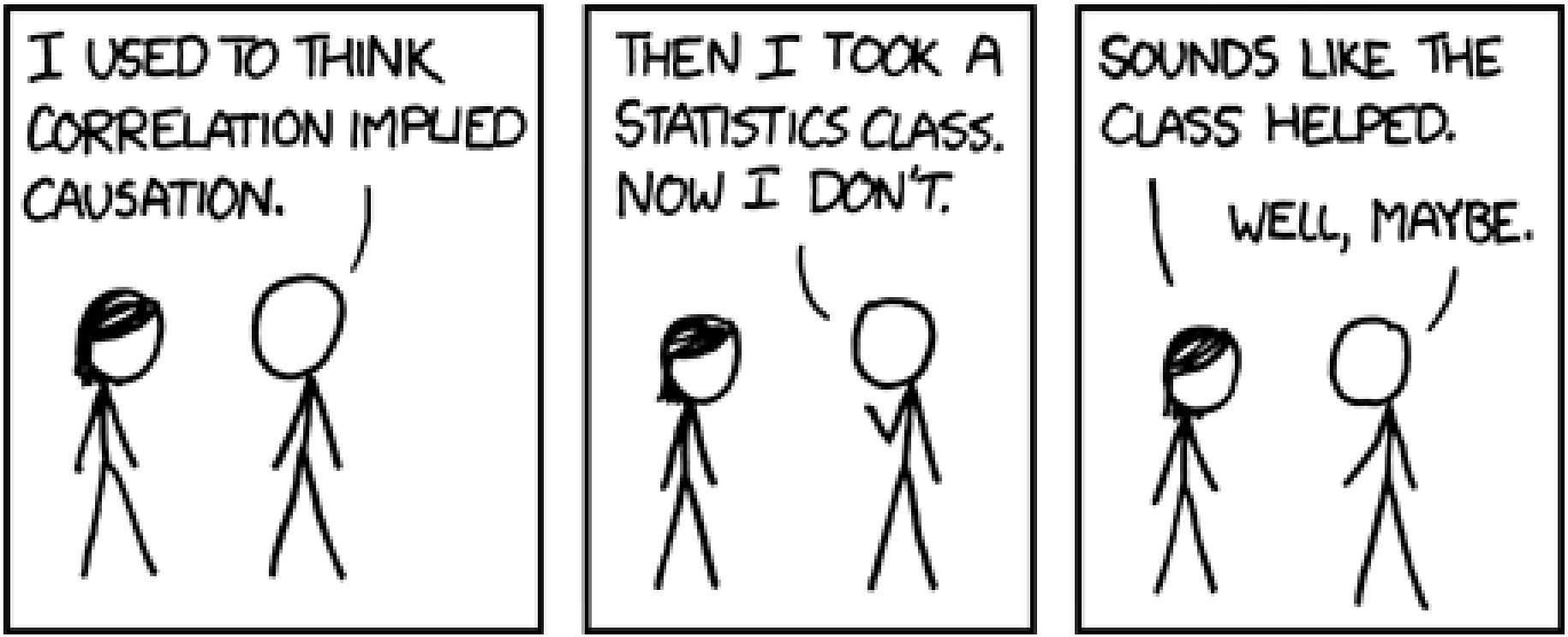
Introduction to Causality

Table of Contents

Motivation

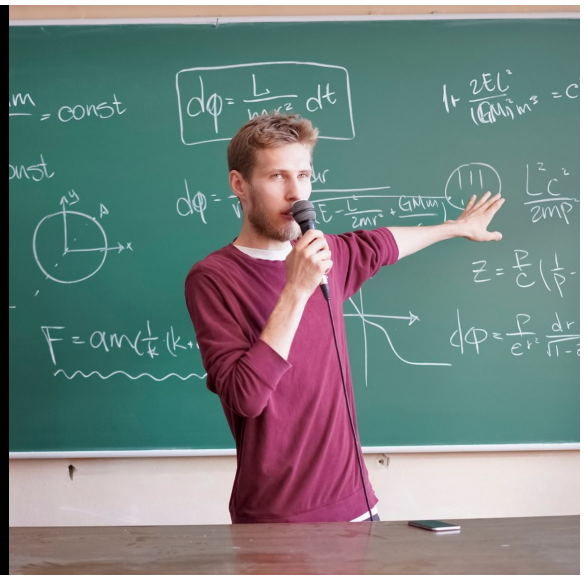

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Motivation

Why is this such a key question?

Stability: Causal relationships are robust to external change.

How does the human brain think about causality?



Motivation


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Example: Distribution shift

Suppose we are trying to classify images $\mathbf{x}$ of cats or dogs (label $\mathbf{y}$), and we learn a NN $p(\mathbf{y}|\mathbf{x})$ for this task.

Now let's assume we go to a different "environment". What might have changed?

- $p(y)$: Maybe there are more cats in dogs in some countries...
- $p(x|y)$: Maybe we see different breeds of dog more often. Or we capture the image in different conditions, e.g. day/night
- $p(y|x)$: ??

Questions

- What does "A causes/caused B" mean?
- How can we infer causal relationships?
- How can we represent and use causal information?

Defining causality: Things to consider

- **Type/token level**: Suppose medicine A causes most patients to recover (B), but also had no effect or killed a small minority of patients ($\neg$B).
- **Necessary/Sufficient cause**: Do we assert there is no alternative cause for B, or that A alone can cause B?

Characterizing causality: a first attempt

Name	**Language**	**N/S?**
Association	"A makes B more likely"	
Temporal Precedence	"A comes before B"	N
Counterfactuals	"If A had been different, B might have been different"	S
Physical Mechanism/ Direct Cause	"There is some mechanism through which A influences B"	N, S

Table of Contents

Probabilistic Causality

Long line of attempts in social sciences, economics and statistics.

- Granger causality
- Suppes (1970): "Probability raising" and "screening off using temporal precedence"
- Eells (1991): Contexts, splitting up type/token-level causality

Usually assumes knowledge of **temporal precedence**

Generic framework for probabilistic causality

C is "causally relevant" for E if:

- C precedes E temporally
- $P(E|C) > P(E)$, i.e. "probability raising"
- No common cause S of C and E (preceding both temporally) - *Reichenbach's common cause principle*

The third is where probabilistic accounts of causality differ: some rely on "background contexts", some on enumerating possible causes, etc.

Pros and cons

- Fairly easy to come up with heuristics (and many exist) for using association and temporal information
- Can be of great practical utility in assisting humans in identifying/checking **potential** causes

but ....

- No model-based extension: cannot perform complex queries and reasoning
- Not reliable: all heuristics will misidentify causes sometimes
- Heuristics implicitly code in assumptions
- Reliance on temporal information to break symmetry; this precludes application to many problems
- Struggles to deal with token-level causality

Counterfactual causality

A **counterfactual statement** is one which expresses information about what *did not happen*; i.e. a hypothetical. For example, "if c happened, e would have happened".

Lewis' (1973) counterfactual theory defined this statement using the notion of "possible worlds". Either:

- There are no possible c-worlds;
- There exists an c-world where e holds, which is closer to the actual world than any c-world in which e does not hold

Then, Lewis defines e to be **causally dependent** on c if:

- "If c was true, e would be true"
- "If c was not true, then e would not be true"

Pros and Cons

- Axiomatic framework available which explicitly codes in assumptions;
- Can be extended compositionally with the idea of a "causal chain";
- Explicitly works on the unit/token level

but...

- No obvious way to define "closest world", without invoking causality in a circular manner;
- Original formulation not probabilistic;

Table of Contents

Why do we need a model?

Definitions of causality are good ... but we want to be able to reason/answer causal queries.
Here are some natural language queries:

- Does A cause B?
- What caused B?
- Did A cause B?
- What distribution over B (lung cancer) does forcing someone to smoke (A $= 1$) cause compared to the general population?

Associations: Given A is 1, what is the distribution on B?
Interventions: If we fix A to be 1 ($do(A = 1)$), then what is the distribution on B?
Counterfactuals: In a specific situation where we observed $C = c$, if we fix A to be 1, then what is the distribution on B?

Modelling

We need a **model**: a representation of reality which allows us to assign truth values to relevant statements through some computational procedure. e.g.

- **Truth tables** are a model for evaluating Boolean expressions
- **Joint probability distributions** are a model for evaluating conditional probabilities, conditional independences

A **causal** model needs to encode the truth values of causal statements, e.g. counterfactuals. Probability distributions over variables are not sufficient for a causal model: need additional **assumptions**. Two main frameworks:

Potential Outcomes (PO) and **Structural Causal Models (SCM)**.

Counterfactual Notation

Idea: Define counterfactuals explicitly as counterfactual **random variables** on a probability space $(U, \mathcal{F}, p)$, even though we can never observe them.

Here $u \in U$ represents "all relevant randomness".

Y_x is the counterfactual random variable "the value of Y, had X been x". That is, we set/intervene X to be x. The randomness is over u.

We can then define queries such as:

- $\boldsymbol{p(Y_x = y)}$: "probability that Y is y if we set X to be x"
- $\boldsymbol{p(Y_{x'} = y' | X = x, Y = y)}$: "probability that Y would have been y' if X were x', given that we actually observed $X = x$ and $Y = y$"
- $\boldsymbol{p(Y_x = y, Y_{x'} = y')}$

Rubin Causal Model (Potential Outcomes): A statistical/algebraic approach to counterfactals

In PO framework, we define counterfactual variables Y_x as primitives. Thus, it is simple to answer causal queries of the type we have described.

Causal knowledge is represented as knowledge of about the probability distribution over counterfactual variables, e.g. $p(X, Y_{z=0}, Y_{z=1})$ where Z is a binary treatment.

However, it can be very difficult to specify such as distribution, especially when there are many variables involved; might need to work with conditional independences and algebraic manipulation.

Pearl's SCMs (Structural Causal Models): A structural approach to counterfactuals

Model represented by a graph and set of **structural equations**:

$$v_i = f_i(\text{pa}_i, u_i)$$

where v_i are variables, pa_i represent the parents of v_i in the graph, and u_i are "background variables". We make the model probabilistic by additionally including a distribution $p(u)$.

This representation is sufficient to answer all of the causal queries we have previously mentioned.

Takeaways

- Causal relationships are useful because they are **stable**;
- All causal questions can be expressed in terms of counterfactuals;
- In order for a machine to reason about causal queries, they need a **causal model**;
- A **causal model** requires assumptions which cannot always be inferred from observational data
- Pearl's SCMs are widely used in causal inference and machine learning