STATSIG

Timothy Chan
Head of Data

Beyond Simple A/B Testing: Advanced Experimentation Tactics

Data Council 2024

March 27th, 2024

Statsig.com

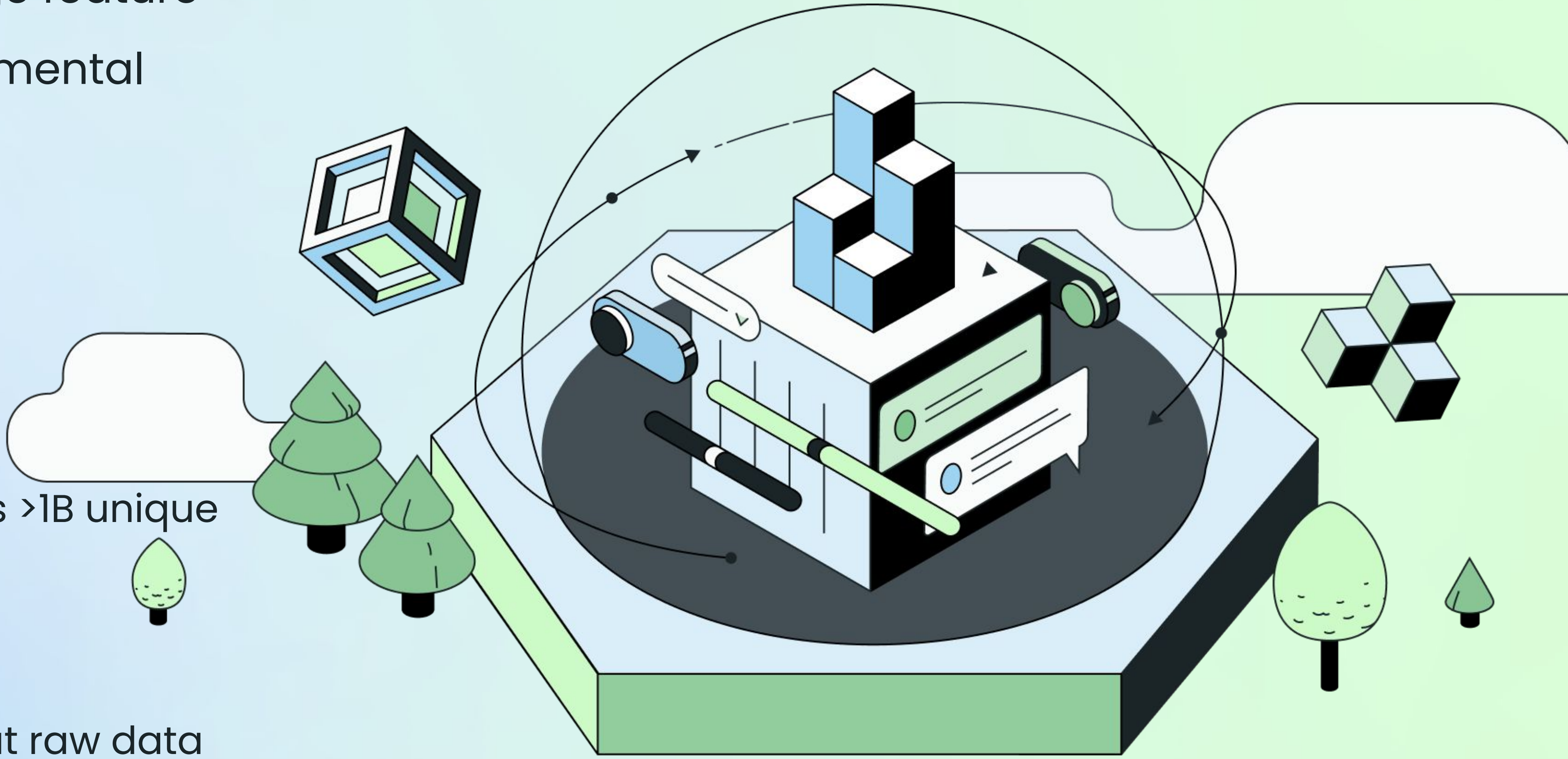
The Statsig Team

STATSIG

SEQUOIA

MADRONA
VENTURE GROUP

Statsig is a modern experimentation and feature flagging platform. We help companies like Notion, OpenAI, Figma, and Atlassian manage feature rollouts and compute experimental results.



Statsig Cloud

- >600B events a day
- >20k total experiments across >1B unique user identifiers.

Statsig Warehouse Native

- Full power of Statsig Cloud but raw data never leaves your data warehouse.

Overview

Review of Experimentation 101

1. AB Testing Basics

Experimentation 201

1. CUPED
2. Holdouts
3. The Peeking Problem and Sequential Testing
4. Stratified Sampling
5. Switchback Experiments
6. Multi-Armed Bandits
7. Heterogeneous Treatment Effects
8. Experimental Meta Analysis

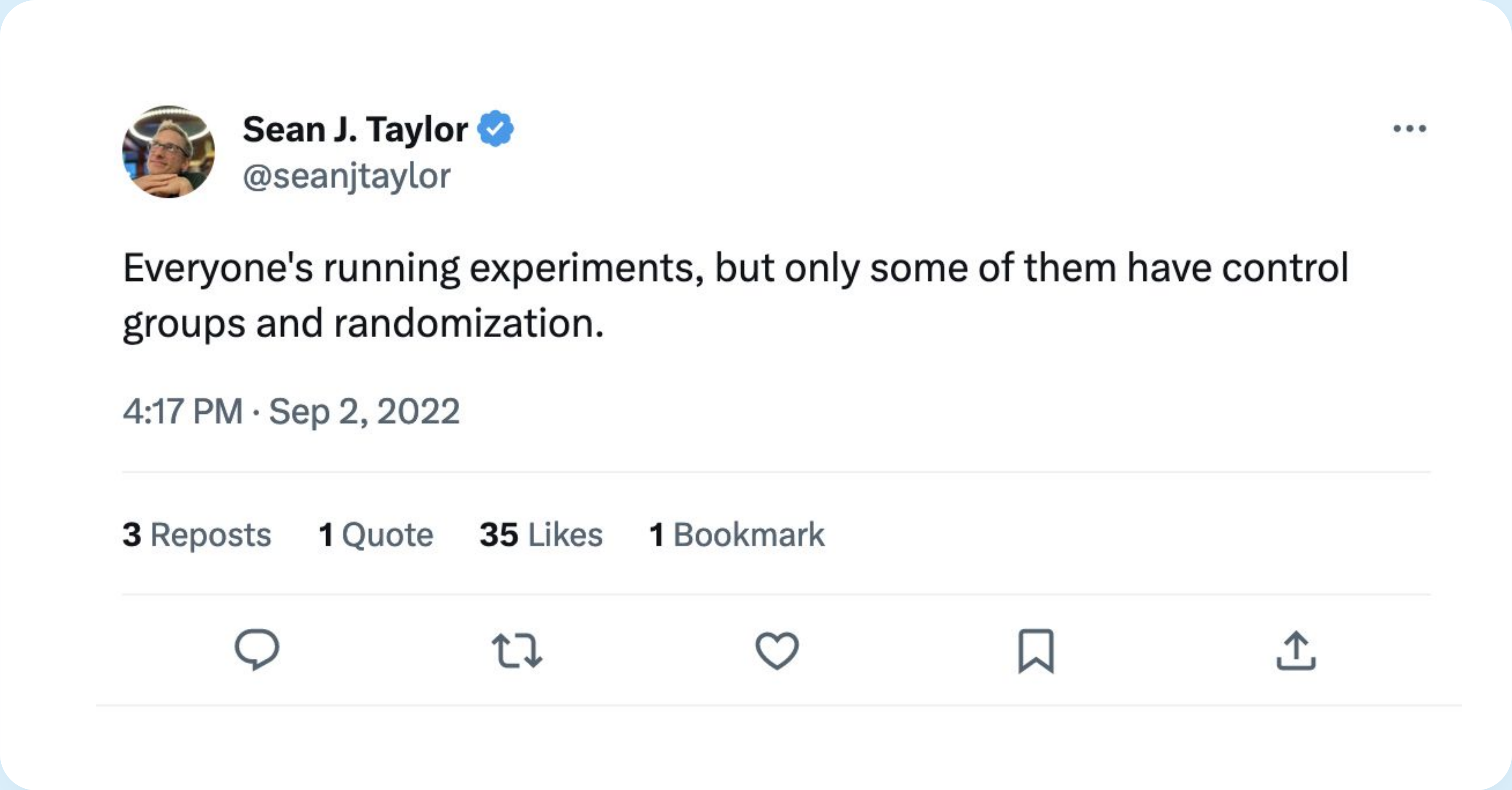
Experimentation 101: Why A/B Test?

Scientific gold standard for measuring causality

Ideas are evaluated by causal user data not opinions

Product development becomes a scientific, evidence-driven process

Building products is hard



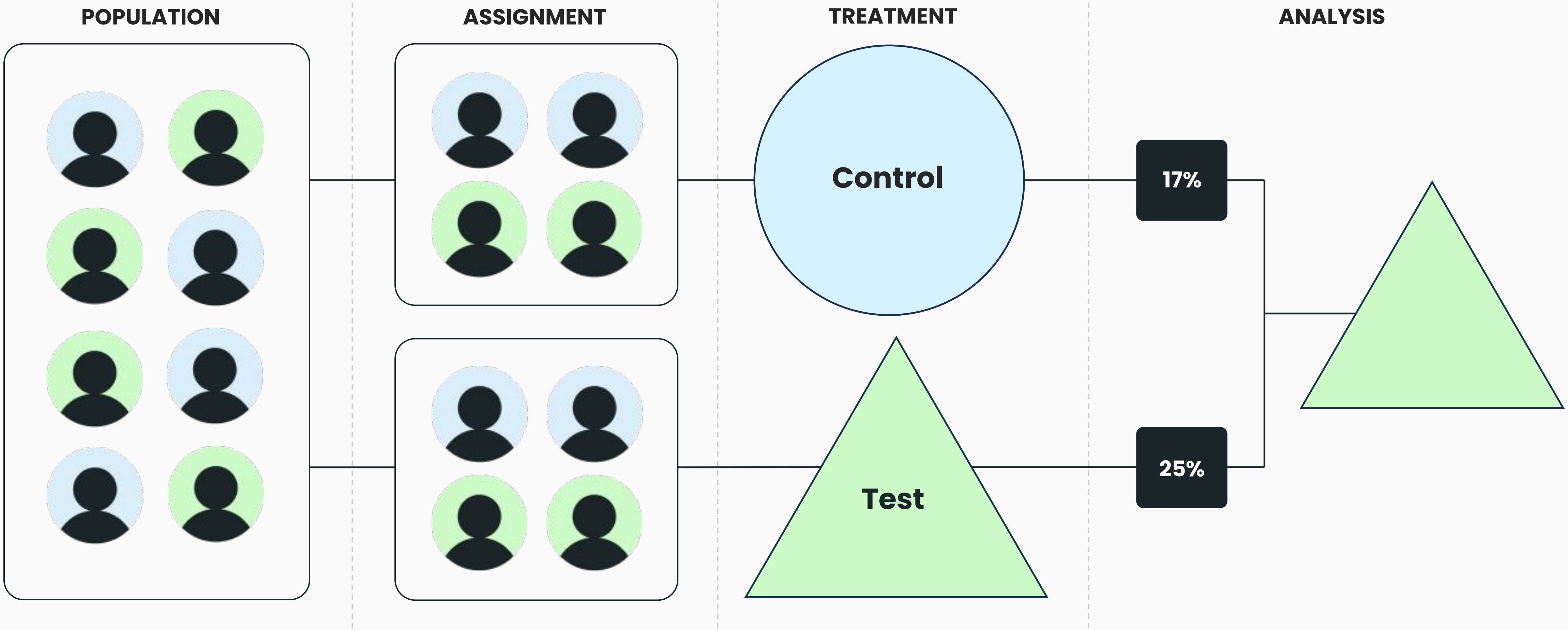
Sean J. Taylor
@seanjtaylor

Everyone's running experiments, but only some of them have control groups and randomization.

4:17 PM · Sep 2, 2022

3 Reposts **1** Quote **35** Likes **1** Bookmark

How Does Testing Work?



Experimentation Best Practices

❑ Start with a hypothesis

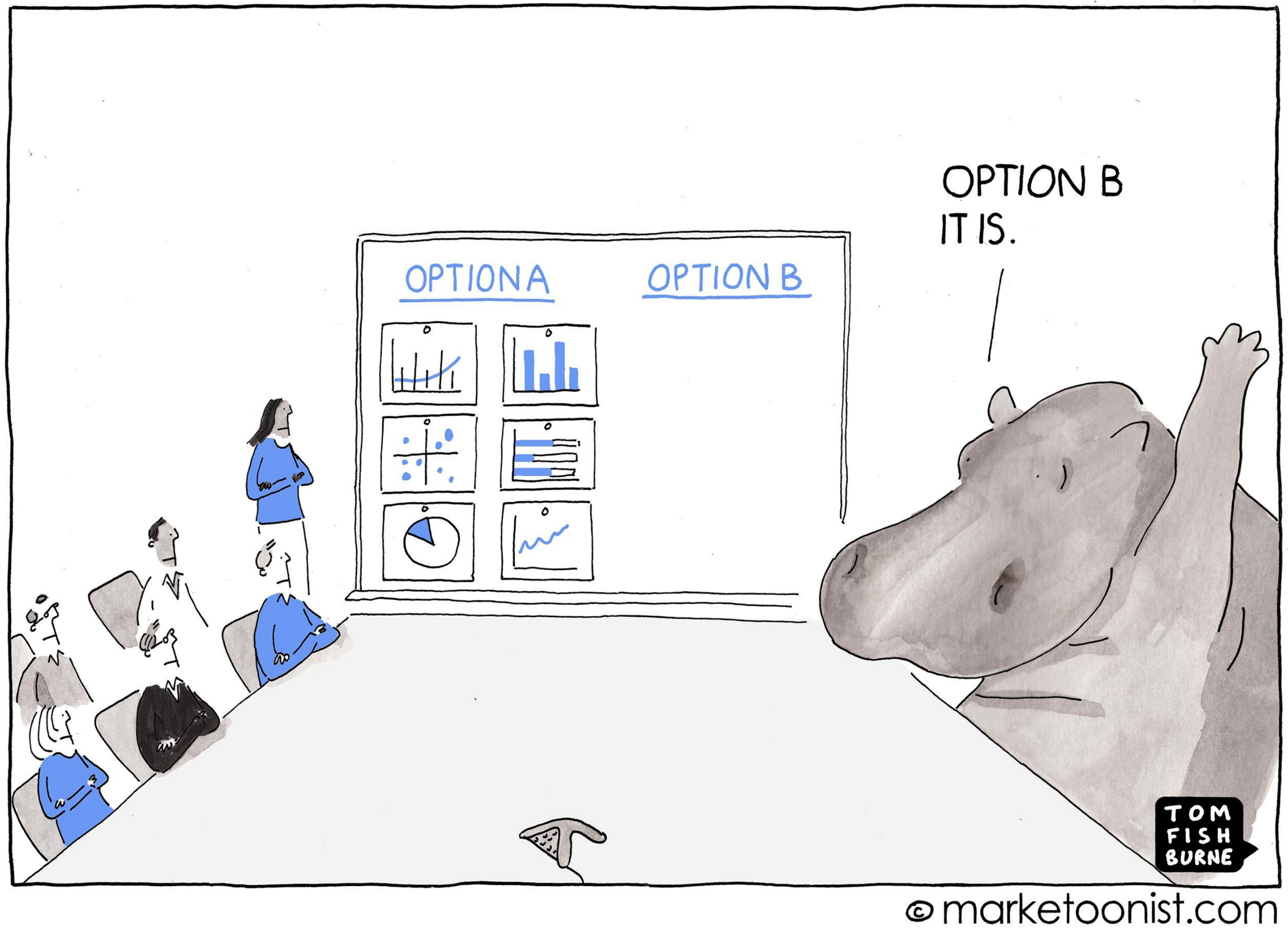
❑ Power Analysis (tradeoff between sample size, statistical power, and time)

❑ Standardized methodology

❑ Use 95% confidence intervals by default

❑ Don’t fret about interaction effects

The HiPPO



OPTION A
OPTION B
OPTION B IT IS.
TOM FISH BURNE
© marketoonist.com

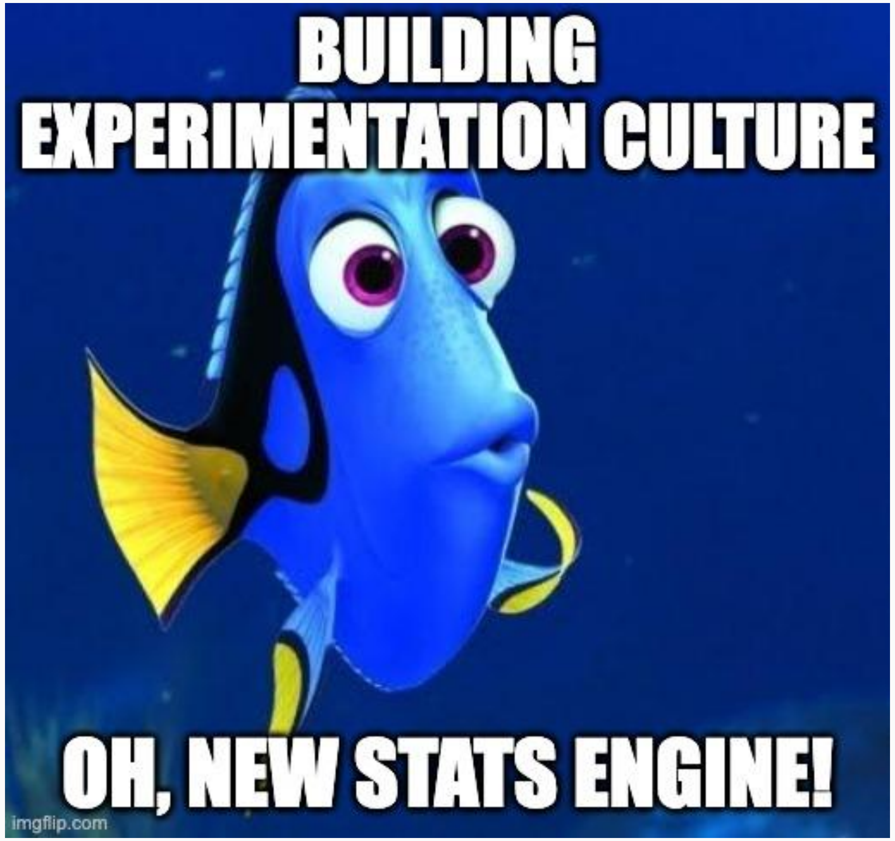
Stats Engines Don’t Build Culture

Experimentation should be easy and automatic

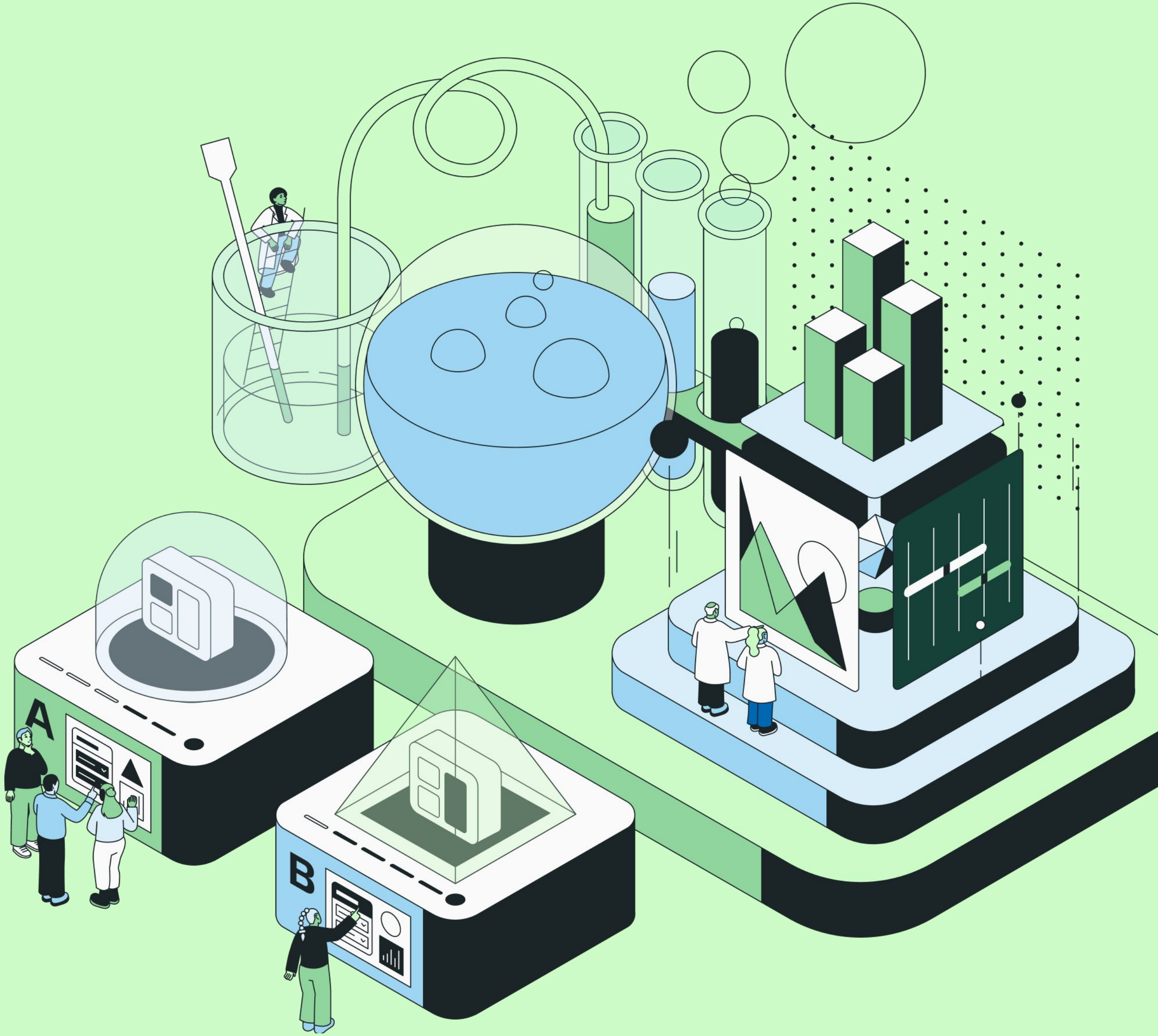
Experimentation is a team sport,
the entire product team is on the field

Experiment Review

Optimize for velocity

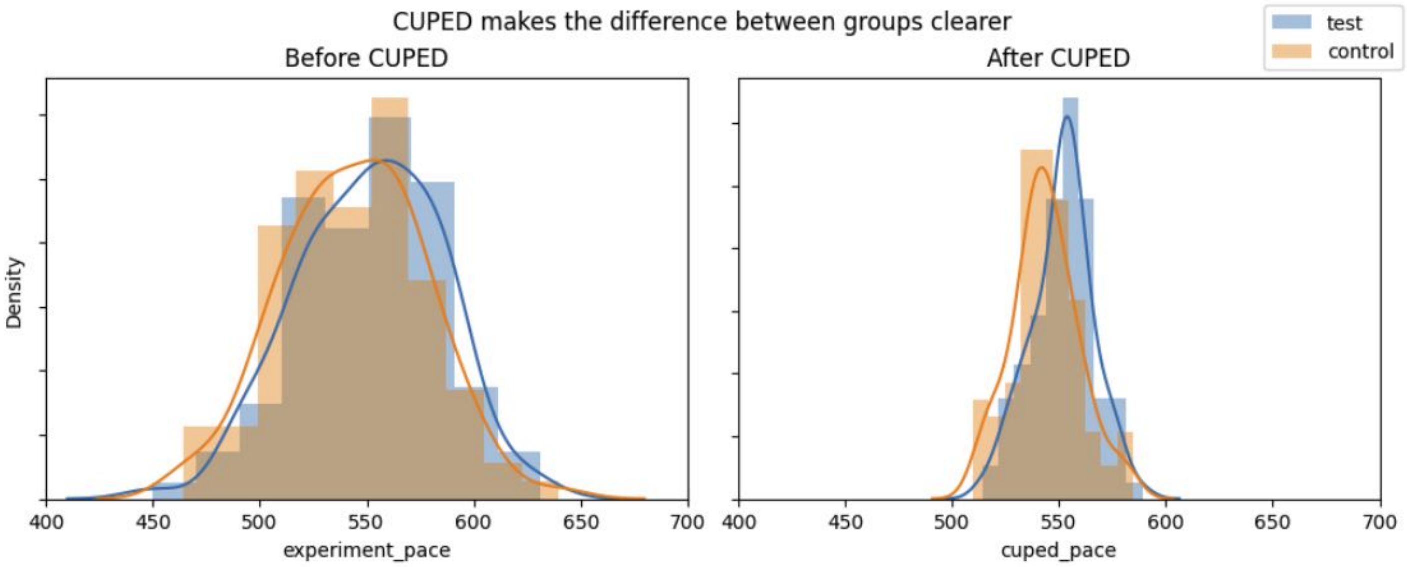


Welcome to Experimentation 201



Controlled Experiment Using Pre-Experimental Data (CUPED)

Can reduce confidence intervals by 30-60%, resulting in more statistical power in less time.



Craig Sexauer https://www.statsig.com/blog/cuped

! Problem: The Winner’s Curse

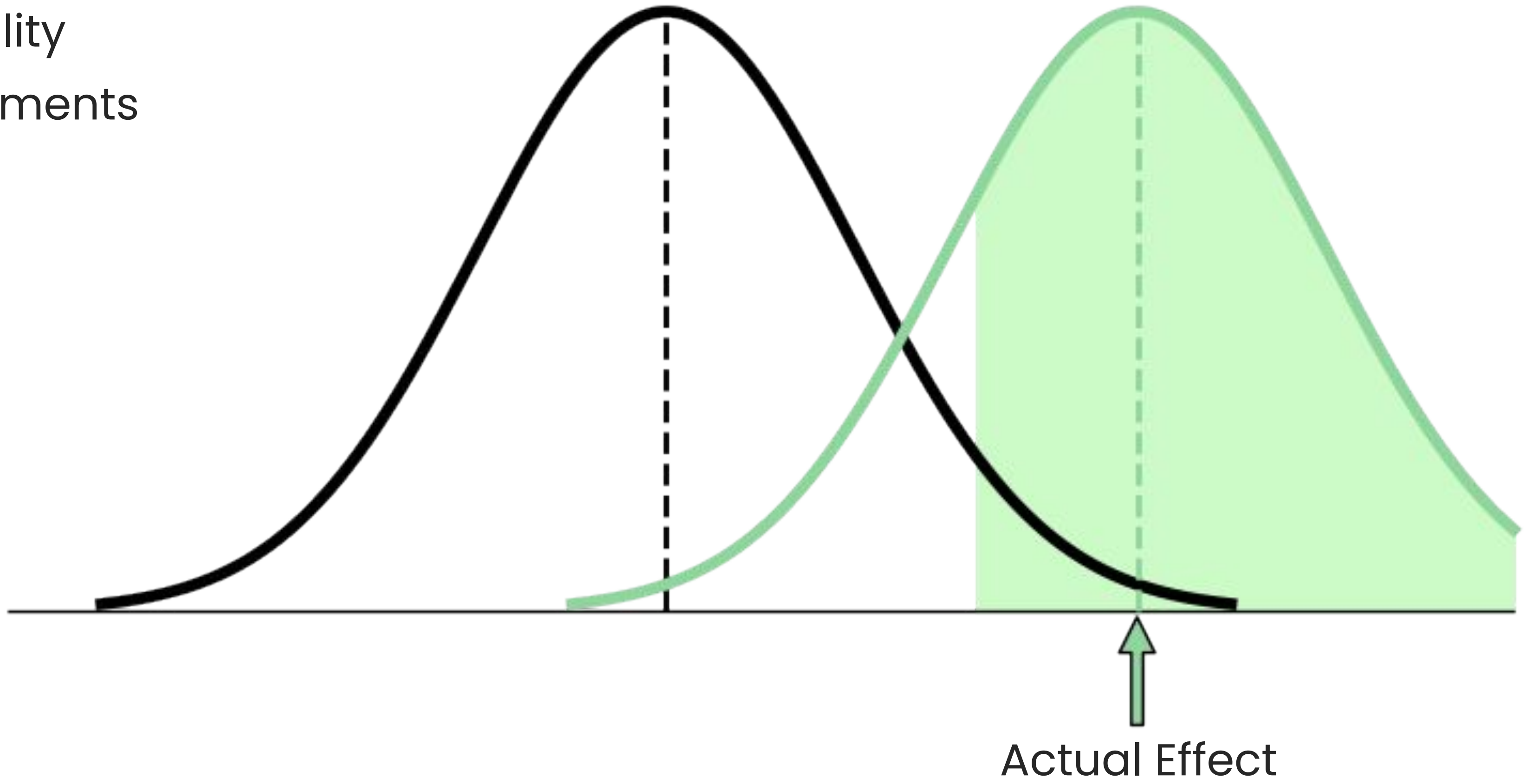
Definition

The phenomenon where estimates from AB tests do not hold up to their expectations.

! Problem: The Winner's Curse

Possible Causes

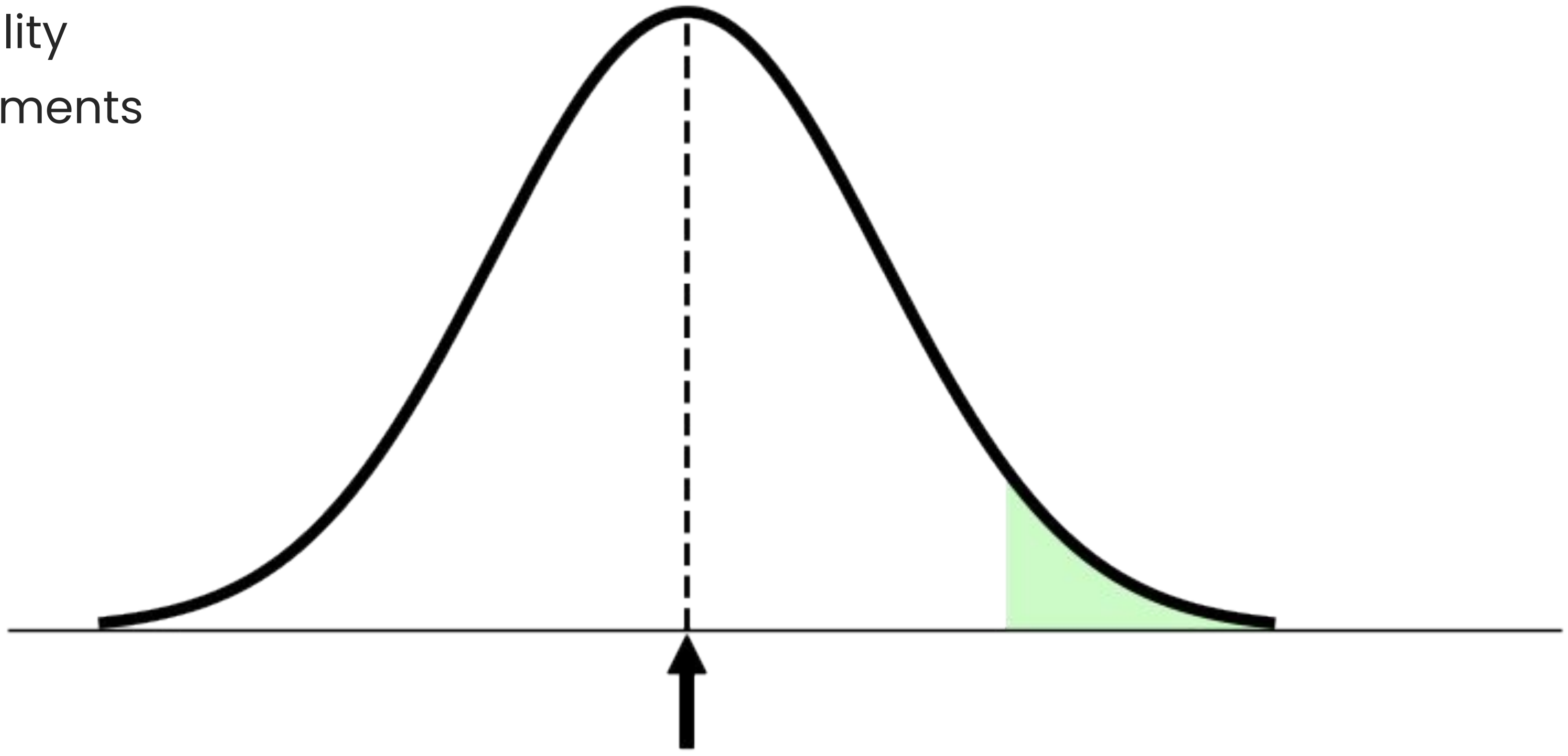
1. Long-term sustainability
2. Underpowered experiments



! Problem: The Winner's Curse

Possible Causes

1. Long-term sustainability
2. Underpowered experiments
3. False positives

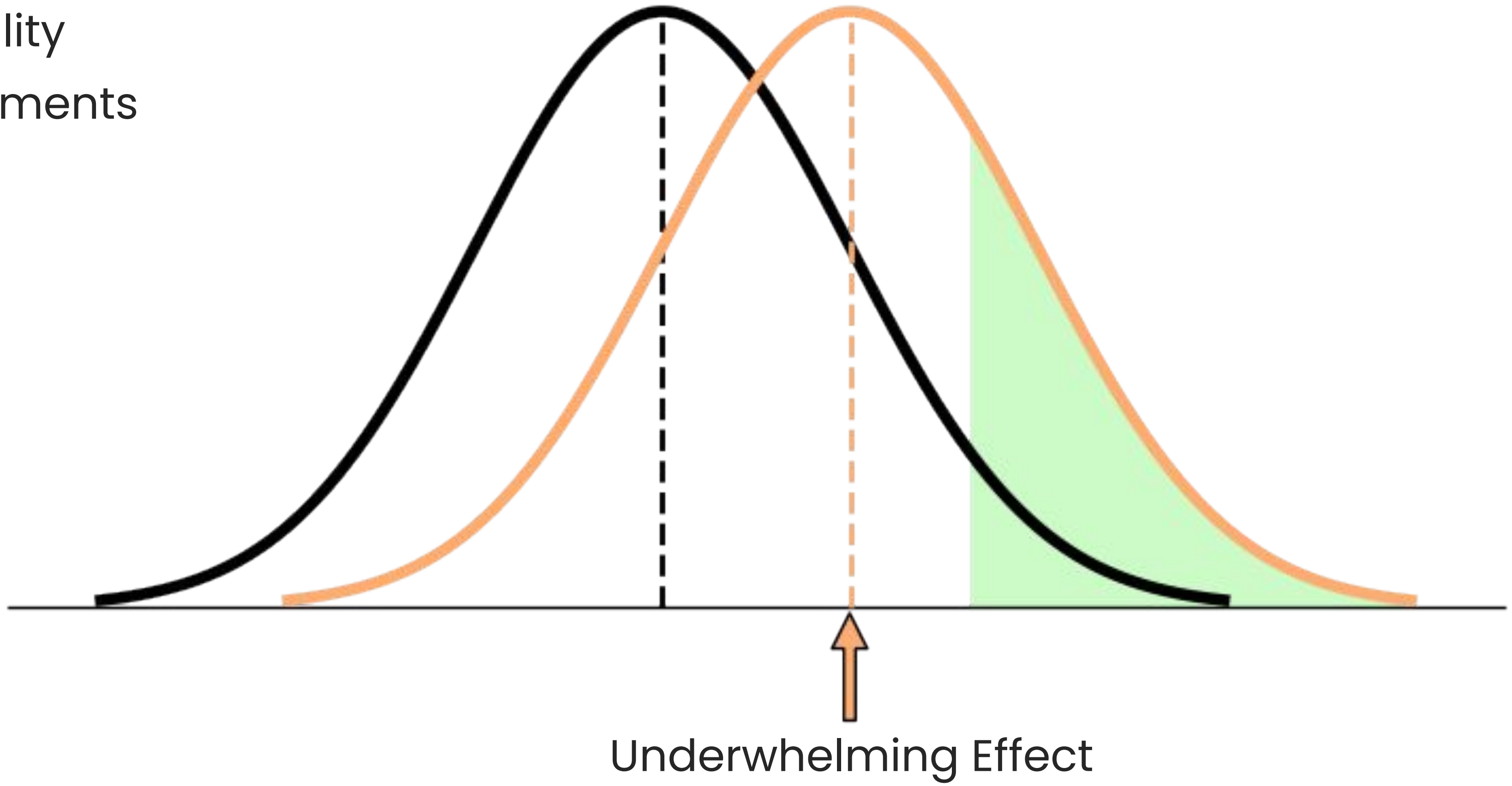


No Actual Effect

Problem: The Winner's Curse

Possible Causes

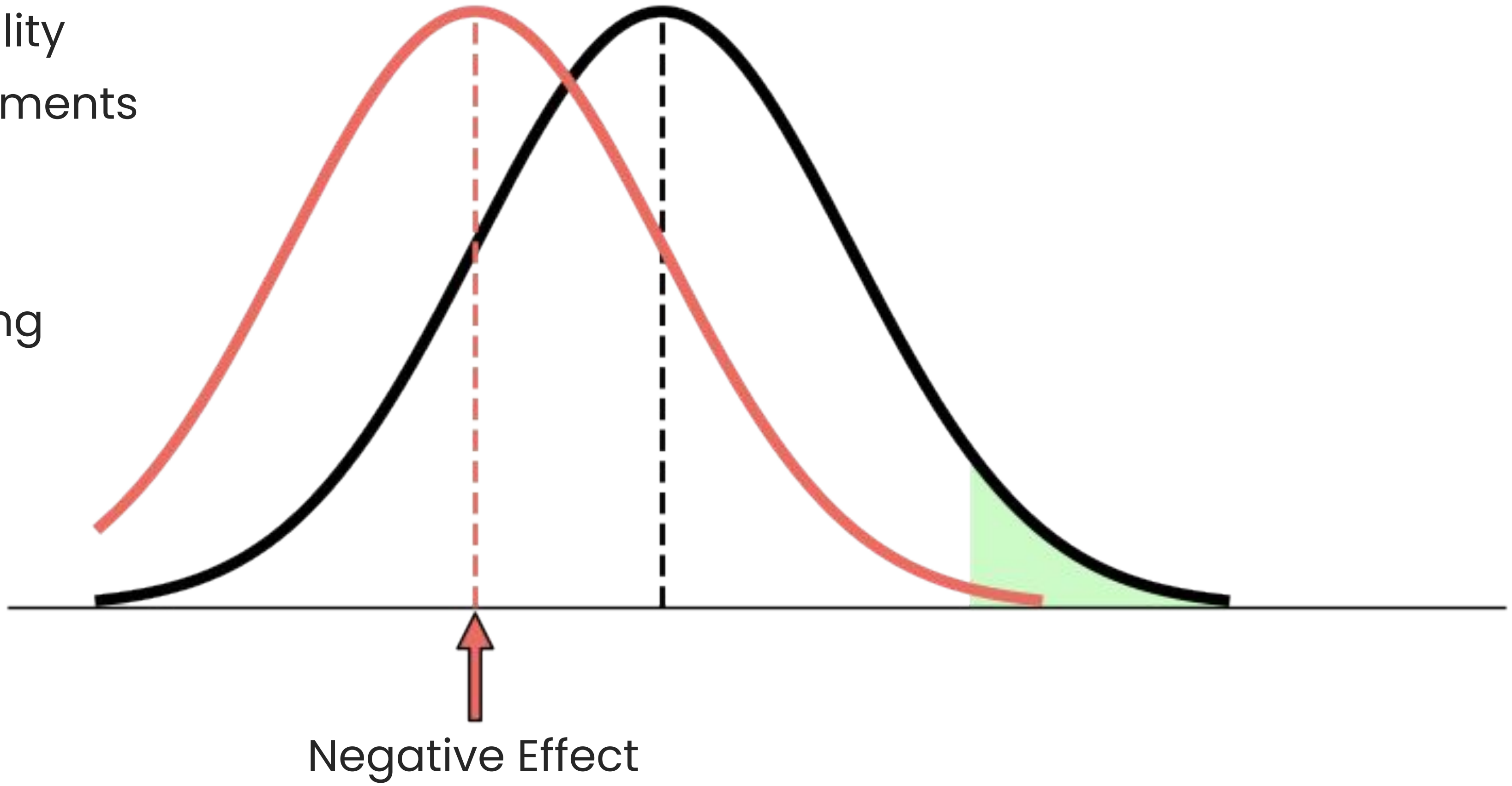
1. Long-term sustainability
2. Underpowered experiments
3. False positives
4. Over-estimations



❗ Problem: The Winner’s Curse

Possible Causes

1. Long-term sustainability
2. Underpowered experiments
3. False positives
4. Over-estimations
5. Biased Decision Making



Solution: Holdouts

Definition

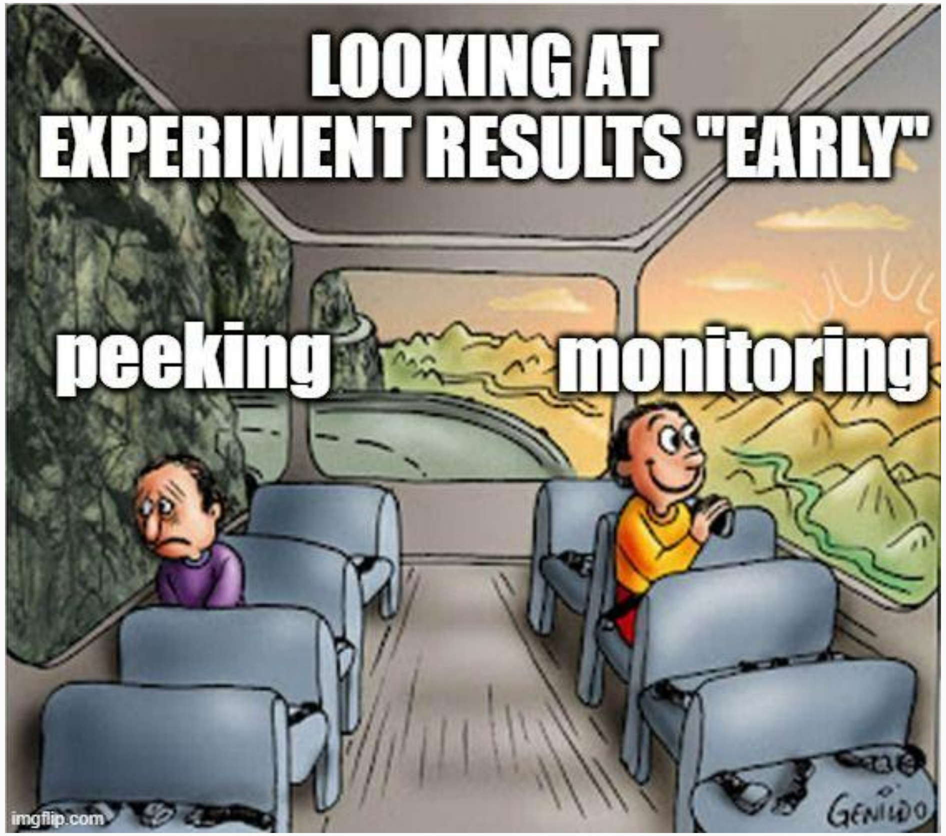
A small % of users who are intentionally withheld from a feature or features after rollout, for a longer-than-normal period.

Several Types

- Team-wide
- Feature-specific
- Hypothesis-based

- **Powerful**
- **Deceptively expensive**

Problem: The Peeking Problem

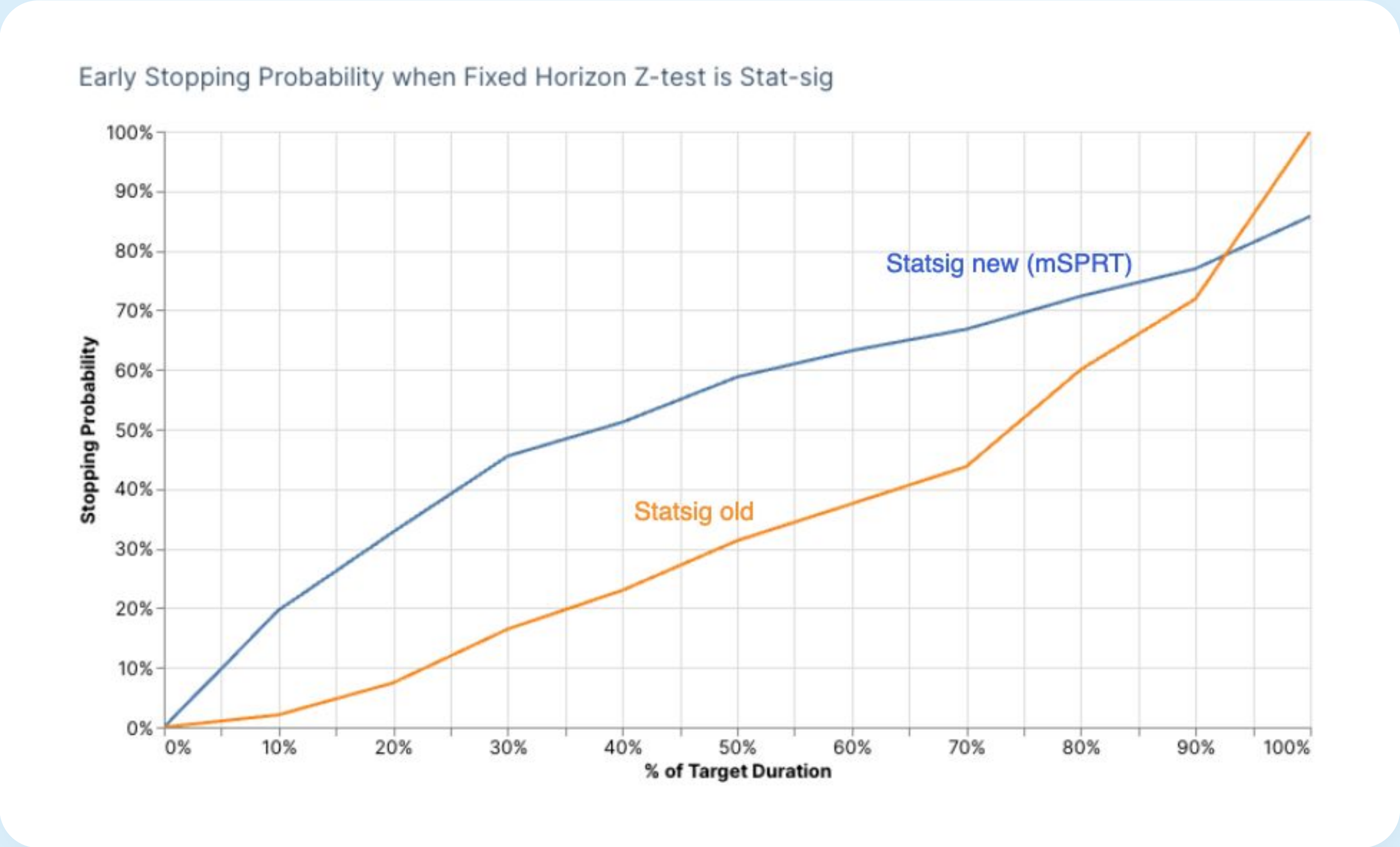


Solution: Sequential Testing

Tradeoffs

- Statistical Power
- Sensitivity
- Speed

What about multiple metrics?



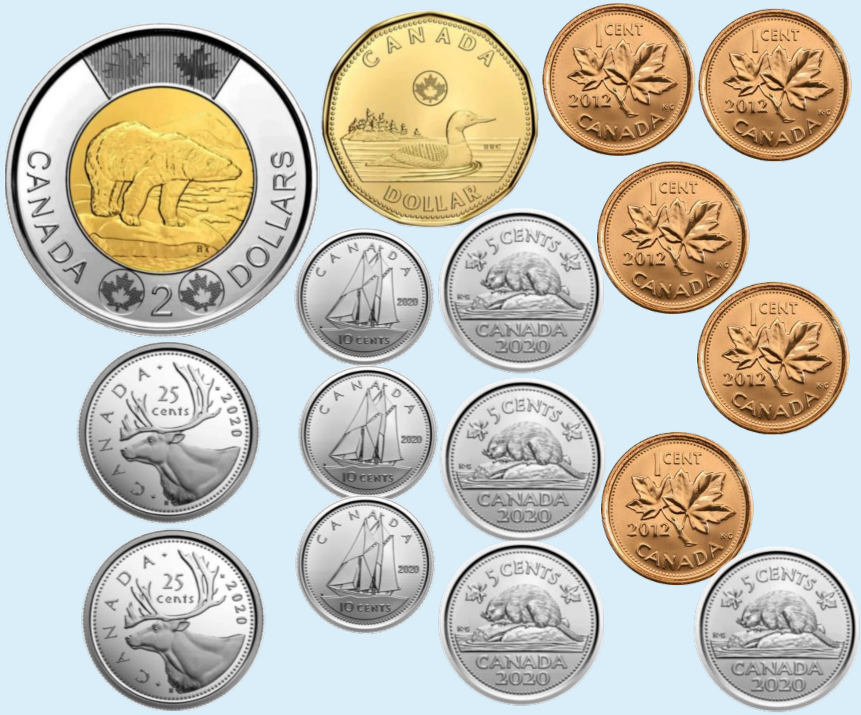
Maggie Stewart https://www.statsig.com/blog/sequential-testing-on-statsig

! Problem: Randomization is Random

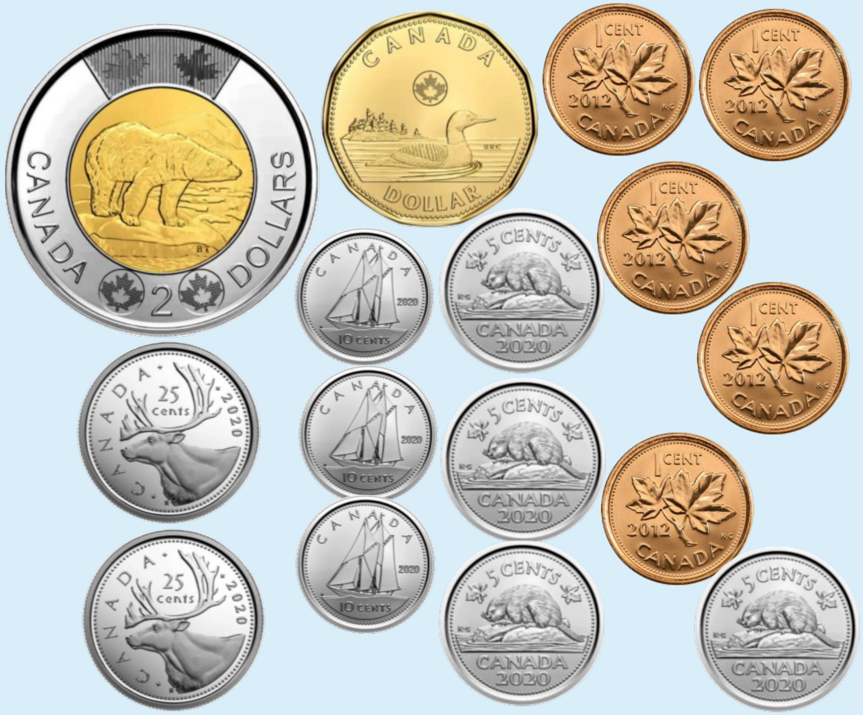
$5.78

$2.32

Solution: Stratified Sampling



$4.05



$4.05

✓ Solution: Stratified Sampling

B2B Experimentation

- High heterogeneity
  - High variance users, by orders of magnitude
  - Subgroups are important to track and compare
- Impact on whales are very important to accurately track
- Limited sample size

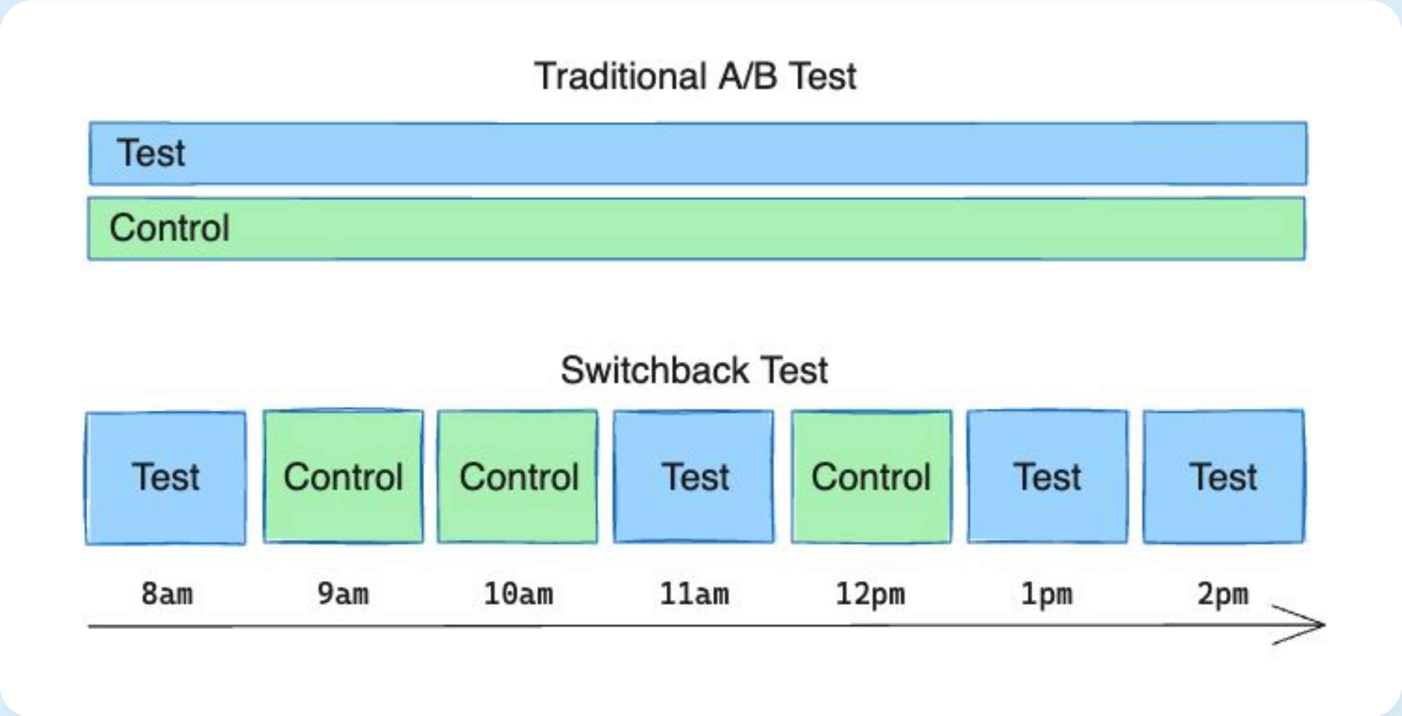
! Problem: Network Effects

Experimental groups can affect each other

- Eg. Social networks, two-sided marketplaces, messaging apps
  - Violation of independence assumption
- Cannot accurately measure individual impact of change, nor project total impact.

Solution: Switchback Tests

- Testing the entire network, by switching states over different time periods.
- Interval Selection is critical
- Assumes long-term impact and residual effects are minimal.



! Problem: Fixed Allocation

Learning can be expensive—Experiments take awhile to reach “certainty”

Inferior options are given equal traffic for a lengthy period

More variants markedly impact statistical power and experiment duration

Non-stationary effects

Examples

- Holiday Sale periods
- Non-durable goods (eg. news)
- Low statistical power

✓ Solution: Multi-Armed Bandit

Pros

- Automated decision making
- Good in situations with multiple options
- Great at eliminating “bad” options

Cons

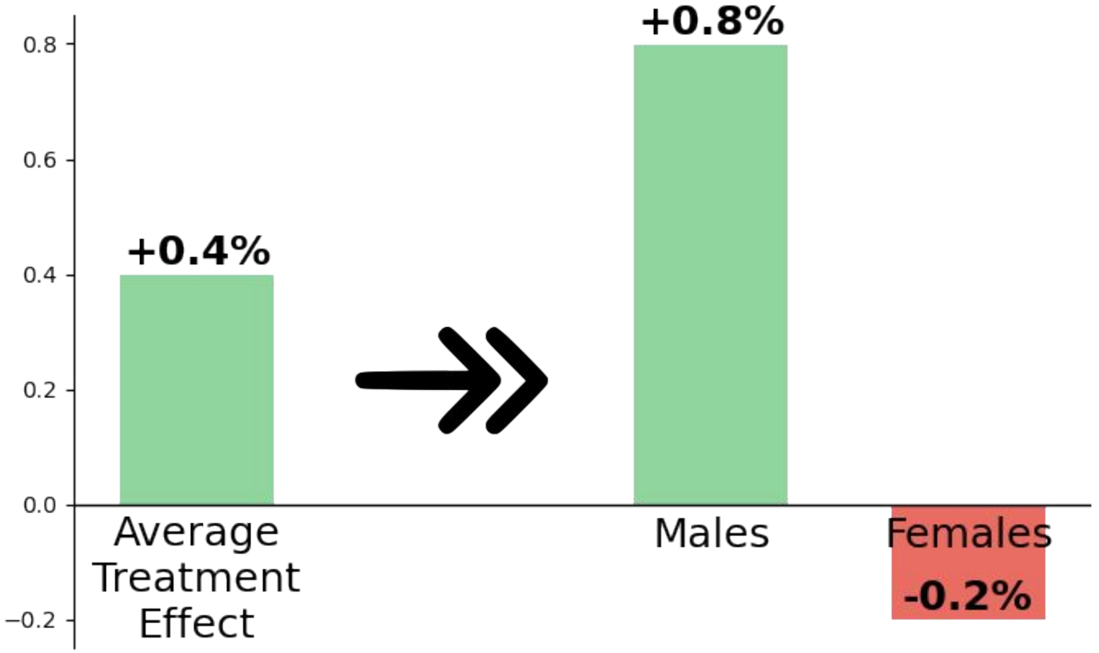
- Learning opportunities are limited
- Cannot handle nuanced decision-making

Heterogeneous Treatment Effects

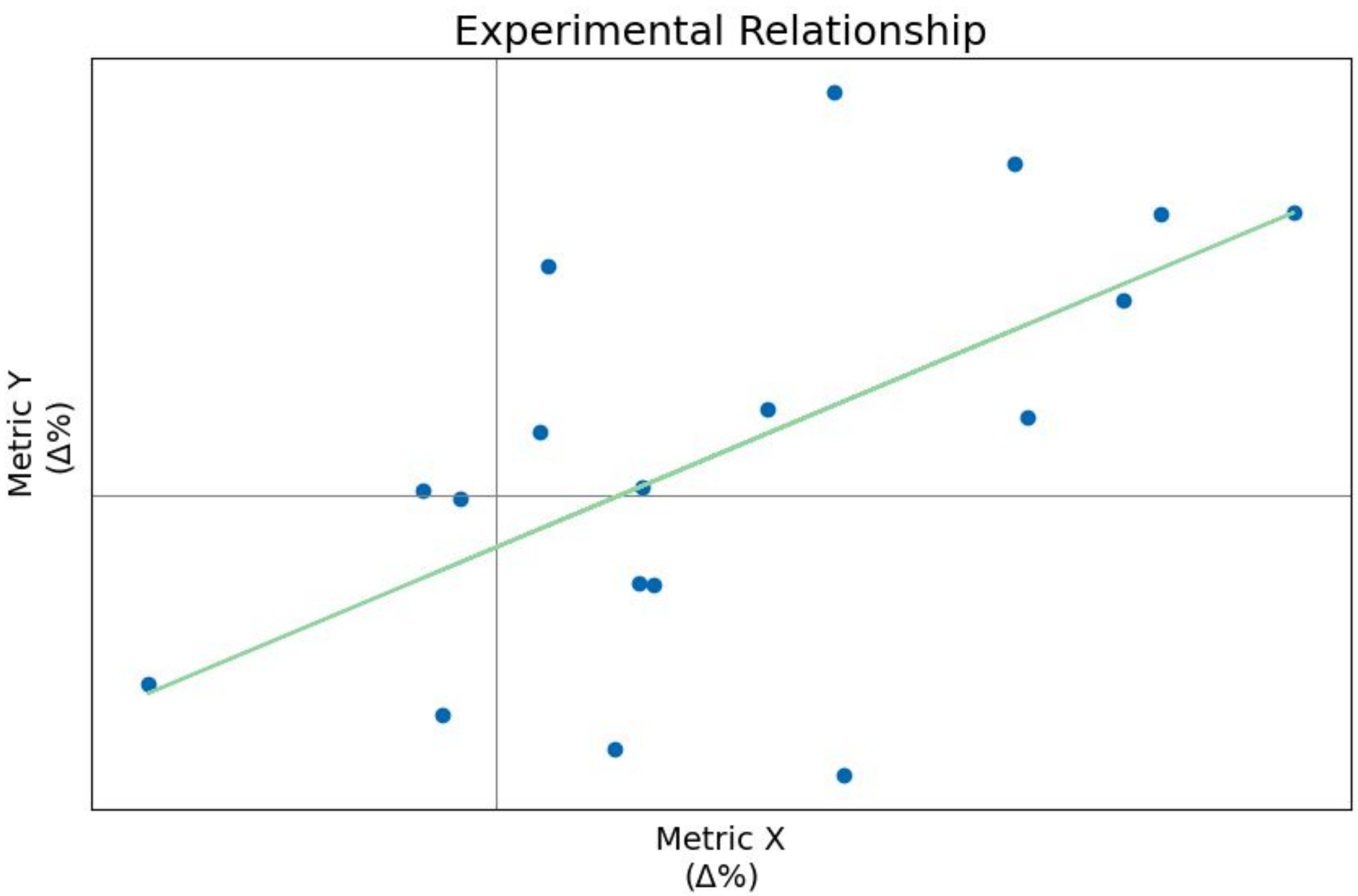
Average Treatment Effect vs Heterogeneous Treatment Effects

Detection

- Hypothesis-driven
- Automation across multiple attributes



Experimental Meta Analysis



Conclusion

Limitations		Solution
Experiments take too long	➢	CUPED
Winner’s Curse	➢	Holdouts
Peeking Problem	➢	Sequential Testing
Randomization Sucks	➢	Stratified Sampling
Network Effects	➢	Switchback Testing
Fixed Allocation	➢	Multi Armed Bandits
No Average User	➢	Heterogeneous Effects Detection
Only Specific Findings	➢	Experimental Meta Analysis

STATSIG

Statsig.com

linkedin.com/in/trchan

tim@statsig.com

@trchan1

statsig.com.pets

Thank you

Randomization is the Secret Sauce

Soft White

R250 G250 B250 | #FAFAFA
C2 M2 Y2 K0 | Paper color

Navy Black

R27 G37 B40 | #1B2528
C80 M65 Y60 K70 | PMS2965 C

Growth Gradient

Lime + Light Blue

Lime

R179 G251 B199 | #CCFBC7
C28 M0 Y32 K0 | PMS365 C

Light Blue

R217 G238 B249 | #D9EEF9
C13 M1 Y1 K0 | PMS657 C

Blue

R157 G213 B242 | #9DD5F2
C35 M3 Y1 K0 | PMS277 C

Dark Blue

R0 G104 B179 | #0068B3
C90 M60 Y1 K0 | PMS3506 C

Green

R144 G212 B158 | #90D49E
C44 M0 Y50 K0 | PMS360 C

Dark Green

R22 G65 B57 | #164139
C86 M50 Y69 K51 | PMS3435 C

Lavender

R189 G189 B255 | #BDBDFF
C24 M23 Y0 K0 | PMS2635 C

Dark Lavender

R108 G108 B188 | #6C6CBC
C64 M60 Y0 K0 | PMS2095 C

Yellow

R255 G248 B186 | #FFF8BA
C1 M0 Y33 K0 | PMS0131 C

Orange

R254 G173 B114 | #FEAD72
C0 M42 Y66 K0 | PMS1375 C