Causal Inference from Complex Observational Data

Samantha Kleinberg
Stevens Institute of Technology
samantha.kleinberg@stevens.edu

Three key points

- We need causal knowledge
- Causes are hard to find and we need domain expertise
- It’s not hopeless!

“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).

Causal claims abound

Sleep Apnea Tied to Increased Cancer Risk

By ANAHAD O'CONNOR

FACEBOOK
TWITTER
GOOGLE+
E-MAIL
SHARE
PRINT

Two new studies have found that people with sleep apnea, a common disorder that causes snoring, fatigue and dangerous pauses in breathing at night, have a higher risk of cancer. The new research marks the first time that sleep apnea has been linked to cancer in humans.

Ryan Collerd for The New York Times

A CPAP device is used to treat sleep apnea.

About 28 million Americans have some form of sleep apnea, though many cases go undiagnosed. For sleep doctors, the condition is a top concern because it deprives the body of oxygen at night and often coincides with cardiovascular disease, obesity and diabetes.

"This is really big news," said Dr. Joseph Golish, a professor of sleep medicine with the MetroHealth System in Cleveland who was not involved in the research. "It's the first time this has been shown, and it looks like a very solid association," he said.

Dr. Golish, the former chief of sleep medicine at the Cleveland Clinic, said that the cancer link may not prove to be as strong as the well-documented relationship between sleep apnea and cardiovascular disease, "but until

Causal claims abound

Sleep Apnea Tied to I Risk

By ANAHAD O'CONNOR

FACEBOOK
TWITTER
GOOGLE+
E-MAIL
SHARE
PRINT

Two new studies have found with sleep apnea, a common causes snoring, fatigue and pauses in breathing at night, risk of cancer. The new rese first time that sleep apnea h to cancer in humans.

About 28 million Americans form of sleep apnea, though undiagnosed. For sleep doct condition is a top concern b deprives the body of oxygen often coincides with cardiov disease, obesity and diabete

"This is really big news," sai Golish, a professor of sleep the MetroHealth System in "It's the first time this has b association," he said.

Dr. Golish, the former chief that the cancer link may not relationship between sleep apnea and cardiovascular disease," but until

May 17, 2012 11:24 AM

Two cups of coffee a day cuts risk of dying by 10 percent, research shows

By CBS News Staff

(Credit: istockphoto)

(CBS/AP) How good is coffee for your health? For years, research has gone both ways, with some studies finding it boosts risk for heart disease, while other studies find it could be protective against breast and skin cancers.

Green coffee beans may lead to weight loss, study shows
Coffee helps prevent skin cancer? What study shows
PICTURES: Coffee and your health: Latest findings

A large-scale study of 400,000 people offers good news for coffee-drinkers: you might just live longer.

The study is the largest ever done on the issue, and the results should reassure any coffee lovers who think it's a guilty pleasure that may do harm. And whether it's regular or decaf doesn't even matter.

"There may actually be a modest benefit of coffee drinking," said lead researcher Neal Freedman of the National Cancer Institute.

The study, published online in the May 16 issue of the New England Journal of Medicine, kicked off in 1995 and involved 402,260 AARP members ages 50 to 71 who lived in California, Florida, Louisiana, New Jersey, North Carolina, Pennsylvania and Atlanta and Detroit. People who already

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.

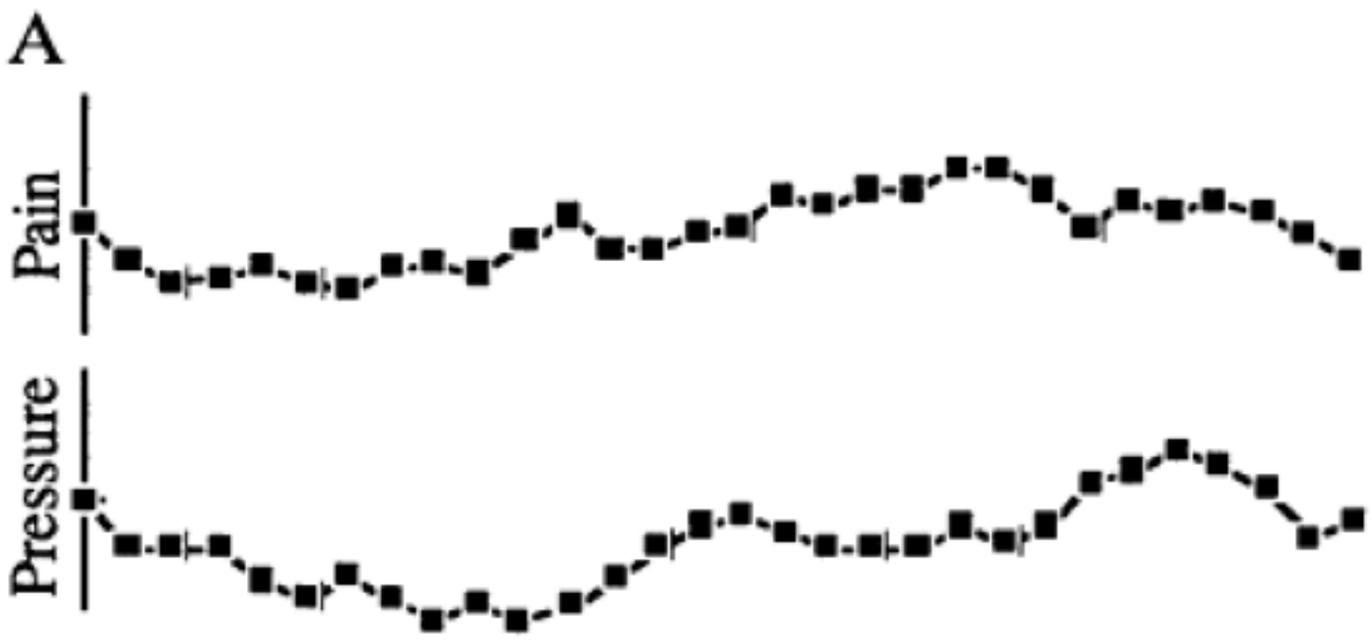
People who ate more red meat were less physically active and more likely to smoke and had a higher body mass index, researchers found. Still, after controlling for those and other variables, they found that each daily increase of three ounces of red meat was associated with a 12 percent greater risk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

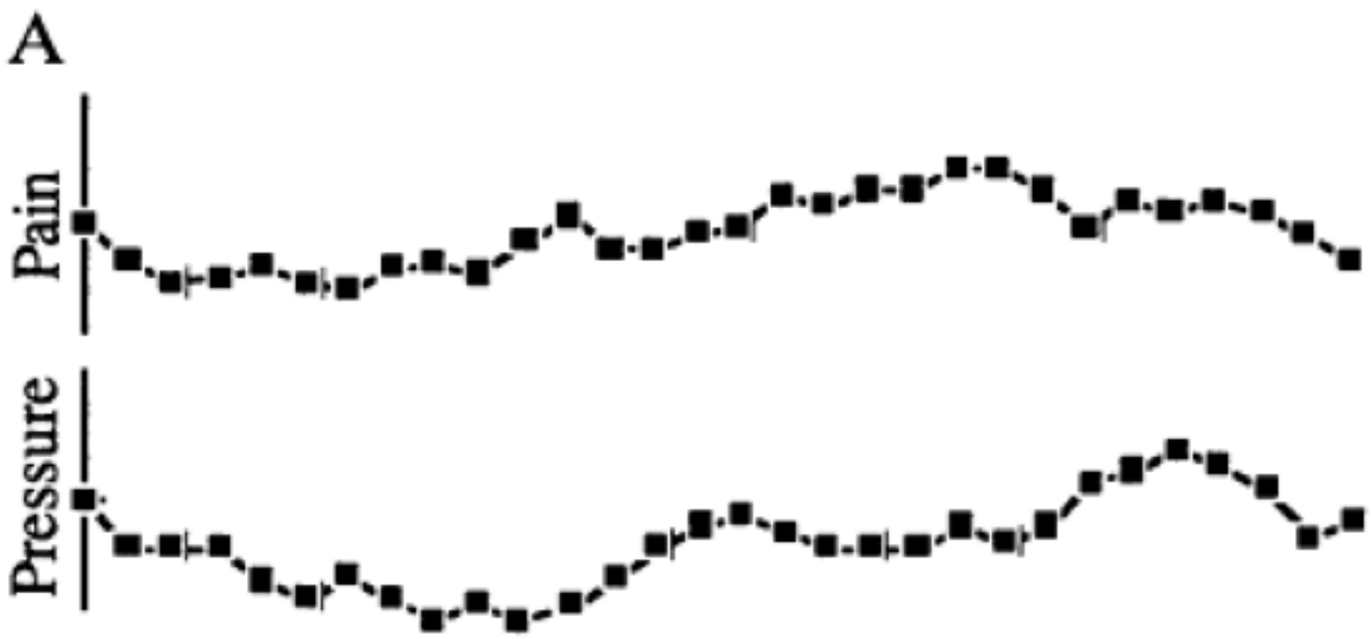
If people in the study had eaten half as much meat, the researchers estimated, deaths in the group would have declined 9.3 percent in men and 7.6 percent in women.

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.



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.



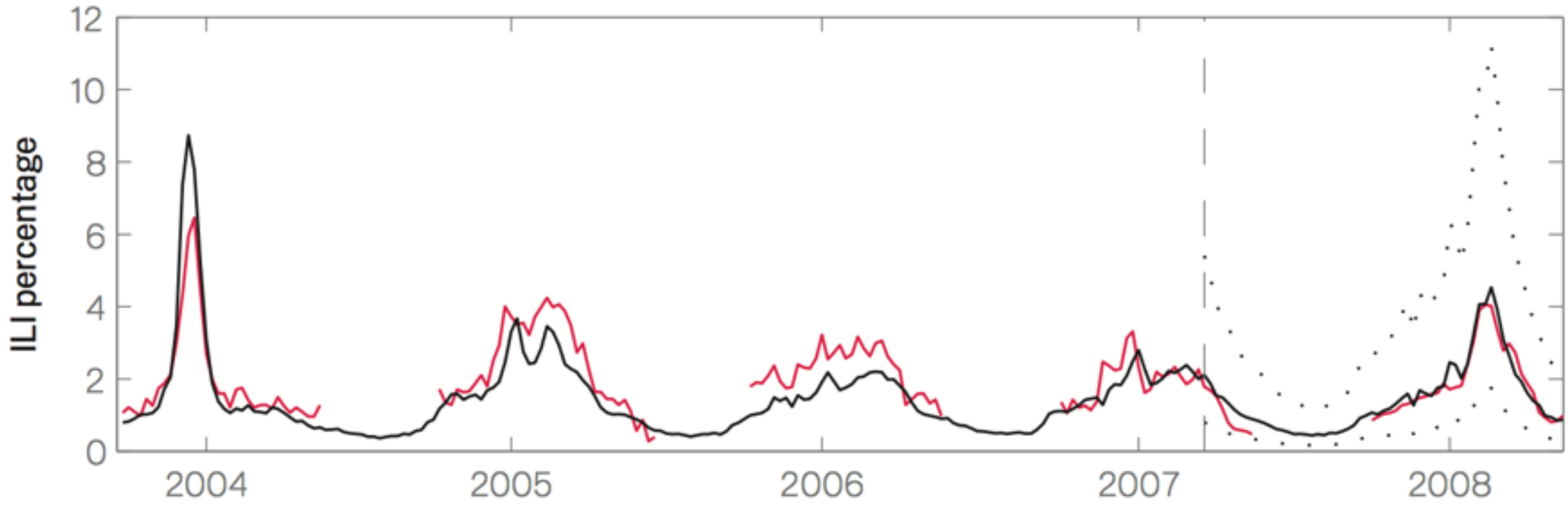
Correlation = 0!

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.

Why do we need causes?

- Prediction
- Explanation
- Intervention

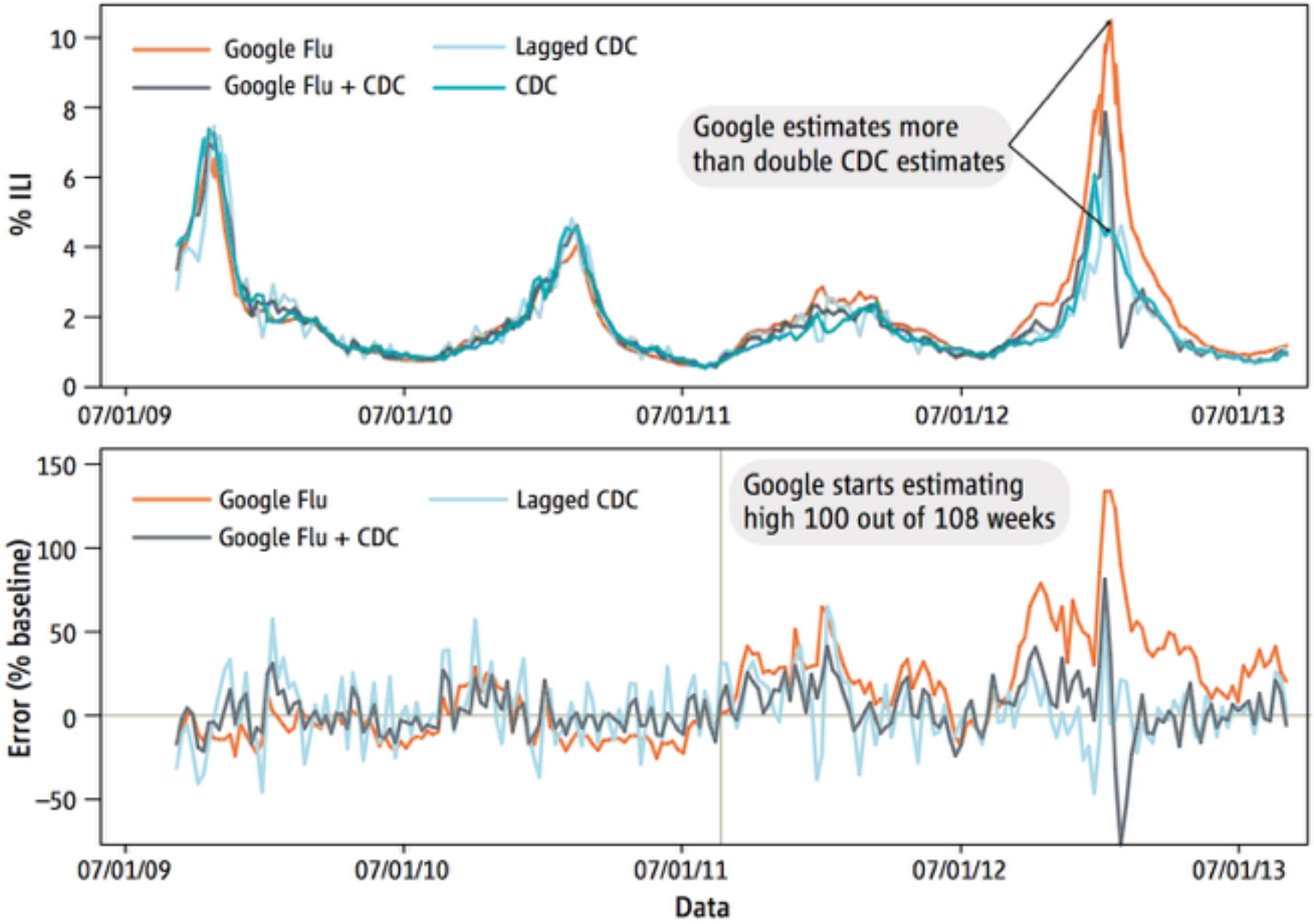
Google flu



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

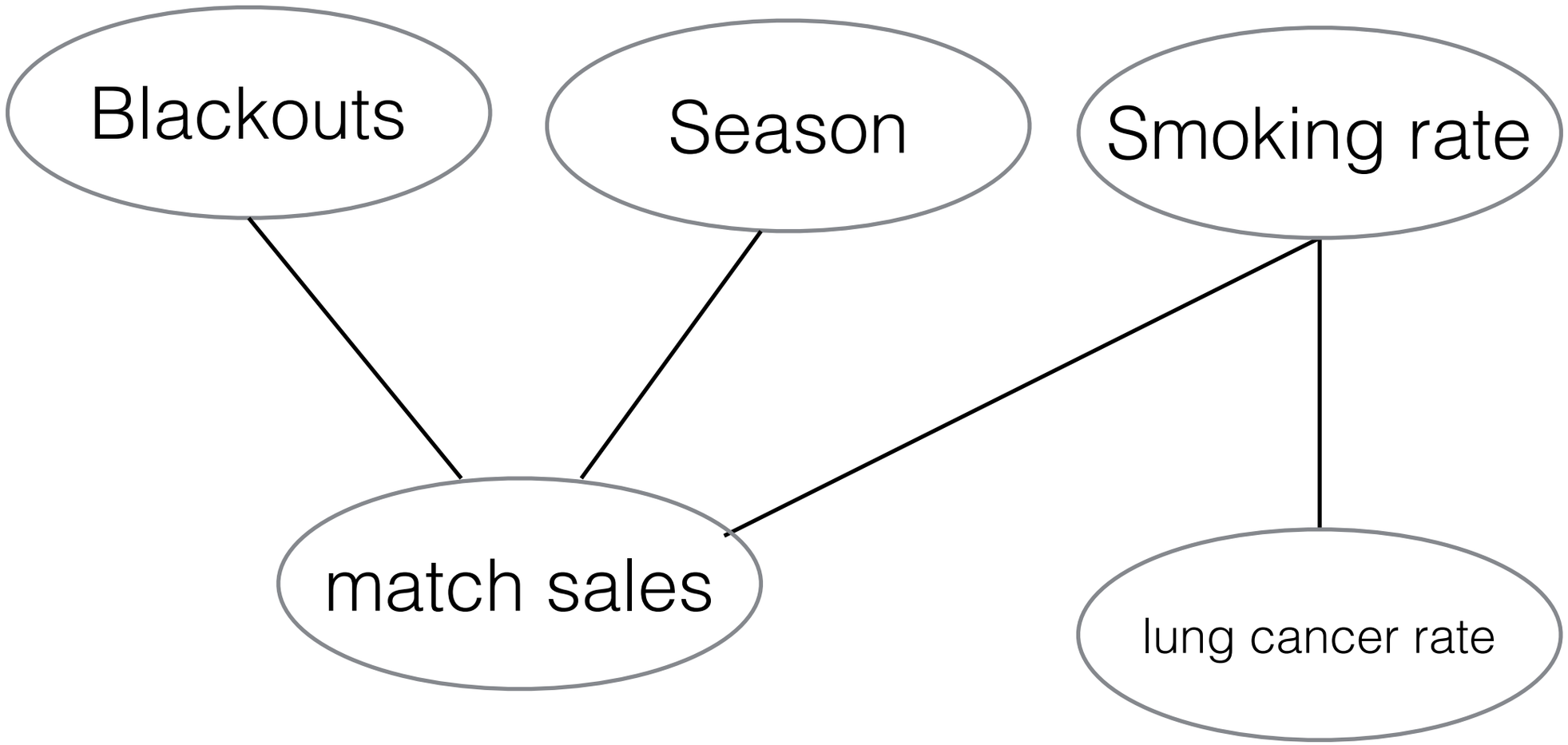


BIG DATA

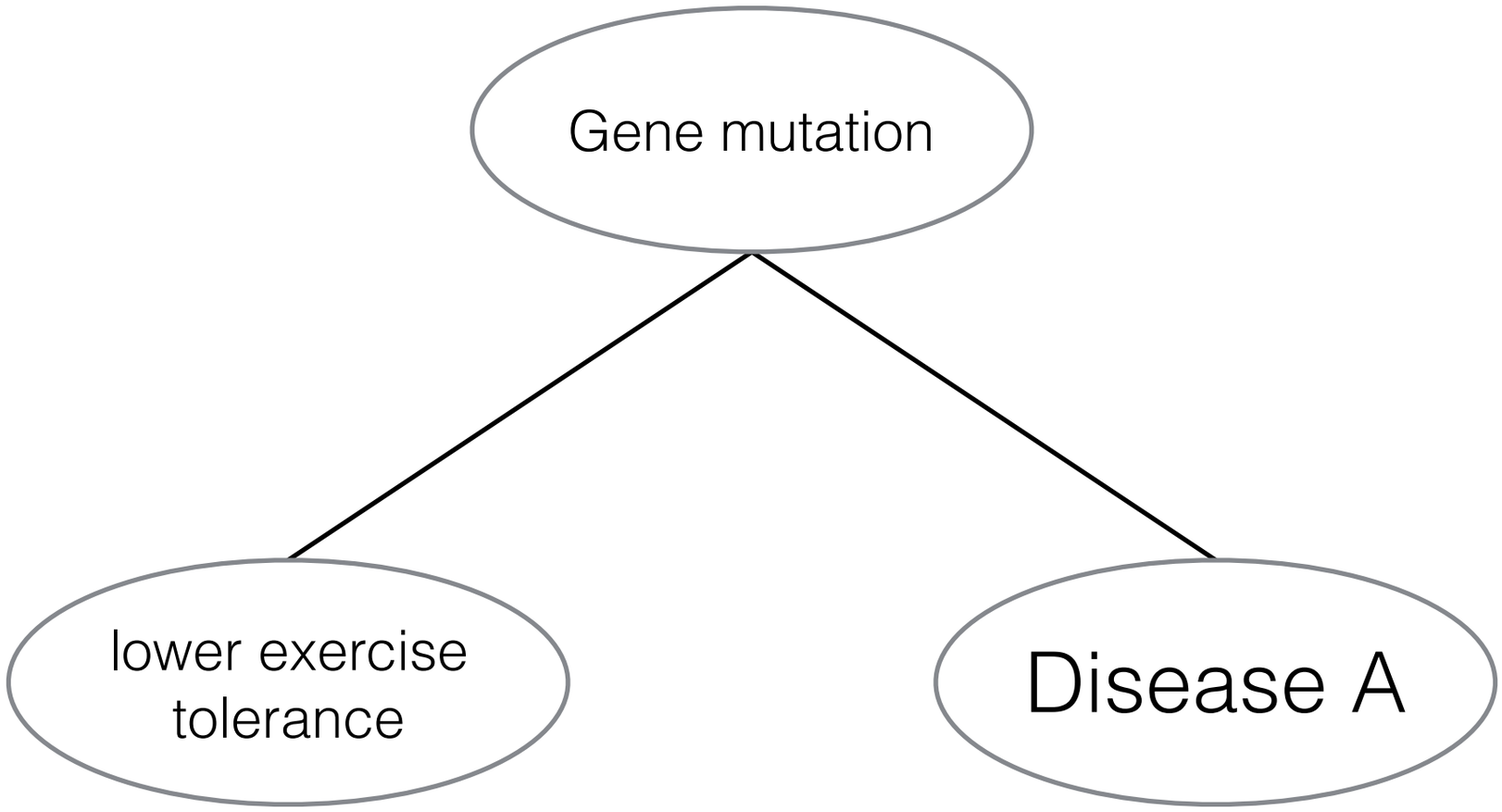
The Parable of Google Flu: Traps in Big Data Analysis

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

Prediction

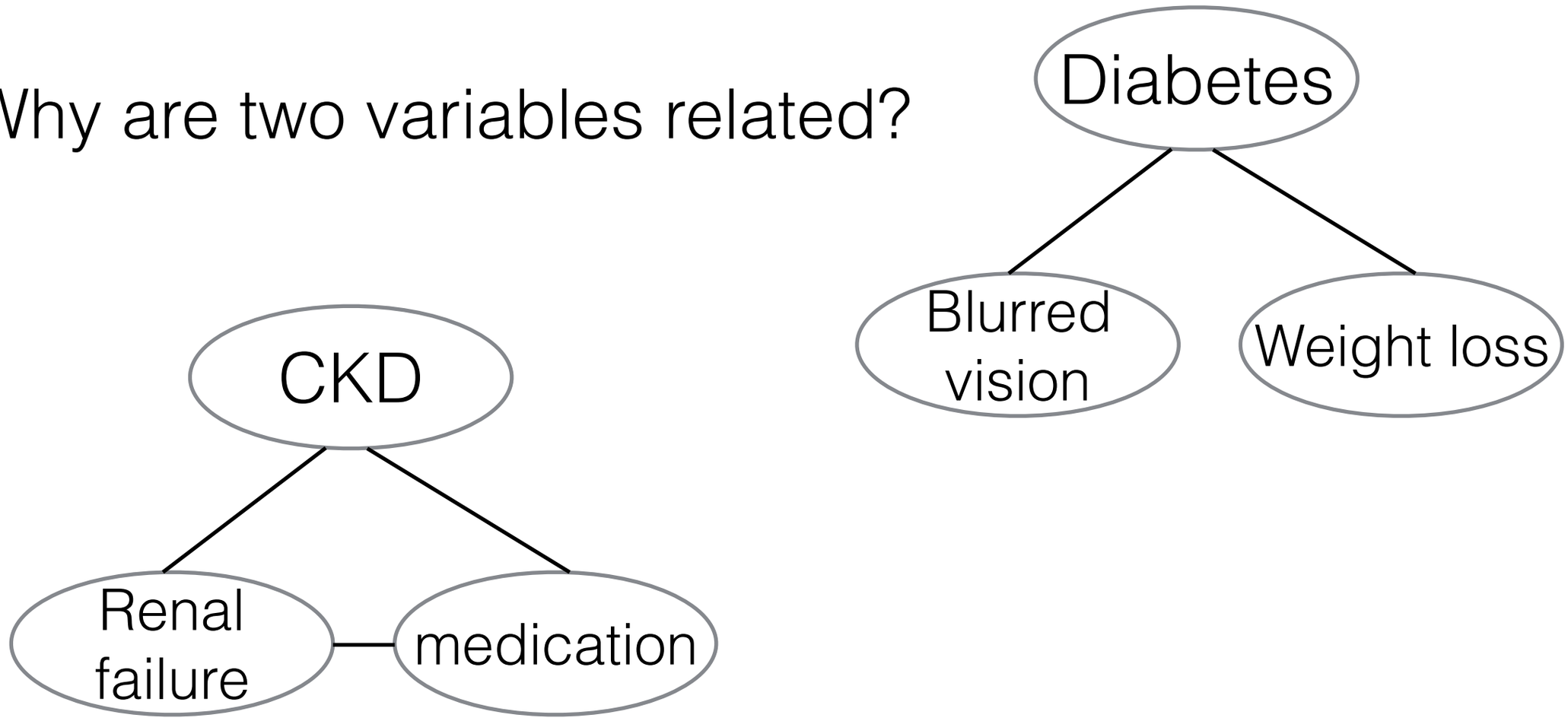

Blackouts
Season
Smoking rate
match sales
lung cancer rate

Prediction, continued



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

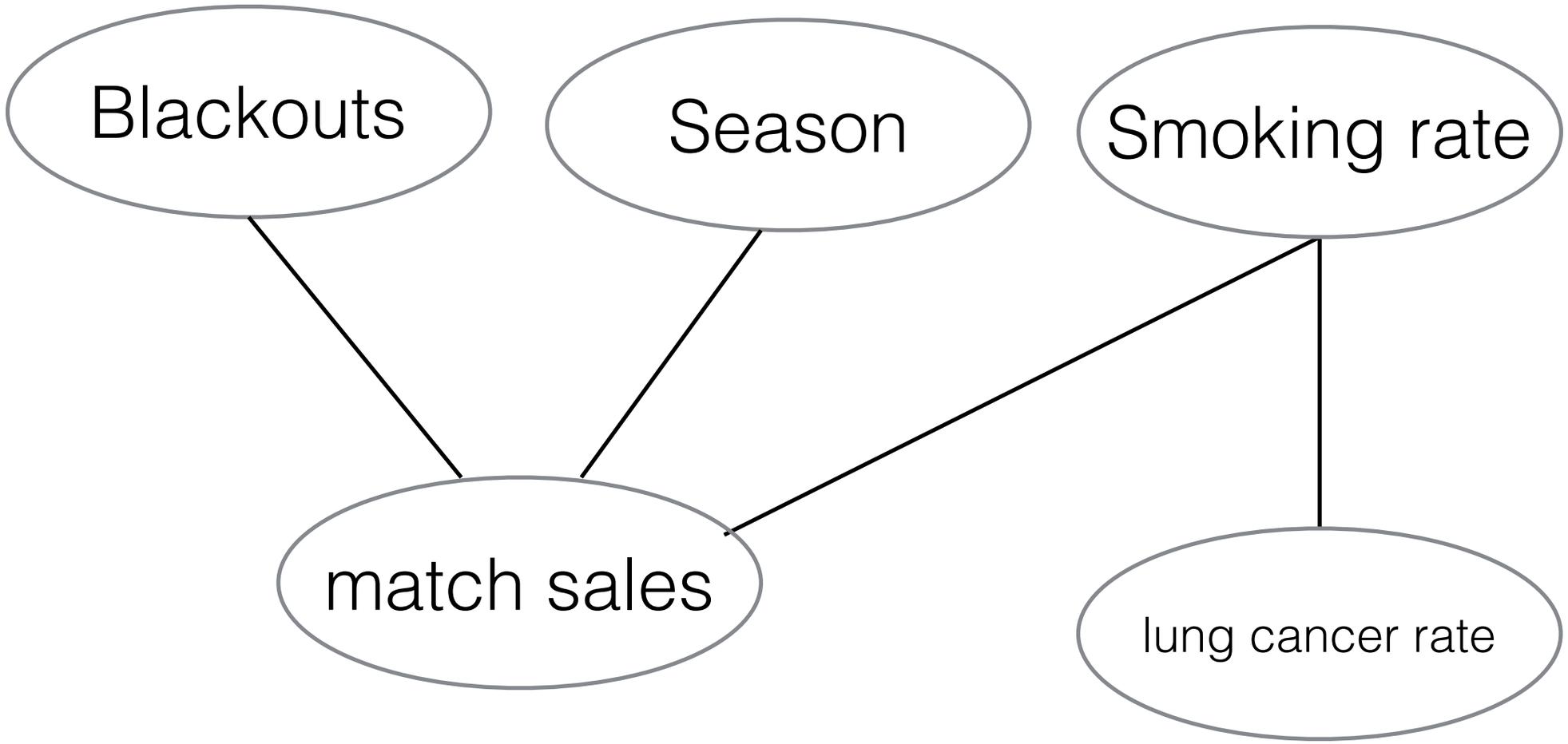
- Methods for finding causes from data, but what about explaining events?
- Practical problem, but challenging
  - Information incomplete
  - Where do explanations come from?
  - General and singular can differ

Intervention

- Why do we need causes to take action?
  - Buying stocks
  - Taking vitamins
  - Decreasing sodium to prevent hypertension

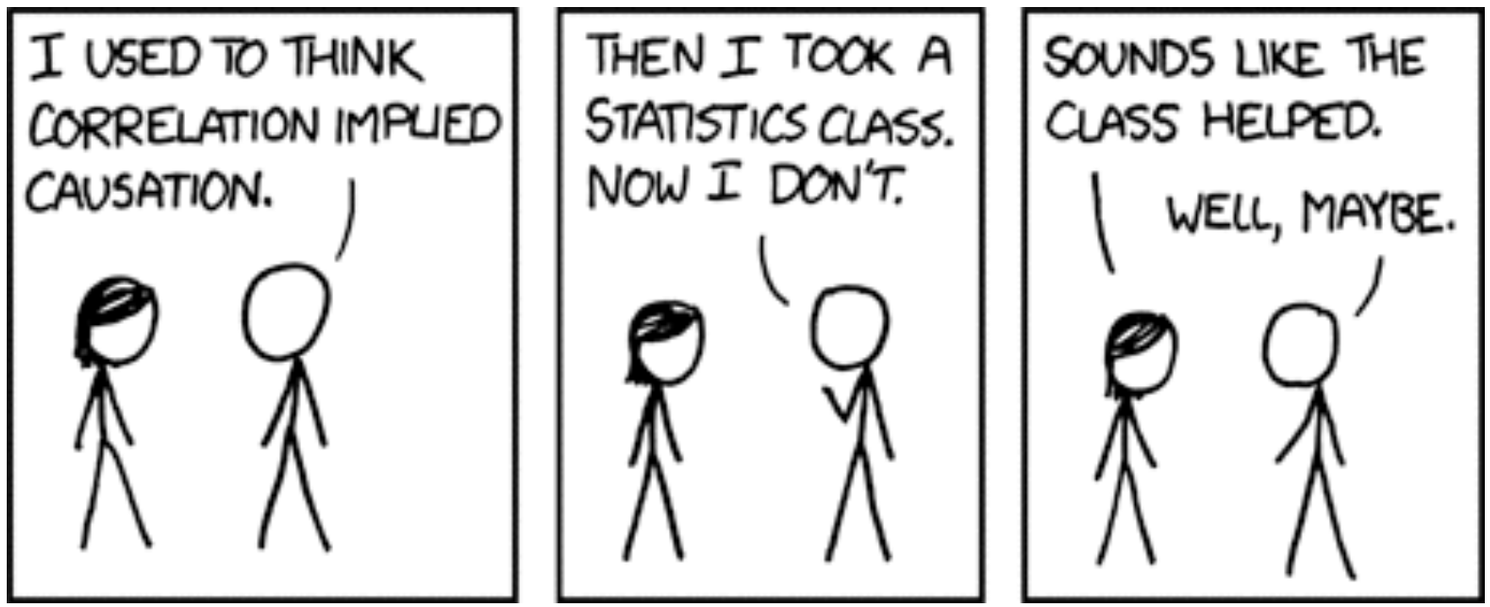
- What happens if we intervene on a correlated factor?

Using causes to guide intervention



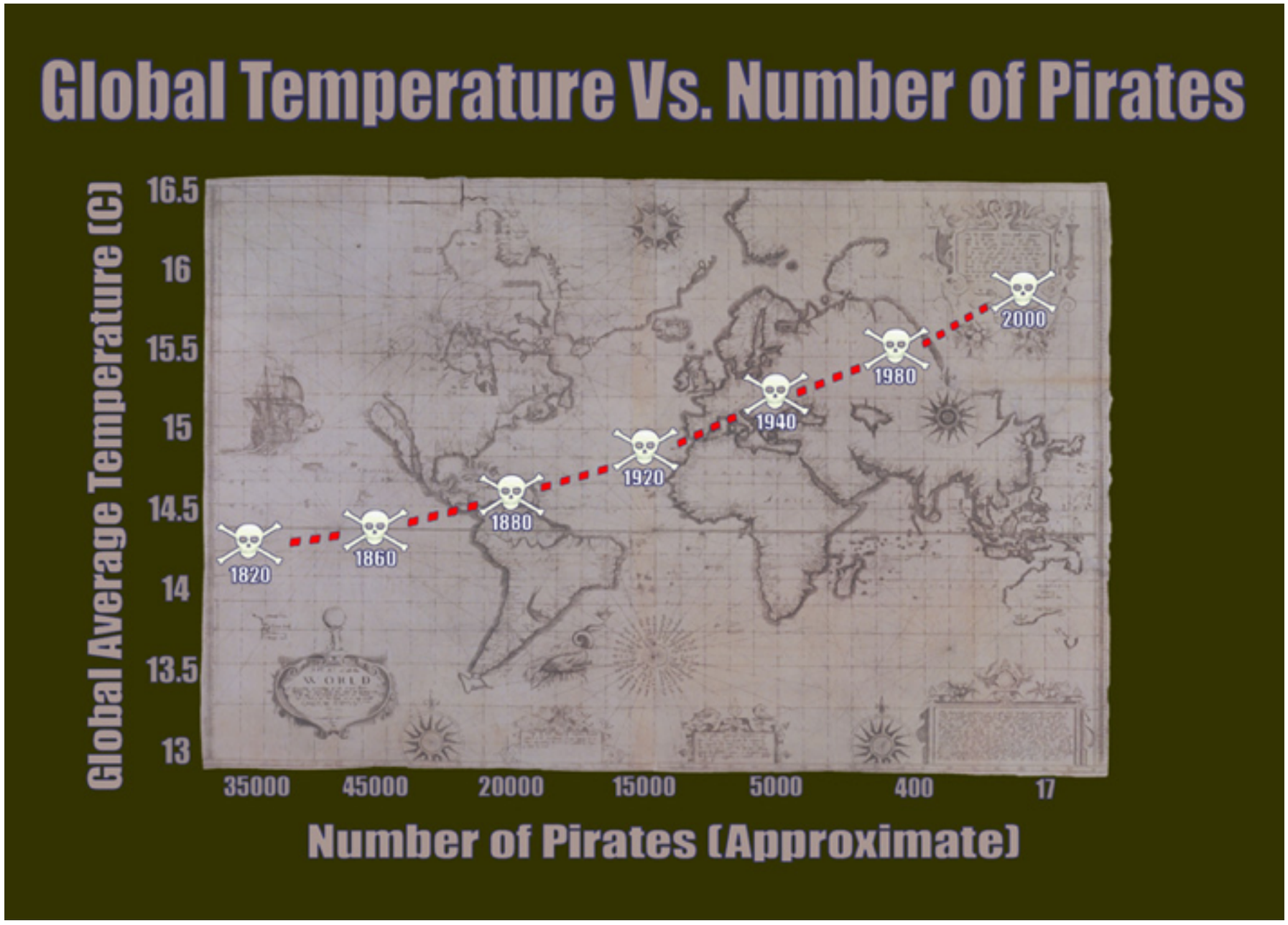
Using interventions to find causes

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



http://xkcd.com/552/

Nonstationary time series

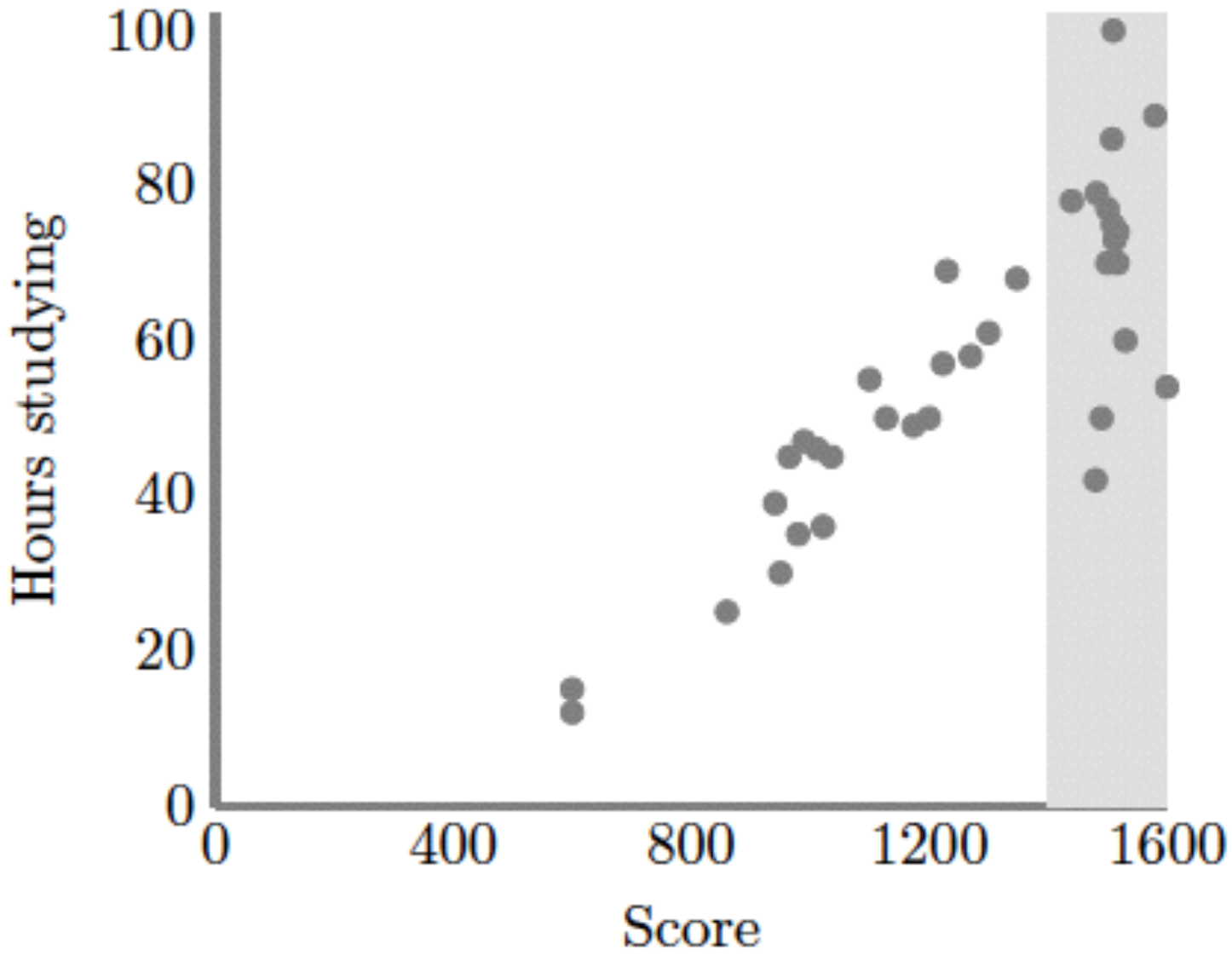


http://bama.ua.edu/~sprentic

Nonstationary time series

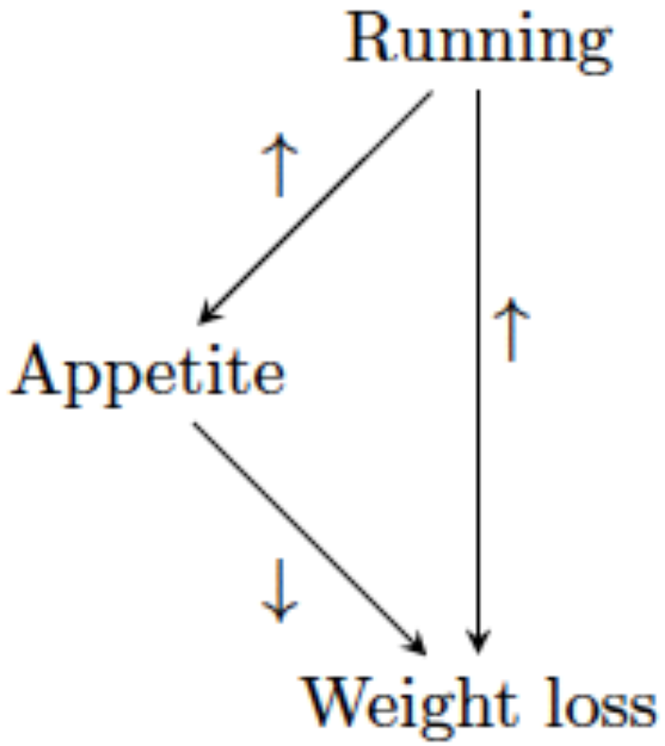
S. Kleinberg. (2015) Why: A Guide to Finding and Using Causes. O'Reilly Media.

Restricted range



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

Canceling out



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

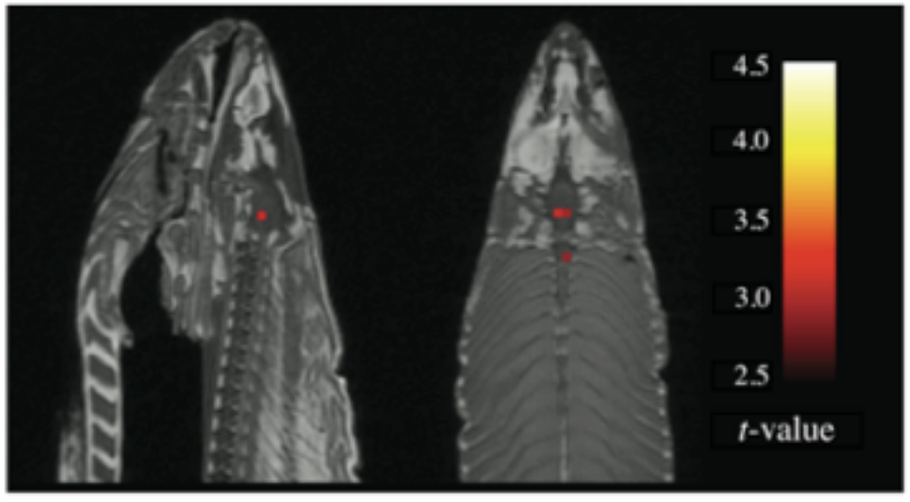
Multiple testing

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.



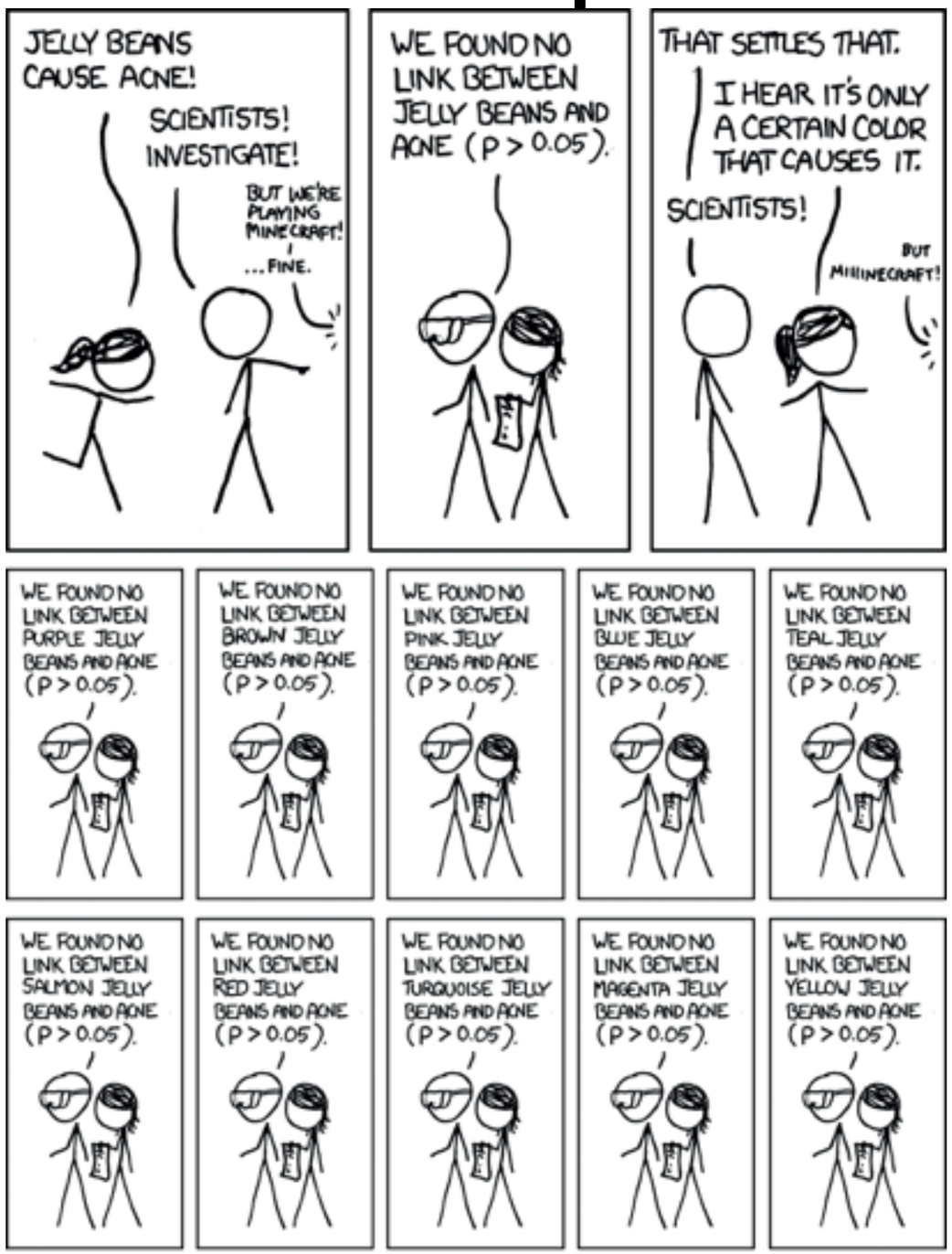
A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

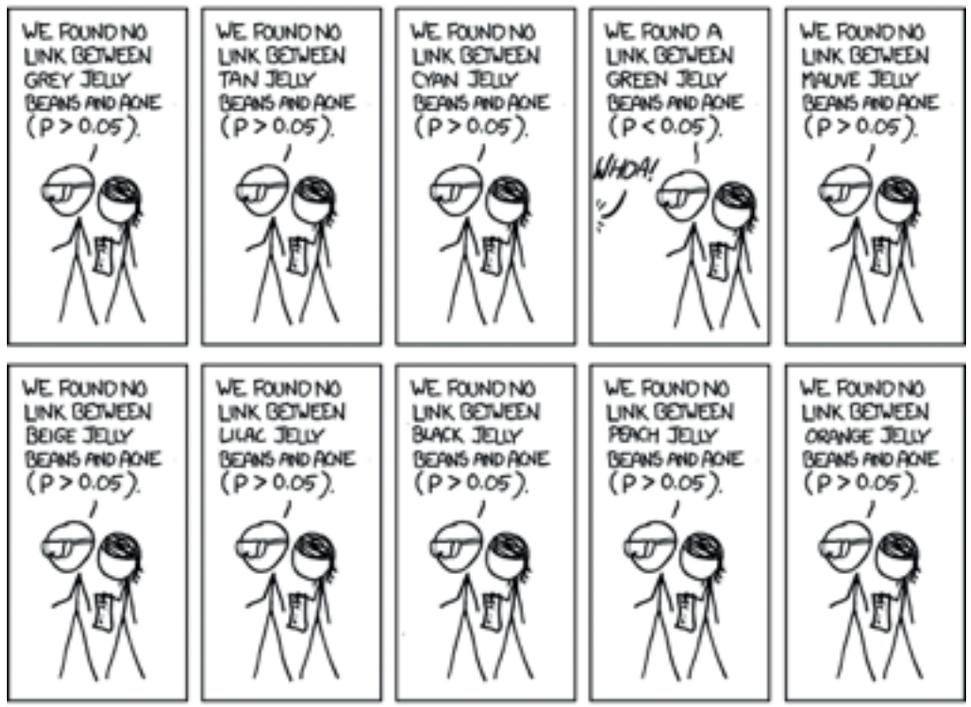
Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

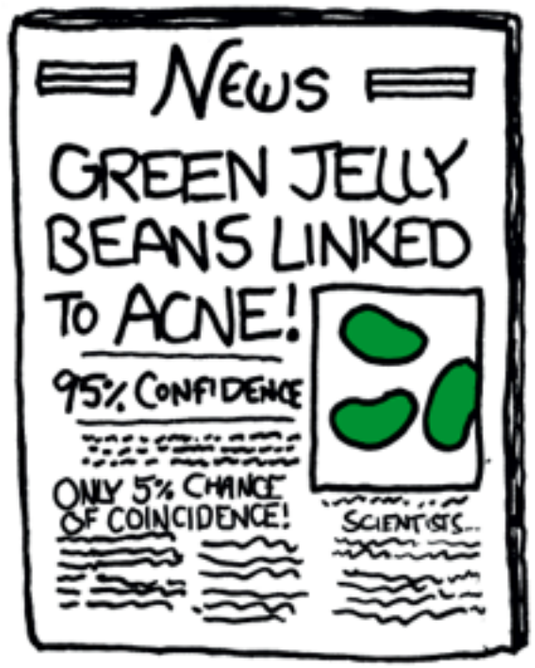
Identical *t*-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

Bennett, C. M., Miller, M. B., & Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for multiple comparisons correction. NeuroImage, 47(1), 125.

Multiple comparisons







http://xkcd.com/882/

Causation without correlation: Simpson's paradox

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

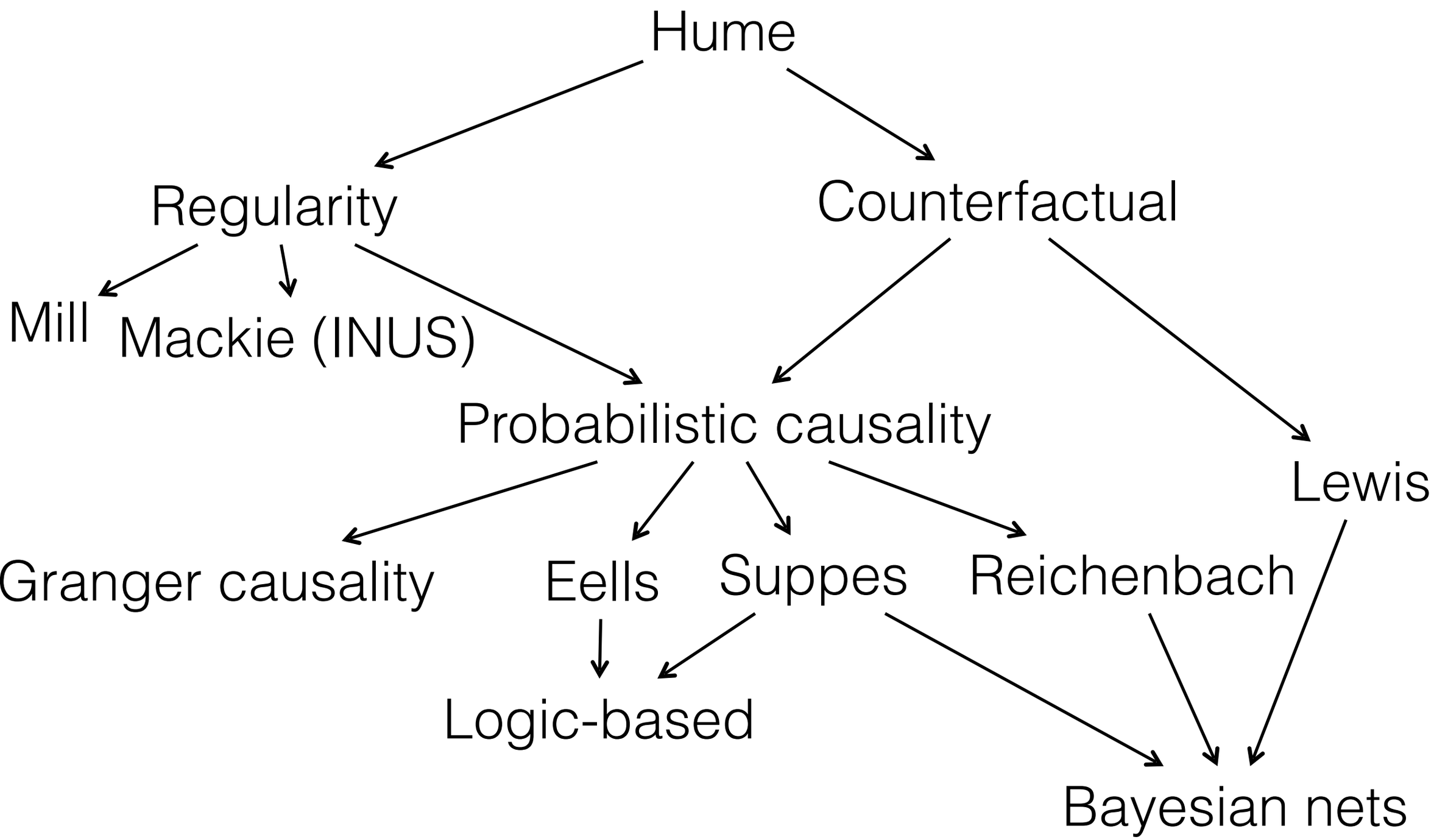
Causation without correlation: Simpson's paradox

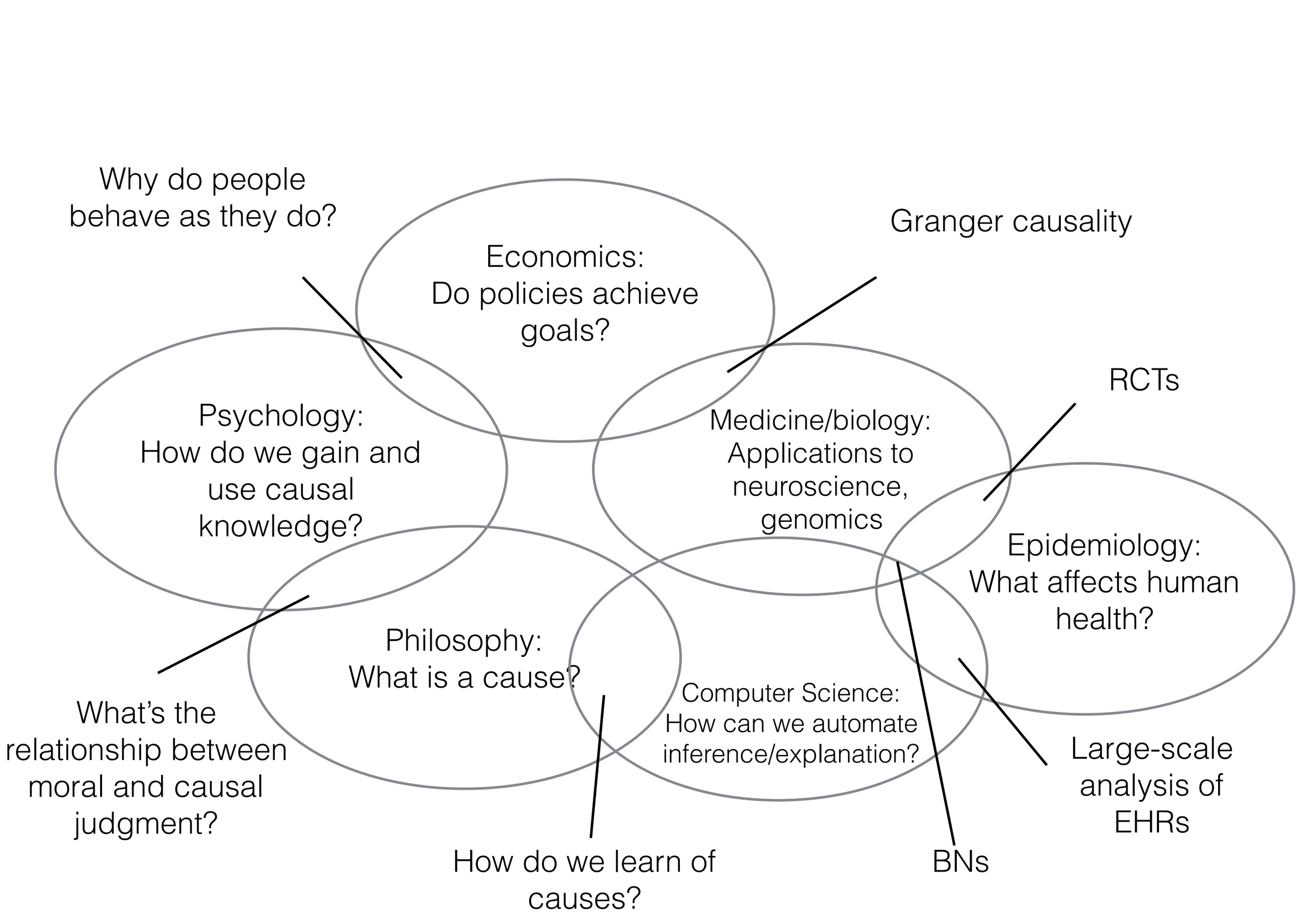
Treatment	**Men**		**Women**		**Combined**	
	Dead	Alive	Dead	Alive	Dead	Alive
A	80	120 (60%)	5	95 (95%)	85	215 (72%)
B	20	20 (50%)	39	221 (85%)	59	241 (80%)
Total	100	140	44	316	144	456

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

Why is causal inference hard?

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


Hume
Regularity
Counterfactual
Mill
Mackie (INUS)
Probabilistic causality
Lewis
Granger causality
Eells
Suppes
Reichenbach
Logic-based
Bayesian nets


Why do people behave as they do?
Economics: Do policies achieve goals?
Granger causality
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

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

Data to causes: a few overlooked assumptions

- No hidden common causes
- Data represents true distributions
- Right variables

Representative data (faithfulness)

- If no alarm system -> robberies, data should reflect dependence
- May not if…
  - Canceling out (doesn't have to be exact)
  - Selection bias

Right variables

- What are we finding causes between?
- Measure weight..
  - Use continuous value? BMI? group obesity/ morbid obesity together?
  - Or should we use weight change+other features?

Why observational data?

- Routinely collected in many situations
  - Electronic health records in hospitals
  - ICU data streams
  - Body-worn sensors and mobile devices

- Experiments often infeasible, unethical, or too expensive

Some recent work

- We can make some progress in getting causes from medical data
- Explanation can be automated (sometimes)
- Understanding chronic disease requires data from daily life
    - Automated dietary monitoring (a fitbit for nutrition)

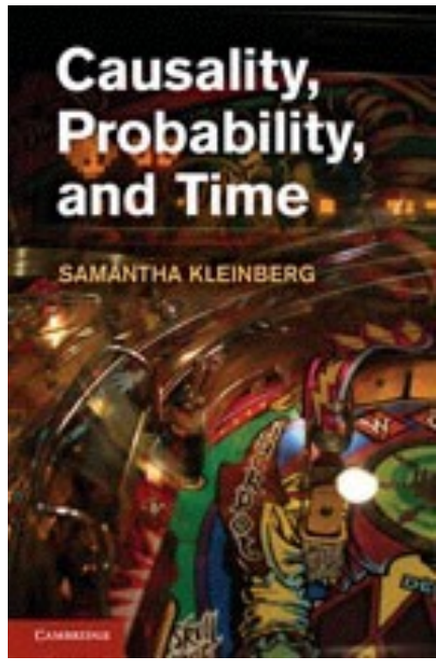
Logic-based causal inference

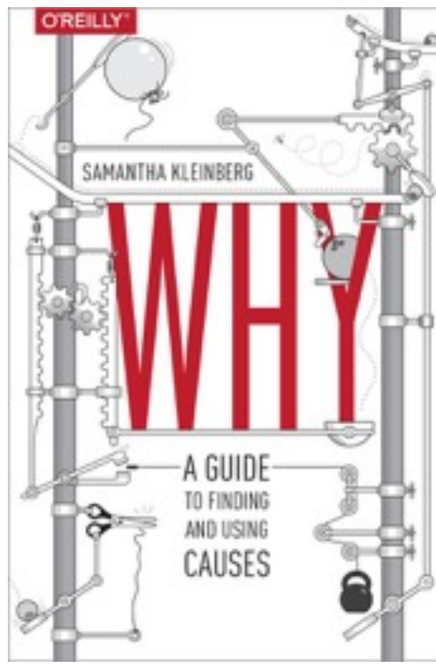
- Complex, temporal relationships

$$v \rightsquigarrow^{\geq 15, \leq 40}_{\geq 0.4} g$$

- Assess average difference cause makes to probability of effect

$$\varepsilon_{avg}(c, e) = \frac{\sum_{x \in X \ c} P(e|c \wedge x) - P(e|\neg c \wedge x)}{|X \backslash c|}$$





Kleinberg, S. (2012) Causality, Probability, and Time. Cambridge University Press.
Kleinberg, S. (2015) Why: A Guide to Finding and Using Causes. O'Reilly Media

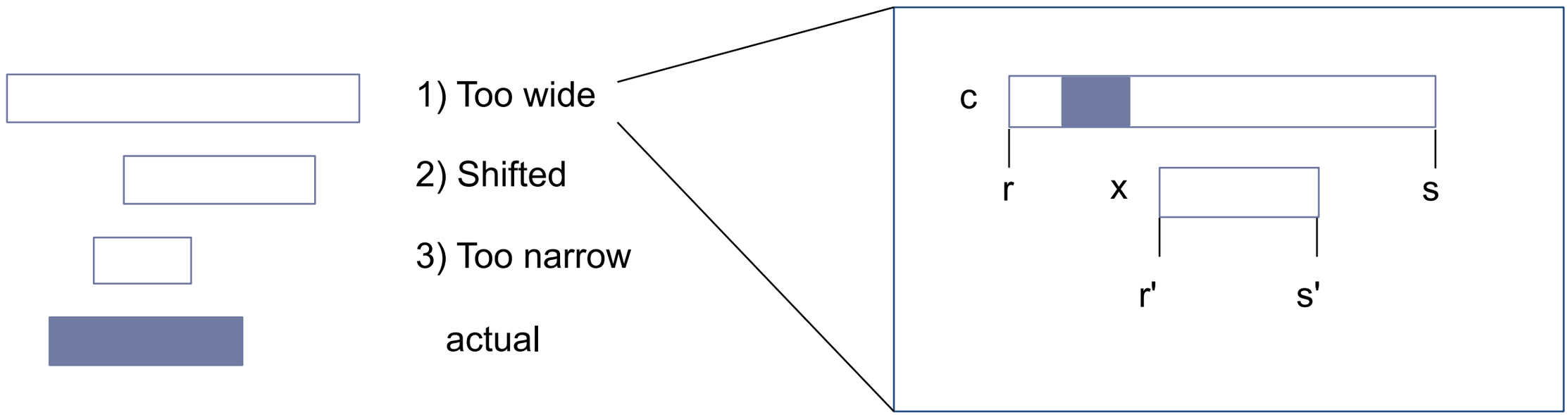
- Main idea: looking for better explanations for the effect

- Inferring timing

- Instead of accepting/rejecting hypotheses, refine them from data

- Can start by testing relationships between all variables and CHF in 1-2 weeks, and ultimately infer "high AST leads to CHF in 4-10 days”

Finding timings: greedy search



$$\varepsilon_{avg}(c,e) = \frac{\sum_{x \in X\ c} P(e|c \wedge x) - P(e|\neg c \wedge x)}{|X \backslash c|}$$

$$P(e|c \wedge x) = \frac{\#(c \wedge x \wedge e)}{\#(c \wedge x)}$$

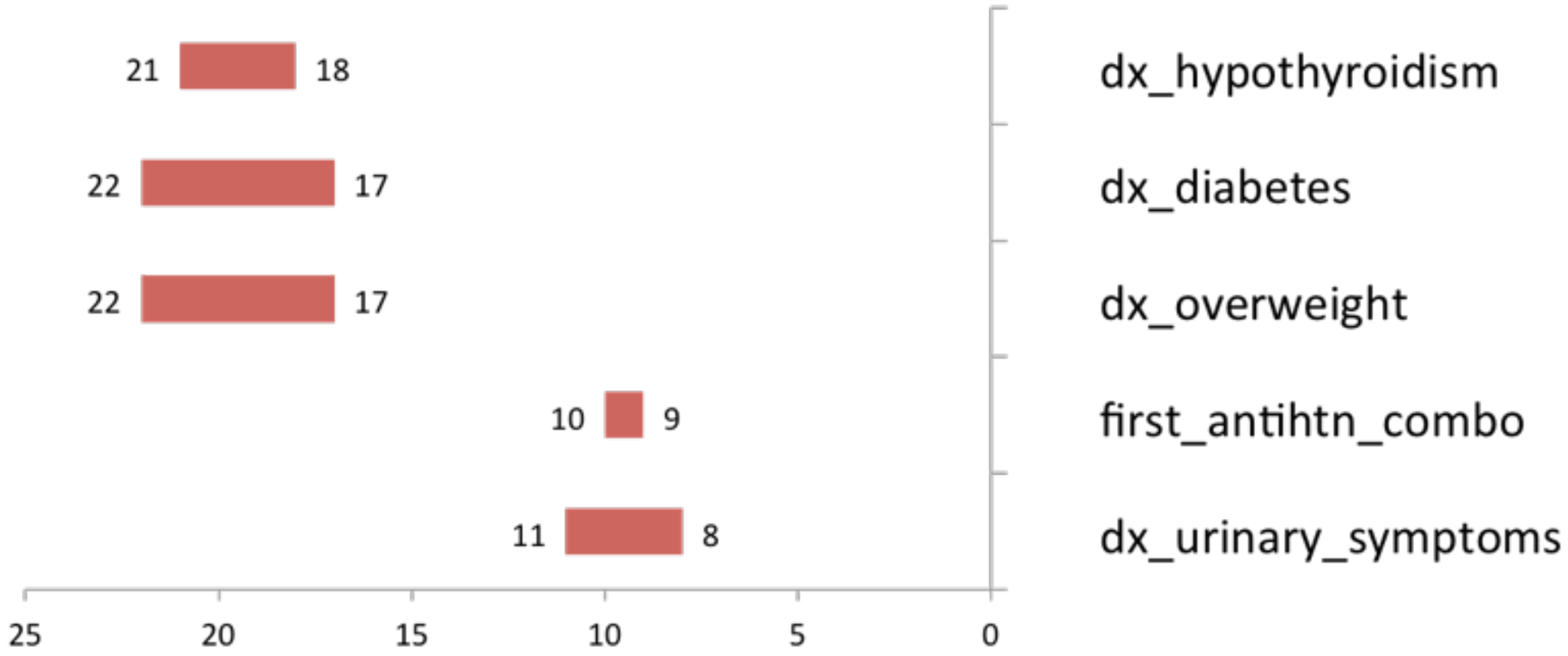
- Key assumptions
    - Stationarity, no latent confounders
- Main advantages
    - Exact inference, time window (vs. lag), complex relationships

Application: finding risk factors for heart failure

- Which patients have heart failure?

- When does the heart failure start?

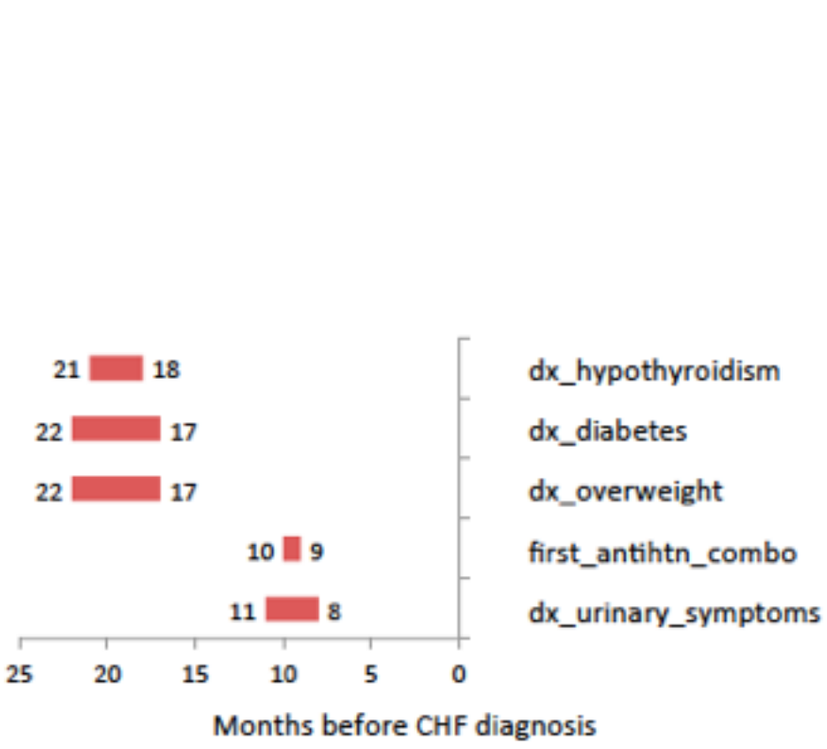
CHF - Geisinger



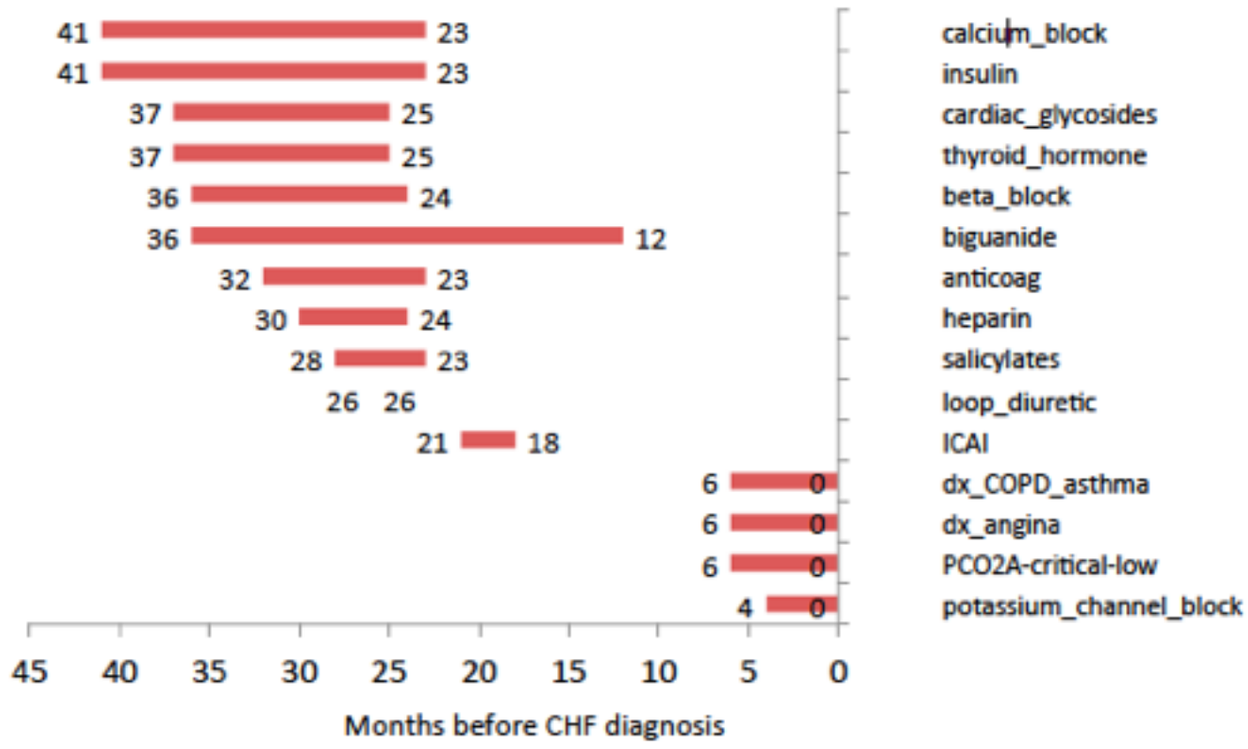
Months before CHF diagnosis

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

CHF – Geisinger and CUMC



(a) Geisinger



(b) CUMC

“Same” study, multiple populations

-Different data

-Different types of error

-Many decisions, replicating method vs testing same hypothesis

What do differing results mean?

Application: stroke

Massive amounts of data collected in ICU (>100,000 measurements per person) and by body-worn sensors

How do patients change over time in ICU?

NICU dataset

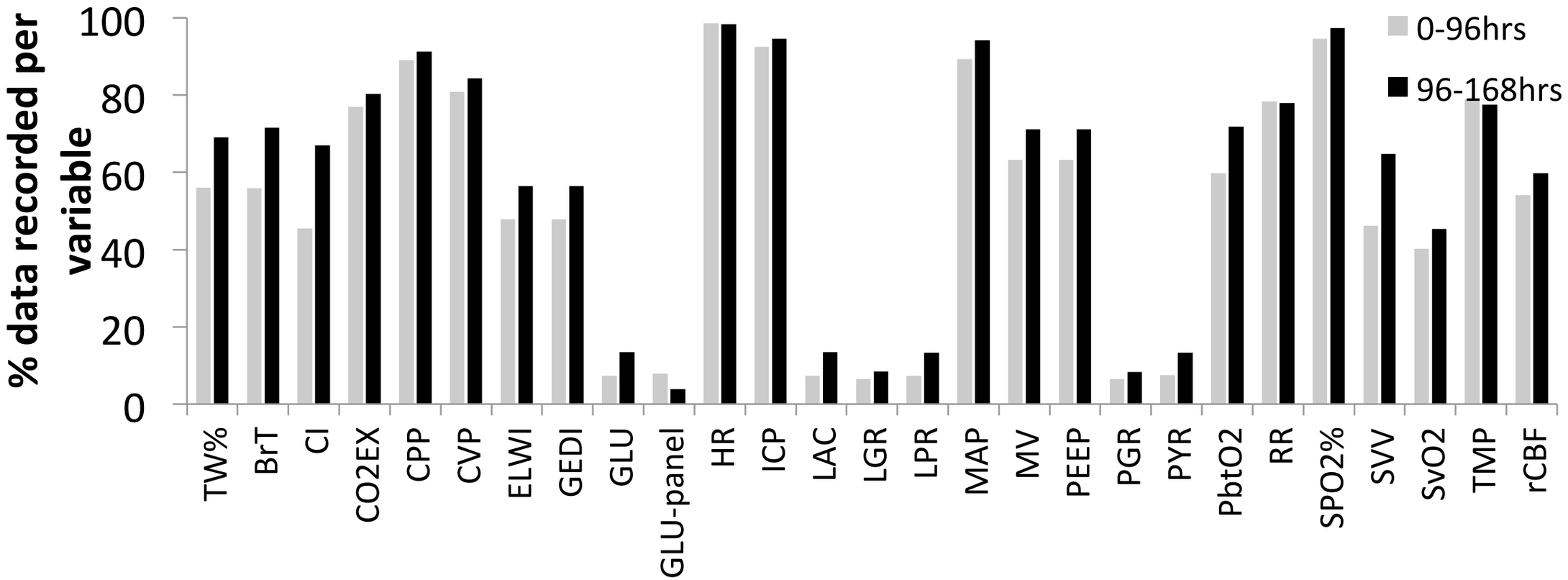
- 98 patients with subarachnoid hemorrhage
- Monitoring included
  - Depth and surface EEG
  - Microdialysis
  - Physiologic measurements

  (no data on procedures)

Lots of data, but lots of missing data

- Device malfunctions
- Device connected to perform a procedure
- Different monitors started at different times
- Different recording frequencies

Lots of data, but lots of missing data



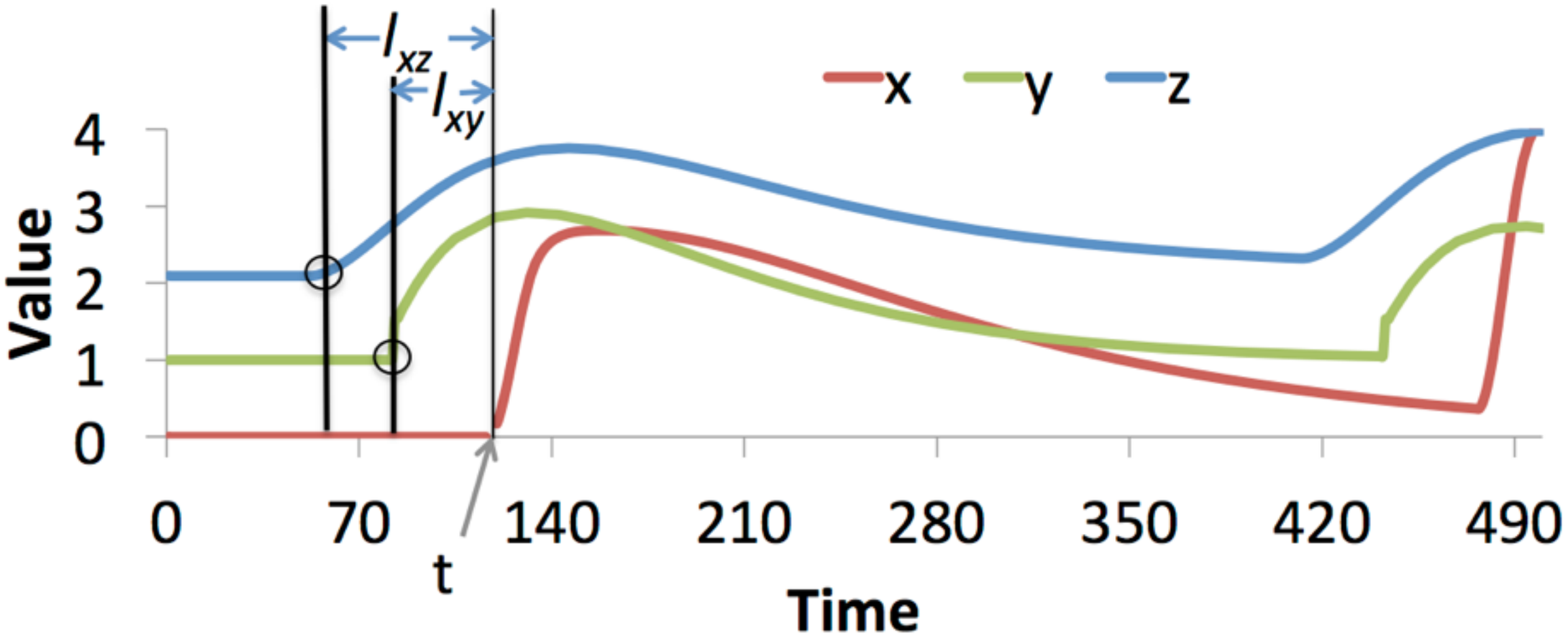
And…

- All variables may be missing at once (if measured by single device)
  - Can’t use imputation methods that assume some present values

- Missing values depend on variable + other variables
  - e.g. BG depends on itself, as well as insulin

- Variables are correlated across time

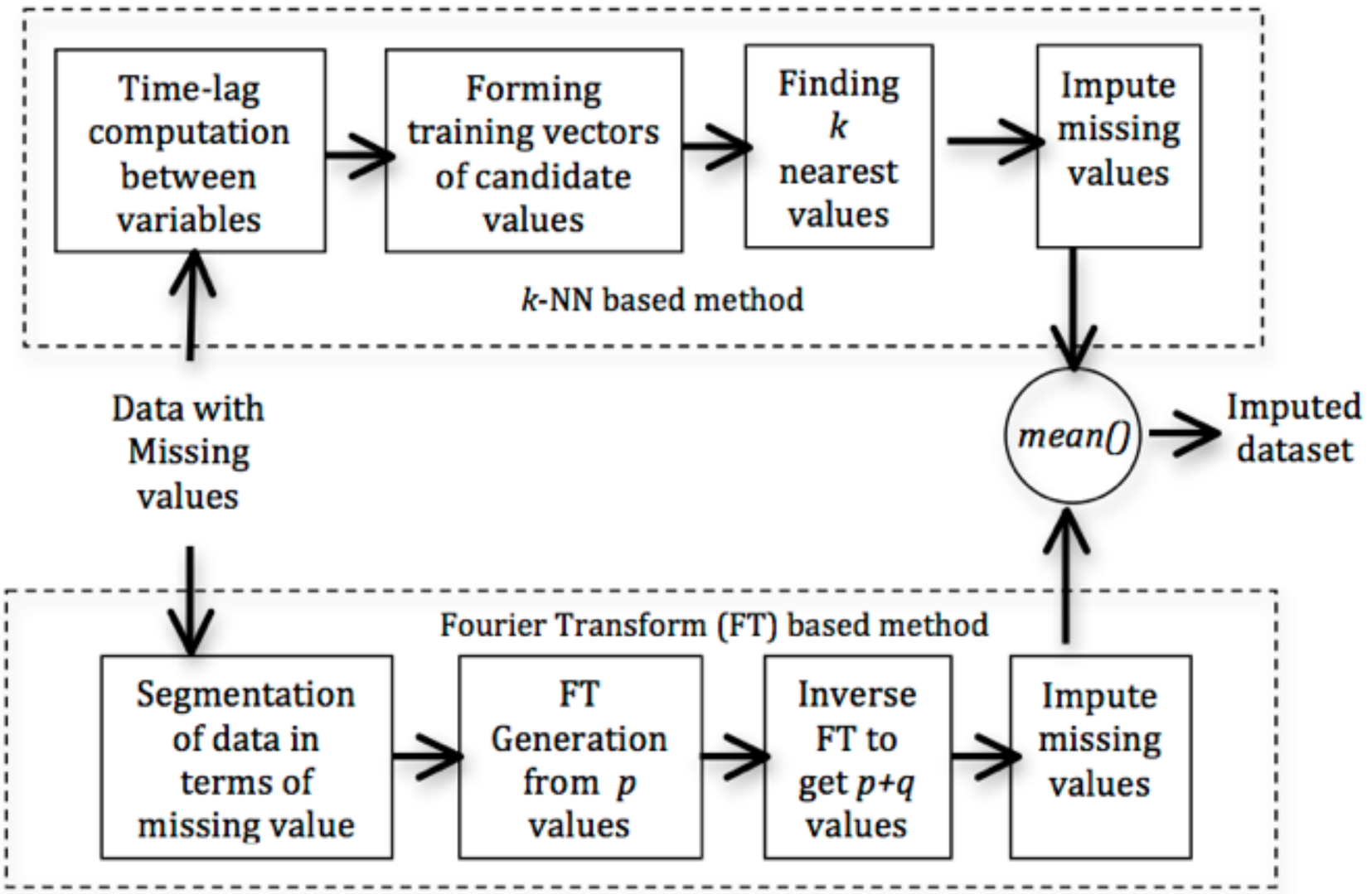
Imputation w/lagged correlations



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)

https://github.com/kleinberg-lab/FLK-NN

Imputation w/lagged correlations



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)

https://github.com/kleinberg-lab/FLK-NN

Evaluation

Mean of NMAE for DSIM dataset. Bold values indicate highest accuracy.

Method	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
BPCA	0.046	0.047	0.049	0.051	0.052	0.053	0.055	0.057	0.059	0.061
EM	0.057	0.054	0.053	0.053	0.053	0.053	0.055	0.056	0.058	0.060
Hot deck	0.053	0.055	0.057	0.059	0.063	0.069	0.081	0.095	0.108	0.116
Inpaint	73.4	79.8	82.0	81.7	88.4	89.9	98.8	100.9	105.3	109.1
k-NN	0.044	0.046	0.047	0.049	0.051	0.056	0.064	0.076	0.088	0.098
MEI	0.179	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178
MICE	0.063	0.065	0.065	0.065	0.068	0.069	0.072	0.074	0.076	0.081
Fourier	0.048	0.049	0.050	0.049	0.051	0.051	0.053	0.053	0.056	0.058
L*k*-NN	**0.041**	0.042	0.043	0.044	0.046	0.046	0.047	0.048	0.051	0.056
FL*k*-NN	**0.041**	**0.041**	**0.042**	**0.043**	**0.044**	**0.044**	**0.045**	**0.046**	**0.048**	**0.051**

Evaluation

Mean of NMAE for DSIM dataset. Bold values indicate highest accuracy.

Method	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
BPCA	0.046	0.047	0.049	0.051	0.052	0.053	0.055	0.057	0.059	0.061
EM	0.057	0.054	0.053	0.053	0.053	0.053	0.055	0.056	0.058	0.060
Hot deck	0.053	0.055	0.057	0.059	0.063	0.069	0.081	0.095	0.108	0.116
Inpaint	73.4	79.8	82.0	81.7	88.4	89.9	98.8	100.9	105.3	109.1
k-NN	0.044	0.046	0.047	0.049	0.051	0.056	0.064	0.076	0.088	0.098
MEI	0.179	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178
MICE	0.063	0.065	0.065	0.065	0.068	0.069	0.072	0.074	0.076	0.081
Fourier	0.048	0.049	0.050	0.049	0.051	0.051	0.053	0.053	0.056	0.058
L*k*-NN	**0.041**	0.042	0.043	0.044	0.046	0.046	0.047	0.048	0.051	0.056
FL*k*-NN	**0.041**	**0.041**	**0.042**	**0.043**	**0.044**	**0.044**	**0.045**	**0.046**	**0.048**	**0.051**

Mean of NMAE for NICU dataset.

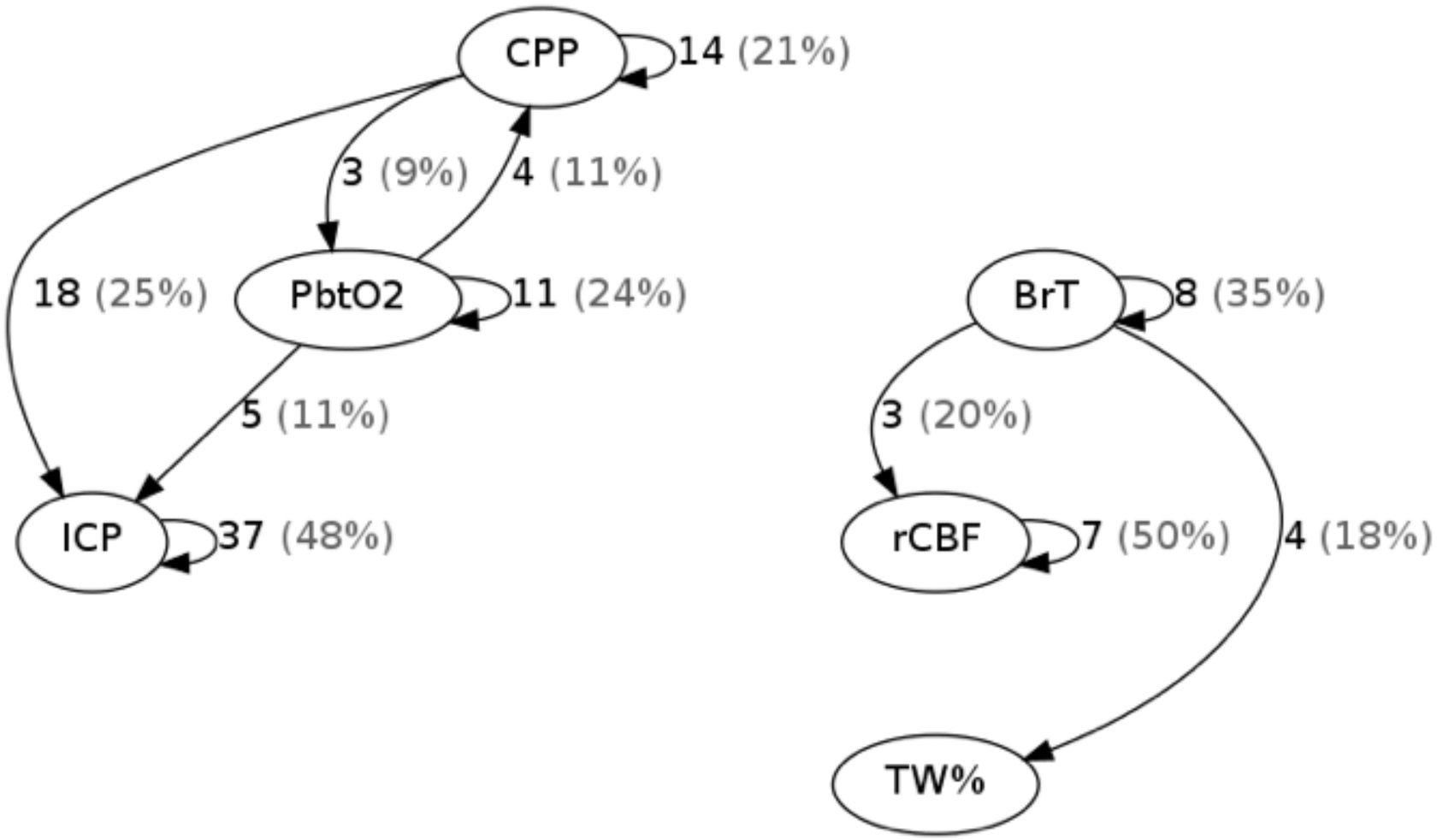
Method	10%	15%	20%	25%	30%	35%	40%	45%	50%
BPCA	0.082	0.124	0.146	0.177	0.197	0.203	0.190	0.190	0.199
EM	0.046	0.049	0.049	0.051	0.053	0.055	0.057	0.060	0.062
Hot deck	0.026	0.031	0.039	0.049	0.064	0.078	0.091	0.101	0.110
Inpaint	1.410	1.491	1.569	1.642	1.731	1.798	1.852	1.950	2.017
k-NN	0.024	0.027	0.031	0.036	0.045	0.055	0.066	0.076	0.084
MEI	0.089	0.092	0.091	0.091	0.091	0.091	0.091	0.091	0.091
MICE	0.057	0.062	0.064	0.066	0.069	0.073	0.076	0.080	0.084
Fourier	0.025	0.025	0.027	0.028	0.030	0.030	0.032	0.035	0.040
L*k*-NN	**0.019**	0.022	0.024	0.027	0.029	0.032	0.035	0.039	0.044
FL*k*-NN	0.020	**0.021**	**0.022**	**0.024**	**0.026**	**0.027**	**0.030**	**0.033**	**0.038**

Evaluation

Mean of NMAE for additional 5% simulated data on NICU and DMITRI dataset.

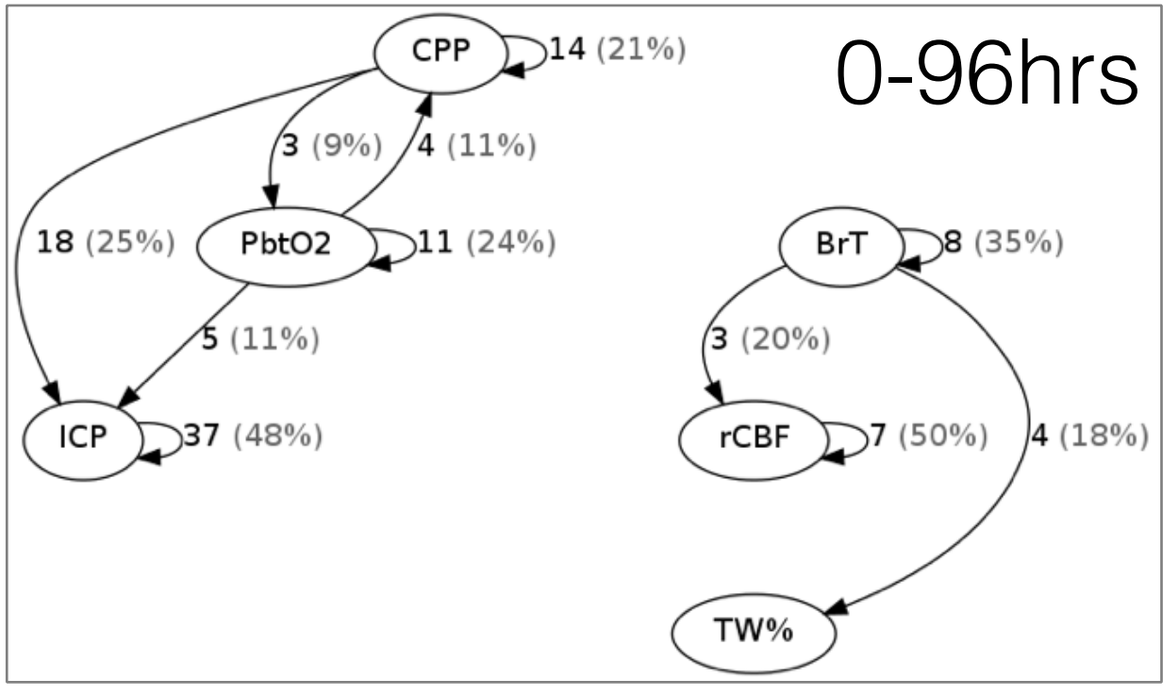
	NICU	DMITRI
BPCA	0.047	0.072
EM	0.048	0.07
Hot deck	0.029	0.073
Inpaint	1.42	42.95
k-NN	0.026	0.064
MEI	0.092	0.106
MICE	0.06	0.101
Fourier	0.0258	**0.04**
L*k*-NN	0.019	0.063
FL*k*-NN	**0.018**	0.045

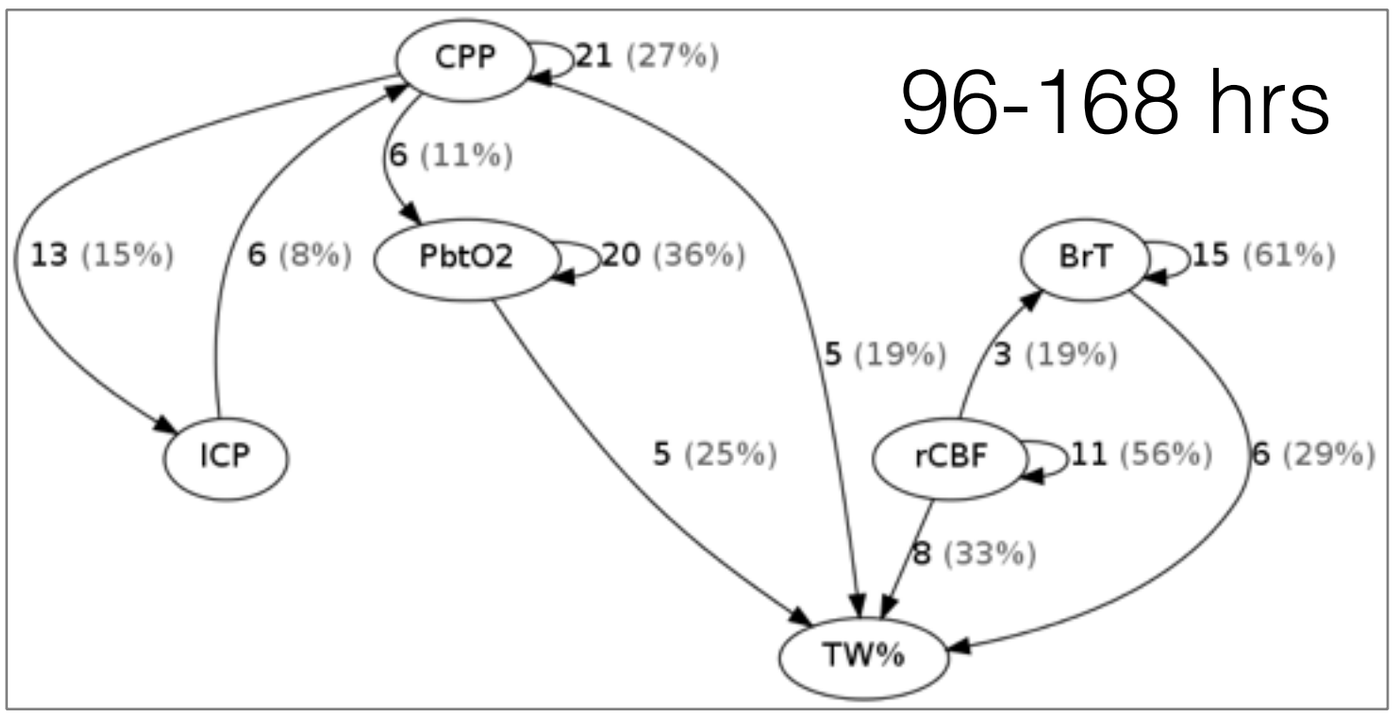
Different data, different physiology



Claassen J, Rahman SA, Huang Y, Frey H, Schmidt M, Albers D, Falo CM, Park S, Agarwal S, Connolly ES, Kleinberg S (2016) Causal structure of brain physiology after brain injury. PLoS ONE (in press).

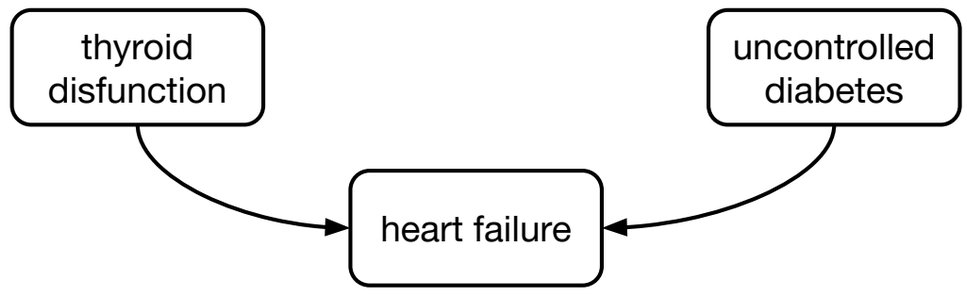
Regulation changes over time



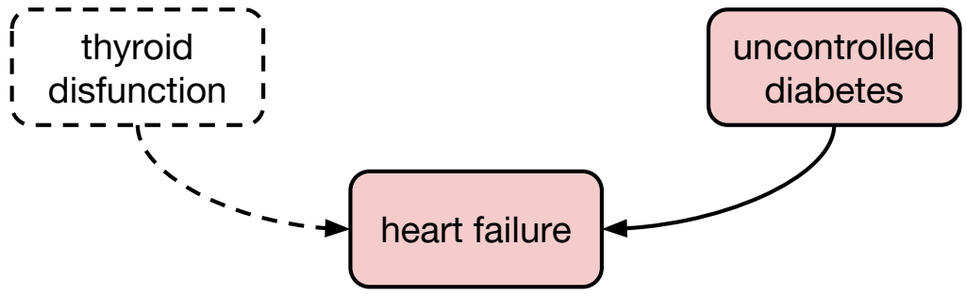


Explanation and Inference

- Causal inference operates on the type-level

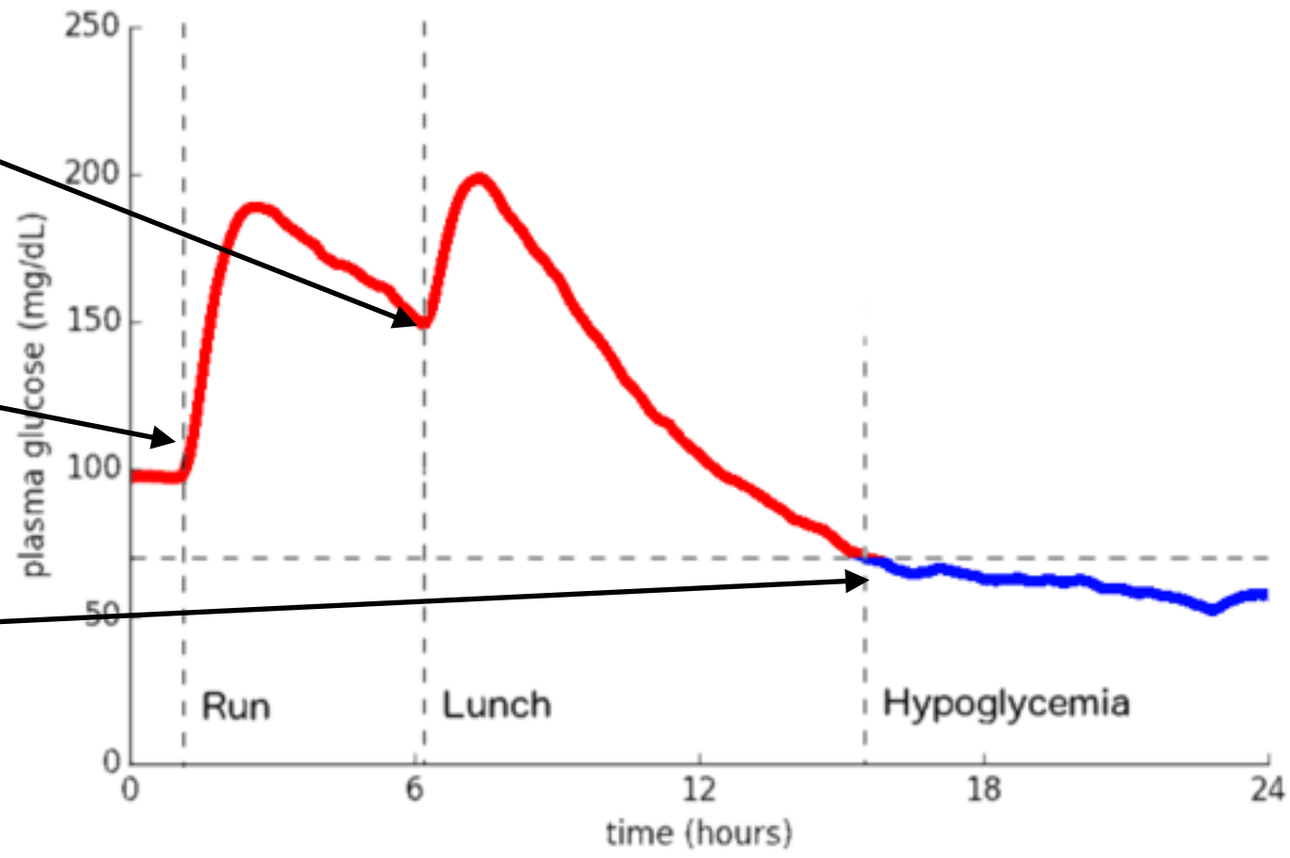


- Causal explanation explains particular events



Example

- Frank uses a CGM to help manage type 1 diabetes
- He goes for a run first thing in the morning
- He has lunch at 12pm, with a normal insulin bolus
- Several hours later, he has unexpected low blood sugar
- **Did the morning run cause the low blood sugar?**



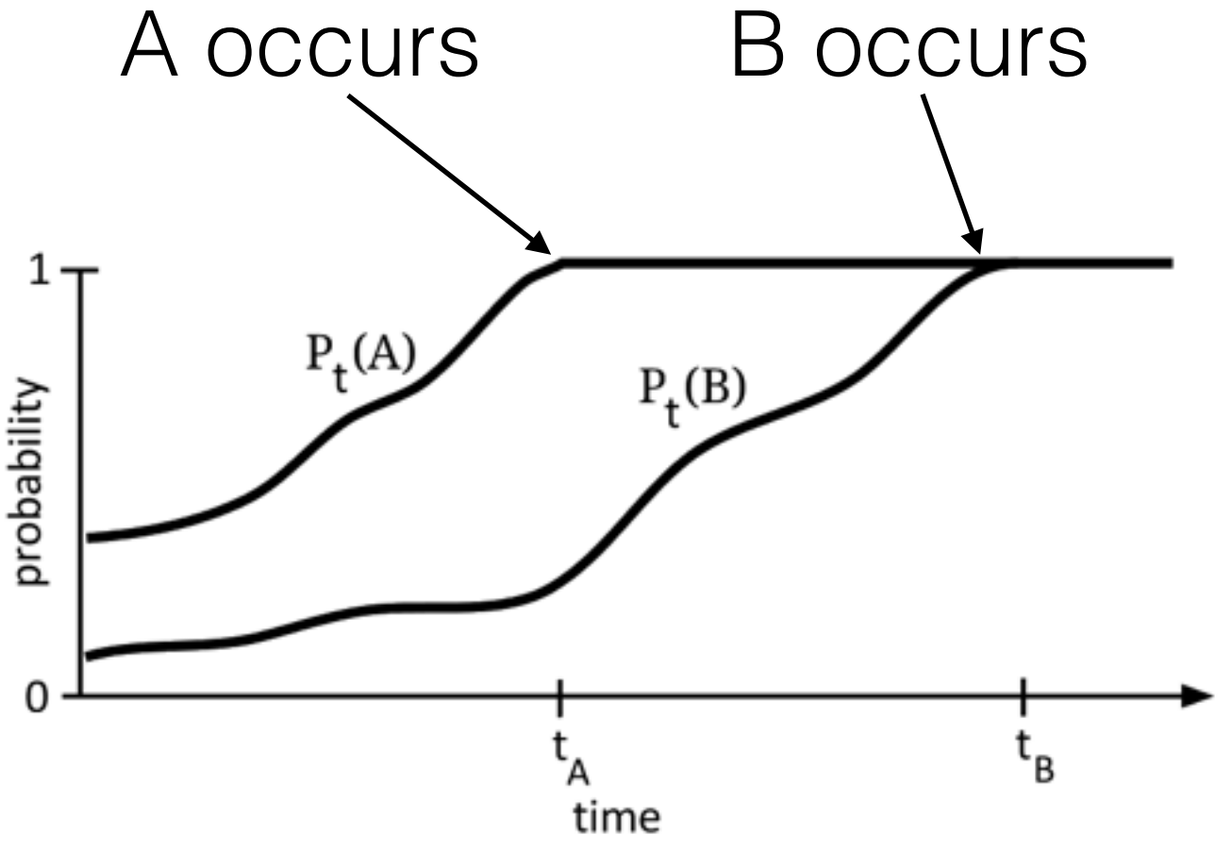
- Goals for explanation
    - Find causes of specific events automatically (no human in the loop)
    - Find causes of when, whether and how events occur
- Approach: simulation to answer counterfactual queries

C. Merck and S. Kleinberg. Causal explanation under indeterminism: A sampling approach. AAAI, 2016.

Events and Probability Trajectories

- events = sets of possible worlds

- $P_t(A)$ is a *probability trajectory*



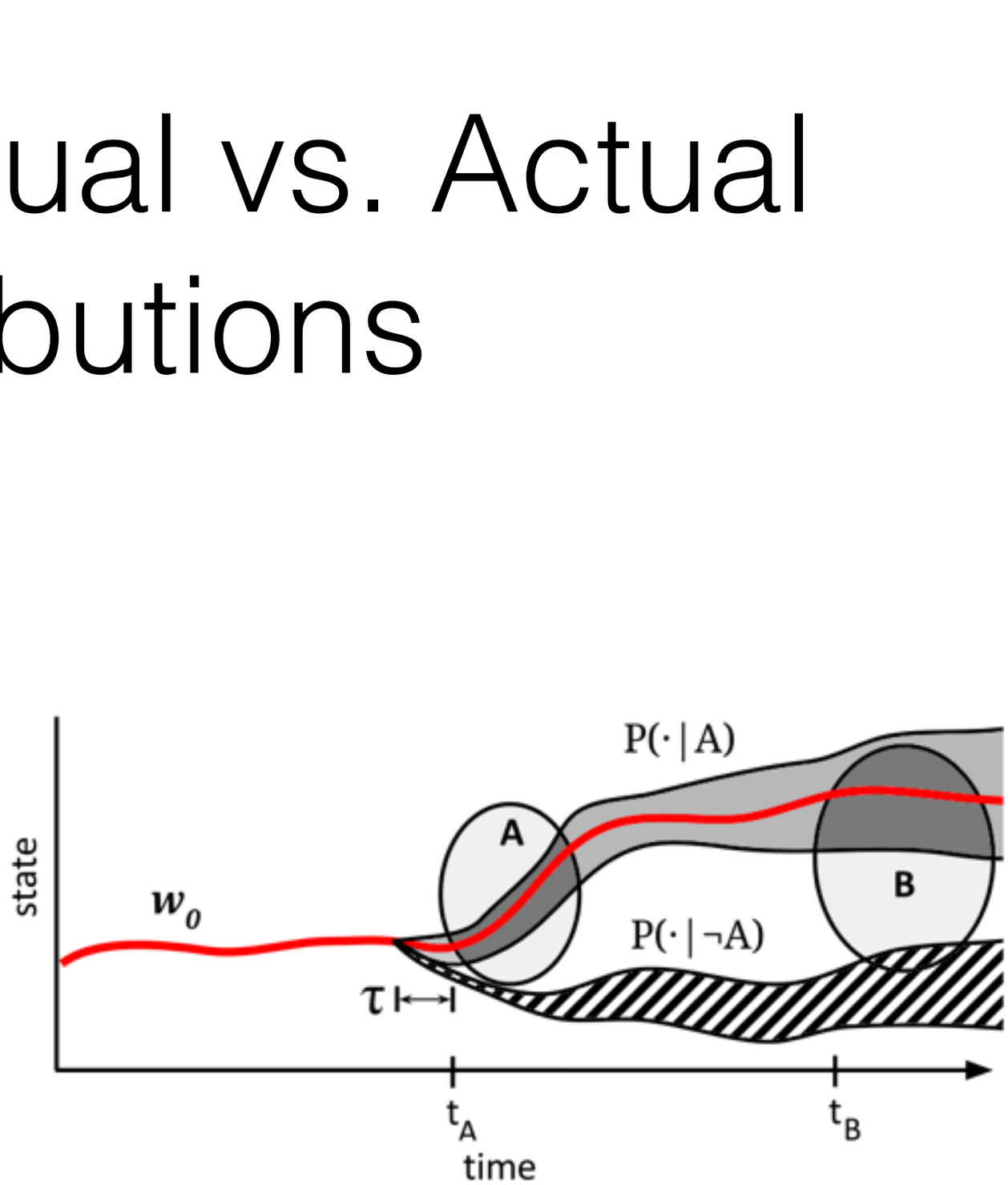
Counterfactual vs. Actual Distributions

- P(•|¬A) is the *counterfactual distribution* of A

$$P(B|\neg A) = \frac{P_{t_A-\tau}(B \cap \neg A)}{P_{t_A-\tau}(\neg A)}$$

- P(•|A) is the *actual distribution* of A

$$P(B|A) = \frac{P_{t_A-\tau}(B \cap A)}{P_{t_A-\tau}(A)}$$



Three Types of Explanation

probability:	B **because of** A iff P (B|A) >> P (B|¬A)	B **despite** A iff P (B|A) << P (B|¬A)
timing:	B **hastened by** A iff $E[t_B|A] << E[t_B|\neg A]$	B **delayed by** A iff $E[t_B|A] >> E[t_B|\neg A]$
intensity:	B **intensified by** A iff $E[m_B|A] >> E[m_B|\neg A]$	B **attenuated by** A iff $E[m_B|A] << E[m_B|\neg A]$

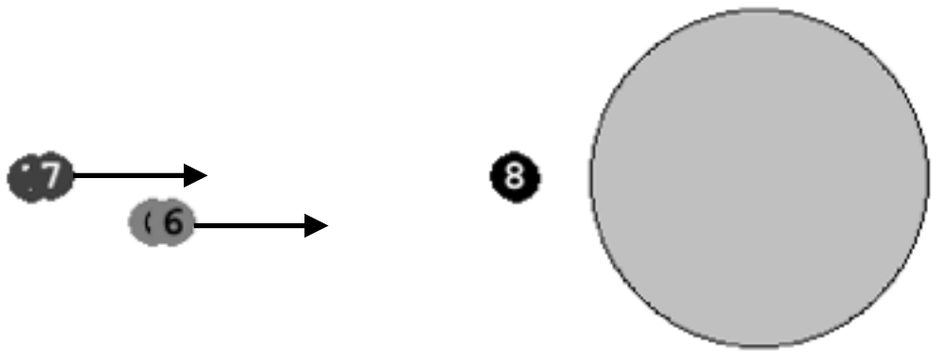
t_B = time of B occurring

m_B = intensity (manner) of B occurring

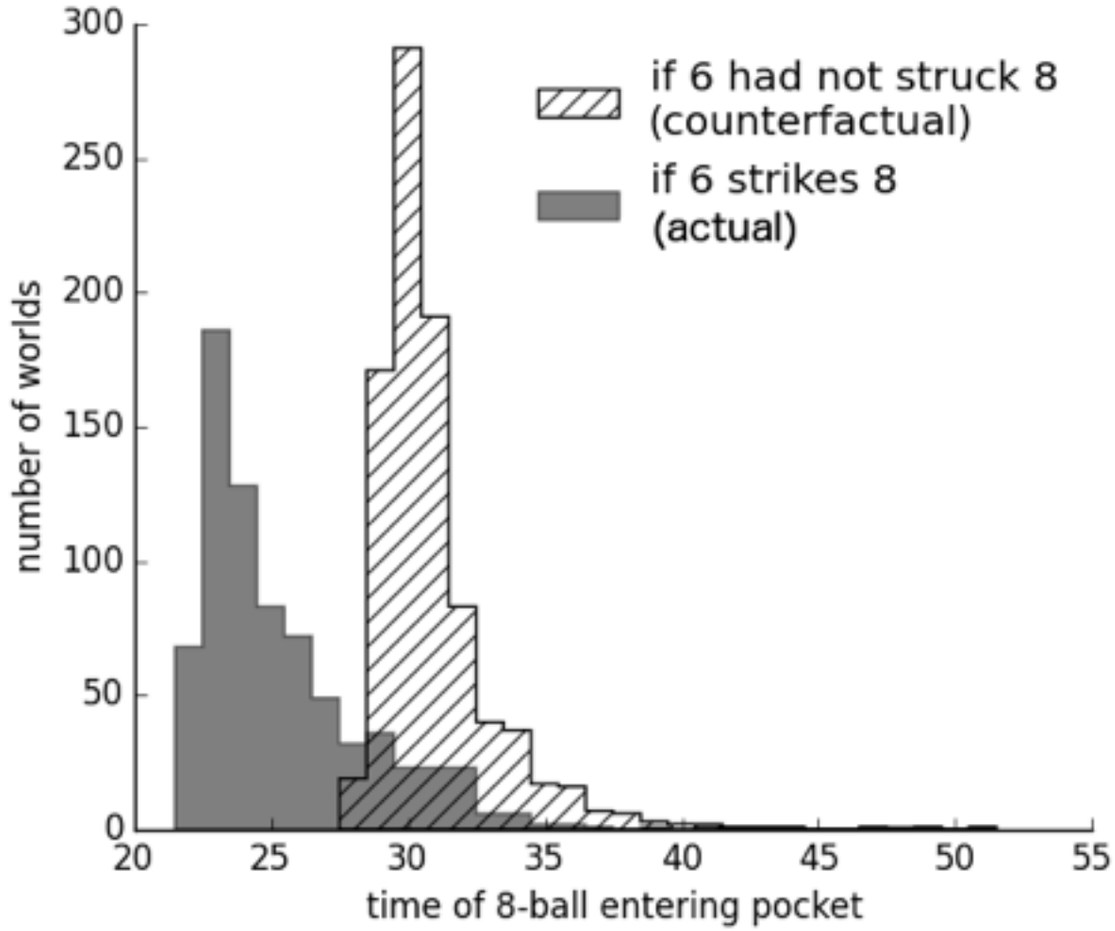
Causal Chain with Billiard Balls

Causal Chain with Billiard Balls

Hastening



- 6->8, then 8->P
- but 7->8 is a more reliable backup
- probability raising finds
  “8->P despite 6->8”
- but by analyzing timing we find
  “6->8 hastened 8->P”



Diabetes Simulation

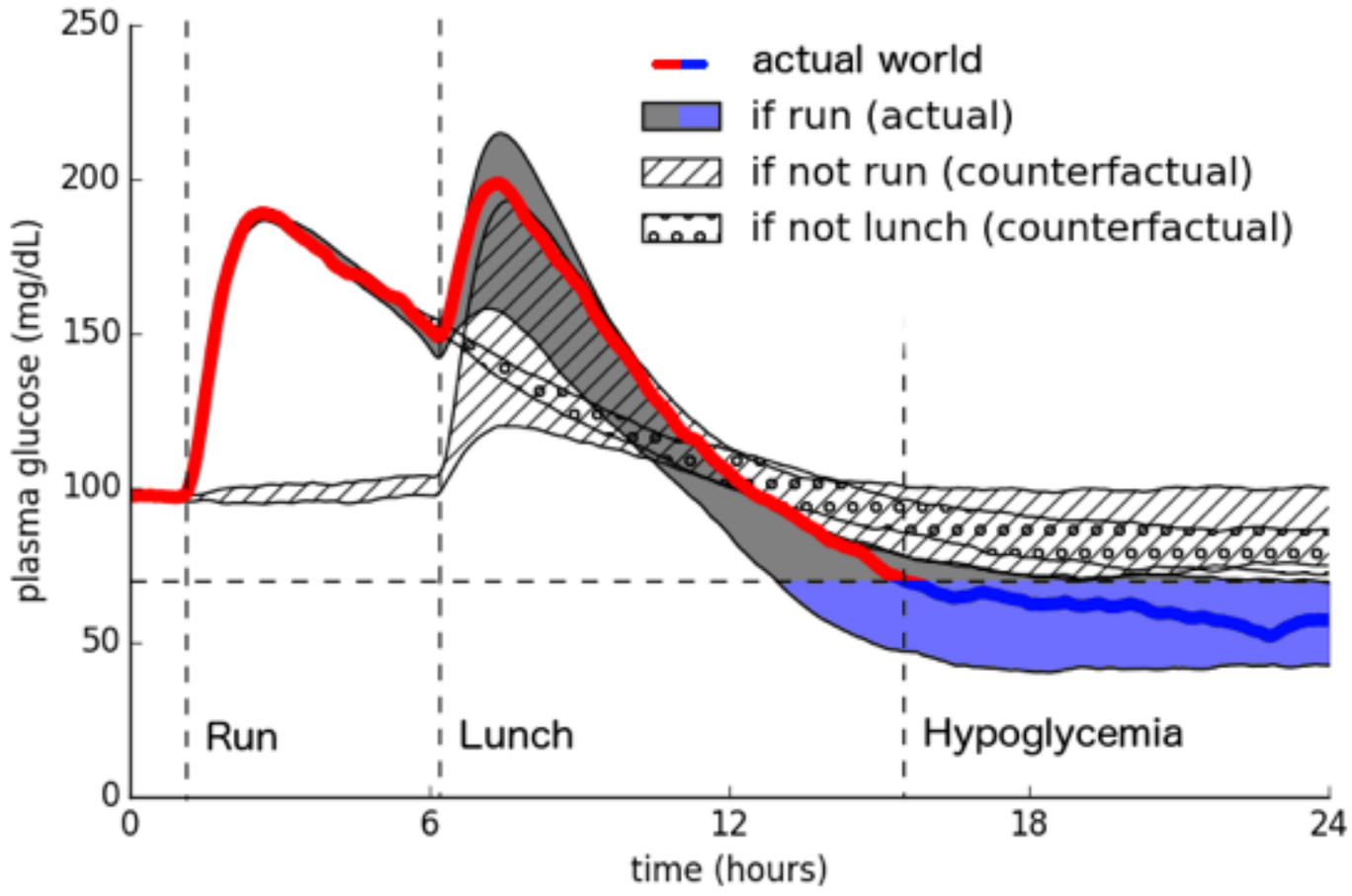
- Run or Lunch alone would not have caused Hypoglycemia (see counterfactual dists)

- Yet together they explain the Hypoglycemia (see actual distribution)

- We see beyond the most recent event (Lunch)

- We can measure quantitative strength of effect in mg/dL:

  E[glu | R] - E[glu | ¬R]



Results

Scenario	Relationship	Probability	Timing
common cause	true	+0.334*	-
	spurious	-0.011	-0.08
causal chain	direct	+0.527*	-8.91*
	indirect	+0.262*	-7.56*
backup cause	true	-0.147*	-5.50*

Results

Scenario	Relationship	Probability	Timing
common cause	true	+0.334*	-
	spurious	-0.011	-0.08
causal chain	direct	+0.527*	-8.91*
	indirect	+0.262*	-7.56*
backup cause	true	-0.147*	-5.50*

Relationship	Probability	Timing	Intensity
$R \rightarrow H$	+0.642*	-3.14h*	+27.5 mg/dL*
$L \rightarrow H$	+0.721*	-10.3h*	+28.0 mg/dL*

Mo data mo problems

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

Open problems

- Finding the “right” variables
- Causality beyond variables
- Nonstationarity at multiple scales
- Combining results + datasets
- Uncertainty in data and timing
- Uncovering hidden assumptions