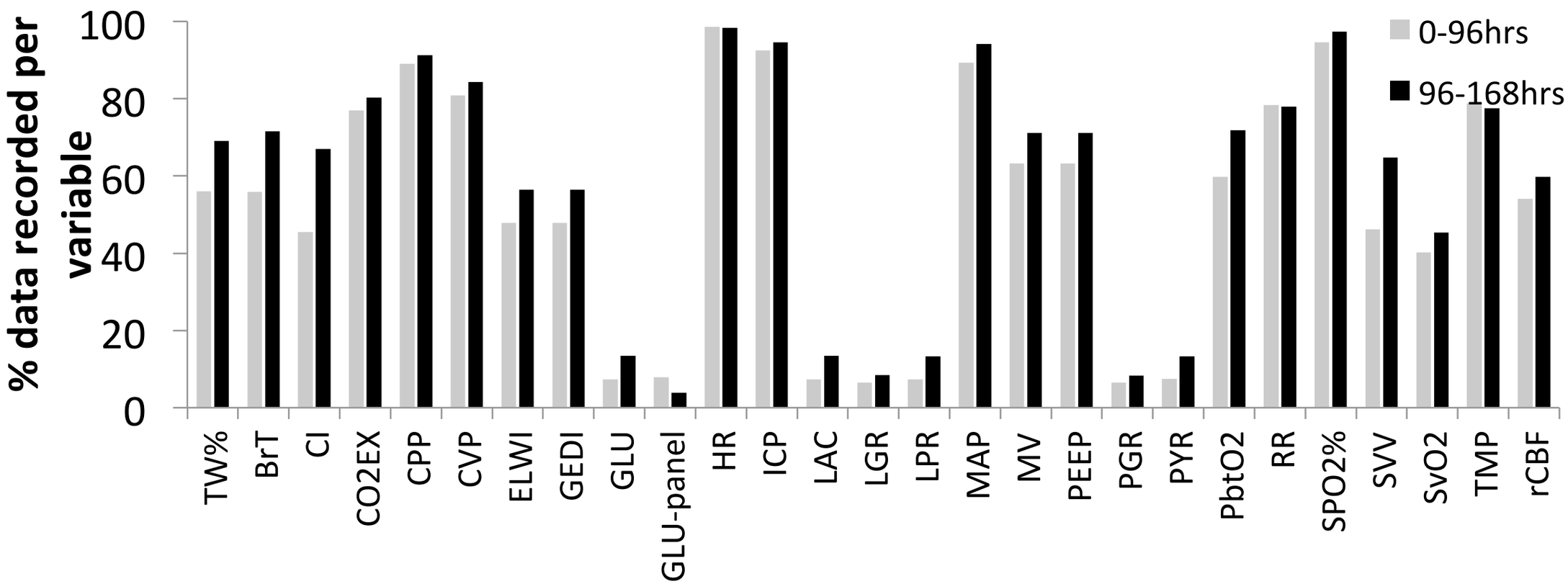
Diabetes

- Used causal inference methods + body-worn sensors to find cause of changes in glycemia
  - intense activity leads to hyperglycemia



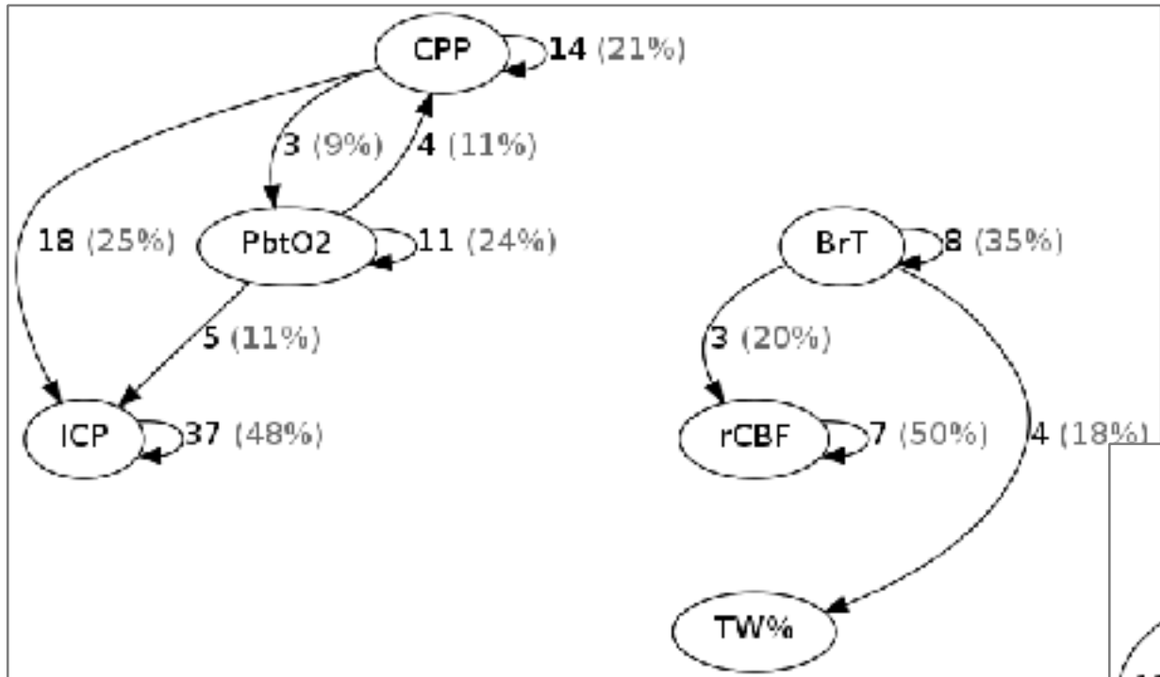
N. Heintzman and S. Kleinberg. Using Uncertain Data from Body-Worn Sensors to Gain Insight into Type 1 Diabetes. Journal of Biomedical Informatics (2016)

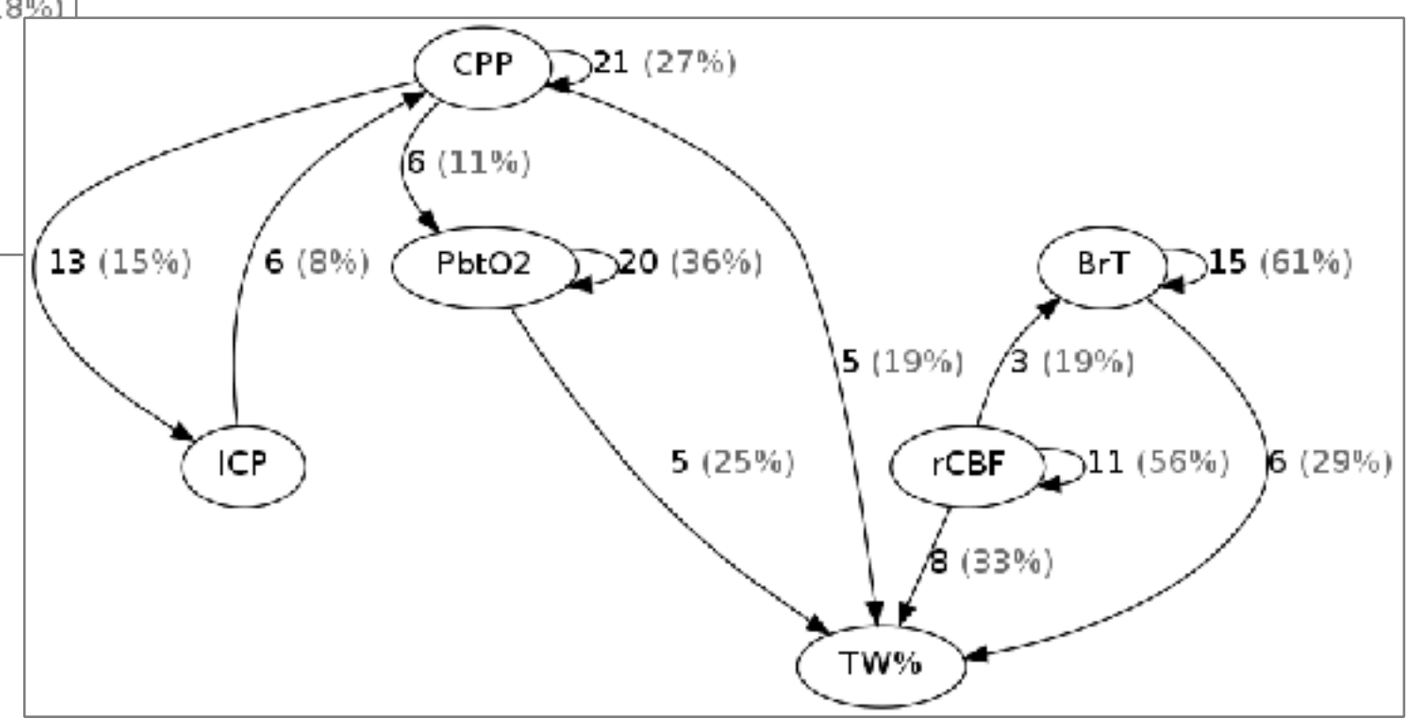
ICU data: high frequency, often missing



S. A. Rahman, Y. Huang, J. Claassen, N. Heintzman, and S. Kleinberg. Combining Fourier and Lagged k-Nearest Neighbor Imputation for Biomedical Time Series Data. Journal of Biomedical Informatics (2015)

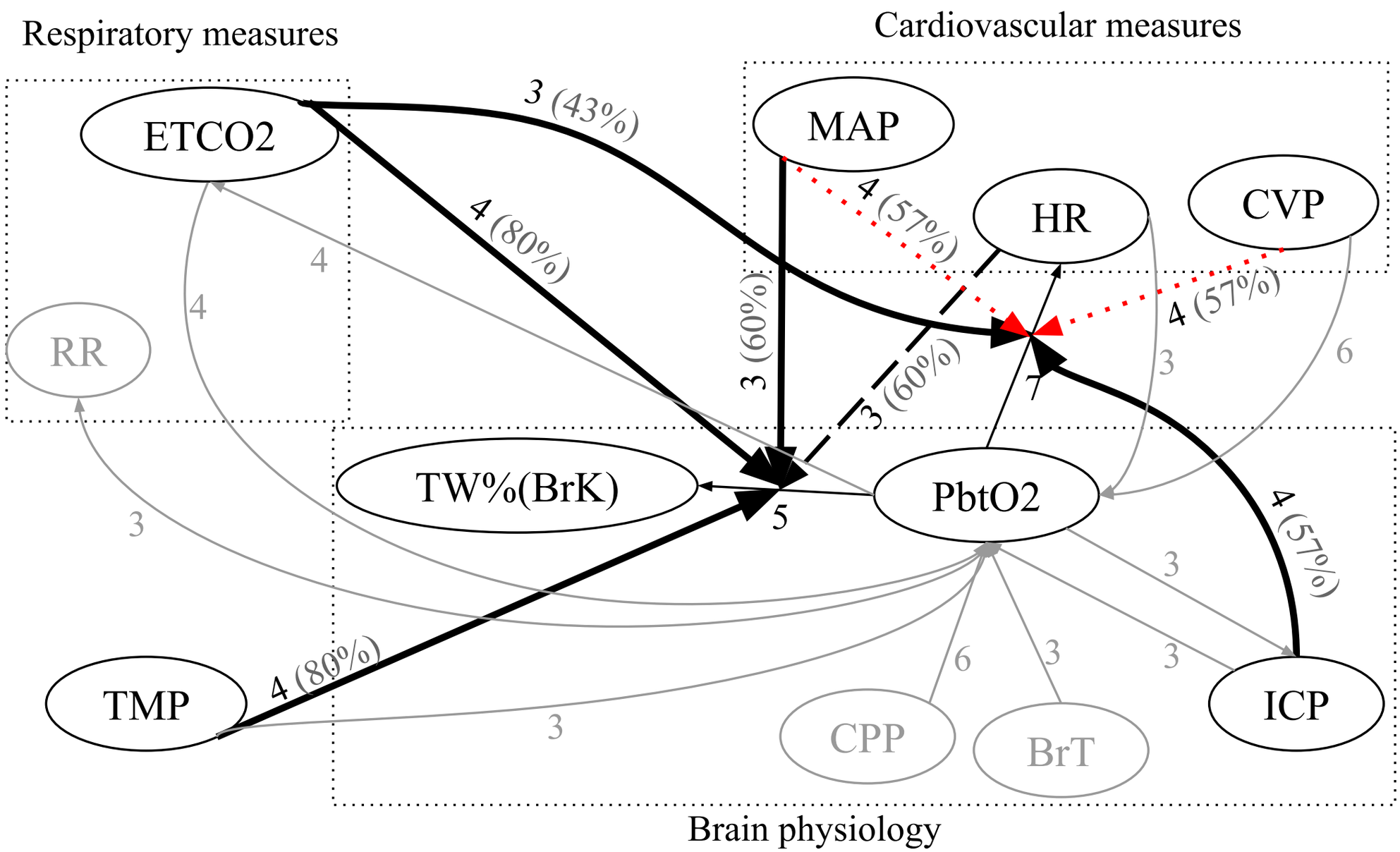
Stroke





Claassen J, Rahman SA, Huang Y, Frey H, Schmidt M, Albers D, Falo CM, Park S, Agarwal S, Connolly ES, Kleinberg S (2015) Causal structure of brain physiology after brain injury. PLoS ONE.

More descriptive structures



Zheng, M, Claassen J, Kleinberg S (2018) Automated Identification of Causal Moderators in Time-Series Data. ACM SIGKDD Workshop on Causal Discovery.

Please introduce yourself

Name
major
why you’re here

Administrative stuff

- Course website: http://www.skleinberg.org/teaching/CI18
- Prerequisites: none
- Textbook: **none**, articles
- Everything on syllabus CAN CHANGE (will make announcement in class)
- Workload/grading:
  - midterm exam (30%), final project (50%), participation (15%), homework (5%)
- Participation includes weekly reading discussions

Some previous final projects

- Does popularity cause campaign contributions?
- What causes flight delays at Newark airport?
- Is the value of bitcoin driven by exchange rates?
- Does fracking cause earthquakes?

Key policies

1. No late work. There are few deadlines, but if the deadline's 11:59pm, work submitted at 12:01am isn't accepted.

   Do not email me saying the time changed as you were submitting. That is the definition of late.

   Why? Try submitting an NIH grant or conference paper 2 minutes late!

2. Plagiarism = F

What’s plagiarism?

Presenting someone else’s work as your own

- Copying an entire paper
- Copying parts of other works, without attribution
  - E.g. quotes without citations, images
- Changing a few words, but keeping ideas and structure, without acknowledging source

Easy to avoid! Do your own work and acknowledge all sources. Quotes must be in quotes. Don’t submit a collage.

1. Quotes belong in quotation marks

- WRONG!
  - Blah blah. Text from other papers. Blah Blah. [Citation]
  - **Section 1** [Citation]. Text you did not write.
  - **Section 1** [Full text of someone else's paper] **References** Else, Someone. "Paper you copied from"

- Right ☺️
  - My text "Quote from awesome work." [citation] my text

Seriously, papers are not collages

- “But the other paper said what I wanted to!”
  - Put it in your own words
- “It’s just reference material”
  - Then add a reference to it
- “I copied from your book because you said it so well!”
  - I am not senile and will recognize my own words

3. No copying/close paraphrasing

Don't Let This Be You!

2006: Kaavya Viswanathan, Harvard sophomore, 2-book deal from Little, Brown

Plagiarized!

Sloppy Firsts (2001), page 7: "Bridget is my age and lives across the street. For the first twelve years of my life, these qualifications were all I needed in a best friend. But that was before Bridget's braces came off and her boyfriend Burke got on, before Hope and I met in our seventh grade Honors classes" (McCafferty, quoted in Zhou, "Examples…").

Opal Mehta (2006), page 14: "Priscilla was my age and lived two blocks away. **For the first** fifteen years of my life**, those were the only qualifications I needed in a best friend**. … **But that was before** freshman year, when Priscilla's glasses **came off**, and the first in a long string of boyfriends **got on**" (Viswanathan, quoted in Zhou, "Examples…").

"an act of literary identity theft"
(Steve Ross, qtd. in Zhou, "Publisher…")

http://researchguides.stevens.edu/plagiarism

If you’re having trouble…

- Come to office hours! Monday 1:45-2:45 in North 208 and by appointment
- Email me

Causal claims abound

March 12, 2012

Risks: More Red Meat, More Mortality

By NICHOLAS BAKALAR

Eating red meat is associated with a sharply increased risk of death from cancer and heart disease, according to a new study, and the more of it you eat, the greater the risk.

The analysis, published online Monday in Archives of Internal Medicine, used data from two studies that involved 121,342 men and women who filled out questionnaires about health and diet from 1980 through 2006. There were 23,926 deaths in the group, including 5,910 from cardiovascular disease and 9,464 from cancer.

People who ate more red meat were less physically active and more likely to smoke and had a higher body mass index, researchers found. Still, after controlling for those and other variables, they found that each daily increase of three ounces of red meat was associated with a 12 percent greater risk of dying over all, including a 16 percent greater risk of cardiovascular death and a 10 percent greater risk of cancer death.

The increased risks linked to processed meat, like bacon, were even greater: 20 percent over all, 21 percent for cardiovascular disease and 16 percent for cancer.

If people in the study had eaten half as much meat, the researchers estimated, deaths in the group would have declined 9.3 percent in men and 7.6 percent in women.

Previous studies have linked red meat consumption and mortality, but the new results suggest a surprisingly strong link.

“When you have these numbers in front of you, it’s pretty staggering,” said the study’s lead author, Dr. Frank B. Hu, a professor of medicine at Harvard.

What is a cause?

“Most striking, society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing why but only what.”

Mayer-Schonberger, V. and K. Cukier. (2013) Big Data: A Revolution That Will Transform How We Live, Work, and Think. Earnon Dolan/Houghton Mifflin Harcourt, (page 7).

Why is causality hard?

- No single definition
- No fail-proof method for finding it
- Observational data

Causality has consequences

- Sally Clark's 1st son died in 1996, as a result of SIDS
- Her 2nd son died in 1999, also as a result of SIDS

- Prosecutors argued too unlikely to both be SIDS, must be murder.
  - Chance of SIDS = 1/8,543 so chance of 2 deaths = 1/(8,543*8,543) (approx 1 in 73 million)
  - What's wrong with this?

XMRV leads to CFS?

Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome

Vincent C. Lombardi,[1]* Francis W. Ruscetti,[2]* Jaydip Das Gupta,[3] Max A. Pfost,[1] Kathryn S. Hagen,[1] Daniel L. Peterson,[1] Sandra K. Ruscetti,[4] Rachel K. Bagni,[5] Cari Petrow-Sadowski,[6] Bert Gold,[2] Michael Dean,[2] Robert H. Silverman,[3] Judy A. Mikovits[1]†

Chronic fatigue syndrome (CFS) is a debilitating disease of unknown etiology that is estimated to affect 17 million people worldwide. Studying peripheral blood mononuclear cells (PBMCs) from CFS patients, we identified DNA from a human gammaretrovirus, xenotropic murine leukemia virus–related virus (XMRV), in 68 of 101 patients (67%) as compared to 8 of 218 (3.7%) healthy controls. Cell culture experiments revealed that patient-derived XMRV is infectious and that both cell-associated and cell-free transmission of the virus are possible. Secondary viral infections were established in uninfected primary lymphocytes and indicator cell lines after their exposure to activated PBMCs, B cells, T cells, or plasma derived from CFS patients. These findings raise the possibility that XMRV may be a contributing factor in the pathogenesis of CFS.

Chronic fatigue syndrome (CFS) is a disorder of unknown etiology that affects multiple organ systems in the body. Patients with CFS display abnormalities in immune system function, often including chronic activation of the innate immune system and a deficiency in natural killer cell activity (*1*, *2*). A number of viruses, including ubiquitous herpesviruses and enteroviruses, have been implicated as possible environmental triggers of CFS (*1*). Patients with CFS often have active β herpesvirus infections, suggesting an underlying immune deficiency.

The recent discovery of a gammaretrovirus, xenotropic murine leukemia virus–related virus (XMRV), in the tumor tissue of a subset of prostate cancer patients prompted us to test whether XMRV might be associated with CFS. Both of these disorders, XMRV-positive prostate cancer and CFS, have been linked to alterations in the antiviral enzyme RNase L (*3*–*5*). Using the Whittemore Peterson Institute's (WPI's) national

[1]Whittemore Peterson Institute, Reno, NV 89557, USA. [2]Laboratory of Experimental Immunology, National Cancer Institute–Frederick, Frederick, MD 21701, USA. [3]Department of Cancer Biology, The Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, OH 44195, USA. [4]Laboratory of Cancer Prevention, National Cancer Institute–Frederick, Frederick, MD 21701, USA. [5]Advanced Technology Program, National Cancer Institute–Frederick, Frederick, MD 21701, USA. [6]Basic Research Program, Scientific Applications International Corporation, National Cancer Institute–Frederick, Frederick, MD 21701, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: judym@wpinstitute.org

XMRV leads to CFS?

August 23, 2010

Study Links Chronic Fatigue to Virus Class

By DAVID TULLER

When the journal Science published an attention-grabbing study last fall linking chronic fatigue syndrome to a recently discovered retrovirus, many experts remained skeptical — especially after four other studies found no such association.

Now a second research team has reported a link between the fatigue syndrome and the same class of virus, a category known as MLV-related viruses. In a paper published Monday by The Proceedings of the National Academy of Sciences, scientists found gene sequences from several MLV-related viruses in blood cells from 32 out of 37 chronic-fatigue patients but only 3 of 44 healthy ones.

The researchers did not find XMRV, the specific retrovirus identified in patients last fall. But by confirming the presence of a cluster of genetically similar viruses, the new study represents a significant advance, experts and advocates say.

"I think it settles the issue of whether the initial report was real or not," said K. Kimberly McCleary, president of the CFIDS Association of America, the leading organization for people with chronic fatigue syndrome.

Leonard A. Jason, a professor of psychology at DePaul University and a leading researcher on the syndrome, agreed. "This class of retroviruses is probably going to be an important piece of the puzzle," he said.

Chronic fatigue syndrome, estimated to afflict at least one million Americans, has no known cause and no accepted diagnostic tests, although patients show signs of immunological, neurological and endocrinological abnormalities. Besides profound exhaustion, symptoms include sleep disorders, cognitive problems, muscle and joint pain, sore throat and headaches.

The new paper, by researchers from the National Institutes of Health, the Food and Drug Administration and Harvard Medical School, was accepted for publication in May. Social networks and online communities soon learned the general findings and were eagerly awaiting the paper.

But in July, researchers from another federal agency, the Centers for Disease Control and Prevention, published a

"I think it settles the issue of whether the initial report was real or not"

But ...

RETRACTION

Post date 23 December 2011

Science is fully retracting the Report "Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome" (*1*). Multiple laboratories, including those of the original authors (*2*), have failed to reliably detect xenotropic murine leukemia virus–related virus (XMRV) or other murine leukemia virus (MLV)–related viruses in chronic fatigue syndrome (CFS) patients. In addition, there is evidence of poor quality control in a number of specific experiments in the Report. Fig. 1, table S1, and fig. S2 have been retracted by the authors (*3*). In response to concerns expressed about Fig. 2C [summarized in (*4*)], the authors acknowledged to *Science* that they omitted important information from the legend of this figure panel. Specifically, they failed to indicate that the CFS patient–derived peripheral blood mononuclear cells (PBMCs) shown in Fig. 2C had been treated with azacytidine as well as phytohemagglutinin and interleukin-2. This was in contrast to the CFS samples shown in Figs. 2A and 2B, which had not been treated with azacytidine.

Given all of these issues, *Science* has lost confidence in the Report and the validity of its conclusions. We note that the majority of the authors have agreed in principle to retract the Report but they have been unable to agree on the wording of their statement. It is *Science*'s opinion that a retraction signed by all the authors is unlikely to be forthcoming. We are therefore editorially retracting the Report. We regret the time and resources that the scientific community has devoted to unsuccessful attempts to replicate these results.

BRUCE ALBERTS
Editor-in-Chief

References

1. V. C. Lombardi *et al.*, *Science* **326**, 585 (2009); 10.1126/science.1179052.
2. G. Simmons *et al.*, *Science* **334**, 814 (2011); 10.1126/science.1213841.
3. R. H. Silverman *et al.*, *Science* **334**, 176 (2011); 10.1126/science.1212182.
4. J. Cohen, *ScienceInsider* (4 October 2011); http://scim.ag/_Mikovits.

NATURE NEWS BLOG

Another XMRV study retracted

28 Dec 2011 | 13:23 GMT | Posted by Heidi Ledford | Category: Biology & Biotechnology, Health and medicine

Five days after *Science* withdrew a controversial study linking chronic fatigue syndrome (CFS) to a virus, another research paper on the topic has been retracted.

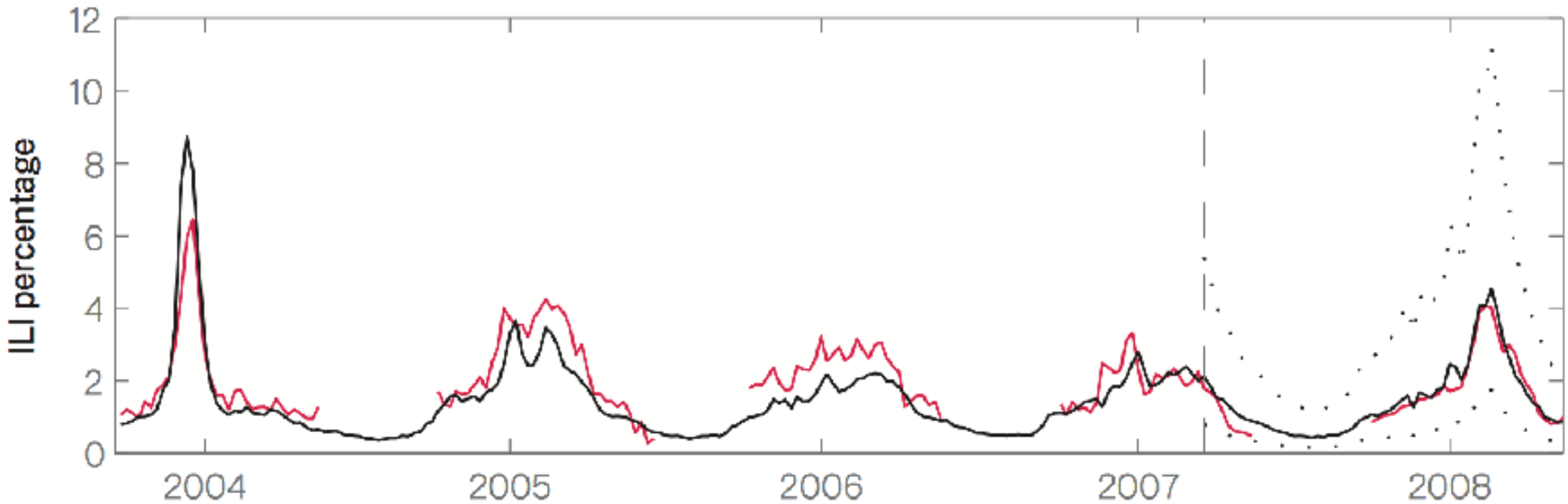
This study, retracted by the *Proceedings of the National Academy of Science* on 27 December, was widely seen as supportive of the proposed link between a virus and CFS, so its retraction is another blow to those who back that connection. It had found that people with chronic fatigue syndrome were more likely than healthy controls to harbour MLV-like viruses in their blood. (MLV stands for murine leukaemia virus; and the virus singled out for attention by the *Science* paper, XMRV, is a member of the MLV-like family of retroviruses).

Publication of the PNAS study had been delayed last year, while it underwent a last-minute additional round of review (see 'Chronic fatigue findings were held back'). It was eventually published in August 2010 (see 'Delayed chronic fatigue syndrome paper to be published'), but even then the authors acknowledged that they were unable to provide the additional evidence requested by the last panel of reviewers.

Why do we need causes?

- Prediction
- Explanation
- Intervention

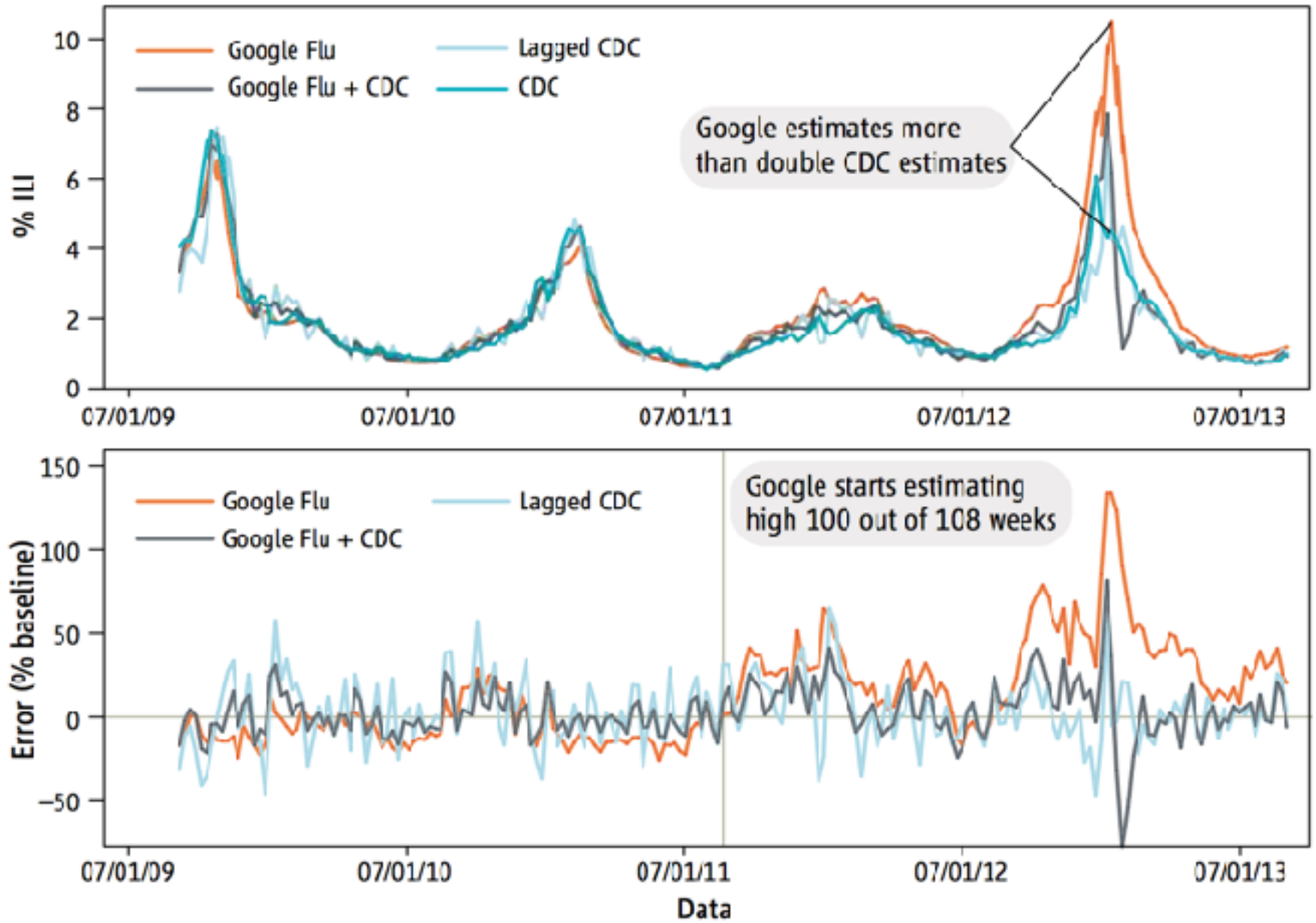
Google flu



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]

[1]Google Inc. [2]Centers for Disease Control and Prevention

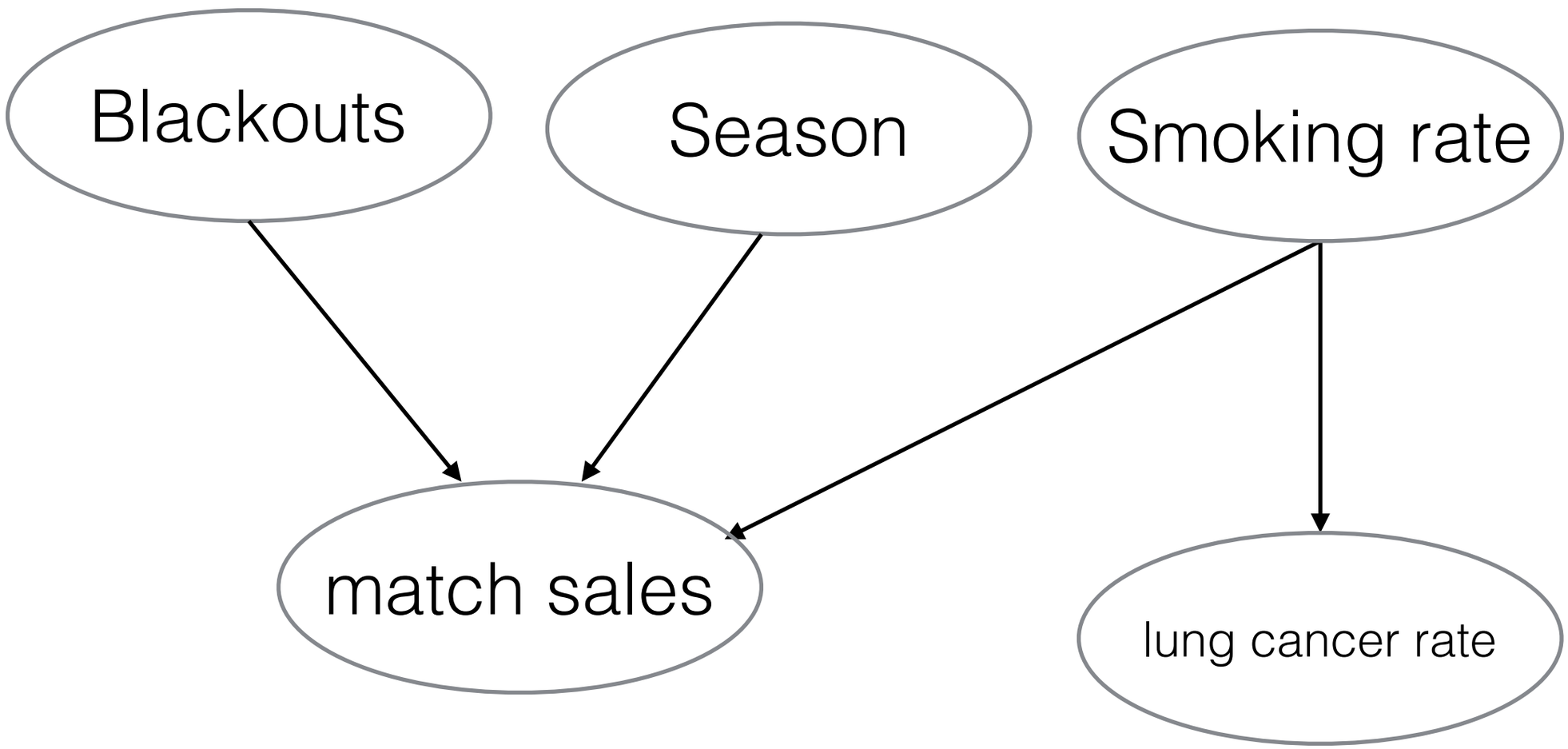


BIG DATA

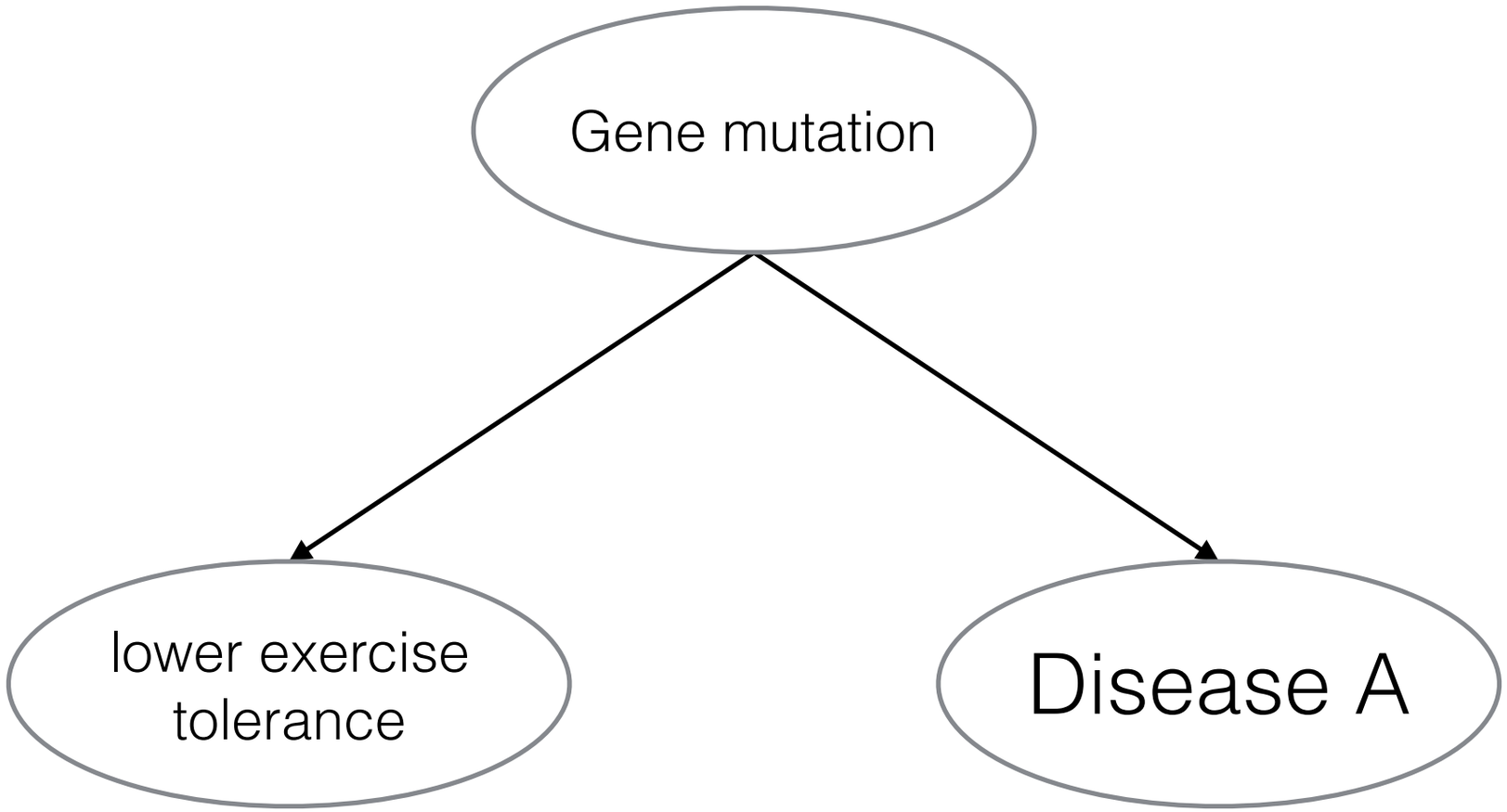
The Parable of Google Flu: Traps in Big Data Analysis

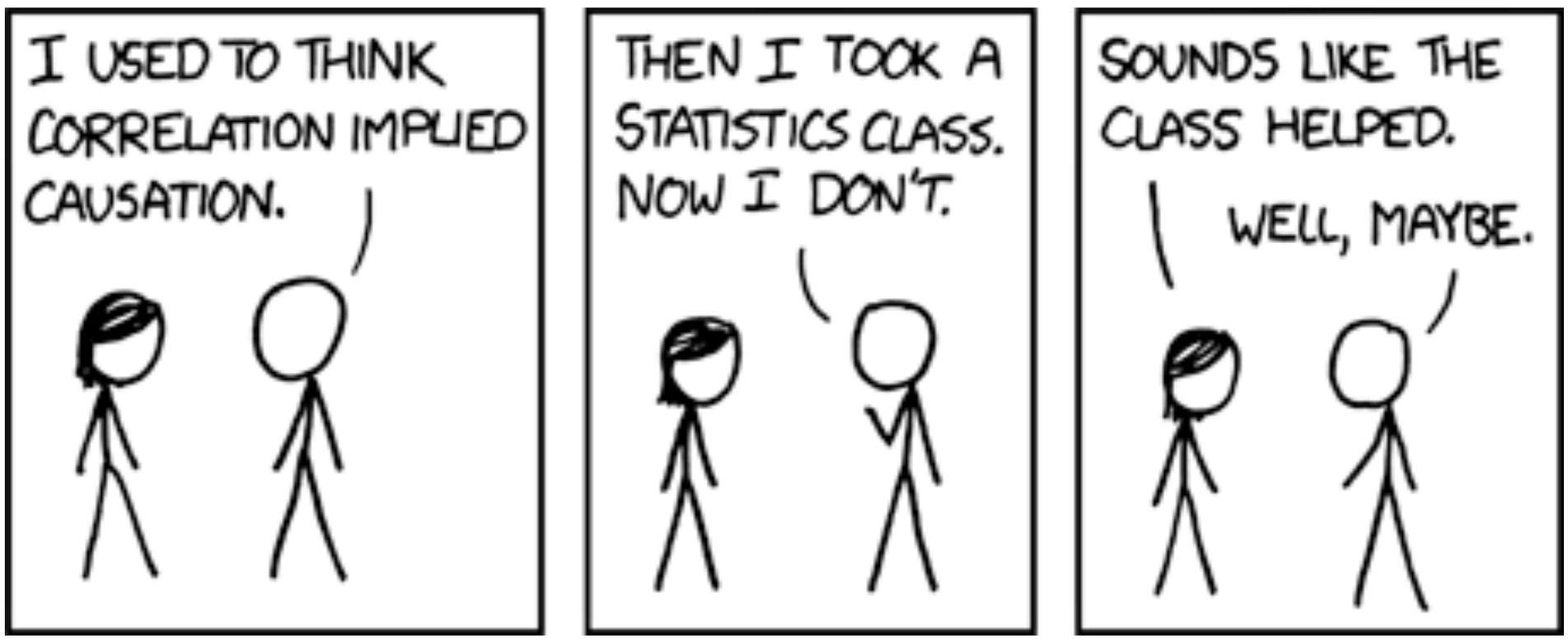
David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

Prediction


Blackouts
Season
Smoking rate
match sales
lung cancer rate

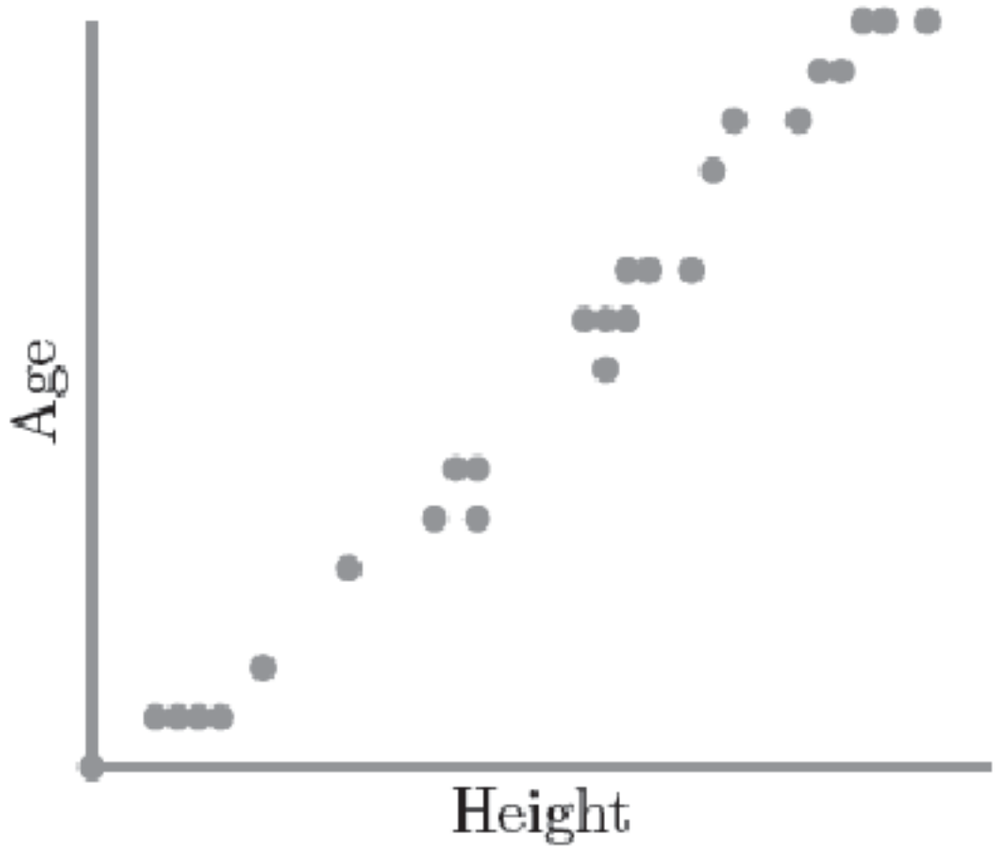
Prediction, continued



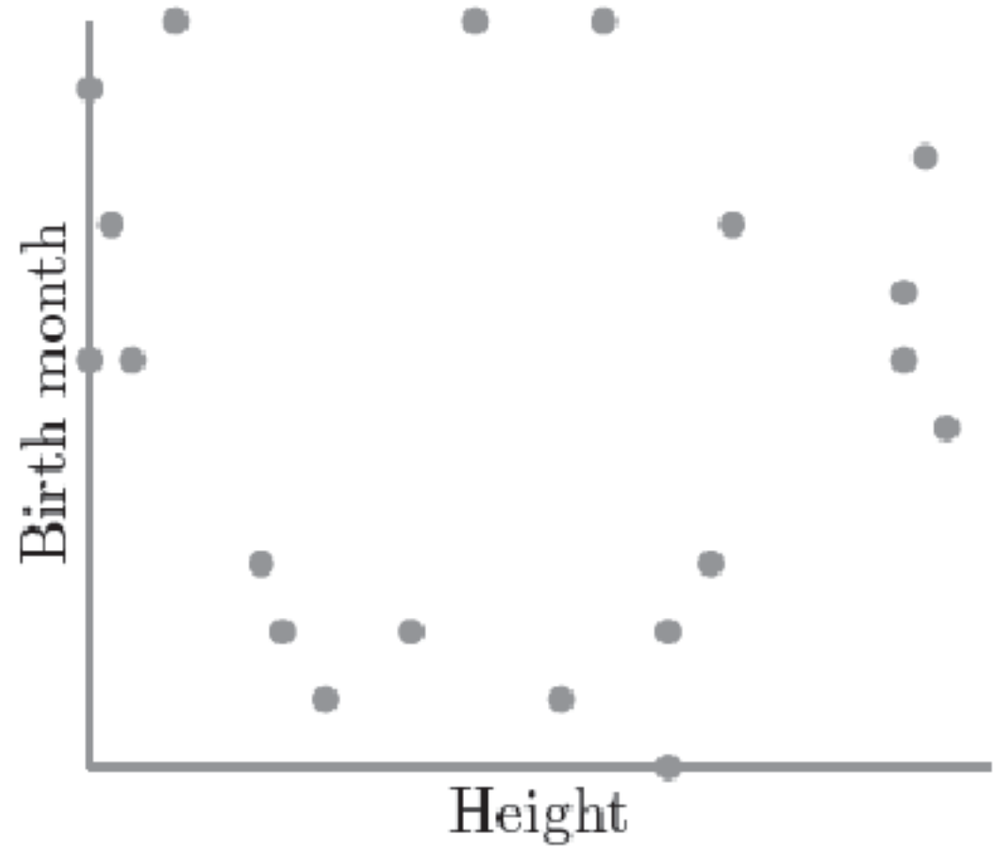


http://xkcd.com/552/

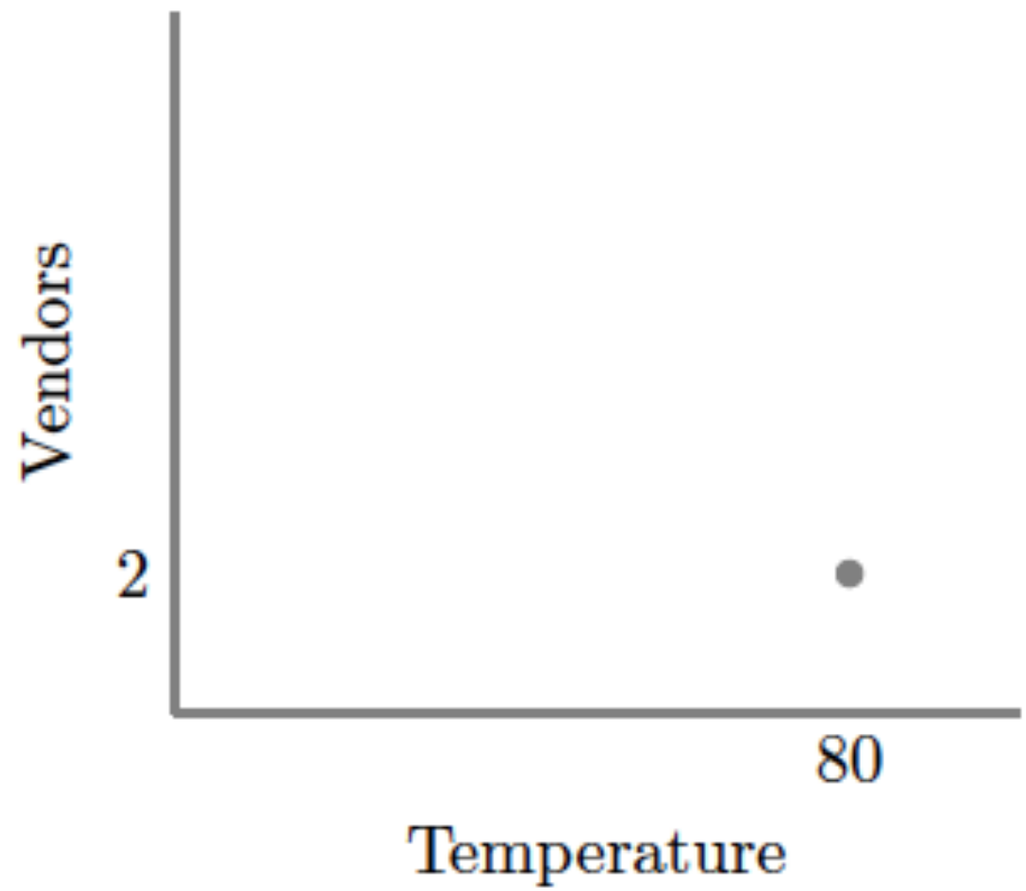
Correlation



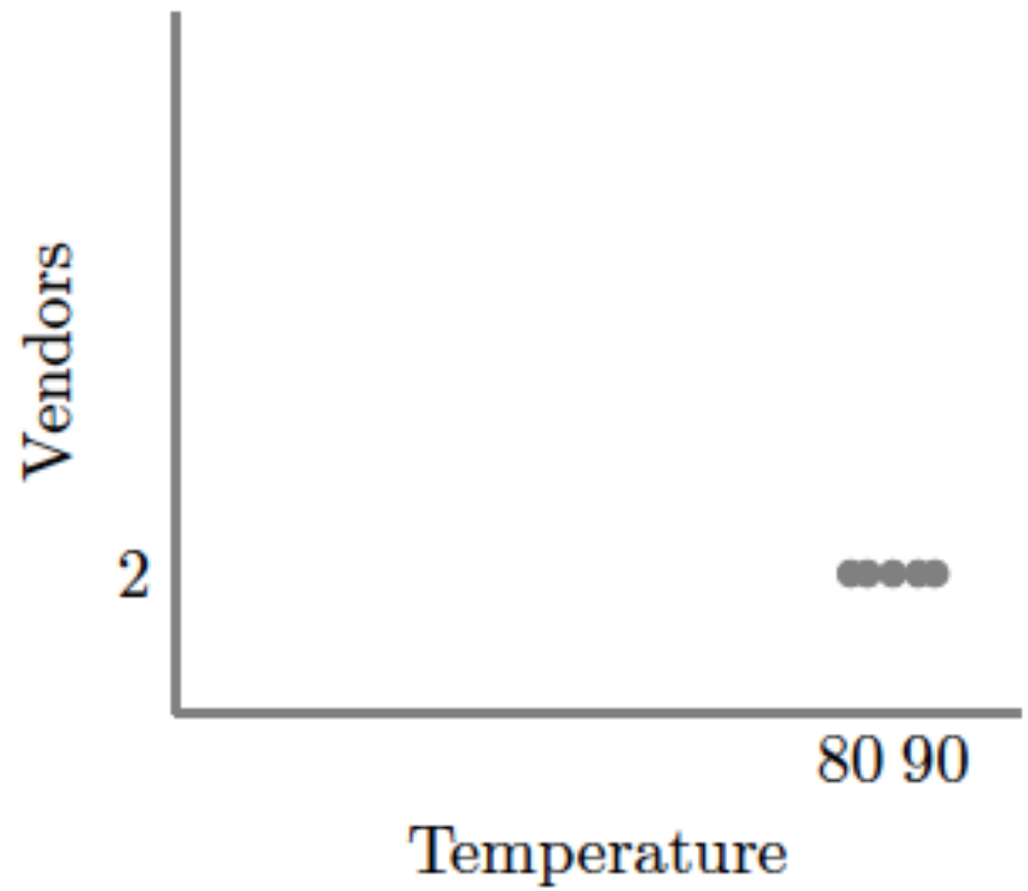
(a) Positively correlated



(b) Uncorrelated



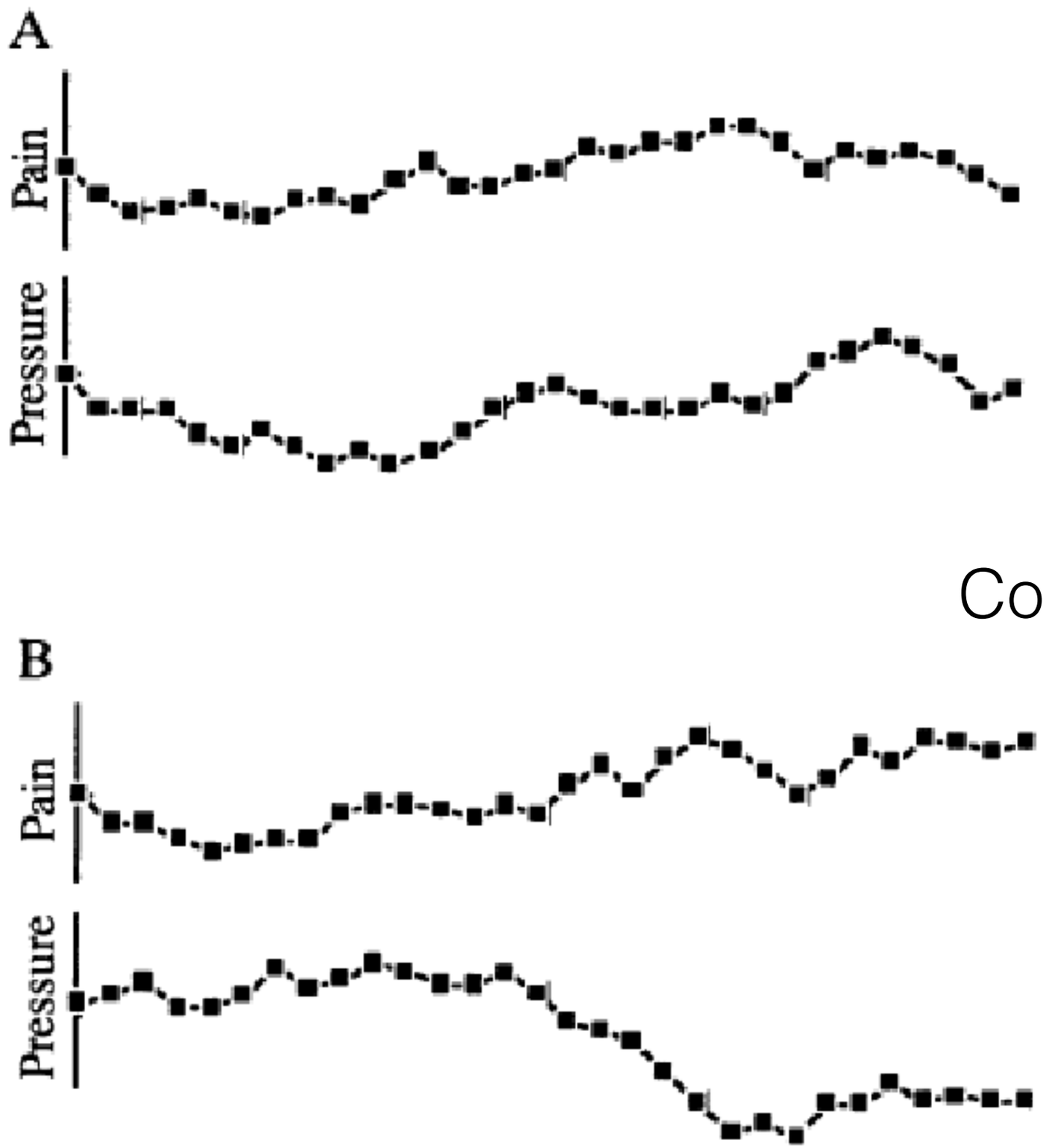
(a)



(b)

Correlations abound

- High HDL is related to lower heart disease
- Height and age
- Tides and traffic on the west side highway
- XMRV and CFS



Correlation = 0!

Redelmeier DA, Tversky A (1996) On the belief that arthritis pain is related to the weather. Proceedings of the National Academy of Sciences 93(7):2895-2896.

Correlation and Causation

- What's a correlation?
  - Relatedness of variables across samples or time
- Common measure: Pearson's correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Measuring correlation

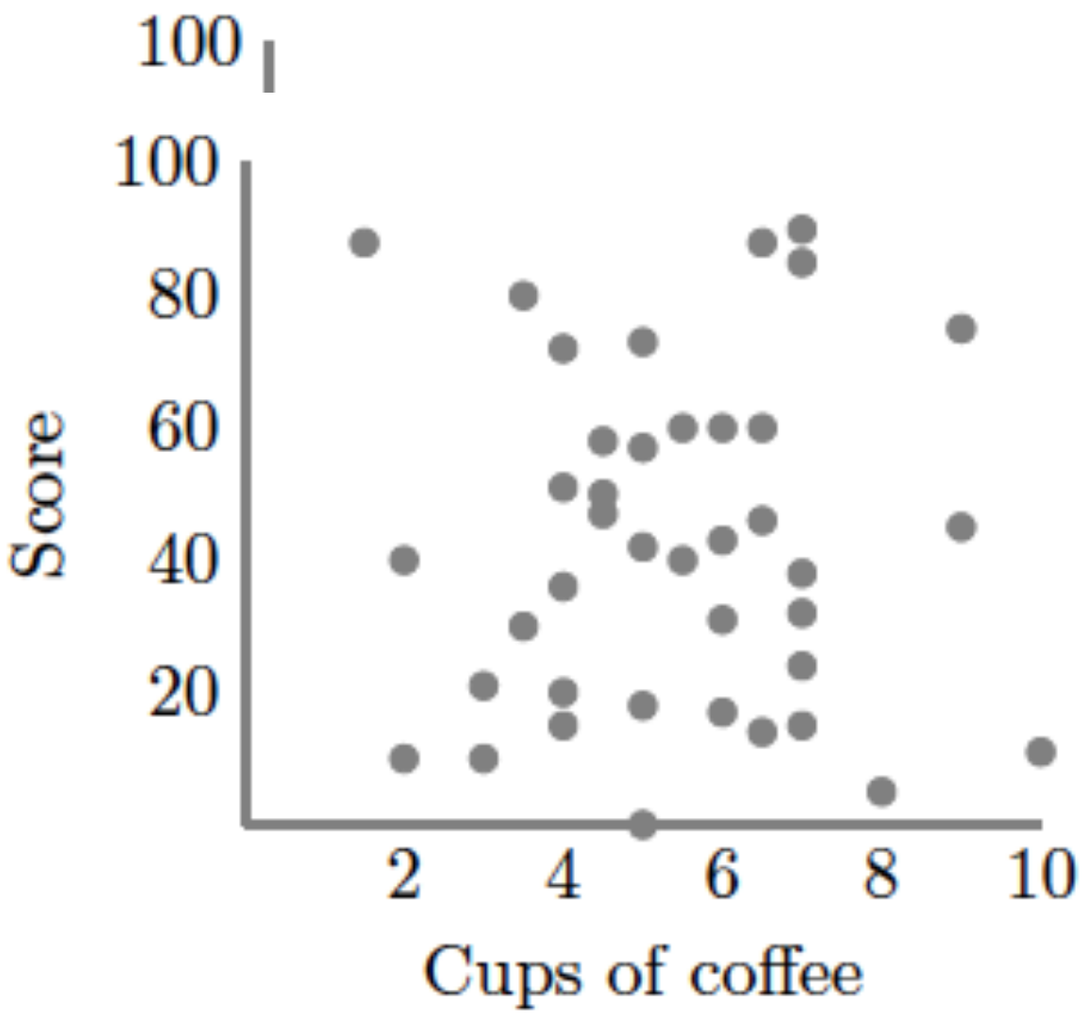
X	Y
0	1
1	1
3	3
4	5
5	5

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

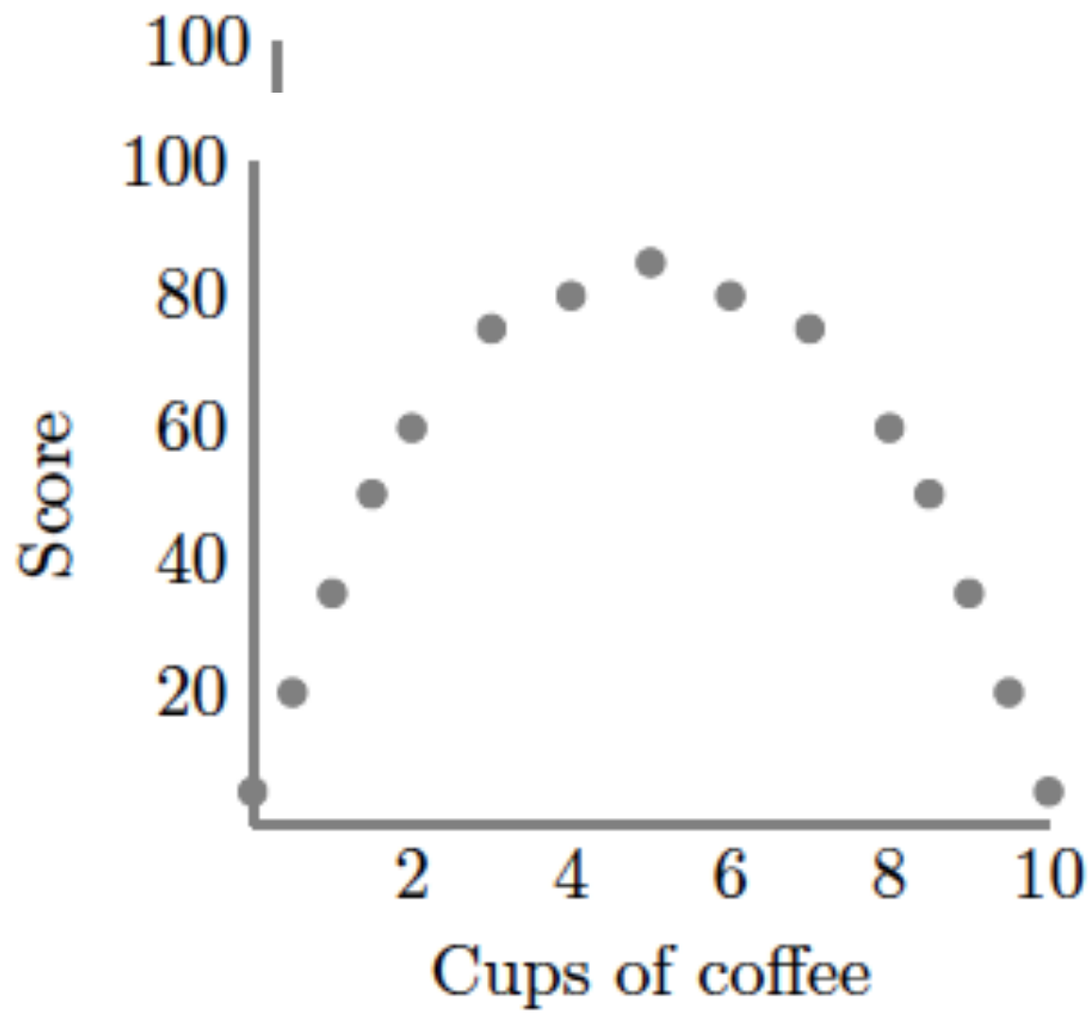
Cups of coffee (X) vs. correct test answers (Y)

$$r = (5.2 + 3.2 + 0 + 2.8 + 4.8) / 4.1473 \times 4$$

$$r = 0.9645$$

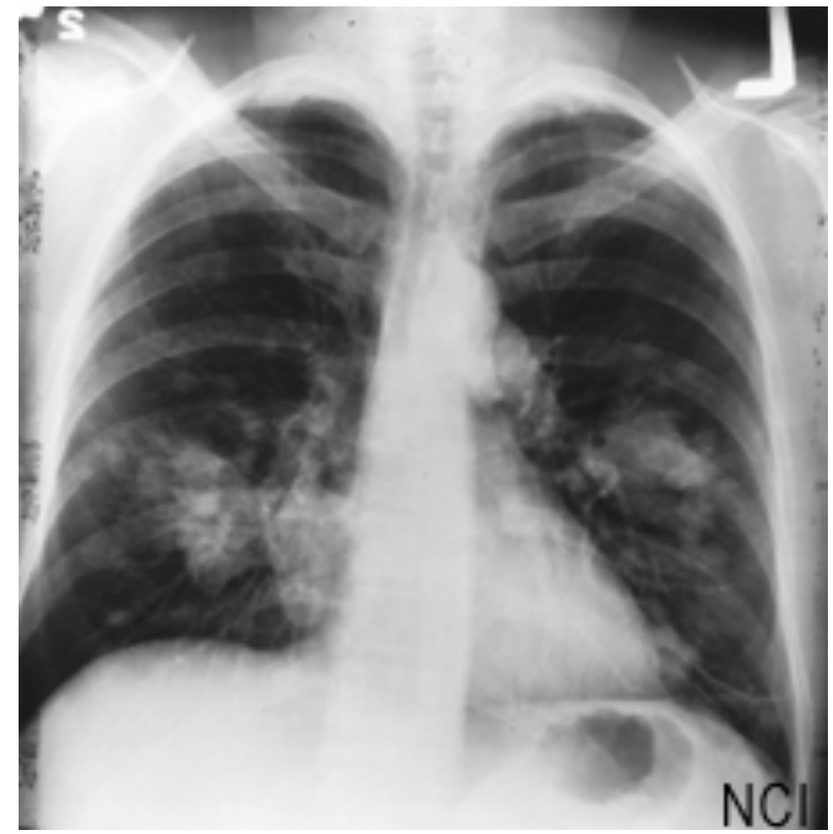


(e) No correlation (r=0.000)



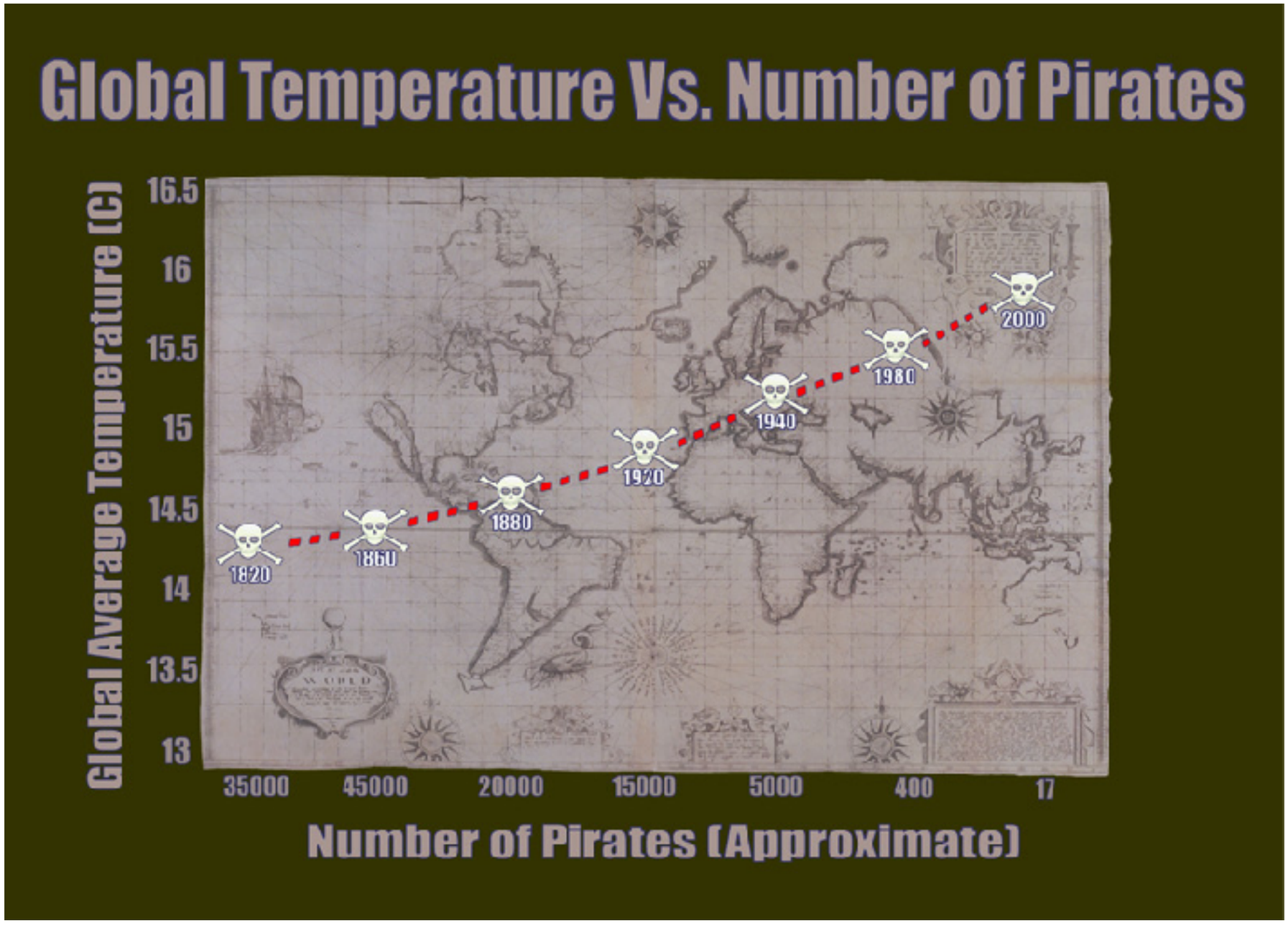
(f) Nonlinear relationship (r=0.000)

Hidden Common Causes



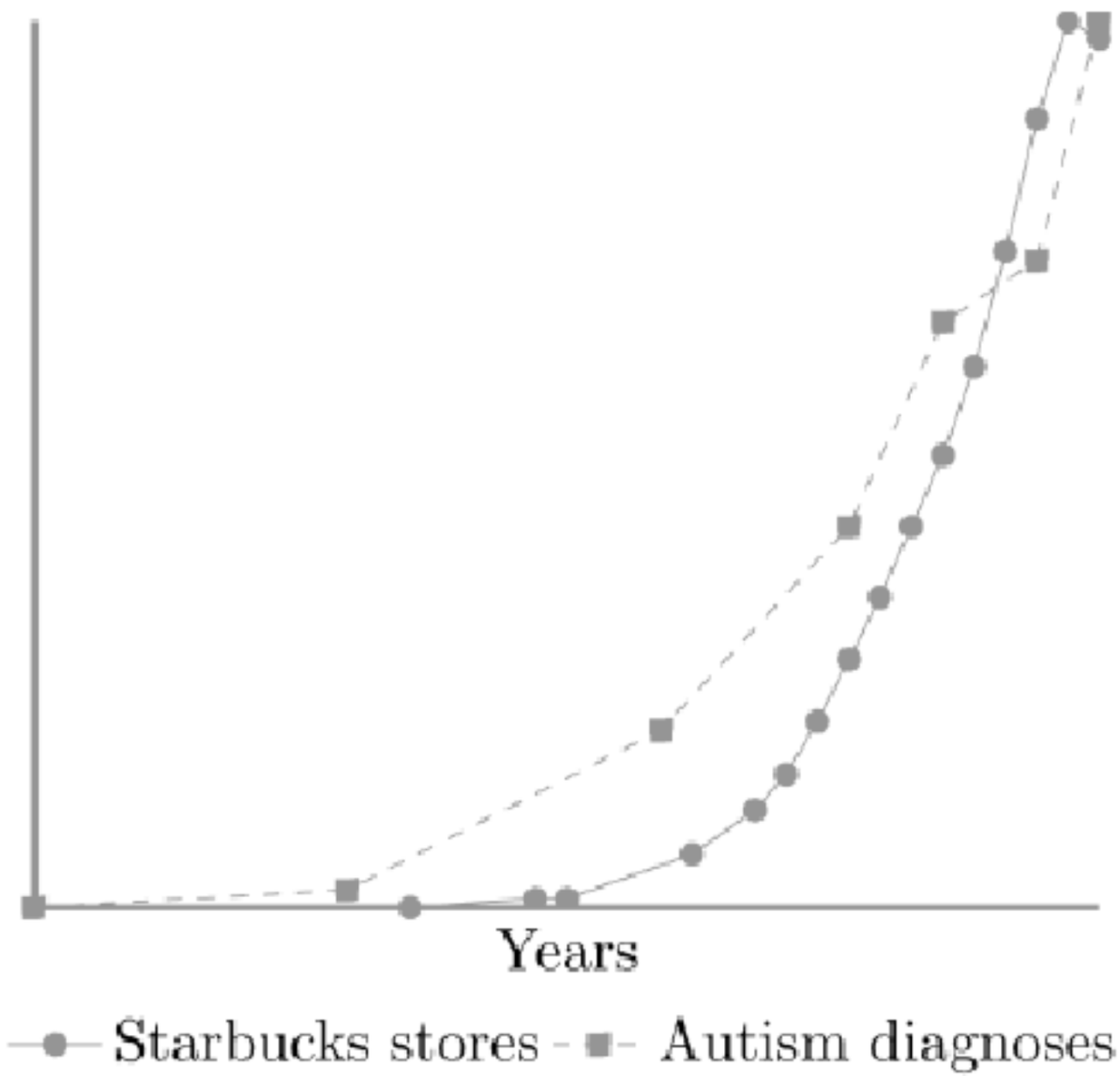
Correlation with some causation

Nonstationary time series



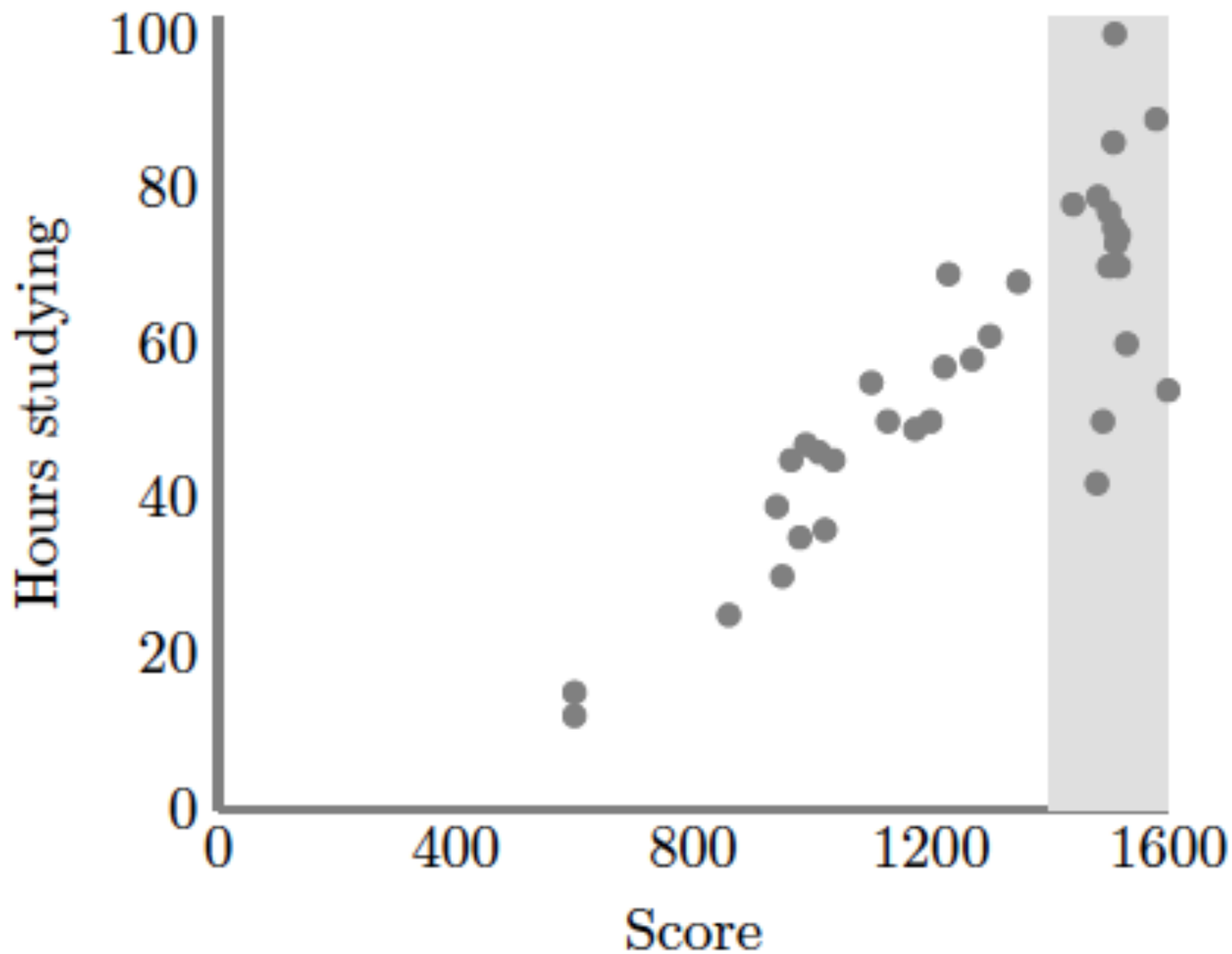
http://bama.ua.edu/~sprentic

Nonstationary time series

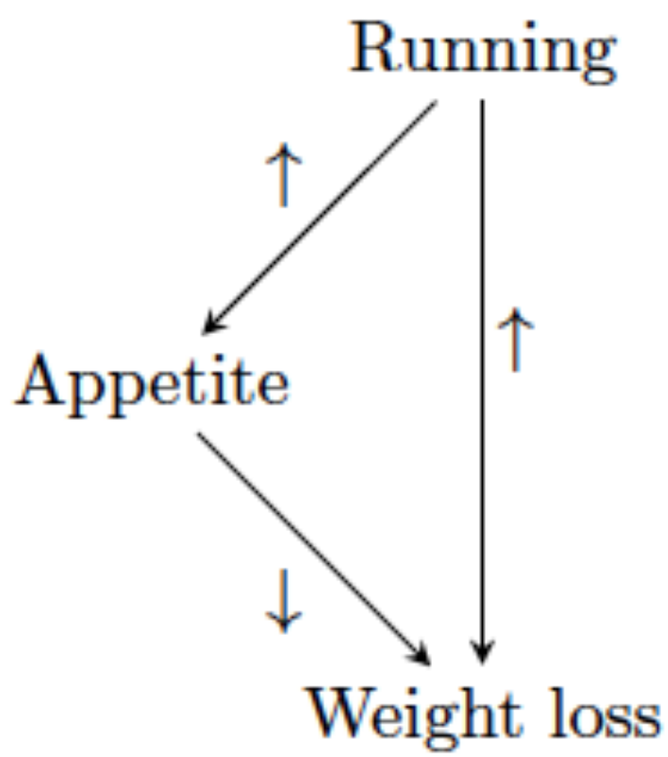


S. Kleinberg. (2015) Why: A Guide to Finding and Using Causes. O'Reilly Media.

Restricted range



Canceling out



Simpson's paradox

Treatment		
	Dead	Alive
A	85	215 (72%)
B	59	241 (80%)
Total	144	456

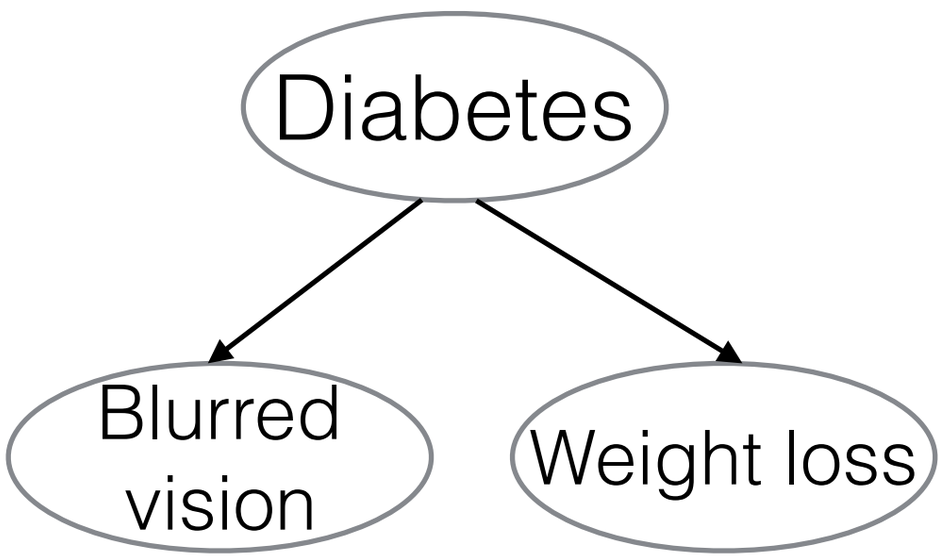
Causation without correlation: Simpson's paradox

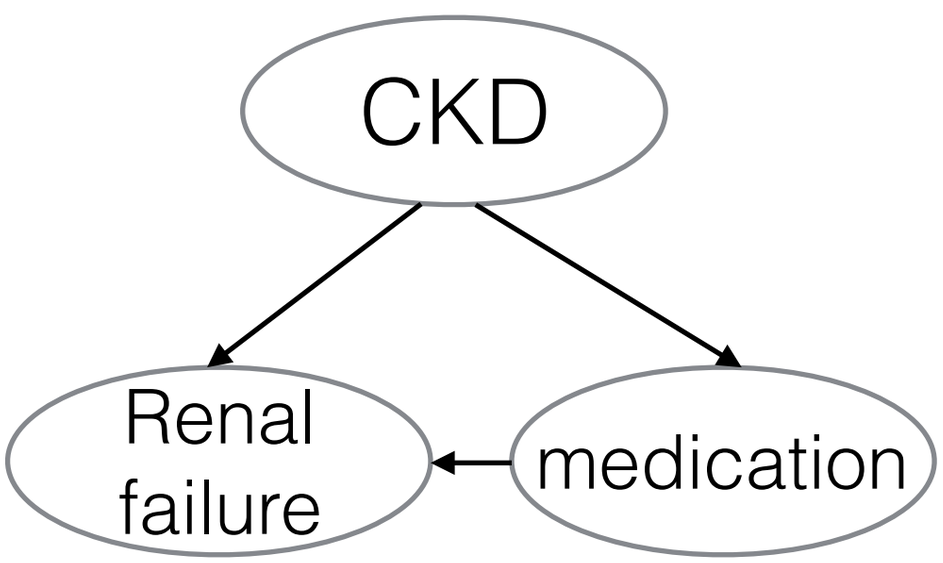
	Men		Women		Combined	
Treatment	Dead	Alive	Dead	Alive	Dead	Alive
A	80	120 (60%)	5	95 (95%)	85	215 (72%)
B	20	20 (50%)	39	221 (85%)	59	241 (80%)
Total	100	140	44	316	144	456

Baker SG, Kramer BS (2001) Good for women, good for men, bad for people: Simpson's paradox and the importance of sex-specific analysis in observational studies. Journal of women's health & gender-based medicine 10: 867-872

Explanation (1)

- Why are two variables related?





Explanation (2)

- General causes of illness vs. cause of a specific patient's illness

- Why did an event happen?
  - Why did a particular person develop lung cancer at age 42?
  - What led to the U.S. recession in 2007?
  - Is a stroke patient's secondary brain injury due to seizures?

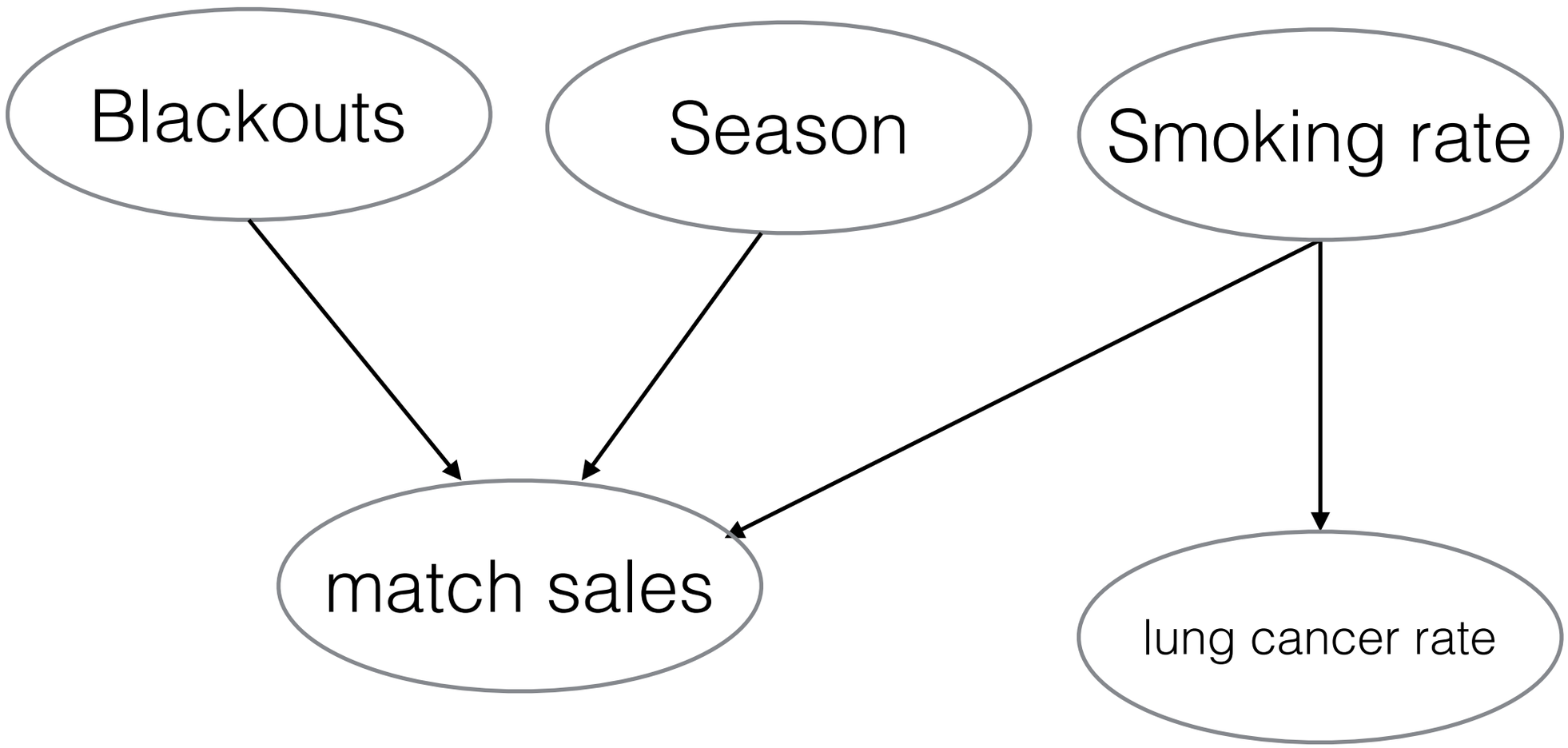
Automating explanation

- Methods for finding causes from data, but what about explaining events?
- Practical problem, but challenging
  - Information incomplete
  - Where do explanations come from?
  - General and singular can differ

Intervention

- Why do we need causes to take action?
  - Buying stocks
  - Taking vitamins
  - Decreasing sodium to prevent hypertension
- What happens if we intervene on a correlated factor?

Using causes to guide intervention



Using interventions to find causes

- Does playing violent video games make children violent?
- Does too little sleep increase mortality rate?
- Does medication cause side effects?

Recap

A cause

….allows prediction of future events

….can explain connections

….can explain occurrences

….enables interventions to prevent/produce outcomes

BUT! Not every cause does and the story is more complicated

Causality & time

Interpretation in the absence of time

- Compare:
  - A. Smoking causes lung cancer with probability ≈ 1 after 90 years
  - B. Smoking causes lung cancer with probability = ½ in less than 10 years.

Probability

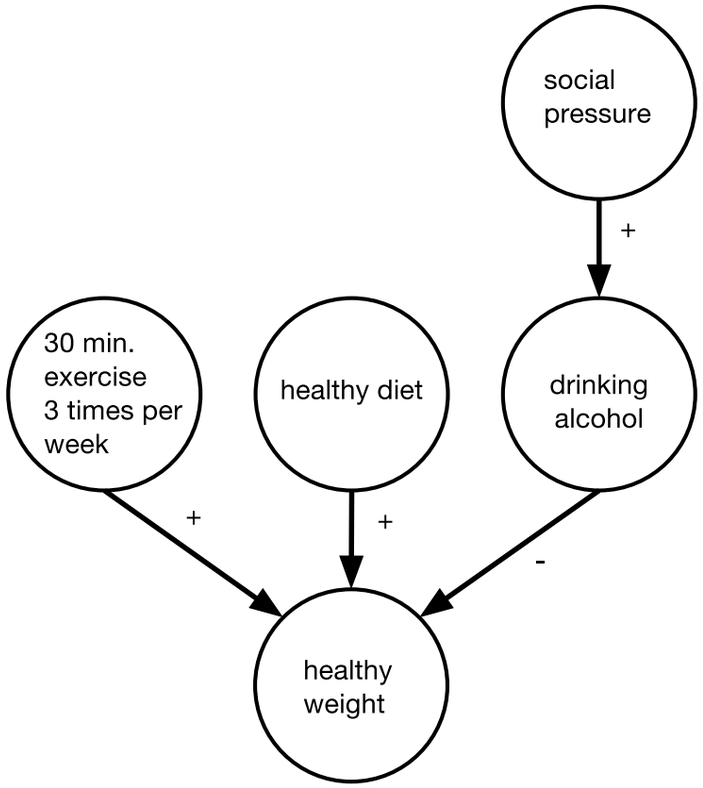
- Few relationships deterministic
  - Relationship vs. limits of knowledge
- Understanding risk
- Choosing intervention target
  - Medication efficacy vs. side effects
- Can also measure strength of relationship

Complexity

- Interactions
  - Smoking + lung cancer: other conditions affecting probability and time: genetics, environment
- Planning effective interventions
  - Political Speeches
- Durations, conjunctions, sequences of events

Decision-making

- Should I have oatmeal or a fruit salad?
- When is the best time to run?
- How should I invest my retirement savings?



Course overview

Weeks 1-3	Weeks 4-8	Weeks 9-12	Weeks 13-14
What Is a cause?	How can we find causes?	When can we find causes?	Projects + Special topics

Three main questions

- What is a cause?
  - Theories of what distinguishes them from correlations and how we can identify them
- How can we find causes?
  - Features of causes that allow us to learn about them
- When can we infer causes?
  - Methods for inference from data
  - Study design
  - Applications to challenging cases

- **Causal inference**: finding causal relationships from data

- **Causal explanation**: finding reason for a specific event that occurs at a particular time and place

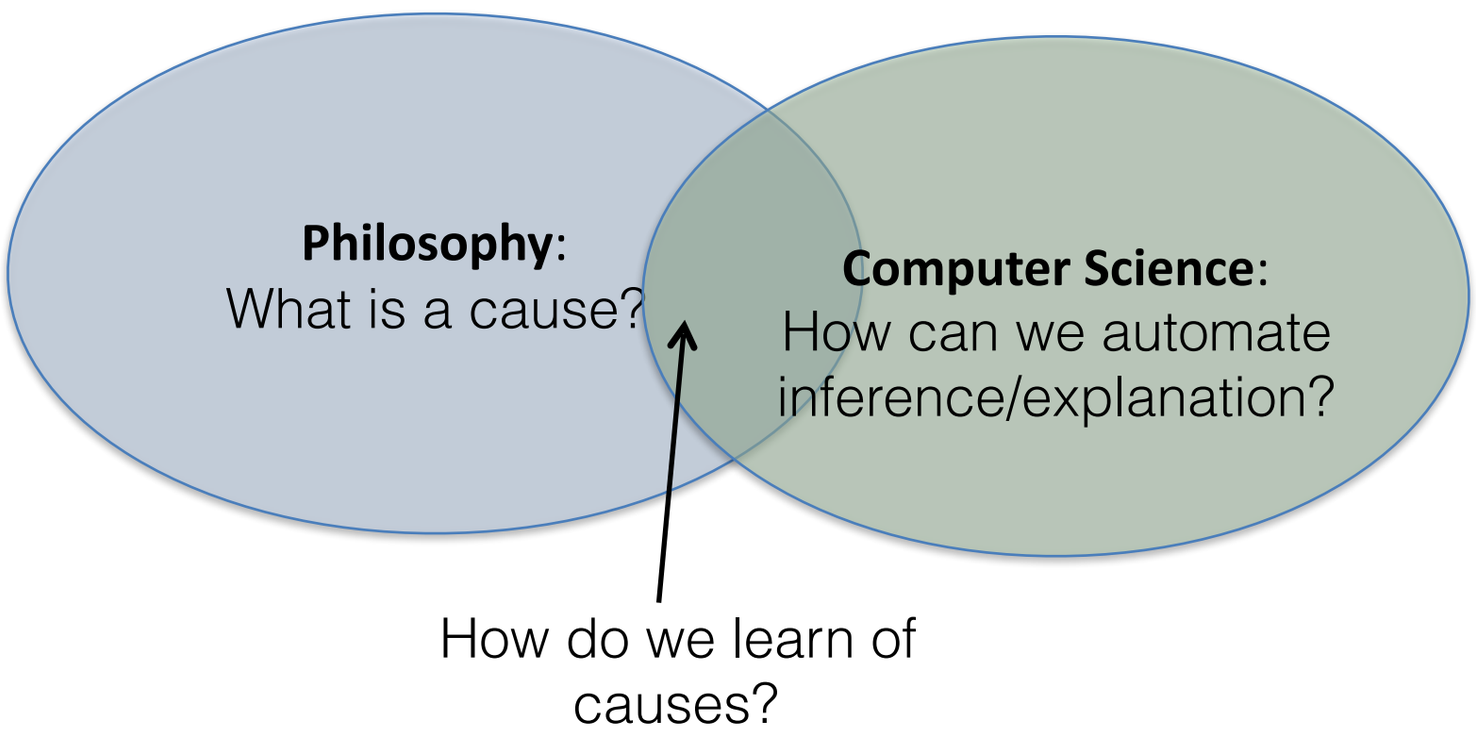
Mo data, mo problems

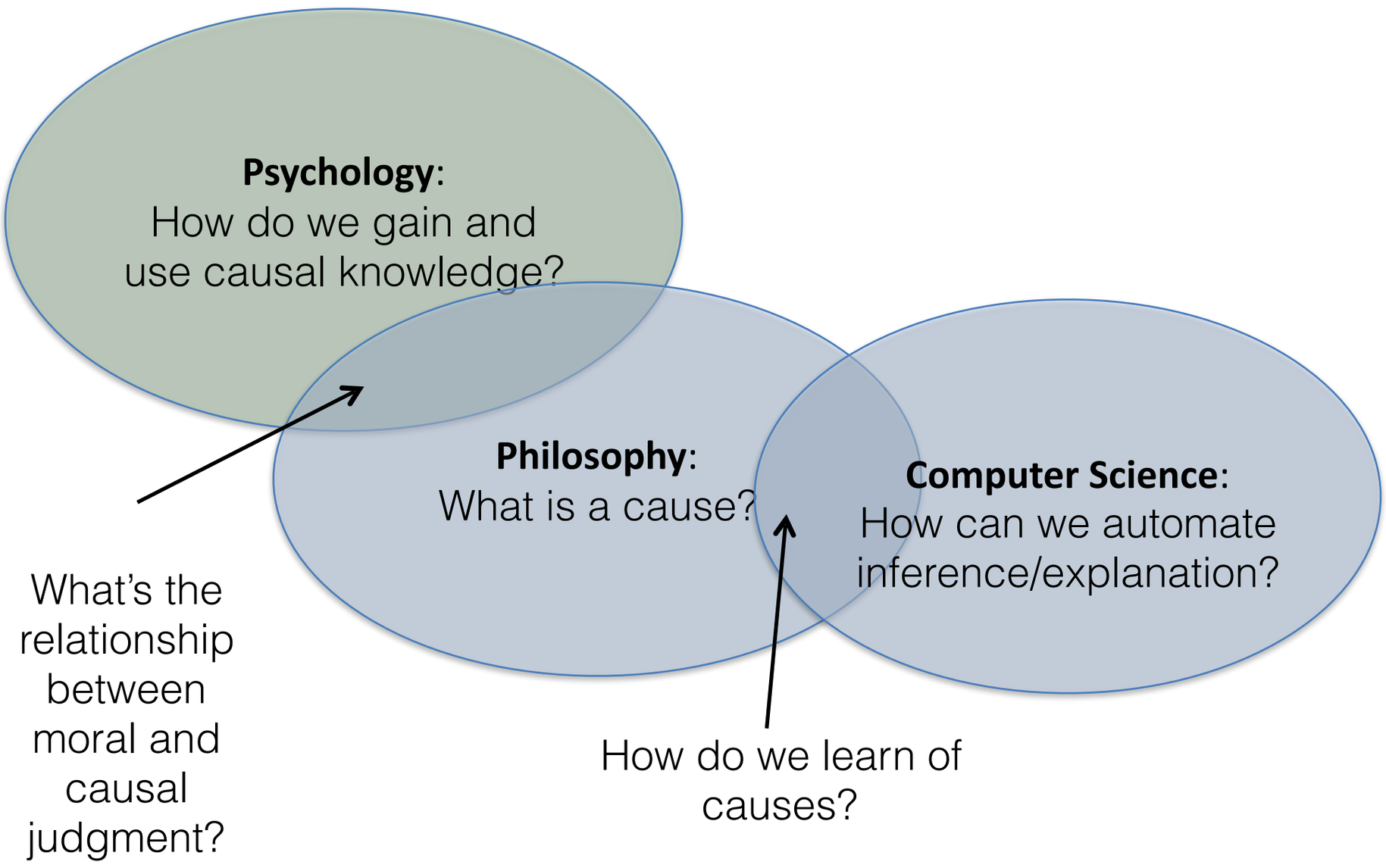
- Big ≠ good
- Uncertainty
- Selection bias
- Signal:noise
- Interpretation
- Time
- Ground truth

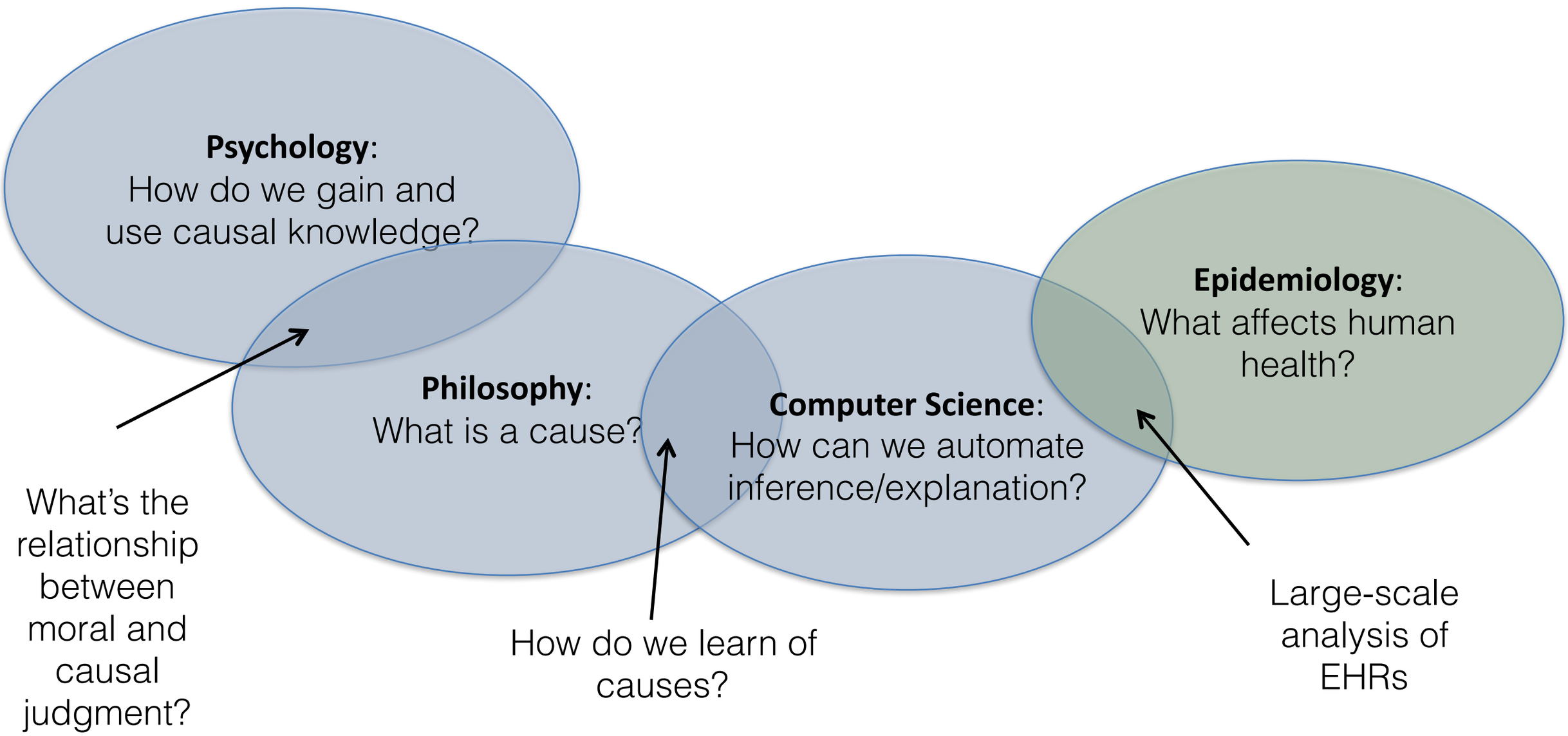
Many unsolved problems

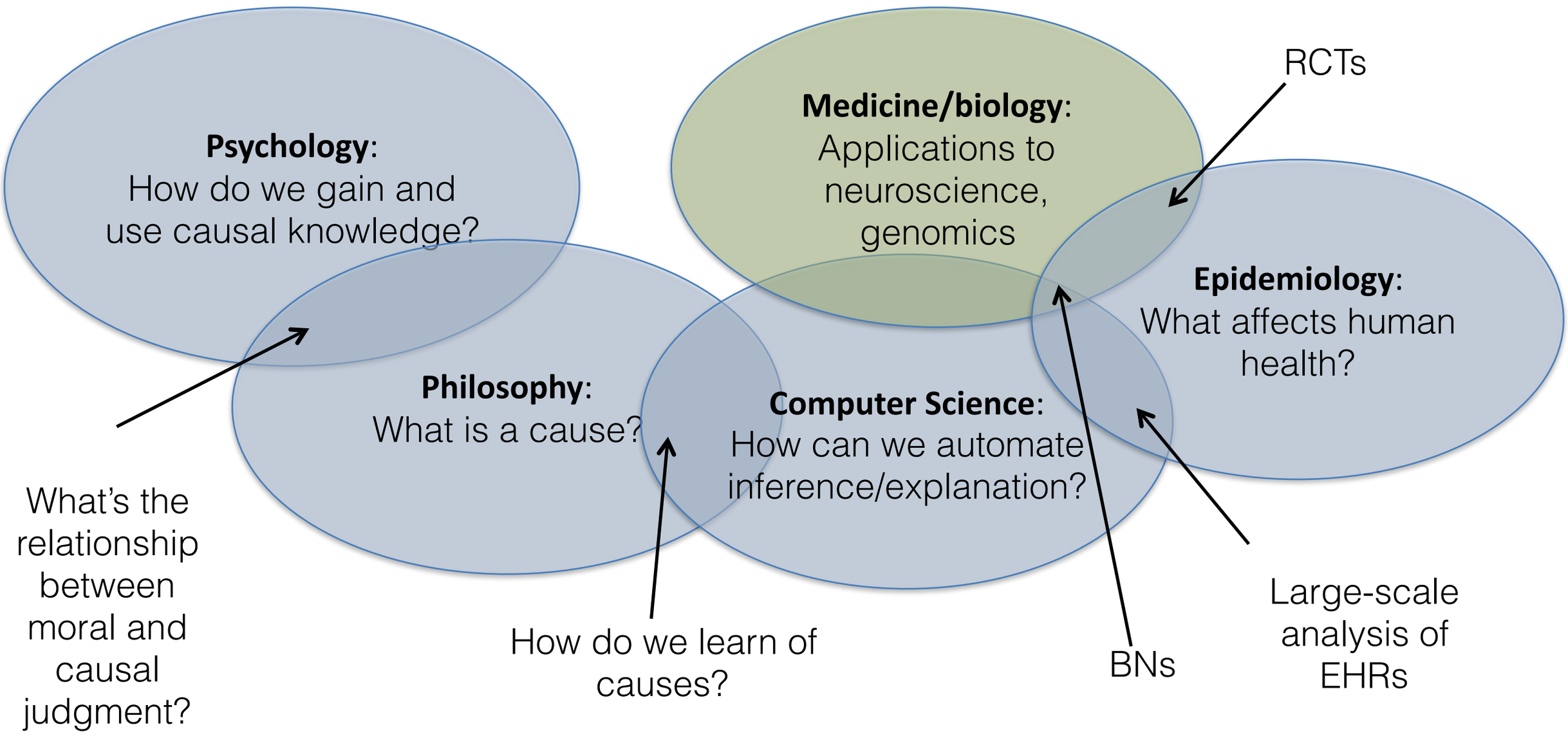
- Hidden variables
- Relationships that change over time
- Hypothesis generation
- Estimating uncertainty
- Testing assumptions
- Automating explanation

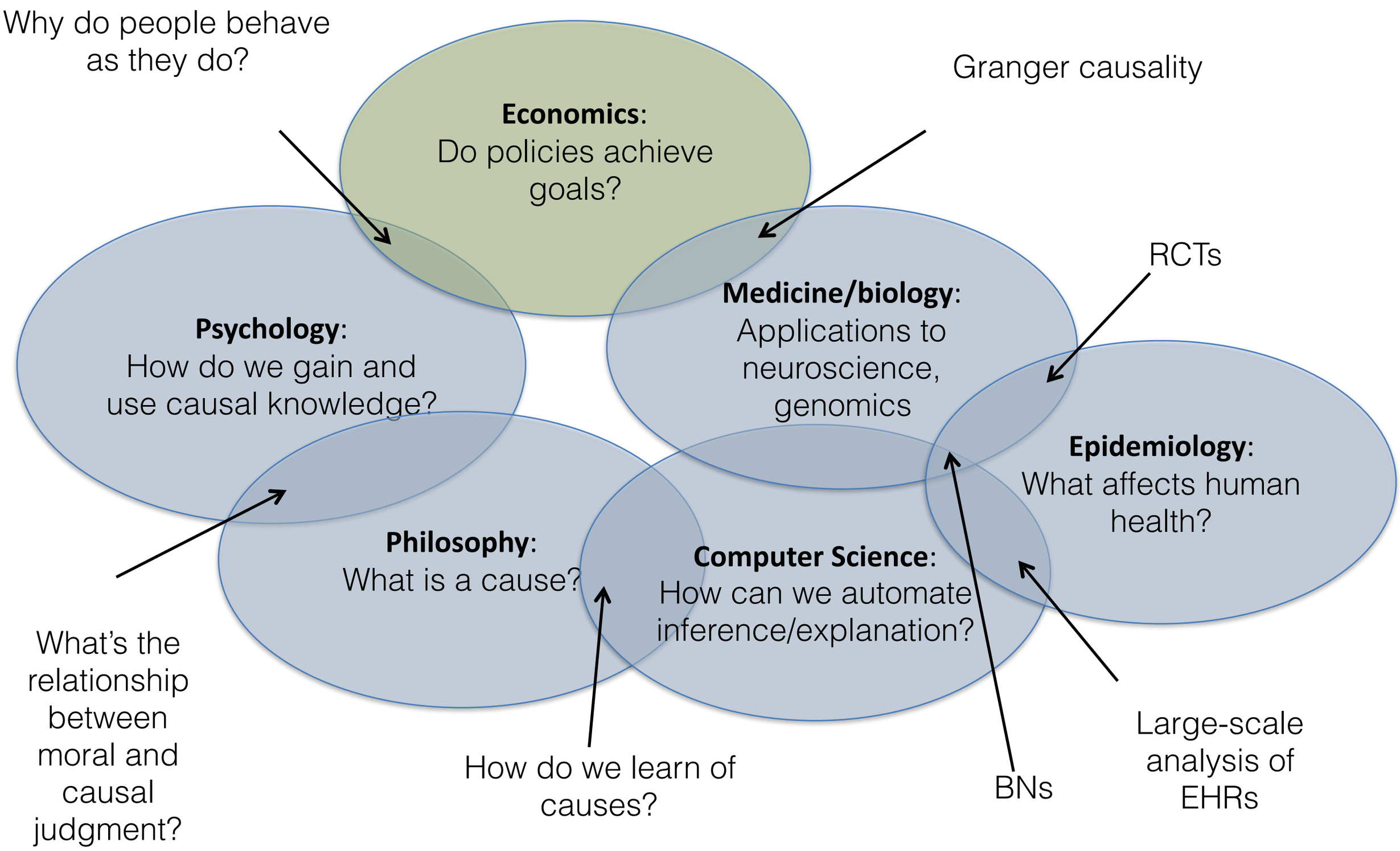
Philosophy:
What is a cause?


Philosophy:
What is a cause?
Computer Science:
How can we automate inference/explanation?
How do we learn of causes?


Psychology:
How do we gain and use causal knowledge?
Philosophy:
What is a cause?
Computer Science:
How can we automate inference/explanation?
What's the relationship between moral and causal judgment?
How do we learn of causes?


Psychology:
How do we gain and use causal knowledge?
Philosophy:
What is a cause?
Computer Science:
How can we automate inference/explanation?
Epidemiology:
What affects human health?
What's the relationship between moral and causal judgment?
How do we learn of causes?
Large-scale analysis of EHRs


Psychology:
How do we gain and use causal knowledge?
Medicine/biology:
Applications to neuroscience, genomics
RCTs
Epidemiology:
What affects human health?
Philosophy:
What is a cause?
Computer Science:
How can we automate inference/explanation?
What's the relationship between moral and causal judgment?
How do we learn of causes?
BNs
Large-scale analysis of EHRs


Why do people behave as they do?
Granger causality
Economics: Do policies achieve goals?
Psychology: How do we gain and use causal knowledge?
Medicine/biology: Applications to neuroscience, genomics
RCTs
Epidemiology: What affects human health?
Philosophy: What is a cause?
Computer Science: How can we automate inference/explanation?
What's the relationship between moral and causal judgment?
How do we learn of causes?
BNs
Large-scale analysis of EHRs

Granger causality beyond economics

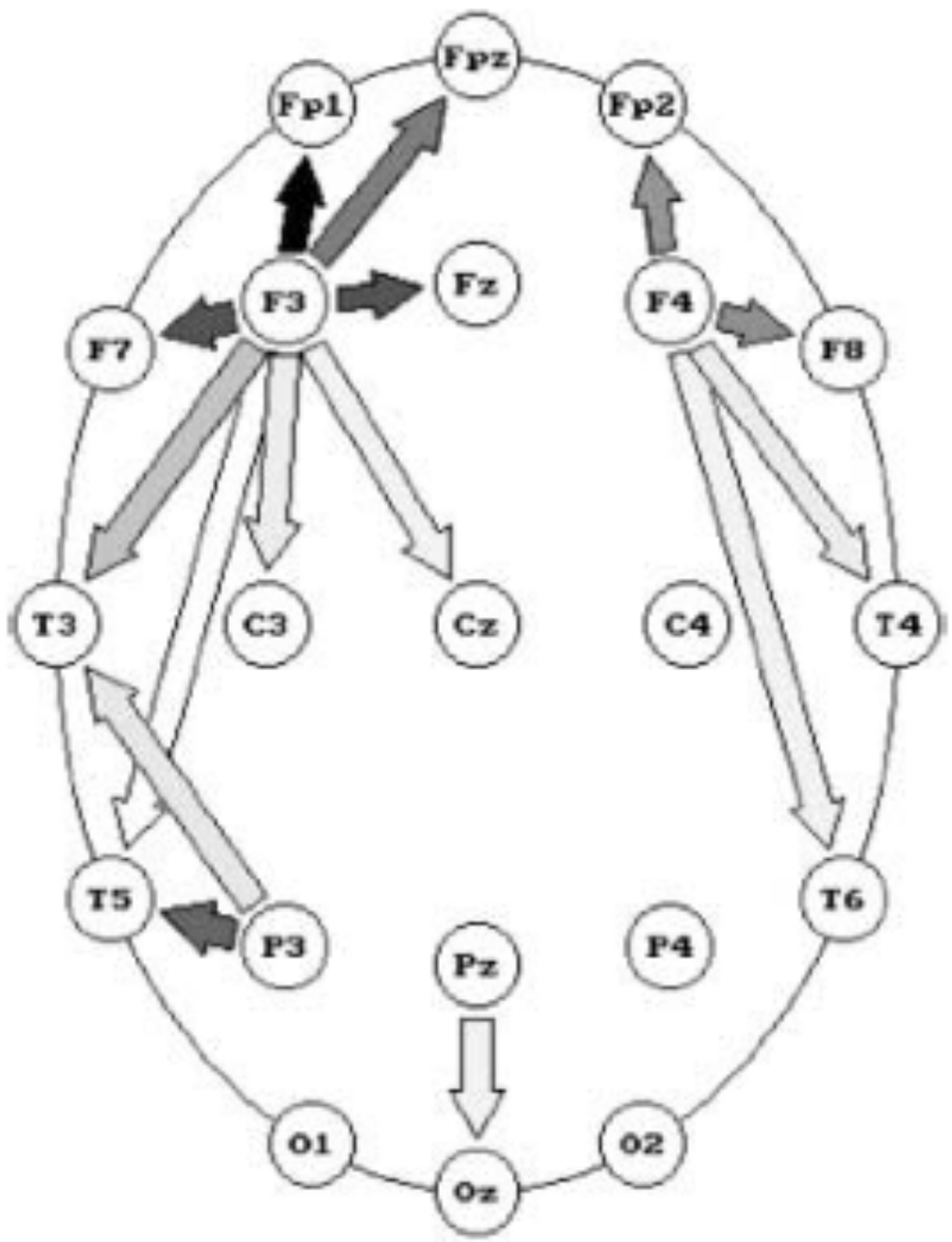
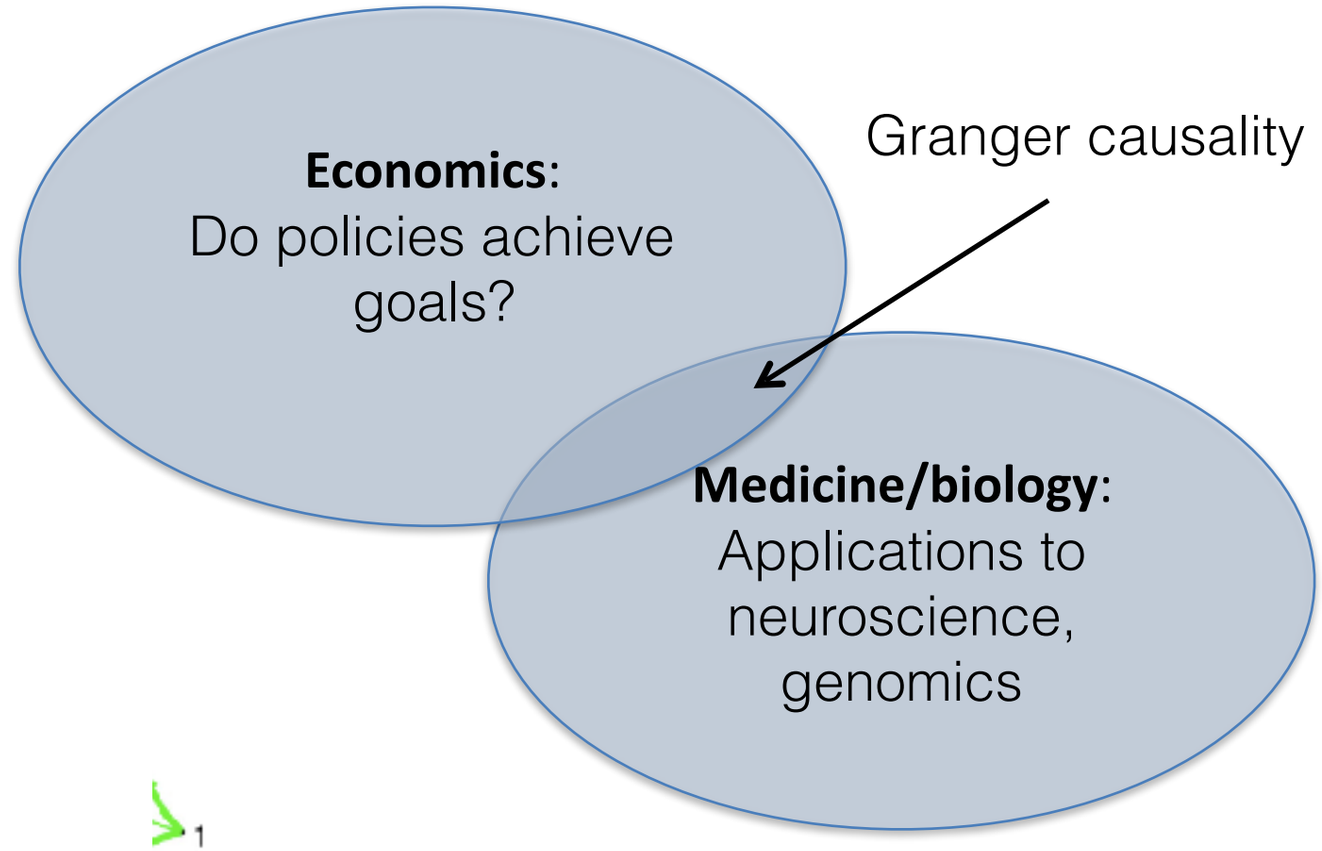


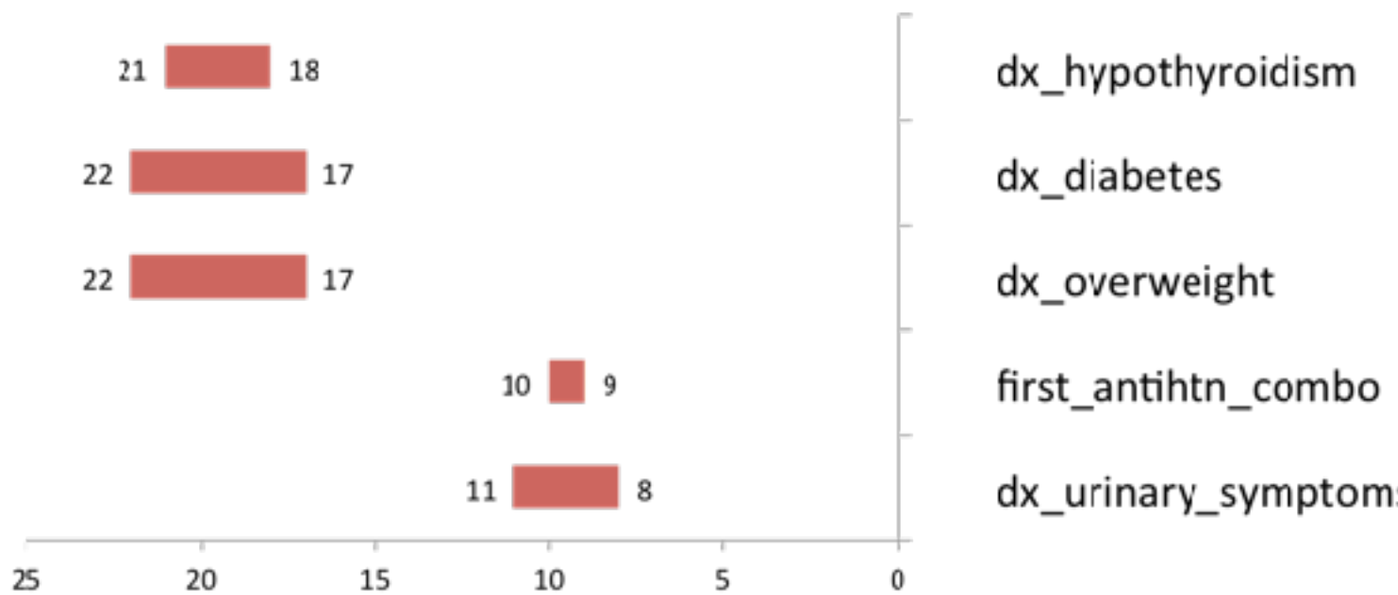
Fig. 10. Patterns of causal influences during stage 2 sleep

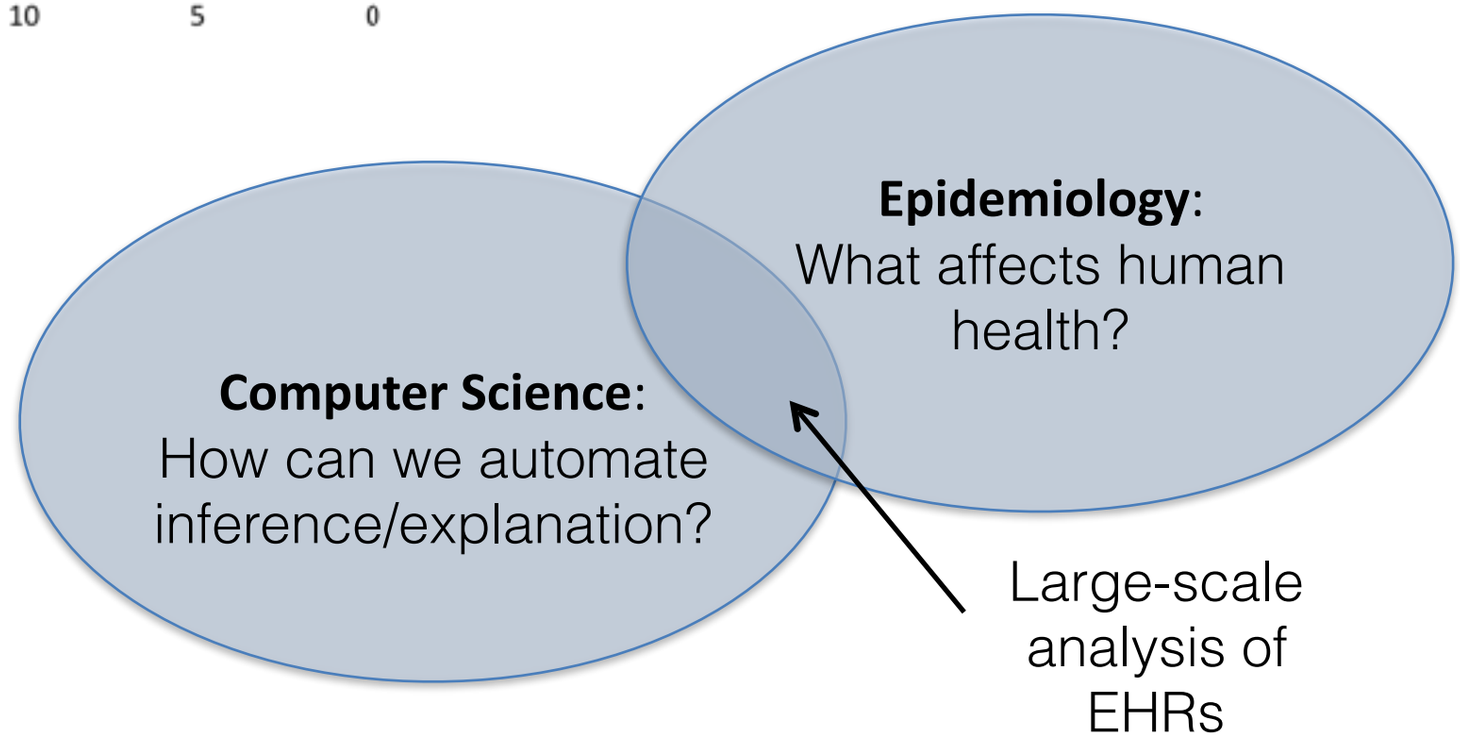


Kamiński M, Ding M, Truccolo WA, Bressler SL (2001) Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. Biological Cybernetics 85: 145-157.

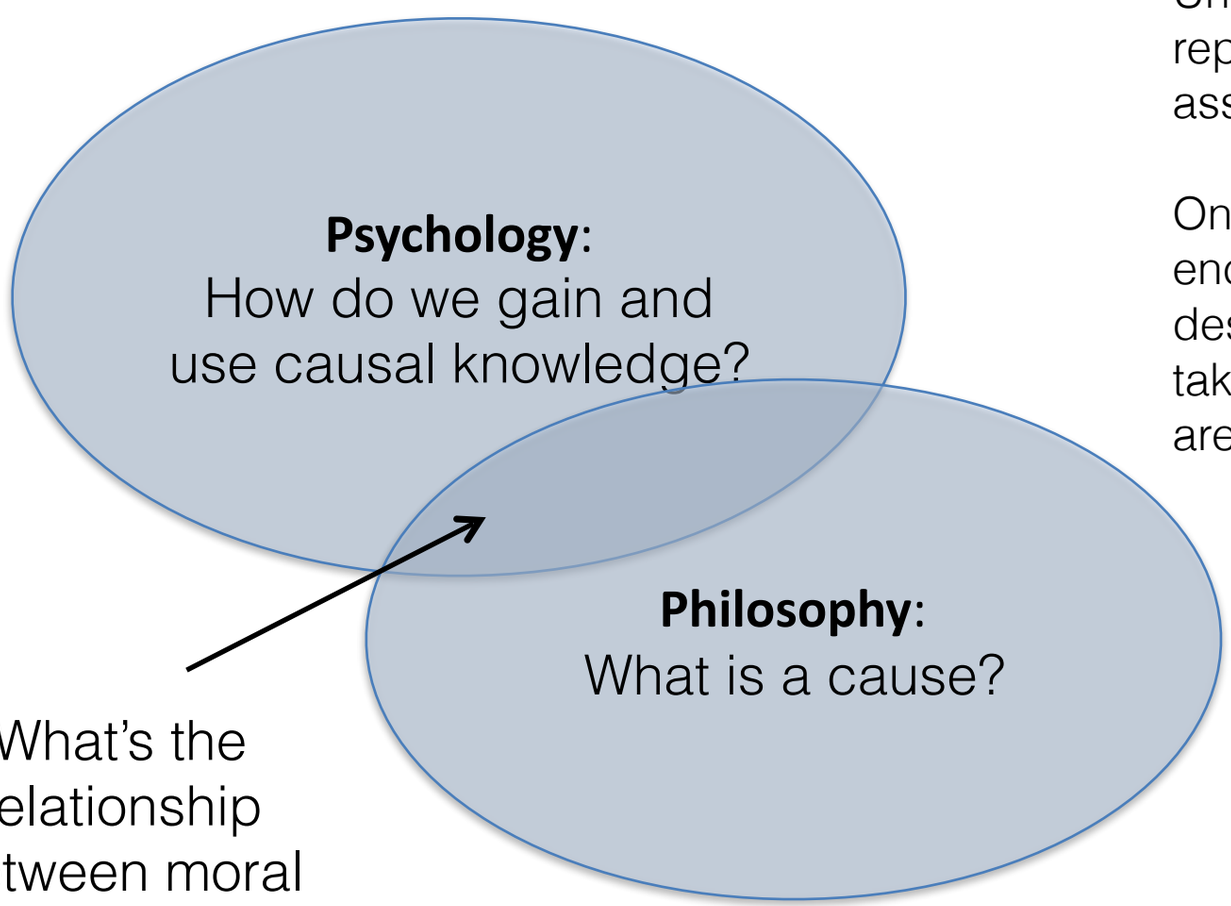
Seth AK (2010) A MATLAB toolbox for Granger causal connectivity analysis. J Neurosci Methods 186: 262-273.

EHR analysis





Causal judgment



What's the relationship between moral and causal judgment?

The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own.

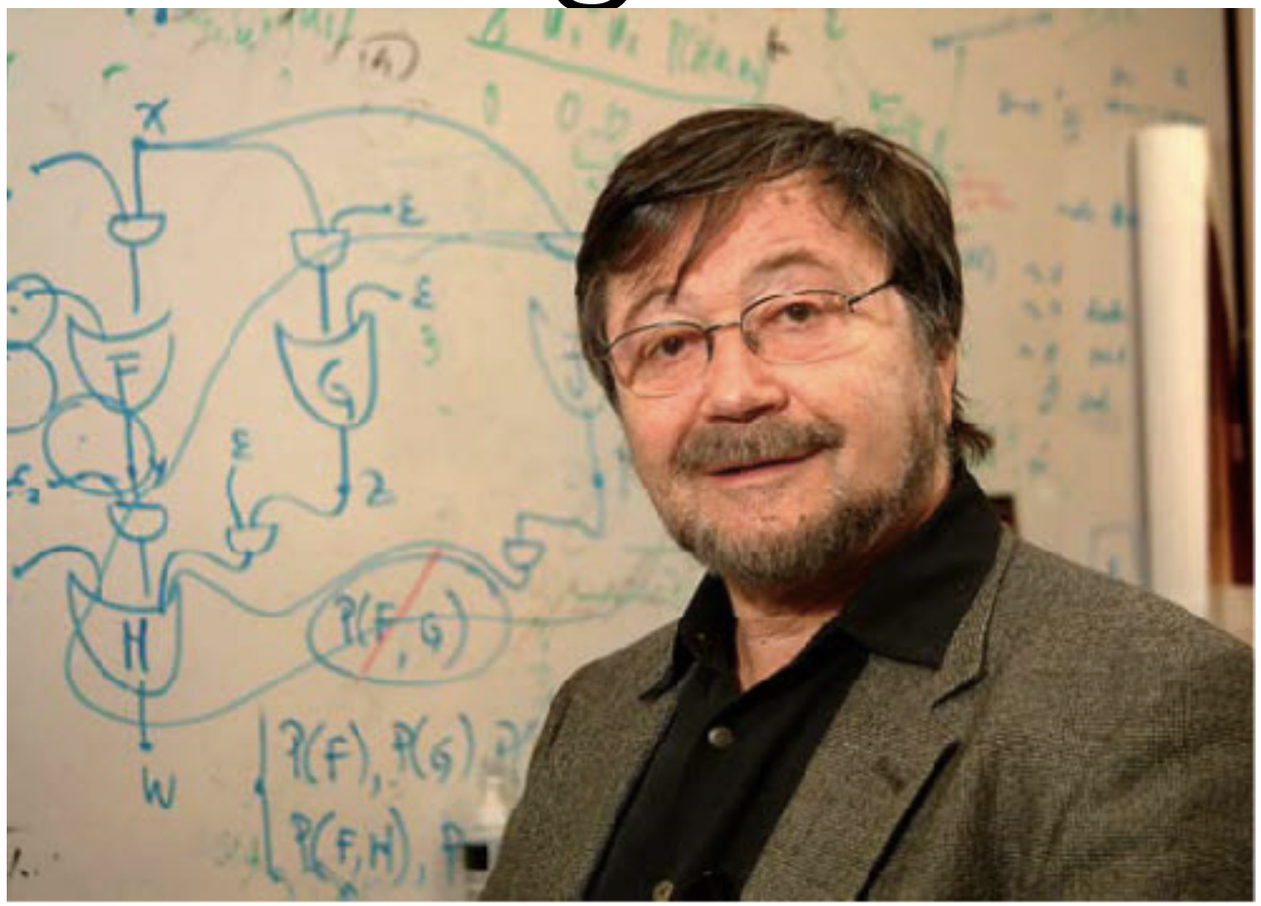
The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly emailed them reminders that only administrative assistants are allowed to take the pens.

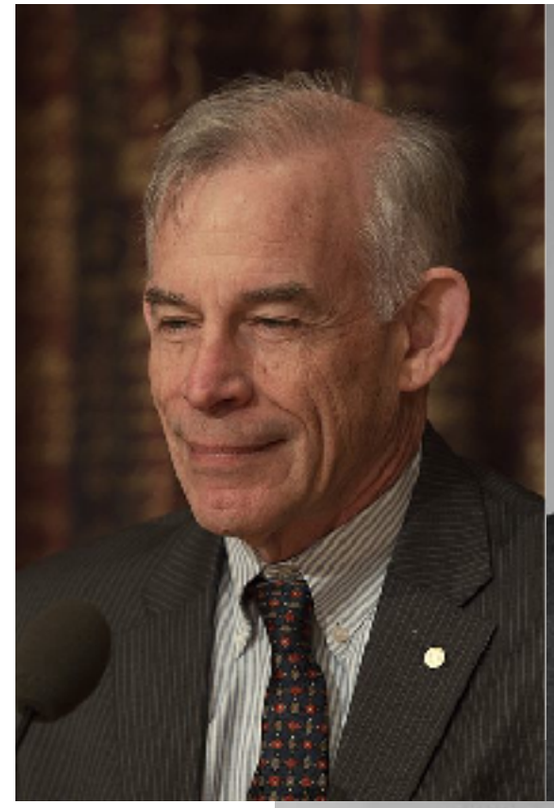
On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message… but she has a problem. There are no pens left on her desk.

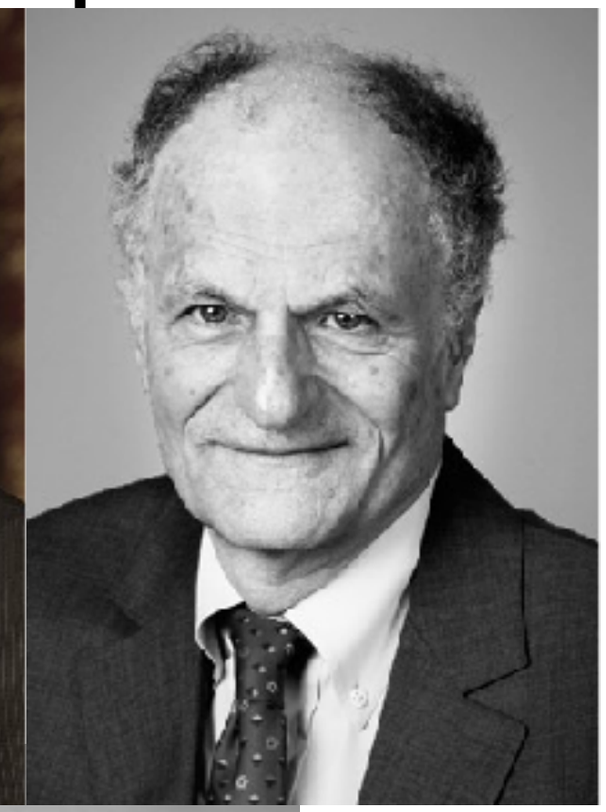
-Professor caused it?
-Assistant caused it?

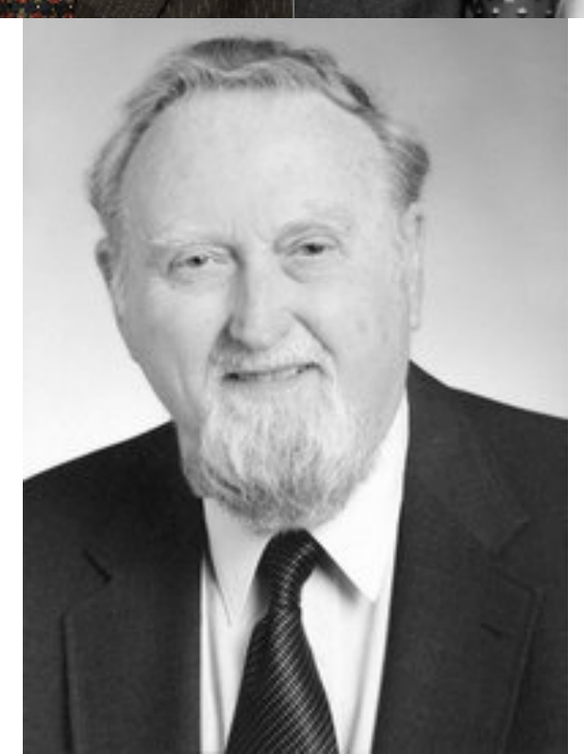
18 students, -3 to 3 scale
Professor: 2.2
Assistant: -1.2

Turing award + Nobel prize









For next week

Rasputin was led to a cellar and served cake and wine laced with cyanide – enough poison to kill 5 men.

This failed, so he was shot in the back.

Once again Rasputin survived…only to be shot again.

They then bound him and threw him in an icy river.

He escaped the bonds, but drowned.

What caused his death?

See syllabus for reading

Reading responses due Friday at noon!