

# Applied Causal Inference Powered by ML and AI

Vasilis Syrgkanis  
MS&E, Stanford

## Instructors



**Vasilis Syrgkanis**  
vsyrgk stanford edu

## Course Assistants



**Hui Lan**  
huilan stanford edu



**Johannes Ferstad**  
jof stanford edu

# A Data Science Tail

Credit: [joint blogpost with Scott Lundberg, Eleanor Dillon, Jacob LaRiviere and Jonathan Roth](#)

# Somewhere in the world right now...

- (M)anager: “Build a model that predicts whether a customer will renew their product subscription”
- (D)ata (S)cientist: “I’ll collect many factors from our database that I believe are predictive of renewal”
- $X = \{customer\ discount, ad\ spending, customer's\ monthly\ usage, last\ upgrade, bugs\ reported\ by\ a\ customer, interactions\ with\ a\ customer, sales\ calls\ with\ a\ customer, and\ macroeconomic\ activity\}$
- DS: “Using last year’s data, I fitted a state-of-the-art ML model (xgboost; gradient boosted forest) to predict  $y = \{renewal\}$  from  $X$ !”



```
[2]: X, y = user_retention_dataset()  
      model = fit_xgboost(X, y)
```

# So what...

- DS: “It learned a function  $f: X \rightarrow y$  that represents the relationship between the variables  $X$  and the outcome  $y$ !”
- M: “Fantastic! How accurate does it predict when given new data it hasn’t seen?”
- DS: “It gives the correct answer 99% of the time!”
- M: “Fantastic! It’s a great model! We can use it to project next year’s revenue!”
- M: “Oh; Maybe we can also see what it learned and try to prevent churn proactively!”
- DS: “Yeah! I know an amazing new explainable machine learning tool called SHAP; you give it any model and it returns what variables were important and in which direction they influence the outcome.”
- M: “Let’s see it!”

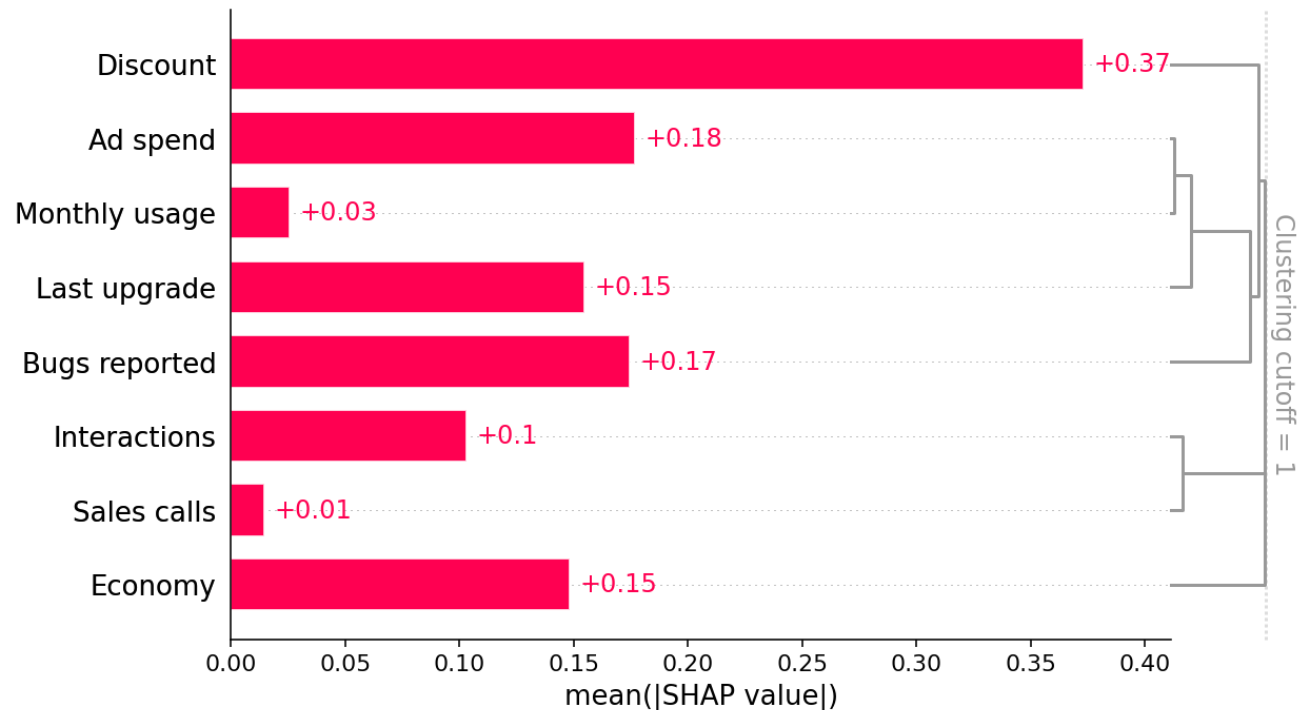
# The important factors

```
import shap

explainer = shap.Explainer(model)
shap_values = explainer(X)

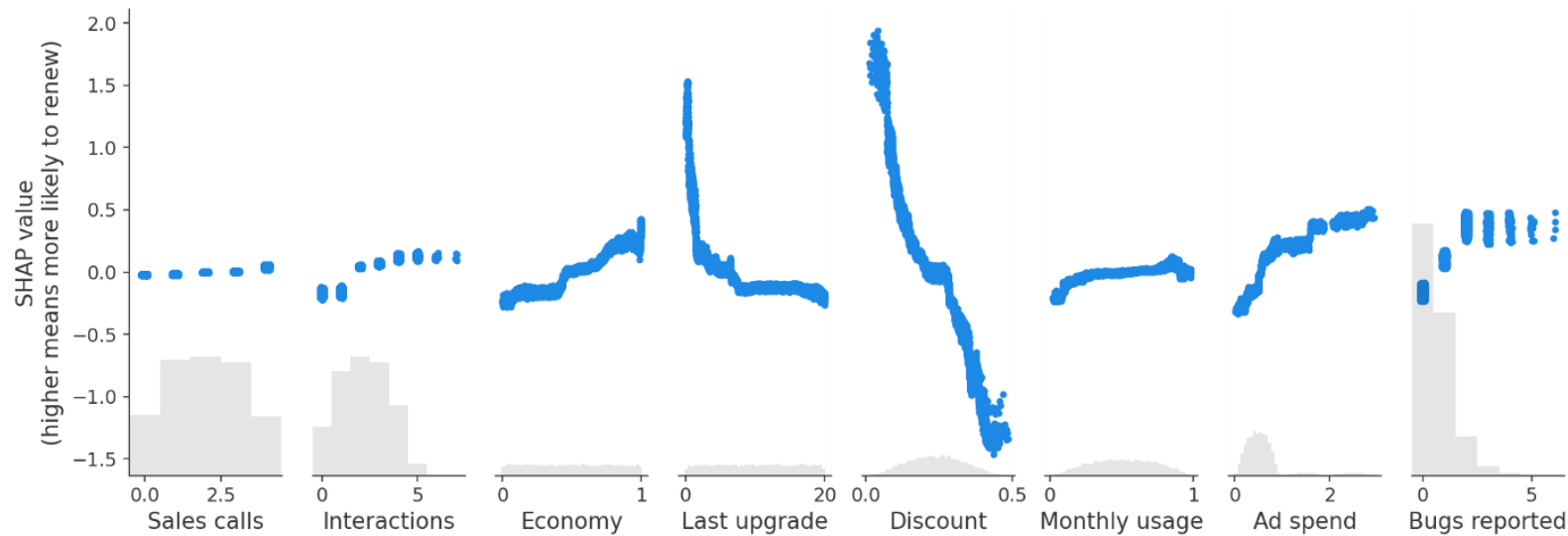
clust = shap.utils.hclust(X, y, linkage="single")
shap.plots.bar(shap_values, clustering=clust, clustering_cutoff=1)
```

- DS: “It seems that discounts and ad spend are important! Also bugs!”
- M: “Great let’s see how much each one affects the outcome?”



# The awkwardness

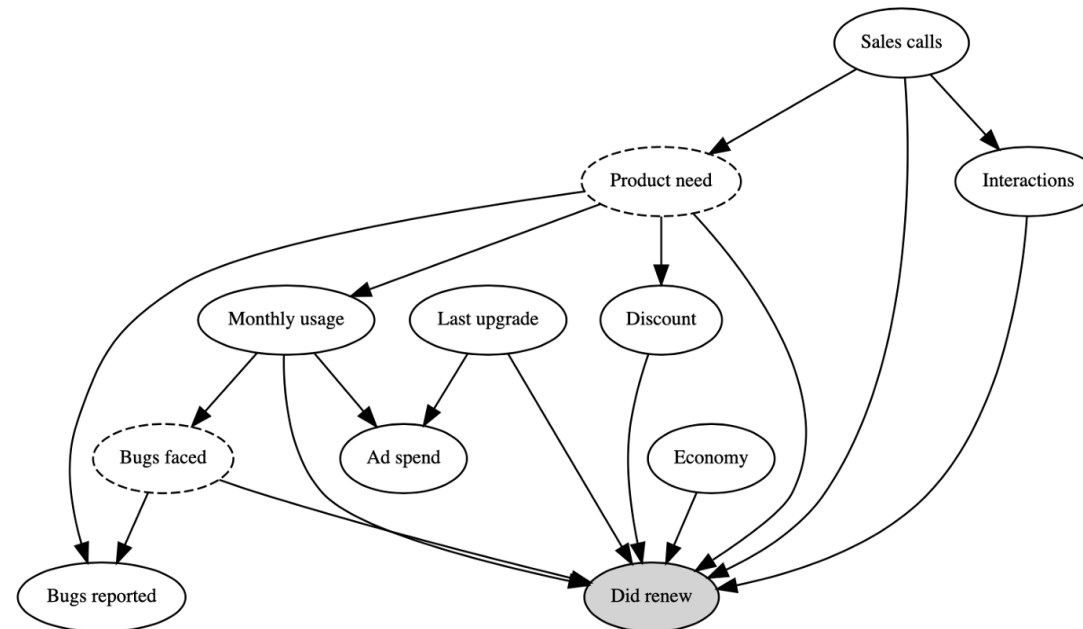
```
shap.plots.scatter(shap_values)
```



- DS: “So larger discounts reduces renewal! Also, more bugs lead to renewal! Oh, and ads are very important!”
- M: “Great let’s increase prices, add more bugs and spam everyone!”

# What happened?

- Business expert:
  - “Users with high usage who value the product are more likely to report bugs and to renew their subscriptions.”
  - “The sales force tends to give high discounts to customers they think are less likely to be interested in the product, and these customers have higher churn.”



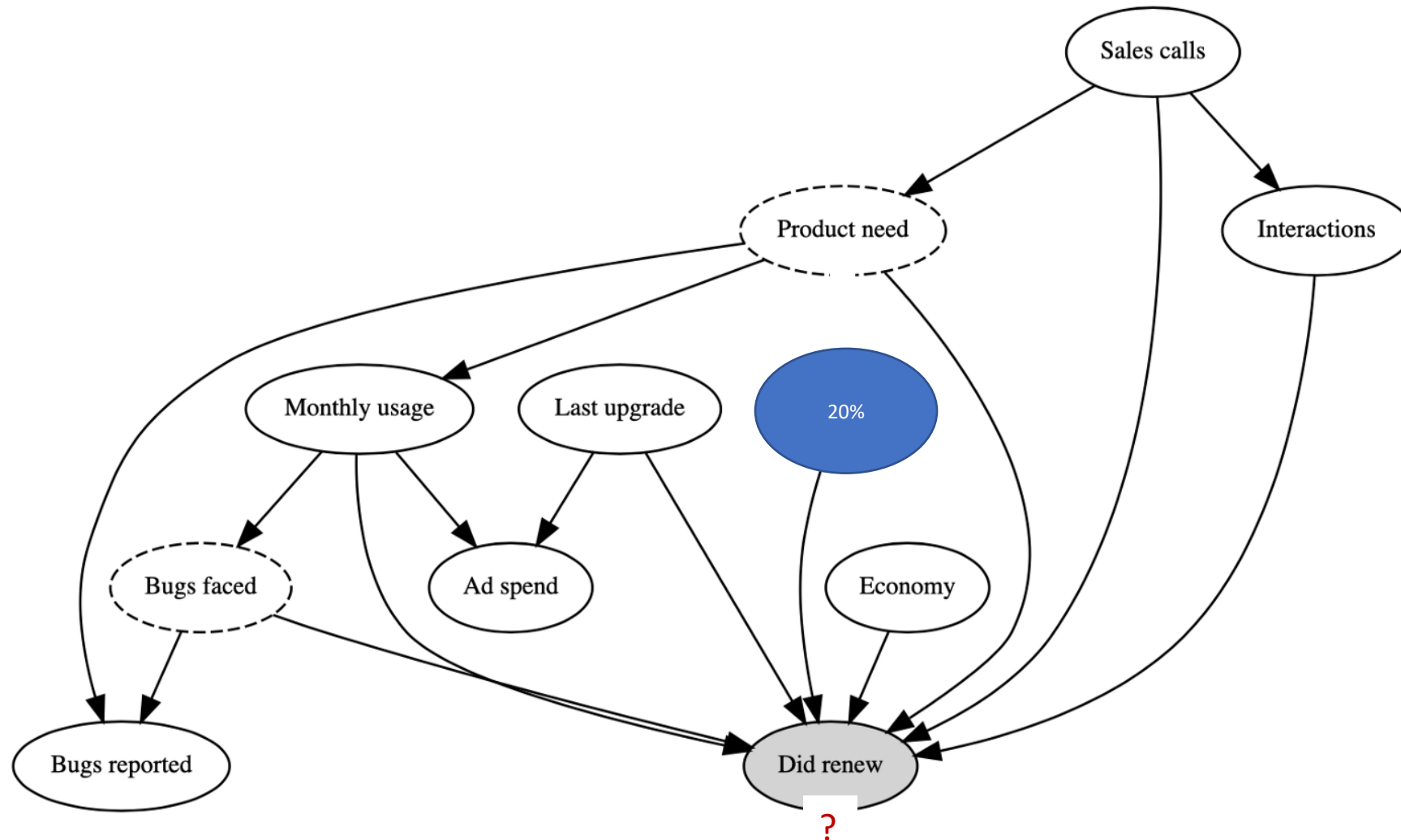
Are the counter-intuitive  
relationships that the model  
learned problematic?



# It depends

- If our goal was to simply project next year's revenue (without any intervention), then these relationships are not problematic
- Such tasks that ask for “projecting” some outcome variable in the absence of any intervention are “predictive tasks”
- If our goal is to understand what would happen if we intervene in one of the variables to increase retention, then these relationships are problematic
- Such tasks that ask for “what-if” or “counterfactual” values of an outcome under some intervention are “causal tasks”

# Causal/Interventional Question



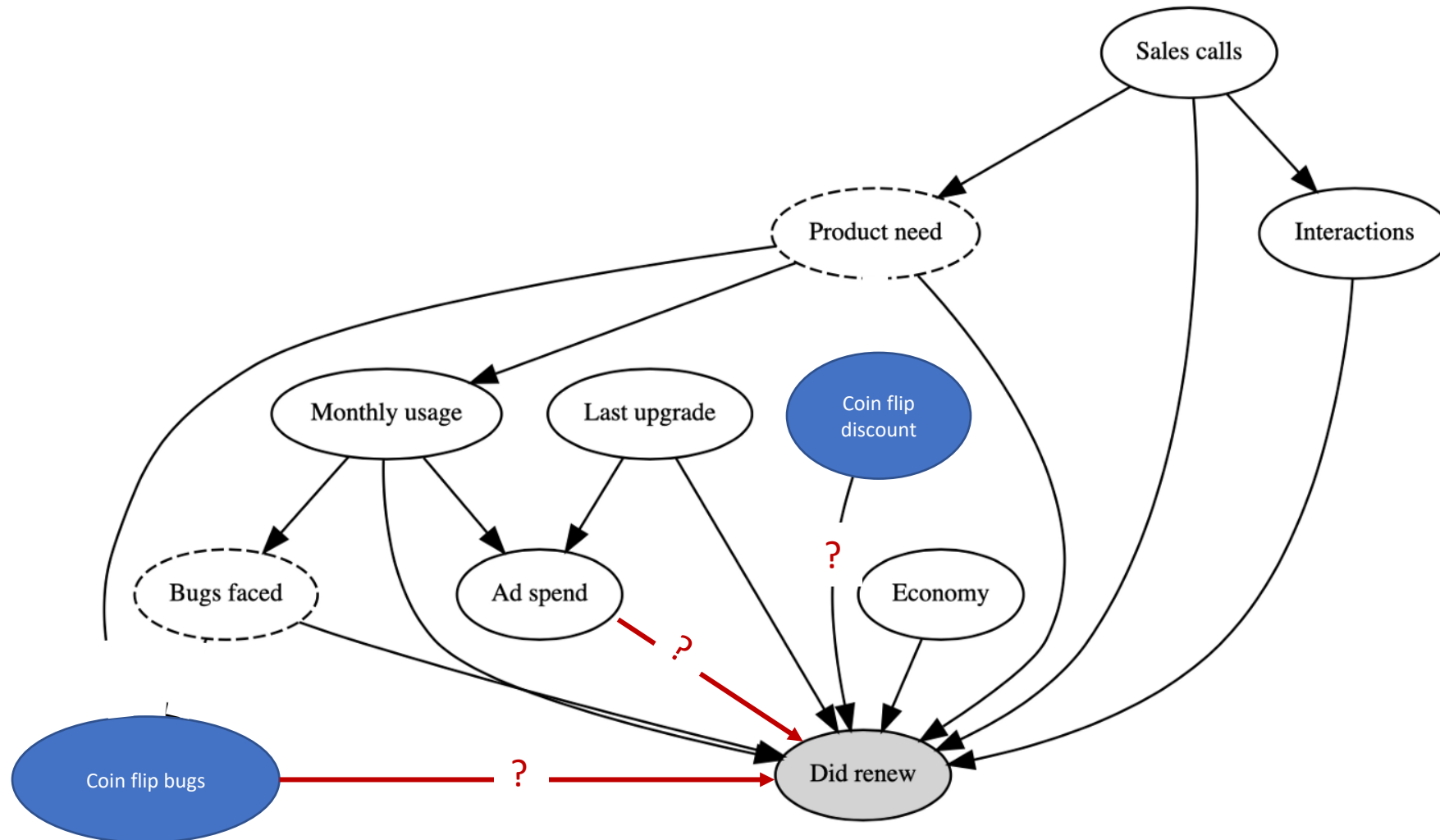
How do we answer causal questions?

Experiments:  
The ideal solution

# Why the ideal

- By randomizing the treatment have two populations that are statistically indistinguishable other than that they differ in the assigned treatment
- Any statistical differences in the outcome between the two populations can then safely be attributed to the treatment

# What would an A/B test do?



# Limitations

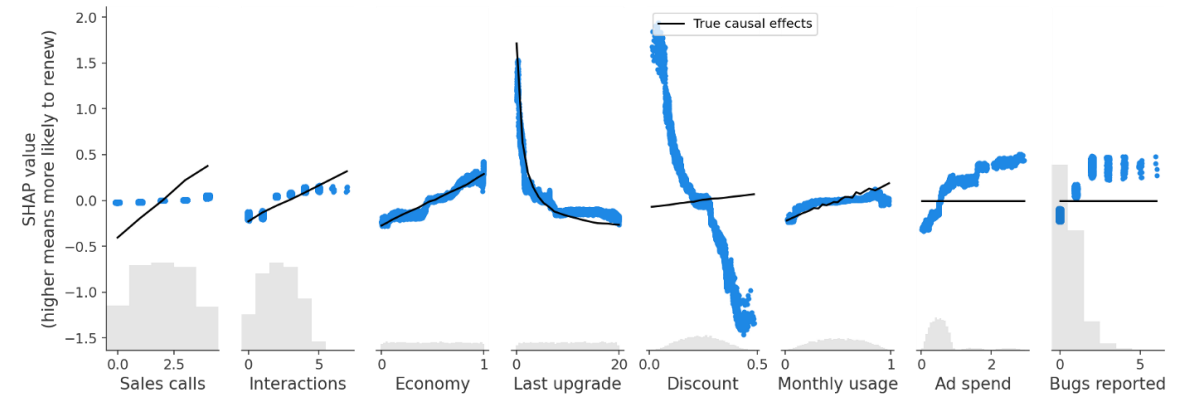
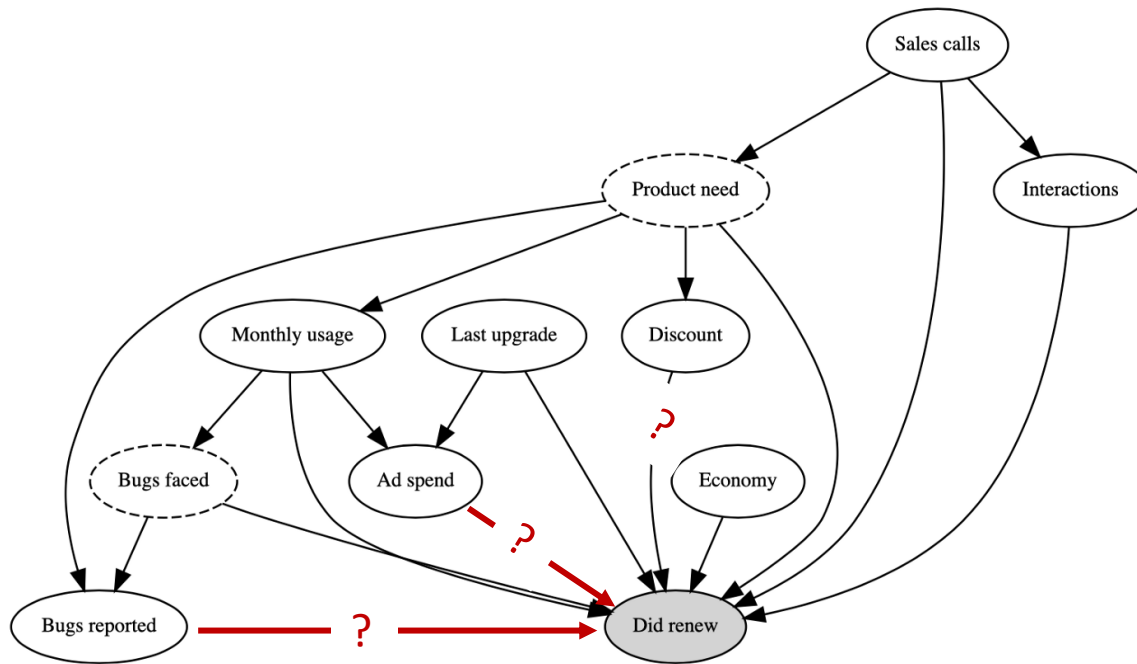
- Ethical
- Practical
- Generalizability

Observational data and studies



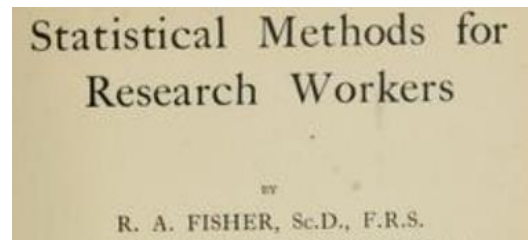
Require domain knowledge

of the high-level mechanisms that underlie the data collection process



# Causal Inference

- Addresses interventional (what-if) statistical questions and the identification of causal relationships from data



*Journal of Educational Psychology*  
1974, Vol. 66, No. 5, 688-701

## ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN RANDOMIZED AND NONRANDOMIZED STUDIES<sup>1</sup>

DONALD B. RUBIN<sup>a</sup>

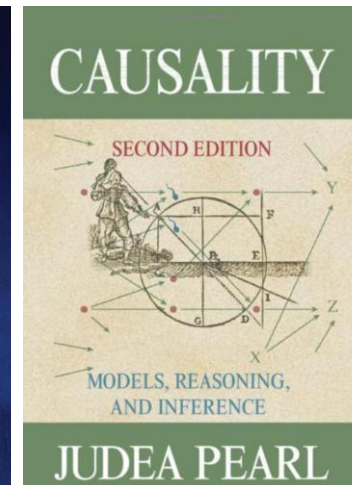
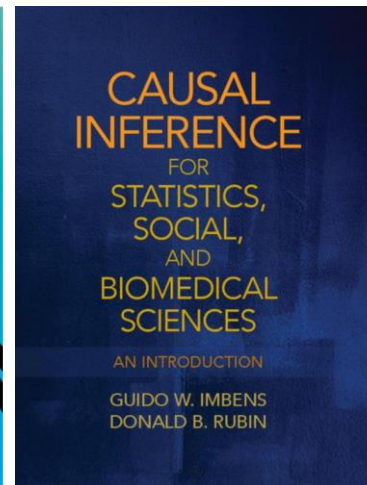
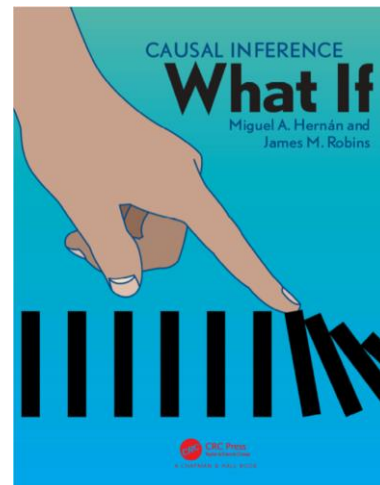
*Educational Testing Service, Princeton, New Jersey*

*Statistical Science*  
1996, Vol. 11, No. 4, 455-480

## On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.

Jerzy Splawa-Neyman

Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original, which  
appeared in *Roczniki Nauk Rolniczych* Tom X (1923) 1-51 (*Annals of Agricultural Sciences*)



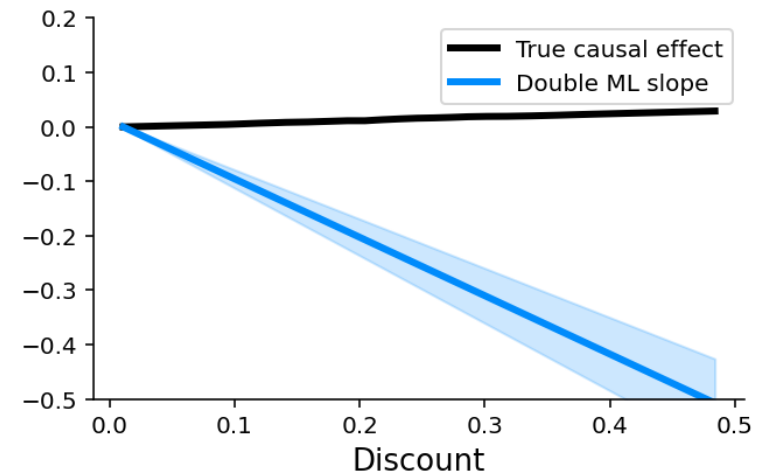
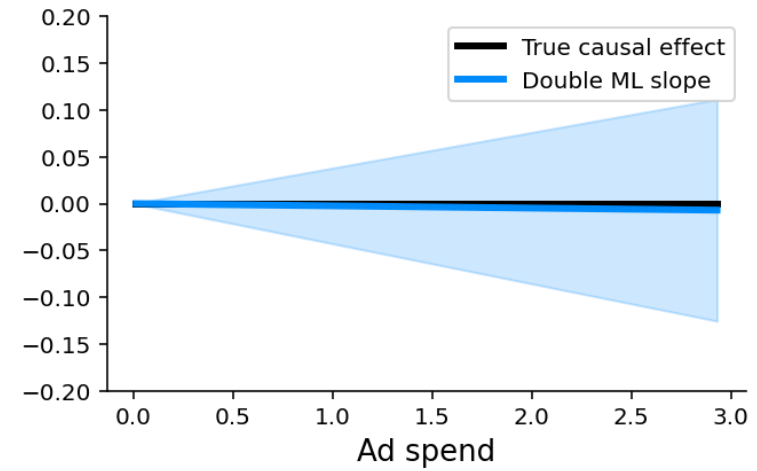
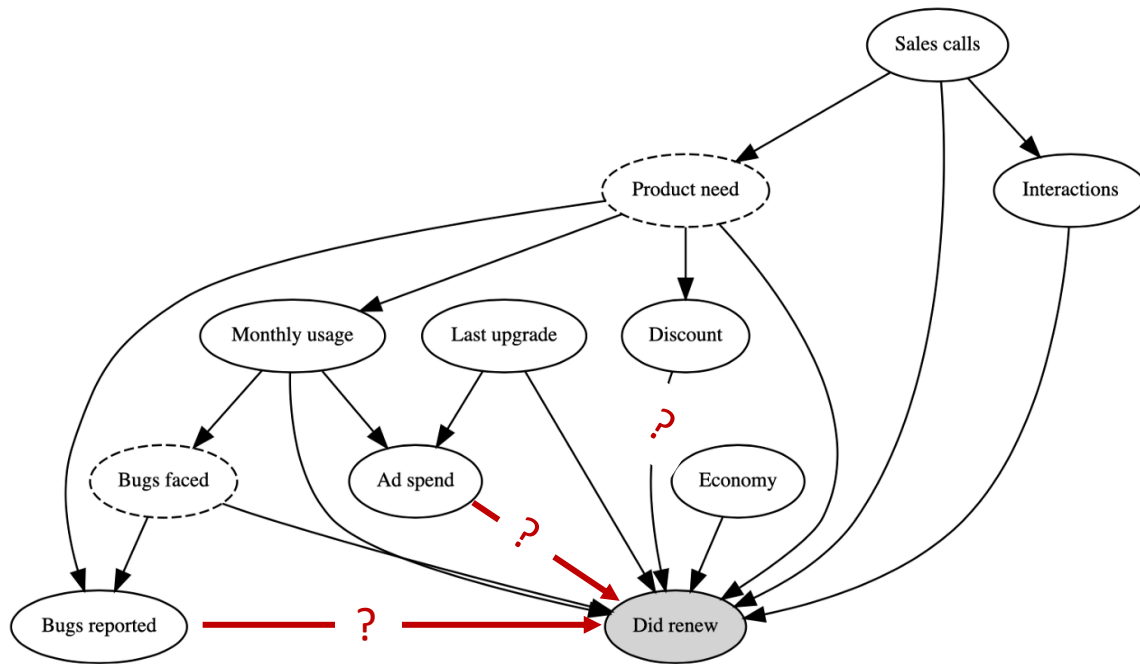
ECONOMETRICA  
VOLUME 11 JANUARY, 1943 NUMBER 1

## THE STATISTICAL IMPLICATIONS OF A SYSTEM OF SIMULTANEOUS EQUATIONS

By TRYGVE HAAVELMO

Require domain knowledge

of the high-level causal mechanisms that underlie the data collection process



# The hierarchy of evidence

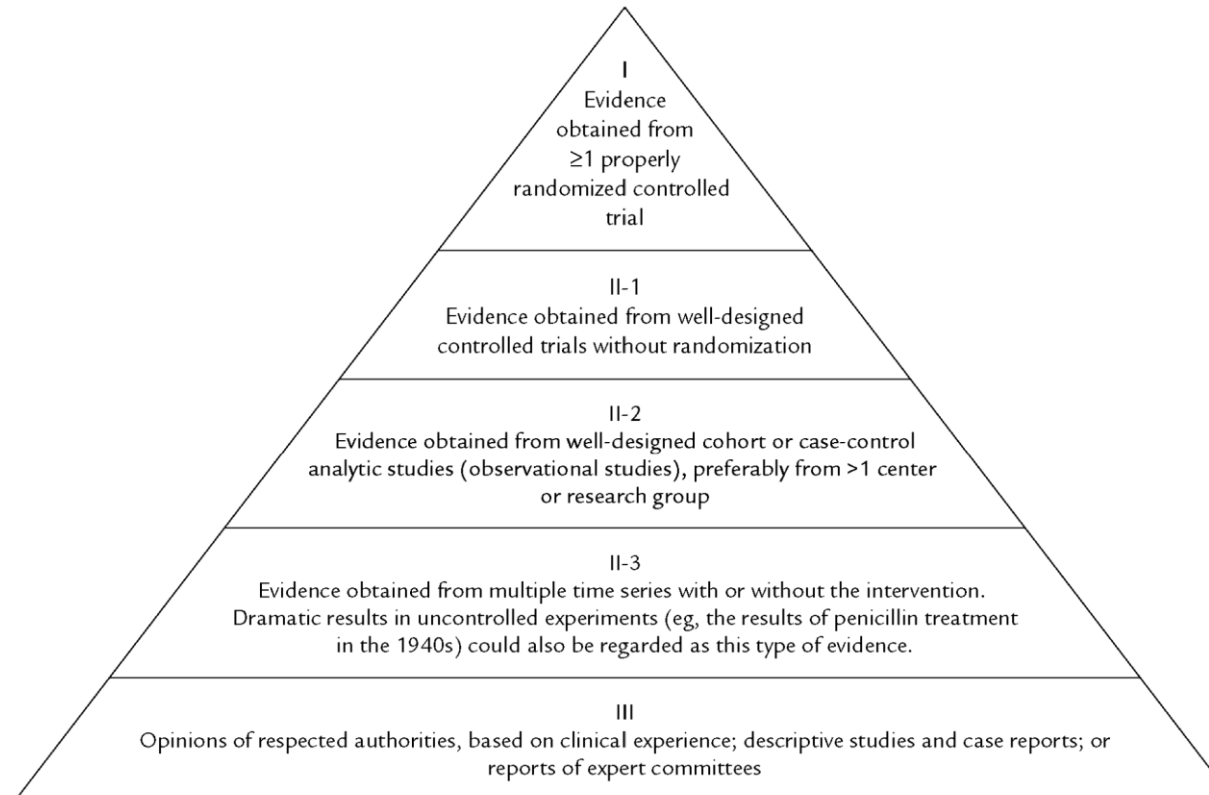


Figure 1. The evidence-grade hierarchy as set out by the US Preventive Services Task Force.<sup>5</sup> Observational studies are ranked as II-2 on the evidence scale.

## Importance of Observational Studies in Clinical Practice

Robert J. Ligthelm, MD<sup>1</sup>; Vito Borzi, PhD<sup>2</sup>; Janusz Gumprecht, MD, PhD<sup>3</sup>; Ryuzo Kawamori, MD<sup>4</sup>; Yang Wenying, MD<sup>5</sup>; and Paul Valensi, MD<sup>6</sup>

# The immense improvement in observational techniques

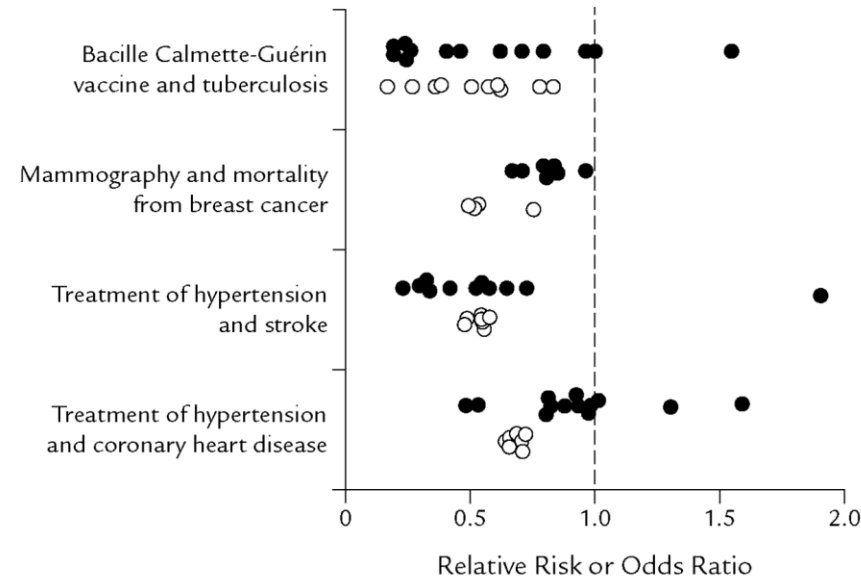


Figure 2. Comparison of odds ratio ranges for observational studies (○) and randomized controlled trials (●). Adapted with permission.<sup>13</sup> Copyright © 2000 Massachusetts Medical Society. All rights reserved.

## Importance of Observational Studies in Clinical Practice

Robert J. Ligthelm, MD<sup>1</sup>; Vito Borzi, PhD<sup>2</sup>; Janusz Gumprecht, MD, PhD<sup>3</sup>; Ryuzo Kawamori, MD<sup>4</sup>; Yang Wenying, MD<sup>5</sup>; and Paul Valensi, MD<sup>6</sup>

# The pitfalls

- Conclusions hinge on validity of assumptions
- Data quality is much more questionable in observational studies
- Definitions of “treatment” and analysis should emulate a “target ideal trial”



## Avoidable flaws in observational analyses: an application to statins and cancer

Barbra A. Dickerman<sup>1\*</sup>, Xabier García-Albéniz<sup>1,2</sup>, Roger W. Logan<sup>1</sup>, Spiros Denaxas<sup>3,4,5</sup> and Miguel A. Hernán<sup>1,6,7</sup>

The increasing availability of large healthcare databases is fueling an intense debate on whether real-world data should play a role in the assessment of the benefit-risk of medical treatments. In many observational studies, for example, statin users were found to have a substantially lower risk of cancer than in meta-analyses of randomized trials. Although such discrepancies are often attributed to a lack of randomization in the observational studies, they might be explained by flaws that can be avoided by explicitly emulating a target trial (the randomized trial that would answer the question of interest). Using the electronic health records of 733,804 UK adults, we emulated a target trial of statins and cancer and compared our estimates with those obtained using previously applied analytic approaches. Over the 10-yr follow-up, 28,408 individuals developed cancer. Under the target trial approach, estimated observational analogs of intention-to-treat and per-protocol 10-yr cancer-free survival differences were  $-0.5\%$  (95% confidence interval (CI)  $-1.0\%$ ,  $0.0\%$ ) and  $-0.3\%$  (95% CI  $-1.5\%$ ,  $0.5\%$ ), respectively. By contrast, previous analytic approaches yielded estimates that appeared to be strongly protective. Our findings highlight the importance of explicitly emulating a target trial to reduce bias in the effect estimates derived from observational analyses.

What is new

- Richer datasets (small data) and high-dimensionality
  - Larger datasets (big data) and the desire for personalization
- 
- Advancement of modern predictive machine learning and large-scale computation
  - Desire for more robust and flexible analysis even in classical datasets



# Causal Machine Learning

Re-directing the ability of machine learning estimators to bypass the curse of dimensionality, from the current focus of solving prediction problems to solving statistical problems that arise in causal inference.

# Many industrial and scientific use cases

- Return-on-investment, pricing, customer segmentation and personalization
- Digital experimentation, online ad targeting
- Personalized medicine
- Heterogeneity of effect in social science studies

Modern case studies



# Example 1: Digital Recommendation A/B Tests

Through the lens of a Case-Study at TripAdvisor

# TripAdvisor Membership Problem

- What is the causal effect of becoming a member on TripAdvisor on downstream activity on the webpage?
- How does that effect vary with observable characteristics of the user?
- Useful for understanding the quality of membership offering/improvements/targeting

# TripAdvisor Membership Problem

- What is the causal effect of becoming a member on TripAdvisor on downstream activity on the webpage?
- How does that effect vary with observable characteristics of the user?
- Useful for understanding the quality of membership offering/improvements/targeting

**Standard approach:** Let's run an A/B test!

**Not applicable:** We cannot enforce the treatment!

- We cannot take a random half of the users and make them members
- Membership is an action that requires user engagement!

# Recommendation A/B Tests

- In optimizing a service we want to understand the causal effects of actions that involve user engagement (e.g. becoming a member)

# Recommendation A/B Tests

- In optimizing a service we want to understand the causal effects of actions that involve user engagement (e.g. becoming a member)
- We can run a **recommendation A/B test**:
  - “recommend/create extra incentives” to half the users to take the action/treatment
- *Example at TripAdvisor*: enable an easier sign-up flow process for a random half of users



# Recommendation A/B Tests

- In optimizing a service we want to understand the causal effects of actions that involve user engagement (e.g. becoming a member)
- We can run a **recommendation A/B test**:
  - “recommend/create extra incentives” to half the users to take the action/treatment
- *Example at TripAdvisor*: enable an easier sign-up flow process for a random half of users
- **Non-Compliance**: ``user’s choice to comply or not`` can lead to biased estimates

# Instrumental Variables (IV)

- **Instrumental Variable:** any random variable  $\mathbf{Z}$  that affects the treatment assignment  $\mathbf{T}$  but does not affect the outcome  $\mathbf{Y}$  other than through the treatment [Wright'28, Bowden-Turkington'90, Angrist-Krueger'91, Imbens-Angrist'94]
- Cohort assignment in recommendation A/B test is an instrument
- We can apply IV methods to estimate average treatment effect  $\theta$

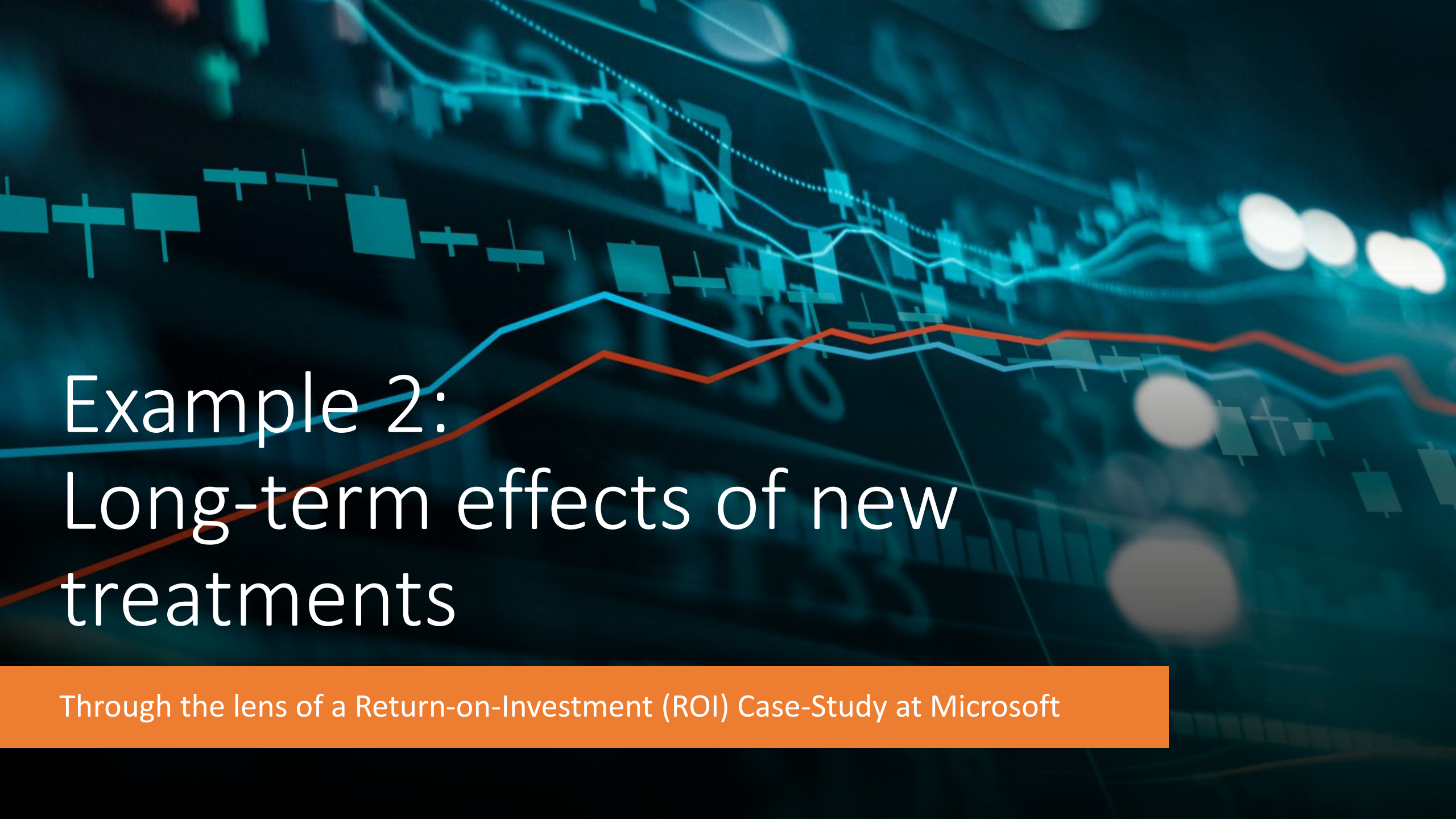
# TripAdvisor Experiment

**For random half of 4 million users, easier sign-up flow was enabled**

- Easier sign-up incentivizes membership

**For each user we observe**

- Instrument  $Z$ : whether the easier sign-up flow was enabled
- Variables  $X$ : observed characteristics of each user: e.g. prior history on platform, location
- Treatment  $T$ : whether the user became a member
- Outcome  $Y$ : number of visits in the next 14 days

The background of the slide is a dark teal color with a complex financial chart overlay. The chart includes a candlestick pattern in the upper left, several overlapping line graphs in light blue and white, and a prominent red line that trends upwards from the bottom left towards the center. The overall aesthetic is high-tech and data-driven.

# Example 2: Long-term effects of new treatments

Through the lens of a Return-on-Investment (ROI) Case-Study at Microsoft

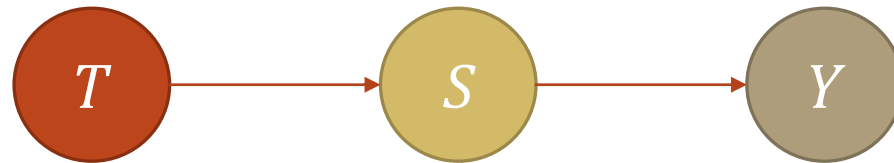
# Estimating Long-Term Returns on Investment

- Companies frequently deploy new discount or customer support programs
- Which of these programs (“investments”) are more successful than others?
- Success is a **long-term** objective: what is the effect of the program on the two-year customer journey (e.g., effect on two-year revenue)
- We cannot wait two years to evaluate a program
- **Main Question.** Can we construct estimates of the values of these programs with **short-term** data, e.g. after 6 months?



# Long-Term Effects from Short-Term Surrogates

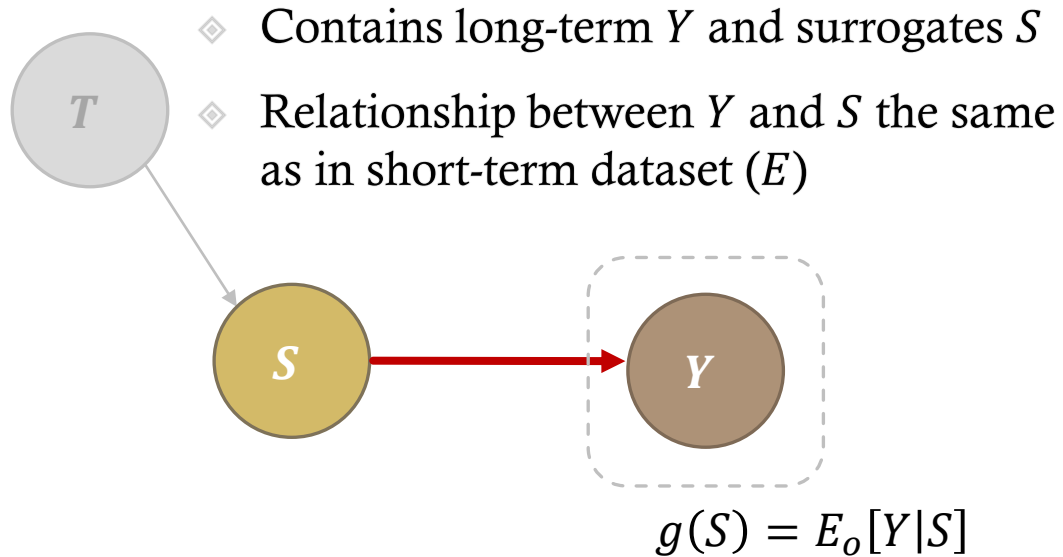
- ◇ Suppose that there are many short-term signals  $S$  that are indicative of a customer's long-term reward  $Y$  (e.g. the next 6-month purchase patterns of a customer could be indicative of their long-term spend)
- ◇ Suppose that investment program  $T$  affects long-term rewards if and only if it affects these short-term signals



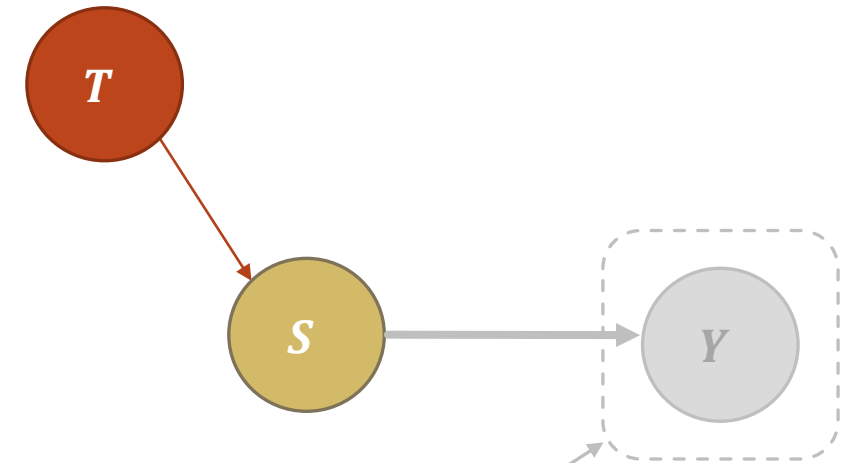
- ◇ We will call these short-term signals  $S$  surrogates

# Causal Inference with Surrogates 101

historical/long-term (O)



recent/short-term (E)



1. Estimate  $g(S) := E[Y|S]$  (surrogate index) from (O) by regressing  $Y \sim S$

2. Impute expected long-term outcomes in (E)
3. Regress  $g(S) \sim T$  to estimate effect of  $T$  on  $Y$  from (E)



# Key Assumptions

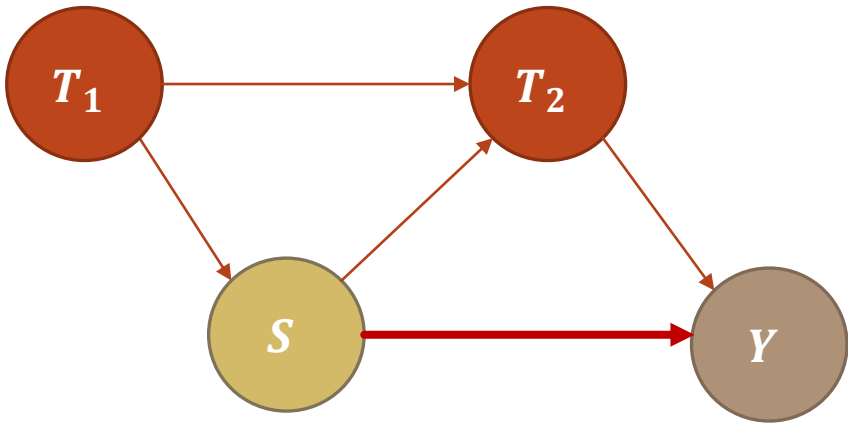
- ◆ Long-term effect only goes through surrogates
- ◆ Expected relationship between surrogates and long-term reward is the same long-term setting ( $O$ ) and in short-term setting ( $E$ )



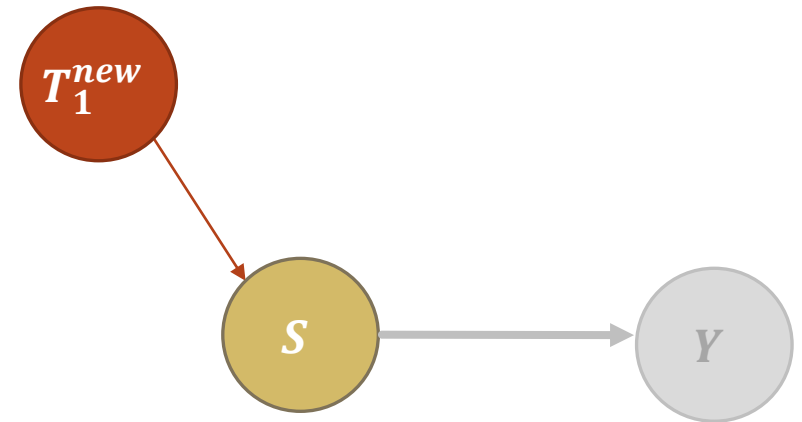
# Key Assumptions can be Easily Violated

## Investment policies are dynamic and change

historical/long-term (O)



recent/short-term (E)



- ◇ We deployed older/deprecated investments
- ◇ In a potentially long-term highly auto-correlated manner
- ◇ Investments are potentially adaptive
- ◇ Investment policies change

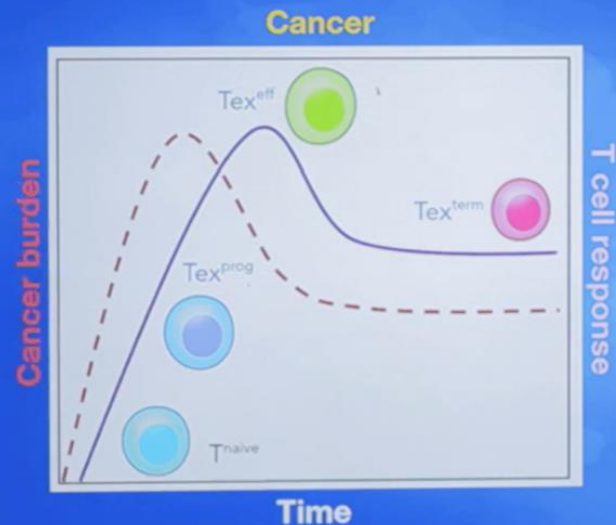


# Example 3: Genomic data

Effect of genetic interventions on effectiveness of cancer therapies

# Cancer Immunotherapy Data Science Grand Challenge

**Cancer evades T cell killing  
by driving T cells to exhaustion.**



$T_{ex}^{prog}$  : progenitor T cell  
 $T_{ex}^{eff}$  : effector T cell  
 $T_{ex}^{term}$  : exhausted T cell

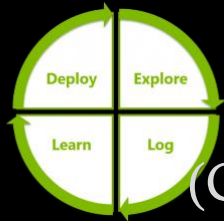
McLane, L. M. et al *Annual Review of Immunology* (2019).

- Centers on the question of how to make T cells better at killing cancer cells
- The Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard and other collaborators
- While scientists have tested some genetic modifications to T cells in the lab, there are too many possible
- Bring ML to this problem to help identify which “perturbations” could make cancer immunotherapy more effective

# Data

- 4,978 unperturbed cells and 26,031 perturbed cells
- Each perturbation is one of 73 different CRISPR single gene knockouts
- For each cell: a 15,077-dimensional gene expression vector and its condition ('unperturbed' or the name of the gene knockout)
- Each cell classified (by experts) into one of 5 cell states ('progenitor', 'effector', 'terminal exhausted', 'cycling', 'other')
- State is related to the effectiveness of a knockout for cancer immunotherapy

# A Growing Software Tool EcoSystem



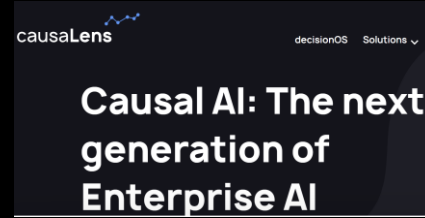
Microsoft

Decision Service  
(Contextual Bandits)



Microsoft

ShowWhy



Microsoft

DoWhy



Microsoft

DoWhy



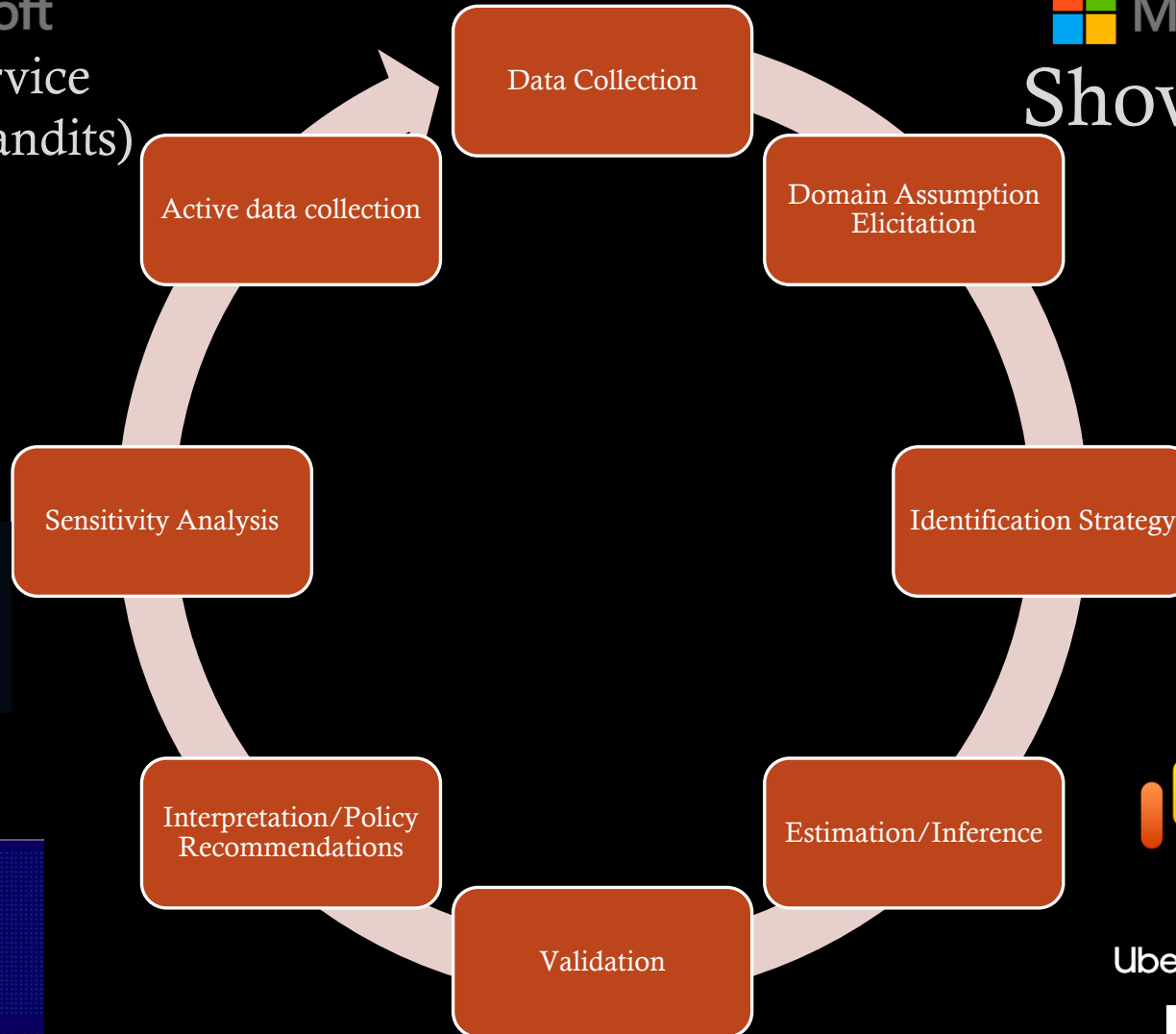
Unlock the Power of Event Sequences  
to Answer the Why



PUBLISHED ON DECEMBER 8, 2021 IN NEWS

**Microsoft Introduces  
New Resources &  
Tools To Help  
Implement AI  
Responsibly**

Microsoft has launched new tools and guidelines to enable product leaders build AI responsibly from research to practice



Auto-Causality



grf-labs

Uber



CausalML

Booking.com



UpliftML



tlverse

What we hope you'll take away

# Goals of the class

- How do you identify if what you want to estimate is feasible from your data and what data you need to bring to make it feasible
- What quantities can you identify from the data you have
- How do you properly use ML based estimation in your observational and experimental causal inference pipeline
- How do you construct confidence intervals for your estimate

Structure and rough outline



# Lecture Outline

- Section 1: Experiments
- Section 2: Estimation and inference with linear models in high-d
- Section 3: Observational data, causality, DAGs, structural equations
- Section 4: Estimation and inference with modern non-linear ML methods
- Section 5: Unobserved confounding and instruments
- Section 6: Heterogeneous Effects and Policy Learning
- Section 7: Further topics

# Logistics

- 7-8 roughly weekly homework assignments (90%)
- Class participation (10%)
- Main text-book: (un-published manuscript shared through Canvas)
- Webpage: <https://stanford-msande228.github.io/winter23/>
- Discussion and homework material: <https://canvas.stanford.edu/>
- Submissions: <https://www.gradescope.com/courses/486969/>

Office Hours: (Starting Week 2)

	Time	Location
Vasilis Syrgkanis	Thursday 3-4pm	Huang 252
Johannes Ferstad	Tuesdays 3-4pm	TBA
Hui Lan	Wednesdays 9:30 - 10:30 am	TBA