Machine learning for causality?

@kordinglab

The bitter Lesson (Sutton, 2019)

- Clever human solutions are eventually beat by general purpose machine learning
    - Chess
    - Go
    - Speech recognition
    - Image recognition
    - Natural Language Processing

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably

How good we are not after ML but after causality

- Medicine is almost always after causality
- And ML merely models correlations

3 Schools

- RCT or bust
  - Randomize, avoid threads
  - Very strong strategy for large audience, high value
  - Weak for underserved populations
  - Gold standard (also C. elegans)
- Observational
  - Fit complex models, Correct for various variables
  - Lack of causal validity
- Quasiexperimental
  - Find a place where the world contains randomization
  - Usually not possible

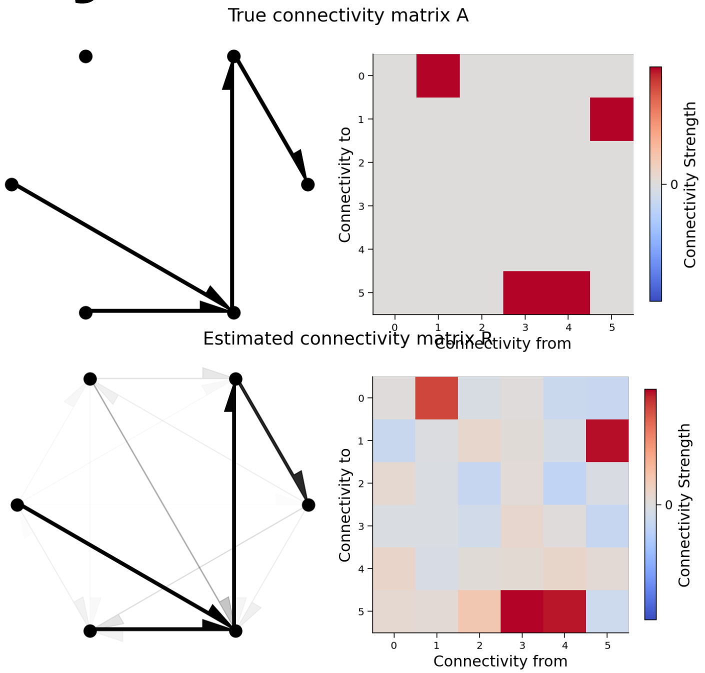
Intuition: Simulate a trivial causal system

$$\vec{x}_{t+1} = \sigma(A\vec{x}_t + \epsilon_t)$$

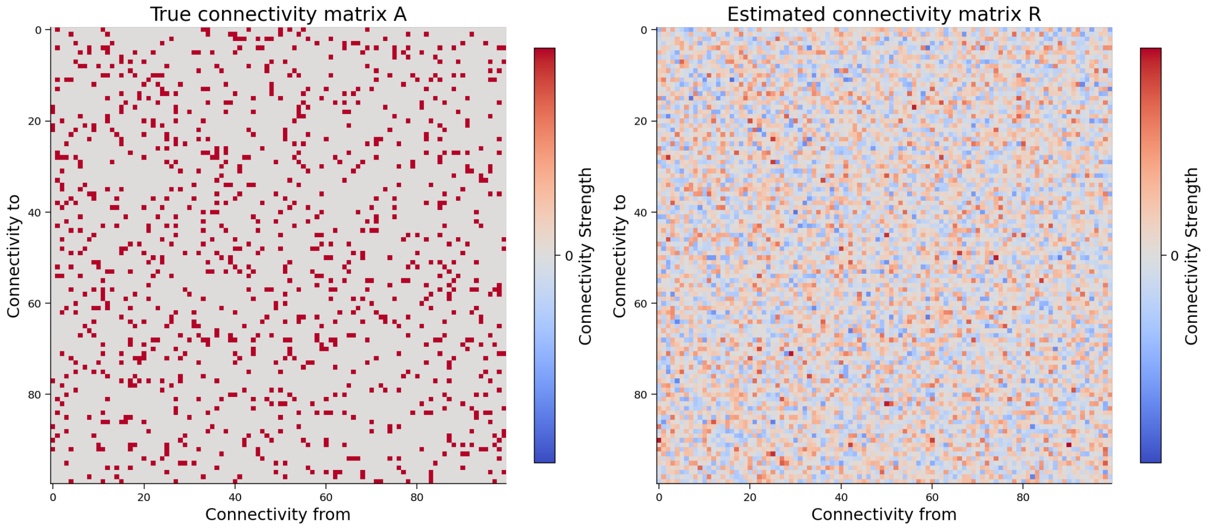
- $\vec{x}_t$ is an n-dimensional vector representing our n-neuron system at timestep t
- σ is a sigmoid nonlinearity
- A is our $n \times n$ *causal ground truth connectivity matrix* (more on this later)
- ϵ_t is random noise: $\epsilon_t \sim N(\vec{0}, I_n)$
- $\vec{x}_0$ is initialized to $\vec{0}$

Is correlation ~ causation?

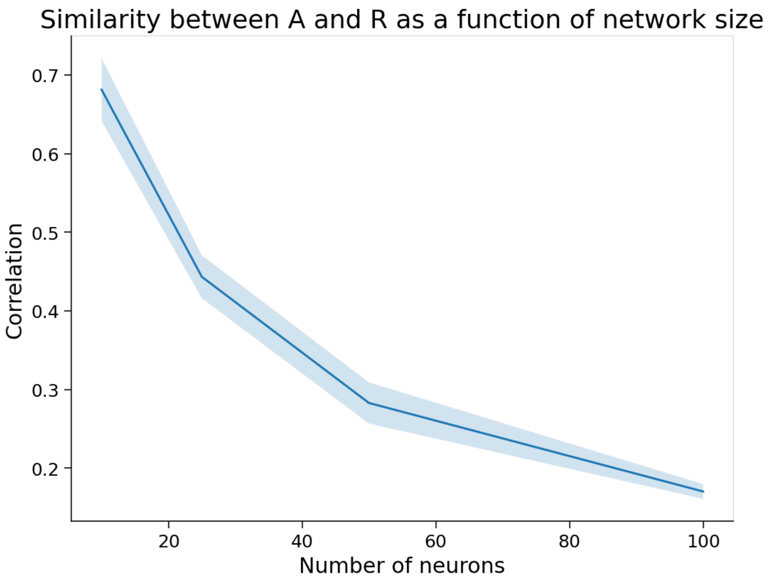
Great in small system



Not so great in big system



Delayed Correlation vs Causation



Omitted Variable Bias

$$y_i = x_i \beta + z_i \delta + u_i$$

$$\widehat{\beta} = (X'X)^{-1} X'(X\beta + Z\delta + U)$$

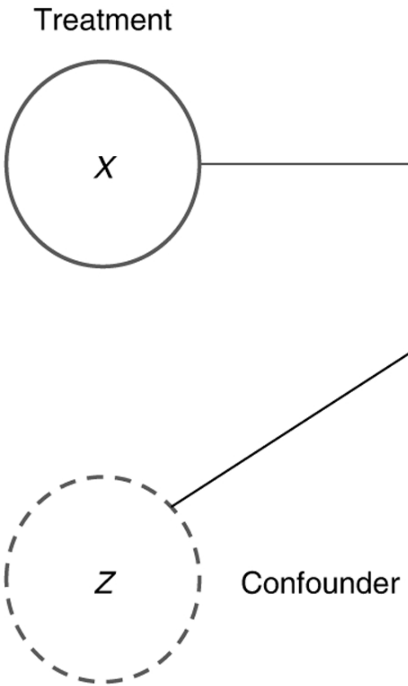
$$E[\widehat{\beta} \mid X] = \beta + (X'X)^{-1} E[X'Z \mid X]\delta$$

The bias should be arbitrarily big relative to the signal
This problem does not go away with more data

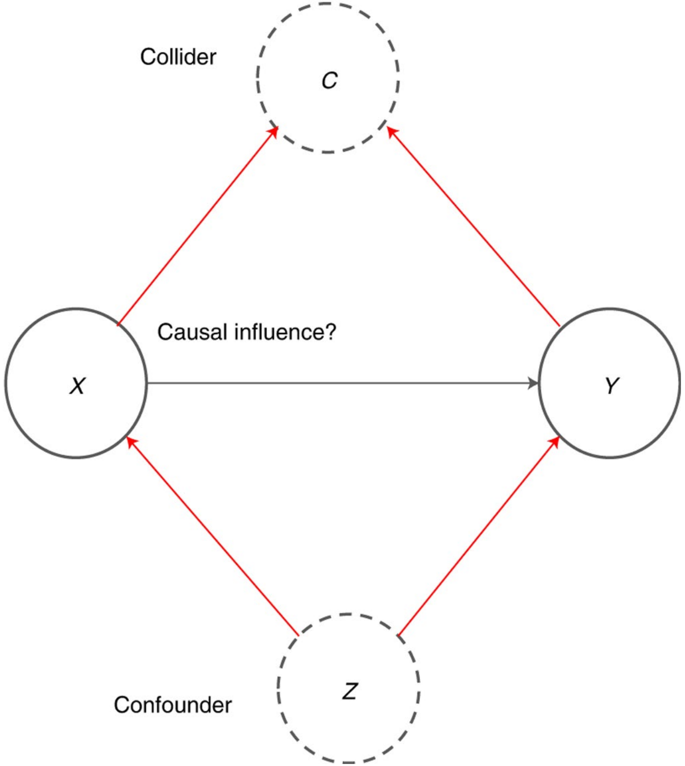
Measuring and interpreting neuronal correlations

Marlene R. Cohen[1] and Adam Kohn[2]

An ideal RCT allows unbiased measurements of causal effects

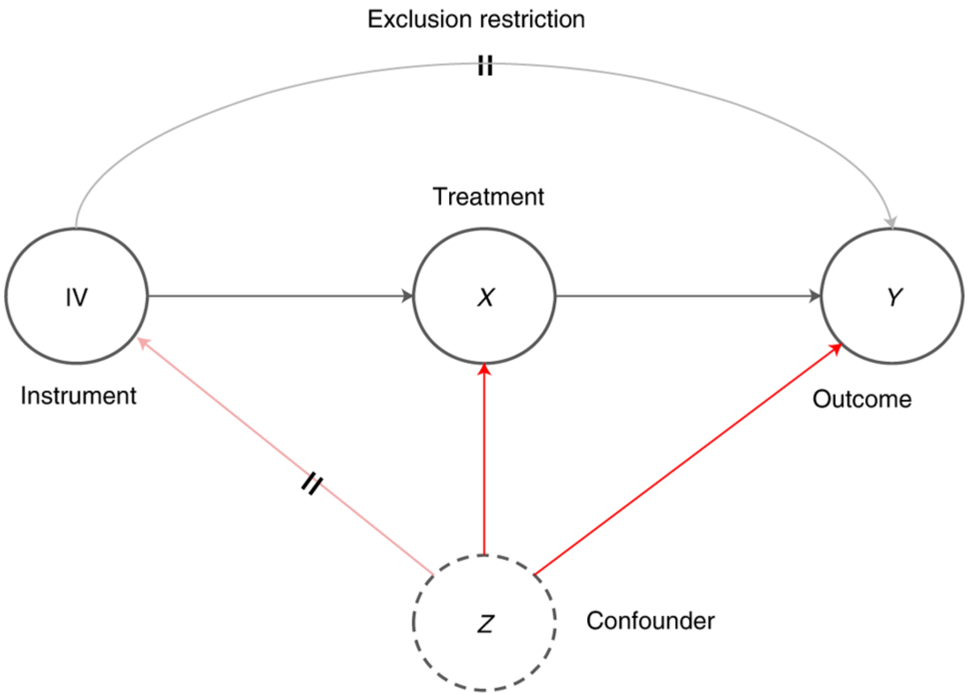


But alas, the world is not perfect: observational studies



Why is it so hard to obtain causality?

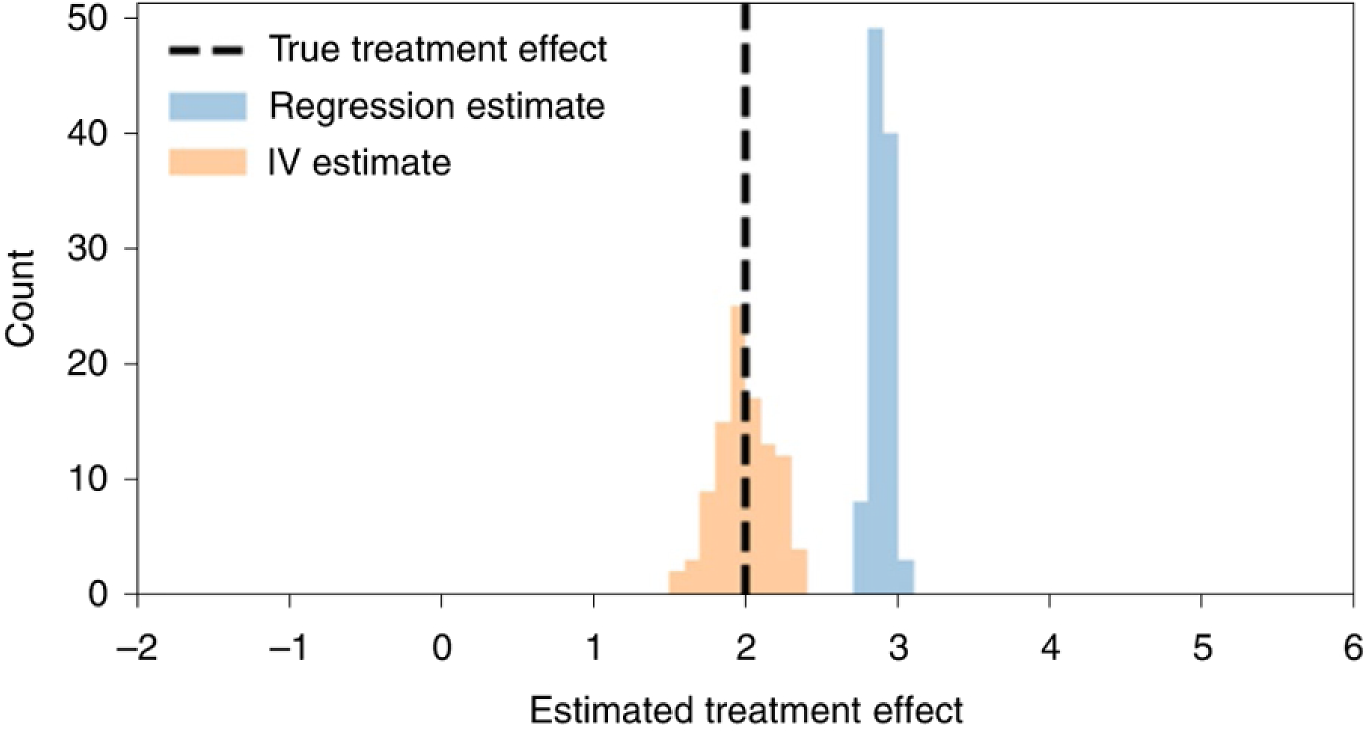
Instrumental variables



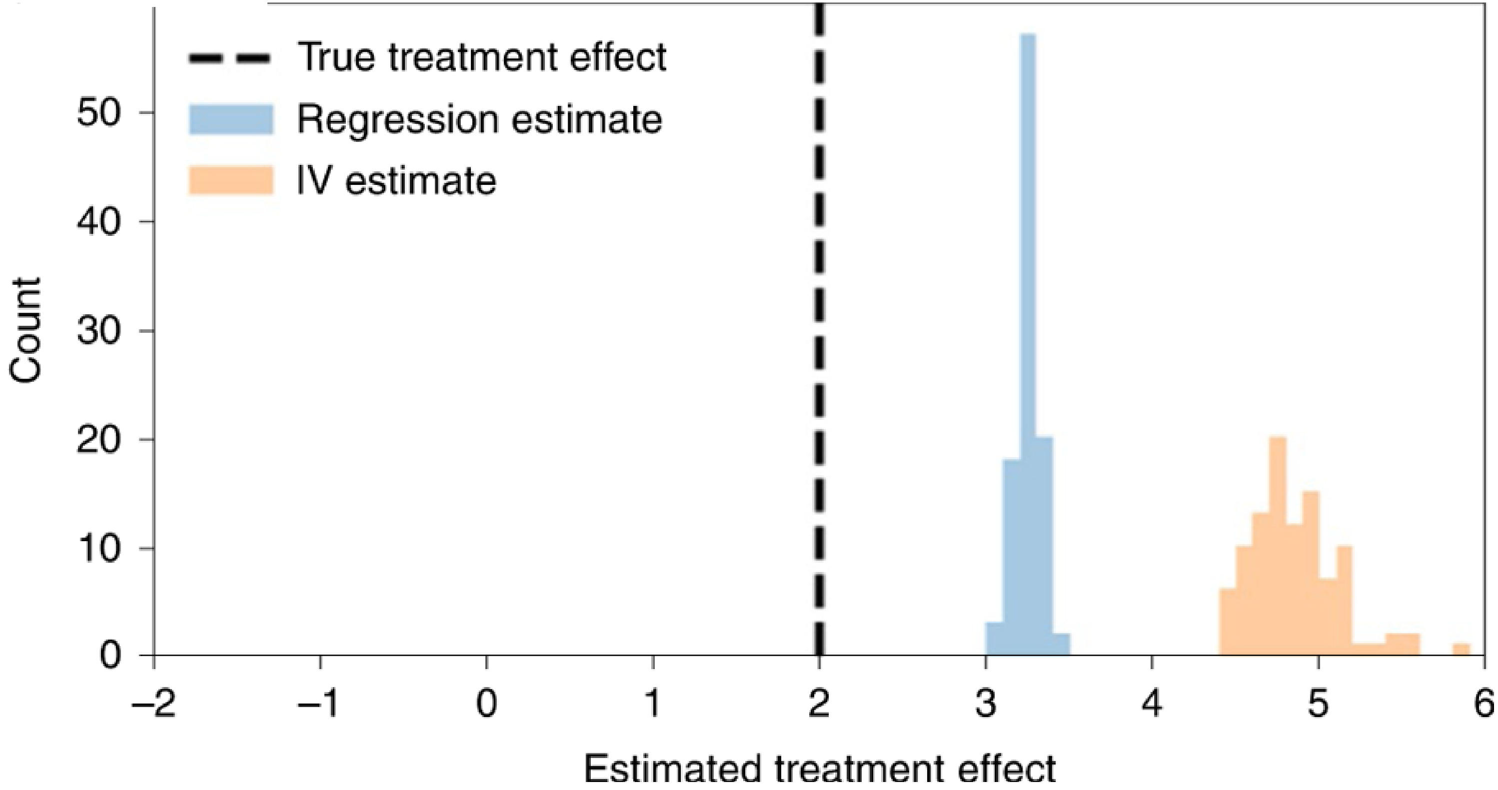
Quantifying causality in data science with quasi-experiments

Tony Liu,[1] Lyle Ungar,[1] and Konrad Kording[2,3,✉]

If exclusion restriction correct



If it is violated

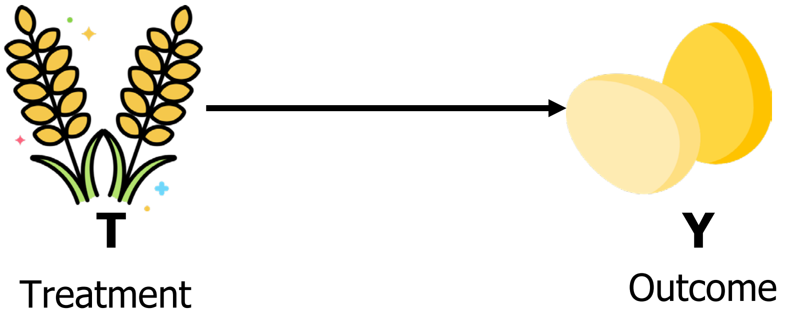


There are lots of quasiexperiments

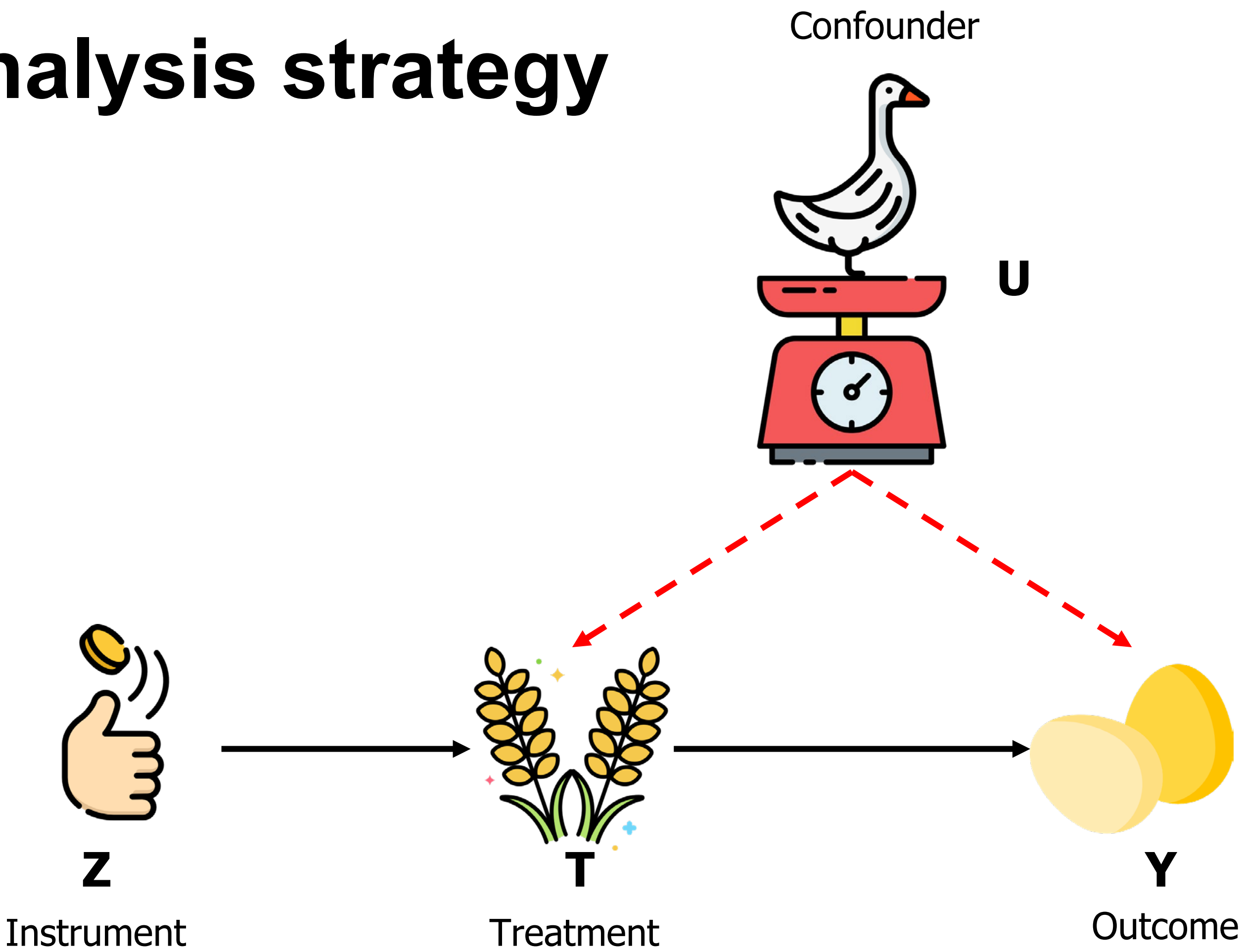
- Check out Tony's tutorials

https://github.com/tliu526/causal-data-science-perspective

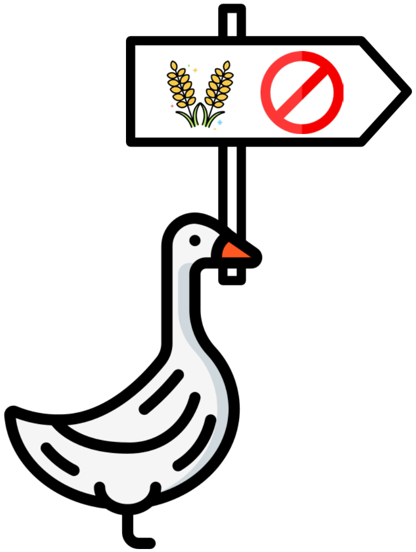
Does eating grains make ducks lay more eggs?



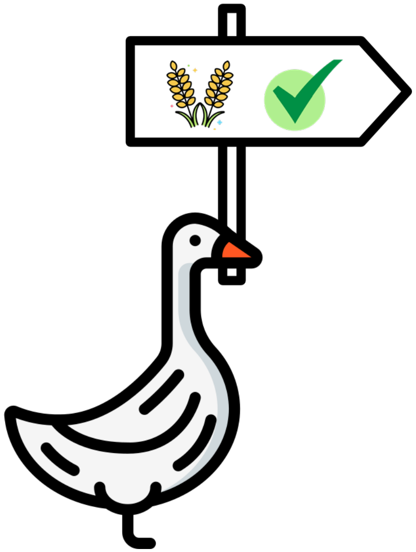
IV analysis strategy



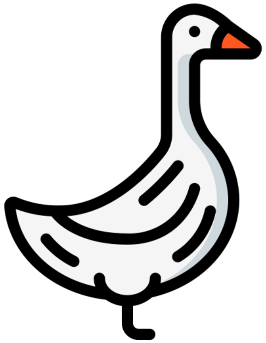
Problem: non-compliance



Never-takers

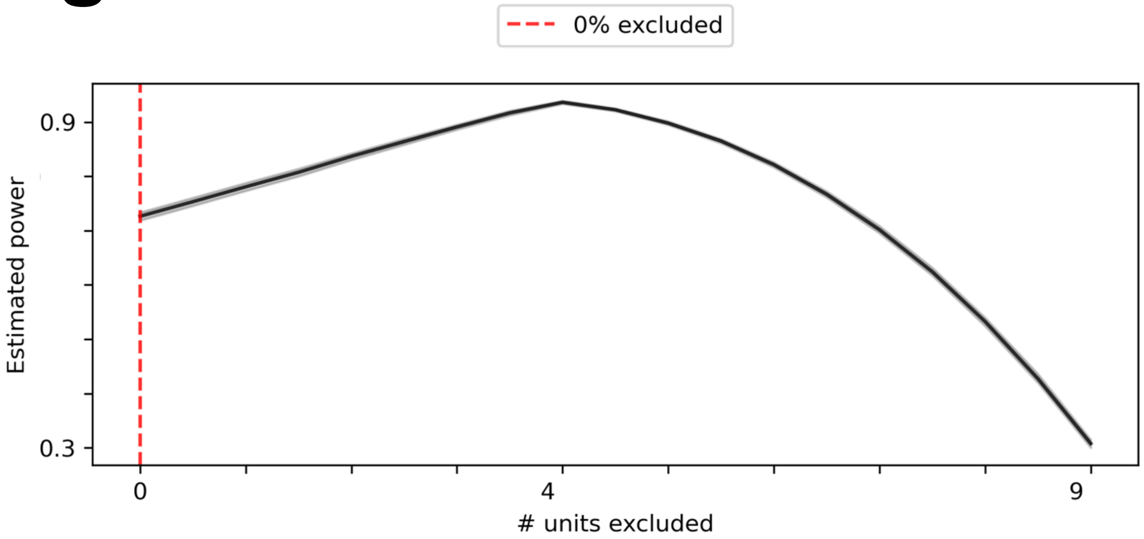


Always-takers

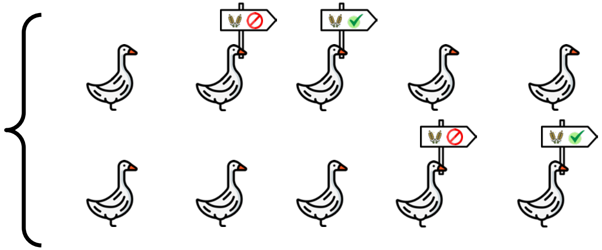


Compliers

Improving exclusion criteria for IV studies

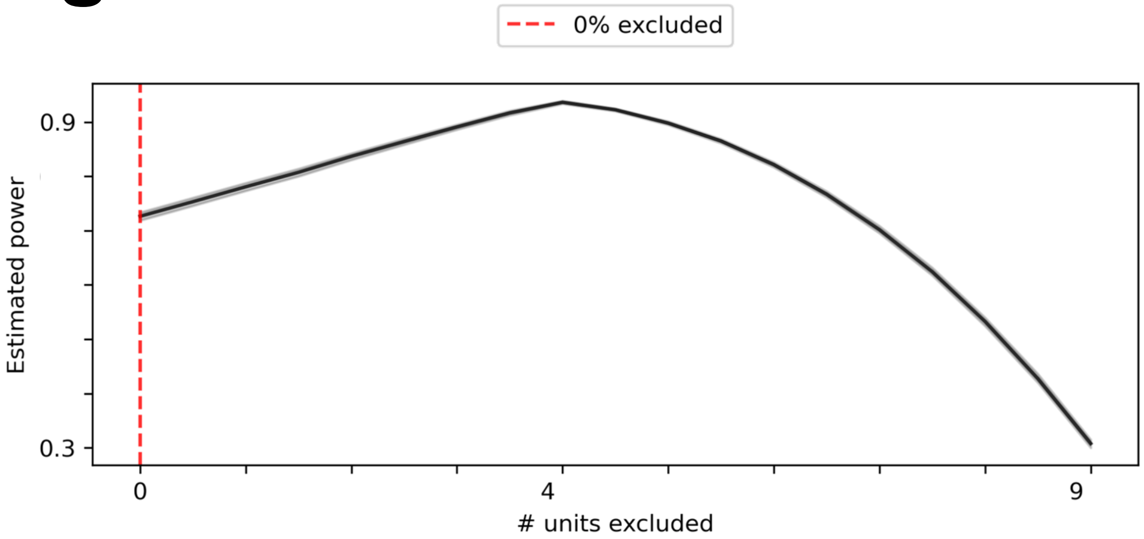


n = 10

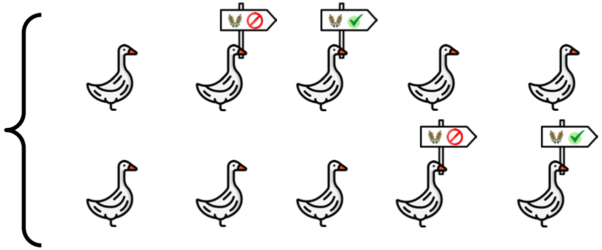


$$Var_{IV} = \frac{Var[Y|Z,\text{compliers}]}{Np^2_{\text{comply}}E[Z](1-E[Z])}$$

Improving exclusion criteria for IV studies

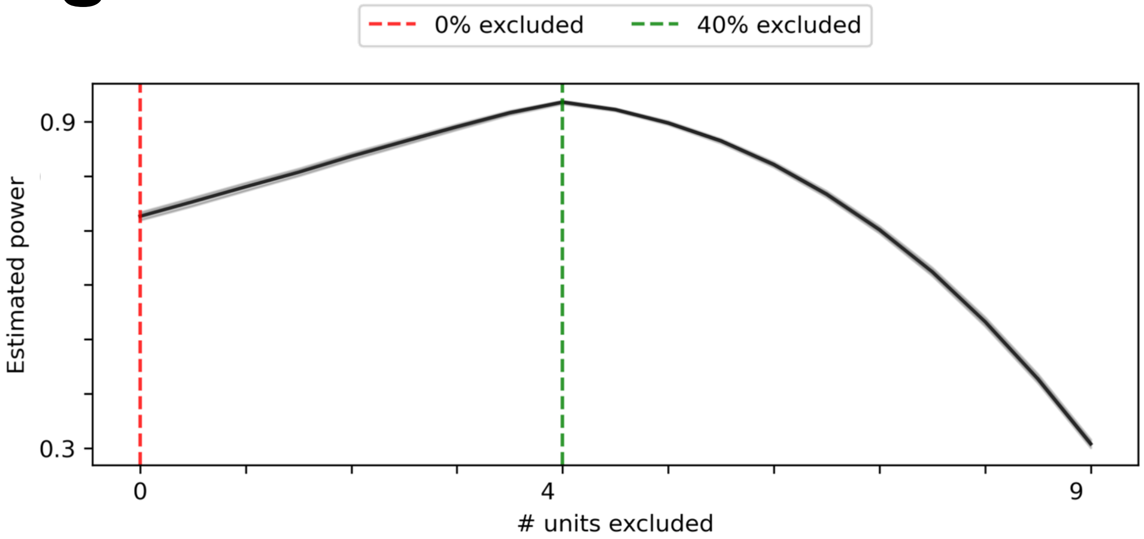


n_{eff} =
3.6

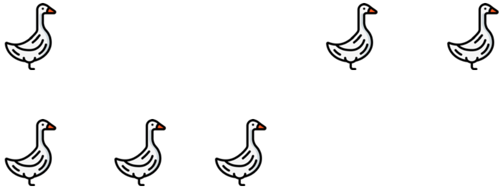


$$N_{\text{eff}} = N \times p^2_{\text{comply}}$$

Improving exclusion criteria for IV studies



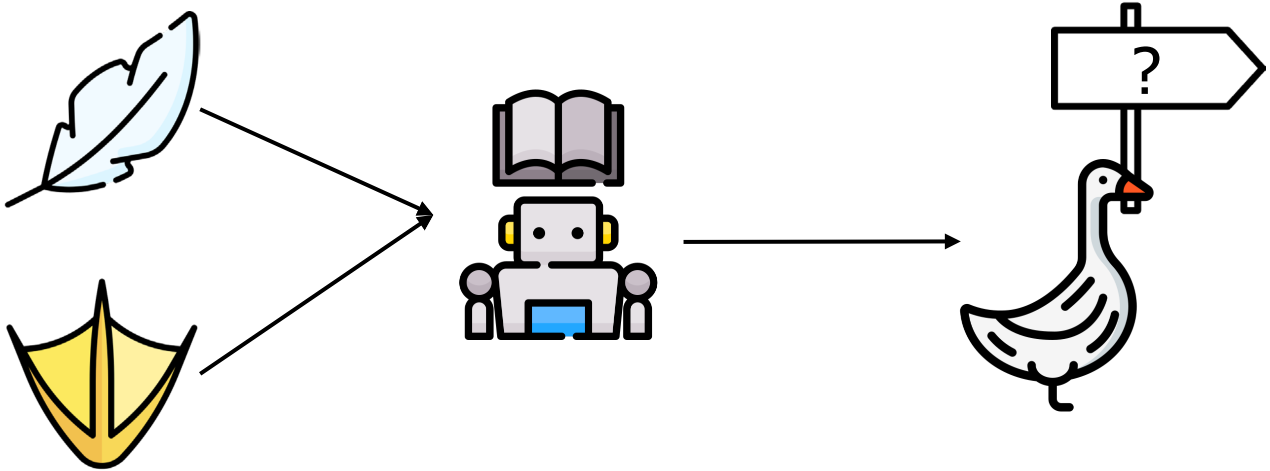
n_{eff} = 6



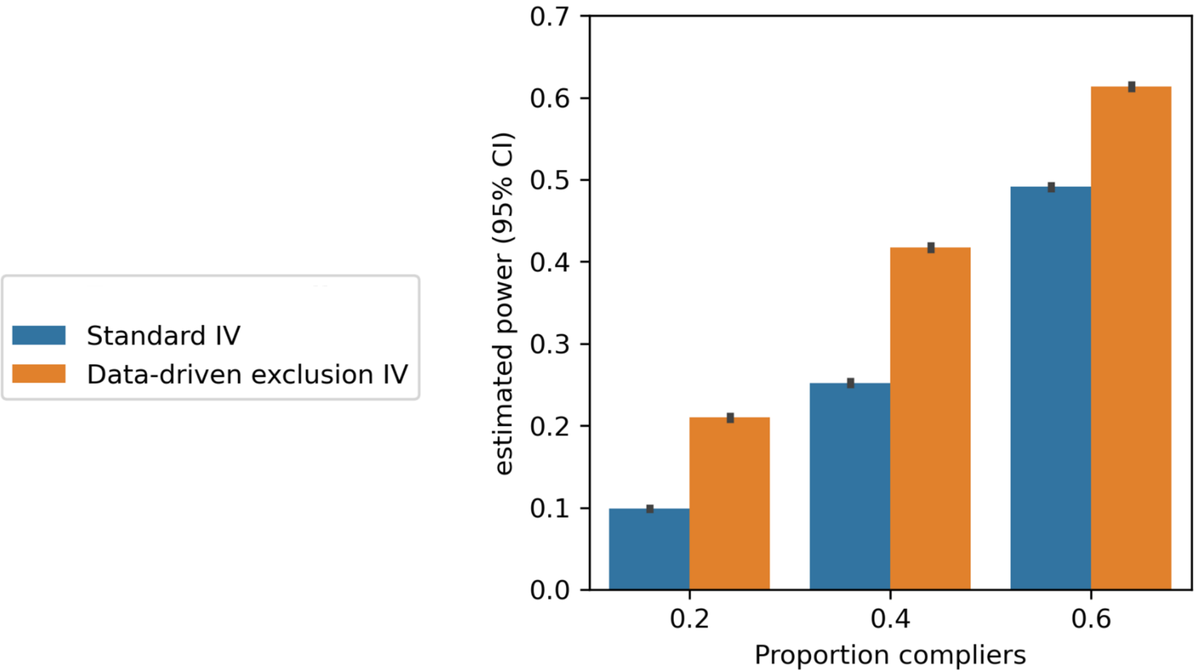
$$N_{\text{eff}} = N \times p^2_{\text{comply}}$$

Data-driven exclusion criteria

Use observed features and machine learning to predict out-of-sample compliance



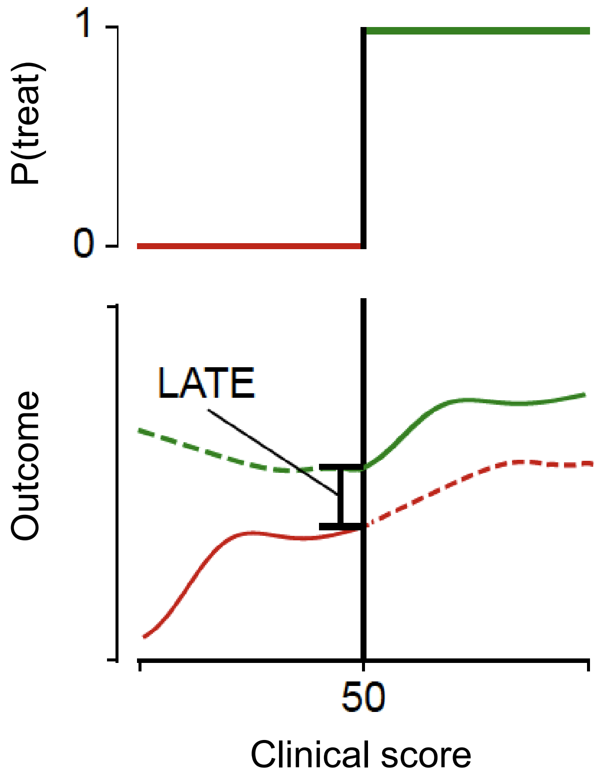
Data-driven exclusion improves power

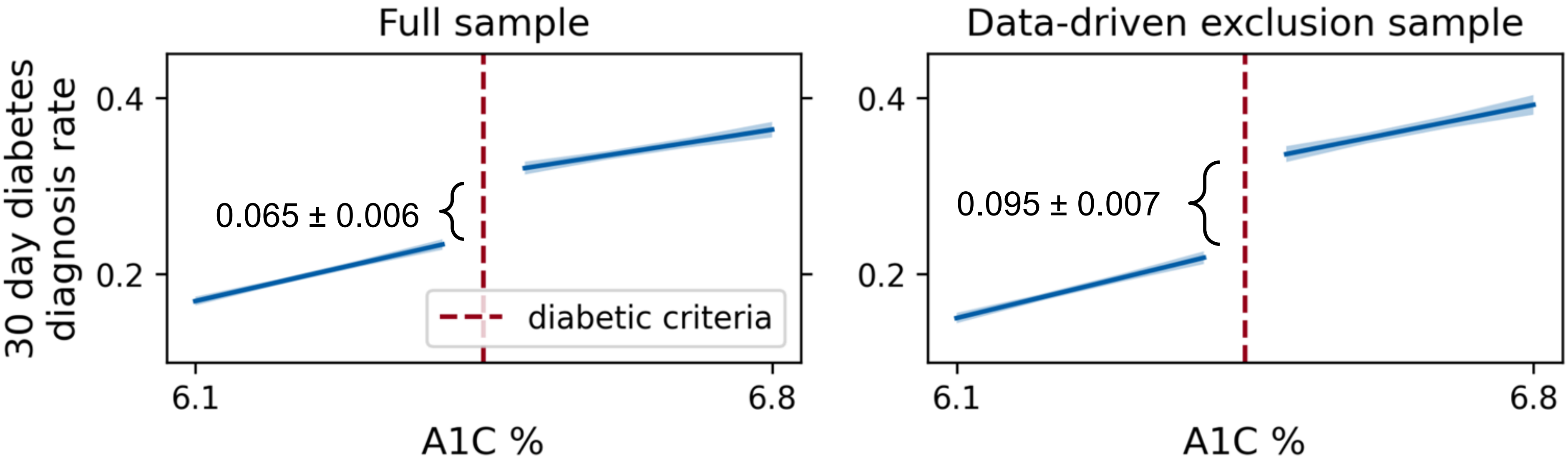


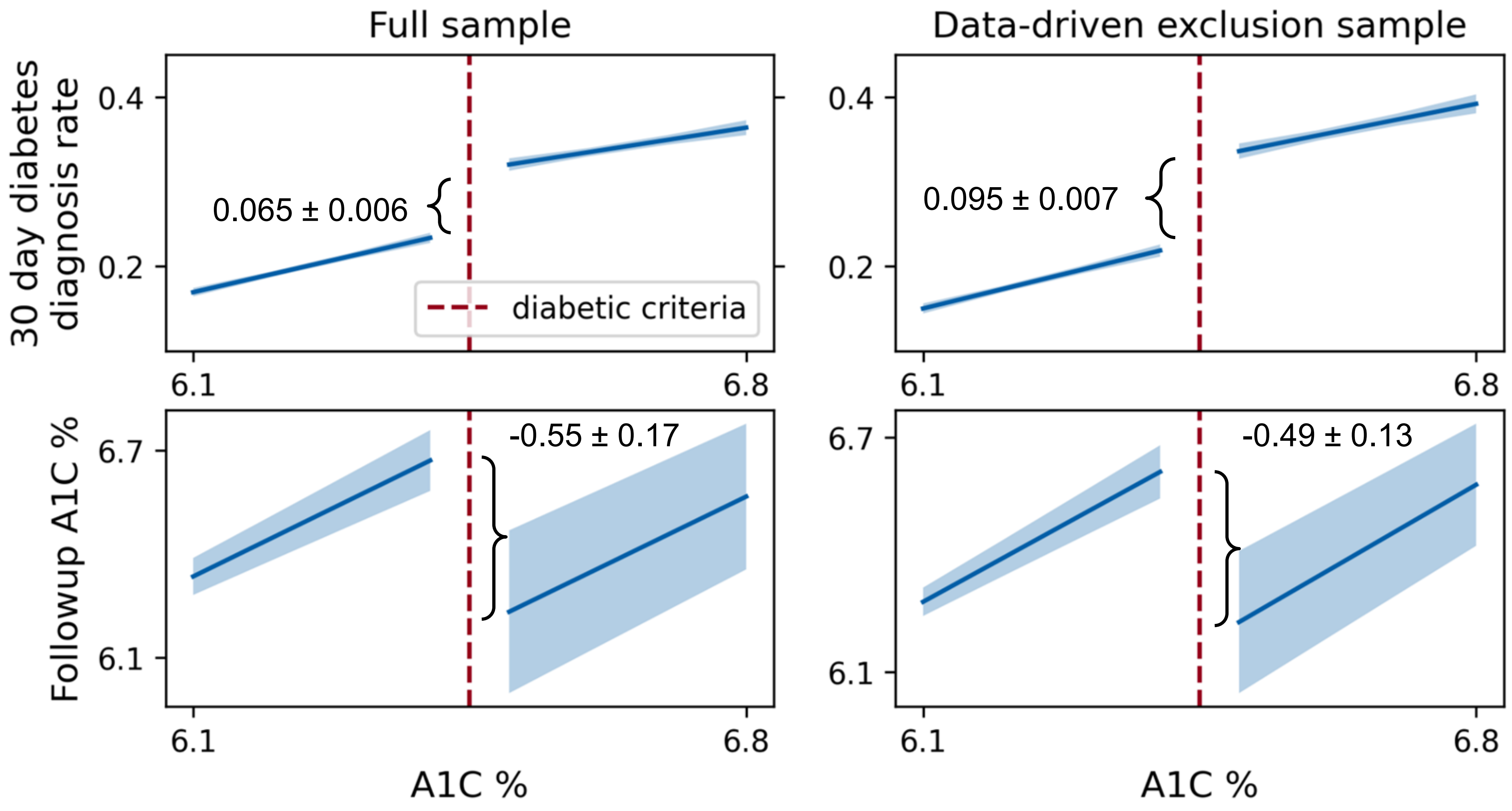
Applying this to diabetis

- Optum data
- Above A1C=6.5 you should be diabetic
- But it is fuzzy (hence IV)
- Follow-up A1C should be better

Let us use regression discontinuity design




Full sample
Data-driven exclusion sample
30 day diabetes diagnosis rate
0.4
0.2
0.065 ± 0.006
0.095 ± 0.007
diabetic criteria
6.1
6.8
A1C %



Full sample
Data-driven exclusion sample
30 day diabetes diagnosis rate
0.4
0.2
0.065 ± 0.006
0.095 ± 0.007
diabetic criteria
6.1
6.8
Followup A1C %
6.7
6.1
-0.55 ± 0.17
-0.49 ± 0.13
A1C %

A nonstandard logic

- ML tells us whom we are talking about
- Our predictions then only apply to these items
- We can then purposefully say nothing about the expected noncompliers
- (which seems unavoidably true)

Find all the RDDs

- Currently working to
  - Apply this idea to every threshold
  - Continuously sweep to find all thresholds
  - Optum and other datasets

A bitter lesson approach

Published in Transactions on Machine Learning Research (09/2023)

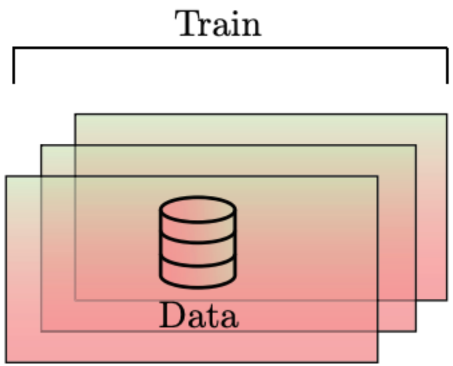
Learning domain-specific causal discovery from time series

Xinyue Wang *wsinyue@seas.upenn.edu*
Department of Bioengineering
University of Pennsylvania

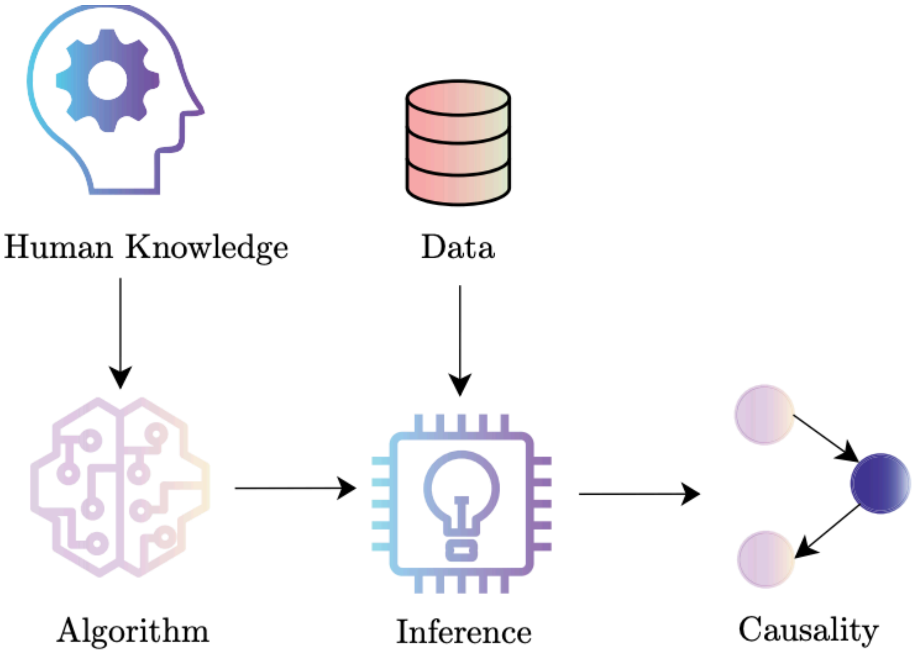
Konrad Kording *koerding@gmail.com*
Department of Bioengineering
University of Pennsylvania

Supervised learning

- Know inputs and true output
- Get good at mapping from input to output
- Causal inference is not like this!

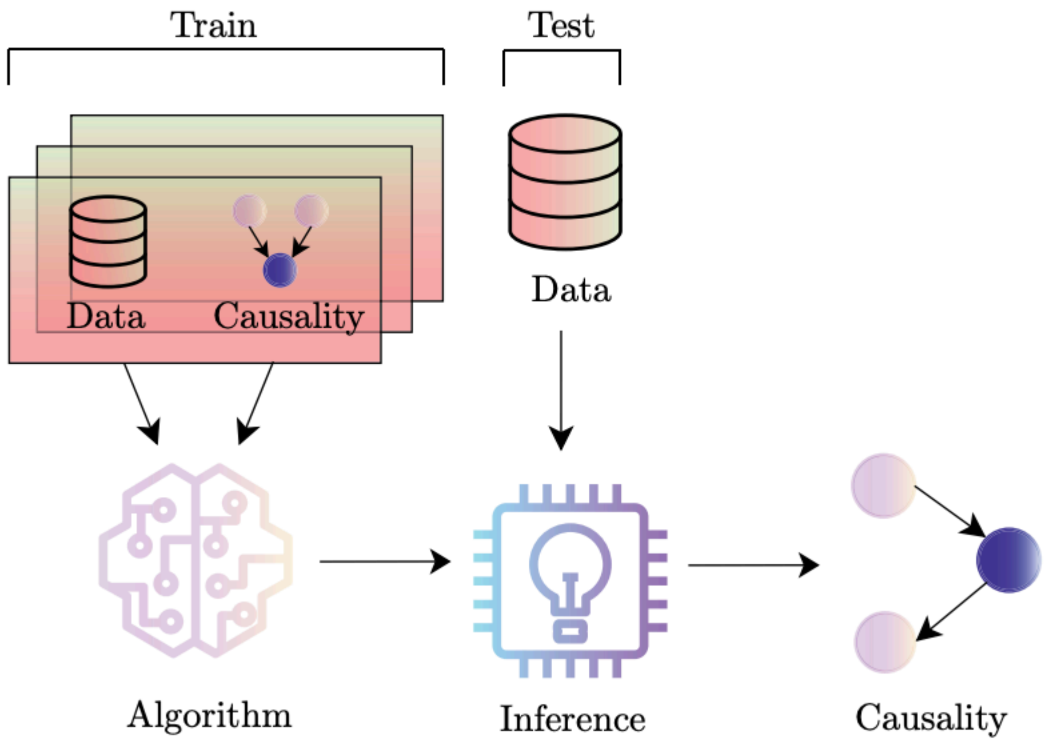


Where do CI/ CD algorithms come from?



Just like in ML

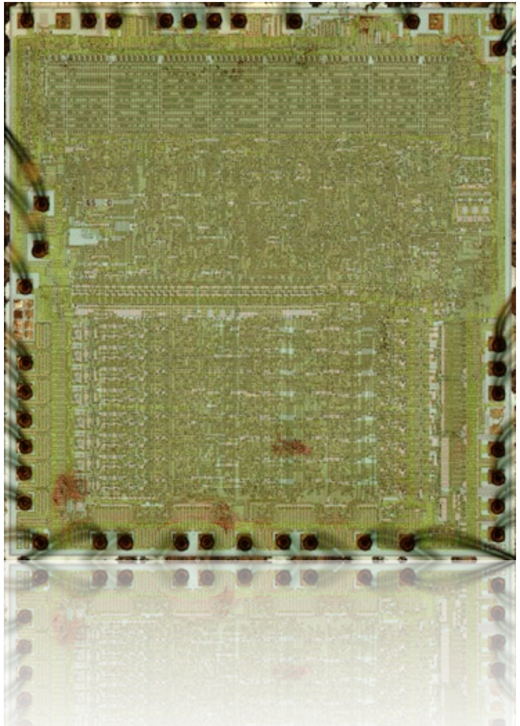
Learning causal discovery



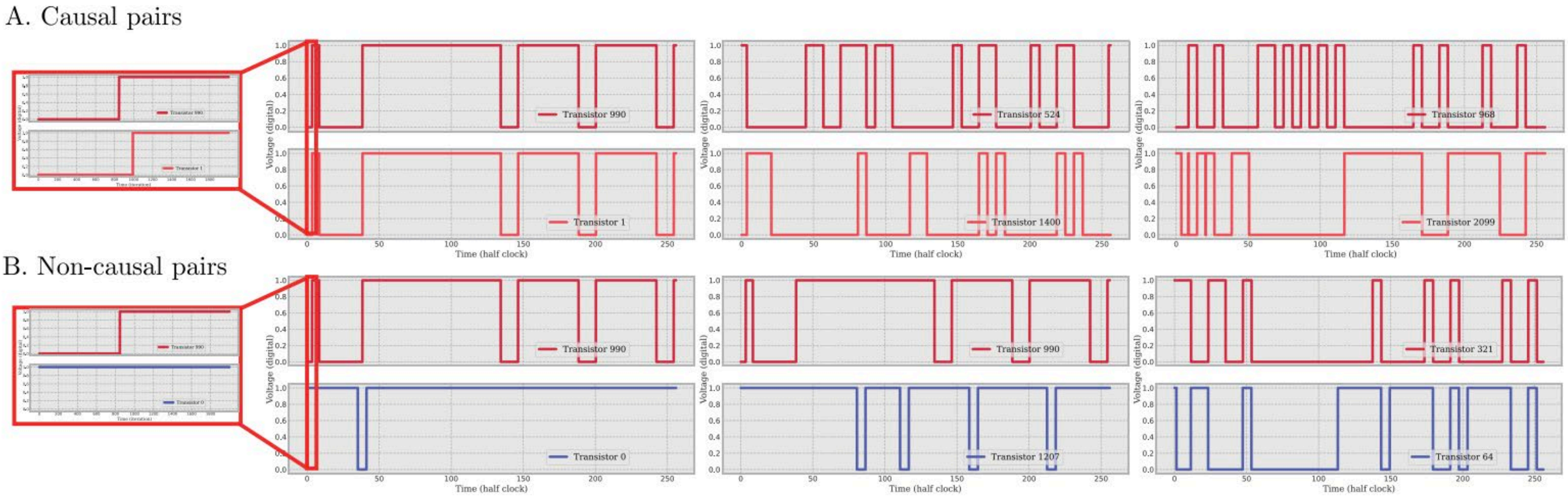
Causal inference is supervised learning, after all.

We need something to test this idea

- Something with known causality
- And nontrivial data

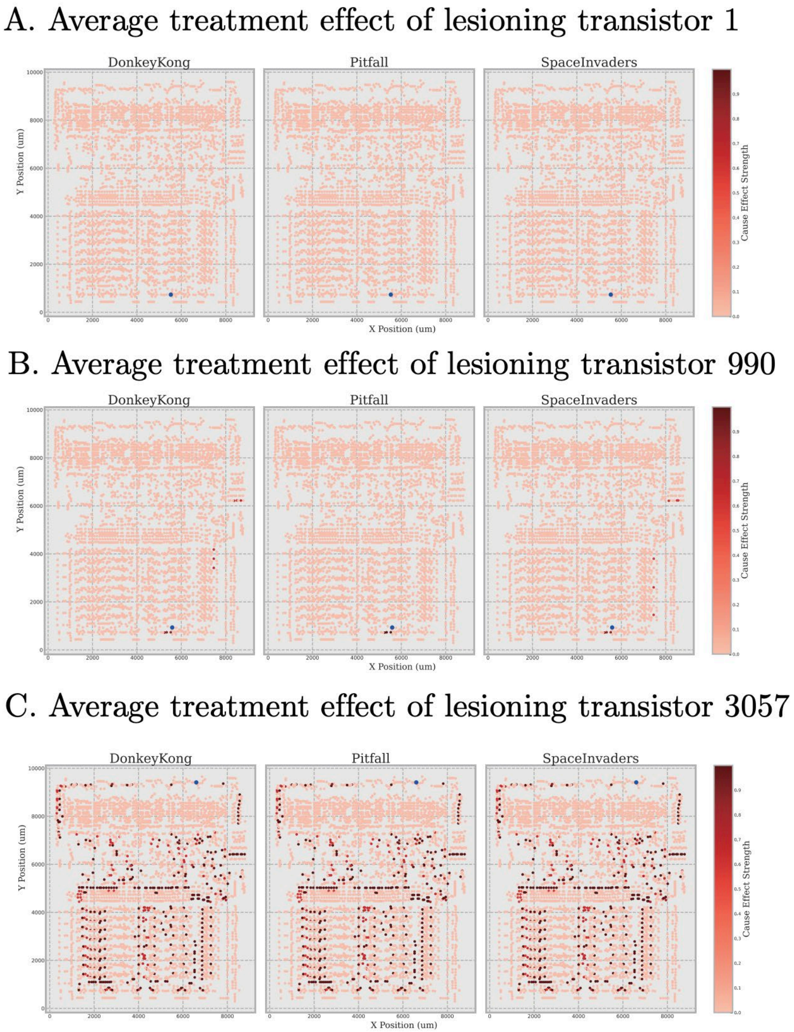


A causal discovery problem with the microprocessor



- Read through two channels. Predict if they are connected

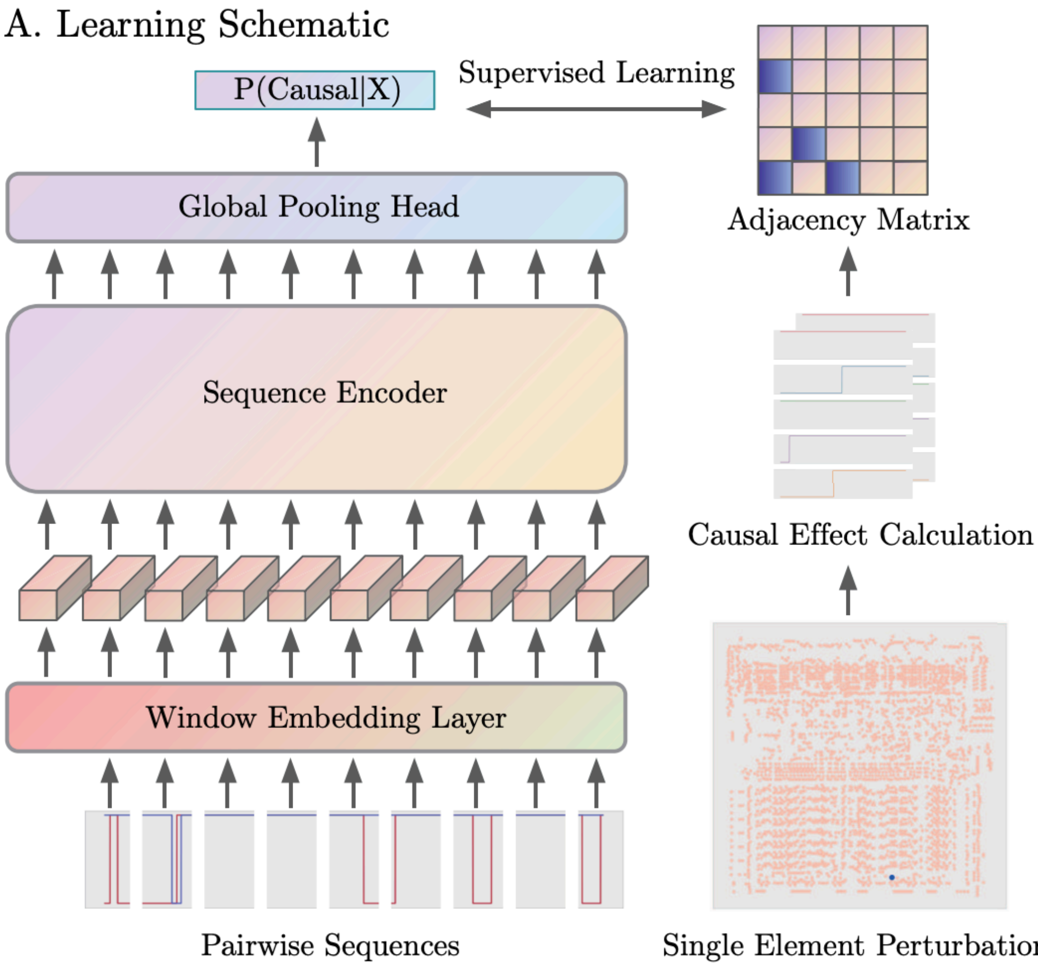
Perturbations to define causality



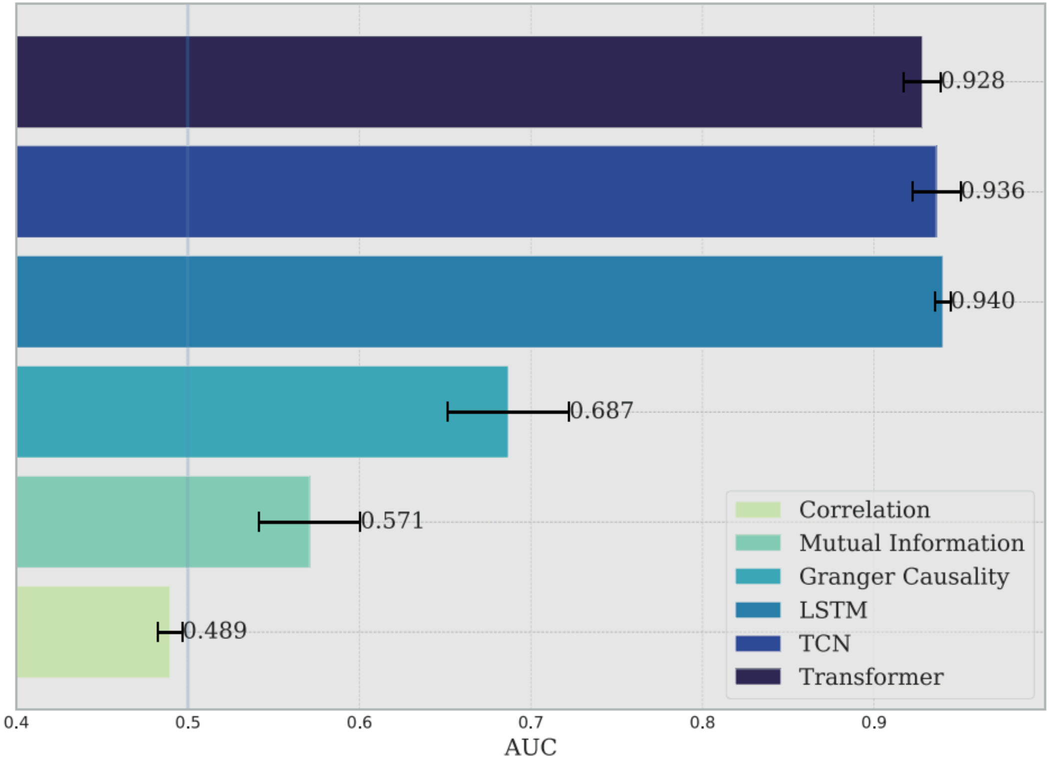
Honest validation

- Use 1 half of the data to train a causality detector
- Use other half to test if it works

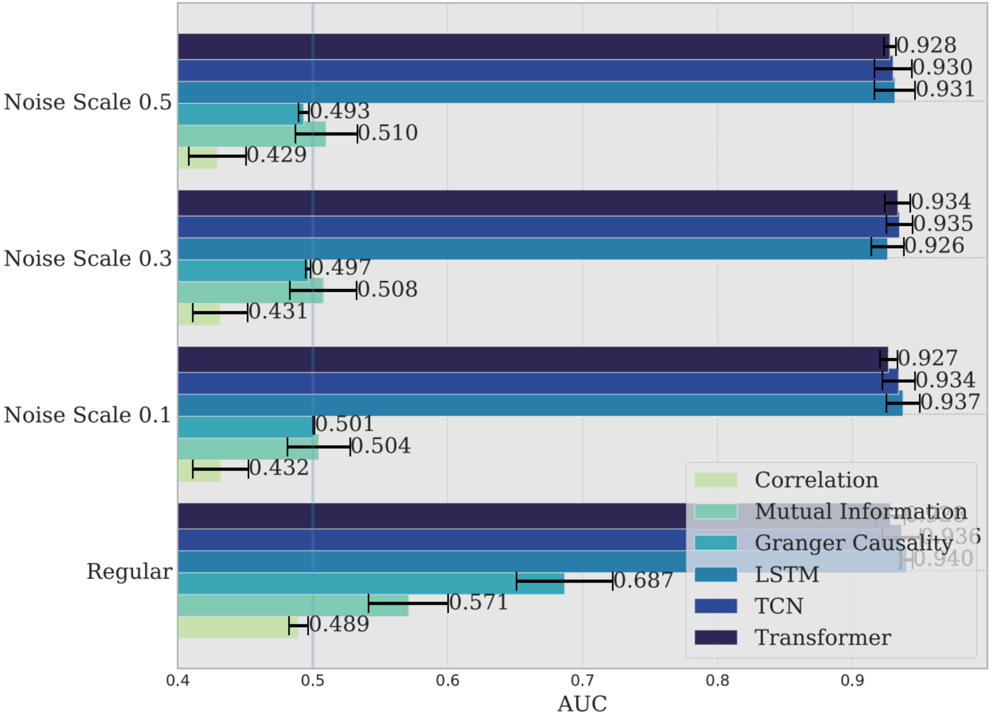
Standard Transformer ML system



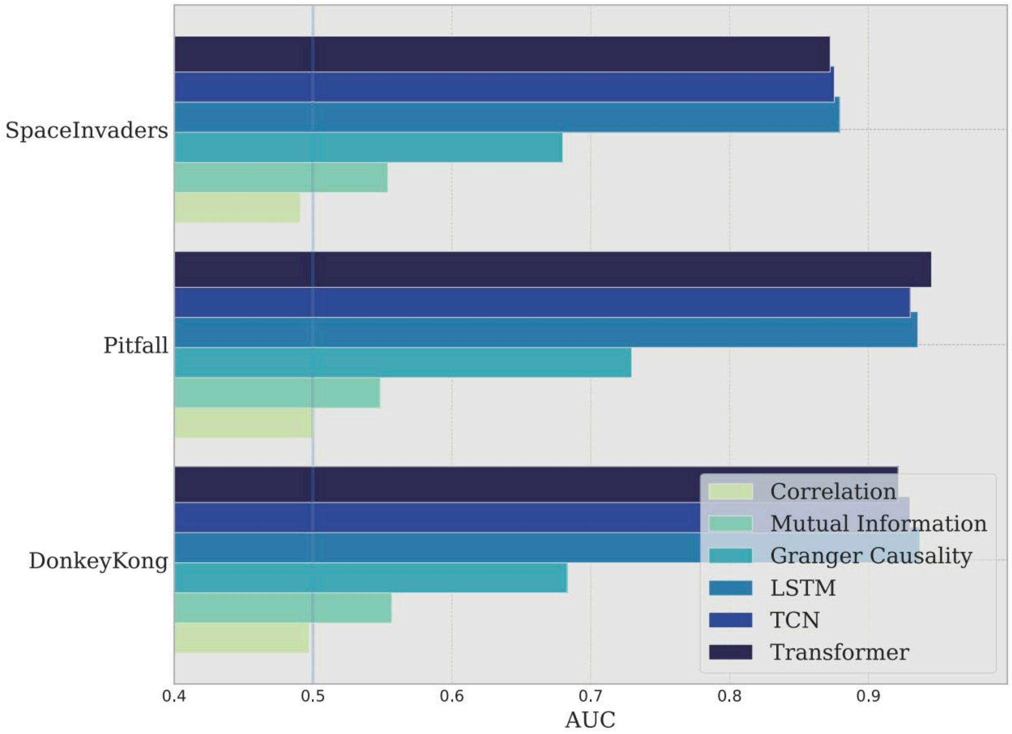
Works pretty great



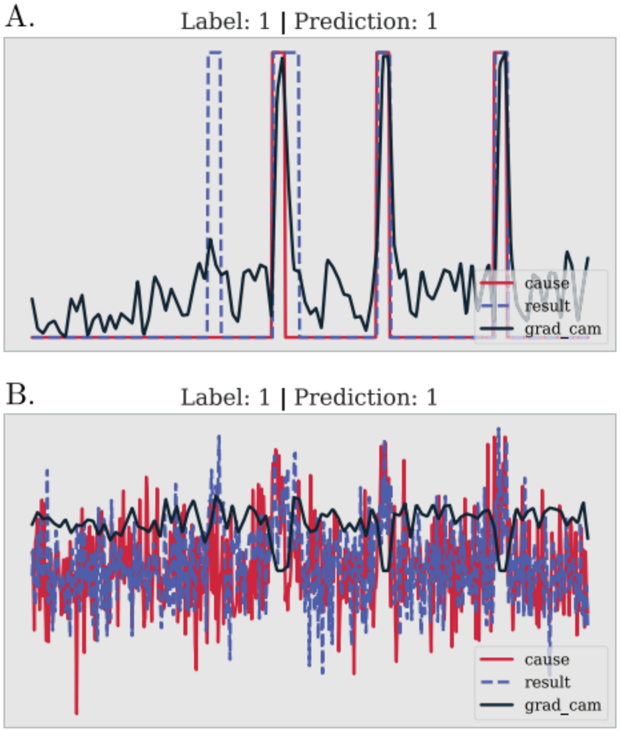
Also robust to noise



Generalizes well across games: train on Donkey Kong, test others



Focuses attention where it matters (transients)



Simulated fMRI

Table 3: AUPRC comparison on different simulations of NetSim (mean ± std).

Dataset	Corr	DYNO	GC	MI	PCMCI+	LiNGAM	cMLP	cLSTM	eSRU	SRU	Transformer
Sim1	0.47 ± 0.02	0.41 ± 0.08	0.40 ± 0.08	0.39 ± 0.08	0.39 ± 0.09	0.43 ± 0.15	0.42 ± 0.15	0.41 ± 0.14	0.40 ± 0.14	0.39 ± 0.14	**0.97±0.06**
Sim2	0.45 ± 0.03	0.33 ± 0.12	0.32 ± 0.12	0.31 ± 0.11	0.29 ± 0.11	0.30 ± 0.11	0.29 ± 0.11	0.28 ± 0.11	0.27 ± 0.11	0.26 ± 0.10	**0.94±0.05**
Sim3	0.44 ± 0.03	0.32 ± 0.13	0.29 ± 0.14	0.27 ± 0.12	0.26 ± 0.12	0.27 ± 0.13	0.26 ± 0.12	0.24 ± 0.12	0.23 ± 0.12	0.22 ± 0.12	**0.89±0.06**
Sim8	0.41 ± 0.07	0.36 ± 0.08	0.38 ± 0.11	0.37 ± 0.10	0.36 ± 0.10	0.40 ± 0.14	0.40 ± 0.14	0.39 ± 0.14	0.39 ± 0.14	0.38 ± 0.13	**0.85±0.13**
Sim10	0.48 ± 0.02	0.38 ± 0.10	0.39 ± 0.12	0.40 ± 0.11	0.40 ± 0.12	0.42 ± 0.16	0.42 ± 0.16	0.42 ± 0.15	0.42 ± 0.15	0.42 ± 0.15	**0.96±0.03**
Sim11	0.28 ± 0.03	0.26 ± 0.04	0.26 ± 0.06	0.26 ± 0.06	0.25 ± 0.07	0.25 ± 0.08	0.25 ± 0.08	0.24 ± 0.08	0.24 ± 0.08	0.23 ± 0.08	**0.71±0.10**
Sim12	0.43 ± 0.02	0.36 ± 0.08	0.33 ± 0.11	0.31 ± 0.10	0.29 ± 0.11	0.30 ± 0.11	0.28 ± 0.11	0.27 ± 0.11	0.26 ± 0.11	0.26 ± 0.11	**0.90±0.06**
Sim13	0.48 ± 0.04	0.47 ± 0.05	0.48 ± 0.07	0.49 ± 0.10	0.47 ± 0.10	0.48 ± 0.10	0.47 ± 0.10	0.47 ± 0.10	0.47 ± 0.11	0.46 ± 0.11	**0.76±0.10**
Sim14	0.48 ± 0.02	0.41 ± 0.08	0.41 ± 0.09	0.40 ± 0.08	0.38 ± 0.09	0.42 ± 0.13	0.41 ± 0.13	0.40 ± 0.13	0.39 ± 0.13	0.39 ± 0.13	**0.93±0.08**
Sim15	0.45 ± 0.03	0.38 ± 0.07	0.40 ± 0.09	0.41 ± 0.08	0.41 ± 0.10	0.48 ± 0.21	0.47 ± 0.20	0.45 ± 0.20	0.44 ± 0.19	0.43 ± 0.19	**0.72±0.09**
Sim16	0.48 ± 0.01	0.44 ± 0.05	0.45 ± 0.07	0.44 ± 0.06	0.44 ± 0.06	0.46 ± 0.10	0.46 ± 0.10	0.45 ± 0.10	0.45 ± 0.10	0.45 ± 0.10	**0.96±0.03**
Sim17	0.47 ± 0.01	0.39 ± 0.09	0.36 ± 0.10	0.36 ± 0.09	0.35 ± 0.10	0.42 ± 0.19	0.40 ± 0.19	0.37 ± 0.19	0.35 ± 0.19	0.34 ± 0.19	**0.98±0.02**
Sim18	0.48 ± 0.03	0.42 ± 0.07	0.42 ± 0.12	0.41 ± 0.10	0.40 ± 0.11	0.43 ± 0.16	0.42 ± 0.16	0.41 ± 0.16	0.40 ± 0.15	0.39 ± 0.15	**0.99±0.02**
Sim21	0.48 ± 0.03	0.42 ± 0.08	0.41 ± 0.08	0.40 ± 0.08	0.38 ± 0.09	0.42 ± 0.15	0.41 ± 0.14	0.40 ± 0.14	0.39 ± 0.14	0.38 ± 0.13	**0.95±0.06**
Sim22	**0.41±0.04**	0.38 ± 0.06	0.38 ± 0.08	0.39 ± 0.07	0.37 ± 0.08	0.37 ± 0.09	0.35 ± 0.09	0.34 ± 0.09	0.34 ± 0.09	0.34 ± 0.09	0.31 ± 0.11
Sim23	0.40 ± 0.04	0.35 ± 0.06	0.40 ± 0.12	0.39 ± 0.10	0.41 ± 0.14	**0.47±0.21**	0.45 ± 0.20	0.43 ± 0.19	0.42 ± 0.19	0.41 ± 0.19	0.41 ± 0.05
Sim24	0.36 ± 0.06	0.31 ± 0.07	0.34 ± 0.10	0.35 ± 0.10	0.35 ± 0.11	0.35 ± 0.11	0.34 ± 0.11	0.34 ± 0.11	0.34 ± 0.11	0.33 ± 0.11	**0.37±0.11**

Gene networks

Table 5: AUPRC comparison on the Dream3

Dataset	**Corr**	**DYNO**	**GC**	**MI**	**PCMCI+**	**cMLP**	**cLSTM**	**eSRU**	**SRU**	**Transformer**
Ecoli2	0.01	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	**0.04**
Yeast2	0.01	0.01	0.02	0.01	0.01	0.04	0.04	0.04	0.04	**0.06**
Yeast3	0.01	0.01	0.01	0.01	0.01	0.06	**0.07**	0.06	0.06	0.06

What does this mean

- Cons
  - Requires a lot of ground truth data
  - Impossible for humans to understand
  - Biased
- Pros
  - If we have enough ground truth it will beat all humans
  - Meaningful proofs (inherited from empirical risk minimization)
  - A new approach to get into messy observational spaces

Take home message

- Causality and ML really have overlapping problems
- So much to learn
- Check out Clear conference

- Exclusion can be a meaningful data problem
- Maybe we should look for more ways to learn how to do CD/ CI, and prepare ourselves for a future where the bitter lesson is key to medicine