

# Experimentation in the Wild

A FIELD GUIDE TO A/B TESTING & MORE





## SPEAKER INTRO

- Well-rounded jackass
- Data science and engineering, in the SaaS, e-commerce industry working across NLP, recommendations and experimentation topics.
- Currently, at Priceloop

An algorithmic pricing platform that makes use of a combination of machine learning and experimentation to dynamically price products for our client portfolio.

As a market, language and industry-agnostic platform, Priceloop is looking for pilot customers...globally!

# Preamble

- No pens, paper or other note taking apparatus required. All slides will be made available for consumption at your leisure.
- Feel free to ask me any industry or market specific examples; this talk is meant to be as interactive as possible and if clarity can be provided through illustration, I'll strive to do so.

# Agenda

1. Getting familiar with A/B testing
2. Real-world challenges with experimentation
3. Overcoming limitations of A/B testing with a new paradigm
4. Wrap-up

## ***EXTRA BONUS!***

- Going Bayesian!

# An Introduction to A/B Testing

So your company is building a product? Ever been in one of those endless meetings that seem to go-on in circles?



# An Introduction to A/B Testing

So your company is building a product? Ever been in one of those endless meetings that seem to go-on in circles?



Sometimes all we have is a collection of opinions and no concrete strategy towards a decision. So what's *the solution*?

# An Introduction to A/B Testing

So your company is building a product? Ever been in one of those endless meetings that seem to go-on in circles?



Sometimes all we have is a collection of opinions and no concrete strategy towards a decision. So what's *the solution*? A/B Testing that's right.

# A brief interlude

In the 18th Century, scurvy ran rampant. It was a major cause of death in the British Fleet, almost 1400/1900 men in Baron Anson's royal navy supposedly died from contracting scurvy.

- In 1753, James Lind, a Scottish doctor devised a treatment that involved the use of citrus fruits (heavy in Vitamin C) and tested it on 6 pairs of patients
- He tested not one but 6 different variations (types) of the treatment
- While every pair of patients were given a similar diet, each individual pair was provided different supplements
- The pair of sailors that were given 2 lemons and 1 orange a day, made incredible recoveries!

His work published in *Treatise of the Scurvy* may be considered as one of the first **controlled experiments**!



## And back to the 21st Century...

An A/B test is a controlled experiment. We have a lot of commonplace idioms:

- You can't compare apples to oranges
- Context is King

Believe it or not, these have roots (or links to) in the field of experimentation. The underlying principle here being: to decide on a winner between two options, they must,

- A. Be comparable to each other
- B. Have a pre-defined metric, that is used to declare the winner

# Who is A/B Testing?



- Measuring different flavors of recommendation strategies
- Measuring impact of different content on home/ category pages



- Testing efficacy of different layouts, colors and display styles



- Testing copy and content in email newsletters leading to higher review rates on the website



- Determining which marketing campaigns result in gain of more users on their mobile app



- Evaluating impact of registration page vs. the app homepage in driving registrations



- What drives users to order from a restaurant - review ratings or shorter delivery times?

---

# Now let's start getting technical

How do you run a simple A/B test?

1. Create two variations of a solution
2. Create two groups of your audience/testers; ensure that they are comparable
3. Decide on a metric
4. Run your test for a period of time to collect data from both groups
5. Terminate your test and declare the group that scored higher on your pre-decided metric - the winner!



Step I

## Formulate a hypothesis

What are we testing here? Is recommendation strategy B superior to existing strategy A?  
Layout B v/s Layout A?

# Test Design

**SIGN UP**

**SIGN UP**

*EXAMPLE: STRONG ASSUMPTION*

**SIGN UP**

**SIGN UP**

*EXAMPLE: FRIVOLOUS ASSUMPTION*

# Test Design



## Step II

### Define a test horizon

How long do we want to run the test? This is usually informed by the arrival rate/ traffic that is received by your website/application.

Approx. daily traffic

5000

Variations to test

2

Sample Size (per variant)

10000

Planned Test Horizon

4



Planned Test Horizon

5





## Test Design

### Step III

### Setting statistical significance

An important design parameter. The impact of a poor choice can make or break the test.

If the sample size is too small, our test results can be misleading. Statistical significance infers and allows one to control the exact probability of the winner being declared by an effect of random chance.

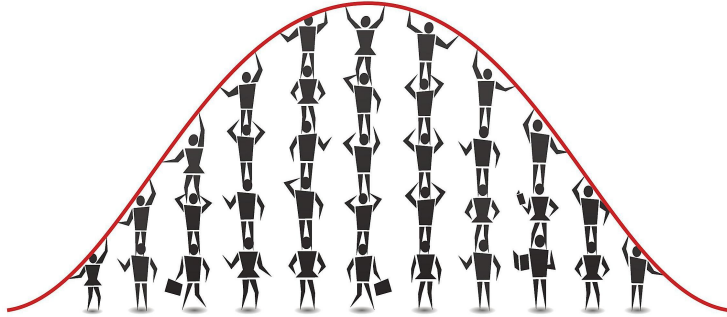
#### **INSIGNIFICANT RESULTS**

Conversions	5	6
Visits	10	10
CTR	50%	60%

#### **CONCLUSIVE TEST**

Conversions	5000	4800
Visits	20000	20400
CTR	25%	23.5%

# Test Design



## Step IV

### One-tailed vs. Two-tailed

Factor in the desired effect, do we merely want to observe the effect of the change? Is the direction of the effect also relevant or simply the magnitude?

#### **ONE-TAILED**

The kind of test when you only care about detecting a significant change.

#### **TWO-TAILED**

When both a significant change and the direction (positive/ negative) has to be determined.

#### **WHY NOT RUN TWO-TAILED TESTS ALL THE TIME?**

Valid question! Two-tailed tests require 2x the time to run.

A smart strategy is to assume two-tailed and re-consider at time of test design what suits the need.

# Challenges in A/B Test Design

1. Not enough traffic/ small sample sizes
2. Coming up with strong/sensible hypothesis
3. Tests cannot be changed in between
4. Testing too many variants and simultaneous testing
5. Dealing with failed tests - how to iterate



# Best Practices

## Dealing with low traffic

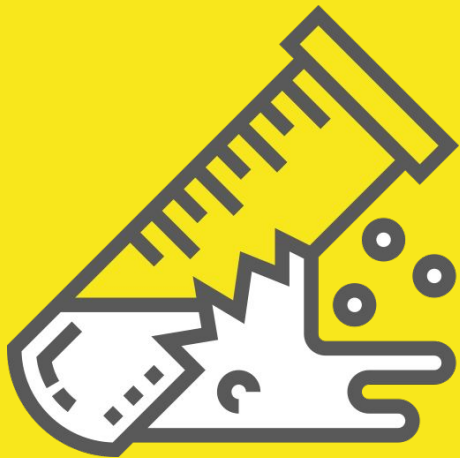
Run shorter tests with higher probability of visible change. When dealing with this scenario, terminate long tests as failed earlier rather than later. Analyze and have a good estimate of your traffic and site metrics before starting tests blindly.

## Coming up with strong/sensible hypothesis

Avoid frivolous tests to confirm a hunch. Save your bandwidth for high-impact tests.

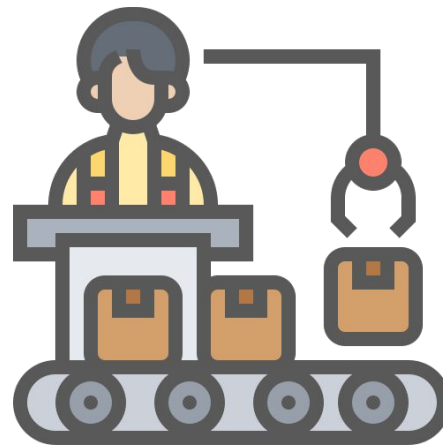
## Tests cannot be changed in between

This is a good principle that you should adhere to. If there is was a flaw in design parameters, it's always better to terminate rather than modify. Use past tests to inform test design. If necessary, run A/A tests to build confidence in experiment infrastructure.



## Dealing with failed tests - how to iterate

Always try and learn even from your mistakes, Re-design parameters and do not be afraid of re-running tests when necessary.

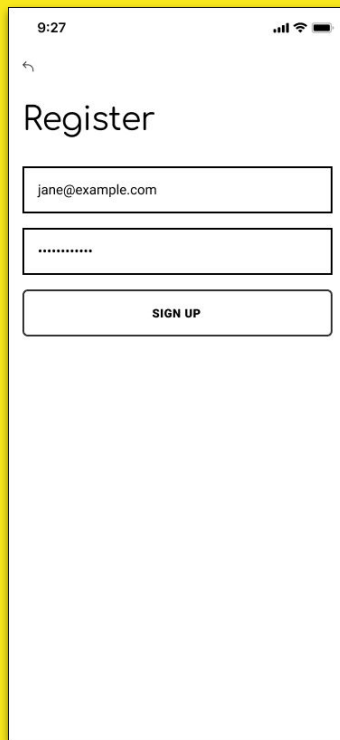


## Testing too many variants and simultaneous testing

Tooling is your friend here. Allocate traffic in a way that ensures no overlap between different tests.

Test your knowledge

# Is this an A/B test?



9:27

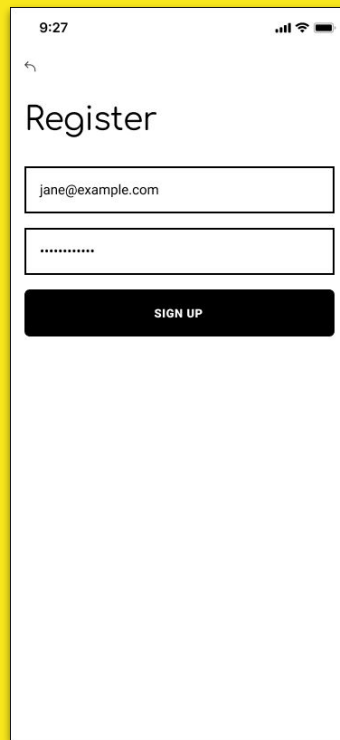
Register

jane@example.com

\*\*\*\*\*

SIGN UP

This is a mobile app registration screen. It features a white background with a status bar at the top showing the time as 9:27 and signal/battery icons. Below the status bar is a back arrow icon. The title "Register" is centered. There are two input fields: the first contains the email "jane@example.com" and the second contains masked text "\*\*\*\*\*". At the bottom is a white button with the text "SIGN UP".



9:27

Register

jane@example.com

\*\*\*\*\*

SIGN UP

This is a mobile app registration screen, identical to the one on the left but with a different button style. It features a white background with a status bar at the top showing the time as 9:27 and signal/battery icons. Below the status bar is a back arrow icon. The title "Register" is centered. There are two input fields: the first contains the email "jane@example.com" and the second contains masked text "\*\*\*\*\*". At the bottom is a black button with the text "SIGN UP" in white.

# Is this an A/B test?



YES

9:27

Register

jane@example.com

\*\*\*\*\*

SIGN UP

9:27

Register

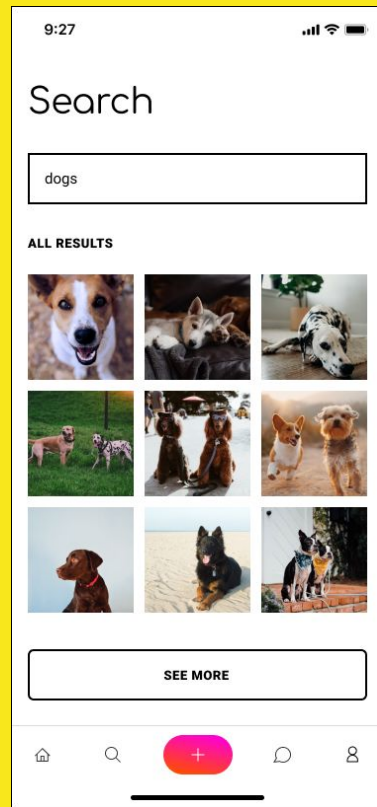
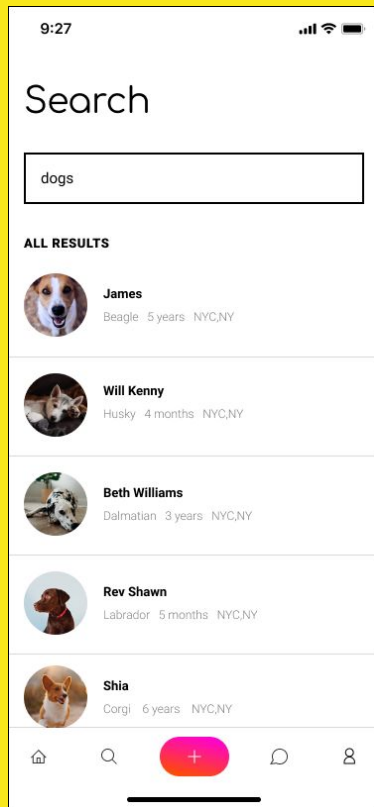
jane@example.com

\*\*\*\*\*

SIGN UP

Sensible Assumption: Sign-up is low due to non-obvious registration button

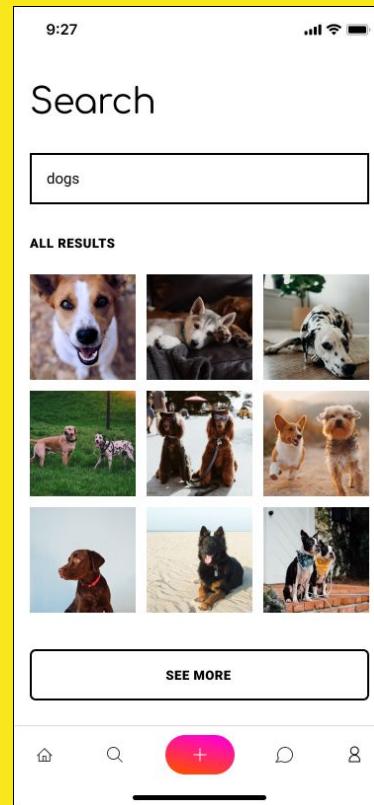
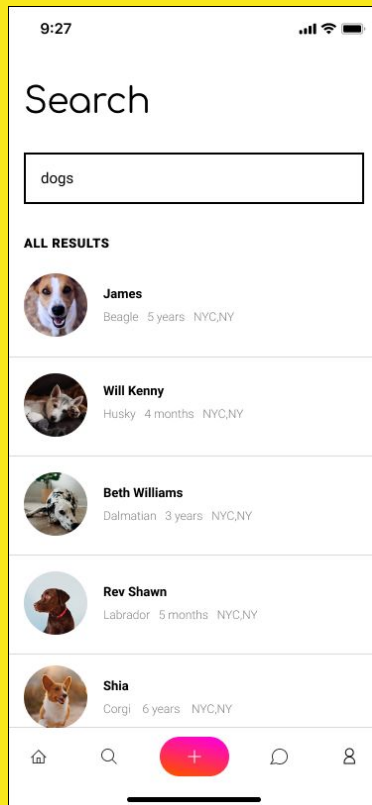
# Is this an A/B test?



# Is this an A/B test?

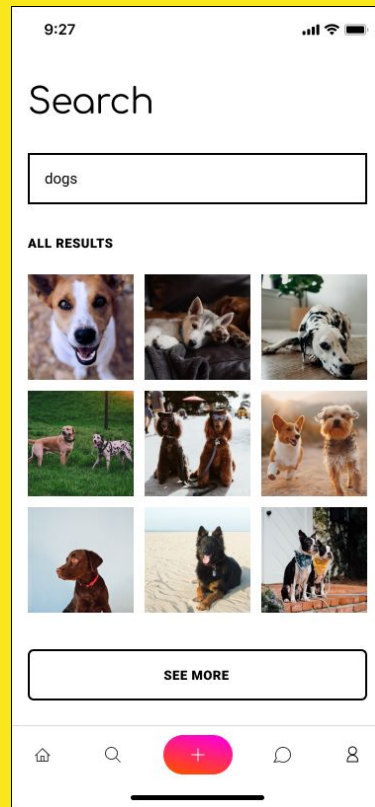
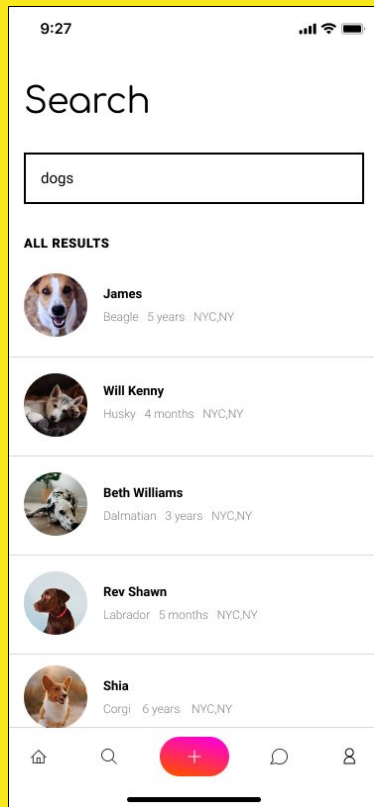


# YES



# Is this a “Good” A/B test?

NO





# Is this an A/B test?

## Last Seen



Creativity stimulating lotion.  
Drink every morning to  
generate better ideas!

**\$12.48**

Worldwide shifting available  
Buyers protection possible!

★★★★☆ 4.99

♡ Watch



Prototyping items to create  
a lot of unnecessary  
things...

**\$128.99**

★★★★☆ 4.87

♡ Watch



John Von Ebalkin SPRING

**\$13.95**

1258 bids, 359 watchers  
\$5.95 for shipping

★★★★☆ 4.56

♡ Watch



Vintage Typewriter to post  
awesome stories about UI  
design and webdev.

**\$49.50**

Eligible for Shipping To Mars or  
somewhere else

★★★★☆ 4.05

♡ Watch



Lee Pucker design. Leather  
botinki for handsome  
designers. Free shipping.

**\$13.95**

1258 bids, 359 watchers  
\$5.95 for shipping

★★★★☆ 4.56

♡ Watch



Timesaving kitten to save  
months on development.  
Playful, cute, cheap!

**\$128.99**

★★★★☆ 4.87

♡ Watch

# Is this an A/B test?

NO



## Last Seen



Creativity stimulating lotion.  
Drink every morning to  
generate better ideas!

**\$12.48**

Worldwide shifting available  
Buyers protection possible!

★★★★☆ 4.99

♡ Watch



Prototyping items to create  
a lot of unnecessary  
things...

**\$128.99**

★★★★☆ 4.87

♡ Watch



John Von Ebalkin SPRING

**\$13.95**

1258 bids, 359 watchers  
\$5.95 for shipping

★★★★☆ 4.56

♡ Watch



Vintage Typewriter to post  
awesome stories about UI  
design and webdev.

**\$49.50**

Eligible for Shipping To Mars or  
somewhere else

★★★★☆ 4.05

♡ Watch



Lee Pucker design. Leather  
botinki for handsome  
designers. Free shipping.

**\$13.95**

1258 bids, 359 watchers  
\$5.95 for shipping

★★★★☆ 4.56

♡ Watch



Timesaving kitten to save  
months on development.  
Playful, cute, cheap!

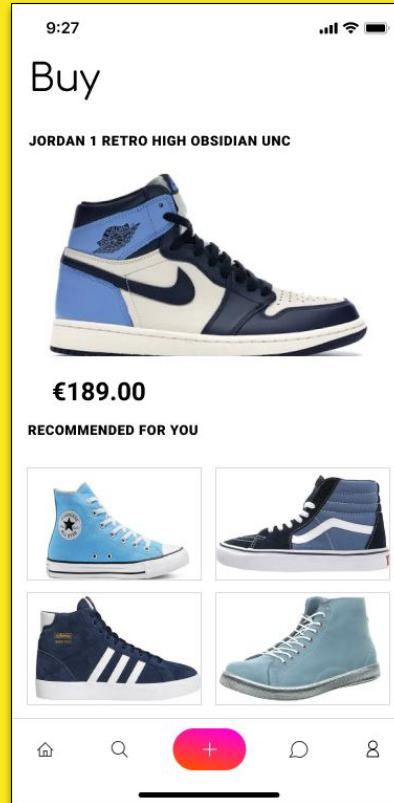
**\$128.99**

★★★★☆ 4.87

♡ Watch

## Last Seen

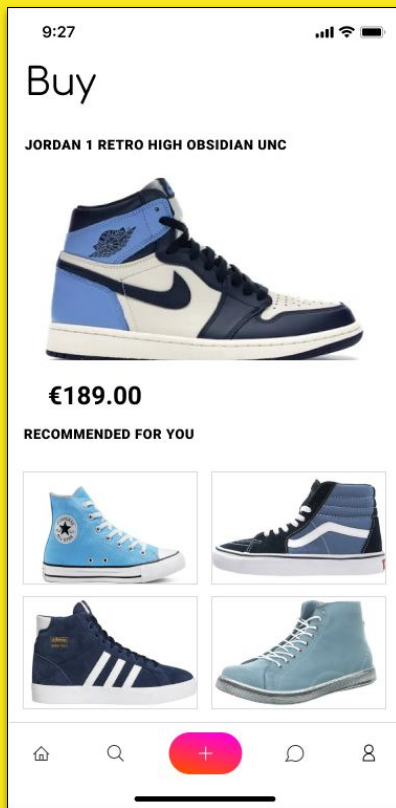
# Is this an A/B test?



# Is this an A/B test?



# YES



When Data gets Dirty

# Real World Problems #1

Are groups used in testing/experimentation really comparable?

- It is quite possible to collect certain user-level information: location, postal code.
- However, we can never *truly* determine gender, ethnicity or the age of the visitor/shopper
- Explicitly request this information?
  - Age (Maybe)
  - Gender (Poor Practice)
  - Ethnicity (Controversial!!)

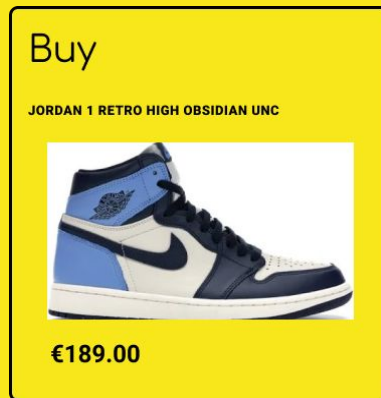
## The Solution

User segmentation! Segments are product-level groups of customers that display certain behavioral preferences. This can be based on frequency, device type, browser type, location etc

# Real World Problems #2

When split testing is impossible:

- Causes a scenario of differential pricing



## The Solution

Well quite simply, this is a tale in caution while designing the test. You want to ensure that customers under same conditions never experience differing experiences.

If you do have to test prices, be smart and find work-arounds!

# Real World Problems #3

Unrealistic/ magnified effects

- Observing uplift the seem unrealistic?
- Do your tests perform much better than expected?

This is is often a scenario of mixed effects in play. Promotions, cannibalization and sales need to be handled.

**Best Practices:**

1. Model mixed effects with a linear model to help estimate the effect of external variables
2. Use latent variables
3. Matching and stratified testing to ensure visible properties and distributions are comparable



# Real World Problems #4

Are there any situations/ scenarios where not testing is a good idea? Why?

- Examining seasonality
- Losing money during peak traffic seasons



Best Practice: Testing blackouts!

# Limitations with A/B Testing

# Back-end testing

We just ran in to our first major roadblock with A/B testing. How do we deal with testing changes to our back-end?

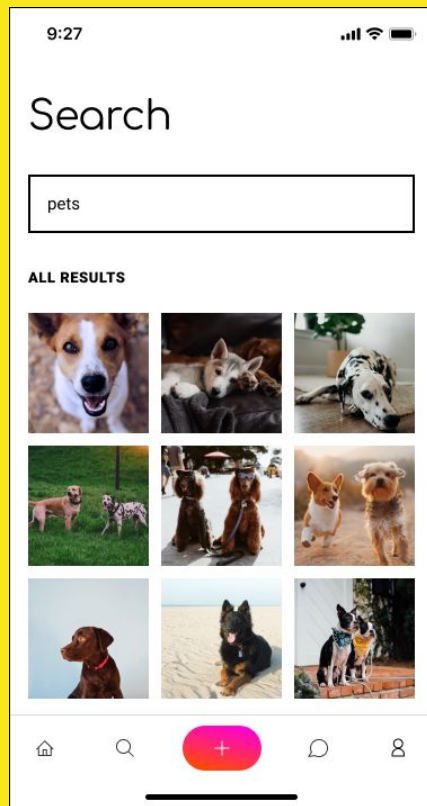
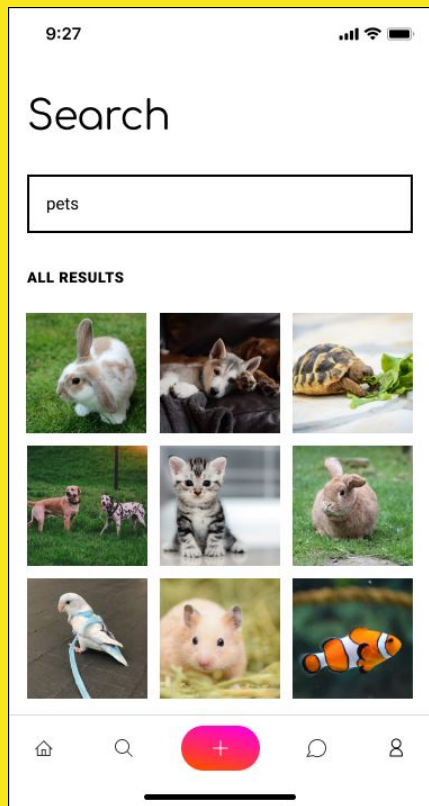
- What's a back-end change?

Simply put any change that is not visible to your customer is a back-end change.

- Right, so the customer is supposed to be indifferent to the change?

Yes and no. While the customer should never know what changes you rolled out in your back-end. Certain improvements to the website; a change in search indexing, different recommendation strategy will/ and should impact your customer's behavior.

# You searched for.....

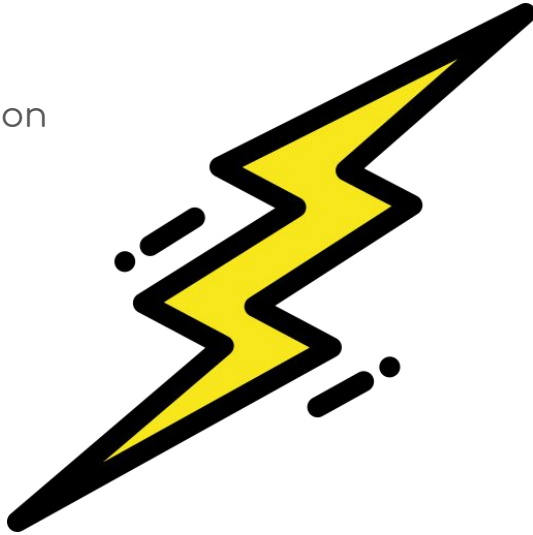


# Black-box testing

A black-box is used to define an object/mechanism, of which the viewer has no idea of the inner workings. The user can observe an input and the resultant output.

## Step 1 (input)

Ask the 8-ball a question and shake.



## Step 2 (output)

Read response to your question

# A Concrete Example - SEO Testing

When Larry Page and Sergey Brin started Google in 2001, it was well documented that they made use of the PageRank algorithm to score the relevance of a page to a search query.

Nowadays, with affiliate marketing (promoted results) , natural language processing in search, localization and personalization - it is no longer possible to gauge how a page will rank on a query.

Ongoing challenge for SEO teams to determine what keywords and titles lead to better ranking for this page. Google is a black-box.

# D-in-D testing to the Rescue

What to do when you cannot create comparable groups?

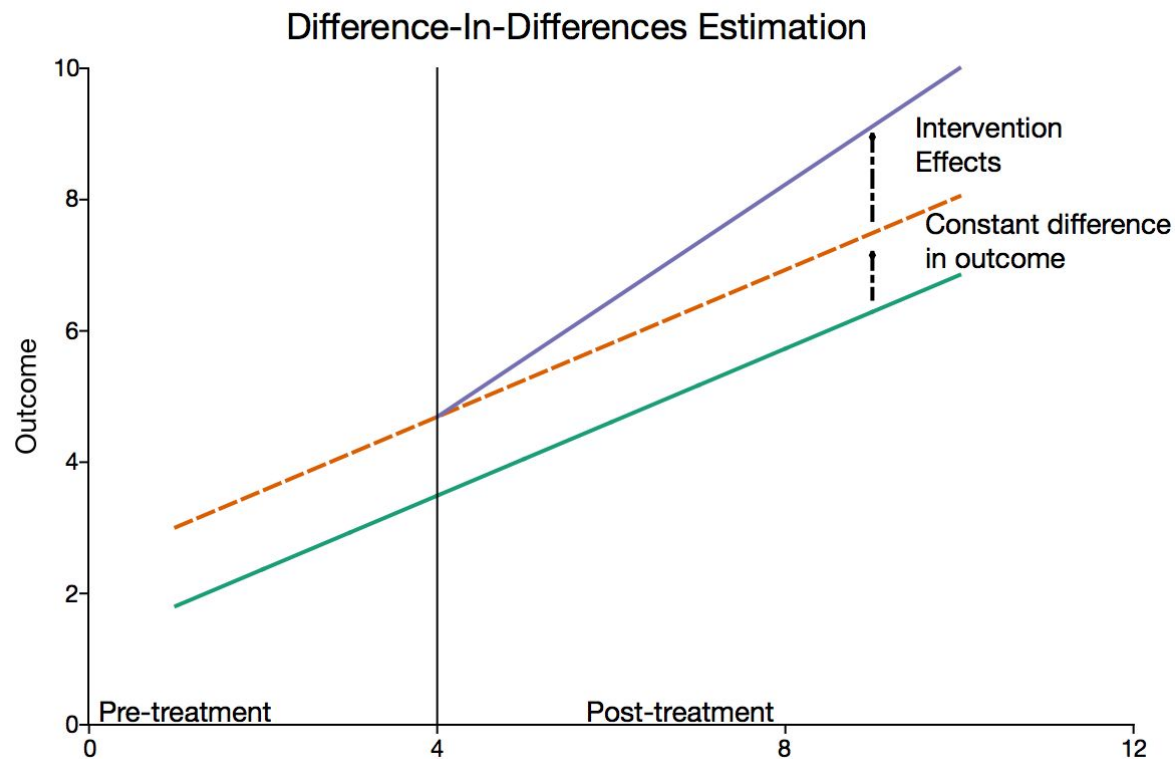
- Does one tag-title style lead to better ranking on average?
- Different users will input different search queries to land on the page
- What are the metrics in such a test?

Fortunately, difference-in-difference estimation, another form of parallel testing changes, used in econometrics and social science research has recently made its way over to the product scientist's experimentation toolbox.

What is DiD testing and how does it work?

The goal is to mimic design of a controlled test using observational study data. By identifying two groups with parallel trends, a “control” vs a “variant”, the natural effect of a treatment can be estimated using pre and post-treatment difference of mean in the outcome variable.

# Visually....





## Who is using the DiD paradigm?



- For long-tail categories and products, A/B testing recommendation strategies often lead to insignificant results due to low traffic/ sales. DiD measures the effect of two groups of products powered by varying recommendation strategies.
- Pricing strategies can be tested similarly to recommendation strategies.

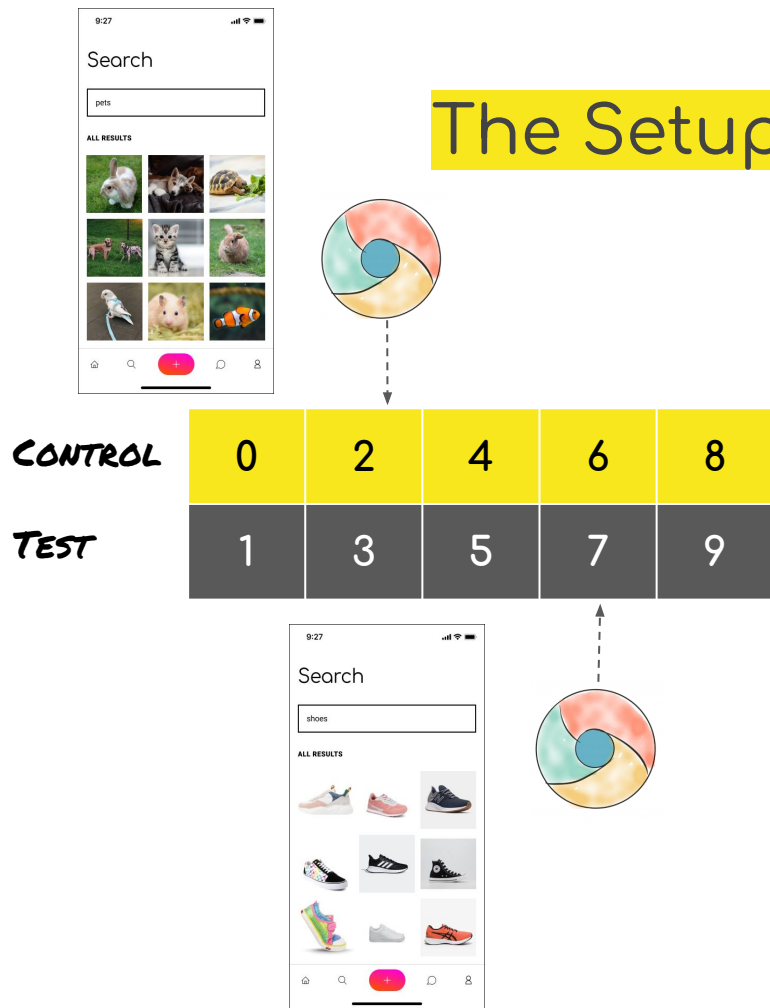


- SEO optimization is an ongoing topic. The team at Etsy constantly tests different title lengths, style, keywords, metadata and copy changes to improve ranking of product and shop pages on Google search.

# The Setup of DiD Experiments

## Step 1

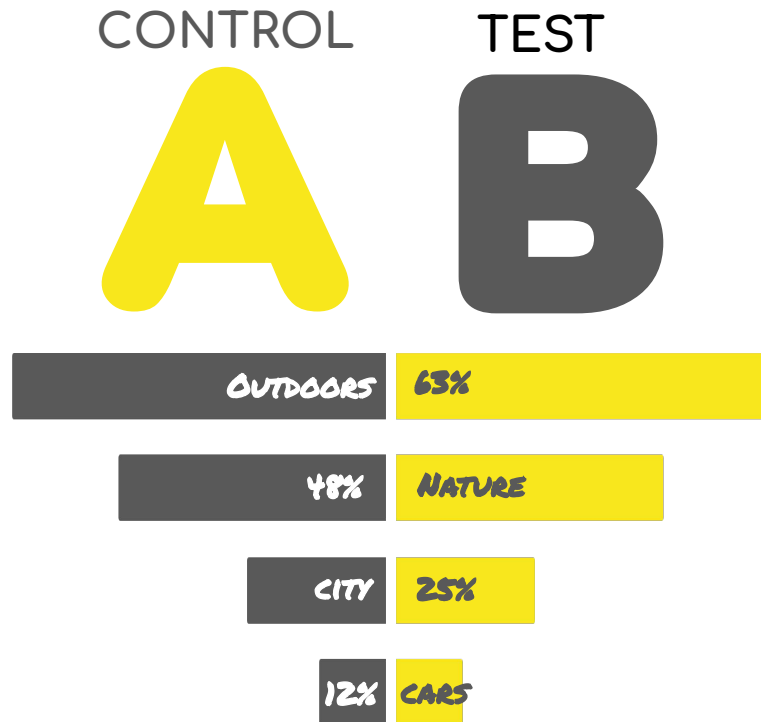
- While A/B testing is performed on user groups, DiD testing requires grouping a subset of pages in to “control” and “test”. Determine the fraction of pages that will be included in the test.



# The Setup of DiD Experiments

## Step 2

- Randomly assign pages in this subset to either test or control. Ensure that groups are comparable on underlying properties (similar means). If not, repeat till we have desired comparable groups.



# The Setup of DiD Experiments

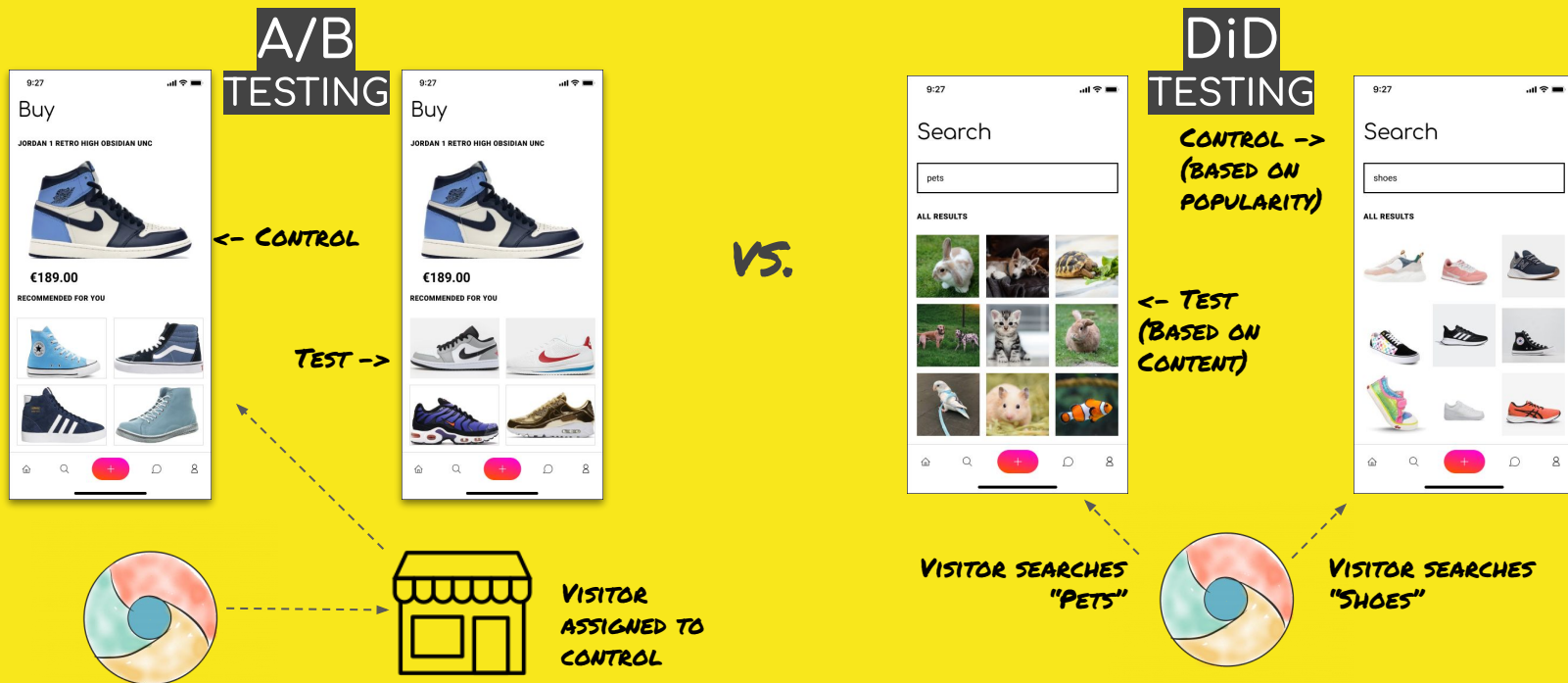
## Step 3

- Determine the horizon for your test. Apply treatment to the “test” group (commencing the test).

## Step 4

- On termination of the test, the difference of the means of control and test group provides an estimate of treatment “impact”.

# In Production



A visitor arrives on the website, and in real-time, is assigned to either CONTROL or TEST group.

Visitor arrives via search engine results, depending on the query (and which group the resultant page belongs to) is served either TEST or CONTROL.

# Evaluation

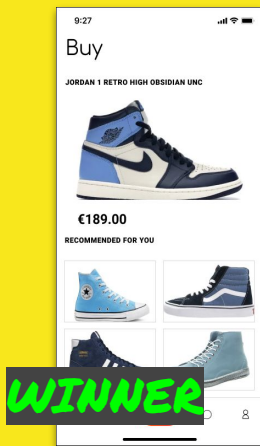
## A/B TESTING

At conclusion of the test, compare performance of CONTROL vs. TEST on the basis of predetermined metrics. If there is a significant lift in metrics, this should be clearly visible

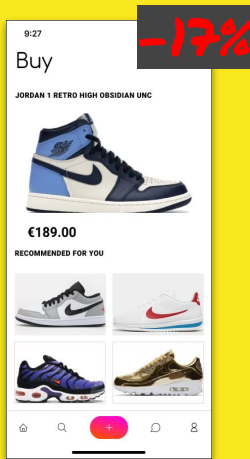
VS.

## DiD TESTING

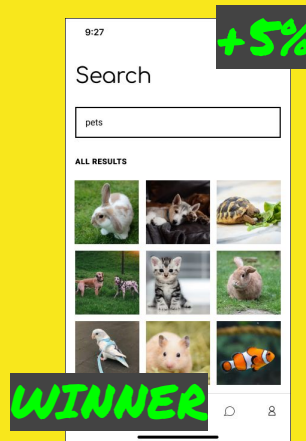
At conclusion of test, compare the means of the CONTROL group with the mean of the TEST group. The difference in means is an estimate of the effect or impact of the treatment (applied exclusively to the test group).



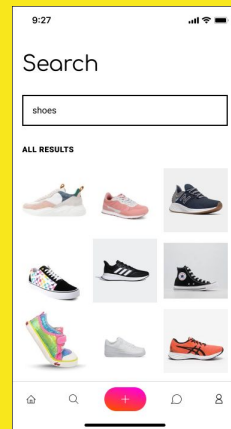
CONTROL



TEST



TEST  
(BASED ON CONTENT)



CONTROL  
(BASED ON POPULARITY)

# Outro

A summary of  
learnings/takeaways from  
this talk

- We examine two experimentation frameworks that are commonly used in practice to test decisions and evaluate the impact of the planned change.
- We walkthrough the general test design and setup of these experiments, drawing comparisons between the two.
- Finally, we look at real world examples of common challenges and pitfalls, suggesting best practices to navigate the perilous journey that is product experimentation.

# SHOUT-OUTS & CREDITS

DESIGN, GRAPHICS + THEME COURTESY:

[Saumya Bhatt](#)

IMAGES/ICONS SOURCED FROM:





Going Bayesian!

# A Tale of Two Methodologies

- Frequentists vs. Bayesians

## Common Example:

*Imagine a trial of tossing an **unbiased** coin 5 times. You observe the following sequence **HTHHT**. Before tossing the coin a sixth time, if you were asked to place a bet. Would you call heads or tails?*

### The Conundrum:

Given the series of occurrences, we might be biased towards believing that heads is more likely (0.6) vs. tails.

### Reality

Today, it is common knowledge that the probability of an unbiased coin landing heads or tails is 50/50.

This is the Bayesian Way!

# Bayesian Tooling

## The Beta Distribution

- So it's a distribution. Big deal!

PLAYER	RUNS	BALLS PLAYED	MATCHES
S. Tendulkar	17000	35000	721
V. Kohli	700	1000	89

\*ASSUME DATA  
IS FROM 2012 PDF



If we were to calculate a batting average, Kohli would easily beat Tendulkar. However, as we know, a player's performance deteriorates with age. Additionally, the larger sample size of Tendulkar's history does not allow for direct comparison. What then? Are we to conclude that the two can never truly be compared?

The Beta distribution incorporates *prior expectation*. Using this we can estimate what Kohli's average would be, even before he decides to pick up a bat.

### FORMULATION

**Beta( $\alpha_0$  + hits,  $\beta_0$  + misses)**

# Putting it all to test

So how do we combine our learnings from the Beta distribution towards Bayesian A/B testing?

- Priors

There's an underlying distribution to every natural process. You may have learned the arrival rate of these processes in an advanced college statistics course.

- The underlying probability of rare events is modeled by the Chi distribution
- Arrival rate of calls is modeled by a Poisson
- The Beta is the general form of a binomial distribution, this generalizability offers it great versatility in modeling unknown priors.

# The Bayesian Testing Paradigm

VARIANT	VISITS	CONVERSIONS	CONVERSION RATE	FIXED CONV. RATE
CONTROL	12000	600	0.05	0.45
TEST	12500	612	0.049	0.46

- Let's observe the results of the assumed test, looking simply at the conversion rate, we would conclude that the CONTROL is a superior variant.
- However, the TEST variant clearly displays more conversions!
- Assuming that we know the priors of the beta distribution that models this business ( $\alpha_0=81$ ,  $\beta_0=219$ ).
- With the assistance of the formulation we examined earlier we compute the bayesian conversion rate as indicated in the final column.
- The results are now closer to expectation!