Causal Inference & Simulation

Kenneth (Kenny) Joseph

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

Overview

Three things you need for your project/to do good [computational social] science (IMO)

1. The ability to **understand/reason about** the social world
2. The ability to use that understanding to **design interventions** on the world
3. The ability to **conduct measurements** that assess the [potential] **effect** of that intervention

Overview

Three things you need for your project/to do good [computational social] science (IMO)

1. The ability to **understand/reason about** the social world
2. The ability to use that understanding to **design interventions** on the world
3. The ability to **conduct measurements** that assess the [potential] **effect** of that intervention

To help us with this, I'm going to talk briefly, at a very high level, about two things:

1. Causal inference
2. Simulation

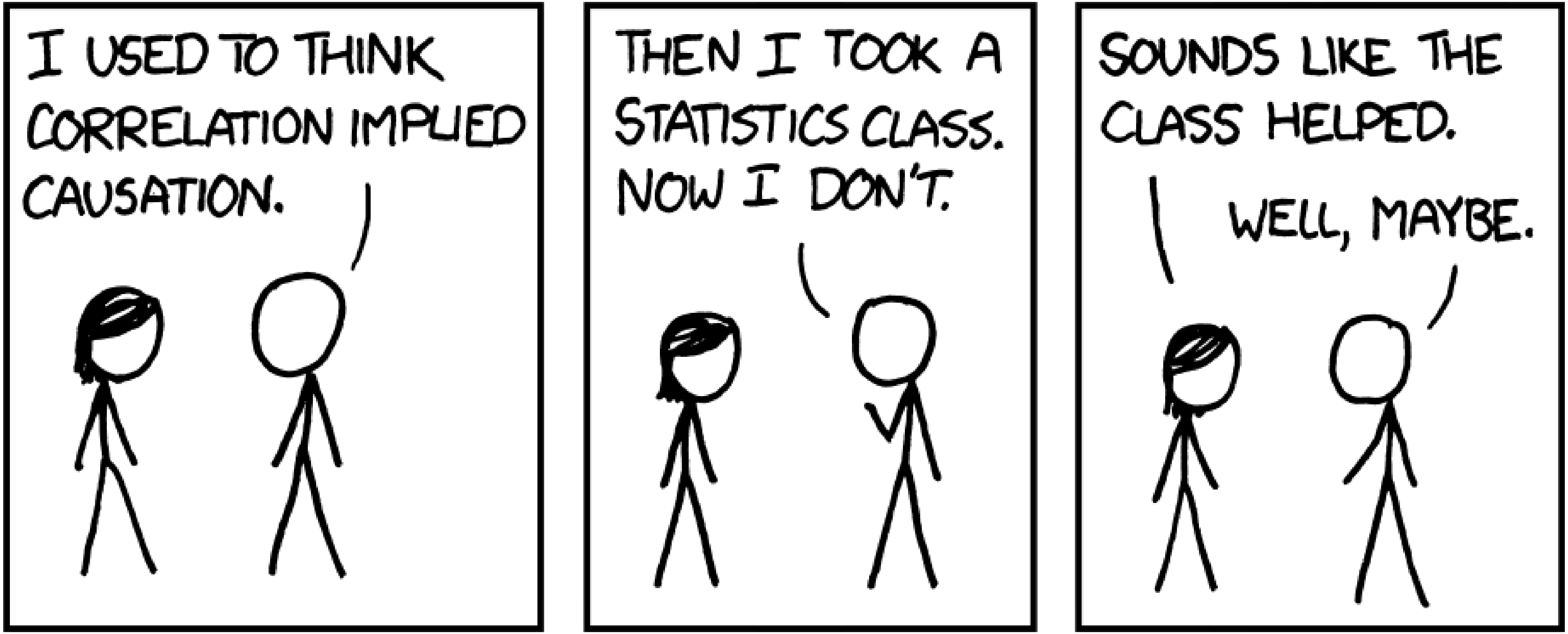
Overview of this lecture

Three things you need for your project/to do good [computational social] science (IMO)

1. The ability to **understand/reason about** the social world
2. The ability to use that understanding to **design interventions** on the world
3. The ability to **conduct measurements** that assess the [potential] **effect** of that intervention

Note: we're going to do this through the use of *diagrams*, or, less fancy – pictures.

But first!


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

Correlation vs. Causation

- Correlation: A **measure of the relationship between two variables**
- Causation
  - Informally: “if I change A, then B will also change”
  - Formally, two ways to think about this:
    - “Do calculus”
    - Potential outcomes
    - People argue about which of these is better. Irrelevant for our purposes

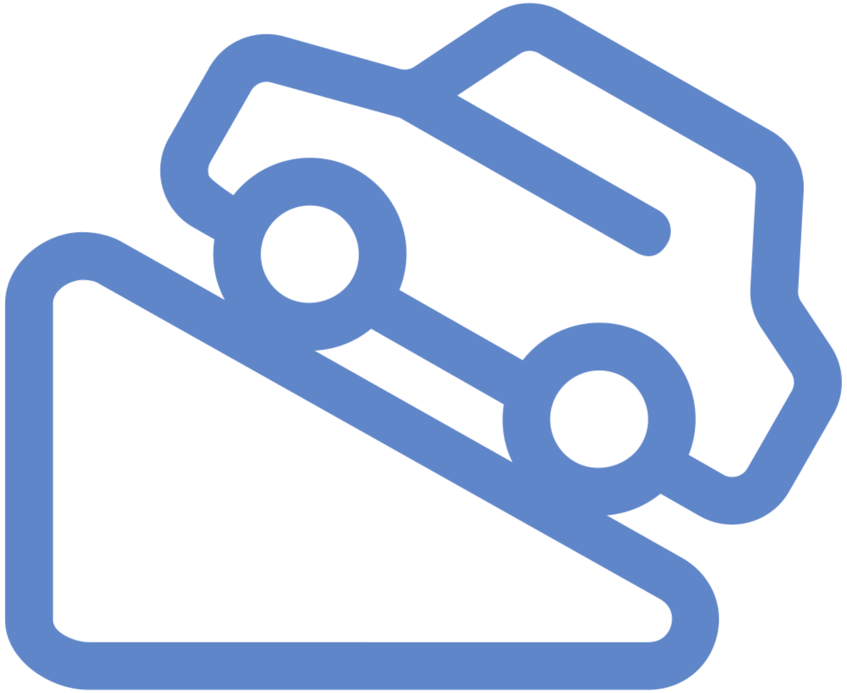
Correlation vs causation

https://tylervigen.com/spurious-scholar

- Can you summarize the difference between correlation and causation?
- Why do we care about causation and not correlation?
- Does correlation always mean causation?
- Does *no* correlation always mean *no* causation?

No correlation != No causation

- What does pressing down the gas pedal do to speed?
- What does going up/down a hill do to speed?
- If we didn't know that, what conclusions might we draw from a skilled driver?



https://goodauthority.org/news/milton-friedmans-thermostat/

bq. Watch what happens on a really steep uphill bit of road. Watch what happens when the driver puts the pedal to the metal, and holds it there. Does the car slow down? If so, ironically, that confirms the theory that pressing down on the gas pedal causes the car to speed up! Because it means the driver knows he needs to press it down further to prevent the speed dropping, but can't. It's the exception that proves the rule.

Here, for example, a naive observer might take a particular campaign action, which is associated with the candidate's defeat, as evidence of incompetence by the campaign. It's not – it may be correlated only because it is the best thing that the campaign can do under particularly difficult external circumstances.

In general, correlations are fragile beings

https://www.youtube.com/watch?v=DbJyPELmhJc&ab_channel=AutodeskResearch

Why might we see a correlation where there is no (direct) causation?

What did we just do?

- We drew an *assumed* causal graph
- This diagram we drew is reminiscent of two things
  - Probabilistic graphical models/directed acyclic graphs (or in a restricted setting, causal graphical models)
  - System dynamics models
- But first – how do we move from *assumed* to *known*?

Introducing causal inference

What if we had a method that could **explicitly account for these challenges and help us think through *real causation*?**

Enter causal inference!

https://microsoft.github.io/dowhy/example_notebooks/tutorial-causalinference-machinelearning-using-dowhy-econml.html

Sounds great! But there’s a catch…

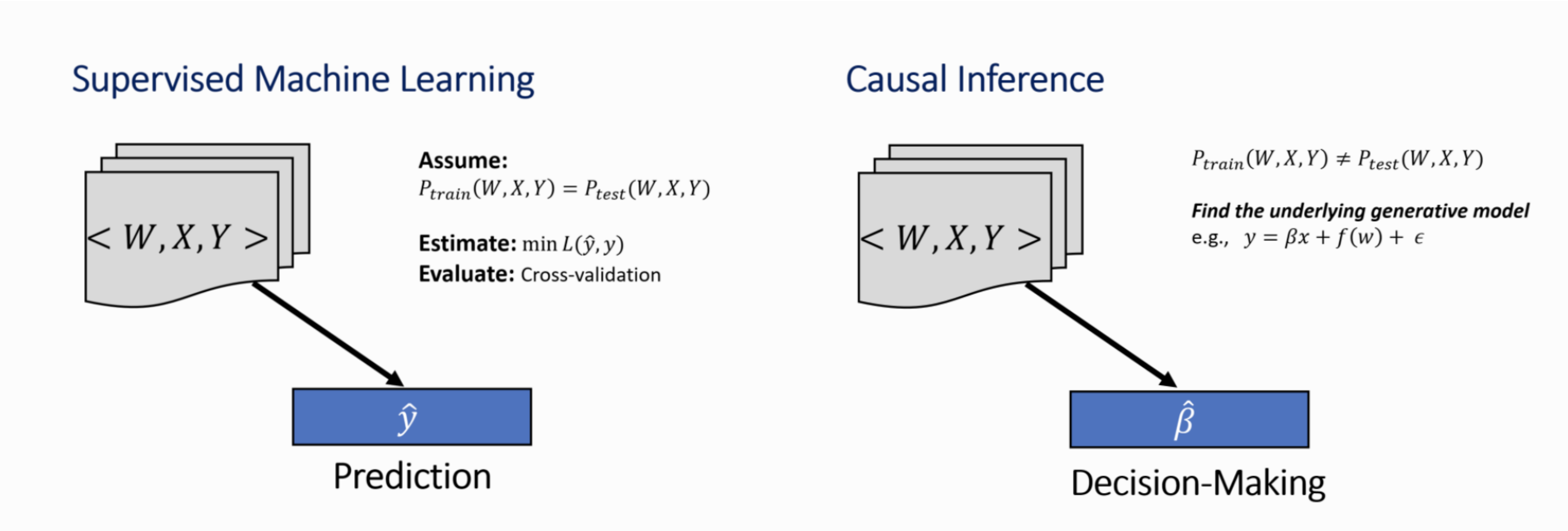
I want to know if telling Alvin he is going to get an A no matter what will make his participation better or worse

Real World: **do(T=1)** Counterfactual World: **do(T=0)**

I can’t both give Alvin an A AND not give it to him!
This is the **Fundamental Problem of Causal Inference**

https://microsoft.github.io/dowhy/example_notebooks/tutorial-causalinference-machinelearning-using-dowhy-econml.html

Aside: Causal inference vs. ML



What can we do, then?

1. Target averages/expectations instead of an individualized effect

$$E[Y|do(A=1)] - E[Y|do(A=0)]$$

2. Experiment!

Note: do operator indicates an intervention, in an RCT we intervene by randomly assigning treatment and control to comparable groups.

Problems with the simple story

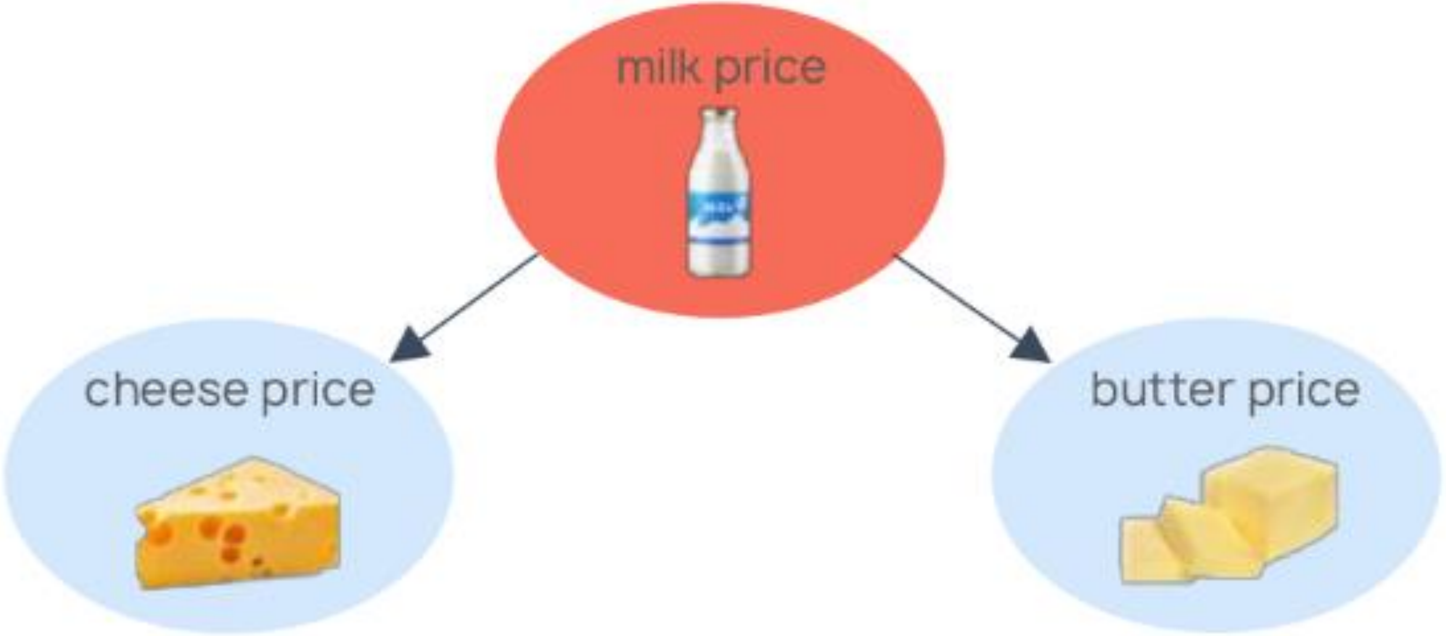
- Sometimes, experimentation is unethical
  - I suspect half of you would be very angry if I gave you a placebo study guide ☺
- Other times, we might have wanted to experiment but simply couldn't, and are left with a bunch of observational data
- Since we do not always have access to experimental data, we rely on observational data for estimating causal effect

Uh-oh

- Now:
  - We have an idea how to evaluate an intervention, but when we can't experiment, we have to **control for factors associated with both the intervention and the outcome**
  - Related: how do we decide on a treatment in the first place?
- Enter DAGs [and then, causal graphical models]. Informally, there are two "kinds" of probabilistic DAGs
  - Bayes Nets encode the factorizations of any joint probability distribution
  - Causal Graphical Models put explicit assumptions about causation into DAGs.
  - We'll focus on the latter

Causal DAGs– a tool to explain causal reasoning

https://causalens.com/resources/white-papers/why-correlation-based-machine-learning-leads-to-bad-predictions/

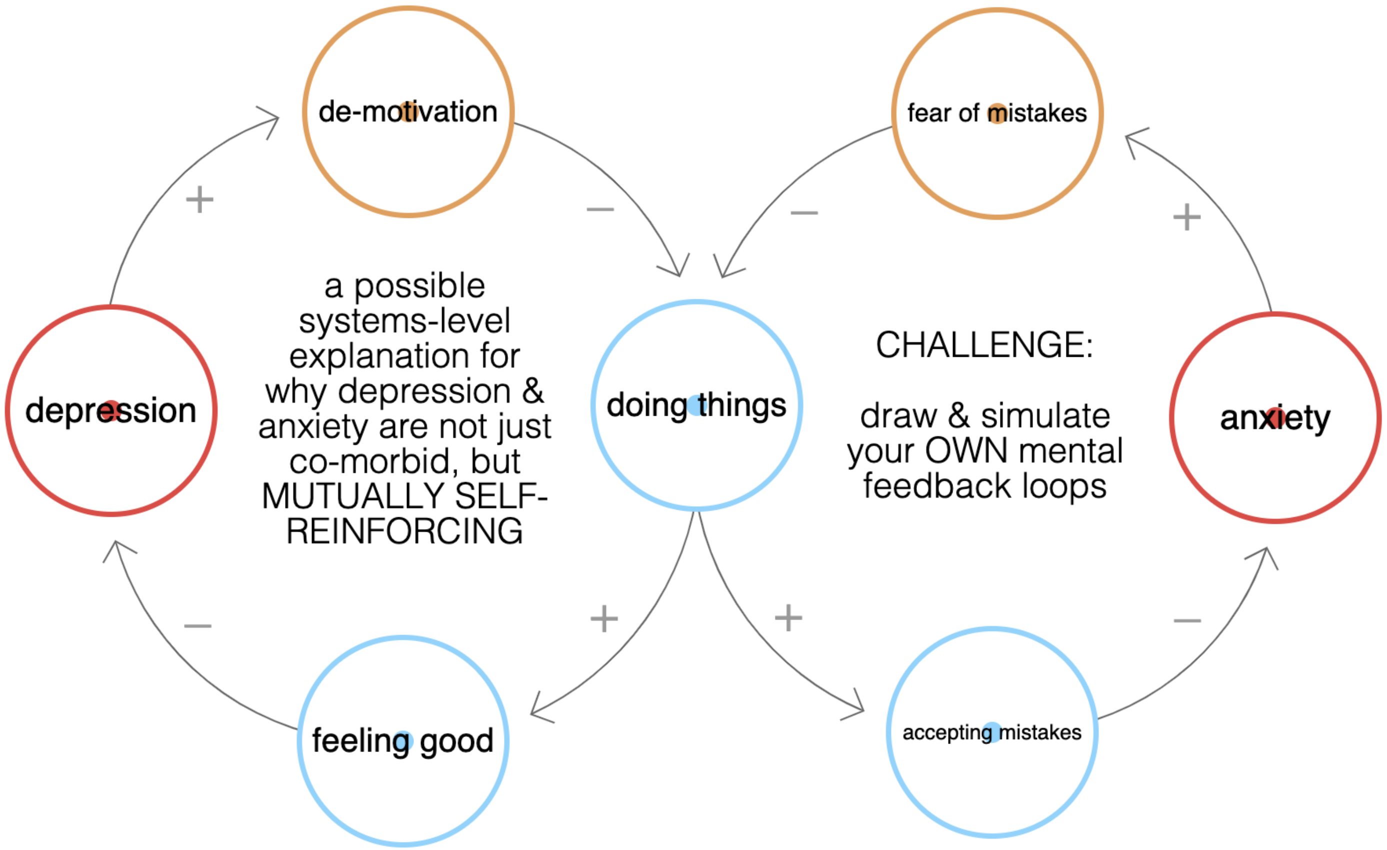


What is a causal graphical model?

- A **diagram** that helps us **explain** our **assumed causal** relationships between two things
  - Nodes are **random variables**
  - Edges are **causal relationships between RVs**
- Let's practice!
- Draw a DAG that represents the immigration and crime example.

Problems with DAGs

- Not everything is a directed and acyclic graph
- We don’t know probabilistic distributions for everything, sometimes we just know high-level relationships
- The real world is **messy and complex**
- **… enter simulation**



https://ncase.me/loopy/