Applied Machine Learning for Business Analytics

Lecture 11: Causal Inference for Decision Making

Lecturer: Zhao Rui

Logistics

- Thanks for all of your work in the kaggle competition.
    - We will decide the threshold based on the leaderboard snapshotted on Mar-27. And use the threshold to grade your kaggle competition performance in the final leaderboard.
- Appreciate if you keeps video on!

Agenda

1. Why do we need causal inference?
2. A/B Testing
3. Causal Inference
4. Summary

1. Why Causal Inference

From Data to Insights

- From Data, we can draw various kinds of insights
  - Compare groups of **subjects** on **important metrics**
- Subjects could be customers, products, team members and etc
  - Customers who used a "free shipping for orders over 80 sgd" coupon spent 20% more than shoppers who did not use the coupon
  - Products displayed in the front of the store were bought 12% more often than products in the back of store
  - Sales agents in the Boon Lay delivered 9% higher GTV-per-agent than agents in the Kent Ridge.
  - Customers who clicked on our recent launched product feature "financial news" transact 30% more monthly than users who did not click

Comparisons could give us insight into how the system really works

From Insights to Actions

- From Insights, we can take actions to improve the outcomes we would like to optimize
- **Insights**:
  - Products displayed in the front of the store were bought 12% more often than products in the back of store
- **Actions**:
  - Move weakly-selling products from the back of the store to the front, maybe their sales will increase by 12%?

From Insights to Actions

- From Insights, we can take actions to improve the outcomes we would like to optimize
- **Insights**:
  - Customers who clicked on our recent launched product feature "financial news" transact 50% more monthly than users who did not click
- **Actions**:
  - If we could incentive our users who did not click on financial news to click on the financial news the next time, maybe they will spend 50% more the following month?

# Will those actions have the desired effects?

Misleading Comparisons

- If the comparisons are calculated from experimental data, those actions should be solid
  - A/B Testing
- If the comparisons are calculated from observational data, those actions might not work
  - Confounding is often the reason
  - Causal inference is the tool to answer whether we can take actions based on the comparisons from the observational data

Toy Example

- We found
  - People who sleep with shoes on have a 60% higher chance to get headache more that those who sleep without shoes.
- Can we take actions as:
 - Suggest those people sleep without shoes to prevent headache

Toy Example

- We found
  - People who sleep with shoes on have a 60% higher chance to get headache more that those who sleep without shoes.
- Can we take actions as:
  - Suggest those people sleep without shoes to prevent headache
- Confounding factor:
  - Got drunk last night

Got Drunk last night

Causation

Sleep With Shoes On

Causation

Correlation

Headaches
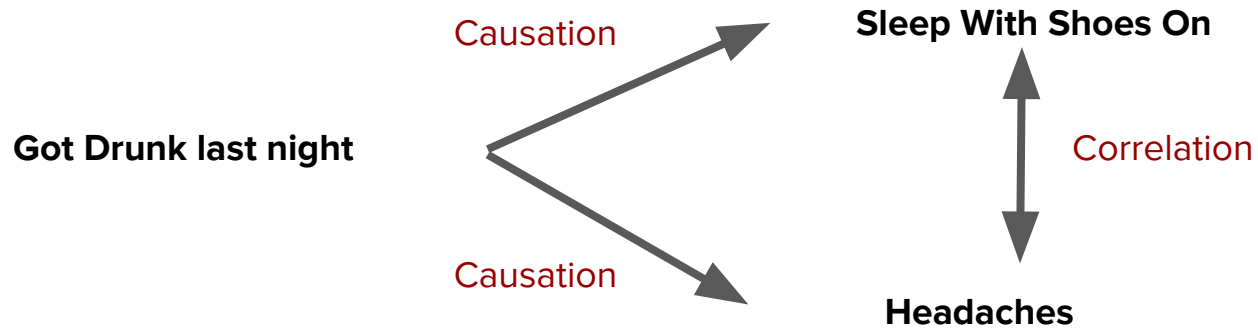
Biased Causal Conclusions from Observational Data

Health Expense
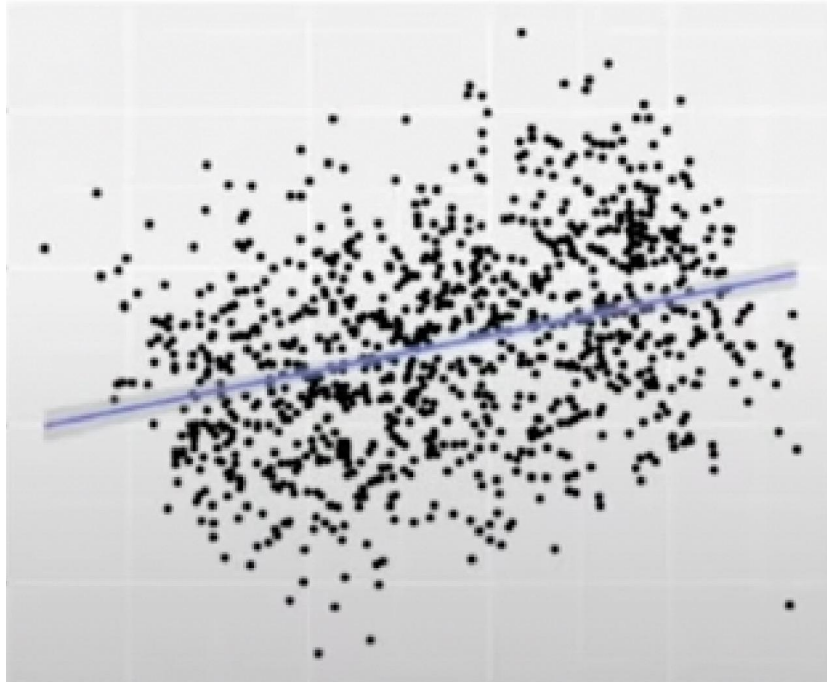
Exercise Level

**Exercise Level**

↓

Health Expense

Biased Causal Conclusions from Observational Data

- Based on observational data:
    - More exercise, higher health expense
    - Something must be wrong

Accounting for Confounders

- Control the age among samples
- In each age group, more exercise, lower health expense

Health
Expense

Exercise Level

age
— 10–19
— 20–29
— 30–39
— 40–49
— 50–59
— 60–69
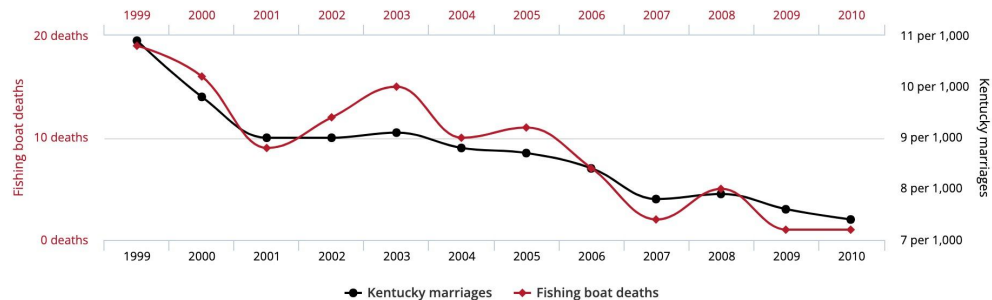
Exercise Level

Age

Health Expense
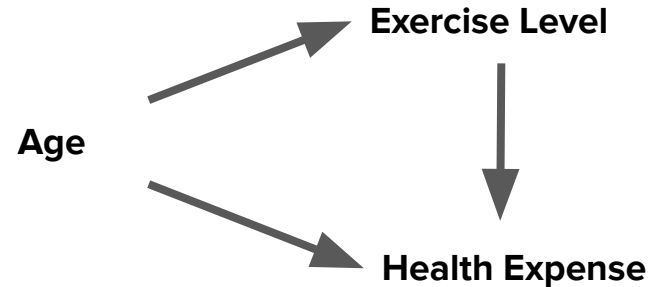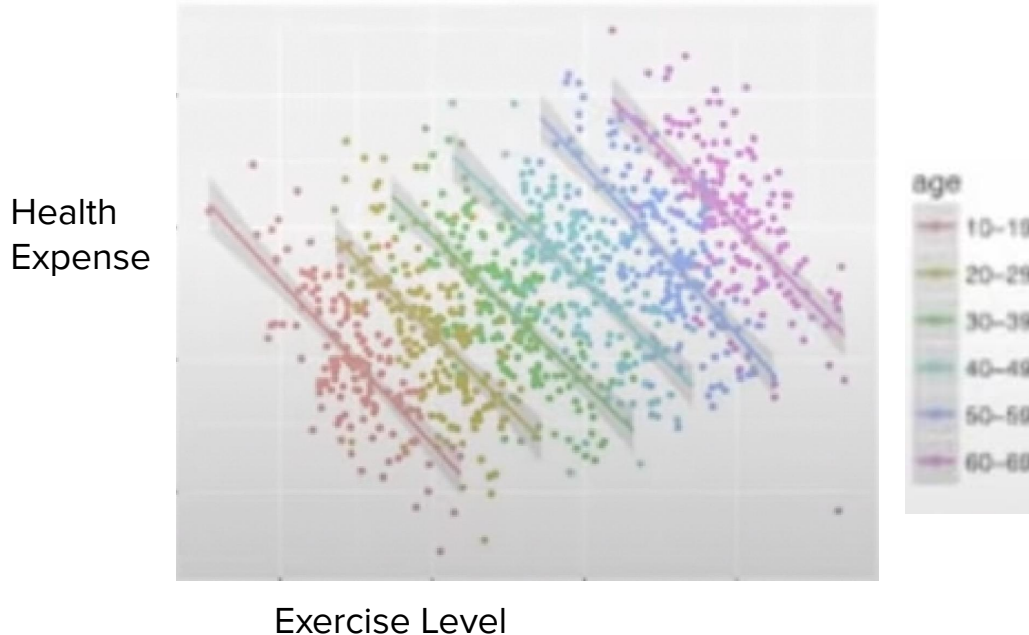
Correlation is not Causation

People who drowned after falling out of a fishing boat
correlates with
Marriage rate in Kentucky

Correlation: 95.24% (r=0.952407)

http://www.tylervigen.com/spurious-correlations

16

One Equation

Correlation - Confounders = Causation

We need to control for confounders. This is the key in Causal Inference

Applications of Causal Inference

- In Internet companies, causal inference is gaining increasing attention
- Users Growth
    - Better Product features and services
    - Monetary Incentives
    - Content Push
- Intelligent Marketing
    - Paid-ads
    - Billboard Campaign
- Flexible Pricing

Causal Inference Drives Better Decisions

- Decision Making is Challenging
  - Should we release this new product design?
    - Minor changes: button position
    - Major changes: homepage revamp
  - Should we give this user vouchers or not?
  - Should we run the crypto trading competition?
- Decentralized Execution
  - Get ideas quickly implemented
  - Lets data speak
- Learning and Prioritization
  - Learning could be compounded.

Establish Causality

- Association
  - Two variables change together
  - Due to Confounding Factors, correlation is not causation
- From Correlation to Causation
  - Remove all possible confounders
  - After controlling for confounders, association might be causation.
- How to control confounding factors
  - A/B Testing (Randomized Control Trial)
  - Quasi-experiments
  - Counterfactuals Inference

2. A/B Testing

A/B Testing

- KYC is one of important onboarding process at Pluang
    - Some of users tried to conduct verification but somehow did not complete
    - Growth team may propose: should we send emails to those users who tried but not completed KYC process?

We noticed that you have not completed your KYC.

For a limited time only, we are giving away 30 USD worth of ETH to every NEW trader on our platform.

However, we do not know whether those users just forgot to complete or have lost interest in the product

Identity Verification

To ensure your account security, an identity verification is required.

🔒 Data Protected Through Security Measures
Your data will be kept confidential and used for verification only in accordance with the applicable legal regulations.

Basic Verification
Unlocks Gold and Crypto Asset products
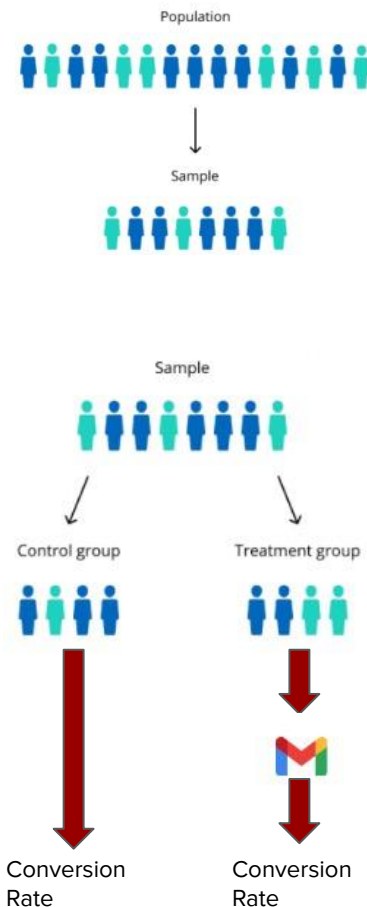Estimated completion ~2 minute

Additional Verification

A/B Testing

- From the whole population:
  - Get Samples randomly
- From samples:
  - divide into two groups randomly: treatment and control
- After experimentation period:
  - Compare the KYC conversion rate between control and treatment groups

Stable Unit Treatment Value Assumption

- SUTVA
  - The assumption should be hold during A/B Testing
  - If it is not true, your experimental results might be biased

- Definition (from Wiki)

Stable unit treatment value assumption (SUTVA)

We require that **"the [potential outcome] observation on one unit should be unaffected by the particular assignment of treatments to the other units"** (Cox 1958, §2.4).

Sampling

- Before experiments, we are not sure whether the impacts of treatments are positive or negative?
  - Sampling can control the potential risk
- Samples and its distribution should be representative of the whole population
 - Random sampling
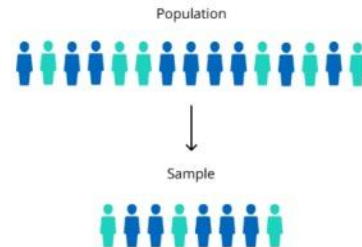
Hypothesis Testing

- Pre-Test
  - Choose the evaluation metric (Always CTR?)
  - Define **experiment parameters**
- During-Test
  - Data collection
  - No early stopping and no p-hacking!
- After-Test
  - Data analysis
    - Visualization
    - Significance Tests
      - T-test for continuous evaluation metrics
      - Z-test for rate evaluation metrics
      - Should we refer to bootstrap statistics?

Bayesian Testing should also be considered which requires less sample size
https://towardsdatascience.com/bayesian-a-b-testing-and-its-benefits-a7bbe5cb5103

Sometimes we can not do A/B Testing

- It is impossible to run A/B testing
  - Email vs Billboard
- Experimentation is with a high cost
  - Take months to draw conclusions
- Personalized Treatment
  - A/B testing can only estimate ATE: average treatment effect
  - Uplift modeling: select the best treatment for **each user** in order to maximize target business metric (estimation of individual treatment effect)
  - This is also the strength of machine learning models
- Can we just use observational data in the past to make decisions?
  - Causal Inference
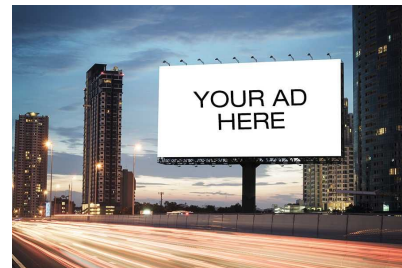
Treatment vs Control

Treatment

Control

Give vouchers or See new UI

Individual Treatment Effect

Potential Outcome if untreated

Potential Outcome if treated

Individual Treatment Effect

	Y(1)	Y(0)	Y(1) - Y(0)
	1	1	0
	0	0	0
	1	0	1
	0	1	-1

Average Treatment Effect

	Y(1)	Y(0)	Y(1) - Y(0)
	1	1	0
	0	0	0
	1	0	1
	0	1	-1

ATE: E[Y(1) - Y(0)]

Conditional Average Treatment Effect

Pre-exposure covariates

	Y(1)	Y(0)	Y(1) - Y(0)	X
	1	1	0	1
	0	0	0	0
	1	0	1	0
	0	1	-1	1

CATE: E[Y(1) - Y(0)|X=x]

What is the treatment effect for male users?

Estimation of Treatment Effects

- Depends on the own applications:
  - ATE
    - Global Change
  - CTE
    - Specific change for user groups
  - ITE
    - Personalized Service

We only have factual outcome

Unfortunately, we are not naruto

	Y(1)	Y(0)	Y(1) - Y(0)
	1	?	?
	0	?	?
	?	0	?
	?	1	?

Can we estimate the treatment effect from data

	Y(1)	Y(0)	Y(1) - Y(0)	Received Treatment T	Observed Outcome Y = Y(T)
👩	1	?	?	1	1
👨	0	?	?	1	0
🧑	?	0	?	0	0
👩‍🦰	?	1	?	0	1

E[Y(1) - Y(0)]

Can we estimate the treatment effect from data

	Y(1)	Y(0)	Y(1) - Y(0)	T	Y = Y(T)
👩	1	?	?	1	1
👨	0	?	?	1	0
🧑	?	0	?	0	0
👩‍🦰	?	1	?	0	1

E[Y|T=1] — E[Y|T=0] = E[Y(1) - Y(0)]

This only holds in randomized experiments

A/B Testing is not a personalized solution

Estimate Treatment Effects for all users

Estimate Treatment Effects for users living in JKT, age is around 25-28 and the recent tranx. asset is gold

ATE $\equiv$ E[Y|T=1] $-$ E[Y|T=0]

Can use simple averages

CATE $\equiv$ E[Y|T=1, X=x] $-$ E[Y|T=0, X=x]

Can't use simple averages anymore!

3. Causal Inference

Challenges for Causal Inference

- Confounders
  - In the past observational data, confounders are not controlled
- Goal: Does the new feature: feed page make users transact more?
- After the feed page feature has been launched for one month, we sample two group of users: clickers and non-clickers. Check whether they transact?

Challenges for Causal Inference: Confounders

- Goal: Does the new feature: feed page make users transact more?

Sample	Usages of feed page features	Make transactions
Alice	1	0
Bob	1	1
Sam	1	1
Tony	1	1
Jerry	1	0
Clay	0	1
Frank	0	1
William	0	0
Tom	0	0
David	0	0

Treatment Group: 60% transacted

Control Group: 40% transacted

Challenges for Causal Inference: Confounders

- Goal: Does the new feature: feed page make users transact more?

Sample	Usages of feed page features	Make transactions
Alice	1	0
Bob	1	1
Sam	1	1
Tony	1	1
Jerry	1	0
Clay	0	1
Frank	0	1
William	0	0
Tom	0	0
David	0	0

- The "treatment effect" is 0.2 uplift. Should we take the following actions?
  - Incentive more users to click the financial news
  - Make financial news tab as the default one
- The answer is no since treatment and control groups here are not randomly splitted.

Challenges for Causal Inference: Confounders

- Goal: Does the new feature: feed page make users transact more?

Sample	Usages of feed page features	Make transactions
Alice	1	0
Bob	1	1
Sam	1	1
Tony	1	1
Jerry	1	0
Clay	0	1
Frank	0	1
William	0	0
Tom	0	0
David	0	0

- ○ The confounding factor might be:
  - ■ User onboarding time
  - ■ New user will have less chance to explore all information and click on news tab. Then, they will also not make transactions.

Challenges for Causal Inference: Confounders

- Goal: Does the new feature: feed page make users transact more?

Sample √	Usages of feed page features	Make transactions
Alice	1	0
Bob	1	1
Sam	1	1
Tony	1	1
Jerry	1	0
Clay	0	1
Frank	0	1
William	0	0
Tom	0	0
David	0	0

Treatment Group: 60% transacted
    Average onboarding time: 9 months

Control Group: 40% transacted
    Average onboarding time: 7 months

Challenges for Causal Inference: Selection Bias

- Treatment group is not a good representation of all users in population
    - Here, treatment group only represents returning users
    - This is the selection bias due to confounders

Usage of feed page

Users Onboard Time
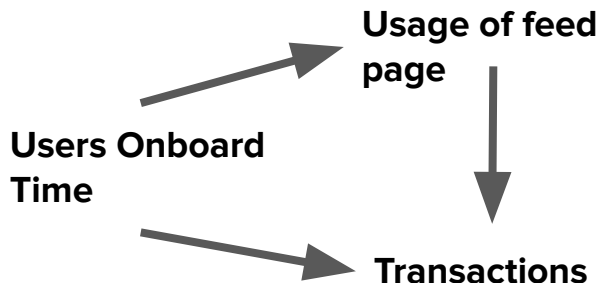
Transactions

Randomization is important

Challenges for Causal Inference: Counterfactuals

- Counterfactuals
  - Fact are truth: Bob made transaction
  - Counterfactuals are assumptions:
    - What if Bob did not use feed page, will he also transact or not?

Sample	Usages of feed page features	Make transactions
Alice	1	0
Bob	1	1
Sam	1	1
Tony	1	1
Jerry	1	0
Clay	0	1
Frank	0	1
William	0	0
Tom	0	0
David	0	0

Challenges for Causal Inference: Counterfactuals

- Counterfactuals
  - Fact are truth: Alice made transaction
  - Counterfactuals are assumptions:
    - What if Alice did not use feed page, will she transact or not?
- How to estimate counterfactual outcomes?

Sample	Control Outcome	Treatment Outcome
Alice	?	0
Bob	?	1
Sam	?	1
Tony	?	1
Jerry	?	0
Clay	1	?
Frank	1	?
William	0	?
Tom	0	?
David	0	?

Methods for Causal Inferences

- Matching
- Propensity Scores
- Bayesian Structural Time Series
- Other methods:
  - Stratification Analysis
    - https://towardsdatascience.com/the-magic-of-stratification-in-data-analysis-f1ee4800a283
  - Difference in Difference
    - https://medium.com/bukalapak-data/difference-in-differences-8c925e691fff
  - Fixed-effect Regression
    - https://towardsdatascience.com/fixed-effect-regression-simply-explained-ab690bd885cf
  - https://github.com/microsoft/EconML
  - https://github.com/uber/causalml

3.1 Matching

Matching

- Find confounders: x for each sample
- Identify pairs of treated (T=1) and control (T=0) units whose confounders x are similar or even identical to each other.
  - Distance (xi, xj) < e
- Small e: less bias but high variance

Confounder: x

Treatment: T
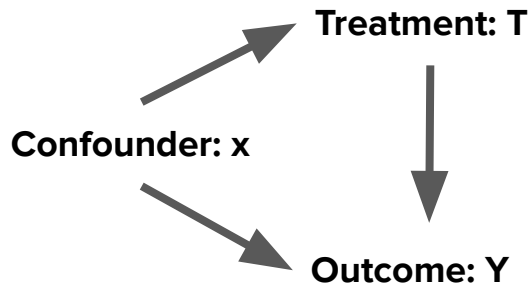
Outcome: Y

Matching

- Compare onboarding time between control users and treatment users

Sample	Onboarding Time	Control Outcome	Treatment Outcome
Alice	15	1	0
Bob	11	1	1
Sam	6	0	1
Tony	10	1	1
Jerry	3	0	0
Clay	13	1	0
Frank	10	1	1
William	6	0	1
Tom	5	0	1
David	1	0	0

1. Alice is matched to Clay
2. Bob is matched to Frank

Matching

- Individual treatment effect: Y(T=1) - Y(T=0)
- Average Treatment Effect:
  - Average of ITE over all samples
  - (-1+0+1+0+0-1+0+1+1+0)/10 = 0.1

Sample	Onboarding Time	Control Outcome	Treatment Outcome	ITE
Alice	15	1	0	-1
Bob	11	1	1	0
Sam	6	0	1	1
Tony	10	1	1	0
Jerry	3	0	0	0
Clay	13	1	0	-1
Frank	10	1	1	0
William	6	0	1	1
Tom	5	0	1	1
David	1	0	0	0

Matching

- If we only want to estimate Average Treatment Effect:
  - We can match each treated subject to a control subject
  - Then, we have a perfect balance on the covariante/confounders between treatment and control group
  - Then, we can compute E(Y|T=1) - E(Y|T=0_matched) as the estimation of ATE

Matching

- Based on confounders, we try to find similar pairs of data samples
- What if we have more than one confounders:
  - Curse of dimensionality
  - There will be few matches

| Edu Level	Income Level	Prior-period GTV	Onboarding Time	Feed Click Indicator	Transaction
High School	Medium	2000	5-10 sessions/daily	1	1
Colleague	Low	1500	20-50 sessions/daily	1	1
Primary School	High	500	50-60 sessions/daily	0	0

Confounders

3.2 Propensity Score

Propensity Score

- Propensity score e(x) is the probability of a unit/sample to get treated

$$e(x) = P(T = 1|x)$$

- Propensity score can not be observed, need to be estimated

Propensity Score

- Estimating propensity score: $\hat{e}(x) = P(T = 1|x)$
    - Classification problem: predicting the label (treatment vs control) based on observed features x
    - In the context, based on users features, predict whether this user will click feed page
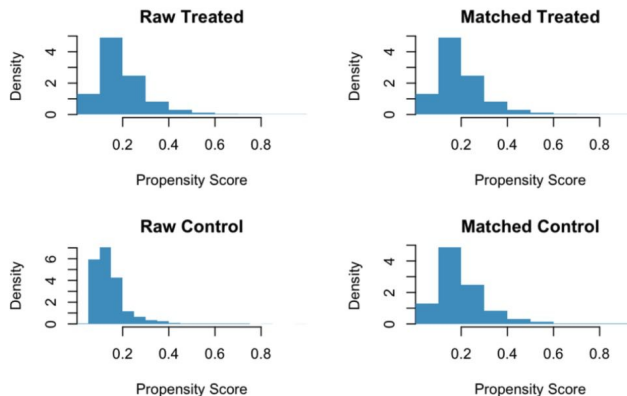   - Machine learning models can be used here

Confounders: X $\longrightarrow$ ML Models $\longrightarrow$ **Treatment: T**
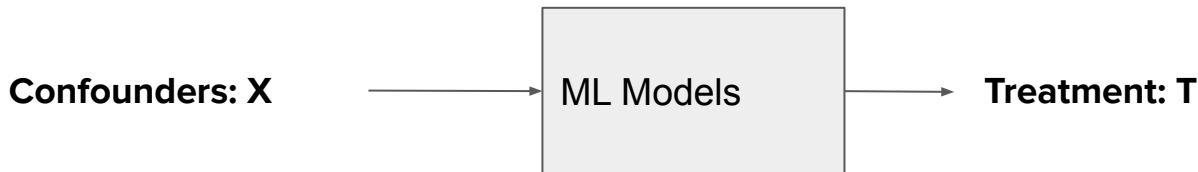
Propensity Score Matching

- Matching paris by distance between propensity score:

$$Distance(x_i, x_j) = |\hat{e}(x_i) - \hat{e}(x_j)|$$

- Compared to Matching, PSM is also kind of dimensionality reduction
  - From confounding factors to the estimated propensity scores

Inverse of Propensity Weighting (IPW)

- There is a biased comparison between control and treatment groups
- Based on propensity scores, we can re-weight each groups to reflect the true global distribution

Unit	$e(X)$	$1 - e(X)$	#units	#units (T=1)	#units (T=0)
A	0.7	0.3	10	7	3
B	0.6	0.4	50	30	20
C	0.2	0.8	40	8	32

Unit	#units (T=1)	#units (T=0)
A	10	10
B	50	50
C	40	40

Confounders are the same!

Distribution Bias

Reweighting by inverse of propensity score: $w_i = \dfrac{T_i}{e_i} + \dfrac{1 - T_i}{1 - e_i}$
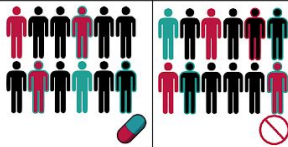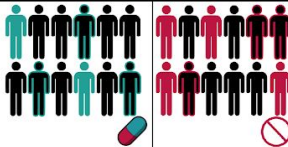
P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

Inverse of Propensity Weighting (IPW)

- Estimate ATE by IPW:
  - Equation:

$$ATE = \frac{1}{n} \sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}(x_i)} - \frac{1}{n} \sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\hat{e}(x_i)}$$

  - IPW creates a pseudo-population where the confounders' distribution are the same between treated and control groups
  - Compared to PSM, IPW use all data samples

3.3 Bayesian Structural Time Series

| Method	Experimental setup		Description	Causal evidence
Statistical Experiment			Control and treatment are not identical but divided at random. This makes it possible to build a precise estimate of the causal effect of treatment. *A/B testing, Central Limit Theorem, Bayesian Statistics*	⊕ ⊕ ⊕
Quasi-experiment			Control and treatment are not identical and divided by a "natural" criterion. Depending on "internal" and "external" quality of the criterion, it is possible to build a good estimate of the causal effect of treatment. *Differences-in-differences, Regression Discontinuity, Instrumental variables, Matching, Controlled Regression*	⊕ ⊕
Counterfactuals		?	Control group does not exist, instead its behaviour is estimated with a predictive model of what would have happened without the treatment (= counterfactual). *Synthetic Differences-in-Differences, Athey & Imbens, CausalImpact*	⊕
Descriptive statistics	**None**		There is no control group, no experimental setup. Trends are surfaced with the caveat that correlation is not causation.	⊖

Levels of evidence ladder for causal inference methods

Context

- Counterfactual Inference
  - The data only consists of observations of the treatment
  - Control group does not exist
  - New features can only be released to all the user base
    - TV Ads
    - Trading Competition
    - Blanket Campaign
- Estimation Comes
  - Predict what would happen had this feature not existed
  - It is not easy! (Find the confounding variables)
  - [Available Packages](#) (from google)-Bayesian Structural Model

Toy Examples
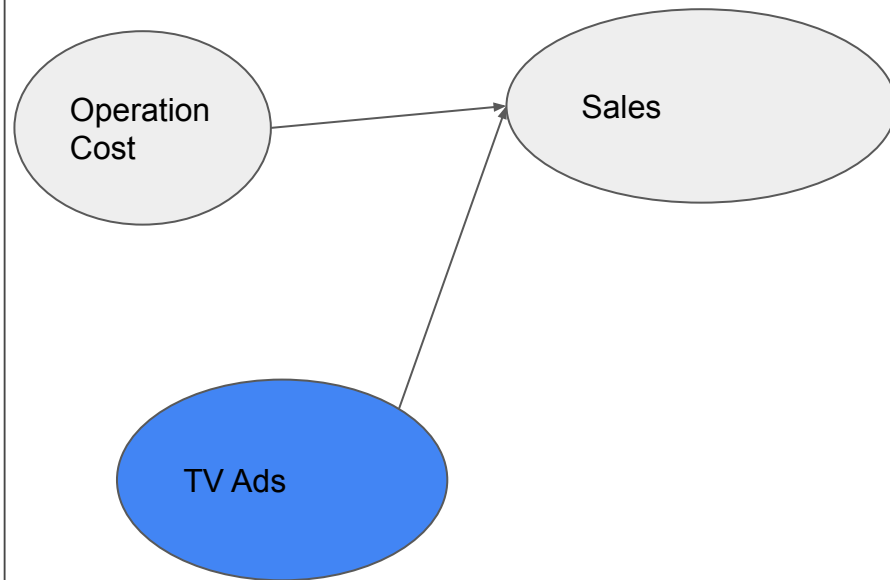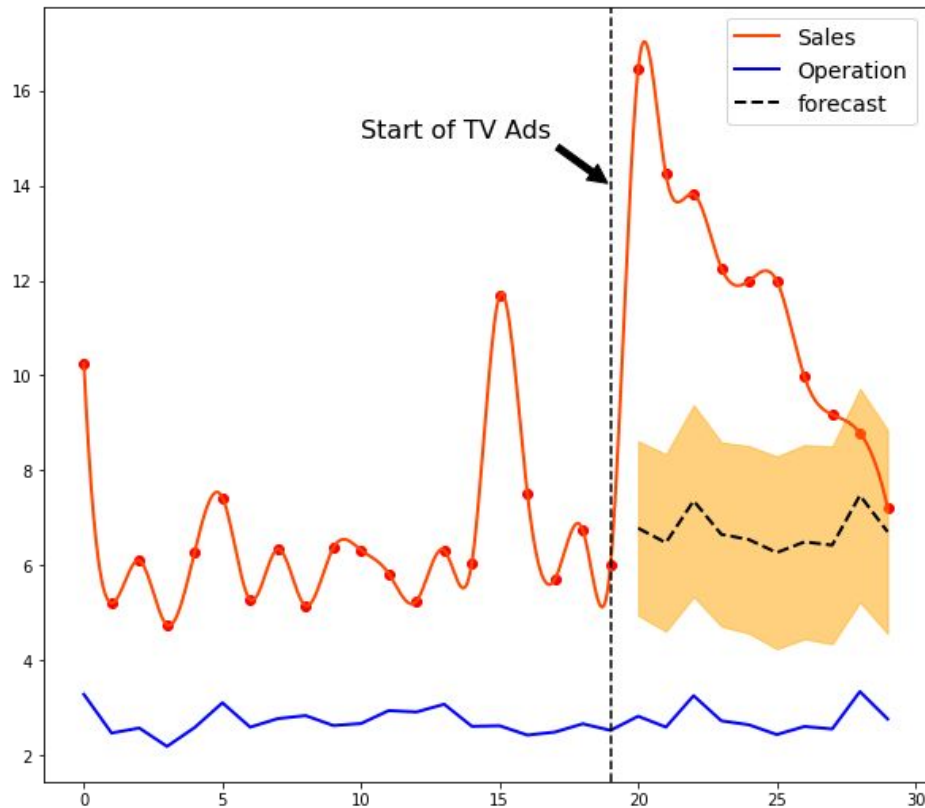
4. Summary

Correlation is not Causation

- To make decisions, we can use evidences from
    - A/B tests (experimental data)
    - Causal Inference (observational data)

Causal Inference + Machine Learning

- There is a trend in combining machine learning and causal inference
    - Machine Learning models are unable to generalize past the domain of examples present in a dataset. By capturing the causal relation, machine learning models can be more generalizable.

Causal Inference + Machine Learning

- There is a trend in combining machine learning and causal inference
  - Machine Learning models are unable to generalize past the domain of examples present in a dataset. By capture causal relationships, machine learning models can be more generalizable.
  - By utilizing machine learning models, causal inference can be more accurate and personalized.
    - Machine learning can bring personalization: from ATE to more accurate ITE
    - Machine learning can bring flexibility: learn which confounding factor to include

Confounders: X

Treatment: T

ML Models

Outcomes: Y