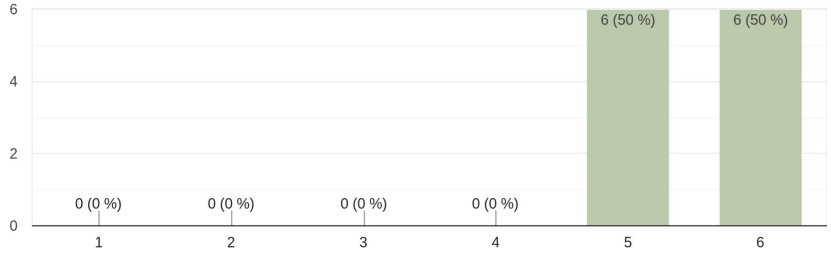
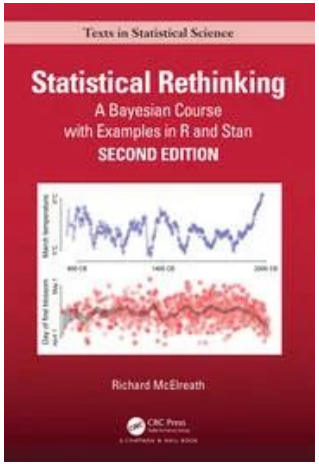
Feedback of lecture 6

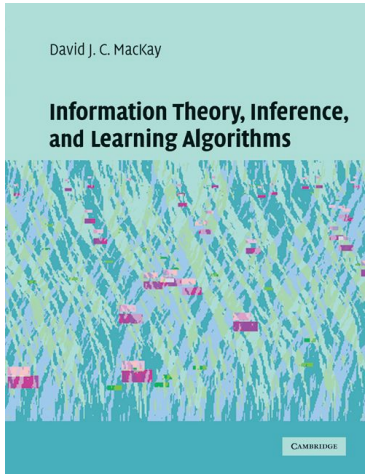
How was your overall impression of the sixth lecture?
12 Antworten



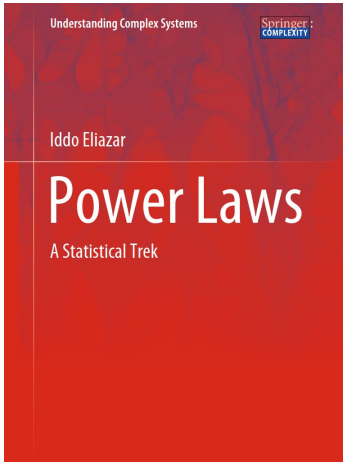
\+ Team discussions were good
\+ More examples and time for distributions would be better



McElreath, Richard. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. 2nd ed. CRC Texts in Statistical Science. Taylor and Francis, CRC Press, 2020. YouTube, GitHub



MacKay, David J. C. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003. http://www.inference.org.uk/mackay/itila/book.html, recordings on YouTube



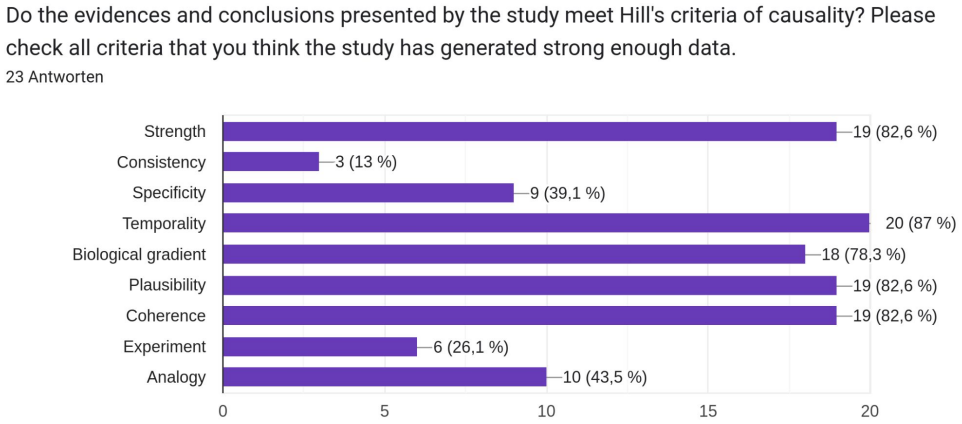
Eliazar, Iddo. Power Laws: A Statistical Trek. Springer Complexity. Springer Nature, 2020. https://doi.org/10.1007/9783030332358.

Offline activities of lecture 6 (23 participants)

Controversial/counterintuitive views from A.B. Hill's speech:

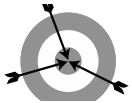
- Causation cannot rely on statistics or p-values alone.
- Scientific knowledge is incomplete; research may still be needed.
- Open-mindedness toward improbable or unconventional evidence is essential.
- Human and contextual factors complicate causal interpretation.
- Hill's criteria guide judgment, not serve as strict rules.

It is only interesting how much is a chance that if "Exposure to sugar rationing in the first 1000 days of life protected against chronic disease" was done infinite times, what would be the result? Is is expected answer or just random values?



AMIDD 2024 Lecture 7: causal inference


Strength


Consistency


Specificity


Temporality


Dose response


Plausibility


Coherence


Experiment


Analogy


Reversibility

Adapted from “The Environment and Disease: Association or Causation?” (1965) by A. B. Hill.

Dr. Jitao David Zhang, Computational Biologist

[1] Pharmaceutical Sciences, Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche

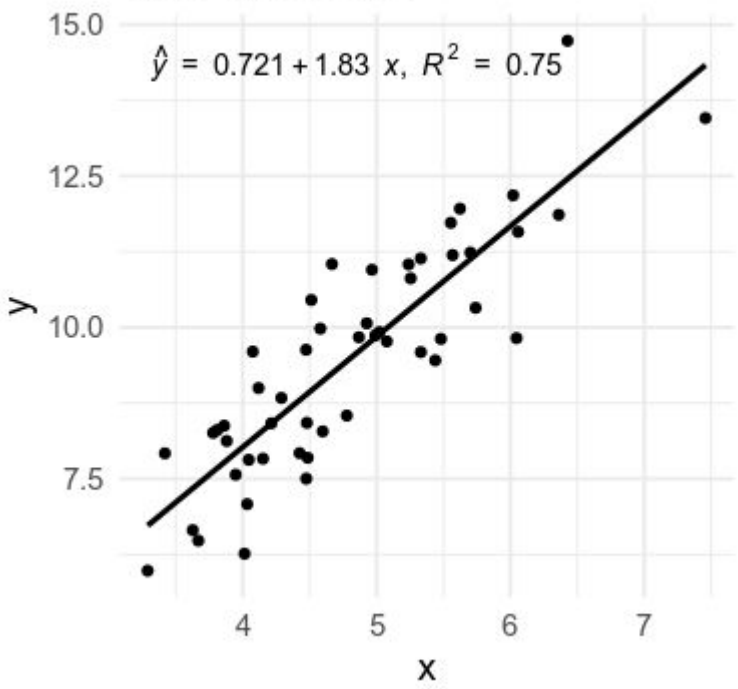
[2] Department of Mathematics and Informatics, University of Basel

The simplest linear model has three components: the intercept, the slope, and a measure of fit

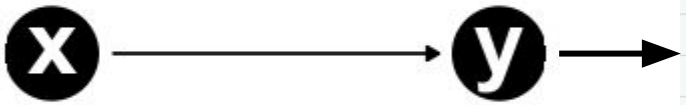
In this example, the coefficient of determination (R^2) is used as the measure.

R^2 measures the relative fit of the linear model with regard to a baseline model, where the mean value of *y* is used as a fit.

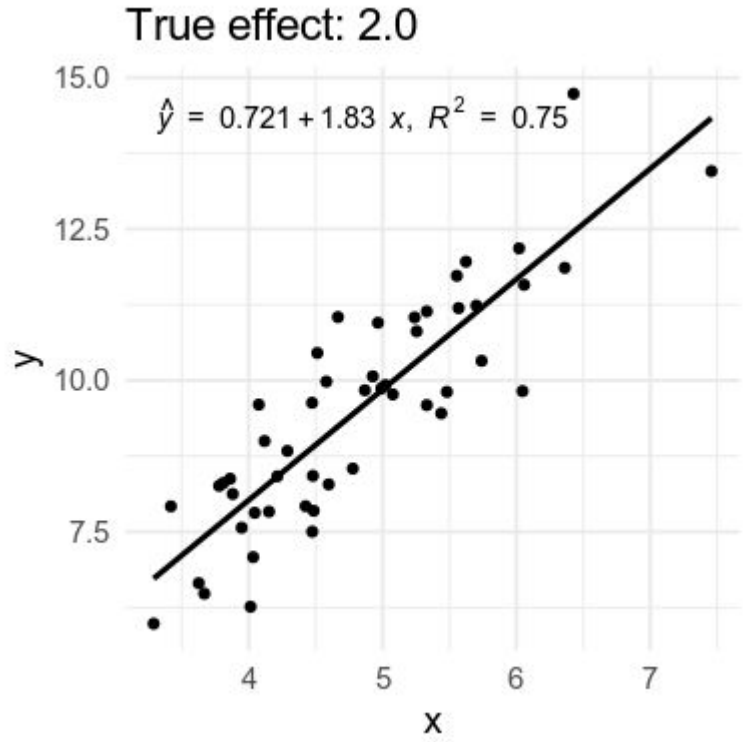
	x	y
1	4.926791	10.067779
2	4.479734	8.424283
3	4.289686	8.835629
4	4.474023	9.630499
5	4.214551	8.416680
6	6.057431	11.578080
7	4.597903	8.283025
8	5.021571	9.922731
9	3.627323	6.651222
10	5.622794	11.959972
11	5.555025	11.727815
12	4.966007	10.951562
13	5.076791	9.768299



Generative models shed light on correlation and causality



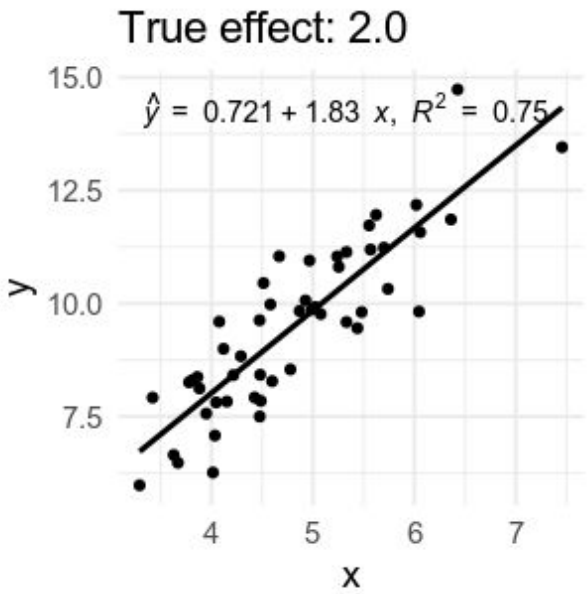
	x	y
1	4.926791	10.067779
2	4.479734	8.424283
3	4.289686	8.835629
4	4.474023	9.630499
5	4.214551	8.416680
6	6.057431	11.578080
7	4.597903	8.283025
8	5.021571	9.922731
9	3.627323	6.651222
10	5.622794	11.959972
11	5.555025	11.727815
12	4.966007	10.951562
13	5.076791	9.768299

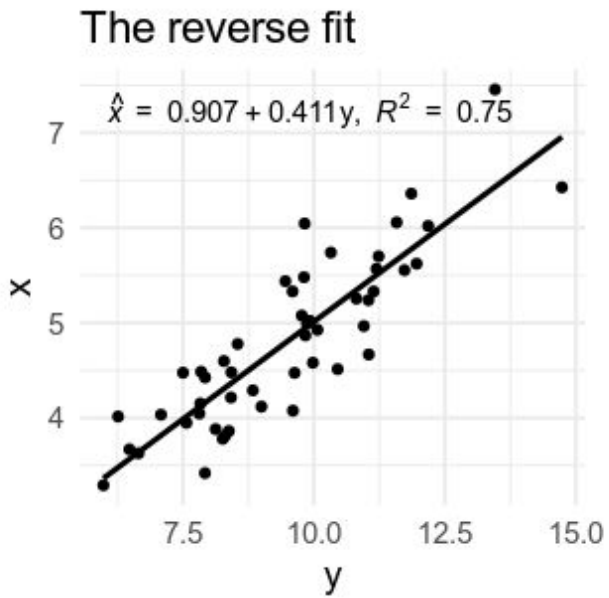


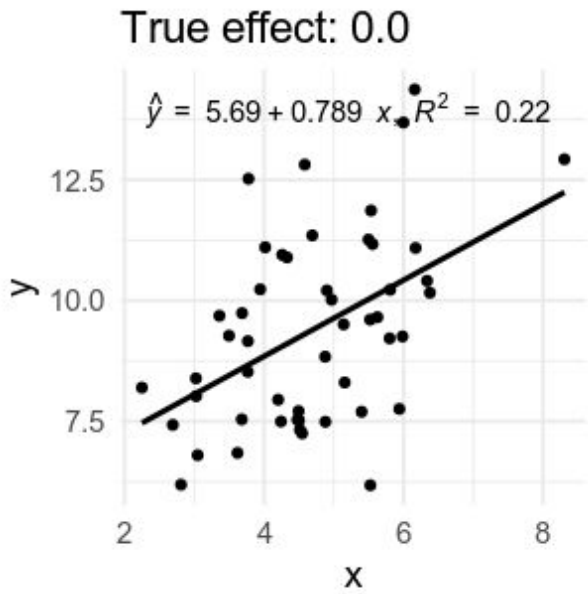
Assumptions of the **generative model**:

1. X is a random variable;
2. Every unit change of X induces a change of 2 units in Y.

Correlation may be coincidence, or causation, or confounding (common cause)

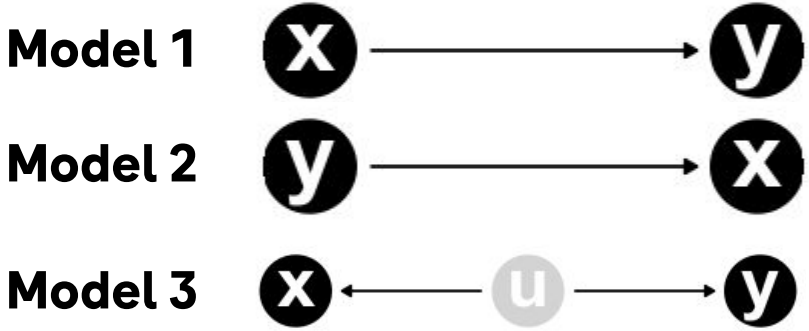






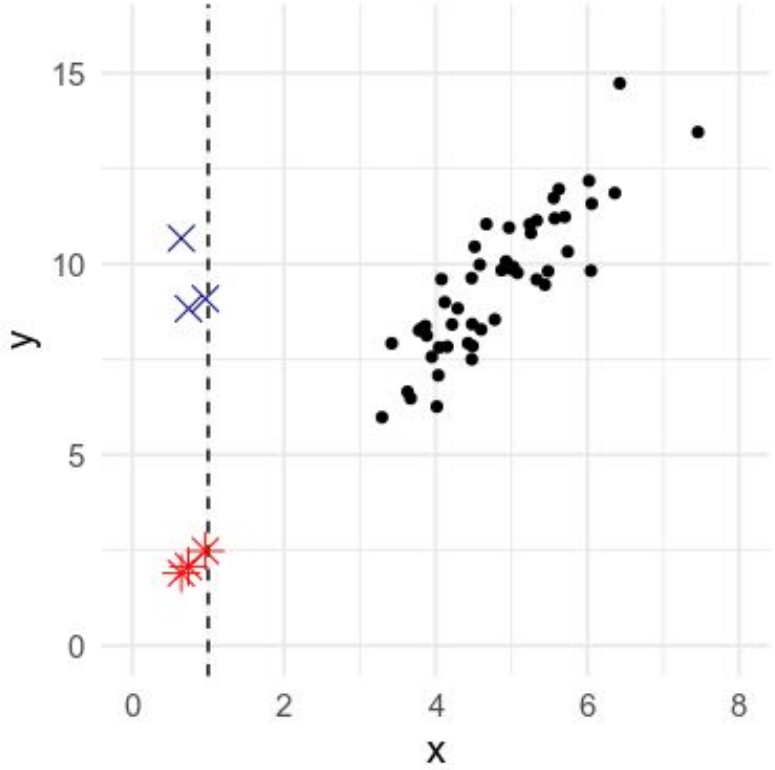
Statistical models alone cannot derive causality from correlation

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



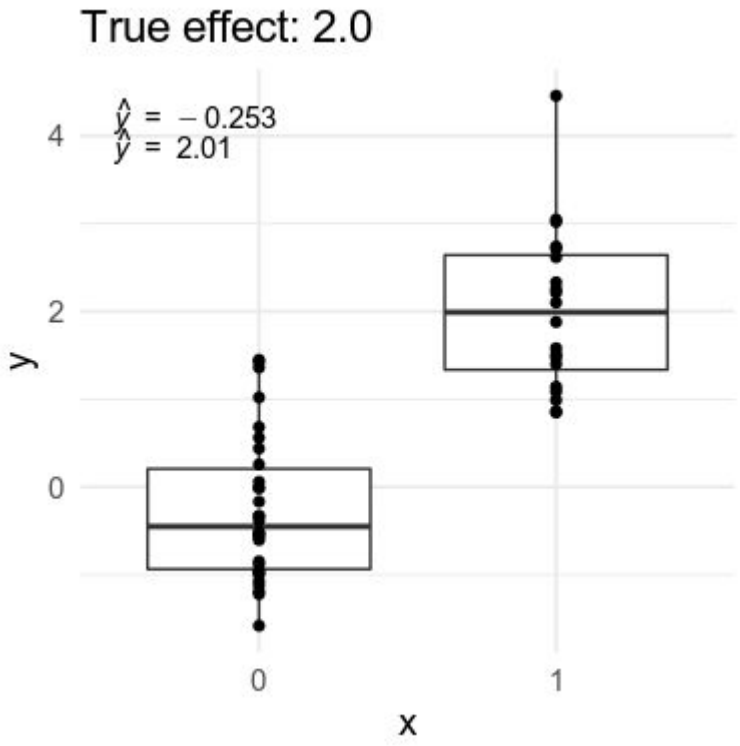
Assume that the data is generated by either Model 1, or Model 2, or Model 3. And assume that we can manipulate the value of X by setting it to 1.0 (the dash line).

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



Variables in models can be either continuous or discrete

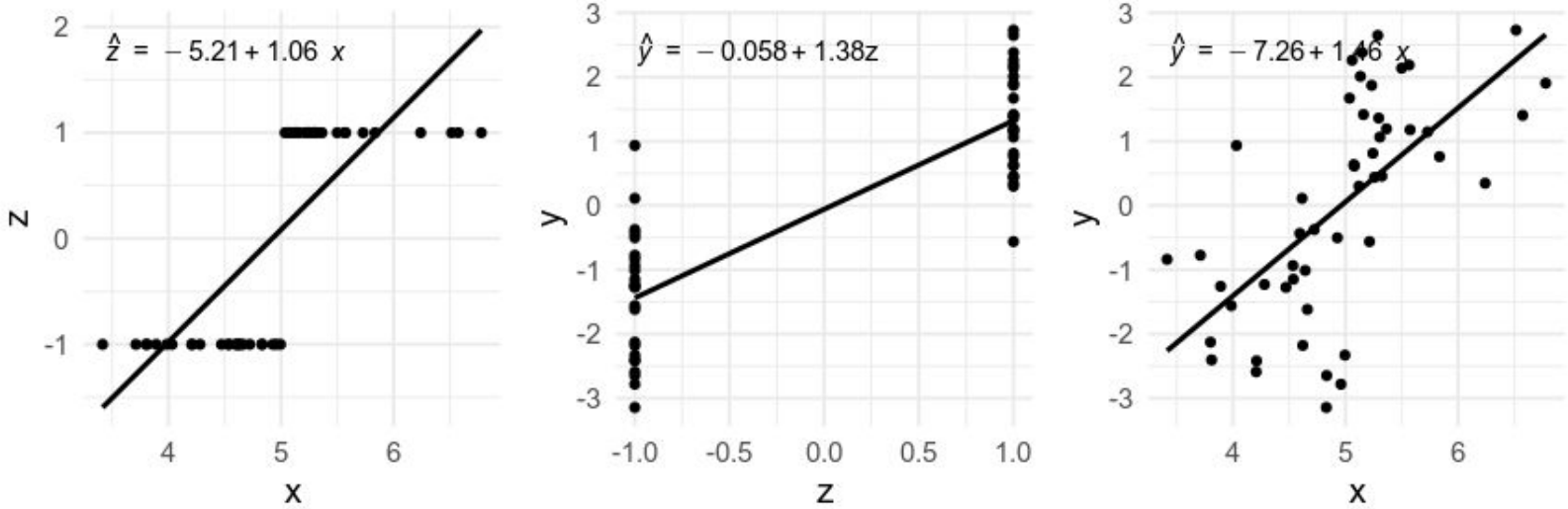
Model 1



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): the Pipe



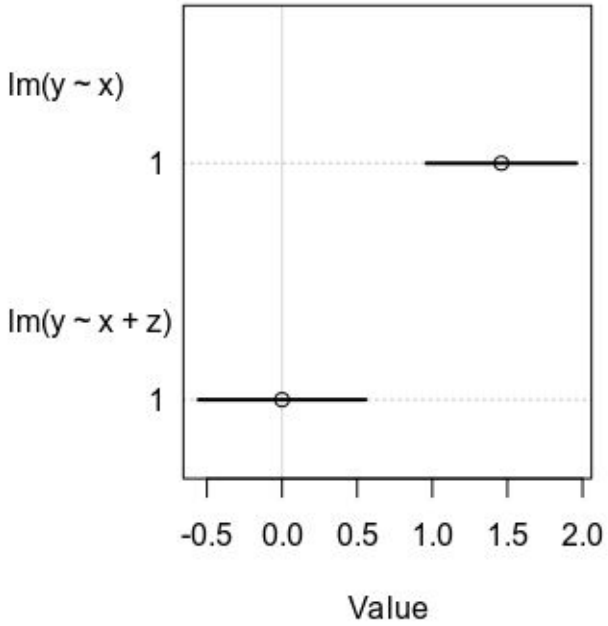
Assumptions of the **generative model**:

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.

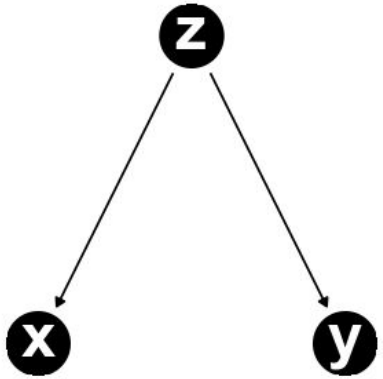
Conditional on the mediator in a pipe, the effect of the cause is blocked

Assumptions of the **generative model**:

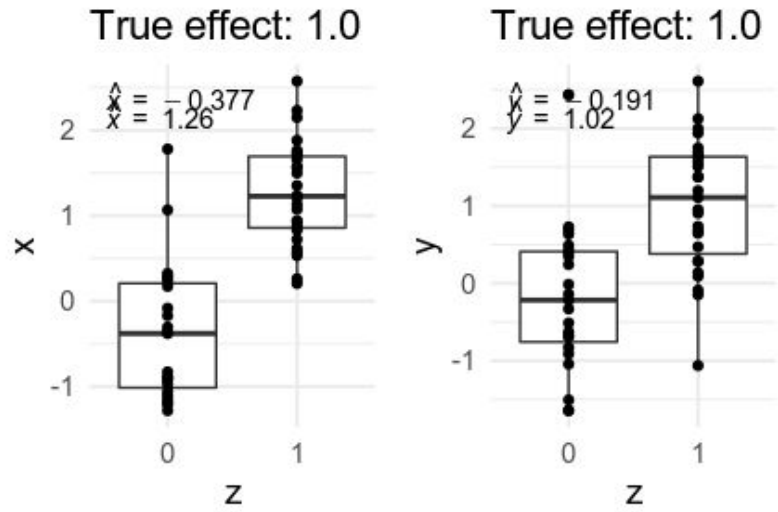
1. X is a random variable following Gaussian distribution N(5,1)
2. Z takes the value of -1 if X is smaller than 5, and X is equal to or larger than 5.
3. Y is a random variable with mean defined by Z*1



Common DAG structures (2): The Fork



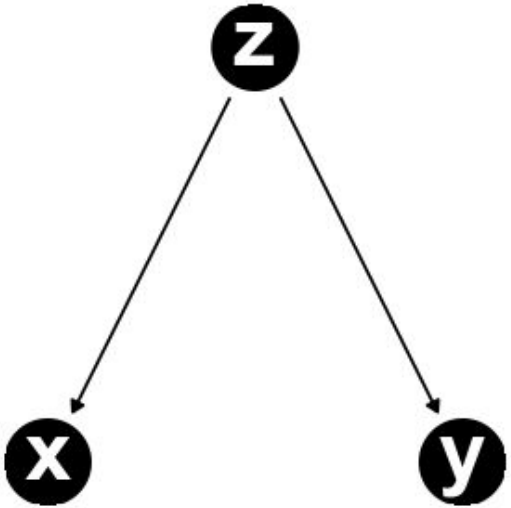
	x	z	y
1	0.20589018	1	-0.08893334
2	0.83248885	1	-0.14740781
3	-0.17082670	0	-1.64397573
4	-0.96908825	0	2.43509459
5	0.61354174	1	2.61131387
6	1.23211496	1	1.37149411
7	0.32263058	0	-1.04306293
8	0.24292789	0	-0.68668697
9	-1.10981001	0	0.24181405
10	0.72138854	1	2.12531733

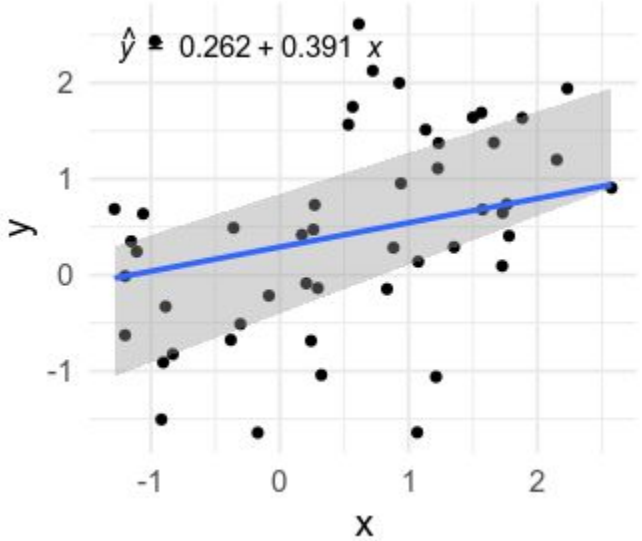


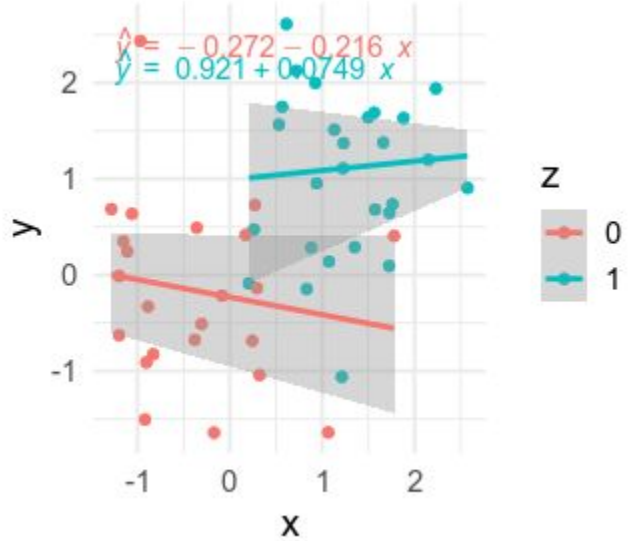
Assumptions of the **generative model**:

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

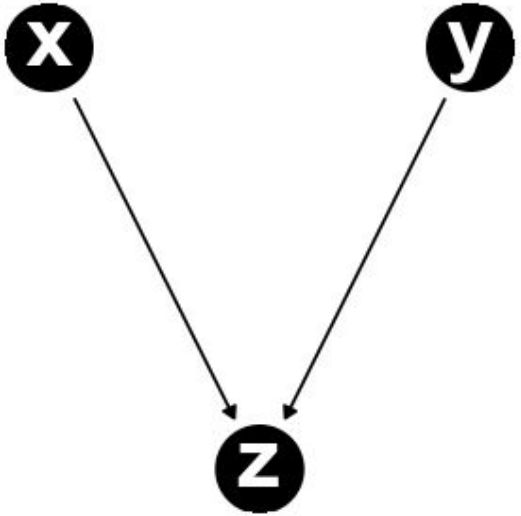




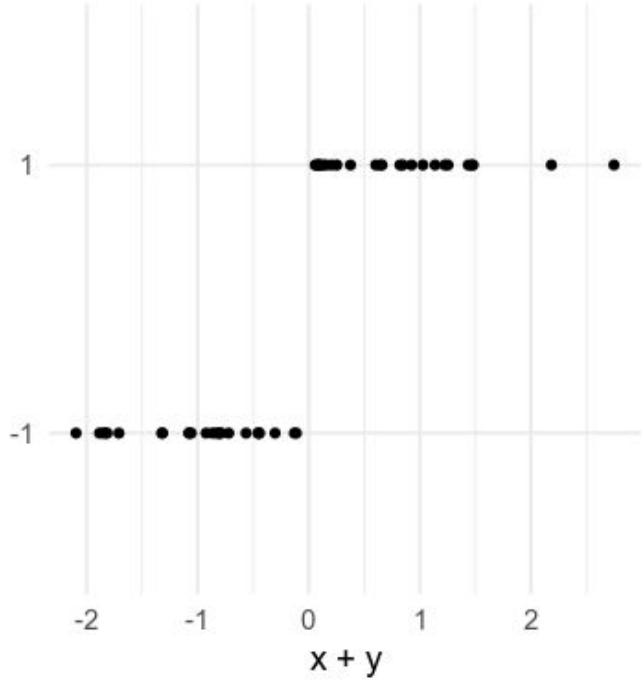


Given a fork structure, both children of the common cause are correlated. The correlation disappears when we condition on the common cause (i.e. stratification by the common cause in the case of discrete variables, or including the variable in the regression in the case of continuous variables).

Common DAG structures (3): The Collider



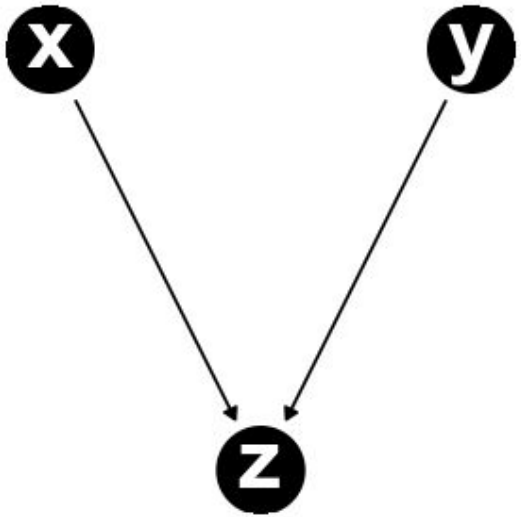
	x	z	y
1	0.835386320	1	-0.73897252
2	-0.005354014	-1	-0.82972315
3	0.058788286	1	0.76213369
4	-1.015602246	-1	-0.05951719
5	-0.339569780	-1	-0.11745910
6	-0.041077979	-1	-1.28243716
7	0.363740407	1	-0.30570762
8	0.119496314	-1	-1.19932461
9	0.257108454	-1	-1.06044066
10	0.304537158	-1	-0.43396492

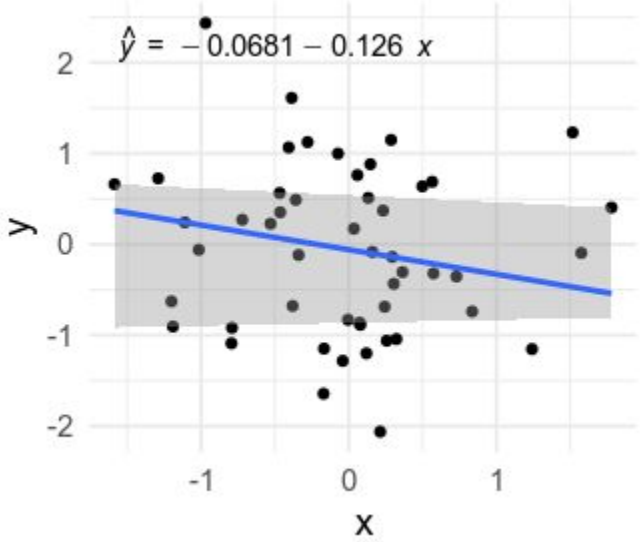


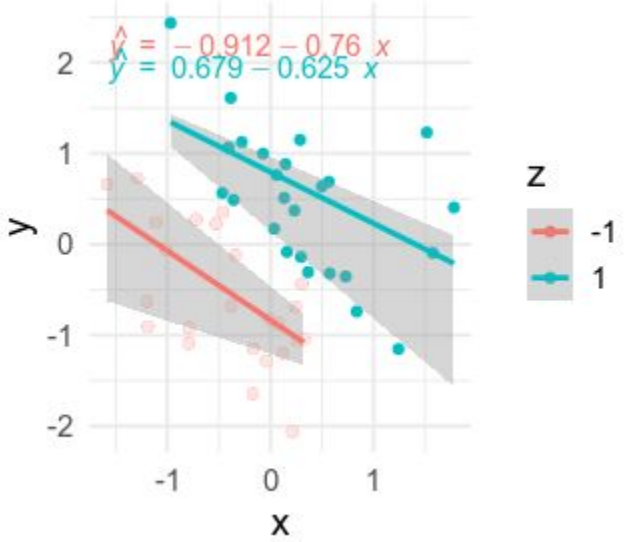
Assumptions of the **generative model**:

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.

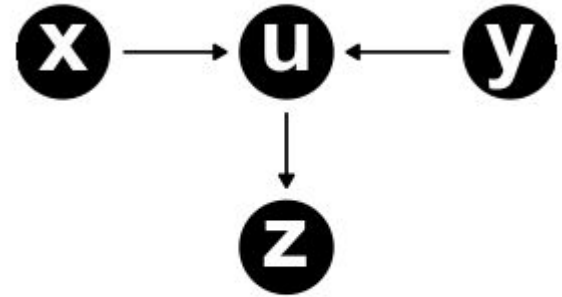
	X correlated with Y?	When conditioning on Z?
The pipe 	**Yes**	**No**
The fork 	**Yes**	**No**
The collider 	**No**	**Yes**

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

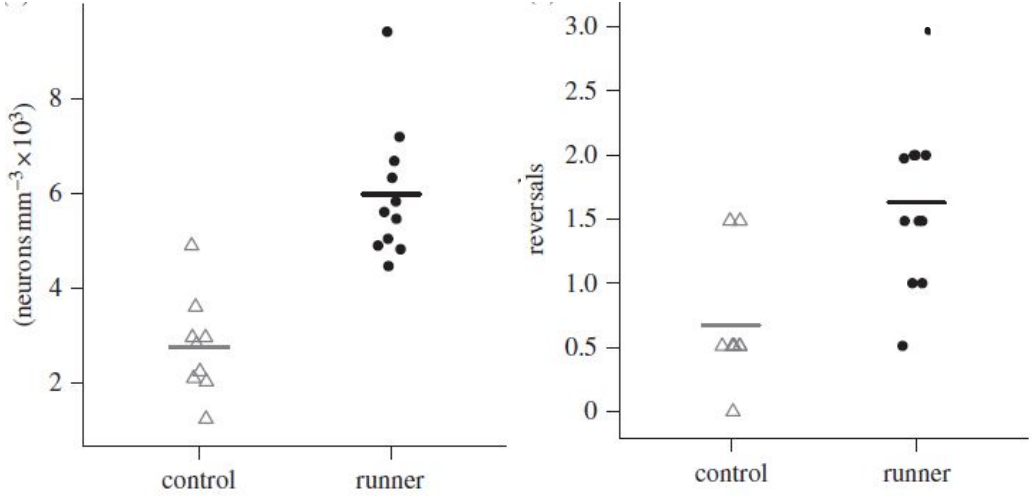


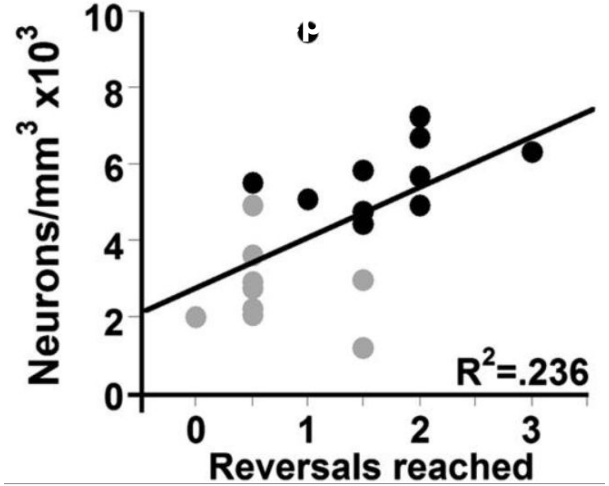
We need to build models (knowledge + assumptions) to infer causality

End of lecture on October 31st, 2025

Claim: running enhances spatial pattern separation in mice

Creer et al., PNAS 2010

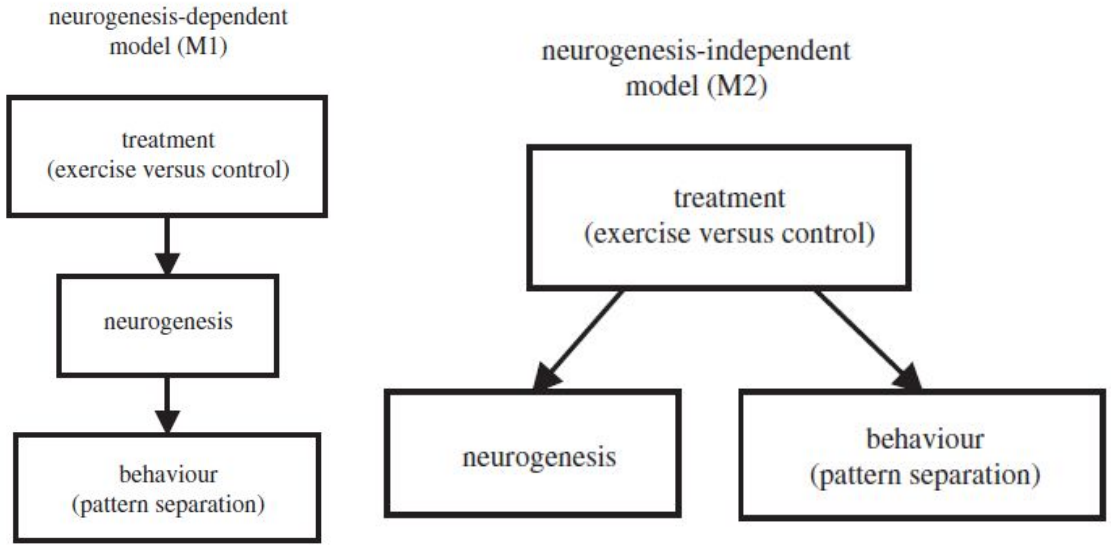




Creer, David J., Carola Romberg, Lisa M. Saksida, Henriette van Praag, and Timothy J. Bussey. “Running Enhances Spatial Pattern Separation in Mice.” Proceedings of the National Academy of Sciences 107, no. 5 (February 2, 2010): 2367–72. https://doi.org/10.1073/pnas.0911725107.

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?

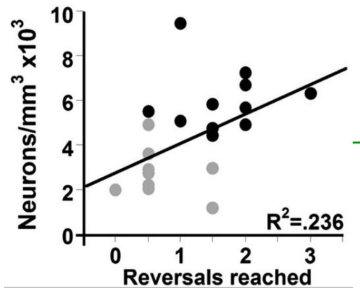


M1 (the **pipe** model) suggests that conditioned on neurogenesis, exercise and behaviour are independent (not correlated).

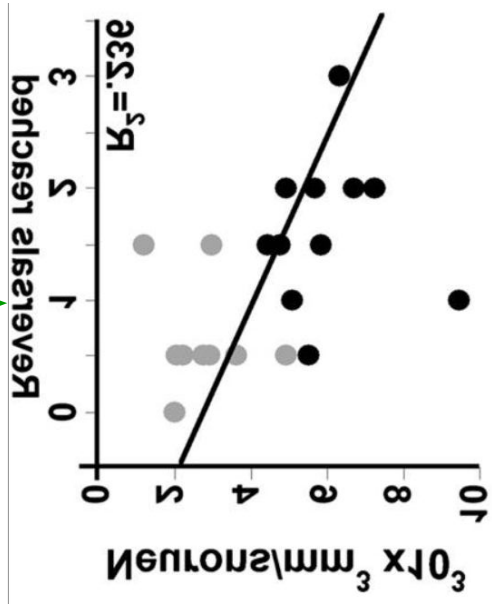
M2 (the **fork** model) suggests that conditioned on exercise, neurogenesis and behaviour are independent.

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.

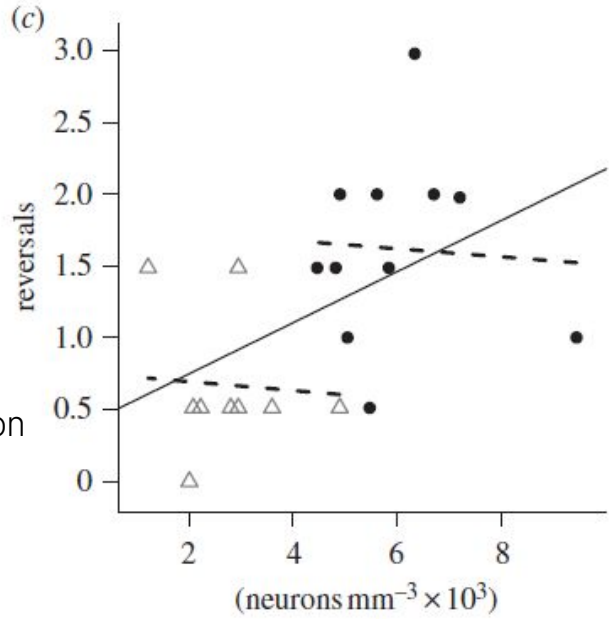
Behaviour and neurogenesis even shows *negative* correlation *conditional on* exercise- an example of Simpson's Paradox



Rotation/ mirroring



Regression conditional on treatment

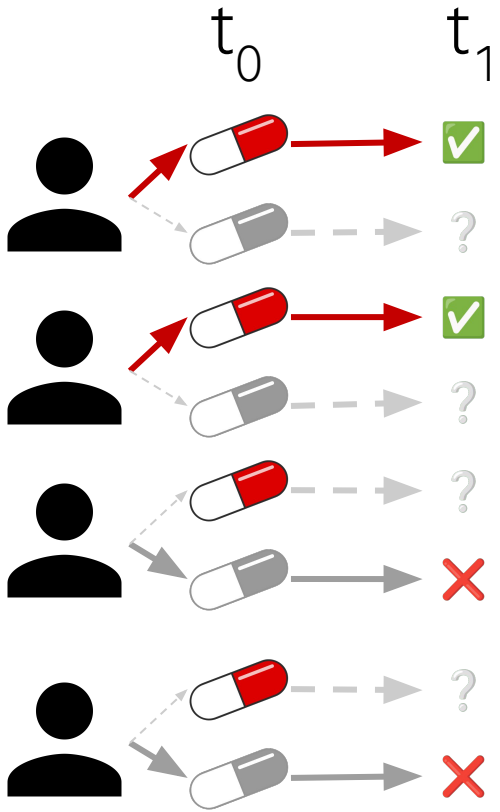


Based on the analysis, I believe model M2 is more likely to be true than M1.

Causal inference reduces bias in analysis by listing models explicitly

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.



Causal inference is a missing data problem

Individual	Treatment	Value (AU)
1	Control	75
2	Control	73
3	Control	74
4	Treatment	55
5	Treatment	45
6	Treatment	60

A classical textbook

Individual	Value (AU) with Control	Value (AU) with Treatment
1	75	?
2	73	?
3	74	?
4	?	55
5	?	45
6	?	60

A classical textbook in 50 years

Assignment mechanism determines which data are missing, and determine the statistical technique to be used

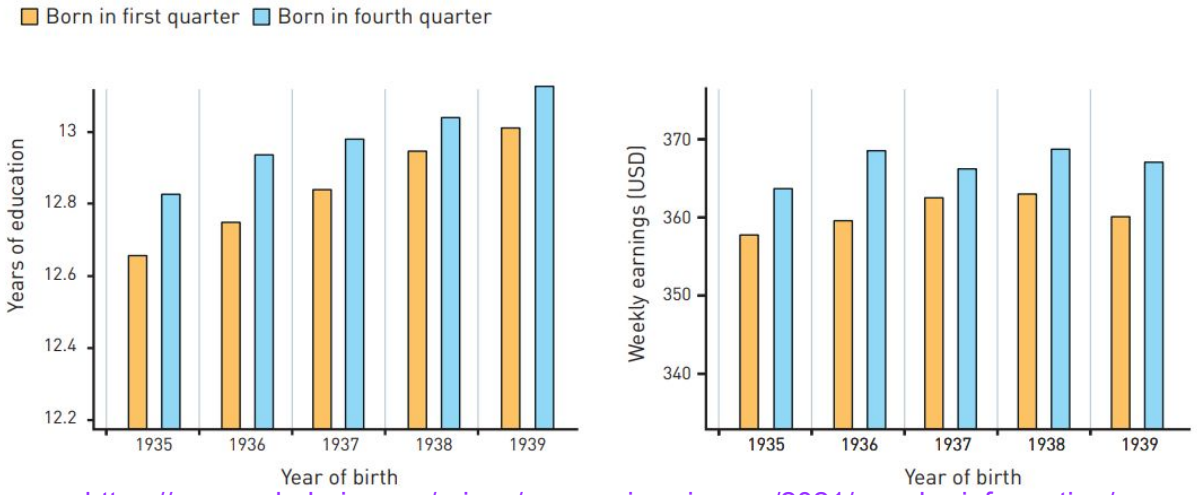
Individual	Value (AU) with Control	Value (AU) with Treatment
1	75	?
2	73	?
3	74	?
4	?	55
5	?	45
6	?	60

- **Classical Randomized Experiments**: we control the assignment mechanism;
- **Regular assignment (*observational studies)***: we know part but not all of the assignment mechanism;
- **Regular assignment with non-compliance**: we need an *instrumental variable*.

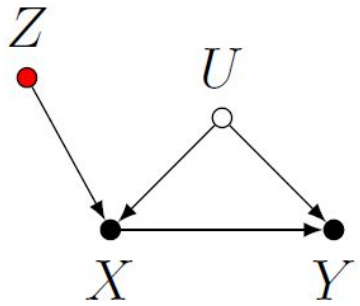
Instrumental variable helps to dissect causal from confounding effects

People born late in the year have more years of education and higher incomes

Additional years of education have a positive effect on income. The figure uses data from Angrist and Krueger (1991).

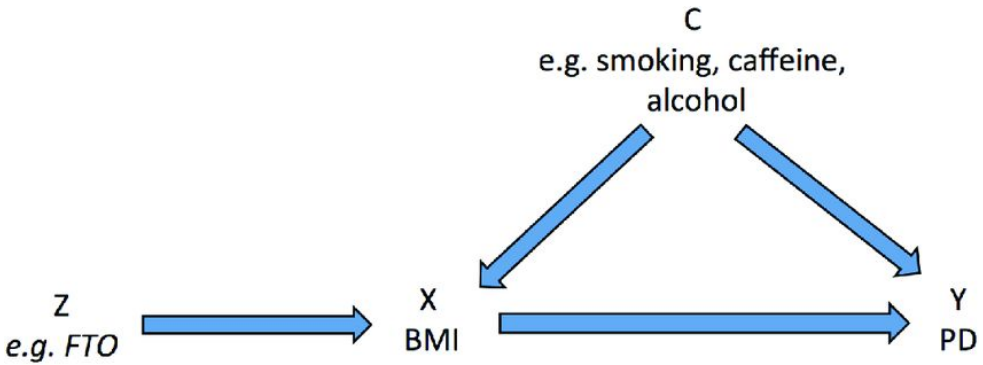


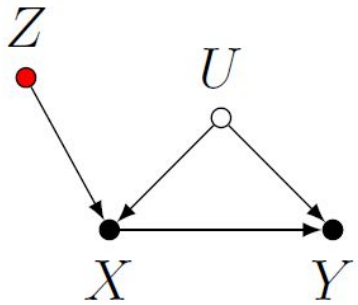
https://www.nobelprize.org/prizes/economic-sciences/2021/popular-information/



Z: Birthday
X: Education
Y: Income
U: Socioeconomic and individual factors

Instrumental variable is critical for both Mendelian Randomization (MR) and handling with non-compliance





Z: Assignment (Placebo/Drug)
X: Treatment
Y: Value of interest
U: unobserved factors

Noyce, Alastair J., Demis A. Kia, Gibran Hemani, Aude Nicolas, T. Ryan Price, Eduardo De Pablo-Fernandez, Philip C. Haycock, et al. “Estimating the Causal Influence of Body Mass Index on Risk of Parkinson Disease: A Mendelian Randomisation Study.” PLoS Medicine 14, no. 6 (June 2017): https://doi.org/10.1371/journal.pmed.1002314.

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.

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 when necessary:**
   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)

Ten Simple Rules of Causal Inference

1. Clarify whether correlation or causation is of interest.
2. Draw models of knowledge, assumptions, and data-generation processes as graphs.
3. Formulate the causal question by identifying the target estimand.
4. Collect, check the quality of, and filter data.
5. Estimate the causal effect with software.
6. Challenge and refute the causal model.
7. Compare results with estimates from alternative methods.
8. Share model, data, and analysis.
9. Design, perform, and analyze new experiments.
10. Apply learnings from causal inference in the real world.

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

Tabular Prior-data Fitted Network (TabPFN) uses causal models to generate training data for machine-learning models

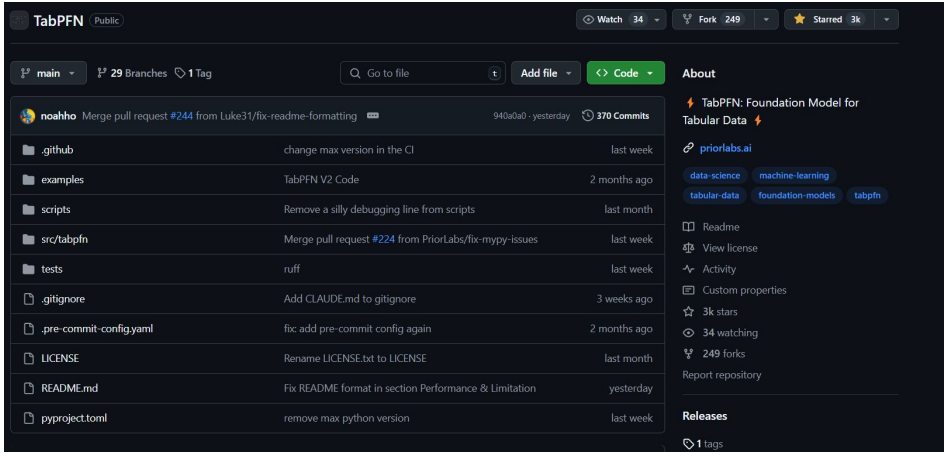
Article

Accurate predictions on small data with a tabular foundation model

https://doi.org/10.1038/s41586-024-08328-6
Received: 17 May 2024
Accepted: 31 October 2024
Published online: 8 January 2025
Open access
Check for updates

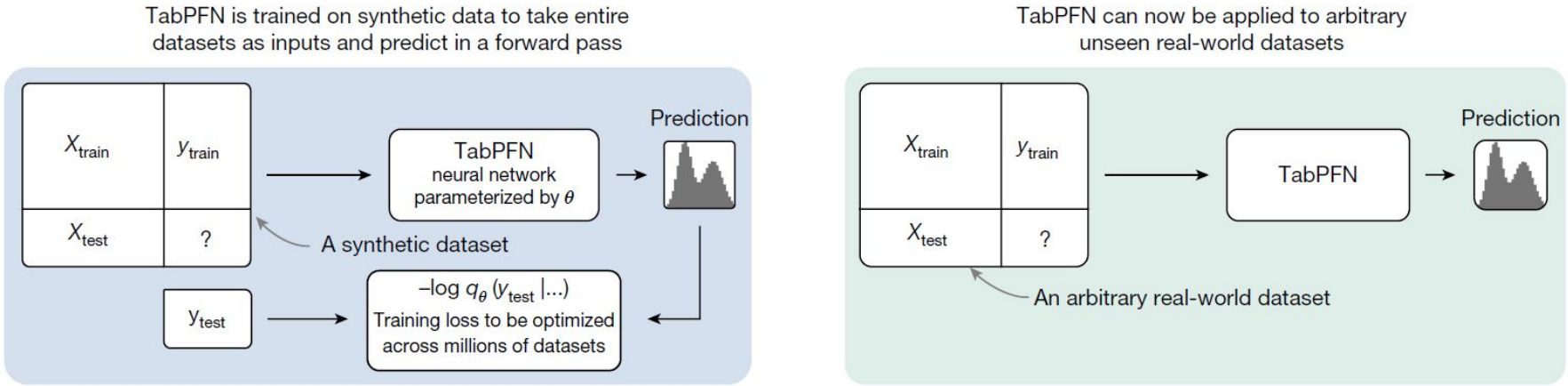
Noah Hollmann[1,2,3,7], Samuel Müller[1,7], Lennart Purucker[1], Arjun Krishnakumar[1], Max Körfer[1], Shi Bin Hoo[1], Robin Tibor Schirrmeister[4,5] & Frank Hutter[1,3,6]

Tabular data, spreadsheets organized in rows and columns, are ubiquitous across scientific fields, from biomedicine to particle physics to economics and climate science[1,2]. The fundamental prediction task of filling in missing values of a label column based on the rest of the columns is essential for various applications as diverse as biomedical risk models, drug discovery and materials science. Although deep learning has revolutionized learning from raw data and led to numerous high-profile success stories[3–5], gradient-boosted decision trees[6–9] have dominated tabular data for the past 20 years. Here we present the Tabular Prior-data Fitted Network (TabPFN), a tabular foundation model that outperforms all previous methods on datasets with up to 10,000 samples by a wide margin, using substantially less training time. In 2.8 s, TabPFN outperforms an ensemble of the strongest baselines tuned for 4 h in a classification setting. As a generative transformer-based foundation model, this model also allows fine-tuning, data generation, density estimation and learning reusable embeddings. TabPFN is a learning algorithm that is itself learned across millions of synthetic datasets, demonstrating the power of this approach for algorithm development. By improving modelling abilities across diverse fields, TabPFN has the potential to accelerate scientific discovery and enhance important decision-making in various domains.



Source code at https://github.com/PriorLabs/tabpfn, released in a license analogous to Apache 2.0 (commercial use friendly).

TabPFN is trained with synthesized data generated by DAGs, and predicts missing value in user's input data

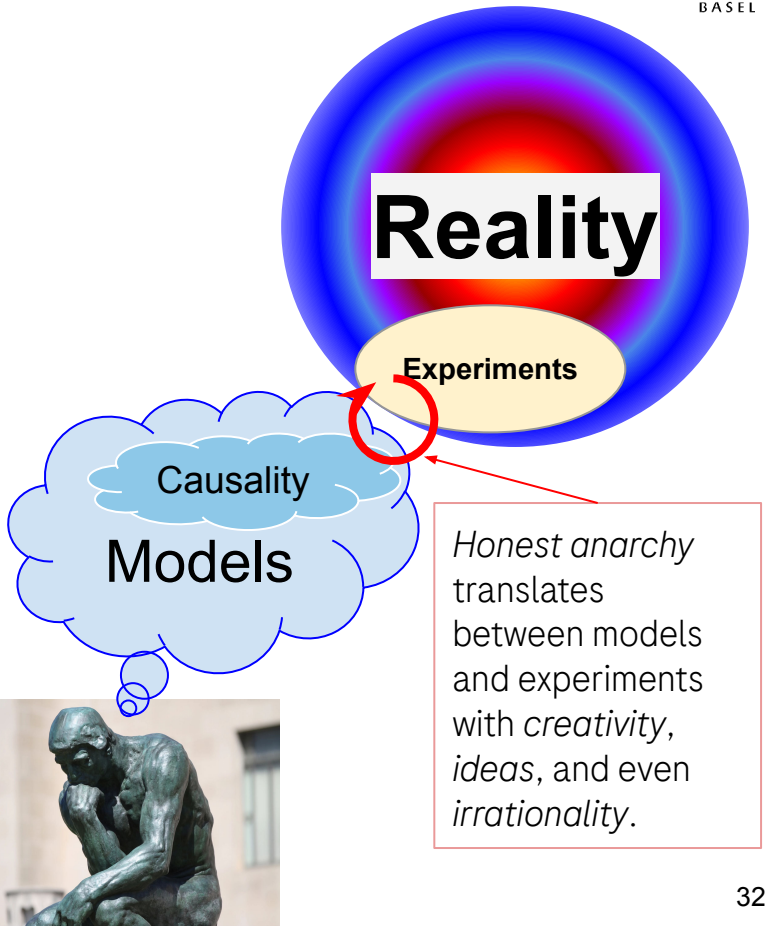


A metaphor: Imagine a lab where millions of billiard games are played simultaneously: in each game, different numbers of balls are placed randomly, and a white ball starts with a random velocity at a random position. By learning the trajectory of all balls of all games, one may learn to predict the trajectory of any real billiard game, as long as the positions of all balls and the initial velocity of the white ball is known.

Causal inference is not always easy

Credit: Ulrich Certa

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



Honest anarchy translates between models and experiments with *creativity*, *ideas*, and even *irrationality*.

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

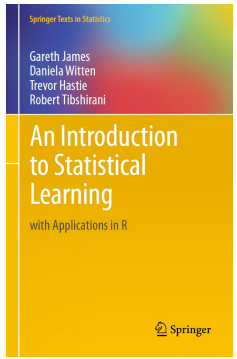
- Richard McElreath

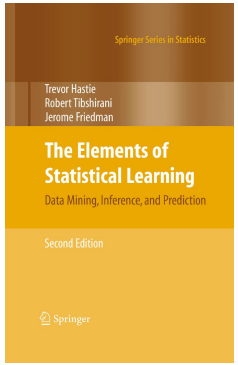
Conclusions

1. Statistical and machine learning models can model linear and nonlinear relationships between variables.
2. Applying statistical and ML models in drug discovery needs to consider the facts that we always work on something new, structure similarity does not warrant activity similarity, and correlation is not causation.
3. Causal inference combines prior knowledge and statistical/ML modelling to answer *what-if* questions.

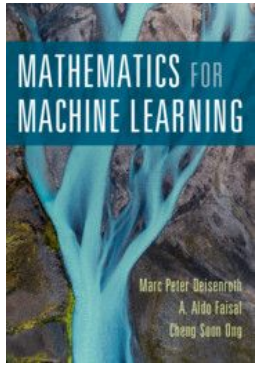
Resources for learning about machine learning

ESL and ISL: From a frequentist view (almost)

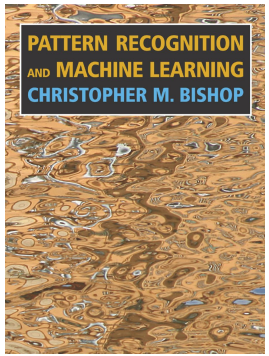


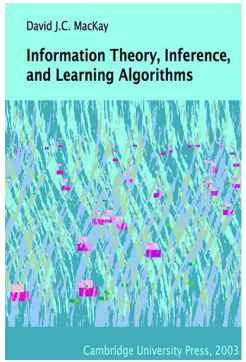


Mathematical foundations

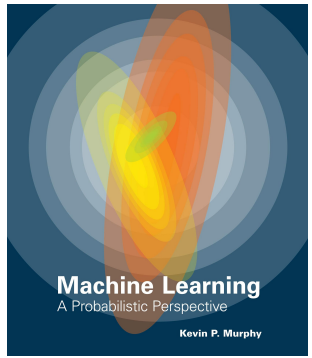


PRML and ITILA: From a Bayesian view



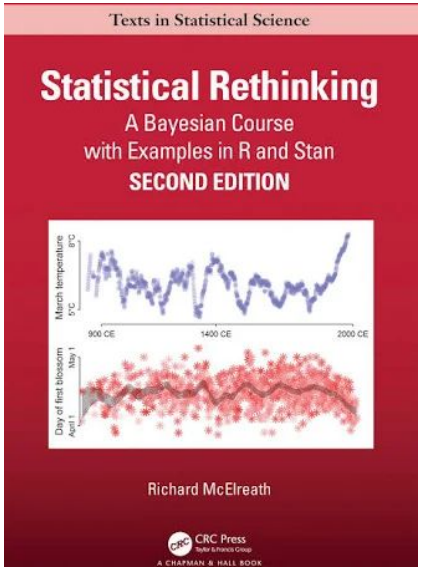


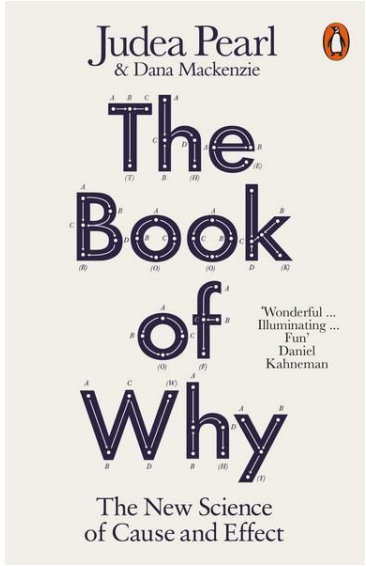
MLaPP: Application oriented, more accessible, and balanced views

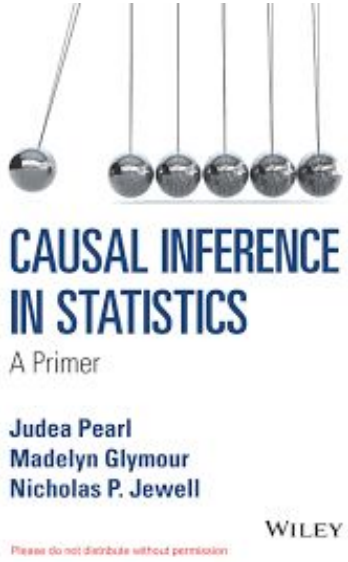


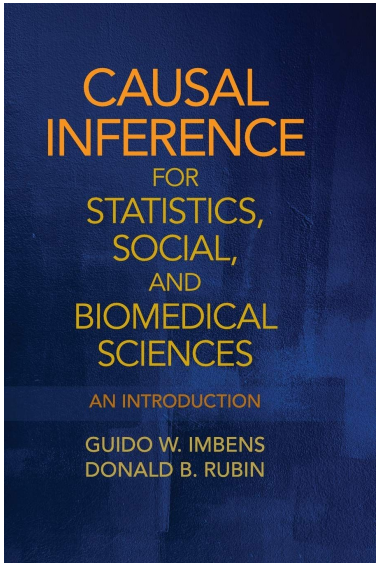
Resources for learning about causal inference

Causal inference in drug discovery and development, Michoel and Zhang, 2022









Lectures available on YouTube

Answers

Red stars are supported by Model 1.

Blue crosses are supported by both Model 2 and Model 3.

Reason: causality (C→E, from cause to effect) is directional. Manipulating C has an effect on E, while manipulating E has no effect on C. Blue crosses are around mean values of Y. If Y causes X, manipulating X has no effect on Y. Then the most likely values of Y will be around the mean of existing samples.