AMIDD Lecture 3: Statistical models and causal inference

Dihydrofolic acid

MTX

Dihydrofolate reductase (DHFR)
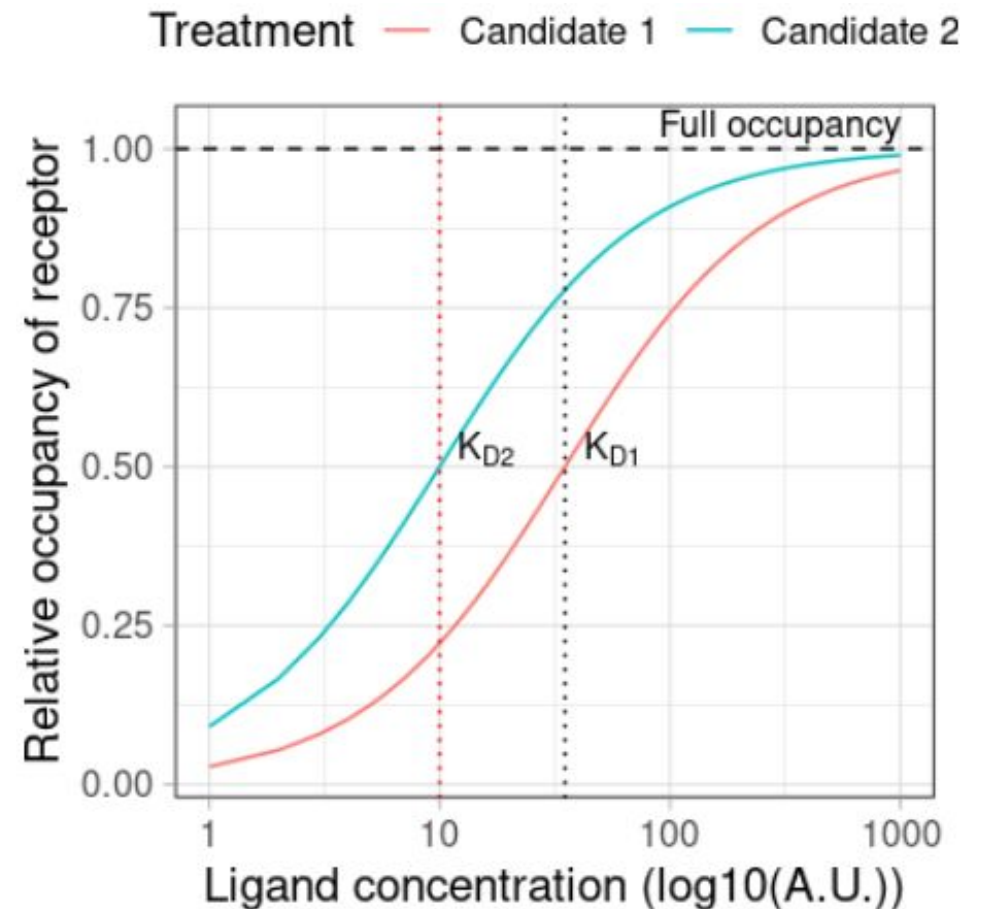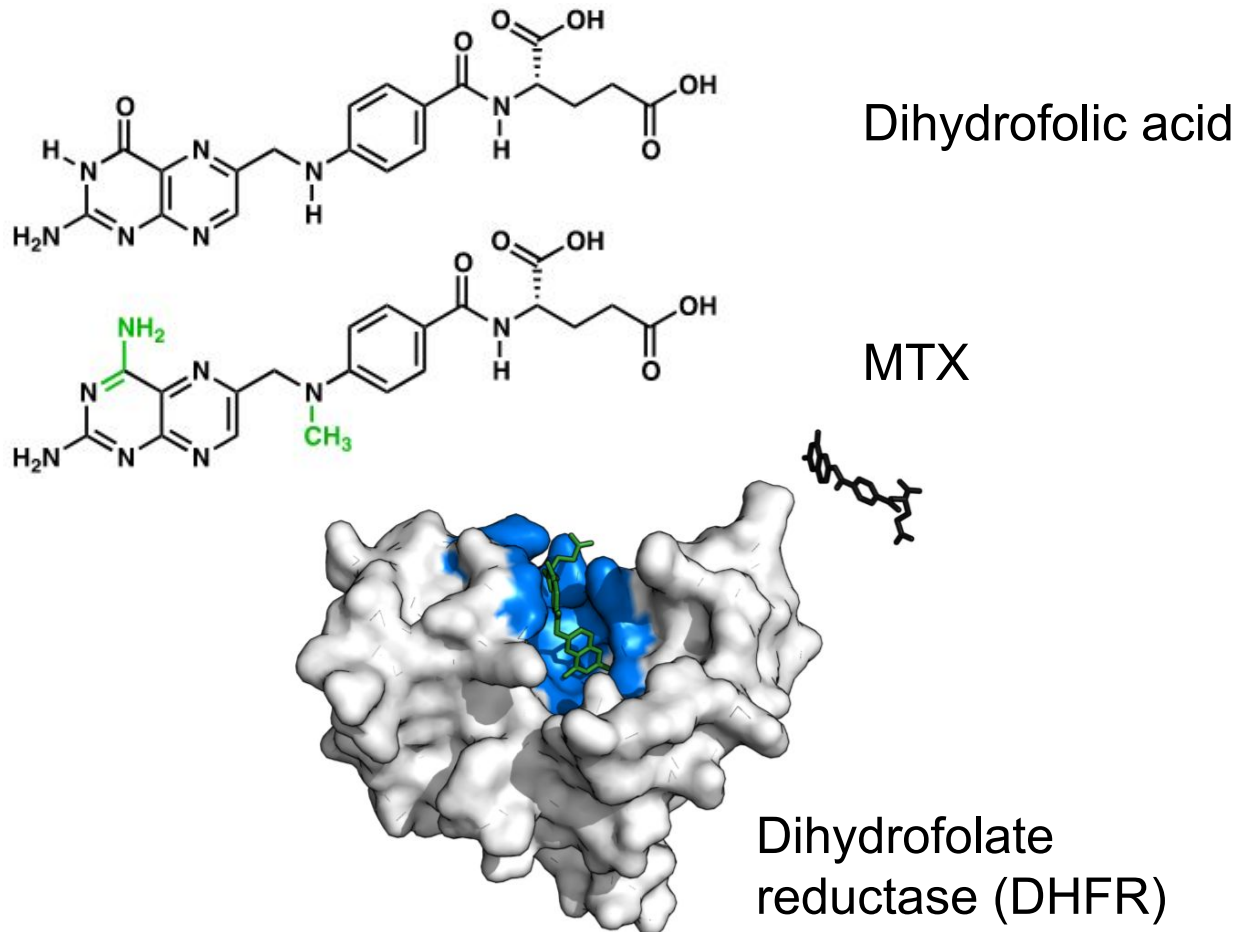
Dr. Jitao David Zhang, Computational Biologist

[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*

Quantitative Structure-Activity Relationships (QSARs) as an example of statistical modelling

QSAR is a statistical modelling of correlation between biological activity and physicochemical properties, or $\Delta\phi=f(\Delta S)$, where ϕ indicates a biological activity and S indicates a chemical structure (1868-1869).
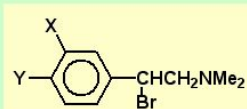
An example: **The Free-Wilson analysis.** The assumption: the biological activity for a set of analogues could be described by the contributions that substituents or structural elements make to the activity of a parent structure.

Molecular Descriptors (MD)

	Target property	MD_1	MD_2	...	MD_M
C_1	y_1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,M}$
C_2	y_2	$x_{2,1}$	...	...	...
C_3	y_3	...	...	...	...
C_4	y_4	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
C_N	y_N	$x_{N,1}$	$x_{N,2}$	...	$x_{N,M}$

Compounds (C)

The basic form of a QSAR model: find a function f that predicts y from x, $y{\sim}f(x)$

Multivariate regression analysis

$$\log(1/ED_{50}) = -0.301[m{-}F] + 0.27[m{-}Cl] + 0.434[m{-}Br] + 0.579[m{-}I]$$
$$+ 0.454[m{-}Me] + 0.340[p{-}F] + 0.768[p{-}Cl] + 1.020[p{-}Br]$$
$$+ 1.429[p{-}I] + 1.256[p{-}Me] + 7.821$$
$$n = 22, r^2 = 0.94, s = 0.194, F = 17.0$$

Unsupervised versus supervised models
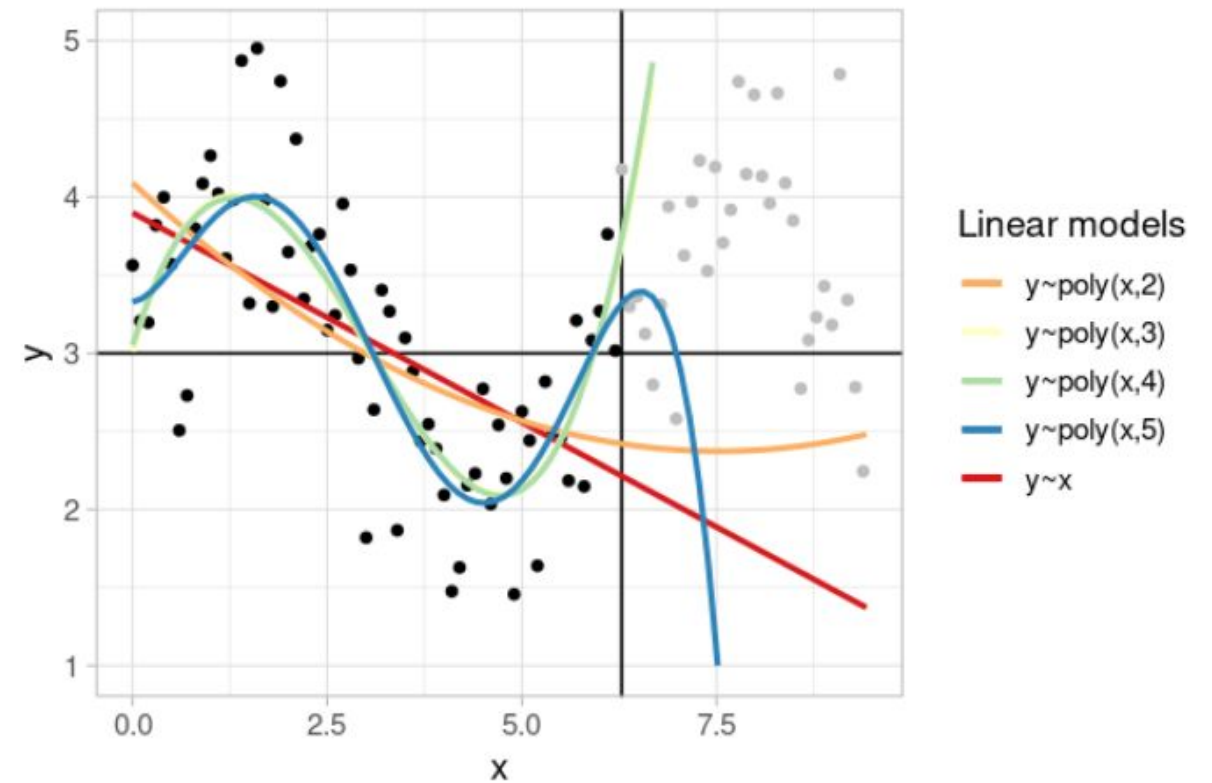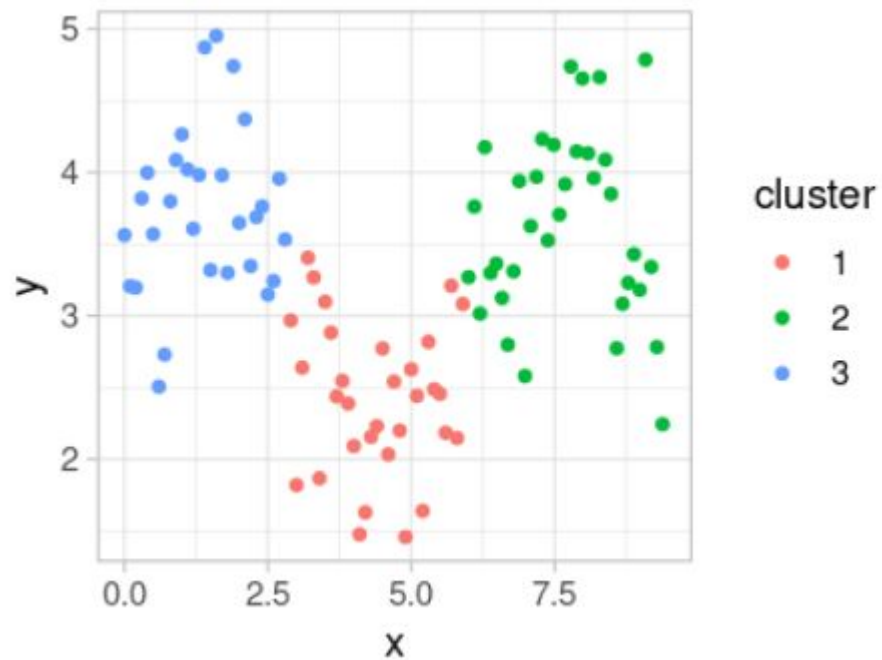
Linear model can be used to model local effects of non-linear models
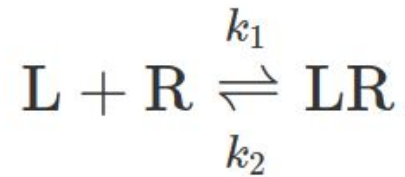
$$\mathbf{L} + \mathbf{R} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \mathbf{LR}$$

$$K_D \equiv k_2/k_1$$

$$[LR] = [R_{total}]\frac{[L]}{[L] + K_D}$$
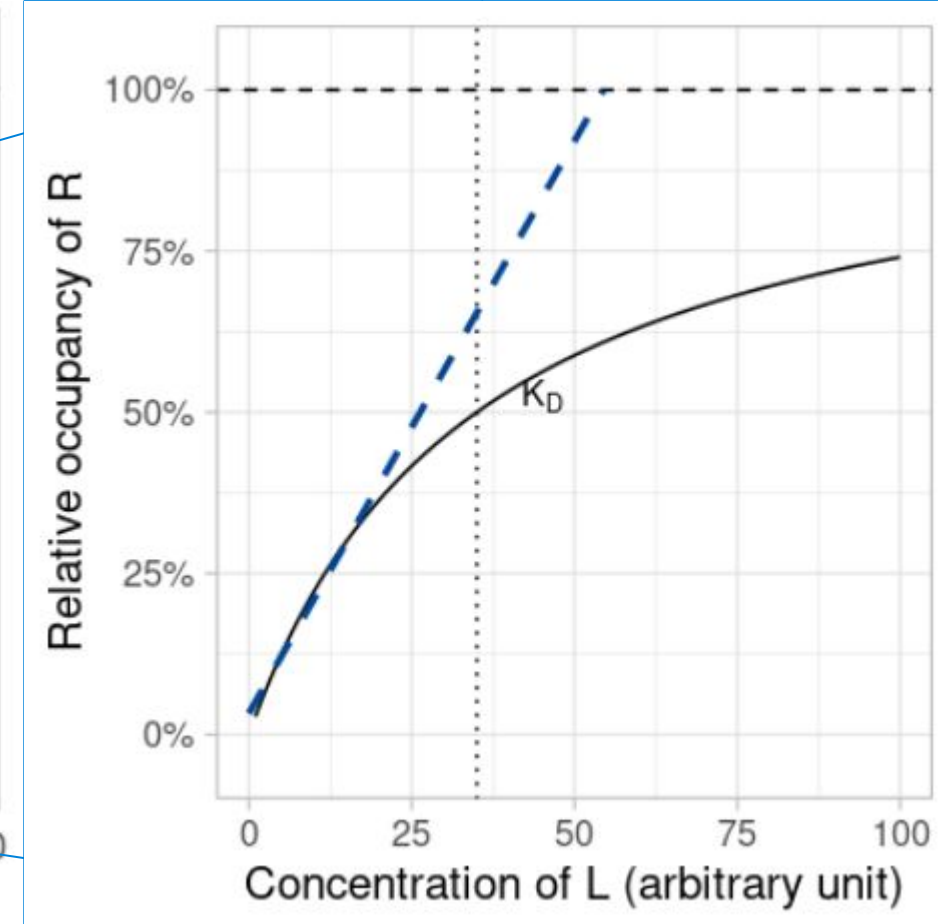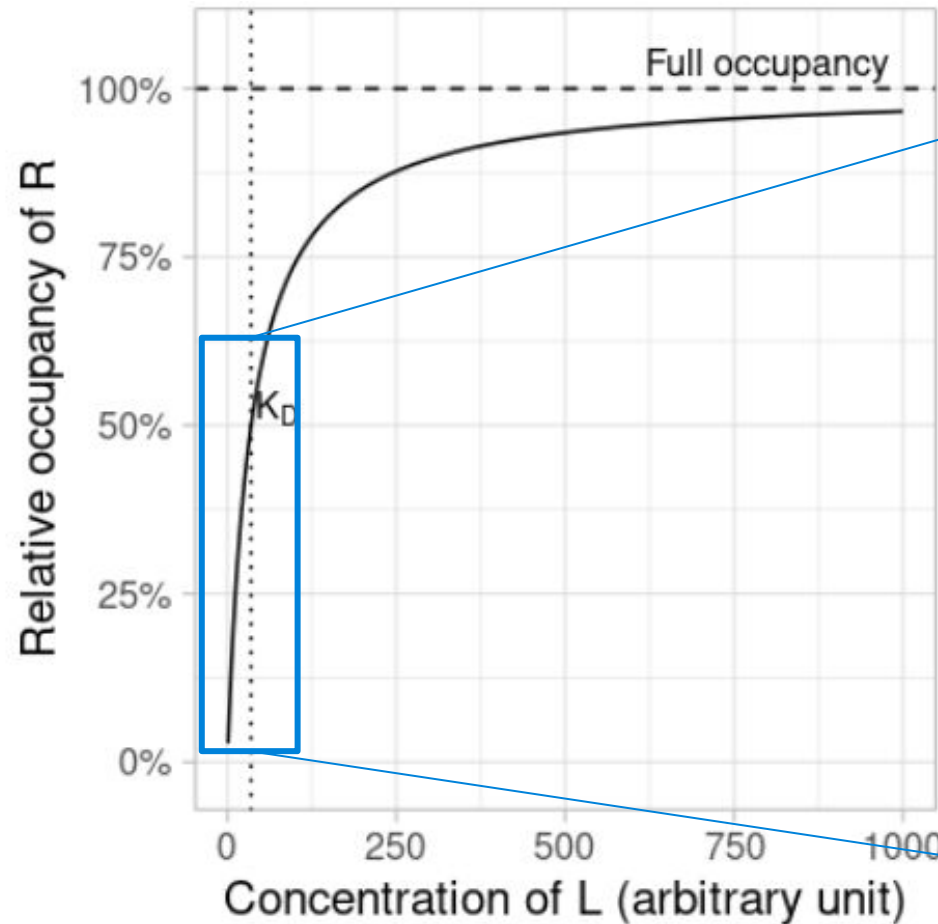
The simplest linear model has three components: the intercept, the slope, and a measure of fit

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

$$\hat{y} = 0.721 + 1.83\ x,\ R^2 = 0.75$$
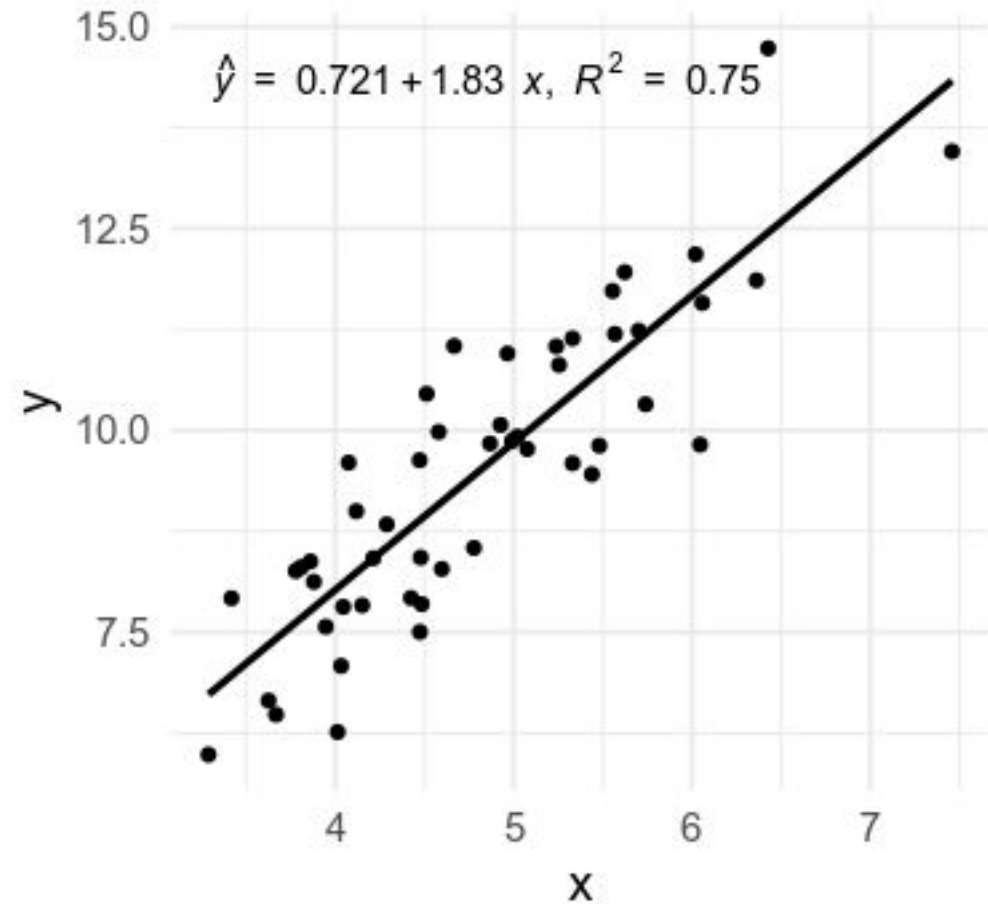
A visual explanation of R^2

The better the linear regression (right) fits the data in comparison to the average (left), the closer the value of R^2 is to 1.

The areas of the blue squares represent the squared residuals with respect to the linear regression. The areas of the red squares represent the squared residuals with respect to the average.

R^2 is defined as 1-(blue area)/(red area).
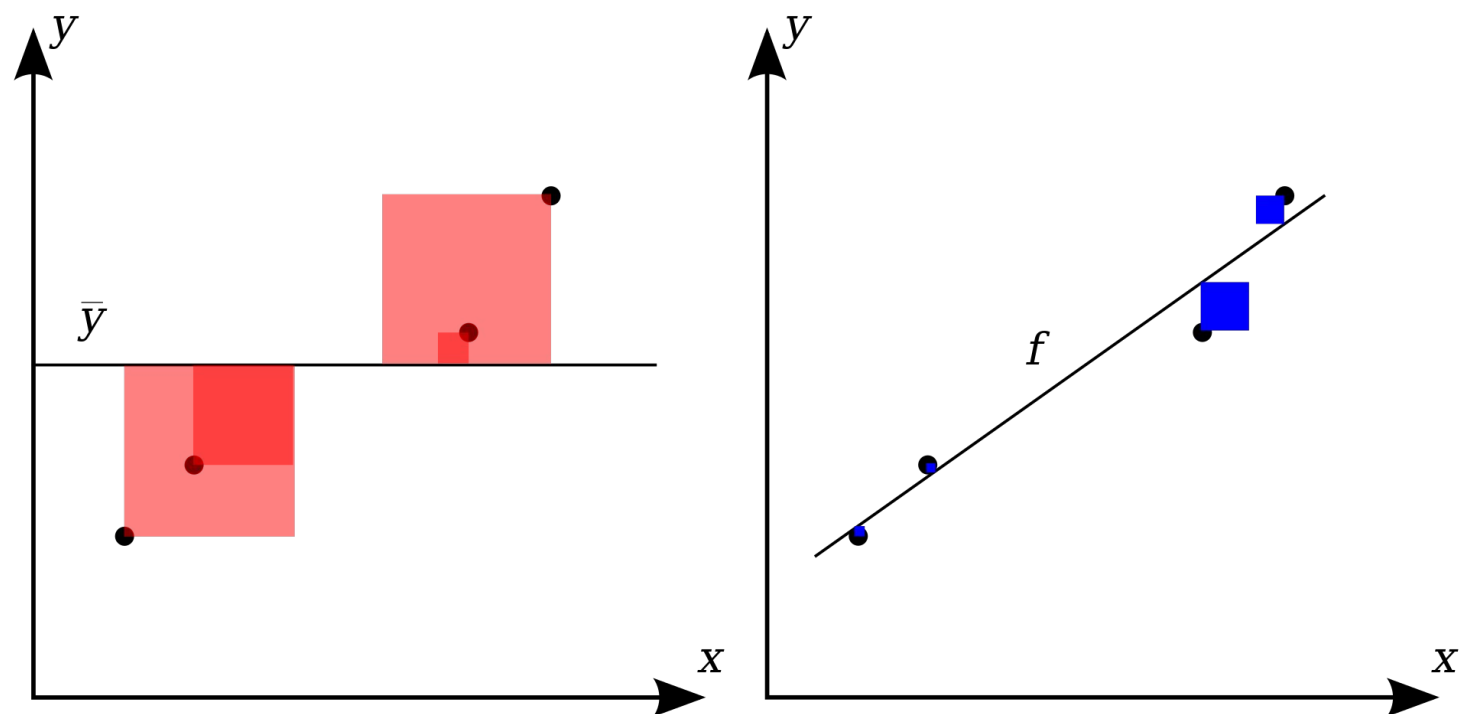
Question: can R^2 be negative?

Work by Orzetto, CC-SA 3.0, from Wikimedia

6

Logistic regression is a example of *generalized* linear model, which allows dependent variable defined other than real numbers

- Dependent variable of a linear regression model is defined on ℝ.

- *Generalized* linear models allow the dependent variable to be defined on other domains than real numbers, for instance binary (0/1), counts (non-negative integers), etc.

- Logistic regression maps input real numbers to the range between 0 and 1 in two steps: (1) building a simple linear regression, (2) applying the *logistic function* to map the intermediate dependent variable to the desired domain (0,1).
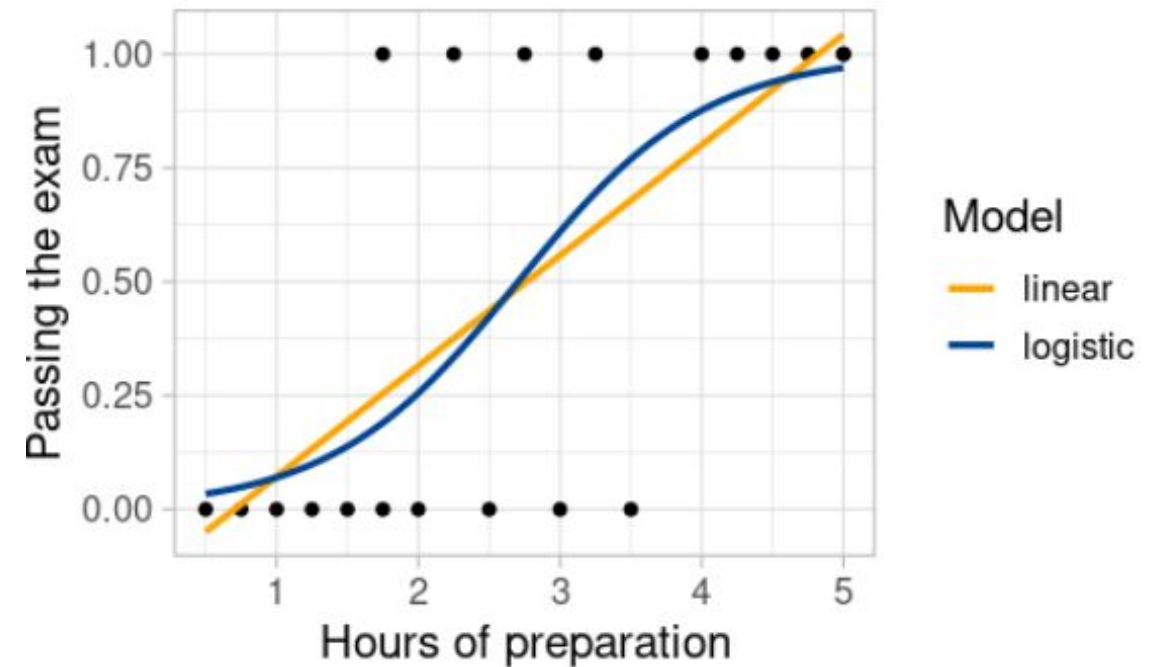
$$t = \alpha + \beta x \qquad y = \frac{1}{1 + e^{-t}}$$

Data come from Wikipedia's item on logistic regression

7

Multiple regression with regularization

- We may have multiple independent variables. For instance, in the example on the right side, we want to predict *which topics contribute to passing the exam.* In such cases, we apply *multiple regression*.

- In multiple regressions, we often wish for a sparse solution: i.e. we wish to know the few most important features that contribute to the prediction. A technique to achieve this is *regularization.*

- Regularization penalizes large coefficients. It effectively push coefficients towards zero. For instance, the equation below shows the *error function* of *ridge regression*.

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

f(Y): a binary label to indicate whether someone pass an exam

g_1-g_9: students

$\square_1$-$\square_8$: topics

$\beta_{\square 1}$-$\beta_{\square 8}$: coefficients of topics

Equation: Bishop, Christopher M. [Pattern Recognition and Machine Learning](#), page 10

8

We next address two problems: (1) unknown performance of model when new data are met, and (2) non-linearity
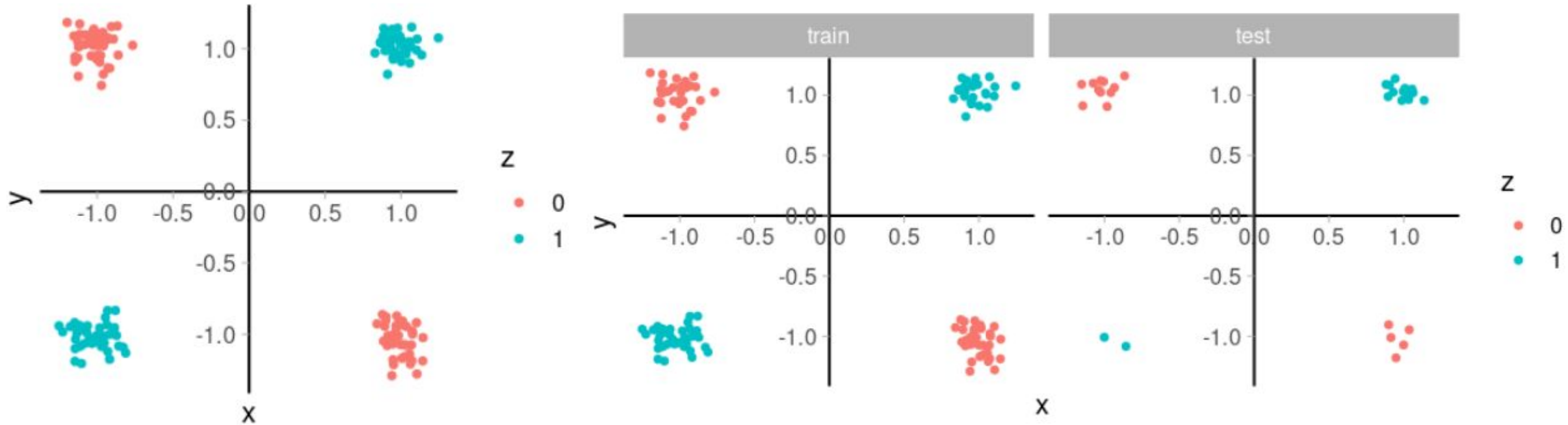
- Left: a simple example of non-linearity: linear models cannot predict *z* well based on values of *x* and *y*. We need something else.

- Right: a model is usually trained in some data, and the performance is assessed in unseen test data.

We can use random forest to model non-linearity
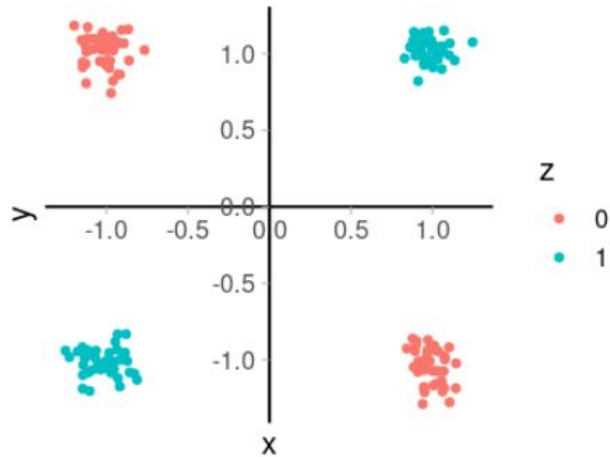
Tree #1
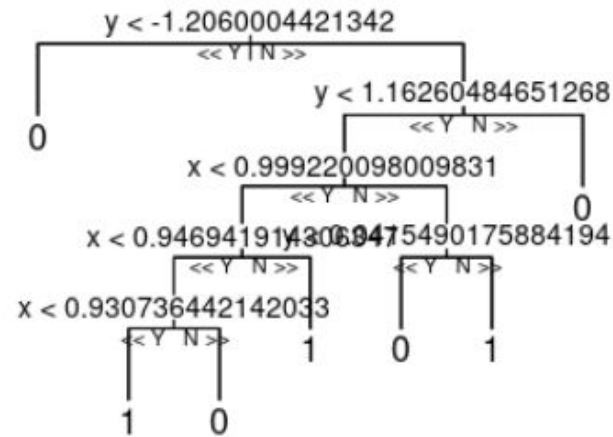
Tree #100
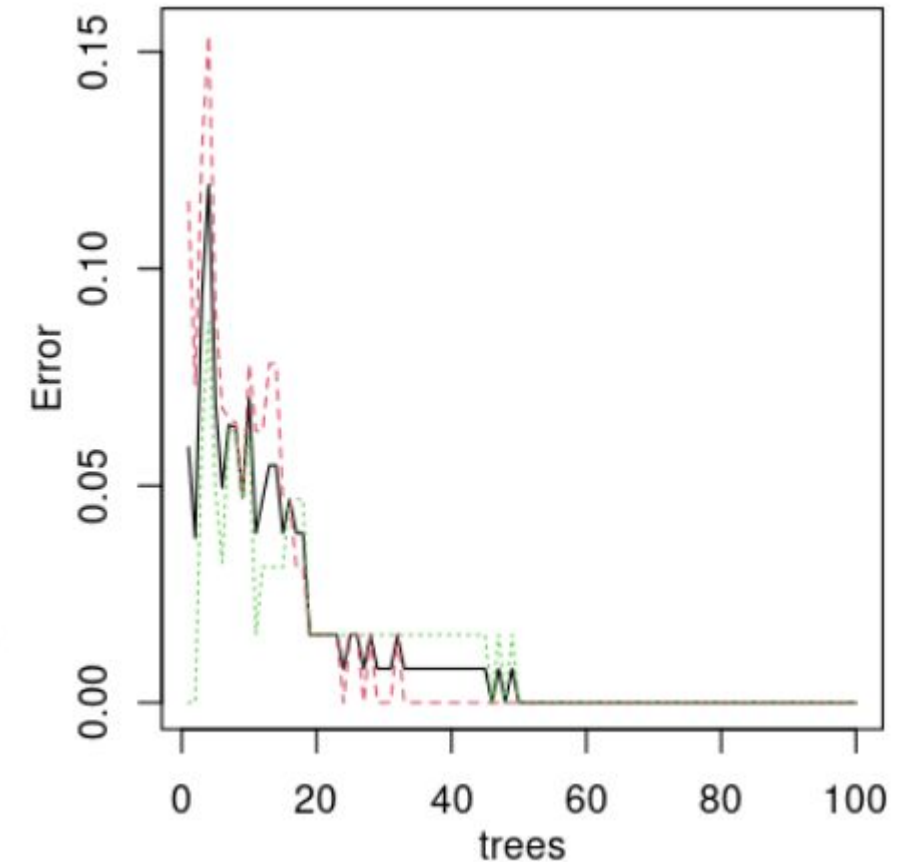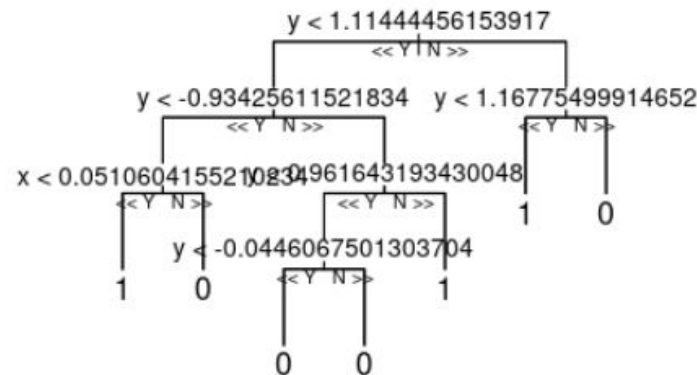
- Random forest is a collection of *decision trees*. Each tree partitions the input data to make predictions.

- Random forest is an example of *ensemble methods*: each tree has weak performance, however the consensus can perform surprisingly well.

Neural network can be used to model non-linearity, too

- Neural network models non-linearity by applying multiple linear combinations subsequently in the forward propagation.

- Once the architecture (# hidden layers, # nodes, etc.) is fixed, weights of edges neural network are initialized with random numbers, and then optimized by iterative forward and reverse propagation to minimize the error.

- Right figure: the trained neural network with the example data. Blue nodes indicate intercepts.

Hidden layer
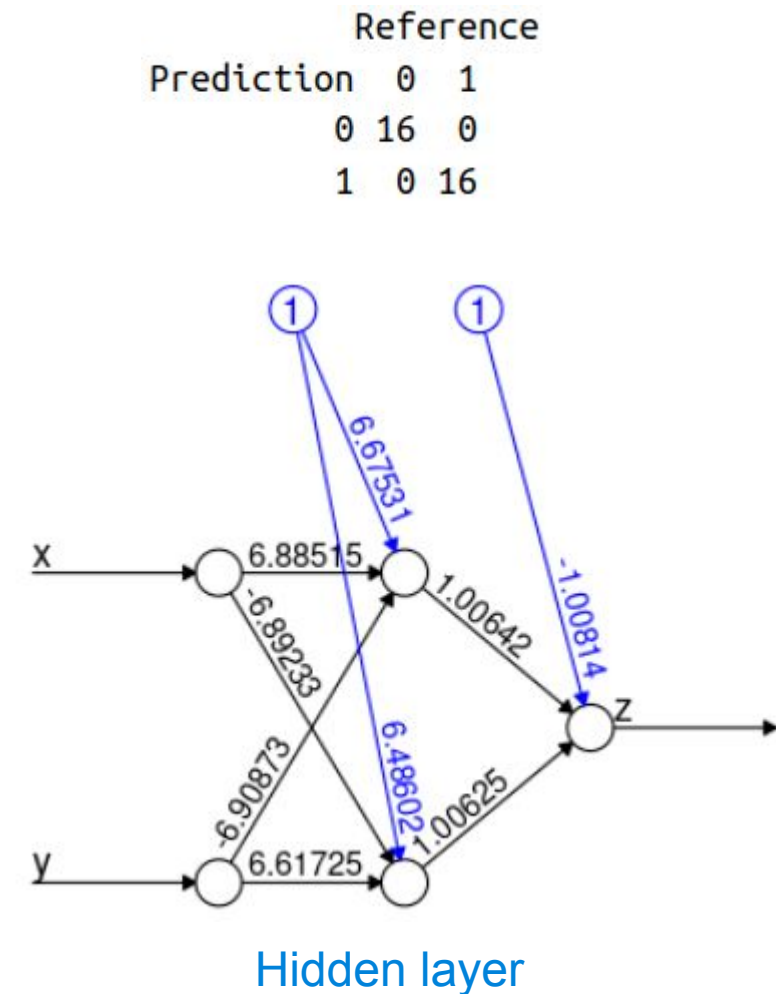
Error: 0.000172   Steps: 1085

Generally, well-performing models tend to be less interpretable
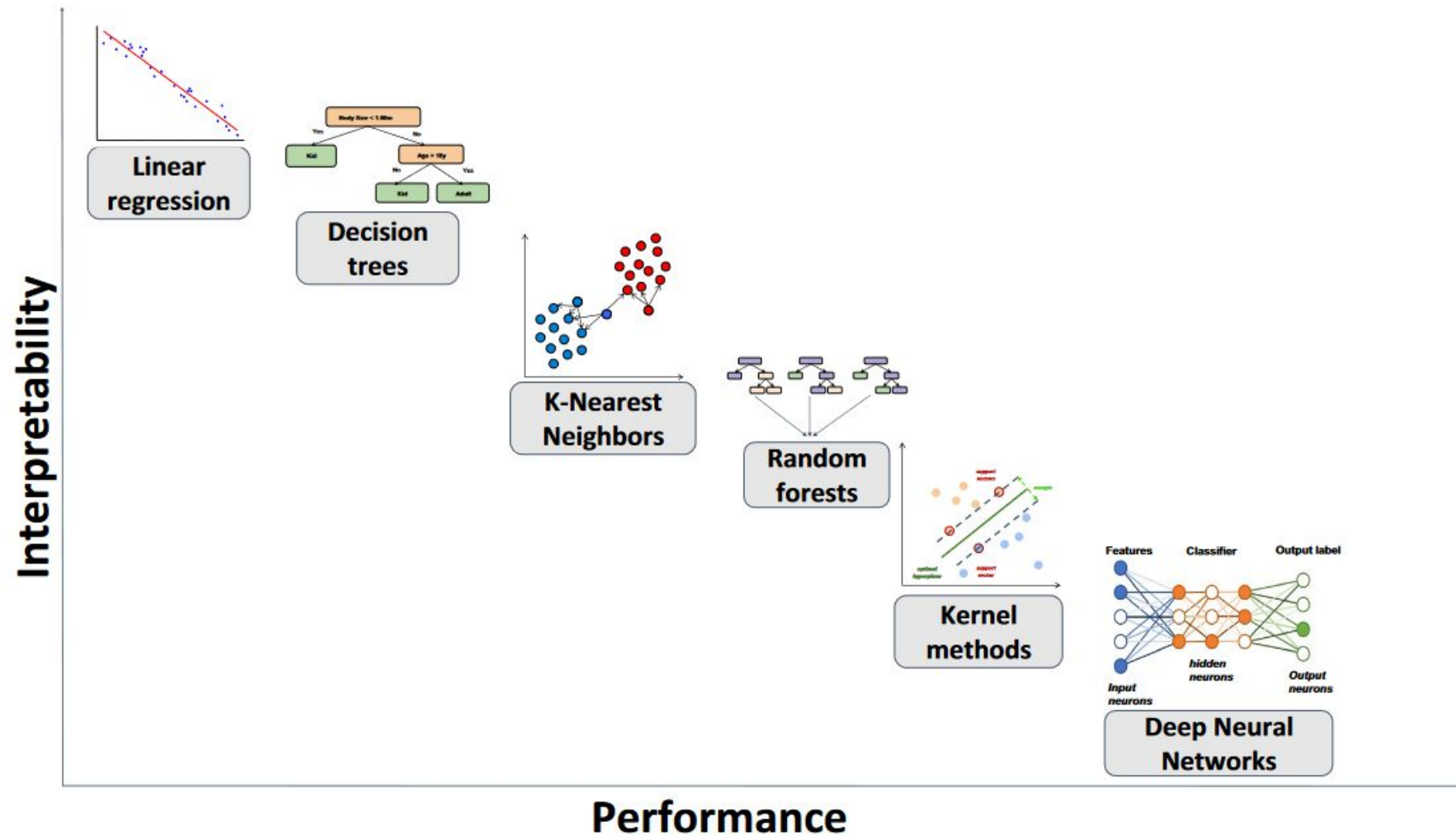
Tree-based methods are well interpretable and yet generally outperform deep learning on tabular data

The authors collected 45 tabular datasets from varied domains.

They found that tree-based models remain state-of-the-art on medium-sized data (~10k samples), even without considering the speed.

Conclusion: when working with tabular data, consider tree-based methods.
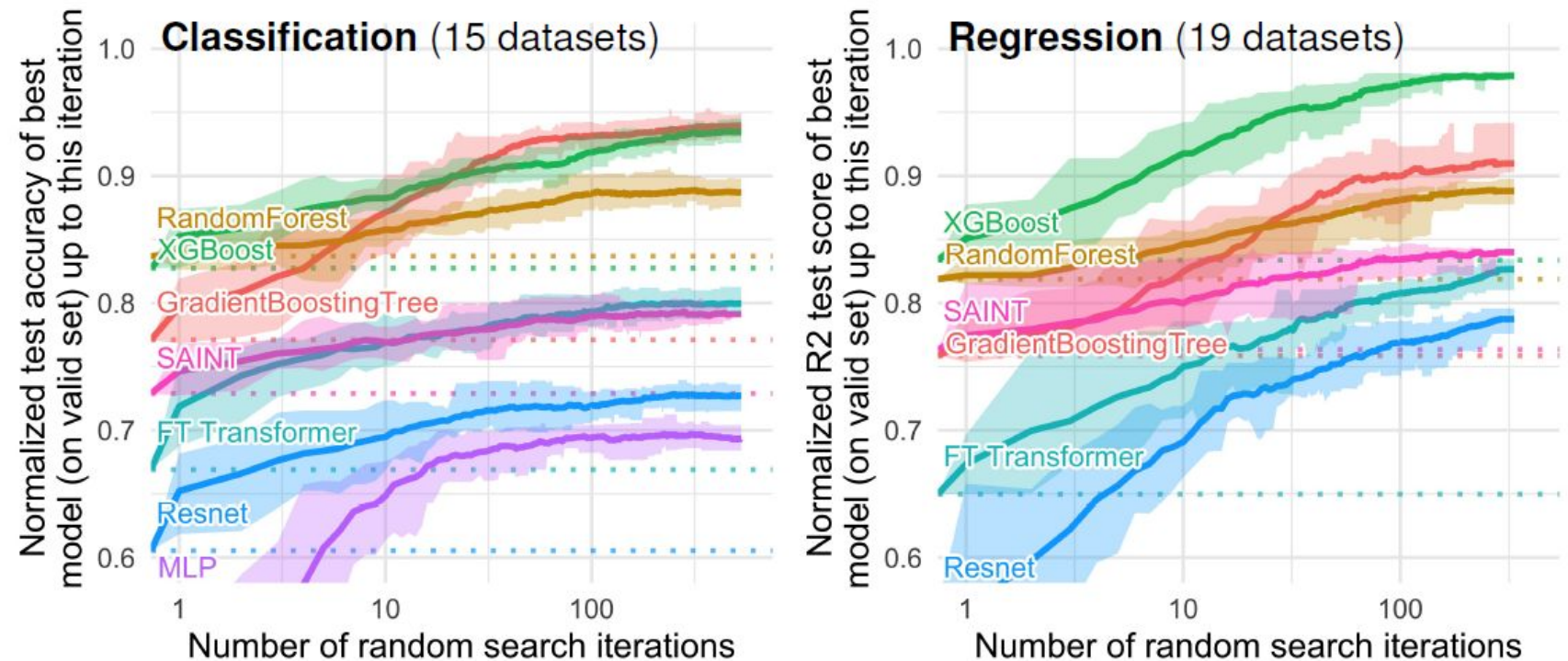
Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. "Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?" arXiv, July 18, 2022. https://doi.org/10.48550/arXiv.2207.08815. GitHub repository: https://github.com/LeoGrin/tabular-benchmark

Watchout 1: Temporal validation is essential for drug discovery
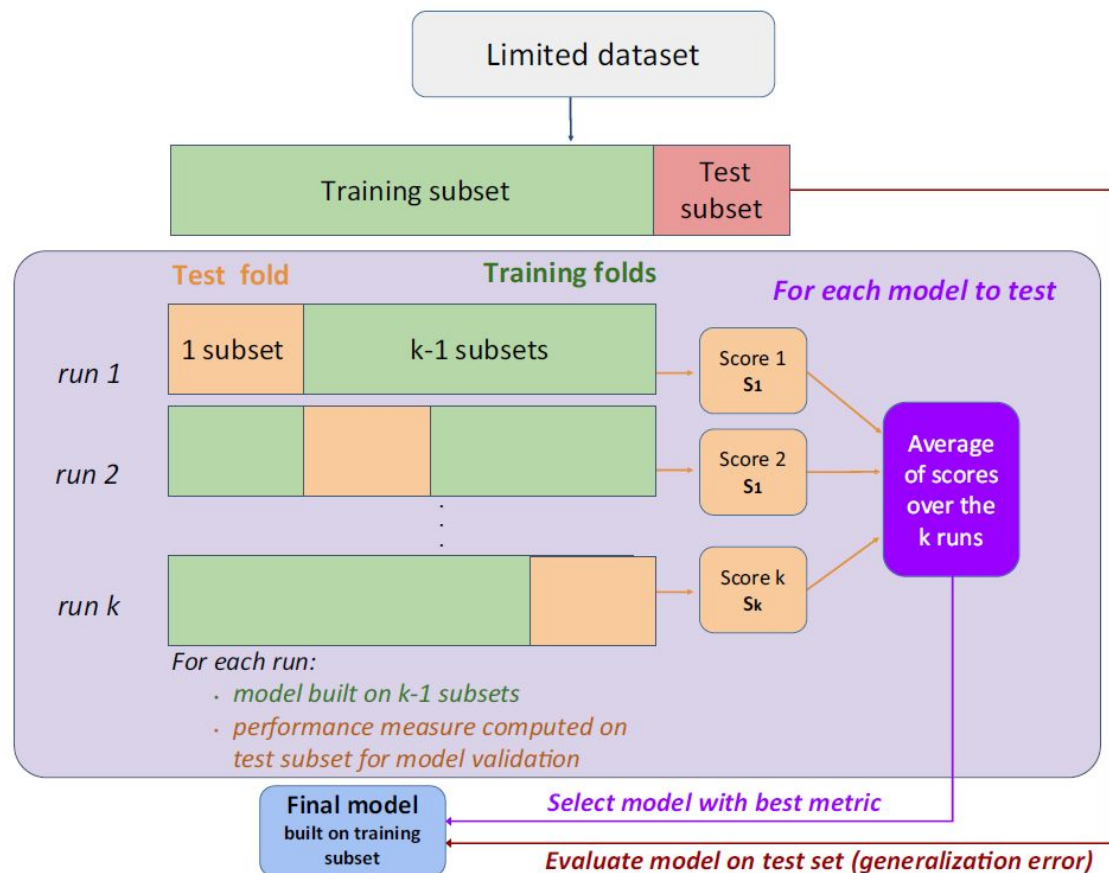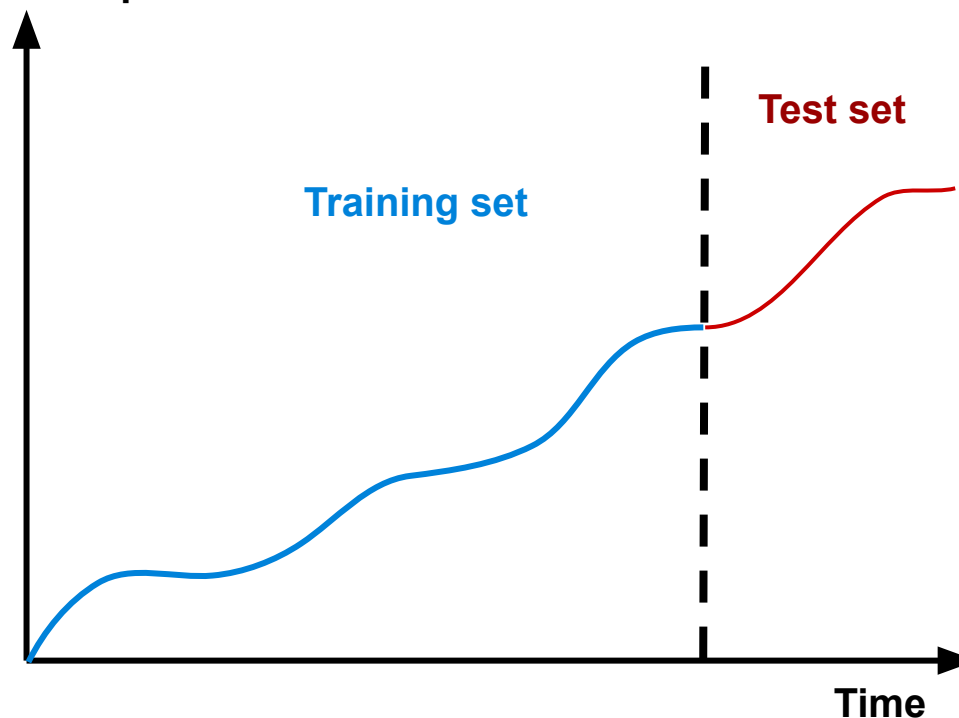
(Left) To assess the generalization ability of a supervised learning algorithm, data are separated into a training subset used for building the model and a test subset used to assess the generalization error.

(Right) Temporal validation is especially important for drug discovery, because chemical structures used in the training set may differ substantially from those that will be tested.

Watchout 2: Molecular similarity does not equal biological similarity

	1	2	3	4	5
A					
B					
C					
D					
E					

A1: 2D fingerprints A2: 3D fingerprints A3: Scaffolds A4: Structural keys A5: Physicochemistry
B1: Mechanisms of action B2: Metabolic genes B3: Crystals B4: Binding B5: HTS bioassays
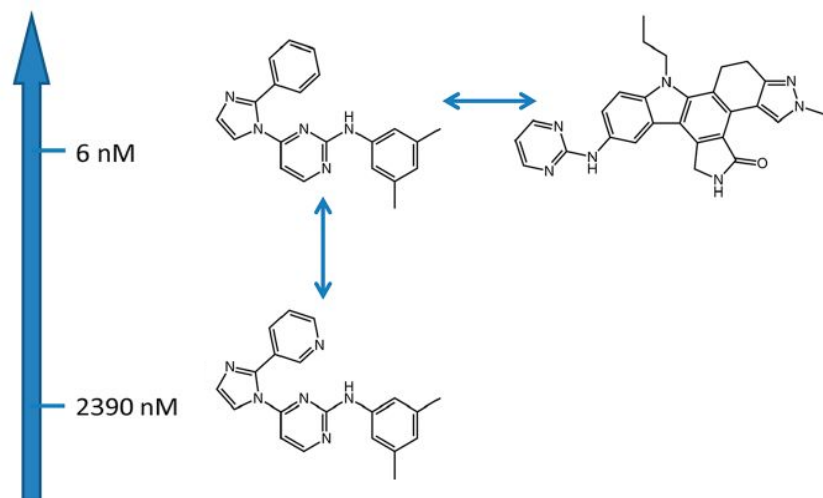C1: Small molecule roles C2: Small molecule pathways C3: Signaling pathways C4: Biological processes C5: Interactome
D1: Transcription D2: Cancer cell lines D3: Chemical genetics D4: Morphology D5: Cell bioassays
E1: Therapeutic areas E2: Indications E3: Side effects E4: Diseases & toxicology E5: Drug–drug interactions

A: Chemistry
B: Targets
C: Biological network
D: Cells
E: Clinical readout

Watch out biological activity cliffs:
Structural similarity does not imply similar activity. Top: three vascular endothelial growth factor receptor 2 (VEGFR2) ligands that represent different similarity−activity relationships.
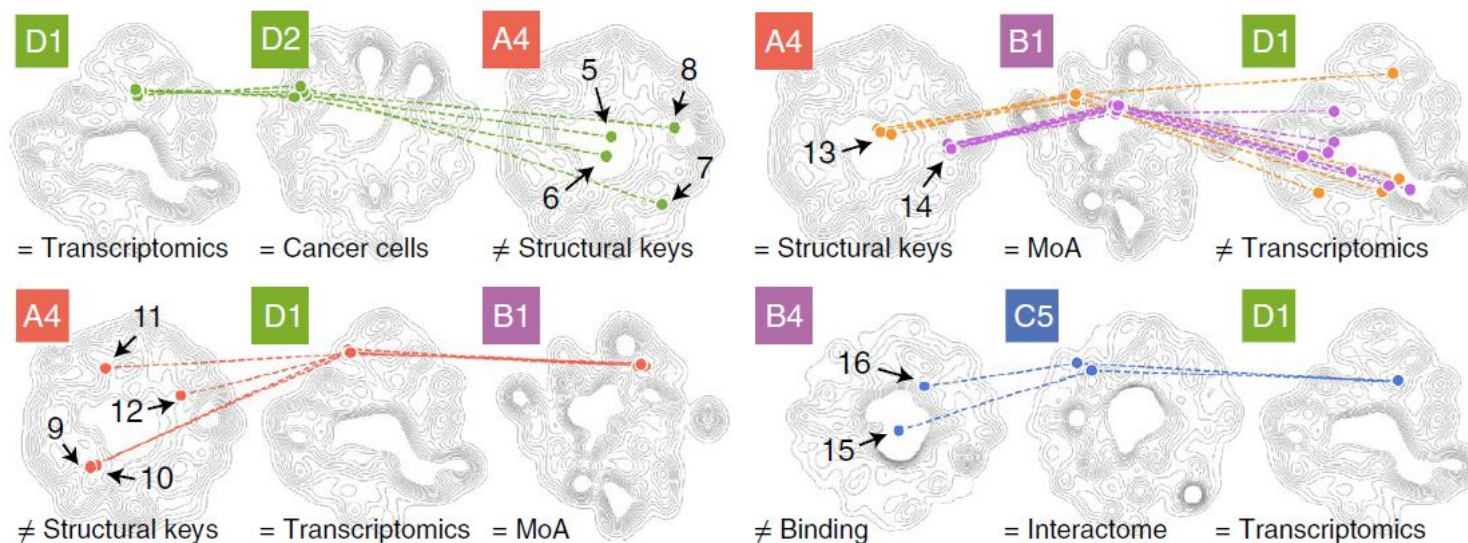
Duran-Frigola, Miquel, Eduardo Pauls, Oriol Guitart-Pla, Martino Bertoni, Víctor Alcalde, David Amat, Teresa Juan-Blanco, and Patrick Aloy. 2020. "Extending the Small-Molecule Similarity Principle to All Levels of Biology with the Chemical Checker." Nature Biotechnology, May, 1–10.

Watchout 3: Do we need correlation or causation?

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

$\alpha = 0.70704013\cdots$ $\alpha = 0.28495951\cdots$
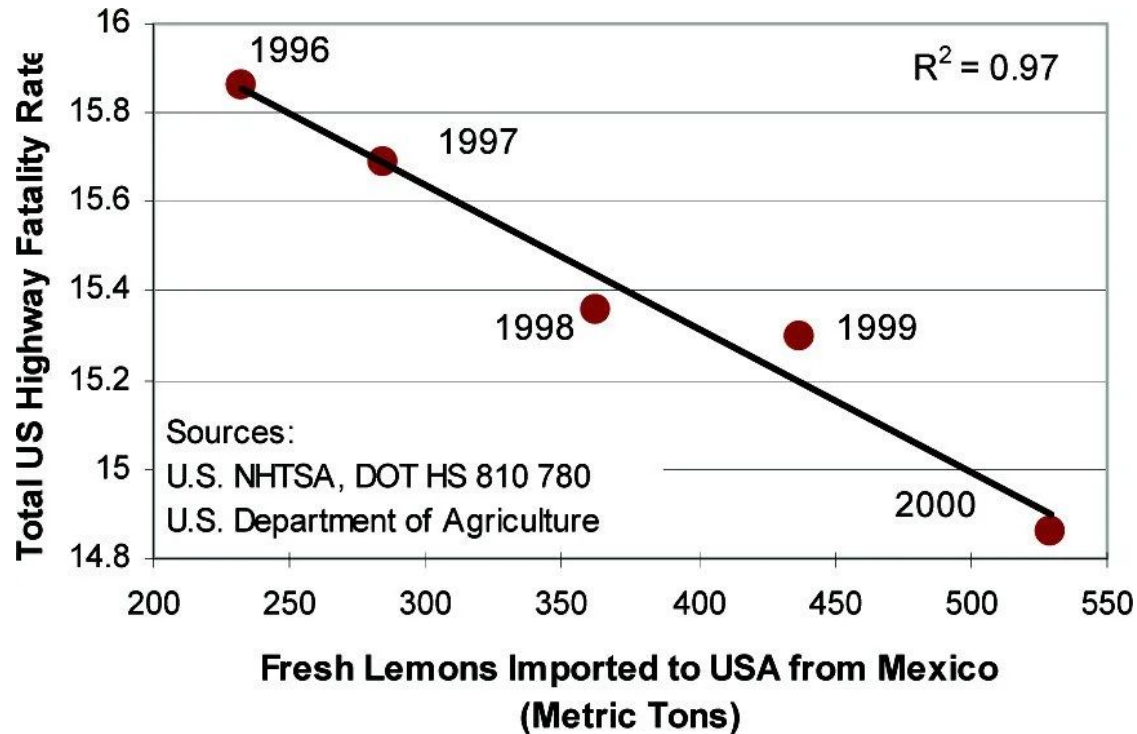
Johnson, Stephen R. "The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy)." Journal of Chemical Information and Modeling 48, no. 1 (January 1, 2008): 25–26. https://doi.org/10.1021/ci700332k
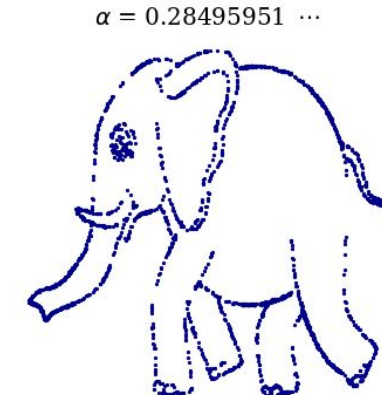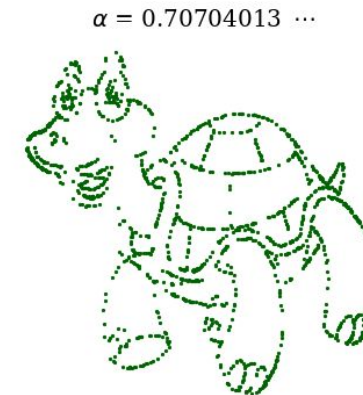
Boué, Laurent. "Real Numbers, Data Science and Chaos: How to Fit Any Dataset with a Single Parameter." ArXiv:1904.12320 [Cs, Stat], April 28, 2019. http://arxiv.org/abs/1904.12320.GitHub Repo. Also see: Drawing an elephant with four complex parameters

Generative models shed light on correlation and causality

UNI BASEL

True effect: 2.0

$\hat{y} = 0.721 + 1.83\ x,\ R^2 = 0.75$

X → y →

	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

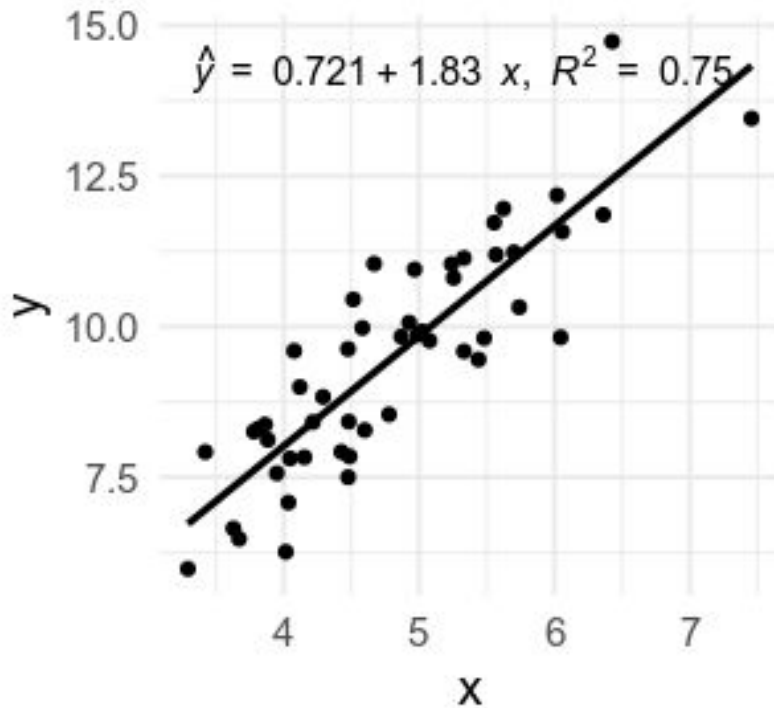Assumptions of the **generative model**:
1. X is a random variable;
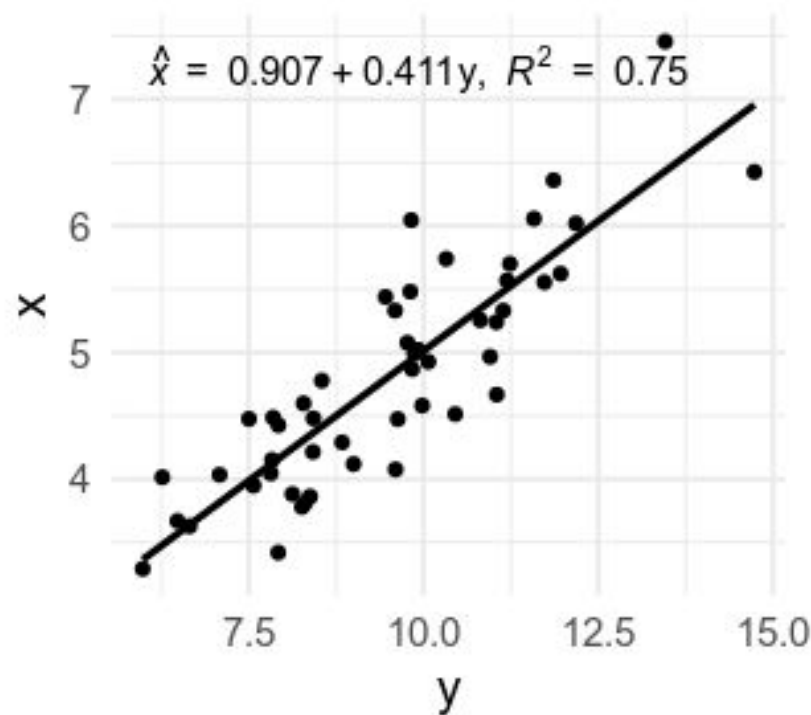2. Every unit change of X induces a change of 2 units in Y.

Correlation is caused by causation or confounding

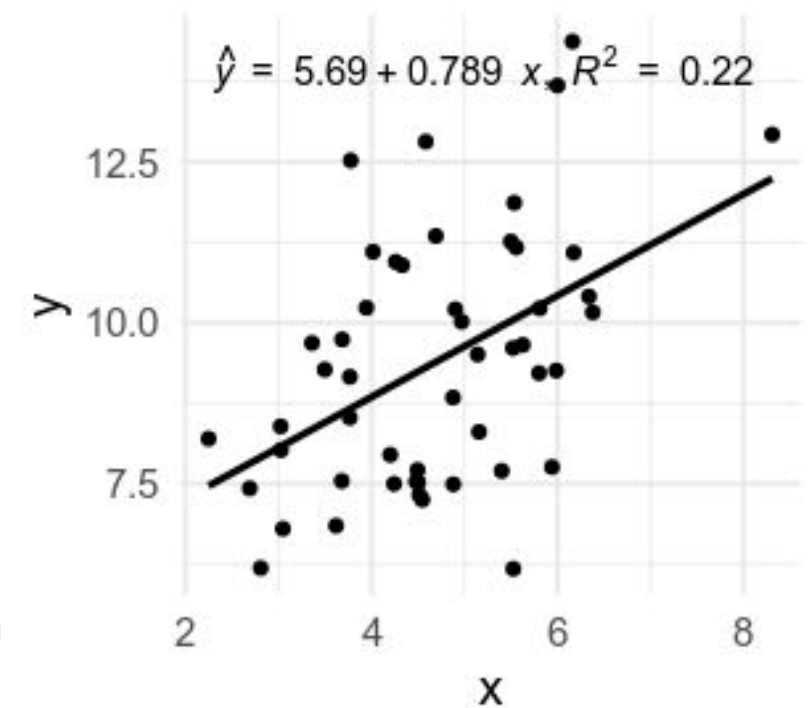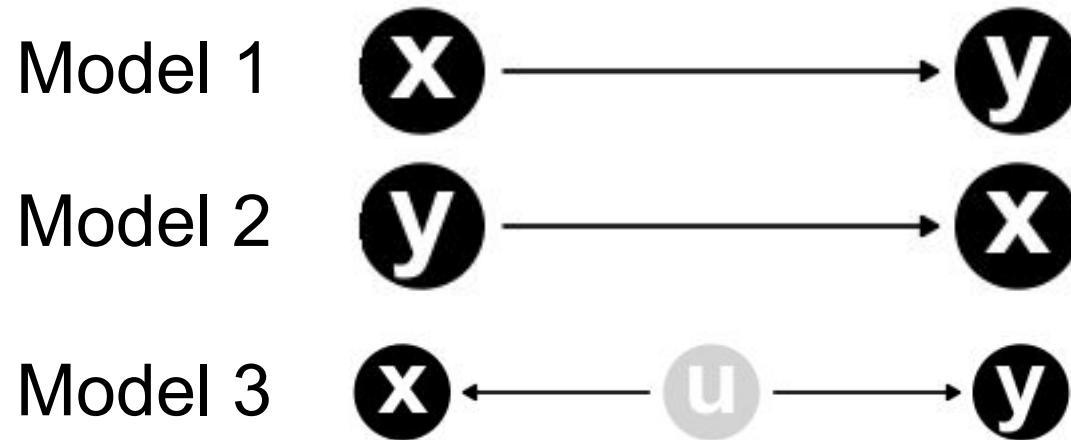**Statistical models alone cannot derive causality from correlation**

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

Model 1

Model 2

Model 3

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?

Generative models shed light on correlation and causality

True effect: 2.0

	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

$\hat{y} = 0.721 + 1.83\ x,\ R^2 = 0.75$

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.
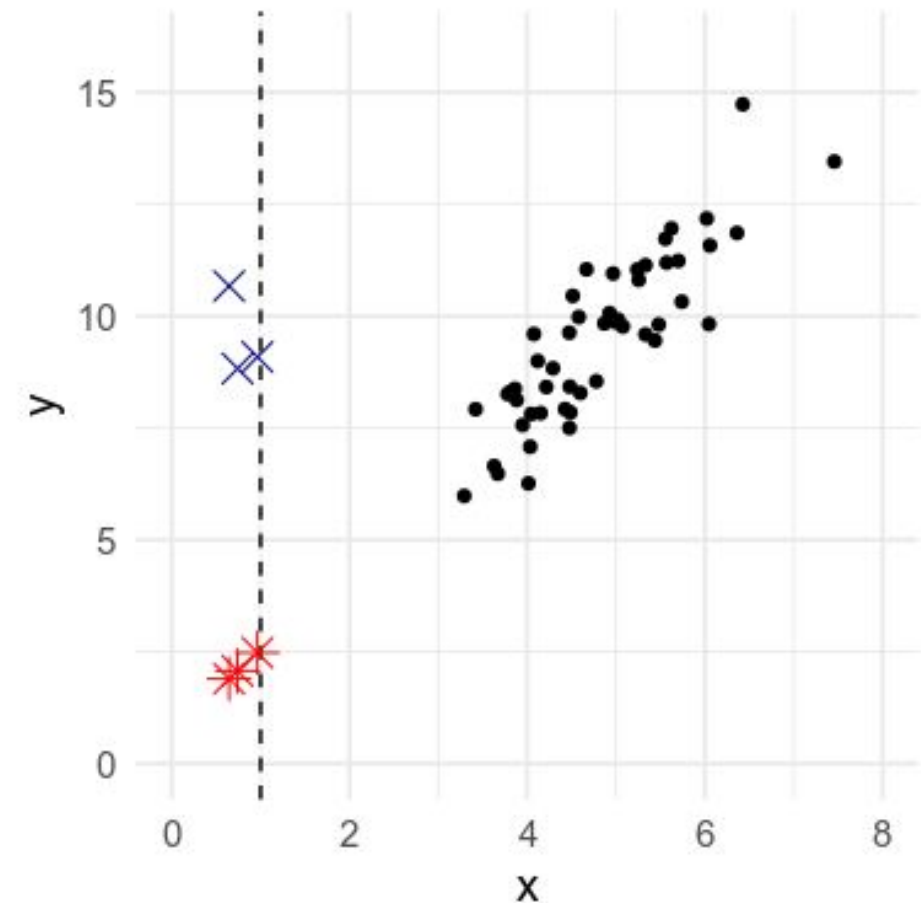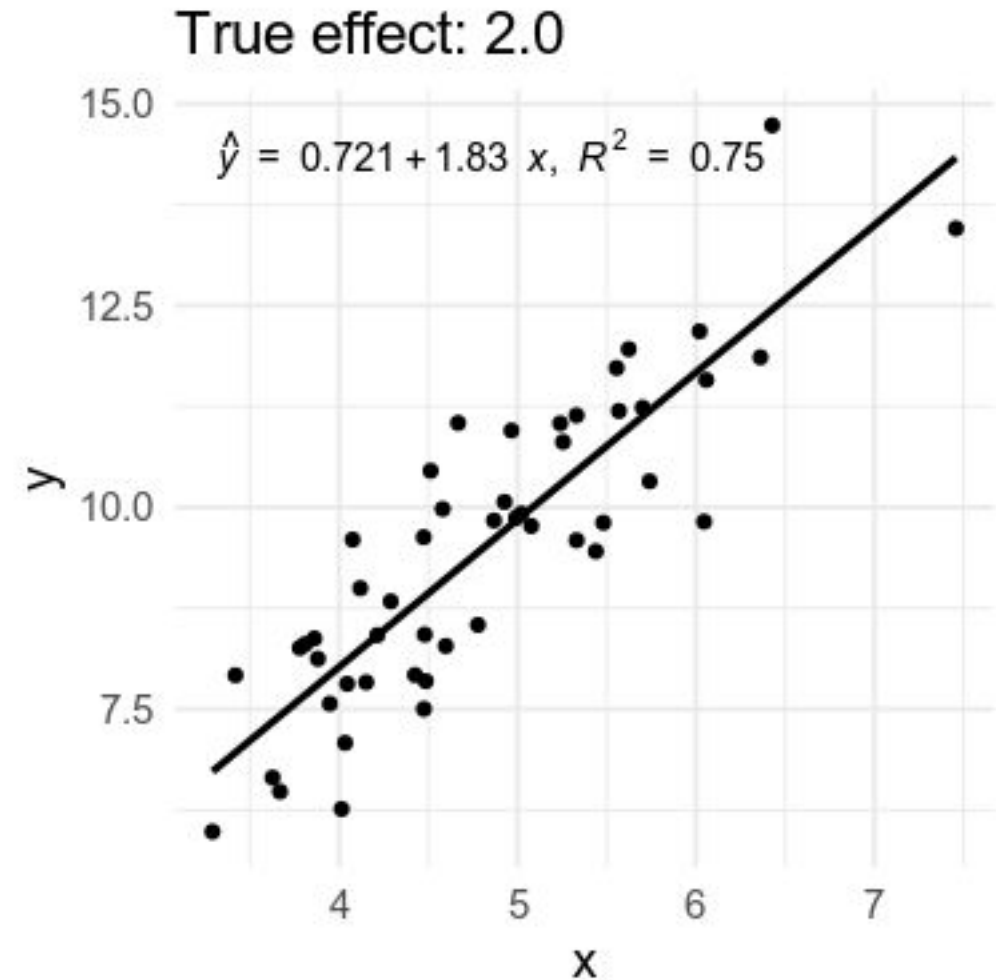
Correlation is caused by causation or confounding

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

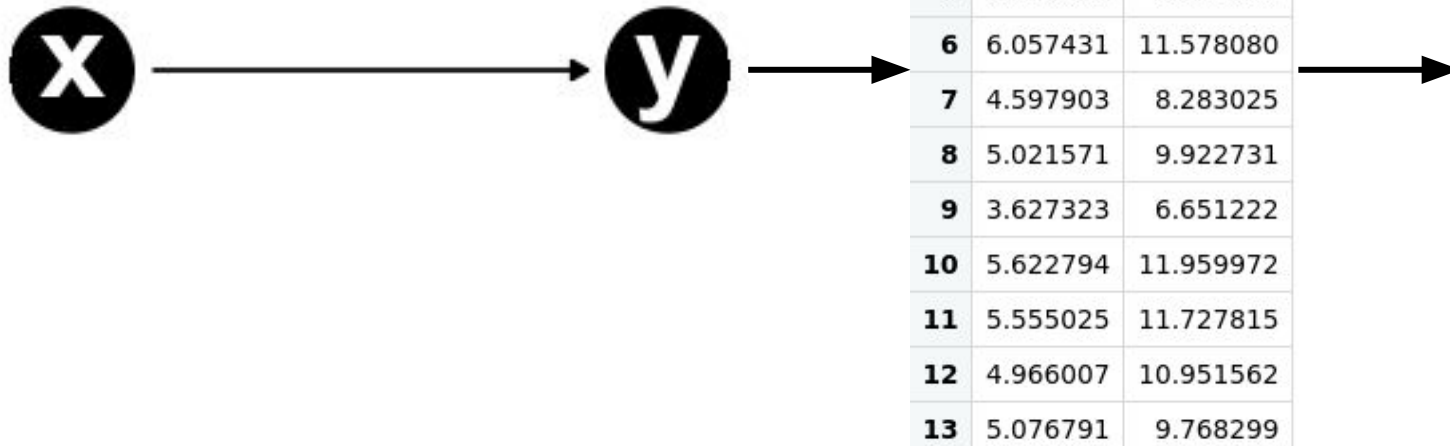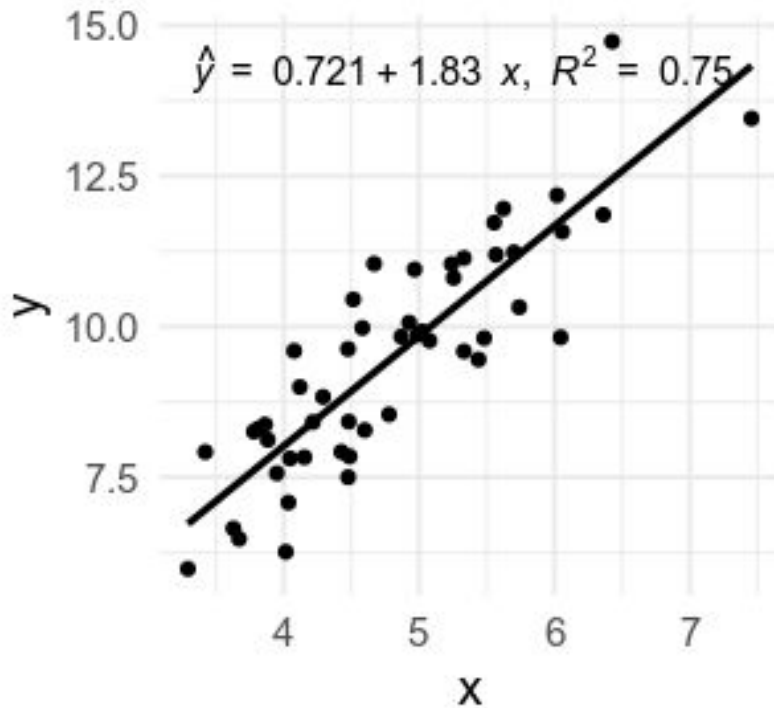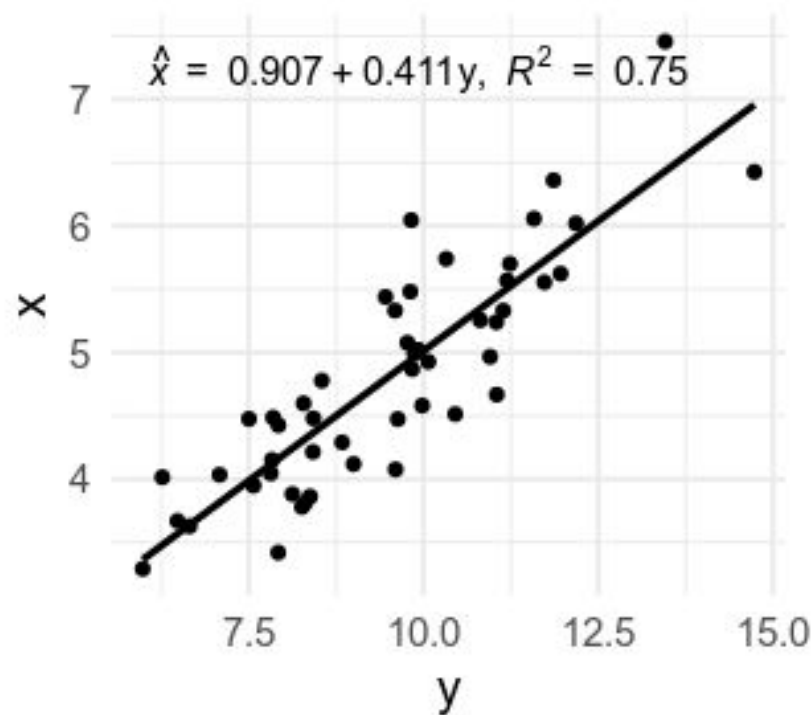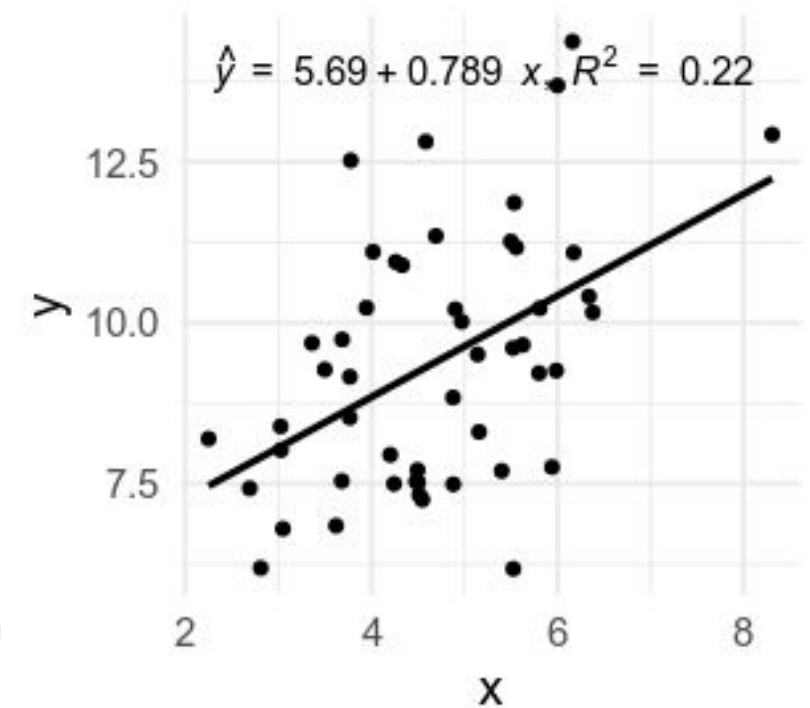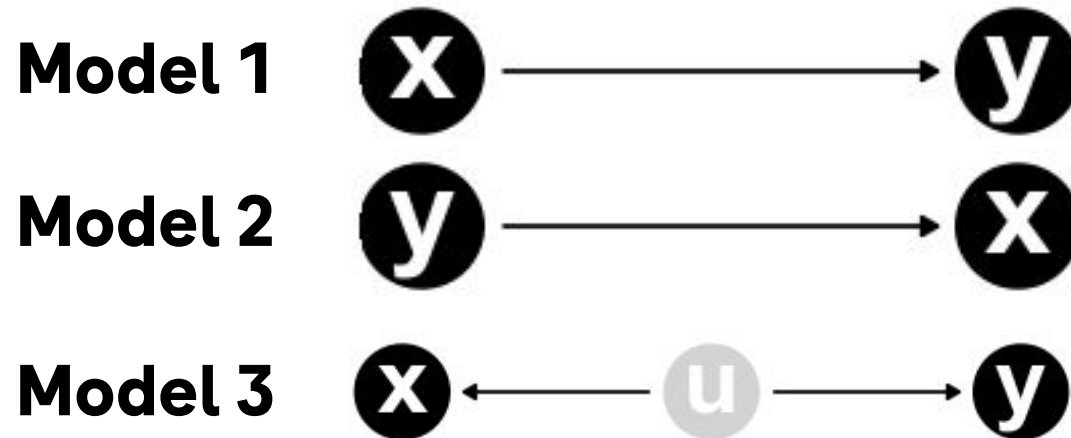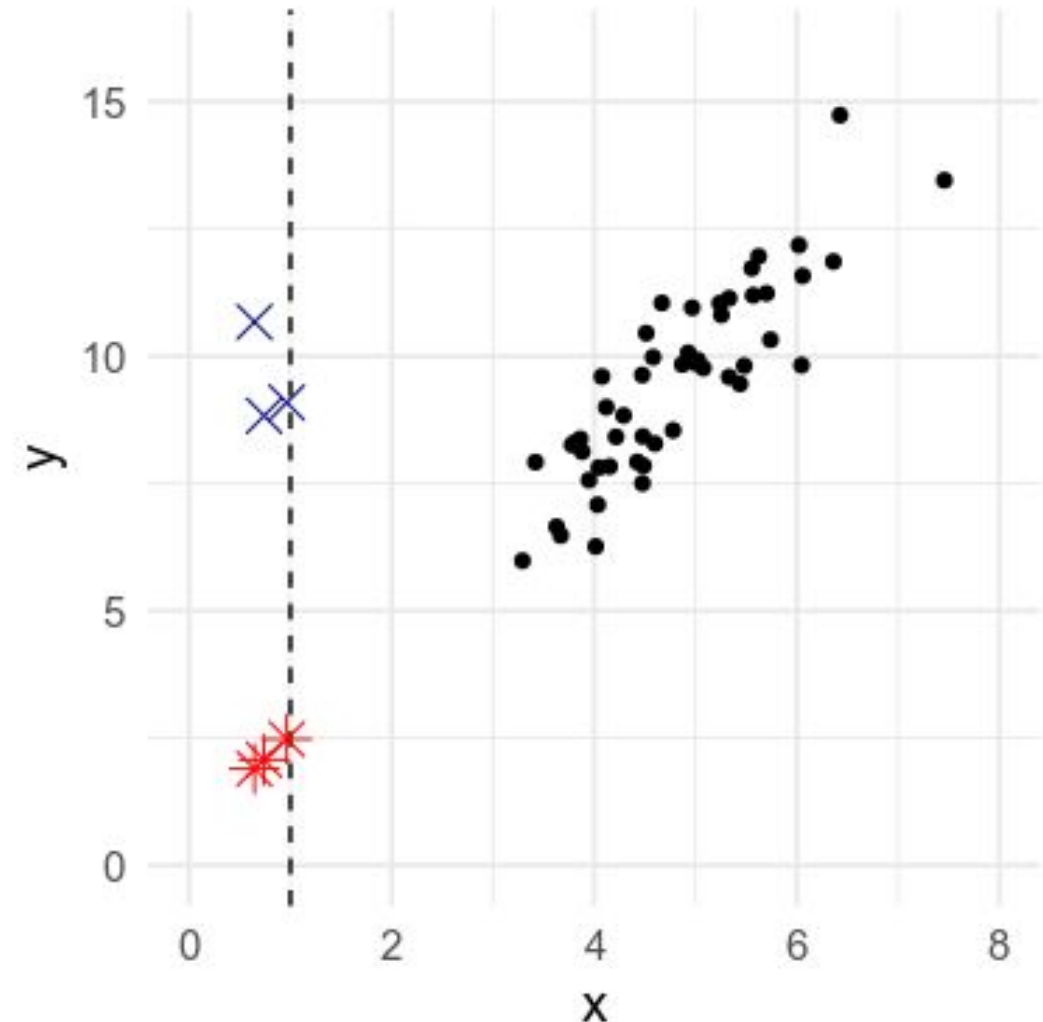**Model 1** x → y

Model 2 y → x

Model 3 x ← u → y

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?
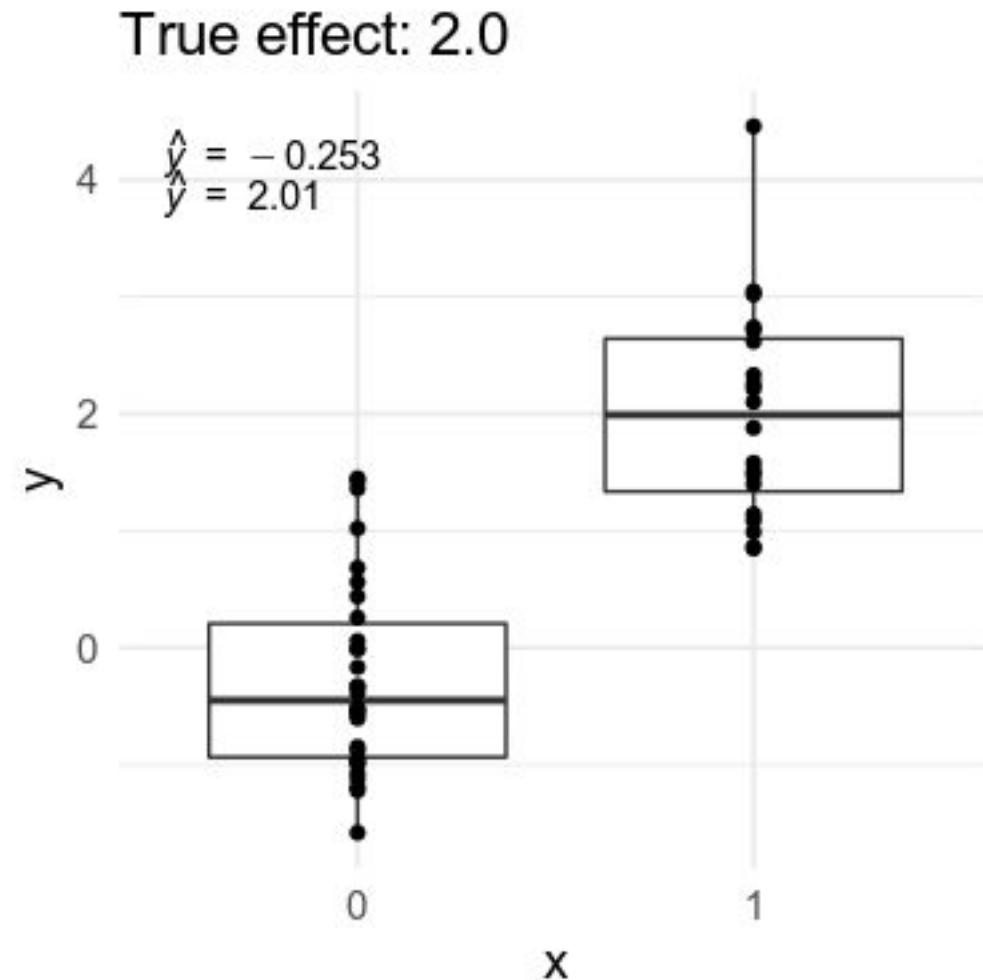
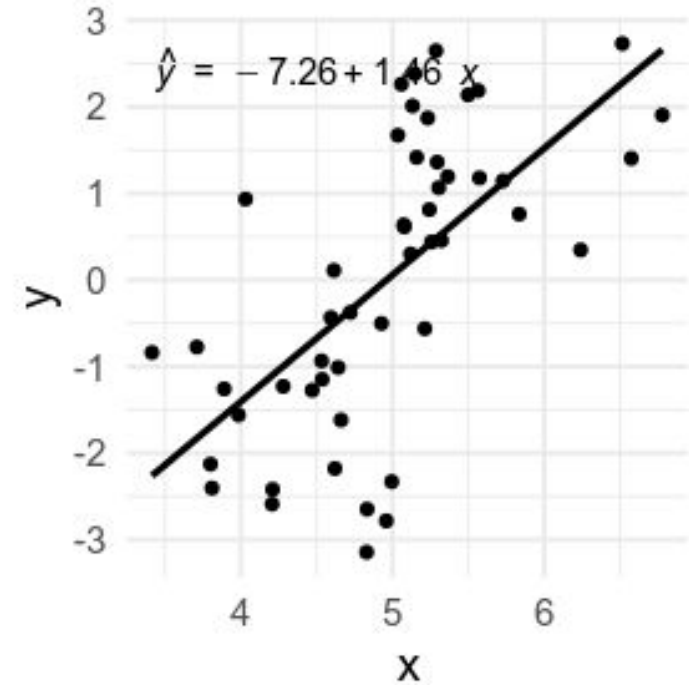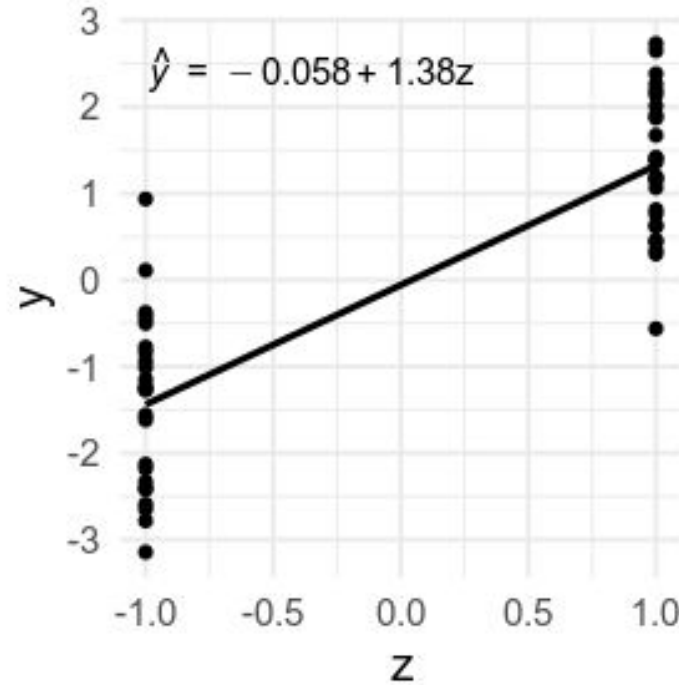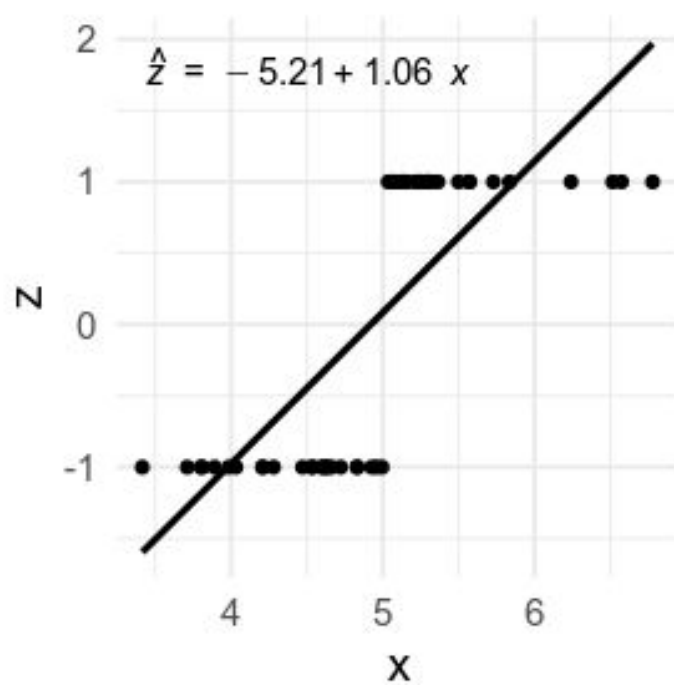# Variables in models can be either continuous or discrete

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
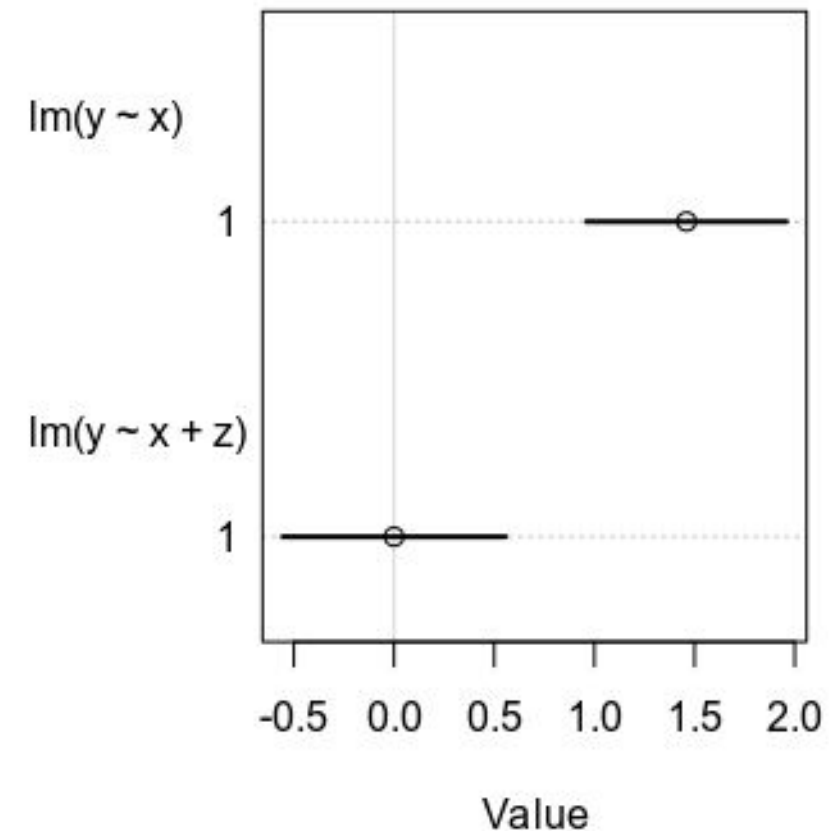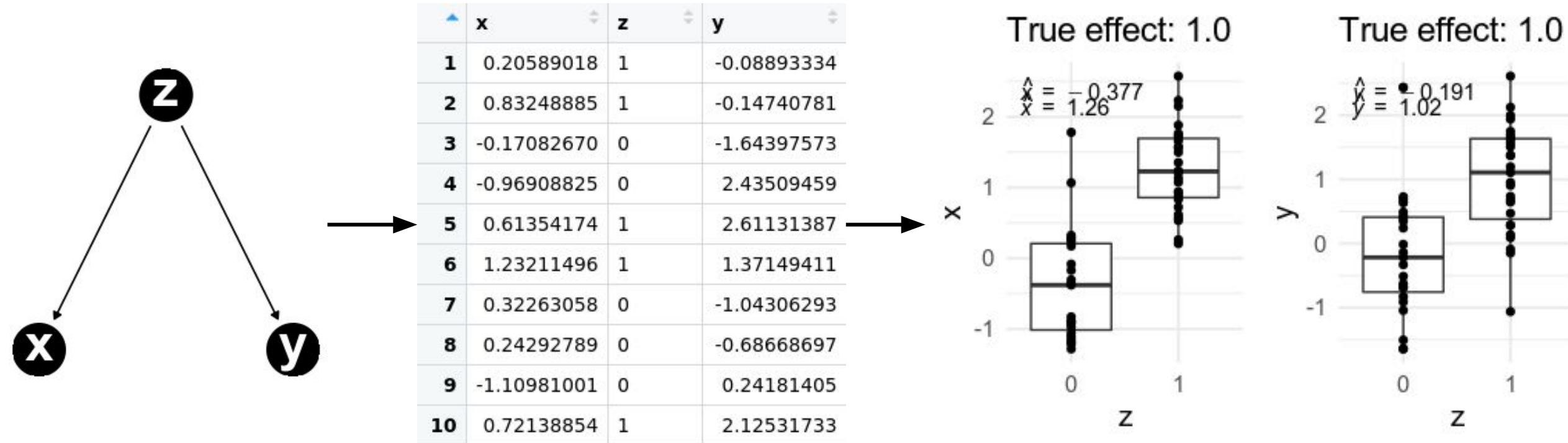

Assumptions of the **generative model**:
1. X is a random variable following Gaussian distribution N(5,1)
2. Z takes the value of -1 if X is smaller than 5, and 1 if X is equal to or larger than 5.
3. Y is a random variable with mean defined by Z*1.5.

Common DAG structures (2): Fork

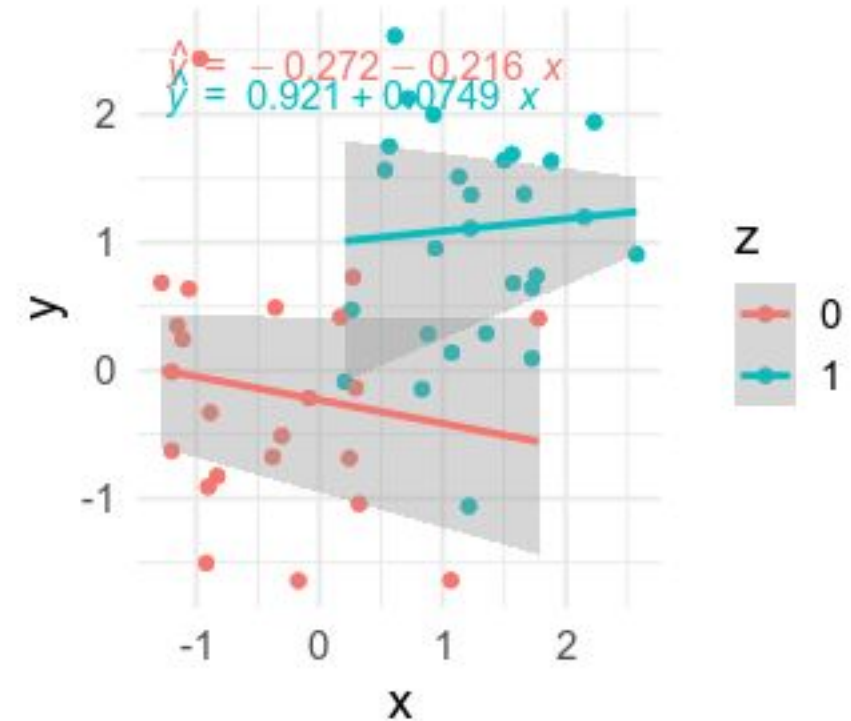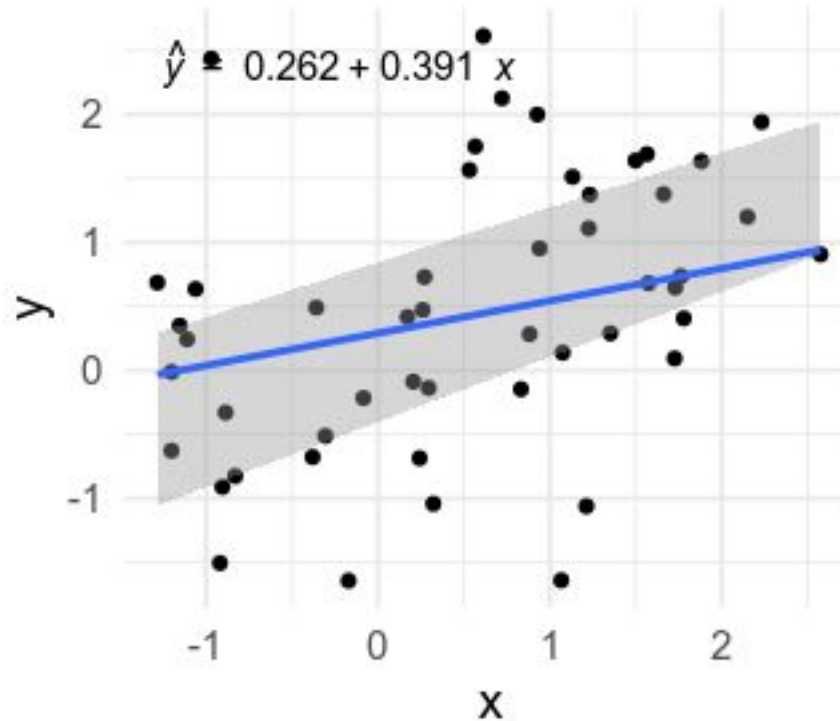Assumptions of the **generative model**:

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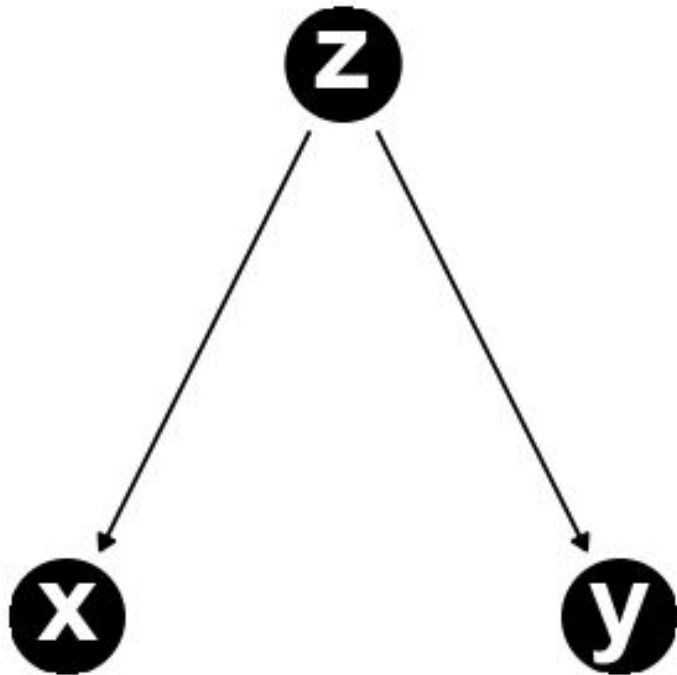
Given a fork structure, both children of the common cause are correlated. The correlation disappears when we condition on the common cause (i.e. stratification by the common cause in the case of discrete variables, or including the variable in the regression in the case of continuous variables).

Common DAG structures (3): Collider

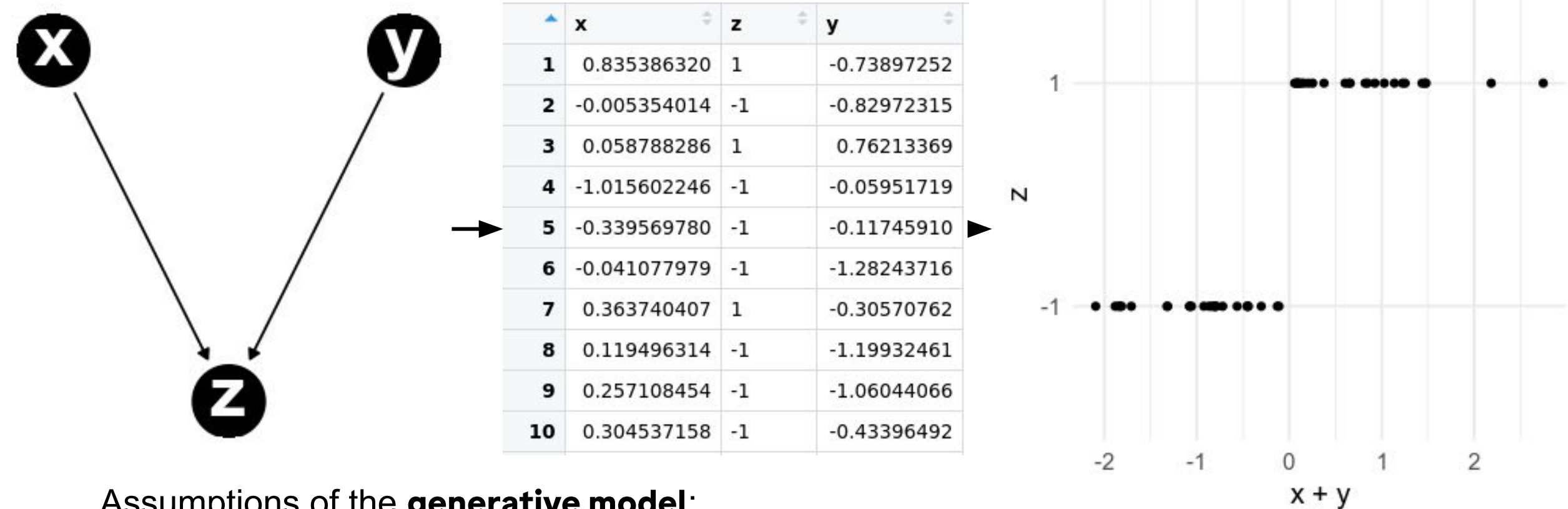|	x	z	y
1	0.835386320	1	-0.73897252
2	-0.005354014	-1	-0.82972315
3	0.058788286	1	0.76213369
4	-1.015602246	-1	-0.05951719
5	-0.339569780	-1	-0.11745910
6	-0.041077979	-1	-1.28243716
7	0.363740407	1	-0.30570762
8	0.119496314	-1	-1.19932461
9	0.257108454	-1	-1.06044066
10	0.304537158	-1	-0.43396492

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.

28

Conditioning on the collider introduces *spurious correlations*
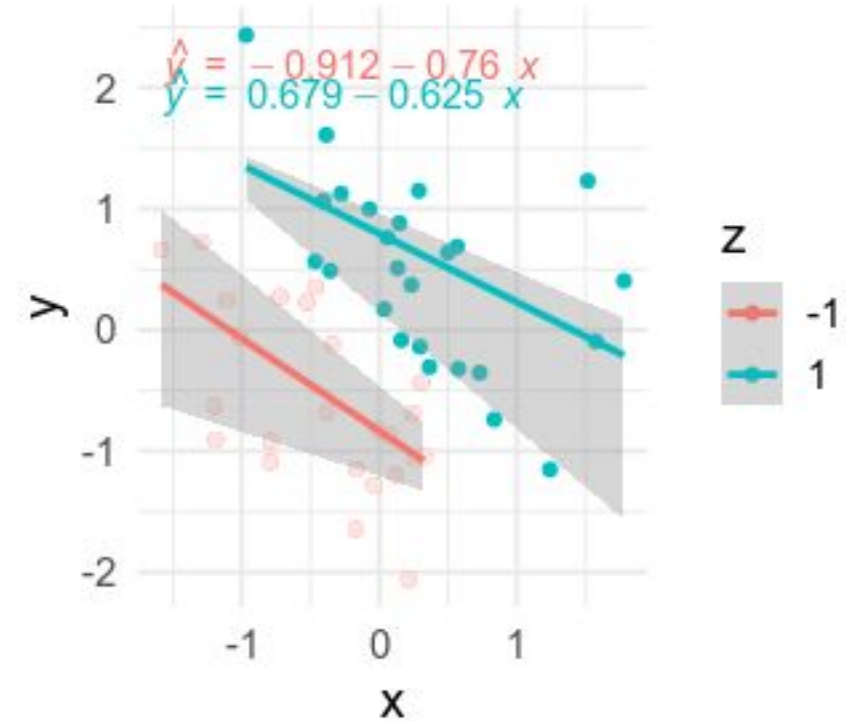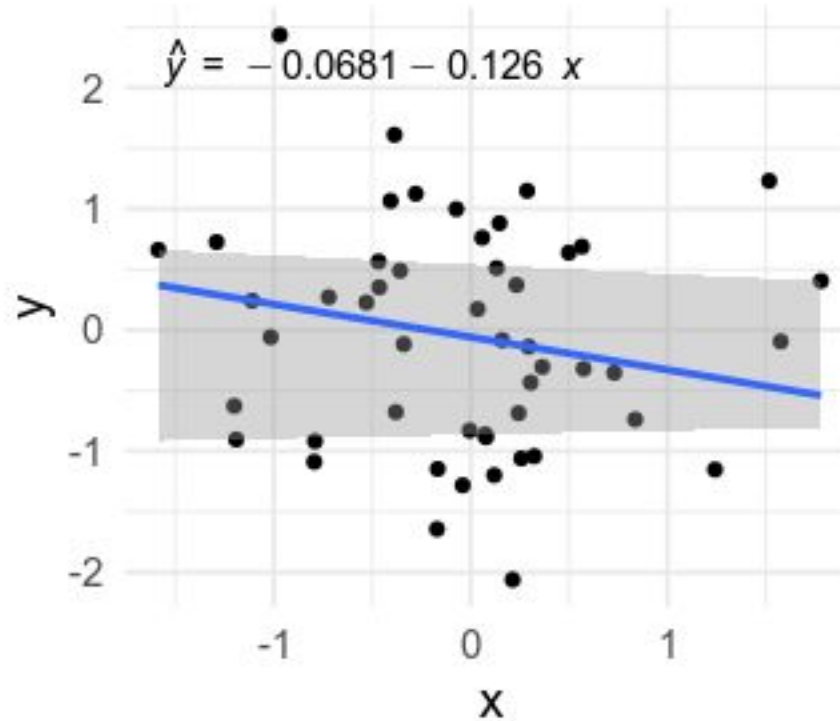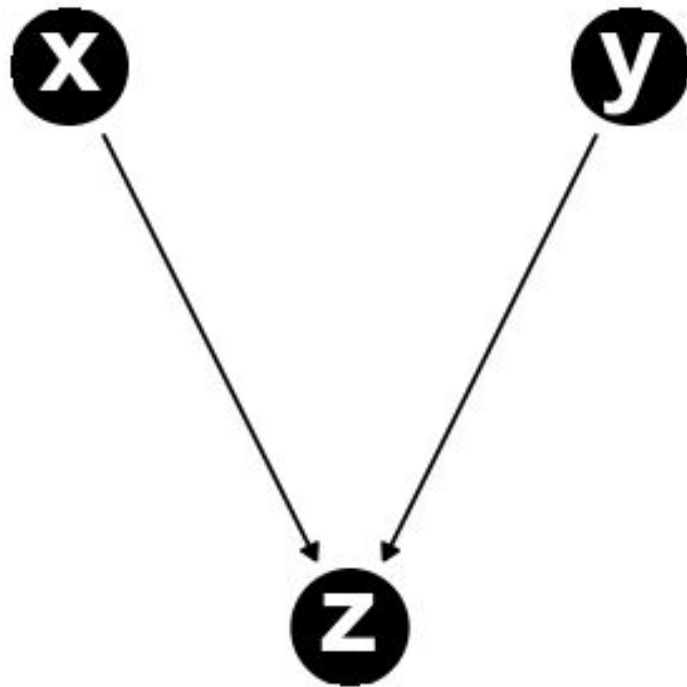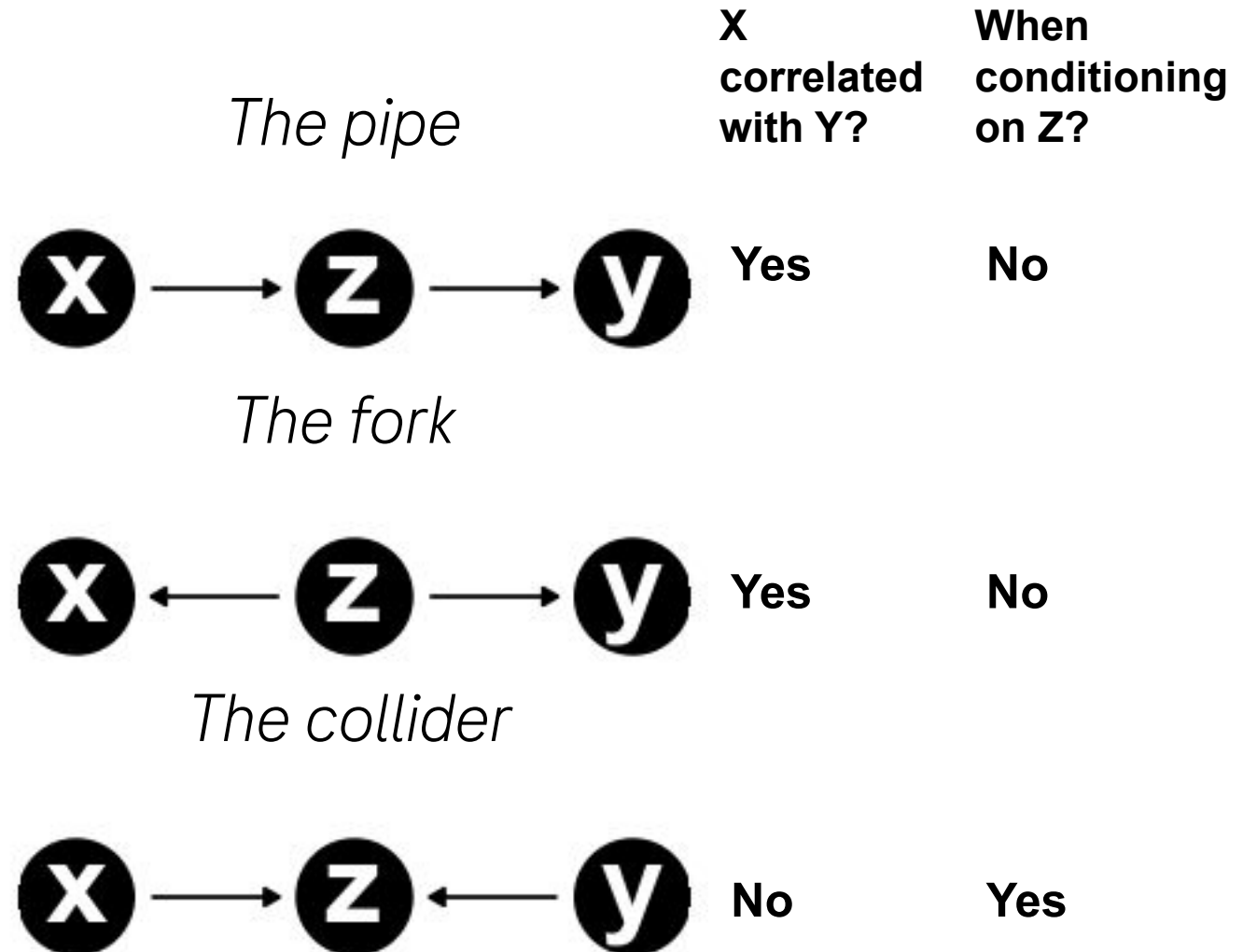

In a collider structure, the parents of the collider can be independent from each other. However, they become correlated when we condition on the collider. **Collider is everywhere!**

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		
$x \rightarrow z \rightarrow y$	Yes	No
The fork		
$x \leftarrow z \rightarrow y$	Yes	No
The collider		
$x \rightarrow z \leftarrow y$	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
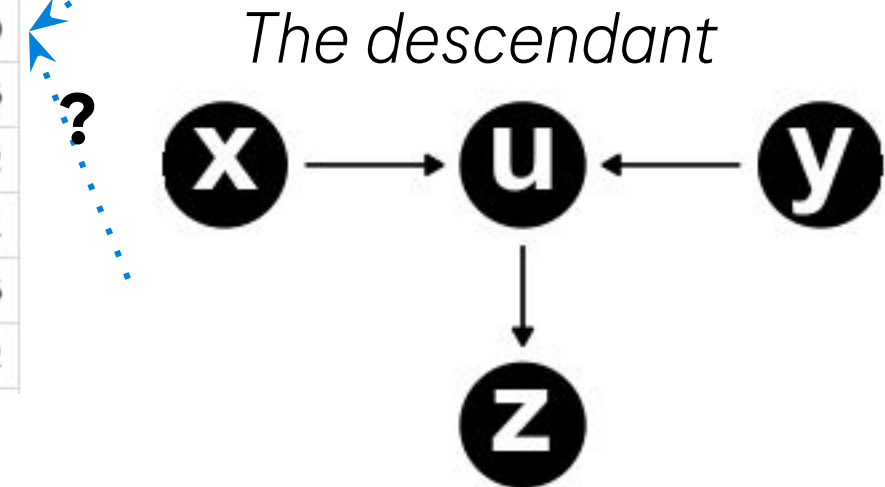
We need to build models (knowledge + assumptions) to infer causality

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.
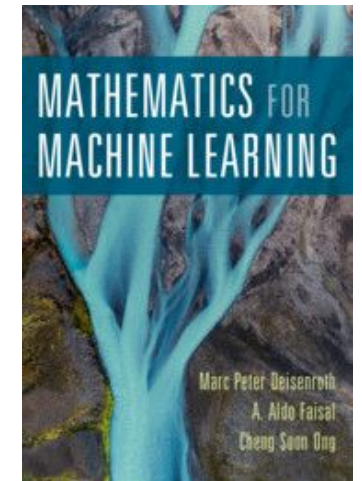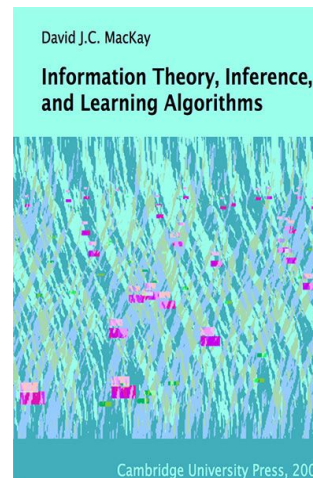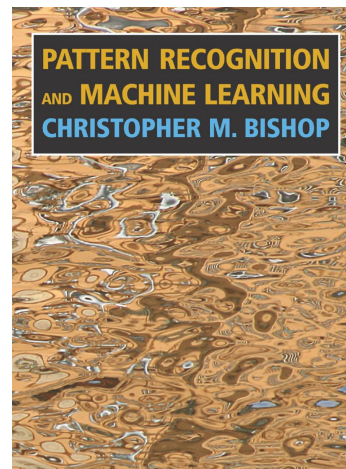
Resources for learning about machine learning

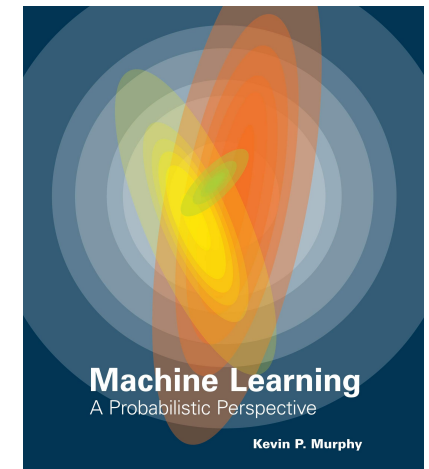*ESL* and *ISL*: From a frequentist view (almost)
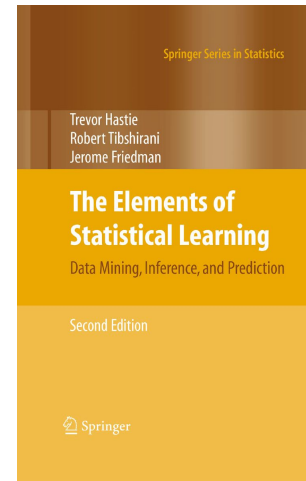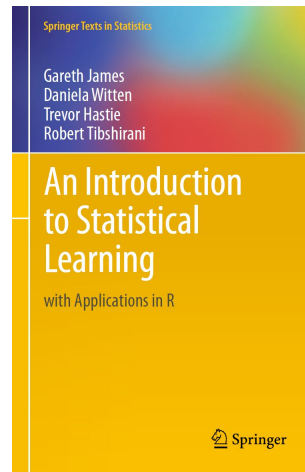
Mathematical foundations

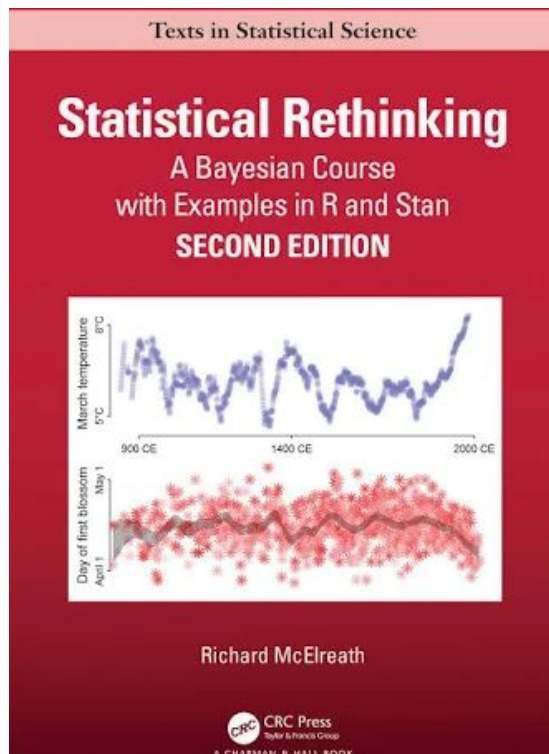*PRML* and *ITILA*: From a Bayesian view

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