Health +AILab

Grand Challenges for Causal Inference

Samantha Kleinberg
Stevens Institute of Technology

SECTIONS

The New York Times

Well

EAT

Coffee May Protect Against Cancer, W.H.O. Concludes

By ANAHAD O'CONNOR JUNE 15, 2016 8:58 AM 104

Email

Share

Tweet

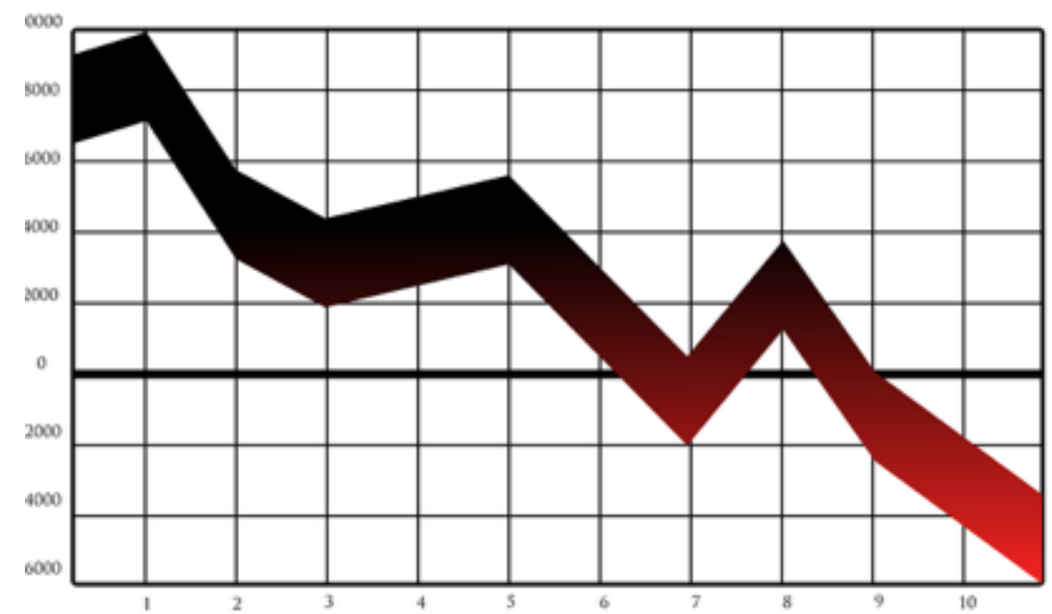
Save

More

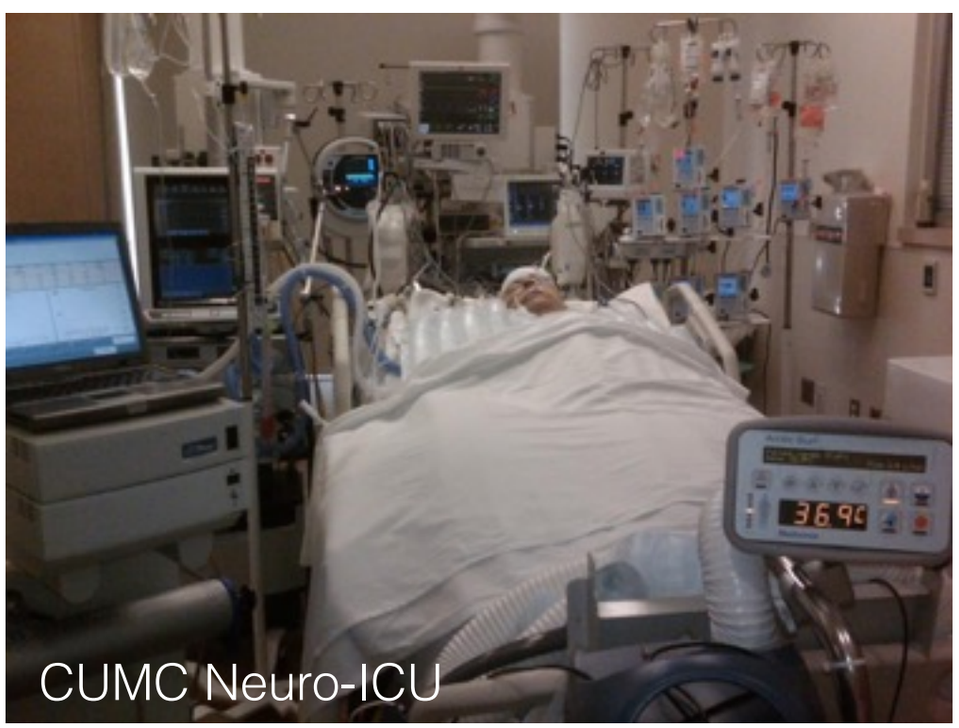
An influential panel of experts convened by the World Health Organization concluded on Wednesday that regularly drinking coffee could protect against at least two types of cancer, a decision that followed decades of research pointing to the beverage's many health benefits. The panel also said there was a lack of evidence that it might cause other types of cancer.

Fred R. Conrad for The New York Times

Does coffee prevent cancer?



Why did the market crash?



What causes stroke?

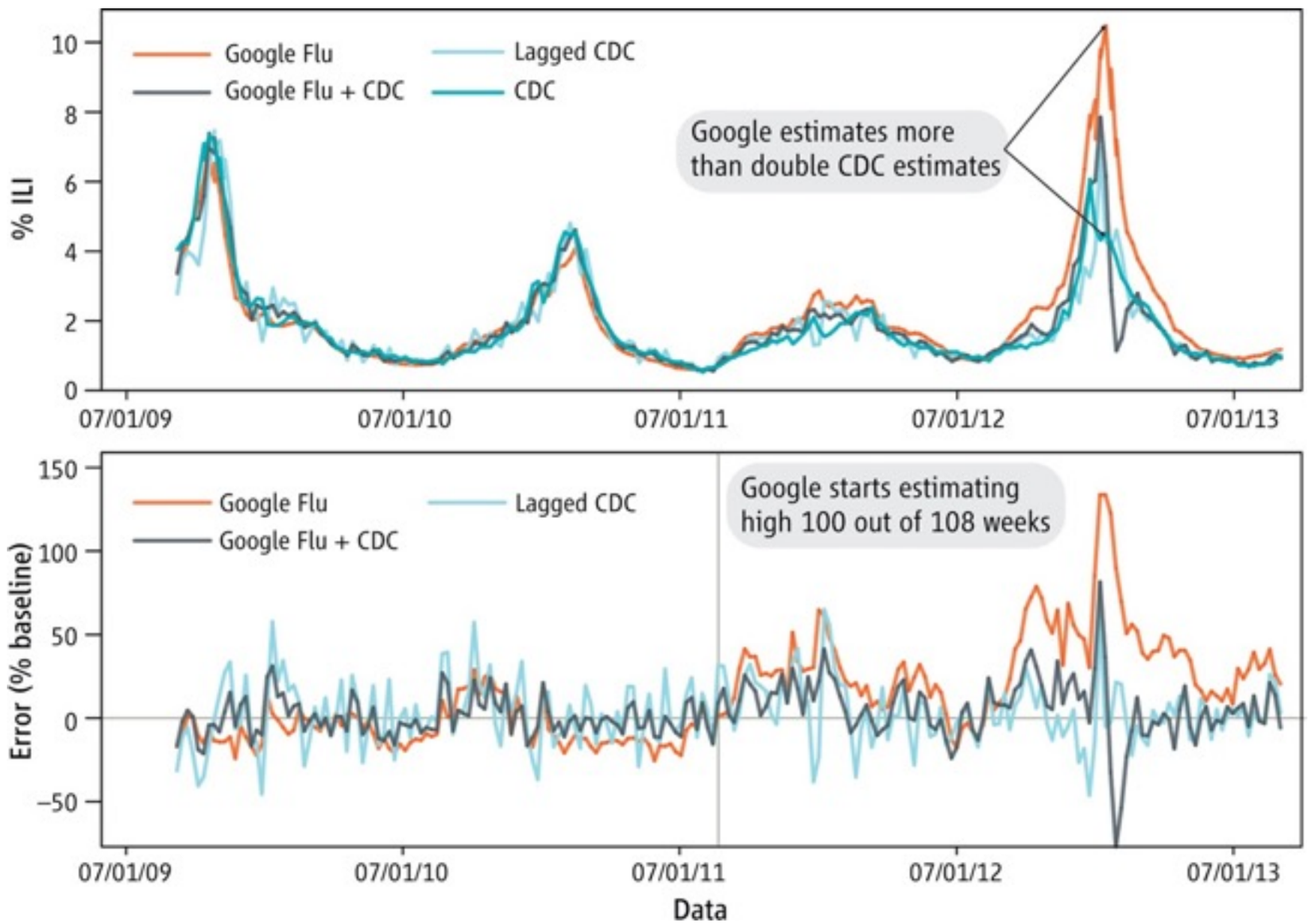
Who's at fault?

Why causality now?

“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).

Predict



David Lazer et al. Science 2014;343:1203-1205

Published by AAAS

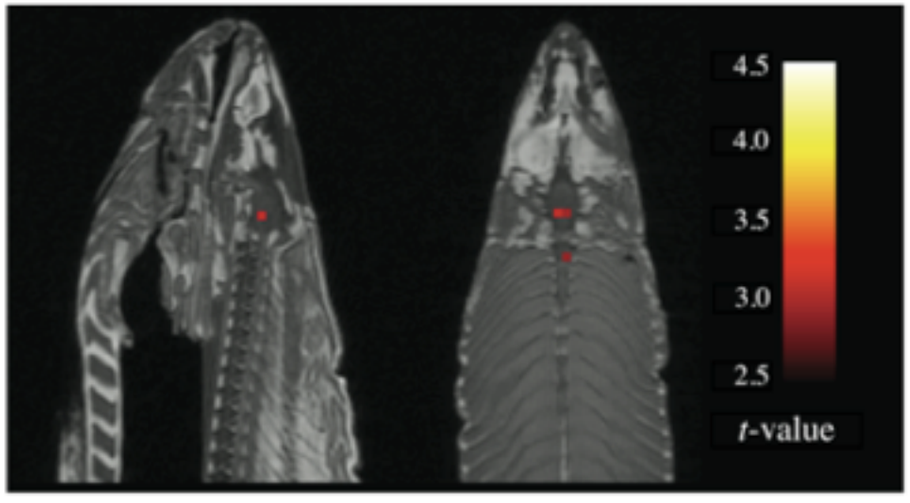
Multiple testing

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.



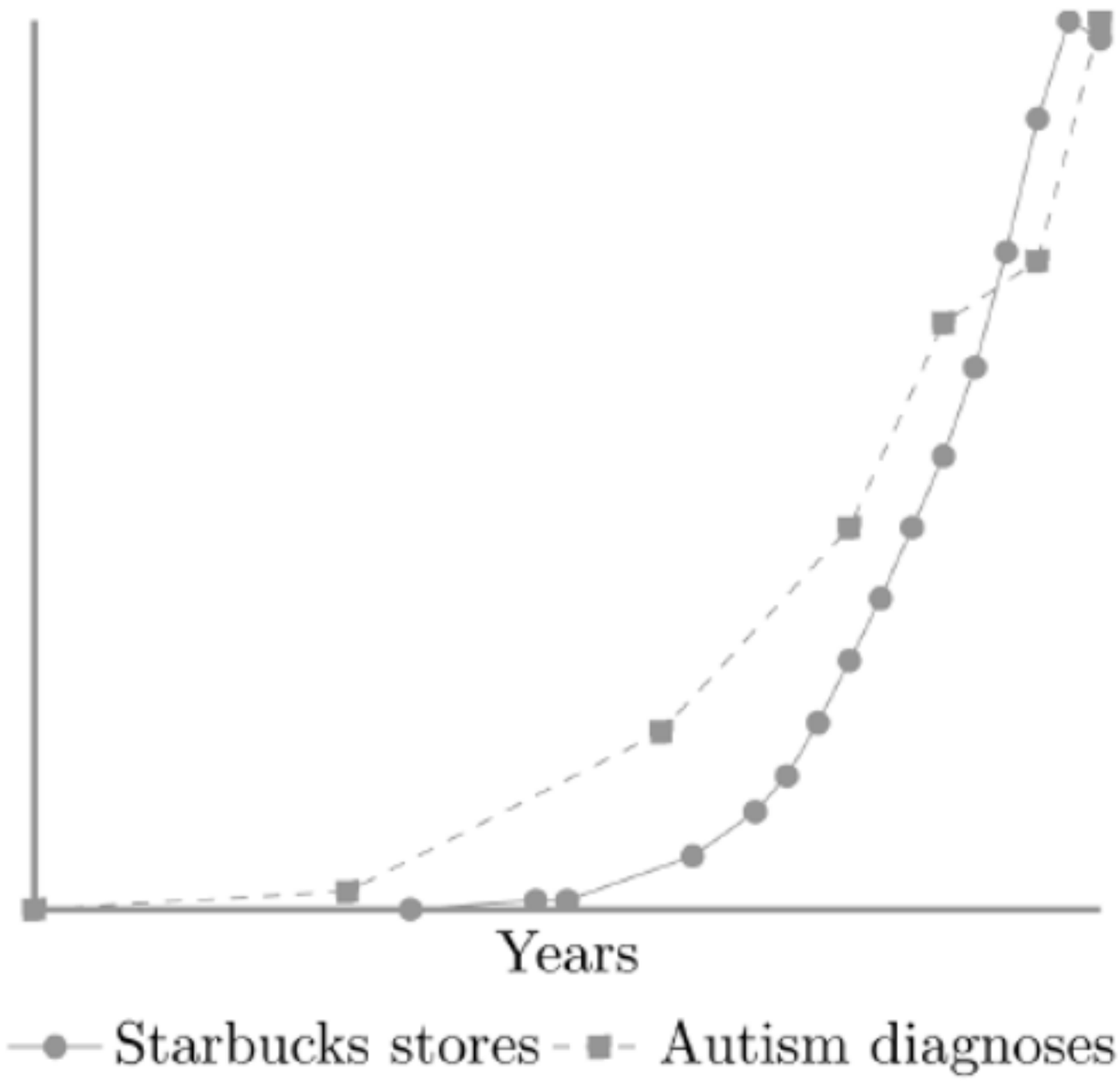
A t-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

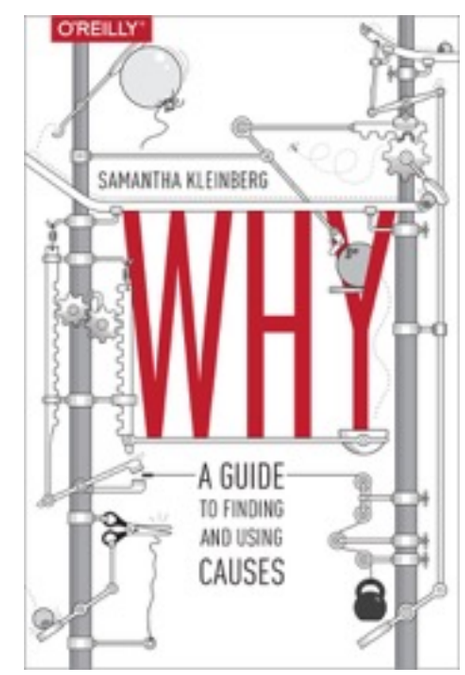
Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of 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 ($p = 0.25$).

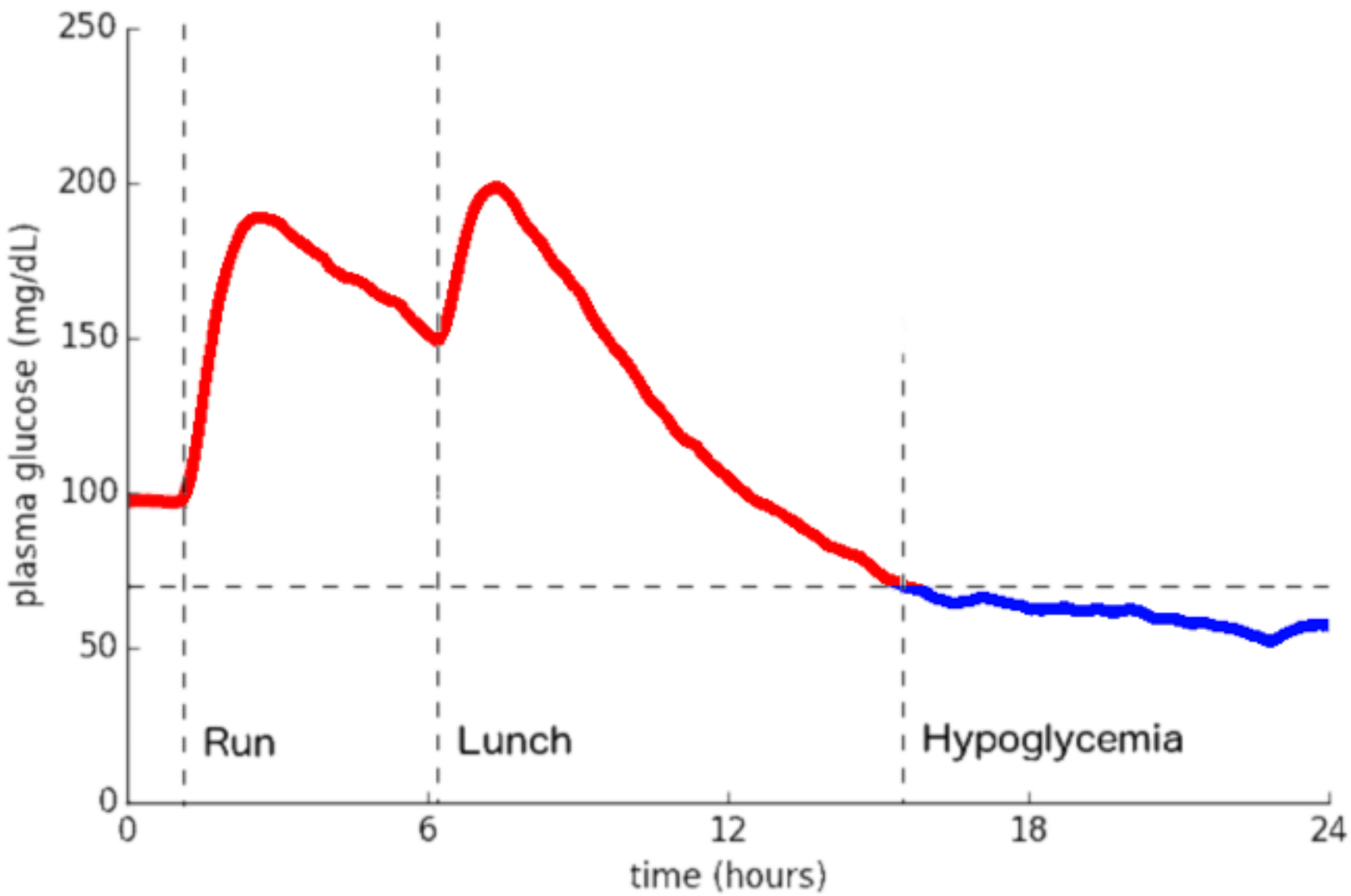
Bennett, C. M., Miller, M. B., & Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for multiple comparisons correction. NeuroImage, 47(1), 125.

Explain


Years
Starbucks stores
Autism diagnoses


O'REILLY
SAMANTHA KLEINBERG
WHY
A GUIDE
TO FINDING
AND USING
CAUSES

Explain



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

Control

- Reducing class sizes in Tennessee led to better scores on standardized tests

- Will it work in California?

Bohrnstedt, G. W. and Stecher, B. M. (eds.) (2002). What We Have Learned about Class Size Reduction in California. American Institutes for Research, Palo Alto, CA.

Observational health data

- Experiments often infeasible, unethical, or too expensive
- Routinely collected in many situations
  - Electronic health records in hospitals
  - ICU data streams
  - Body-worn sensors and mobile devices

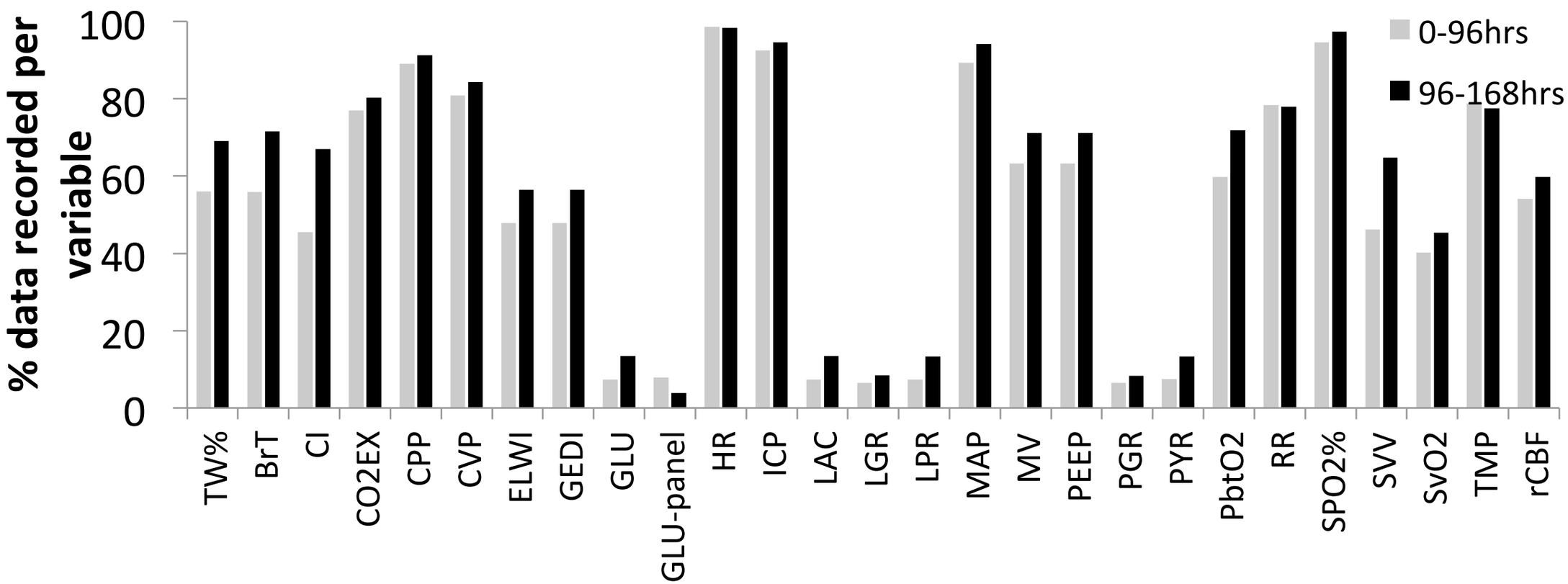
Much big data is uncertain

- What is a diagnosis or causes? Uncertainty in groupings and causal relata
- When is a measurement or causal relationship true? Uncertainty over time
- How do we know what's correct? Uncertainty in ground truth

$$r_{i,t} = \sum_{j} \beta_{ij} f_{j,t'} + \varepsilon_{i,t}.$$

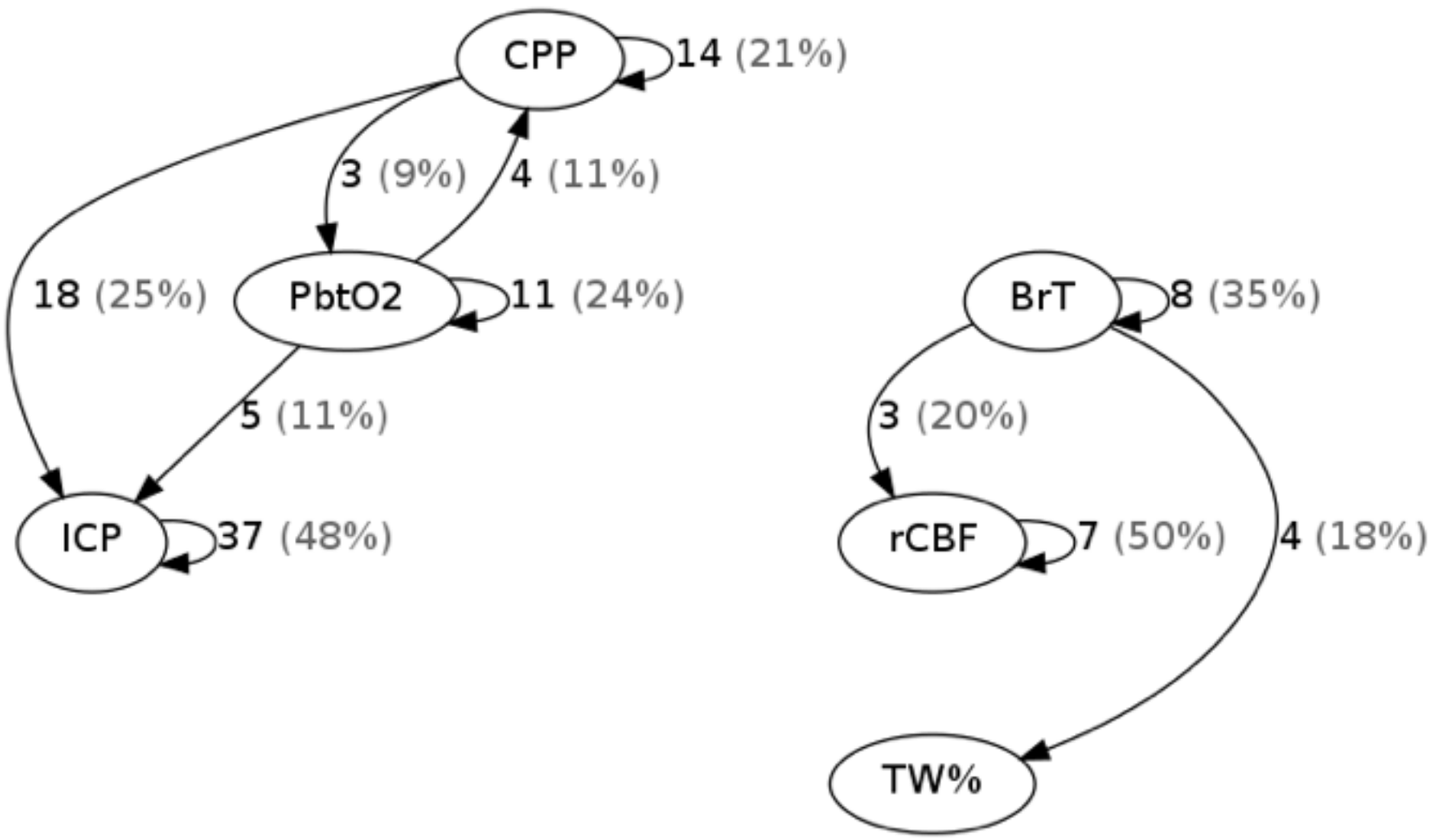
CUMC Neuro-ICU

Lots of data, but lots of missing data



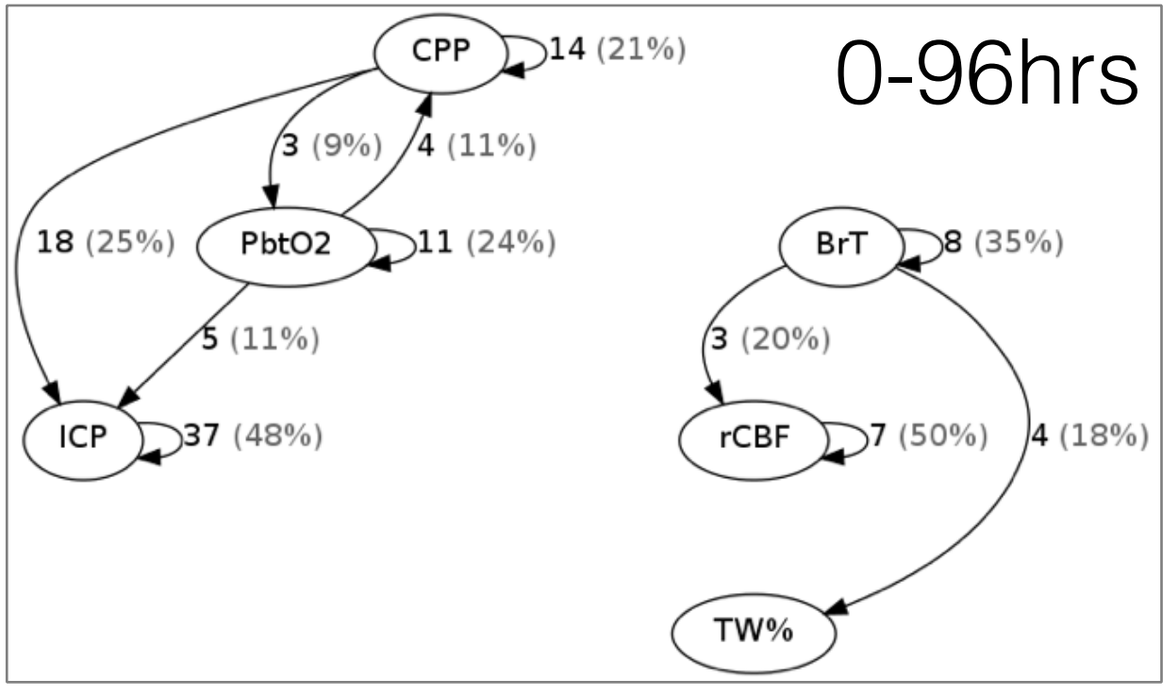
S. A. Rahman, Y. Huang, J. Claassen, N. Heintzman, and S. Kleinberg. Combining Fourier and Lagged k-Nearest Neighbor Imputation for Biomedical Time Series Data. Journal of Biomedical Informatics (2015)

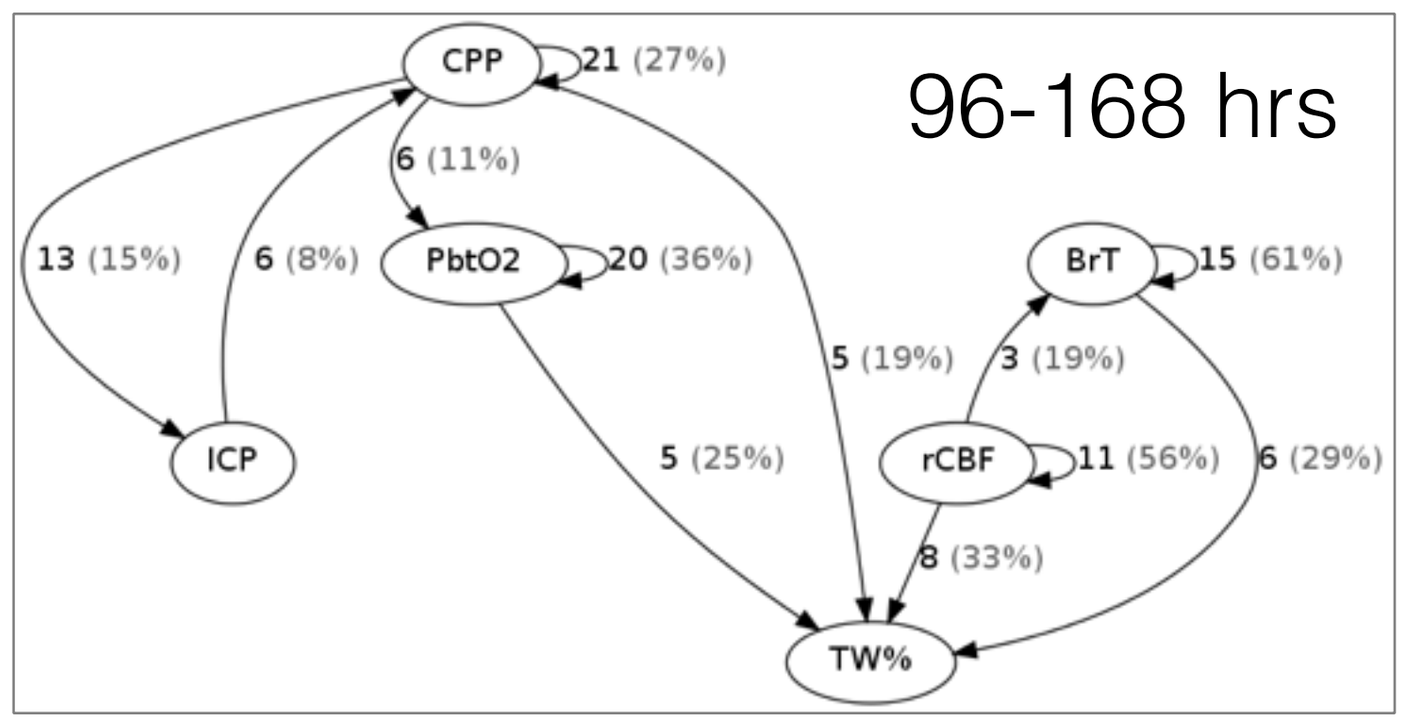
Different data, different physiology



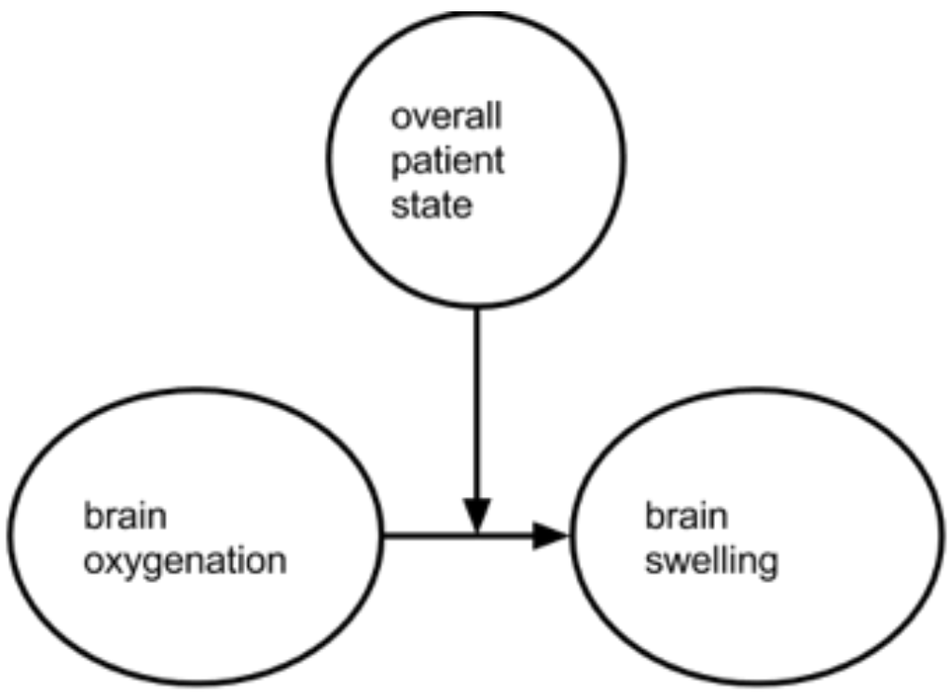
Claassen J, Rahman SA, Huang Y, Frey H, Schmidt M, Albers D, Falo CM, Park S, Agarwal S, Connolly ES, Kleinberg S (2015) Causal structure of brain physiology after brain injury. PLoS ONE.

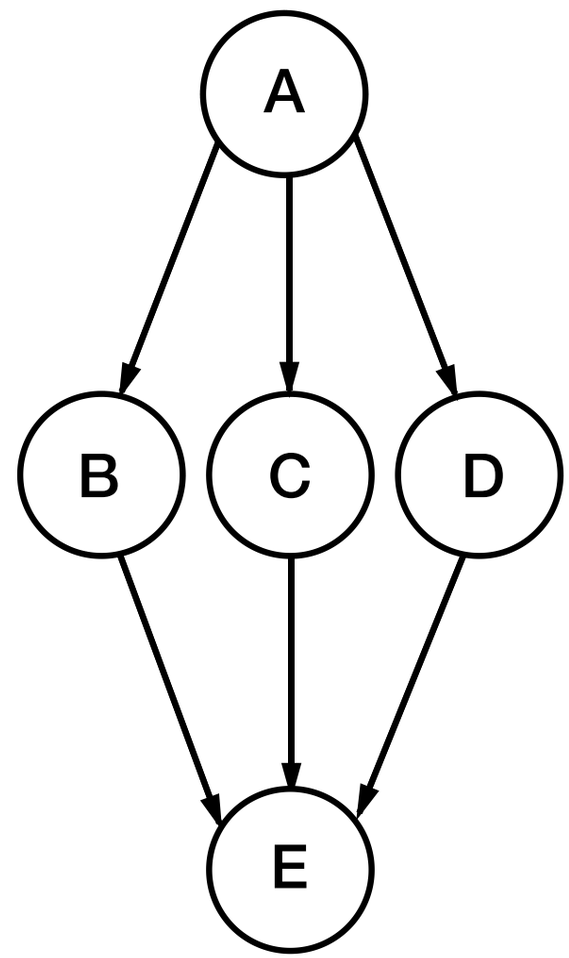
Regulation changes over time

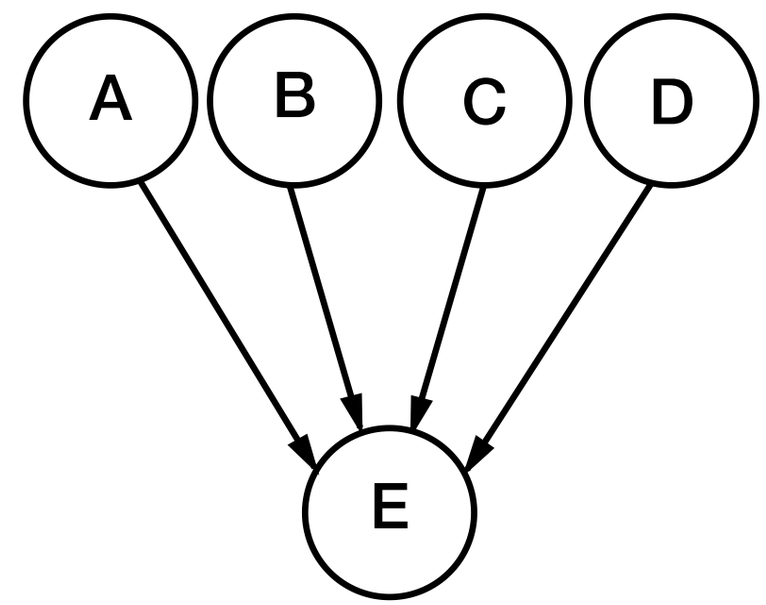


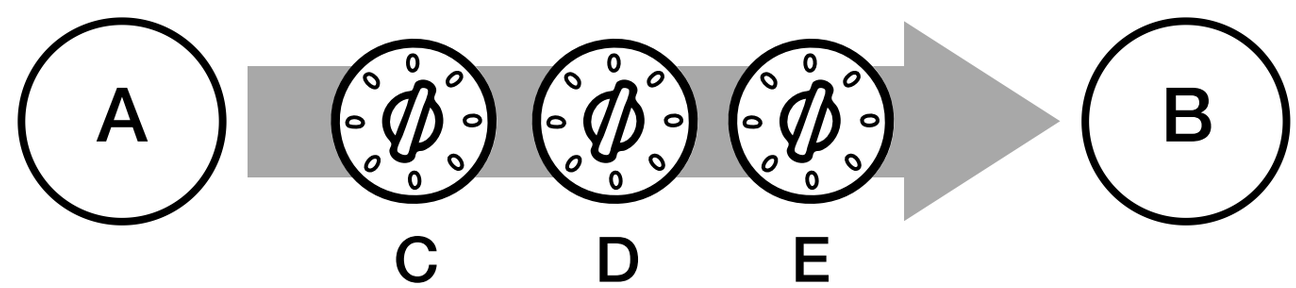


Moving beyond variables





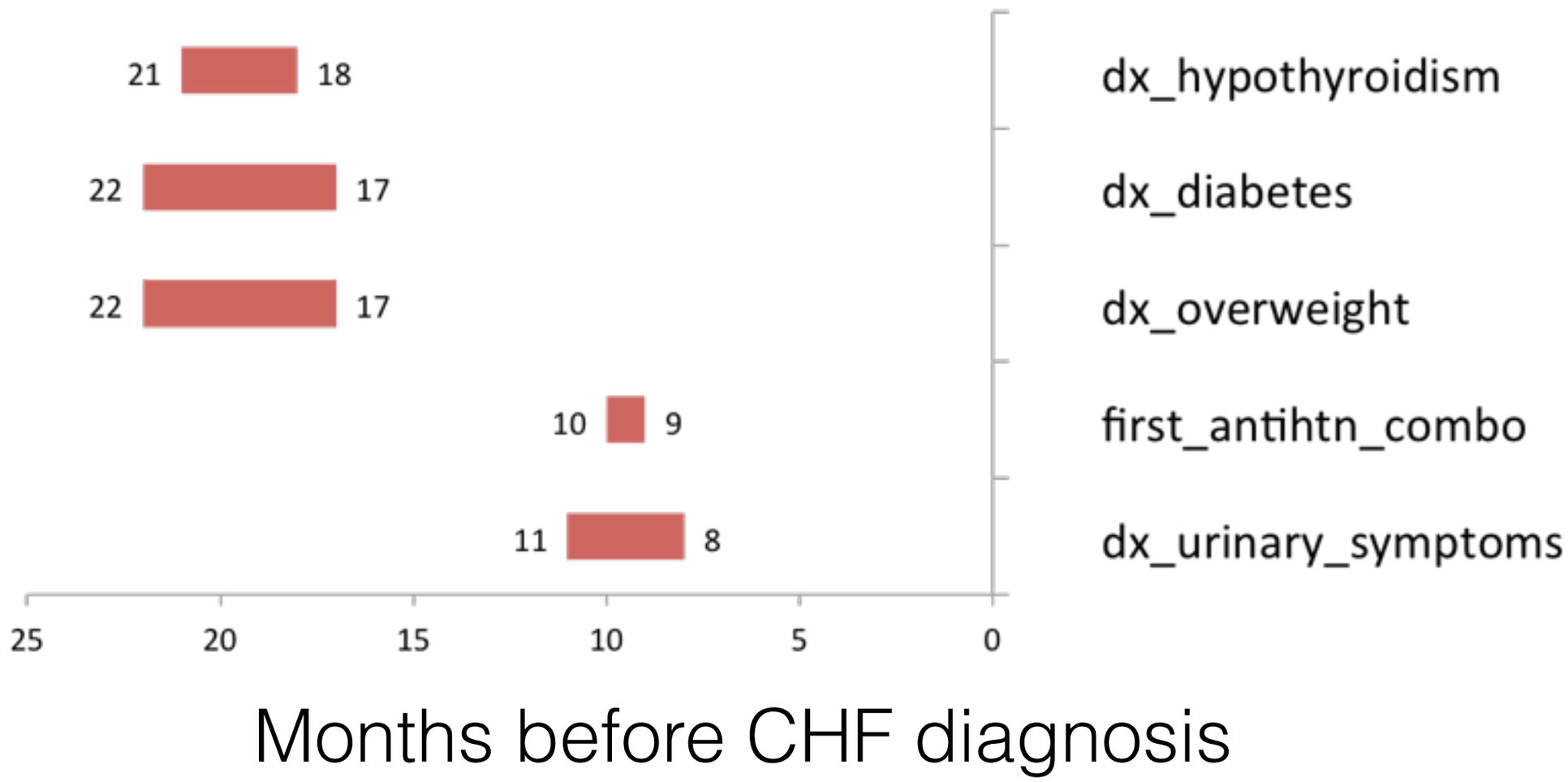




Learning from EHR data

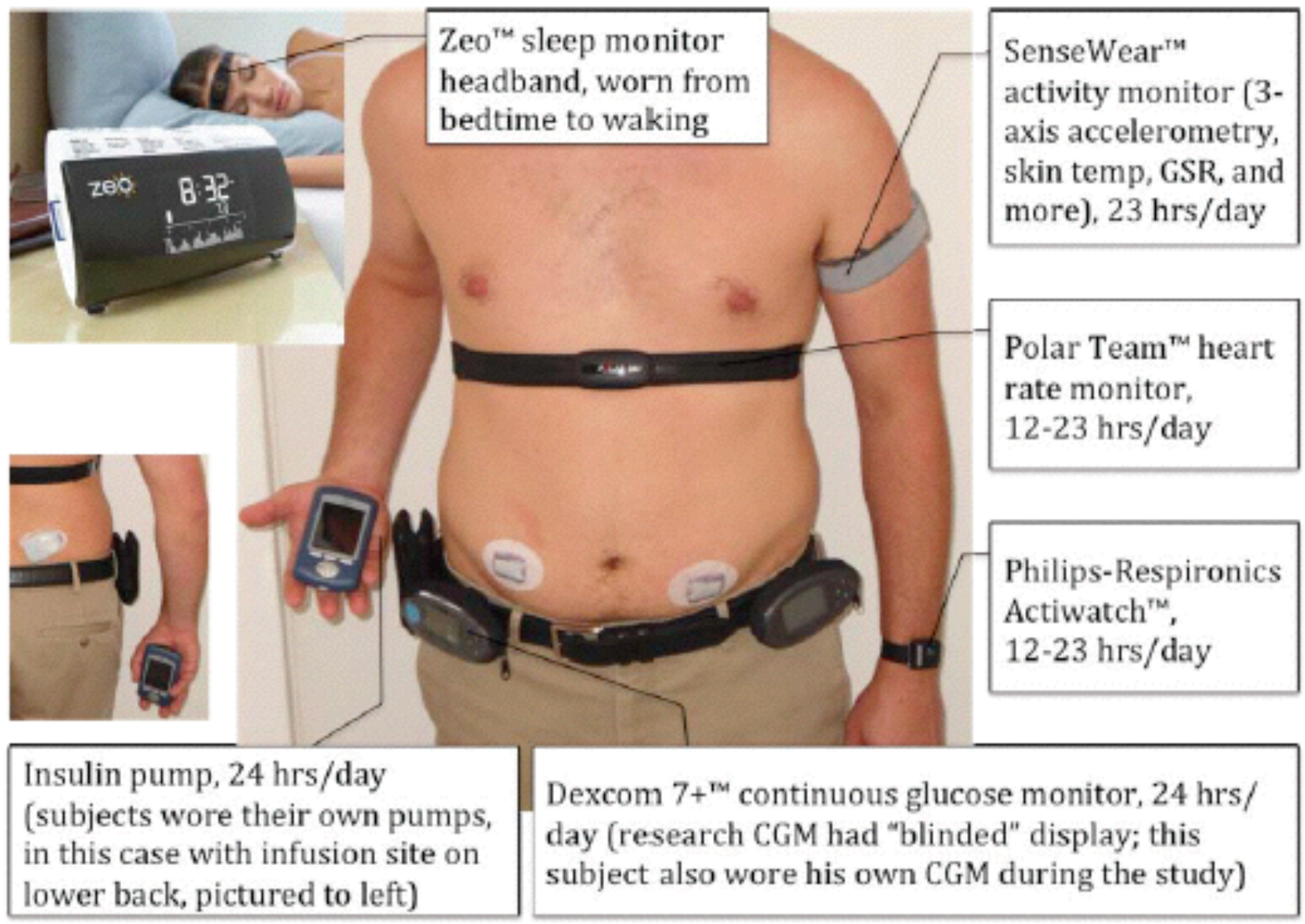
- Which patients have heart failure?
- When does the heart failure start?

Need for expert knowledge



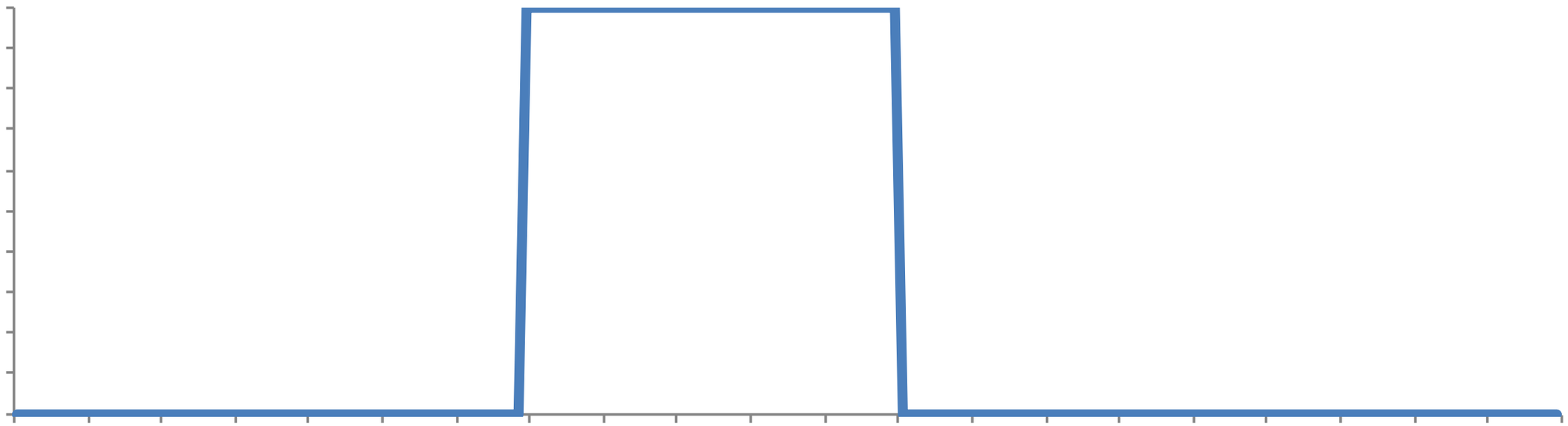
Kleinberg S, Elhadad N (2013) Lessons Learned in Replicating Data-Driven Experiments in Multiple Medical Systems and Patient Populations. In: AMIA Annual Symposium.

Finding causes of hyperglycemia

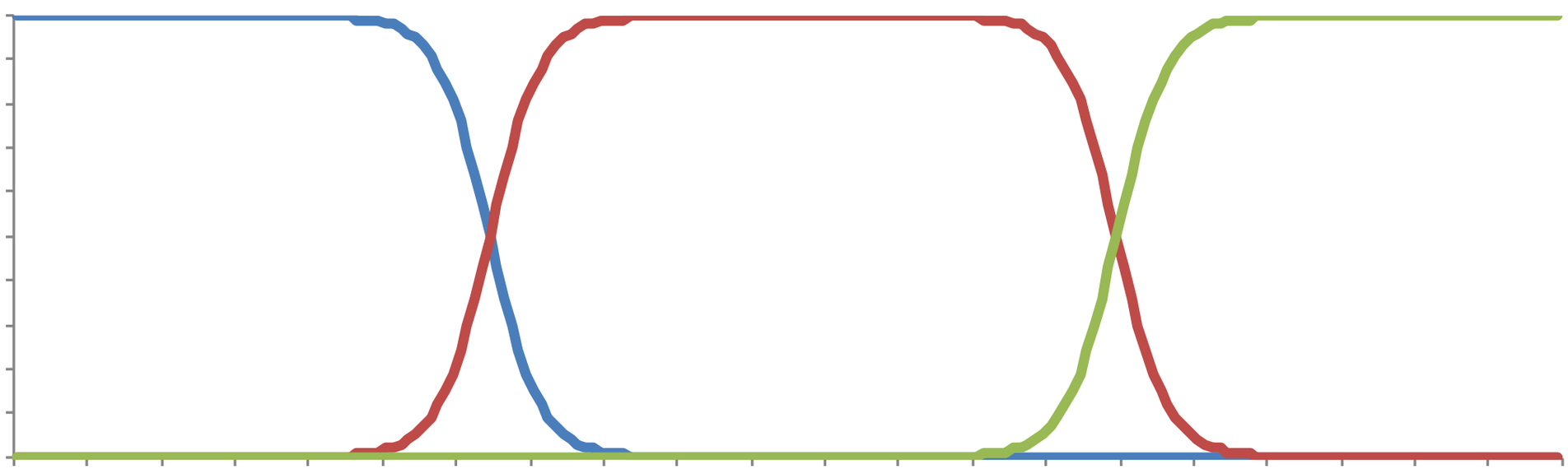


With N. Heintzman (Dexcom) http://dial.ucsd.edu/what-we-do.php

Usual thresholds



Probabilistic discretization



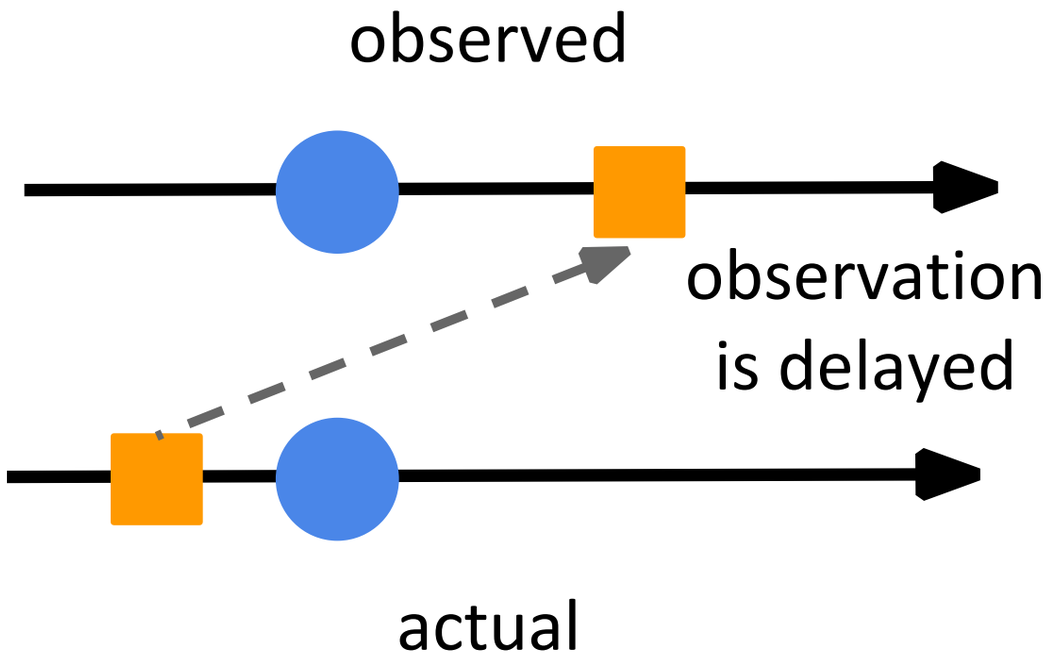
very vigorous exercise leads to hyperglycemia (fdr <.01) in 5-30 minutes

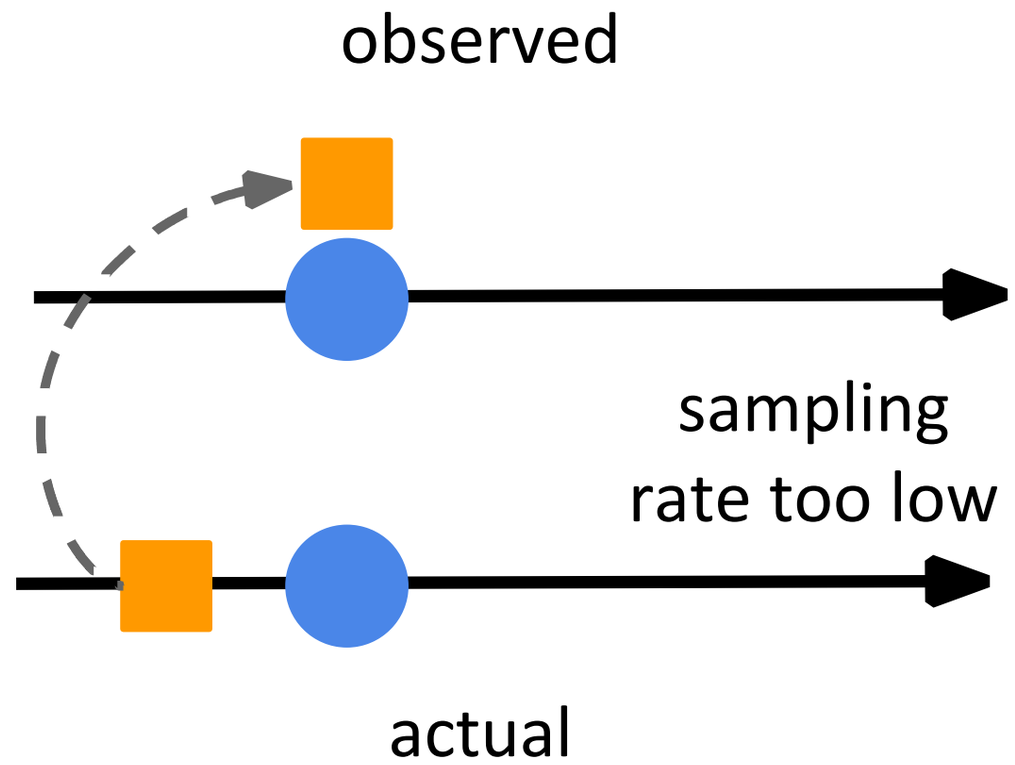
- Found using both HR (anaerobic activity zone) and METs

- Supported by literature (Marliss and Vranic, 2002; Riddell and Perkins, 2006)

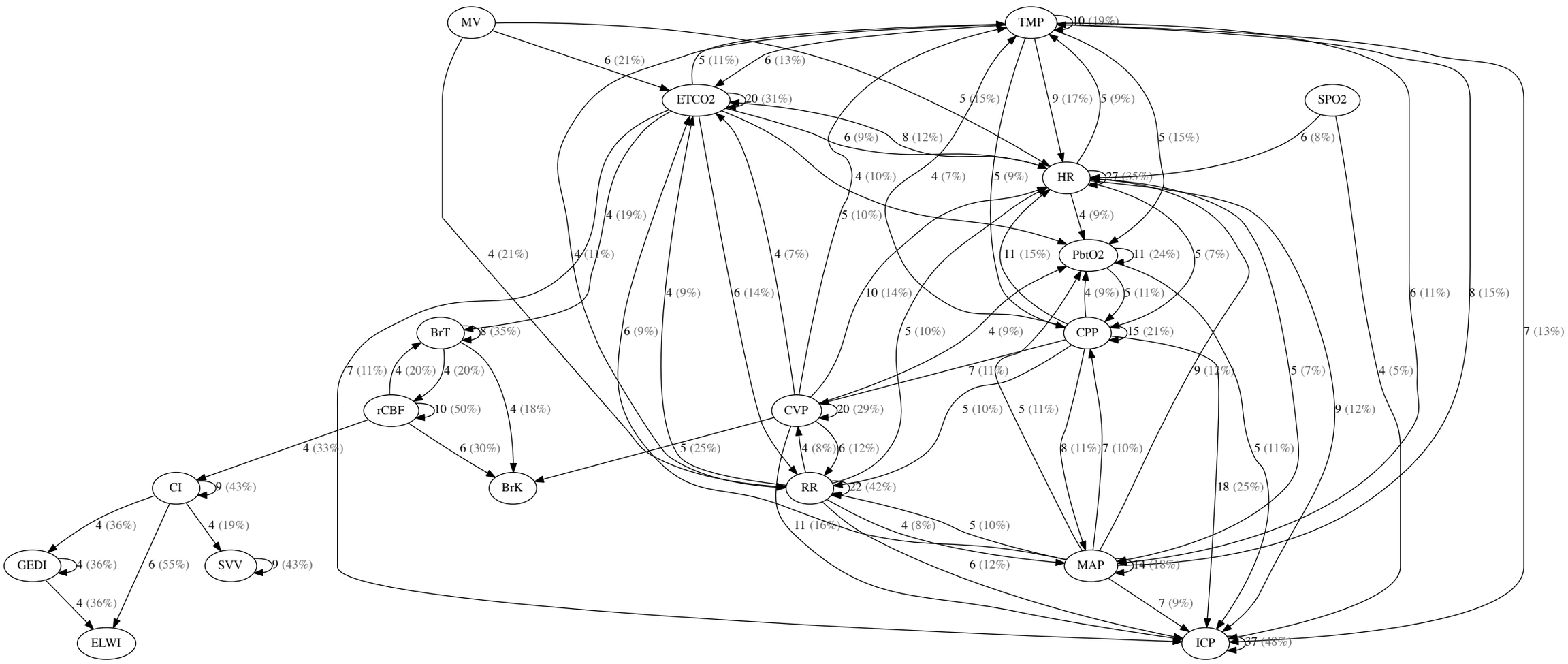
N. Heintzman and S. Kleinberg. Using Uncertain Data from Body-Worn Sensors to Gain Insight into Type 1 Diabetes. Journal of Biomedical Informatics (2016)

What if we can’t trust timing?





Who’s the end user?



A few (of many) challenges

- What are the elements of a causal relationship?
- How can we find causes with uncertain data?
- How to integrate data w/prior knowledge?
- When can we intervene successfully?
- How to communicate complex causes to lay audiences?

Thanks!

Teams @CUMC, Stevens, Geisinger

Funding from: NSF CAREER, NIH R01, JSMF