Economics & Experimentation

Economics
“We love causal inference”

&

Experimentation
“We love causal inference”

ORIGIN

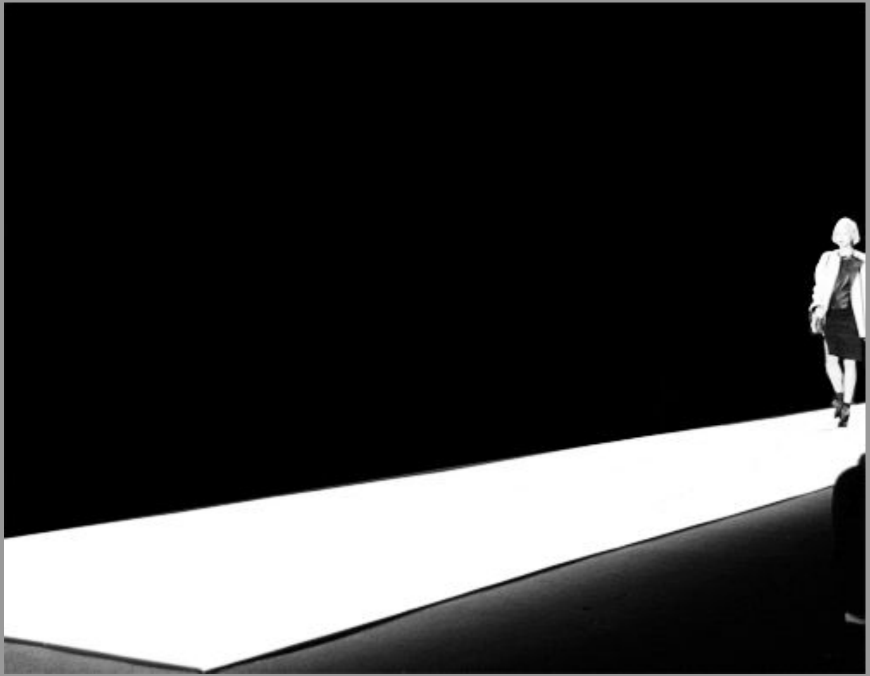
It’s 2019

Internal Customer Need

I believe my product is pretty great, but . . .

Internal Customer Need

. . . how do I *know* if it's impactful for customers?



Incrementality

Teams want to know the incremental impact of their work.

Incrementality is a Counterfactual Comparison

$$\text{Impact} = \text{World With Product} - \text{World Without Product}$$

Fundamental Problem

We can't observe the counterfactual

Ladder of Causation

Association

What if I see . . . ?

Prediction

$$P(y|x)$$

Pearl (2019)

Ladder of Causation

Intervention

What if I do . . . ?

$$P(y|do(x),\ z)$$

Association

What if I see . . . ?

Prediction

$$P(y|x)$$

Pearl (2019)

Ladder of Causation

Counterfactuals
What if I had . . .? Why?

$$P(y|x', y')$$

Intervention
What if I do . . . ?

$$P(y|do(x),\ z)$$

Association
What if I see . . . ?
Prediction

$$P(y|x)$$

Pearl (2019)

Causation!

Teams ask causal questions every day

. . .

Tools

. . . but often lack the tools to answer their causal questions.

Tools: Spectrum of Control

Data Generating Process
(DGP)

High Control
of DGP

Low Control
of DGP

Tools: Spectrum of Control

Data Generating Process (DGP)

High Control of DGP

Low Control of DGP

A/B Test

Causal Identification based on controlled random assignment of treatments

Tools: Spectrum of Control

Data Generating Process (DGP)

High Control of DGP

Low Control of DGP

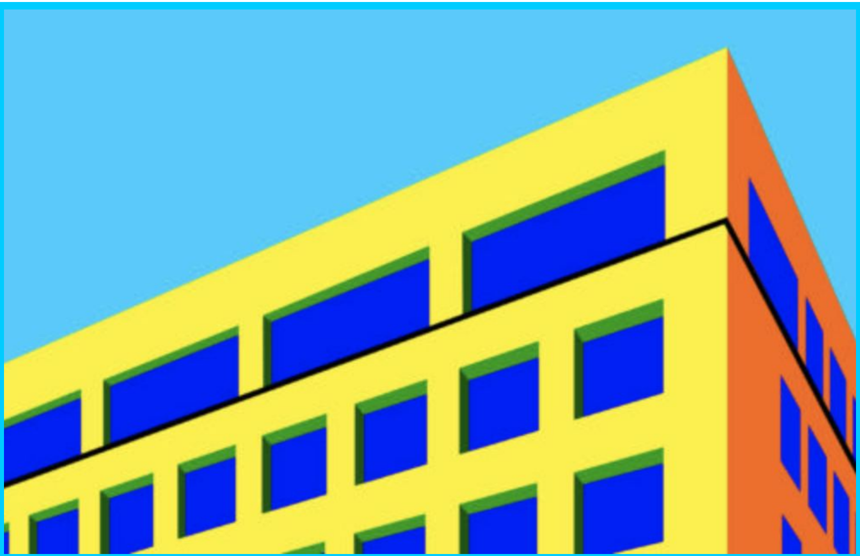
A/B Test

Causal Identification based on controlled random assignment of treatments

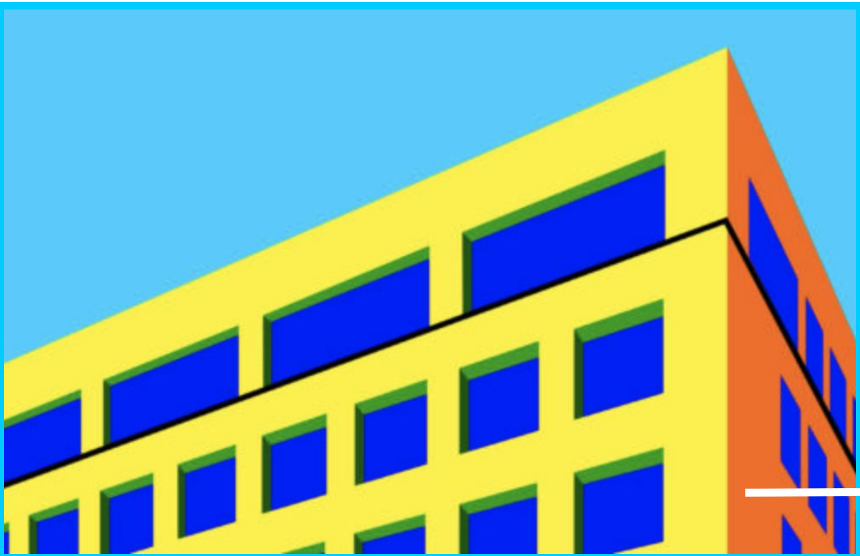
Quasi-experimental

Causal Identification based on statistical adjustment

Two Floors

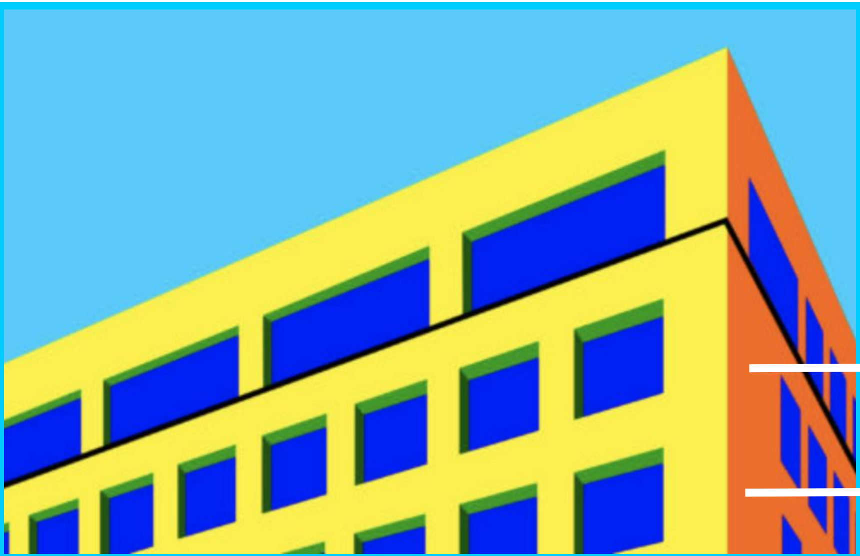


Two Floors



BM1 4th Floor
Octopus Experimentation Team
“We love **causal** inference with experimental methods

Two Floors



BM1 5th Floor
Economics Team
"We love **causal** inference with quasi-experimental methods"

BM1 4th Floor
Octopus Experimentation Team
"We love **causal** inference with experimental methods

Let's join up

Economics
"We love causal inference"

&

Experimentation
"We love causal inference"

Mission

Mission

We empower teams to measure their impact based on evidence they can trust.

Objectives

1. Trusted measurement

2. At Scale

Mechanisms

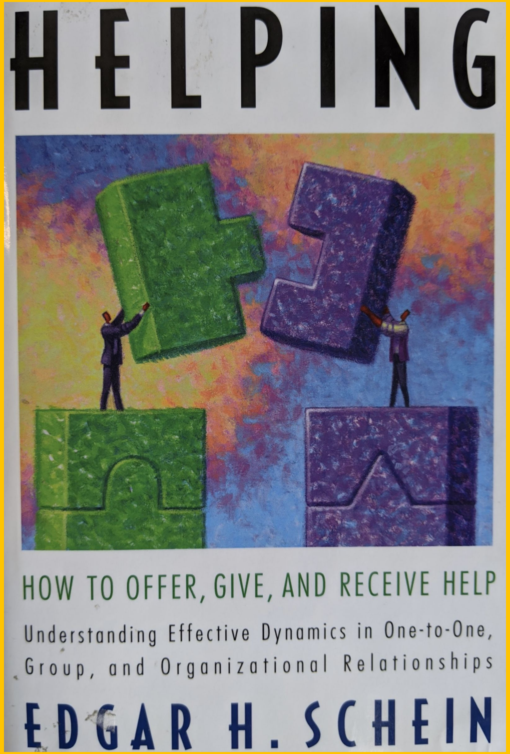
Scaling Trusted Measurement

Trusted Measurement

Custom Research

Helping

In the Centres of Excellence, a lot of our role is **consulting and coaching** others so they can be better researchers



Helping and Domain Expertise

Domain expertise is critical for causal modeling.

Our stakeholders have this expertise.

Need to **help** them access and use this knowledge to develop causal understanding

Scaling Trusted Measurement

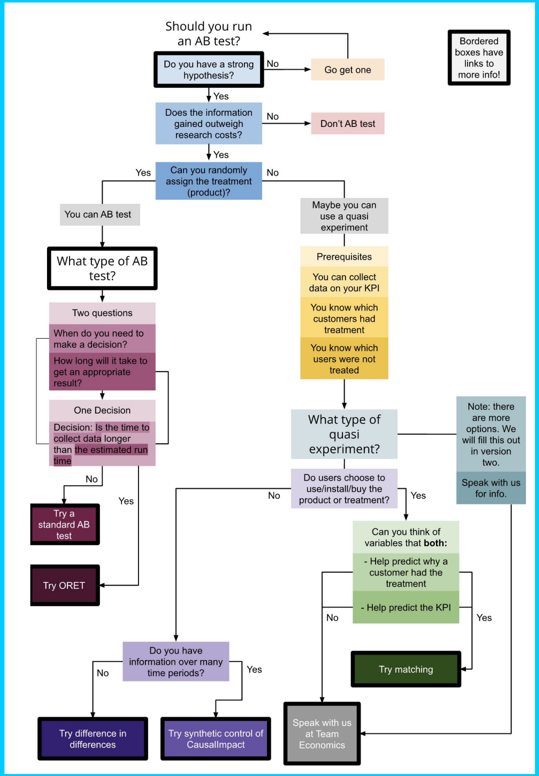
Trusted Measurement

Custom Research

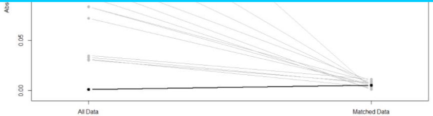
Semi-Automation (software, education, review)

At Scale

Right Tool Initiative [link]



Causal Inference Peer Review [link]



In some matching models, there were a few covariates of which the standardized difference of means increased a bit after matching (indicated by the black lines). It normally happens to covariates that have small differences before matching as they do not factor heavily into the propensity score model. For those well balanced covariates, the standardized difference of means remained less than 0.1. Therefore, we are not too concerned about it.

0.23.3 Effect estimation

After the matching process, we conduct statistical analysis as if the datasets had been generated through randomization (Choirat et al. (2015)). The matched treatment comes from the Plus paying members in the treatment group. ATT (Average Treatment Effects on the Treated) is relevant in this case, which is the expected causal effect of the treatment for individuals in the treatment group, instead of ATE (Average Treatment Effect).

$$E[\delta \mid D=1] = E[Y^1 - Y^0 \mid D=1] = E[Y^1 \mid D=1] - E[Y$$

btanz commented 15 days ago • edited ▾ Member

- [ ] in progress
- [x] done

Overall feedback

I absolutely enjoyed reading the paper and providing a review. My main reasons are the following:

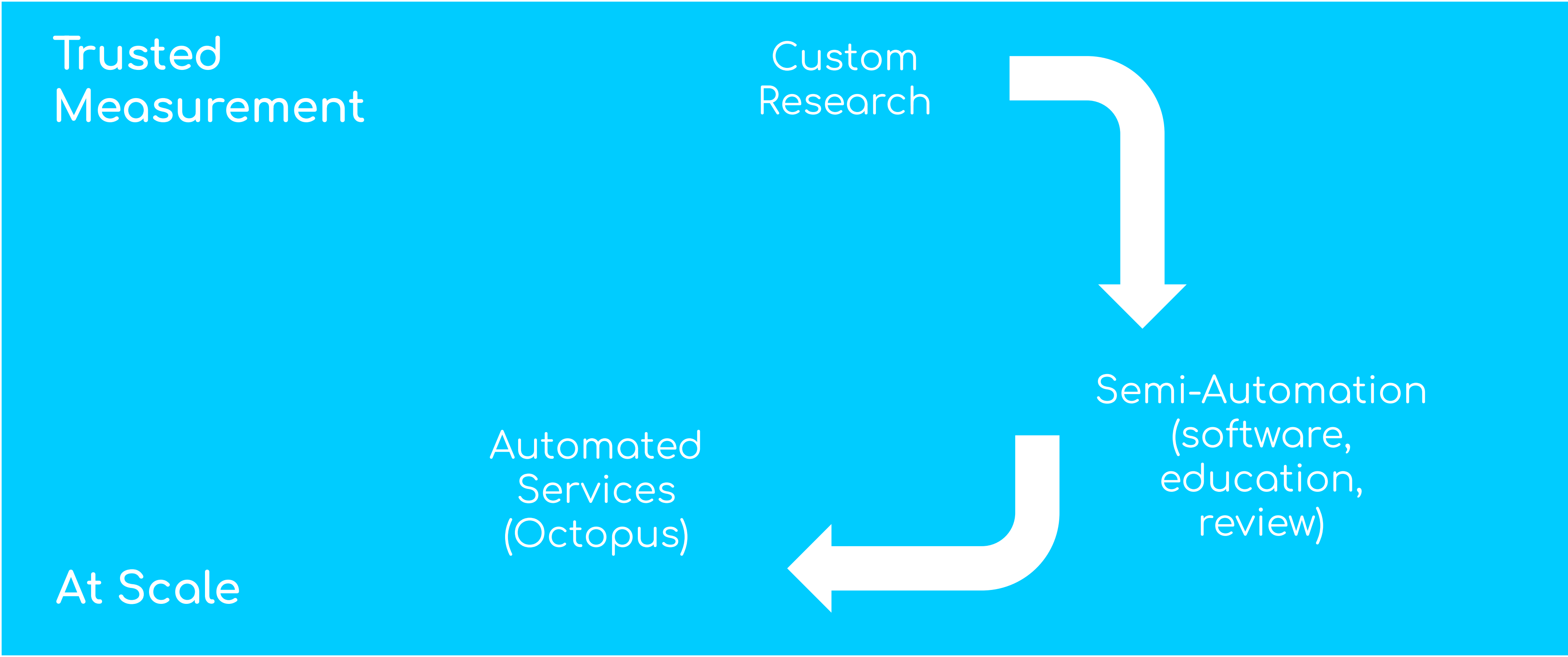
- The paper deals with a problem of first order business relevance to DX, which is understanding the impact of interventions on our North Star KPI CLV from transactions, advertising and subscriptions.
- In terms of research approach and essentially, the paper clearly distinguishes modelling, causal identification and statistical estimation, which is a very good practice to create clarity on logically distinct problems. It helped me understand better assumptions and goals of the work.
- A solution the paper advocates for - using a surrogate index to predict long term outcomes - is in my view indeed very promising for solving the problem the paper poses.

Note: You can find detailed feedback in this PR. The main points are summarised below, following the review template.

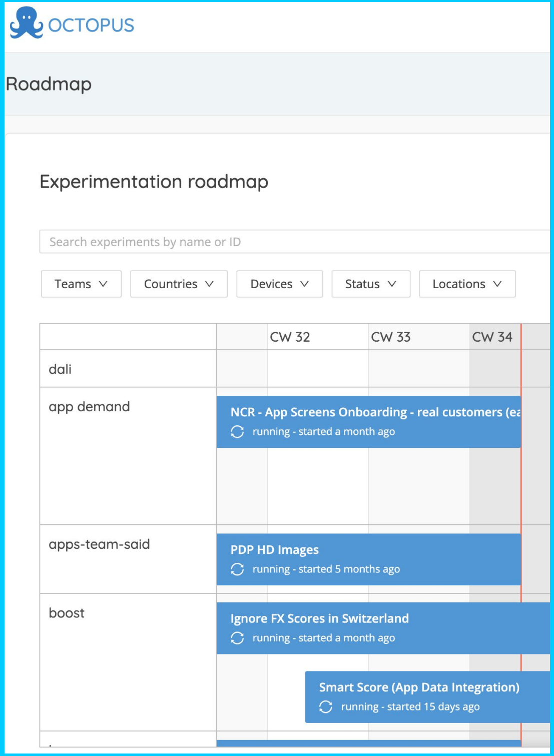
Research purpose

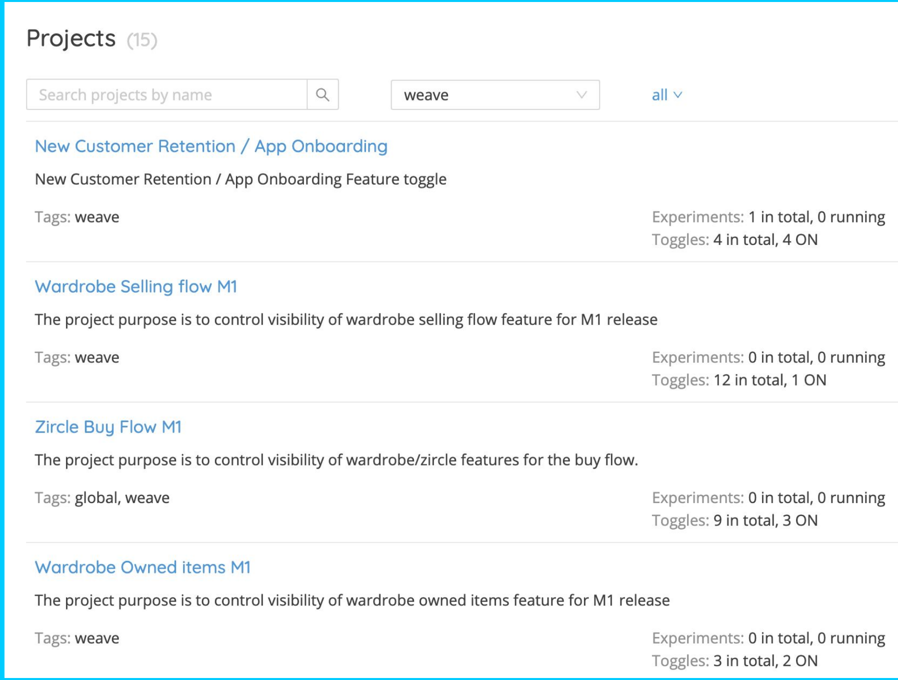
Problem statement

Scaling Trusted Measurement

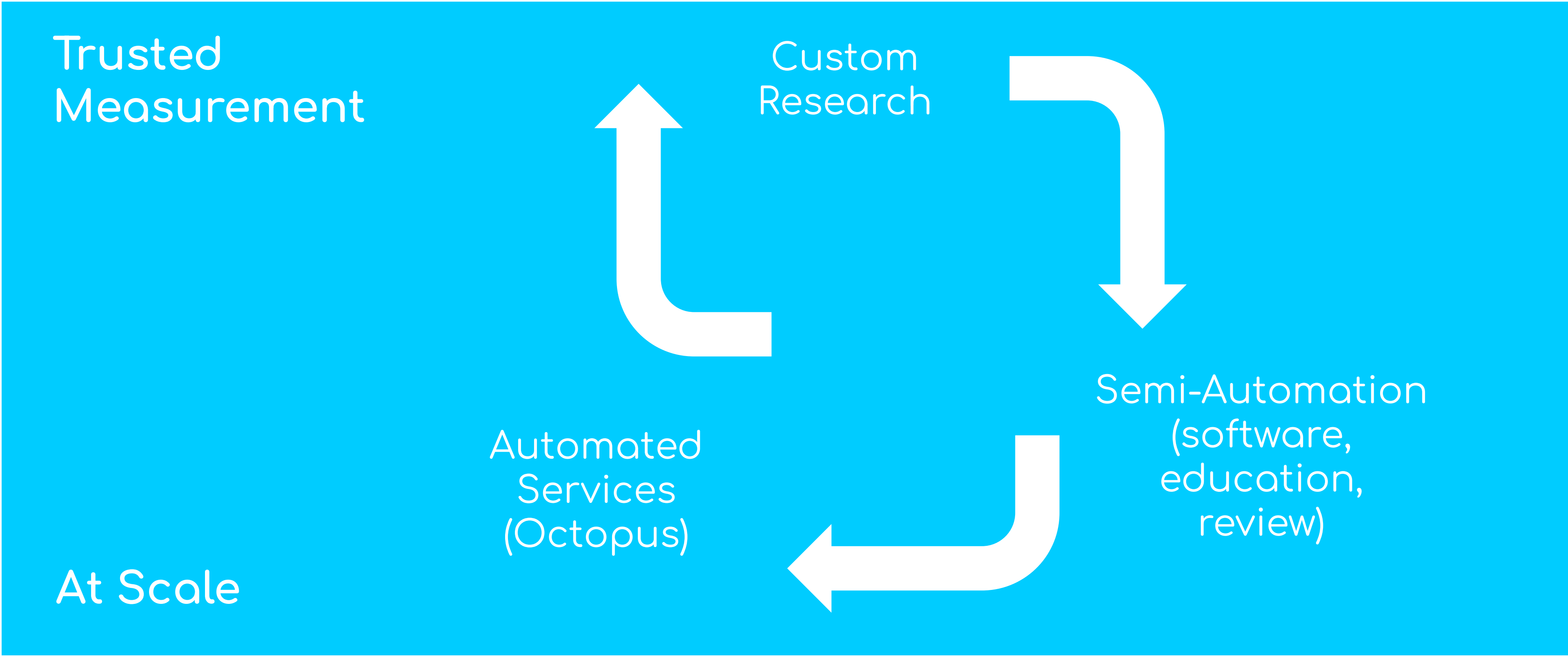


Software & Services





Scaling Trusted Measurement



It’s 2020+

Hot Topics

Aligning Algorithms

Vision: Aligning Algorithms with the Zalando Group Strategy

Problem to Solve

The Zalando customer experience is increasingly created by automated decision-making systems. These systems optimise for different and potentially contradictory goals that in many cases are also not aligned with the Zalando Group Strategy.

The Aligning Algorithms Project aims to solve this problem by providing automated decision-making systems with the customer experience goals to optimise for that best advance the Zalando Group Strategy.

Strategy for Estimating Impact of Customer Events on CLV

Continuously Improving Downstream Impact Estimation

Proposition: optimising for the right thing badly is better than optimising for lots of different wrong things well.

How might we manage this process to move as quickly as possible from quick and dirty estimates of the downstream impact of what we can measure with minimal assumptions/research to a full featured and automatically updating catalog of events' downstream impact

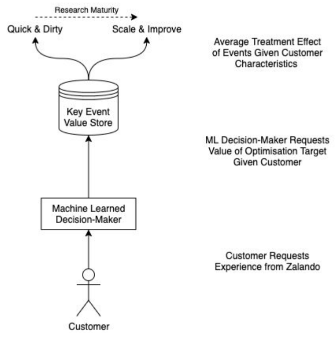
Get Started Quick and Dirty

Ex: Stability Controlled Quas-Experiment (Hazlett 2019). Allows plausible average treatment effect on treated (ATT) with ~4 easily queried numbers. Manual work is required to define "plausible" baseline trend. Limited estimation heterogeneity.

Scale and Improve

Ex: Statistical Surrogacy (Athey et als. 2020 and Eckles et al. 2020). Utilises data generated by known process--short-term A/B tests--to estimate proxy events' long-run impact. Could be automatically updated over time with more experiments (on even "unrelated" products).

In Production for Customers



Causal Machine Learning

Causal ID <> ML

If we can reformulate $P(y|x', y')$

as a set of prediction problems $P(y|x)$

we can take full advantage of machine/deep learning methods.

Ex 1. Controlled Rollouts

For Weave, we gradually expose products to bigger audiences. But no opportunity to A/B test.

Ex 1. Controlled Rollouts

Need some way to measure uplift with historical and cross-market data.

Estimating Weave Uplift

Causal Identification

$$ATT = E[KPI^{W=1}_{t \geq CW35} - KPI^{W=0}_{t \geq CW35} | W = 1]$$

Effect Estimation

$$E[KPI_{t \geq CW35} | W = 0, \{KPIj\}^{CW34}_{j=0}]$$

"I just ran hundreds of prediction models" -- Patrick Doupe

[Paper]

Ex 2. Surrogates of Long-run lift

Causal Identification

$$\text{Estimated lift} = \hat{f}\,(KPI_1) - \hat{f}\,(KPI_2)$$

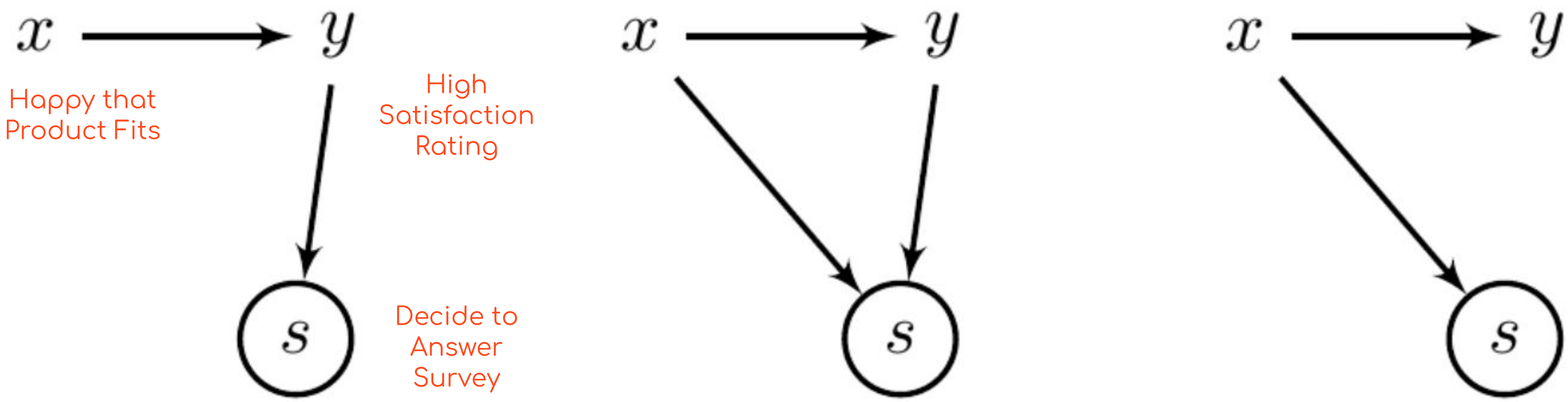
Effect Estimation

$$h_O(s, x) = \mathbf{E}_O[Y_{O,i} \mid S_{O,i} = s, X_{O,i} = x]$$

Paper

Sources of bias

Self-Selection Bias

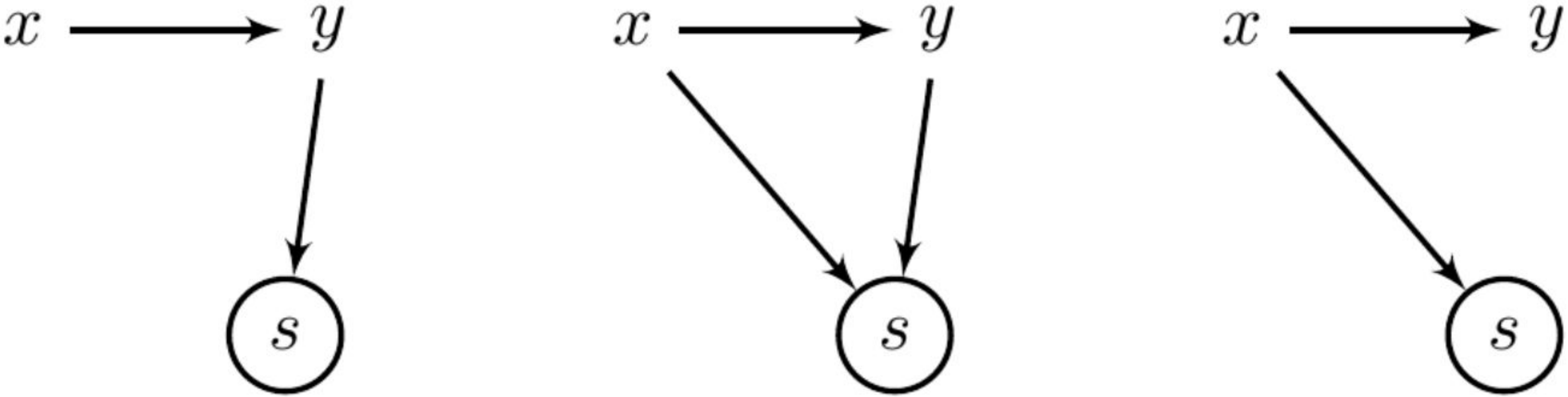


Limitations (see [Bareinboim, Tian, and Pearl] 2014, 2411–2112)

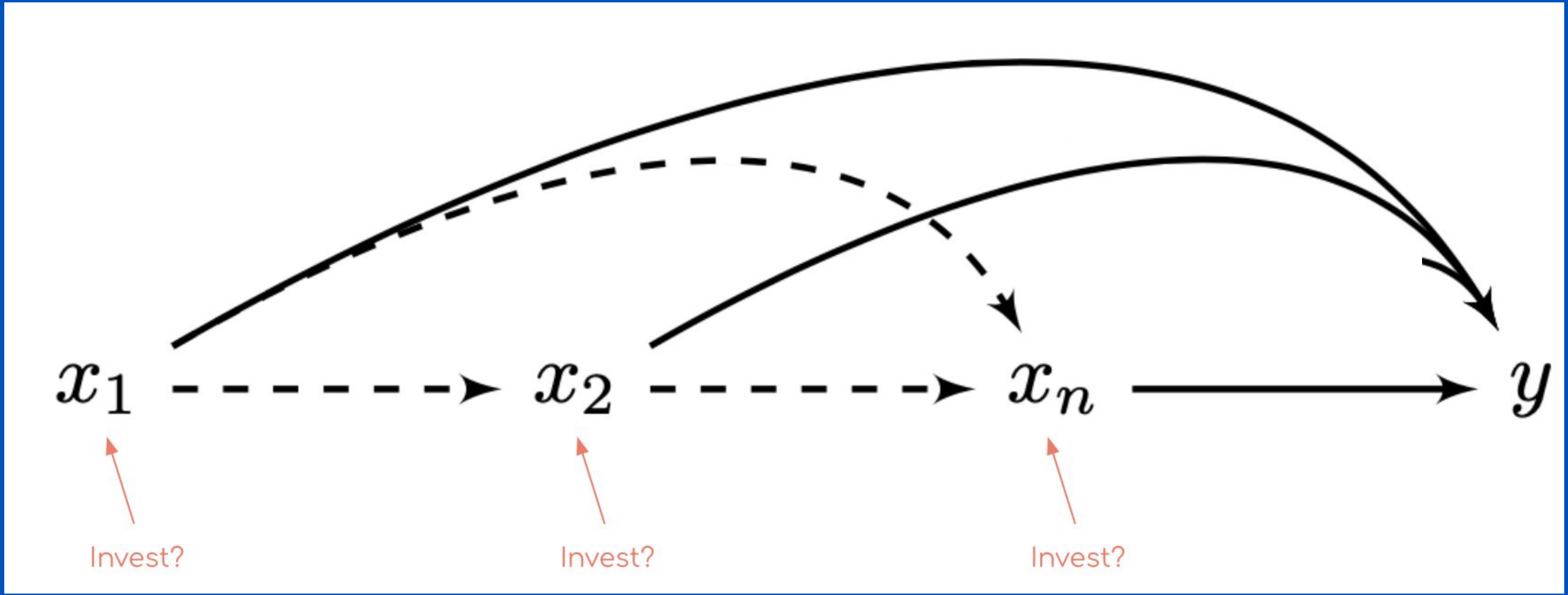
Effect unknowable

Effect unknowable

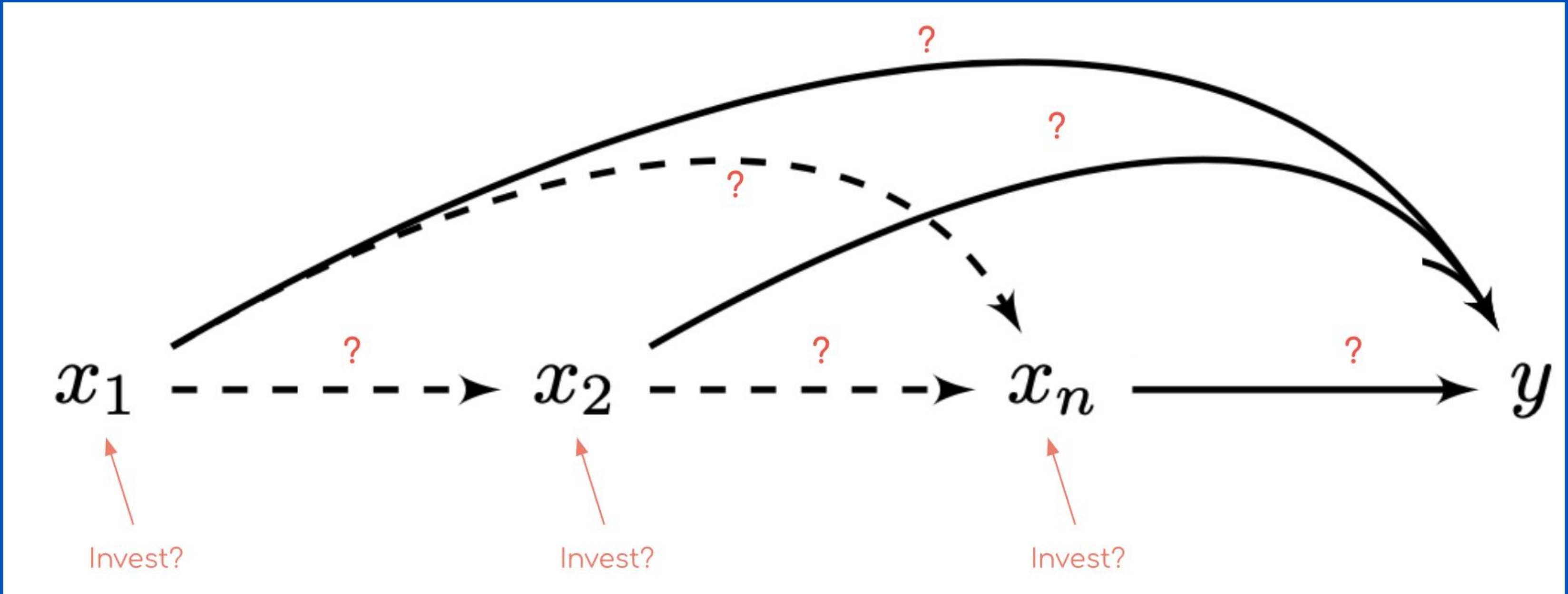
Effect knowable

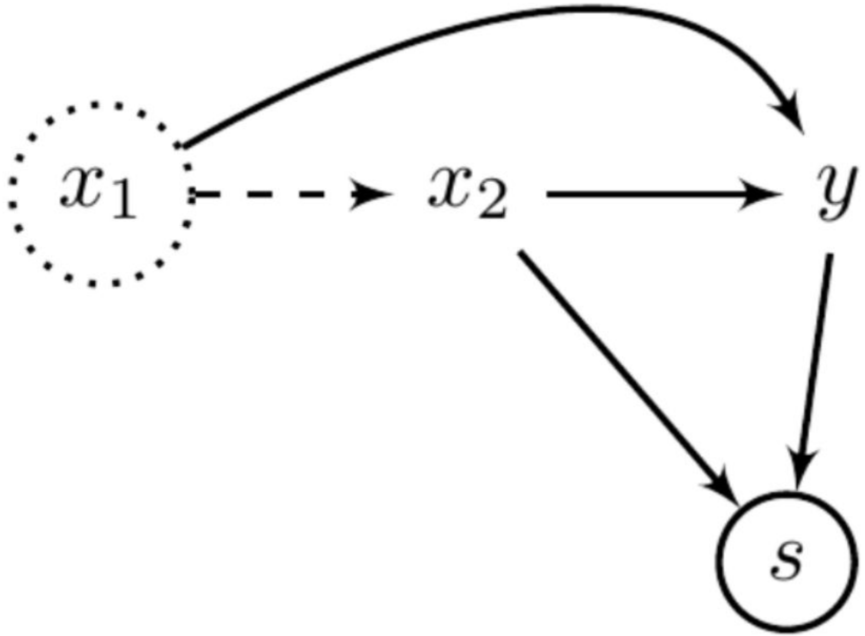


Attribution Modelling



Attribution Modelling





X1 is not observed, even after the customer registers/logs in.

Understand and Mitigate

Our objective is to understand this selection bias and find mitigation strategies (both statistical and product)

SUTVA

Stable Unit Treatment Value Assumption (SUTVA)

Most causal models assume no interference between individuals under treatment.

If there is interference, we get biased causal estimates.

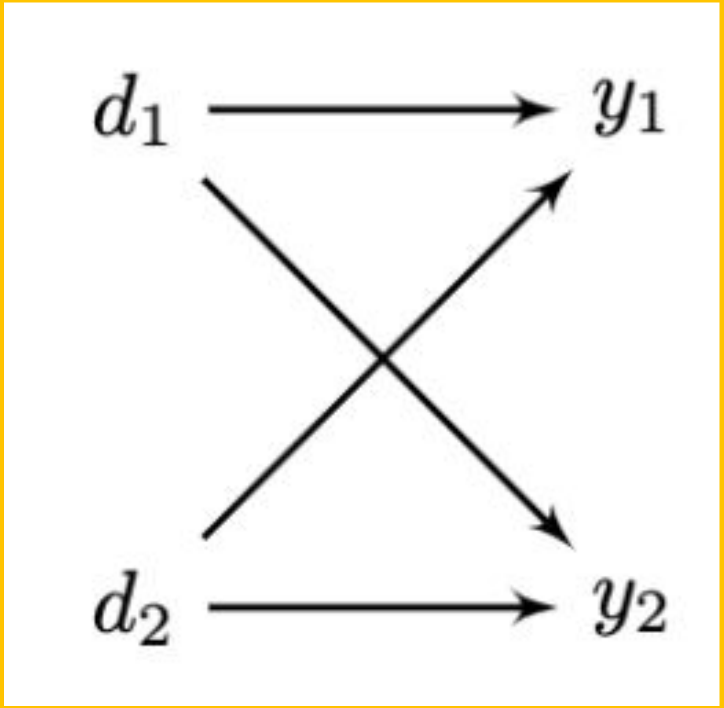
SUTVA Violations

Interference
Networks/Spillovers

Treatment (*d*) impacts outcome for treated unit and others.

Think: social networks, markets

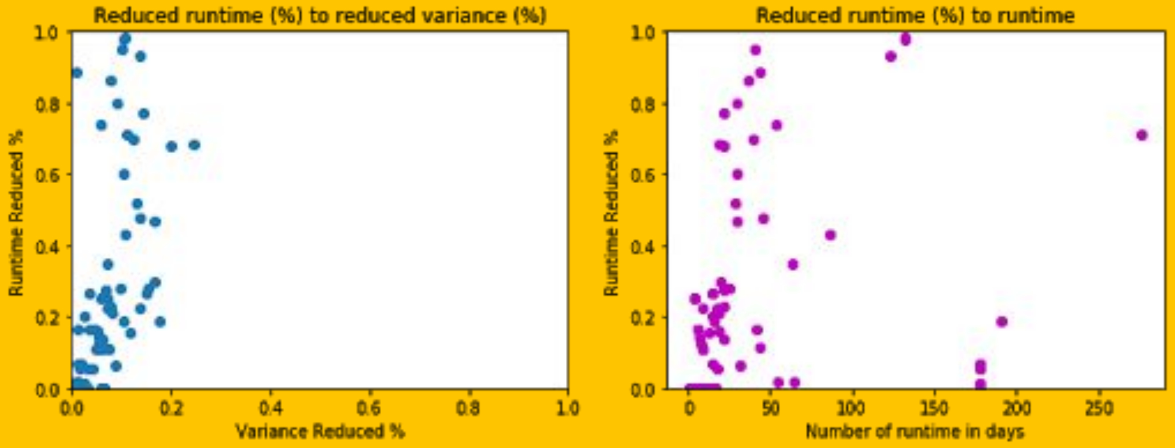
Requires new randomisation schemes and estimators



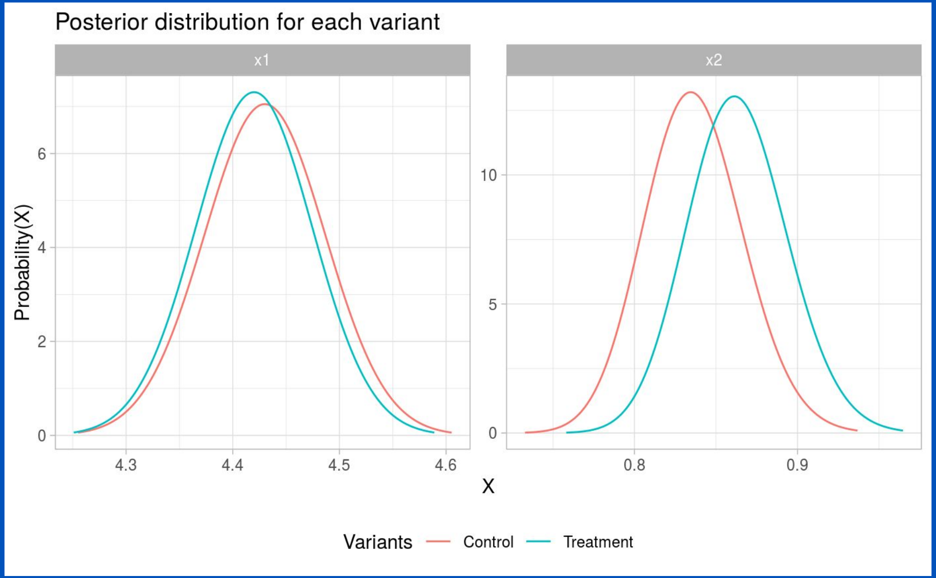
Information efficient inference

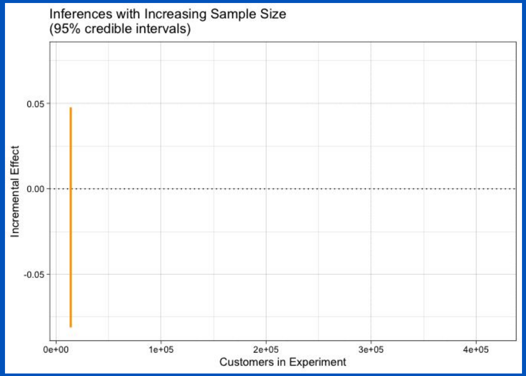
Learning is costly

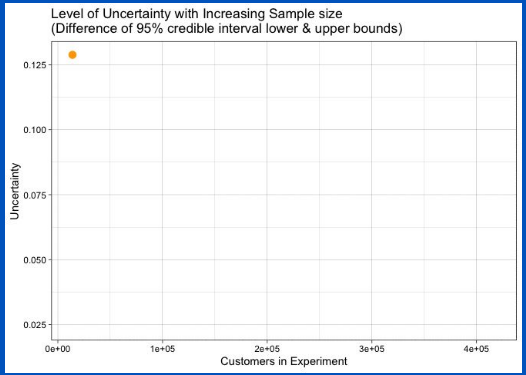
Variance Reduction



Rapid Experiments







Documentation

Adaptive Experiments

Bandits are an **information efficient** way to learn "best" treatment.

But they use **biased assignment, biasing** incremental effect estimates.

→

Adaptive Experimentation (e.g. bandits with causal estimators)

Q & A