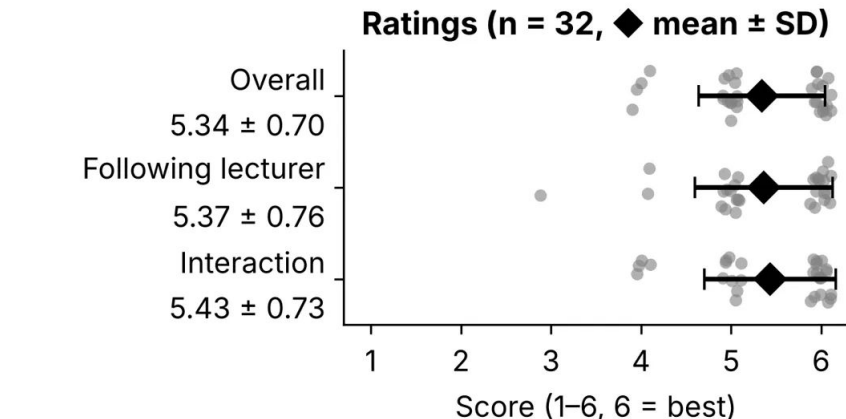
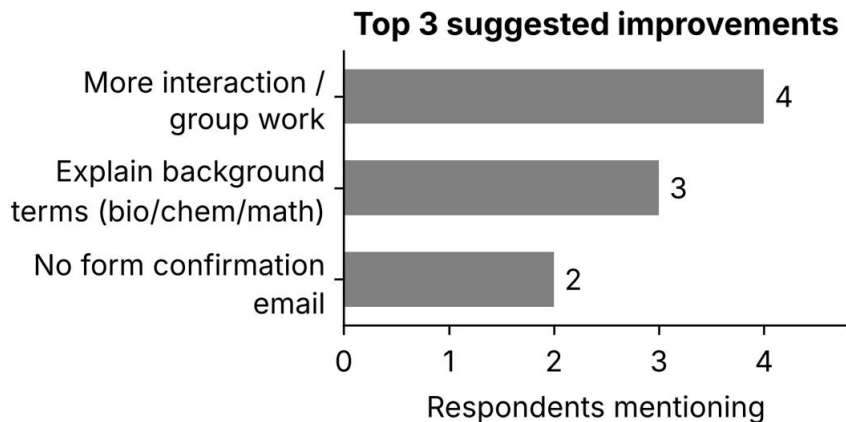


Feedback and offline activities of Lecture 6

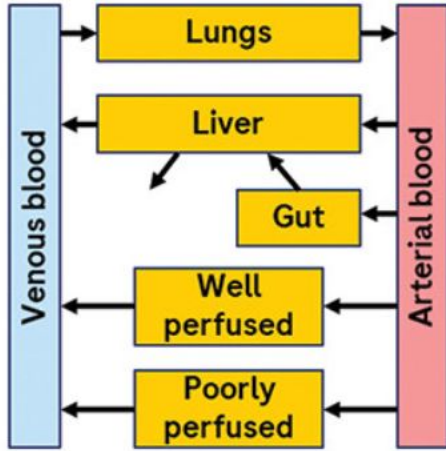


Common questions:

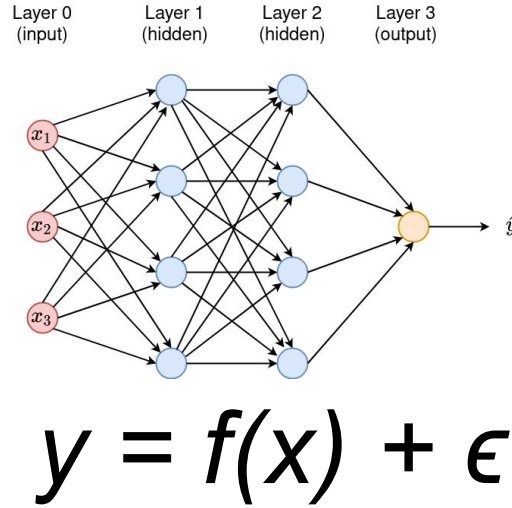
1. **Druggability (10):** Why are some proteins easier to target than others?
2. **Methodology of the review (9):** How are multi-target drugs counted? Why do 24% of drugs still have no known target?
3. **Biologics vs small molecules (8):** Why are biologics rising and developing faster?
4. **Do more targets mean better treatments? (7):** Is it better to pursue new targets or improve drugs for known ones?
5. **Finding new targets, including AI (7):** How are targets prioritised, and how are AI-predicted targets validated?

Suggestions: link reading to lecture, point to key sections or give guiding questions, hands-on tasks.

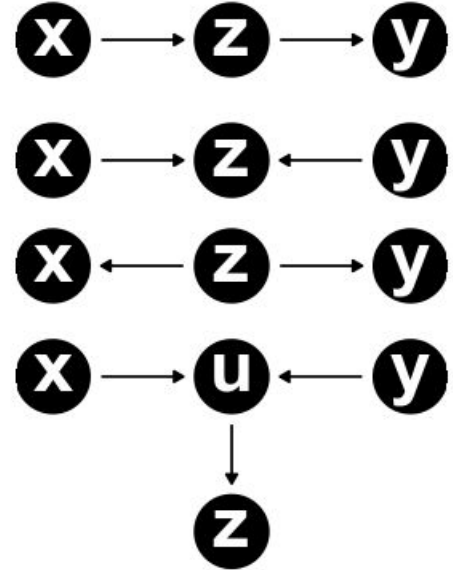
AMIDD 2026 Lecture 7: Three types of computational models



Mechanistic models



Statistical/ML/AI models



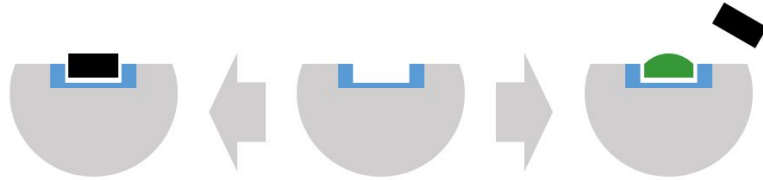
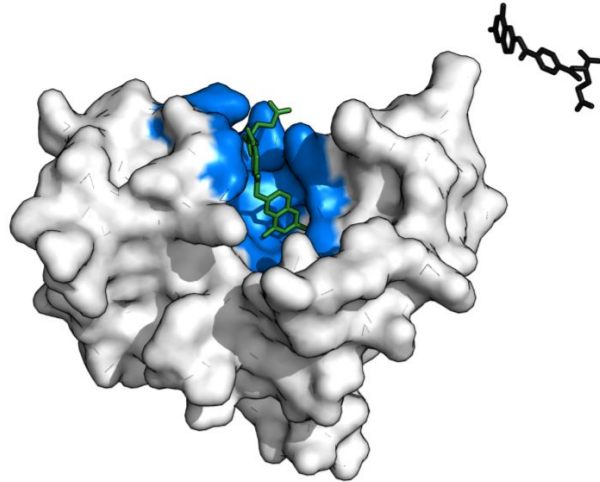
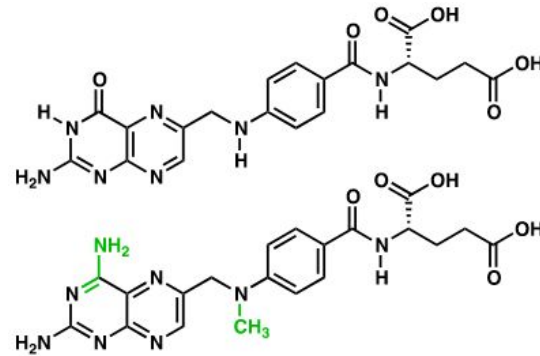
Causal models

PD 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*

Methotrexate is a competitive inhibitor of DHFR



Dihydrofolic acid
(natural ligand of
DHFR)

Methotrexate (MTX,
synthetic ligand)

The protein: Dihydrofolate reductase (DHFR) converts dihydrofolic acid into tetrahydrofolate. The process is important for cell proliferation and cell growth. DHFR is a drug target for oncology (cancer) and autoimmune diseases.

The natural substrate: Dihydrofolic acid (a reduced form of folate, which is also known as vitamin B9), in black. Dihydrofolic acid is the *natural ligand* of DHFR.

The drug: Methotrexate (MTX), in green, is a *synthetic ligand* of DHFR, and it is a *competitive inhibitor* of DHFR.

The binding site: where the enzyme binds its substrate and catalyses the chemical reaction, in blue.

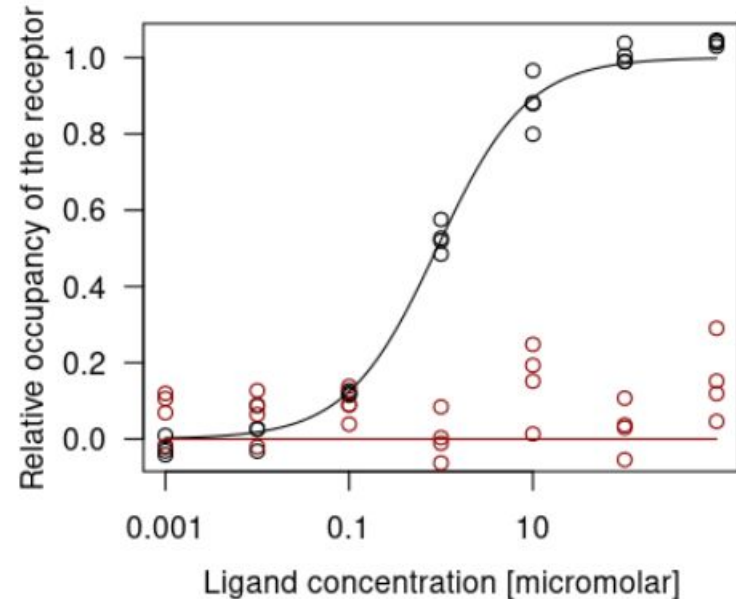
Concentration-occupancy curves characterize protein-ligand binding

X-axis: ligand concentration. Common units: molar (M), micromolar (μM , 10^{-6} M), nanomolar (nM, 10^{-9} M), picomolar (pM, 10^{-12} M).

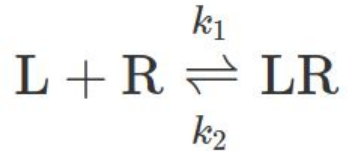
Y-axis: relative occupancy of the receptor. Alternative values are possible, for instance response (more about that later).

Points: individual measurements. In this plot: mean value of replicates with error bars indicating variability.

Lines: fitted sigmoidal curves using the Hill function or its variants.

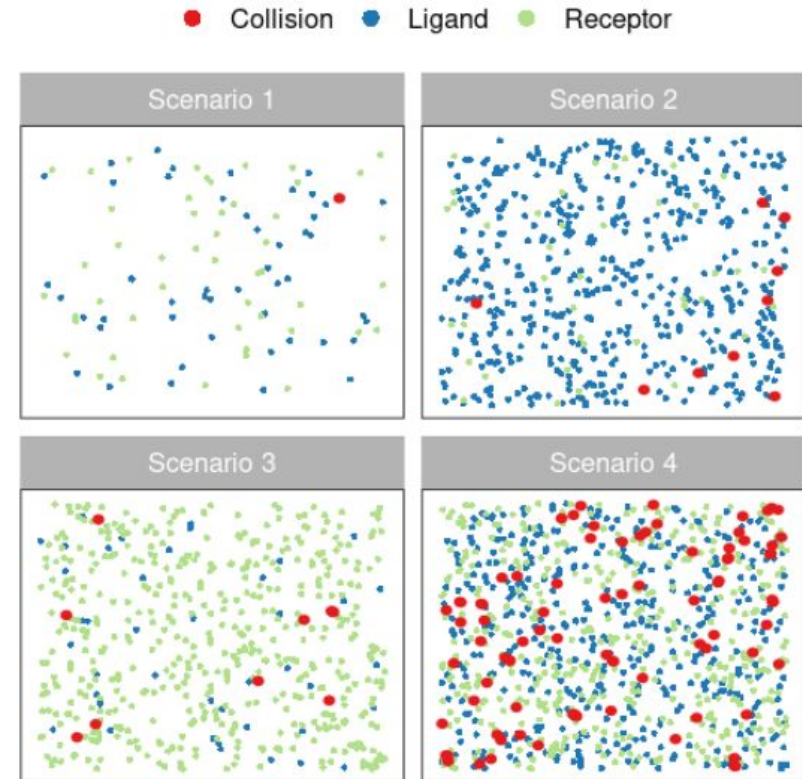


A simple mathematical model of how a receptor is occupied at varying drug concentrations



$$\frac{d[LR]}{dt} = k_1[L][R] - k_2[LR]$$

- Ligand binding to receptor is a reversible reaction.
- **The law of mass action:** the rate of the chemical reaction is directly proportional to the product of the activities or concentrations of the reactants. The proposition can be derived from the *collision theory*. See the right graph for an illustration.



An ordinary differential equation (ODE) model quantifies receptor occupancy by varying concentrations of ligands

$$\begin{array}{ccc}
 \text{L} + \text{R} \xrightleftharpoons[k_2]{k_1} \text{LR} & \xrightarrow{\text{The law of mass action}} & \frac{d[\text{LR}]}{dt} = k_1[\text{L}][\text{R}] - k_2[\text{LR}] \\
 \downarrow & & \downarrow \text{At equilibrium, no net change of [LR]} \\
 & & k_1[\text{L}][\text{R}] = k_2[\text{LR}] \\
 & & \downarrow R_{\text{total}} = [\text{R}] + [\text{LR}] \\
 & & k_1[\text{L}]([\text{R}_{\text{total}}] - [\text{LR}]) = k_2[\text{LR}], \\
 & & \downarrow \\
 [\text{LR}] = [\text{R}_{\text{total}}] \frac{[\text{L}]}{[\text{L}] + K_D} & \xleftarrow{K_D \equiv k_2/k_1} & [\text{LR}] = \frac{k_1[\text{L}][\text{R}_{\text{total}}]}{k_1[\text{L}] + k_2}
 \end{array}$$

The ODE model leads to the simplest form of the *Hill-Langmuir Equation*

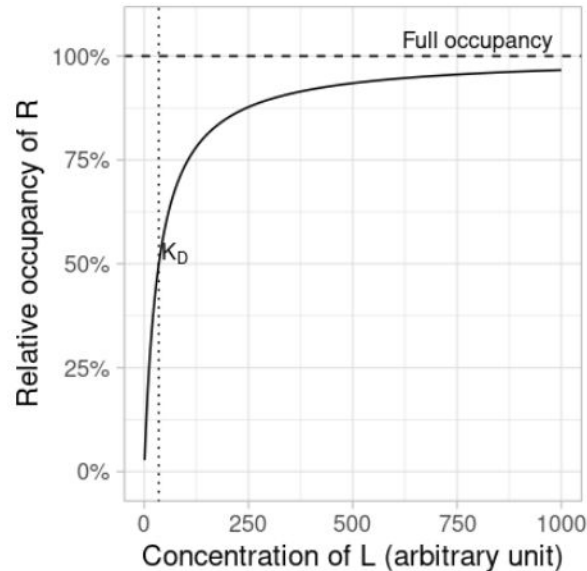
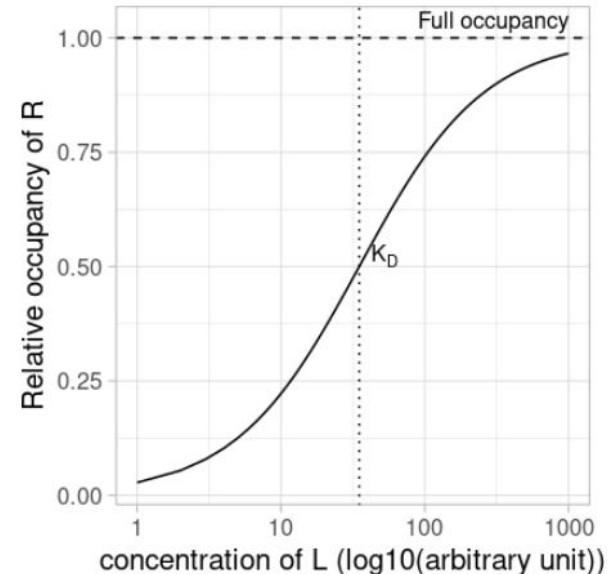


The Hill-Langmuir function describes the occupancy of receptors by natural ligands or drugs. We will meet it again.

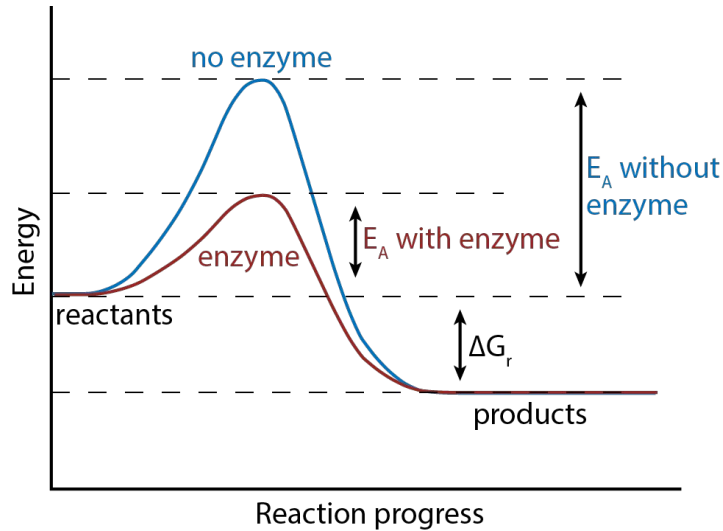
We can interpret K_D (the dissociation constant) both mathematically and physically & chemically. Mathematically, K_D represents (1) the ratio of reaction rate constants, and (2) the concentration required to occupy half of the receptors.

Hill-Langmuir function:

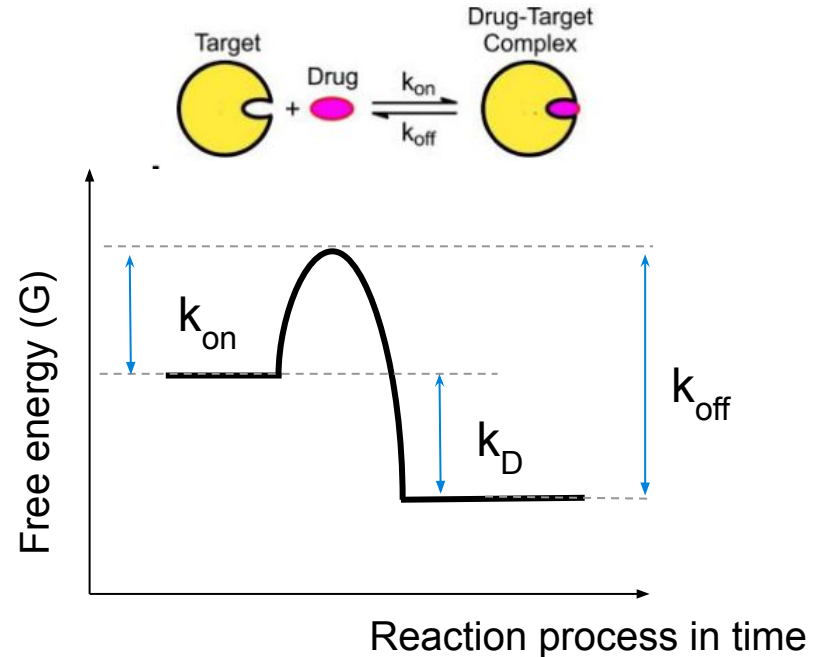
$$[\text{LR}] = [\text{R}_{total}] \frac{[\text{L}]}{[\text{L}] + K_D}$$



The thermodynamic interpretation: K_D is directly associated with the free energy of the reaction



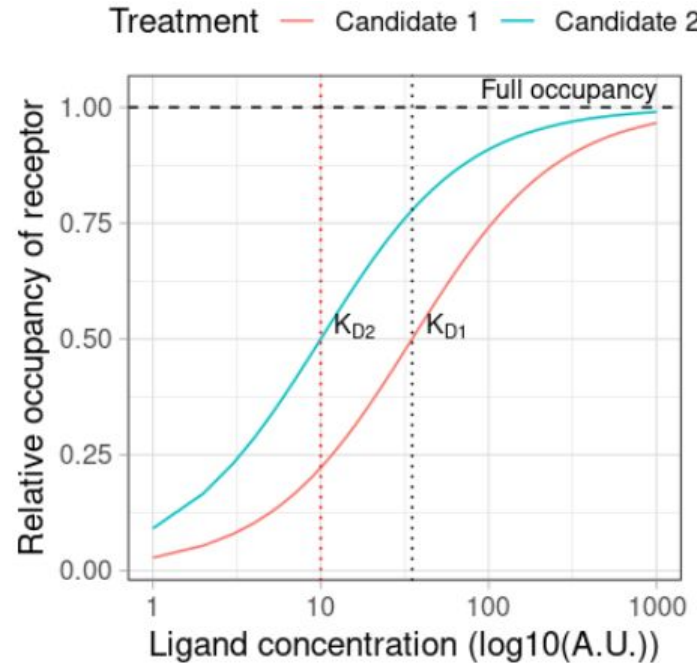
Left: [LibreTexts, Role of Enzymes](#),
 Matthew F Kirk, Kansas State University



$$\Delta G = RT \ln K_D$$

(R: gas constant, T: absolute temperature)

Question: all other conditions the same, which drug candidate is more favorable? Why?



The Lotka-Volterra model of predator-prey relationships

- The Lotka-Volterra equations modelling predator-prey relationships.

$$\frac{dx}{dt} = \alpha x - \beta xy, \quad (1)$$

$$\frac{dy}{dt} = -\gamma y + \delta xy, \quad (2)$$

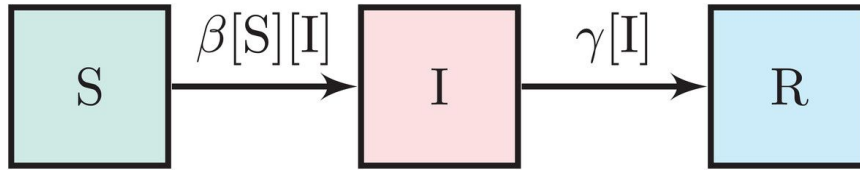
where

- x is the number of prey (*e.g.* rabbits),
- y is the number of predator (*e.g.* foxes),
- $\frac{dx}{dt}$ and $\frac{dy}{dt}$ represent growth rates of the two populations,
- t represents time,
- α , β , γ , and δ are real parameters specifying the interaction of the two species.



Interactive visualization: <https://claude.ai/artifact/BFuzGkY3nd3qV9qSHVKgTH>

The SIR model of epidemiology models population behavior of infection and recovery

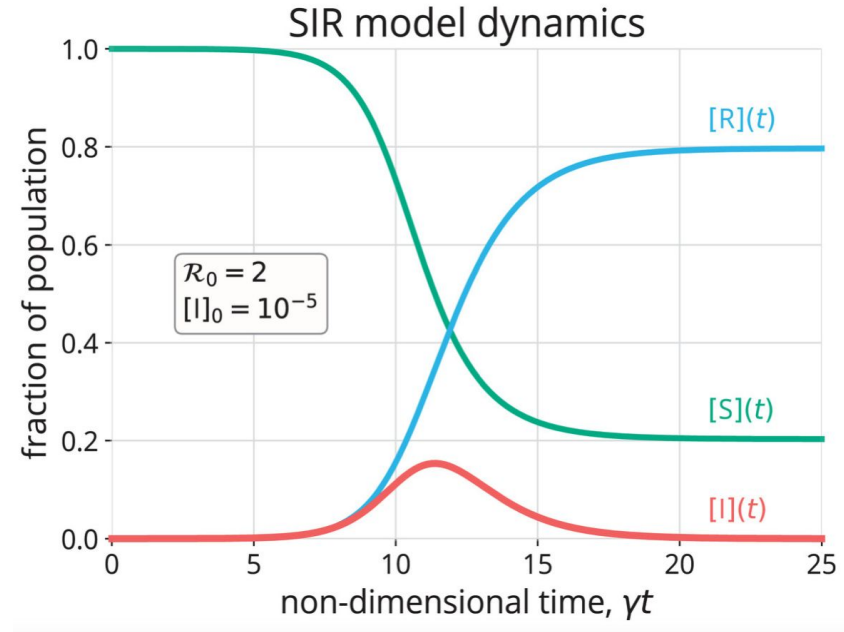


The SIR model of epidemiology

- S: Susceptible
- I: Infectious
- R: Removed

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta IS}{N}, \\ \frac{dI}{dt} &= \frac{\beta IS}{N} - \gamma I, \\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

Simon CM. 2020. The SIR dynamic model of infectious disease transmission and its analogy with chemical kinetics. PeerJ Physical Chemistry 2:e14 <https://doi.org/10.7717/peerj-pchem.14>



R_0 , the basic reproduction number, is the average number of secondary infections caused by one infectious individual in a fully susceptible population. It is defined as β/γ .

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

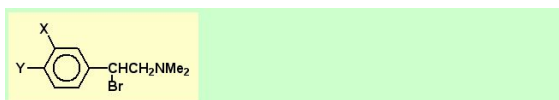
QSAR is a statistical modelling of correlation between biological activity and physicochemical properties, or $\Delta\phi=f(\Delta S)$, where ϕ indicates a biological activity and S indicates a chemical structure (1868-1869).

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}

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

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.



	meta	para	meta-				para-						
(X)	(Y)	F	Cl	Br	I	Me	F	Cl	Br	I	Me	obsd.	calc.a)
H	H											7.46	7.82
H	F						1					8.16	8.16
H	Cl							1				8.68	8.59
H	Br								1			8.89	8.84
H	I									1		9.25	9.25
H	Me										1	9.30	9.08
F	H	1										7.52	7.52
Cl	H		1									8.16	8.03
Br	H			1								8.30	8.26
I	H				1							8.40	8.40
Me	H					1						8.46	8.28
Cl	F		1				1					8.19	8.37
Br	F			1			1					8.57	8.60
Me	F					1	1						
Cl	Cl		1					1					
Br	Cl			1					1				
Me	Cl					1			1				
Cl	Br		1										
Br	Br			1									
Me	Br					1							
Me	Me						1						
Br	Me			1									

$\log (1/ED_{50}) = -$

 $+ ($

 $+ ($

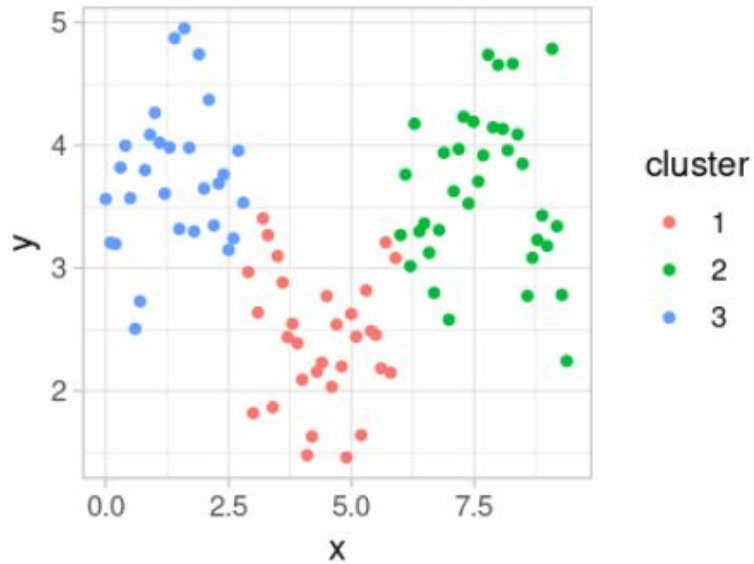
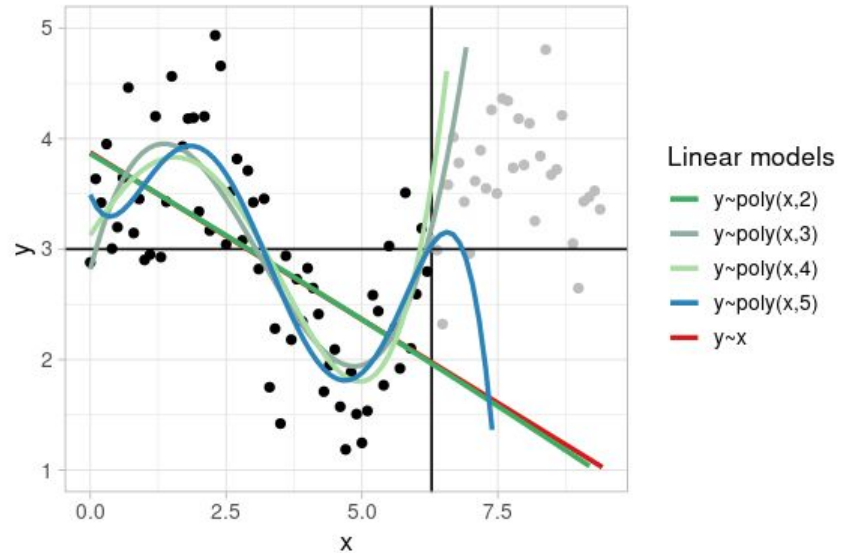
 $n =$

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

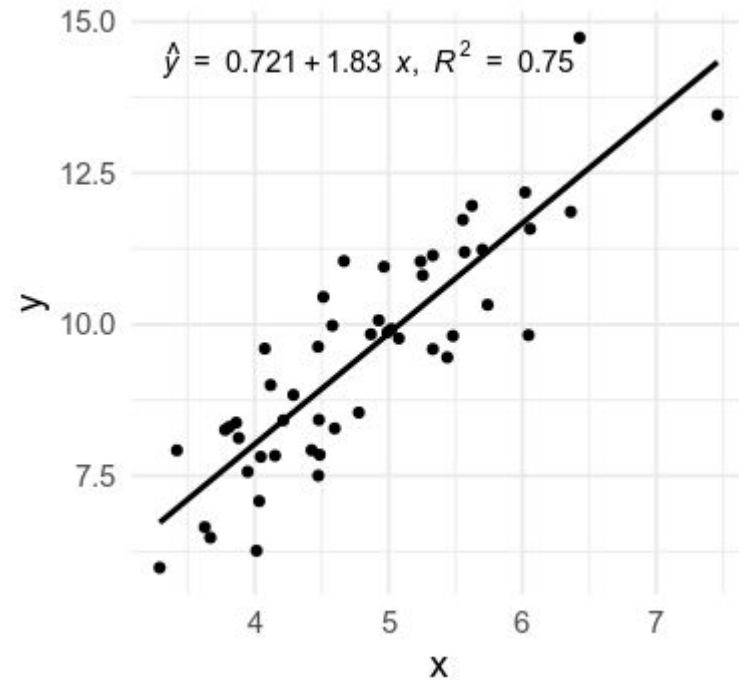


The simplest linear model has three components: the intercept, the slope, and a measure of fit

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

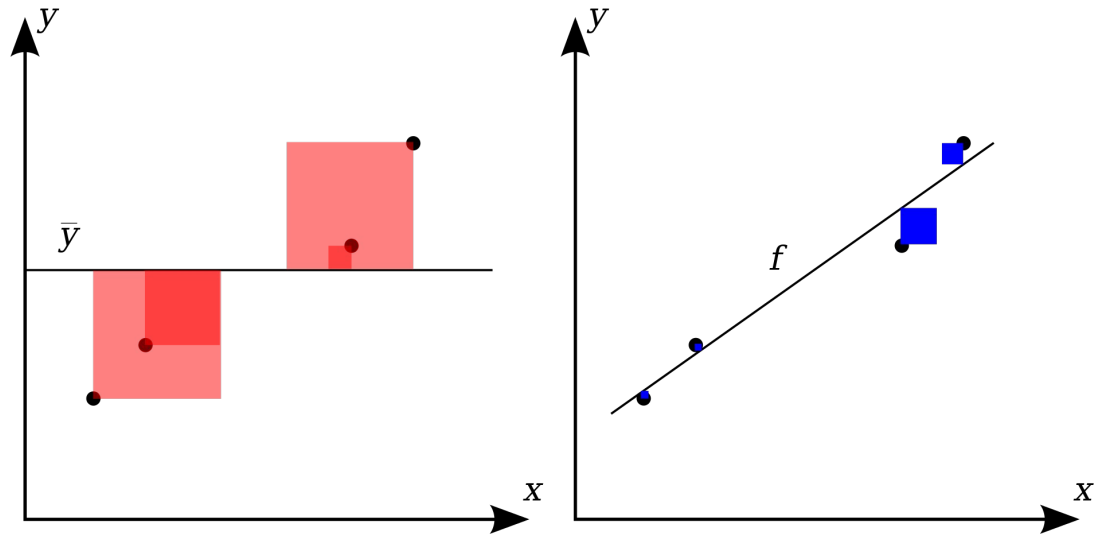


A visual explanation of R^2

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 $1 - (\text{blue area}) / (\text{red area})$.



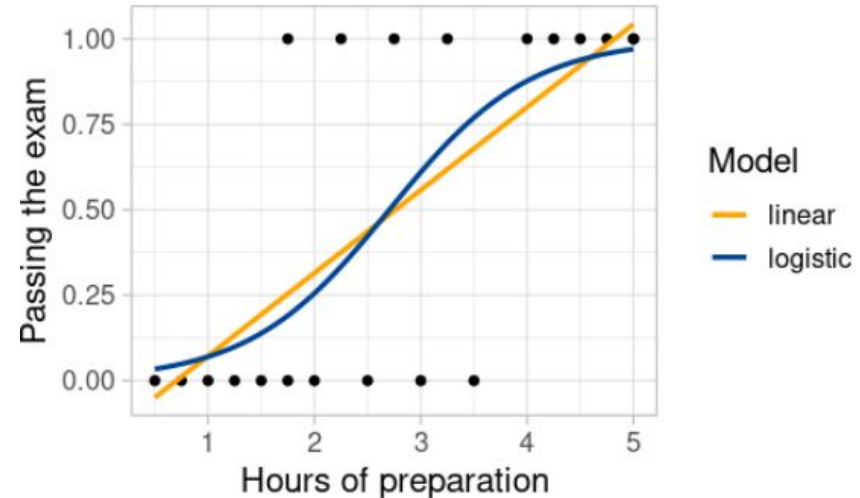
Work by Orzetto, CC-SA 3.0, from [Wikimedia](#)

Question: can R^2 be negative?

Logistic regression is an example of a generalized linear model, which allows the dependent variable to take values other than real numbers

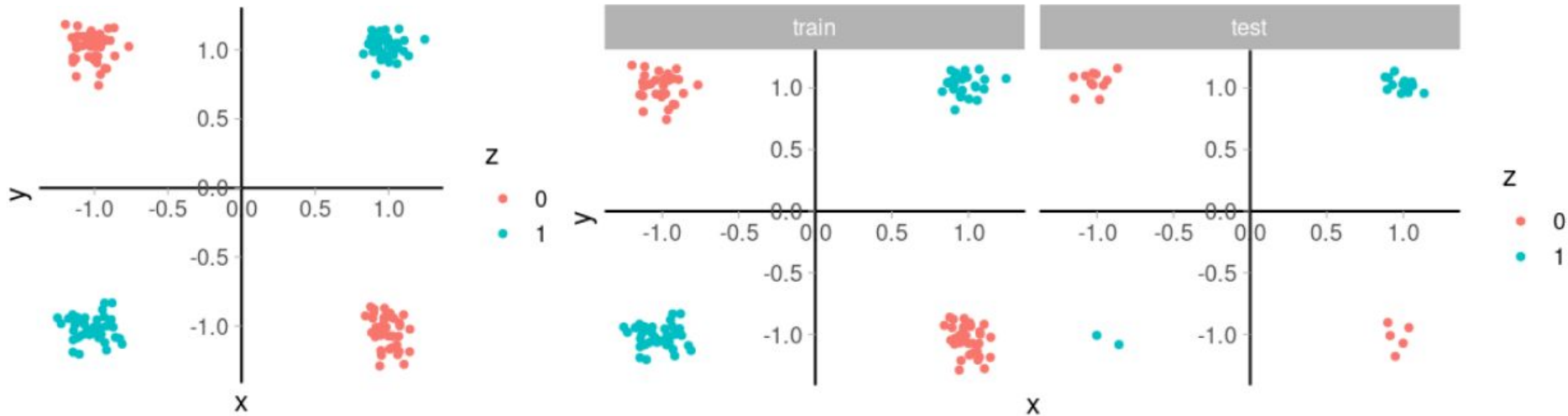
- Dependent variable of a linear regression model is defined on $\mathbb{R}$.
- *Generalized* linear models allow the dependent variable to be defined on other domains than real numbers, for instance binary (0/1), counts (non-negative integers), etc.
- Logistic regression maps input real numbers to the 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).

$$t = \alpha + \beta x \quad y = \frac{1}{1 + e^{-t}}$$



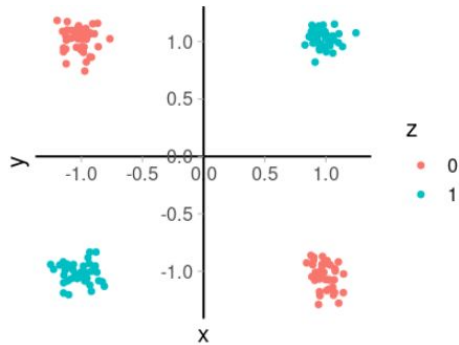
Data come from [Wikipedia's item on logistic regression](#)

We next address two problems: (1) how a model performs on new data, and (2) non-linearity

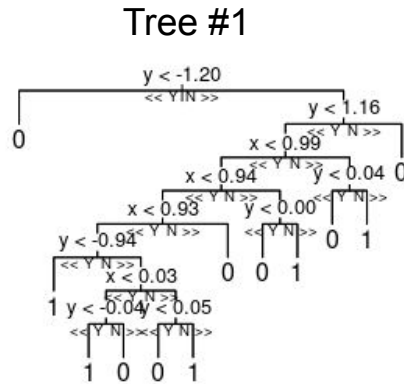
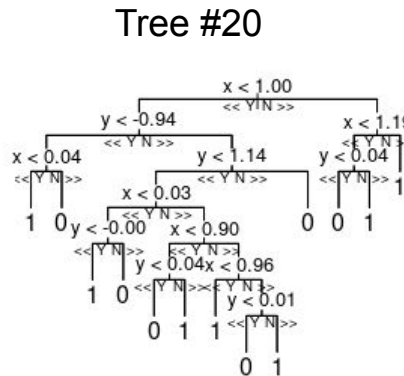
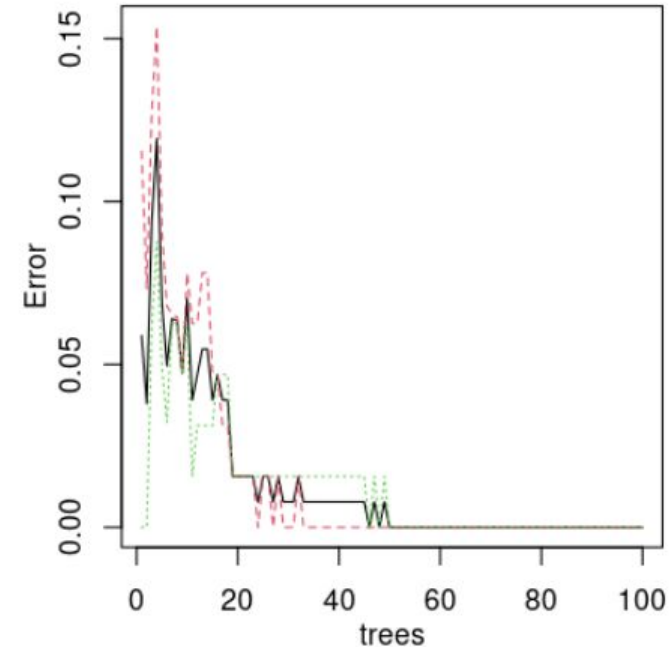


- Left: a simple example of non-linearity: linear models cannot predict z well based on values of x and y . We need something else.
- Right: a model is usually trained on some data, and the performance is assessed on unseen test data.

We can use random forest to model non-linearity



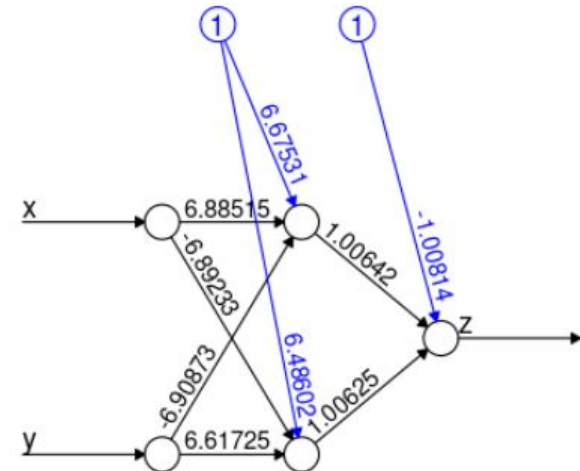
- Random forest is a collection of *decision trees*. Each tree partitions the input data to make predictions.
- Random forest is an example of *ensemble methods*: each tree has weak performance; however, the consensus can perform surprisingly well.



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

- Neural network models non-linearity by applying multiple linear combinations and (usually) non-linear activation functions in the forward propagation.
- Once the architecture (# hidden layers, # nodes, *etc.*) is fixed, the edge weights of the network are initialized with random numbers, and then optimized by iterative forward and backpropagation to minimize the error.
- Right figure: the trained neural network with the example data. Blue nodes indicate intercepts.

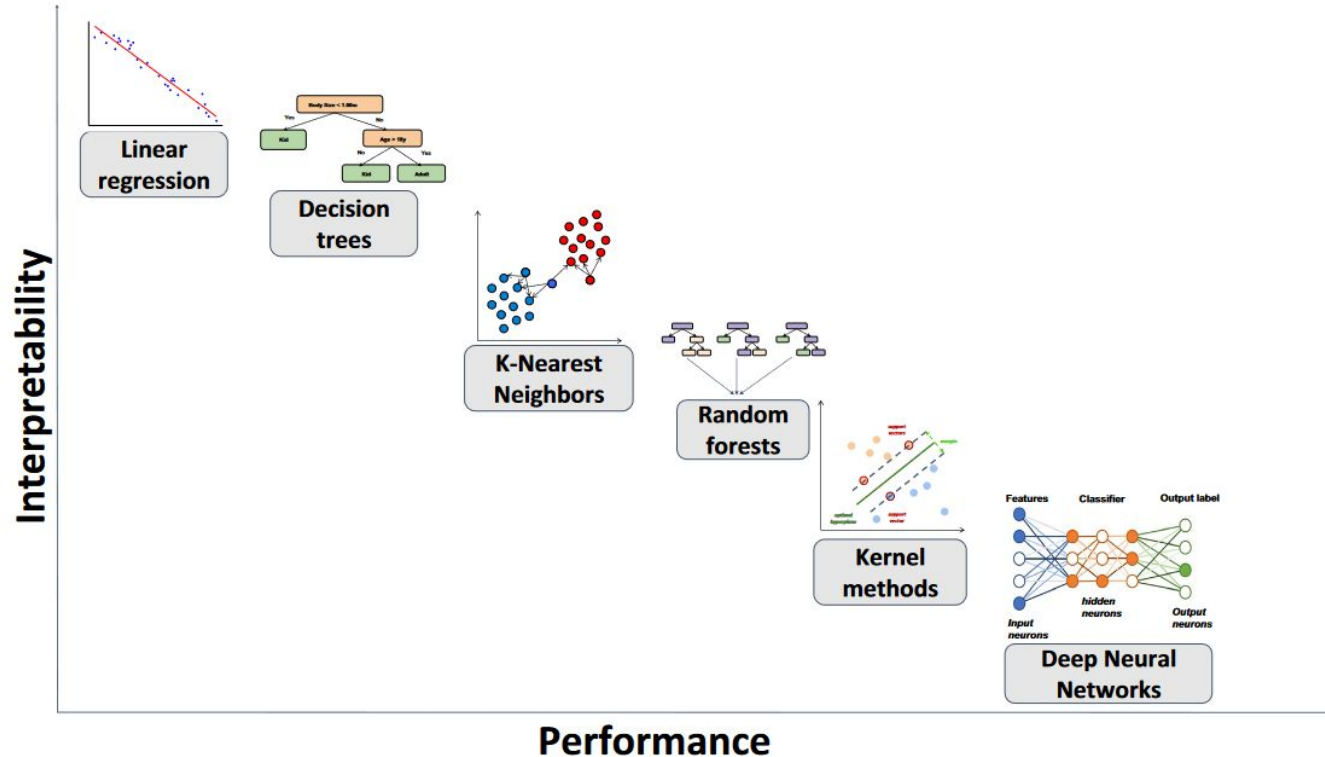
	Reference	
Prediction	0	1
0	16	0
1	0	16



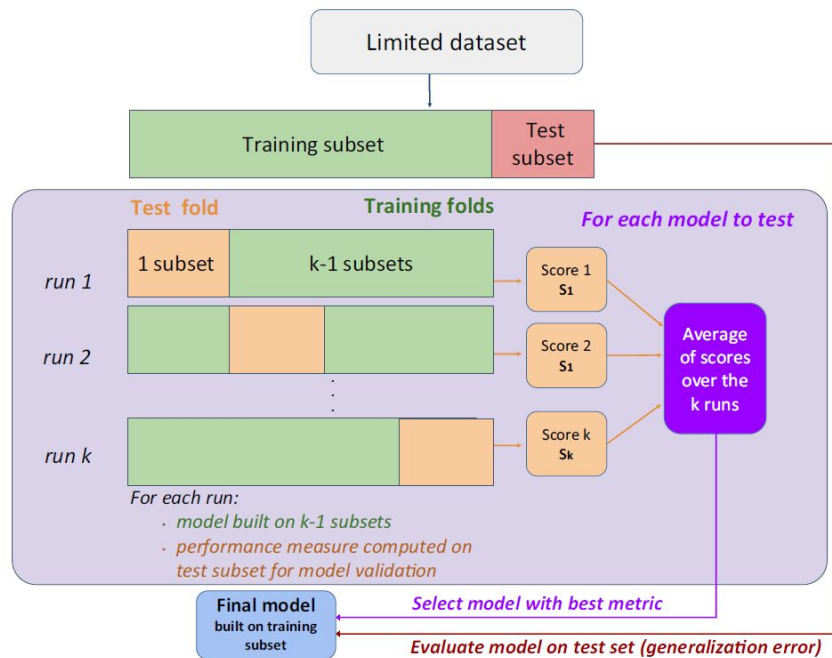
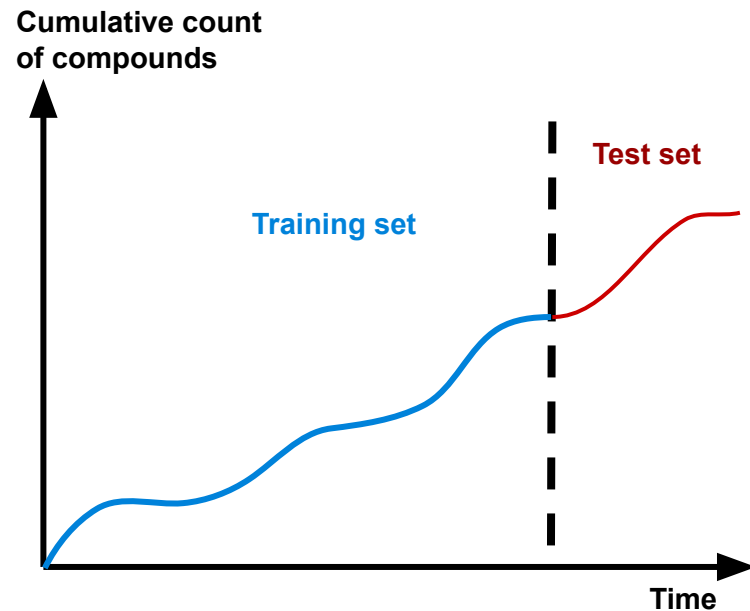
Hidden layer

Error: 0.000172 Steps: 1085

Generally, well-performing models tend to be less interpretable



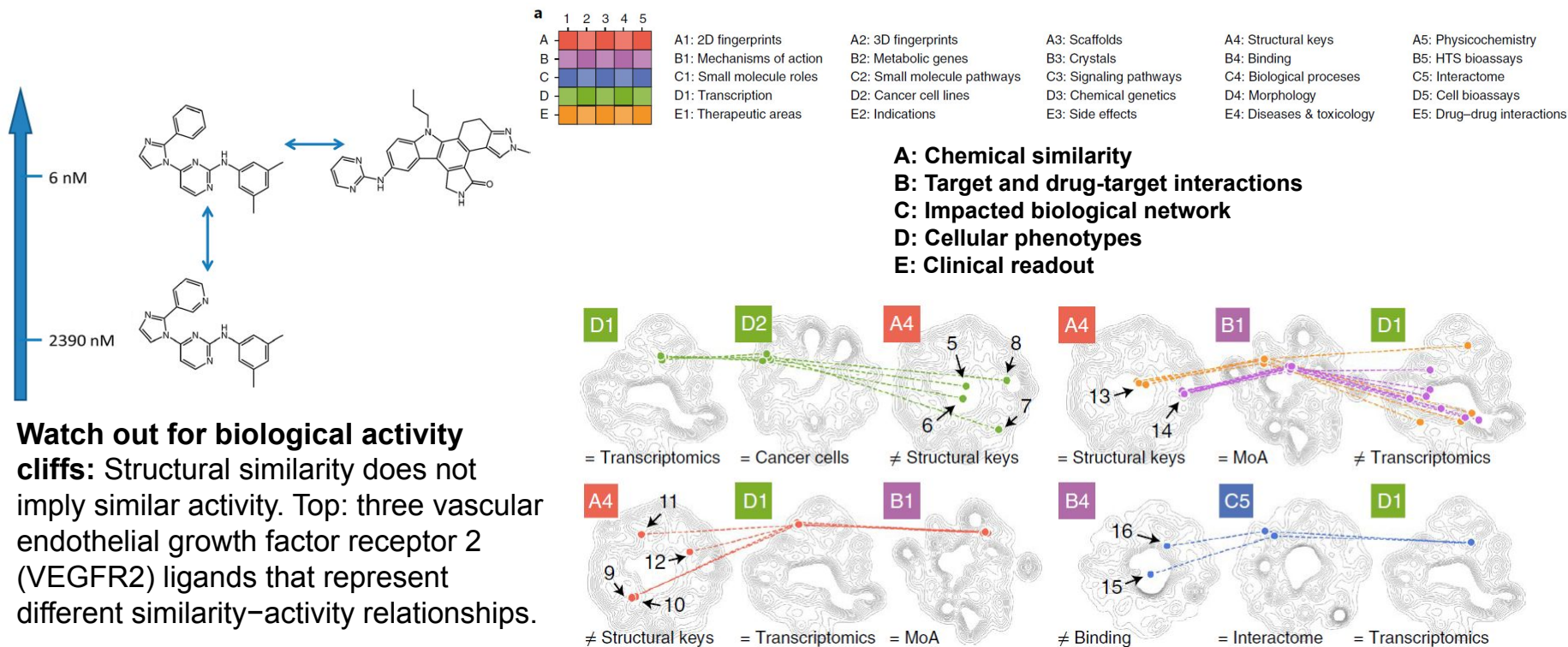
Watch out 1: Temporal validation is essential for drug discovery



(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.

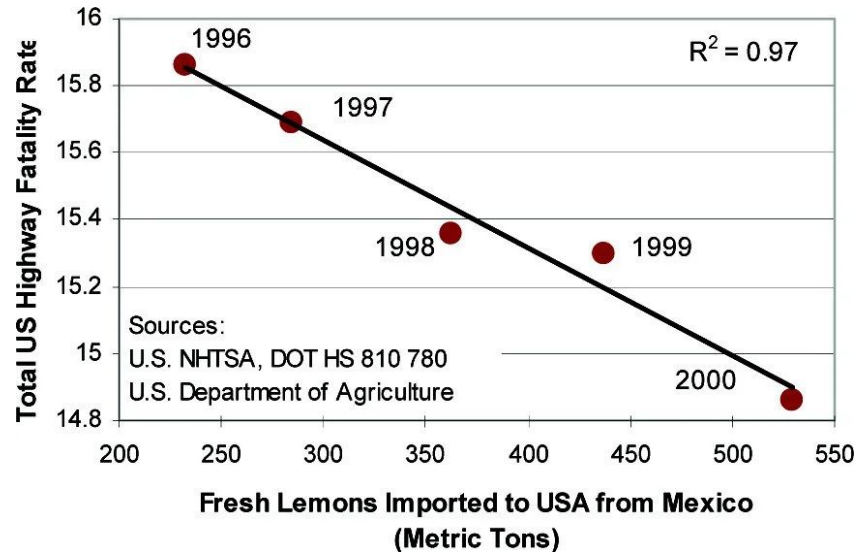
(Right) Temporal validation is especially important for drug discovery, because chemical structures used in the training set may differ substantially from those that will be tested.

Watch out 2: Molecular similarity does not equal biological similarity



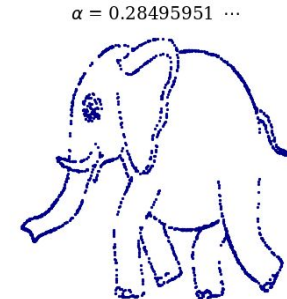
Duran-Frigola, Miquel, Eduardo Pauls, Oriol Guitart-Pla, Martino Bertoni, Víctor Alcalde, David Amat, Teresa Juan-Blanco, and Patrick Aloy. 2020. [“Extending the Small-Molecule Similarity Principle to All Levels of Biology with the Chemical Checker.”](#) Nature Biotechnology, May, 1–10.

Watch out 3: Do we need correlation or causation?



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>

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



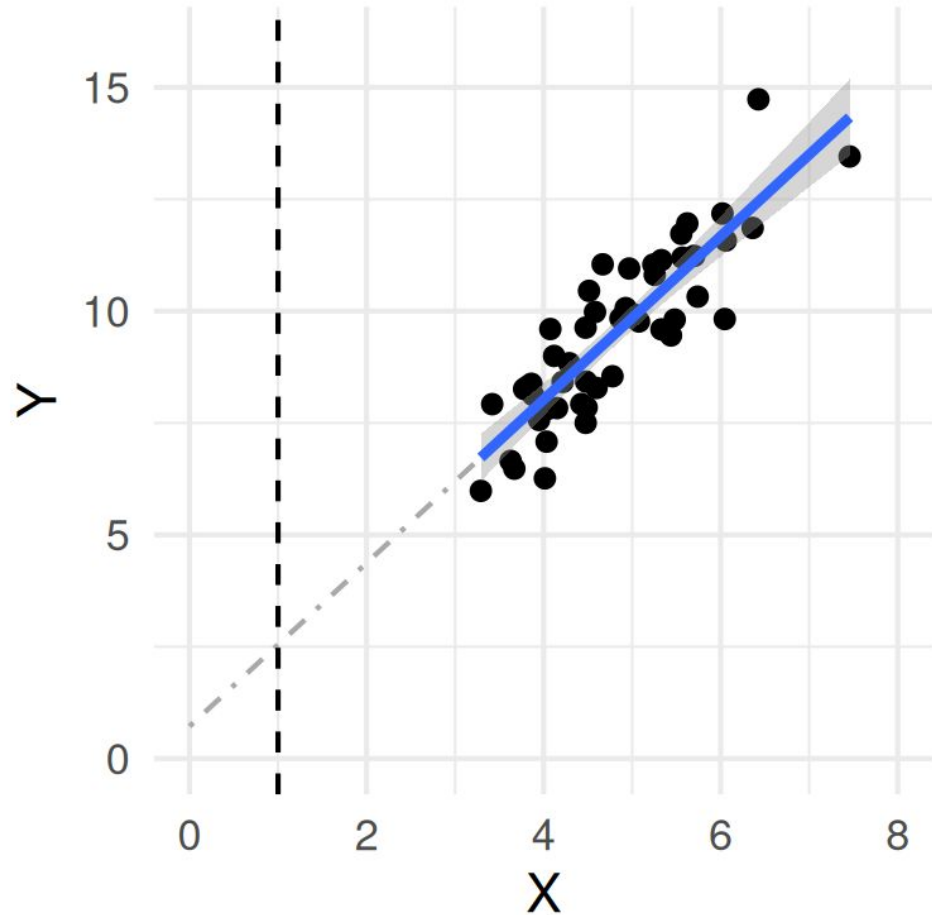
Boué, Laurent. "Real Numbers, Data Science and Chaos: How to Fit Any Dataset with a Single Parameter." *ArXiv:1904.12320 [Cs, Stat]*, April 28, 2019. <http://arxiv.org/abs/1904.12320>. [GitHub Repo](#).

Also see: [Drawing an elephant with four complex parameters](#)

We collected data from 50 independent observations, each consisting of paired measurements of X and Y . Assume that the relationship between X and Y is linear.

In a new observation, we manipulate the value of X by setting it to 1.

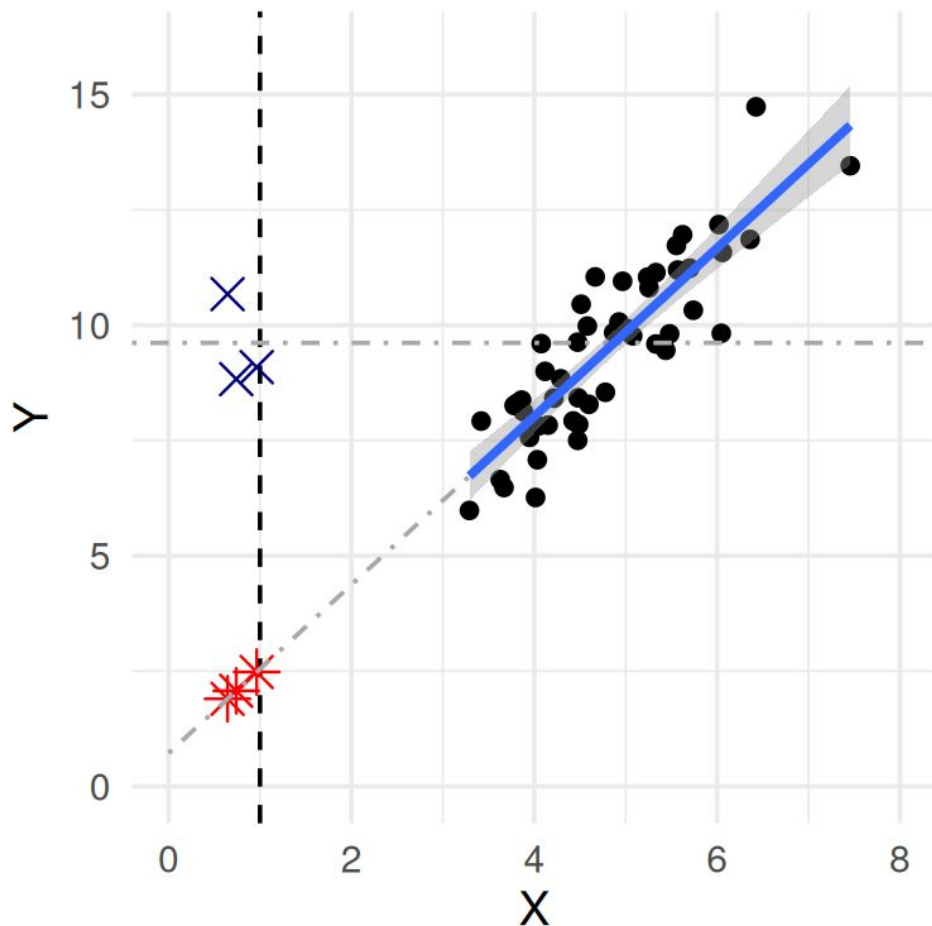
What value of Y do we expect?



Let's vote

We make three new observations where X is set to 1, the outcome would be:

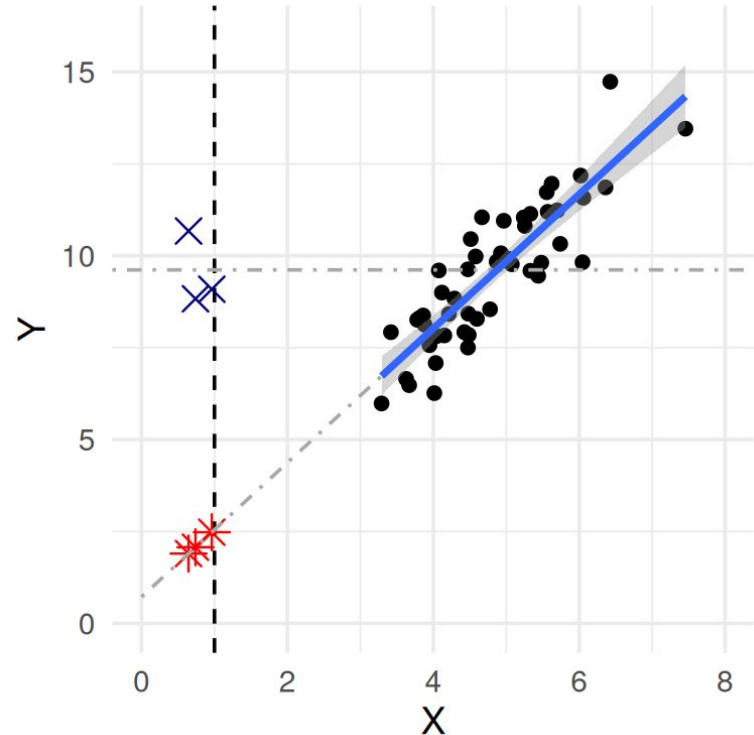
- (A) Red stars;
- (B) Blue crosses;
- (C) Both red stars and blue crosses are possible;
- (D) I am not sure.



Predicting intervention outcome requires causal understanding

- Scenario 1:
 - X: highest temperature minus 30°C
 - Y: incidence of heat stroke
 - Likely outcome of the intervention: red stars
- Scenario 2:
 - X: sales of ice-cream
 - Y: highest temperature minus 20°C
 - Likely outcome of the intervention: blue crosses
- Scenario 3:
 - X: sales of ice-cream
 - Y: incidence of drowning
 - Likely outcome of the intervention: blue crosses

Causation, reverse causation, and confounding all produce correlation. Predicting intervention outcome requires *causal understanding*, which data and statistical models alone cannot provide.

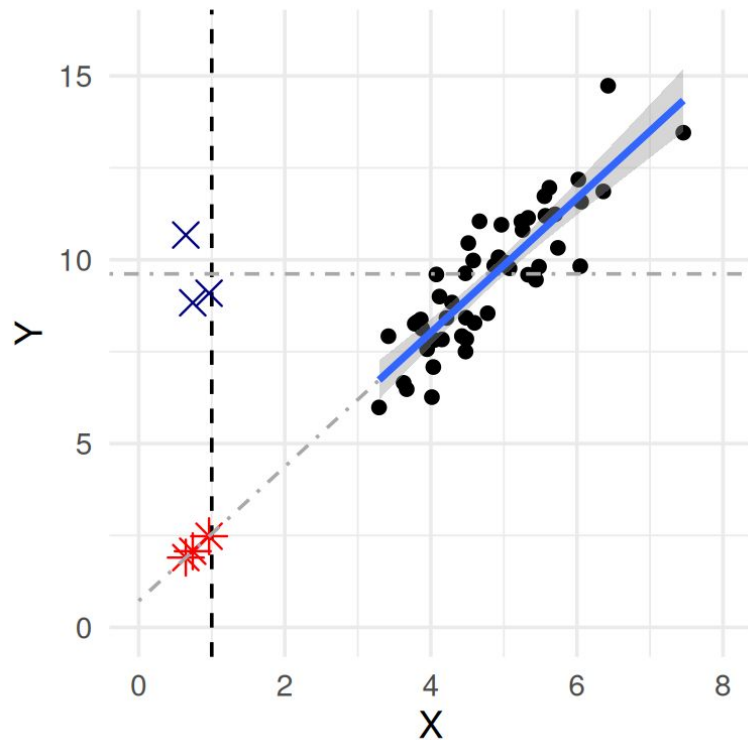


Causal models are critical when we intend to intervene

Imagine that X is the amount of a particular protein in the blood plasma of a patient with a disease, and Y is the disease severity of that patient.

A central question: *should we develop a drug to reduce the levels of this protein?*

We can use Directed Acyclic Graphs (DAGs) as causal models to identify causality.



Conclusions and outlook

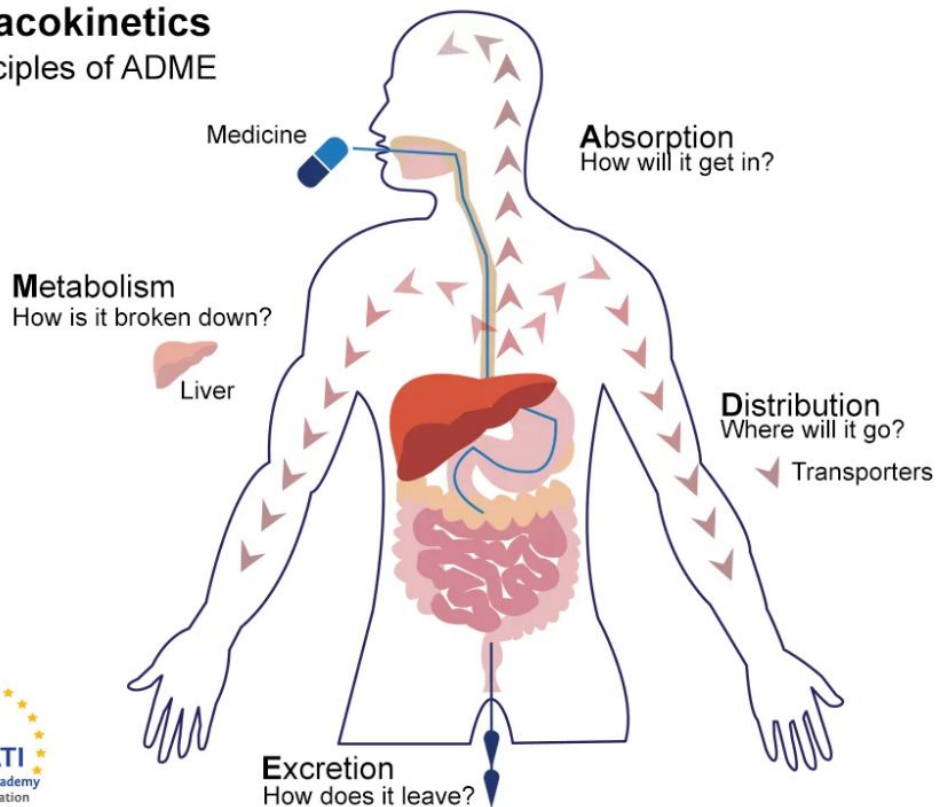
- **We learned examples of ODE-based mechanistic models, statistical/ML/AI models, and causal models.**
- **In the next lecture, we will look at examples of their applications, and how different methods can be combined to solve real-world problems.**

Backup slides

ODE-based models are powerful tools to model ADME/PK profiles of drugs

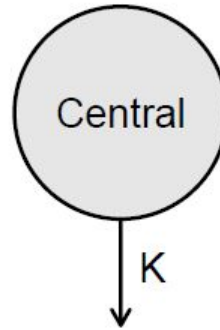
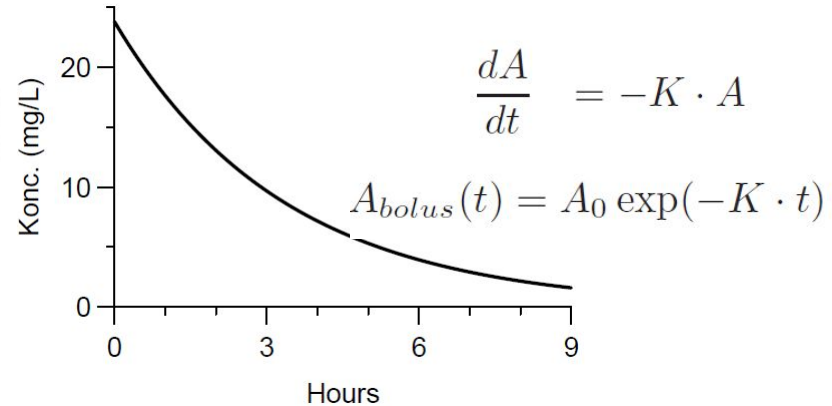
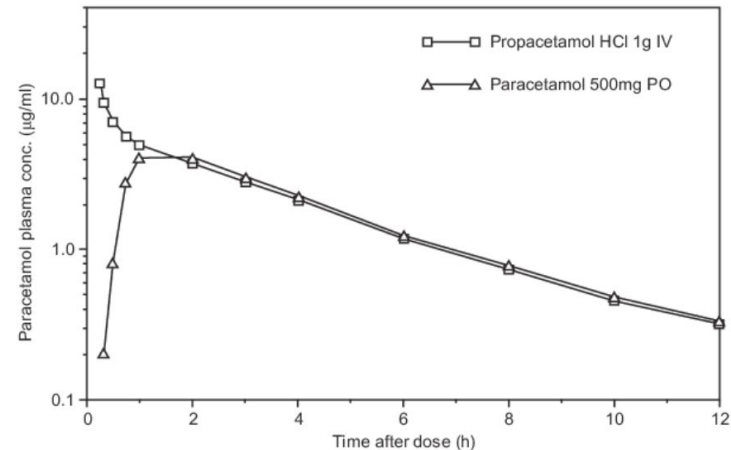
Pharmacokinetics

The principles of ADME



ODE-based mechanistic models are often used in pharmacokinetic modelling

- Pharmacokinetics (PK) describes how the drug is absorbed, distributed, metabolised, and excreted by the body.
- A basic mathematical model of PK is a compartment model, i.e. one or more ordinary differential equations that describe the relationship between drug concentration and time. The simplest model is the decay model of bolus (injection).
 - A_0 : initial concentration
 - $A(t)$: drug concentration at time t
 - K : rate of clearance
- A real-world example: PK of propacetamol, a pro-drug of paracetamol, given intravenously.



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 pushes coefficients towards zero. For

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

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8		
1	g_1	1		1	1				1	β_{ω_1}
	g_2									$>0 \beta_{\omega_2}$
1	g_3		1			1	1	1	1	$>0 \beta_{\omega_3}$
	g_4	1							1	$>0 \beta_{\omega_4}$
	g_5								1	β_{ω_5}
1	g_6	1					1			$>0 \beta_{\omega_6}$
	g_7		1	1	1				1	β_{ω_7}
	g_8	1				1				β_{ω_8}
1	g_9						1	1	1	

$f(\mathbf{Y})$: a binary label to indicate whether someone passes an exam

g_1 - g_9 : students

ω_1 - ω_8 : topics

β_{ω_1} - β_{ω_8} : coefficients of topics

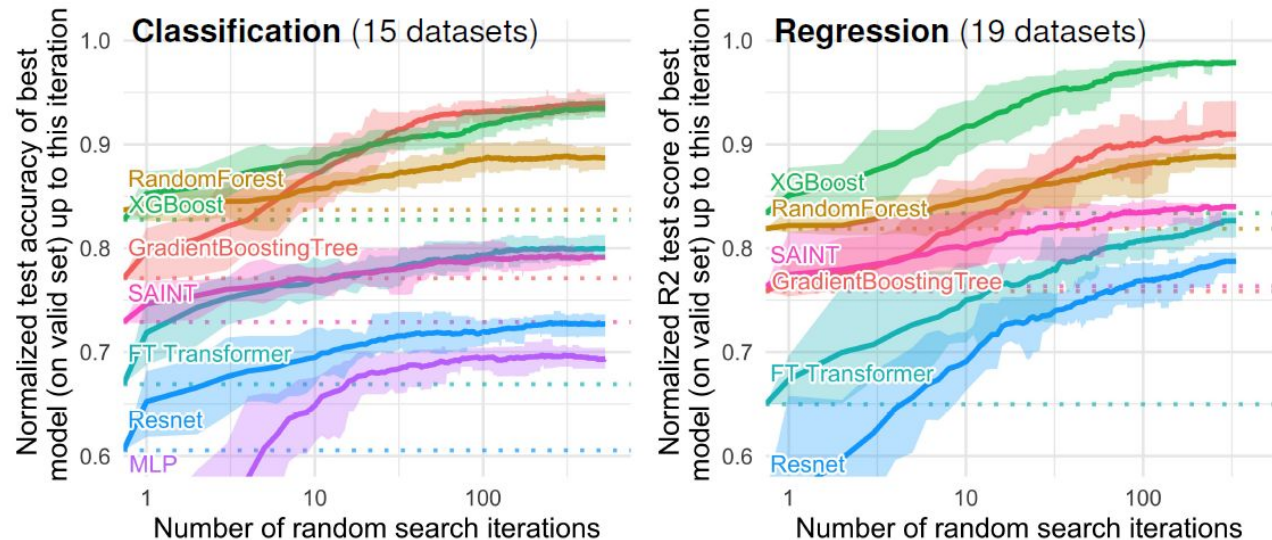
Equation: Bishop, Christopher M. [Pattern Recognition and Machine Learning](#), page 10

For tabular data: tree-based methods are well interpretable and generally outperform deep learning... until TabPFN (2022-25)

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, ... or TabPFN (next lecture)

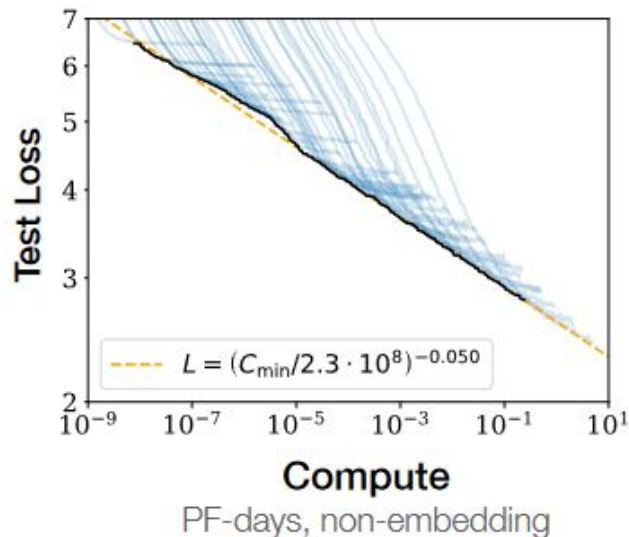
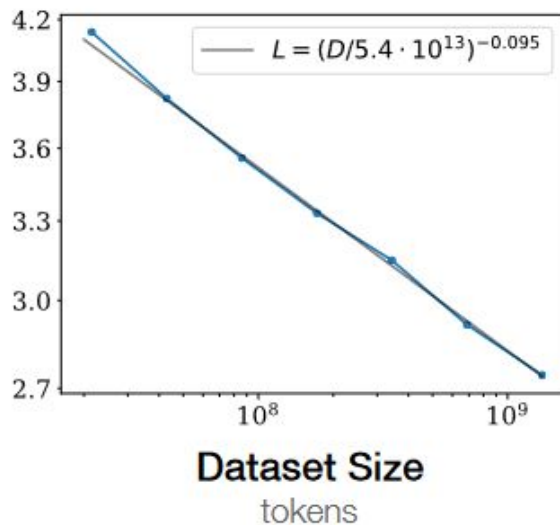
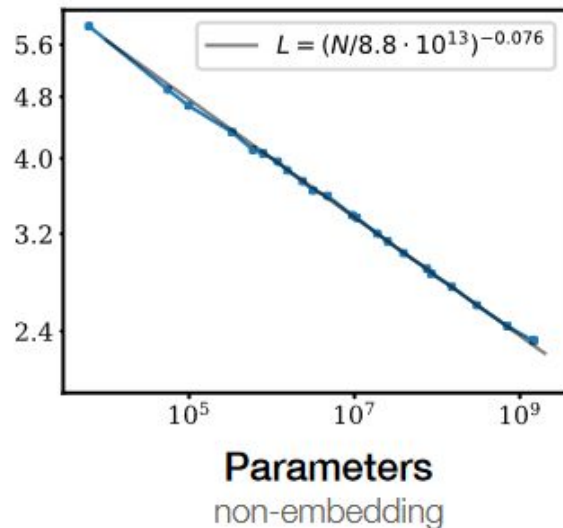


Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. "Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?" arXiv, July 18, 2022.

<https://doi.org/10.48550/arXiv.2207.08815>. GitHub repository:

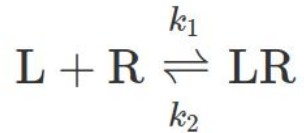
<https://github.com/LeoGrin/tabular-benchmark>

Performance of neural networks improves as data amount, computational power, and model size increase



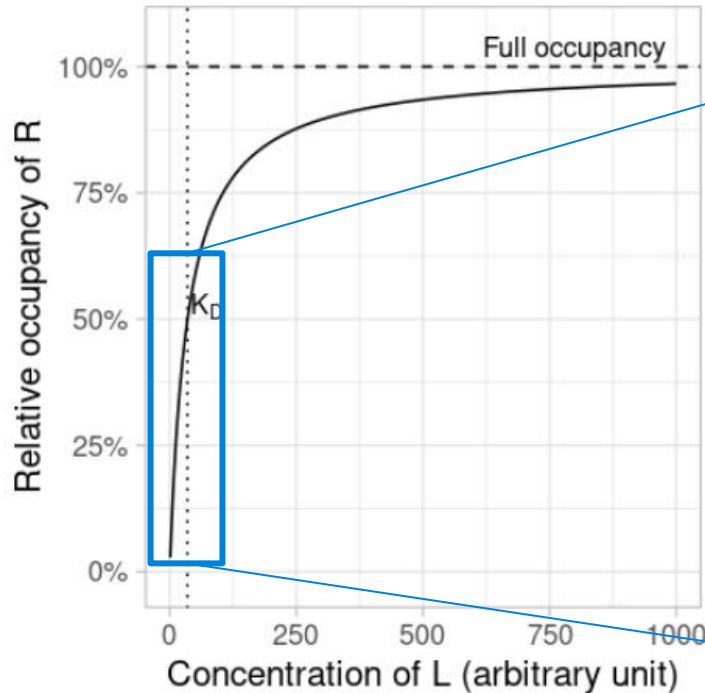
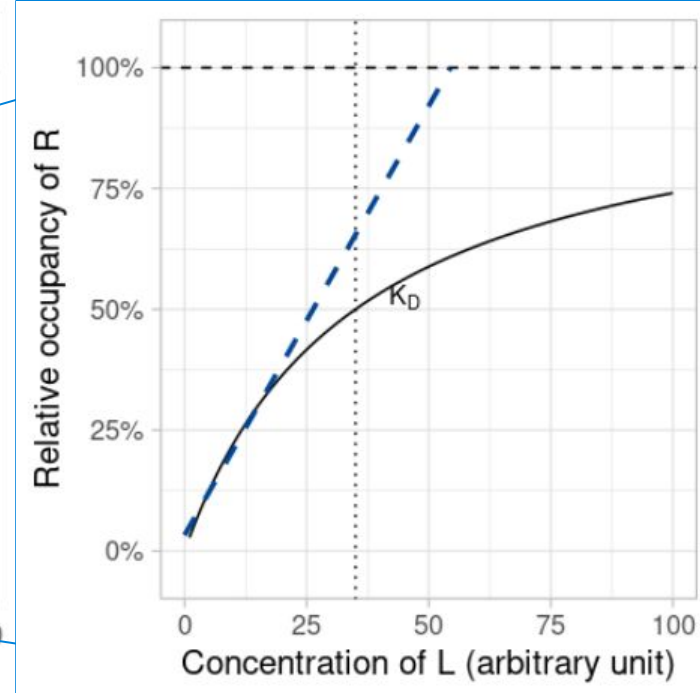
Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute used for training. PF-days: Peta-FLOPs. From: Kaplan, J. et al. [Scaling Laws for Neural Language Models](#) (2020).

Linear models can be used to model local effects of non-linear models

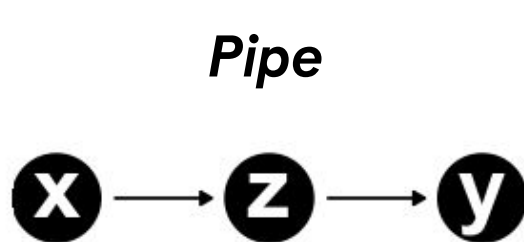


$$K_D \equiv k_2/k_1$$

$$[LR] = [R_{total}] \frac{[L]}{[L] + K_D}$$

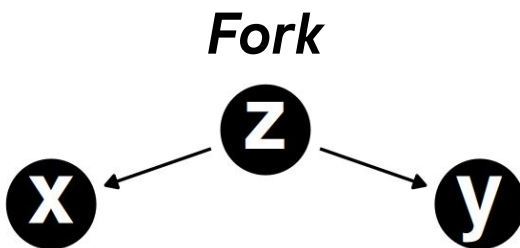


Common structures of Directed Acyclic Graphs (DAGs)



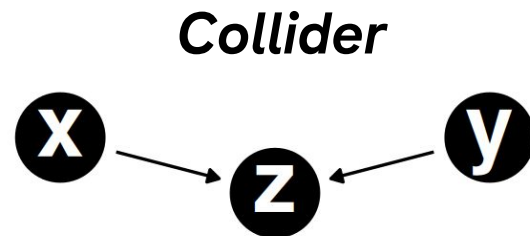
X: The quality of a movie
Z: Review of an influencer
Y: Follower watching the movie

X and Y are marginally correlated. Intervening Z changes Y (but not X), removing the correlation between X and Y.



Z: Dog ownership
X: Number of daily walks
Y: Poop bags used

X and Y are correlated. Intervening Z changes both X and Y. Within each subgroup of Z, X and Y are not correlated.



Z: Late arrival at work
X: Traffic congestion status
Y: **Amount of oversleeping**

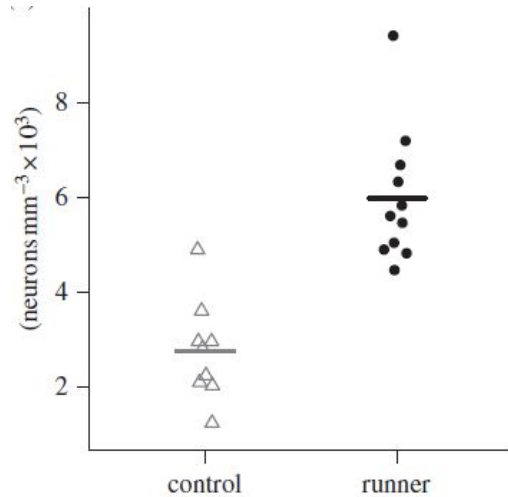
X and Y are not correlated. Intervening Z has no effect on X or Y. Within each subgroup of Z, X and Y are (anti-)correlated.

Causal structures and interventions can create or abolish correlations

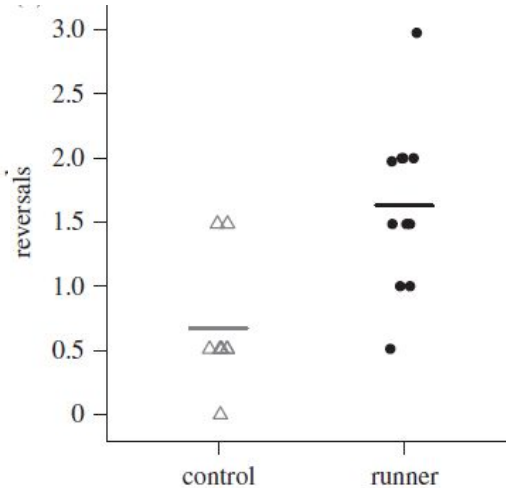
Case study: Running enhances spatial pattern separation in mice

Creer et al., PNAS 2010

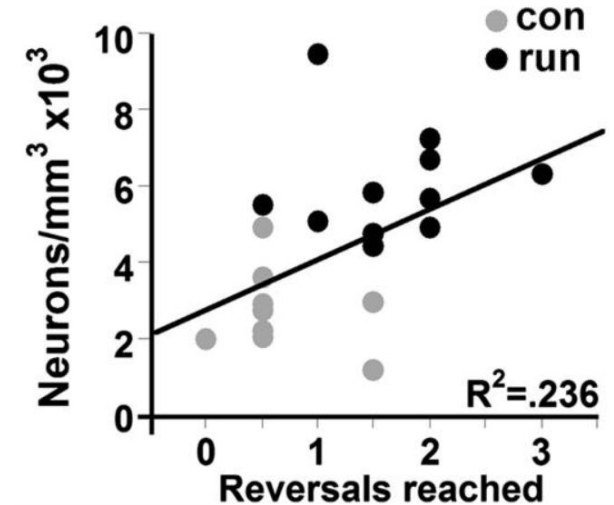
Neurogenesis in mouse hippocampus



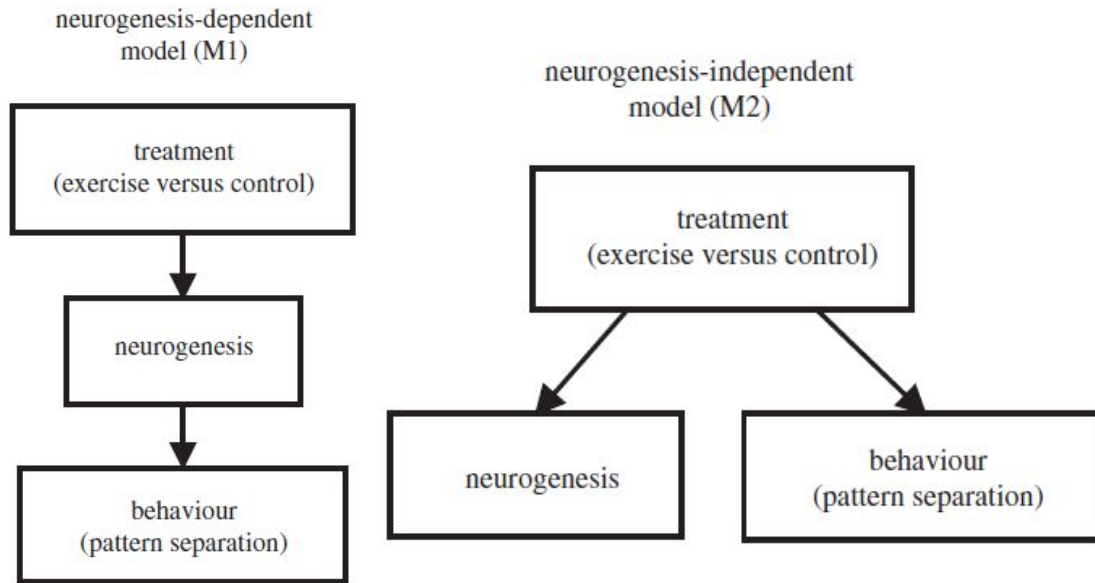
Ability to discriminate locations of adjacent stimuli



Correlation between the two measurements



The billion-dollar question: should we develop a drug enhancing neurogenesis to improve brain health?

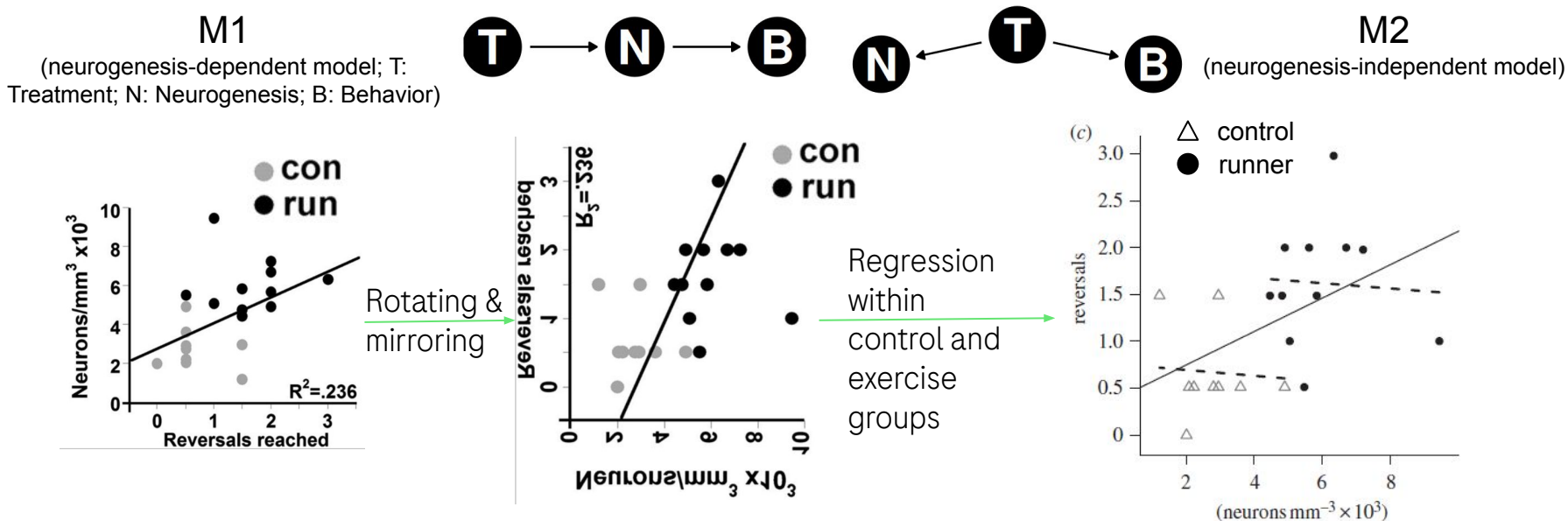


M1 (the **pipe** model) implies that neurogenesis and behavior would be *positively* correlated within each subgroup of treatment.

M2 (the **fork** model) implies that neurogenesis and behavior would not be correlated within each subgroup of treatment.

While M1 supports initiating a discovery program, M2 discourages us from doing so.

Causal thinking reduces cognitive bias by weighing models explicitly



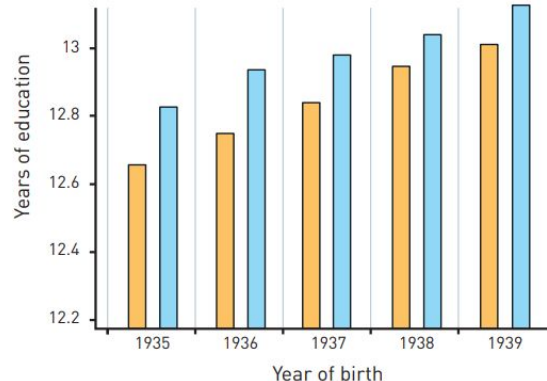
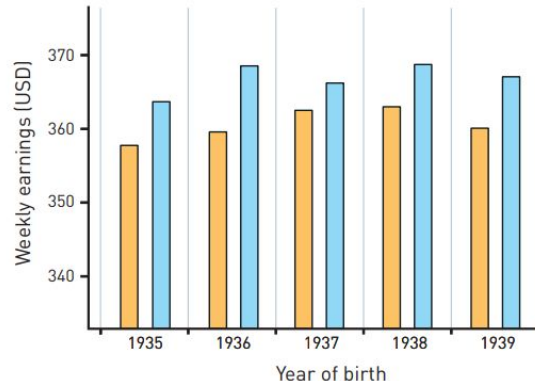
Neurogenesis (N) and behavior (B) are barely correlated (if anything, negatively) in both control and exercise groups, which refutes M1 (the pipe model) while strengthening M2 (the fork model). This piece of evidence speaks *against* starting a discovery program.

We note that several industrial projects aiming at promoting neurogenesis have been terminated.

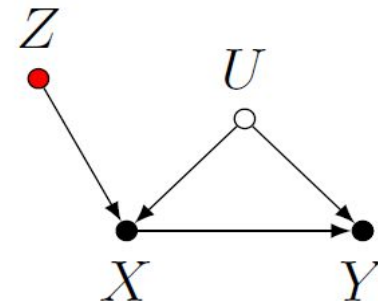
We can quantify causal relationships that are otherwise difficult to measure experimentally

In the US, children can choose to leave school when they turn 16 or 17. Therefore, children born early in the year can legally leave school earlier than those born late. Angrist and Krueger (1991) observed that people born late in the year have more years of education and higher incomes.

■ Born in first quarter ■ Born in fourth quarter



[Natural experiments help answer important questions](#); the Prize in Economic Sciences in Memory of Alfred Nobel, 2021, which was awarded to David Card, Joshua Angrist and Guido Imbens. J. Angrist and Alan B. Krueger. (1991), “[Does Compulsory School Attendance Affect Schooling and Earnings?](#),” The Quarterly Journal of Economics.



Z: Birthday (the *instrumental variable*)

X: Education

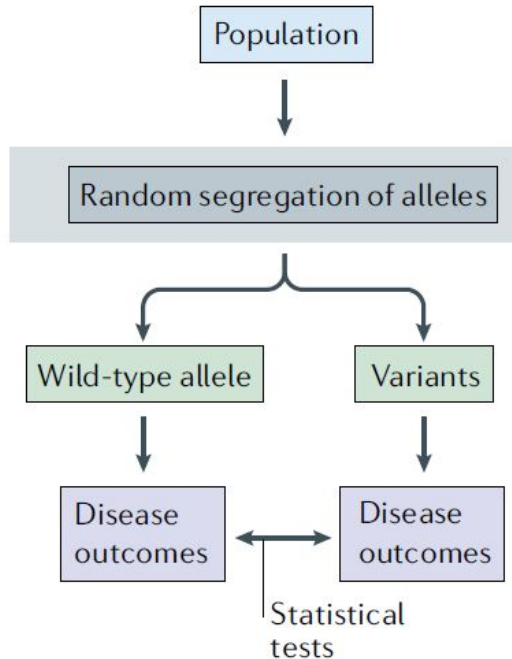
Y: Income

U: Socioeconomic/ individual factors

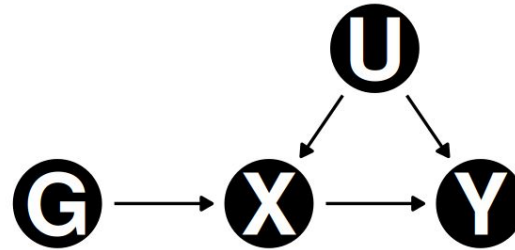
The causal effect is estimated by the ratio between the difference in earnings between late-born and early-born, and the difference in schooling (~9% per additional year of education).

Mendelian Randomization (MR) applies causal thinking to identify targets for drug discovery

Mendelian randomization



Each of us carries millions of genetic variants (SNPs) that are randomly assigned. We can use them as instrumental variables to test the causality between genes and diseases.



G: Genetic variant(s) associated with changes in X, the *instrumental variable*

X: Exposure, e.g. protein level

Y: Outcome, e.g. disease incidence/severity

U: Confounding factors, e.g. sex, diet, age

Example: Circulating FGF21 protein levels are correlated with symptoms of the liver disease MASH (metabolic dysfunction-associated steatohepatitis). MR confirms a causal relationship between FGF21 and several risk factors of MASH, supporting FGF21 as a potential therapeutic target.