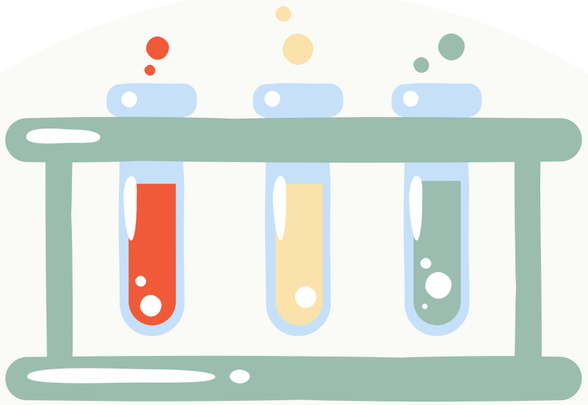
A/B TEST EXPERIMENT

A Brief Introduction and Best Practices



Introduction and motivation

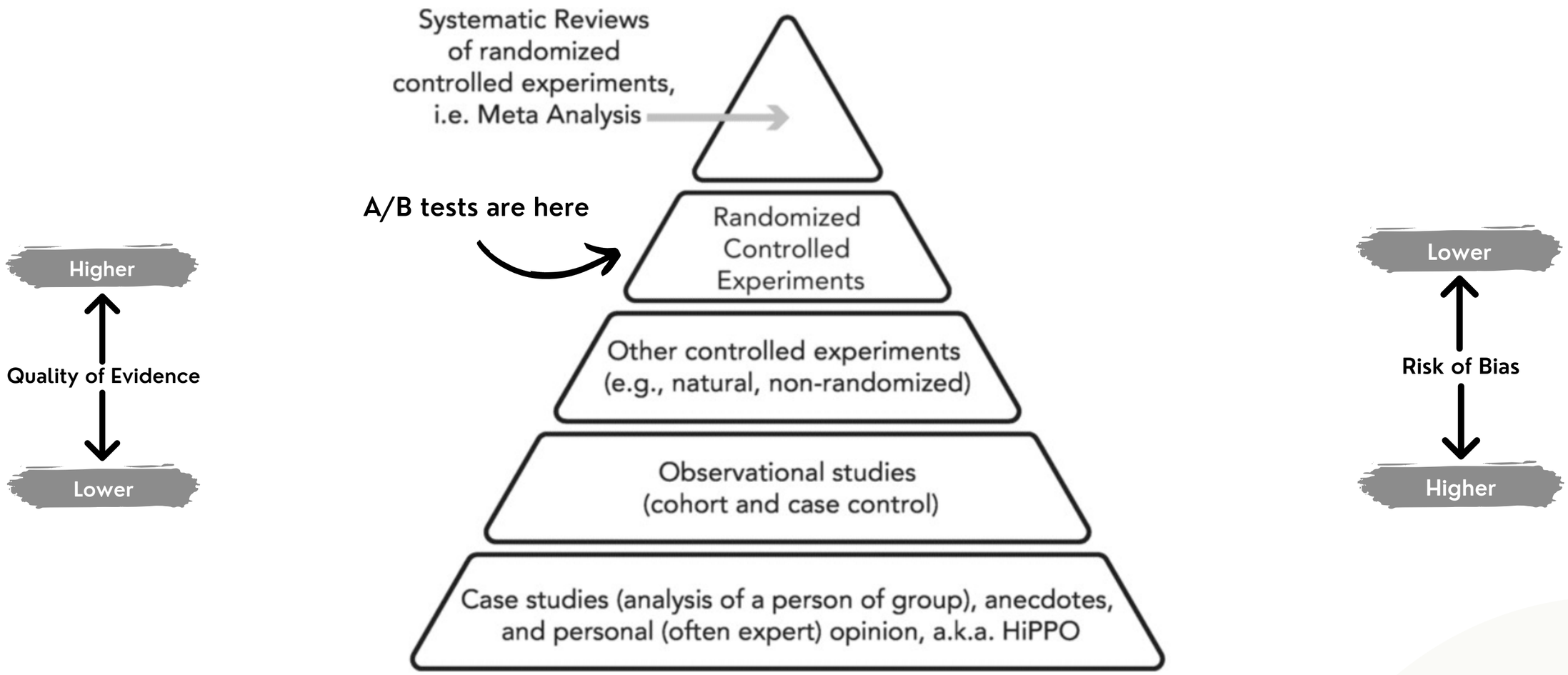
Which one is better?

Which one is better?

Secara umum, kita **tidak bisa menebak** dengan pasti mana ide yang lebih baik.
Oleh karena itu, **pengujian dan observasi** perlu dilakukan.

Hierarchy of Evidence

Greenhalgh, 2014

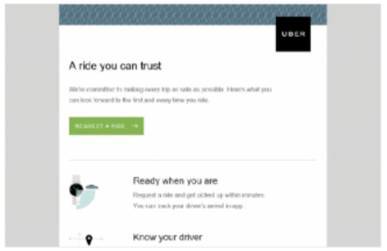


"A/B test experiment is the gold standard for establishing causality"

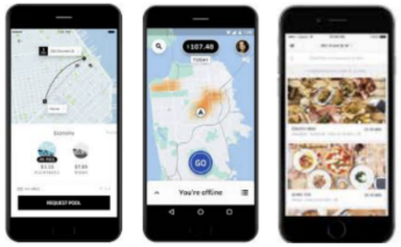
Kohavi, Tang, Xu (2020)

A/B practice is here and there (e.g. Uber & Netflix)

Where they do A/B?



Backend
Python, Java, Go



Mobile
iOS, Android

What they do A/B?

User facing features

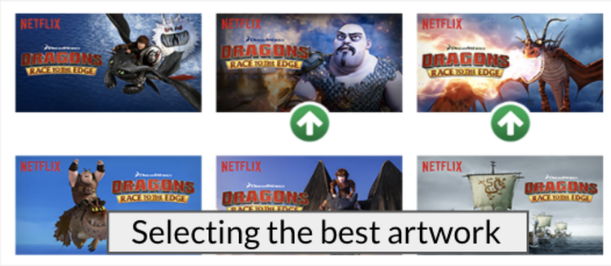
Bug fixes

Where they do A/B?

Backend

Homepage

What they do A/B?



Selecting the best artwork

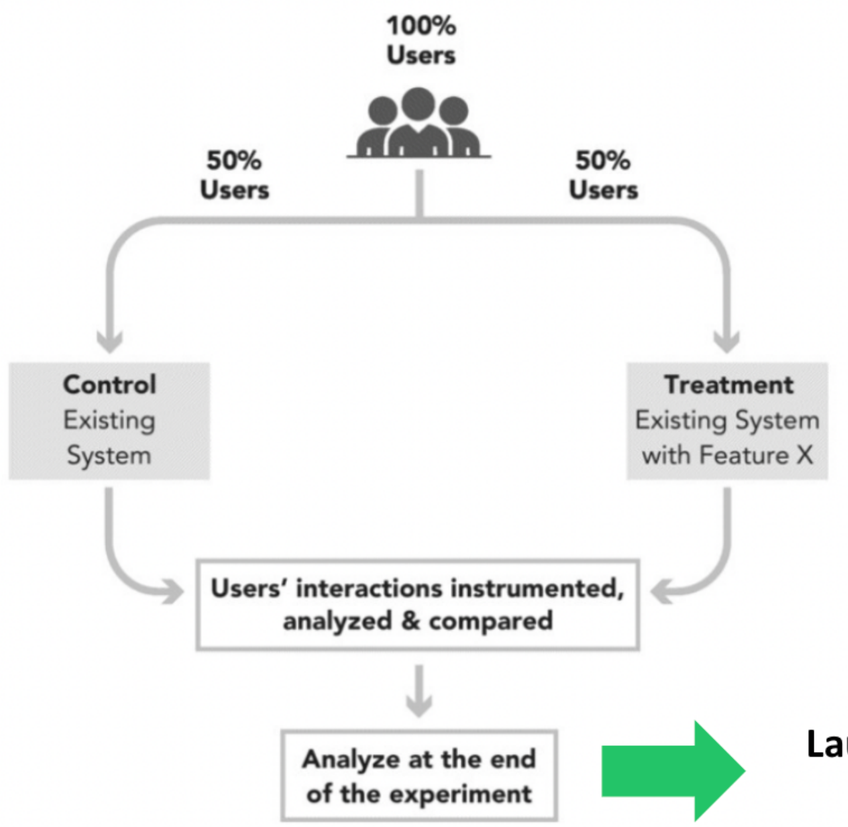
Streaming Video

1 personalized page

10-40 rows

per device

What is A/B Testing?



Launch or **don't Launch?**

- User dibagi **secara acak** menjadi 2 grup (atau lebih) → memastikan tidak ada bias antara control dan treatment
- Pengujian dilakukan **secara bersamaan dalam periode** waktu yang sama → memastikan tidak ada seasonality yang mempengaruhi data

Why A/B Testing?

Case:
Misalkan sebuah perusahaan mengeluarkan fitur **Subscription**. Dengan Subscription, user akan mendapatkan benefit tertentu.
Kemudian dari hasil analisis menunjukkan bahwa:

- Subscriber **memiliki *performance* yang lebih tinggi dibandingkan dengan Non-subscriber.**
- Subscriber **memiliki *performance* yang lebih tinggi dibandingkan dengan sebelum** periode Subscription.

Apakah kita bisa menyimpulkan **bahwa Subscription menghasilkan dampak positif?**

Why A/B Testing?

Dari hasil tersebut, kita **tidak bisa menyimpulkan sebab-akibat** dari fitur Subscription. Karena:

- *Imbalance users or biased* → Subscriber kemungkinan punya intensi yang lebih tinggi untuk dibandingkan dengan Non-subscriber
- *Seasonality* → seasonality mempengaruhi perilaku sebelum dan sesudah periode Subscription
- *Correlation <> Causality*

Why A/B Testing?

Company	Success Rate
Microsoft	33%
Bing	15%
Google Ads	10%
Netflix	10%
Airbnb Search	8%

Booking.com
1,000 tests simultaneously

airbnb
700 tests per week

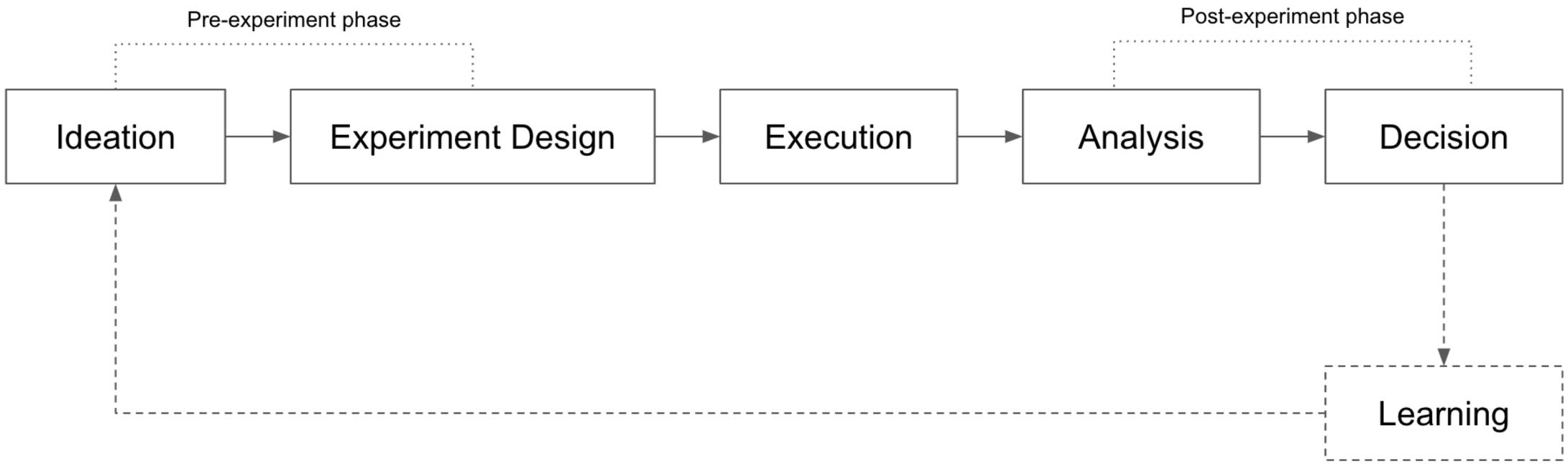
Google
>10,000 tests a year

facebook
>10,000 tests a year

Microsoft
>10,000 tests a year

We don't know how bad the changes are and how much the changes impact our metrics.
Even in Big Tech Companies, not all ideas work. Without A/B Testing, many failed initiatives will be launched

A/B Testing Cycle



Pre-experiment Phase

Experiment Design

1. **Define metrics**
   Primary metrics, Guardrail metrics, Secondary metrics

2. **Define how many variants**

3. **Power Analysis**
   Define the sample size and experiment duration

4. **Launch criteria**

Experiment Design: Define Metrics

- Primary metric → metrik utama yang menentukan kesuksesan eksperimen
- Guardrail metrics → metriks yang tidak boleh turun
- Secondary metrics → metriks penunjang yang digunakan sebagai *supporting metrics*

Story: We build a new model that showing product recommendation to engage users to repeat purchase certain products.

Metric Type	Metric Name	Reason
Primary Metric	• CVR	The goal is making users repeat purchase so we expect the CVR to increase
Guardrail Metric	• Revenue	We don't want revenue decrease
Secondary Metric	• CTR • #Product Impressions • #Transaction • #Click	The metrics help us understand the impact of the model. e.g the CVR decrease because of the CTR decrease, so that users have less interested in the products

Experiment Design: Sample Size & Experiment Duration

Power Analysis

- **Power analysis** adalah metode untuk **mengestimasi jumlah sampel dan durasi eksperimen** yang diperlukan.
- **Minimum sample size per varian** dapat dihitung menggunakan persamaan berikut *(best practice formula)*:

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

Dengan:

- Statistical Power 80%
- Alpha (significance level) 5%
- σ: Standard deviation
- Δ: Difference between control and treatment

Experiment Design: Sample Size & Experiment Duration

Minimum Detectable Effect (MDE)

- Minimum Detectable Effect (MDE) adalah **minimum uplift yang ingin kita ukur.**
- Contoh: *by changing the layout, we expect the CTR metric to increase by at least 3%.*
- Tinggi/rendahnya MDE bersifat relatif, antar perusahaan dan tim mungkin berbeda.

MDE	#Sample per Variant	Duration (days)	Duration (weeks)
1%	6,122,126	248	35
2%	1,530,532	62	9
3%	680,236	28	4
4%	382,633	16	2
5%	244,885	10	1
6%	170,059	7	1
7%	124,941	5	1
8%	95,658	4	1
9%	75,582	3	0
10%	61,221	2	0

Experiment Design: Launch Criteria

We decide to roll out the Variant IF

1

Scenario 1

- Primary metric statistically significant **increase**
- Guardrail metrics **increase or at least flat** (no significant decrease)

2

Scenario 2

- The primary metric is flat
- Guardrail metrics are flat (no significant decrease)
- Secondary metrics are a statistically significant increase

Experiment Design

Let's Try: Sample size calculator

Post-experiment Phase

Post-experiment Analysis

1. Sample Sufficiency
2. Sample Ratio Mismatch (SRM)
3. Descriptive & Inferential Analysis
4. Conclusion & Recommendation

Post-experiment Analysis: Sample Ratio Mismatch (SRM)

- SRM adalah bias yang terjadi apabila **proporsi user** antara varian **saat eksperimen berbeda dengan proporsi user dalam desain eksperimen.**
- **Chi-squared Goodness-of-fit test** dapat digunakan untuk mendeteksi SRM.
- Contoh:
  - Dalam desain eksperimen, proporsi antara Control dan Treatment adalah 50%:50%.
  - Saat eksperimen, jumlah Control sebanyak 10K, sedangkan Treatment sebanyak 8K. Sehingga proporsi antara Control:Treatment adalah 55%:45%
  - Berdasarkan Chi-squared Goodness-of-fit test, proporsi user saat eksperimen berbeda dengan desain eksperimen sehingga eksperimen tersebut mengalami bias SRM.

Read more: Diagnosing Sample Ratio Mismatch in Online Controlled Experiments: A Taxonomy and Rules of Thumb for Practitioners

Post-experiment Analysis: Descriptive & Inferential Analysis

Descriptive Anlysis

understand and summarize the characteristics of the data
(e.g calculate mean, median, check for outliers, etc)

Inferential Analysis

to draw a conclusion about the population based on the sample
(e.g t-test, z-test, chi-squared test)

Post-experiment Analysis: Conclusion & Recommendation

Dalam menyusun Conclusion & Recommendation, ada hal yang perlu diperhatikan:

- Berdasarkan Launch Criteria pada desain eksperimen, **apakah Treatment variant dapat kita *roll out?***
- Apa ***learning* yang kita dapat** dari eksperimen tersebut?

Experimentation Trustworthiness

1

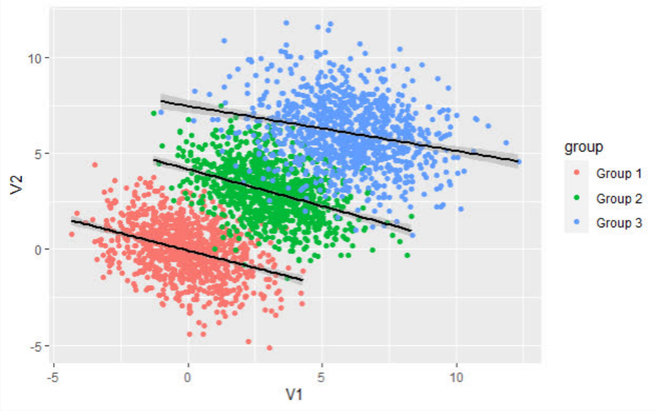
Tyman's Law

the more unusual or interesting the data, the more likely they are to have been the result of an error of one kind or another

2

Simpson's Paradox

when data is put into groups that reverses or disappears when the data is combined



3

Novelty Effect

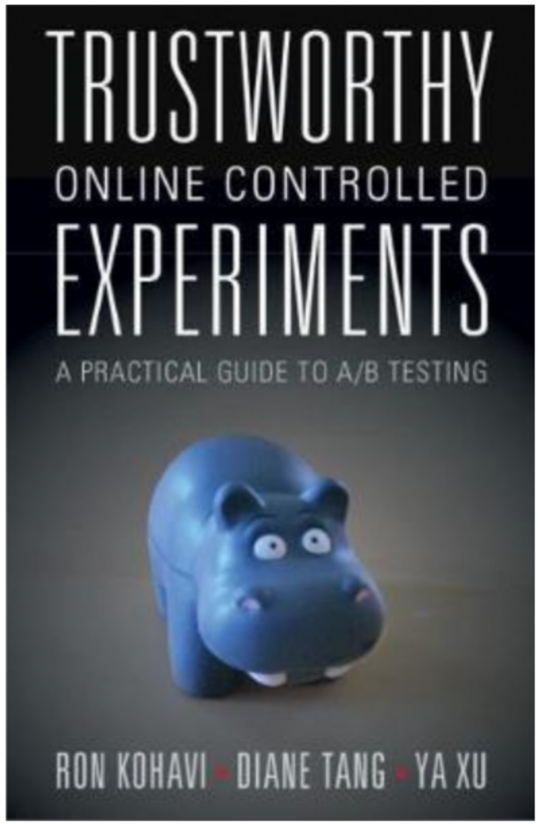
happens when users interact with new feature

4

Lack of Statistical Power

the experiment is underpowered to detect the effect size we are seeing, there are not enough users in the test

Learning References



- A/B testing papers: https://exp-platform.com/
- Expert of A/B testing: Ron Kohavi
- A paper about SRM: Diagnosing Sample Ratio Mismatch in Online Controlled Experiments: A Taxonomy and Rules of Thumb for Practitioners
- A paper about the choice of randomization unit: Choice of the Randomization Unit in Online Controlled Experiment

https://medium.com/@ahmadnuraziz3 -- follow ya, hehe

Example of Post-experiment Analysis

Example of A/B Testing Calculation

Thank You