

A/B Testing Stories

Agnes Urbanics-Salanki

February 2020, CEU

Agenda

- My journey to analytics
- A/B testing stories

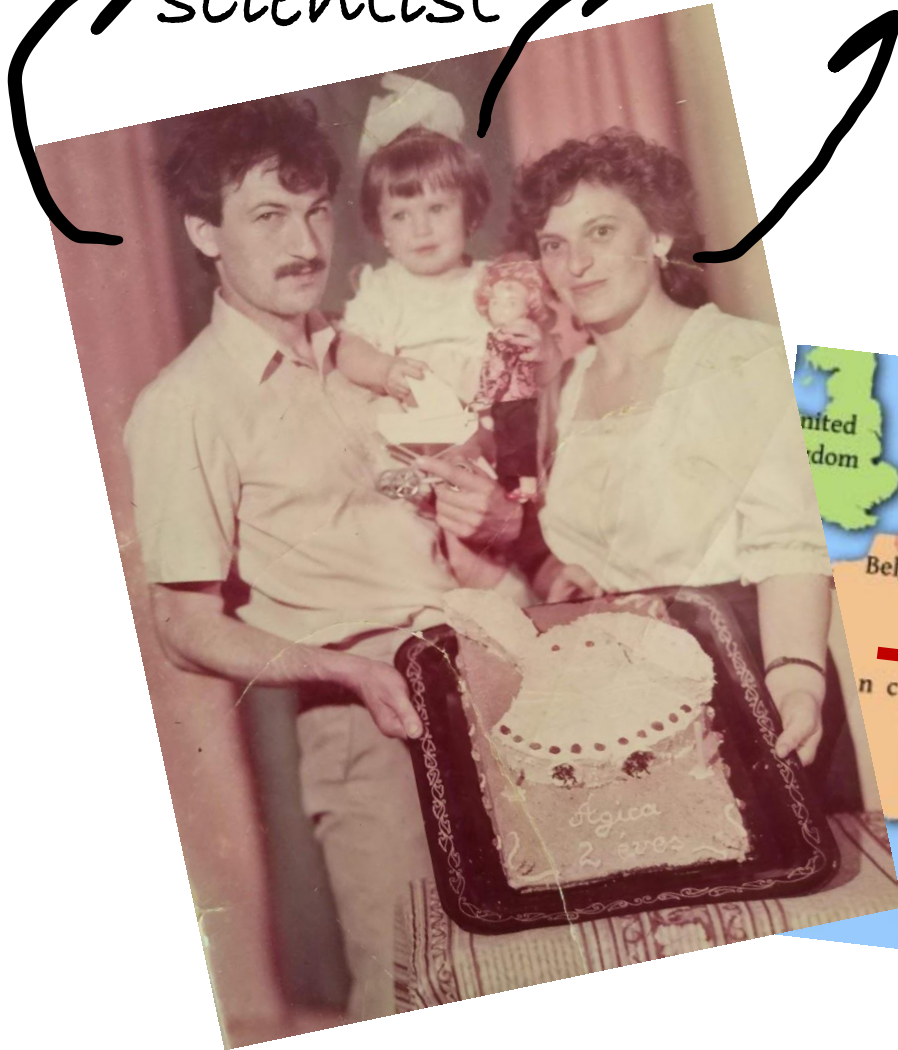
MY JOURNEY TO ANALYTICS

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



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



Hotels.com™

- Programming
 - R ♥
 - Internships
 - Traveling
- Importance of QA and good engineers
 - 'business intuition'
- 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

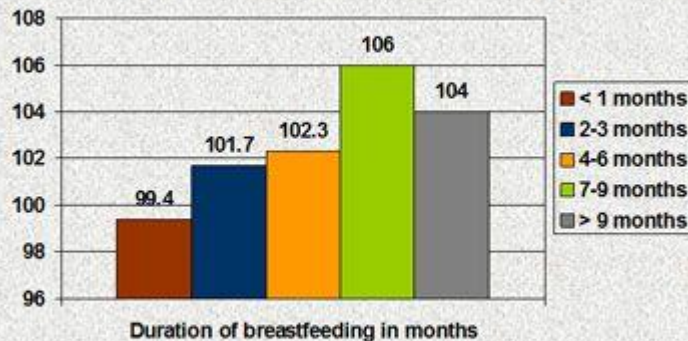
- Why A/B testing?
- A/B tests in product management
- A/B testing in theory
- Challenges

WHY TESTING?

Observational studies vs controlled experiments

■ Breast is best!

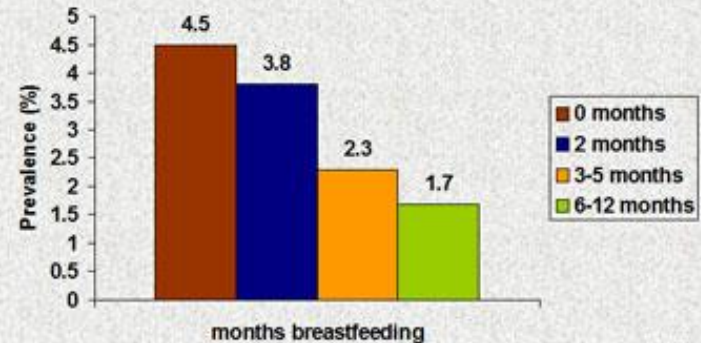
Duration of breastfeeding associated with higher IQ scores in young adults, Denmark



Adapted from: Mortensen EL, Michaelsen KF, Sanders SA, Reinisch JM. The association between duration of breastfeeding and adult intelligence. *JAMA*, 2002; 287: 2365-2371.

Slide 2.22

Breastfeeding decreases the prevalence of obesity in childhood at age five and six years, Germany



Adapted from: von Kries R, Koletzko B, Sauerwald T et al. Breast feeding and obesity: cross sectional study. *BMJ*, 1999; 319:147-150.

Slide 2.19

Breastfeeding guilt experienced by half of mothers - BBC survey

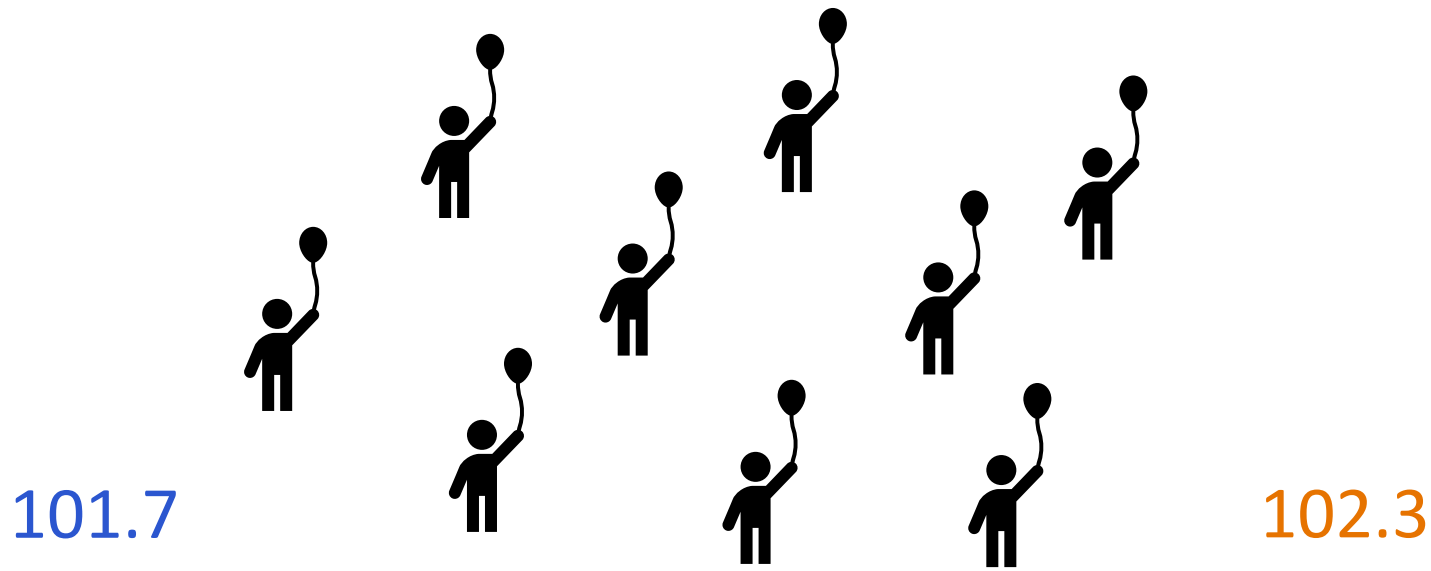
🕒 29 January 2019



🔗 Share

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

Observational studies vs controlled experiments



- Breastfed \leftrightarrow higher IQ ✓

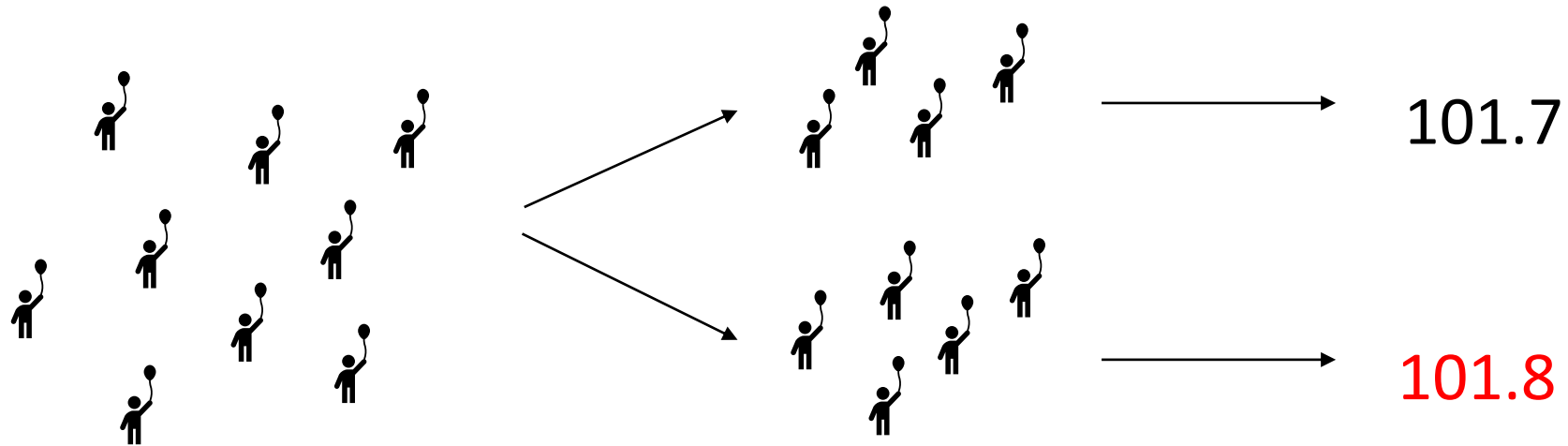
- Breastfed \rightarrow higher IQ
You cannot determine causality based on an observational study

- Higher IQ $\leftarrow \dots \rightarrow$ Breastfed

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

■ Highest Paid

Person's Opinion

- Knows the customer
- Knows the product
- Knows the strategy
- etc.

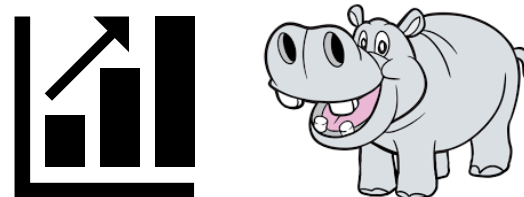
■ Testing

- 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*

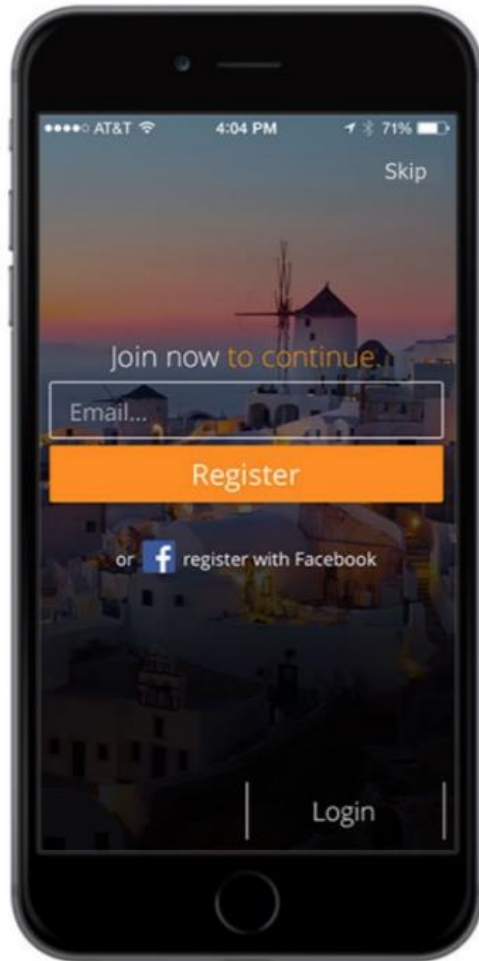
Testing in Product Management

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

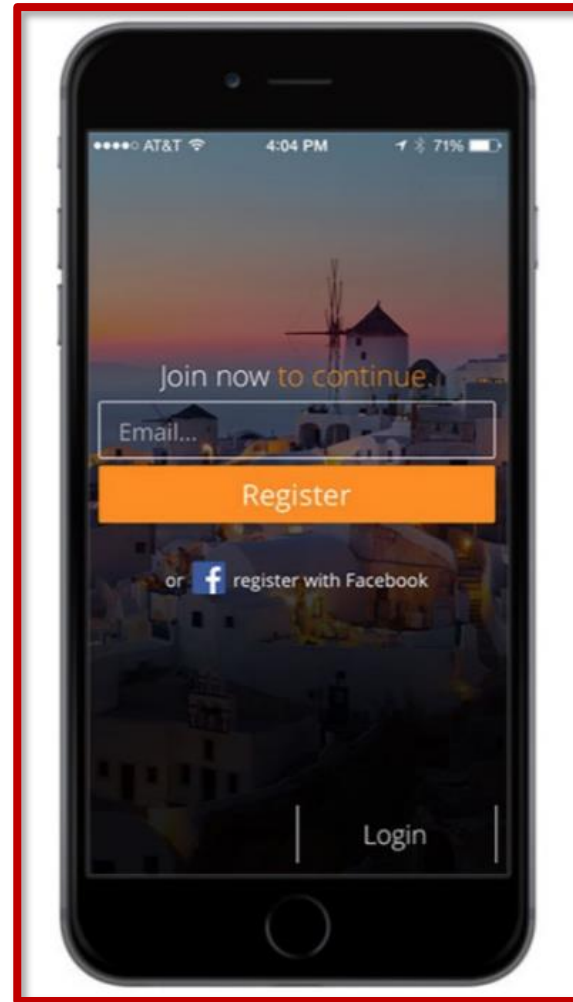
From now on, we will test
features ON people

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

The image shows a mobile app interface for Variant B. It features a blue header with the text "You could win dinner with Barack Obama". Below this, a paragraph of text reads: "Make a donation today and be automatically entered to win dinner with President Obama. Once the deadline's passed, you may not have this chance again, so enter today—we'll cover your airfare and hotel." Below the text are three form fields: "*First name:", "*Last name:", and "*Address:". Each field has a corresponding input box. At the bottom of the app, there is a navigation bar with four icons: a back arrow, a forward arrow, a share icon, and a book icon.

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

Visit Our Other Sites: [WINE ENTHUSIAST CATALOG](#) [WineExpress.com WINE SHOP](#) [WINE ENTHUSIAST MAGAZINE](#)

800.962.8443 View Cart Track Order My Account Customer Service

WineExpress.com

A Classic Blend of Value & Service

Shipping is just \$1.95 per bottle and our select Wine of the Day **ALWAYS** ships for 99¢ per bottle!

Browse By Price **GO** SEARCH Enter Keyword or Item # **GO** [Advanced Search](#) [Email Specials](#)

Wine Express® Wine Clubs | Wine Gift Baskets | Gift Samplers | Sale | Our Guarantee | Video Tastings

Browse by: [Wines by Type](#) | [Wines by Varietal](#) | [Wines by Region](#)

Wine of the Day

Wine of the Day is your daily selection of delicious WineExpress.com wines that ship for just 99¢, for 24 hours only.

Villa Antinori 2005 Toscana IGT - [Read the Description](#)

Item Number: 29 30 191 05
Our Price: **\$21.95**

Discounted Case Price: **\$237.99** - 22% Off! (\$29.75 per bottle)

Availability: **IN STOCK**

Ship to State: **California**

Buy: **1** Bottle(s)

ADD TO CART

CUSTOMER RATING
Based on the averaged scale of 1 to 5 glasses
★★★★★
[Write a Review](#) [Read 1 Review](#)
[About Our Ratings](#)

Video Tasting: Virtually Taste Before You Buy

You May Also Like:

- Chianti Classico 2007 Straccali**
Item Number: 29 30 191 07
Our Price: **\$14.95**
- Chateau de Beze 2007**
Item Number: 29 30 191 08
Our Price: **\$29.95**
- Chateau de Beze 2008**
Item Number: 29 30 191 09
Our Price: **\$29.95**

Variant B

WineExpress.com
A Classic Blend of Value & Service

Order in the next [3:43:32] to get 99 cent shipping!

Wine of the Day

Your daily selection of delicious WineExpress.com wines that ship for just 99 cents, for 24 hours only

Chianti Classico DOCG 2007 Straccali
Item Number: 29 30 191 07

CUSTOMER RATING Based on a scale of 1 to 5 glasses
★★★★★ [Read 1 Review](#)

Chianti Classico is not the same as mere Chianti. "Classico" means that the grapes were grown in the oldest delimited zone in the region, and the production code is quite strict. Only vineyards situated on hillsides above 700 meters with advantageous orientation may be included. Vine density is spelled out as are maximum yields, and for DOCG even the maximum amount of wine per ton of grapes is set by law. Chianti Classico is no ordinary wine and this is no ordinary Chianti Classico. Straccali sourced their grapes from some of the best vineyards in the zone, vineyards up to 1,800 feet above sea level, where the cool temperatures combine with the direct sunlight to bring out all of the complexity and nuance that Sangiovese can deliver. 90% Sangiovese blended with 10% Canaiolo and Merlot aged in oak casks and then in bottle, this wine offers ripe cherry, anise, tobacco and cedar notes in a supple easy-to-enjoy style. Best of all it's a tremendous value.

EXPERT RATING
Based on the 100 Point Scale
WE: 92
WS: 92
WA: 92
[About Our Ratings](#)

Customer Reviews [Write a Review](#) [About Our Ratings](#)

Displaying Review 1 of 1

★★★★★ **Kicked our dinner up a notch!**

By **Napahol** [Write a Review](#) from On the Hudson on 6/6/2010

Gift: No
Pairs Well With: Beef
Price: Balanced, Earthy
Best Used: Entertaining
Describe Yourself: Aspiring Enthusiast
Bottom Line: Yes, I would recommend this to a friend

This was the highlight of our dinner party! With notes of spice and forest floor the Sangiovese grape really showed! The integration and flavor profile in the oak really made the wine hold up after being opened for several hours (it was our 3rd bottle). Enjoy now or hold on to for a couple of years as I'd be curious to see how this wine shows after some time in my European.

Was this review helpful to you? Yes / No - You may also like this product.

Displaying Review 1 of 1

You May Also Like

- Chianti Classico 2007 Straccali**
Item Number: 29 30 191 07
Our Price: **\$14.95**
- Chianti Classico 2008 Straccali**
Item Number: 29 30 191 08
Our Price: **\$14.95**
- Chianti Classico 2009 Straccali**
Item Number: 29 30 191 09
Our Price: **\$14.95**
- Chianti Classico 2010 Straccali**
Item Number: 29 30 191 10
Our Price: **\$14.95**

KPI: Revenue per visitor

Windows Office Online Feedback

Variant A

Please let us know if this content was helpful.

Rate this content:



Tell us why you rated the content this way (optional):

Remaining characters: 650

Submit

Variant B

How helpful was this information?

Click a star.

Not helpful  Very helpful

Click to rate: 3 out of 5 stars



How helpful was this information?

Click a star.

Not helpful  Very helpful

Why did you rate the information this way?

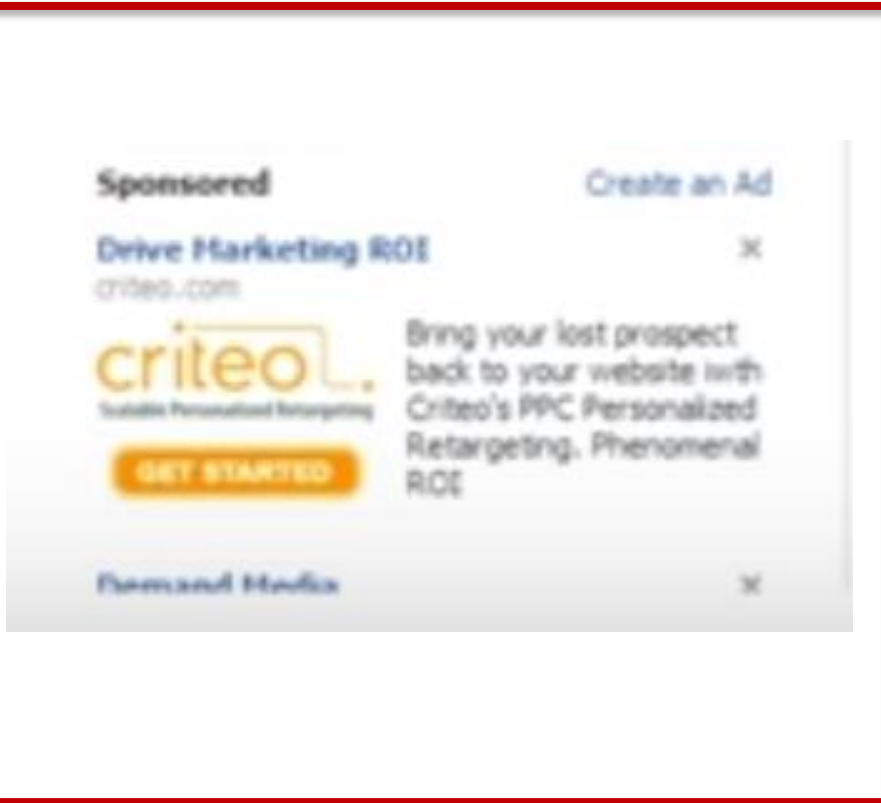
Remaining characters: 650

Submit

KPI: Response rate

Criteo FB advertiser sign ups

Variant A



Variant B



KPI: Conversion

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”*

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?

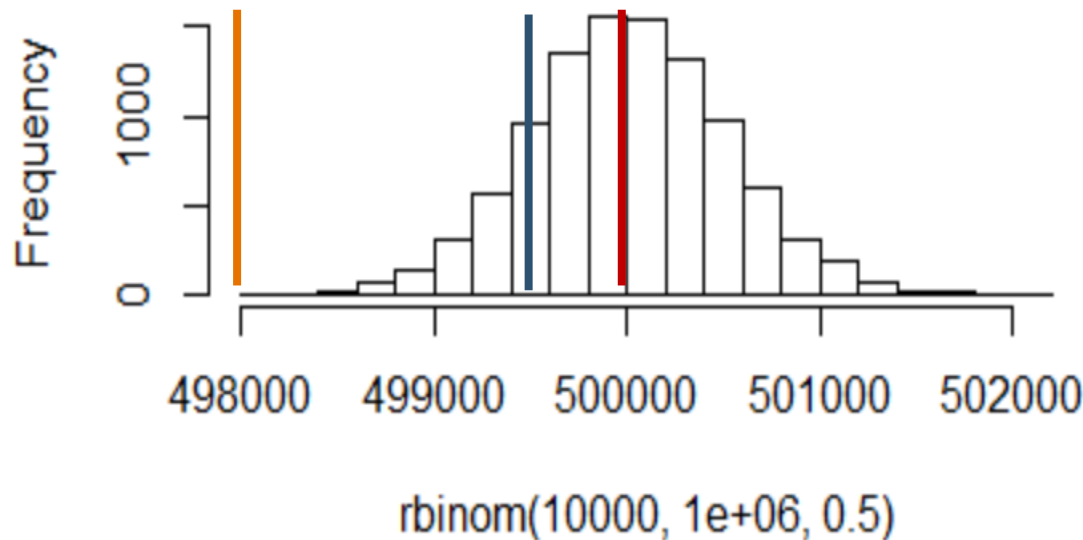
Split the audience

- 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

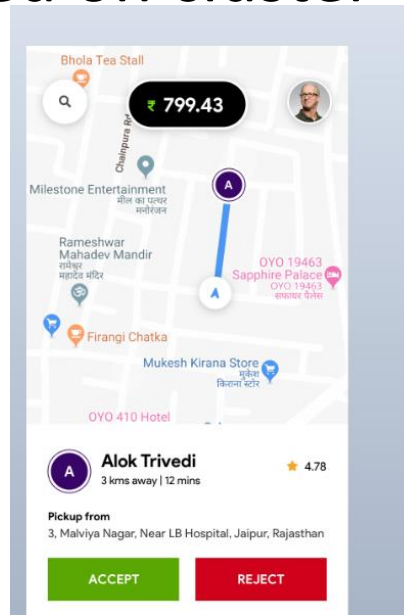
■ Basic checks to run

- 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

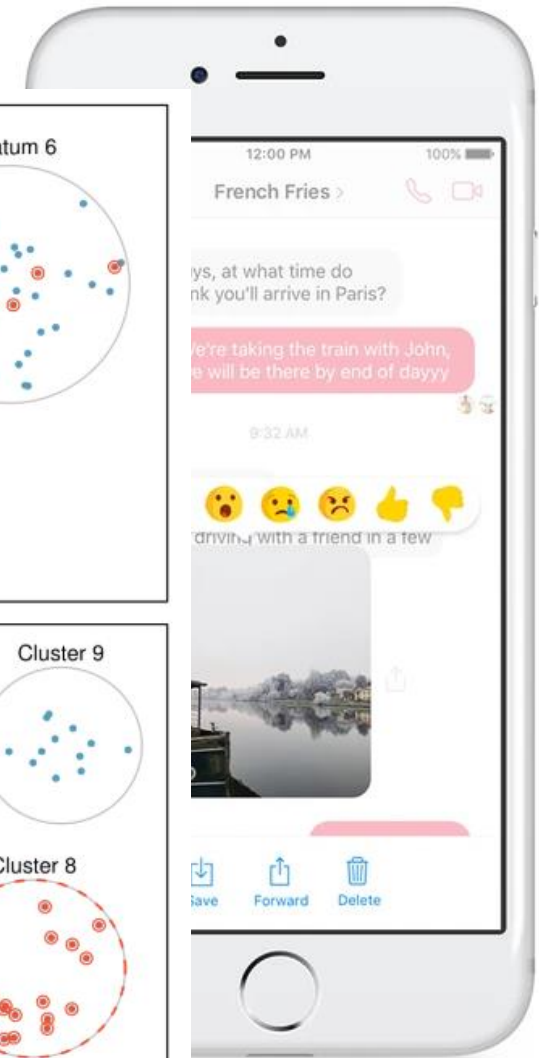
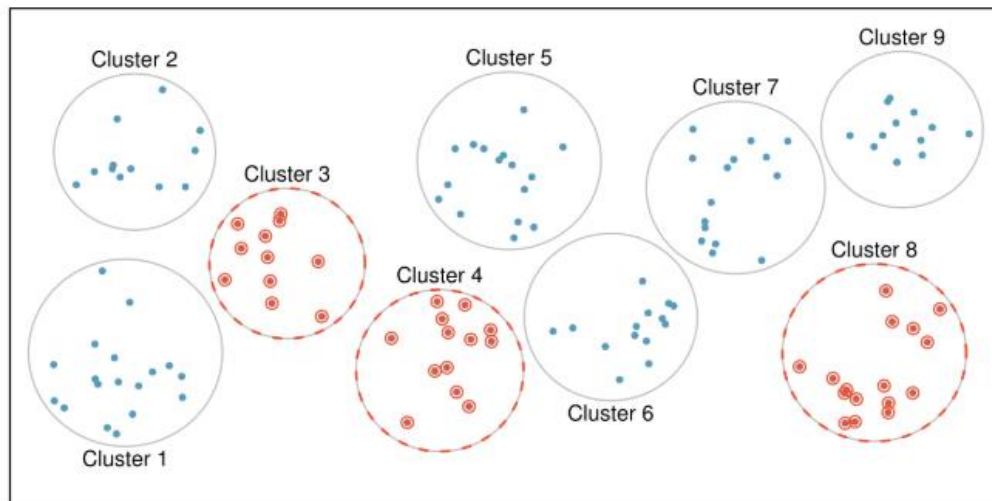
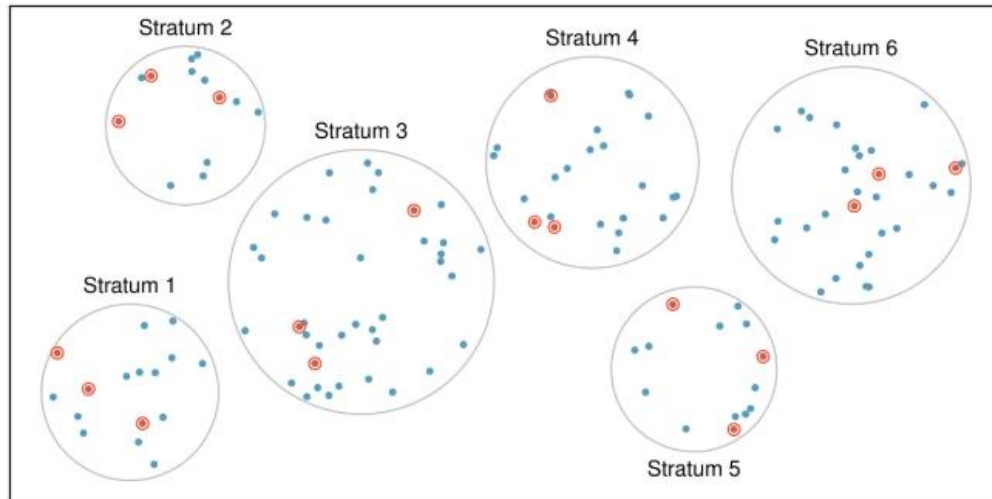
- People are not independent
 - E.g., your users are members in a network or otherwise related, e.g. in a marketplace
 - Workaround: special allocation algorithm, usually based on cluster sampling



Split the audience: challenges

- People are not independent

- E.g., your network of a marketplace
- Workaround: algorithmic sampling



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

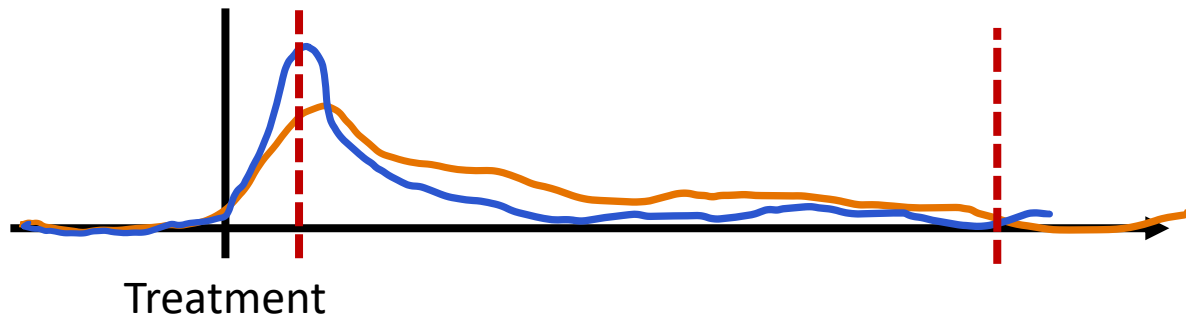
Clicks on the email

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

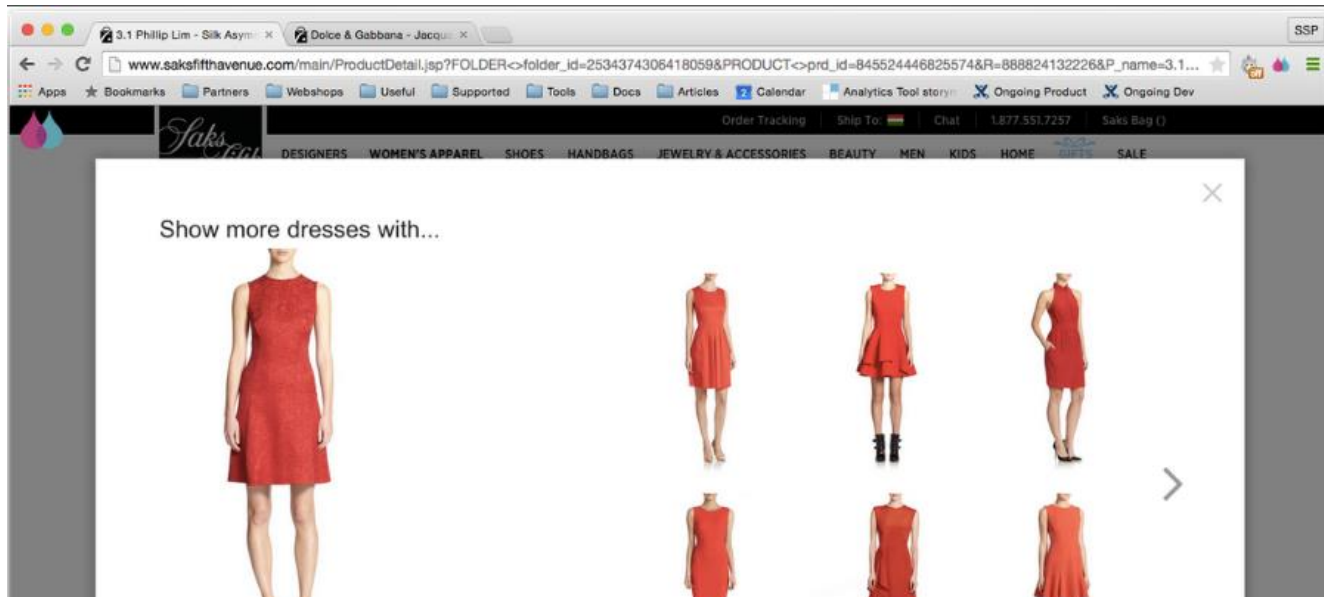
- Increased click rate is great! Or.. is it?



Choose a KPI: challenges

- Engagement

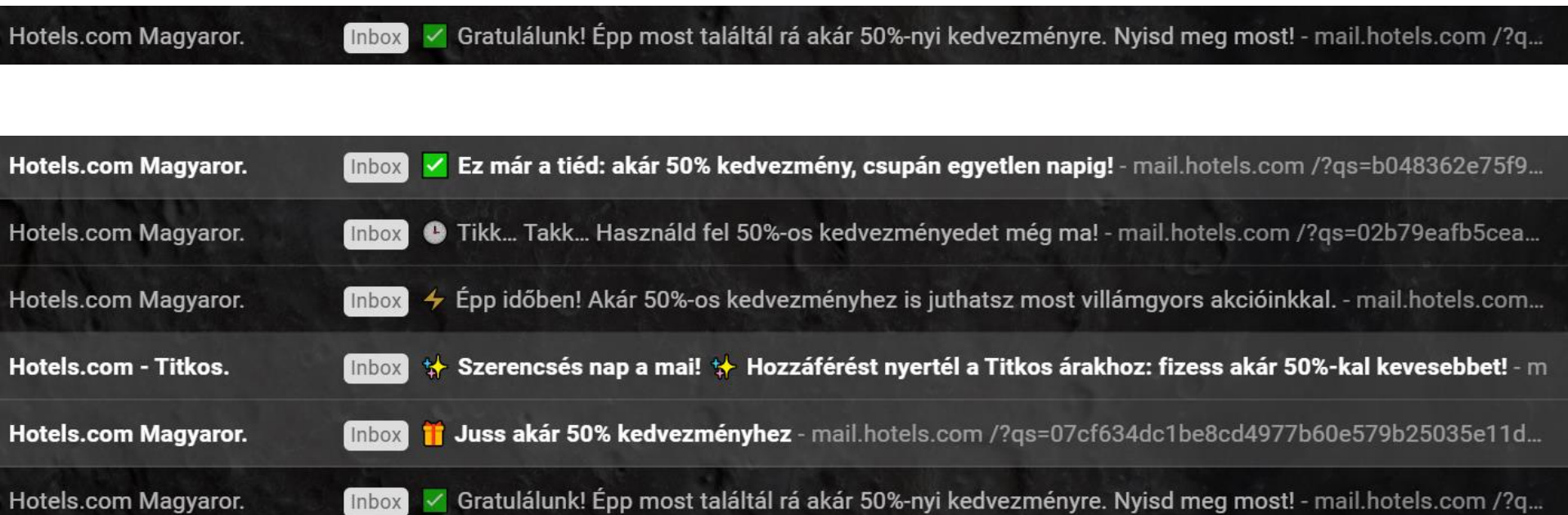
- Increased click rate is great! Or.. is it?



Engagement on its own is
rarely useful

Choose a KPI: challenges

- Novelty effect
 - Emojis in a subject line



Workadound: create a cut for new customers!

Choose a KPI: challenges

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

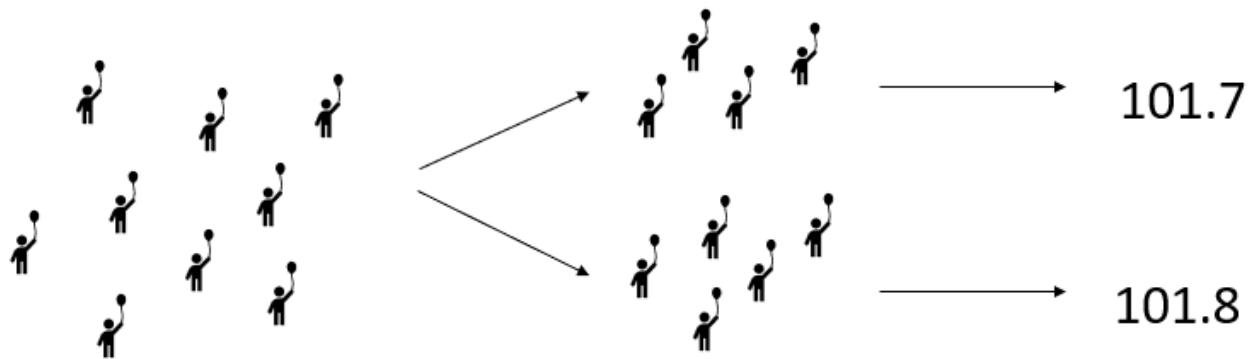
Testing setup should
be representative



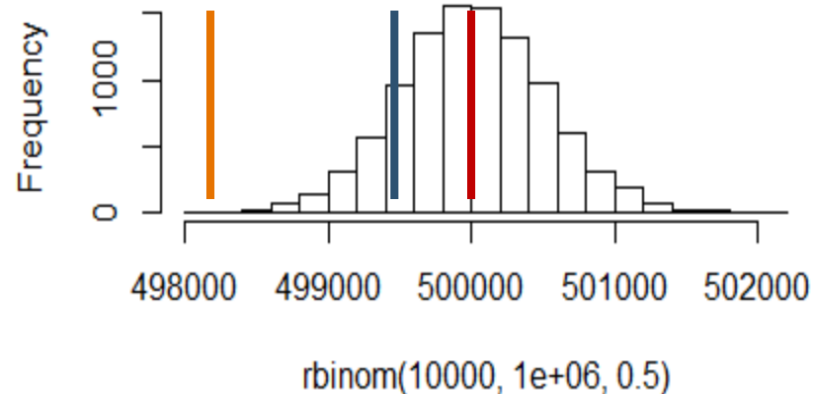
MEASURE THE DIFFERENCE

Calculate the difference

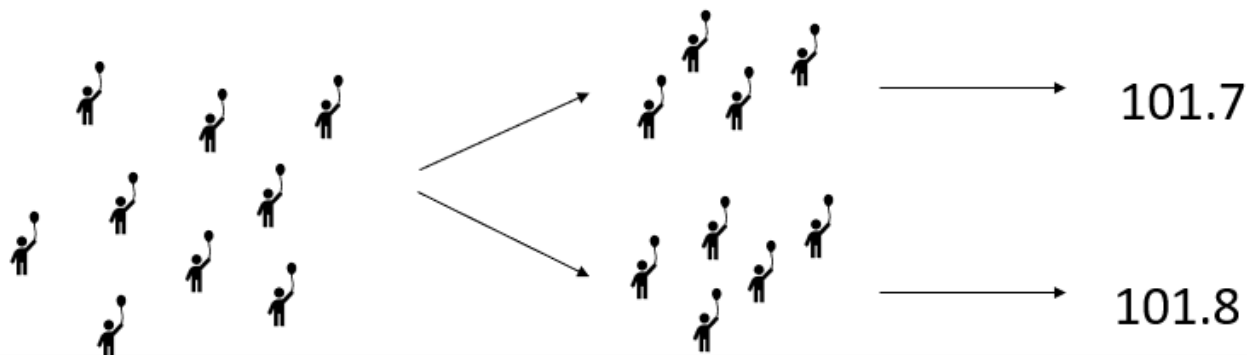
- It is different.. by how much?



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 \rightarrow reject null hypothesis
- If the estimate is not very far \rightarrow 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

- H_0 : the two proportions are the same:

- $p_{test} = p_{control}$

- H_A : the two proportions are different:

- $p_{test} \neq p_{control}$

- Point estimate: $p_{test} - p_{control}$

- Draw the distribution:

- under H_0 , there is a *pooled proportion*

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

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

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%

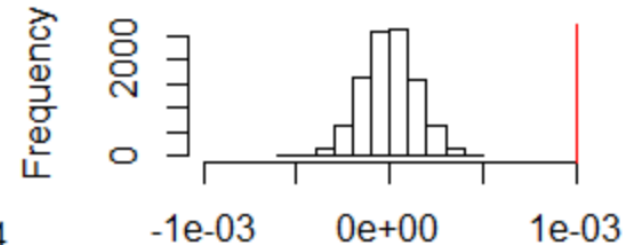
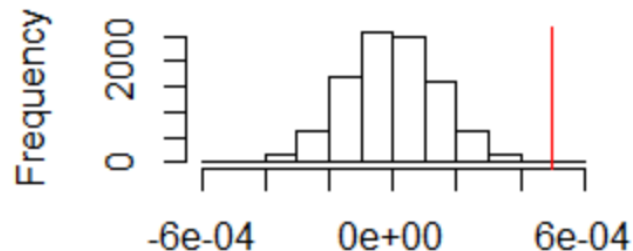
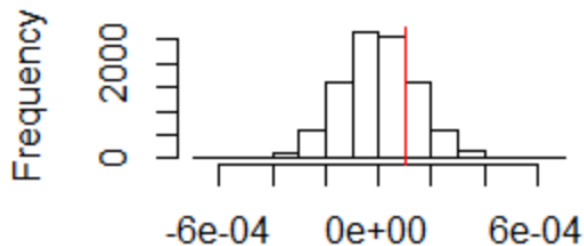
Control: 1%
Test: 1.1%

■ Point Estimate

0.0001

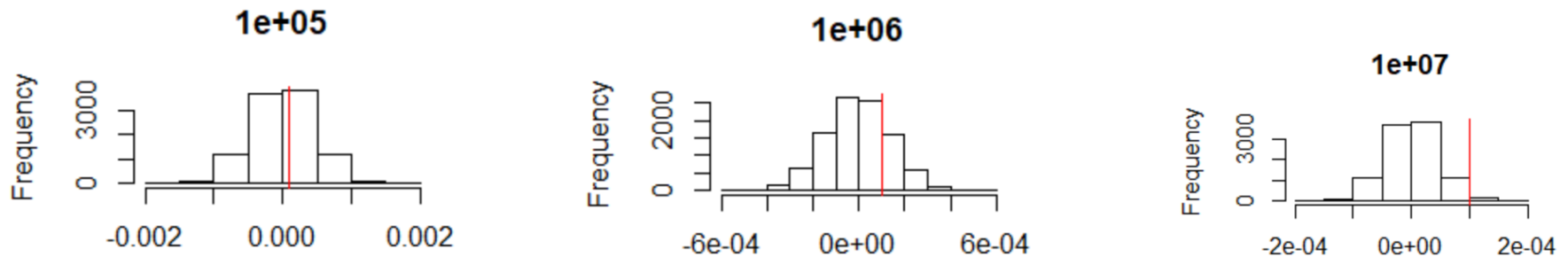
0.0005

0.001



How to get better results?

- Point estimate is relatively far
 - Larger sample \rightarrow smaller SD
 - 1% conversion, 1% impact \rightarrow 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

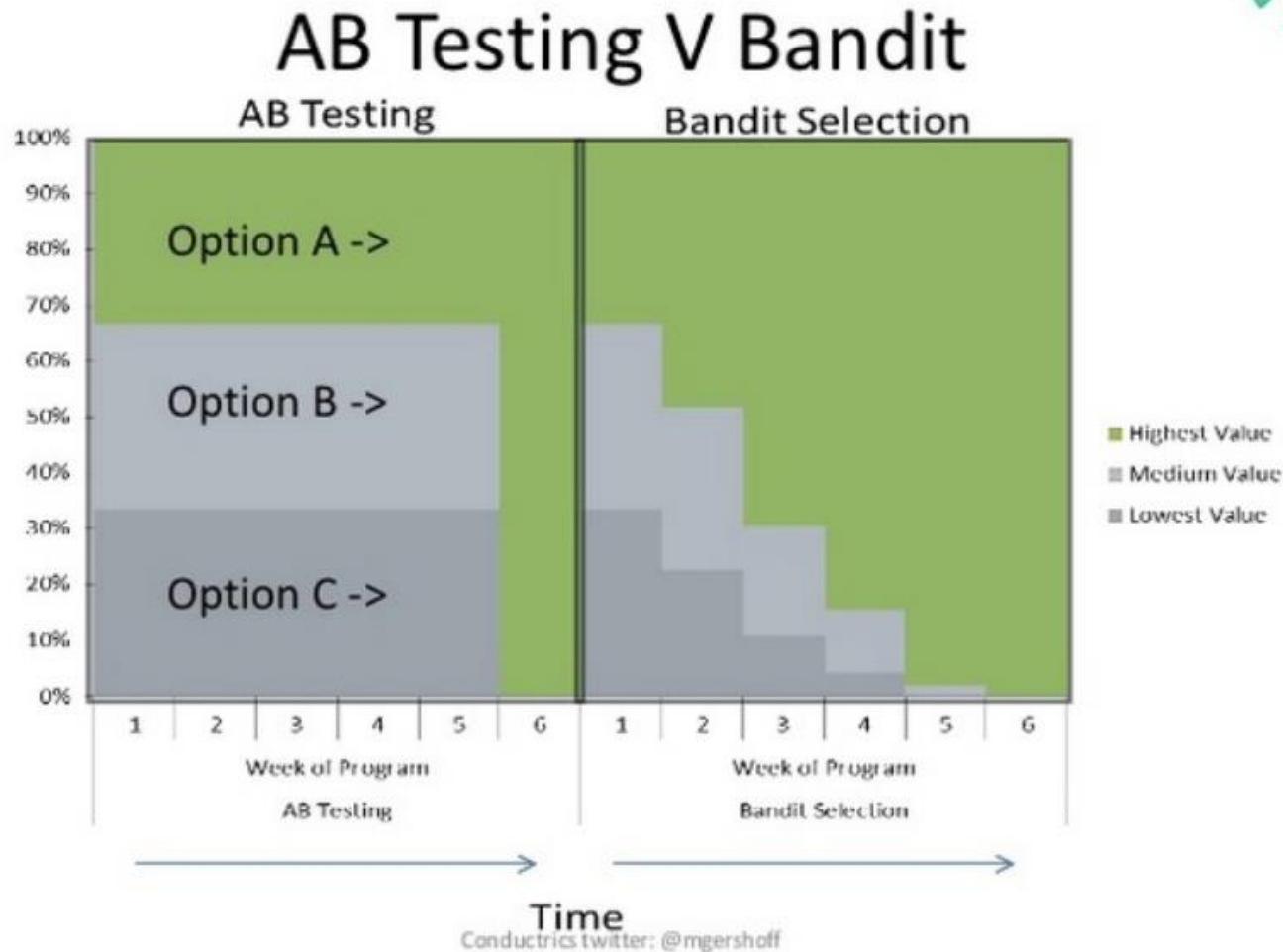
Before the test

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

After the test

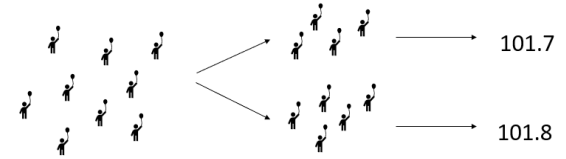
- 'Annual impact'
- Deep dive
 - Optimization, optimization, optimization
- Things change..
 - People, customer base

One word about MABs



Summary

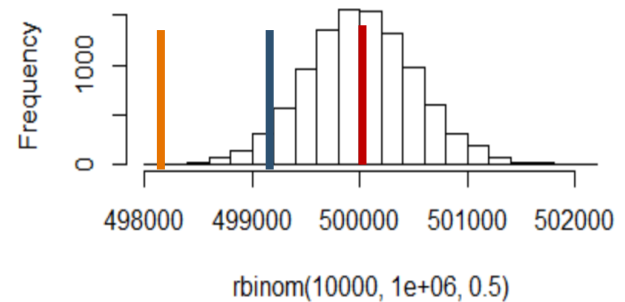
1. Split the audience



2. Choose a KPI



3. Calculate the difference



4. Make a decision

