

An Industry Perspective on A/B Testing

Huizhi (Kenny) Xie
2024.03.09

CONTENTS

01 Why A/B Testing?

02 What is A/B Testing?

03 Top Practical Challenges

04 Takeaway

/01 Why A/B Testing?



Let's Start With The Growth Logic Of An Internet Company.

- An internet company builds product/service for its users.
- The product/service has its own value proposition.
- Users use the product/service and provide feedback.
- The internet company improves its product/service based on users' feedback for continuous growth.

User's Feedback Is Important Because It Is Not Easy To Come Up With Ideas That Work.

Microsoft

At Microsoft, **1/3** of the ideas improved business metrics, **1/3** had no impact on business metrics, **1/3** degraded business metrics.

- Kohavi 2009

Google

In well-optimized search engine, only **10%-20%** of ideas improved business metrics.

- Jim Manzi 2012

Slack

70% of our ideas were abandoned.

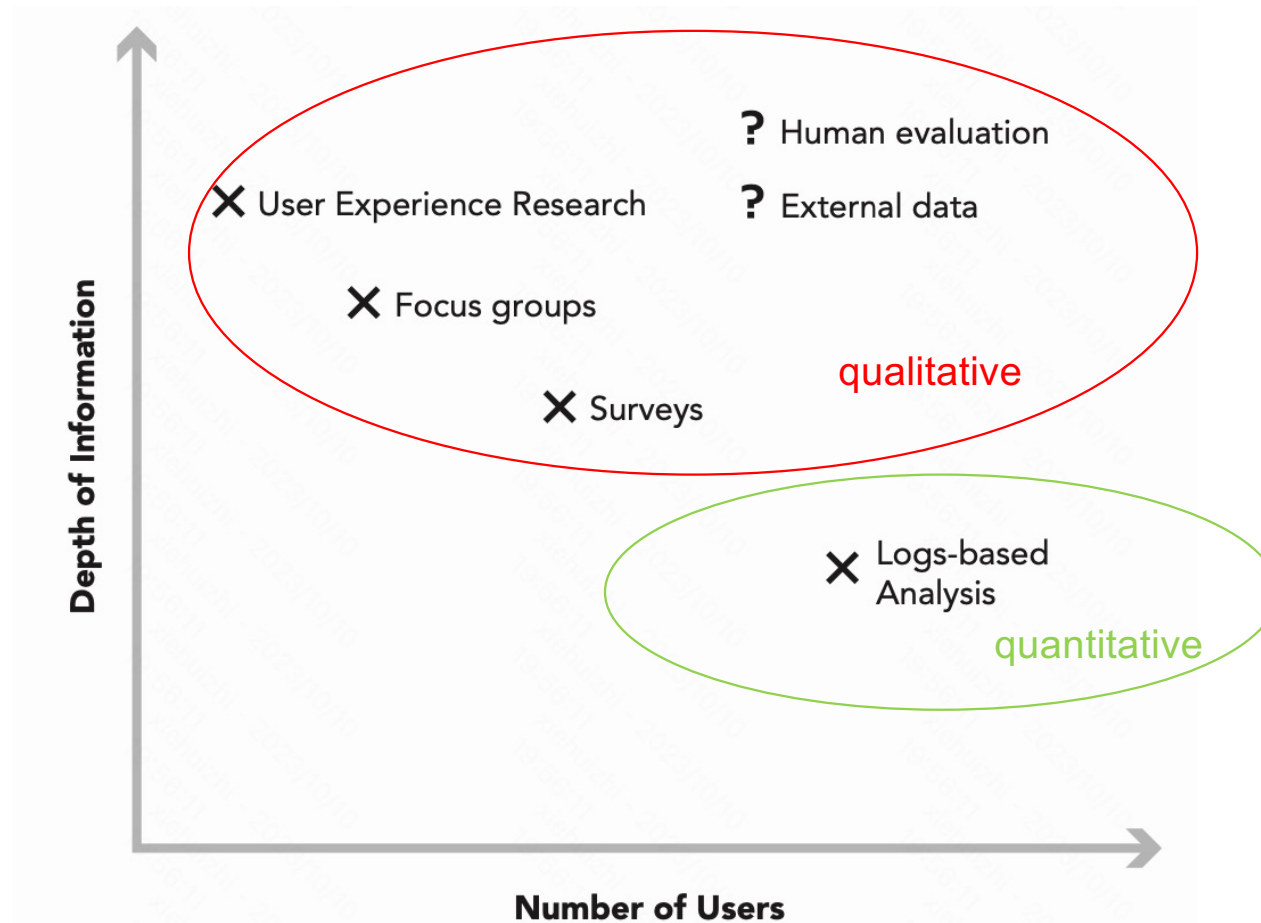
- Mosavat 2019

Netflix

90% of our ideas were wrong.

- Mike Moran 2007

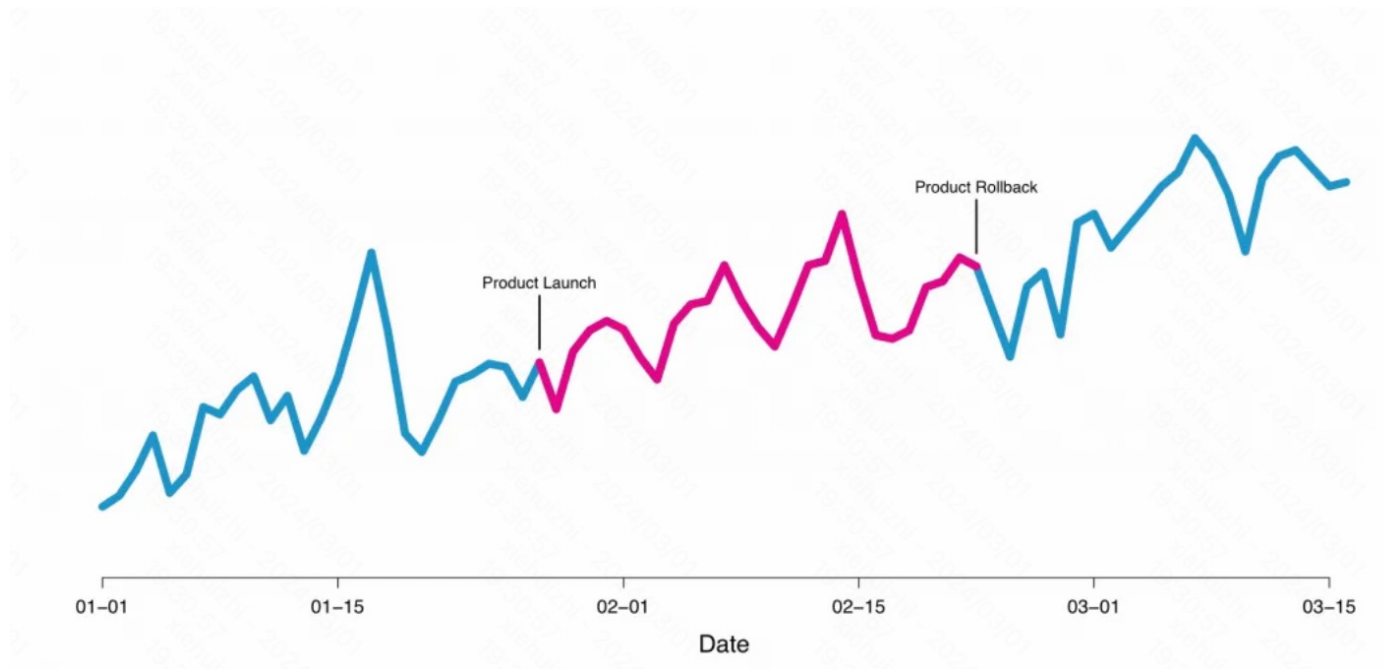
User's Feedback Can Be Qualitative Or Quantitative.



Data-driven Product Development Based On Quantitative Feedback Is The Buzzword These Days.

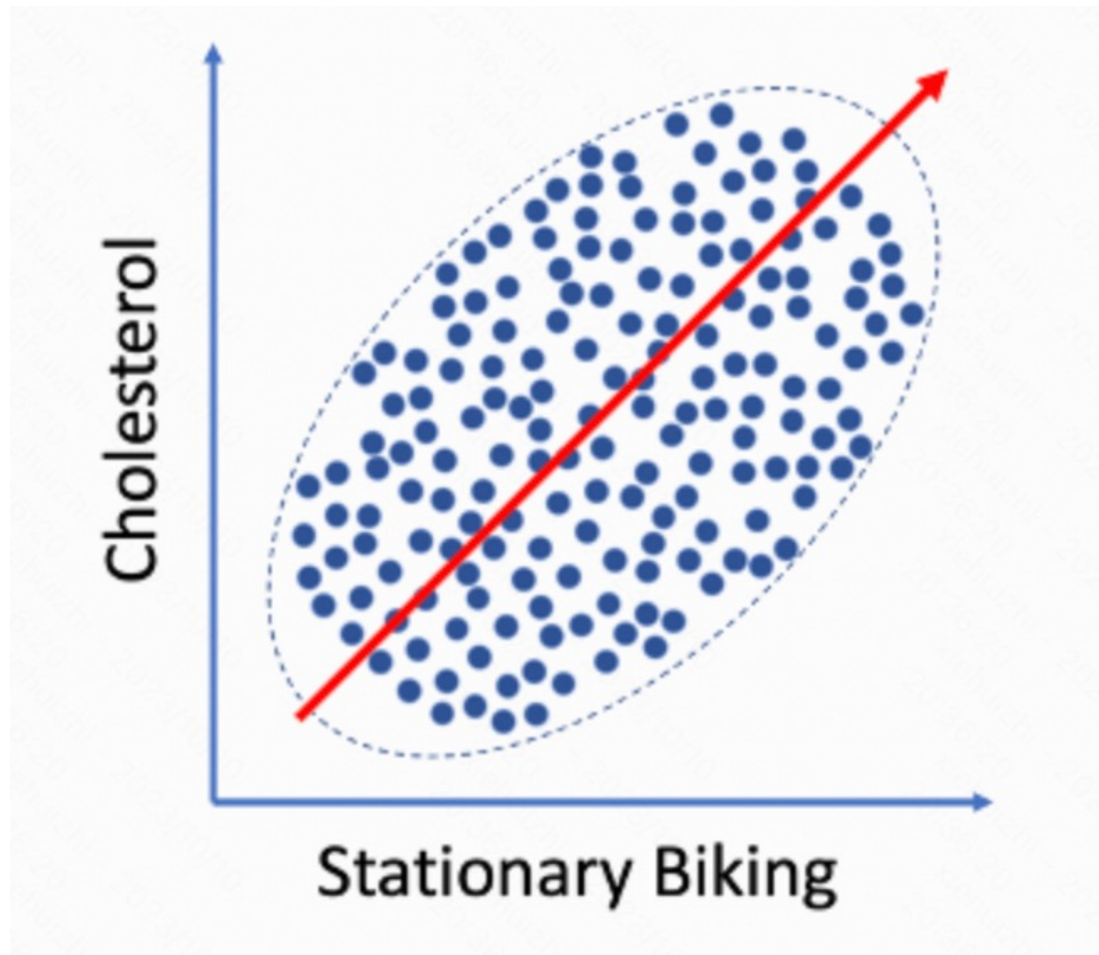
- Data-driven product development is more efficient.
- For data-driven product development to work, there are two key elements:
 - Metric: measure value delivered to users
 - Causal inference: link what a company does to improvement in the metric, or incremental value delivered to users

Causal Inference Is Not Easy: An Airbnb Example

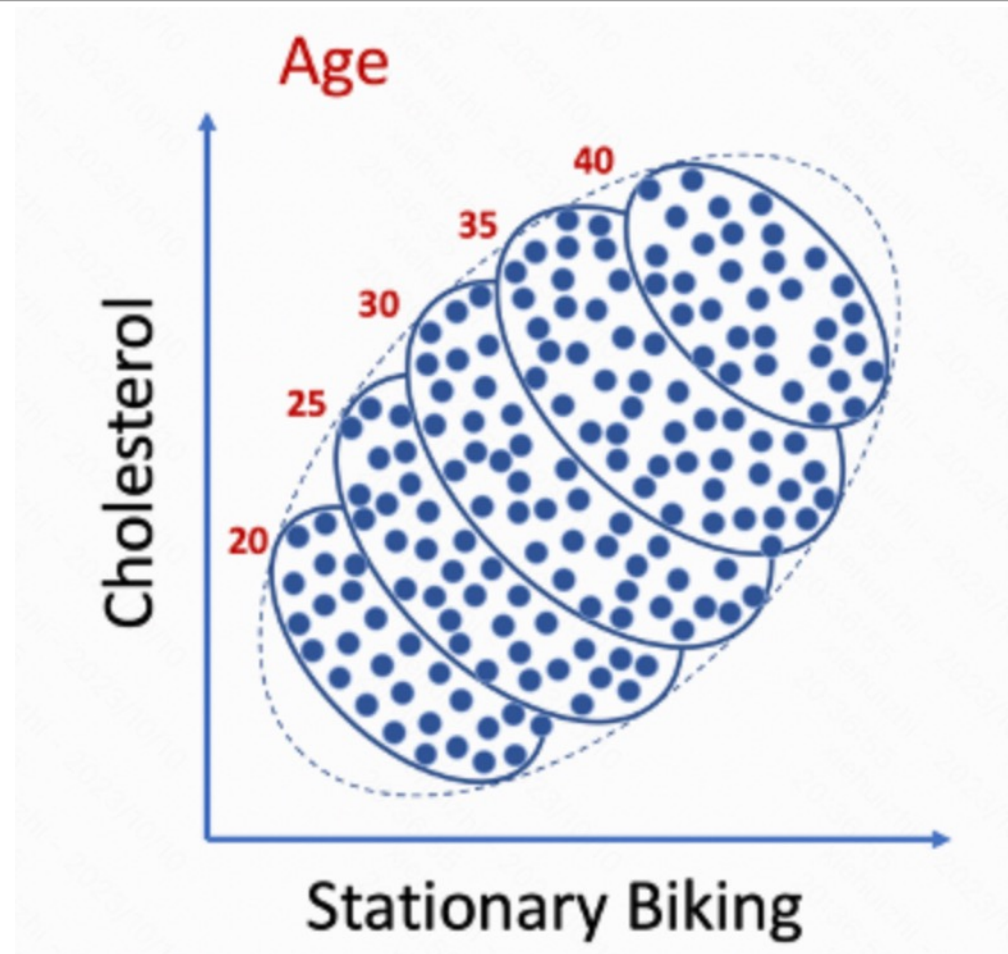


<https://medium.com/airbnb-engineering/experiments-at-airbnb-e2db3abf39e7#.miqyczkzb>

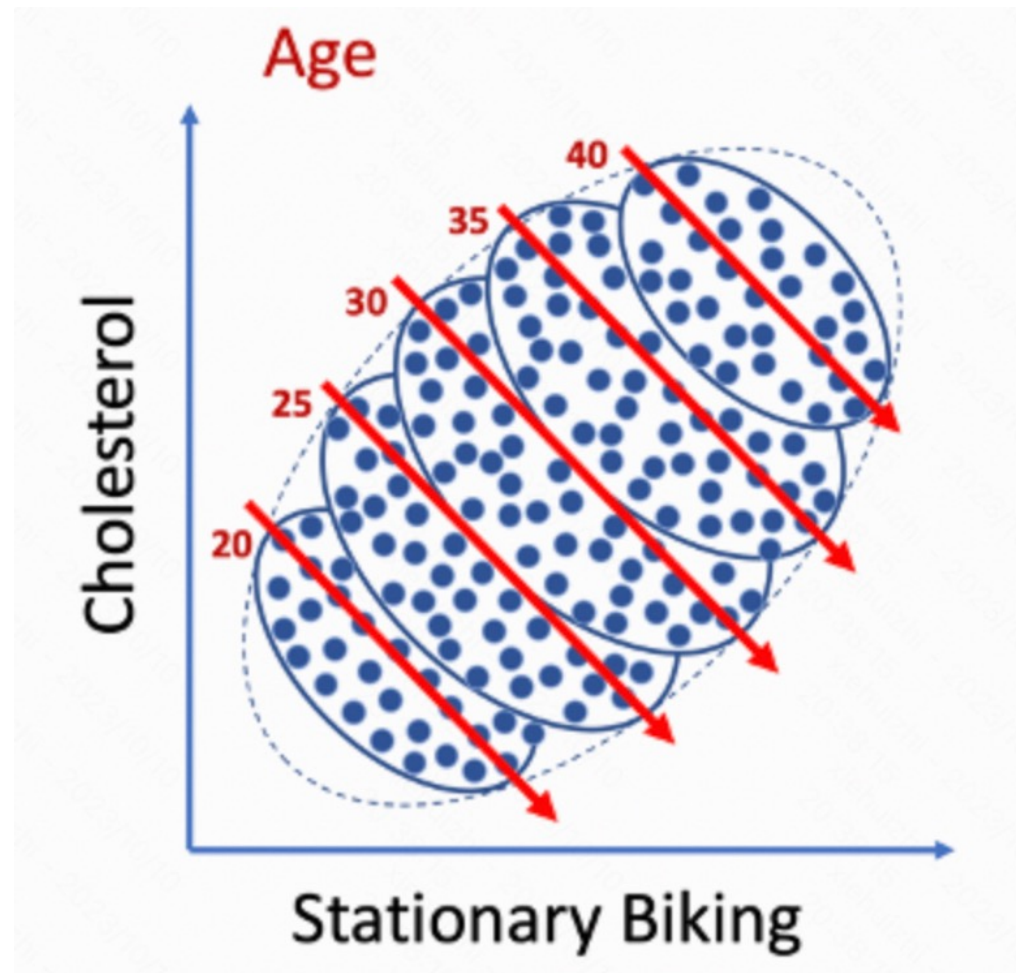
Causal Inference Is Not Easy: A Real Life Example



Causal Inference Is Not Easy: A Real Life Example



Causal Inference Is Not Easy: A Real Life Example



A/B Testing Is The Gold Standard For Causal Inference.

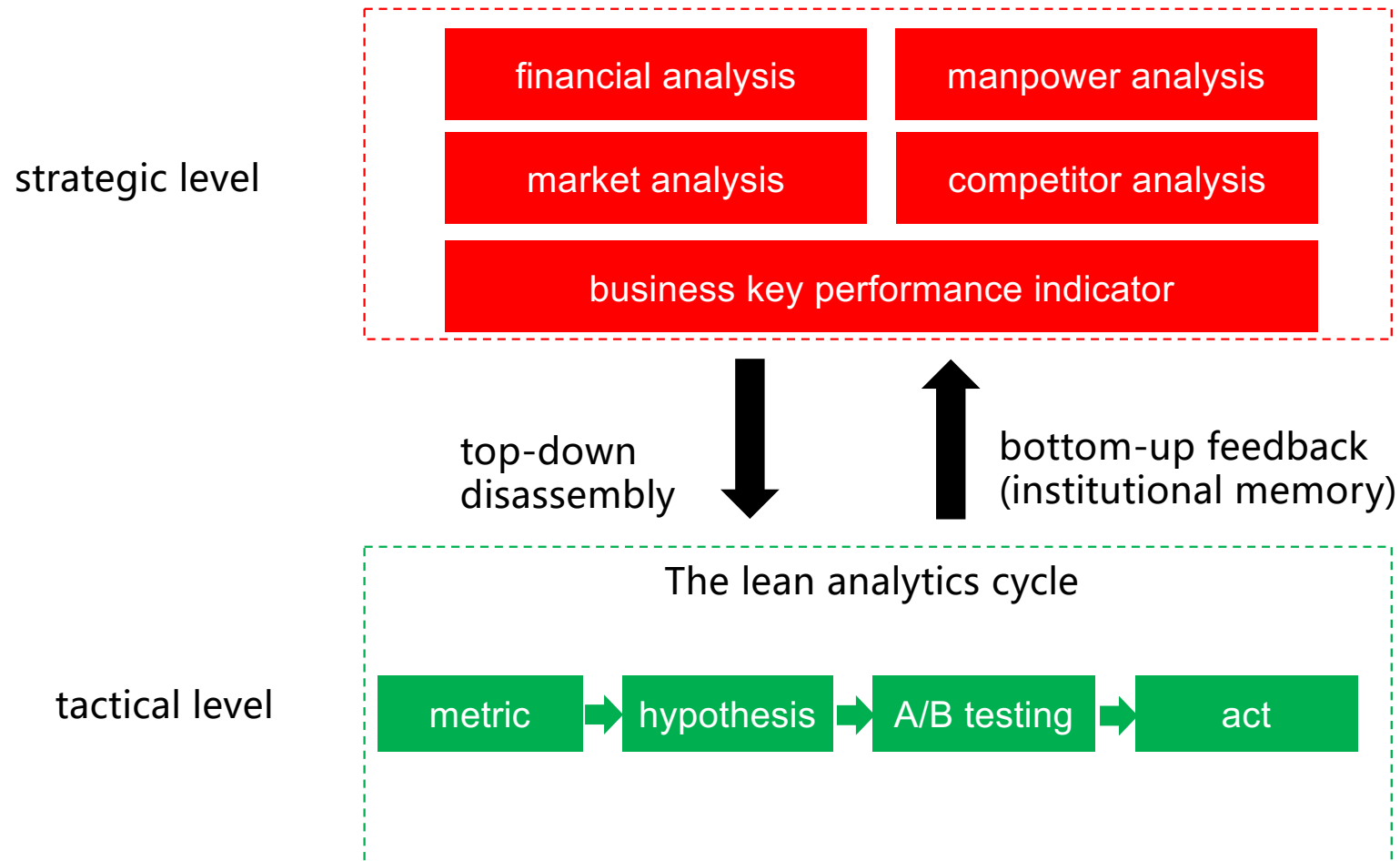


An A/B Testing Pyramid



Courtesy of Brooks Bell

A High-level View Of An Internet Company's Data-driven Operation



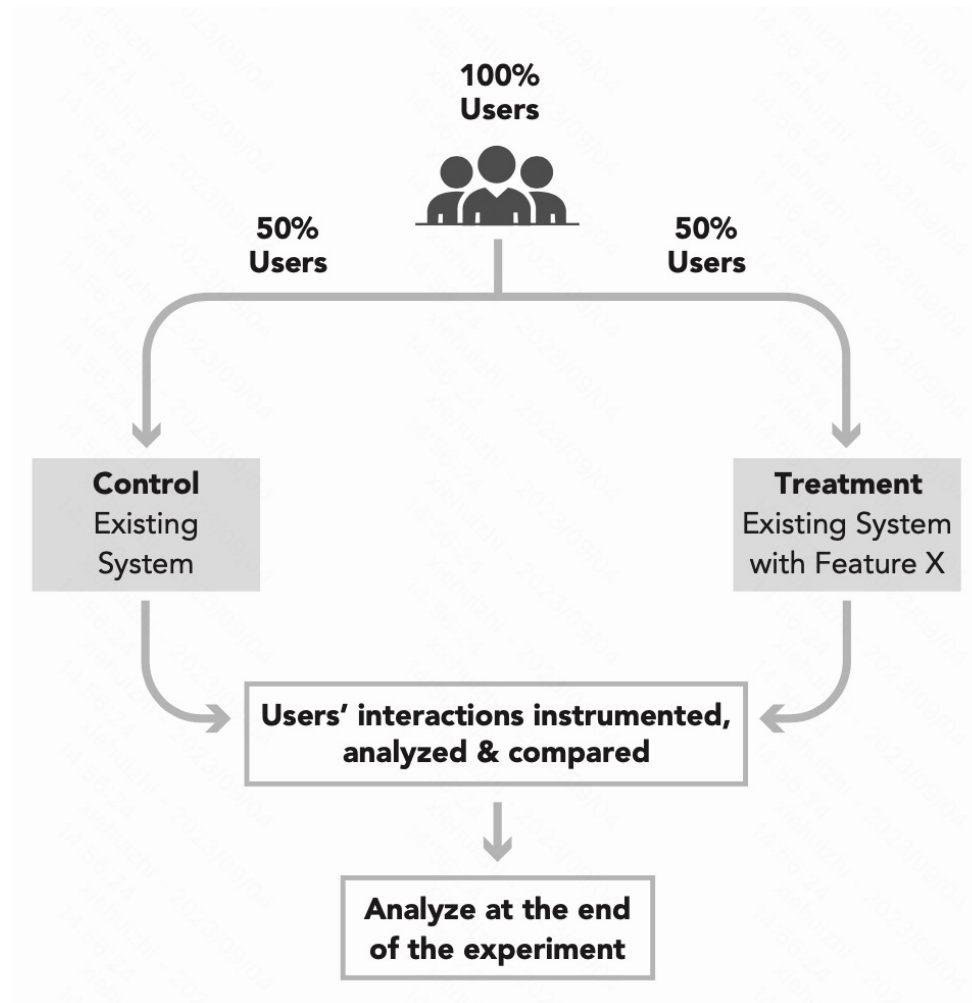
Internet Company Tests Almost Everything Online.

- Relevance algorithms (search, ads, personalization, recommendation, and so on)
- User interface (UI)
- Latency/performance
- Content management systems
- Customer support systems
- Backend code changes
- ...

/02 What is A/B Testing?



A/B Testing In One Slide (Kohavi, Tang and Xu 2020)



Online Experiment Is A Unique Type of Experiment: Same Science, Different Application

- Large traffic volume
- Low cost of data collection
- Lots of evaluation criteria
- Relatively small treatment effect
- Concurrent experiments on the same user set
- ...

Statistical Inference For A/B Testing In One Slide

A pair of hypotheses:

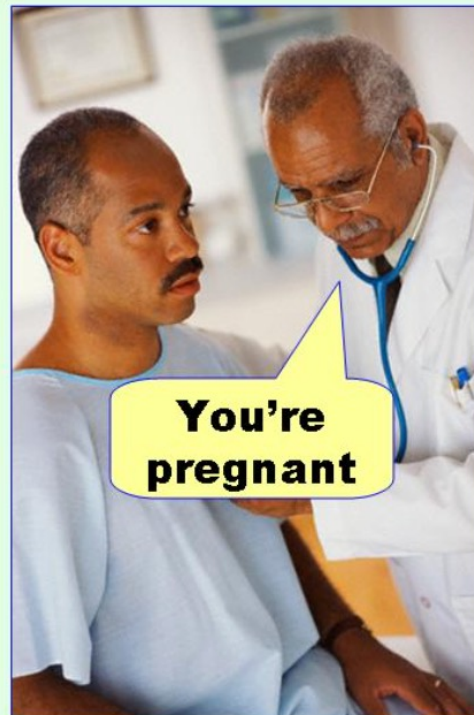
$$H_0 : \mu_A = \mu_B \quad \text{VS} \quad H_1 : \mu_A \neq \mu_B$$

Two types of error:

	H0 is true	H1 is true
reject H0	Type I error	correct decision
reject H1	correct decision	Type II error

Understanding The Two Types of Error: Analogy To Pregnancy Test

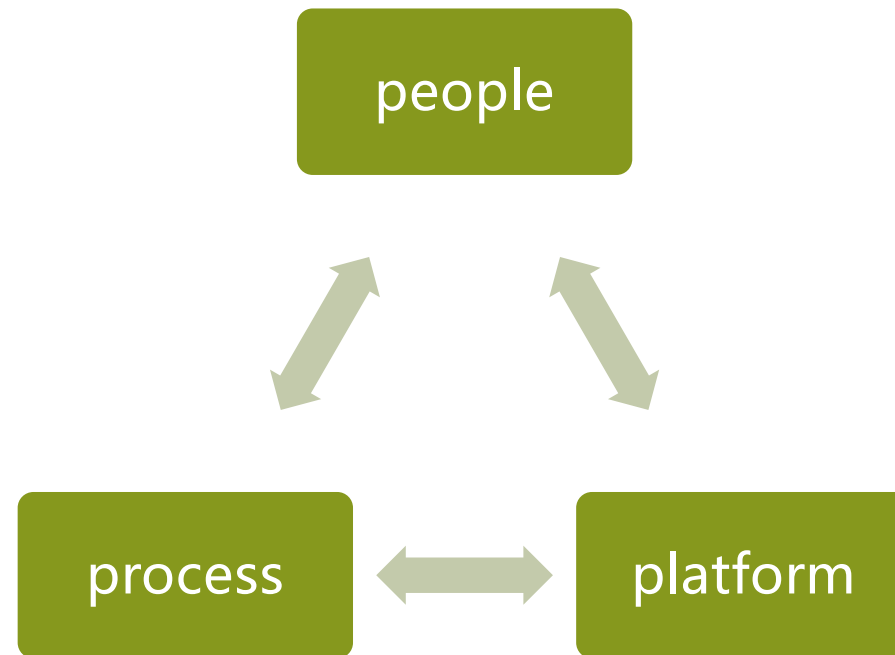
Type I error
(false positive)



Type II error
(false negative)

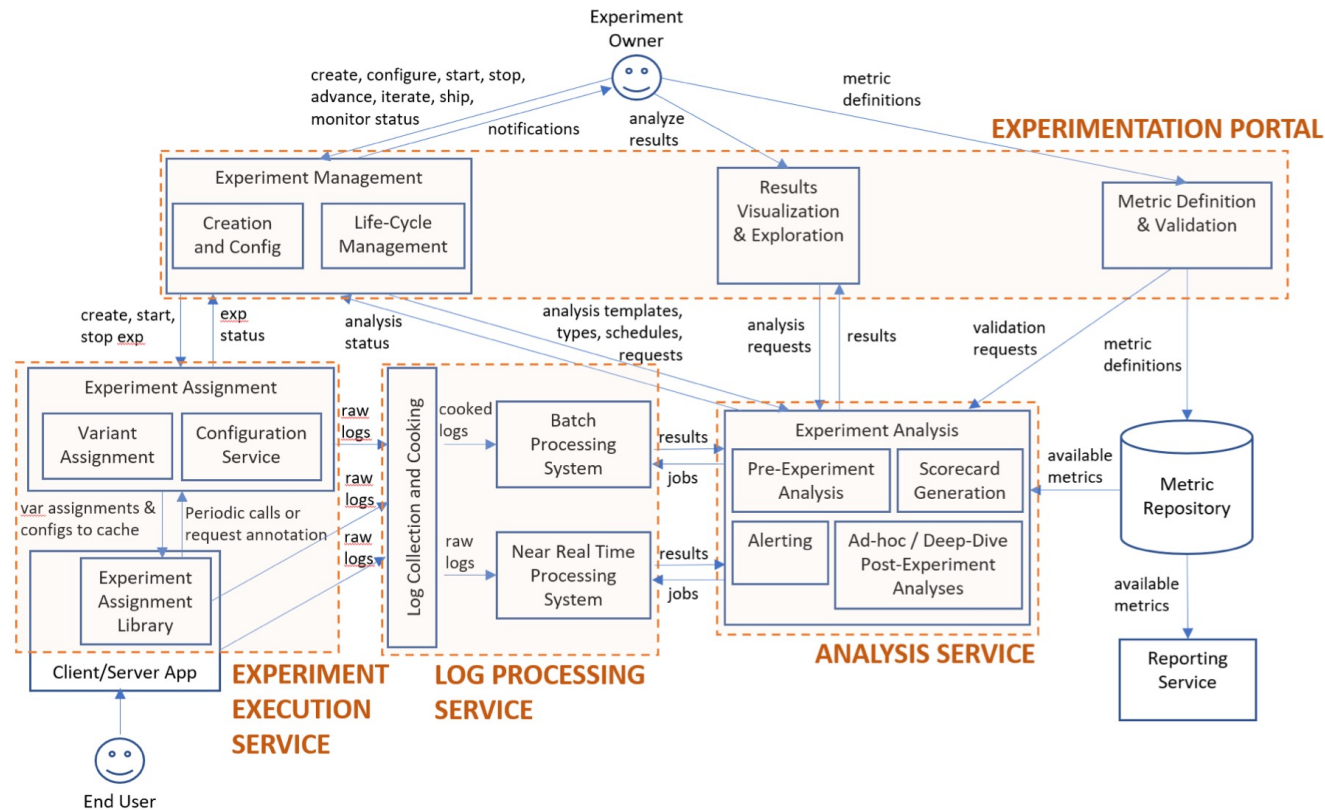


Internet Company's A/B Testing Ecosystem



<https://engineering.linkedin.com/ab-testing/why-experimentation-so-important-linkedin>

Architecture Of A Large-scale Online Experiment Platform (Gupta et al. 2018)



/03 Top Practical Challenges

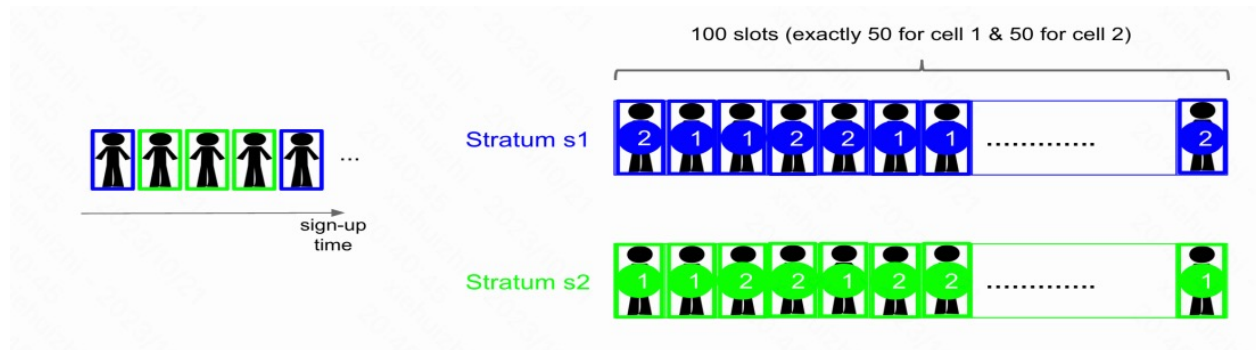


Improving The Sensitivity Of Experiments: Problem

- Recall the two types of errors in frequentist statistical inference
- Two types of error conflict with each other. With type I error controlled at a nominal value, sample size, true delta, and sampling variance of a metric affect type II error.
- Reducing type II error is to improve the sensitivity of experiments.
- Sampling/randomization or analysis stage to reduce type II error?

Netflix's Real Time Stratified Sampling System (Xie & Aurisset 2016)

(a)	1	2	3	4	5	6	100
(b)	25	57	9	12	95	64	43
(c)	1	2	1	1	2	2	1

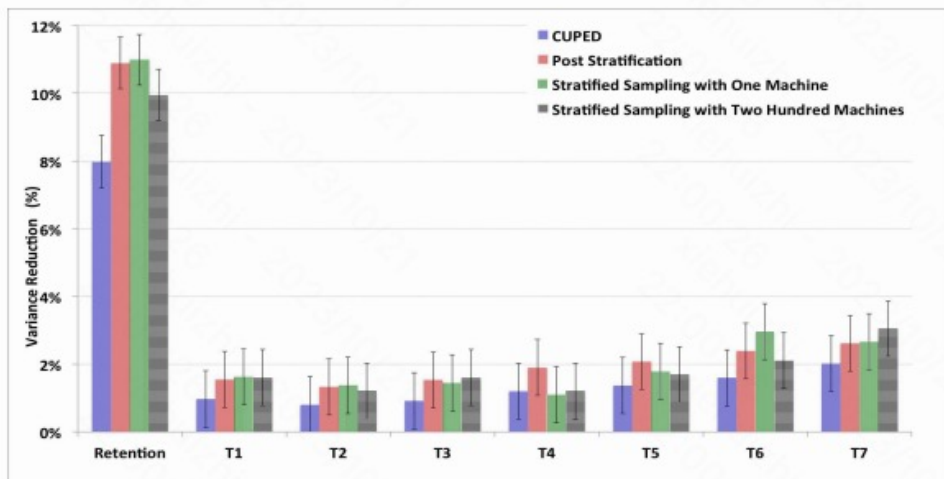


Stratified Sampling, Post Stratification And CUPED Are Asymptotically Equivalent. (Xie & Aurisset 2016)

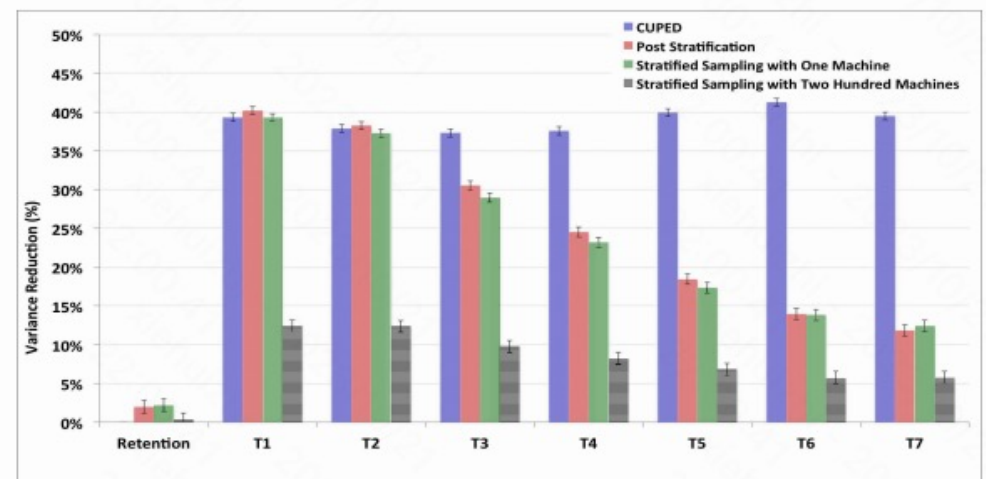
- \bar{Y} : simple average
- \widehat{Y}_{strat} : weighted average based on strata proportion
- $Var_{srs}(\bar{Y})$: variance of simple average under simple random sampling
- $Var_{strat}(\bar{Y})$: variance of simple average under stratified sampling
- $Var_{srs}(\widehat{Y}_{strat})$: variance of weighted average under simple random sampling
- $Var_{strat}(\widehat{Y}_{strat})$: variance of weighted average under stratified sampling
- n : sample size
- It can be shown:

$$var_{strat}(\hat{Y}_{strat}) = var_{srs}(\hat{Y}_{strat}) + O\left(\frac{1}{n^2}\right) = var_{srs}(\bar{Y}) + O\left(\frac{1}{n}\right),$$
$$var_{strat}(\hat{Y}_{strat}) \leq var_{srs}(\hat{Y}_{strat}) \leq var_{srs}(\bar{Y}).$$

Numerical Results Based On Real Experiment Data (Xie & Aurisset 2016)



new users



old users

Estimating The Long-term Effect: Problem

- The impact of a change may take a long time to materialize in terms of key product metrics.
 - Impact of change of ranking results in an online travel service materialize after customers stay in a vacation rental or hotel room months after booking
 - Increasing number of ads in search results bring in more revenue at first but might have the opposite effect due to user attrition
 - Introduction of clickbaits on a content provider service may cause increase in clicks due to novelty effect but may induce larger dissatisfaction in the long term

Estimating The Long-term Effect: Common Solutions and Challenges

- Long-term experiments or holdouts
- Proxies
- User learning modelling
- Surrogates

A Short Discussion On Long-term Experiments Or Holdouts

- Software development cycle is usually short.
- Engineering cost to maintain a code fork that is not updated for a long time
- Non-persistent user tracking and network interactions
- The above said, it is worth learning novelty effect or primacy effect.

A Short Discussion On Proxies

- Head company's approaches:
 - Netflix used logistic regression and survival analysis to find good predictors of user retention.
 - LinkedIn created metrics based on lifetime value model.
 - Uber found some macro-economic models to be useful in finding good proxies.
 - Bing and Google have found proxies for user satisfaction and retention by having a mental causal structure model that estimates the utility of an experience to users.

Model User Learning (Hohnhold, O'Brien and Tang 2015)

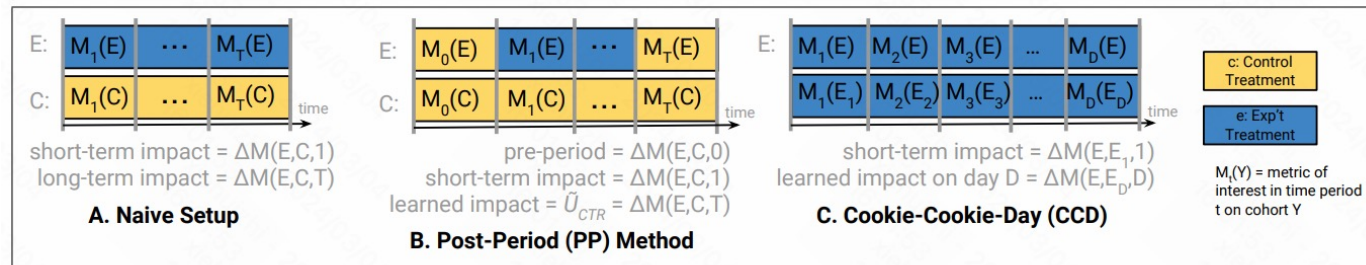


Figure 1: Graphical depiction of the Naive Setup, Post-Period, and Cookie-Cooke-Day methods.

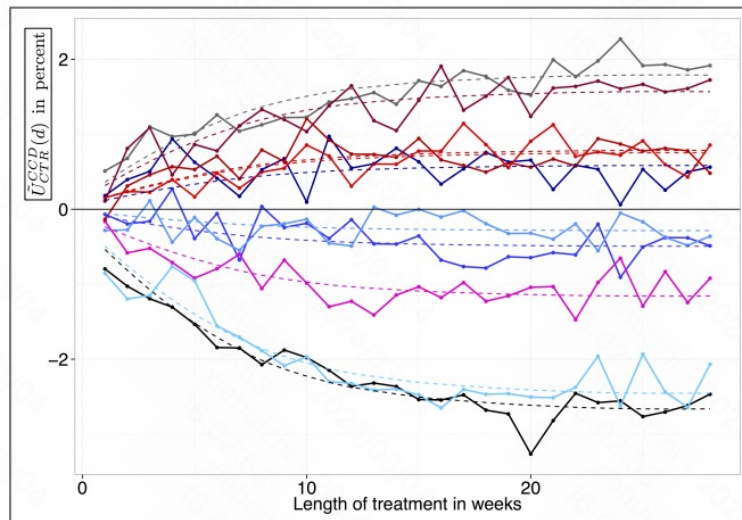


Figure 2: $\tilde{U}_{CTR}^{CCD}(d)$ for a grid of 10 different system parameter changes for mobile devices. The dashed lines give a simple exponential learning model, see Equation (4).

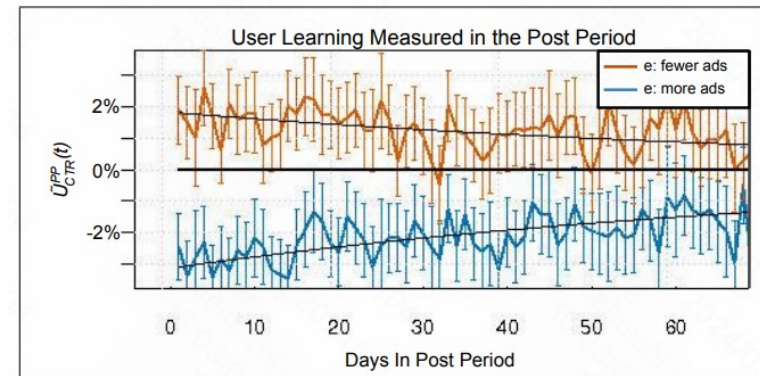


Figure 3: Two Post-Period comparisons to the control.

Model User Learning Cont'd

(Hohnhold, O'Brien and Tang 2015)

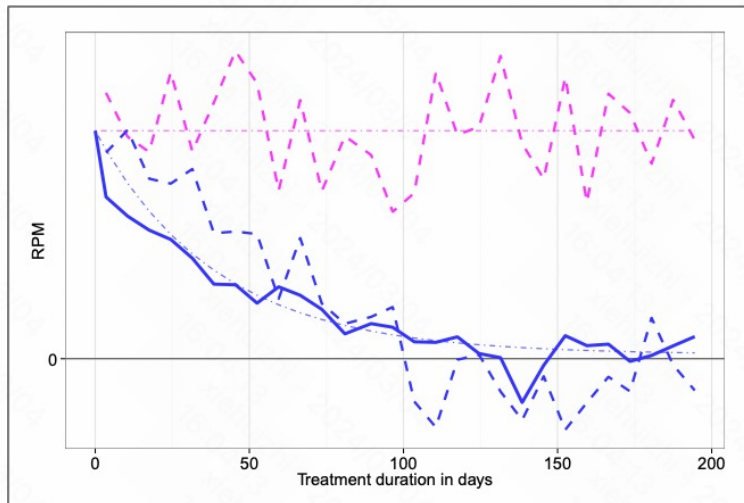


Figure 5: Short-term (pink) and long-term (blue) Δ RPM metrics for simple ad load changes on mobile Google search, restricted to old cookies, 6/26/2013 – 1/9/2014.

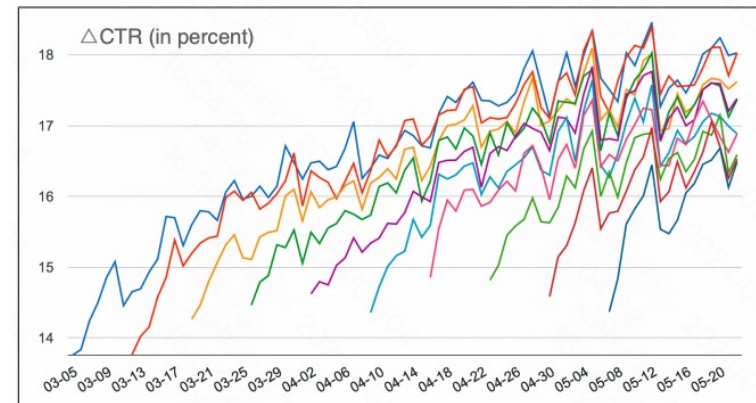


Figure 6: Δ CTR time series for different user cohorts in the launch. (The launch was staggered by weekly cohort.)

A Short Discussion On Surrogates

- A statistical surrogate lies on the causal path between the treatment and the long-term outcome. Condition on the surrogate, the treatment and the long-term outcome are independent.
- Both observational data and experimental data can be used to find good surrogates.
- Example work:
 - The surrogate index: combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Athey et al. 2019. Facebook used this approach to find good surrogates of the 7-day outcome of an experiment by just using 2-3-day experiment results.
 - Estimating effects of long-term treatments. Huang et al. 2023
- Having too many surrogates may make this approach less interpretable.

Overall Evaluation Criterion Metric (OEC): Problem

- Key properties of a good OEC:
 - Indicative of the long-term gain in key business metric
 - Hard to game and incentivize the right set of actions in the product team
 - Sensitive: easy to move in experiment
 - Cheap to compute
 - ...

Finding A Good Metric Is Hard: A Real Life Example

In 1902, the French quarter in Hanoi was overrun with rats. A "deratization" scheme paid citizens for each rat they captured (the proof requested for payment was rat's tail).



But they barely made a dent in the problem!

Two phenomena happened:

1. Tailless rats started appearing
2. A thriving rat farming industry emerged in the city

Finding A Good Metric Is Hard: The Microsoft Bing Example

- OEC for search
 - A ranking bug in an experiment resulted in very poor search results
 - Degraded (algorithmic) search results cause users to search more to complete their task, and ads appear more relevant
 - Distinct queries went up over 10%, and revenue went up over 30%
- Analyzing queries per month, we have

$$\frac{\text{Queries}}{\text{Month}} = \frac{\text{Queries}}{\text{Session}} \times \frac{\text{Sessions}}{\text{User}} \times \frac{\text{Users}}{\text{Month}}$$

where a session begins with a query and ends with 30-minutes of inactivity.
(Ideally, we would look at tasks, not sessions).

- In a controlled experiment, the variants get (approximately) the same number of users by design, so the last term is about equal
- Key observation: we want users to find answers and complete tasks quickly, so queries/session should be smaller
- The OEC should therefore be based on the middle term: Sessions/User

KPIs And Experiment Metrics Are Different!

	KPIs	Experiment Metrics
Purpose	Measure how effectively key business objectives are being achieved Lagging Metric	Evaluate specific hypotheses and make ship decisions based on indicators for improvement in KPI Leading metric
Format	Often displayed as a single uncontrolled time series in a dashboard	Often displayed as a comparison between treatment and control in a scorecard
Focus	Focus on trends, anomalies, and comparisons against targets	Focus on statistically significant deltas between treatment and control groups
Time period	MoM, QoQ, YoY	1 week, 2 week
Granularity	Measured in aggregate	Measured for each user in the experiment

Heterogeneity In Treatment Effects (HTE): Problem

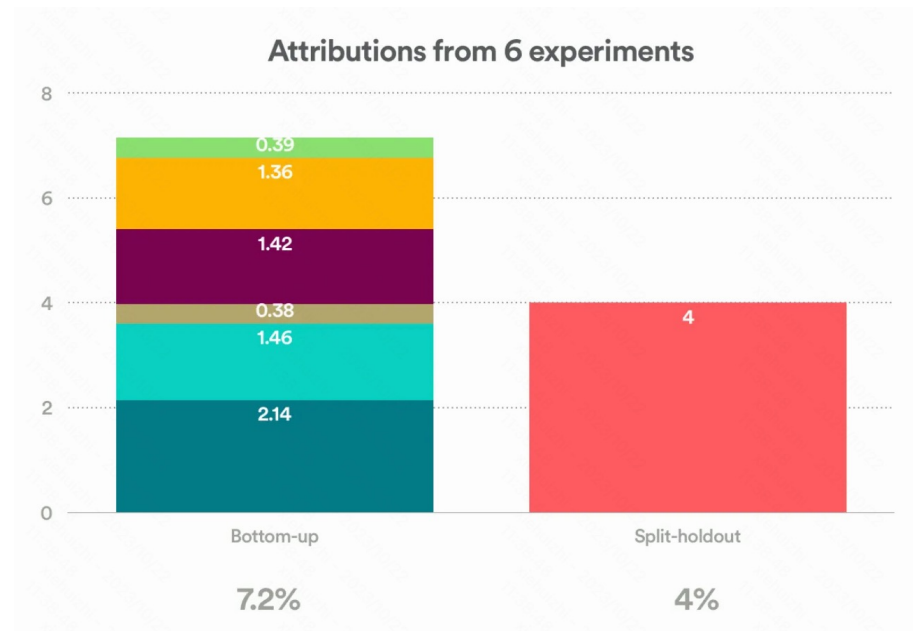
- The primary goal of an A/B test is to understand the average treatment effect.
- It is ideal to know individual treatment effect.
- Knowing individual treatment effect is not possible because we cannot observe the counterfactual.
- The closest to individual treatment effect is conditional average treatment effect.

Heterogeneity In Treatment Effects (HTE): Common Solutions And Challenges

- Solution:
 - Common user segments
 - Market/country
 - User activity level
 - Device and platform
 - Time and day of week
 - ...
 - Machine learning models: Athey et al.'s honest decision tree and random forest
- Challenge:
 - Computation scale
 - Low signal-to-noise ratio
 - Multiple testing
 - Interpretable and memorable results

Aggregated Effect Of Multiple Experiments: Problem (Shen and Lee 2018)

Winner's curse: estimated effect of launched experiments is higher than the true effect.



Aggregated Effect Of Multiple Experiments: Solution (Shen and Lee 2018)

- Set of experiments with statistically significant effects: $A = \{i | X_i / \sigma_i > b_i\}$
- True aggregated effect of experiments in A : $T_A = \sum_{i \in A} a_i$
- Estimate of true aggregated effect of experiments in A : $S_A = \sum_{i \in A} X_i$
- A is a random set, so in general $ES_A \neq ET_A$
- In fact, $ES_A \geq ET_A$, the proof is as follows:

$$\begin{aligned}\mathbb{E}[S_A - T_A] &= \mathbb{E}\left[\sum_{i \in A} (X_i - a_i)\right] = \mathbb{E}\left[\sum_{i=1}^n I(i \in A)(X_i - a_i)\right] \\ &= \sum_{i=1}^n \mathbb{E}[I(i \in A)(X_i - a_i)] = \sum_{i=1}^n \mathbb{E}\left[I\left(\frac{X_i}{\sigma_i} > b_i\right)(X_i - a_i)\right] \\ &= \sum_{i=1}^n \mathbb{E}[I((X_i - a_i) > (b_i \sigma_i - a_i))(X_i - a_i)].\end{aligned}$$

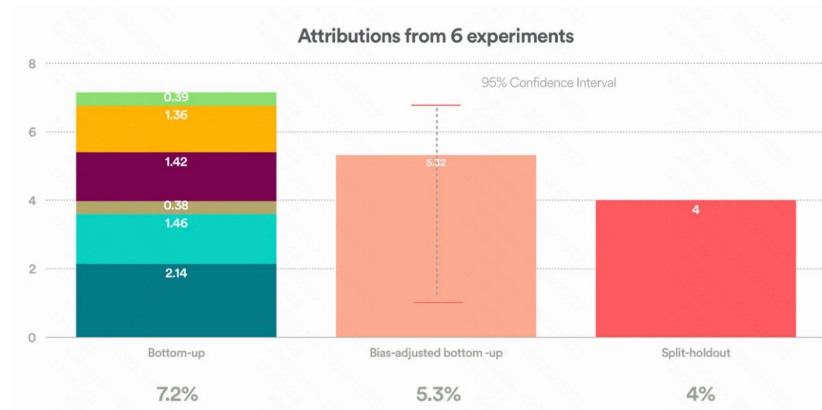
Aggregated Effect Of Multiple Experiments: Solution Cont'd (Shen and Lee 2018)

- Bias:

$$\beta = \mathbb{E}[S_A - T_A] = \sum_{i=1}^n \sigma_i \phi\left(\frac{\sigma_i b_i - a_i}{\sigma_i}\right)$$

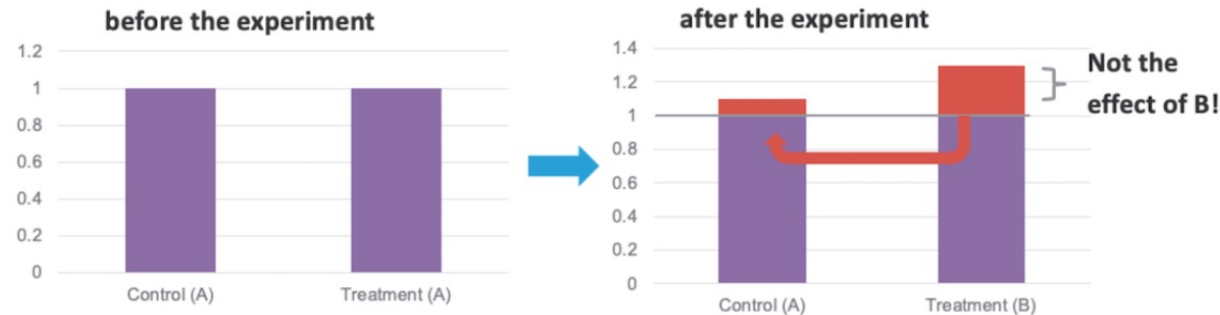
- Estimate of bias:

$$\hat{\beta} = \sum_{i=1}^n W_i \phi\left(\frac{W_i b_i - X_i}{W_i}\right)$$



Network Effects: Problem

- The stable user treatment value assumption (SUTVA) is violated because of user non-independence.



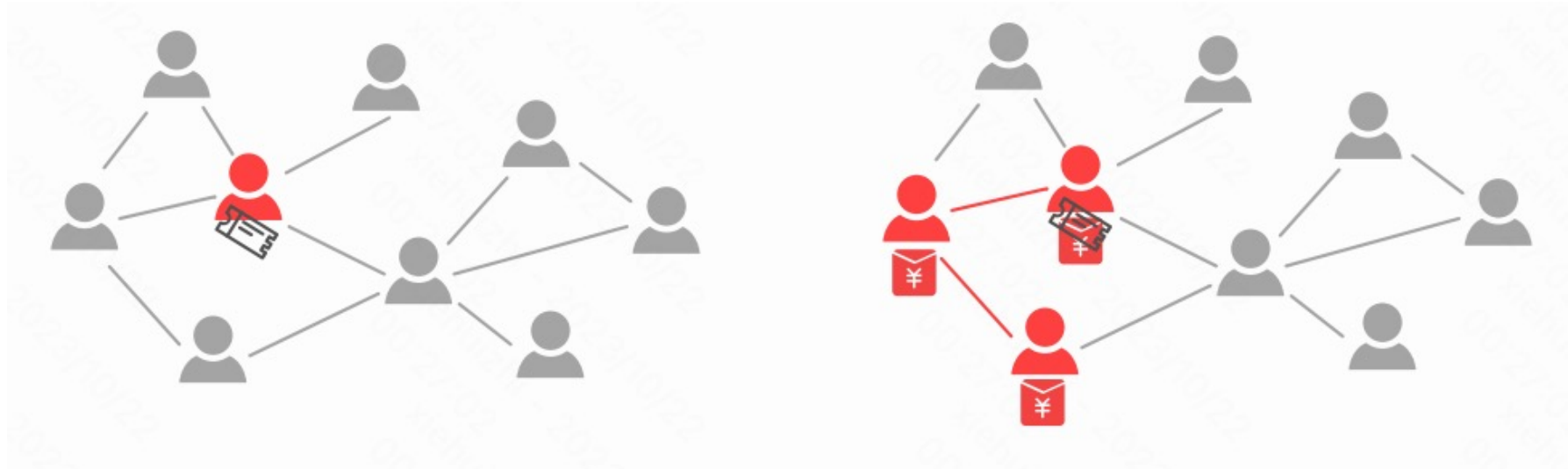
A/B testing with interference. When treatment leaks into control, we can no longer rely on computing $\text{mean}(B) - \text{mean}(A)$.

<https://engineering.linkedin.com/blog/2019/06/detecting-interference--an-a-b-test-of-a-b-tests>

Network Effects: Common Solutions

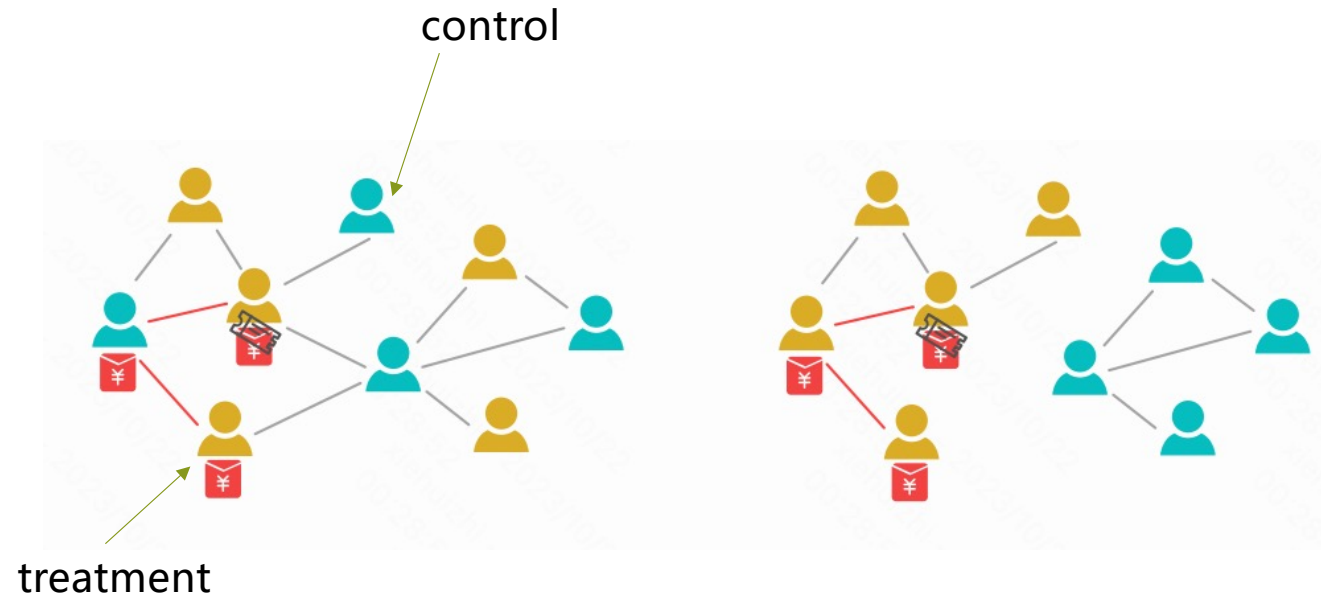
- Producer and consumer model: two-sided randomization
- Known influence network model: randomization at user-cluster level
- One-to-one communication: randomization at call level, causal inference based on observational data
- Market effects: randomization at user-cluster level, equal budget split for ads experiments
- Multiple identities for the same person

Known Influence Network Model: Ant Group's Person-to-person Spread Marketing Strategy (Cai et al. 2021)



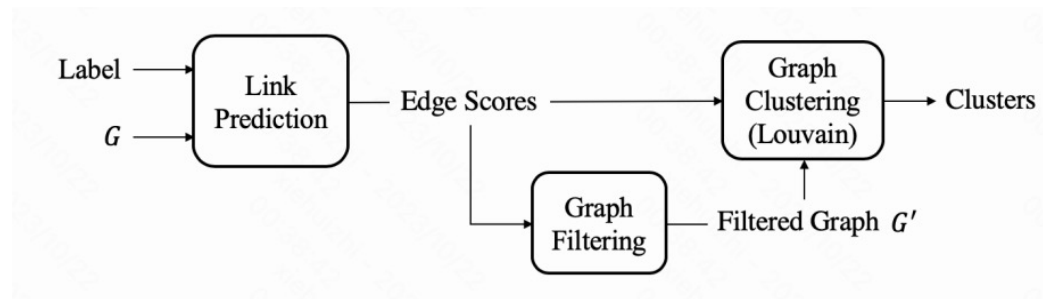
Inviter and invitee both get cash incentive.

Treatment Effect Based On User-level Randomization Is Biased. (Cai et al. 2021)



Cluster Level Randomization Improves Accuracy Of The Treatment Effect Estimate By **50%**. (Cai et al. 2021)

1. LinkLouvain clustering:



2. Cluster level randomization

3. Delta method to estimate the variance of user-level metrics

Interaction Between Experiments: Problem

- Let T_1 and T_2 be two treatments in two experiments. If $ATE(T_1) + ATE(T_2) \neq ATE(T_1T_2)$, then we say there is interaction between T_1 and T_2 .
- Textbook example: T_1 changes the foreground color to blue, T_2 changes the background color to blue.

Interaction Between Experiments: Common Solution

- The layer system: experiments that interact with each other run on the same layer.
- Automatic detection of interaction effects between experiments that run on different layers concurrently

Developing Experimentation Culture: Problem

- It can be hard to subject one's idea to experimentation and receive negative feedback. It feels like someone telling you that your baby is ugly.
- Experimentation culture development stages:
 - Hubris: every idea is considered a winner
 - Skepticism: some experimentation is run and intuition challenged
 - Humility: coming up with ideas that work is not easy
- It is ideal to see experiments with positive results. But experiments with negative results are also of great value:
 - Avoid rolling out worse experience to users
 - Better understanding of users and product

Developing Experimentation Culture: Common Solution

- Experimentation platform and tools: improve experimentation speed
- Practices, policies, and capabilities:
 - High touch: by quarter, experimentation team works with critical business team
 - Top down buy in
 - Negative and positive case studies
 - Safe rollout: work with engineers
 - Report cards and gamification
 - Education and support

Netflix's Learning Culture

Co-CEO Reed Hastings teaches Netflix employees the basics of A/B testing

Our Data & Insights team has shaped our company culture by chasing opinions for strategy and across a huge volume of levels of the company.

- Elizabeth Stone, VP of Data and Insights

An under the morning already-gatekeeper stakeholder clear decision framework, and a suite of intuitive tools for them to interpret straightforward tests, while being more hands-on in riskier spaces. This scales our experimentation volume while maintaining decision quality.

- Mihir Tendulkar



Training Others To Scale Experimentation: Problem

- The concept of A/B testing is simple. But there can be complex practical issues that cannot be standardized and solved by platform or a set of FAQs.
- Centralized team does not scale very well.
- Practical challenges in spreading expertise about experimentation that enable experimentation at scale:
 - How do we set up a community to support experimenters?
 - How do we incorporate them in the experiment lifecycle?
 - How do we incentivize these people?
 - How do we quantify their impact?
 - How do we train them?
 - How do we maintain quality standards?

Training Others To Scale Experimentation: Common Solution

- Yandex's "*Experts on Experiment*" program: handpicked from product team by the central experimentation group
- Amazon's "*Weblab Bar Raisers*" program: high-judgment and experienced experimenters
- Twitter's "*Experiment Shepherds*" program: including CTO, around 50 members, pre-test and pre-launch review
- Booking.com's "*Experimentation Ambassadors*" program: handpicked by the central experimentation group, around 15 members, work with a support ticketing system to provide experimentation consulting service
- Microsoft's "*Center of Excellence Model*" : the data science team in the central experimentation group provide consulting service
- Google's "*Just-in-time Education Model*" : checklist, experiment council

There Are More Practical Challenges.

- Continuous decision-making: sequential testing
- Multiple metrics, experiment groups inference: multiple testing
- Intelligent traffic allocation: multi-arm bandits, connected to reinforcement learning
- Causal inference scenarios where experimentation fails or is too costly: natural experiments, observational data causal inference
- The application of large language models in experimentation
- ...

/04 Takeaway



Key Takeaway

- Internet company builds product/service for its users. Data-driven operation improves the efficiency of value delivering to users, thus the efficiency of continuous growth.
- Metric and causal inference are two key elements for data-driven operation to work.
- Causal inference is not easy. A/B testing is the gold standard for causal inference.
- The concept of A/B testing is simple, but the practice is hard, 1/3 is about business and user understanding, 1/3 is about engineering, 1/3 is about science.

Let's End With Kohavi's Quote.



Why I love controlled experiments

In many data mining scenarios, interesting discoveries are made and promptly ignored. In customer-driven development, the mining of data from the controlled experiments and insight generation is part of the critical path to the product release