runtastic

Generating Business Value from A/B Testing

Workshop
November 29th, 2017

Hilda Kosorus
Data Scientist

About me

2010

JKU Linz: Research Assistant

time series analysis, text mining, semantic similarity, knowledge bases, ontologies, recommender systems

Doctoral studies

2013

JKU Linz: Senior Researcher

MDL, sequential patterns mining, compressing sequential semantic patterns

Post-doctoral studies

Teaching

2015

Runtastic: Data Scientist

data concepts & quality

KPI prediction

business analytics

data-driven product development

2017

Runtastic: Data Scientist

machine learning

A/B testing processes

forecasting at scale

DS mentoring

About Runtastic

runtastic

WHO WE ARE

We are

4

founders

We are

8

years old

We were profitable after just

20

months

We are over

235

employees

We come from

36+

unique countries

We have

3

offices in Linz, Vienna and Salzburg

Our products are available in up to

15

languages

We are

1 TEAM

and one shared vision

OUR VISION

We want every individual to live a more aware and active lifestyle with Runtastic, leading to a longer, happier life!

runtastic

24h

EACH DAY WE HAVE

150,000

NEW DOWNLOADS

WE HAVE MORE THAN

120 MILLION

REGISTERED USERS

WE BOAST OVER

230 MILLION

TOTAL APP DOWNLOADS

WE HAVE MORE THAN

3,472,757

FANS AND FOLLOWERS

OUR PORTFOLIO

The Runtastic Portfolio offers users a comprehensive and easy-to-use portfolio of health & fitness products, services and content.

About you

Let’s get to know each other

- What is your current position?
- What kind of topics are you working on?
- What do you expect from this workshop?
- Have you done any experimentation before?
  If so, what was your approach?

runtastic

A/B Testing Workshop

How we leverage experimentation

Before

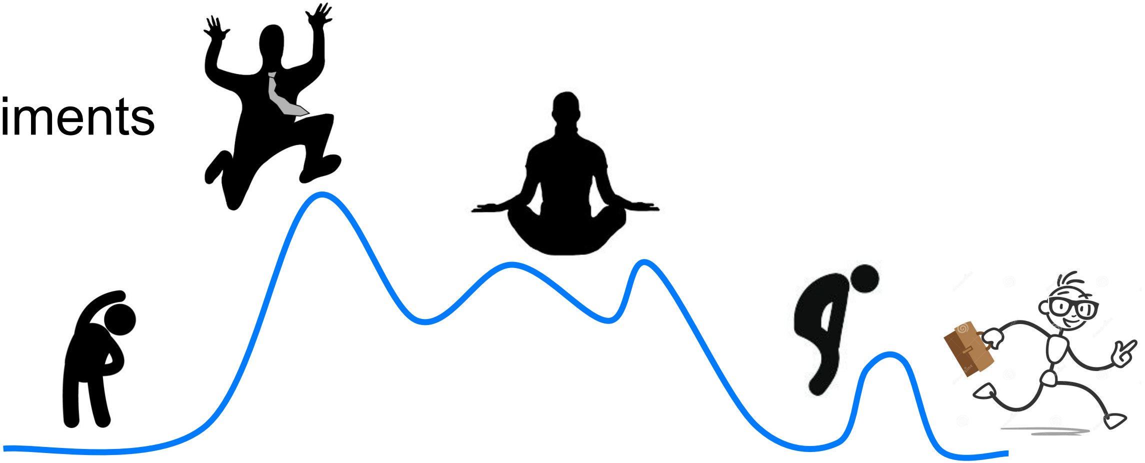
- Who: a few PMs
- What: selected app views
- Tools: limited functionality
- No learning process

Now

- Who: PMs, CRM, content marketing
- What: products, features, communication, blog
- Coordinated by data scientists
- Tools: external & in-house
- Knowledge sharing & learning
- Community

What you should expect

- Basic intro and examples
- Theory
- Checklist & guidelines
- Learning from past experiments
- Common pitfalls
- Practice



What is A/B testing?

A/B testing is NOT…

Validation of guesswork

Consumer psychology tactics

GI → GO

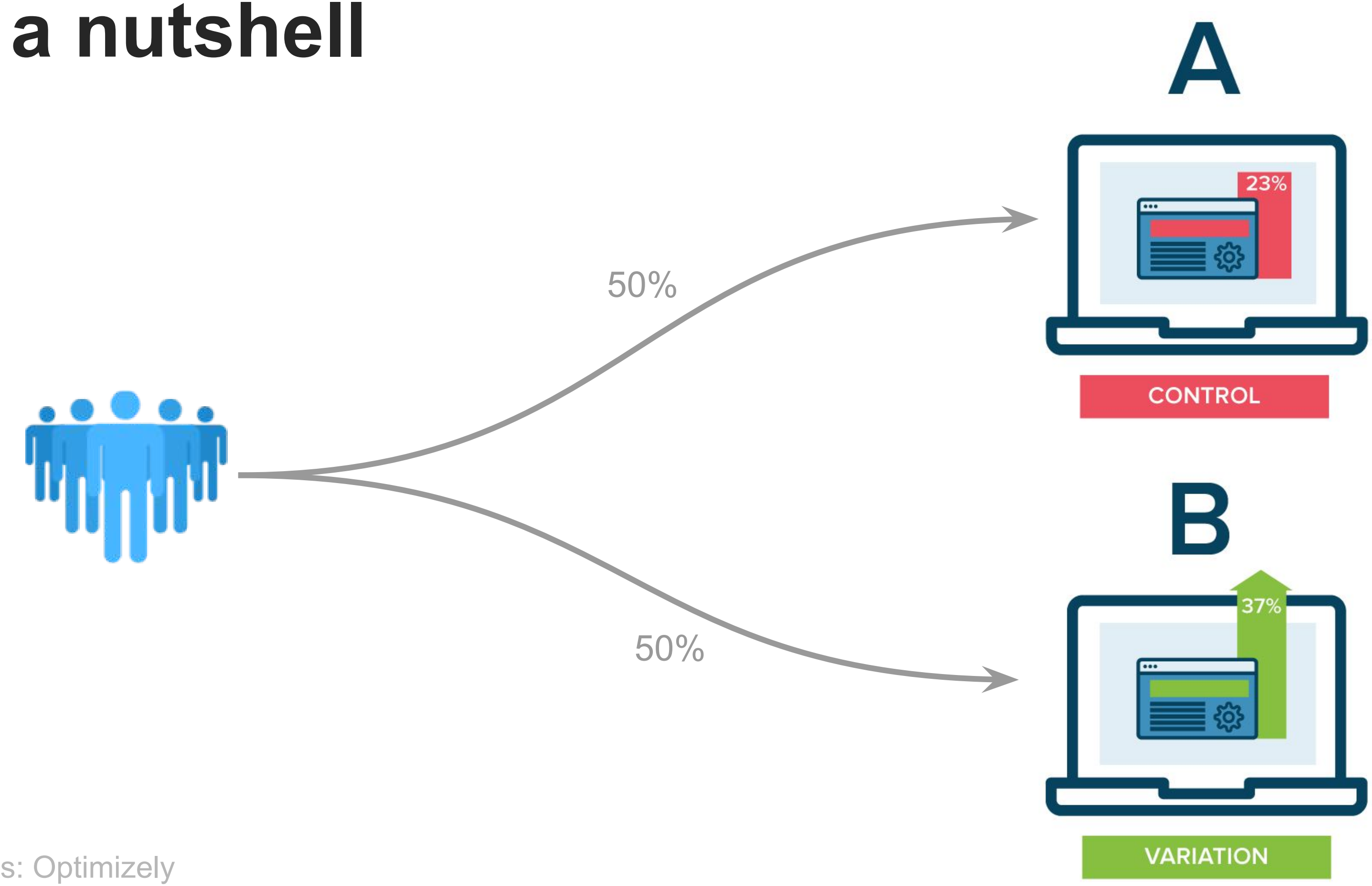
Meek tweaking

What is A/B testing?

Conducting experiments to optimize customer experience

OR

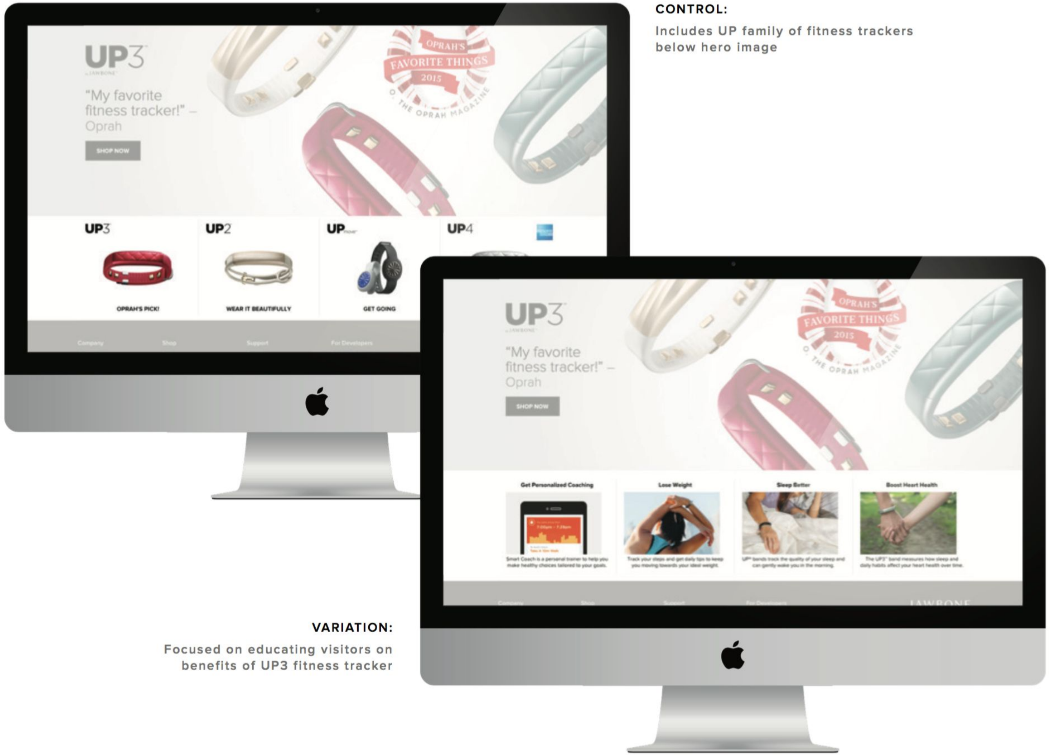
In a nutshell



*Credits: Optimizely

Examples

Jawbone’s homepage



CONTROL:

Includes UP family of fitness trackers below hero image

VARIATION:

Focused on educating visitors on benefits of UP3 fitness tracker

Variation includes educational information regarding the high-level benefits of fitness trackers.

43%

Increase in Revenue Per Visit

24%

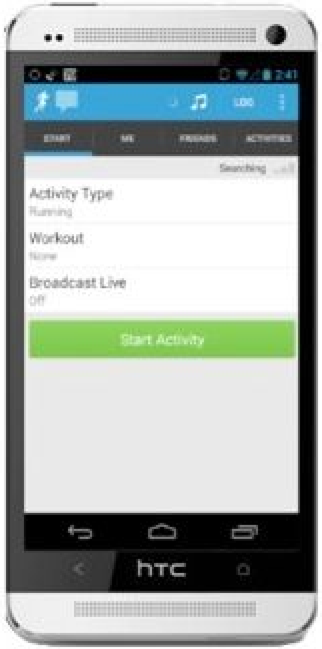
Increase in Revenue Per Visit on Mobile

*Credits: Optimizely

Runkeeper

Goal: increase retention on sport activities other than running.

Variant B increased retention on non-running activities with **235%**.

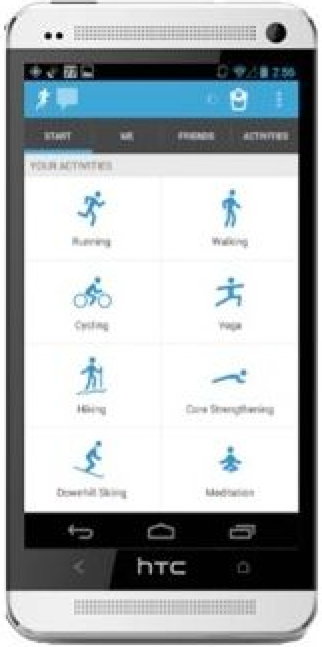


Activity Type
Running

Workout
None

Broadcast Live
Off

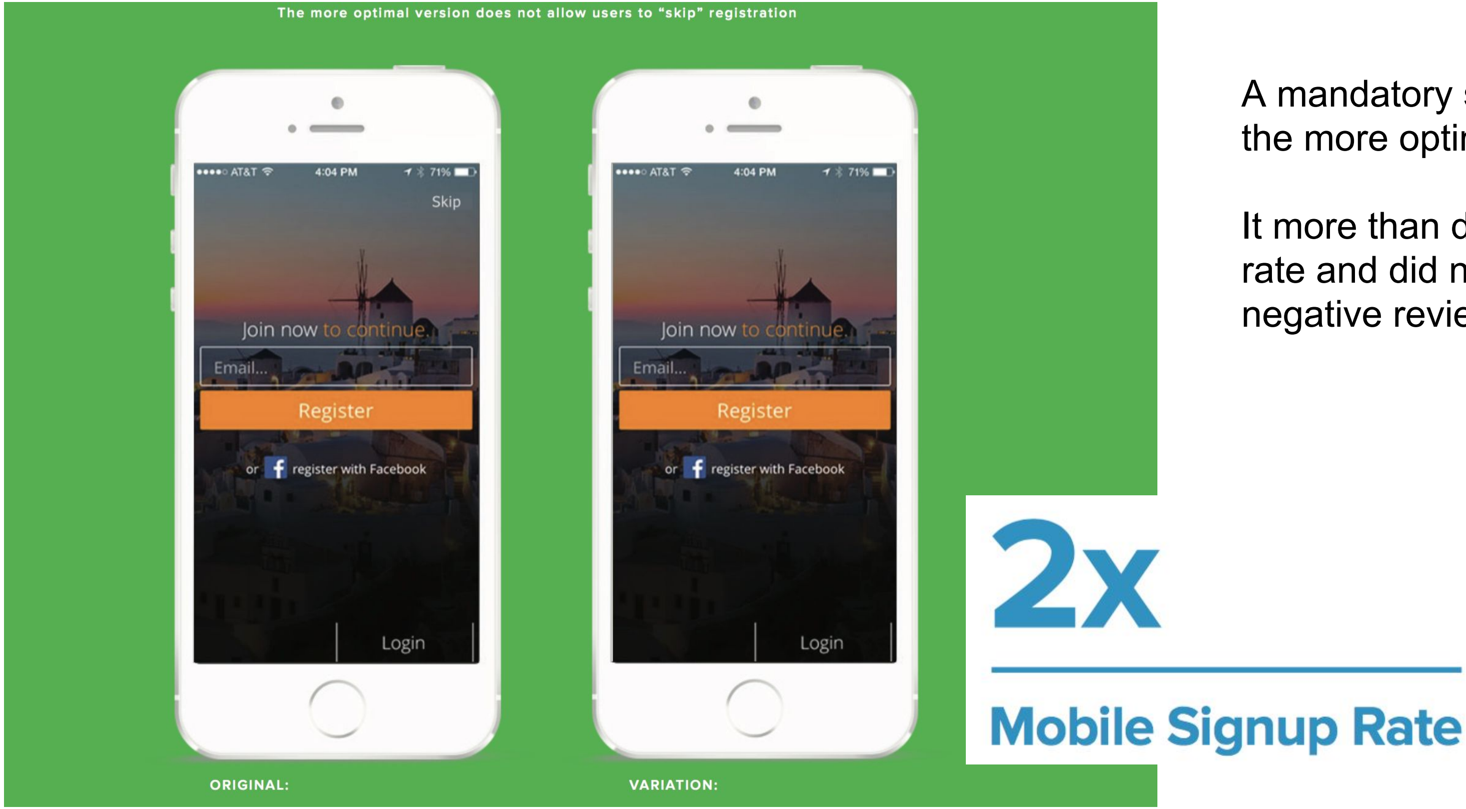
Version A



Version B (Winner)

*Credits: Optimizely

Secret escapes: registration



The more optimal version does not allow users to “skip” registration

2x

Mobile Signup Rate

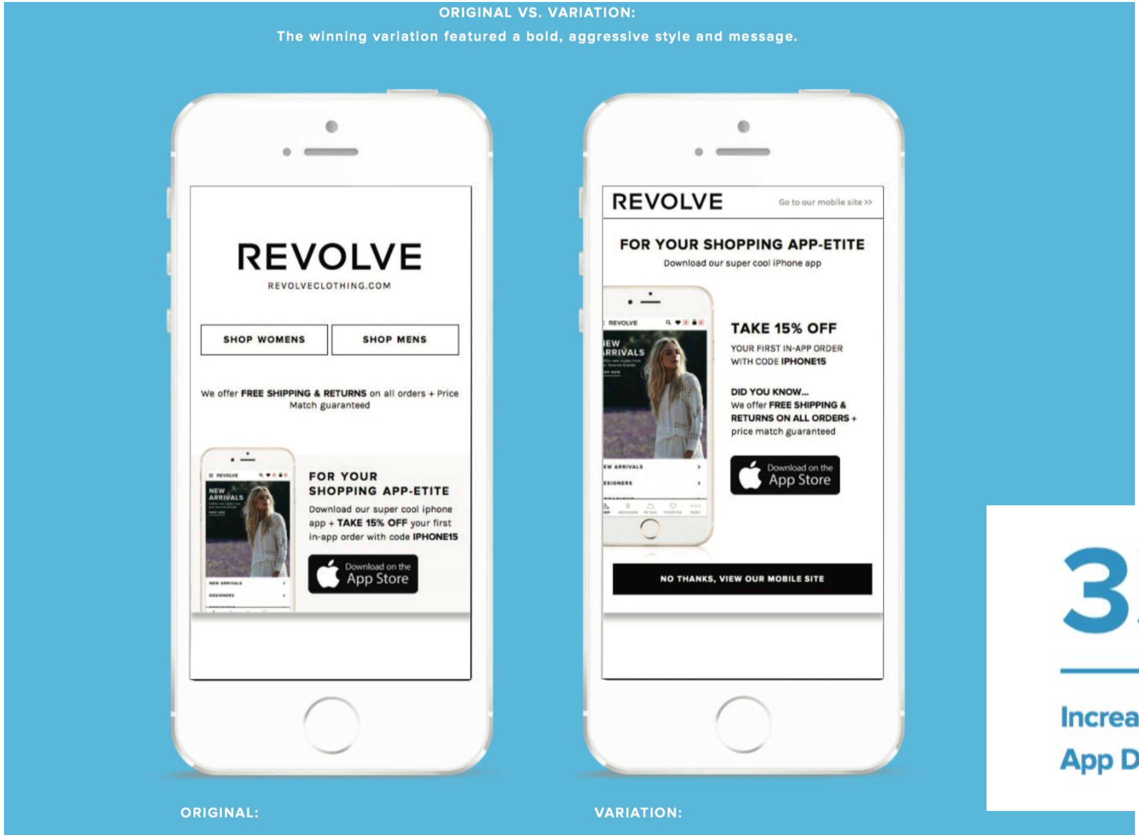
A mandatory signup gate was the more optimal experience.

It more than doubled the signup rate and did not lead to negative reviews or comments.

*Credits: Optimizely

runtastic

Revolve: mobile app downloads



ORIGINAL VS. VARIATION:

The winning variation featured a bold, aggressive style and message.

REVOLVE

REVOLVECLOTHING.COM

We offer FREE SHIPPING & RETURNS on all orders + Price Match guaranteed

FOR YOUR SHOPPING APP-ETITE

Download our super cool iphone app + TAKE 15% OFF your first in-app order with code IPHONE15

REVOLVE

FOR YOUR SHOPPING APP-ETITE

Download our super cool iPhone app

TAKE 15% OFF

YOUR FIRST IN-APP ORDER WITH CODE IPHONE15

DID YOU KNOW...

We offer FREE SHIPPING & RETURNS ON ALL ORDERS + price match guaranteed

ORIGINAL:

VARIATION:

The variation with the aggressive splash page increased app downloads from the mobile site.

350%

Increase in Mobile App Downloads

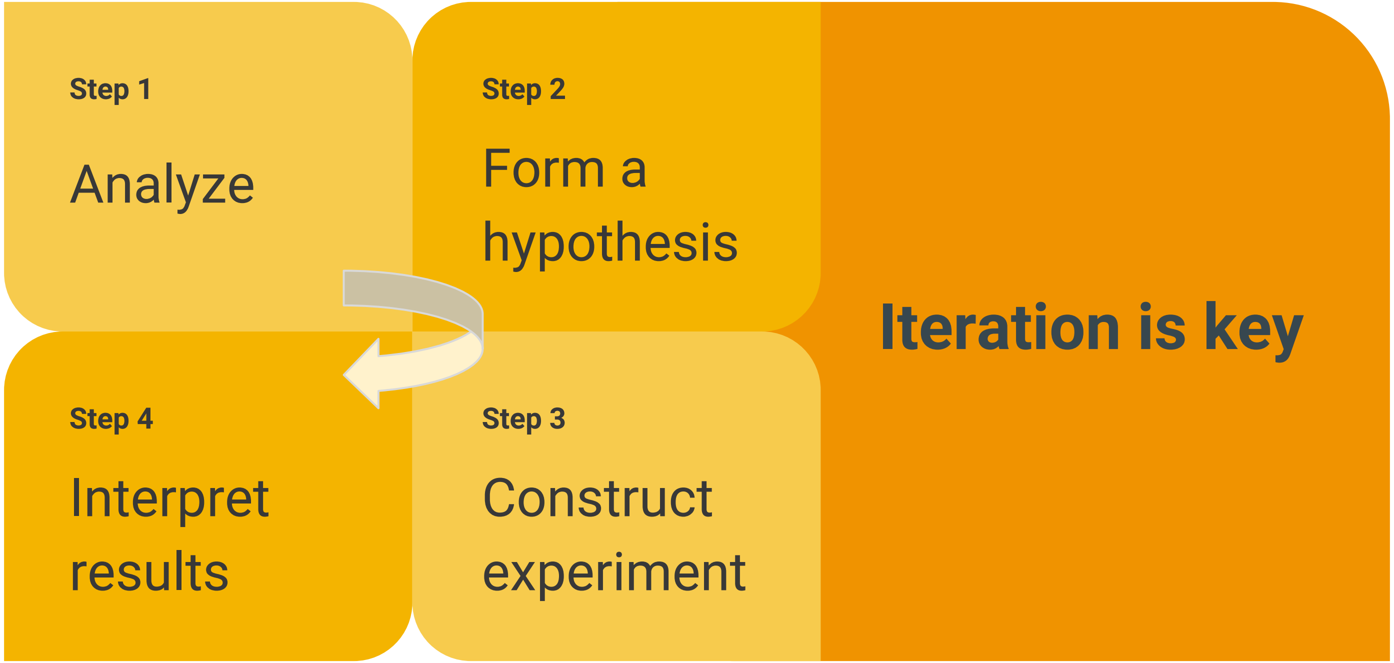
*Credits: Optimizely

Use case collection

- What would be an interesting experimentation use case?
- What goal(s) would you like to achieve?
- Why do you think your treatment would work better?
- How do you plan to make use of the result?

Overview

The 4 steps of A/B testing



Step 1

Analyze

Step 2

Form a hypothesis

Step 3

Construct experiment

Step 4

Interpret results

Iteration is key

Statistical background

Frequentist Approach

- Make predictions on underlying truths of the experiment using **only data from the current experiment**.
- **Goal of an A/B test**: determine whether the data collected during the experiment can conclude that one variation is **measurably different** from the other.

Terminology

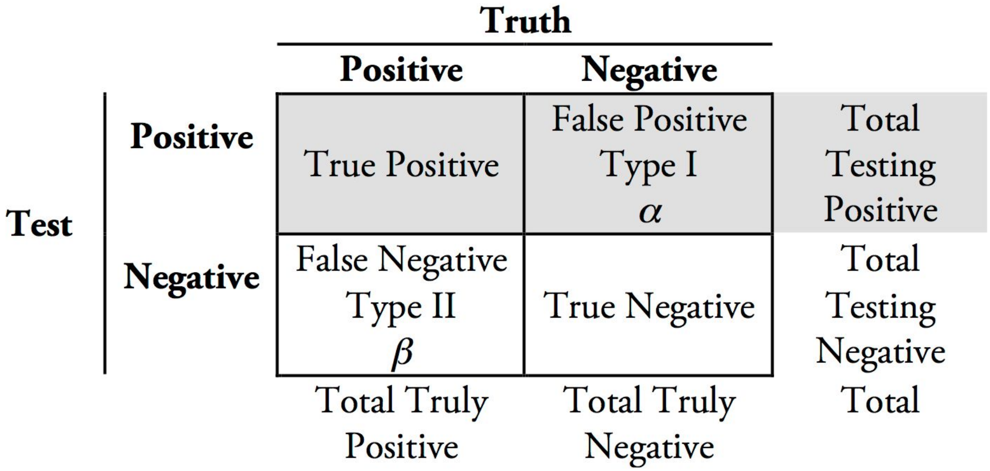
- Control (A), treatment (B), variants (B1, B2, …)
- Sample size (n)
- Baseline conversion rate
- Minimum detectable effect (MDE)
- Statistical significance level (α)
- Statistical power (1-β)
- Error types: false positive, false negative
- Hypothesis (null & alternative)
- P-value

Statistical Significance Level (α)

The probability of detecting a false positive (type I error).

The probability of detecting a difference, when there is actually no difference.

(in practice, $\alpha = 5\%$ or $\alpha < 5\%$)



		Truth		
		Positive	**Negative**	
Test	**Positive**	True Positive	False Positive Type I α	Total Testing Positive
	Negative	False Negative Type II β	True Negative	Total Testing Negative
		Total Truly Positive	Total Truly Negative	Total

Statistical Power (1-β)

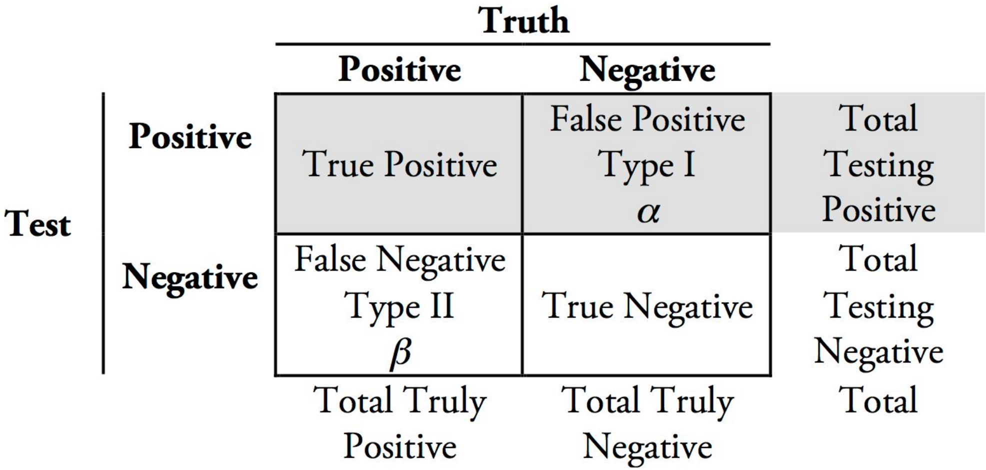
The probability that a statistical test will detect a difference between the control and variant, if there really is such a difference.

The probability to detect a true positive result.

Power = 1 - β

β = probability of detecting a type II error

(in practice, power = 80% or > 80%)



		Truth		
		Positive	Negative	
Test	Positive	True Positive	False Positive Type I α	Total Testing Positive
	Negative	False Negative Type II β	True Negative	Total Testing Negative
		Total Truly Positive	Total Truly Negative	Total

Start with a hypothesis

A **hypothesis** is an assumption made with little/no evidence.

Null hypothesis (H0) = “There is NO difference between A and B”

Alternative (Ha) = “There is a difference between A and B”

Hypothesis testing:

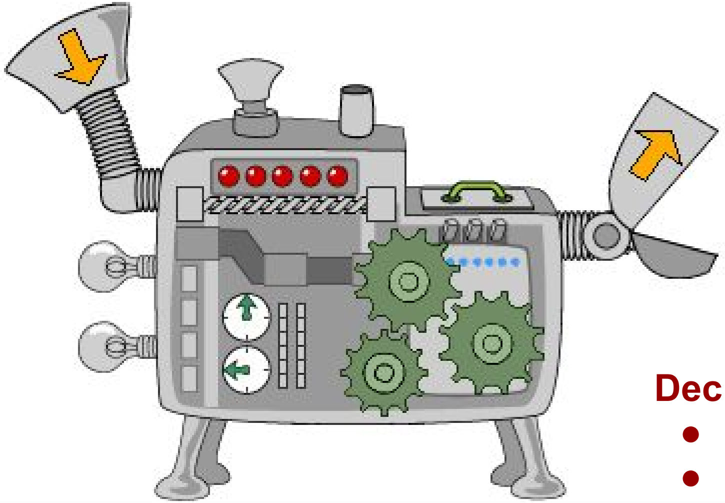
1. Assume H0 is true
2. Collect data for A and B
3. Reject H0 if A is *significantly* different from B
   "My data *looks* different than the null hypothesis."

Hypothesis testing

H0: (A = B)

Data (e.g. conversions)

Error threshold (α, β)



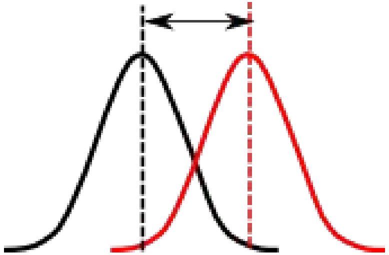
P-value

Objective indicator: how consistent your data are with the H0

Decision:

- If $\boldsymbol{p < \alpha}$, then **reject** H0
- Otherwise, accept H0

T-test

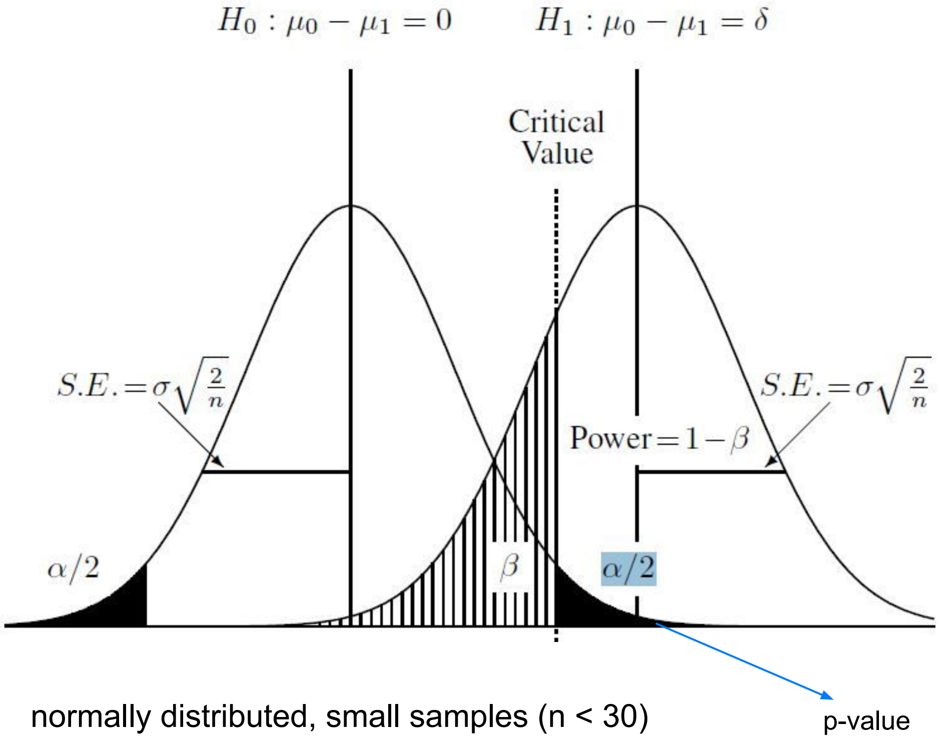


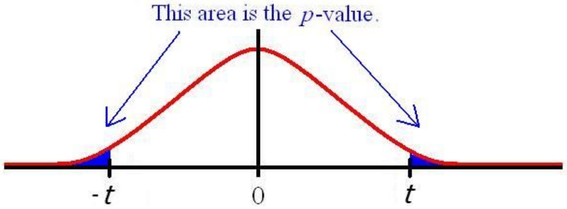
- Compares two **averages** (means) and tells you if they are different from each other.
- The t-test also tells you how **significant** the differences are:
  whether those differences could have happened by chance (p-value).
- Note: Assumes that the two samples are **normally distributed**
  (alternative: Mann-Whitney U test)

Example: Membership up-selling

- Control: no offer
- Treatment: -40% offer, 15 mins after paywall visit with no resulting purchase
- Target metric: conversion rate

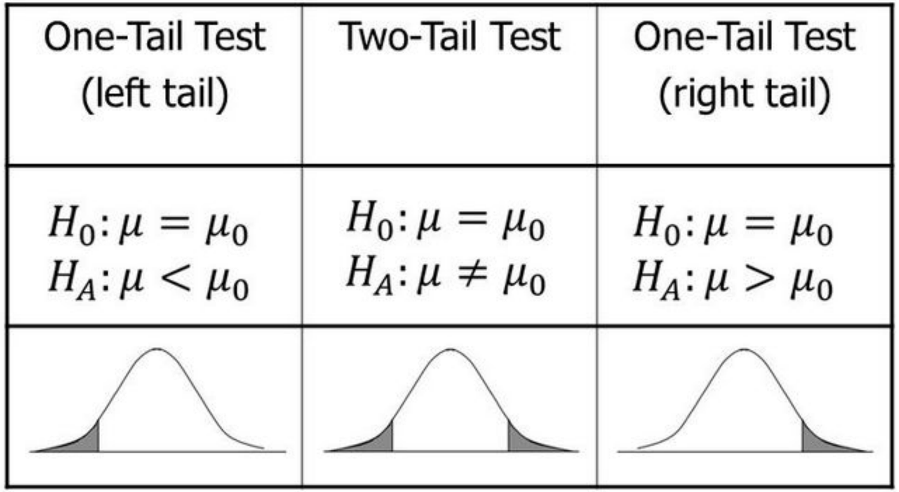
T-test


$H_0 : \mu_0 - \mu_1 = 0$
$H_1 : \mu_0 - \mu_1 = \delta$
Critical Value
$S.E. = \sigma\sqrt{\frac{2}{n}}$
$S.E. = \sigma\sqrt{\frac{2}{n}}$
Power $= 1 - \beta$
$\alpha/2$
β
$\alpha/2$
p-value



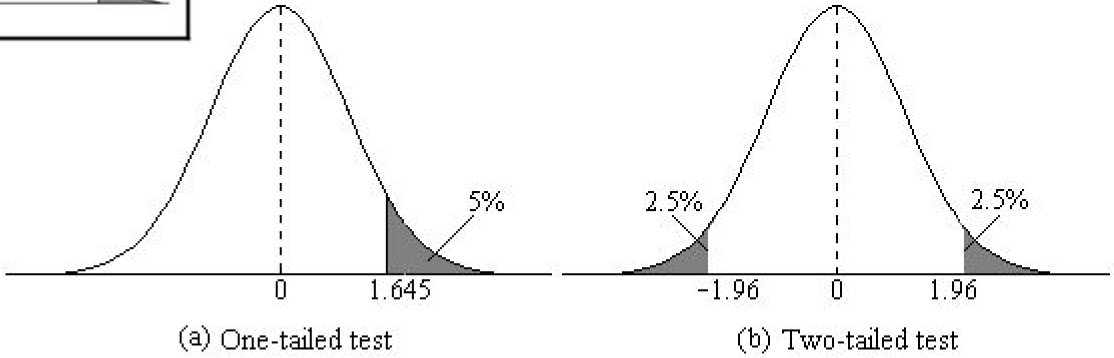
- normally distributed, small samples (n < 30)
- standard deviation unknown

One-tailed vs. two-tailed tests



One-Tail Test (left tail)	Two-Tail Test	One-Tail Test (right tail)
$H_0: \mu = \mu_0$ $H_A: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$	$H_0: \mu = \mu_0$ $H_A: \mu > \mu_0$

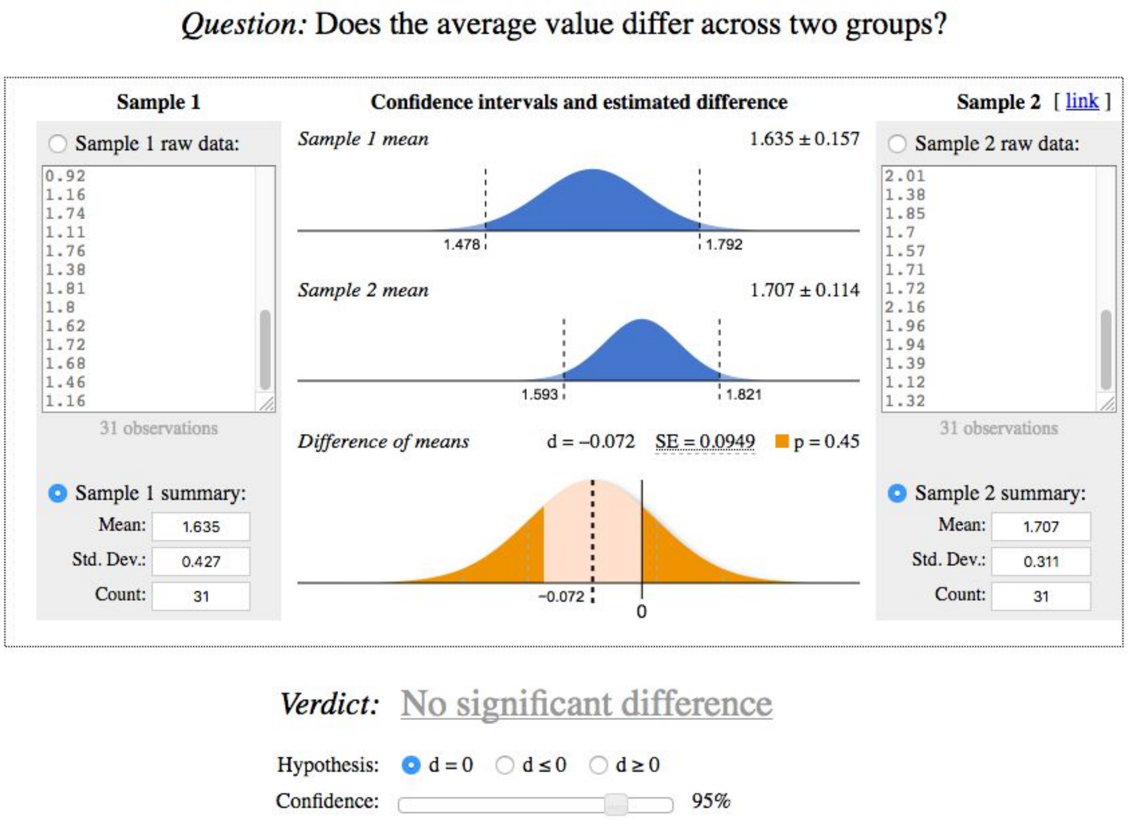
- one-tailed tests look for effects only in one direction (can miss opposite effects)
- are more prone to false positive errors



(a) One-tailed test

(b) Two-tailed test

Two-tailed t-test: Example



Question: Does the average value differ across two groups?

Sample 1

○ Sample 1 raw data:

```
0.92
1.16
1.74
1.11
1.76
1.38
1.81
1.8
1.62
1.72
1.68
1.46
1.16
```

31 observations

● Sample 1 summary:

Mean:	1.635
Std. Dev.:	0.427
Count:	31

Confidence intervals and estimated difference

Sample 1 mean 1.635 ± 0.157

Sample 2 mean 1.707 ± 0.114

Difference of means d = −0.072 SE = 0.0949 🟧 p = 0.45

Sample 2 [link]

○ Sample 2 raw data:

```
2.01
1.38
1.85
1.7
1.57
1.71
1.72
2.16
1.96
1.94
1.39
1.12
1.32
```

31 observations

● Sample 2 summary:

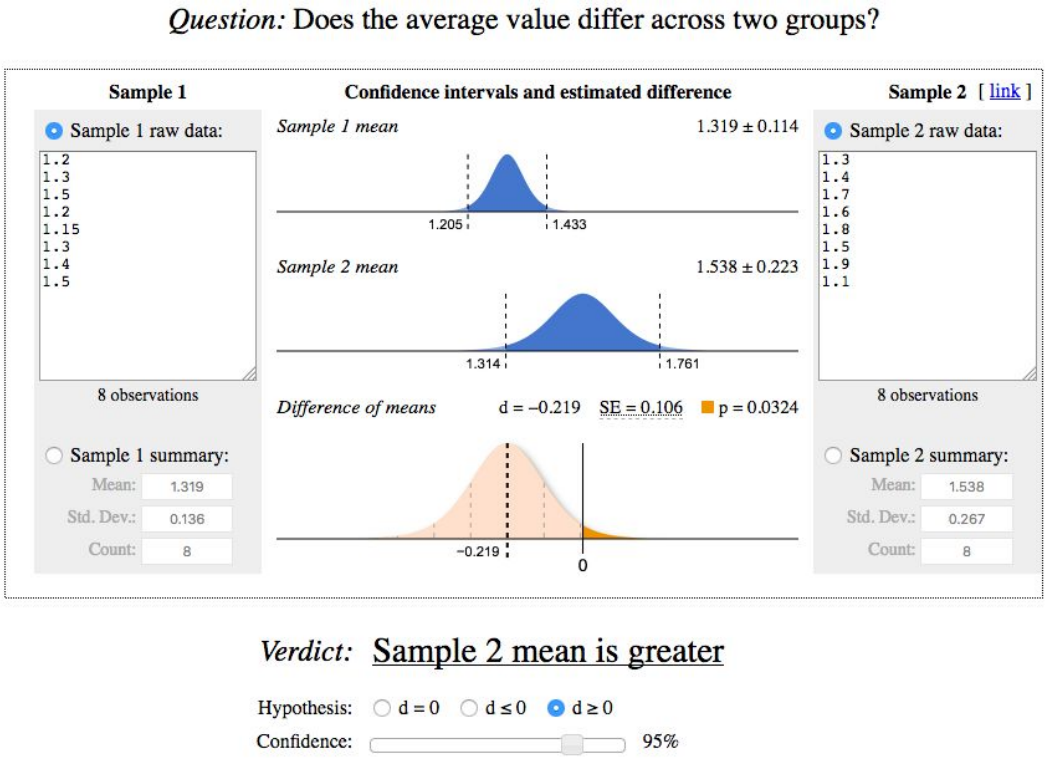
Mean:	1.707
Std. Dev.:	0.311
Count:	31

Verdict: No significant difference

Hypothesis: ● d = 0 ○ d ≤ 0 ○ d ≥ 0

Confidence: 95%

Experimenting with t-tests



Question: Does the average value differ across two groups?

Sample 1

● Sample 1 raw data:

```
1.2
1.3
1.5
1.2
1.15
1.3
1.4
1.5
```

8 observations

○ Sample 1 summary:

Mean:	1.319
Std. Dev.:	0.136
Count:	8

Confidence intervals and estimated difference

Sample 1 mean 1.319 ± 0.114

Sample 2 mean 1.538 ± 0.223

Difference of means d = −0.219 SE = 0.106 🟧 p = 0.0324

Sample 2 [link]

● Sample 2 raw data:

```
1.3
1.4
1.7
1.6
1.8
1.5
1.9
1.1
```

8 observations

○ Sample 2 summary:

Mean:	1.538
Std. Dev.:	0.267
Count:	8

Verdict: Sample 2 mean is greater

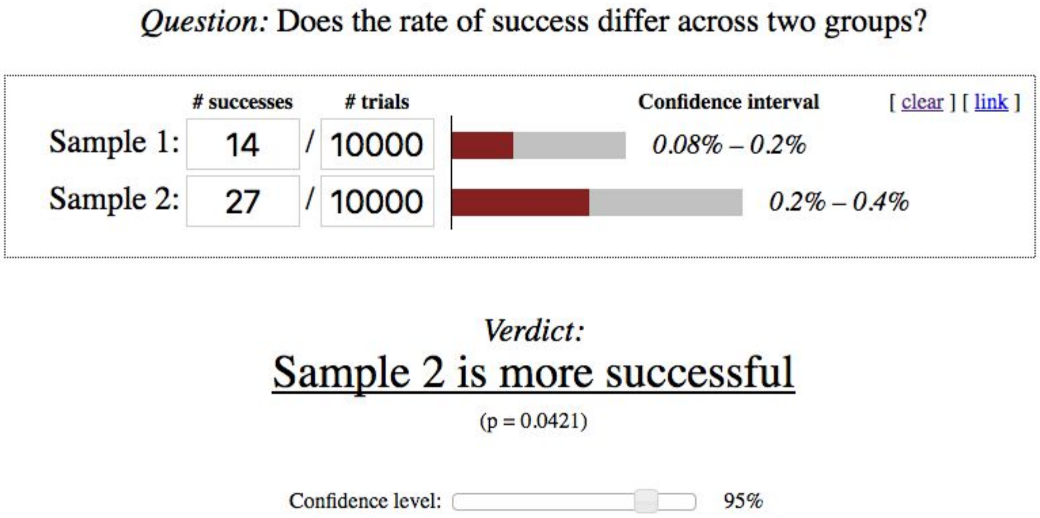
Hypothesis: ○ d = 0 ○ d ≤ 0 ● d ≥ 0

Confidence: 95%

http://www.evanmiller.org/ab-testing/t-test.html

Chi-squared test for independence

- Determine if two discrete variables are **associated** (e.g. gender and political affiliation)
- Used for A/B testing to test whether the rate of success across two or more groups is different
  - Test the **relationship** between treatment and conversion



Question: Does the rate of success differ across two groups?

	# successes		# trials	Confidence interval	[clear] [link]
Sample 1:	14	/	10000	*0.08% – 0.2%*	
Sample 2:	27	/	10000	*0.2% – 0.4%*	

Verdict:

Sample 2 is more successful

(p = 0.0421)

Confidence level: 95%

P-value

The probability that the **observed difference** in conversions would have been observed if there were **no underlying difference** between the control and variant.

or

The probability of obtaining an effect at least as extreme as the one in your sample data, **assuming the truth of the null hypothesis**.

P-values: common misconception

Observation:

- It is *NOT* that probability that the null hypothesis is true!
- P-values are *NOT* the probability of making a mistake by rejecting a true null hypothesis.

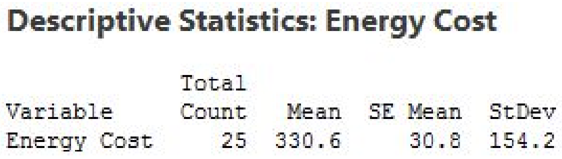
Example: vaccine study produced a p-value of 0.04

- **Correct:** Assuming that the vaccine had no effect, you'd obtain the observed difference or more in 4% of studies due to random sampling error (false positive).

- **Incorrect:** If you reject the null hypothesis, there is a 4% chance that you're making a mistake.

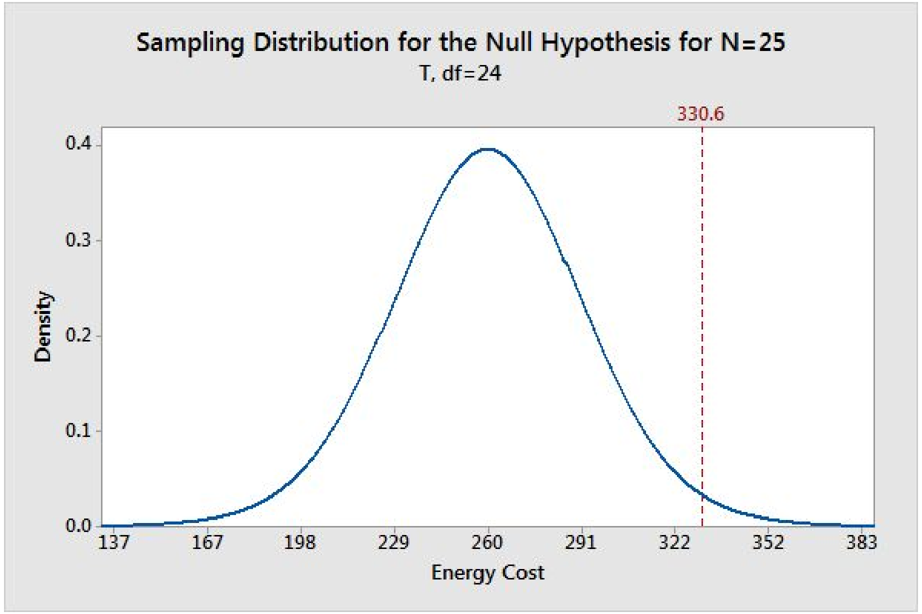
Example: Electricity Cost

We want to determine whether our sample mean ($330.6) indicates that this year's average energy cost is **significantly different** from last year’s average energy cost of $260.



Descriptive Statistics: Energy Cost

```
                   Total
Variable           Count    Mean   SE Mean   StDev
Energy Cost           25   330.6      30.8   154.2
```



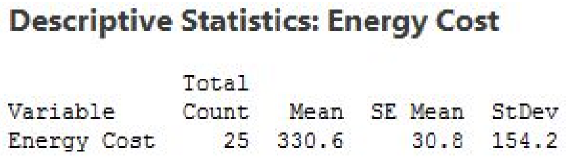
Example: Electricity Cost

<u>**H0**</u>: There is no significant difference between this year's average energy cost of $330.6 and last year's average energy cost of $260.

Our sample mean (330.6) falls **within** the *critical region*, which indicates it is **statistically significant at the 0.05 level**.

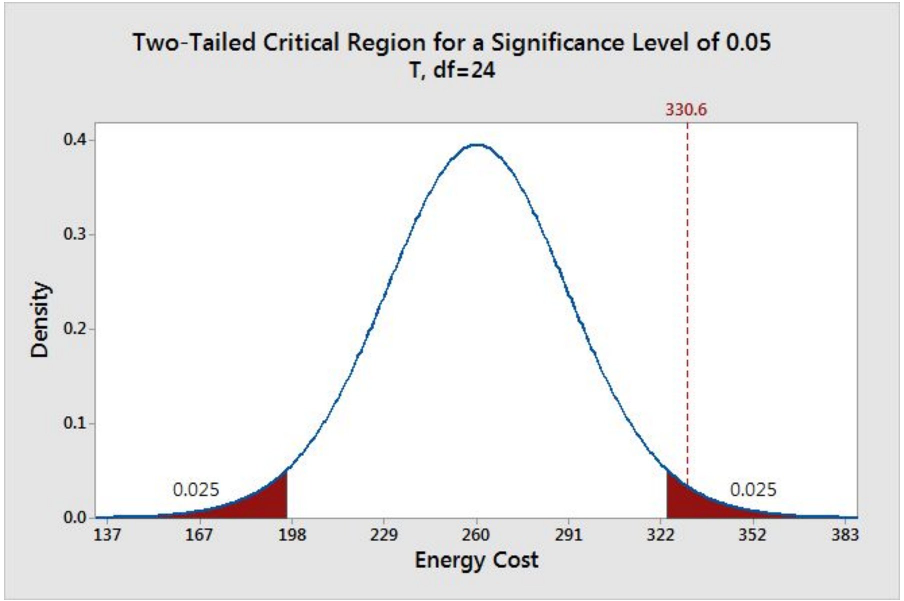
Conclusion: There is **enough evidence** to reject H0.

This year's costs are different than last year's costs.



Descriptive Statistics: Energy Cost

```
                   Total
Variable           Count    Mean  SE Mean  StDev
Energy Cost           25   330.6     30.8  154.2
```



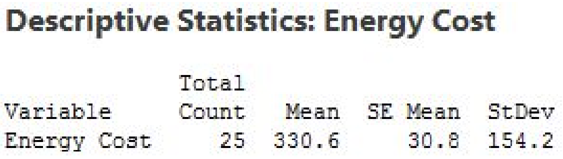
Example: Electricity Cost

<u>**H0**</u>: There is no significant difference between this year's average energy cost of $330.6 and last year’s average energy cost of $260.

Our sample mean (330.6) falls **outside** of the *critical region*, which indicates it is **statistically significant at the 0.01 level**.

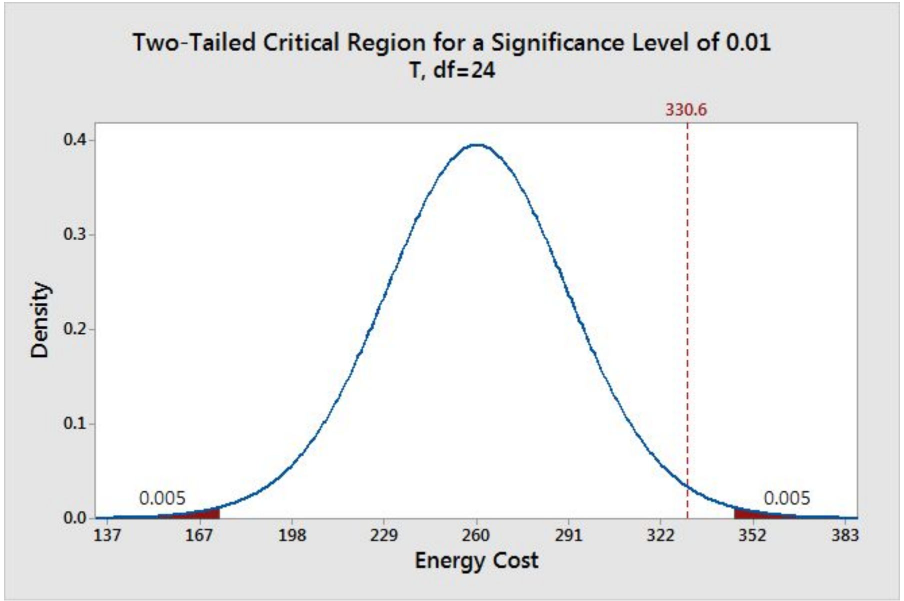
<u>Conclusion</u>: There is **not enough evidence** to reject H0.

This year’s costs are probably not different to year’s costs.



Descriptive Statistics: Energy Cost

```
                    Total
Variable            Count    Mean   SE Mean   StDev
Energy Cost            25   330.6      30.8   154.2
```



Two-Tailed Critical Region for a Significance Level of 0.01
T, df=24

Important

You need to choose your significance level **before** you begin your study!

It protects you from choosing a significance level because it **conveniently** gives you significant results!

How long should I run the test?

Power analysis

Used to determine the **sample size** required to **detect an effect** of a given size with a given degree of **confidence** (α).

What you need:

1. baseline conversion rate
2. minimum effect size (absolute or relative)
3. significance level = α (Type I error) = probability of finding an effect that is not there
4. power = 1 - β (Type II error) = probability of finding an effect that is there

Sample size calculators:

- Evan Miller’s
- Optimizely

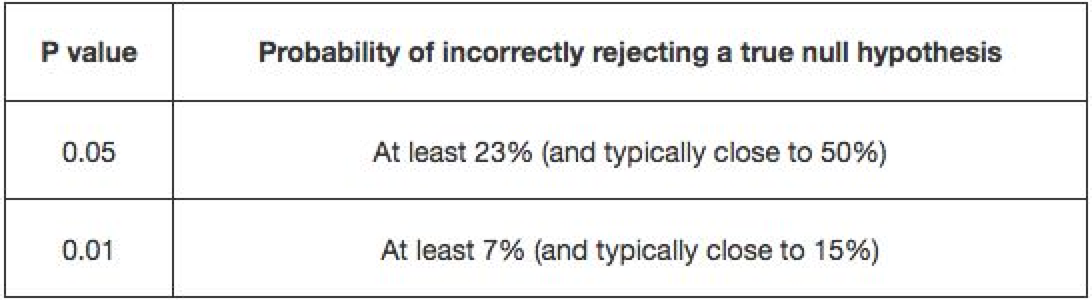
Your turn...

Example: Increase registration rate on your app/website

- Current registration rate: 12%
- Test setup:
  - 1 control & 2 variants (A, B, C)
  - only 20% of visitors
  - avg. # visitors / day: 10k
  - significance level: 98%, power: 80%
  - MDE: 5% (relative vs. absolute)
- How long should you run the test?
- http://www.evanmiller.org/ab-testing/sample-size.html

Disadvantages of frequentist approaches

- Counterintuitive
- The t-test is plagued by **false discoveries**
  - especially when looking at results continuously
  - even with the right sample size of visitors
- True error rate



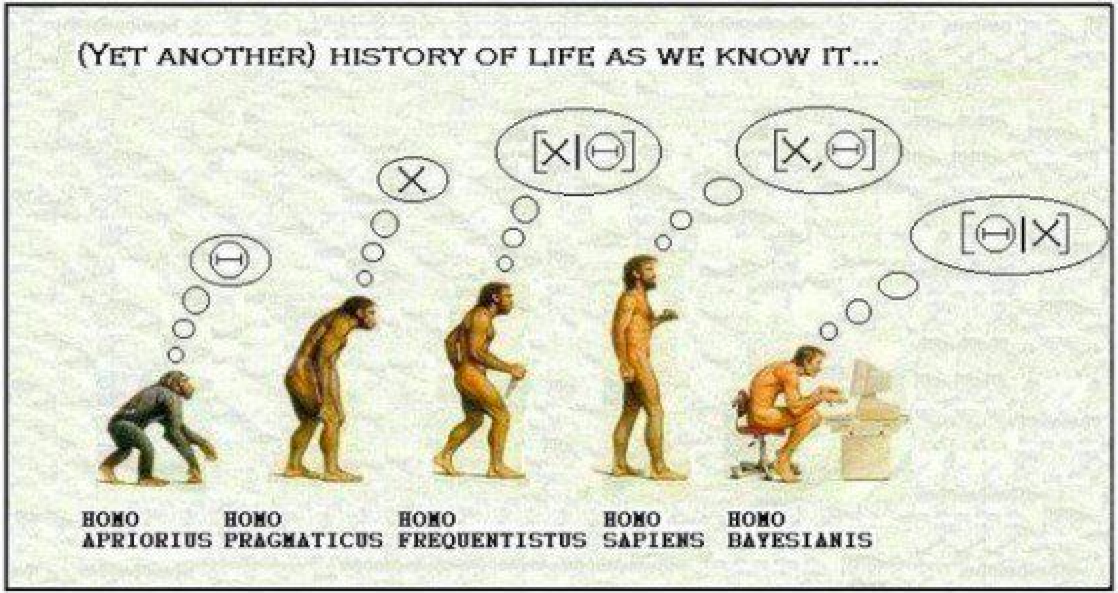
P value	Probability of incorrectly rejecting a true null hypothesis
0.05	At least 23% (and typically close to 50%)
0.01	At least 7% (and typically close to 15%)

Bayesian Approach

- Bayesian statistics take a more *bottom-up* approach to data analysis.
- **Past knowledge** is encoded into a statistical device known as a ***prior.***
- This prior is combined with **current experiment** data to make a conclusion on the test at hand.

*"A fundamental aspect of Bayesian inference is **updating your beliefs** in light of **new evidence**. Essentially, you start out with a prior belief and then update it in light of new evidence. An important aspect of this prior belief is your degree of confidence in it."* (*Boundless Rationality*)

Frequentist vs. Bayesian



The difference is that in the **Bayesian** view, a probability is assigned to a hypothesis.
In the **frequentist** view, a hypothesis is tested without being assigned a probability.

Why Bayesian?

Would you rather say:

“The probability that A > B is 10%”,

or

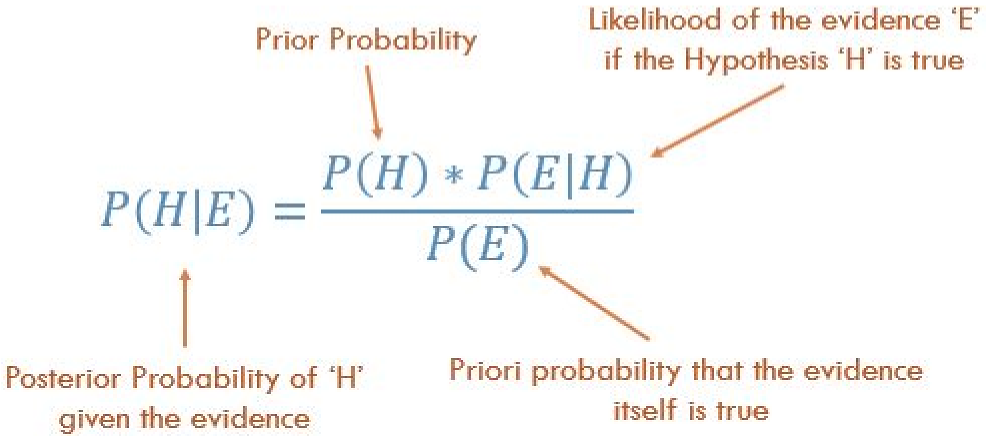
“Assuming the null hypothesis that A and B are equal is true, the probability that we would see a result this extreme in A vs. B is equal to 3%”

?

Bayesian Approach

1. **Define** the **prior** distribution that incorporates your subjective beliefs about a parameter.
2. **Gather** data.
3. **Update** your prior distribution with the data using Bayes’ theorem to obtain a **posterior** distribution.
   The posterior distribution is a probability distribution that represents your updated beliefs about the parameter after having seen the data.
4. **Analyze** the posterior distribution and summarize it (mean, median, sd, quantiles,…).

Bayes' Theorem

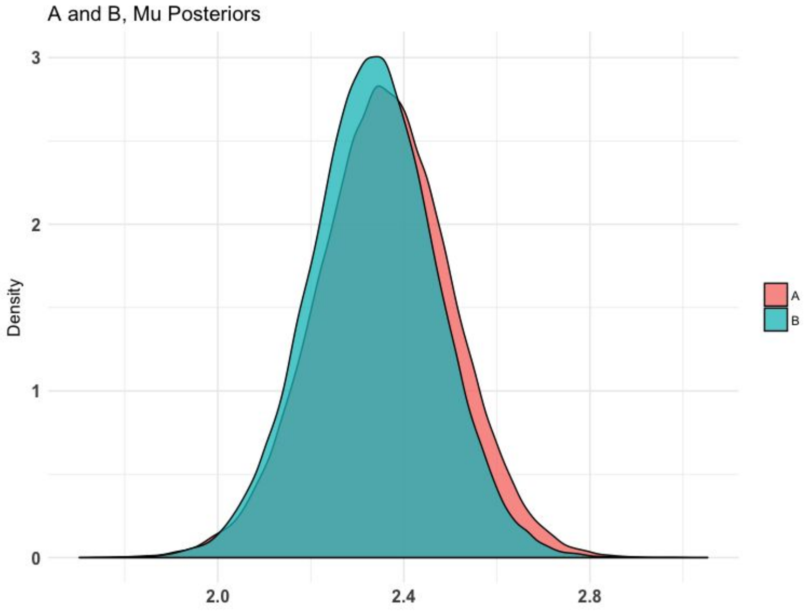


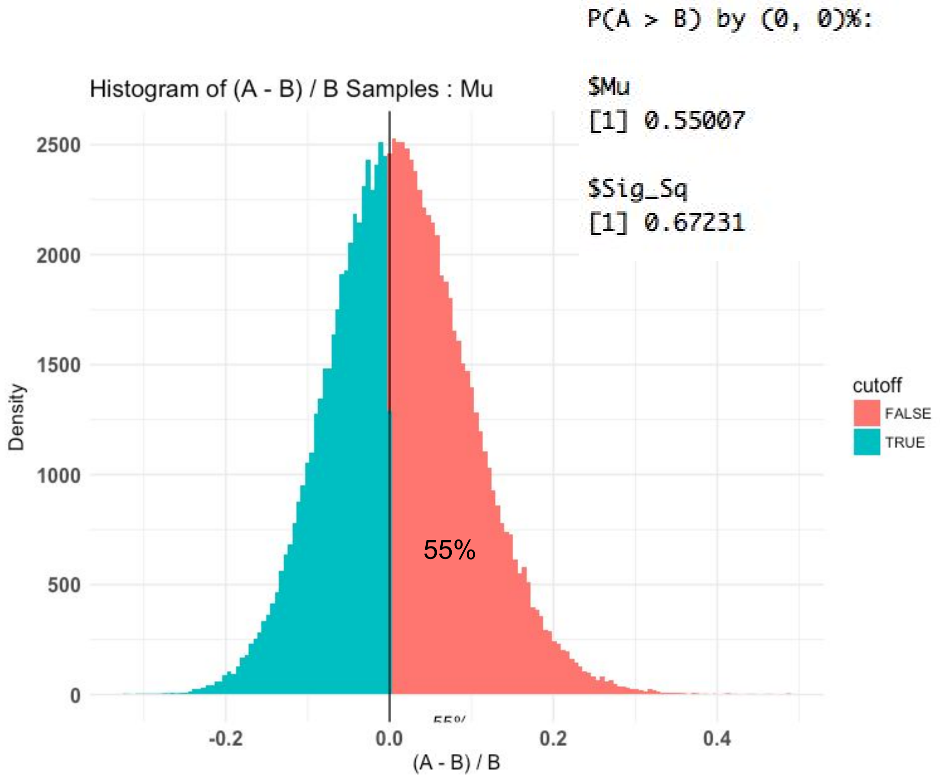
$$P(H|E) = \frac{P(H) * P(E|H)}{P(E)}$$

Prior Probability
Likelihood of the evidence 'E' if the Hypothesis 'H' is true
Posterior Probability of 'H' given the evidence
Priori probability that the evidence itself is true

Example

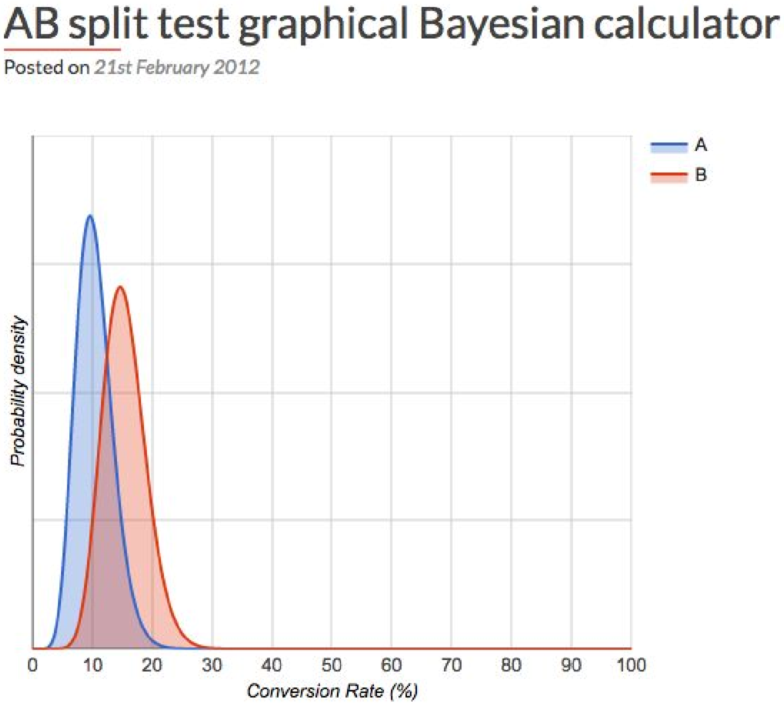
- Control and treatment group distr.



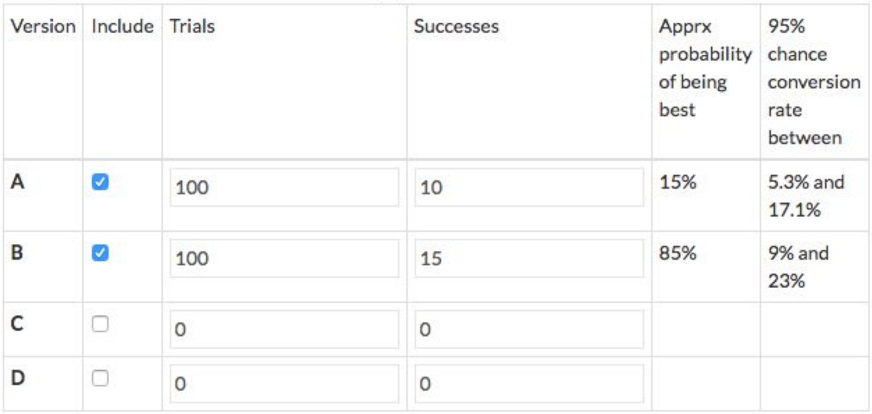

```
P(A > B) by (0, 0)%:

$Mu
[1] 0.55007

$Sig_Sq
[1] 0.67231
```

runtastic

Your turn...


AB split test graphical Bayesian calculator

Posted on *21st February 2012*

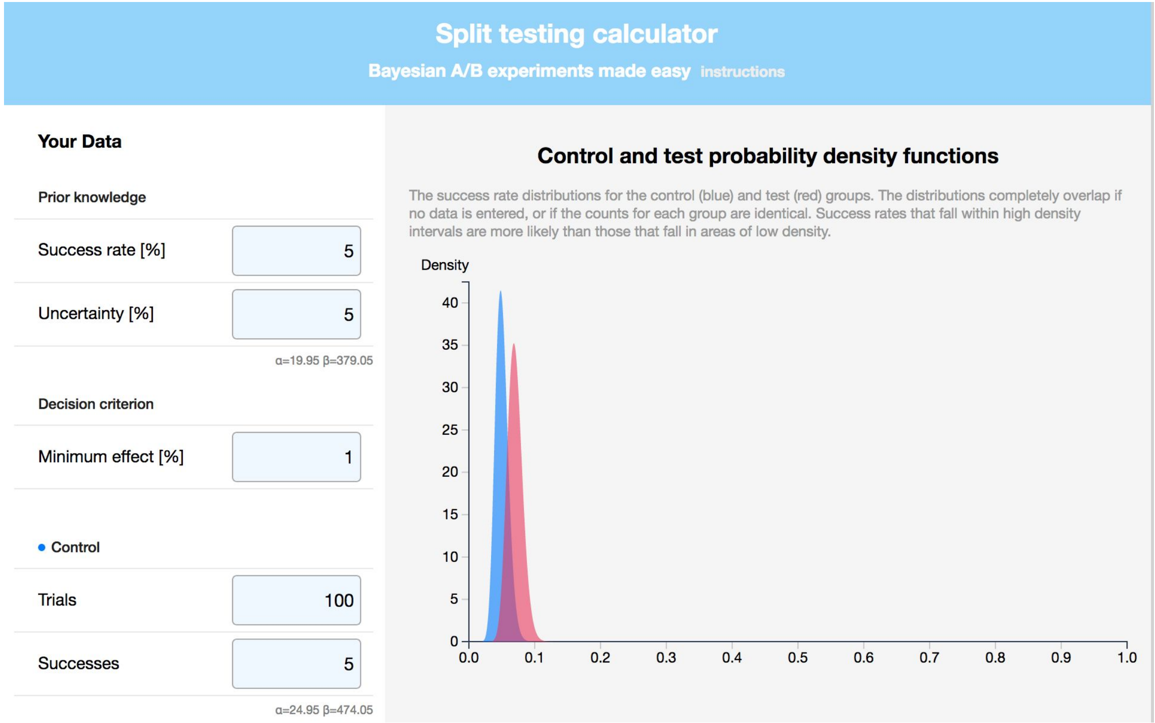


Version	Include	Trials	Successes	Apprx probability of being best	95% chance conversion rate between
A	☑	100	10	15%	5.3% and 17.1%
B	☑	100	15	85%	9% and 23%
C	□	0	0		
D	□	0	0		

*uses the Beta distribution

https://www.peakconversion.com/2012/02/ab-split-test-graphical-calculator/

Another online tool



Split testing calculator

Bayesian A/B experiments made easy instructions

Your Data

Prior knowledge

Success rate [%]	5
Uncertainty [%]	5

α=19.95 β=379.05

Decision criterion

Minimum effect [%]	1

● Control

Trials	100
Successes	5

α=24.95 β=474.05

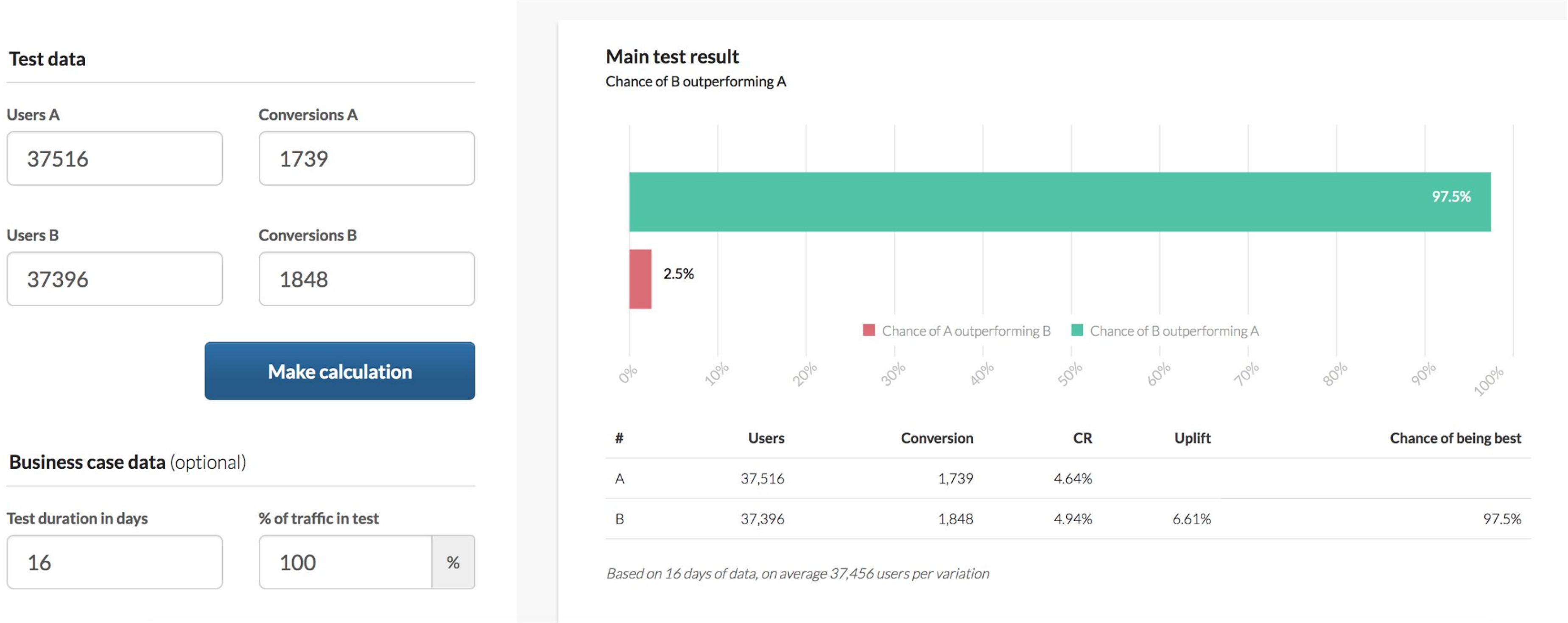
Control and test probability density functions

The success rate distributions for the control (blue) and test (red) groups. The distributions completely overlap if no data is entered, or if the counts for each group are identical. Success rates that fall within high density intervals are more likely than those that fall in areas of low density.

*uses the Beta distribution

https://yanirs.github.io/tools/split-test-calculator

Yet another tool



Test data

Users A	Conversions A
37516	1739
Users B	Conversions B
37396	1848

Business case data (optional)

Test duration in days	% of traffic in test
16	100 %

Main test result

Chance of B outperforming A

#	Users	Conversion	CR	Uplift	Chance of being best
A	37,516	1,739	4.64%		
B	37,396	1,848	4.94%	6.61%	97.5%

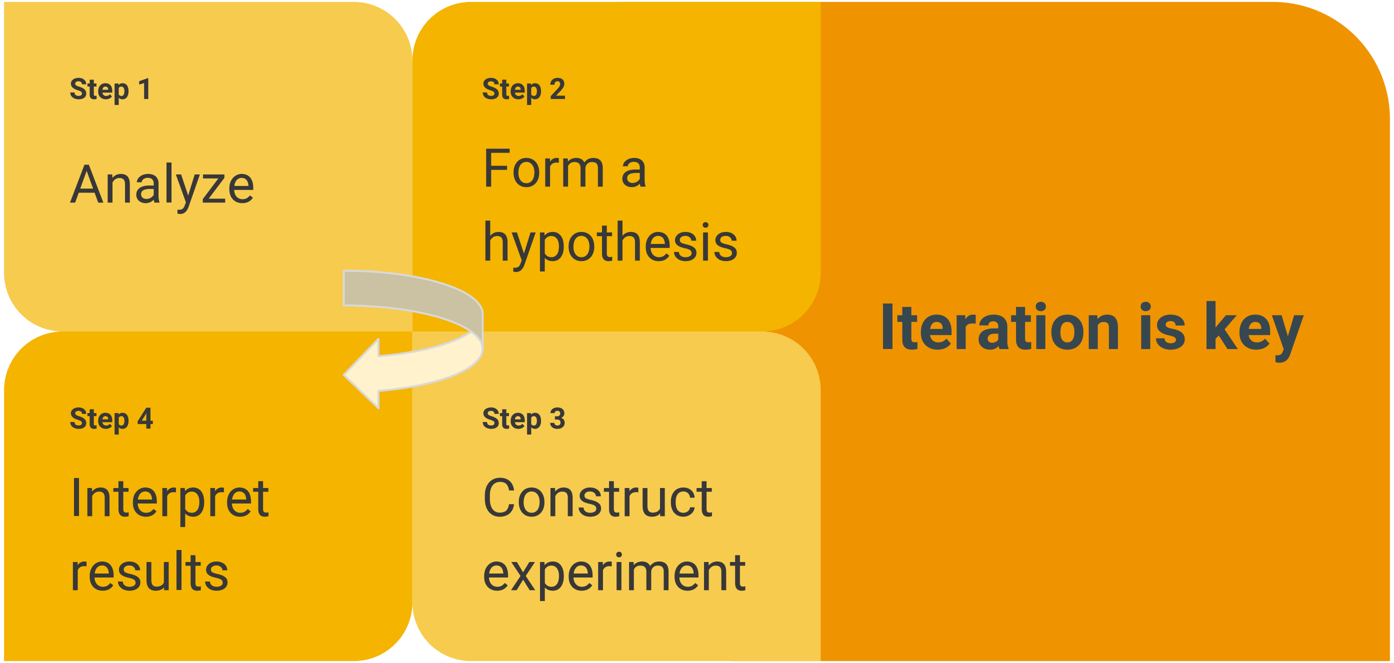
Based on 16 days of data, on average 37,456 users per variation

abtestguide.com/bayesian/

runtastic

The *Definitive* Guide to A/B Testing

The 4 steps of A/B testing



Step 1

Analyze

Step 2

Form a hypothesis

Step 3

Construct experiment

Step 4

Interpret results

Iteration is key

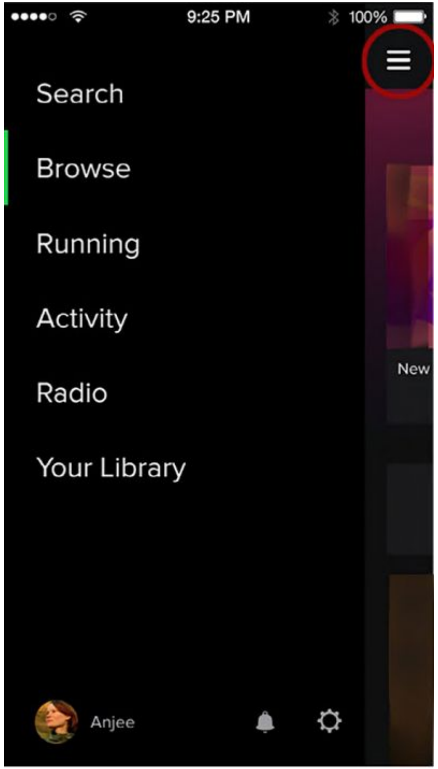
Checklist

1. Define a goal (e.g. increase conversion rate).
2. Investigate current state (gather and analyze data).
3. Identify problem / opportunity area.
4. Form a hypothesis.
5. Identify a control and a treatment (experiment design).
6. Identify key metrics to measure (select one as main metric).
7. *Identify what data needs to be collected.

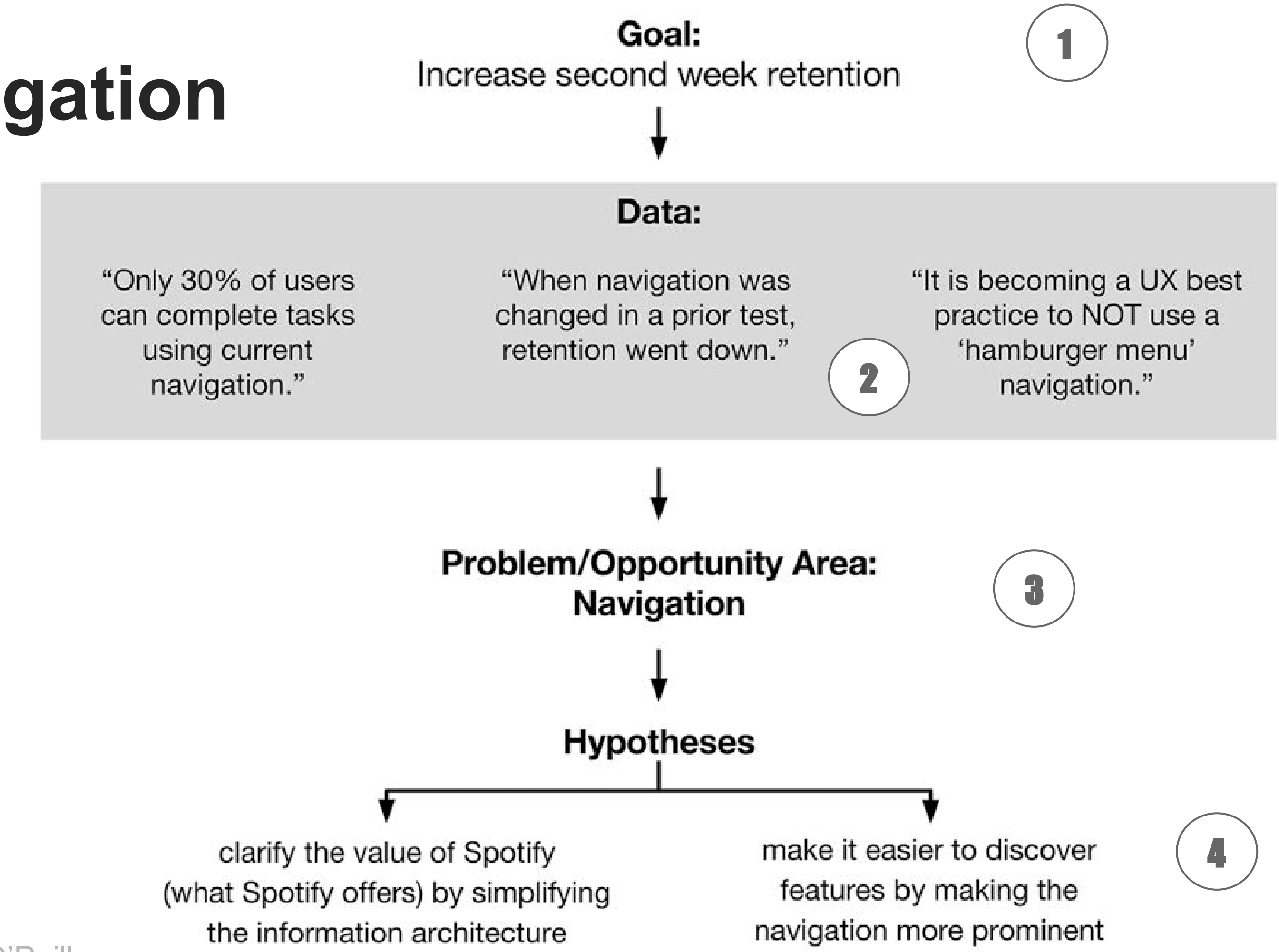
Checklist

8. *Make sure appropriate logging is in place.
9. Determine how small of the difference you would like to detect (MDE).
10. Determine the fraction of users in the treatment group.
11. Run a power analysis: decide how much data you need to collect and how long you need to run the test.
12. Run the test AT LEAST this long (consider seasonality).
13. *First time trying smth. new: run a dummy A/A test to check for systematic biases.

Spotify navigation



*Credits: Designing with Data, O'Reilly



Goal:
Increase second week retention

1

Data:

"Only 30% of users can complete tasks using current navigation."

"When navigation was changed in a prior test, retention went down."

"It is becoming a UX best practice to NOT use a 'hamburger menu' navigation."

2

Problem/Opportunity Area:
Navigation

3

Hypotheses

clarify the value of Spotify (what Spotify offers) by simplifying the information architecture

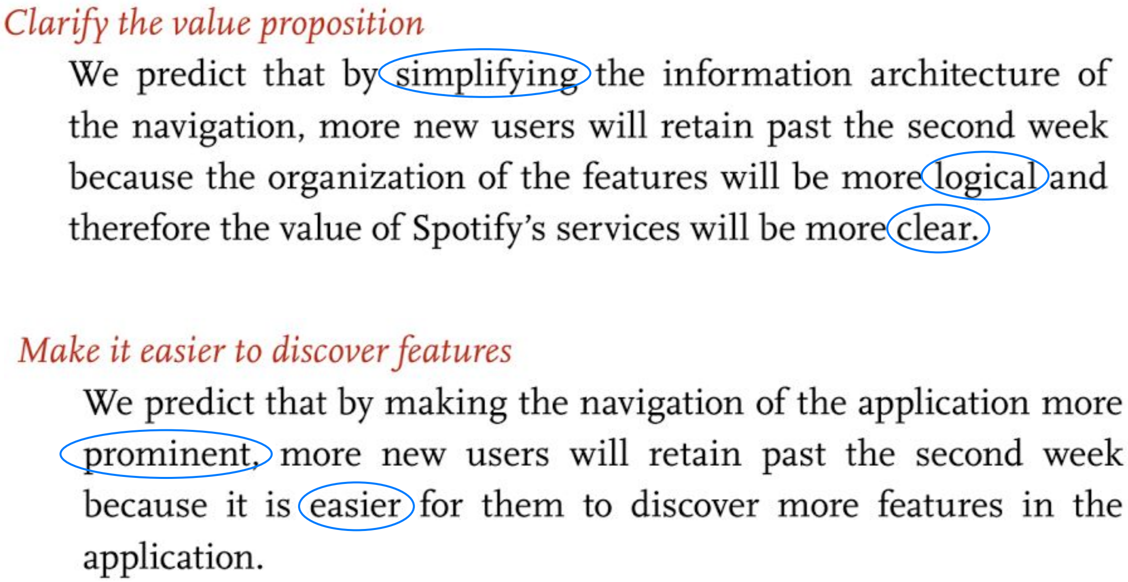
make it easier to discover features by making the navigation more prominent

4

Spotify navigation

4

Possible hypotheses:



Clarify the value proposition

We predict that by simplifying the information architecture of the navigation, more new users will retain past the second week because the organization of the features will be more logical and therefore the value of Spotify's services will be more clear.

Make it easier to discover features

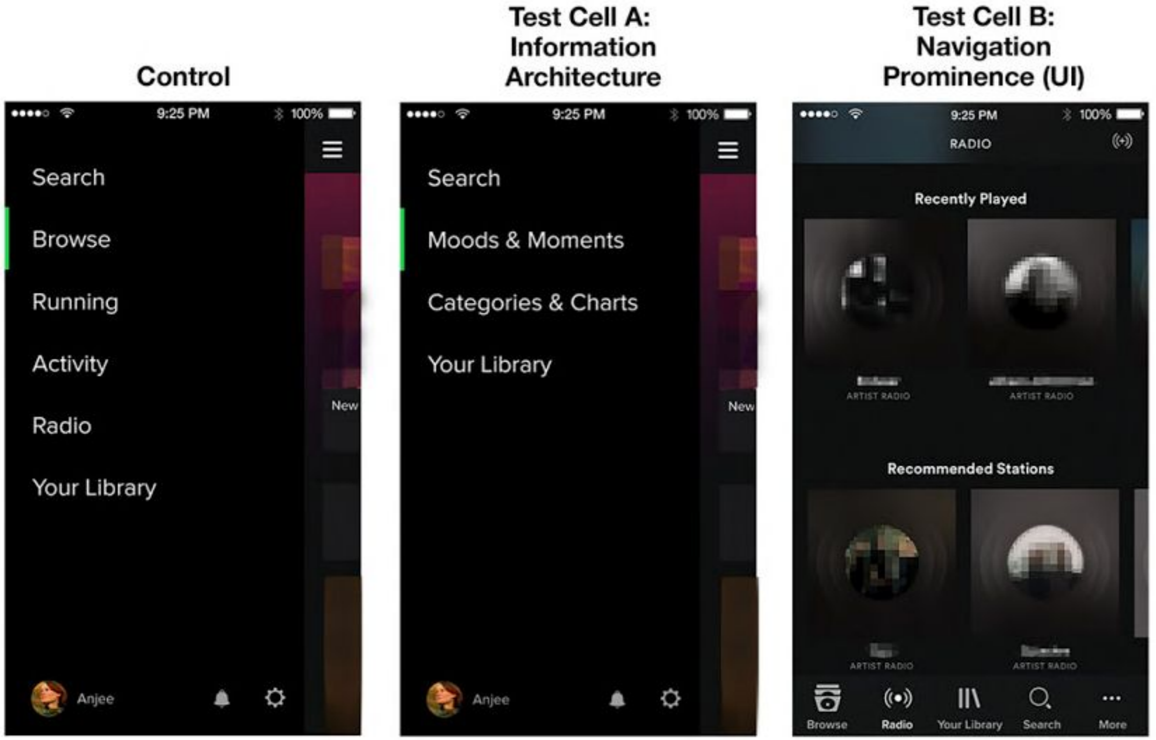
We predict that by making the navigation of the application more prominent, more new users will retain past the second week because it is easier for them to discover more features in the application.

*Credits: Designing with Data, O'Reilly

Spotify navigation

5

- **Variant 1** (cell A): simplified “hamburger” menu
- **Variant 2** (cell B): prominent navigation
- other variants were also tested



Control

Test Cell A: Information Architecture

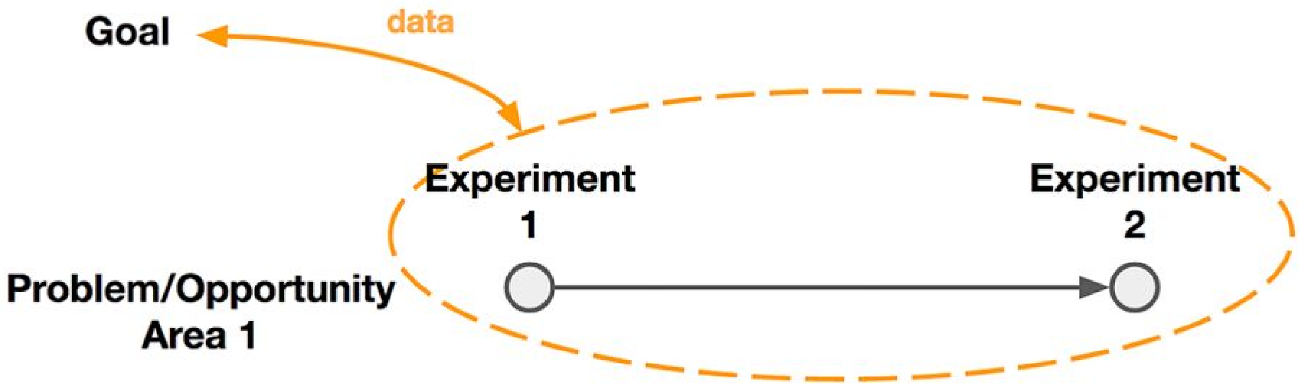
Test Cell B: Navigation Prominence (UI)

*Credits: Designing with Data, O’Reilly

Spotify navigation: learnings

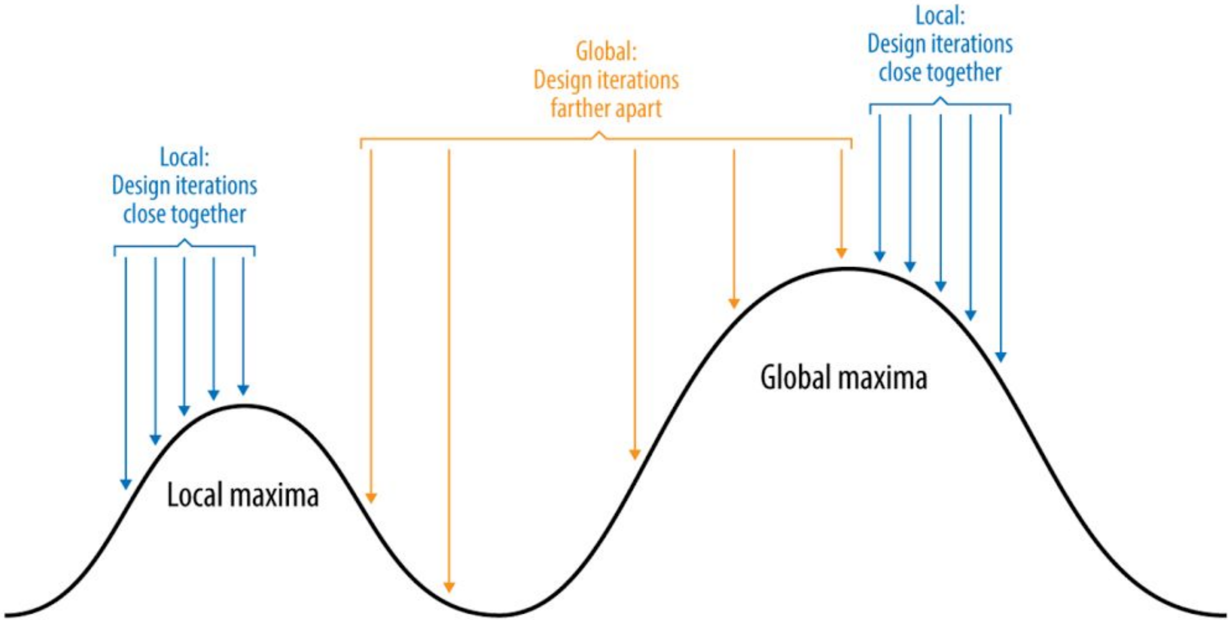
- Preliminary informational A/B test (low investment!) done before to get some **directional learning**
- **No significant effect** on second week retention for either of the two treatment variants
- Cell B showed improvements on some of the **secondary metrics** (i.e. users were exploring more the app)
- **Evidence** from previous analytics: increased exploration of the app was tied to retention (data!)
- <u>Decision</u>: invest further in exploring a more **prominent navigation**

Iteration is key



- for a given opportunity area you might need to run multiple test
- each experiment can provide new data for your goal

Iteration types



Local: Low variation between variants with the goal to optimize current setup.

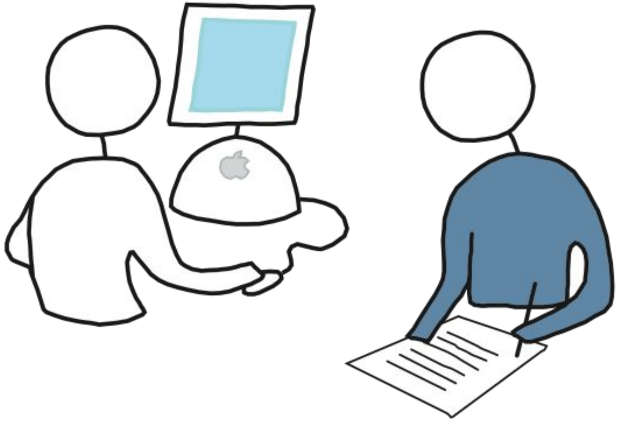
Global: Designs that are very different from each other with potential high impact (new ideas!).

*Credits: Designing with Data, O'Reilly

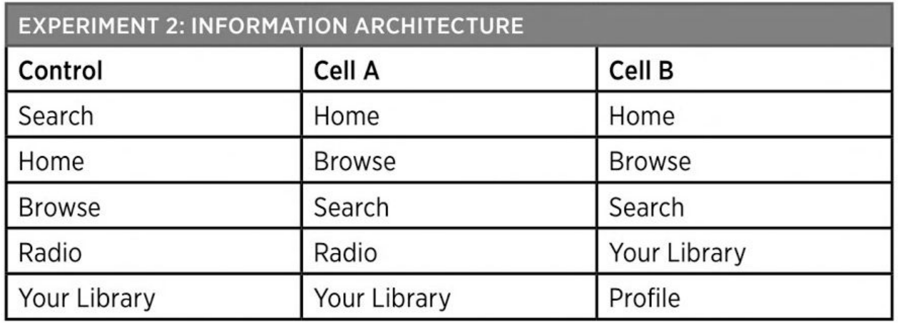
Integrate other methods

- Can bring **new knowledge** to the table that otherwise would not be possible only with A/B testing
- E.g. usability tests, case studies, questionnaires

Spotify example: usability tests showed that the tab bar was generally more successful



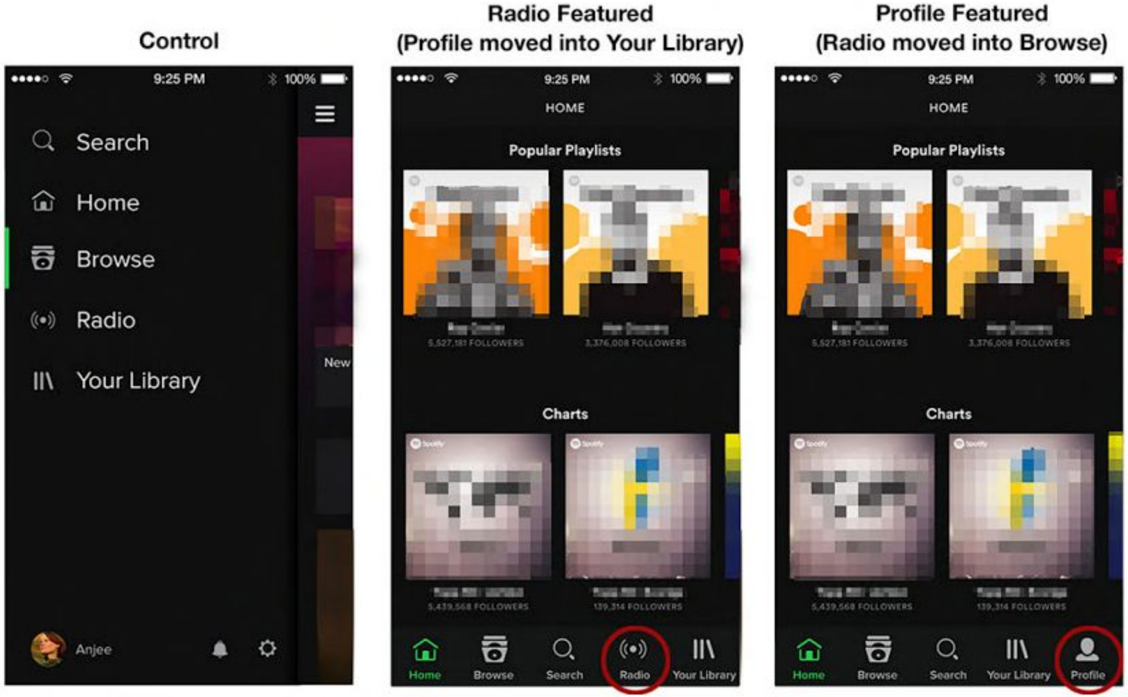
Spotify navigation



EXPERIMENT 2: INFORMATION ARCHITECTURE		
Control	**Cell A**	**Cell B**
Search	Home	Home
Home	Browse	Browse
Browse	Search	Search
Radio	Radio	Your Library
Your Library	Your Library	Profile

Both treatments showed significant **improvement** on second week retention.

Cell A had the added benefit of increasing **another metric** of interest.



Control

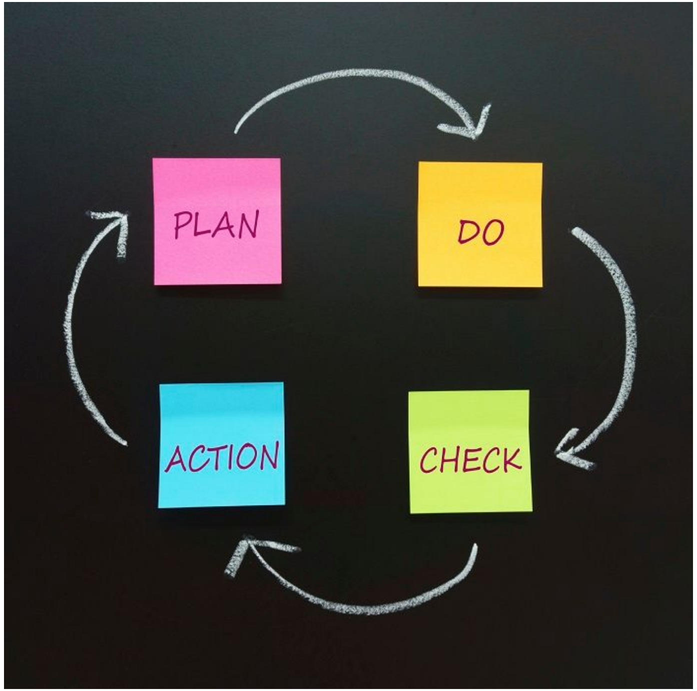
Radio Featured (Profile moved into Your Library)

Profile Featured (Radio moved into Browse)

*Credits: Designing with Data, O'Reilly

Refinement

- Local iterations
- Continued to explore further improvements
  - e.g. numbers of tabs, labels, icons
- Previous learning help design more targeted experiments
- Helps reach a **local maxima**



Test design

5

Define and design variants to clearly and uniquely reflect your question / hypothesis.

Test design

5

Checklist

- Articulate a strong list of things you want to learn (success & failure)
- Ask yourself: “If test X succeeds/fails, what experiment would I run next?”

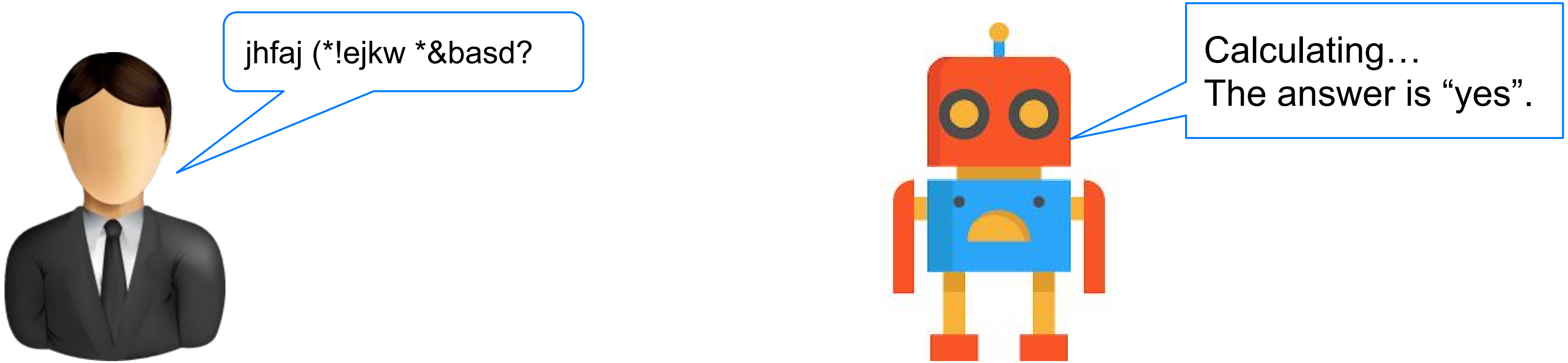
Opportunity: Exaggerate a feature of a test cell with the intention to learn

Questions to ask

- How can I create a **design/experience** that best expresses my hypothesis?
  - How are the **key components** of my hypothesis represented in the design?
  - Are there any **redundant** / **missing** test variants?
  - What will I learn with **this** variant that I wouldn’t be able to learn with the others?
  - Can I “**postpone**” some variants to the next experiment?
- How does the design help **gather the right data** that will provide evidence in favor or against my hypothesis?
- Can the test with these same test variants answer a **different question**?

But remember...

A/B tests are dumm.
Can only answer with “yes” or “no”.
But you pick the question.

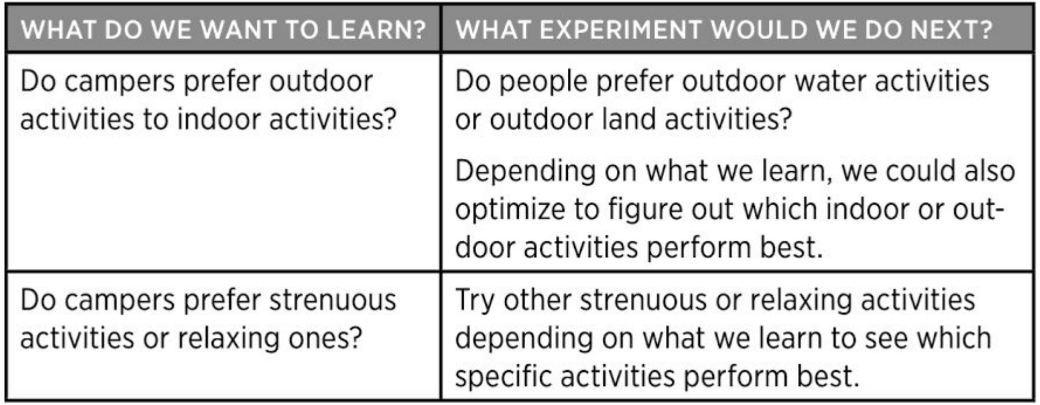


jhfaj (*!ejkw *&basd?

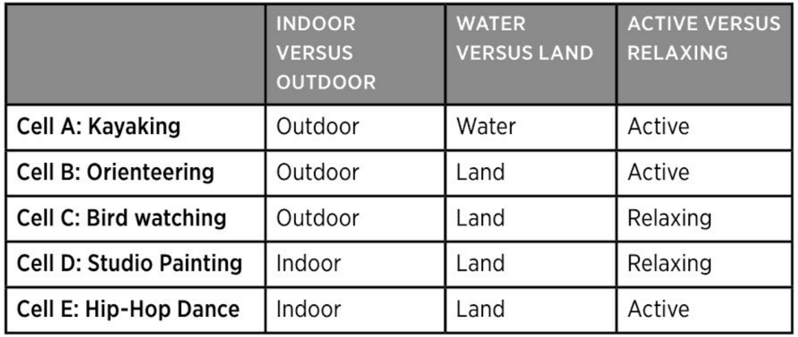
Calculating…
The answer is “yes”.

Camping example

- **Goal**: increase camping enrollment
- **Hypothesis**: by adding new activities to the program we can increase enrollment, because we will engage a wider range of audience



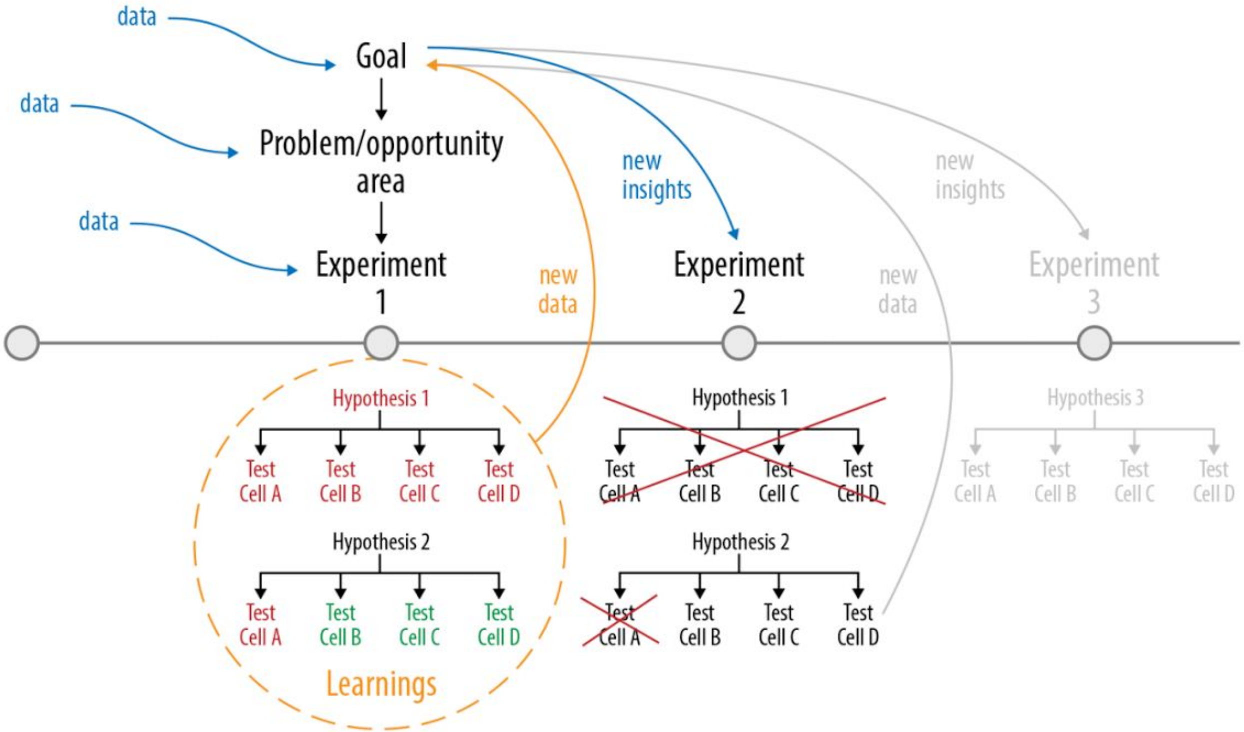
WHAT DO WE WANT TO LEARN?	WHAT EXPERIMENT WOULD WE DO NEXT?
Do campers prefer outdoor activities to indoor activities?	Do people prefer outdoor water activities or outdoor land activities? Depending on what we learn, we could also optimize to figure out which indoor or outdoor activities perform best.
Do campers prefer strenuous activities or relaxing ones?	Try other strenuous or relaxing activities depending on what we learn to see which specific activities perform best.



	INDOOR VERSUS OUTDOOR	WATER VERSUS LAND	ACTIVE VERSUS RELAXING
Cell A: Kayaking	Outdoor	Water	Active
Cell B: Orienteering	Outdoor	Land	Active
Cell C: Bird watching	Outdoor	Land	Relaxing
Cell D: Studio Painting	Indoor	Land	Relaxing
Cell E: Hip-Hop Dance	Indoor	Land	Active

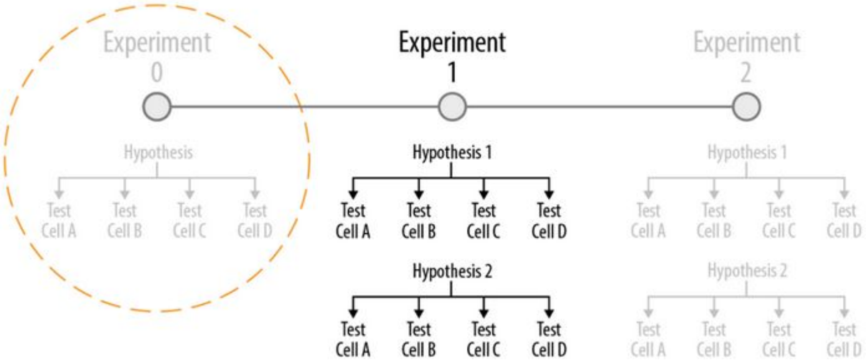
*Credits: Designing with Data, O'Reilly

Learning



*Credits: Designing with Data, O'Reilly

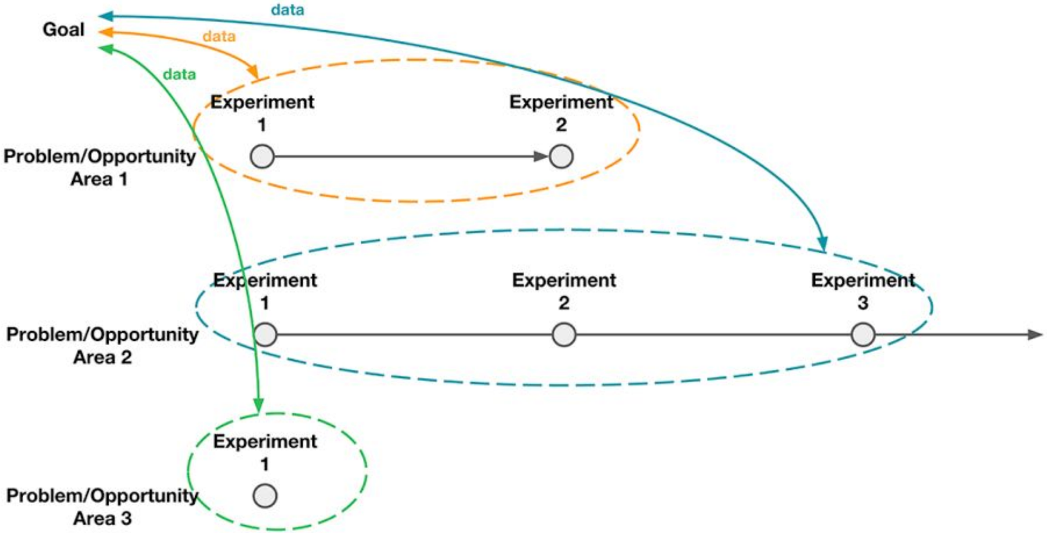
Experiment 0



- “What experiment might I have run ***before*** the one I am planning now?”
- Useful when experimenting with a new problem/opportunity area
- Determine if you jumped too quickly to a set of experiments that are too detailed for the stage you are in.
- Example: remove the feature entirely (skyscanner: “cheapest flights ticker”)
- Benefit: Avoid investing unnecessary resources

When to stop an experiment?

- Should **never** think about stopping and experiment.
- **Instead**: change the **nature** of experiments you are running, if you don't see any gains in the current area
- When learning about **customer behaviour**: explore many problem areas simultaneously
- With experience, your explorations will become more **focused** and **narrow**



*Credits: Designing with Data, O'Reilly

How to evaluate?

Checklist

1. Look at **significance level**.
2. Use the **main KPI** to make a decision.
3. Don't evaluate before
   a. predefined **confidence level** is reached (95% or 99%)
   b. minimum **required sample size** (per group) is reached
   c. the time interval covers all **variations** (e.g. weekly seasonality).

Pitfalls to avoid

Interfering with a running test

Don’t change anything on a running A/B test (except bugs)!

Multiple variants

- Avoid too many variants in a single A/B test

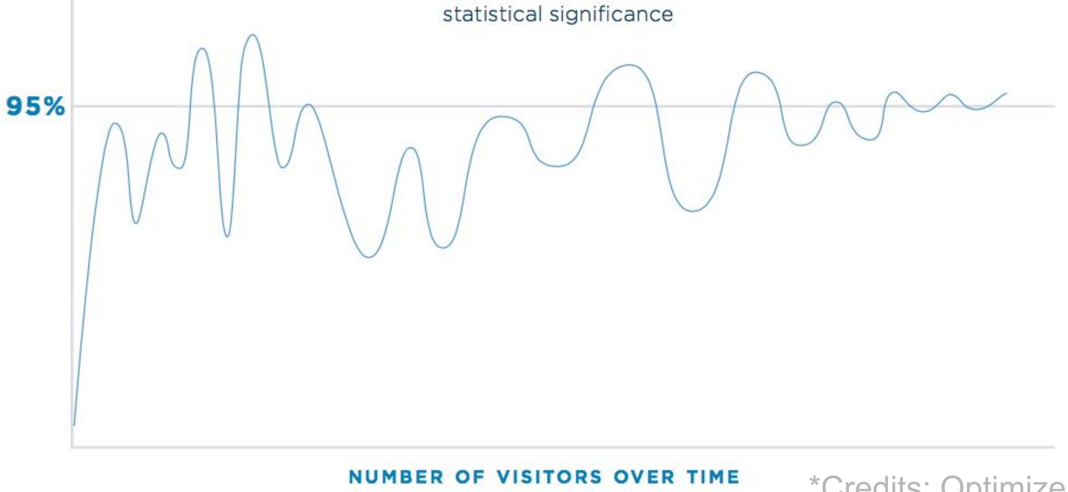
Why?

- Test needs more time to reach significance
- Difficult to manage and learn from

runtastic

Peeking: stopping test early

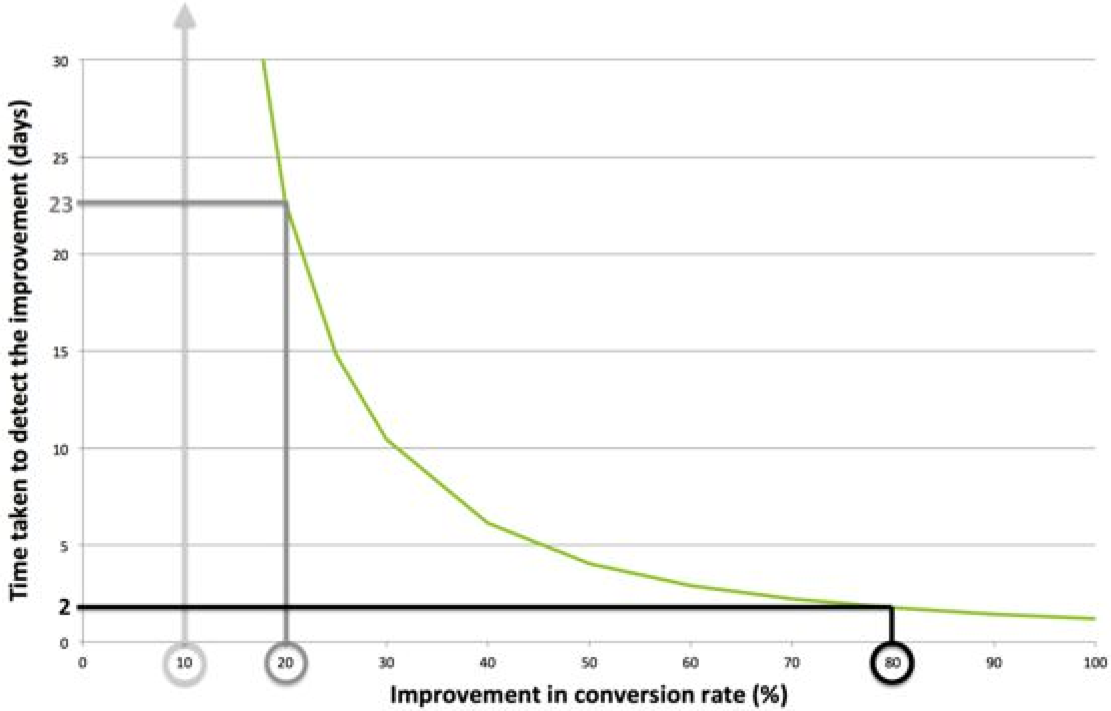
- Stopping a running test that did not reach at least 95% significance.
- Stopping tests early, as soon as it reaches significance.



*Credits: Optimizely

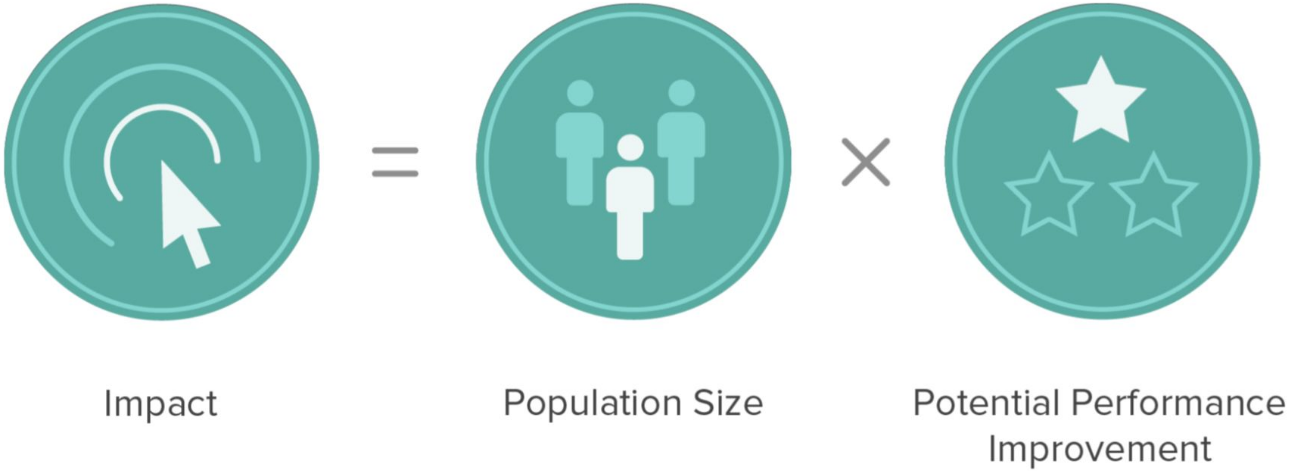
Tweaking

e.g. webpage with
300 views/day



Focusing on low impact

Run high-impact tests.

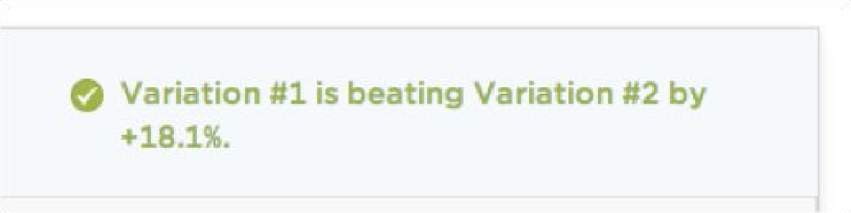


Impact = Population Size × Potential Performance Improvement

*Credits: Optimizely

Ignoring seasonality

- Sample size matters, but seasonality matters more
- Run tests long enough to capture variations



✓ Variation #1 is beating Variation #2 by +18.1%.

Worst-case scenarios

- Ensure product is well-designed for best- and worst-case scenarios
- Consider edge-cases: messy data in designs
- Example: long user names, broken images, etc. twitter

runtastic

Novelty effect

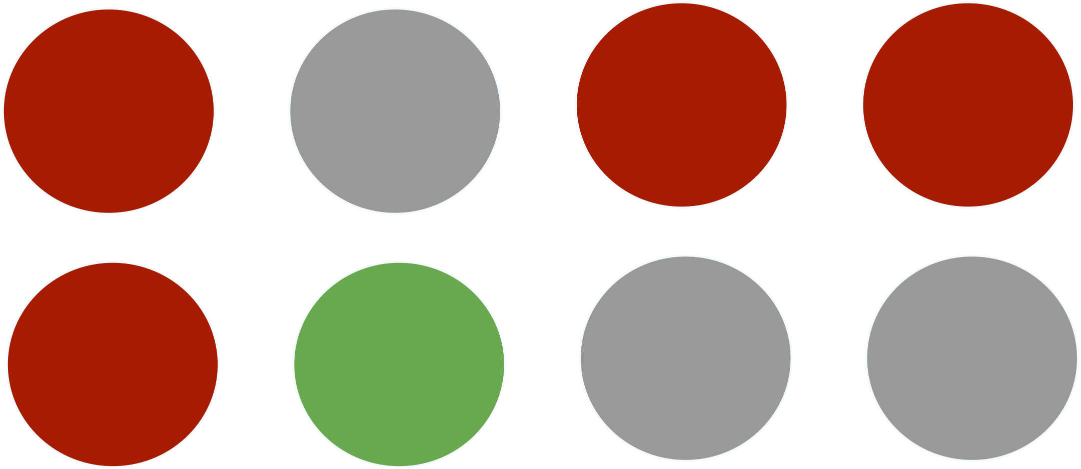
- Changes help disrupt the blindness banner
- Changes can attract more user’s attention
- <u>However</u>: might as well be just temporary effect
- Solution: run tests longer and maybe twice

Not planning ahead

- Thinking about next steps only when you get your results back
- A/B testing should not slow you down
- Anticipate next steps and design next stage of experiments while waiting for results

Final remarks

Be patient: testing is not a free lunch



Learn from experiments and improve!

42

A/B testing is NOT always the best choice to find answers.

Team work

DIY: Setting up an A/B test

1. Find a realistic and meaningful use case
2. Following the checklist, set up your own A/B test
3. Calculate how long the test should run based on traffic, base conversion rate and expected effect (use 3 different MDE values)
4. Define what actions you would take depending on the test outcome
5. What would be your next test?

runtastic

Thank you!

Your feedback is greatly appreciated!

Runtastic.com