Causal Graphs:
Applying PyWhy to Go Beyond Explainability

Michelle Yi & Amy Hodler
October 2024

ODSC

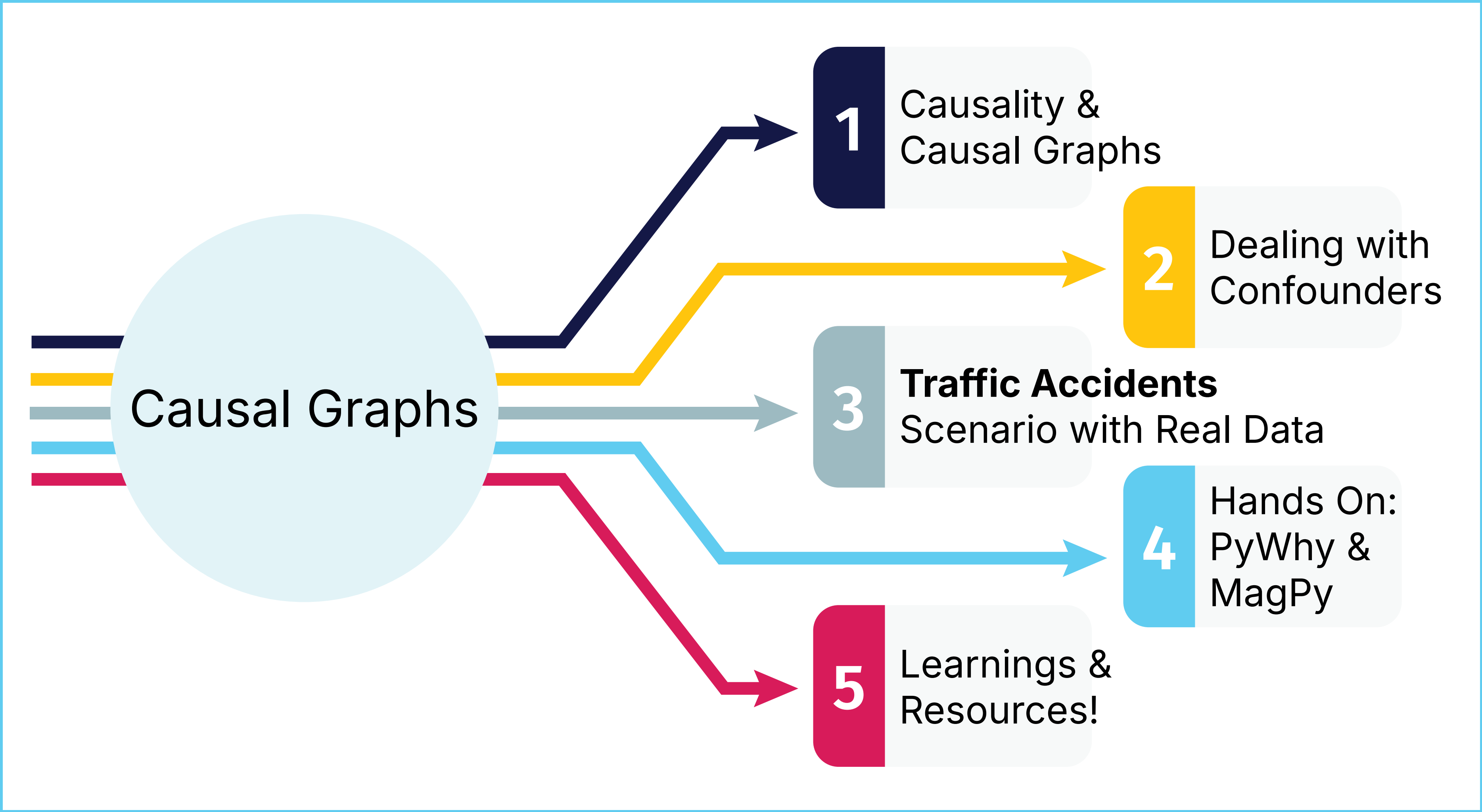
WORKSHOP

MACHINE LEARNING

Amy Hodler
Founder | Consultant | Graph Evangelist
GraphGeeks.org

Michelle Yi
Board Member
Women In Data

Causal Graphs: Applying PyWhy to Go Beyond Explainability

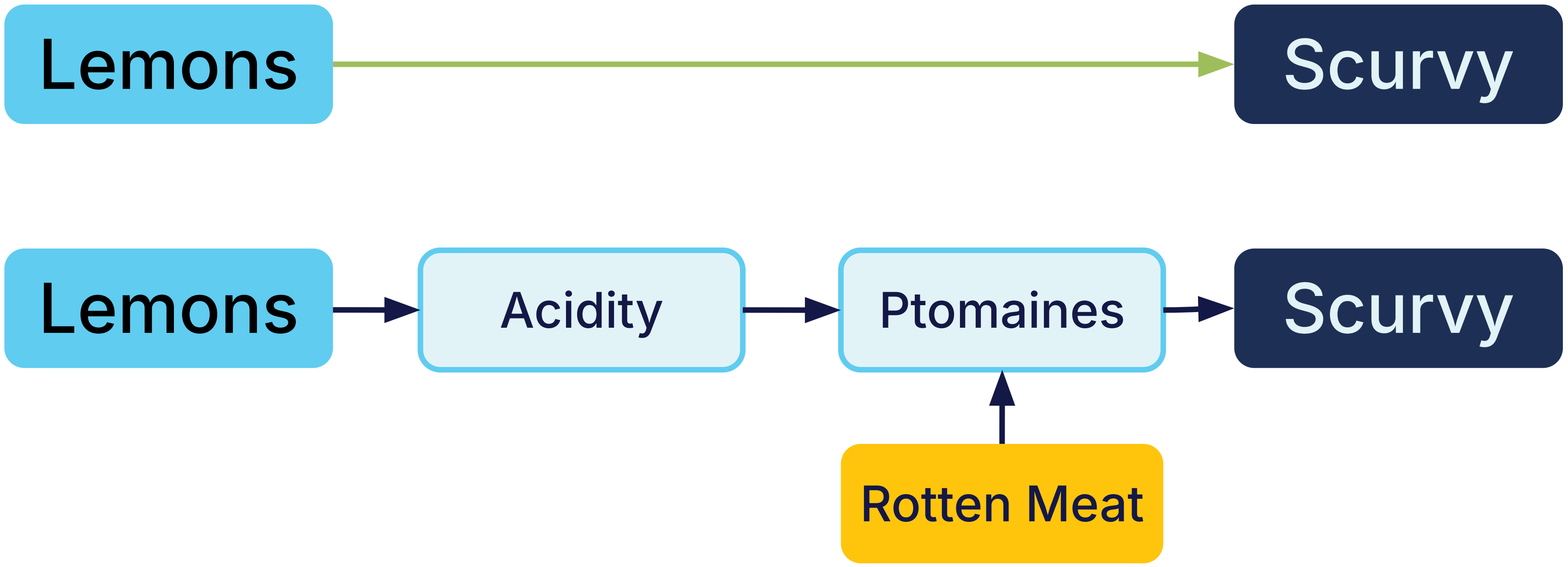

Causal Graphs
1
Causality &
Causal Graphs
2
Dealing with
Confounders
3
Traffic Accidents
Scenario with Real Data
4
Hands On:
PyWhy &
MagPy
5
Learnings &
Resources!

Scurvy Killed Millions of People

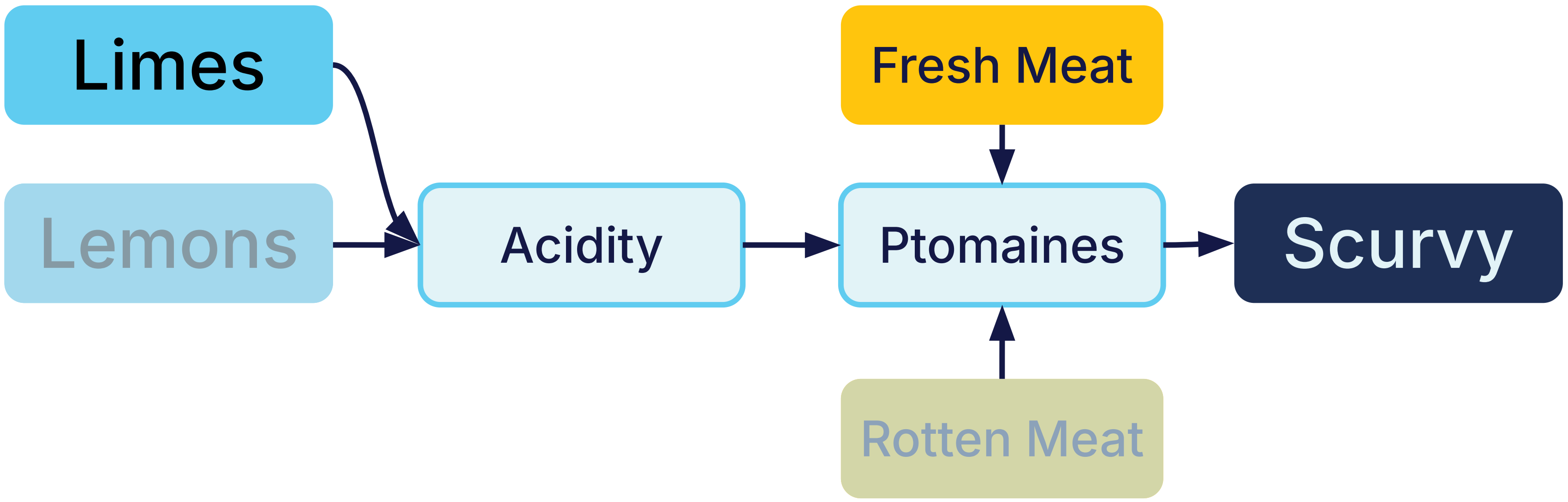
1747 We knew lemons prevented scurvy

Not until 1932 was the cause understood
and we eradicated it as a common disease

James Lind



The leading theory on why lemons cured scurvy was that it neutralized some harmful effects of rotting meat



If acidity is the mediating effect, then we can substitute lemons for limes.

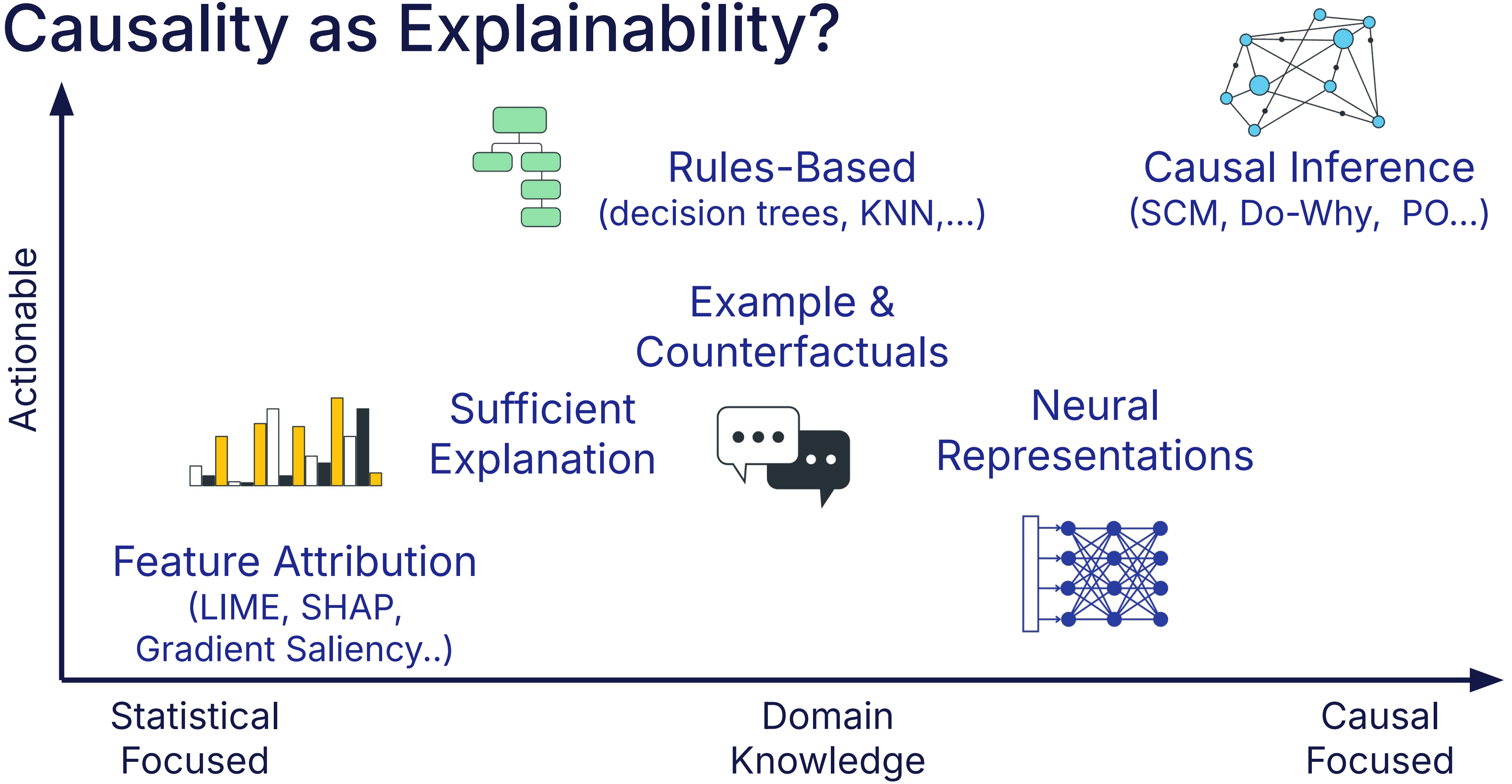
Or we can add fresh sources of meat to avoid the ptomaines in the first place.

"What will we be slapping our foreheads about sixty years from now, wondering how we missed something so obvious?"

Maciej Cegłowski
[Scott and Scurvy]

The Arrow of Why

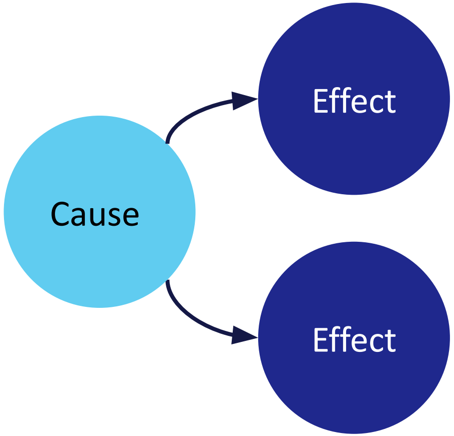
Causality as Explainability?



Causality Goes Beyond Explainability

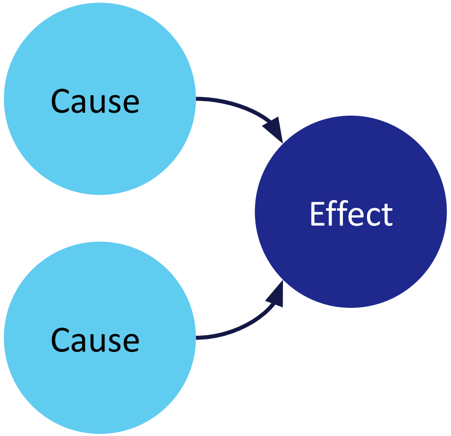
Why does this outcome happen?

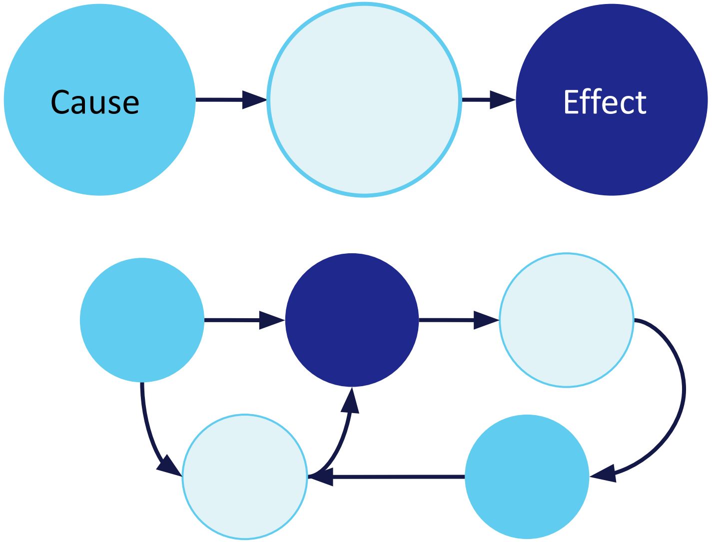
- Finds how a change in one variable causes a change in another
- Causality is inferred using the difference between outcomes



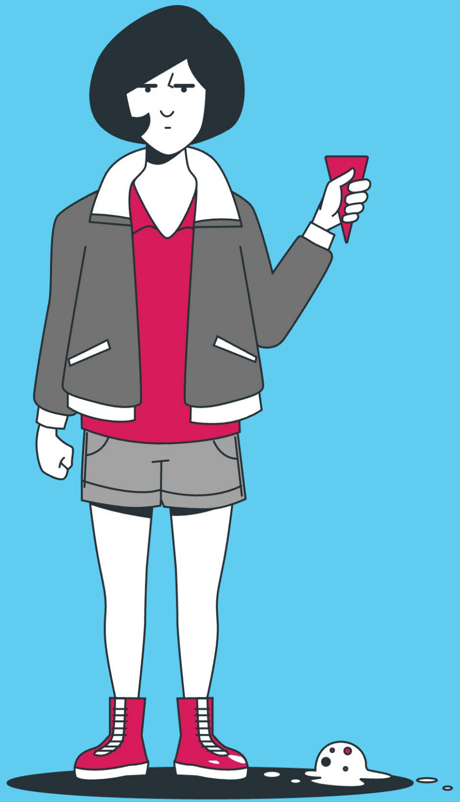
Intervention
Treatment

Outcome





No, You Can't Just Use Plain Machine Learning



Plain ML is Great for Correlations

Observational data tells us about joint distributions

$\mathbf{p(y \mid x)}$ probably of Y given X

$\mathbf{x}$ Anisha buys a padded backpack

$\mathbf{y}$ Anisha buys a laptop

There's a high correlation between Anisha buying a laptop and a fancy padded backpack

But Plain ML is Awful at Causality

Correlations tell us nothing about directionality or interventions

$p(y \mid do(x))$ probably of Y given we **do** X

x Anisha buys a padded backpack

y Anisha buys a laptop

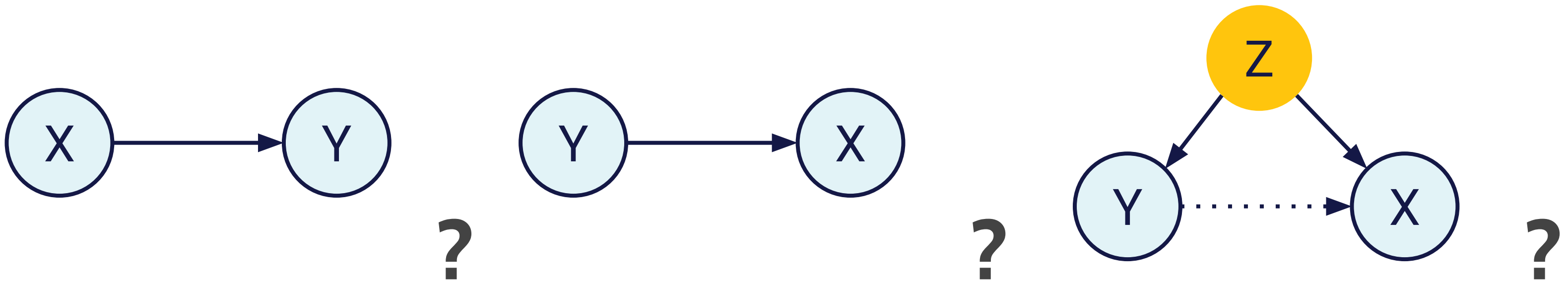
Give her a backpack (intervention) and it's very unlikely she'll buy a laptop

$p(y \mid do(x)) \sim 0$

Give her a laptop (intervention) and she'll likely want to buy a backpack to keep it safe

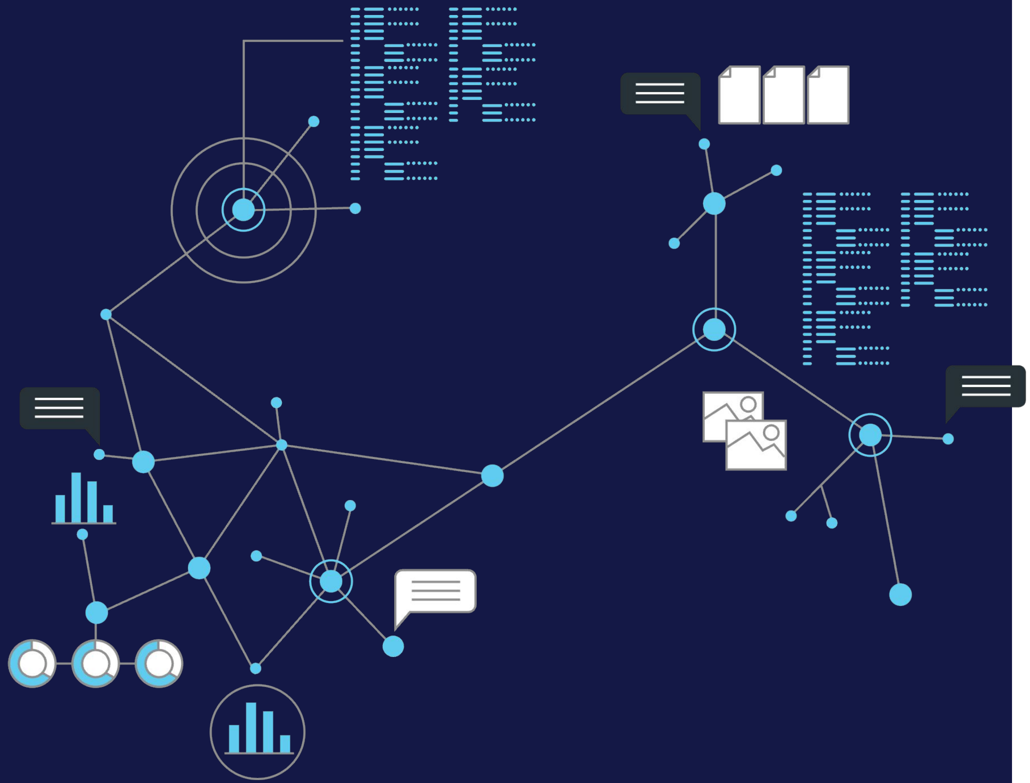
$p(x \mid do(y)) \sim \text{high}$

Causal Structures Are Inductive Biases



These structures are not present in typical ML but are absolutely necessary to make the right decisions!

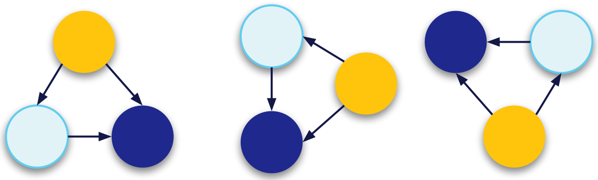
A Better Model



Breakthrough and Turning Point

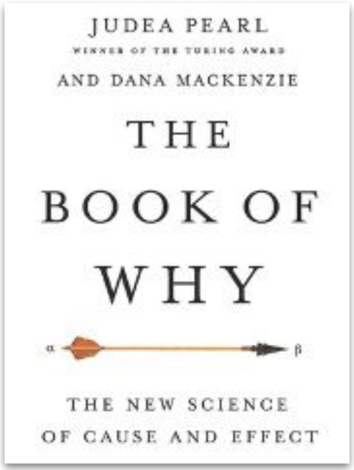
Accessible arguments on causation

Using graphs to model relationships changed the way we think about interventional questions



Creation of a *do*-calculus provided a mathematical way to calculate causal strength of *"doing"* something

Megapost summary:
https://engineeringideas.substack.com/p/megapost-about-causality-the-summary



$P(Y|X)$

to

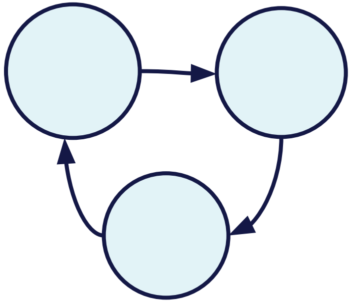
$P(Y|do(X))$

Causal Graphs as a Unifying Model

Directed Acyclic Graphs (DAG) ONLY

No feedback loops: no intervention can cause itself!

Arrow direction illustrates causal influence



Cycles NOT Allowed

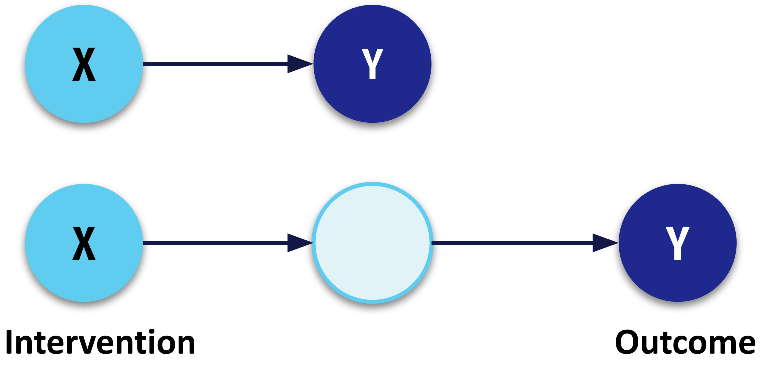
Causal Graphs as a Unifying Model

Directed Acyclic Graphs (DAG) ONLY

No feedback loops: no intervention can cause itself!

Arrow direction illustrates causal influence

Front Door Paths



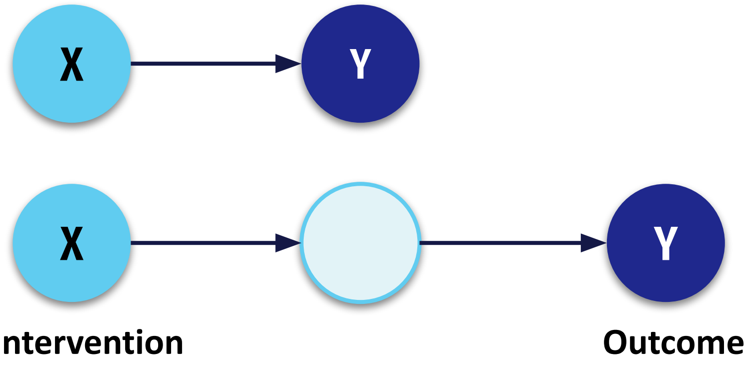
Causal Graphs as a Unifying Model

Directed Acyclic Graphs (DAG) ONLY

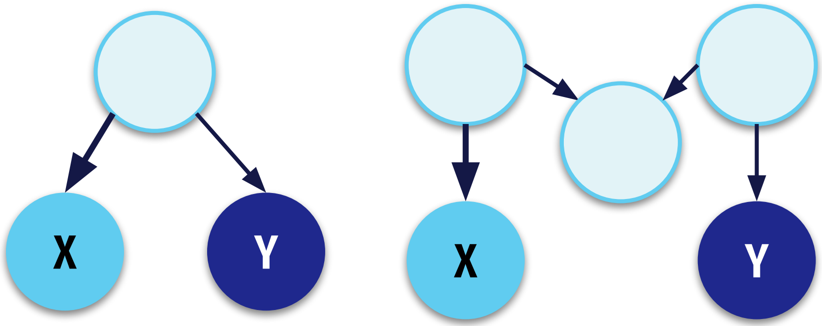
No feedback loops: no intervention can cause itself!

Arrow direction illustrates causal influence

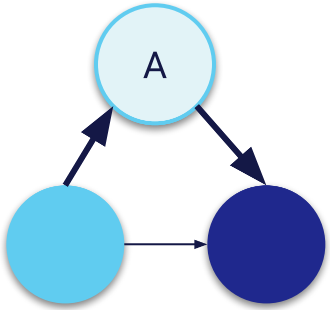
Front Door Paths



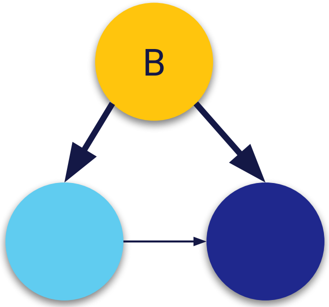
Back Door Paths



Basic Structures of Causal Graphs



A mediator

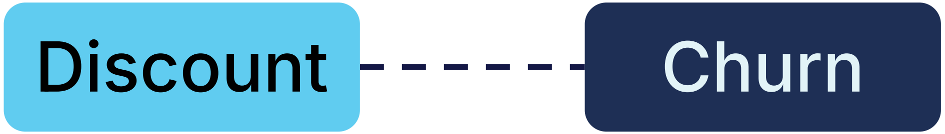


B confounder

Control For
if pointing to
treatment & outcome

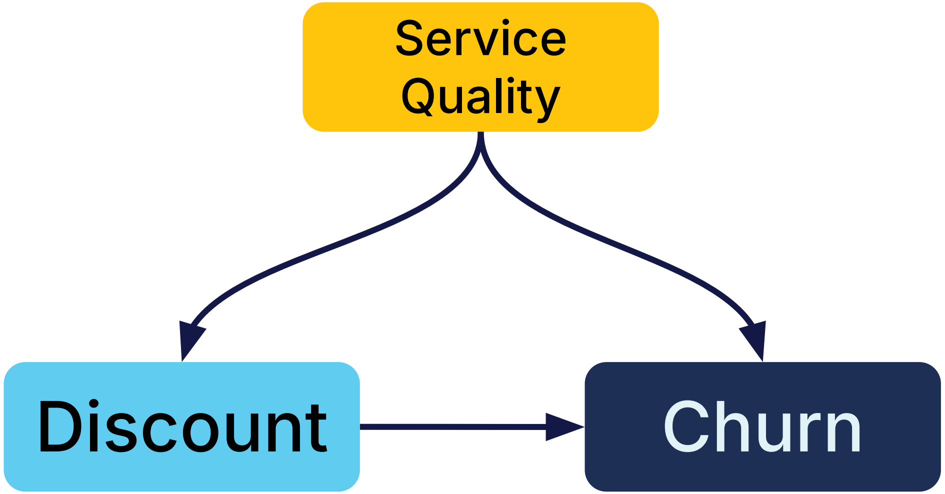
Confounder - Should We Stop Discounts?

Teleco company found the higher discounts offered, the more churn



Confounder - Should We Stop Discounts?

Teleco company found the higher discounts offered, the more churn



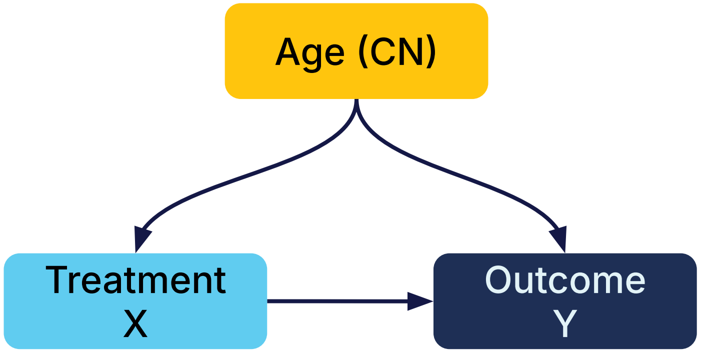
Simpson's Paradox
Customers with problems complain more and receive higher discounts. These customers are also the first ones to cancel their membership when they can.

This creates a positive correlation between our discounts and churn but the causal effect is negative.

Confounders - Back Door Adjustments

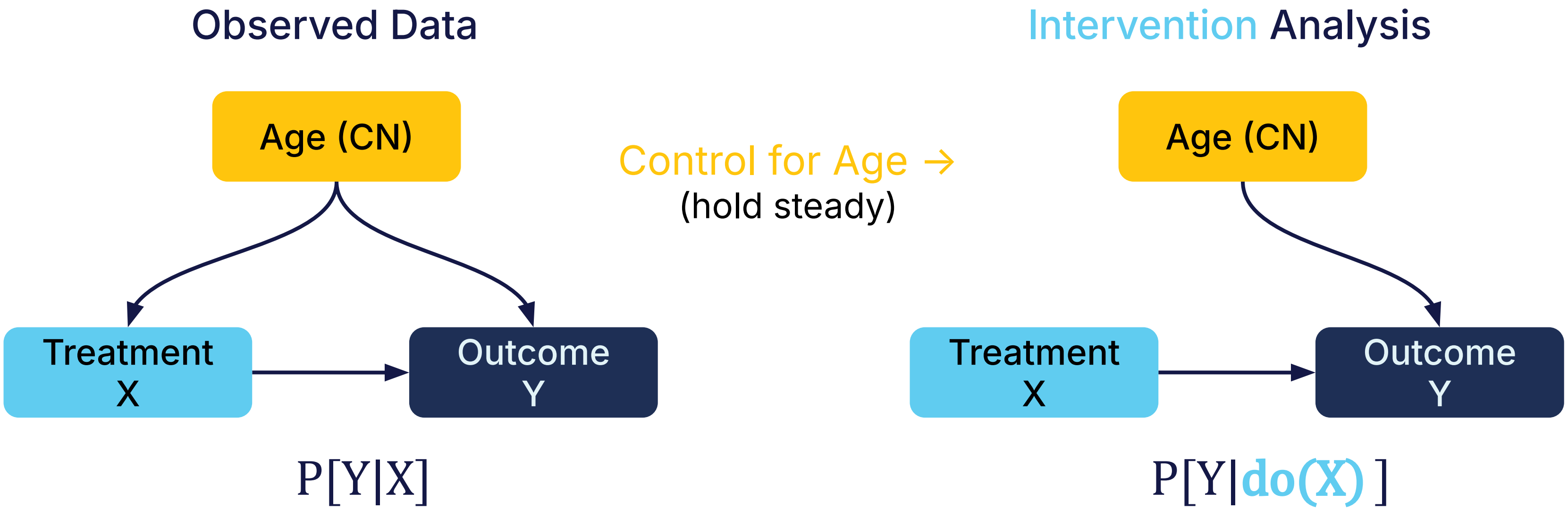
Control Node (CN) 1) Must Not be a child of the treatment your estimating and 2) Must block the path between treatment and outcome.

Observed Data



Confounders - Back Door do-calculus

Interventions represented by the operator, **do(x)**, which "erases" the function while keeping the rest of the model unchanged

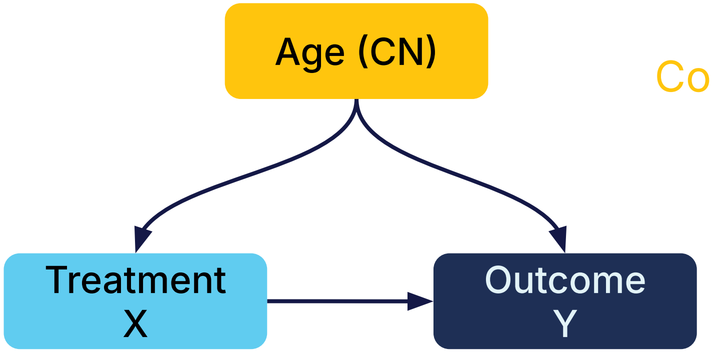


Confounders - Back Door do-calculus

$$P(Y|do(X)) - P(Y)$$

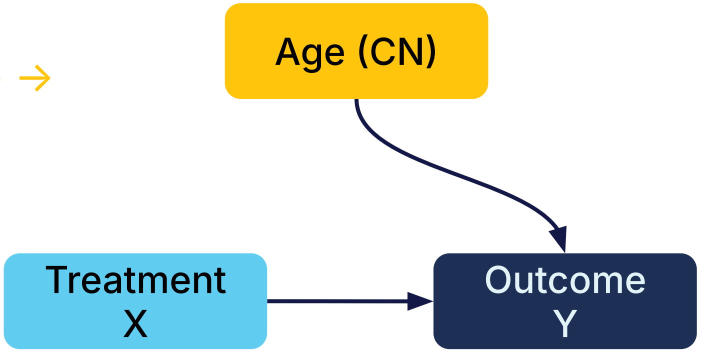
Replaces probability formulas with constant variables to simulate a change in a controlled manner and "closes" the backdoor.

Observed Data



Control for Age →
(hold steady)

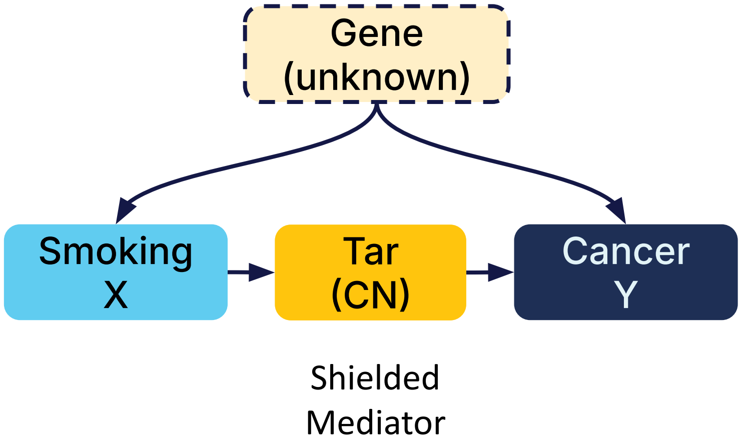
Intervention Analysis



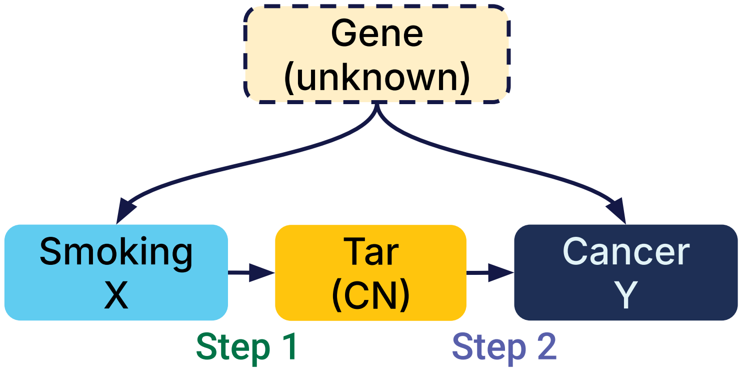
Confounders - Front Door Adjustments

Control Node (CN) 1) Must intercept all directed paths from treatment to outcome 2) No unblocked backdoor path from treatment to CN 3) all backdoor paths from CN to outcome are blocked by the treatment.

Observed & Unknown Data



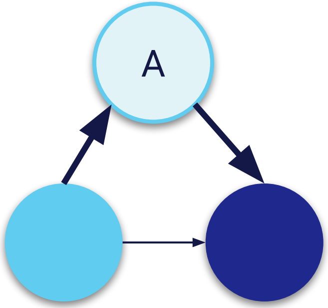
Intervention Analysis



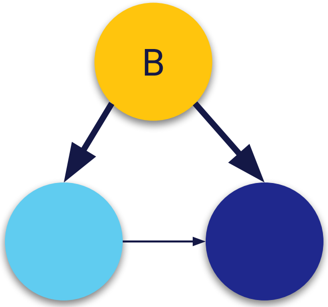
Step 3 - Sum over Mediator and chain together

$$P[Z|do(X)]\ P[Y|do(Z)]$$

Basic Structures of Causal Graphs

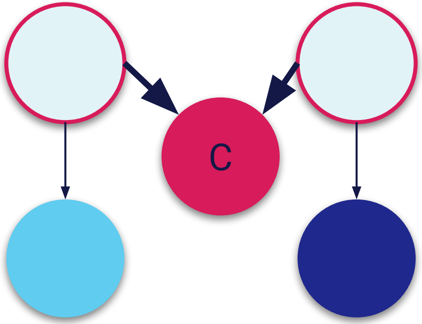


A mediator



B confounder

Control For
if pointing to
treatment & outcome

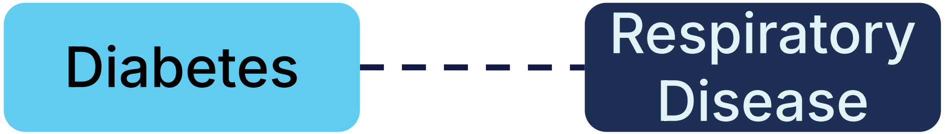


C collider

Do Not Control
As it introduces correlations
that require further controls

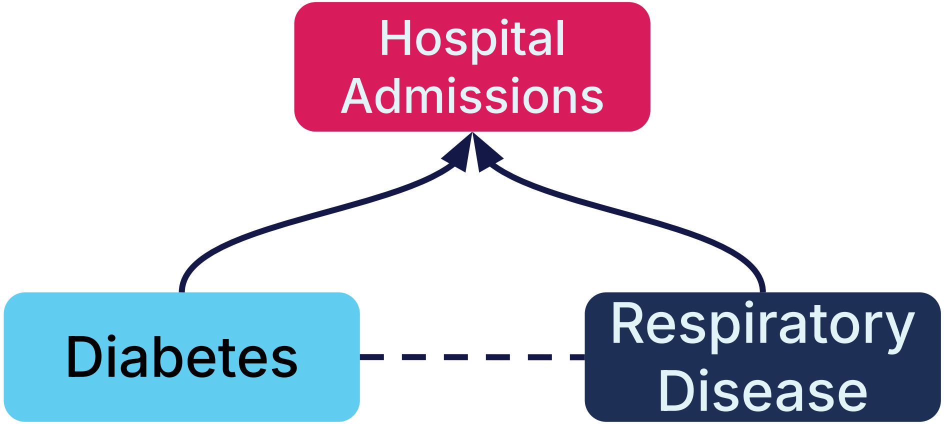
Collider - Should We Get the Flu?

Hospital admissions showed a negative correlation between diabetes and respiratory illness



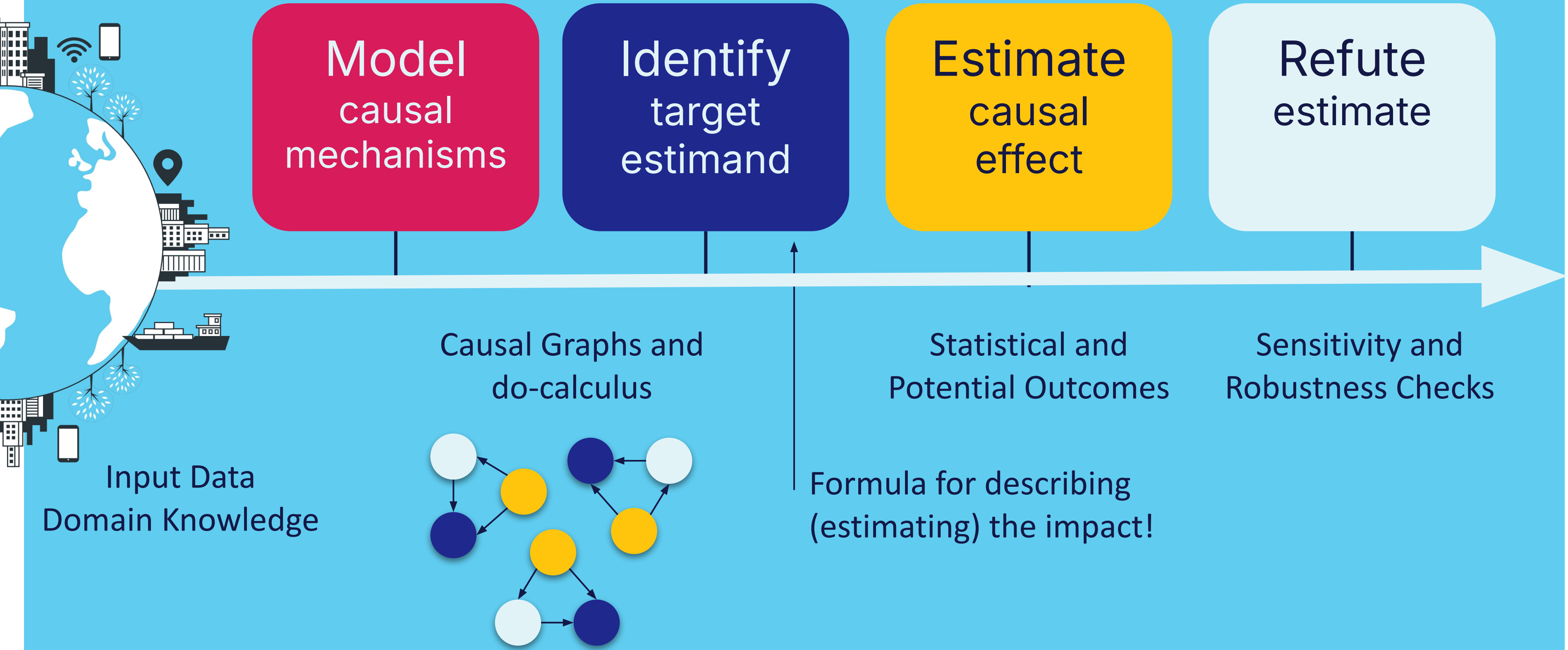
Collider - Should We Get the Flu?

Hospital admissions showed a negative correlation between diabetes and respiratory illness

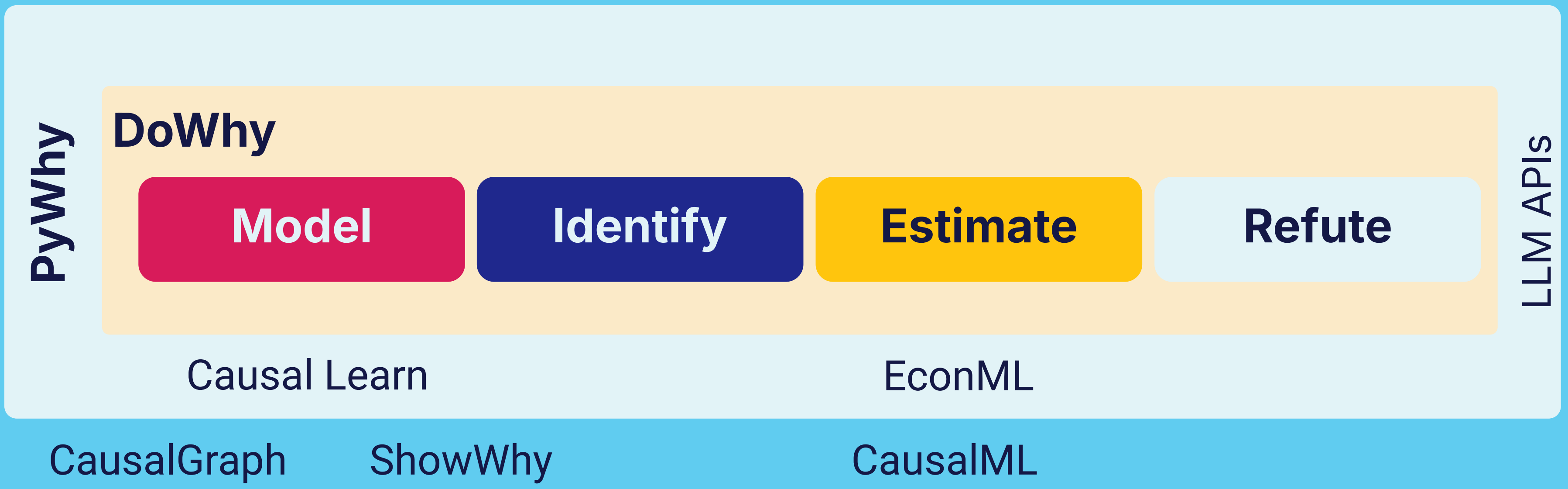


Conditioning on a collider, the hospital admissions, created a correlation between both risk factors, even though there's a priori no causal relationship between them.

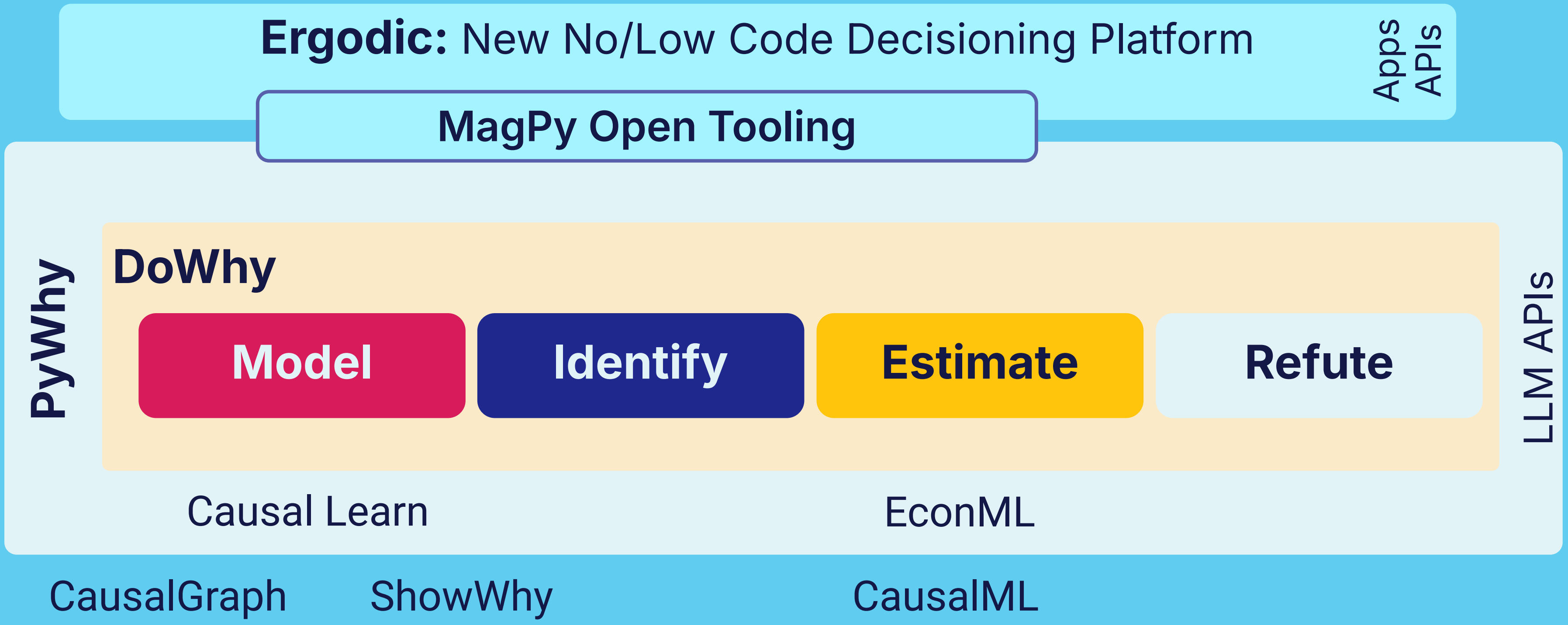
DoWhy Process for Causal Inference



DoWhy, PyWhy, MagPy, and Other Tools



DoWhy, PyWhy, MagPy, and Other Tools

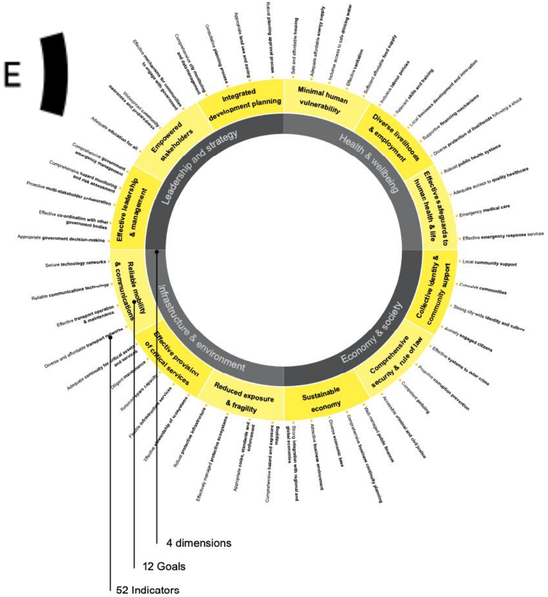


Can We Make Streets Safer?

From City Resilience to Traffic Accidents

City Resilience Index - Practical Framework
www.cityresilienceindex.org

CITY RESILIENCE INDEX



Pared down big questions

- Subcomponents: Leadership, Infrastructure, **Health**, Economy
- Used a credible measurement framework
- Evaluated various data sources
- Map available data to measurable goals
  San Francisco Open City Data www.sf.gov/data/

Reducing the impact of traffic accidents as a **desirable, measurable, actionable** (hopefully) goal

SF.GOV


The San Francisco Standard

Crazy driving, road deaths are rampant. SF candidates vow to crack down

By Gabe Greschler

Published Oct. 16, 2024 • 5:42pm

What would have the biggest impact to traffic safety in San Francisco?

Is there a clear intervention that would reduce accident severity?

What policy changes might help save lives?

Cleaned Accident Data

Vehicle Group

Road Conditions

Time of Day

Collision Type

Weather Conditions

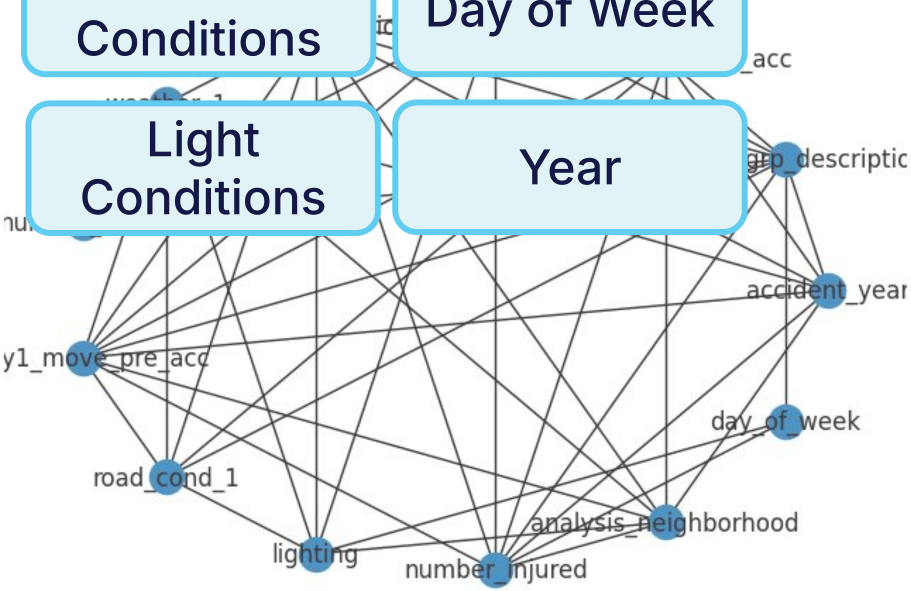
Day of Week

Proceeding Action of 1

Light Conditions

Year

Proceeding Action of 2



Fatalities

Collision(injury) Severity

Injured

Our Approach

Model	Identify	Estimate	Refute
Model causal relationships as a graph comparing different methods Consider/update for confounders	Evaluate different data types to measure the effect of variables on injury severity and fatalities	Find the causal strength of interventions on injury severity and fatalities	Test robustness of findings by adding randomness, checking new vs estimated effect and the p values
PyWhy: GCM MagPy: PC skeleton (A* not for categorical)	MagPy: Mixed Data Oracle	PyWhy: Causal Arrow MagPy: Random Forest	PyWhy

github.com/yulleyi/odsc-west-2024

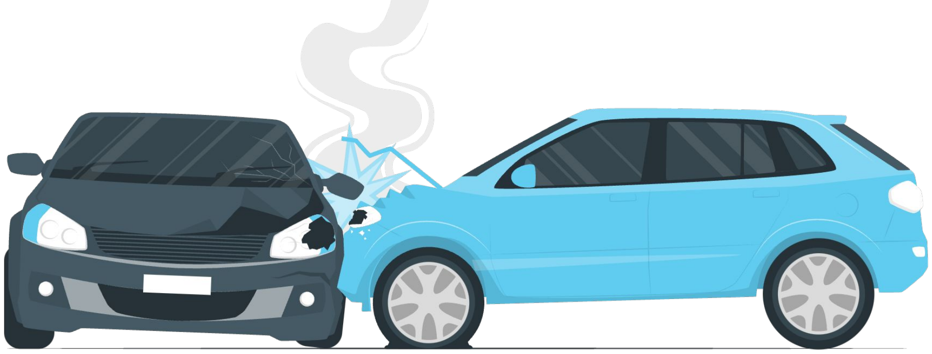
Causal Insights: Reducing Accident Severity

Lighting is a major influence on severity of injuries and fatalities

A simulation showed the potential to save over 700 lives if lighting issues were eliminated (5% reduction)

- Consider the additional injuries, services, and economic impacts!

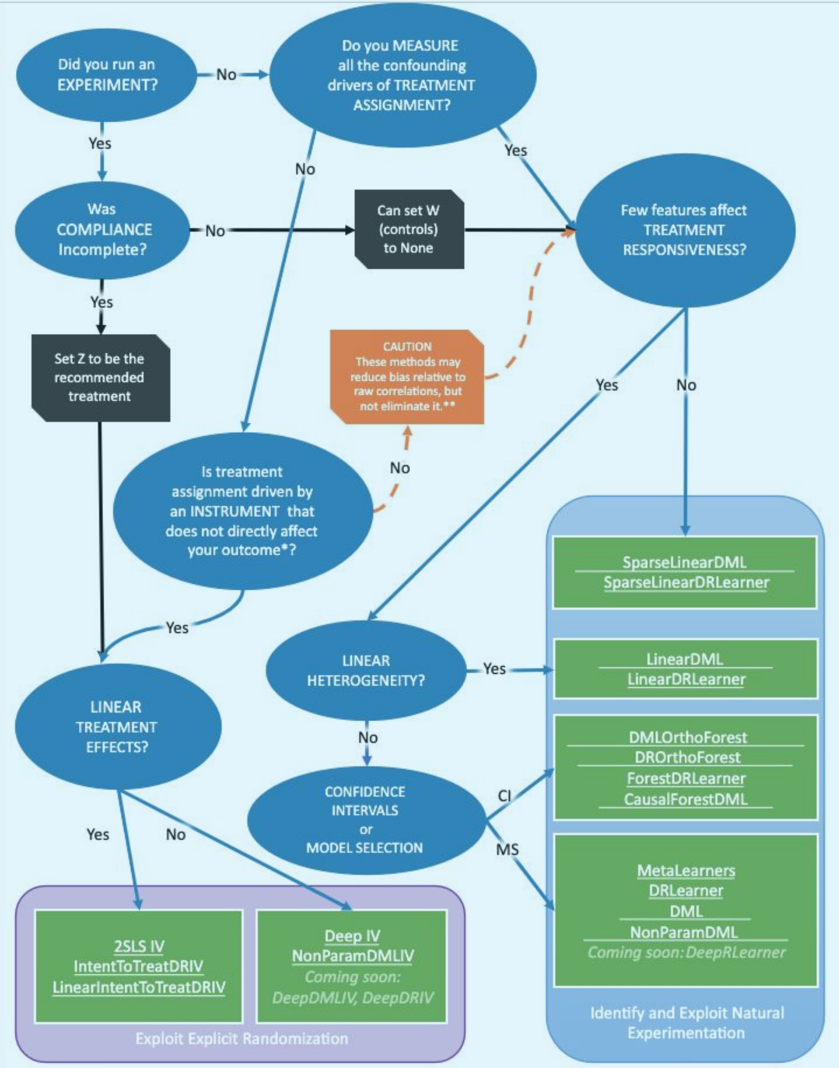
Biggest reduction in high severity accidents is car-only accidents vs pedestrians (U-Turns & Driving Off the Road)



Lots of opportunity for further analysis, connecting more data sources and evaluating change over time

What We Learned

Um...Yeah, This Isn't Helpful



Start Simple and Add Layers

DoWhy

- Graph Causal Model (GCM) - New
- Find relationship strengths
- Simple models (linear regression)
- Limited to averages, not a lot of tuning

MagPy

- Causal discovery & modeling
- Testing of the causal graph
- Estimate interventions
- Mixed variable types

EconML

- Play with other models (🌲🌲, DML, +)
- More & deeper controls, cross validation
- No robustness testing → So fit the log back in DoWhy

Don’t Forget

Categorical Data is Cool but...
More difficult/slower to evaluate
Mixed variables easier in MagPy

Include Domain Experts!
Confirm assumptions on causal links, missing data, confounders

Causal Graphs Take Time
Use tools like MagPy, CausalGraph, Casual Learn

Questions Matter A Lot
Changes data used & causal graph
You will iterate!

One Outcome at a Time
Can’t observe the treatment and not having a treatment in one pass

Missing Data
Beware of spurious correlations and deal with confounders

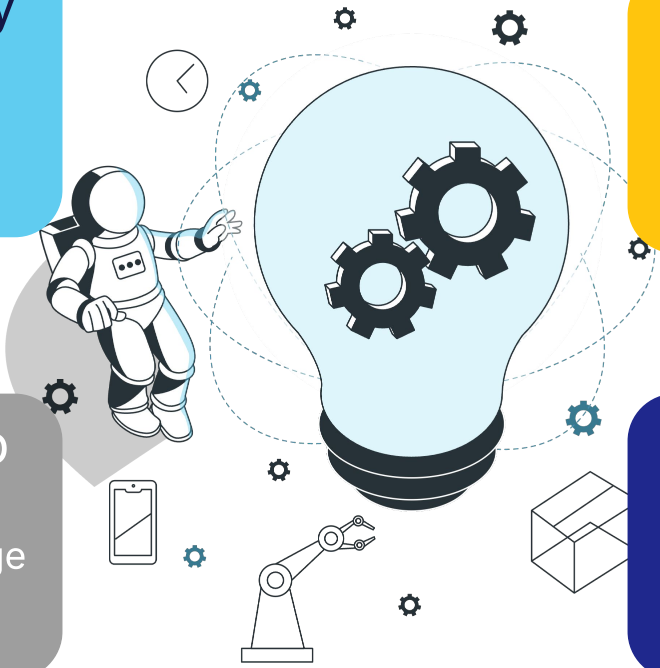
Causal Inference is Required for Better Decisions

Answering Why

Distinguish cause and effect
Find biggest influencers
Inherent explainability

Actionable Results

What changes outcomes?
Estimate impact of an action
Supports decision intelligence



Human-in-the-Loop

Closest to human inference
Integrates domain knowledge
Emphasize human action

Detecting Problems

Uncover poor associations
Is more data needed?
Highlight biased features

Resources

- Don’t be intimated, the maths are done for you!
  - There are multiple ways to do things, just try it out
- Finding good data is a challenge but DoWhy has some curated
  - Synthetic Healthcare Data Example - Synthea
  - SF Open City Data www.sf.gov/data/

Slides and Notebook -github.com/yulleyi/odsc-west-2024

- PyWhy pywhy.org/
- MagPy github.com/ergodic-ai/magpy/
  - Ergodic sign up - ergodic.so/
- Helpful causal inference guide tinyurl.com/5e4auenh
- PyWhy Discord tinyurl.com/365d7e37
- GraphGeeks Discord tinyurl.com/hrjanc3p

GraphGeeks

Start Asking *Why*

WOMEN IN DATA

GraphGeeks

Special Thanks 🙏

Andre Franca
Patrick Blöbaum
Jason Grafft

Michelle@Generationship.ai
Amy.Hodler@GraphGeeks.org

Available for Training