Emily Robinson
@robinson_es

A/B Testing in the Wild

Disclaimer:
This talk represents my own views, not those of Etsy

Overview

INTRODUCTION

Etsy

A/B Testing

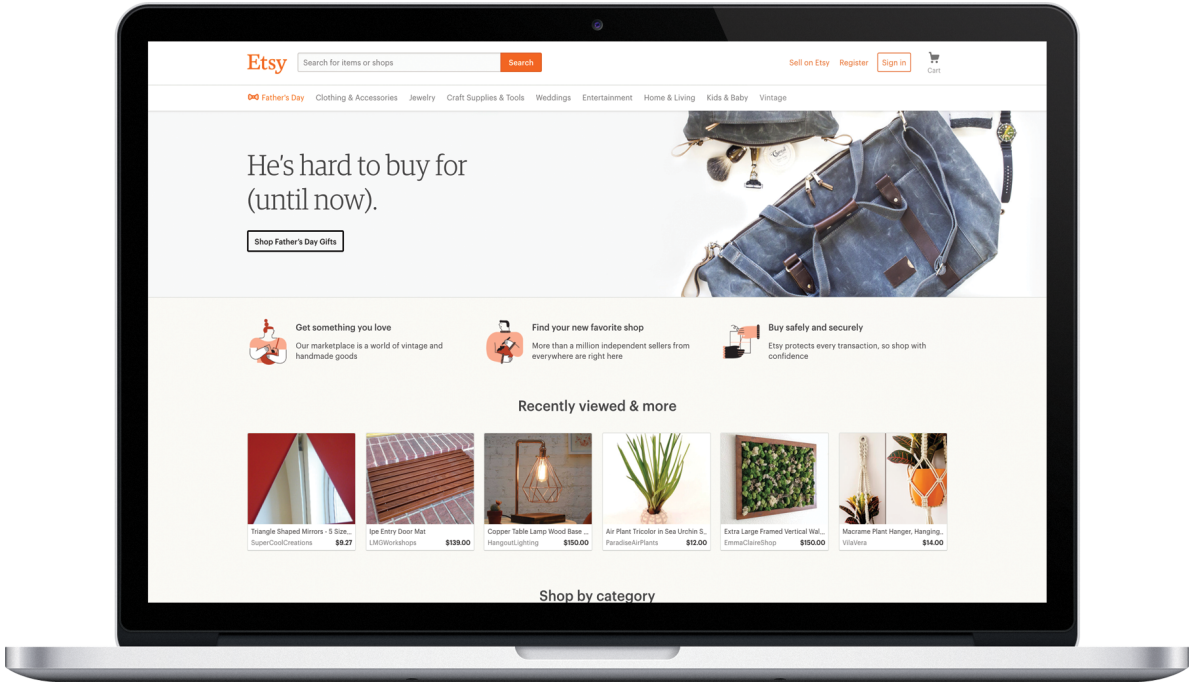
CHALLENGES & LESSONS

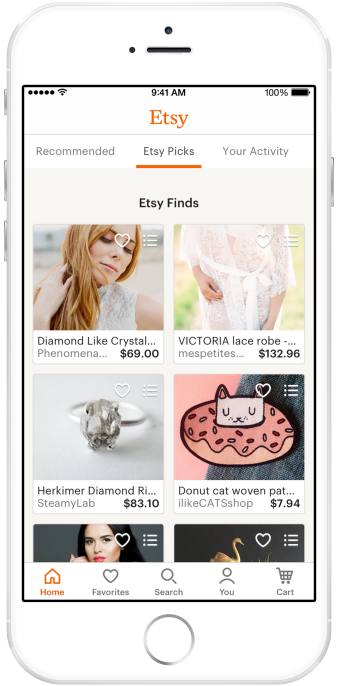
Statistical

Business

Etsy

Etsy is a global creative commerce platform. We build markets, services and economic opportunity for creative entrepreneurs.





Our Items

5 oz. Unicorn Shimmer Silly Pu...

MonstrousThings

★★★★★ (580)

$5.00

Groomsmen Gifts - Leather Toi...

FelixStreetStudio

★★★★★ (6,327)

$65.00

Antique Engagement Ring - An...

TheCopperCanary

★★★★★ (122)

$35,000.00

By The Numbers

1.8M
active sellers
AS OF MARCH 31, 2017

29.7M
active buyers
AS OF MARCH 31, 2017

$2.84B
annual GMS
IN 2016

45+M
items for sale
AS OF MARCH 31, 2017

Photo by Kirsty-Lyn Jameson

A/B Testing

What is A/B Testing?

Scotty Dog Green Pendant - Original Illustration - Vintage Inspired Silver Charm with Ball Chain Necklace

By birdyandbeedesi...

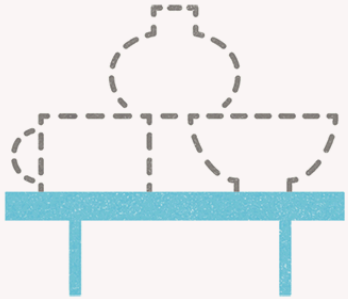
$10.00 Add to cart

Scotty Dog Green Pendant - Original Illustration - Vintage Inspired Silver Charm with Ball Chain Necklace

By birdyandbeedesi...

$10.00 Add to cart

Old Experience



We couldn't find any results for *small cute sneakers for puppies*

Try one of these instead

cute sneakers puppies | small cute puppies | small cute sneakers | small sneakers puppies

New Feature

Showing results for **cute sneakers for puppies.**

small cute sneakers for puppies had no results.

Search instead: small cute sneakers sneakers for puppies cute sneakers sneakers

All categories › **“cute sneakers for puppies”** (15 Results)

Cute Pug Shoes - Fun Shoes P...
MrsCopyCat
★★★★★ (8)
$69.00

Cute, Puppy with sneakers, Do...
TSoriginals
★★★★★ (550)
$11.99

101 Dalmations Converse STYL...
Stompingwithpride
★★★★★ (1)
$47.32

A/B Testing: It's Everywhere

Highly Researched

From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks

What works in e-commerce - a meta-analysis of 6700 online experiments

Overlapping Experiment Infrastructure: More, Better, Faster Experimentation

Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained

Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, Ya Xu

Microsoft, One Microsoft Way, Redmond, WA 98052
{ronnyk, alexdeng, brianfra, rogerlon, towalker, yaxu}@microsoft.com

My Perspective

Millions of visitors daily

Data Engineering Pipeline Set-Up

Generating numbers is easy; generating numbers you should trust is hard!

Why Statistics Anyway?

- "Election surveys are done with a few thousand people"[1]
- Targeting small effects
  - A .5% change in conversion rate (e.g. 6% to 6.03%) on a high traffic page can be millions of dollars annually

[1]Online Experimentation at Microsoft

Example Experiment

Listing Card Experiment

Baseball Hat with Embroidered Patch - Your Choice of Patch and h...
WildflowerandCompany
$32.00

Baseball Hat with Embroidered Patch - Your Choice of Patch & Cap Color! Dad Hat - Wildflower + Co.
WildflowerandCompany
★★★★★ (123 ratings)
$32.00

Result

Baseball Hat with Embroidered Patch - Your Choice of Patch and h...
WildflowerandCompany $32.00

Baseball Hat with Embroidered Patch - Your Choice of Patch & Cap Color! Dad Hat - Wildflower + Co.
WildflowerandCompany
★★★★★ (123 ratings)
$32.00

👎

Listing Card Experiment: Redux

Baseball Hat with Embroidered Patch - Your Choice of Patch & Cap Color! Dad Hat - Wildflower + Co.

WildflowerandCompany

$32.00

👎

Baseball Hat with Embroidered Patch - Your Choice of Pa...

WildflowerandCompany

★★★★★ (123 ratings)

$32.00

👍 🎉 💯

Statistical Challenges

Level of Analysis

Visit:
activity by browser over a defined time period (30 minutes)

Browser:
cookie or device ID (for apps)

User:
Signed-in user ID

Browser vs Visit: An Example

Summer sale Anakin/Luke Sky...
TraywickDesigns
★★★★★ (476)
$34.00

Vader Custom Lightsaber - wit...
ACLightsabers
★★★★★ (20)
$109.00

Ahsoka Tano Rebels Katana an...
GreyJediSabers
★★★★★ (2)
$2,999.00

Jedi lightsaber box
WoodenPretties
★★★★★ (146)
$45.00

Hasbro Black Series Force FX L...
TheSaberSupply
★★★★★ (62)
$36.75

SALE // KOTOR Lightsaber Wit...
GeekyArtifacts
★★★★★ (41)
$138.58

Thai SaberCraft Redemption v...
ThaiSaberCraft
★★★★★ (14)
$194.99

Custom Lightsaber
SaltLakeSaberCo
★★★★★ (710)
$130.00

Next Day

SALE // KOTOR Lightsaber Wit...

GeekyArtifacts

★★★★★ (41)

$138.58

Pros and Cons

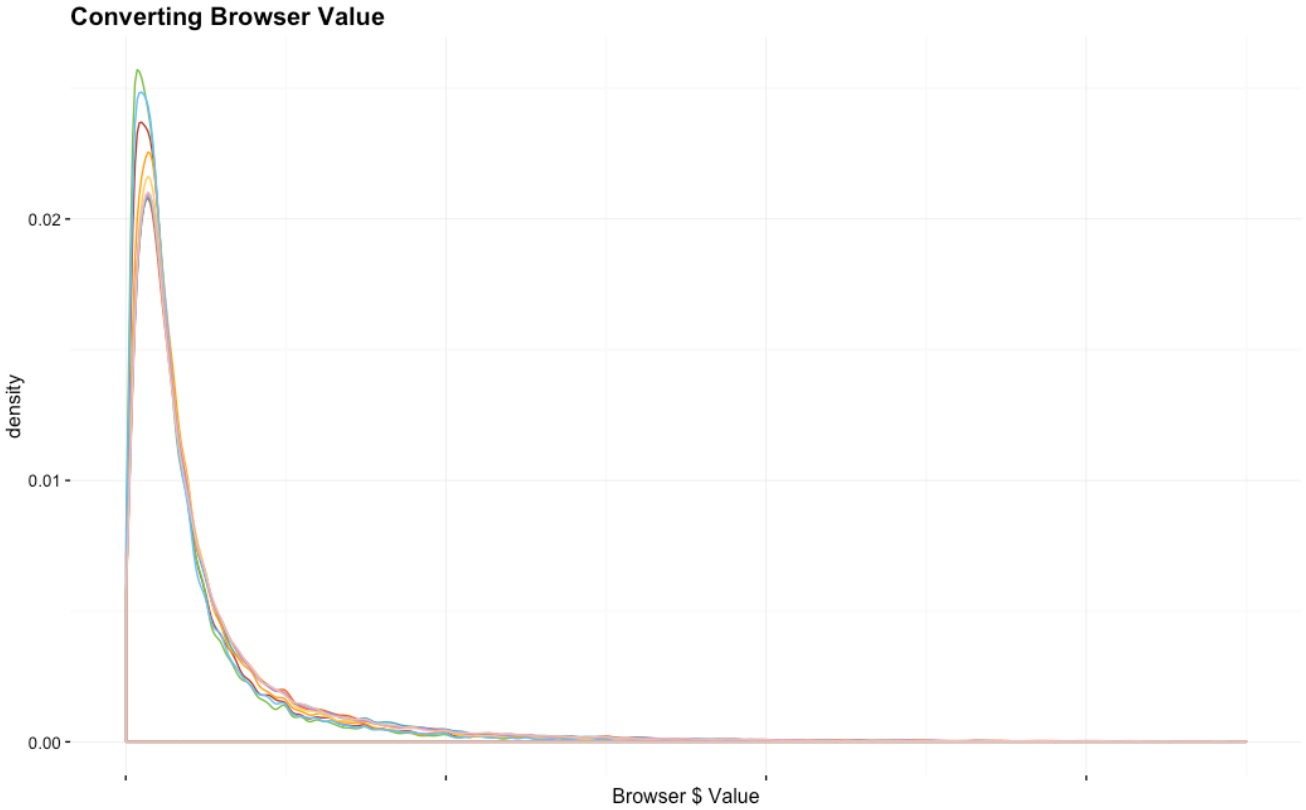
Visit	Browser
Tighter attribution	Captures relevant later behavior
Independence violation assumption	Introduces noise
Cannibalization potential	Misses multiple events for proportion metrics

Our conclusion: offer both, browser generally better

GMS per User

- Generally this is key metric
- But it’s a very badly behaved distribution
  - Highly skewed and strictly non-negative: can’t use t-test
  - Many zeros: can’t log numbers

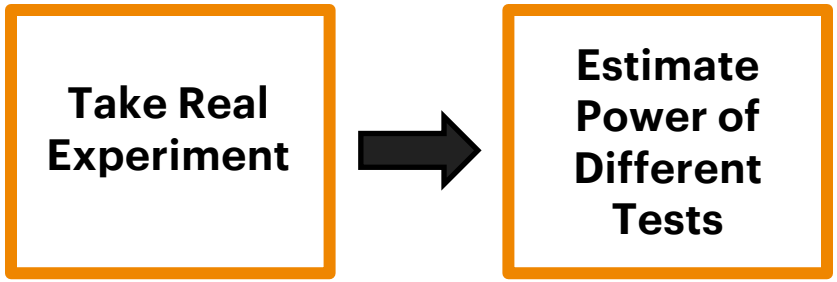
ACBV/ACVV


Converting Browser Value
0.02
0.01
0.00
density
Browser $ Value

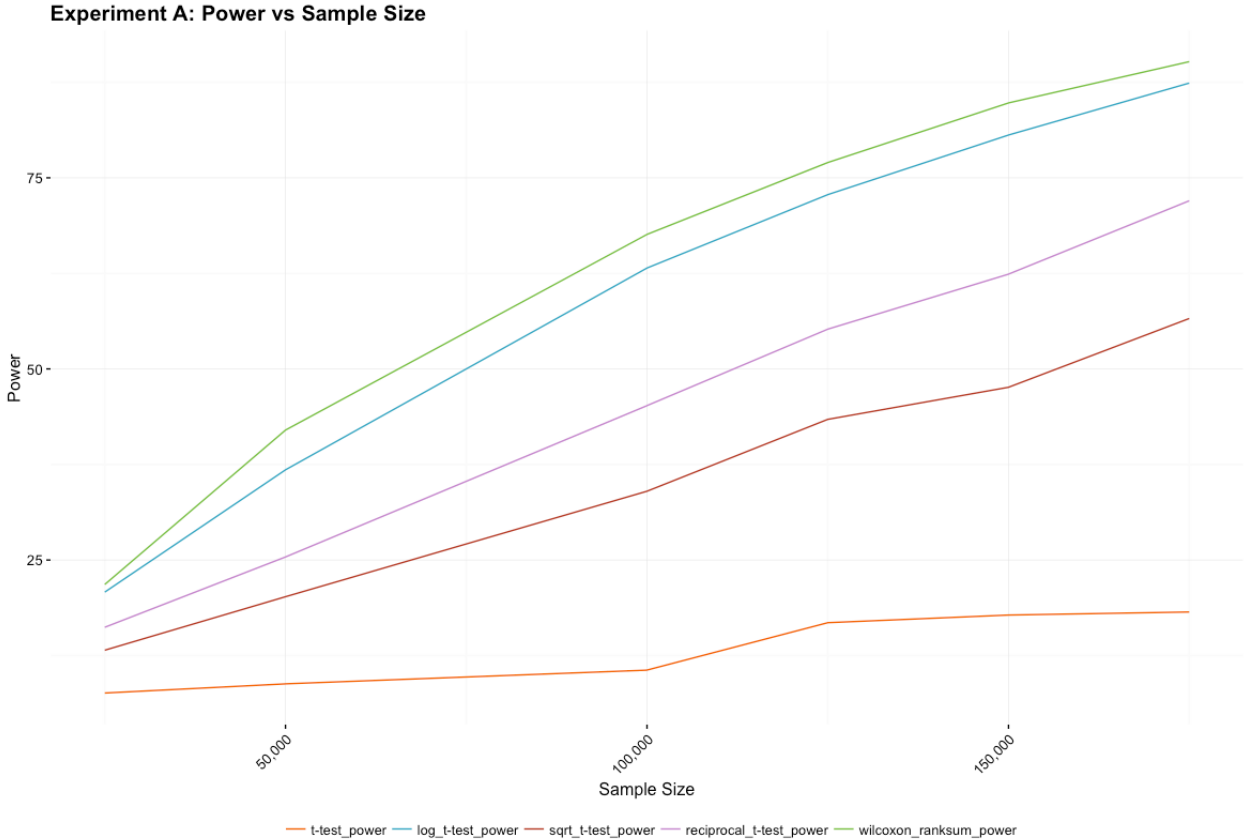
Definitions

- **Power:** Probability if there is an effect of a certain magnitude, we will detect it
- **Bootstrap:** random sampling with replacement
- **Simulation:** modeling random events

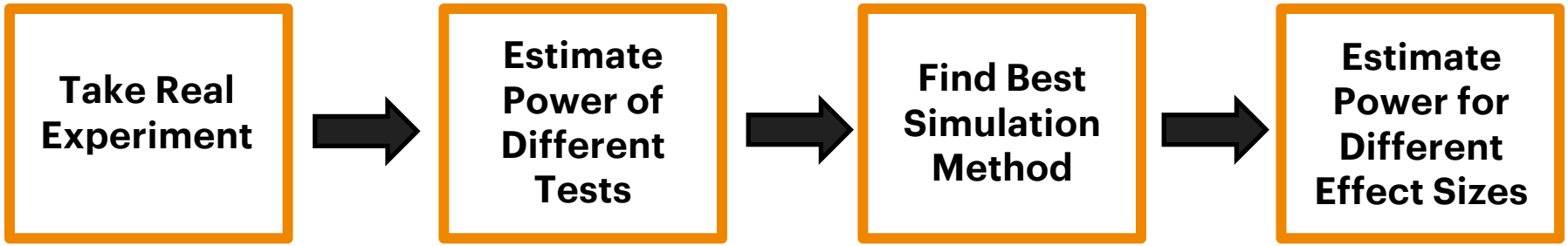
Test Selection Process



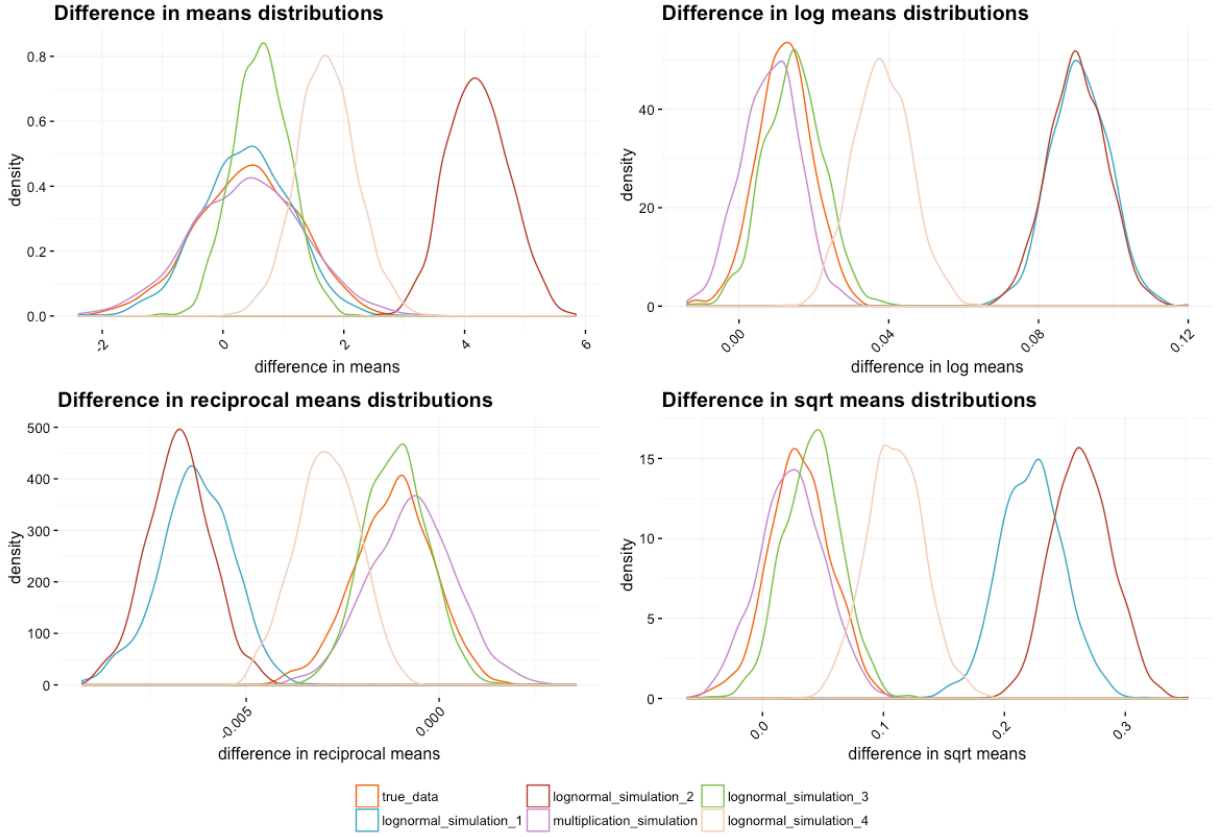
Estimating Power



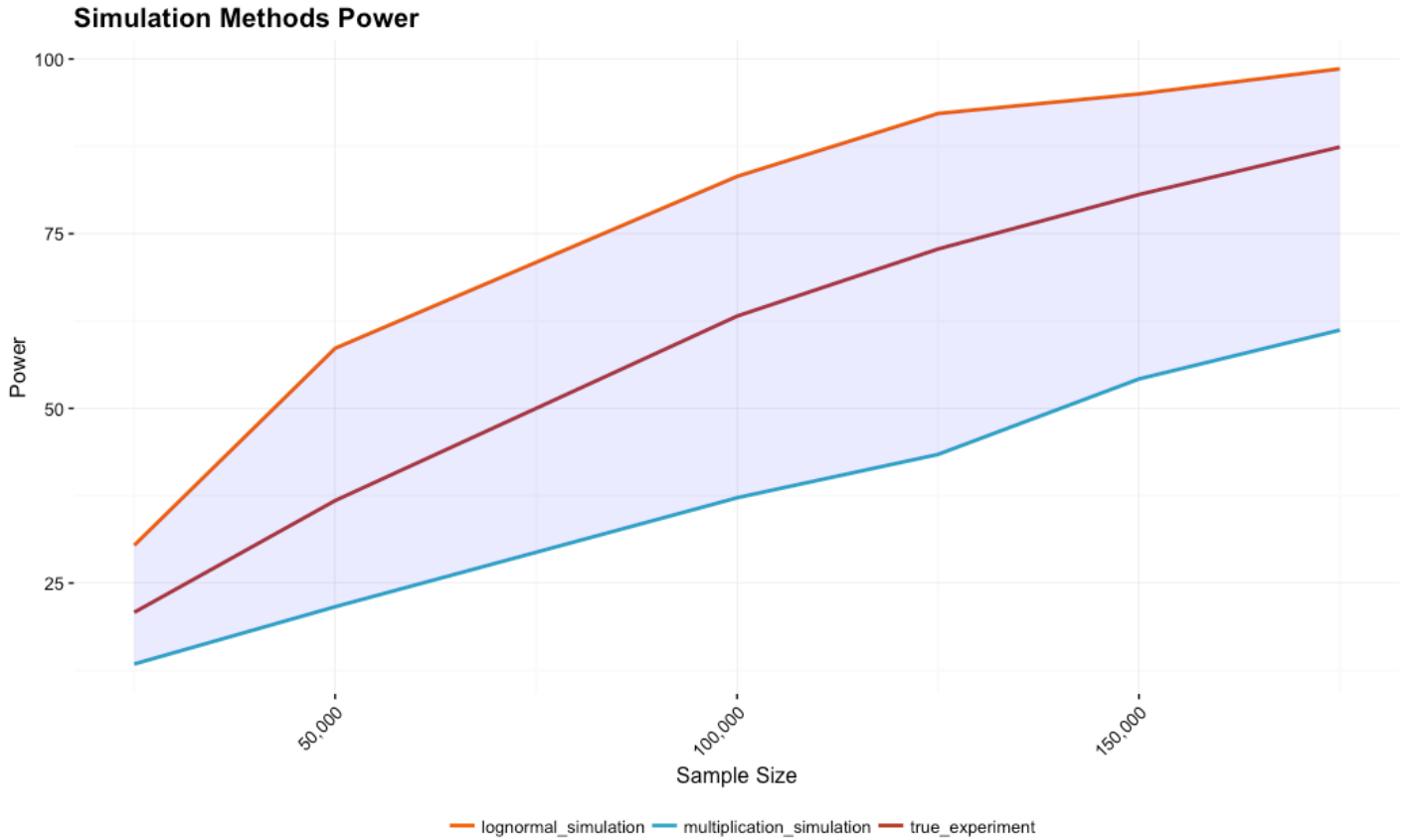
Test Selection Process



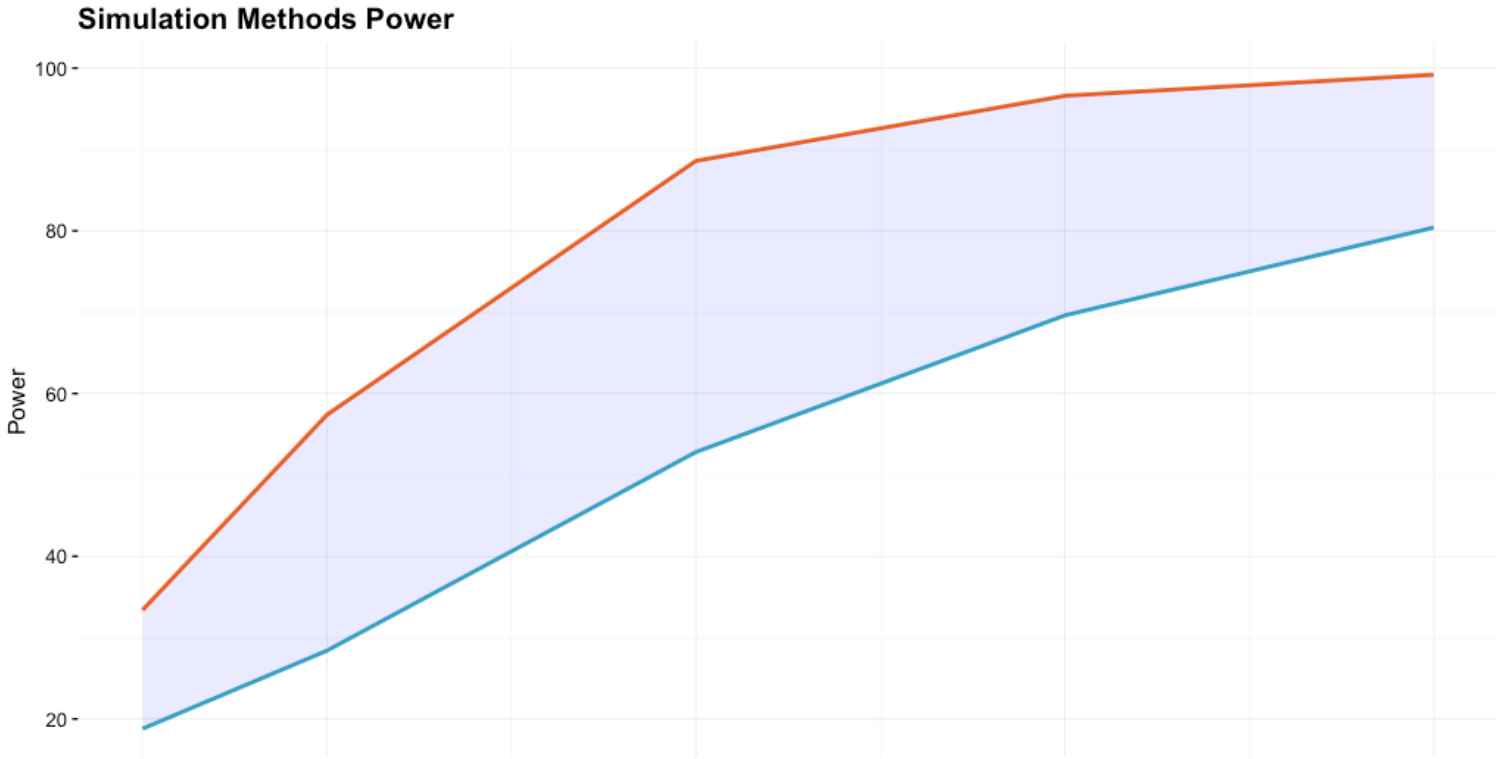
Simulation Method Comparison



Estimating Power



Power at 1% Increase in ACBV



Business Challenges

Working with Teams

Proactive Communication

Develop relationship:
Understand teammates

Demonstrate value:
Prioritization, feasibility, sequencing

Early involvement:
No post-mortems

Dealing with Adhoc Questions

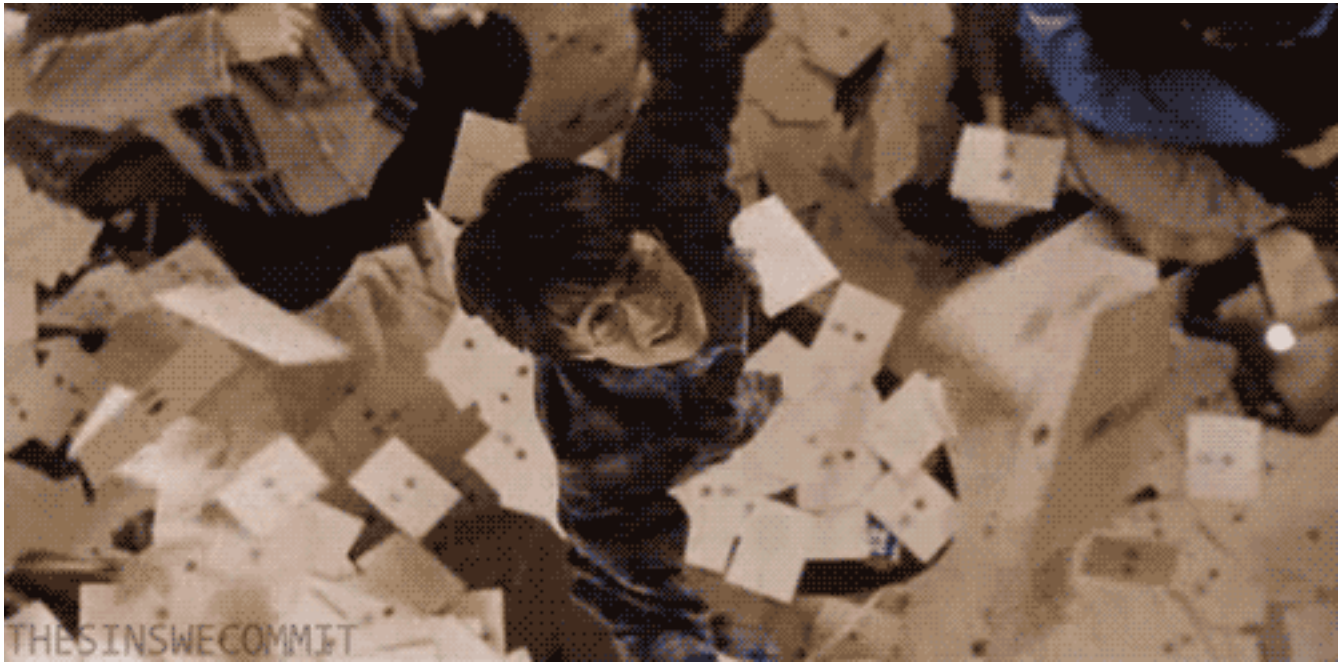
Question:

What's the conversion rate of visitors in Estonia on Saturday looking in the wedding category?

First Response:

What decision are you using this for?

Helps Avoid This



Checks Translation

Data Science Renee
@BecomingDataSci

Following

Replying to @IsabellaGhement

Yep, that's in a slide in a talk I give: business question -> data question -> data answer -> business answer 😊

9:39 AM - 25 Jun 2017

3 Retweets **2** Likes

1 3 2

We often joke that our job … is to tell our clients that their new baby is ugly

Business Partners & Experiments

- Financial and emotional investment
- Inaccurate expectations:
  - Features are built because team believes they're useful
  - But experiment success rate across industry is (sometimes far less) than 50%

Peeking

Question:

“What do the results mean?”

Answer:

“It’s been up for 15 minutes…”

David Robinson
@drob

Following

What I tell people: Don't peek at your A/B test early, it's meaningless.

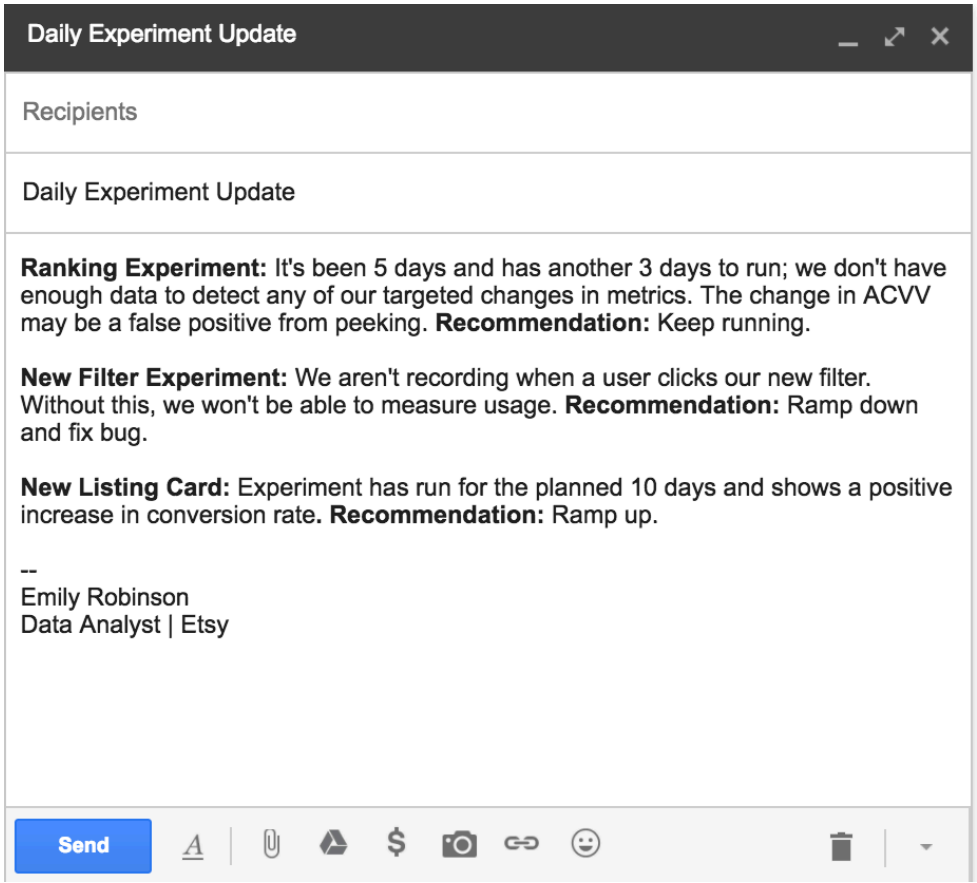
What I do: Come on, B! It's been twenty minutes, you can beat A!

1:58 PM - 29 Feb 2016

93 Retweets **144** Likes

4 93 144

Daily Experiment Updates

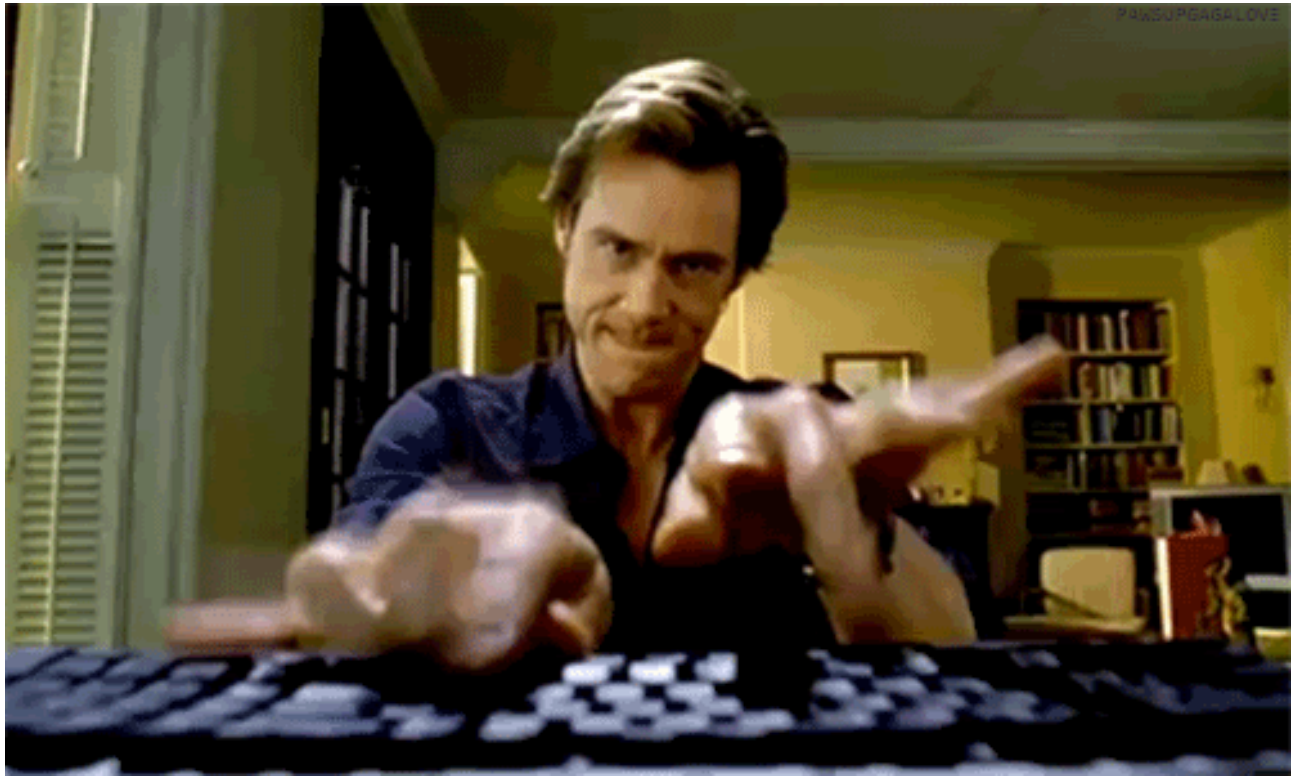


Offers Interpretation

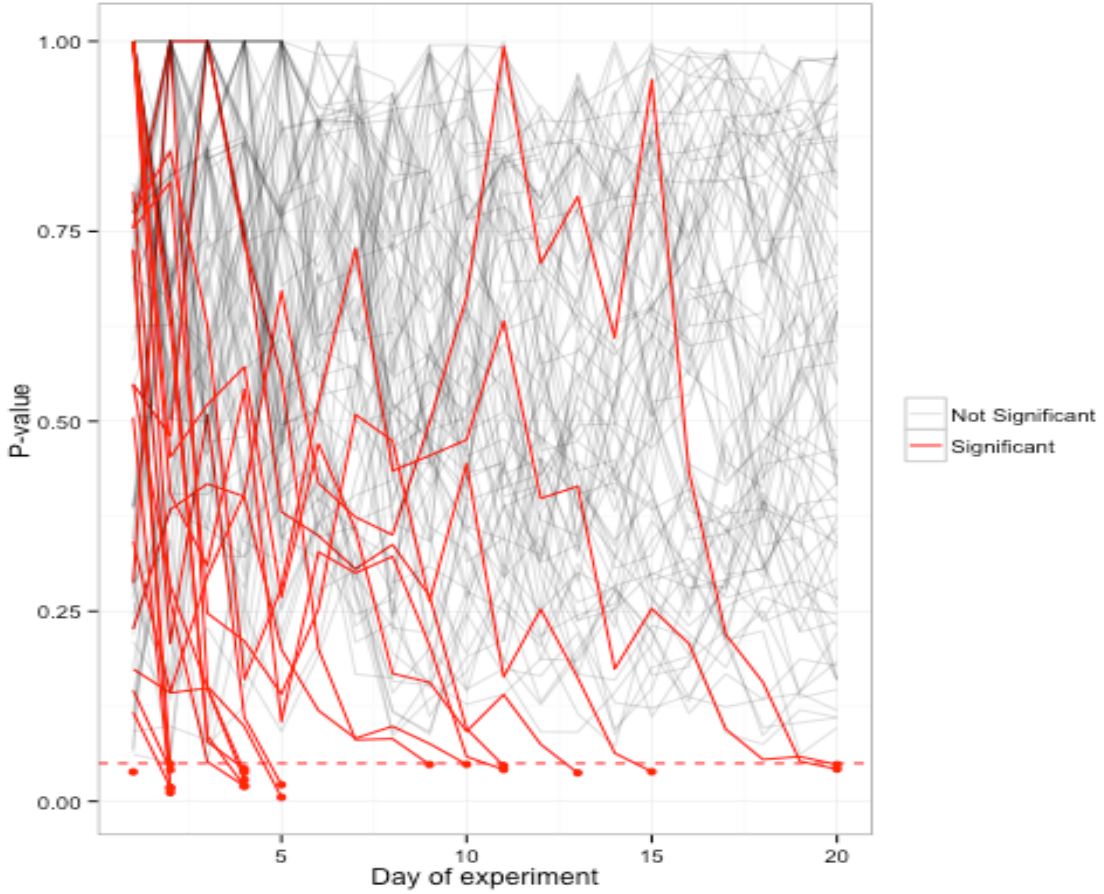
Shows You're Monitoring

*This is a Made-up Example

Want Fast Decision Making



Cost of Peeking: 5% FPR to 20%!



Solution 1: Adjust P-Value Threshold

Easy to Interpret

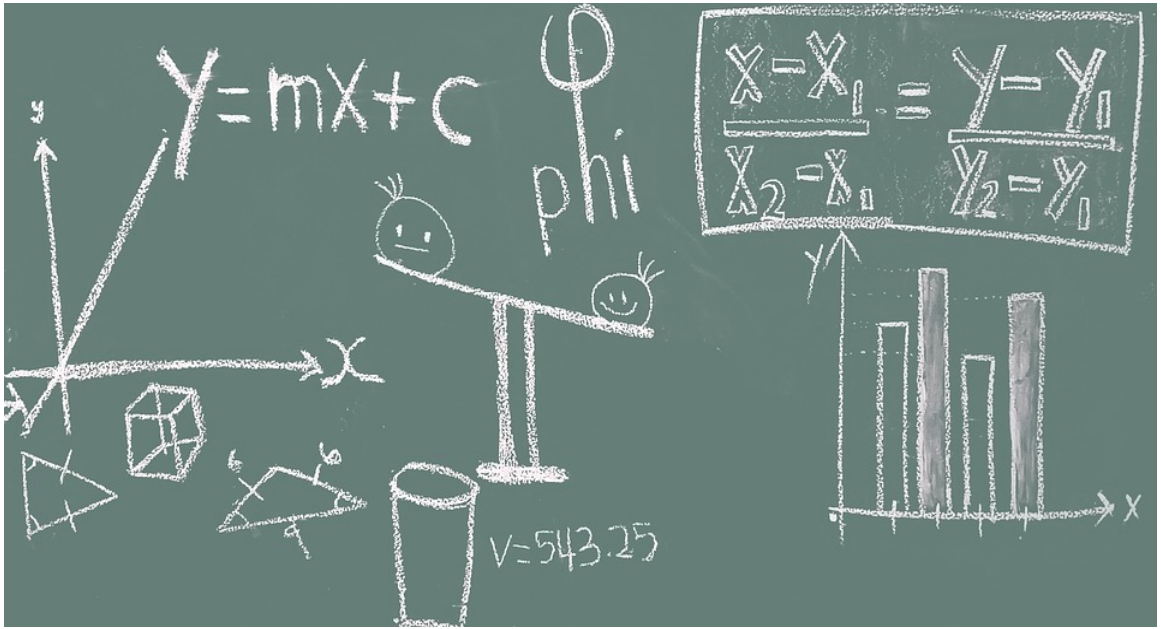
Not Rigorous

Solution 2: “Outlaw” Peeking

Correct Way

Miss Bugs

Solution 3: Continuous Monitoring



Peek and Stay Rigorous

Complicated to Implement & Explain

And at the End of the Day …

The more programming I do, the more thing I run into where

- I don't know
- Google doesn't know
- my colleagues don't know
- we gotta do it anyway

When this happens, I think "Right, this is why they pay a human with a brain who can investigate and learn"

From Julia Evans, @b0rk
"How to be a Wizard Programmer"

Resources

- Controlled Experiments on the Web: Survey and Practical Guide
- Overlapping Experiment Infrastructure: More, Better, Faster Experimentatio
- From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks
- What works in e-commerce – a meta-analysis of 6700 online experiments
- Online Controlled Experiments at Large-Scale
- Online Experimentation at Microsoft
- Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained

Acknowledgments

- Evan D'Agostini (for ACBV development & slides)
- Jack Perkins & Anastasia Erbe (former & fellow search analysts)
- Michael Berkowitz, Callie McRee, David Robinson, Bill Ulammandakh, & Dana Levin-Robinson (for presentation feedback)
- Etsy Analytics team
- Etsy Search UI & Search Ranking teams

Thank You

tiny.cc/abslides

robinsones.github.io

@robinson_es