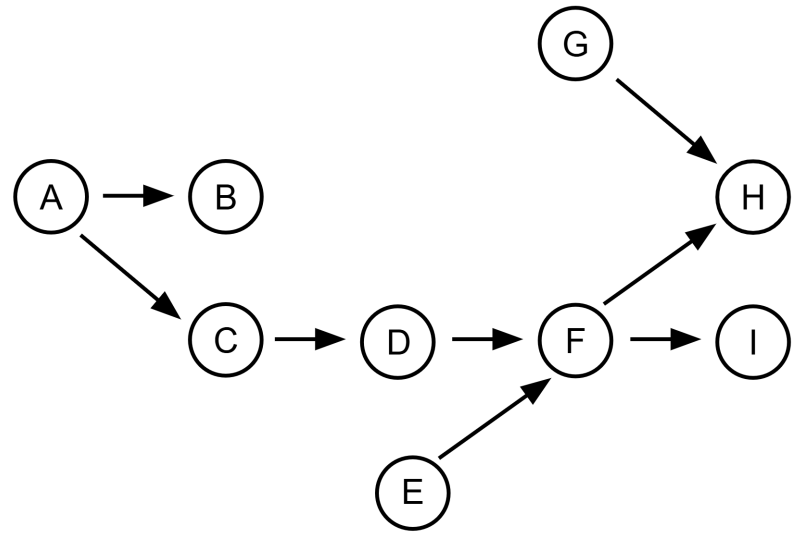
Introduction to Causal Inference

SciPy 2026
Roni Kobrosly Ph.D.



Press Space to begin →

Netflix Technology Blog
May 21 · 6 min read · Listen

A Survey of Causal Inference Applications at Netflix

At Netflix, we want to entertain the world through creating engaging content and helping members discover the titles they will love. Key to that is understanding causal effects that connect changes we make in the product to indicators of member joy.

To measure causal effects we rely heavily on AB testing, but we also leverage quasi-experimentation in cases where AB testing is limited. Many scientists across Netflix have contributed to the way that Netflix analyzes these causal effects.

To celebrate that impact and learn from each other, Netflix scientists recently came together for an internal Causal Inference and Experimentation Summit.

DJ Rich
Jun 28 · 8 min read · Listen

Causal Forecasting at Lyft (Part 1)

By Duane Rich and Sameer Manek

Efficiently managing our marketplace is a core objective of Lyft Data Science. That means providing meaningful financial incentives to drivers in order to supply affordable rides while keeping ETAs low under changing market conditions — no easy task!

Lyft's tool chest contains a variety of market management products: rider coupons, driver bonuses, and pricing, to name a few. Using these efficiently requires a strong understanding of their downstream consequences — everything from counts of riders opening the Lyft app ("sessions") to financial metrics.

Ocelot: Scaling observational causal inference at LinkedIn

December 13, 2022

Share Tweet Share

Co-authors: Kenneth Tay and Xiaofeng Wang

At Linkedin, we constantly evaluate the value our products and services deliver, so that we can provide the best possible experiences for our members and customers. This includes understanding how product changes impact key metrics related to those experiences. However, simply looking at connections between product changes and key metrics can be misleading. As we know, correlation does not always imply causation. When making decisions about the path forward for a product or feature, we need to know the causal impact of that change on our key metrics.

Etsy Code as Craft

Categories

How to Pick a Metric as the North Star for Algorithms to Optimize Business KPI? A Causal Inference Approach

By Xuan Yin
Aug 2, 2020

Learning Objectives

By the end of this tutorial, you should be able to:

- Understand the pitfalls of observational data analysis
- Know the key types of causal relationships
- Understand AI/ML vs causal inference vs and experiments
- Start conducting preliminary causal analyses
- Confidently explore the topic on your own

Is this plot useful? 🏖️

As a vacationer looking to avoid a crowded hotel? This is fine 👍

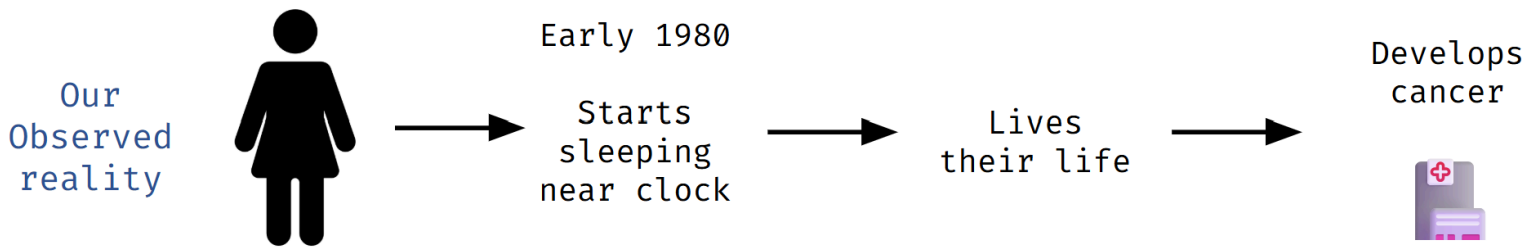
As a hotel owner trying to optimize your pricing with this plot is useless ❌


ARTCO
Radium

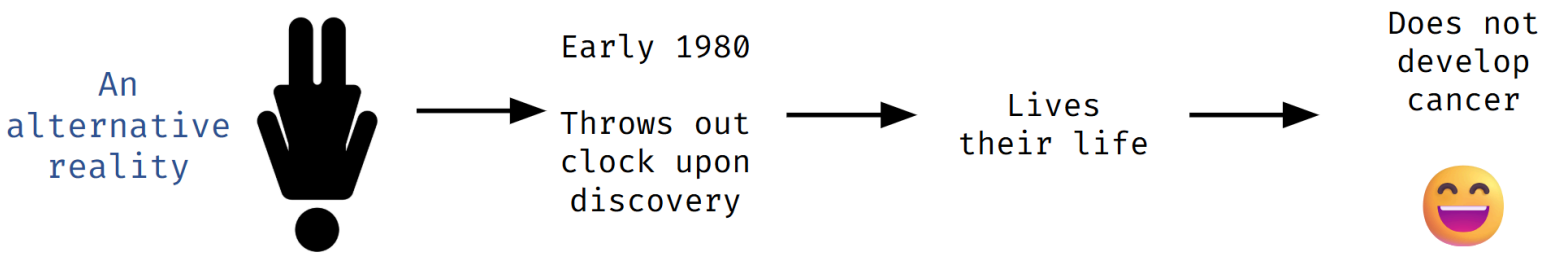

ARTCO
Radium

Does exposure to this radium clock cause cancer?

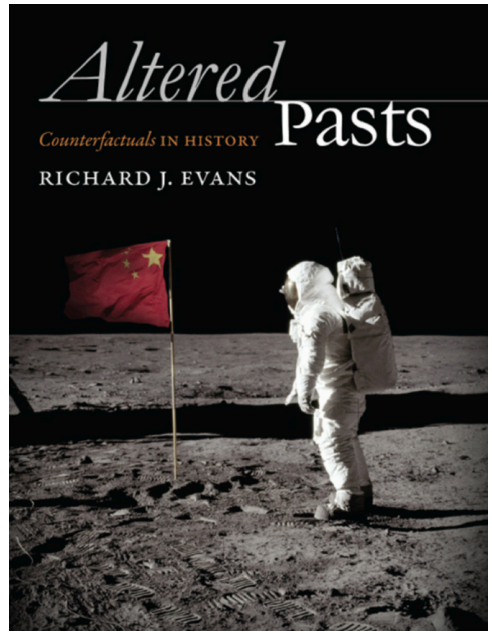
… what happens in an alternative universe?



… what happens in an alternative universe?



Counterfactuals (“Counter to fact”)



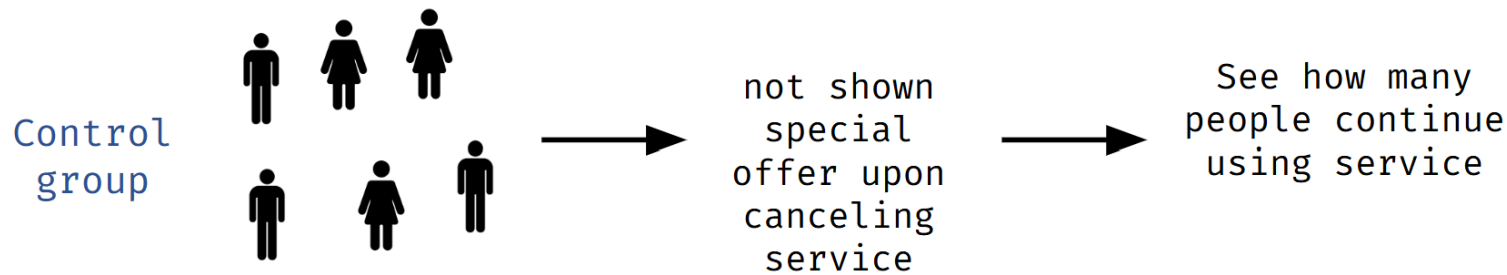
Counterfactuals

You can also think of counterfactuals as a **missing data problem**

ID#	SPECIAL OFFER	AGE	DEVICE	CHURN?
1	Y	40	iphone	Y
2	Y	35	android	N
3	N	77	iphone	N
4	Y	18	android	N

ID#	OBSERVED?	SPECIAL OFFER	AGE	DEVICE	CHURN?
1	✓	Y	40	iphone	Y
1	✗	N	40	iphone	???
2	✓	Y	35	android	N
2	✗	N	35	android	???
3	✗	Y	77	iphone	???
3	✓	N	77	iphone	N
4	✓	Y	18	android	N
4	✗	N	18	android	???

Experiments / A/B Tests / Randomized Controlled Trials



Experiments / A/B Tests / Randomized Controlled Trials

Treatment group

shown special offer upon canceling service

See how many people continue using service

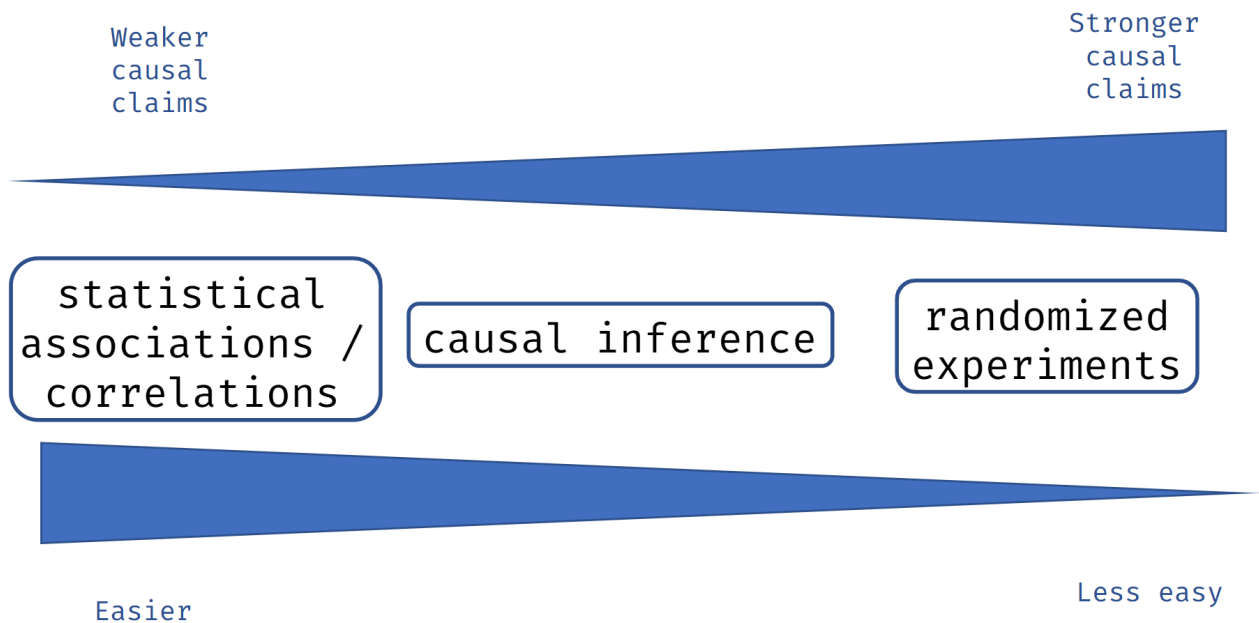
When Experiments Aren’t Feasible

- Understanding how a user’s behavior changes when they switch from an iPhone to the newest Samsung phone
- Too few units, such as in a Merger and Acquisition scenario (there is one event that may or may not happen)
- Modify household incomes in neighborhoods, to see if reducing a neighborhood’s income inequality reduces the local crime rate

When Experiments Aren’t Ethical

- Randomly assign some people to be exposed to lead paint while others are not, then see which group is more likely to develop neurological disorders
- Assigning some social media users to receive more psychologically dark posts to understand how it impacts engagement

The Hierarchy of Evidence



Important Note on Correlations

I’m referring to **RAW associations and correlations**. Calculating correlations is **indispensable** in causal inference work, but we make intelligent adjustments to make them useful.

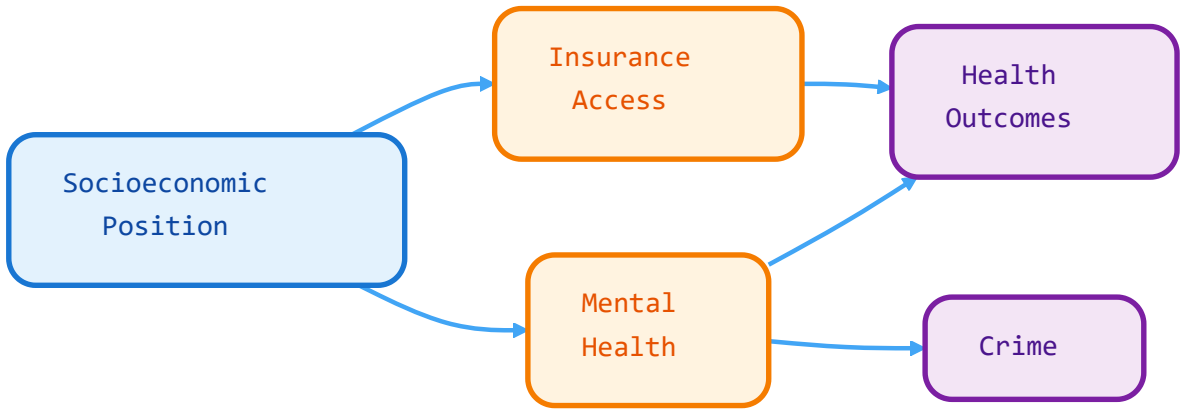
Causal Inference Questions

- How does improving neighborhood income inequality **reduce** neighborhood crime rate?
- How does **increasing or decreasing** the price of a product impact demand?
- What would be the **impact** on diabetes if we reduced average sugar consumption by X grams?

Standard ML Questions

- Can I **cluster** neighborhoods by their characteristics?
- Can I **predict** whether someone will convert from a lead to a customer?
- How well can I **predict** whether a patient will be diagnosed with diabetes later in life?

A Causal Graph (DAG)



Directed Acyclic Graphs (DAGs) help us visualize causal relationships

Exercise Time! 🎯

Let’s practice creating causal graphs

Car Insurance and causality

Make & model

Theft history

Car value

Advanced airbag

Antilock brakes

Driving course completion

Vehicle year

Car safety rating

Accident history

Age

Medical cost of accident

Good student status

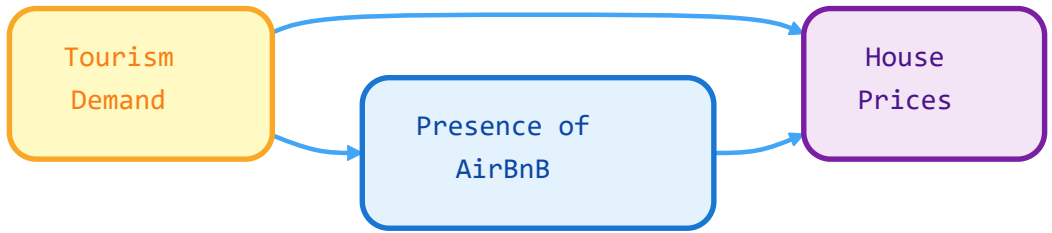
Risk aversion

Three Important Types of Causal Relationships

1) Confounders

- Always want to **control for** confounders in inferential modeling
- Confounding changes the effect size and possibly statistical significance
- Confounders can also **flip the direction** of your association of interest
- Leftover confounding is called "residual confounding"

Confounding Example: AirBnB



Tourism demand is a confounder:

- It increases AirBnB presence
- It increases house prices
- Creates a modifies any true relationship between AirBnB and prices

Types of Confounding

Positive Confounding

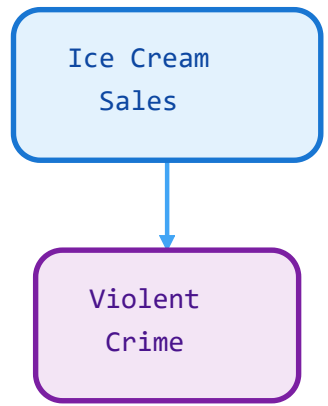
Confounder introduces a bias that pushes association **away from zero**

Negative Confounding

Confounder biases association **towards the zero**

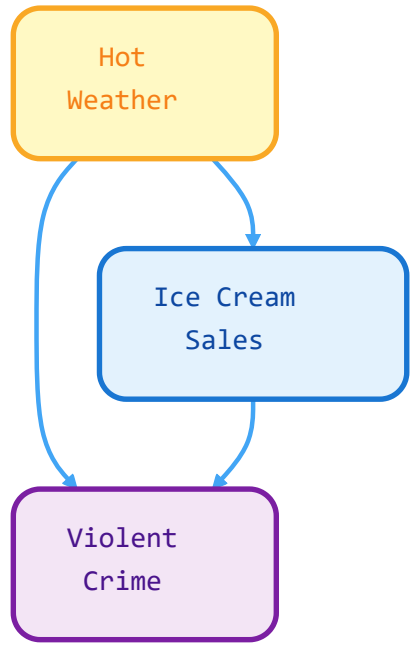
Classic Example: Ice Cream & Crime

Do ice cream sales cause violent crime? 🍦 → 🔫



Not what it seems

Do ice cream sales cause violent crime? 🍦 → 🔫



Controlling for Confounders

After controlling for season/weather, the ice cream-crime association disappears!

How do we "control" for things?

Option 1: Stratification (simple/naive way)

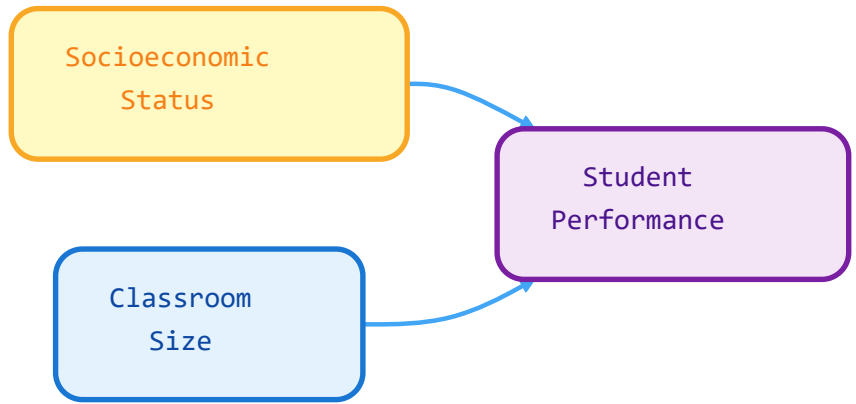
- Filter your dataset so the confounder only takes on 1 value
- Example: `p(violent_crime = 1 | hot_weather = 0)`

Option 2: Use a model!

- We'll go deep on this in the second half of the tutorial

How Experiments Break Confounding

How Experiments Break Confounding



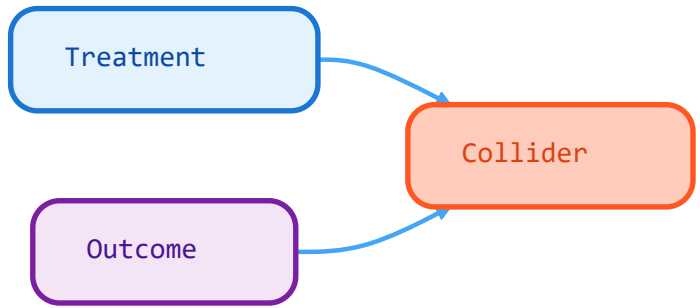
In experiments, **randomization breaks** the association between confounders and treatment! The randomization ensures classroom size is independent of socioeconomic status.

Circling back to experiments vs causal inference

Experiments are wonderful because randomization breaks all confounding

Causal inference is when we take non-experimental (observational) data and carefully try to pick apart the confounding ourselves

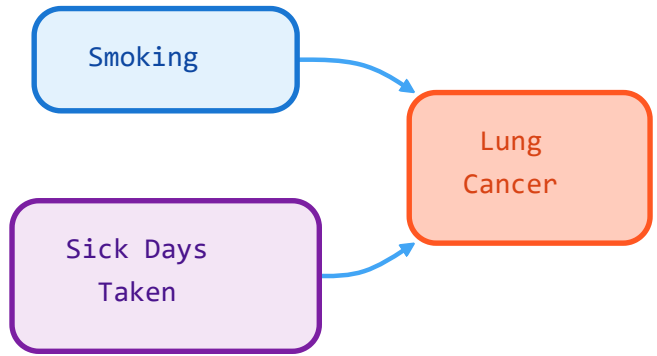
2) Colliders



Key points:

- **Never** want to control for colliders!
- Conditioning on a common effect causes **collider bias**
- Can bias results in positive or negative direction

Collider Example: Sick Days



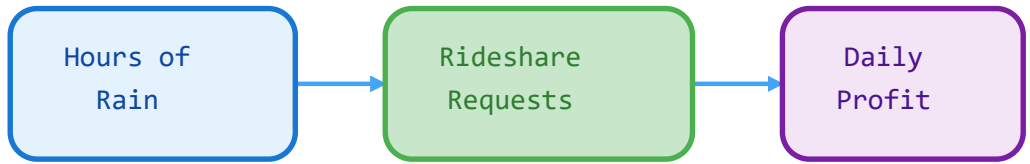
If you control for lung cancer (the collider), you'll create a spurious association between smoking and sick days taken!

3) Mediators

Key points:

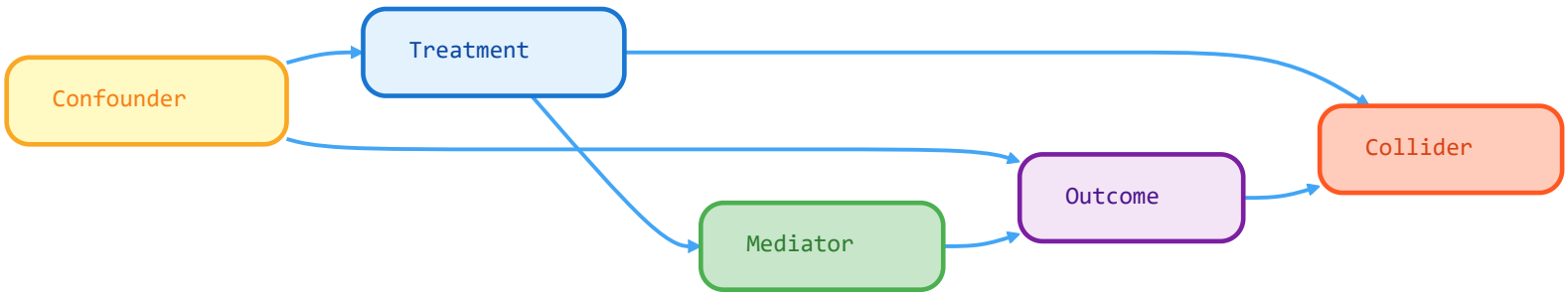
- Controlling for a mediator will **nullify** any relationship between treatment and outcome
- Helps determine causal pathways in observational data

Mediator Example: Rideshare



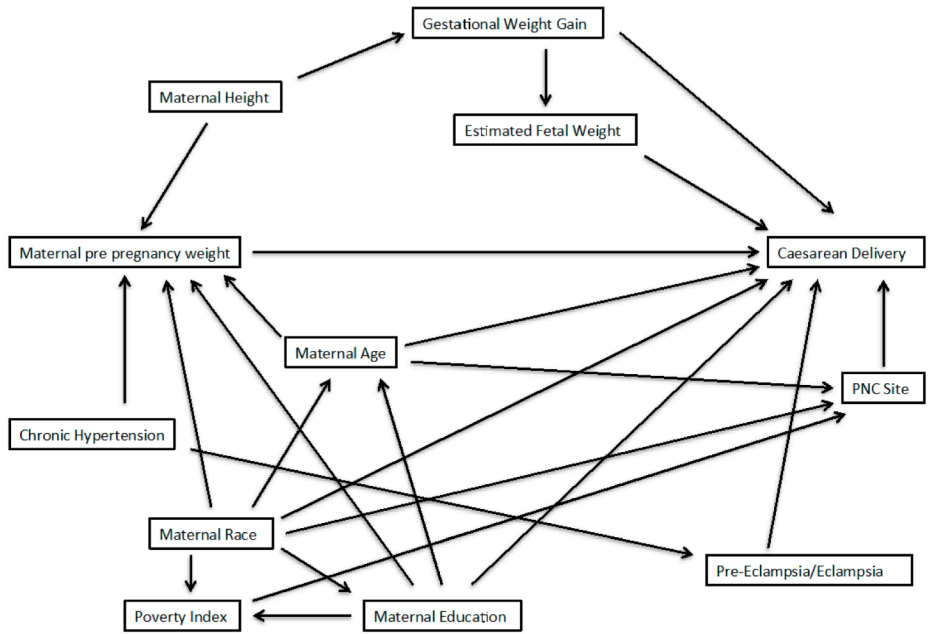
If you control for rideshare requests (the mediator), you'll eliminate the effect of rain on profit! The requests ARE the mechanism by which rain affects profit.

Putting It All Together



✅ Control for confounders
❌ Don't control for colliders
⚠️ Be careful with mediators

Reality is Complicated! Real-world causal graphs can be extremely complex.



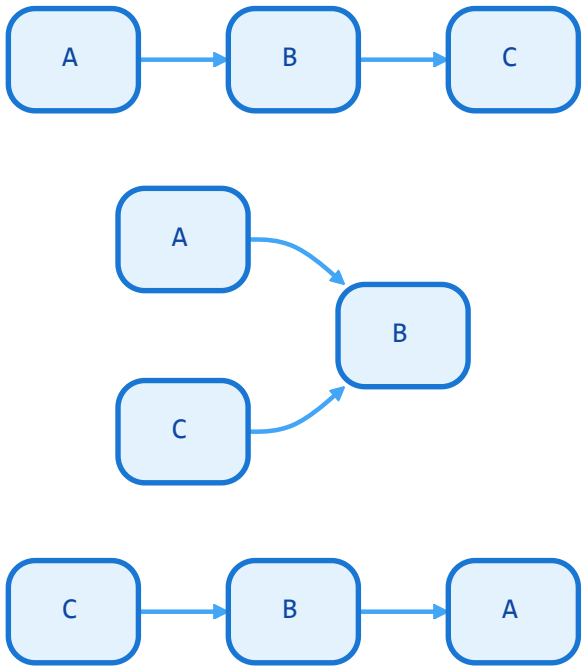
Notebook Exercise #1: Causal Graphs

Time to practice! 🚀

```
from causalgraphicalmodels.csm import StructuralCausalModel, linear_model, logistic_model

confounding_example = StructuralCausalModel({
    "temperature": lambda n_samples: np.random.normal(loc = 23, scale = 3, size=n_samples),
    "price": linear_model(parents = ["temperature"], weights = [2], noise_scale = 5),
    "bookings": linear_model(parents = ["price", "temperature"], weights = [-1, 5], noise_scale = 5),

ce_cgm = confounding_example.cgm
ce_cgm.draw()

data = confounding_example.sample(n_samples=10000)
})
```

Important Asides

Avoid Automated Causal Discovery



These three graphs belong to the same "Markov Equivalence Class" and are indistinguishable with observational data!

❌ Don’t rely on automated causal graph structure learning algorithms

✅ Stick with good domain knowledge

LLMs and Causality

EFFICIENT CAUSAL GRAPH DISCOVERY USING LARGE LANGUAGE MODELS

Thomas Jiralerspong *
Mila, Université de Montréal
thomas.jiralerspong@mila.quebec

Xiaoyin Chen *
Mila, Université de Montréal
xiaoyin.chen@mila.quebec

Yash More
Mila, McGill University

Vedant Shah
Mila, Université de Montréal

Yoshua Bengio
Mila, Université de Montréal

ABSTRACT

We propose a novel framework that leverages LLMs for full causal graph discovery. While previous LLM-based methods have used a pairwise query approach, this requires a quadratic number of queries which quickly becomes impractical for larger causal graphs. In contrast, the proposed framework uses a breadth-first search (BFS) approach which allows it to use only a linear number of queries. We also show that the proposed method can easily incorporate observational data when available, to improve performance. In addition to being more time and data-efficient, the proposed framework achieves state-of-the-art results on real-world causal graphs of varying sizes. The results demonstrate the effectiveness and

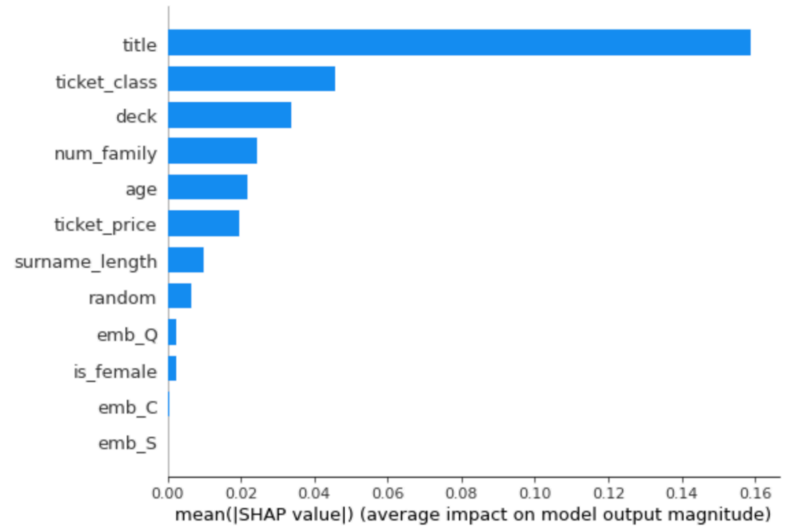
LLMs and Causality

		ASIA (n=8, m=8)				CHILD (n=20, m=25)				NEUROPATHIC (n=221, m=770)			
# of Samples	Method	Acc. (↑)	F Score (↑)	NPE	Ratio (↓)	Acc. (↑)	F Score (↑)	NPE	Ratio (↓)	Acc. (↑)	F Score (↑)	NPE	Ratio (↓)
100	GES	0.23	0.38	8	0.63	0.24	0.38	22	0.62	-	-	-	-
	PC	0.33	0.5	4	0.5	0.19	0.32	19	0.68	0.02	0.04	311	0.96
	NOTEARS	0.5	0.67	4	0.33	0.14	0.25	23	0.75	0.029	0.06	76	0.94
	DAGMA	0.22	0.36	3	0.64	0.14	0.24	41	0.76	0.03	0.05	77	0.95
	Pairwise	0.33	0.5	20	0.5	0.3	0.46	27	0.54	-	-	-	-
	Ours	**0.88**	**0.93**	7	**0.067**	0.3	0.46	27	0.54	**0.22**	**0.35**	331	**0.64**
1000	GES	0.67	0.8	7	0.2	0.31	0.44	34	0.53	-	-	-	-
	PC	0.5	0.67	7	0.33	0.29	0.45	37	0.55	0.04	0.08	559	0.92
	NOTEARS	0.44	0.62	5	0.38	0.2	0.33	17	0.67	0.02	0.04	36	0.96
	DAGMA	0.5	0.67	4	0.33	0.24	0.39	21	0.61	0.03	0.057	316	0.94
	Pairwise	0.38	0.54	14	0.45	0.4	0.57	24	0.43	-	-	-	-
	Ours	0.8	0.89	10	0.11	0.4	0.57	24	0.43	-	-	-	-
10000	GES	0.7	0.82	9	0.18	0.42	0.58	47	0.42	-	-	-	-
	PC	0.55	0.71	9	0.29	0.26	0.42	47	0.58	-	-	-	-
	NOTEARS	0.44	0.62	5	0.38	0.19	0.32	19	0.68	0.03	0.058	53	0.94
	DAGMA	0.5	0.67	4	0.33	0.22	0.36	20	0.64	0.033	0.063	214	0.94
	Pairwise	0.47	0.64	17	0.36	0.14	0.25	86	0.75	-	-	-	-
	Ours	0.8	0.89	10	0.11	**0.46**	**0.63**	25	**0.37**	-	-	-	-

Table 1: Results on the Asia (8 nodes, 8 edges), Child (20 nodes, 25 edges), and Neuropathic (221 nodes, 770 edges) causal graphs. Our method obtains state-of-the-art performance on all 3 causal graphs. More results can be found in the appendix.

⚠️ Traditional variable importance methods don’t tell you anything about causality!

These tools are useful for prediction, but not for causal inference!



We've discussed three types of causal relationships.
Going forward, we're going to assume you identified key **confounders** you want to control for, as you estimate the causal impact between a **"treatment"** and an **"outcome"**…

If You Are Doing Causal Modeling…

Think before looking at data - Carefully consider quantities of interest and their relationships using domain knowledge

Stick with a small set of important variables - Only include variables you have domain knowledge about

Understand bivariate relationships - Before modeling, examine relationships between:

- Independent variables with each other
- Independent variables with dependent variable

Identify potential confounders - Clearly identify covariates to control for and those NOT to control for

Assumptions of Causal Inference

Four Key Assumptions:

Temporality - Causes always occur before effects. Treatment must occur before measured outcome. Covariates should occur before treatment.

SUTVA (Stable Unit Treatment Value) - The treatment status of one individual does not affect the potential outcomes of any other individuals.

Positivity - For each level of each covariate, there needs to be some variability in treatment and outcome variables.

Ignorability - All major confounding variables are included in your data. This is tough but necessary for unbiased treatment effect estimates.

Assumption Violations: Example #1

Temporality Violation

Scenario: I want to understand whether frequent emails to customers might impact customer satisfaction.
I have survey data with customer self-reported satisfaction from a year ago, and I use this past month's number of emails for each customer as a proxy for how often we email them generally.
⚠️ **Problem:** Past satisfaction cannot be caused by future emails!
Temporal ordering is violated.

Assumption Violations: Example #2

Positivity Violation

Scenario: I want to see the causal impact of a neighborhood's cleanliness on crime rates, controlling for 20 known confounders.
I pull up an academic dataset with data on 40 distinct neighborhoods.
So, my sample size is 40.
⚠️ **Problem:** 20 covariates with only 40 observations! Severe overfitting risk and positivity violations are likely.

Assumption Violations: Example #3

SUTVA Violation

Scenario: I want to see how releasing a new in-app, multiplayer game through my social media app impacts user engagement. I only want to give it to some test users initially.
With this multiplayer game you can play with anyone who has the social media app by sending them invites. Accidentally, our test users can invite non-test users.
⚠️ **Problem:** Treatment spillover! Test users affect control users through invites, violating independence.

Assumption Violations: Example #4

Ignorability Violation

Scenario: We're curious how a job training program could impact a person's income 3 years in the future.
Unfortunately we don't have lots of data on the participants so we perform a causal inference analysis only controlling for the person's age.
⚠️ **Problem:** Massive residual confounding! Education, work history, location, industry, etc. are all missing.

Metrics for Causal Effects

Counterfactuals with Binary Treatment (additive)

Observed Reality

Experiences 500ms delay on website
Click-through rate: 60%

Alternative Reality

Experiences no delay on website
Click-through rate: 65%

Average Treatment Effect = 60% - 65% = -5%

Counterfactuals with Binary Treatment (ratio)

Observed Reality

Worked with radium for years
20% probability of developing cancer in the coming year

Alternative Reality

Never worked with radium
15% probability of developing cancer in the coming year

Average Treatment Effect = 20% / 15% = 1.3 times higher risk

Important Note on Units of Analysis

You can apply causal inference to **any unit of analysis**:

- People
- Browser sessions
- Webpages
- Clusters of friends (social media data)
- Neighborhoods
- Buildings
- Pharmacies
- etc.

Common Causal Metrics

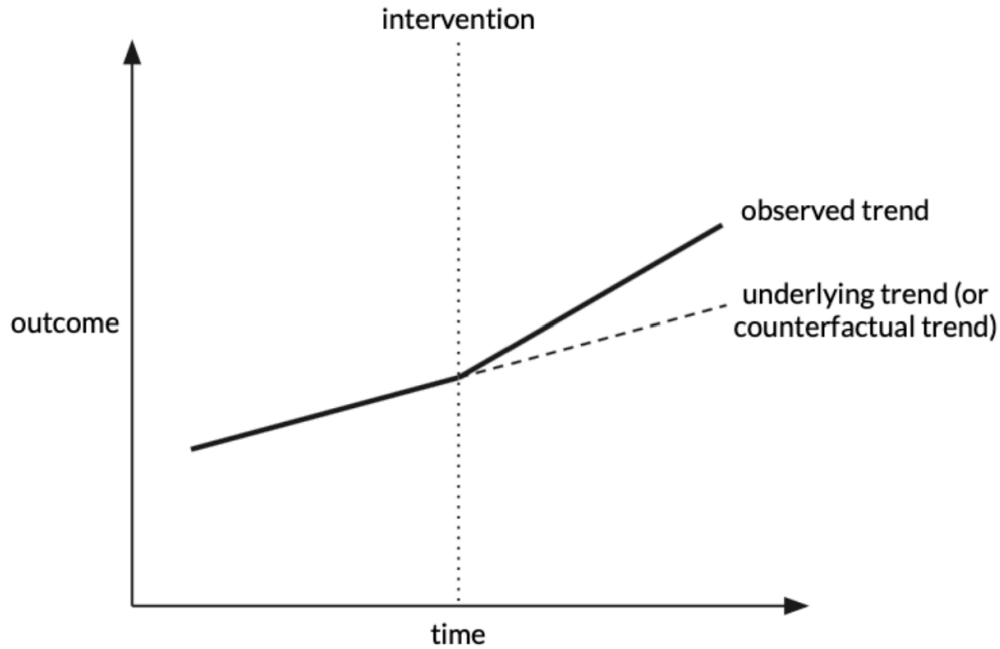
METRIC	POPULATION
ATE - Average Treatment Effect	Effect in entire population
ATT - Average Treatment Effect Among Treated	Effect in treated population
ATU - Average Treatment Effect Among Untreated	Effect in untreated population
ITE - Individual Treatment Effect	Effect for a single unit

Conditional Causal Metrics

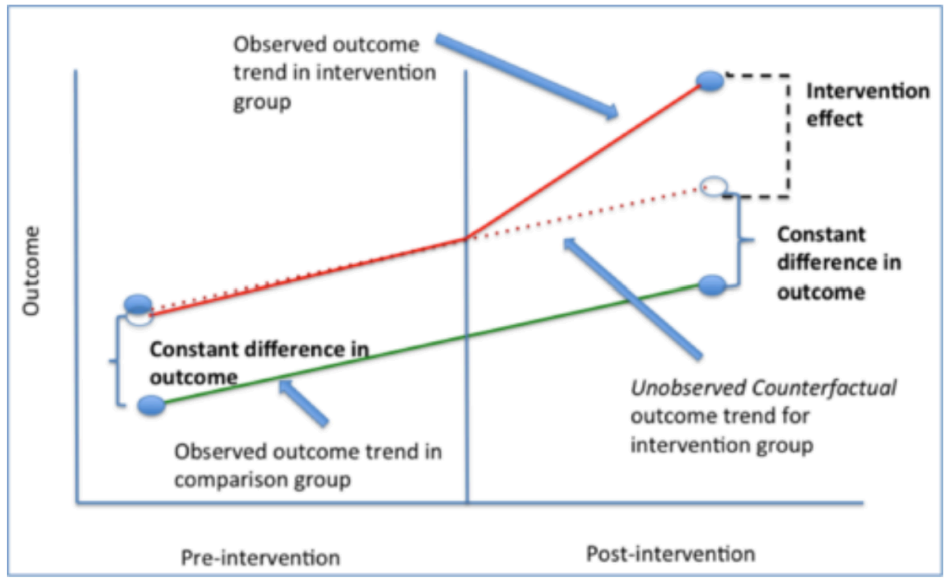
METRIC	POPULATION
CATE - Conditional Average Treatment Effect	Effect segmented by some covariate
CATT - Conditional ATT	Effect in treated, segmented by covariate
CATU - Conditional ATU	Effect in untreated, segmented by covariate

Modeling Approaches for Causal Inference

Interrupted Time Series



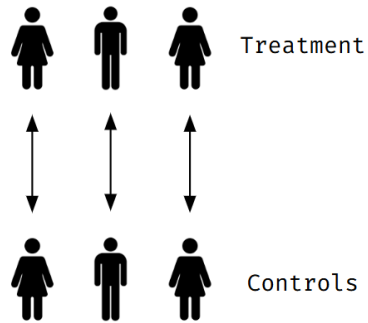
Difference in Differences




```
(Treatment_post - Treatment_pre) - (Control_post - Control_pre)
```

Propensity Score Matching (PSM)

Start with some population from observational data



Divide group into those that received and didn't receive treatment. Match each treatment unit with a similar control unit.

Δ

Compare outcomes values between the two groups

PSM Step 1: Start with Data

ID#	COVAR 1	COVAR 2	TREAT	OUTCOME
1	…	…	1	20
2	…	…	1	15
3	…	…	0	10
4	…	…	0	10
5	…	…	1	20

PSM Step 2: Calculate Propensity Scores

ID#	COVAR 1	COVAR 2	TREAT	**PS**	OUTCOME
1	…	…	1	**0.65**	20
2	…	…	1	**0.33**	15
3	…	…	0	**0.64**	10
4	…	…	0	**0.33**	10
5	…	…	1	**0.97**	20

Use a model to predict `` `treat` `` from covariates

PSM Step 3: Match Units

Match 1

ID#	TREAT	PS	OUTCOME
1	1	0.65	20
3	0	0.64	10

Match 2

ID#	TREAT	PS	OUTCOME
2	1	0.33	15
4	0	0.33	10

Match based on similar propensity scores!

PSM Step 4: Calculate Effect

ID#	TREAT	OUTCOME
1	1	20
2	1	15
3	0	10
4	0	10

Average Treatment Effect = (20 + 15)/2 - (10 + 10)/2 = 7.5

Meta-learners: S-Learner

Key idea: Train a model to predict outcomes, then simulate counterfactuals

S-Learner Step 1: Train model with a set of participants for whom we have complete data

ID#	COVAR 1	COVAR 2	TREAT	OUTCOME
1	…	…	1	20
2	…	…	1	15
3	…	…	0	10
4	…	…	0	10
5	…	…	1	20

Train a model: `` `outcome ~ covariates + treat` ``

S-Learner Step 2: Predict outcome where everyone has treatment = 1

ID#	COVAR 1	COVAR 2	TREAT	OUTCOME	**$\hat{Y}$(TREAT=1)**
1	…	…	**1**	20	**22.5**
2	…	…	**1**	15	**16.0**
3	…	…	**1**	10	**14.0**
4	…	…	**1**	10	**17.0**
5	…	…	**1**	20	**22.5**

S-Learner Step 3: Predict outcome where everyone has treatment = 0

ID#	COVAR 1	COVAR 2	TREAT	OUTCOME	**Ŷ(TREAT=1)**
1	…	…	**0**	20	**18.5**
2	…	…	**0**	15	**14.0**
3	…	…	**0**	10	**11.5**
4	…	…	**0**	10	**13.0**
5	…	…	**0**	20	**19.5**

S-Learner Step 4: Calculate treatment effect

ID#	Ŷ(TREAT=1)	**Ŷ(TREAT=0)**	**CATE**
1	22.5	**18.5**	**4.0**
2	16.0	**14.0**	**2.0**
3	14.0	**11.5**	**2.5**
4	17.0	**13.0**	**4.0**
5	22.5	**19.5**	**3.0**

Average CATE = 3.1

Notebook Exercise #2

Implementing S-Learner By Hand

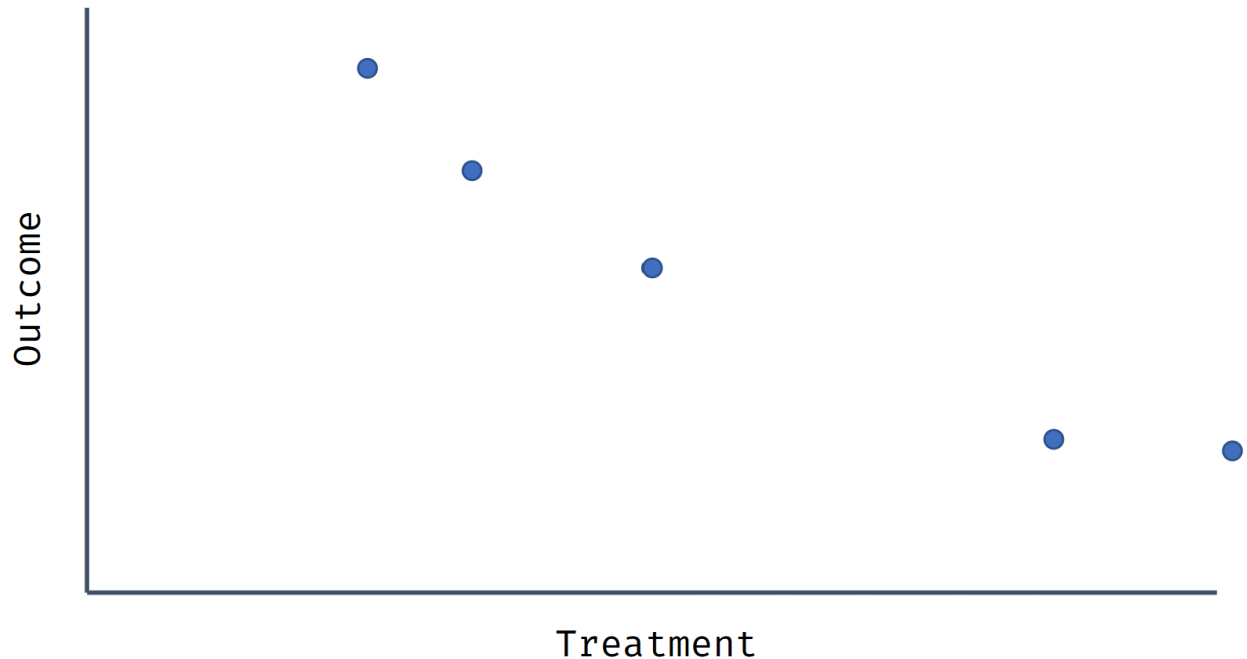
Time to code! 💻

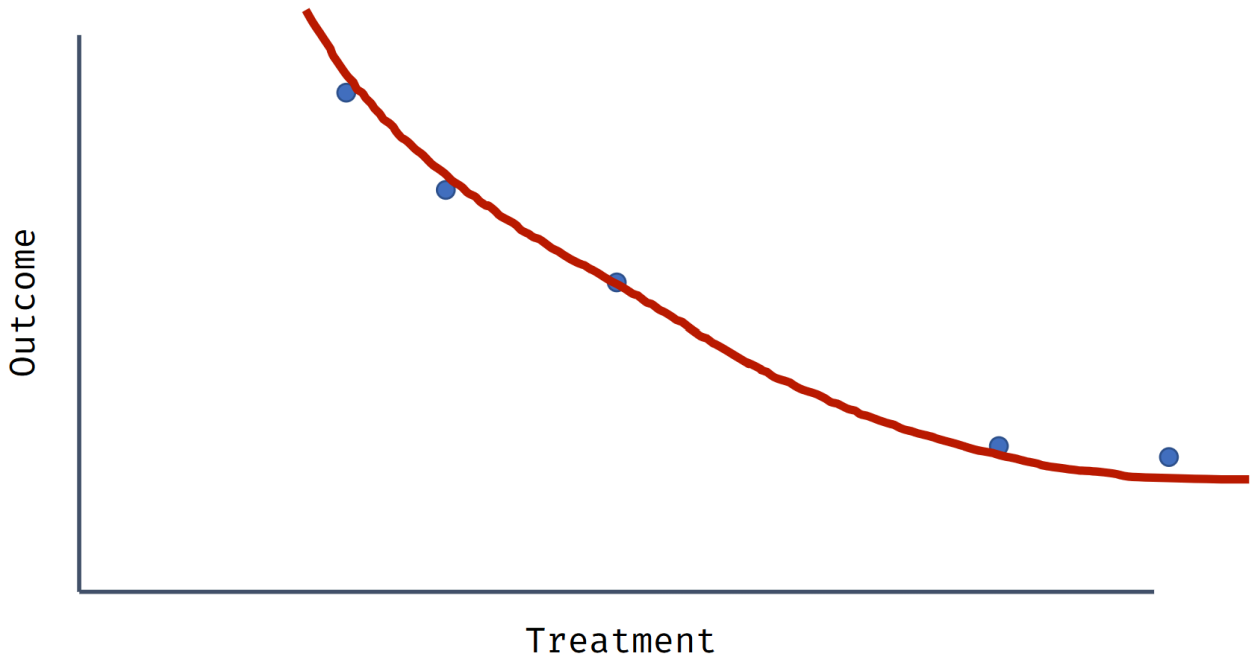
Quick Aside on Continuous Treatments

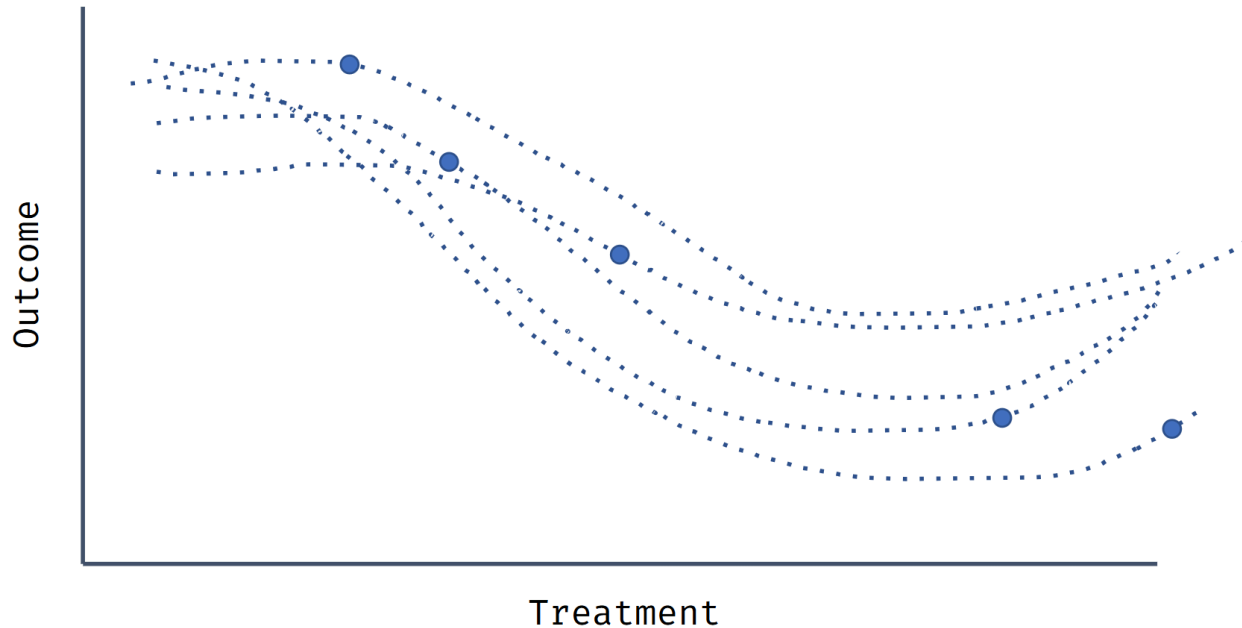
What if treatment isn’t binary?

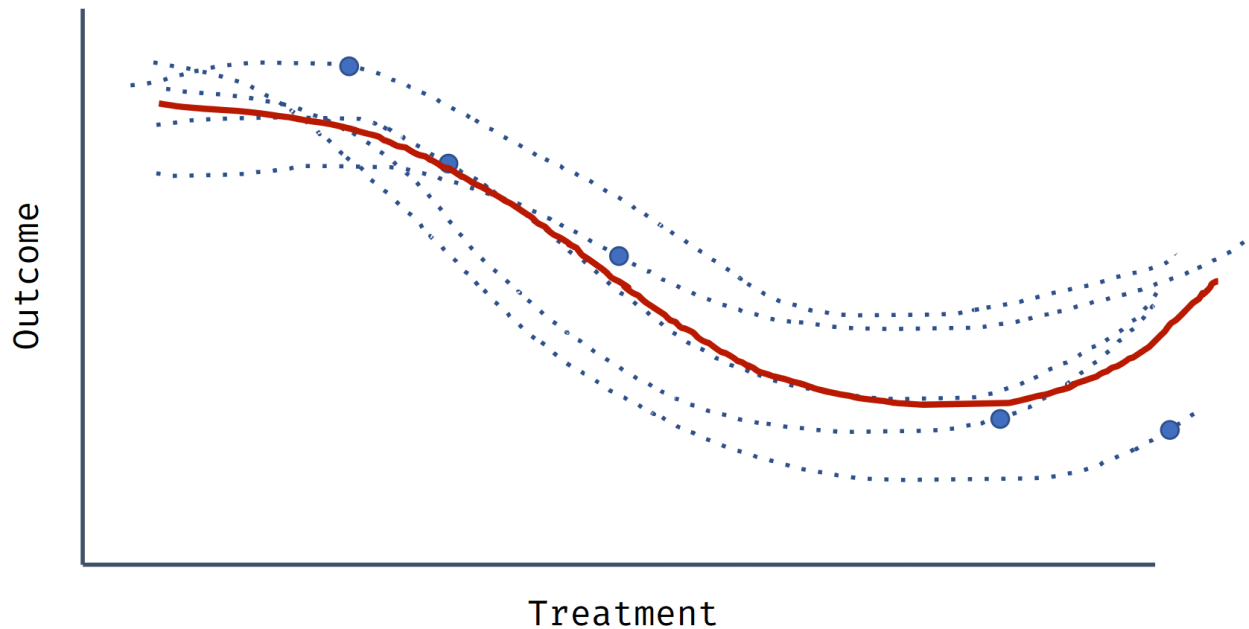
Examples:

- How does the *amount* of advertising spending affect sales?
- How does an increased wait time of X affect satisfaction?


Outcome
Treatment


Outcome
Treatment


Outcome
Treatment


Outcome
Treatment

A demo on `DoWhy` and `BSTS`

Closing Thoughts

The Perils of Multiple Testing

Running many statistical tests inflates your false positive rate!

Statistics

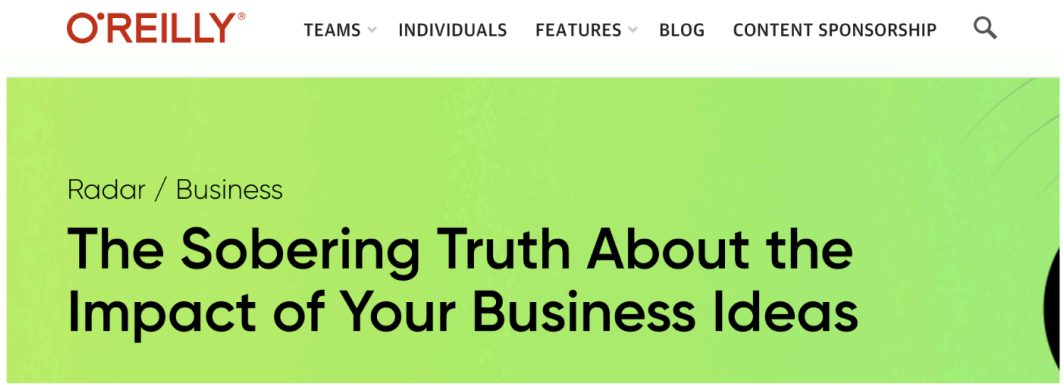
Priya Ranganathan,
C. S. Pramesh[1],
Marc Buyse[2,3]

Department of Anaesthesiology, Tata Memorial Centre, [1]Department of Surgical Oncology, Division of Thoracic Surgery, Tata Memorial Centre, Mumbai, Maharashtra, India, [2]International Drug Development Institute, San Francisco, California, USA, [3]Department of Biostatistics, Hasselt University, Hasselt, Belgium

Common pitfalls in statistical analysis:
The perils of multiple testing

Be Humble!

It’s likely your research or business idea doesn’t work!



Troubleshooting Tips

- Having **domain knowledge** and understanding the data-generating process is often way more productive than just throwing an algo at the problem
- There is value in trying **multiple techniques** to understand their range of estimates (but use p-value correction!)
- You'll **never capture all confounders**, but do aim to capture the major ones
- If your results don't make sense and your code isn't buggy, you're probably **missing a big source of bias**
- Causal inference is powerful but **still not as trustworthy as running a proper experiment**. Approach all results with healthy skepticism.

Thank You! 🎉

Questions?

Resources:

- GitHub: [your-repo-link]
- Marimo Notebooks: [notebook-links]
- Further Reading: Pearl’s "The Book of Why"

Let’s Practice!

Open the Marimo notebooks and let’s get hands-on with causal inference! 🚀

```
cd notebooks
uv run marimo edit 01_causal_graphs.py
```