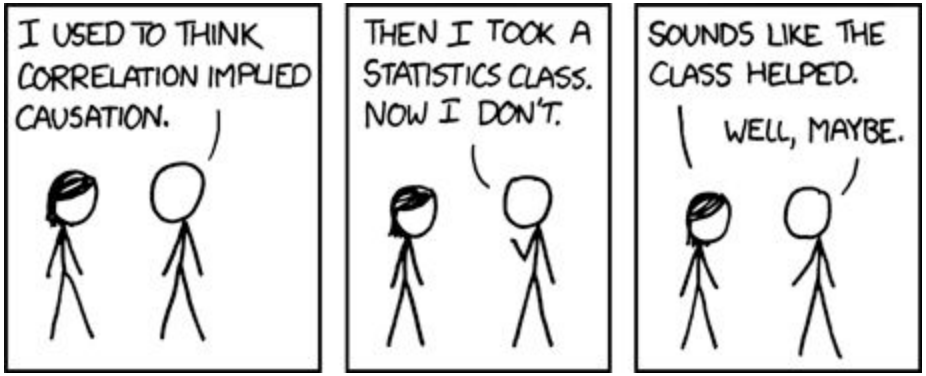


xkcd: correlation
https://xkcd.com/552/

Feedback from you about the lecture

Three main improvements for future lectures

- A road map of overview
- A better structure of exercises
- A better balance between concept and details

End-term project

Students self-organize into a team of two to work on the project. If forming a team is not possible, individual contribution is possible. Choose one top from the two options:

- Option 1: Write a target (or screening) proposal for a disease of your choice, using publicly available data and your analysis to support your arguments.
- Option 2: Write a Rmarkdown/Jupyter report analysing data from the Drug Central database, raising your own scientific questions about drug-target associations and answering them with analysis.

Once the project report is submitted, it will peer-reviewed by another group, which give comments and suggestions:

- Students assigning themselves in groups of two by June 4th (Friday);
- Submission deadline: June 27 (Monday);
- Peer review submission deadline: July 4th (Monday).

Causal Inference in Drug Discovery and Development

Mathematical and Computational Biology in Drug Discovery (MCBDD)
Module V

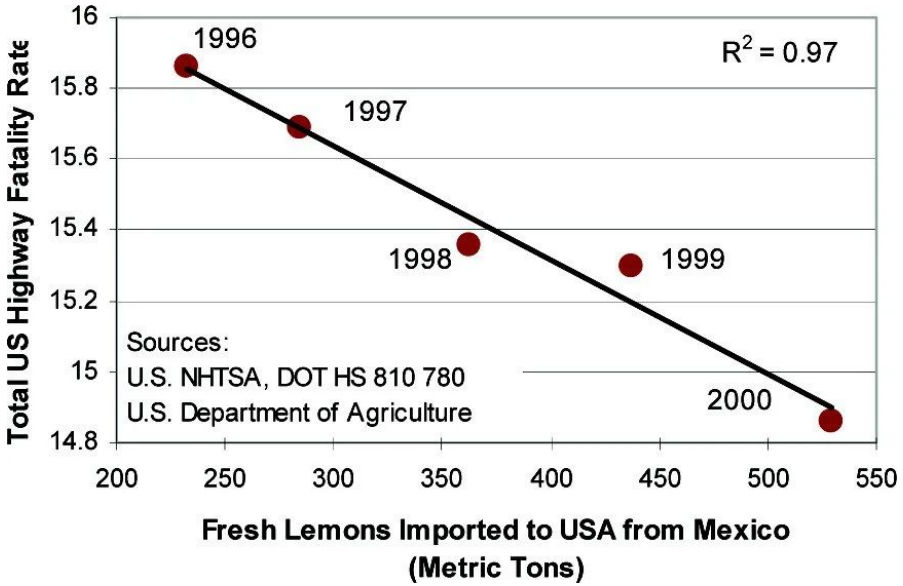
Dr. Jitao David Zhang
June 2022

Outline

- **Speaking the causal language**
- **Examples in drug discovery and development**
- **Consequences**

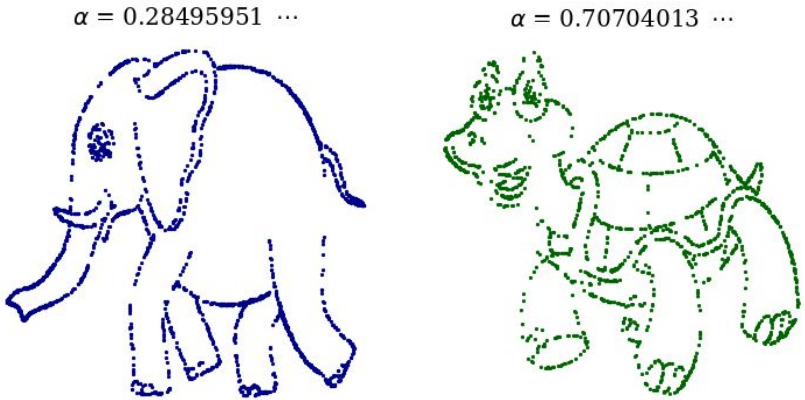
Correlation does not imply causation: so what is causation?

Cum hoc ergo propter hoc?



Johnson, Stephen R. “The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy).” Journal of Chemical Information and Modeling 48, no. 1 (January 1, 2008): 25–26. https://doi.org/10.1021/ci700332k.

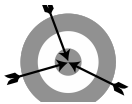
$$f_\alpha(x) = \sin^2\left(2^{x\tau} \arcsin\sqrt{\alpha}\right)$$



Boué, Laurent. “Real Numbers, Data Science and Chaos: How to Fit Any Dataset with a Single Parameter.” ArXiv:1904.12320 [Cs, Stat], April 28, 2019. http://arxiv.org/abs/1904.12320.GitHub Repo

Bradford Hill's Criteria as a working definition of causality


Strength


Consistency


Specificity


Temporality


Dose response


Plausibility


Coherence


Experiment

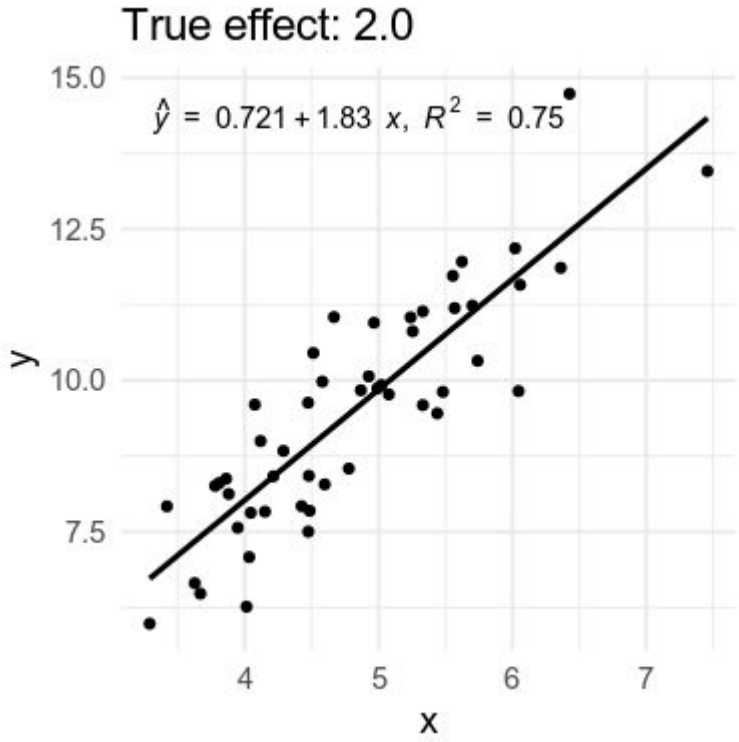

Analogy


Reversibility

Hill, Austin Bradford. "The Environment and Disease: Association or Causation?" Proceedings of the Royal Society of Medicine 58, no. 5 (May 1965): 295–300.

Generative models shed light on correlation and causality

	x	y
1	4.926791	10.067779
2	4.479734	8.424283
3	4.289686	8.835629
4	4.474023	9.630499
5	4.214551	8.416680
6	6.057431	11.578080
7	4.597903	8.283025
8	5.021571	9.922731
9	3.627323	6.651222
10	5.622794	11.959972
11	5.555025	11.727815
12	4.966007	10.951562
13	5.076791	9.768299

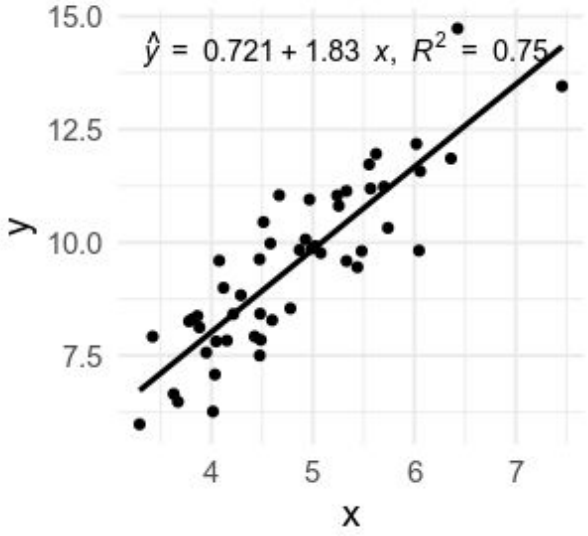


Assumptions of the **generative model**:

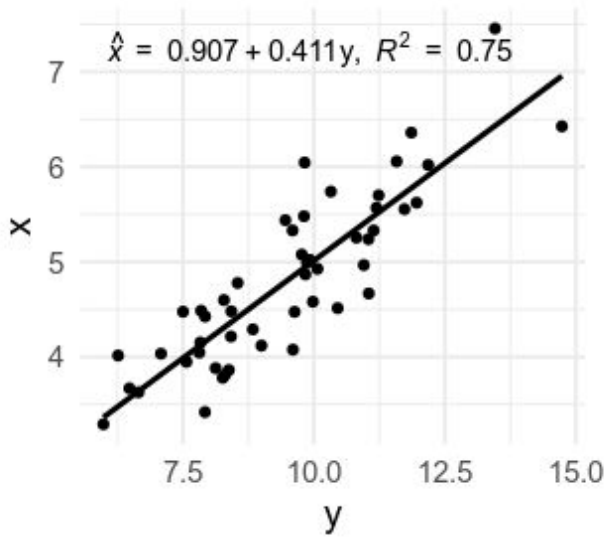
1. X is a random variable;
2. Every unit change of X increases a change of 2 unis in Y.

Correlation is caused by causation or confounding

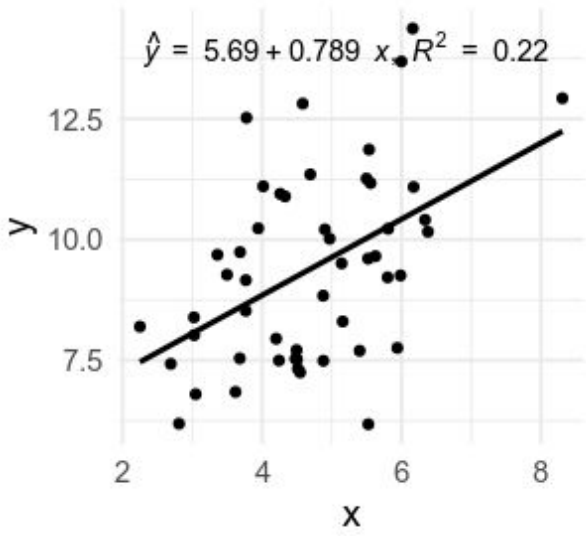
True effect: 2.0



The reverse fit

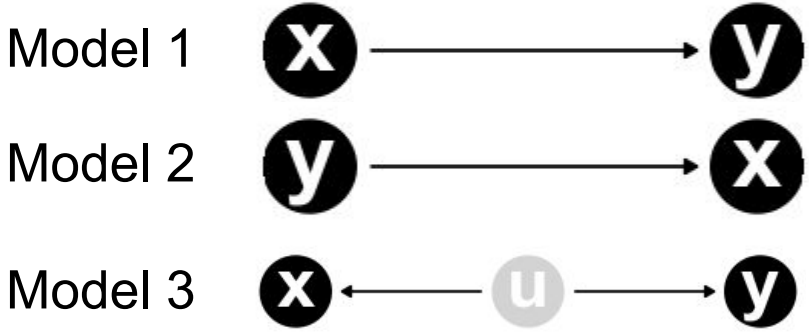


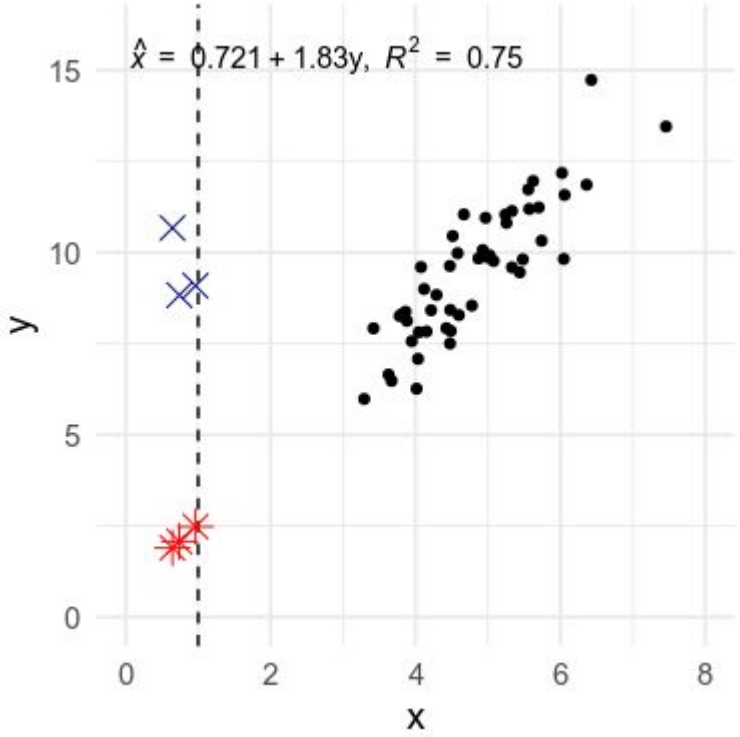
True effect: 0.0



Statistical models alone cannot derive causality from correlation

We learn causality by (1) listing models explicitly and (2) manipulating a variable and observe the outcomes



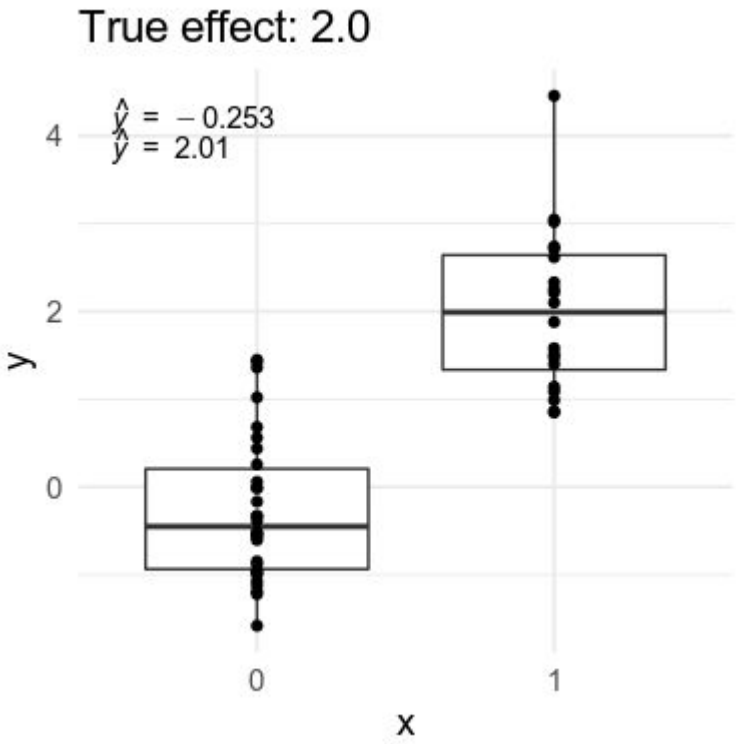


Assume that the data is generated by either Model 1, or Model 2, or Model 3. And assume that we can manipulate the value of X by setting it to 1.0 (the dash line).

Question: which outcomes (red stars or blue crosses) would support which models? Why?

Variables in models can be either continuous or discrete

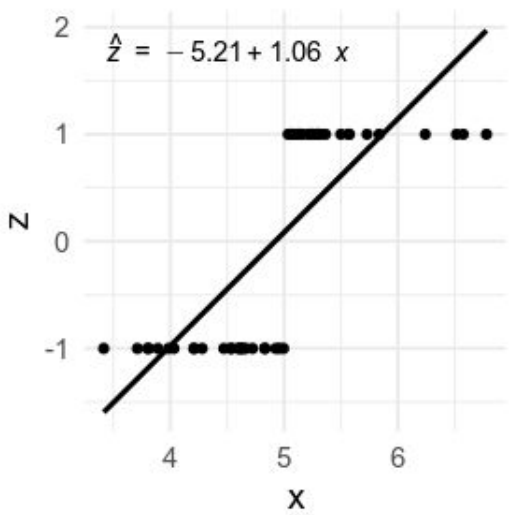
Model 1

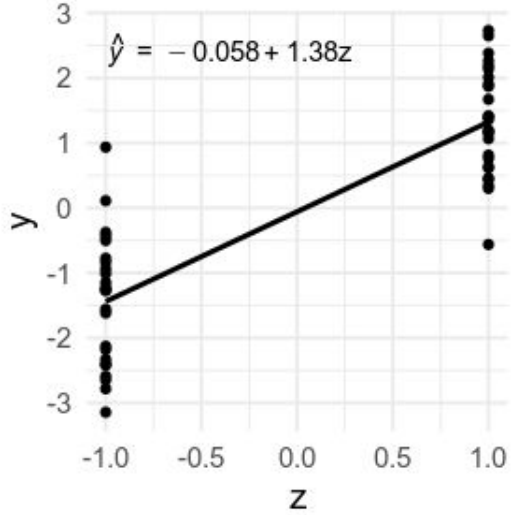


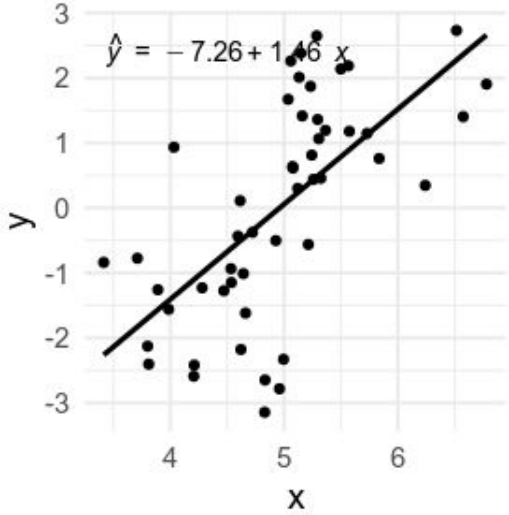
Assumptions of the **generative model**:

1. X is a random variable taking the value of either 0 or 1 with equal probability *p=0.5*.
2. Y is a random variable following Gaussian distributions:
   a. Mean(Y|X=0)=0
   b. Mean(Y|X=1)=2.0

Common Directed Acylic Graph (DAG) structures (1): Pipe







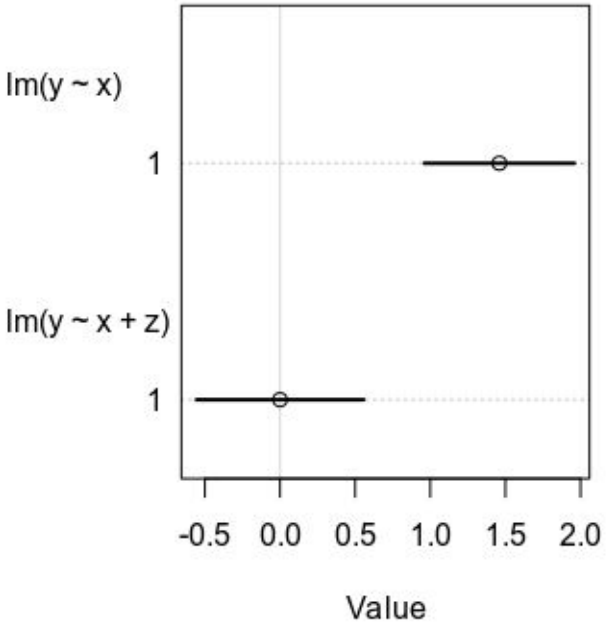
Assumptions of the **generative model**:

1. X is a random variable following Gaussian distribution N(5,1)
2. Z takes the value of -1 if X is smaller than 5, and 1 if X is equal to or larger than 5.
3. Y is a random variable with mean defined by Z*1.5.

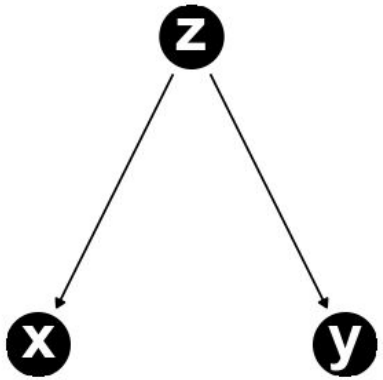
Conditional on the mediator in a pipe, the effect of the cause is blocked

Assumptions of the **generative model**:

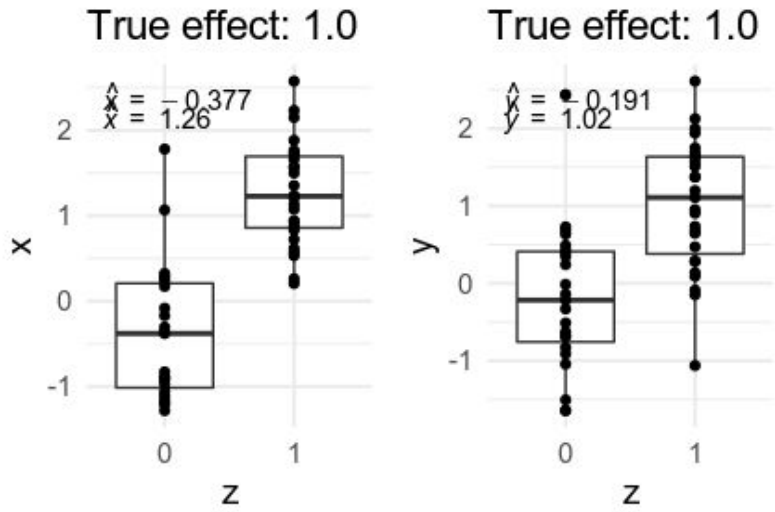
1. X is a random variable following Gaussian distribution N(5,1)
2. Z takes the value of -1 if X is smaller than 5, and 1 if X is equal to or larger than 5.
3. Y is a random variable with mean defined by Z*1.5.



Common DAG structures (2): Fork



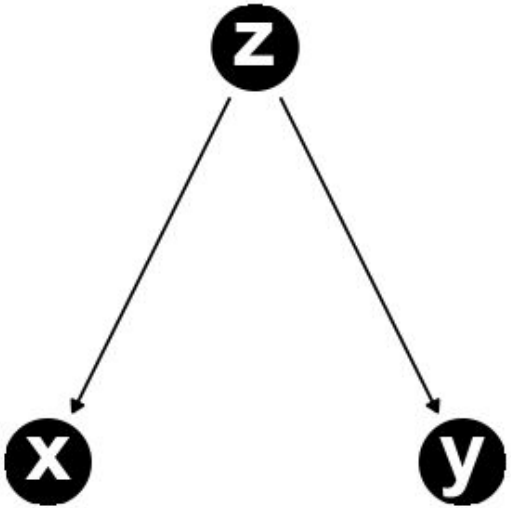
	x	z	y
1	0.20589018	1	-0.08893334
2	0.83248885	1	-0.14740781
3	-0.17082670	0	-1.64397573
4	-0.96908825	0	2.43509459
5	0.61354174	1	2.61131387
6	1.23211496	1	1.37149411
7	0.32263058	0	-1.04306293
8	0.24292789	0	-0.68668697
9	-1.10981001	0	0.24181405
10	0.72138854	1	2.12531733

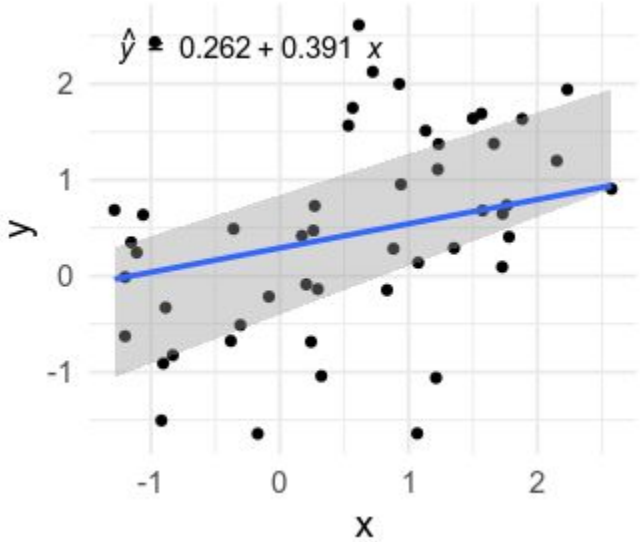


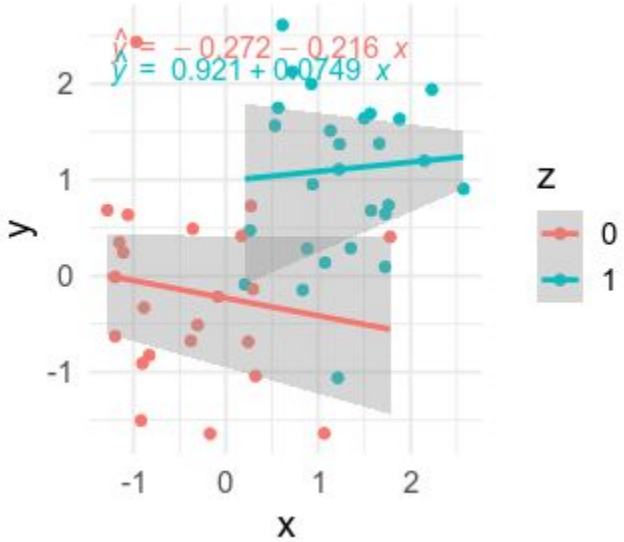
Assumptions of the **generative model**:

1. Z is a random variable taking the value of either 0 or 1.
2. Both X and Y are random variables following Gaussian distribution with mean equal to Z.

Conditioning on the fork *breaks* the correlation

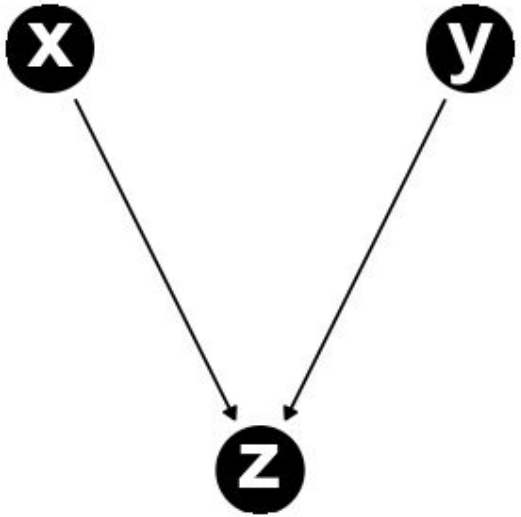




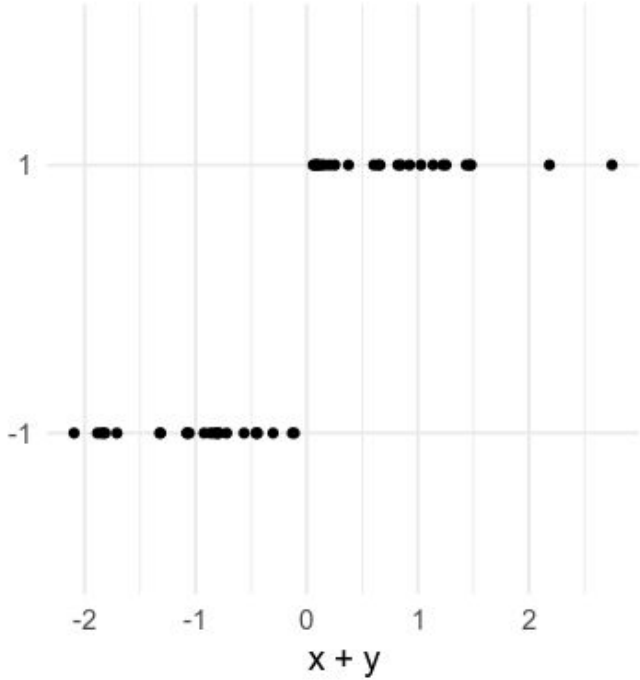


Given a fork structure, both children of the common cause are correlated. The correlation disappears when we condition on the common cause (i.e. stratification by the common cause in the case of discrete variables, or including the variable in the regression in the case of continuous variables).

Common DAG structures (3): Collider



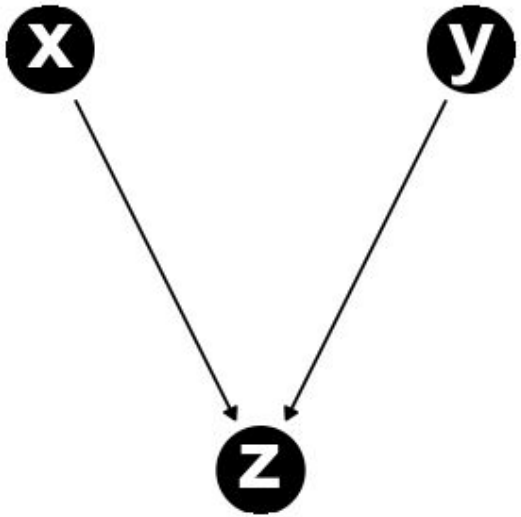
	x	z	y
1	0.835386320	1	-0.73897252
2	-0.005354014	-1	-0.82972315
3	0.058788286	1	0.76213369
4	-1.015602246	-1	-0.05951719
5	-0.339569780	-1	-0.11745910
6	-0.041077979	-1	-1.28243716
7	0.363740407	1	-0.30570762
8	0.119496314	-1	-1.19932461
9	0.257108454	-1	-1.06044066
10	0.304537158	-1	-0.43396492

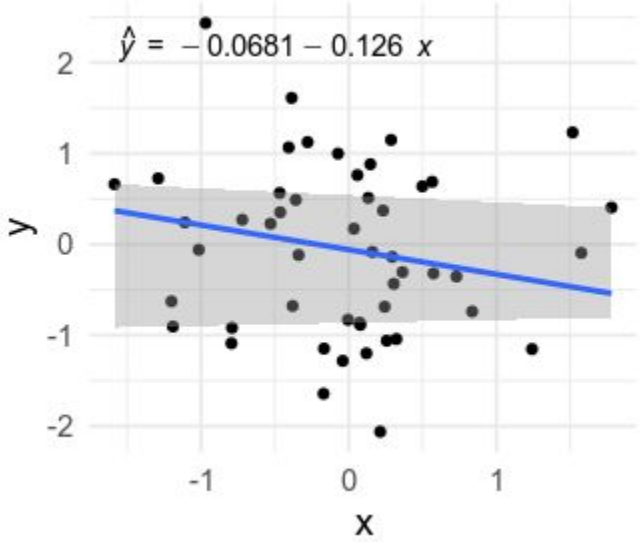


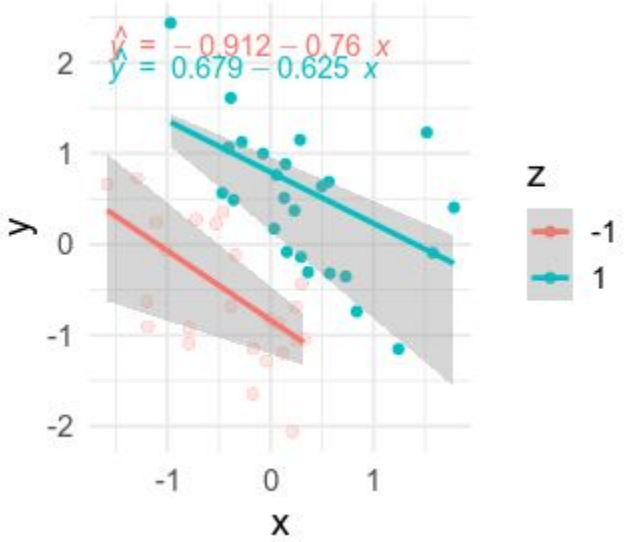
Assumptions of the **generative model**:

1. X and Y are random variables following Gaussian distribution N(0,1)
2. The value of Z is 1 if X+Y>0, and -1 if X-Y<=0.

Conditioning on the collider introduces *spurious correlations*



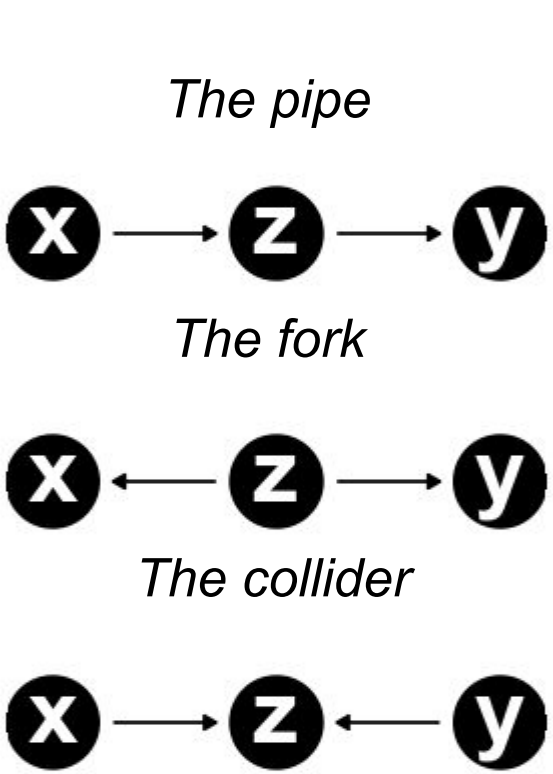




In a collider structure, the parents of the collider can be independent from each other. However, they become correlated when we condition on the collider.
Collider is everywhere!

A summary so far

- Data alone cannot tell causality, though in most cases we are interested in causal questions.
- Correlation between two variables can be caused by coincidence, causality, or common cause.
- Most common structures in a graph causal model are pipes, forks, and colliders. Stratifying by or regressing out variables may **remove** or **create** correlation.



	X correlated with Y?	When conditioning on Z?
The pipe	Yes	No
The fork	Yes	No
The collider	No	Yes

Stop exploitative data analysis, build generative models

?

?

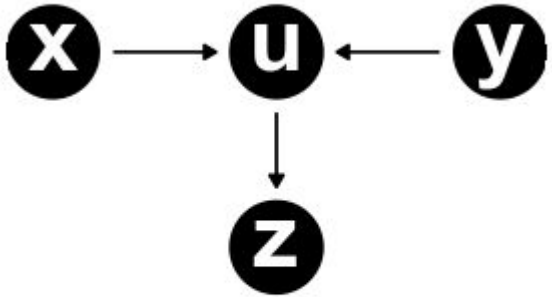
Biomarker, tox study, pathology, omics data, real-world data, EHR, ...

	x	z	y
1	0.835386320	1	-0.73897252
2	-0.005354014	-1	-0.82972315
3	0.058788286	1	0.76213369
4	-1.015602246	-1	-0.05951719
5	-0.339569780	-1	-0.11745910
6	-0.041077979	-1	-1.28243716
7	0.363740407	1	-0.30570762
8	0.119496314	-1	-1.19932461
9	0.257108454	-1	-1.06044066
10	0.304537158	-1	-0.43396492

?

?

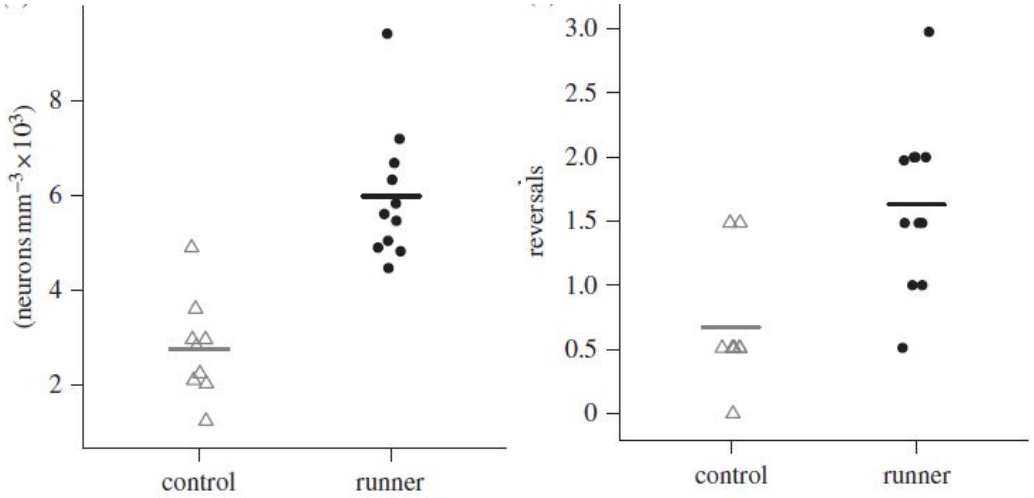
The descendant

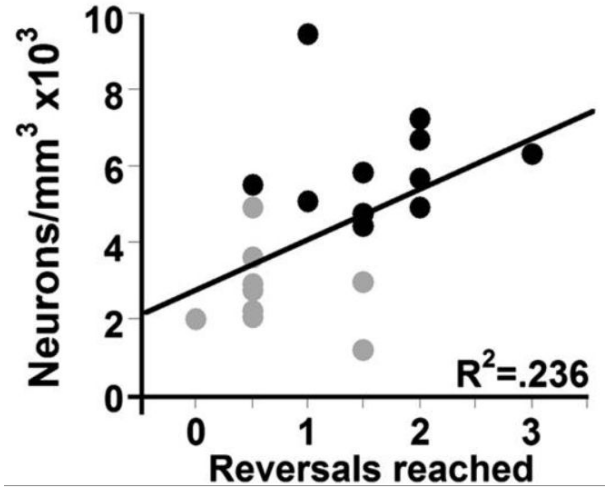


We need to build models (knowledge + assumptions) to infer causality

Running enhances spatial pattern separation in mice

Creer et al., PNAS 2010

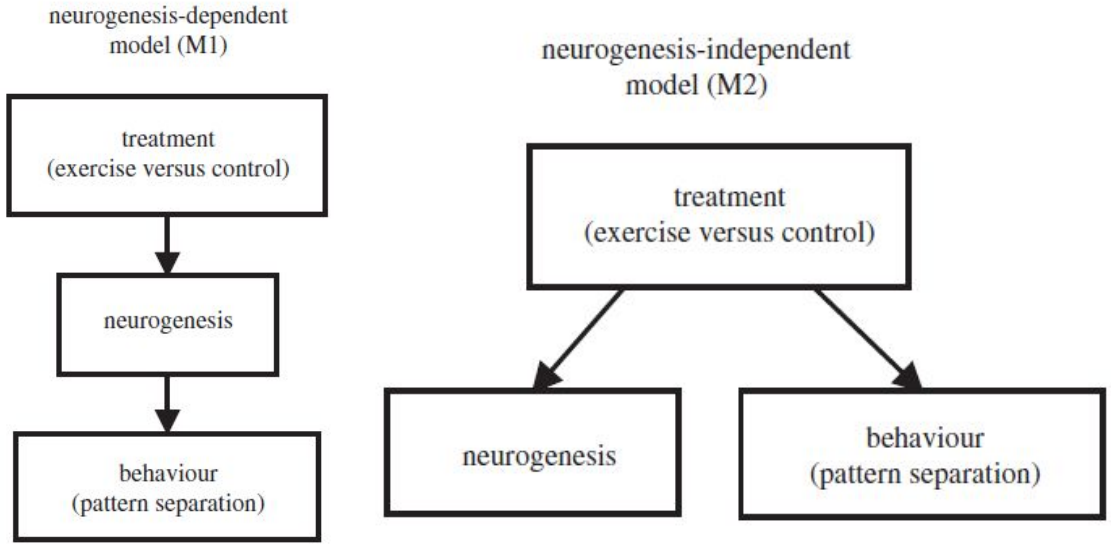




Creer, David J., Carola Romberg, Lisa M. Saksida, Henriette van Praag, and Timothy J. Bussey. “Running Enhances Spatial Pattern Separation in Mice.” Proceedings of the National Academy of Sciences 107, no. 5 (February 2, 2010): 2367–72. https://doi.org/10.1073/pnas.0911725107.

Lazic Stanley E. “Using Causal Models to Distinguish between Neurogenesis-Dependent and -Independent Effects on Behaviour.” Journal of The Royal Society Interface 9, no. 70 (May 7, 2012): 907–17. https://doi.org/10.1098/rsif.2011.0510.

Question: does pharmaceutical modulation of neurogenesis benefit pattern separation?

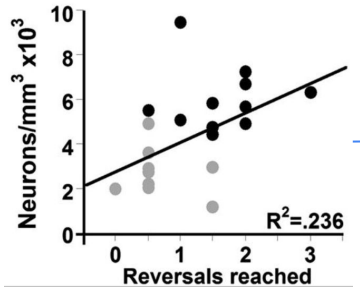


Lazic Stanley E. “Using Causal Models to Distinguish between Neurogenesis-Dependent and -Independent Effects on Behaviour.” Journal of The Royal Society Interface 9, no. 70 (May 7, 2012): 907–17. https://doi.org/10.1098/rsif.2011.0510.

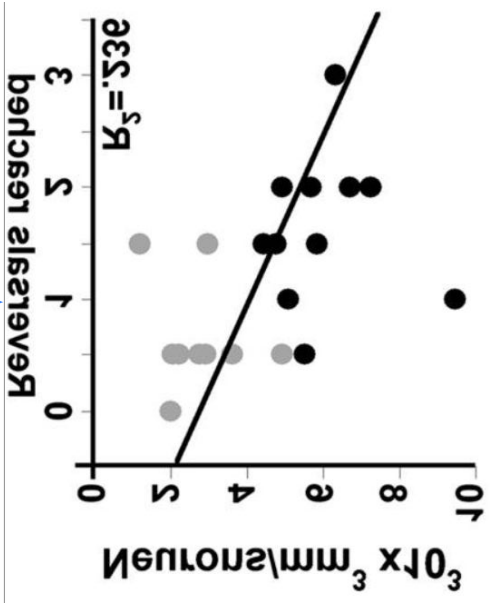
M1 (the **pipe** model) suggests that conditioned on neurogenesis, exercise and behaviour are independent (not correlated).

M2 (the **fork** model) suggests that conditioned on exercise, neurogenesis and behaviour are independent.

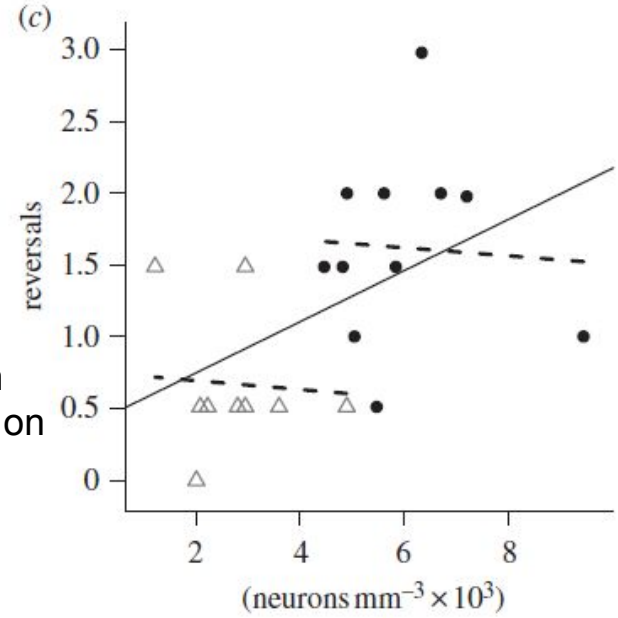
Behaviour and neurogenesis even shows *negative* correlation *conditional on* exercise- an example of Simpson's Paradox



Rotation/ mirroring



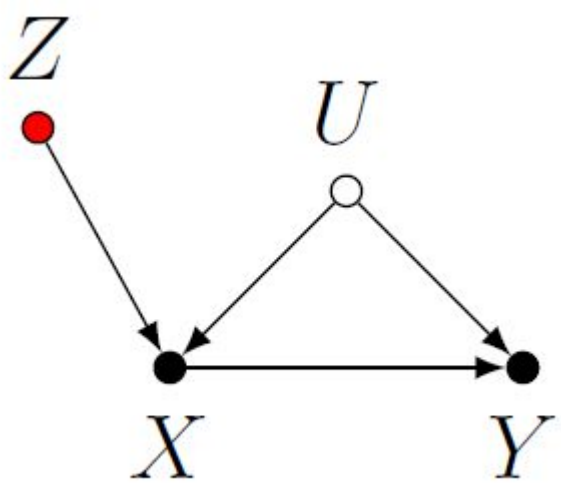
Regression conditional on treatment



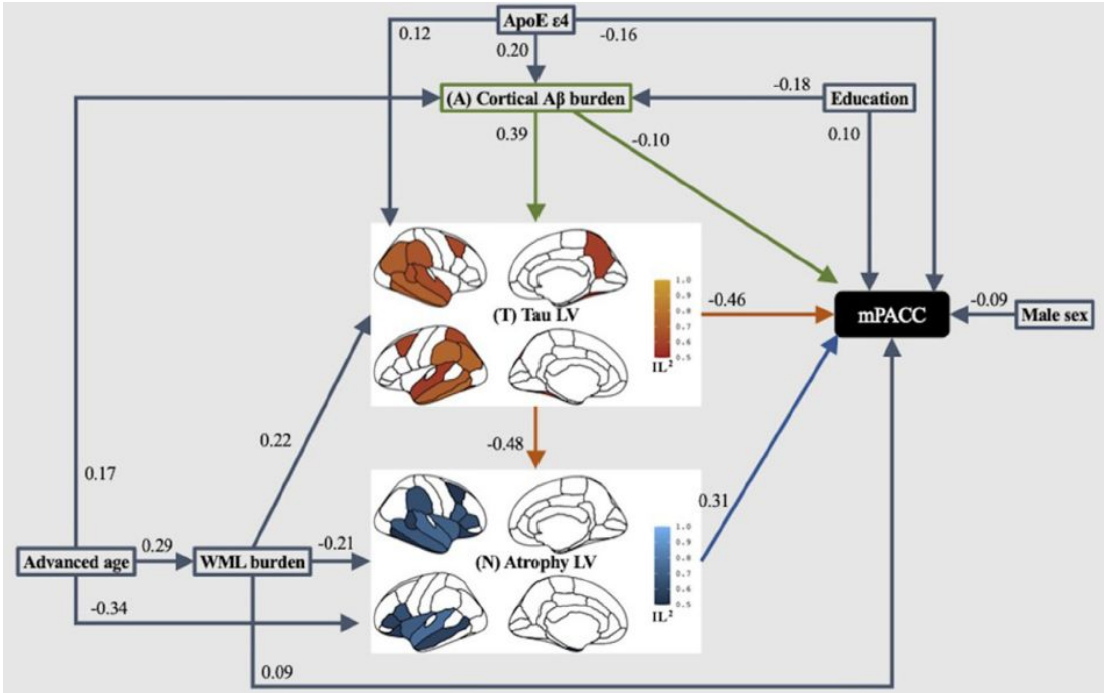
Based on the analysis, I believe model M2 is more likely to be true than M1.

Causal inference reduces bias in analysis by listing models explicitly

More causal models in drug discovery and development



Z is an *instrumental variable*. The model underlies Mendelian Randomization (MR).

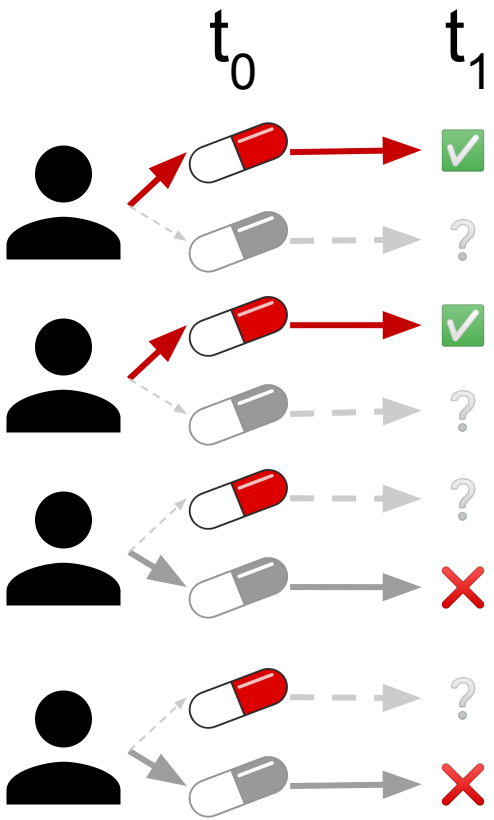


Cinelli, Carlos, Andrew Forney, and Judea Pearl. “A Crash Course in Good and Bad Controls.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, September 9, 2020. https://doi.org/10.2139/ssrn.3689437.

Tosun, Duygu, Zeynep Demir, Dallas P. Veitch, Daniel Weintraub, Paul Aisen, Clifford R. Jack Jr., William J. Jagust, et al. “Contribution of Alzheimer’s Biomarkers and Risk Factors to Cognitive Impairment and Decline across the Alzheimer’s Disease Continuum.” Alzheimer’s & Dementia n/a, no. n/a (2021). https://doi.org/10.1002/alz.12480.

Causal inference is important for both randomized experiments and observational studies

- In drug discovery and development, we often care about ***potential outcomes*** or ***counterfactuals***: what had if the patient received the alternative treatment, keeping everything else constant?
- **Randomized experiments** and **controlled trials** are gold-standard methods to address causal questions. Non-compliance and intermittent events call for causal analysis of the data even in randomized trials.
- Given causal models, it is sometimes possible to learn causal relationships from observational data as well.



Models in disease understanding and drug discovery

Computational simulations

Mechanistic/physical models

Structural causal models

Causal graphical models

Descriptive graphical models

Statistical models

Model	Predict in i.i.d. setting	Predict under changing distr. or intervention	Answer counterfactual questions	Obtain physical insight	Learn from data
Mechanistic/ physical, e.g., Sec. 2.3	yes	yes	yes	yes	?
Structural causal model, e.g., Sec. 6.2	yes	yes	yes	?	?
Causal graphical model, e.g., Sec. 6.5.2	yes	yes	no	?	?
Statistical model, e.g., Sec. 1.2	yes	no	no	no	yes

Hierarchy

Molecular models

Omics and cellular models

Organ- and system-level models

Population models

Enzymic

Cellular models

Microphysiological systems

Animal models

Biological simulations

Consequences

1. **Data alone does not answer causal questions**: whenever we are interested in interventions (modulating a target, changing the structure of a molecule, *etc.*), predictive tools such as linear regression, machine learning, and artificial intelligence models must be embedded in the causal framework.
2. **Addressing causal questions:**
    a. Derive causal models using science, making assumptions transparent
    b. Program the model as a generative simulation
    c. Design research and validate statistical analysis using (b)
    d. Confront the model with data, share both wins and losses transparently with others
    e. Revise and repeat
3. **Model first, data second**: From **DA** (**D**ata and **A**nalytics) to **MADAM** (**M**odel construction, **A**nalysis of the model, **D**ata collection, **A**nalysis of the data with the model, and **M**odel refinement)

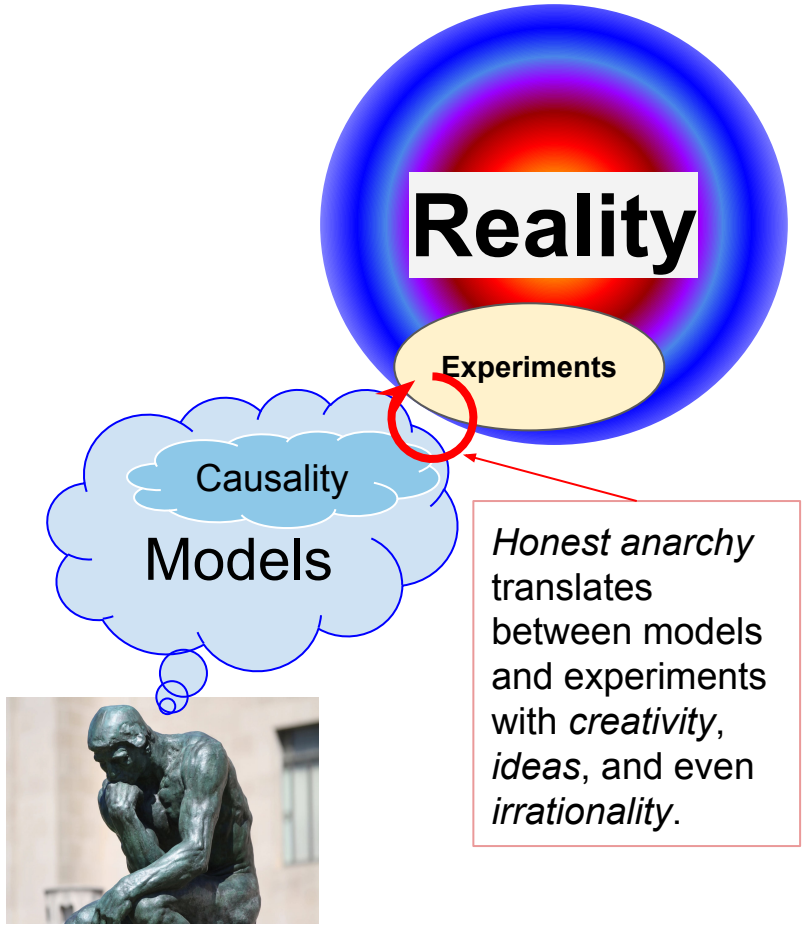
Resources for learning and doing causal inference

- There are many useful resources of learning causal inference, such as *Rethinking Statistics* and *Causal Inference for Statistics, Social, and Biomedical Sciences*, and *Causal Inference in Statistics - A Primer*. Join our reading club to learn more!
- **Put science before statistics. Learn and use both.**
- To start, I recommend Python and R packages:
  - [Python] DoWhy by Microsoft Research: https://github.com/microsoft/dowhy
  - [R] *dagitty* and *ggdag* packages
  - Examples in both Jupyter Notebook and Rmarkdown can be found my code.roche.com repo. It also contains the codes to generate models and simulate data for this presentation.
- Once you are becoming experienced with the causal thinking and need more effective tools
  - [Python] *econml* by Microsoft Research, mainly for econometrics but useful for other fields, too.
  - In most cases, it is better to build bespoke Bayesian models with dedicated tools like OpenBUGS/Stan/PyMC. See my code repo above for examples of using Stan via the *rethinking* package in R by Richard McElreath.

*“There is no method for making causal models other than **science**. There is no method to science other than **honest anarchy**.”*

- Richard McElreath

Science is not about reality.
Science is always about models.



References

1. Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. Causal Inference in Statistics: A Primer. Chichester, West Sussex: Wiley, 2016.
2. Shipley, Bill. Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference with R. Cambridge (GB): Cambridge university press, 2016.
3. Imbens, Guido, and Donald B. Rubin. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. New York: Cambridge University Press, 2015.
4. McElreath, Richard. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. 2nd ed. CRC Texts in Statistical Science. Boca Raton: Taylor and Francis, CRC Press, 2020.