A/B Testing Stories

Agnes Urbanics-Salanki

February 2020, CEU

Agenda

- My journey to analytics
- A/B testing stories

MY JOURNEY TO ANALYTICS

It started pretty early on

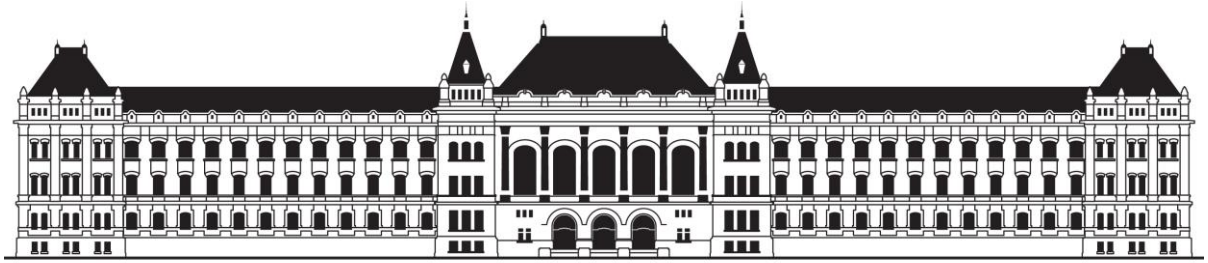
Computer scientist

me

Mathematics teacher

University studies

- BSc/MSc in computer engineering
  - Lots of maths subjects
  - Lots of programming
  - TAing, volunteer research
    - Straightforward transition into a PhD



MŰEGYETEM 1782

A few words about my PhD

- Research area: infrastructure analytics (data analysis on logs from distributed systems)

- Why was I a drop out?
  - Didn’t do my homework before starting
  - Laziness to learn the domain
  - Couldn’t accept that the focus is on the result and not methodology

Early career years

- Secret Sauce Partners
  - In-house engineering team
  - Cross-functional teams
  - 2 analysts
  - Strong statistical background
- Product Analytics
  - Ad-hoc investigations: early in the product, lots of low-hanging fruit (not always immediately actionable though)
  - A/B tests: not just the features but the product itself
  - Lots of iterations on the product

Early career years

- Hotels.com/Expedia Group
  - Almost 100 analysts
  - Out-sourced engineering team, quite a few partners
  - Global business, with lots of channels and competitors

- Product Analytics
  - Textbook A/B testing cases are rare, there is a lot of ‘hacking’
  - Presentation- and report-heavy

Lessons learned

MŰEGYETEM 1782

- Programming
- R ❤
- Internships
- Traveling

secret sauce PARTNERS

- Importance of QA and good engineers
- ‘business intuition’

Hotels.com™

- Stakeholder management
- ‘Executive summary’

Community: conferences, meetups

One more word about analysts vs data scientists

- Data Analyst
  - (DS in Product)
  - Number person
  - SQL, R
- Data Scientist
  - (Core DS, Research S)
  - Algorithm person
  - Python
- BI
  - Dashboard person
  - Tableau, Alteryx

A/B TESTING STORIES

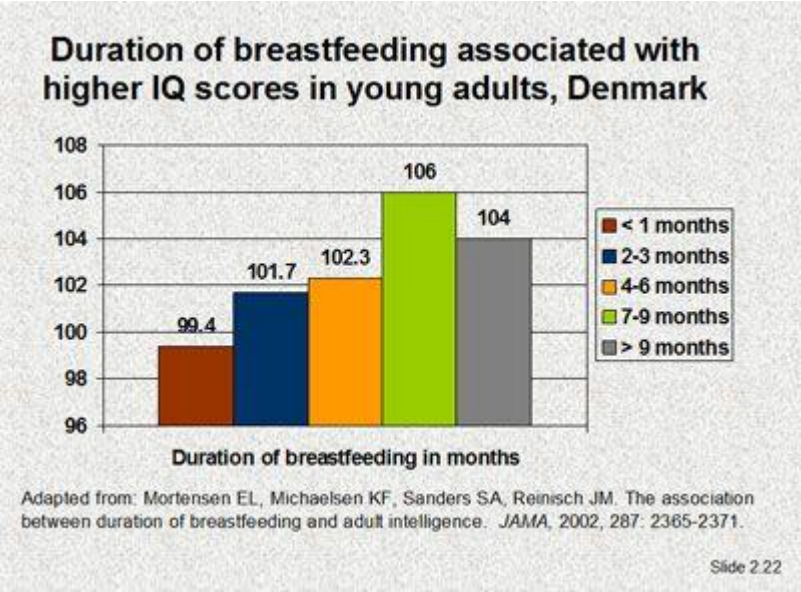
A/B tests

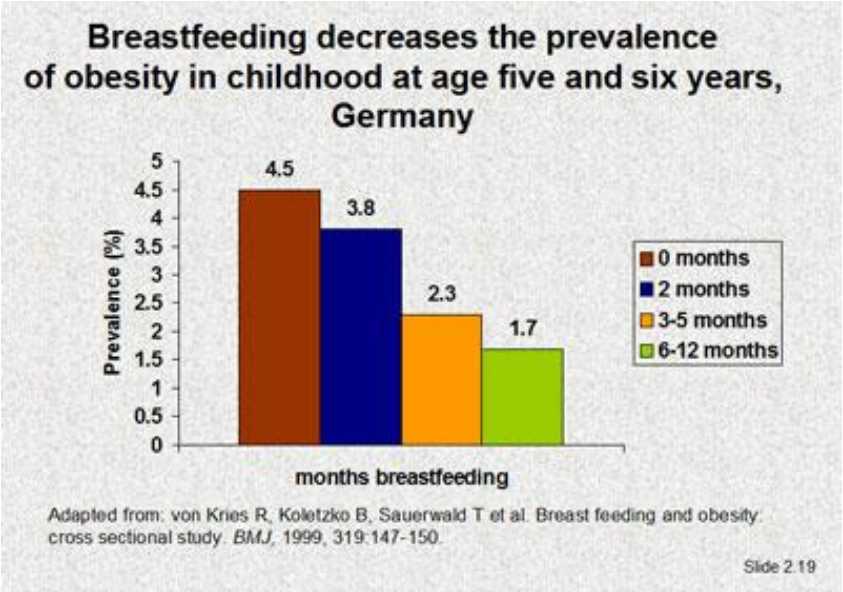
- Why A/B testing?
- A/B tests in product management
- A/B testing in theory
- Challenges

WHY TESTING?

Observational studies vs controlled experiments

- Breast is best!



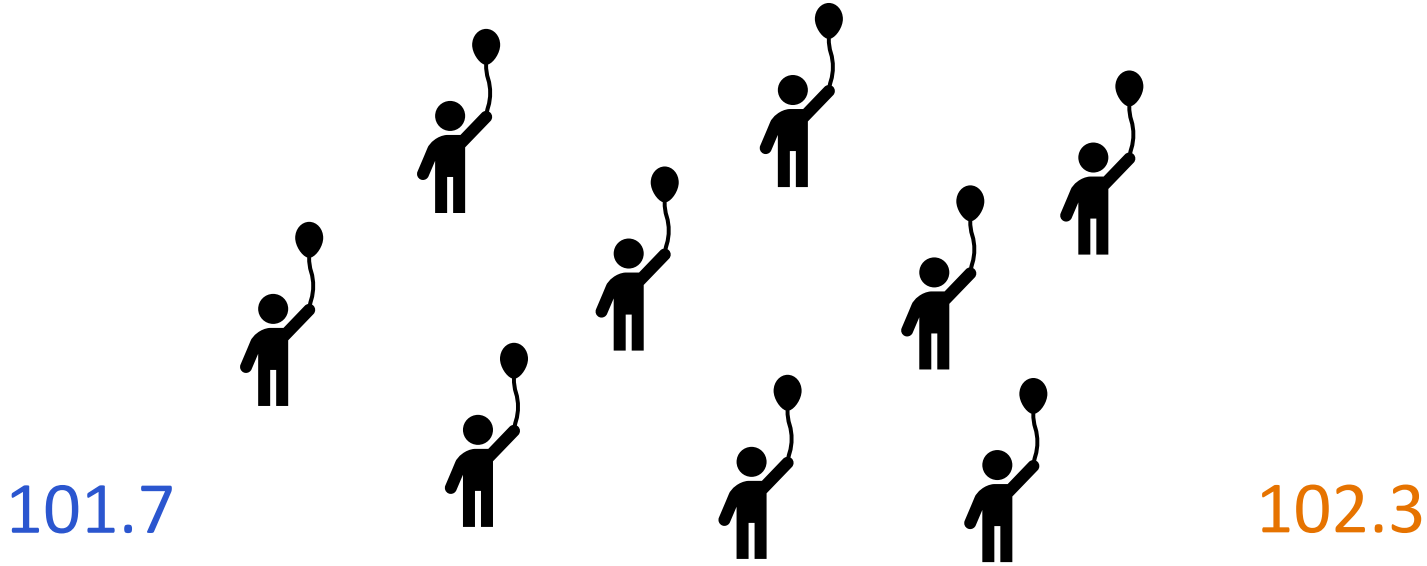


Breastfeeding guilt experienced by half of mothers - BBC survey

29 January 2019 Share

Source: Emily Oster: Cribsheet: A Data-Driven Guide to Better, More Relaxed Parenting, from Birth to Preschool

Observational studies vs controlled experiments



- Breastfed ↔higher IQ ✓
- Brea
- Higher IQ ←…→ Breastfed

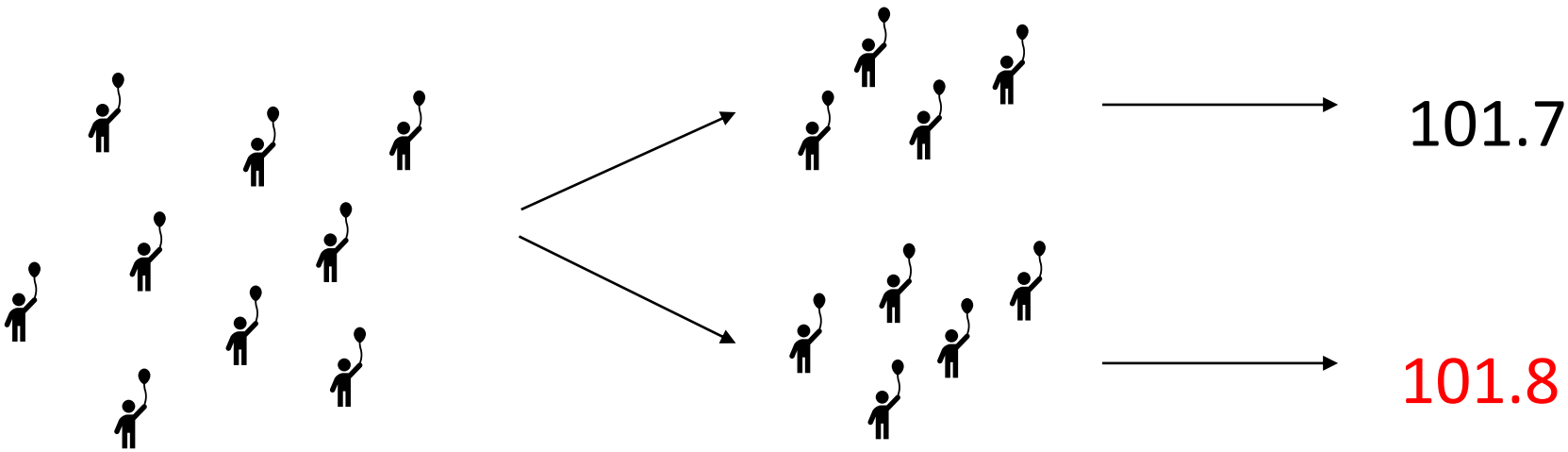
You cannot determine causality based on an observational study

Observational studies vs controlled experiments

- “The members of our loyalty program provide five times as much revenue (…) so we know the program is profitable” (Senior analyst at a conference)

- “Our NPS score increased in the last quarter, so we know our customers are happy with the change” (Analytics manager at a conference)

Observational studies vs controlled experiments



To say anything about causal relationships, you need a proper, random split
BEFORE
you would start the treatment

A/B TESTING IN PRODUCT MANAGEMENT

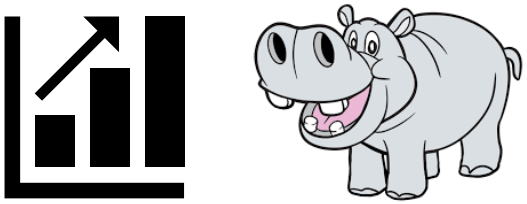
Hippos vs Testing

- **Hi**ghest **P**aid **P**erson's **O**pinion
  - Knows the customer
  - Knows the product
  - Knows the strategy
  - etc.

- Testing
  - Provides numbers
  - Objective
  - A draft is not enough

- Good old times

- Nowadays



Source: Kohavi, Longbotham, Sommerfield, Henne: Controlled experiments on the web: survey and practical guide

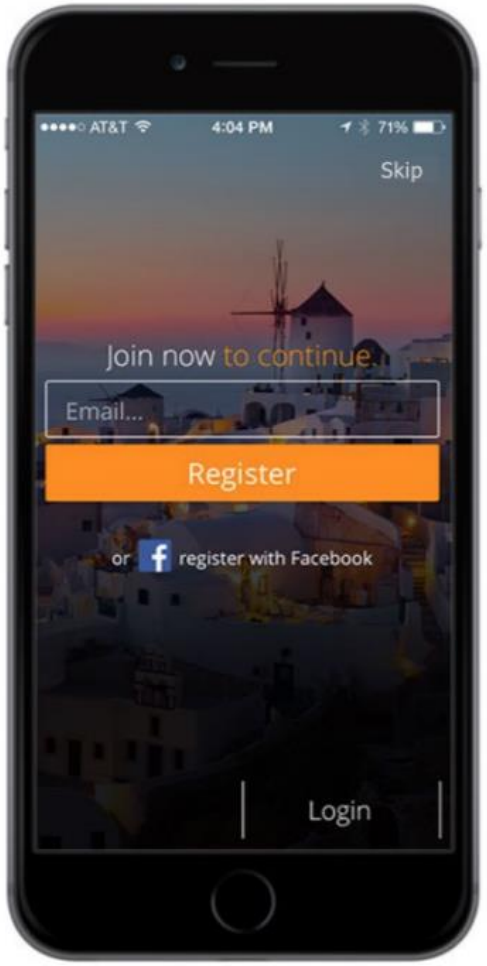
Testing in Product Management

- Online
  - Scales very well
  - Once your have the infrastructure, it is easy
- Lots of customers
  - Mix of different background/motivation etc.
- What to test?
  - UI
  - Algorithms
  - Infrastructure

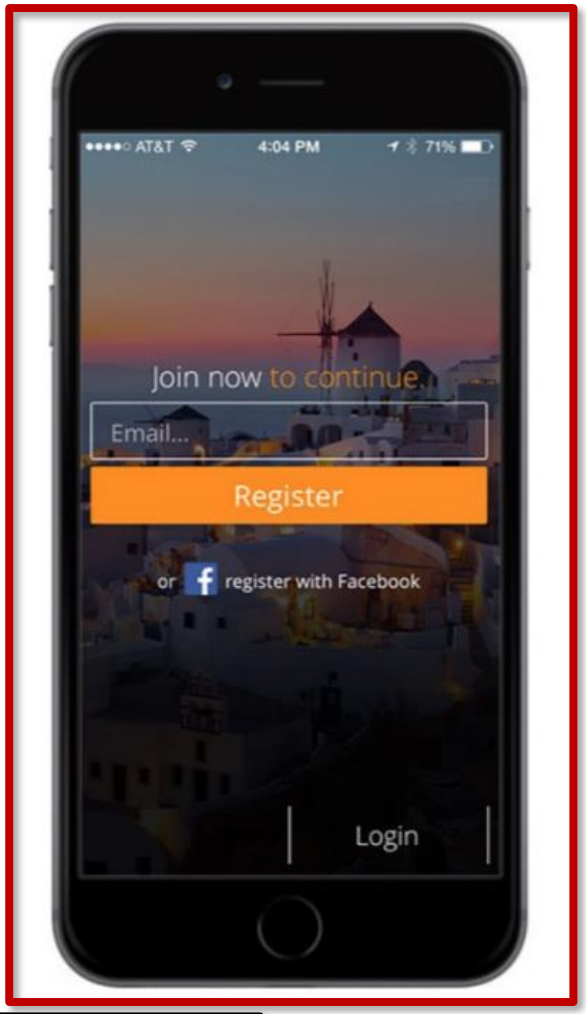
From now on, we will test features ON people

Secret Escapes Mobile Home Page

Variant A



Variant B



KPI: Revenue on app

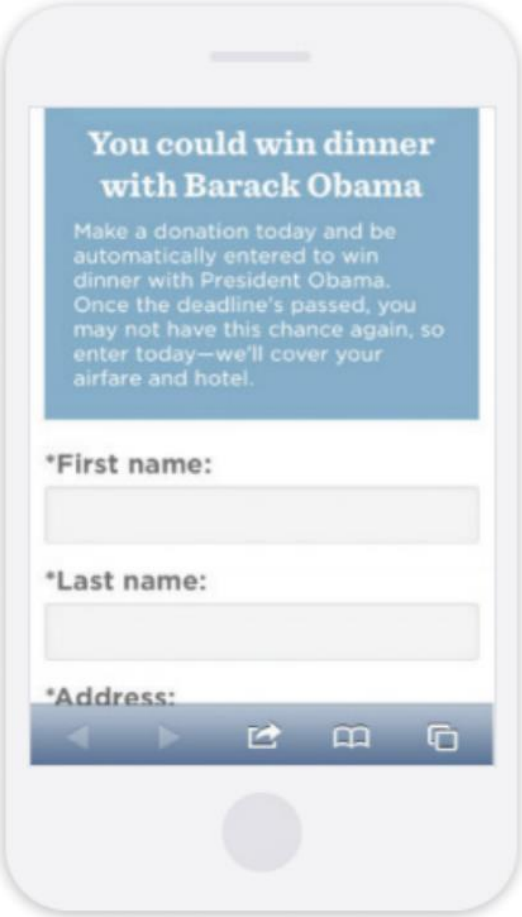
Source: Optimizely Experimentation case studies

https://www.optimizely.com/uk/resources/experimentation-case-studies/

Obama 2012 Campaign

Variant A

Variant B



KPI: Donation conversion

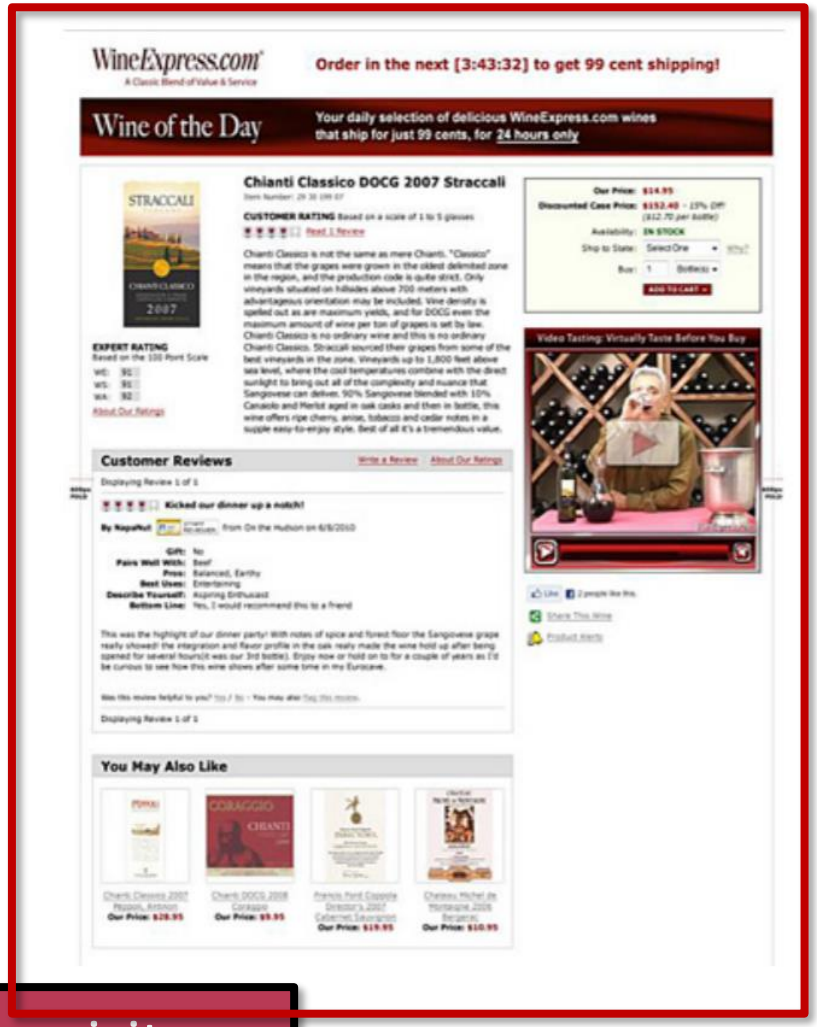
Source: Optimizely Experimentation case studies

https://www.optimizely.com/uk/resources/experimentation-case-studies/

WineExpress.com Wine of the Day

Variant A

Variant B

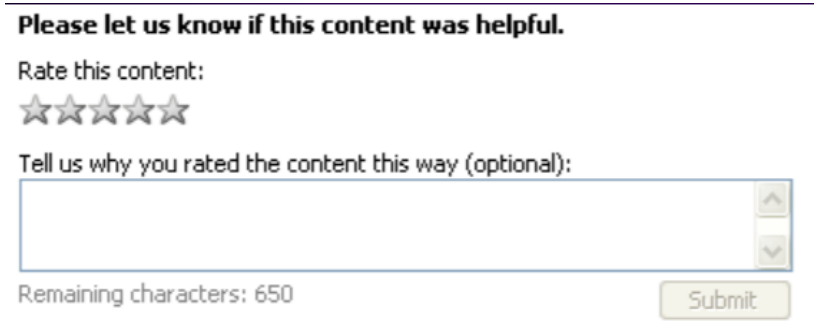


KPI: Revenue per visitor

Source: Croll, Yoskovitz: Lean Analytics: *Use Data to Build a Better Startup Faster*

Windows Office Online Feedback

Variant A



Variant B

KPI: Response rate

Source: https://exp-platform.com/Documents/2007-10EmetricsExperimenation.pdf

Criteo FB advertiser sign ups

Variant A

Variant B

KPI: Conversion

Source: How to get users grow talk by Alex Schultz https://www.youtube.com/watch?v=URiIsrdplbo

Statistics about statistics

- *„only one third of ideas tested at Microsoft improved the metrics they were designed to improve"*
- *"Netflix considers 90% of what they try to be wrong"*
- *"Google ran approximately 12k randomized experiments in 2009, with about 10 percent of these leading to business changes"*

Source: Kohavi, Longbotham, Sommerfield, Henne: Controlled experiments on the web: survey and practical guide

A/B TESTING IN THEORY

1. Split the audience
2. Choose a KPI
3. Calculate the difference
4. Make a decision

A/B TESTING IN PRACTICE

SPLIT THE AUDIENCE

Split the audience

- Person based (~cookie based if online)
- Basic checks to run?

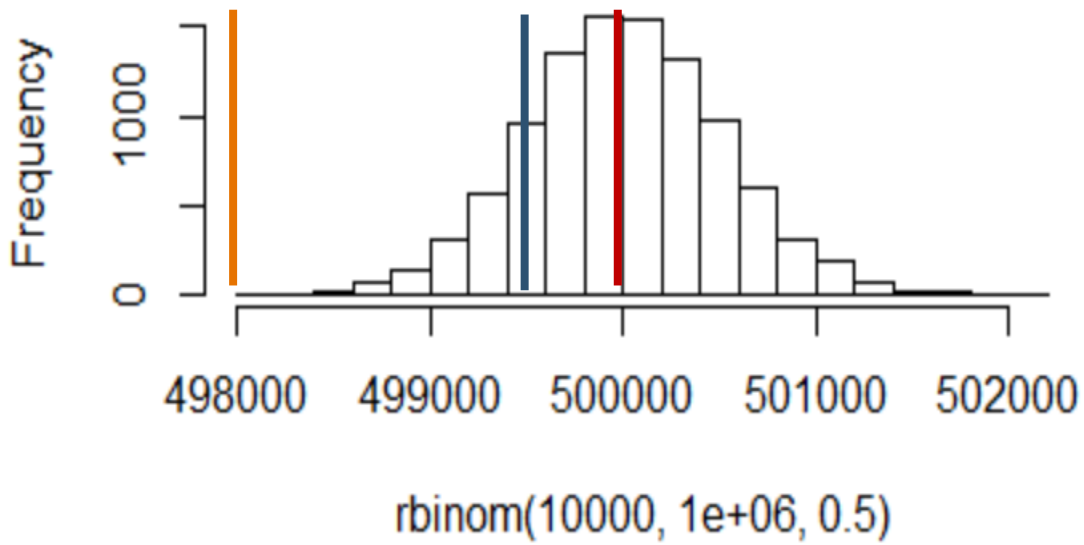
Split the audience

- Basic checks to run
  - Size of the two groups

Control: 500,000
Test: 500,000

Control: 499,500
Test: 500,500

Control: 498,000
Test: 502,000



Split the audience

- Basic checks to run
  - Size of the two groups
  - Control and test are disjunct subsets
    - Workaround suggestion: “could we simply remove these people?”

Scenario 1:
We would like to send out a single email welcoming subscribers who have recently joined our travel newsletter. We test the functionality in 3 different countries.
Check reveals that control and test are not disjunct.
Root cause: it turns out there are 100k customers who have subscribed in more than one country and they got mixed allocation.
What happens if we remove these people?

Split the audience

- Basic checks to run
  - Size of the two groups
  - Control and test are disjunct subsets
    - Workaround suggestion: “could we simply remove these people?”

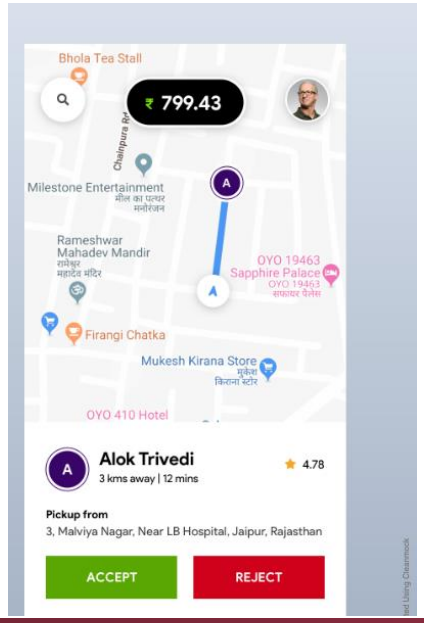
Scenario 2:
We would like to send out a reminder email to users who have reached the booking phase but has not completed their booking.
Check reveals that control and test are not disjunct.
Root cause: it turns out there are 100k customers who have come to the site multiple times and the testing framework reallocated them every time so they got mixed allocation.
What happens if we remove these people?

Split the audience

- Basic checks to run
  - Size of the two groups
  - Control and test are disjunct subsets
    - Workaround suggestion: “could we simply remove these people?”
    - What to check: do we remove the same quality customers from both?
  - Pre-period is comparable
    - Personal horror story: kept allocation from a previous test

Split the audience: challenges

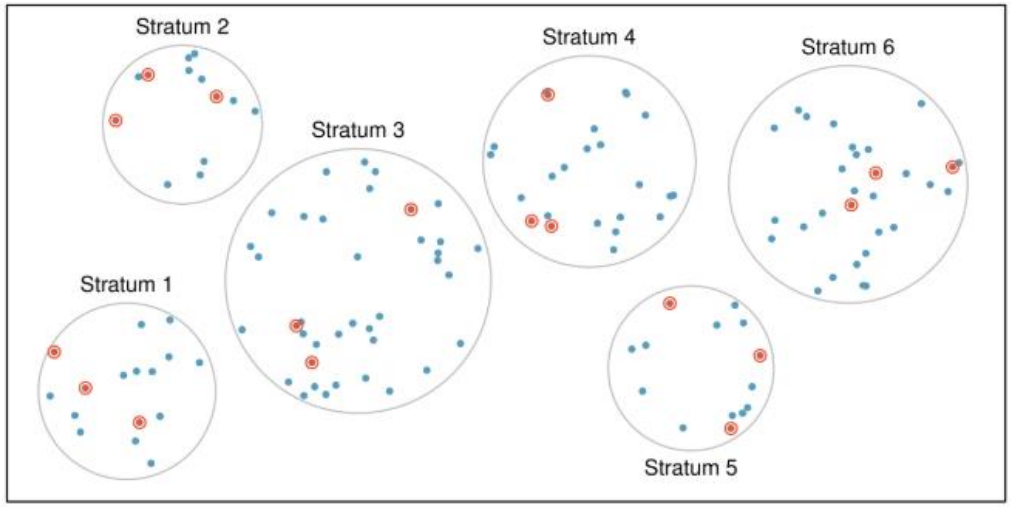
- People are not independent
  - E.g., your users are members in a network or otherwise related, e.g. in a marketplace
  - Workaround: special allocation algorithm, usually based on cluster sampling

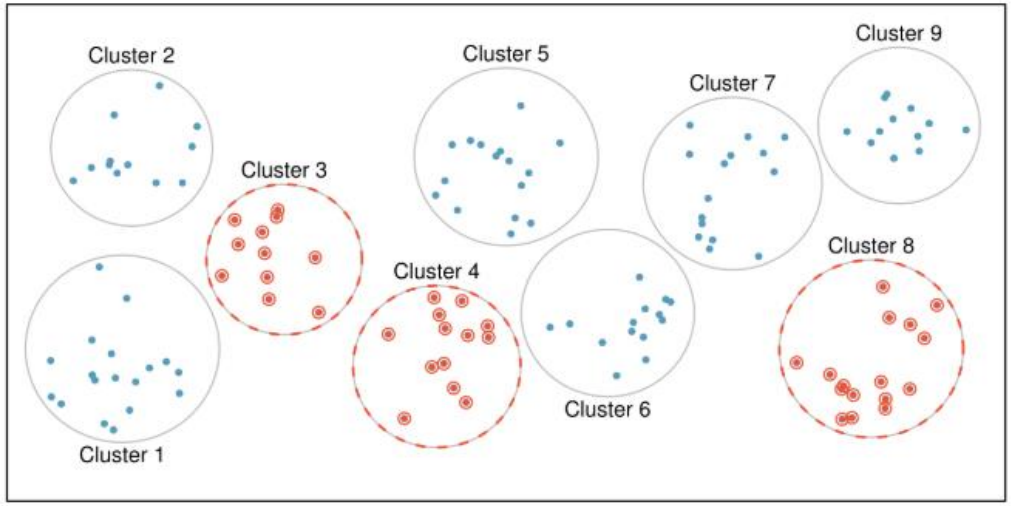


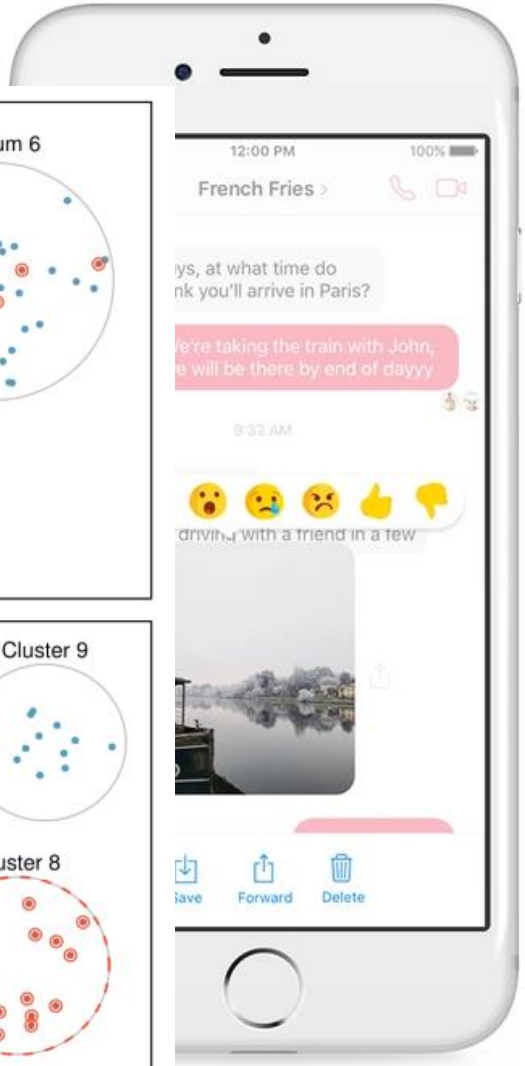
Source: https://developers.facebook.com/videos/f8-2017/how-we-shipped-reactions/

Split the audience: challenges

- People are not independent
  - E.g., your
    network
    a marketp
  - Workarou
    algorithm
    sampling







Source: https://developers.facebook.com/videos/f8-2017/how-we-shipped-reactions/

Split the audience: challenges

- User's authentication is imperfect
  - The allocation is device-based and people can have multiple devices
  - Workaround: special allocation algorithm, usually based on some timing

CHOOSE A KPI

Choose a KPI

- Find something which is measurable
  - Anything out of scope is usually hard
  - Anything ‘sentiment’ is usually hard, find a proxy instead

- A single metric would be nice
  - OEC – Overall Evaluation Criterion
  - ...it is hardly that simple
  - Ultimately: more money, but how?
    - # customers x (# bookings) x $$

Choose a KPI

- Typically something in the booking funnel
  - Except if there is something specific, e.g. joining the loyalty program, downloading the app
  - Conversion/revenue is the best, but the hardest
  - Engagement on its own is rarely useful
  - Don’t mess up the next step

Receives an email

Opens an email

Clicks on the email

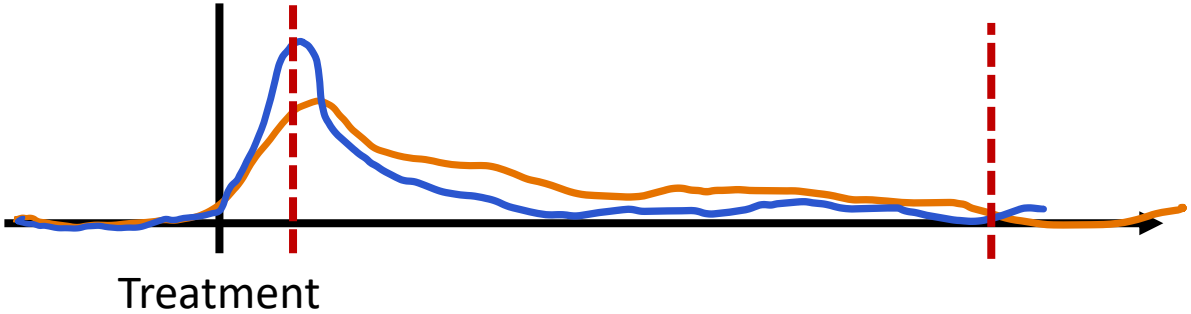
Engagement with the site

Places an order

Choose a KPI

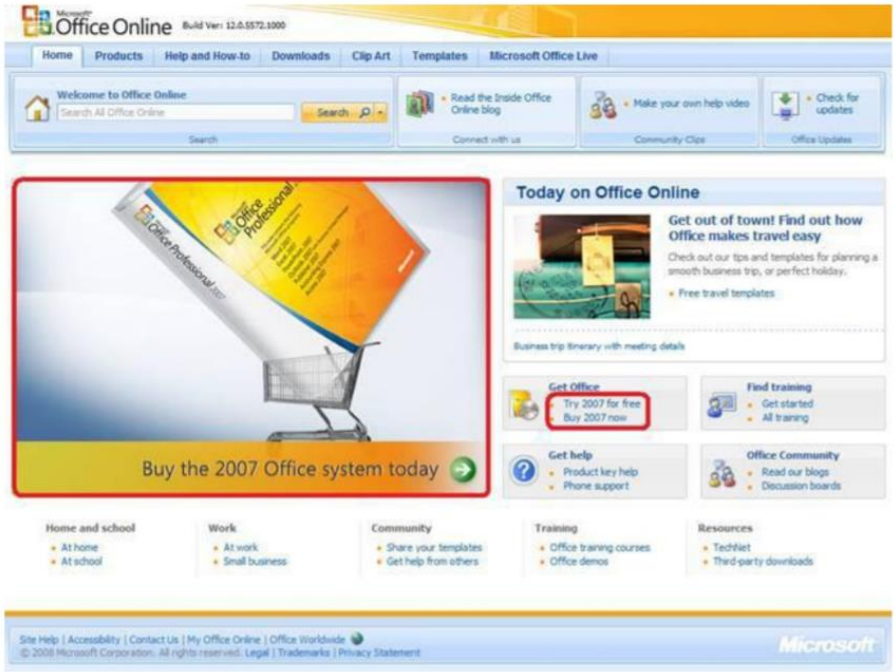
- Focus on customer level impact instead of your product's performance only

- Focus on the timeline if your impact is delayed
  - Too short: you only move the bookings
  - Too long: your impact gets lost



Choose a KPI: challenges

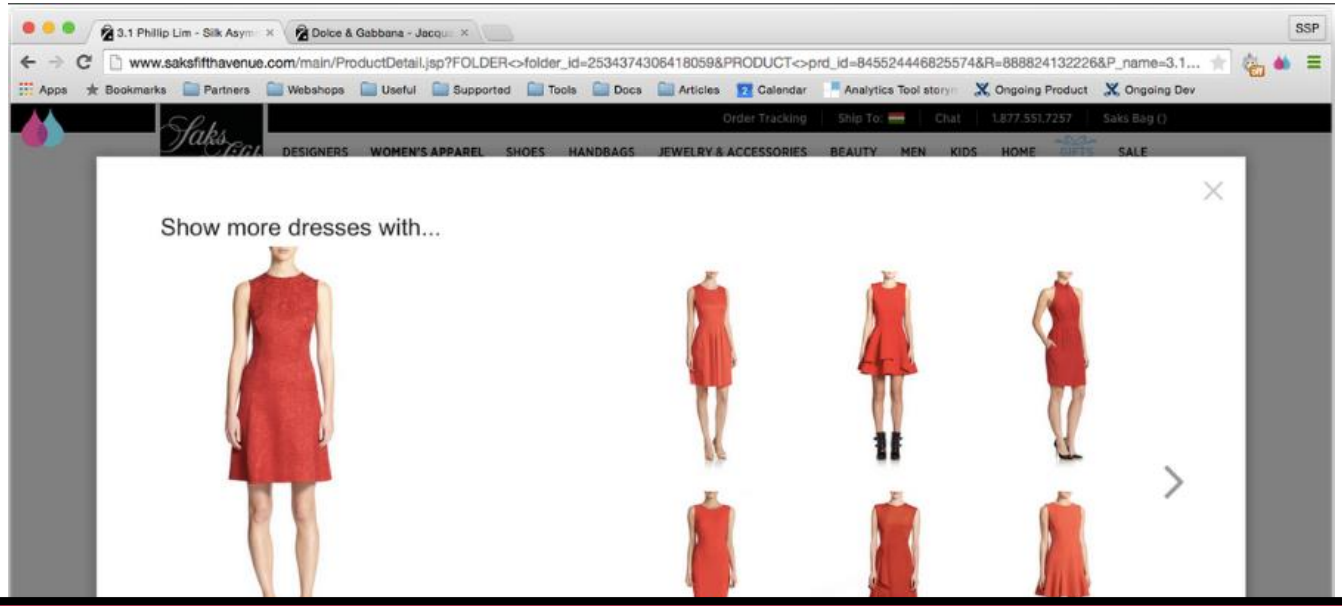
- Engagement
  - Increased click rate is great! Or.. is it?



Source: Kohav, Crook, Longbotham: Online Experimentation at Microsoft

Choose a KPI: challenges

- Engagement
  - Increased click rate is great! Or.. is it?



Engagement on its own is rarely useful

Choose a KPI: challenges

- Novelty effect
  - Emojis in a subject line

Hotels.com Magyaror. Inbox ✅ Gratulálunk! Épp most találtál rá akár 50%-nyi kedvezményre. Nyisd meg most! - mail.hotels.com /?q...

Hotels.com Magyaror. Inbox ✅ **Ez már a tiéd: akár 50% kedvezmény, csupán egyetlen napig!** - mail.hotels.com /?qs=b048362e75f9...

Hotels.com Magyaror. Inbox 🕒 Tikk... Takk... Használd fel 50%-os kedvezményedet még ma! - mail.hotels.com /?qs=02b79eafb5cea...

Hotels.com Magyaror. Inbox ⚡ Épp időben! Akár 50%-os kedvezményhez is juthatsz most villámgyors akcióinkkal. - mail.hotels.com...

Hotels.com - Titkos. Inbox ✨ **Szerencsés nap a mai!** ✨ **Hozzáférést nyertél a Titkos árakhoz: fizess akár 50%-kal kevesebbet!** - m

Hotels.com Magyaror. Inbox 🎁 **Juss akár 50% kedvezményhez** - mail.hotels.com /?qs=07cf634dc1be8cd4977b60e579b25035e11d...

Hotels.com Magyaror. Inbox ✅ Gratulálunk! Épp most találtál rá akár 50%-nyi kedvezményre. Nyisd meg most! - mail.hotels.com /?q...

Workadound: create a cut for new customers!

Choose a KPI: challenges

- Novelty effect
  - „The Pepsi Challenge”
  - Developing New Coke
  - Major crisis → back to Coke

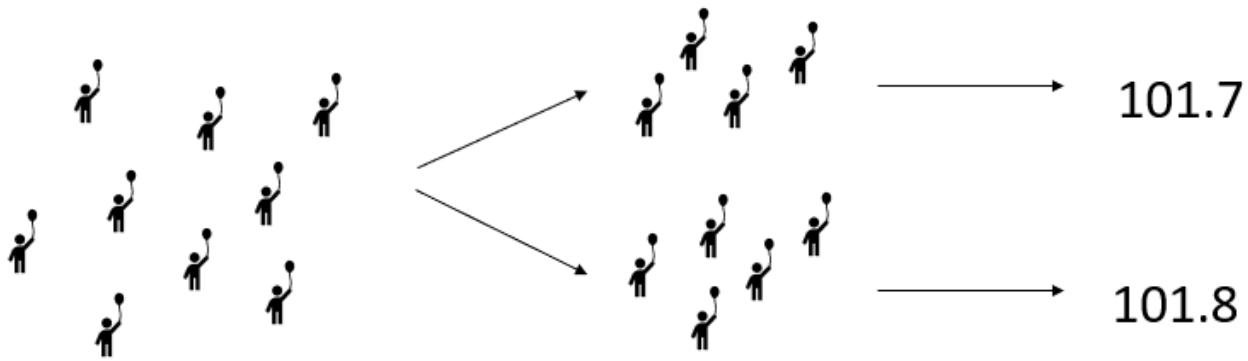
Testing setup should be representative

Source: Gladwell, Malcolm. (2005) Blink :the power of thinking without thinking

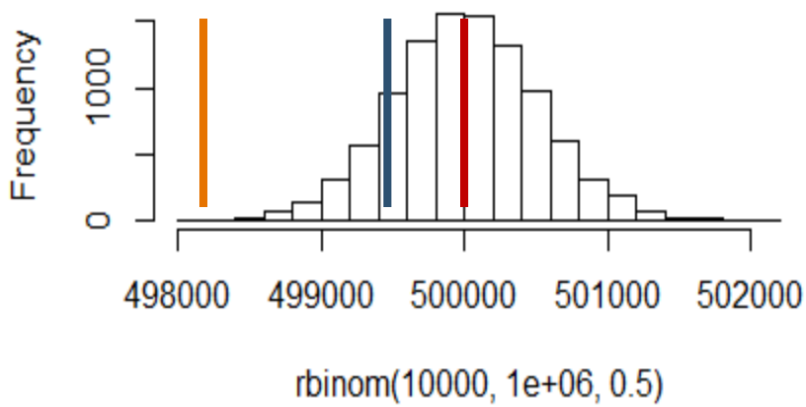
MEASURE THE DIFFERENCE

Calculate the difference

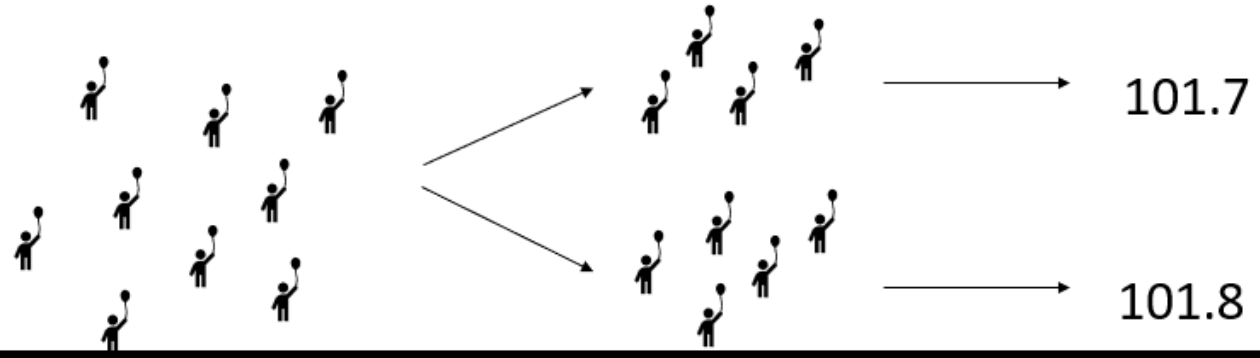
- It is different.. by how much?



Is this ‘a lot’?



Hypothesis testing reminder



1. Hypotheses:
   - **Null hypothesis:** ~Difference appears just by chance
   - **Alternative hypothesis:** ~the values are 'too far' from each other for us believe that they are different only by chance
2. Calculate a point estimate (the difference)
3. Create a distribution (of how the difference should look like)
4. Make a decision:
   - If the estimate is very far → reject null hypothesis
   - If the estimate is not very far → we have no idea

Example: conversion

- #Visitors
  - control: 1M
  - test: 1M
- # Customers

Control: 10k Test: 10.1k	Control: 10k Test: 10.5k	Control: 10k Test: 11k

- Conversion

Control: 1% Test: 1.01%	Control: 1% Test: 1.05%	Control: 1% Test: 1.1%

- Impact

1%	5%	10%

Example: conversion

- H0: the two proportions are the same:
  - $\mathrm{p}_{test} = p_{control}$
- HA: the two proportions are different:
  - $p_{test} \neq p_{control}$
- Point estimate: $p_{test} - p_{control}$
- Draw the distribution:
  - under H0, there is a *pooled proportion*

$$\mathrm{p_p} = \frac{\#customers\ in\ test + \#customers\ in\ control}{\#visitors\ in\ test + \#visitors\ in\ control}$$

  - $N(0, SD = \sqrt{\frac{p_p \times (1 - p_p)}{\#\ visitors\ in\ control + \#\ visitors\ in\ test}})$

Example: conversion

- # Customers

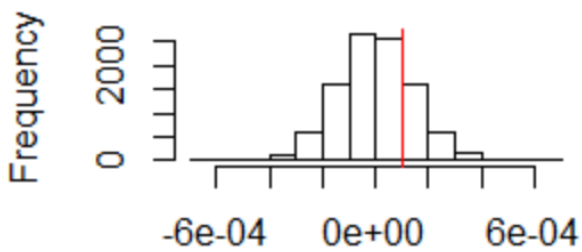
Control: 10k Test: 10.1k	Control: 10k Test: 10.5k	Control: 10k Test: 11k

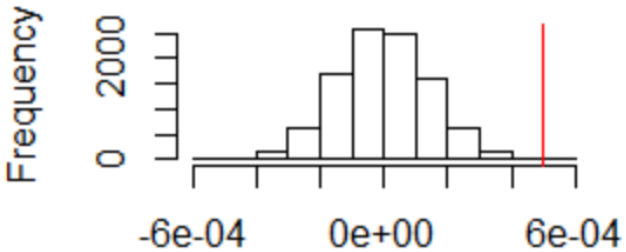
- Conversion

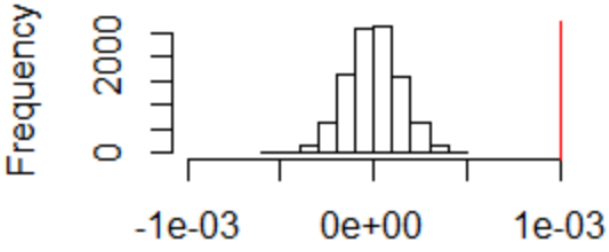
Control: 1% Test: 1.01%	Control: 1% Test: 1.05%	Control: 1% Test: 1.1%

- Point Estimate

0.0001	0.0005	0.001

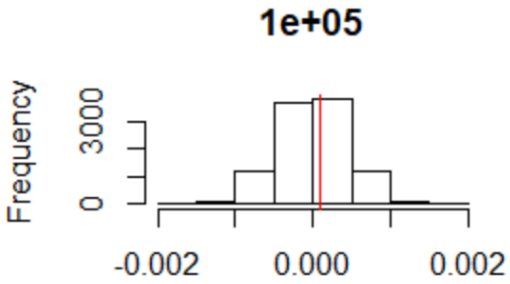


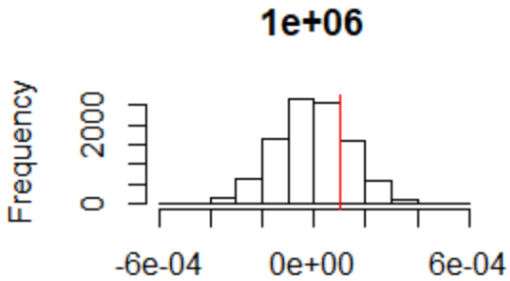


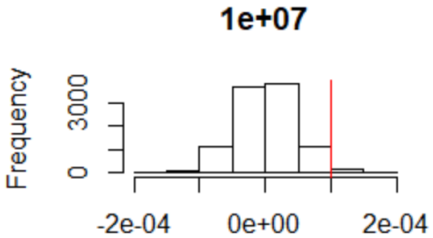


How to get better results?

- Point estimate is relatively far
  - Larger sample → smaller SD
  - 1% conversion, 1% impact → good sample size?







- What if conversion rate is very small?
  - Special workarounds, e.g. Wilson technique

Calculate the difference

- Back to KPI selection
  - Conversion metrics are much better than revenue metrics
  - Customer- level metrics are much better than event-level metrics
    - e.g. click reach and click through rate
    - Workaround: simulation
- What if we want to test stability instead of a change?
  - Equivalence test (new drug is the same but cheaper)

MAKE A DECISION

Make a decision

- If significant → great! 

- What if it is non-significant?
  - Sometimes we roll out anyway because the feature is already implemented

If we have more than one metric?

- Impact on significance
  - What is the chance of a false positive if we have 1, 2, 3, 4, 5 metrics (we test with a 0.95 significance level)?
  - Workaround: Bonferroni correction
- If we have changed more than one thing?
  - Another round? More variants from the beginning?
- If different cuts are in conflict with each other?
  - Like, gold customers enjoy something but non members don't?
  - Workaround: customization

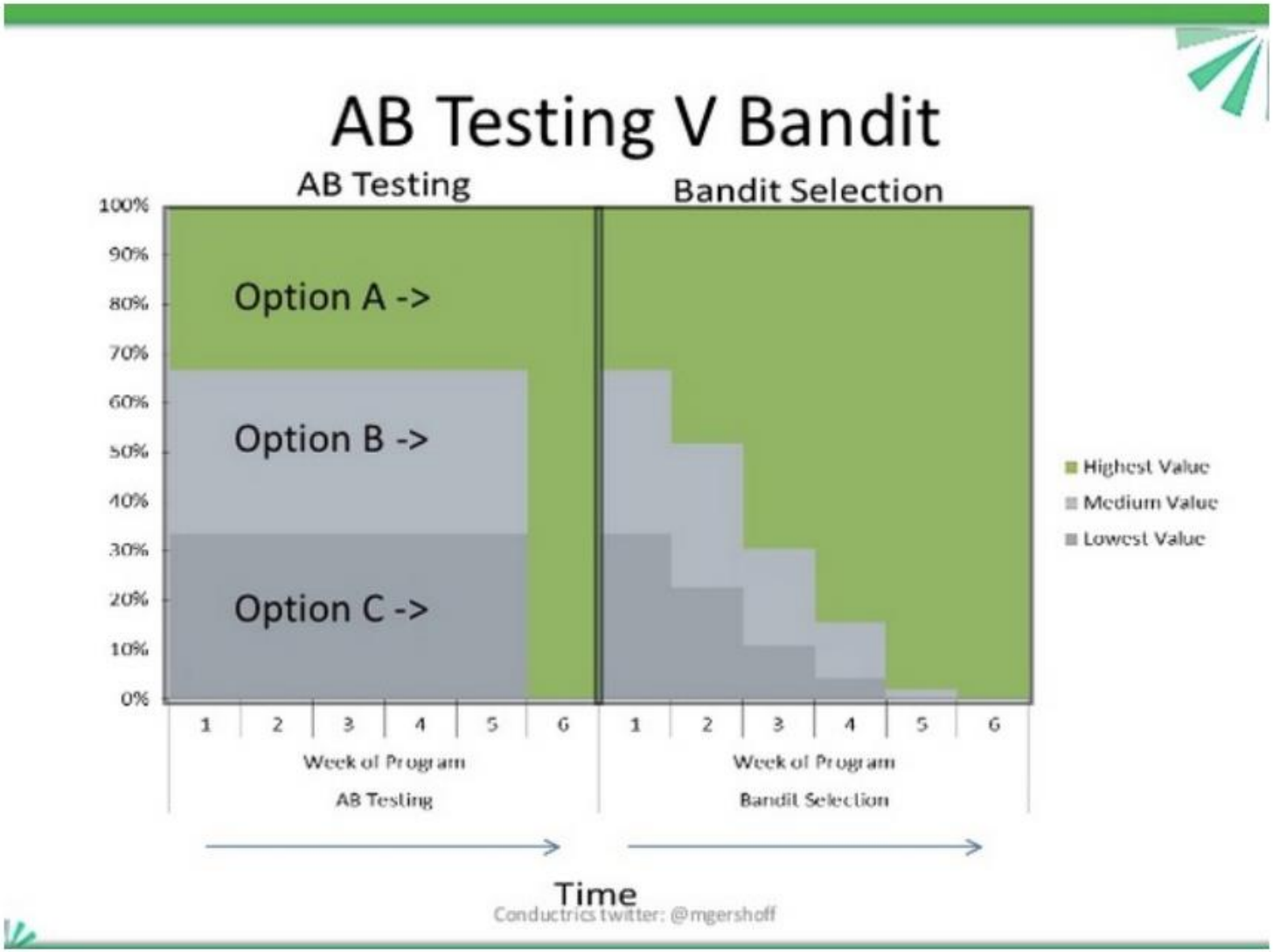
Before the test

- A/A test
  - validation of
    - Data collection infrastructure
    - Test setup
  - Estimation for expected value and variance

After the test

- ‘Annual impact’
- Deep dive
  - Optimization, optimization, optimization
- Things change..
  - People, customer base

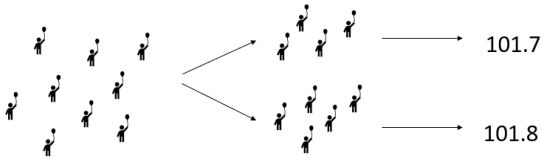
One word about MABs



*Source:*https://cxl.com/blog/bandit-tests/

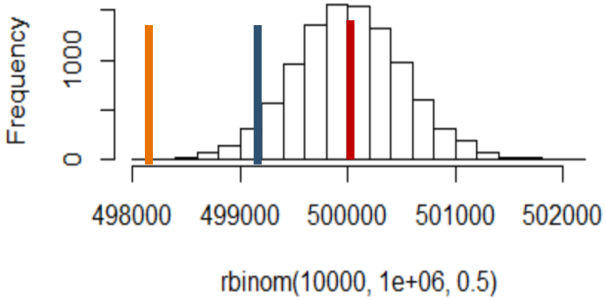
Summary

1. Split the audience



2. Choose a KPI

3. Calculate the difference



4. Make a decision