Machine Learning through the lens of causality

Matthieu Martin
Meetup Grenoble Data Science

January 31, 2020

Outline

Causal inference

Advertising at Criteo

Causality at Criteo

Causal inference

Causal inference: why is it hard ?

Motivation example: [Nobel prizes - choc. consumption]
Messerli 2012

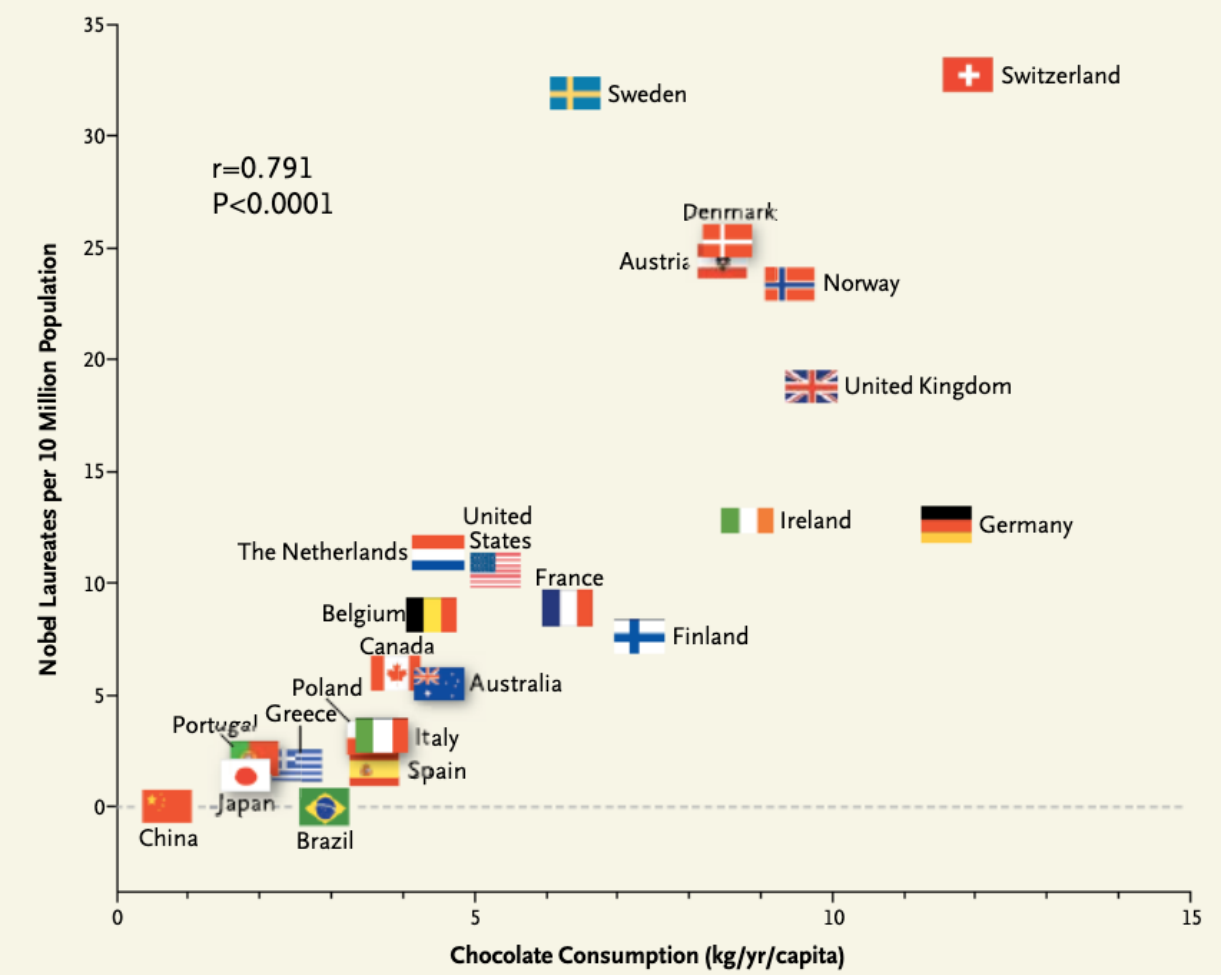


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Causal inference: why is it hard ?

Motivation example: [Nobel prizes - choc. consumption]
Messerli 2012

Causal inference: why is it hard ?

Motivation example: [Nobel prizes - choc. consumption]
Messerli 2012

Genuises are more likely to eat a lot of chocolate?

Causal inference: why is it hard ?

Motivation example: [Nobel prizes - choc. consumption] Messerli 2012

Eating chocolate produces Nobel prize?

Causal inference: why is it hard ?

Motivation example: [Nobel prizes - choc. consumption] Messerli 2012

The country economic strength of the country could explain both!

Causal inference: why is it hard ?

Motivation example: [Nobel prizes - choc. consumption]
Messerli 2012

Causal inference: why is it hard ?

Theorem (Reichenbach's common cause principle)

If a pair of variables X and Y can be embedded into a larger system, then Reichenbach's common cause principle states: if X ***and*** *Y* ***are (unconditionnally) dependent****, then there is:*

1. *either a directed path from X to Y, or*
2. *a directed path from Y to X, or*
3. *there is a node Z with a directed path from Z to X and from Z to Y.*

How can we choose?

Idea: **intervene** on the system, e.g. give more chocolate in some countries and measure the impact on the number of Nobel laureates.

Intervention on a model: Pearl's *do*

Motivation example: [Altitude - Temperature]

X: altitude

Y: temperature

$$X = N_X$$
$$Y = -6X + N_Y$$
$$N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

$$(X, Y) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -6 \\ -6 & 37 \end{pmatrix}\right)$$

Intervention: change **some** distribution variable on the system

Intervention on a model: Pearl's *do*

Motivation example: [Altitude - Temperature]

X: altitude

Y: temperature

~~$X = N_X$~~ $do(X = 2)$

$Y = -6X + N_Y$

$N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$

$$do(X = 2) \Rightarrow Y \sim \mathcal{N}(-12, 1) \neq \mathcal{N}(0, 37)$$

Intervening on X changes the distribution of Y. But...

Intervention on a model: Pearl's *do*

Motivation example: [Altitude - Temperature]

X: altitude

Y: temperature

$X = N_X$

~~$Y = -6X + N_Y$~~ $do(Y \sim \mathcal{N}(2,2))$

$N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$

$$do(Y \sim \mathcal{N}(2,2)) \Rightarrow (X,Y) \sim \mathcal{N}\left(\begin{pmatrix}0\\2\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 2\end{pmatrix}\right)$$

No matter how strongly we intervene on Y, the distribution of X remains as before.

Intuition of "X causing Y".

Questions?

Causal inference: why is it hard ?

- Two treatment variants, prescribed by physician:

  - *Treatment*$_1$: open surgical procedure
  - *Treatment*$_2$: ultrasonic probes and small puncture procedure

	Treatment$_1$	*Treatment*$_2$
Total	273/350= 78%	289/350= **83%**

Table 1: Recovery rate

Causal inference: why is it hard ? (2/2)

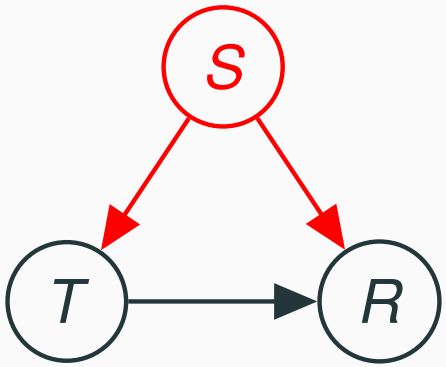
Treatment$_2$ **seems better...**

	Treatment$_1$	*Treatment*$_2$
Total	273/350= 78%	289/350= **83%**

Table 2: Recovery rate, global

But is it?

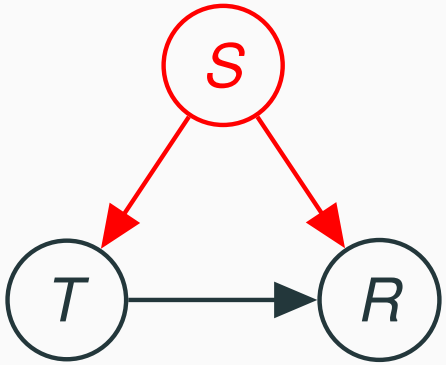
Simpson's paradox (*confounder*: Stone size)



	Treatment$_1$	*Treatment*$_2$
Total	273/350= 78%	289/350= **83%**
Small stone	81/87= **93%**	234/270= 87%
Big stone	192/263= **73%**	55/80= 69%

Table 3: Recovery rate, decomposed

Causal inference: fix #1



	$Treatment_1$	$Treatment_2$
Total	273/350= 78%	289/350= **83%**
Small stone	81/87= **93%**	234/270= 87%
Big stone	192/263= **73%**	55/80= 69%

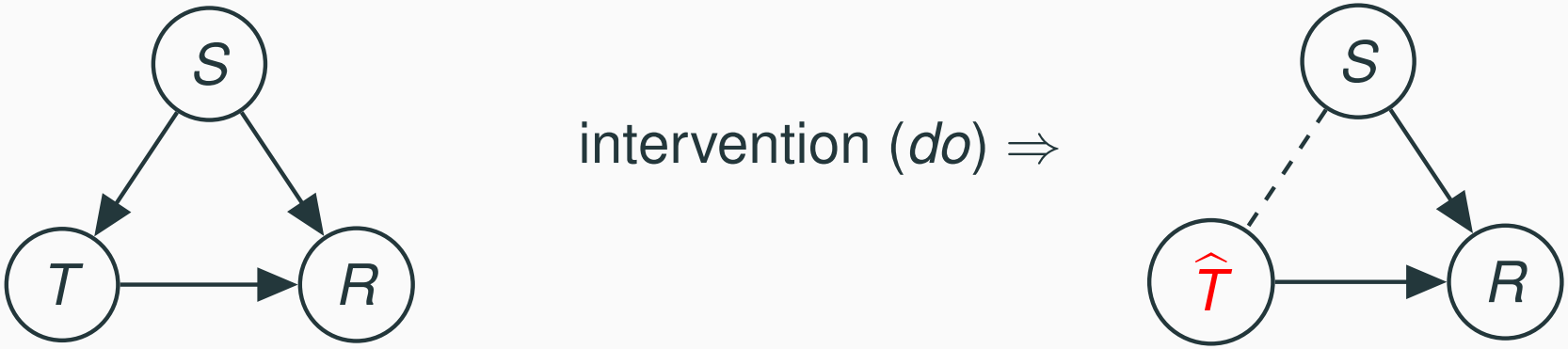
- Here it inverted the true causal effect.
- $P(R = 1|do(T = Tr_1)) \neq P(R = 1|T = Tr_1)$

Causal inference: fix #2

Fix #1: add potential confounders to the model.

- Confounding bias pull the observed causal relation away from the true association.

Fix #2: break confounders to avoid their effects



In practice (if ethical) **assign** treatment at random to avoid selection bias (A/B test)

Questions?

Advertising at Criteo

Advertising at Criteo

Advertising at Criteo

User 123456789

Advertising at Criteo

For Sale

User 123456789

Advertising at Criteo

For Sale

User 123456789

0,15€

0,10€

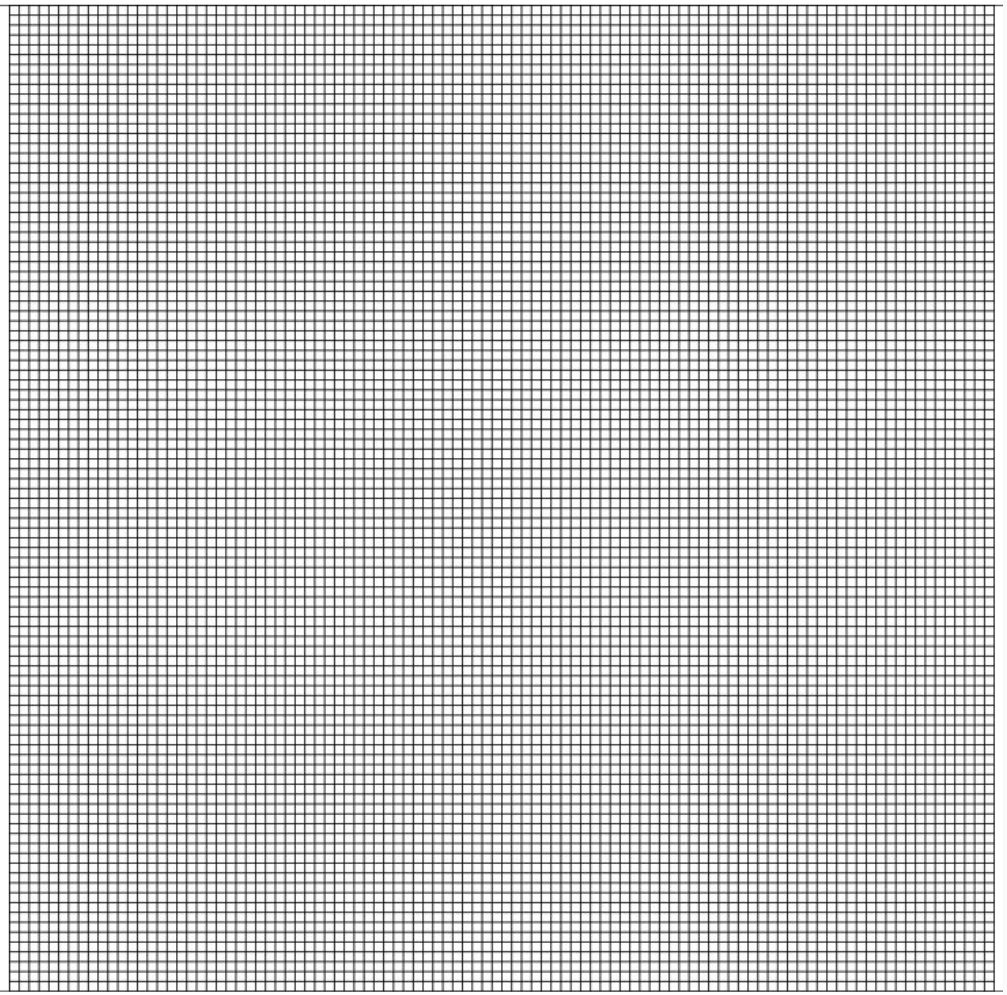
Advertising at Criteo

Advertising at Criteo

Winning Advertiser

Huge volume of data - tiny signal

10 000 displays

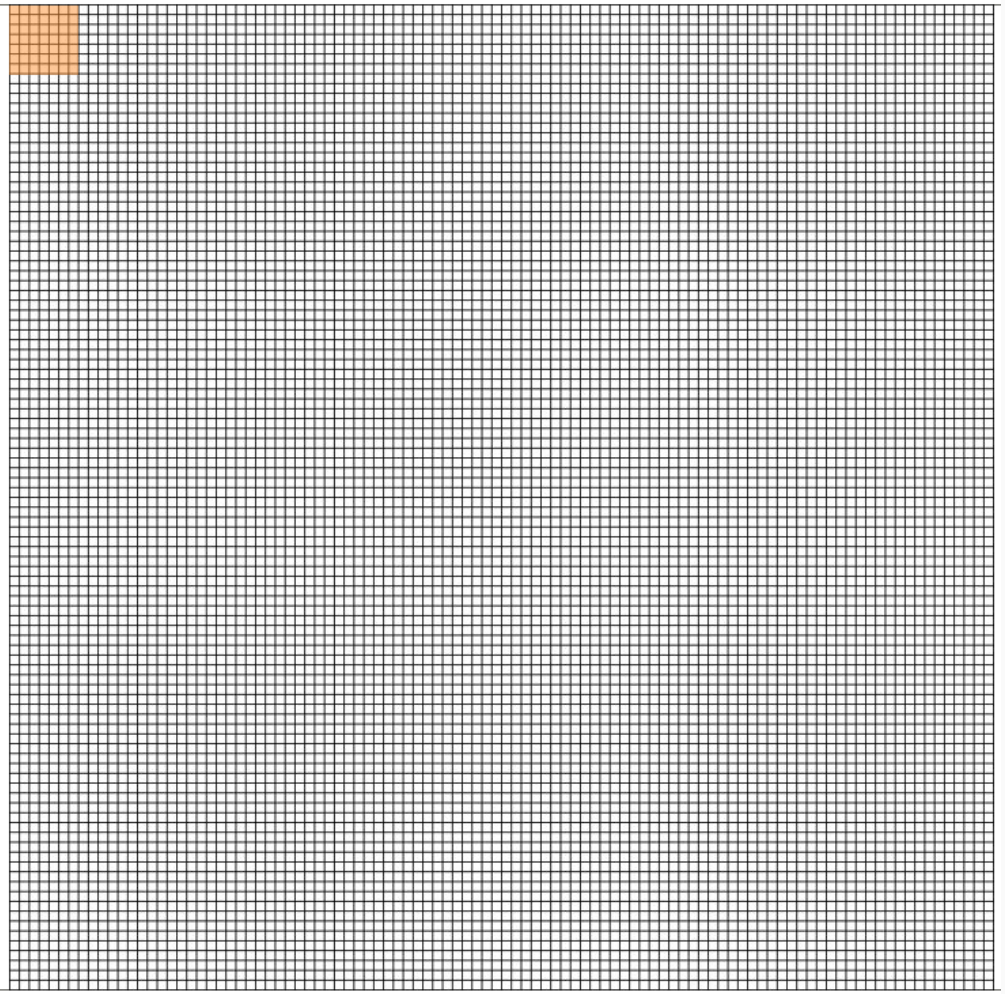


Huge volume of data - tiny signal

10 000 displays

leads to

50 clicks



Huge volume of data - tiny signal

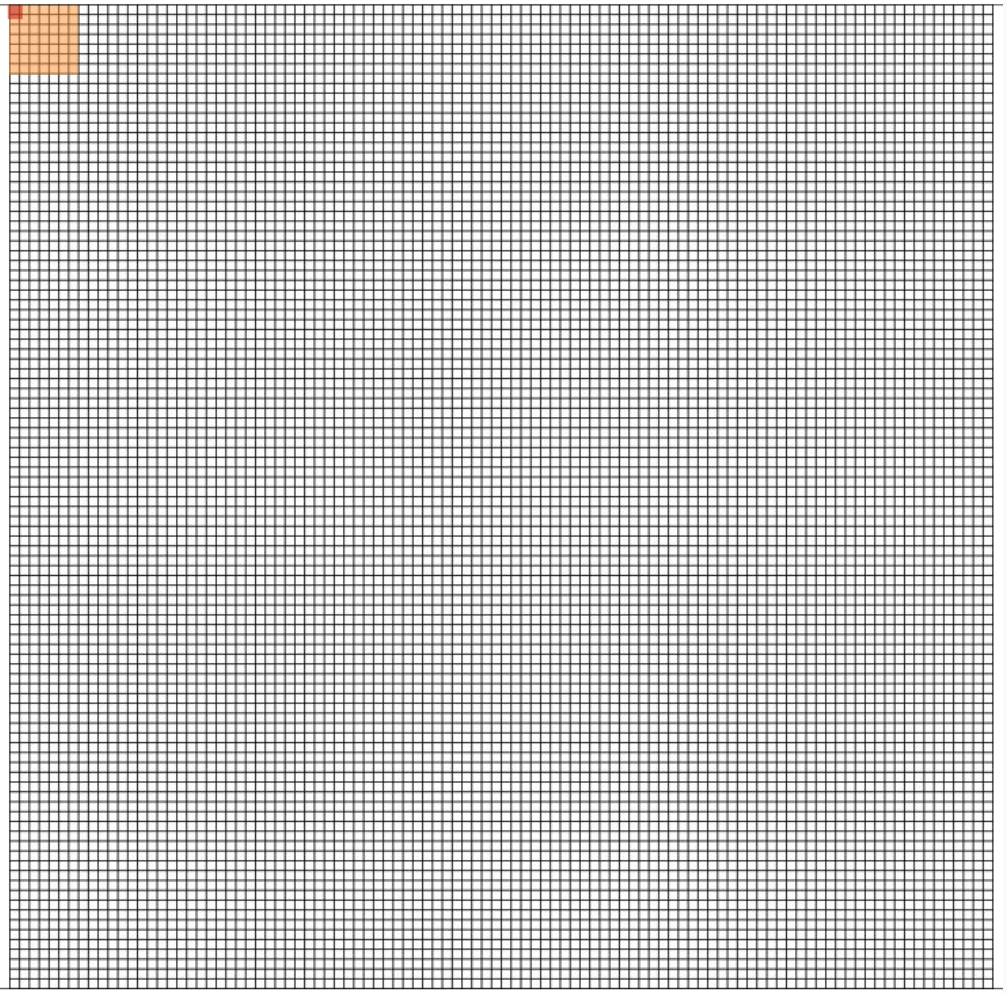
10 000 displays

leads to

50 clicks

leads to

1 sale



Causality at Criteo

Performance advertising metrics

What we measure in practice

Number of sales *attributed* to advertising

- Commonly: last click attribution
- Available on individual units, immediate feedback

What we would like to measure

Number of sales *caused* by advertising for a given budget

- But difficult/costly to measure and optimize
- Only at population level, delayed

Key questions

1. How can we estimate the causal effect of a treatment (e.g. ads) on a given target variable (e.g. sales)?
2. How can we simulate change of the target variable had we acted differently?

Uplift Models

Definition

Uplift: *causal impact of a treatment* at the *individual level*

Purpose

Predict if treating or not an individual *would be more beneficial*

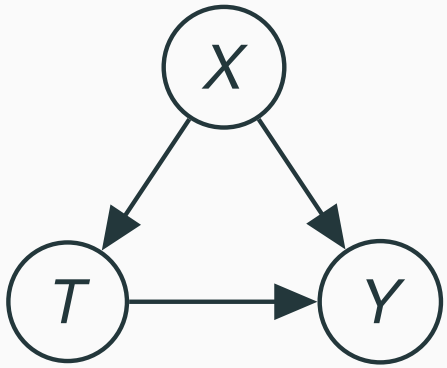
Applications

- Personalized medicine - treatment is a medication
- Digital advertising - treatment is exposure to ads

Simple causal model of advertising

Assumption: i.i.d. users where we observe

- X: users' features
- T: treatment (display)
- Y: response (conversion)



Goal: estimate $U^{do(T)}(x)$

$$U^{do(T)}(x) = \underbrace{P(Y=1|X=x, do(T=1))}_{\text{conversion proba if treated}} - \underbrace{P(Y=1|X=x, do(T=0))}_{\text{conversion proba if } \textbf{not} \text{ treated}}$$

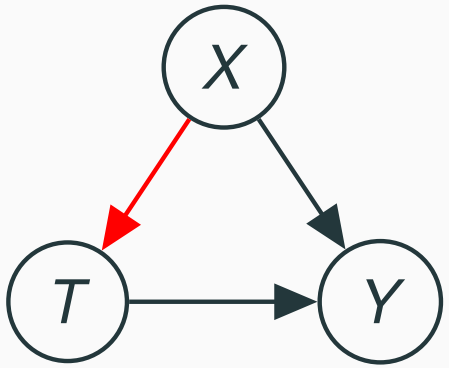
Observe:

$$u^{obs.}(x) = \underbrace{P(Y=1|X=x, T=1)}_{\textbf{observed} \text{ conditional proba}} - \underbrace{P(Y=1|X=x, T=0)}_{\textbf{observed} \text{ conditional proba}}$$

Simple causal model of advertising

Assumption: i.i.d. users where we observe

- X: users' features
- T: treatment (display)
- Y: response (conversion)



Goal: estimate $U^{do(T)}(x)$

$$U^{do(T)}(x) = \underbrace{P(Y=1|X=x, do(T=1))}_{\text{conversion proba if treated}} - \underbrace{P(Y=1|X=x, do(T=0))}_{\text{conversion proba if } \textbf{not} \text{ treated}}$$

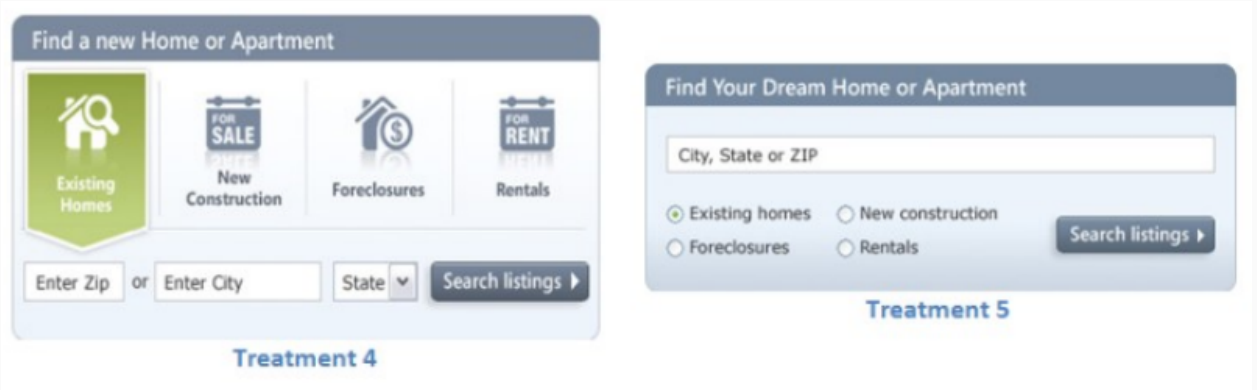
Observe:

$$u^{obs.}(x) = \underbrace{P(Y=1|X=x, T=1)}_{\text{conversion proba if treated}} - \underbrace{P(Y=1|X=x, T=0)}_{\text{conversion proba if } \textbf{not} \text{ treated}}$$

Solution to break all cofounder: A/B test

A/B test: Experiment where

- subjects are randomly assigned to groups
- groups are treated differently
- con's: costly!

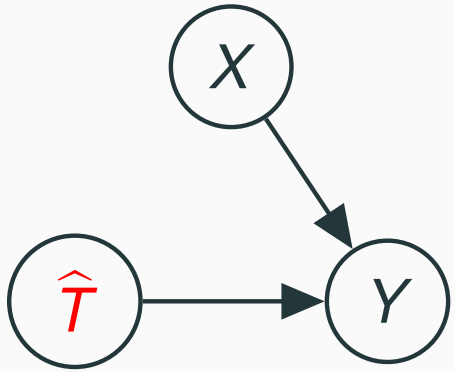


Uplift Prediction

$$U^{do(T)}(x) = P(Y = 1|X = x, do(T = 1)) - P(Y = 1|X = x, do(T = 0))$$

$$u^{obs.}(x) = P(Y = 1|X = x, T = 1) - P(Y = 1|X = x, T = 0)$$

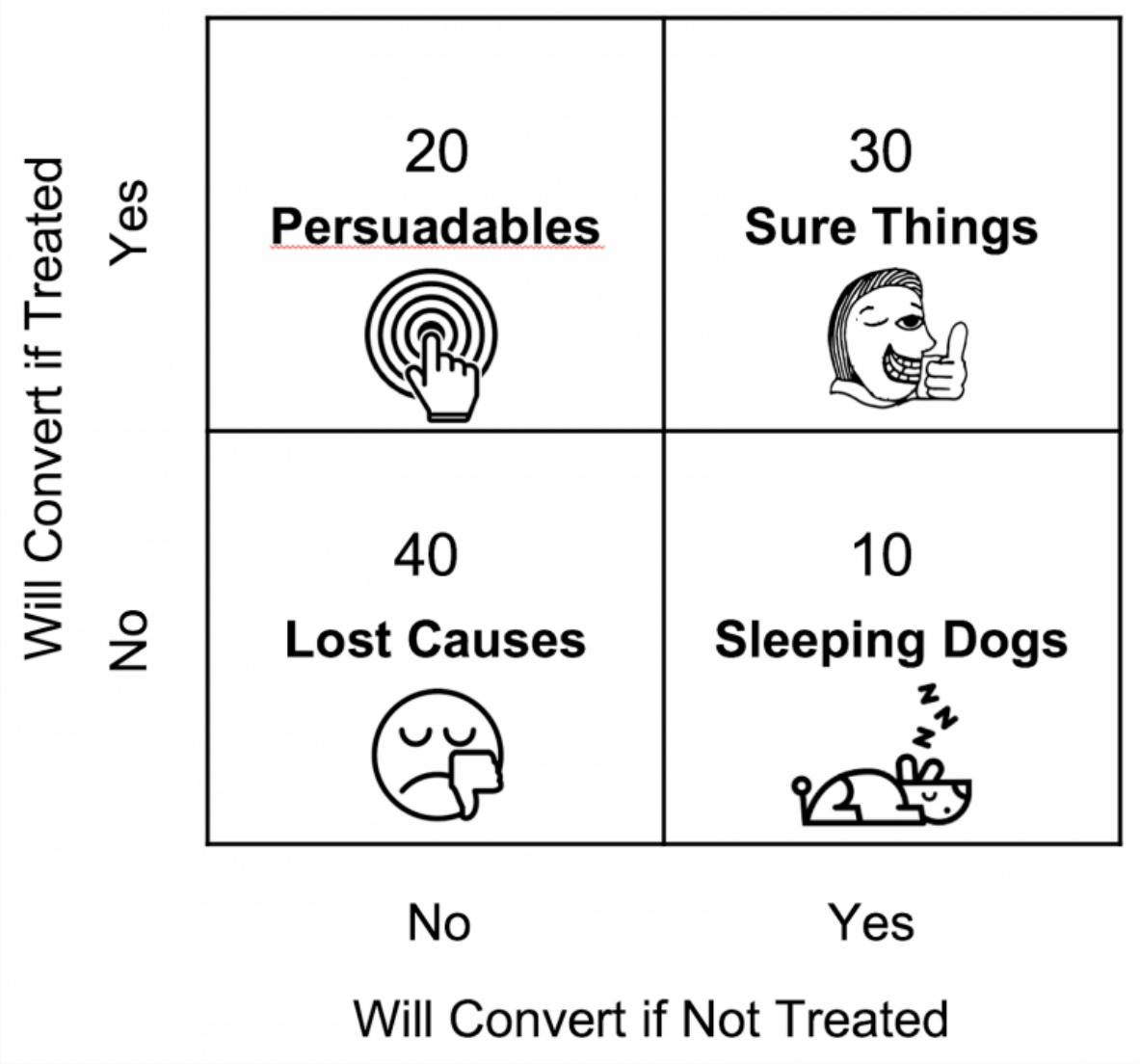
Assume we have a dataset with n points: $\mathcal{D} = \{x_i, y_i, t_i\}_{i=1...n}$



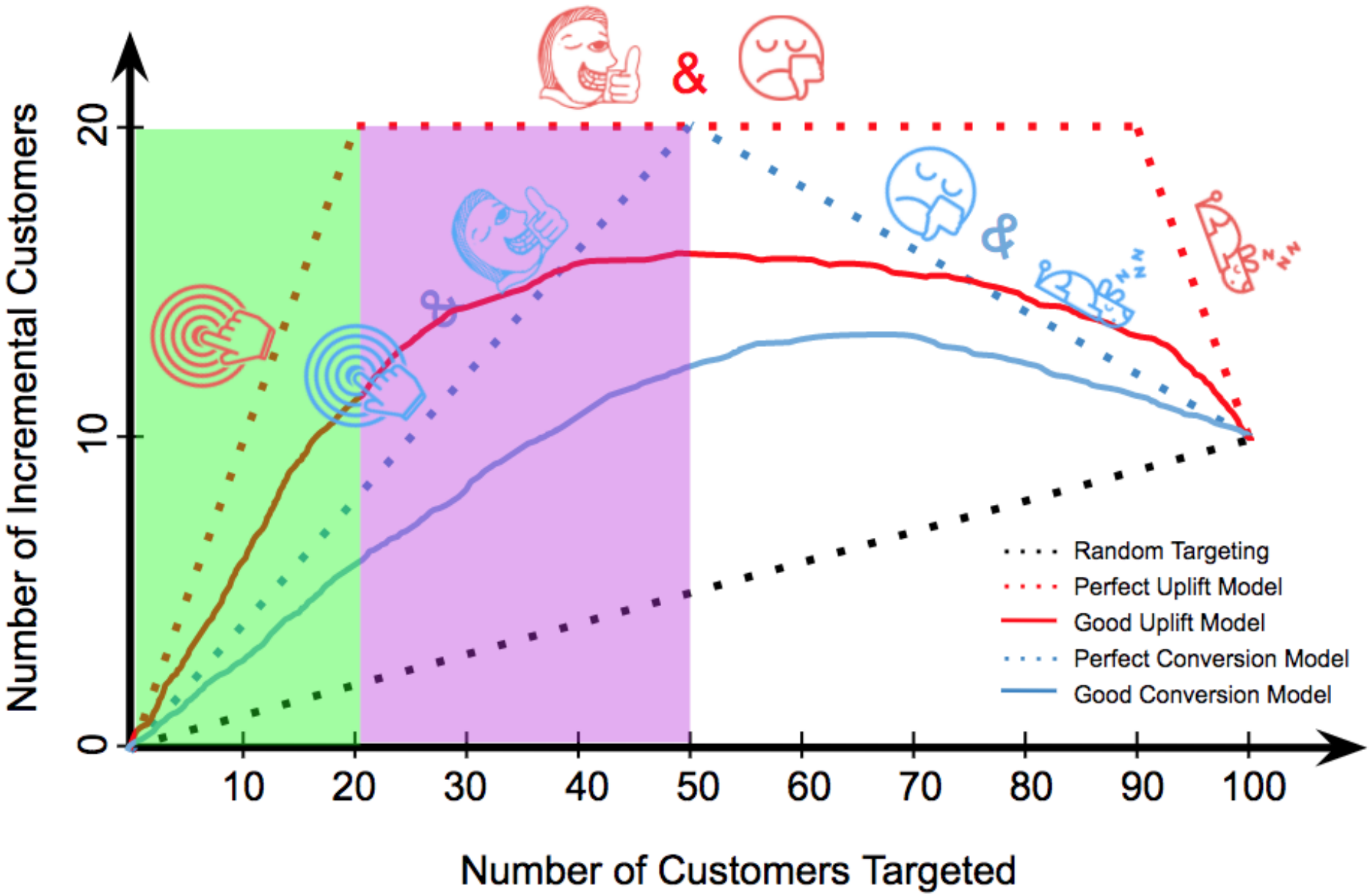
$$T \perp\!\!\!\perp X \Rightarrow U^{do(T)}(x) \equiv u^{obs.}(x)$$

Uplift: Divide and conquer

Imagine we have 100 individuals...



Uplift: Divide and conquer - Model evaluation



Uplift Prediction

- $\not\equiv$ Average Treatment Effect:
  $P(Y = 1|T = 1) - P(Y = 1|T = 0)$
- $\equiv$ Individual Treatment Effect *(= individual level incrementality)*
- $\equiv$ Counter-factual prediction

Fundamental Problem of Causal Inference: *no unit is both treated and untreated*

How to evaluate the uplift?

Several Machine learning models:

- 2-models
- causal trees / forests
- NN based models (work in progress)...

Work done on uplift: takeaways

Uplift Prediction at criteo

- Promising results: one step towards causal advertising
- Challenges: data collection, treatment+class imbalance, evaluation (e.g. low exposure in prod)

Public dataset

`CRITEO-UPLIFT1`: individual level incrementality A/B tests data, with 12 features, treatment and 2 targets benchmark to study uplift. $\hookrightarrow$ (`http://ailab.criteo.com/` (450MB))

Ongoing work

Investigation & testing of 'deep' uplift models:

- Dependant data representation (DDR)
- Shared data representation (SDR)

More details: Betlei, Diemert and Amini (Neurips '19 submission)

Questions?

Counterfactual reasoning

- **Goal**: evaluate multiple possible policies and what would happen if we were to apply them.
- **Idea**: randomize the production/logging policy to estimate all possible variants

Importance Sampling

- For each display draw a random variable for color selection
- Choose one of two colors:
  - Red with probability 0.80
  - Green with probability 0.20

	Proba(red)	Proba(green)
Production	80 %	20 %
Test	20 %	80 %

	Clicks(red)	Clicks(green)
Production	1000	300
Test	?	?

Importance Sampling

	Proba(red)	Proba(green)
Production	80 %	20 %
Test	20 %	80 %

	Clicks(red)	Clicks(green)
Production	1000	300
Test	250	1200

$$Clicks_{Test} \approx Clicks_{Prod}(Red) \times \frac{1}{4} + Clicks_{Prod}(green) \times 4 \quad (1)$$

$$= 1000 \times \frac{0.2}{0.8} + 300 \times \frac{0.8}{0.2} = 1450 \quad (2)$$

with $\frac{1}{4} = \mathbb{P}_{Test}(Red)/\mathbb{P}_{Prod}(Red)$

$4 = \mathbb{P}_{Test}(Green)/\mathbb{P}_{Prod}(Green)$

Importance Sampling

$$Clicks_{Test} = \frac{\mathbb{P}_{Test}(red)}{\mathbb{P}_{Prod}(red)} \times Clicks_{Prod}(red) + \frac{\mathbb{P}_{Test}(green)}{\mathbb{P}_{Prod}(green)} \times Clicks_{Prod}(green)$$

$$\mathbb{E}_{Test}[C] \simeq \sum_{i \in samples} w_i \times C_i$$

where:

- $w_i := \frac{\mathbb{P}_{Test}(A=a_i)}{\mathbb{P}_{Prod}(A=a_i)}$
- a_i the action sampled (under Prod distribution) on line i
- R_i reward observed on sample i

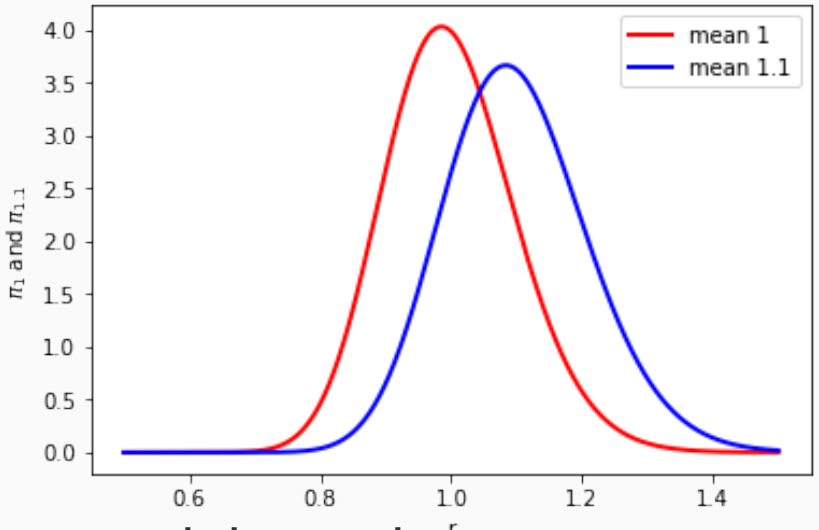
$$\mathbb{E}_{\pi_{test}}[C] = \mathbb{E}_{\pi_{prod}}\left[C\frac{\pi_{test}(r)}{\pi_{prod}(r)}\right]$$

Continuous variables setting

How to estimate the outcome of an other action, i.e. how to explore other bidding policies?

$$bid(X) = eCPM(X)$$

$$bid(R, X) = eCPM(X) \times R \quad \text{with} \quad R \perp\!\!\!\perp X$$

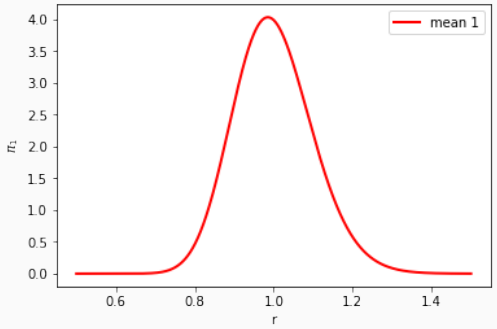


1. Randomize around the action
2. Estimate the outcome over a different (action) policy distribution by IS

Continuous variables

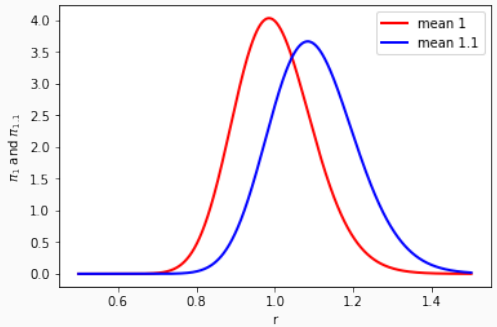
How to estimate the outcome of an other action?

1. Randomize around the action



2. Estimate the outcome over a different action distribution:

$$\mathbb{E}_{\pi_\alpha}[C] = \mathbb{E}_{\pi_1}[C\frac{\pi_\alpha(r)}{\pi_1(r)}]$$



Bid Randomization

criteo AI Lab

Idea

Multiply eCPM by a random factor to **explore slightly different bidding policies**:

$$bid(R, X) = eCPM(X) \times R$$

R: drawn from a known probability distribution with pdf f_R

Outcome

$$Y = \mathbb{E}_{R,X}[y(R, X)]$$

$y(R, X)$: reward (#sales or #visits, total cost, etc.)

Estimation

$$\hat{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i(R_i, X_i)$$

User Bid Randomization at criteo

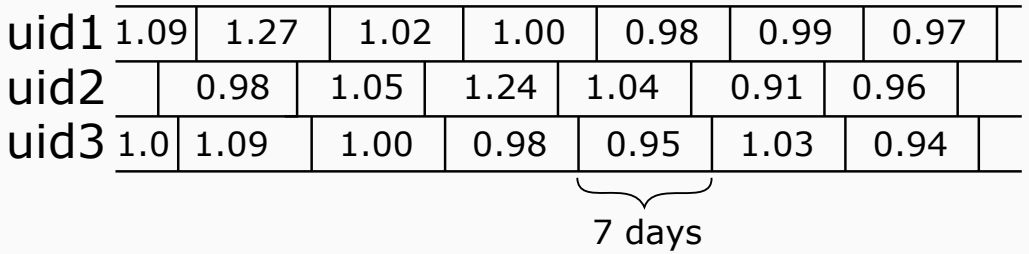
criteo AI Lab

Random bid: Multiply eCPM by a random factor to explore

$$bid(R, X) = eCPM(X) \times R \quad \text{with} \quad R \perp\!\!\!\perp X$$

User timelines and randomization

Sequences of 7-day periods with constant bid randomization



Question: counterfactual

What would have happen if R is randomized with $\pi_\alpha \neq \pi_1$?

Answer: counterfactual estimation using IS

$$\mathbb{E}_{\pi_\alpha}[C] \approx \mathbb{E}_{\pi_1}\left[C\frac{\pi_\alpha(r)}{\pi_1(r)}\right]$$

Offline Estimation

Dataset: $\mathcal{D} = \{x_i, r_i, y_i, c_i\}_{i=1...n} \sim (X, R, Y, C)$

Production policy

Estimation of the *outcome*:

$$\mathbb{E}_{\pi_1}[y(R, X)] \approx \frac{1}{N} \sum_{i=1}^{N} y(r_i, x_i)$$

Test policy

Estimation of the *counterfactual outcome* by IS:

$$\mathbb{E}_{\pi_\alpha}[y(R, X)] = \mathbb{E}_{\pi_1}\left[y(R, X)\frac{\pi_\alpha(X)}{\pi_1(X)}\right] \approx \frac{1}{N} \sum_{i=1}^{N} y_i(R_i, X_i) w(\alpha, R_i)$$

with $w(\alpha, R_i) = \frac{f_R(\frac{R_i}{\alpha})}{\alpha f_R(R_i)}$ (unbiased but variance problems)

Optimization formulation

Target function

We wish to maximize the expected difference in sales:

$$U(\pi_\alpha) = \mathbb{E}_{\pi_\alpha}[y(.)] - \mathbb{E}_{\pi_1}[y(.)]$$

Constrained optimization problem

We want to increase user's actions while controlling our cost (*iso-cost maximization*):

$$\alpha^* \in \arg\max_\alpha U(\pi_\alpha) \text{ s.t. } \left(\frac{\mathbb{E}_{\pi_\alpha}[C]}{\mathbb{E}_{\pi_1}[C]} - 1\right)^2 \leq \epsilon$$

Towards solving this pb: Diemert, Heliou and Renaudin (Neurips '18)

Feature dependent model for R

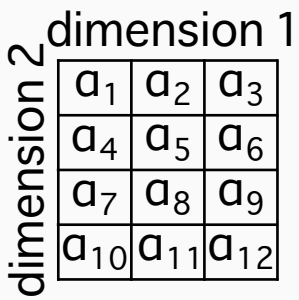
criteo AI Lab

Bid modification

$$bid(R, X = x) = eCPM(X = x) \times R$$

Per 'strata' user representation

Partition the user space in S strata



For $\mathbb{R}^2 \ni x \mapsto s \in S$ where we sample using a policy $R_s \sim \log \mathcal{N}(\mu_\alpha^s, \sigma_\alpha^s)$.

Idea: if strata s represents frequent/good shoppers, choose $\mu_\alpha^s > \mu_1 = 1$.

Offline A/B tests

Validate the method using a dataset of online A/B tests

False negatives are more costly

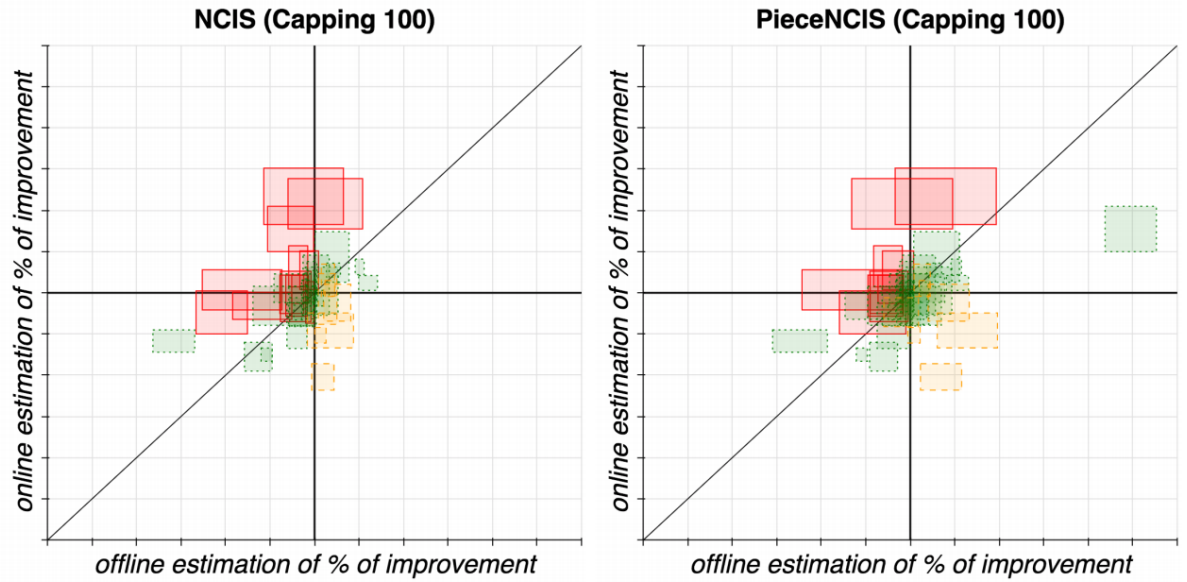


Figure 5: Comparison of online/offline decision. A box is an A/B test. The width (resp. height) of a box is a 90% confidence bound on the offline (resp. online) uplift. The scale is the same for both axis. Green/dotted: right decision. Orange/dashed: false positive. Red/plain: false negative.

References

This presentation contains material from

- Elements of Causal Inference (Peters, Janzing, Schoelkopf- MIT Press - 2017)
- Causality Tutorial (Calauzènes, Gilotte, Diemert, Aslan - Criteo ML Big Days - 2018)
- https://tech.wayfair.com/data-science/2018/05/uplift-modeling-in-display-remarketing/

Thank You!

Questions?

ping me @

mat.martin@criteo.com