# CMPT 733

# Hypothesis Testing

Instructor Zhengjie Miao

Course website https://coursys.sfu.ca/2025sp-cmpt-733-g1/pages/

Based on the slides by Jiannan Wang

# Why Hypothesis Testing?

We want to make a claim from our data

But, data is just a sample

How to prove our claim in this situation?

Using Hypothesis Testing

Example:

- Claim: A data scientist earns more money than a data engineer
- Data: A sample of 50 data scientists and 50 data engineers
  - Result:
    100k vs. 70k

that our claim is correct?

# More examples

Example:

Send out surveys to MPCS students about the program. Are the survey results significant?

Design a new MPCS course on GenAI. 20 students take the course. Is the new course helpful?

# Hypothesis Testing

## Equivalent Terms

- Hypothesis == Claim
- Hypothesis Testing == Claim Proving

## Key Idea

- Prove by contradiction

## Analogy

- <u>How to prove:</u> There exists no smallest positive rational number.
- <u>Hint:</u> a rational number is any number that can be expressed as the fraction a/b of two integers

# Alternative and Null Hypotheses

Alternative Hypothesis ($H_a$)

- The claim that you want to prove it's correct (e.g., the new course is helpful)
- Example $H_a$: $\mu \neq \mu_0$ (two-tailed), $\mu < \mu_0$ (left-tailed), $\mu > \mu_0$ (right-tailed)

Null Hypothesis ($H_0$)

- The opposite side of $H_a$ (e.g., if our results are due to a random chance)
- Example $H_0$: $\mu = \mu_{0,}$ $\mu_1 = \mu_2$

Possible Outcomes

- Reject $H_0$ (a contradiction is found) ➔ Accept $H_a$
- Fail to reject $H_0$ (no contradiction is found)

# Example

Alternative Hypothesis ($H_a$)

- A data scientist earns more money than a data engineer

NULL Hypothesis ($H_0$)

- A data scientist earns less (or equal) money than a data engineer

If $H_0$ is true, what's the probability of seeing:

This is called P-value

- ~~Data Scientist (100 K) vs. Data Engineer (70~~ K)
- Salary(Data Scientist) – Salary(Data Engineer) >= 30 K

# Make a decision based on p-value

We hope that

- p-value is as low as possible so that we can reject $H_0$ (i.e., accept $H_a$)

Level of Significance (e.g., $\alpha = 0.01, 0.05, 0.1$)

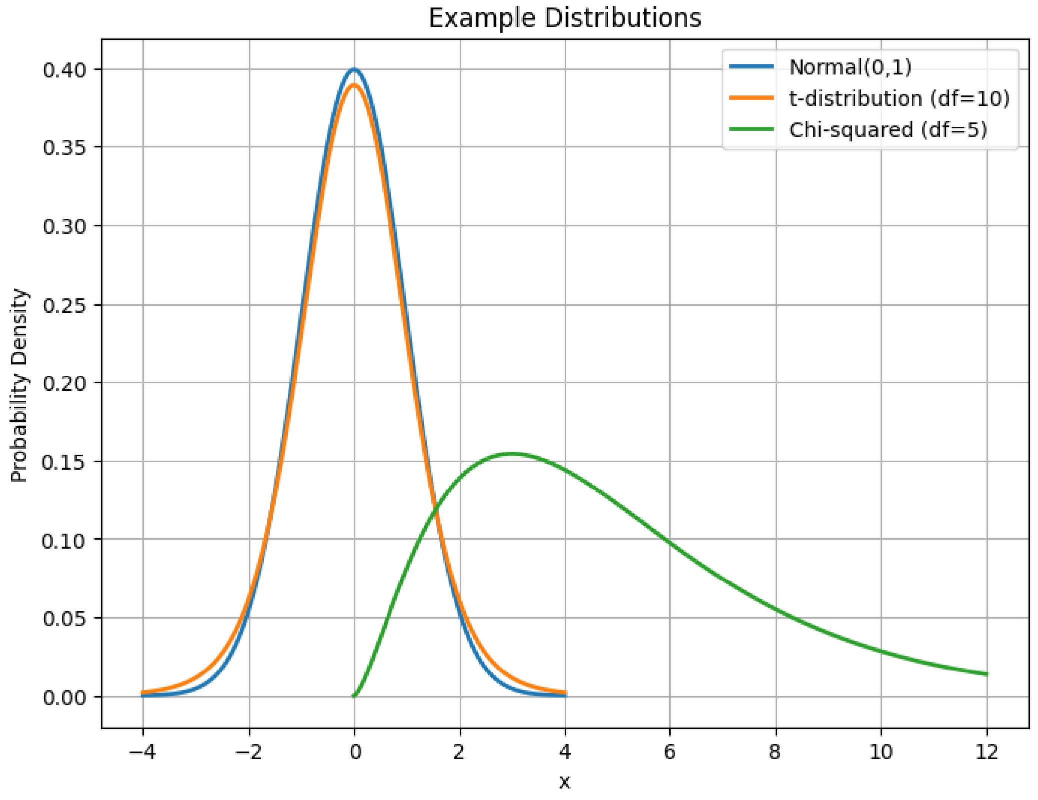
- How low do we want p-value to be?
- P-value < $\alpha$: reject null hypothesis
- P-value > $\alpha$: not enough evidence to reject null hypothesis
  - does not mean we can accept null hypothesis

Level of Confidence (e.g., $c = 1 - \alpha = 99\%$)

- How confident are we in our decision?

# How to calculate p-value

1. Select Alternative Hypothesis
    - Two-tailed, Right-tailed, Left-tailed
2. Determine Test Statistic Distribution
    - e.g., Normal, Student's t, Chi-squared, etc.
3. Specify Degrees of Freedom (if applicable)
    - e.g., t-distribution and Chi-squared tests
4. Calculate the test statistic from sample data
    - Use the CDF to find the prob. of observing an extreme value under the null hypothesis
    - For two-tailed tests, multiply the one-tail probability by 2



# Example

Suppose in the survey, 60% of 100 randomly sampled MPCS alumni said CMPT 733 is helpful for their job

- Null hypothesis $H_0$: at most half of MPCS students find CMPT 733 helpful
- Alternative hypothesis $H_a$: the majority of MPCS students find CMPT 733 helpful
- p-value: the probability that 60% of 100 random students find CMPT 733 helpful ***if*** at most 50% of all students find it helpful

# Example

p = probability that 60% of 100 random students respond CMPT 733 helpful

Population Mean (null hypothesis)

The **z-statistic** formula, using central limit theorem

Sample mean

Population Standard Deviation

$$z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \qquad z \approx \frac{0.6 - 0.5}{0.5/\sqrt{100}} \approx 2.00$$

- What is the probability of drawing > 2.00 from a standard normal distribution?

```
print(1-scipy.stats.norm.cdf(2.00))
✓ 0.0s
0.022750131948179209
```

- p-value is ~ 0.02, we reject the null hypothesis
- Strong support that majority of students find 733 helpful

# Smaller Example

p = probability that 60% of 30 random students respond CMPT 733 helpful

The **z-statistic** formula, using central limit theorem

$$t = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \qquad t \approx \frac{0.6 - 0.5}{0.5/\sqrt{30}} \approx 1.10$$

- What is the probability of drawing > 1.10 from a standard normal distribution?

```
print(1-scipy.stats.norm.cdf(1.10))
✓ 0.0s

0.13566606094638267
```

- p-value is ~ 0.1357, prob under null hypothesis is ~13.57%
- We CANNOT reject null

# Example with unknown standard deviation

p = probability that 60% of 100 random students respond CMPT 733 helpful

The **t-statistic** formula

Population Mean (null hypothesis)

Sample mean

Sample Standard Deviation

$$t = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \qquad t \approx \frac{0.6 - 0.5}{0.49/\sqrt{100}} \approx 2.04$$

- What is the probability of drawing > 2.04 from a Student's t Distribution with n-1 degrees of freedom?

```
print(1-scipy.stats.t.cdf(2.04, 99))
✓ 0.0s
0.022006777804037148
```

Degree of freedom = n-1

- p-value is ~ 0.02, we reject the null hypothesis
- Strong support that majority of students find 733 helpful

# p-Hacking (Cheating on a p-Value)

## Common Mistakes

1. Collect data until the hypothesis testing is passed
2. Keep doing analysis on the same data until you find something significant

## Solution

- You should know what you're looking for ($H_0$ and $H_a$) before you start
- Decrease the level of significance (e.g.,$\alpha/2$ for two hypothesis tests on the same data)

# Pairwise t- tests

Example: You have three treatment groups (A, B, C) and want to determine if the differences in their means are statistically significant

Pairwise Comparisons:

- Compare A vs. B
- Compare A vs. C
- Compare B vs. C

# Multiple Hypothesis Testing

- Multiple simultaneous statistical tests, each of which has a potential to produce a "discovery"
- A stated confidence level (e.g., 1-0.05) generally applies only to each test considered individually
    - When running multiple tests simultaneously, the chance of at least one false positive increases, meaning the overall error rate is higher than 5%.
- Without adjustment, multiple tests can lead to false discoveries
    - Incorrectly identifying a statistically significant effect when there isn't one.

# Multiple Hypothesis Testing

- Example: Imagine a clinical trial where researchers evaluate the effectiveness of a new drug. Instead of testing just one outcome, they examine several outcomes simultaneously
  - e.g., Blood Pressure Cholesterol Levels, Heart Rate, Body Weight
- Each separate hypothesis test is performed to determine if the drug has a statistically significant effect at a 95% confidence level ($\alpha = 0.05$)
- However, because multiple tests are performed, the chance of making at least one false positive (a type I error) across all tests increases.

# Multiple Hypothesis Testing

Example: There are 20 options we are interested in as independent (predictor) features for your model. For each feature, we use a hypothesis test with level of significance 0.05

What's the probability of having one significant result just due to chance? (Recall that probability of making a type I error is $\alpha$)

$$1 - (1-0.05)^{20}=0.64$$

# P-hacking: Pairwise t- tests

```
num_datasets, sample_size = 100, 100
samples = np.random.normal(loc=0, size=(num_datasets, sample_size))
```

```
total_pairs, total_siginificant = 0, 0
for i in range(num_datasets):
    for j in range(i+1, num_datasets):
        p_value = scipy.stats.ttest_ind(samples[i,:], samples[j,:])[1]
        total_pairs += 1
        if p_value < 0.05:
            total_siginificant += 1
print(total_siginificant, total_pairs)
```

✓ 3.6s

```
286 4950
```

# Useful scipy.stats functions

## Independent Sample Tests

Independent sample tests are typically used to assess whether multiple samples were independently drawn from the same distribution or different distributions with a shared property (e.g. equal means).

Some tests are specifically for comparing two samples.

| | |
|---|---|
| `ttest_ind_from_stats`(mean1, std1, nobs1, ...) | T-test for means of two independent samples from descriptive statistics. |
| `poisson_means_test`(k1, n1, k2, n2, *[, ...]) | Performs the Poisson means test, AKA the "E-test". |
| `ttest_ind`(a, b, *[, axis, equal_var, ...]) | Calculate the T-test for the means of *two independent* samples of scores. |

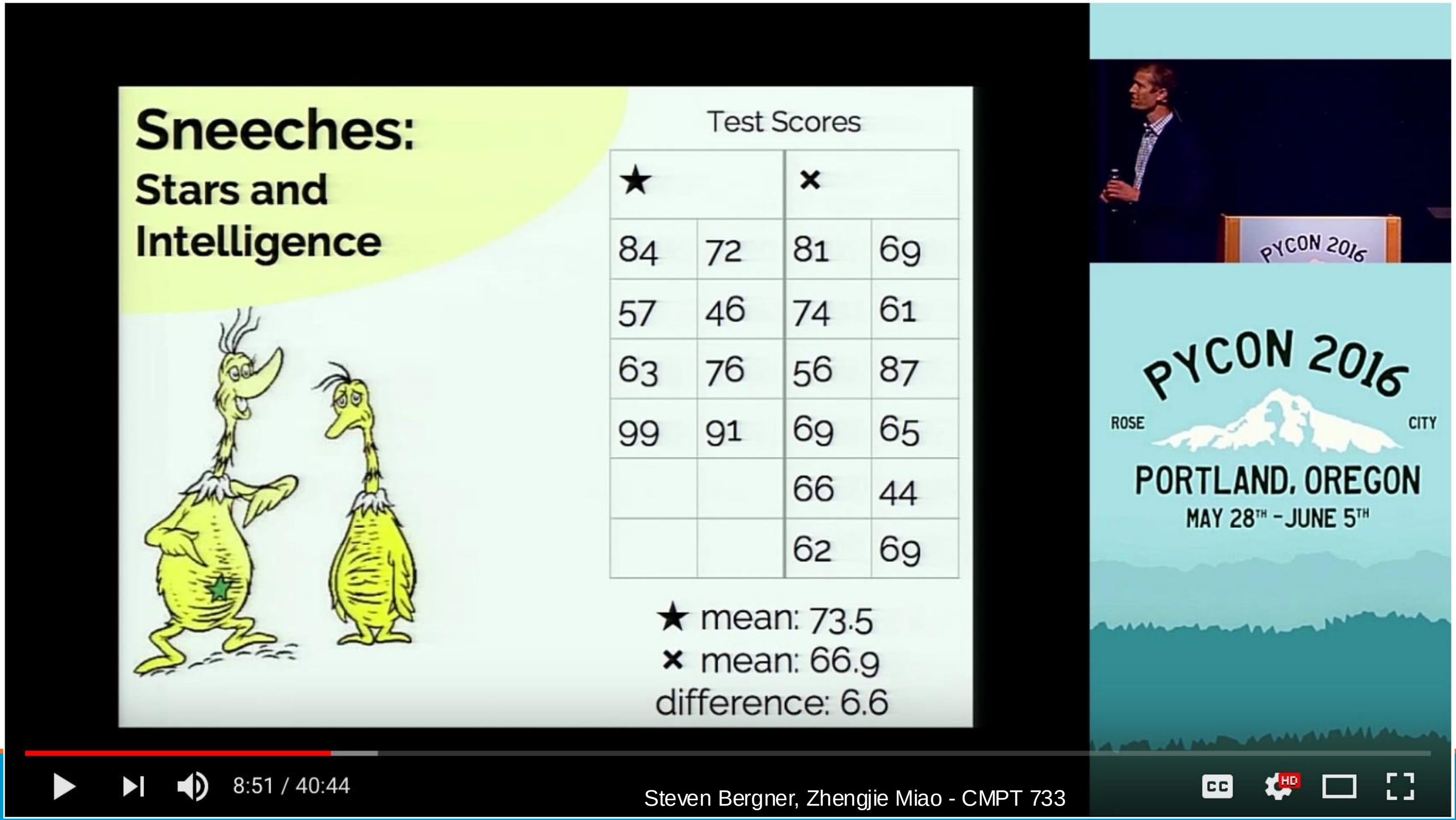
https://docs.scipy.org/doc/scipy/reference/stats.html

# Permutation Test

- Resampling Data: Repeatedly "permute" the labels or values in your dataset
  - E.g., if comparing two groups, shuffle the group labels
- Computing a Test Statistic: For each permutation, calculate a test statistic (e.g., the difference in means between groups)
- Estimating the p-value: compare the observed test statistic from the actual data to this permutation distribution
- When to use: does not require the assumptions of normality or equal variances

# Permutation Test

https://youtu.be/Iq9DzN6mvYA?t=8m9s



# Hypothesis Test Using Simulation

- Generate a large number of datasets under the null hypothesis
  - E.g., drawing samples from a known distribution or resampling (as in permutation testing)
- Calculate the test statistic for observed data
- P-value is the empirical probability of seeing a test statistic as extreme as the observed statistic in the simulated datasets
- Recall Monte Carlo methods / Bootstrap
- **When to use:** no assumptions of normality or equal variances; small sample sizes

# Hypothesis Test Using Simulation

- Generate a large number of datasets under the null hypothesis
  - E.g., drawing samples from a known distribution or resampling (as in permutation testing)
- Calculate the test statistic for observed data
- P-value is the empirical probability of seeing a test statistic as extreme as the observed statistic in the simulated datasets
- Recall Monte Carlo methods / Bootstrap

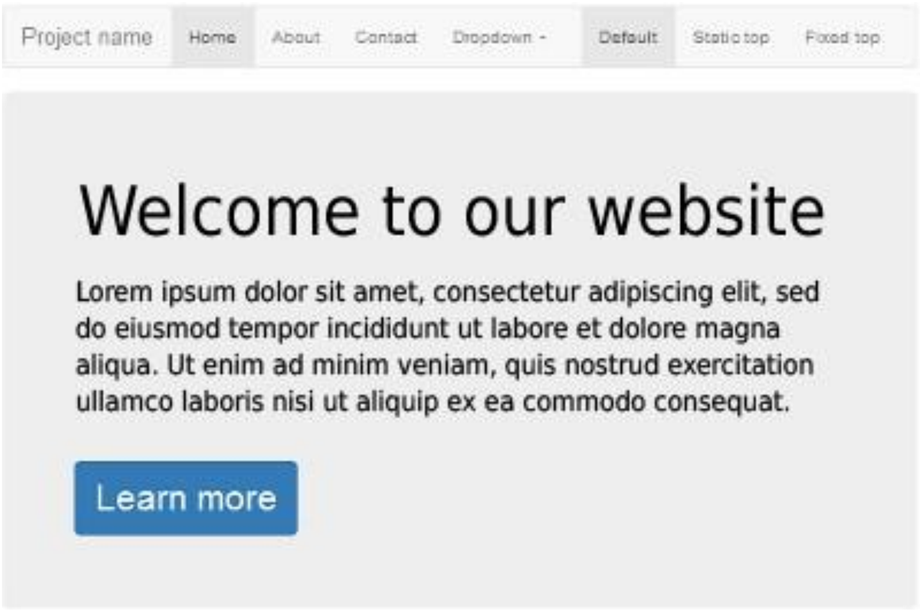
```
num_simulation, sample_size = 1000000, 100
samples = np.random.choice([0, 1], size=(num_simulation, sample_size), replace=True)
sample_means = np.mean(samples, axis=1)
p_value_estimate = np.sum(sample_means >= 0.6) / num_simulation
print(p_value_estimate)
```

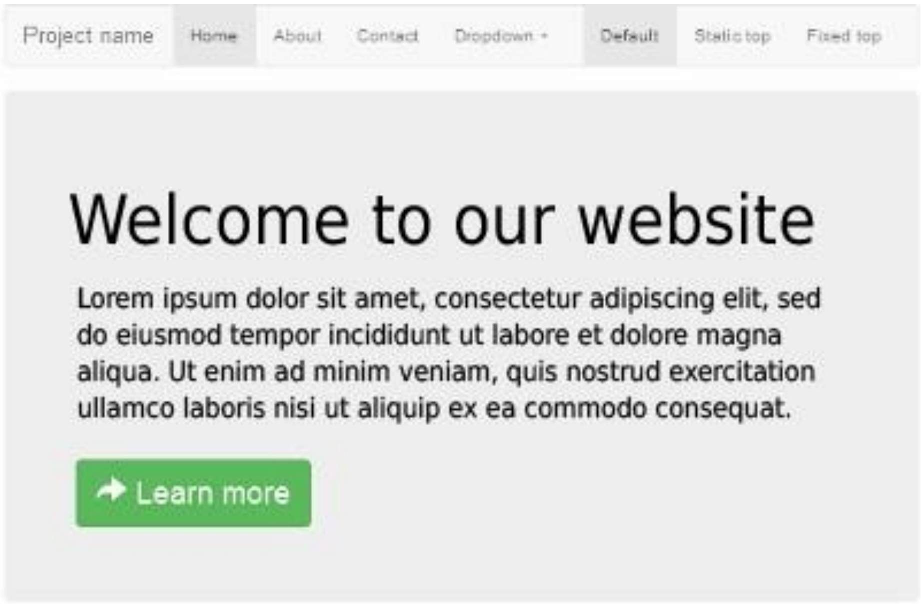
✓ 0.9s

```
0.028133
```

# A/B Testing

What UI is better?



# Surprising A/B Tests

A. Get $10 off the first purchase. Book online now!

B. Get an additional $10 off. Book online now.


https://www.wordstream.com/blog/ws/2012/09/25/a-b-testing

# A/B Testing

- A method to compare two versions (A and B) of a variable to determine which performs better
  - Optimize website design, marketing campaigns, user experience, etc.
- Split users randomly between the two versions and measure the differences in performance

# Why A/B Testing?

- **Data-Driven Decisions:** Replace guesswork with statistical evidence
- **Improved Conversion Rates:** Identify changes that boost user engagement
- **Reduced Risk:** Test ideas on a small scale before full implementation
- **Continuous Improvement:** Iteratively refine products and strategies

# A/B Testing Steps

1. Metrics identification
   1. Invariant metrics (e.g., # users, user demographics, etc.)
   2. Evaluation metrics (what you want to optimize, e.g., CTR, # Orders)
2. Hypothesis definition

| | H0 is True | H0 is False |
|---|---|---|
| Reject H0 | Type I error (False Positive)<br>Prob = $\alpha$ | True Positive<br>Prob = 1-**β** |
| Fail to reject H0 | True Negative<br>Prob = 1-$\alpha$ | Type II error (False Negative)<br>Prob = **β** |

- Usually target $\alpha = 0.05$, **β = 0.8**

# A/B Testing Steps

3. Determine sample size & duration

$$n = \frac{16\sigma^2}{\delta^2}$$

Duration: If the daily active users is greater than 2n, can we finish A/B testing in one day?

- Usually not, because:
  - Bug in your new version
  - Other variants like weekends/holidays/discounts

# A/B Testing Steps

5. Distribute users
6. Invariance test / sanity check
    - Metrics that should not change between variant A & variant B
    - User demographics, traffic sources, device types, …
    - When it fails:
        - Review randomization
        - External factors
        - Drill down to segments

# What can/cannot be A/B Tested?

- New feature of a product
- UI changes
- Backend changes

Question: can we A/B test to decide if ChatGPT should be released?

Question: can we A/B test the new UI of a house selling website

- when the evaluation metric is traffic volume?
- when the evaluation metric is # of transactions?

# Conclusion

- Hypothesis Testing
  - Null Hypothesis ($H_0$) and Alternative Hypothesis ($H_a$)
  - P-value and P-hacking
  - A/B Testing

# CMPT 733
# Causal Inference

Instructor Zhengjie Miao

Course website https://coursys.sfu.ca/2025sp-cmpt-733-g1/pages/

Based on the slides by Yi Xie, Nathan Yan, and Jiannan Wang

# Outline

Why Should Data Scientists Care?

Basic Concepts

Causal Inference

# Questions Data Scientists Can Answer

| | |
|---|---|
| Is This A or B? | Classification |
| How much or How Many? | Regression |
| Is This Weird? | Anomaly Detection |
| How Is This Organized? | Clustering |
| What if? | Causal Inference |

# From Prediction to Causation

Predicting user activity for Xbox

- Y = logins in next month
- X = logins in past month, number of friends,...

What if we increase the number of friends?

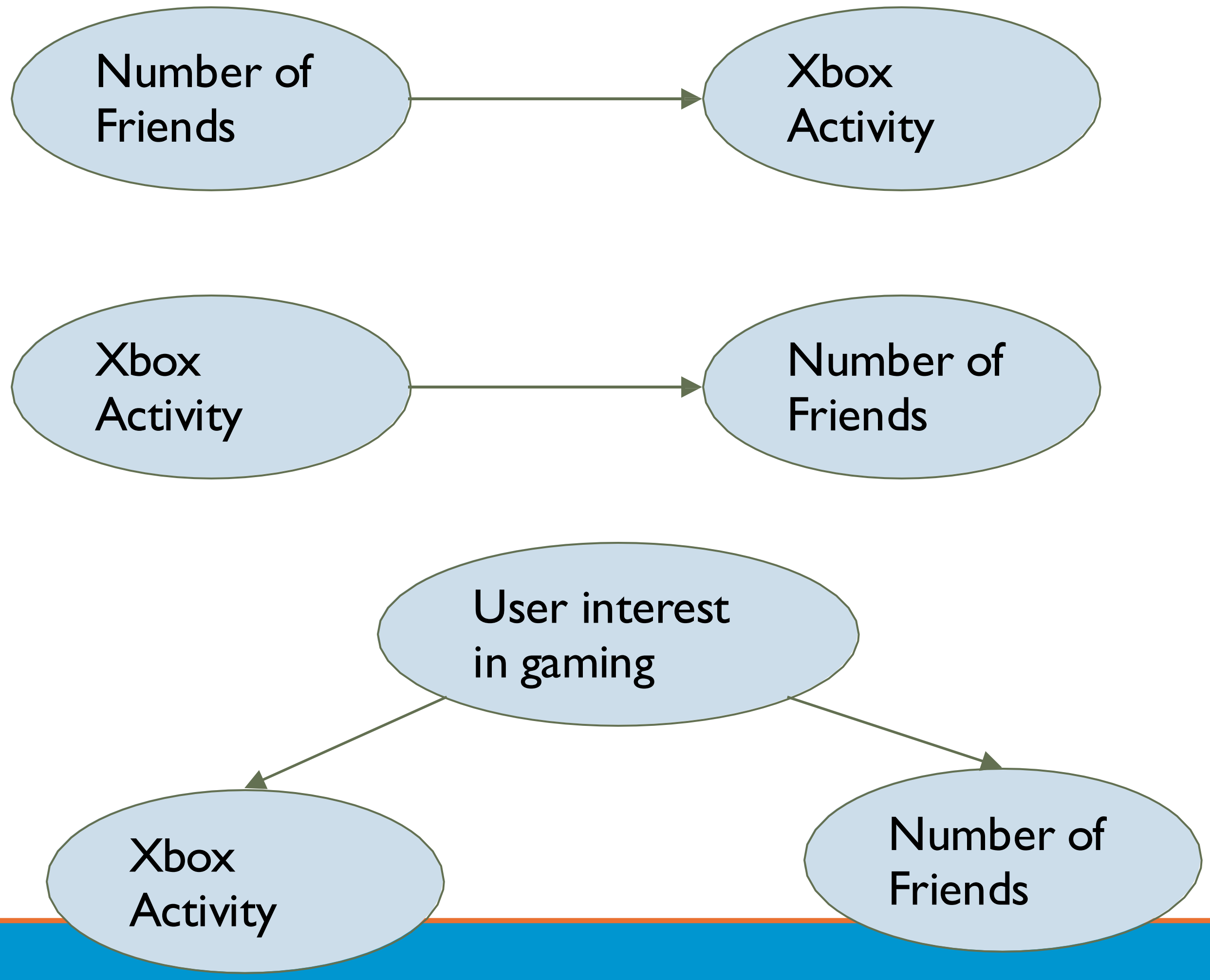
- Would it increase user activity?

# Maybe or May be not (!)

A causes B

B causes A

C causes A and B



# A/B Testing Helps!

## Treatment Group

- A random sample of users
- Launch a campaign to increase friends
- Average activity next month

## Control Group

- A random sample of users
- Not Launch a campaign to increase friends
- Average activity next month

Hypothesis Testing

# A/B Testing Does Not Work

- It is infeasible to do A/B testing
  - What if you went to UBC rather than SFU, would it help you land a better job?

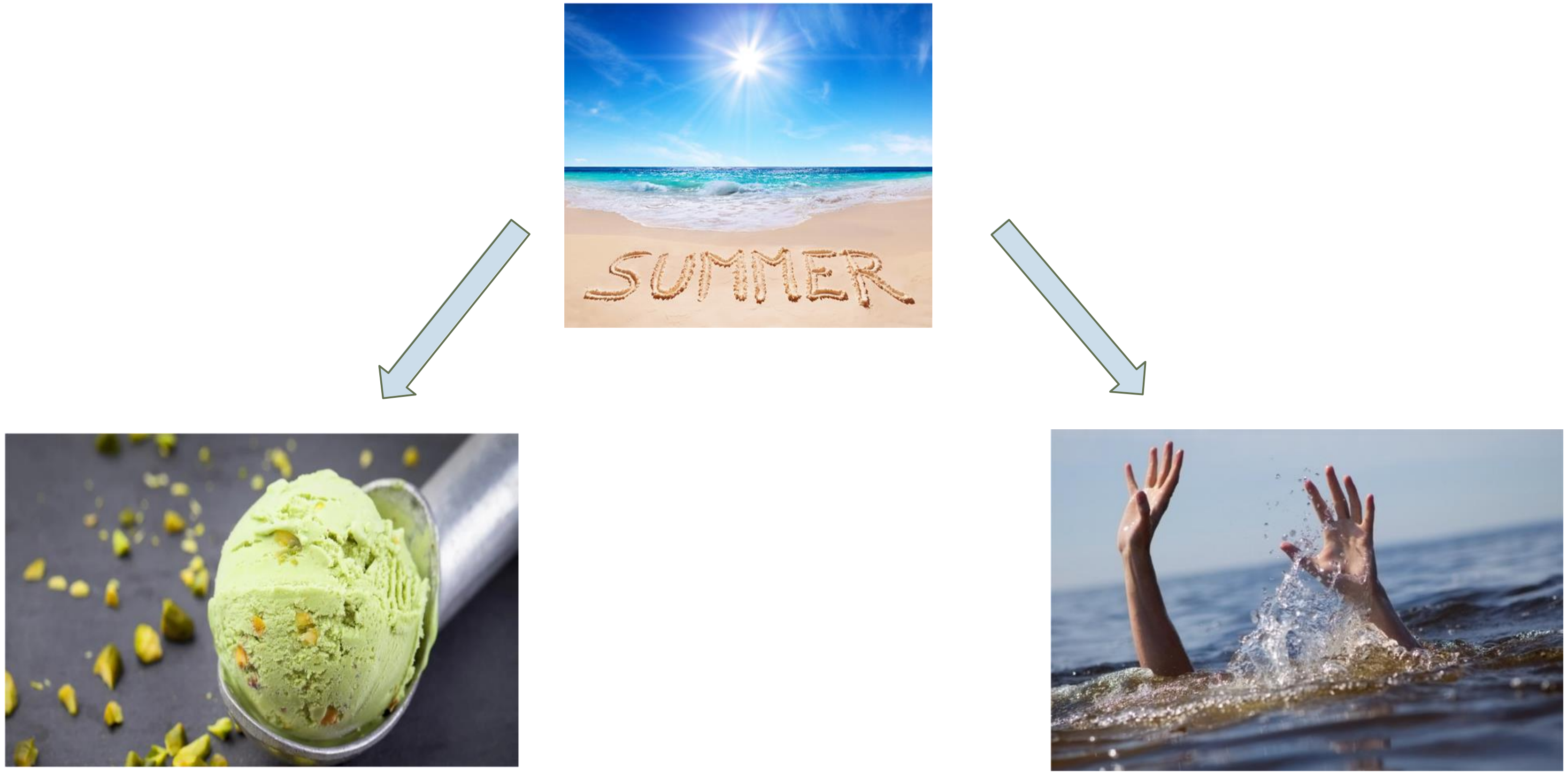
- It is unethical to do A/B testing
  - What if subscription price is set to $69 rather than $99, would it increase revenue?

# Example 1: Causality ≠ Correlation

When a team was investigating the city death rate data, they found that reported cases of teenage drowning death increase while the sales of ice cream also increase.

??

# Example 1: Causality ≠ Correlation



# Example 2: Causality ≠ Correlation

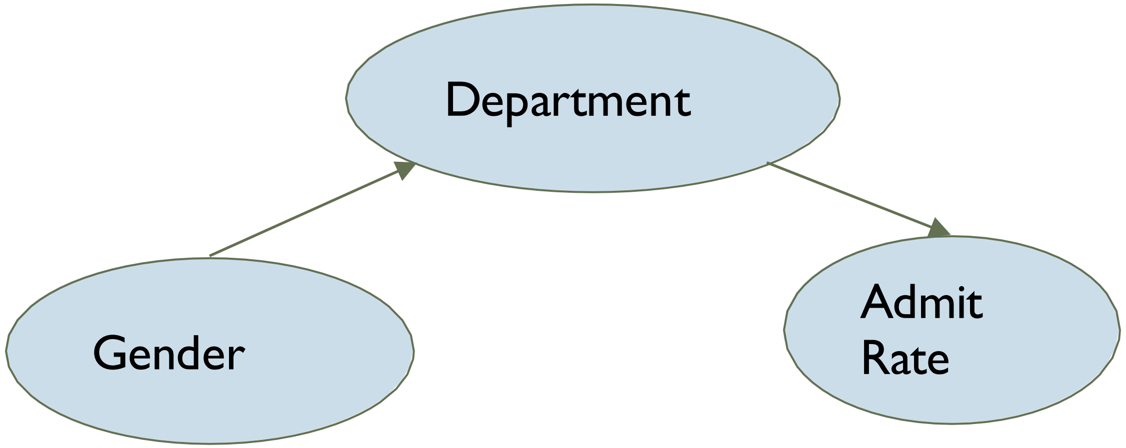
Is Berkeley gender biased?

| | Applicants | Admitted |
|---|---|---|
| **All** | 12,763 | 41% |
| **Men** | 8,442 | **44%** |
| **Women** | 4,321 | 35% |

Simpson's paradox

# Example 2: Causality ≠ Correlation

| Department | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| A | **825** | 62% | 108 | 82% |
| B | **560** | 63% | 25 | 68% |
| C | 325 | 37% | **593** | 34% |
| D | 417 | 33% | 375 | 35% |
| E | 191 | 28% | **393** | 24% |
| F | 373 | 6% | 341 | 7% |
| **Total** | 2691 | 45% | 1835 | 30% |



# Example 2: Causality ≠ Correlation


- In city areas with nearby trees and natural landscapes, there is less domestic violence.
- Apartment complexes with many trees had 52% fewer crimes.
- On tree-lined streets, people drive more slowly, reducing accident risk.

# Outline

## Why Should Data Scientists Care?

## Basic Concepts

- Outcome / Treatment Variables
- Intervention and Do Operation
- Counterfactual
- Causal Graph

## Causal Inference

# Outcome / Treatment Variables

What if participating Study Program, would it improve Grade?

Treatment

Outcome

| Student | Gender | Class | Study Program | Grade |
|---|---|---|---|---|
| Jacky | male | 1 | 0 | 78 |
| Terry | male | 1 | 1 | 82 |
| Mary | female | 1 | 0 | 86 |
| Sarah | female | 2 | 1 | 83 |

# Intervention and Do Operation

- Do operator will signal the experimental intervention (invented by Judea Pearl)

  P(grade|do(enrolled in study program))

  represents the distribution of grad if the person is enrolled in study program

# Intervention and Do Operation

- $P(A \mid B = b)$: probability of A being true given that B is observed as $B = b$
- $P(A \mid do(B) = b)$: probability of A being true given an intervention that sets B to b

# Counterfactual

- What would have happened if I had changed "Treatment Variable"

| Student | Gender | Class | Study Program | Actual Grade |
|---|---|---|---|---|
| Jacky | male | 1 | 0 | 78 |
| Terry | male | 1 | 1 | 82 |
| Mary | female | 1 | 0 | 86 |
| Sarah | female | 2 | 1 | 83 |

# Counterfactual

- What would have happened if I had changed "Treatment Variable"

| Student | Gender | Class | Study Program | Actual Grade | Counterfactual Grade (You cannot get it in reality) |
|---|---|---|---|---|---|
| Jacky | male | 1 | 0 | 78 | 82 |
| Terry | male | 1 | 1 | 82 | 82 |
| Mary | female | 1 | 0 | 86 | 90 |
| Sarah | female | 2 | 1 | 83 | 85 |

# Causality

Causality Definition

- The difference between actual outcome and counterfactual outcome

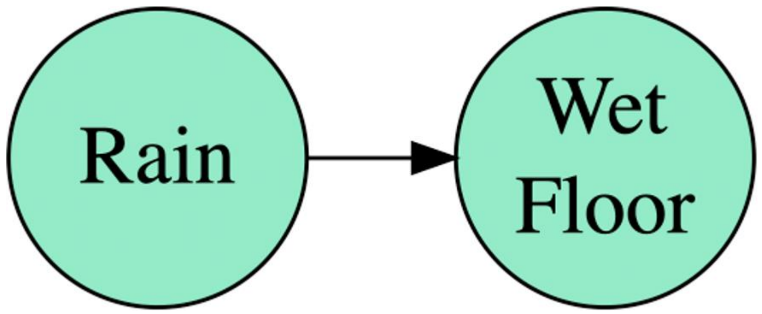
The Fundamental Problem of Causal Inference

- We cannot observe the counterfactual outcome

# Causal Graph

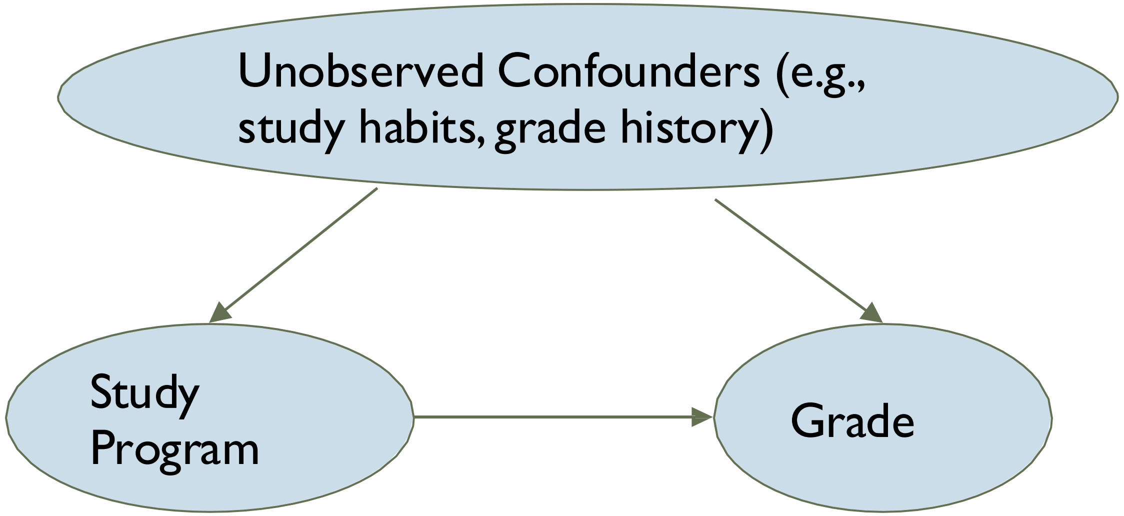
Causal graph is a directed graph

- Nodes: variables
- Directed Edges: X affects Y



# Causal Graph

- Confounding variables: common cause of treatment and outcome




Why Causal Graph?

- Helpful to identify which variables to control for
- Make assumptions explicit

# Outline

Why Should Data Scientists Care?

Basic Concepts

Causal Inference

- Statistical Inference vs. Causal Inference
- Average Treatment Effect (ATE) Estimation
- Causal Inference in Practice

# Statistical vs. Causal Inferences

## Statistical inference

- Data is just a sample
- Your goal is to infer a population
- Think about how to go “backwards” from sample to population

## Causal inference

- Derive a treatment group from Data
- Derive a control group (i.e., without treatment) from Data
- Think about how to infer the actual effect of treatment from the derived treatment and control groups

# Outline

Why Should Data Scientists Care?

Basic Concepts

Causal Inference

- Statistical Inference vs. Causal Inference
- Average Treatment Effect (ATE) Estimation
- Causal Inference in Practice

# Counterfactual

- What would have happened if I had changed "Treatment Variable"

| Student | Gender | Class | Study Program | Actual Grade | Counterfactual Grade (You cannot get it in reality) |
|---|---|---|---|---|---|
| Jacky | male | 1 | 0 | 78 | 82 |
| Terry | male | 1 | 1 | 82 | 82 |
| Mary | female | 1 | 0 | 86 | 90 |
| Sarah | female | 2 | 1 | 83 | 85 |

# Individual Treatment Effect

What is the grade difference between enrolling and not enrolling in the study program?

| Student | Gender | Class | Enroll in Study Program | Not Enroll in Study Program | |
|---|---|---|---|---|---|
| Jacky | male | 1 | 82 | 78 | 4 |
| Terry | male | 1 | 82 | 82 | 0 |
| Mary | female | 1 | 90 | 86 | 4 |
| Sarah | female | 2 | 83 | 85 | -2 |

# Average Treatment Effect (ATE)

The average of all values for individual treatment effects

ATE = (4 + 0 + 4 + -2) / 4 = 1.5

| Student | Gender | Class | Enroll in Study Program | Not Enroll in Study Program | |
|---|---|---|---|---|---|
| Jacky | male | 1 | 82 | 78 | 4 |
| Terry | male | 1 | 82 | 82 | 0 |
| Mary | female | 1 | 90 | 86 | 4 |
| Sarah | female | 2 | 83 | 85 | -2 |

# ATE: Estimation

## ATE estimation methods

- Matching based:
  - Perfect matching
  - Nearest neighbor matching
  - Propensity score matching

- ML based:
  - Regression method
  - Representation learning

# Perfect Matching

| student | gender | class | Study program | grade |
|---|---|---|---|---|
| Terry | male | 1 | 1 | 82 |
| Sarah | female | 2 | 1 | 83 |
| Jacky | male | 1 | 0 | 78 |
| Mary | female | 1 | 0 | 86 |

Find the "perfect matching in counterfactual world"

# Perfect Matching

| student | gender | class | Study program | grade | |
|---|---|---|---|---|---|
| Terry | male | 1 | 1 | 82 | 4 |
| Sarah | female | 2 | 1 | 83 | |
| Jacky | male | 1 | 0 | 78 | |
| Mary | female | 1 | 0 | 86 | |

Find the "perfect matching in counterfactual world"

# Perfect Matching

| Student | Gender | Class | Study Program | Grade |
|---|---|---|---|---|
| Terry | male | 1 | 1 | 82 |
| Sarah | female | 2 | 1 | 83 |
| Jacky | male | 1 | 0 | 78 |
| Mary | female | 1 | 0 | 86 |

By perfect matching, we can’t find ATE over the population, because for tuple Sarah, we can’t find the perfect match in control group

# Nearest Neighbor Matching

| Student | Gender | Class | Study Program | Grade | |
|---|---|---|---|---|---|
| Terry | male | 1 | 1 | 82 | |
| Sarah | female | 2 | 1 | 83 | -3 |
| Jacky | male | 1 | 0 | 78 | |
| Mary | female | 1 | 0 | 86 | |

Find the "nearest matching in counterfactual world"

# Nearest Neighbor Matching

| Student | Gender | Class | Study Program | Grade | |
|---|---|---|---|---|---|
| Terry | male | 1 | 1 | 82 | 4 |
| Sarah | female | 2 | 1 | 83 | -3 |
| Jacky | male | 1 | 0 | 78 | |
| Mary | female | 1 | 0 | 86 | |

ATE = ½ [-3 + 4] = 0.5

# Propensity Score Matching

Steps

- Using logistic regression to infer $e(x)=Pr[T=1|X=x]$

- Match users with treatment 0 with users with treatment 1 based on propensity score, using matching methods

# Propensity Score Matching

| Student | Gender | Class | Study Program | Grade | PSE |
|---|---|---|---|---|---|
| Terry | male | 1 | 1 | 82 | 0.7 |
| Sarah | female | 2 | 1 | 83 | 0.6 |
| Jacky | male | 1 | 0 | 78 | 0.7 |
| Mary | female | 1 | 0 | 86 | 0.55 |

Based on PSM derived by logistic regression, the match we'll find for Terry is Jacky, and the match we'll find for Sarah is Mary

ATE = 0.5, computation is similar with nearest neighbor matching

# Regression Method

Intuition: the distribution of Y given X is different when treatment is different

Train two separate regression models under Treatment = 0 or Treatment = 1, infer $p(Y \mid T = 0, X)$ and $p(Y \mid T = 1, X)$

# Regression Method

| Student | Gender | Class | Study Program | Grade |
|---|---|---|---|---|
| Terry | male | 1 | 1 | 82 |
| Sarah | female | 2 | 1 | 83 |
| Jacky | male | 1 | 0 | 78 |
| Mary | female | 1 | 0 | 86 |

Treat one regression model 1 on [Terry, Sarah] where Study Program = 1

Treat another regression model 2 on [Jacky, Mary] where Study Program = 0

Using model 1 to compute counterfactual outcome of [Jacky, Mary], same for model 2

# Other ML-based Methods

- Representation learning
- Intuition: transform dataset into a space where treatment assignment is more evenly distributed
- For other more techniques, please check latest publications

# Outline

Why Should Data Scientists Care?

Basic Concepts

Causal Inference

- Statistical Inference vs. Causal Inference
- Average Treatment Effect (ATE) Estimation
- Causal Inference in Practice

# Causal Inference Case Study

What if you went to UBC rather than SFU, would it help you land a better job?

1. What data to collect?
2. How to use data to answer this question?

# Allocating Policy for Homelessness

- Researchers allocate different interventions (like emergency shelter, rapid rehousing) for homelessness based on causal inference

- Report published in AAAI 2019

Amanda Kube, Sanmay Das, Patrick J. Fowler: Allocating Interventions Based on Predicted Outcomes: A Case Study on Homelessness Services. AAAI 2019: 622-629

# Social Media

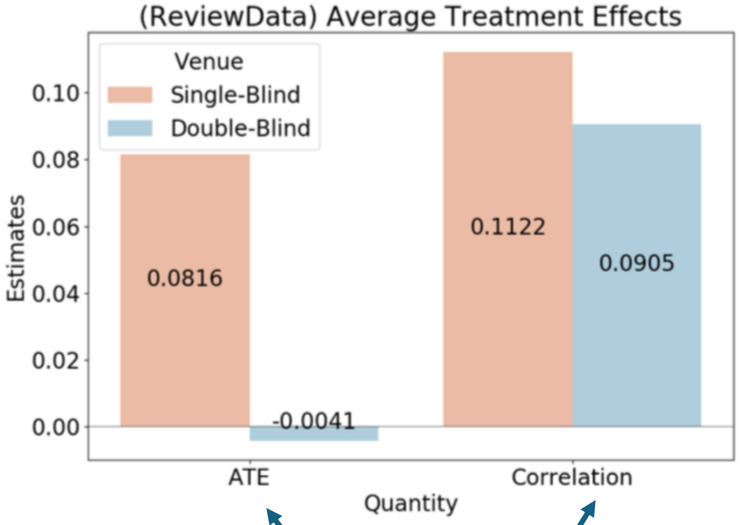
- For Twitter, the impact of race, gender and closeness on persuasion is studied using causal inference

- Result published in NeurIPS 2019

https://cpb-us-w2.wpmucdn.com/sites.coecis.cornell.edu/dist/a/238/files/2019/12/Id_104_final.pdf

# Academic Paper Review

Are reviewers influenced by authors' prestige?




(a)

Causation vs. Correlation

OpenReview.net

High correlation in both single and double blind
High causation only in single blind

[Salimi-Parikh-Kayali-Getoor-Roy-Suciu, SIGMOD'20]

# DOWhy Python Library

- DoWhy is a Python library for causal inference that supports explicit modeling and testing of causal assumptions, developed by Microsoft.

- DoWhy is based on a unified language for causal inference, combining causal graphical models and potential outcomes frameworks.

https://github.com/microsoft/dowhy

# Summary

## Why Should Data Scientists Care?

- "What if" Question ?
- Why not A/B testing?
- Causality ≠ Correlation

## Basic Concepts

- Outcome / Treatment Variables
- Intervention and Do Operation
- Counterfactual
- Causal Graph

## Causal Inference

- Statistical Inference vs. Causal Inference
- Average Treatment Effect (ATE) Estimation
- Causal Inference in Practice