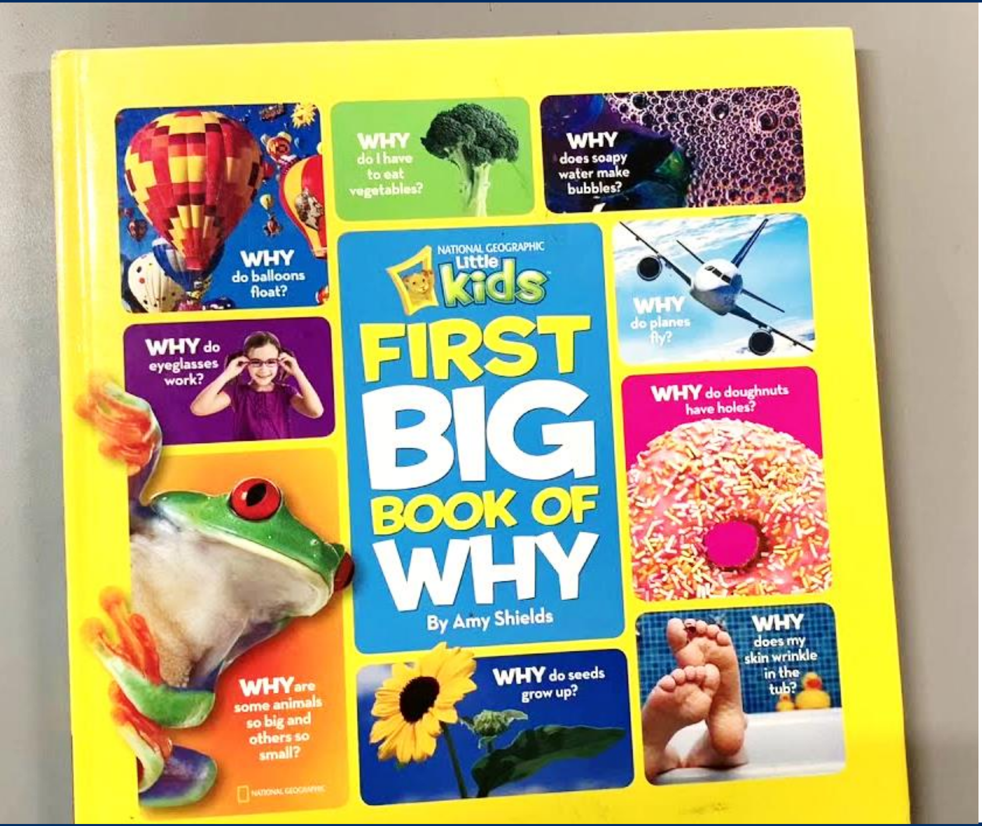
# Causality

## An introduction for ML practitioners

João F. Henriques
Visual Geometry Group

UNIVERSITY OF OXFORD

# Why?

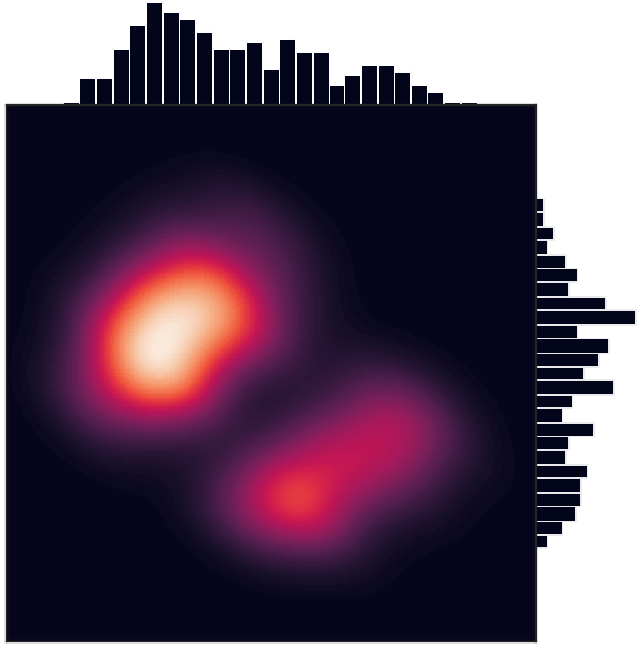



Cause and effect are integral to decision-making:

- What is the cause of these symptoms? → Decide between different treatments.
- Why is the cost of living so high? → Determine which economic intervention is more effective.
- Why did my computer break down? → Decide to update software or replace a part.

…

# Why?

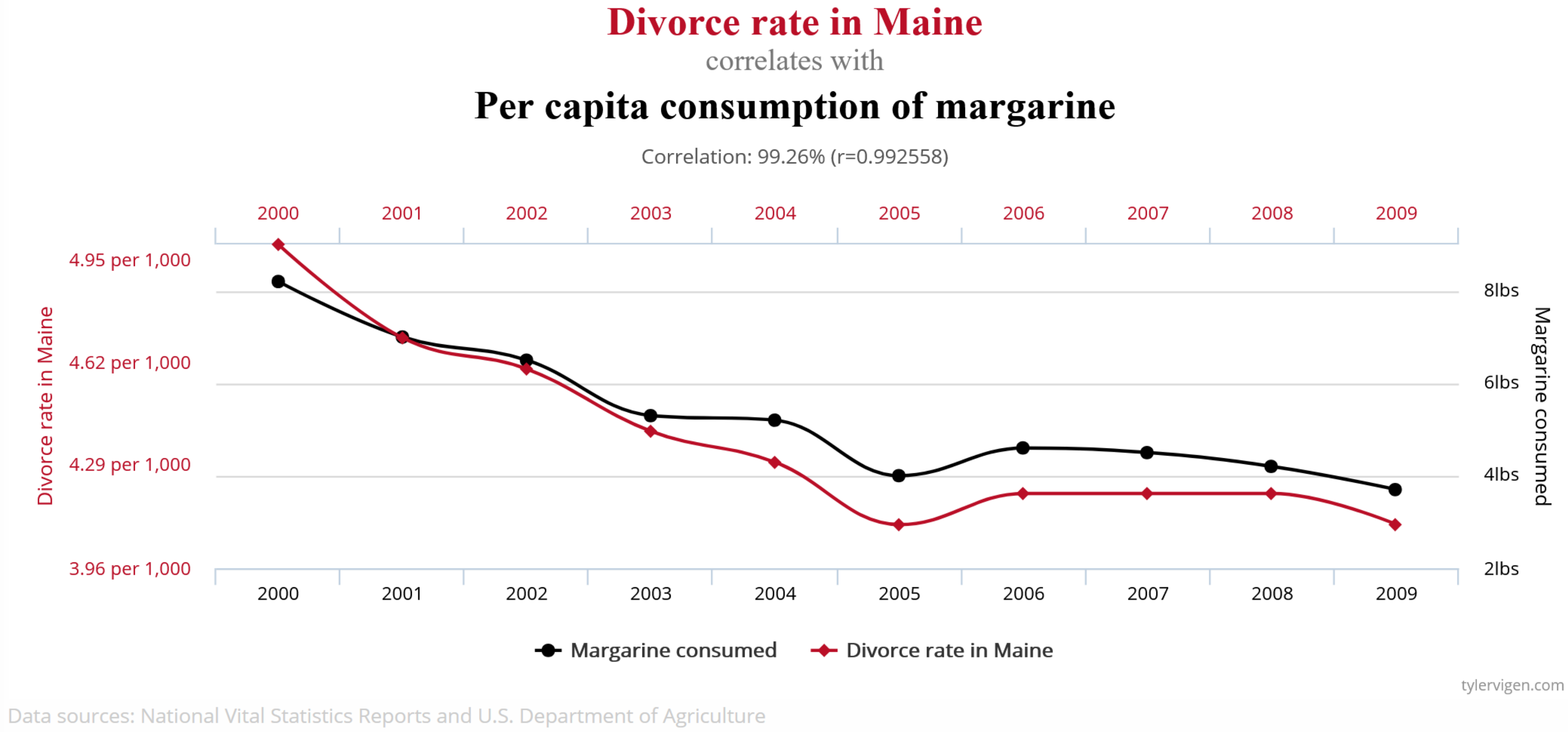


- Statistics and probability is usually concerned with the *description* of data, i.e. inference:

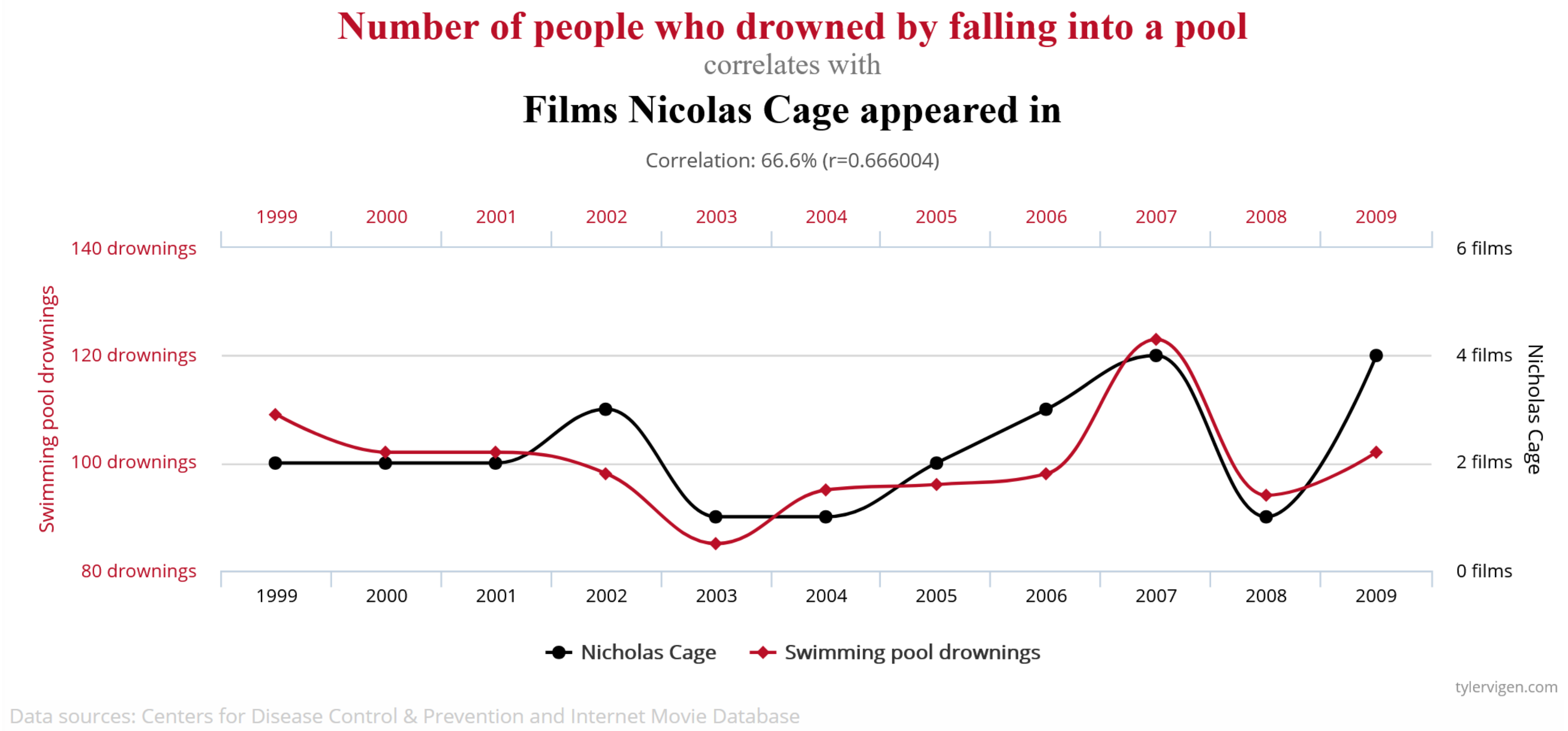
  ↪ Finding a parsimonious description of the **joint probability distribution** of data.

- Inference and hypothesis testing **cannot** answer the questions discussed in the previous slide – “**why**” and “**what if**” **–** used in decision-making.
- However, making **informed decisions** is often why we reach for statistics in the first place!

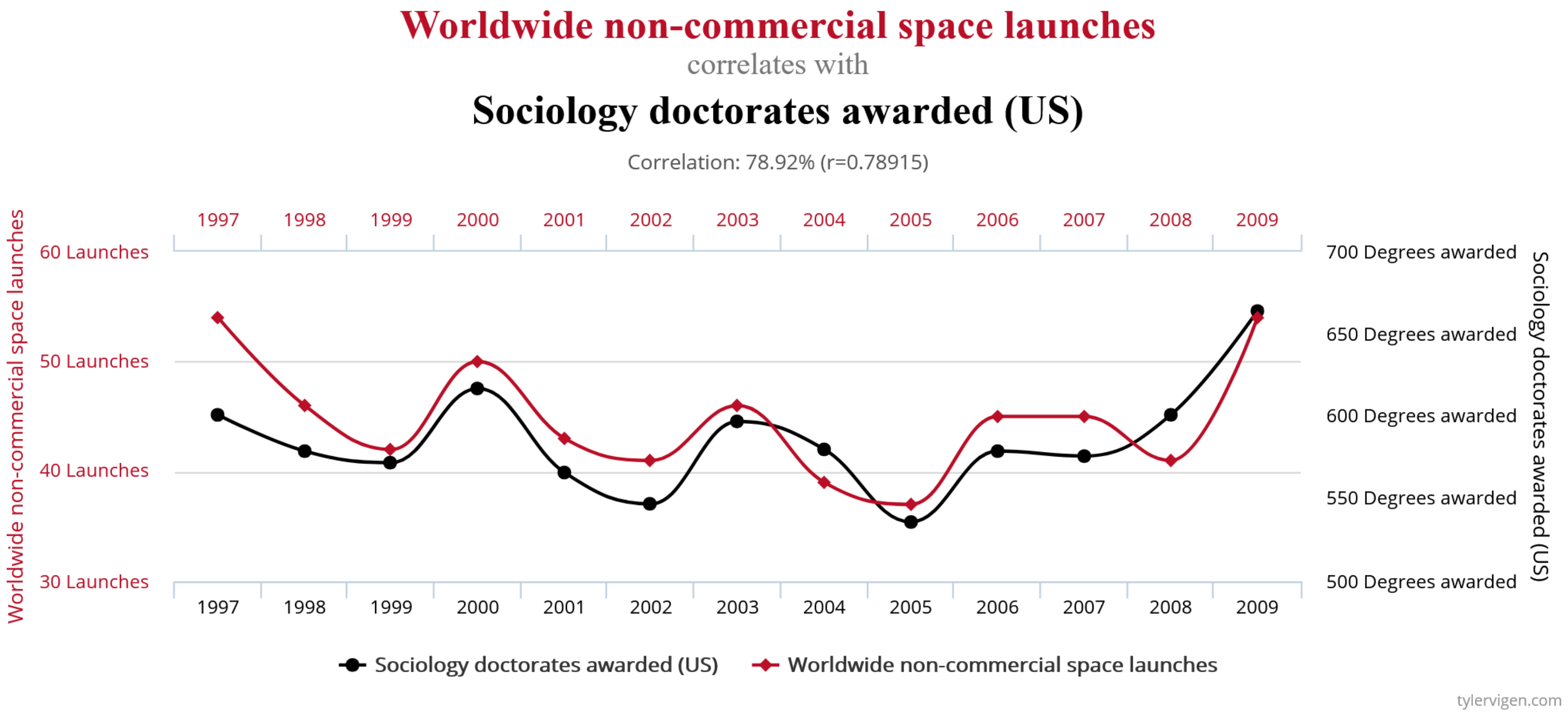
# Correlation ≠ causation



# Correlation ≠ causation

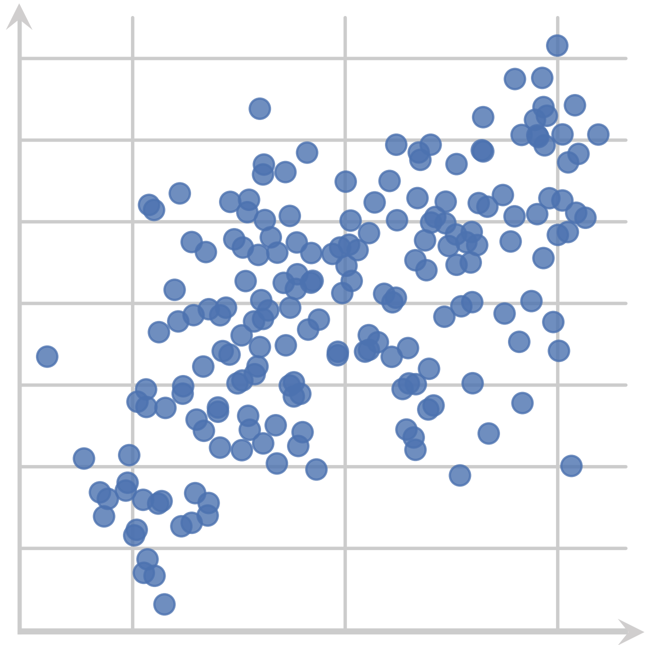


# Correlation ≠ causation



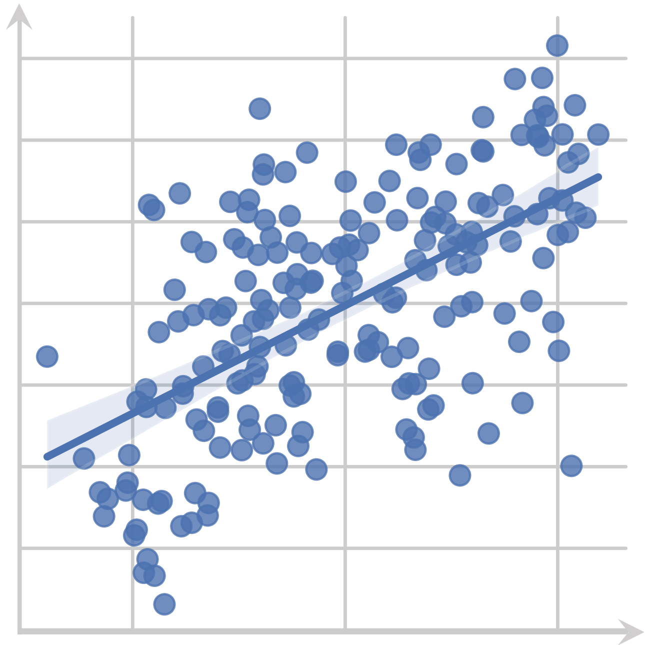

Data sources: Federal Aviation Administration and National Science Foundation

# Simpson's paradox



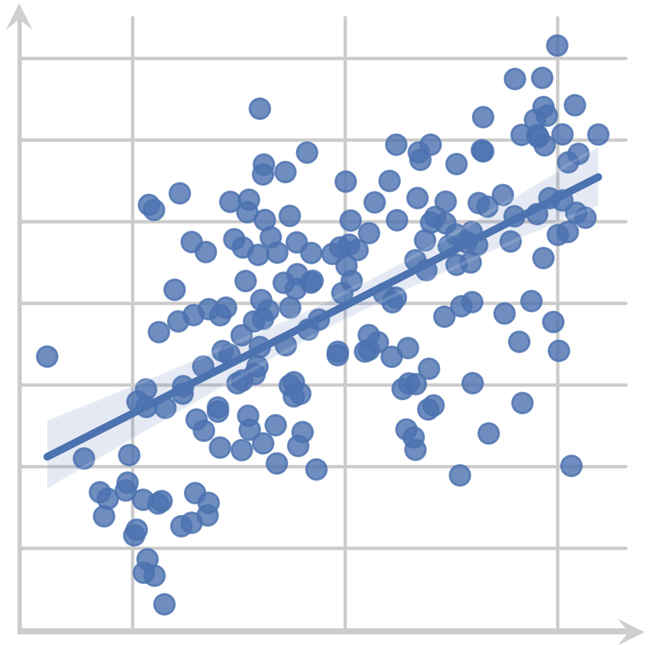
- A dataset of 2 variables, plotted as (X, Y).
- Collected to answer an empirical question (e.g. hypothesis in biology, performance of ML methods vs. some factors…).

# Simpson's paradox



- Both variables seem to be very correlated, with an "ok" linear fit.
- A statistical study could conclude that one variable *may* cause the other (with the usual caveat that correlation ≠ causation).
- Trust the data! Are we done?

# Simpson’s paradox



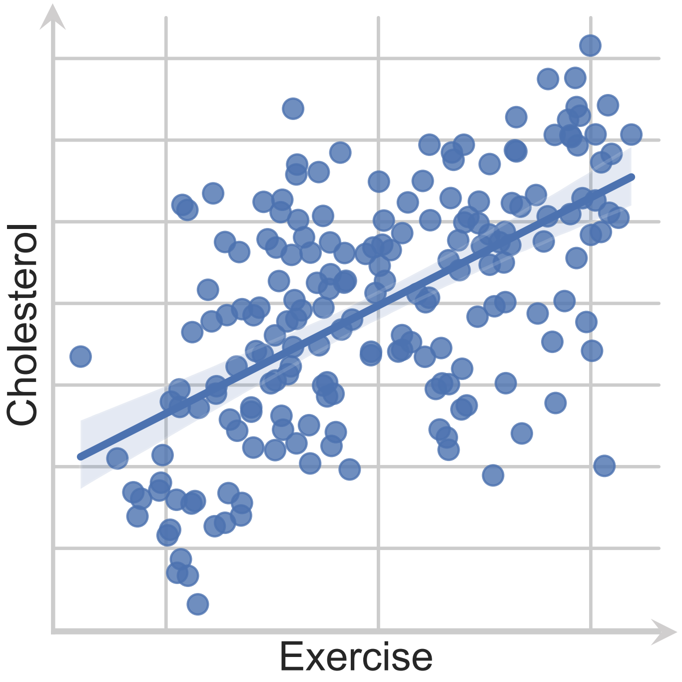
- What if we know more about the domain *(not encoded in the statistics)* that leads us to doubt this conclusion?

Example:

- X axis: hours of weekly exercise.
- Y axis: amount of cholesterol in the blood.

Implies that **more exercise leads to higher cholesterol** (or vice-versa)?

# Simpson's paradox

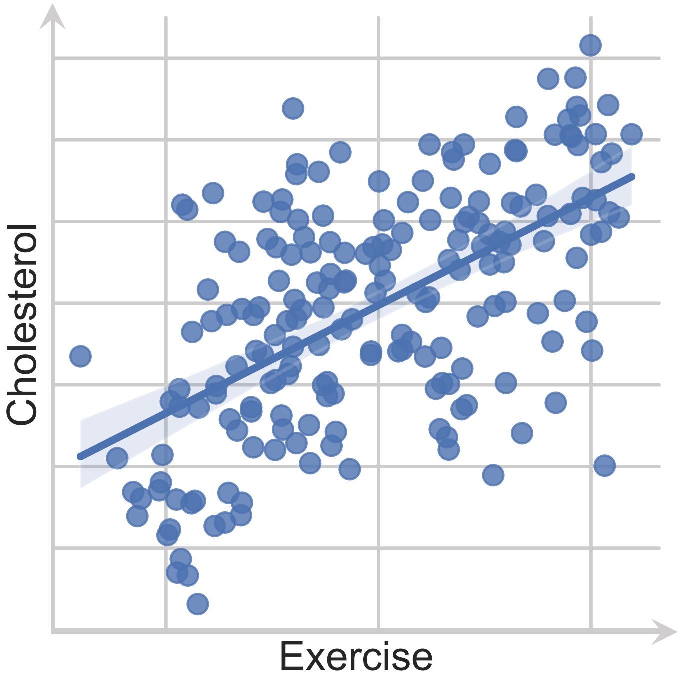



- Don't cancel your gym membership yet.
- What if we group (segregate) the *same* data according to a person's age?
- Also referred to as *conditioning* on that variable.

# Simpson's paradox


- Don't cancel your gym membership yet.
- What if we group (segregate) the *same* data according to a person's age?
- Also referred to as *conditioning* on that variable.

# Simpson's paradox



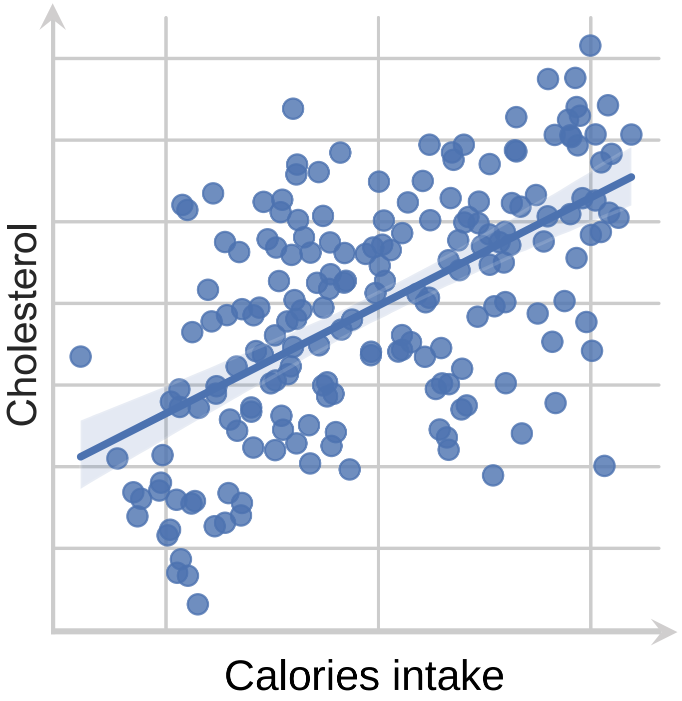

- Suddenly there is a *negative* correlation in each age group.

Explanation:

- As people get older, they tend to exercise more **and** have higher cholesterol (regardless of exercise).
- However, comparing people of the same age, more exercise does lead to lower cholesterol – as expected.

*Paradox:* Ignoring age, the correlation is the *opposite* of what we expect (exercise leads to high cholesterol).

# Simpson's paradox



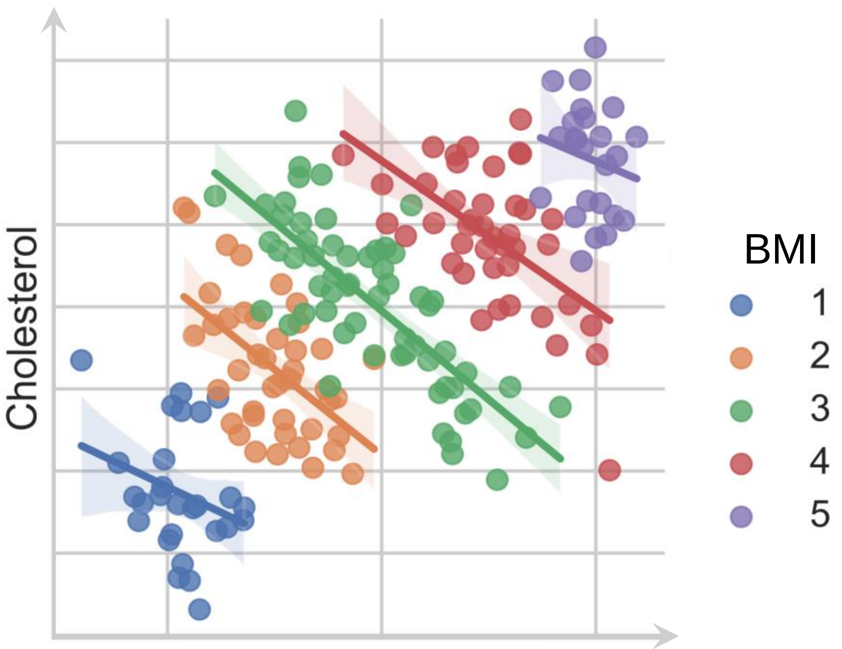

- Does this mean that segregated data always gives us the correct answer?

**No.** Let's imagine that the axes are relabeled:

- X axis: calories per day.
- Y axis: blood cholesterol (as before).

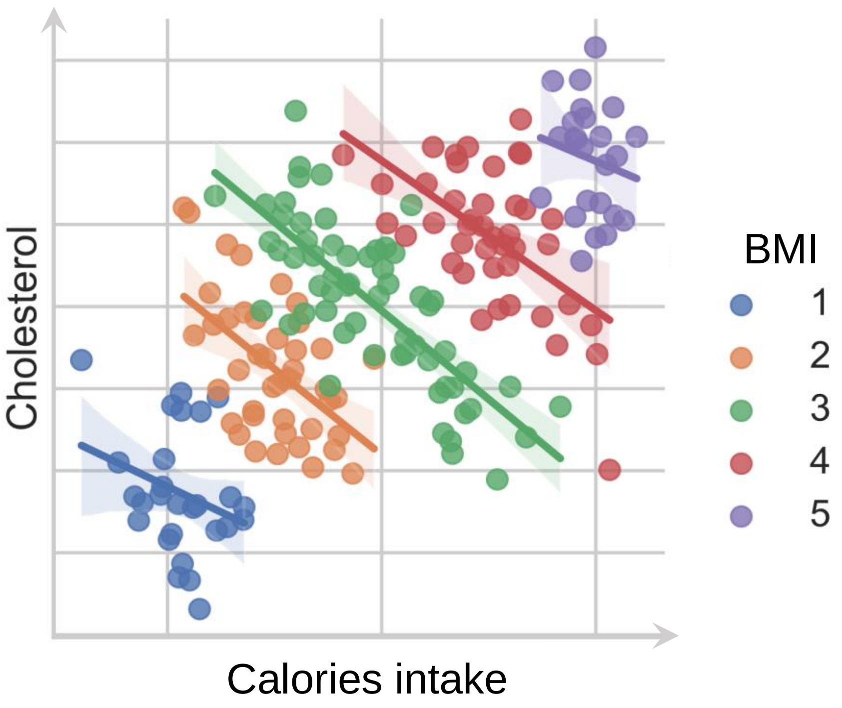
Now, common sense says that there *should* be a positive correlation (eating more food in general leads to higher cholesterol).

# Simpson's paradox




- Now group the data by body mass index (BMI, an indicator of weight relative to height).
- There is now a *negative* correlation in each group!
- Should the conclusion be that more calories leads to lower cholesterol? (Would be true if grouping/conditioning is always the answer, as before.)

# Simpson's paradox

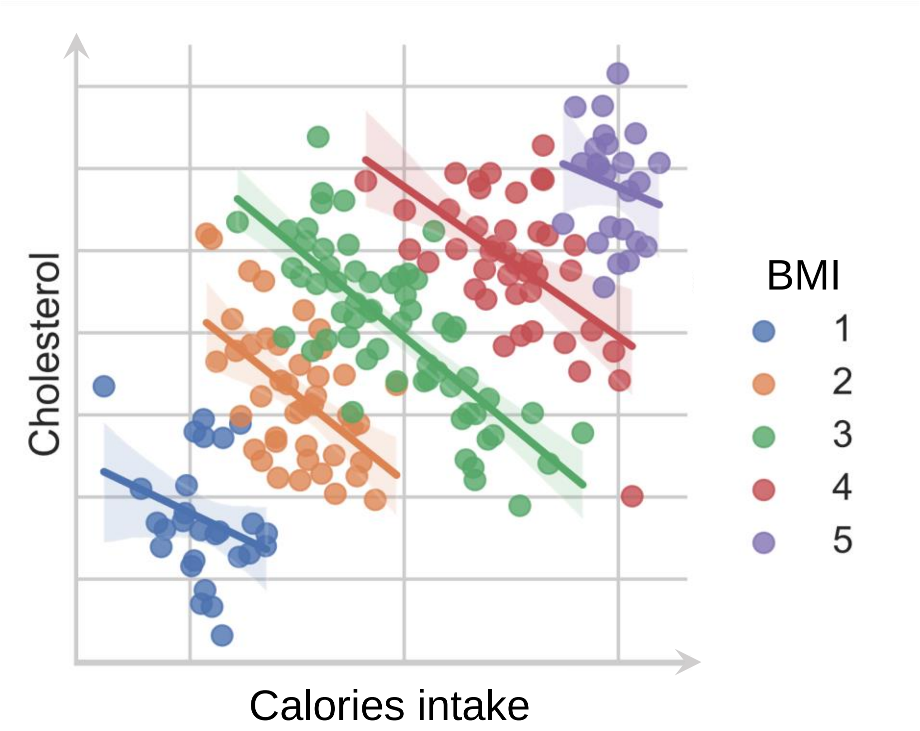



Plausible explanation: *(*I'm not a physician)*

- Increasing calories increases BMI, which in turn is associated with higher cholesterol.
- For a *fixed BMI group*, if the calories intake increases, there must be a corresponding energy expenditure (exercise), otherwise we would move to a higher BMI group. And this exercise lowers cholesterol.

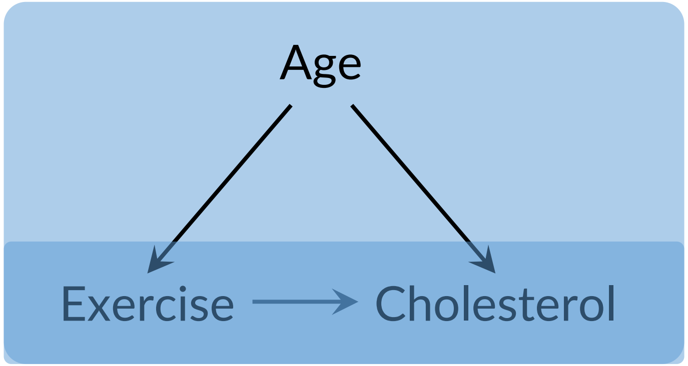
In other words, since BMI is **part of the mechanism** that affects cholesterol, it does **not** make sense to segregate data by it.

# Simpson’s paradox



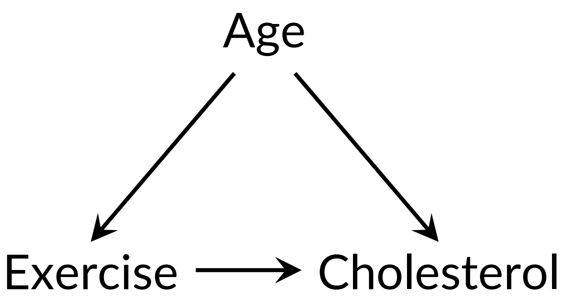

- **Simpson’s paradox** describes situations where grouping reverses correlations *arbitrarily* – and standard statistics cannot tell us which one is correct!
- We needed *domain knowledge* to navigate these examples correctly.
- How can we hope to make correct decisions in much more complex situations (many variables, complex dependencies)?
- We need a *language* to describe this domain knowledge and be systematic.

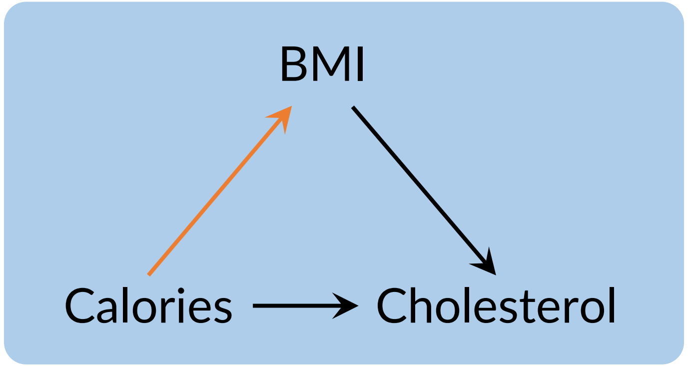
# Graphs of dependencies – intuition




- "Age" influences both other variables.
- In other words, it is a **confounder** when studying exercise vs. cholesterol.
- **Ignoring** age may induce a spurious correlation between them.
- **Conditioning** on age (splitting the data by age group) fixes it.

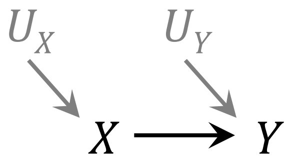
# Graphs of dependencies – intuition




- In the other example, BMI influences cholesterol but is influenced by calories.
- **Conditioning** on BMI would not have the same effect as for age, as it's not a confounder (we will elaborate on this later).

# Structural Causal Model (SCM)



Let us be a bit more formal:

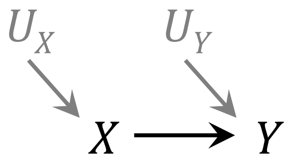
- Define a Structural Causal Model (**SCM**) as a graph.

  ↪ *Think of it like a "blueprint" or "story" of how the data is generated.*

- This graph has a node for each variable of interest (e.g. $X$, $Y$).
- It has a directed edge (arrow) $X \rightarrow Y$ if **knowledge** of $X$ is **necessary** to calculate $Y$.

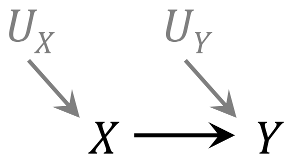
  ↪ In other words, for some function $f$ we have $Y = f(X, \ldots)$.

# Structural Causal Model (SCM)




- This is a Directed Acyclic Graph (DAG) – it has no cycles.
- If $X \rightarrow Y$, we say that $X$ is a **direct cause** of $Y$.
- If $X$ is an ancestor of $Y$, it is an (indirect) **cause** of $Y$.
- Conversely, if two variables are not ancestors of each other, they are **independent** (in the statistical sense).
  *Example:* $U_X$ and $U_Y$.

# Structural Causal Model (SCM)




We split the nodes into 2 sets:

- Exogenous variables $U = \{U_X, U_Y\}$.

  ↪ They are **external** to our model.

  ↪ Represent **unknow** factors: their values are modelled as **probability distributions**.

  ↪ They are also **independent** of each other.

- Endogenous variables $V = \{X, Y\}$.

  ↪ They are **internal** to our model.

  ↪ They are **deterministic**: knowing the values of all exogenous variables, we can set their values with perfect certainty.

# Structural Causal Model (SCM)

$$\begin{tikzcd}
U_X \arrow[dr] & & U_Y \arrow[dr] & \\
& X \arrow[rr] & & Y
\end{tikzcd}$$

*Why are SCMs useful?*

- We can predict patterns of **independence** solely from the graph.

  ↪ Says a lot about the model, even **before** deciding which functions to use or estimating any parameters.

- Predict effect of **interventions** ("what if") with just an SCM and data.

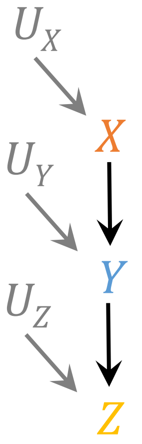
  ↪ Makes statistics "actionable": correctly supports decision-making.

- Use **independence** to help estimate parameters more efficiently:

$$\underbrace{P(A,B)} = \underbrace{P(A)}\ \underbrace{P(B)}$$

$$\begin{matrix}\#\text{parameters}\\ \#\text{samples}\end{matrix} \geq \begin{matrix}\#\text{parameters}\\ \#\text{samples}\end{matrix} + \begin{matrix}\#\text{parameters}\\ \#\text{samples}\end{matrix}$$

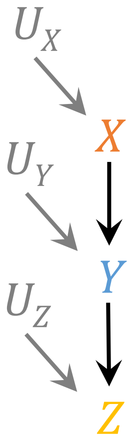
# Independence patterns – chains




- *Example SCM:*
  - School funding $X$, exam scores $Y$, college acceptance $Z$.
  - Exogenous (noise) variables $U$ (all independent).
- All pairs of variables are likely **dependent**:
  $X \rightarrow Y$, $Y \rightarrow Z$, $X \rightarrow Z$
- E.g. a change in $X$ in general results in a change in $Y$ (and indirectly in $Z$).

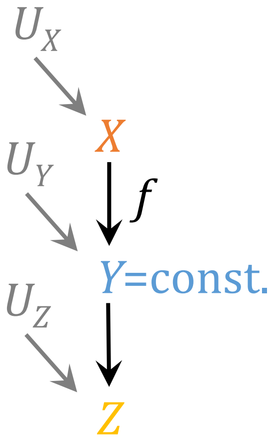
# Independence patterns – chains






- $X$ and $Z$ are **independent**, **conditional** on $Y$.
- What does it mean to condition on $Y$?
- It means we filter the data into groups, for each distinct value of $Y$ (e.g. $Y = a$, $Y = b$, …).
- In practice, this happens when we "know" that for a particular case $Y = a$, (e.g. an exam score was $a$) and we want to model the remaining probabilities.

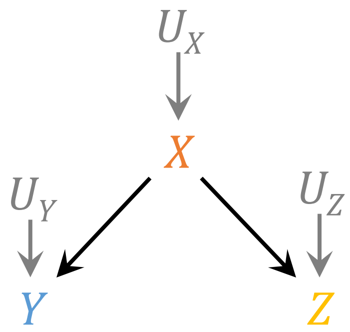
# Independence patterns – chains




- For a **fixed** value $Y = a$ (grouping the data), let us analyse the dependence between $X$ and $Z$.
- As we change $X$ to take on different values, then $U_Y$ will **have** to change to keep the **fixed** value $Y = a$.
  - *Example*: If $Y = f(X) + U_Y$, then in the group $Y = a$ we'd always observe noise set **exactly** to $U_Y = a - f(X)$.
- But $Z$ does not depend on $U_Y$, only on $U_Z$ and $Y$, so it is **independent** of $X$.

- Reiterating: $X$ and $Z$ are **independent**, conditional on $Y$.
- Applies to **any configuration** of variables that **contains a chain**.

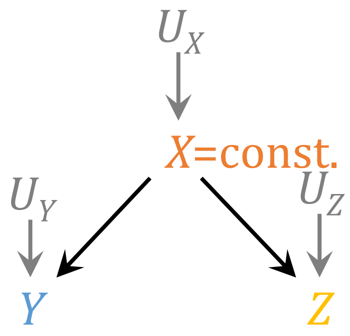
# Independence patterns – forks




- *Example SCM:*
  - Temperature $X$, ice cream sales $Y$, crime $Z$.
  - Exogenous (noise) variables $U$ (all independent).

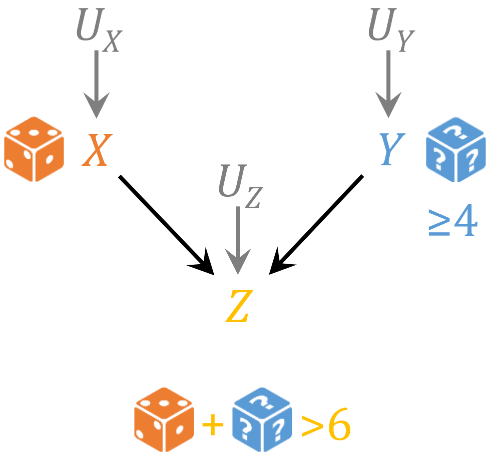
- All pairs of variables are likely **dependent**:
  $X \rightarrow Y$ ,  $X \rightarrow Z$,  $Y$ and $Z$
- $Y$ and $Z$ are dependent because, **some** of the time, **both** will change due to $X$.
- *Example:* Based on this SCM, we would expect correlations between ice cream sales and crime, though there is no direct causal connection between them.

# Independence patterns – forks




- $Y$ and $Z$ are **independent**, **conditional** on $X$.
- To understand this conditioning, again we filter the data into groups ($X = a$, $X = b$, …).
- Now we're comparing only cases where $X$ is constant.
- Since $X$ is constant, $Y$ and $Z$ only depend on their respective noise variables $U$, which are independent.
- So $Y$ and $Z$ are **independent** too.
- This conditional independence applies to **any** configuration of variables with a "common cause" like $X$ (called a "fork").

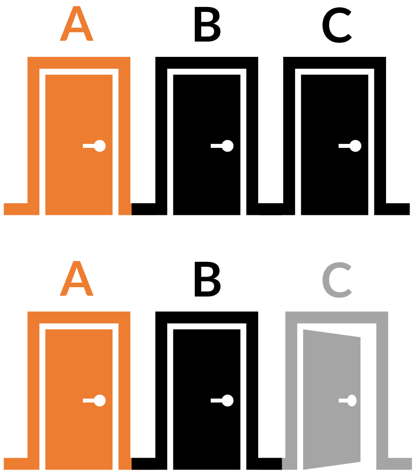
# Independence patterns – colliders




- *Example SCM:*
  - 1st dice $X$, 2nd dice $Y$, sum of dice results > 6 (Boolean) $Z$.
  - Exogenous (noise) variables $U$ (all independent).

- Some pairs of variables are likely **dependent**:
  $X \rightarrow Z$ , $Y \rightarrow Z$
- $X$ and $Y$ are **independent**, **unconditionally**.
- $X$ and $Y$ are **dependent**, **conditionally** on $Z$.
- *Example:* If we know that the first dice result ($X$) is 3, and that the sum is larger than 6 ($Z$), then the second dice result ($Y$) must be 4 or more – the dice results **become dependent**.

# Trouble with colliders



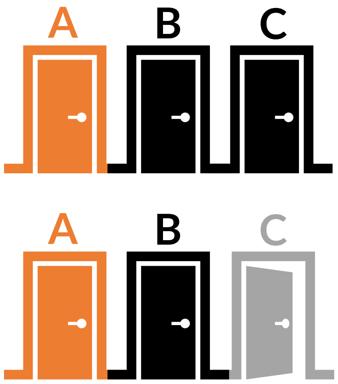

Colliders are very common and create very **unintuitive** results in conditional probabilities.

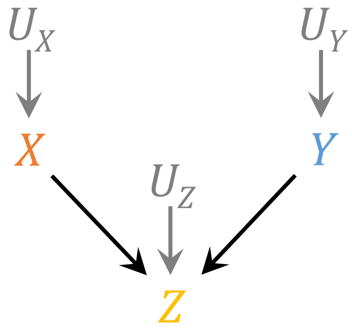
One such example is the "Monty Hall problem":

- There are 3 doors, 1 has a prize behind it (equal probabilities).
- You choose door A.
- Another door, door C, is shown to have no prize behind it.
- Is it better to switch to **door B**, or keep the choice of door A?
  - *Naïve answer*: Both are the same (1/2 chance of winning).
  - *Real answer*: There is a **2/3** chance of winning by switching to **door B**, and **1/3** by staying with door A – so you have a better chance of winning if you switch to **door B**!

# Trouble with colliders





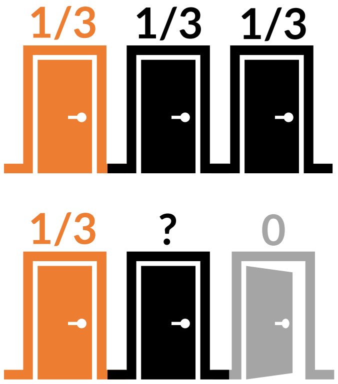



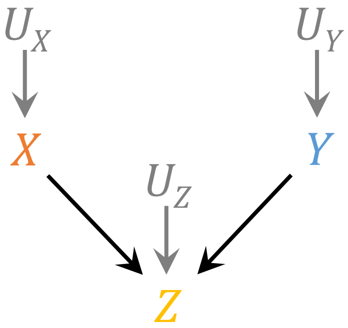

- Why? Let us write it as a SCM:
  - $X$ is your initial choice of door.
  - $Y$ is the door with the prize.
  - $Z$ is the door shown to you with no prize.

    ↪ $Z$ is causally dependent on $X$, and on $Y$. (The door shown must not have the prize, and not be the one you chose.)
- Since this SCM is a **collider**, conditioning on $Z$ (i.e. **knowing** which other door has no prize) makes $X$ and $Y$ **dependent**.

  ↪ This challenges our expectation: Your choice of door ($X$) and the door with the prize ($Y$) were picked **independently**, so you may expect them to remain independent!
  - Realizing that this is a collider helps us avoid this mistake.

# Trouble with colliders

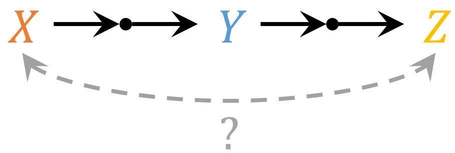


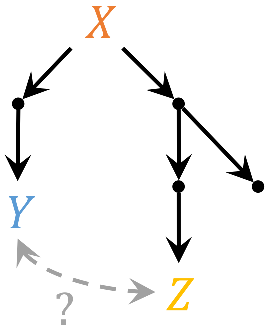



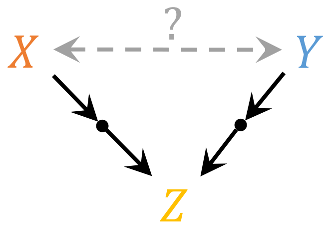

Just to get a quantitative answer to the Monty Hall problem:

- The door you chose has $1/3$ probability of having the prize.
- So when another door is opened showing no prize, your door still has $1/3$ probability of having the prize.
- Since the opened door has $0$ probability of having the prize, and the total probability must sum to $1$, the probability of the **remaining door** must be $1 - 1/3 = 2/3$, making it the best option.

# Generalizing to larger graphs




**Chain** with longer **paths** between the nodes:

- $X$ and $Z$ are **dependent**, unconditionally.
- $X$ and $Z$ are **independent**, conditional on $Y$.

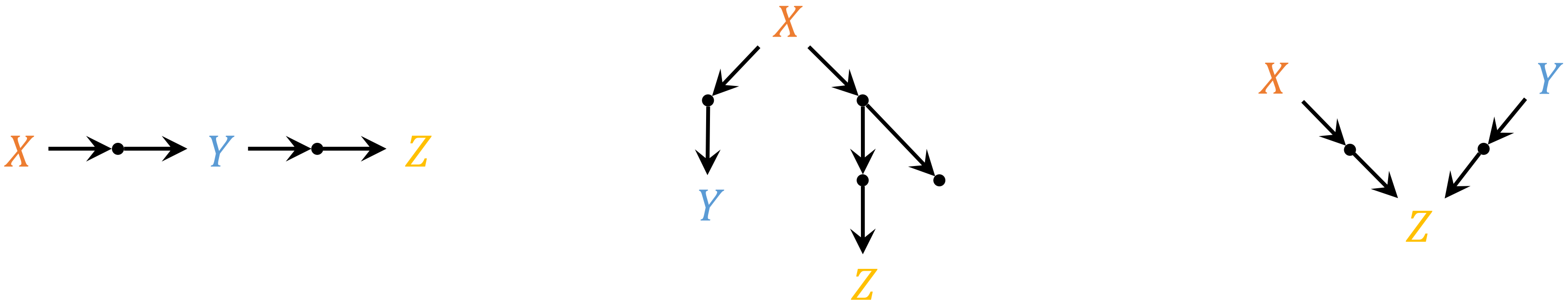
**Fork** where $Z$ and $Y$ are **descendants** of $X$:

- $Y$ and $Z$ are **dependent**, unconditionally.
- $Y$ and $Z$ are **independent**, conditional on $X$.

**Collider** with a **single path** between $X$, $Z$ and $Y$:

- $X$ and $Y$ are **independent**, unconditionally.
- $X$ and $Y$ are **dependent**, conditional on $Z$.

# *d*-separation

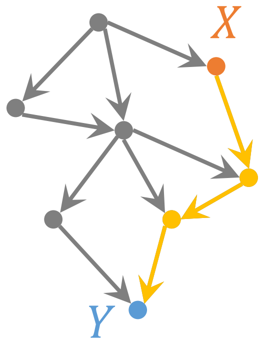



- Can we find dependence/independence patterns more systematically?

  ↪ Yes: We can check for ***d*-separation**.

- "*d*" stands for *directional*: two nodes are "separated" in a graph, taking the edges' directions into account.
- You can think about it as ***d*-separated** = **independent**.

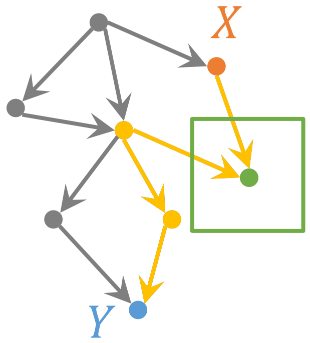
# *d*-separation – unconditional




- Consider an arbitrary graph with two nodes $X$ and $Y$.
- To know whether the two nodes are *d*-separated, consider **all paths** between them (ignoring edge direction).
- If there is any unblocked path between them, they are ***d*-connected** (i.e. not *d*-separated).

  ↪ Then $X$ and $Y$ are likely to be **dependent**.

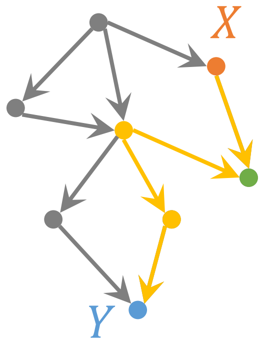
# *d*-separation – unconditional




- Let us delete one edge to see an example of a **blocked path.**
- The highlighted path connects $X$ and $Y$.
- But it is **blocked** because it goes through a collider node (i.e. that node has two incoming edges in the path).
- All paths between $X$ and $Y$ are blocked, so they are ***d*-separated.**

  $\hookrightarrow$ $X$ and $Y$ are **independent.**

# *d*-separation – unconditional




Analogy:

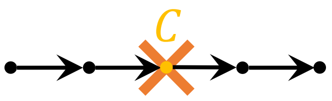
- Think about dependence between two nodes like water flowing.
- A **single blockage** in one path blocks flow through it.
- If **all paths are blocked**, there is no flow – they are *d*-separated.
- But a **single unblocked** path allows some flow to go through.

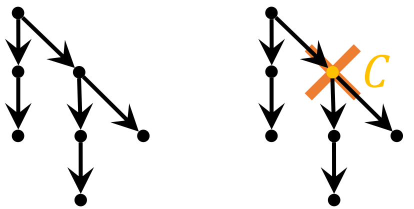
When not conditioning, only colliders block dependence.

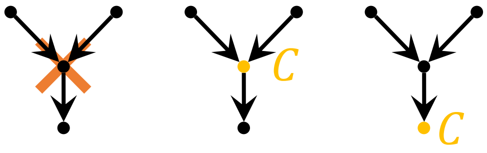
When conditioning on some variables, there are more ways to block.

# *d*-separation – conditional

**Conditioning** on a node $C$:

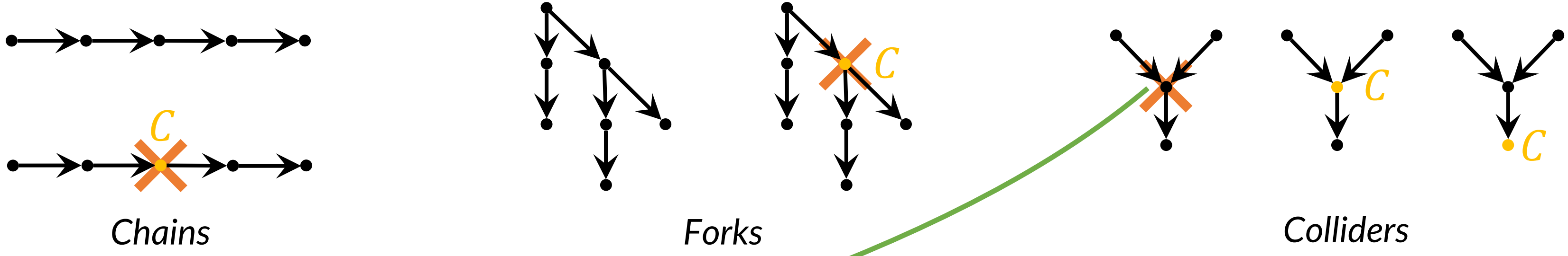


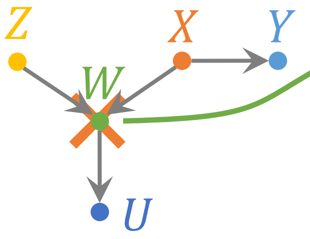




- **Chains** that contain $C$ block dependence.
- **Forks** with $C$ as the middle node block dependence.
- **Colliders** that **do not** have $C$ as the middle node or as a descendant block dependence.

# *d*-separation – example

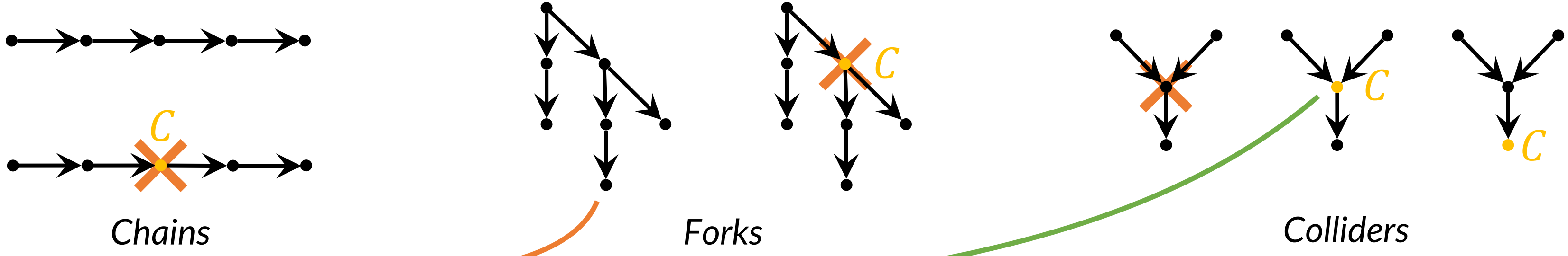


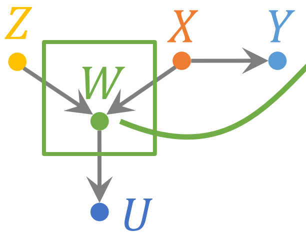




Consider the dependence between $Z$ and $Y$ in this SCM.

↪ **Unconditionally**, they're *d*-separated (**independent**), since the path between them is **blocked** by the collider $W$.

# $d$-separation – example



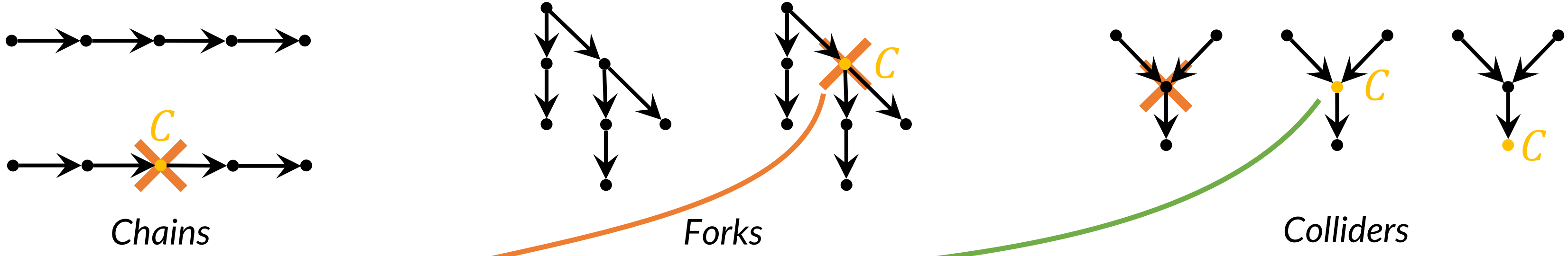

Consider now that we **condition** on $W$.

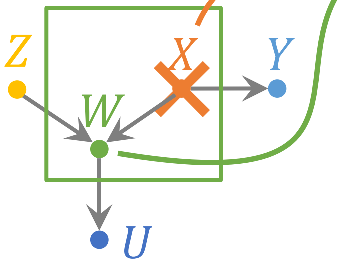
↪ The only path between $Z$ and $Y$ has:

- A **fork** ($X$) that is not conditioned on (**unblocked**);
- A **collider** ($W$) that is conditioned on (**unblocked**).

With an unblocked path, $Z$ and $Y$ are now $d$-connected (**dependent**).

# *d*-separation – example



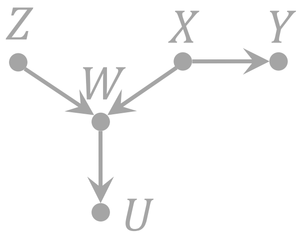

On the other hand, if we **condition** on $W$ and $X$:

↪ $Z$ and $Y$ remain *d*-separated (**independent**), as the path has:

- A **fork** ($X$) that is conditioned on (**blocked**);
- A **collider** ($W$) that is conditioned on (**unblocked**).

But a single blocked node in each path is enough to block dependence completely.

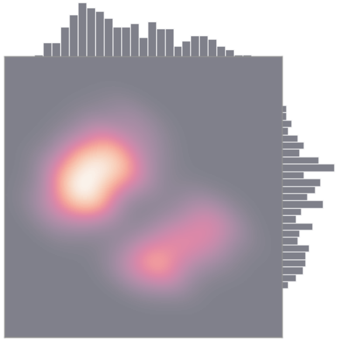
# Model testing




**But why?** What is the usefulness of this in ML practice?

- Causal models have **testable implications** for generated data.
- So if a structural causal model (SCM) is correct, it will **predict** patterns of conditional (in)dependence, which **the data must have.**

  ↪ Otherwise the model is incorrect!

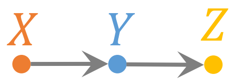


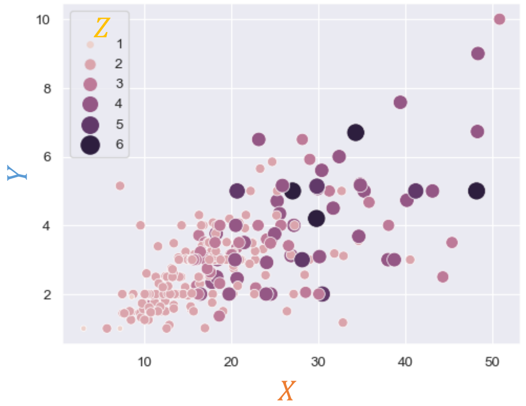
So, given a dataset and a SCM that you think must explain it:

- Use *d*-separation criteria on the SCM to **list** which variables are independent, conditional on other variables.
- Use a statistical independence test to **check** if those variables are independent in the dataset, conditionally (e.g. by grouping).

If a test fails, the SCM does not fit the data, and must be changed.

# Model testing – example



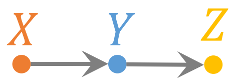

*Example – the graph on the left, with associated dataset:*

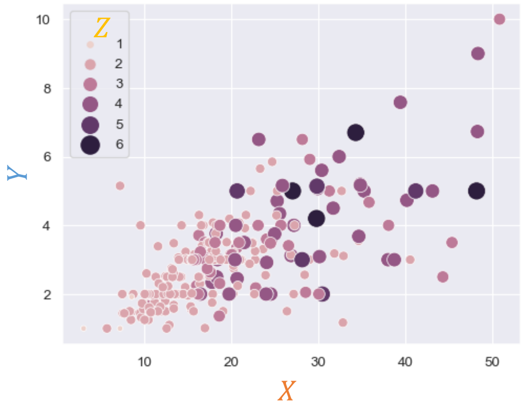
- We know that $X$ and $Z$ are dependent through $Y$, but **independent when conditioned** on $Y$.
- So we regress $Z$ from $X$ and $Y$, for example by fitting the data with a linear model:

$$Z = aX + bY + c$$

- We expect ***a* to be 0**. Otherwise, there is a (linear) dependence between $X$ and $Z$, and the SCM is **wrong**. *(Conditional correlation implies conditional dependence.)*
- We also know how to **repair** it: we must add a path between $X$ and $Z$, that is not *d*-separated by $Y$.

# Model testing – example



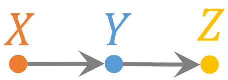

*Example:*

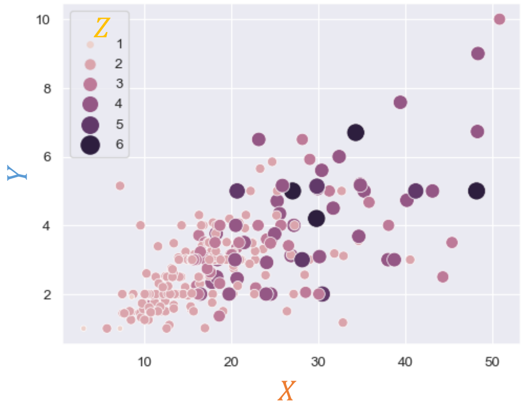
- Alternatively, we could **group the data** based on $Y$ (condition on $Y$).
- Then we expect $X$ and $Z$ to be **independent** in each group.
- To verify this, we use a statistical test of independence.

Use independence tests from standard Python packages:

- $\chi^2$ test (discrete) – `scipy.stats.chi2_contingency`
- Mutual Information (continuous) – `sklearn.feature_selection.mutual_info_regression`

# Model testing



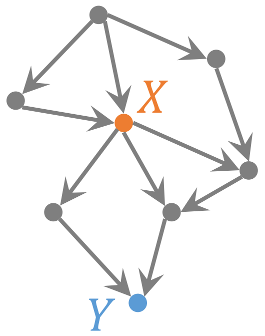

Advantages over fitting a model to data directly:

- *d*-separation is **nonparametric**: it only uses dependency patterns, and results are the same **regardless** of the distributions or functions that relate variables.
- It is **local**: it tells us **where** the model doesn't fit the data, so we can repair it.

  ↪ If instead we fit a model to data directly, and it's a poor fit, there would be no clue as to what went wrong.
- Since it's local, we can get **partial information** even when other parts of the model are unknown, or have parameters that are impossible to estimate from data.

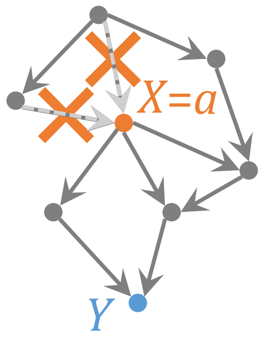
# A brief appetizer – interventions





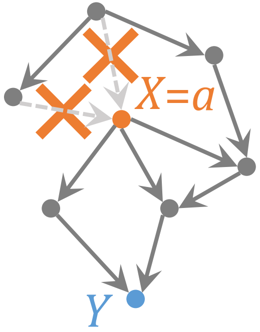

- We can also use SCMs to analyse the effect of **interventions**:
  - What is the outcome if a patient takes a given drug?
  - Would a given policy result in fewer wildfires?
  - Does a given change in a system yield better performance?

# A brief appetizer – interventions




- **Intervening ≠ conditioning.**
- An intervention **changes** the graph:
  - The intervened variable is set **deterministically** to a value (e.g. $X = a$).
  - All incoming edges are **cut** (removed).
- The SCM obtained after the intervention **shares** the remaining structure (functional relations, exogenous distributions) but will generally result in **different** overall probabilities (e.g. $P(Y)$).
- This answers "what if" (decision-making) questions, while avoiding many pitfalls (confounding, Simpson's paradox, etc).

# A brief appetizer – interventions




*"Do"-notation*

- To write the probabilities in a modified graph (obtained by intervening), insert a "do" operator in the conditioned variables:

$$P(Y \mid \mathrm{do}(X = a))$$

- Meaning: Set $X = a$ (deterministically), cutting off incoming edges; then evaluate $P(Y)$ normally in this new graph.

*Average Causal Effect*

- Binary intervention (e.g. take drug $X = 1$ vs. no drug $X = 0$),

$$\mathrm{ACE} = P(Y \mid \mathrm{do}(X = 1)) - P(Y \mid \mathrm{do}(X = 0))$$

- *Interpretation*: **difference** between the **fraction** of the population that recovers ($Y$) when all take drug vs. when none take the drug.

# Parting thoughts

- We still don’t have the whole “book of why”.
- But we can be aware of the kinds of structures that hide behind the data we observe, and how they can trick our intuition.


For a more complete (but still relatively short) picture, check:

*Judea Pearl et al., “Causal Inference in Statistics: A Primer”*

…from which I borrowed many of the examples in these slides.

Thank you!