2021 Wharton Analytics Conference
Supported by Wharton AI for Business, Analytics@Wharton

The Art & Science of A/B Testing

Alex P. Miller
Ph.D. Candidate, Information Systems
Department of Operations, Information, & Decisions

Wharton

Welcome & Introduction

Ph.D. Candidate
Information Systems,
OID Department

Starting June 2021:
Asst. Professor of
Quantitative Marketing,
USC Marshall School of
Business

- Research interests: A/B testing, personalization, e-commerce, algorithmic decision making

- Prior experience: digital marketing, data science/engineering, web analytics consulting

Overview:

1. Core concepts
2. A/B testing paradigms in business
3. Simulation exercise
4. Debrief

What will you get out of this workshop?

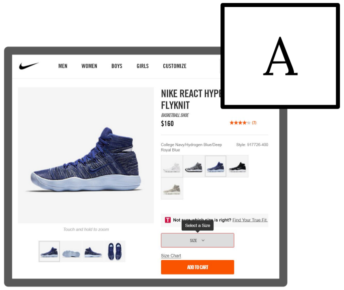
- A hands-on understanding of A/B testing:
  - What is it?
  - What types of business problems can it help you solve?
  - What does it look & feel like to use A/B testing for decision making?
- A high-level understanding of how to use A/B testing tools to solve the **right** problem
  - Key aspects of using statistics for business decision making
  - Without getting bogged down in math

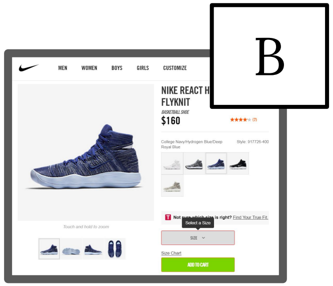
Core Concepts in A/B Testing

Definition:

A/B testing is:

the practice of using of **randomized** experiments for making business decisions

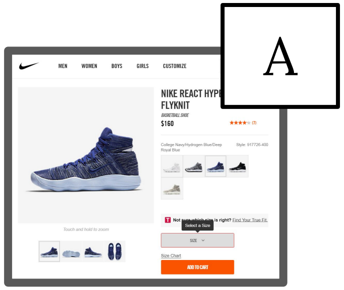


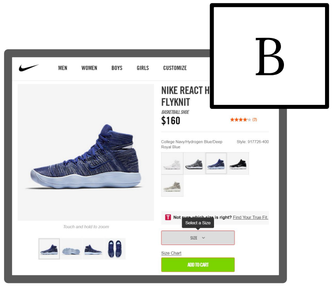


Definition:

A/B testing is:

the practice of using of **randomized** experiments for making business decisions





A/B testing is not:

trying multiple strategies in an *ad hoc* manner and comparing results

People are asking...

Why should you care about A/B testing?

When used properly:

- Randomized experiments are the "gold standard" for measuring cause & effect
    - A/B testing can *help* you predict the future
- Can help you truly understand which components of your products/services drive value
- Can facilitate a culture of empirical measurement & organizational learning

"Experimentation is the least arrogant method of gaining knowledge."

— Isaac Asimov

A/B testing is for everyone

- Tech companies (Microsoft, Google, Amazon, Facebook) are well-known for having intensely experimental organizations

A/B testing is for everyone

- Tech companies (Microsoft, Google, Amazon, Facebook) are well-known for having intensely experimental organizations

- New software companies have opened up rigorous experimentation to even very small companies (or small, non-technical teams at large companies)

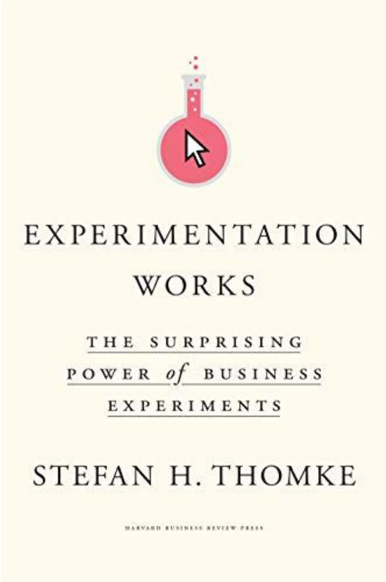
  - Almost every web-analytics platform can be used for experimentation

ADOBE® TEST&TARGET™
Powered by Omniture®
HubSpot
Optimizely
Google Optimize

Recommended Reading

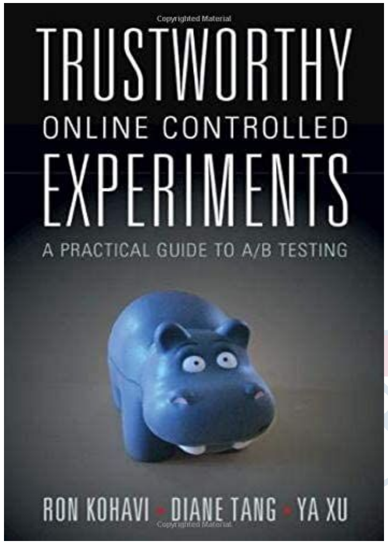
For more details on developing an experimental culture in your organization:

Experimentation Works: The Surprising Power of Business Experiments



For more technical/implementation details about experimentation:

Trustworthy Online Controlled Experiments



A brief introduction to....

The Basics of Business Experiments

Why run experiments?

- Randomized experimentation is a technique of gathering data that is specifically designed as a means of "**causal inference**"

Why run experiments?

- Randomized experimentation is a technique of gathering data that is specifically designed as a means of "**causal inference**"

Causal inference:

The process of understanding and measuring cause & effect

Many (not all) business decisions are problems of causal inference

“Correlation is not causation”

Difference between correlation (or association) and causation:

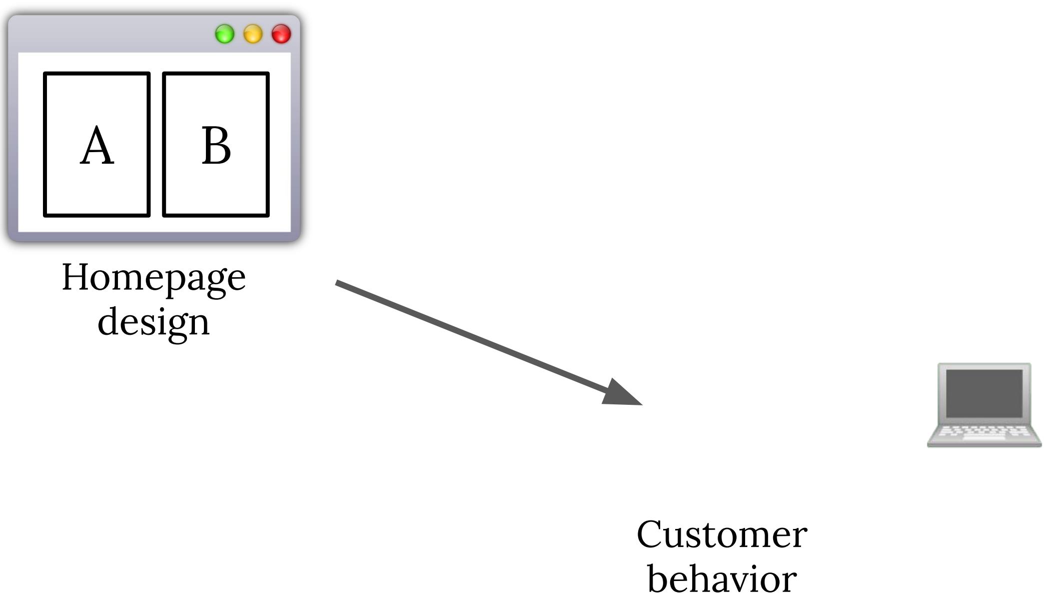
- “We redesigned our homepage last week and customer conversions increased”

- “Customer conversions increased last week **because** of our new homepage design”

How to tell the difference?

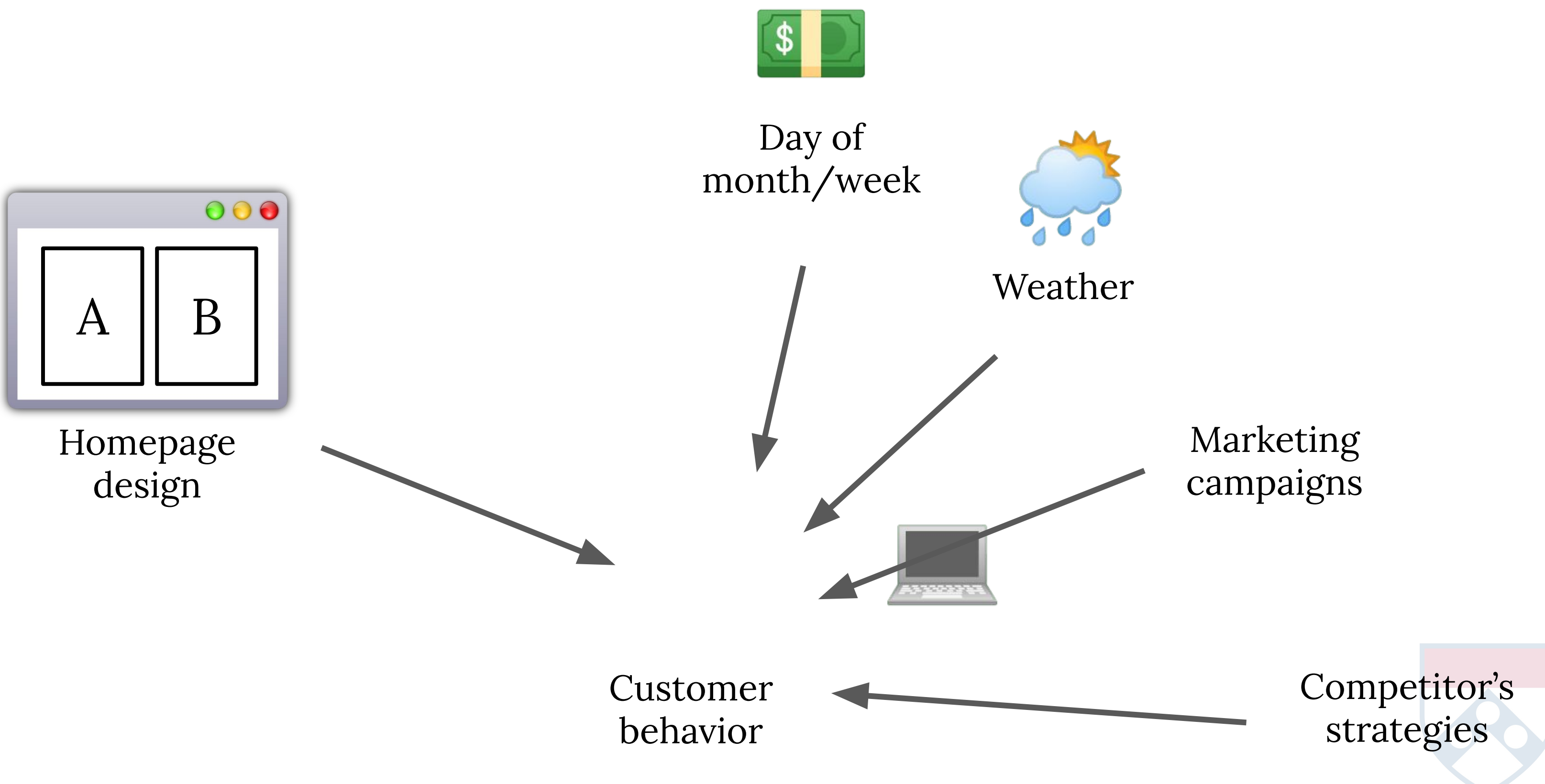
Why is this problem hard?

It’s hard to separate your actions from other factors that could affect customer behavior:

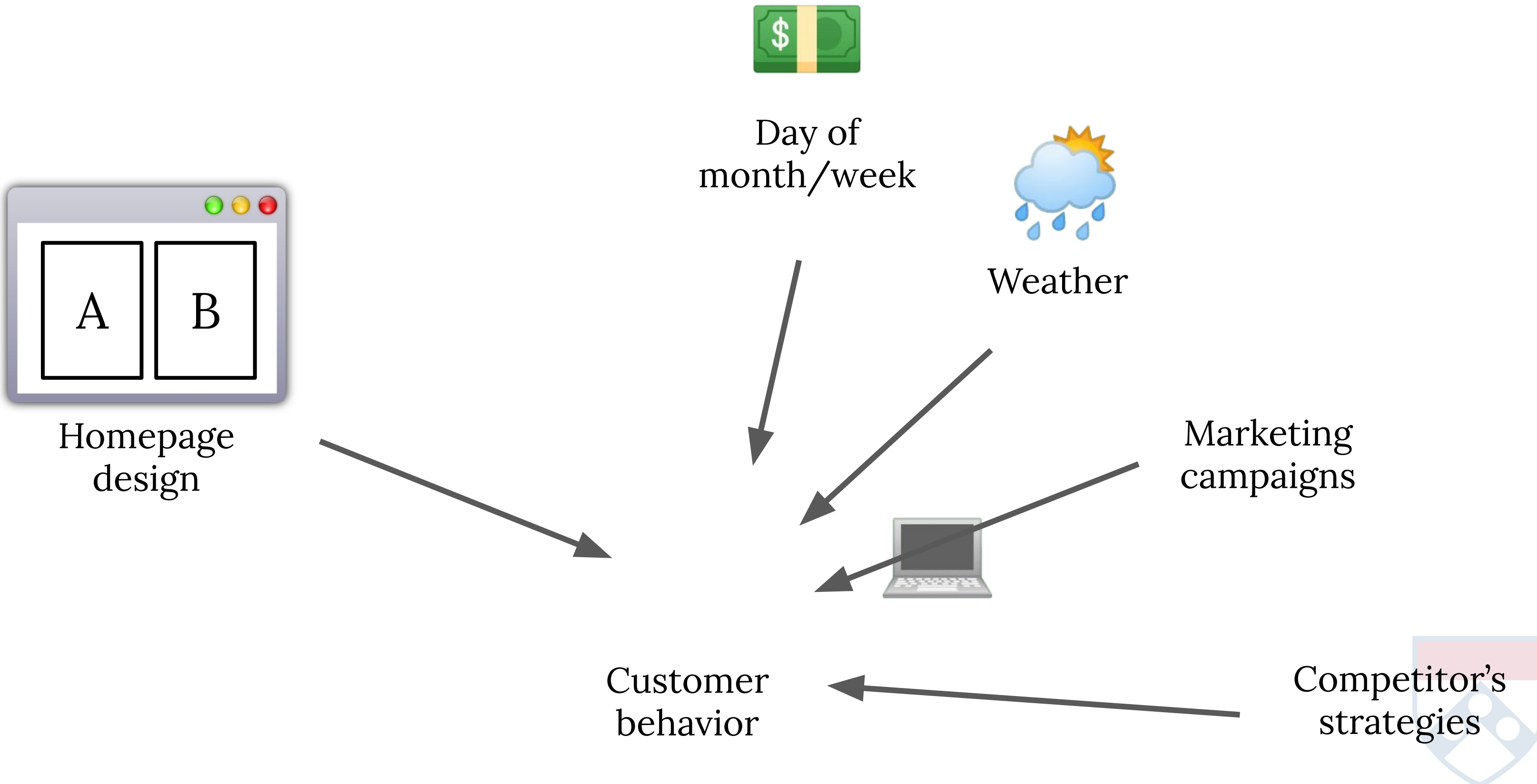


Why is this problem hard?

It's hard to separate your actions from other factors that could affect customer behavior:

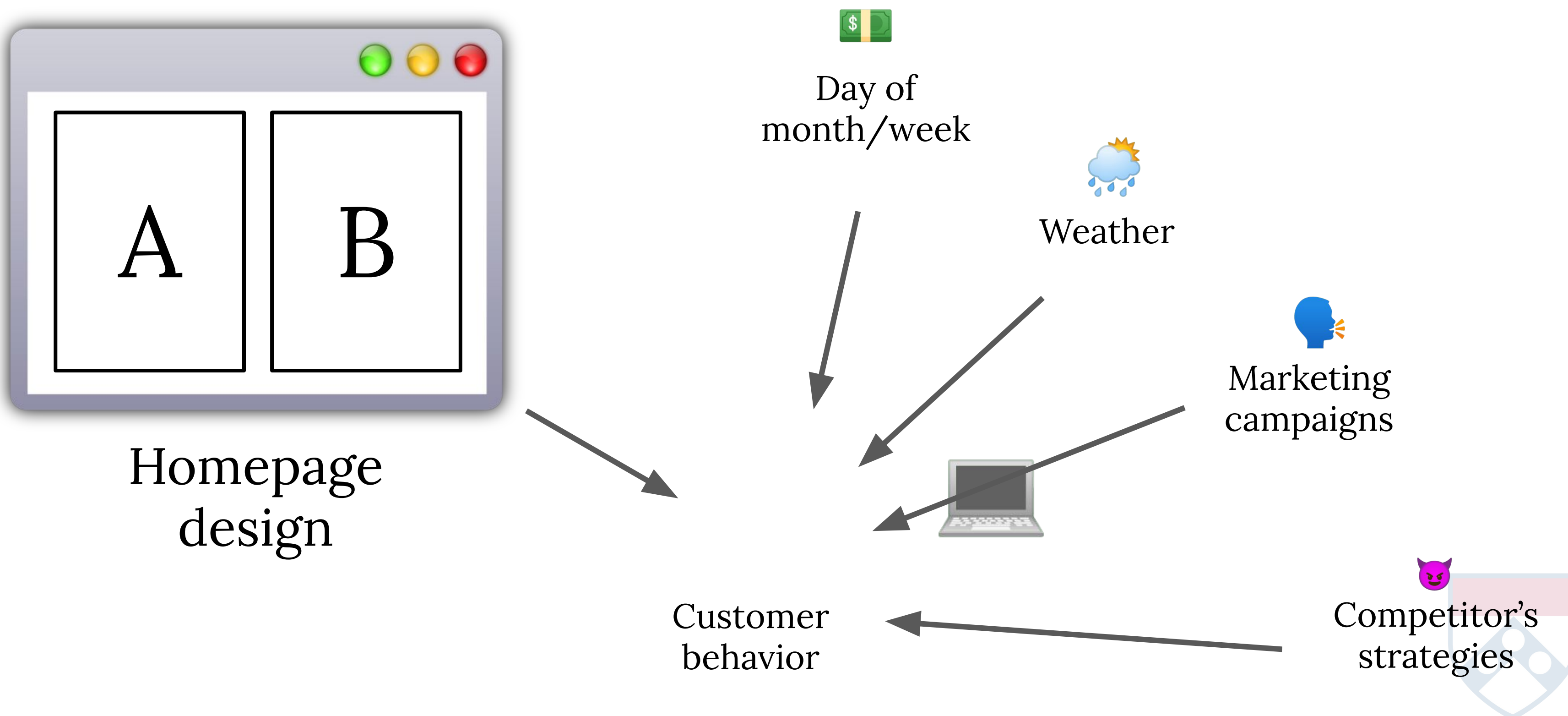


How does randomization help?



How does randomization help?

Randomizing which homepage customers see allows you to isolate the effect of that variable; with enough data, other factors that affect behavior should be balanced



A/B testing is valuable in situations when:

You have multiple strategies/actions you can implement and:

1. [You are willing to admit that] You don't know which one is best

2. You can implement each strategy using randomization

3. You can measure the results of each strategy along dimensions that you care about

A/B testing is a particularly powerful tool in **digital business**, relative to traditional forms of commerce

- Cost of “innovation” relatively low
- Randomization is easy
- Measurement is easy

“Offline” A/B testing can also be valuable, but we will focus on digital experiments today

What should you test?

- This depends critically on your industry/context

- Many online resources and user experience guides exist

- Beware though: What works for one company may not work for yours

  - If you develop a culture of systematic experimentation, you will learn which components of your website/service matter most

Key Steps for Running an A/B Test

1. Develop a set of “hypotheses” to test
   e.g., “variations”, “treatments” “arms”, “strategies”

Key Steps for Running an A/B Test

1. Develop a set of "hypotheses" to test
   e.g., "variations", "treatments" "arms", "strategies"

2. Define your key evaluation criteria

Key Steps for Running an A/B Test

1. Develop a set of “hypotheses” to test
   e.g., “variations”, “treatments” “arms”, “strategies”

2. Define your key evaluation criteria

3. Define your intended sample size & stopping criteria (will revisit)

Key Steps for Running an A/B Test

1. Develop a set of "hypotheses" to test
   e.g., "variations", "treatments" "arms", "strategies"

2. Define your key evaluation criteria

3. Define your intended sample size & stopping criteria (will revisit)

4. Run your experiment: Randomly assign customers to treatment arms

Key Steps for Running an A/B Test

1. Develop a set of "hypotheses" to test
   e.g., "variations", "treatments" "arms", "strategies"

2. Define your key evaluation criteria

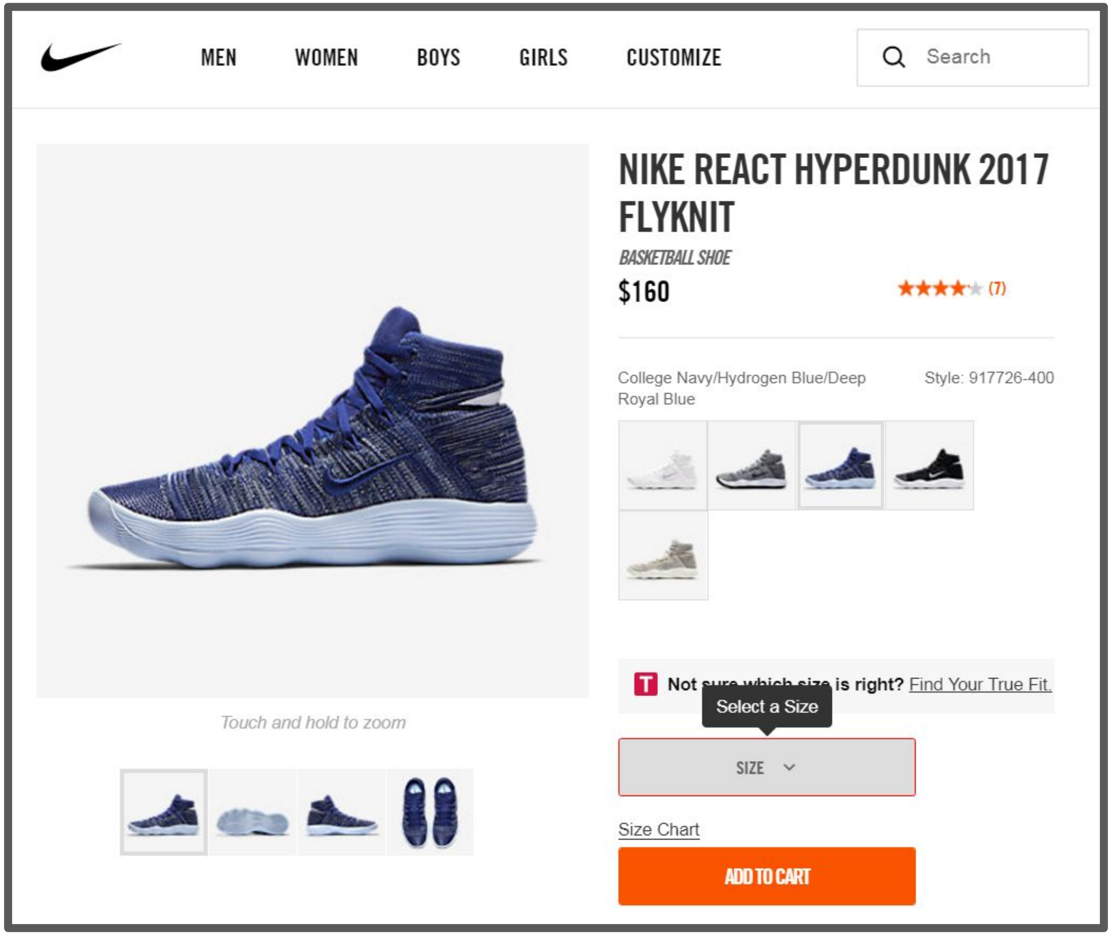
3. Define your intended sample size & stopping criteria (will revisit)

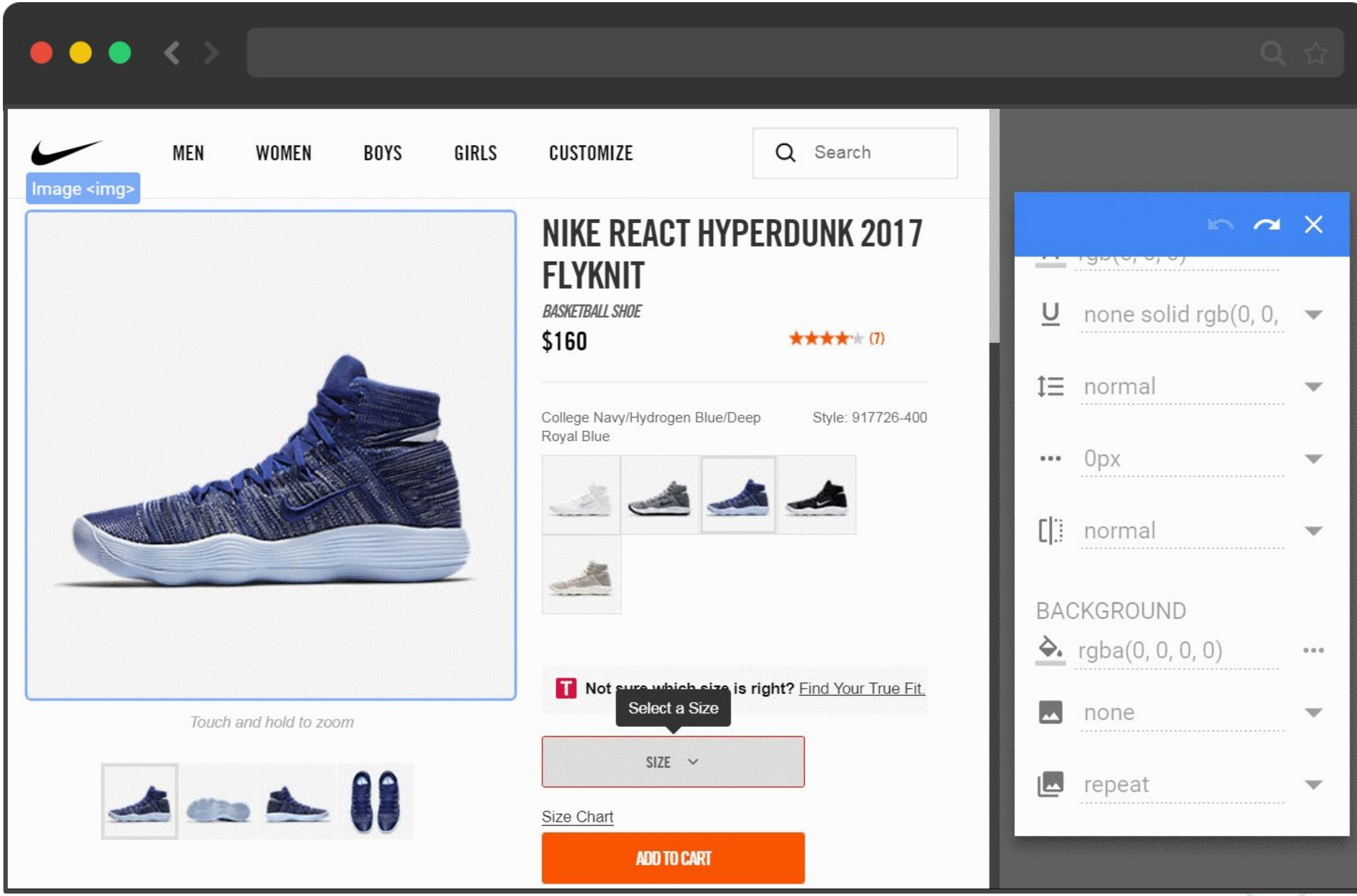
4. Run your experiment: Randomly assign customers to treatment arms

✅ 5. Evaluate your results:
   - Implement the "winning" arm

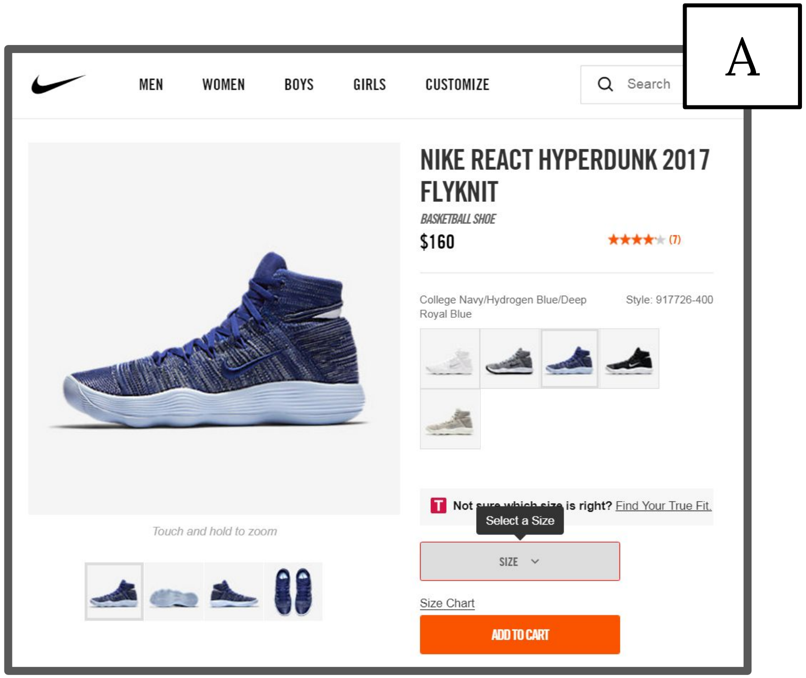
Walkthrough: Optimize Nike product page

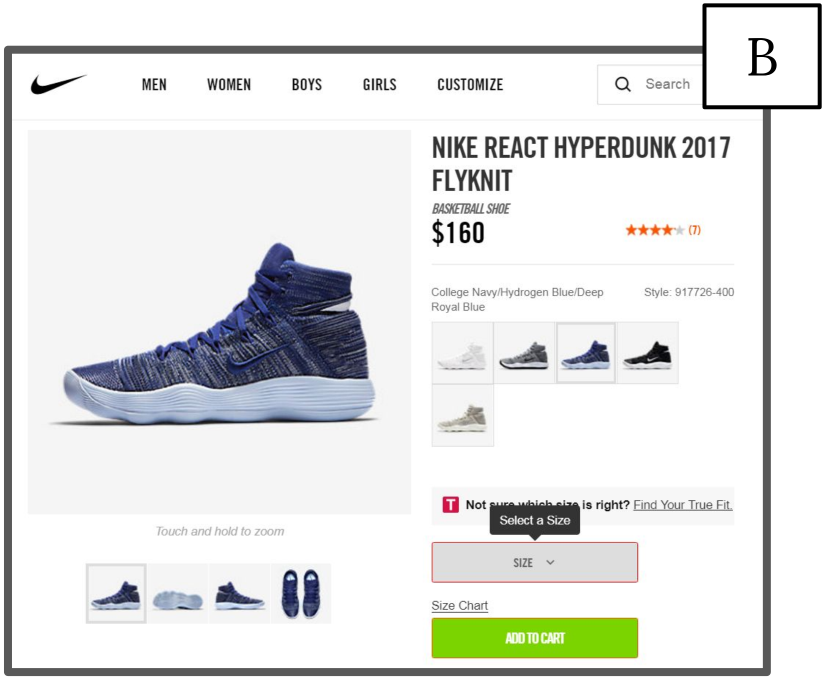
Suppose a UX designer has a new idea for how the product page should look:



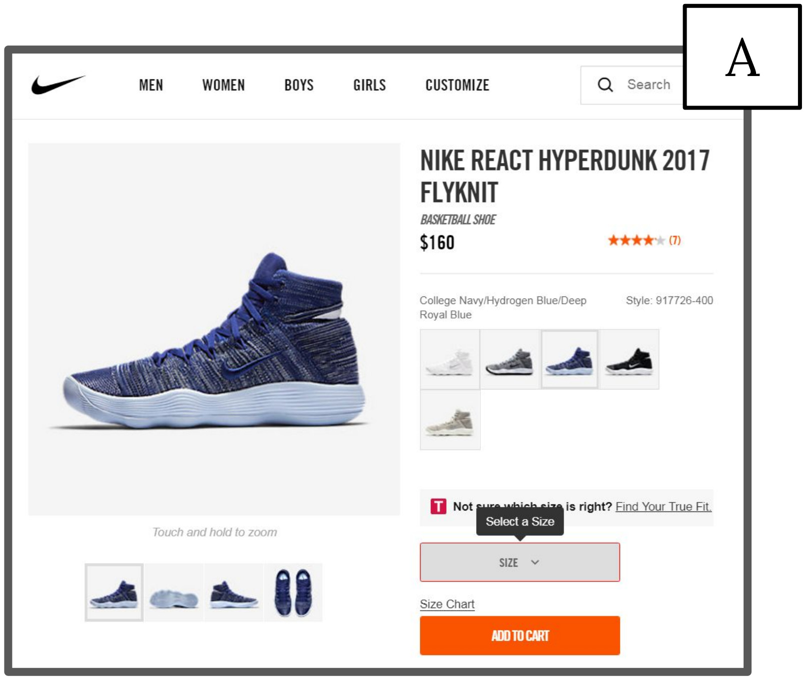

MEN
WOMEN
BOYS
GIRLS
CUSTOMIZE
Search
Image <img>
NIKE REACT HYPERDUNK 2017 FLYKNIT
BASKETBALL SHOE
$160
(7)
College Navy/Hydrogen Blue/Deep Royal Blue
Style: 917726-400
Not sure which size is right? Find Your True Fit.
Select a Size
SIZE
Size Chart
ADD TO CART
Touch and hold to zoom
none solid rgb(0, 0,
normal
0px
normal
BACKGROUND
rgba(0, 0, 0, 0)
none
repeat

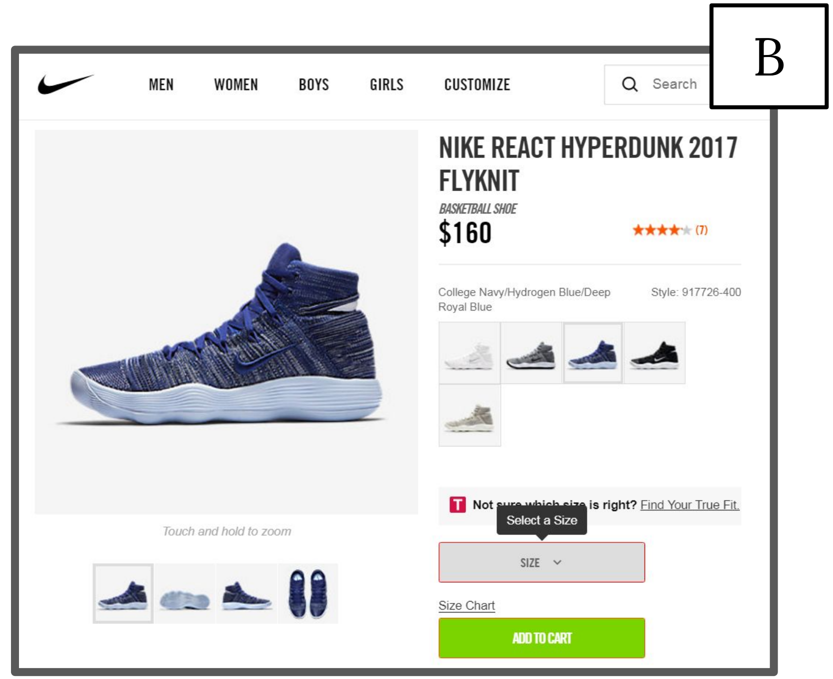
Hypotheses? ✅


A
MEN
WOMEN
BOYS
GIRLS
CUSTOMIZE
Search
NIKE REACT HYPERDUNK 2017 FLYKNIT
BASKETBALL SHOE
$160
(7)
College Navy/Hydrogen Blue/Deep Royal Blue
Style: 917726-400
Touch and hold to zoom
Not
is right? Find Your True Fit.
Select a Size
SIZE
Size Chart
ADD TO CART


B
MEN
WOMEN
BOYS
GIRLS
CUSTOMIZE
Search
NIKE REACT HYPERDUNK 2017 FLYKNIT
BASKETBALL SHOE
$160
(7)
College Navy/Hydrogen Blue/Deep Royal Blue
Style: 917726-400
Touch and hold to zoom
Not
is right? Find Your True Fit.
Select a Size
SIZE
Size Chart
ADD TO CART

Hypotheses? ✅

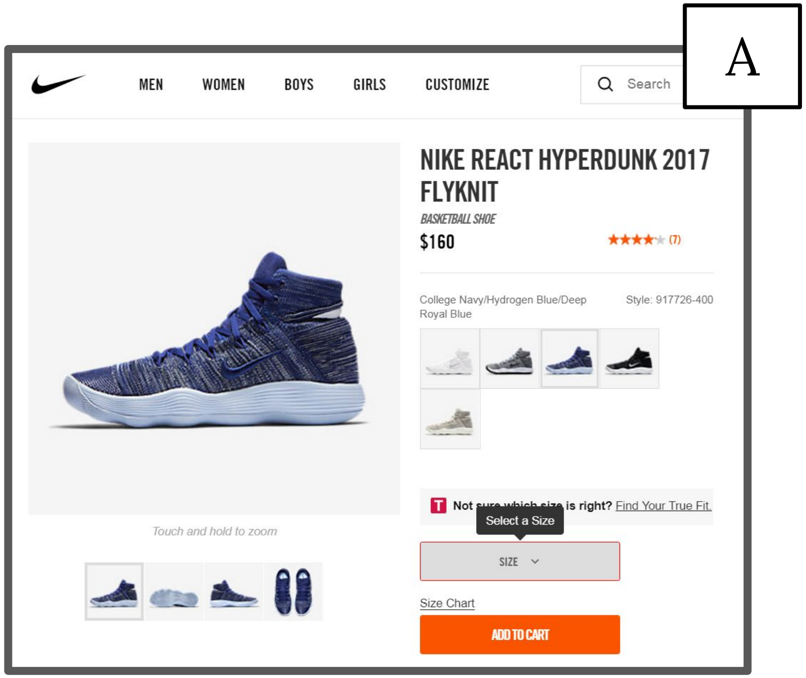


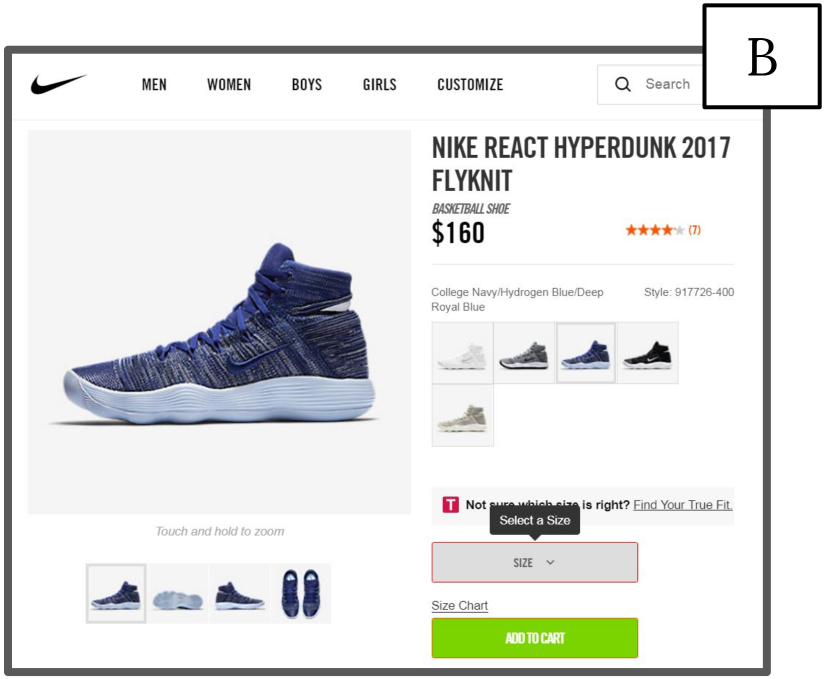


Evaluation criterion?

How long to run?

Hypotheses? ✅

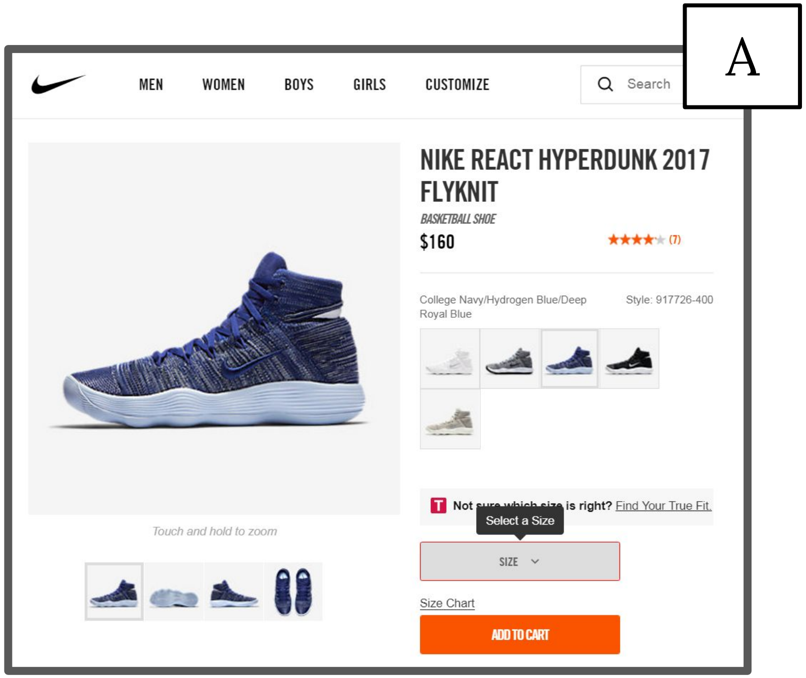


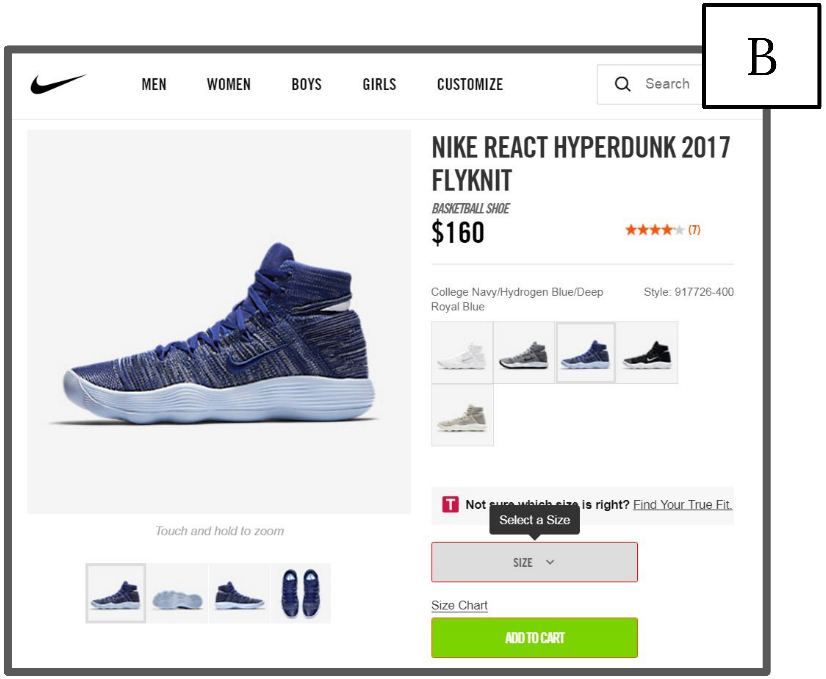


Evaluation criterion? Conversion rate ✅

How long to run?

Hypotheses? ✅



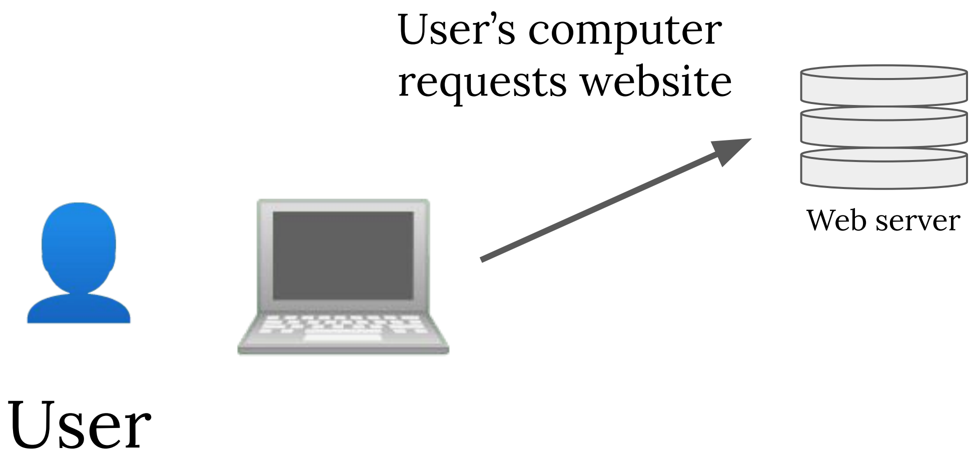


Evaluation criterion? Conversion rate ✅

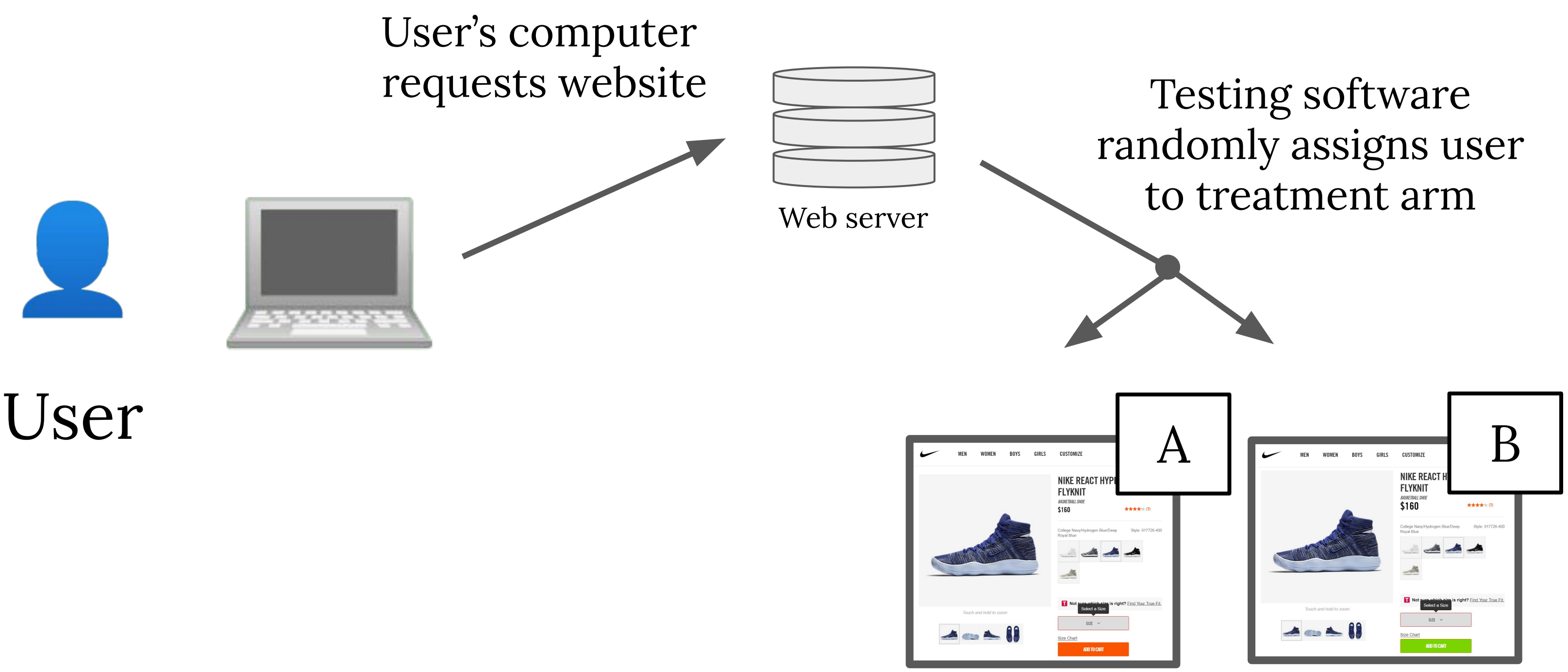
How long to run? 1 week ✅

Run experiment: A/B Test in Action

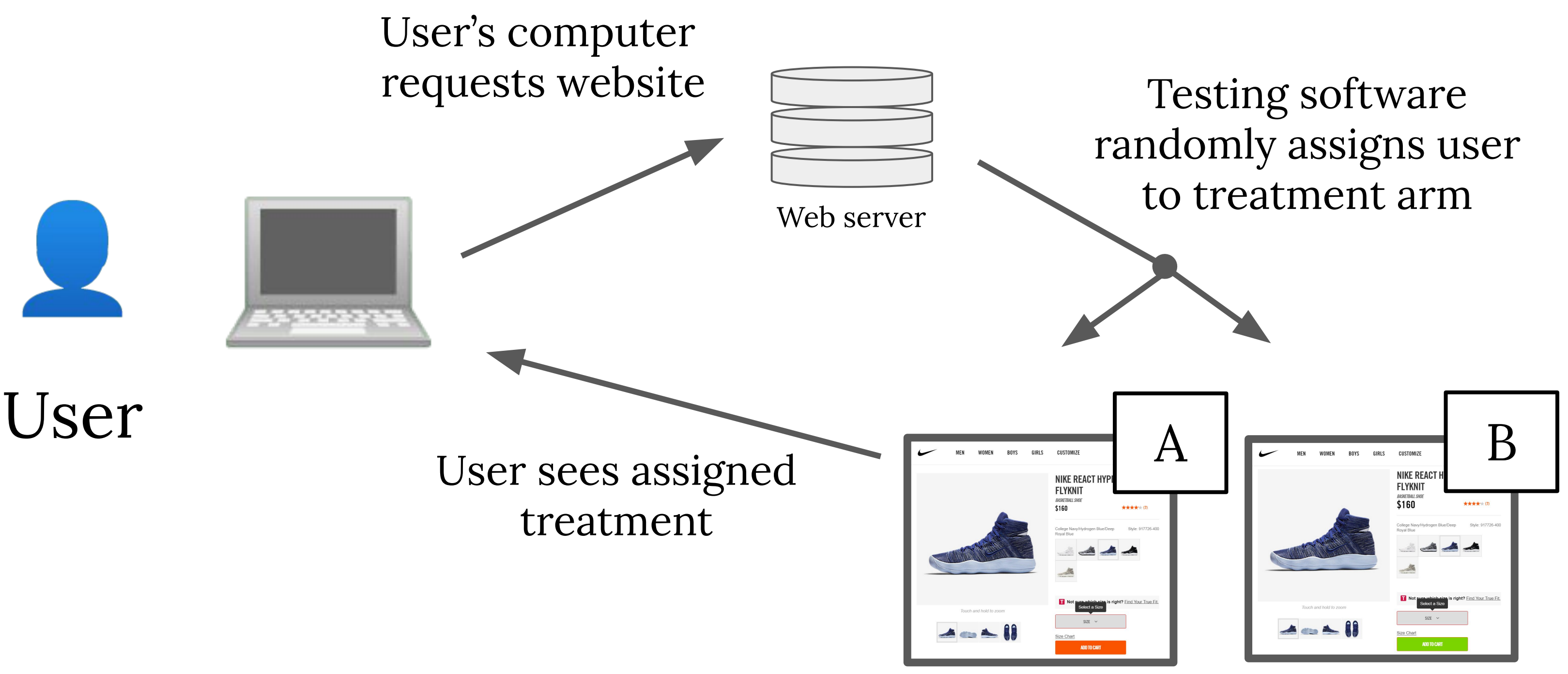
Run experiment: A/B Test in Action



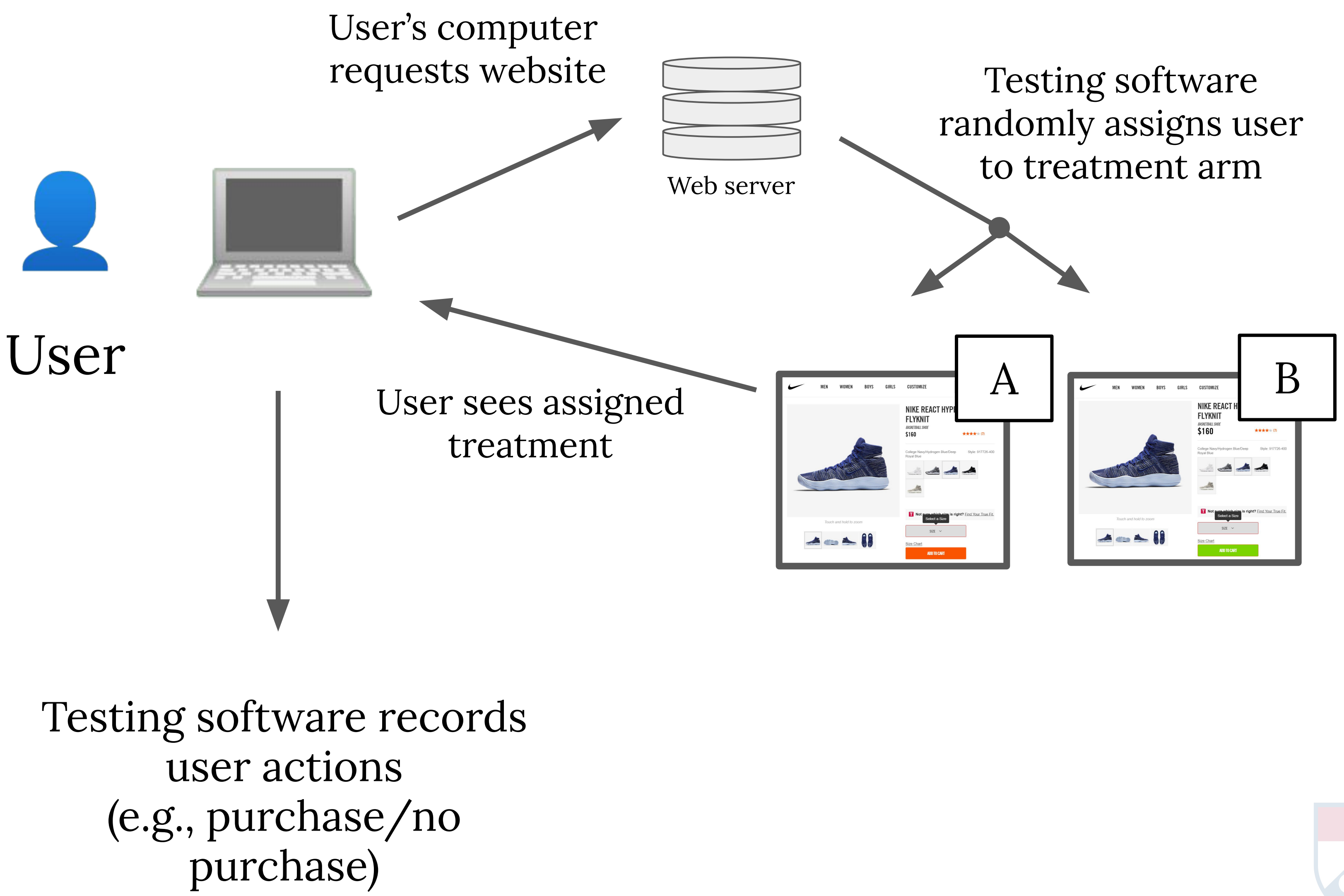
Run experiment: A/B Test in Action



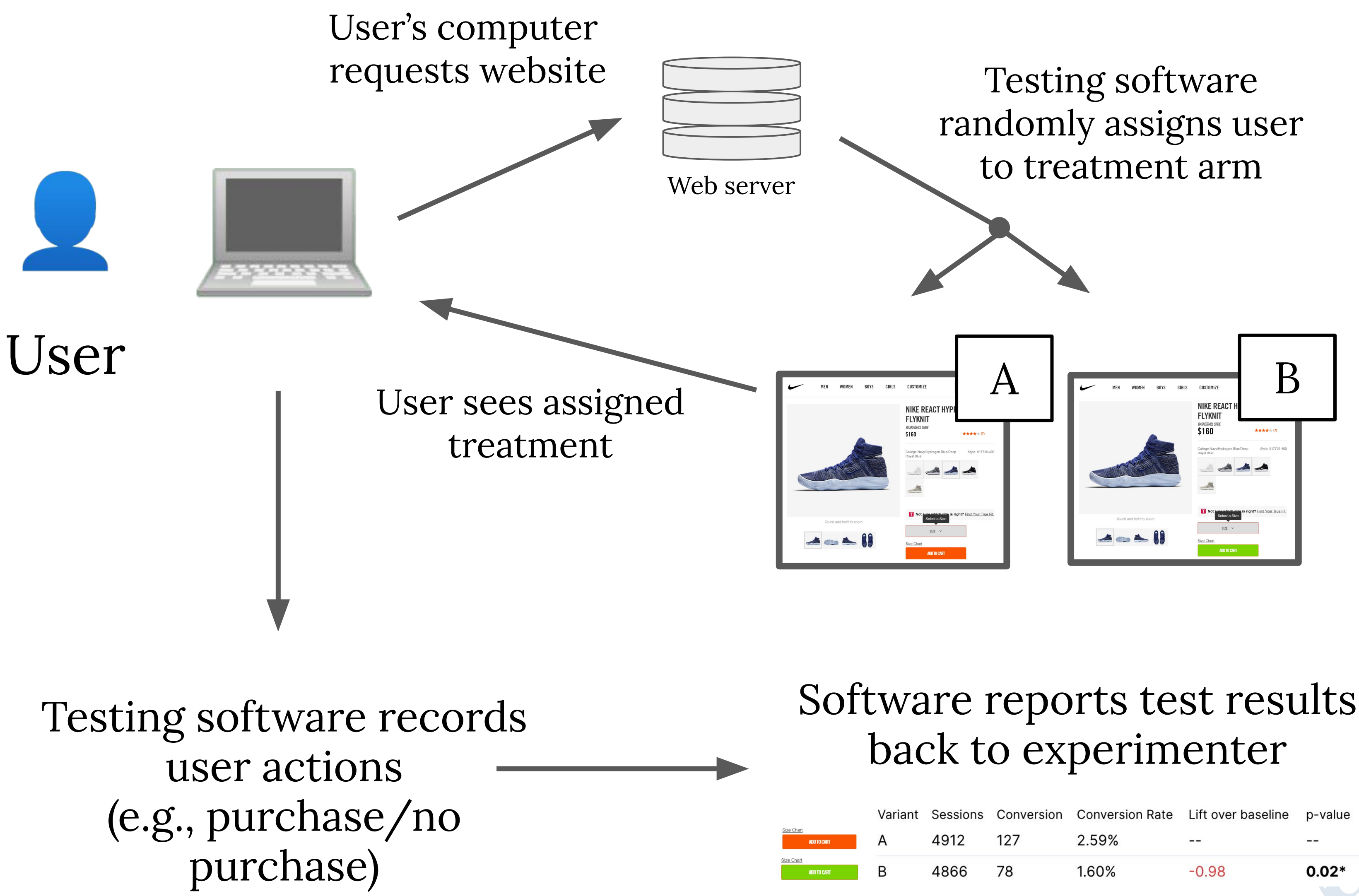
Run experiment: A/B Test in Action



Run experiment: A/B Test in Action



Run experiment: A/B Test in Action



Variant	Sessions	Conversion	Conversion Rate	Lift over baseline	p-value
A	4912	127	2.59%	--	--
B	4866	78	1.60%	-0.98	**0.02***

Evaluating the results from an A/B test

Sample Dashboard (simulated data)

	Variant	Sessions	Conversion	Conversion Rate	Lift over baseline	p-value
	A					
	B					

Evaluating the results from an A/B test

Sample Dashboard (simulated data)

	Variant	Sessions	Conversion	Conversion Rate	Lift over baseline	p-value
	A	4912				
	B	4866				

Evaluating the results from an A/B test

Sample Dashboard (simulated data)

	Variant	Sessions	Conversion	Conversion Rate	Lift over baseline	p-value
	A	4912	127			
	B	4866	78			

Evaluating the results from an A/B test

Sample Dashboard (simulated data)

	Variant	Sessions	Conversion	Conversion Rate	Lift over baseline	p-value
	A	4912	127	2.59%		
	B	4866	78	1.60%		

Evaluating the results from an A/B test

Sample Dashboard (simulated data)

	Variant	Sessions	Conversion	Conversion Rate	Lift over baseline	p-value
Size Chart ADD TO CART	A	4912	127	2.59%	--	
Size Chart ADD TO CART	B	4866	78	1.60%	-0.98	

"Effect size"

Evaluating the results from an A/B test

Sample Dashboard (simulated data)

	Variant	Sessions	Conversion	Conversion Rate	Lift over baseline	p-value
	A	4912	127	2.59%	--	--
	B	4866	78	1.60%	-0.98	**0.02***

Evaluating the results from an A/B test

Sample Dashboard (simulated data)

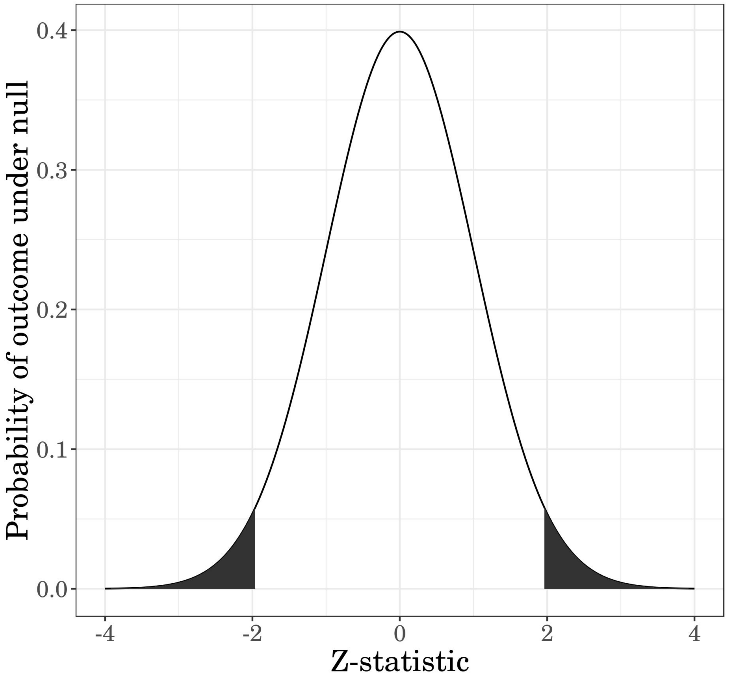
	Variant	Sessions	Conversion	Conversion Rate	Lift over baseline	p-value
	A	4912	127	2.59%	--	--
	B	4866	78	1.60%	-0.98	**0.02***

- This dashboard reports raw "*p*-values"
- It is common to report 1-*p* as "confidence" (e.g., p=0.02 implies "98% confidence")
- Practices are changing, but this is very common paradigm in statistical software

How does statistics help?

Statistics provides a principled way to quantify how certain you should be about your results given:

- **the magnitude of effect you observed** and **your sample size**



In general: More data → more confidence the effect you measured is real

Common statistics can be difficult to interpret

The question you want to answer:

- What is the probability that version A is better than version B?

Common statistics can be difficult to interpret

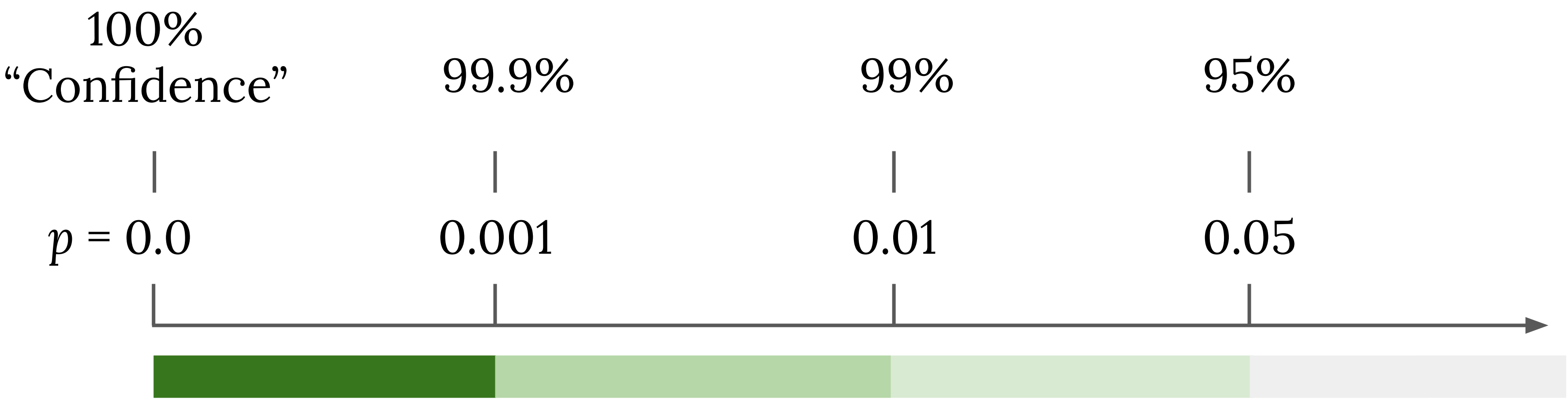
The question you want to answer:

- What is the probability that version A is better than version B?

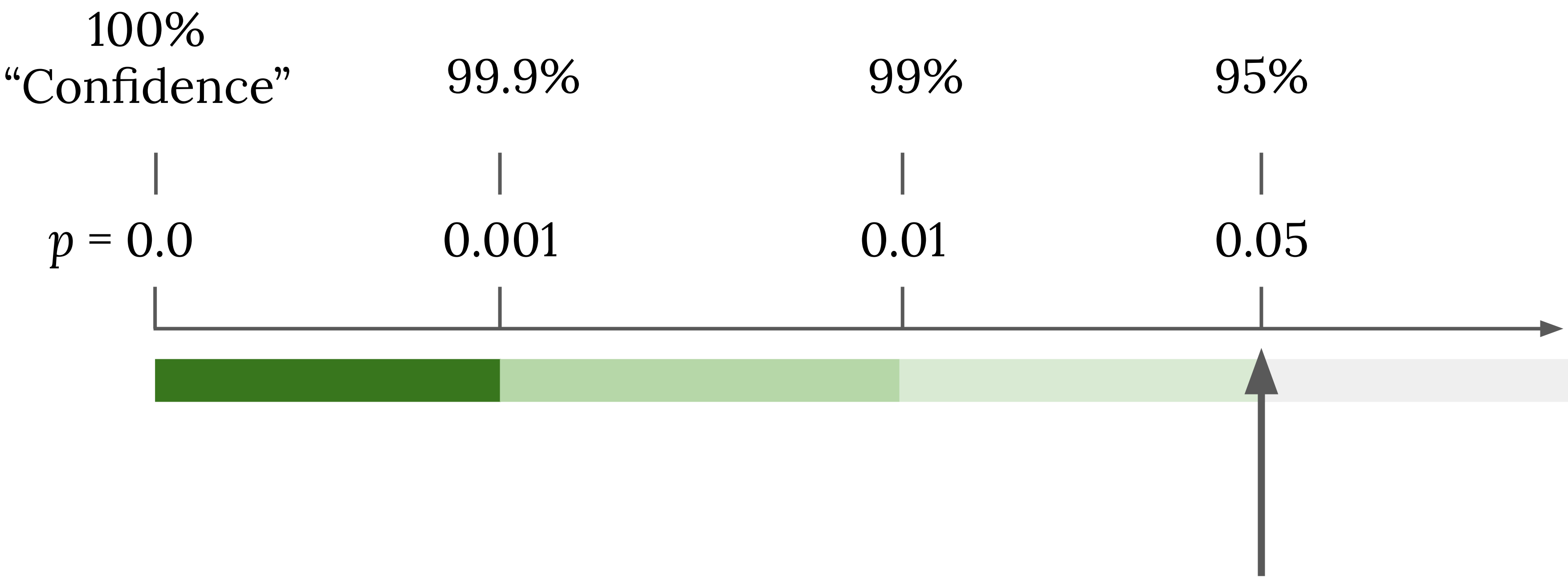
The question most A/B testing tools answer (those based on p-values or "Frequentist" statistics):

- Assuming there were no difference between versions A & B, what is the chance I would have observed a result as (or more extreme) than the result I observed in this experiment?

p-values for humans (rules of thumb)

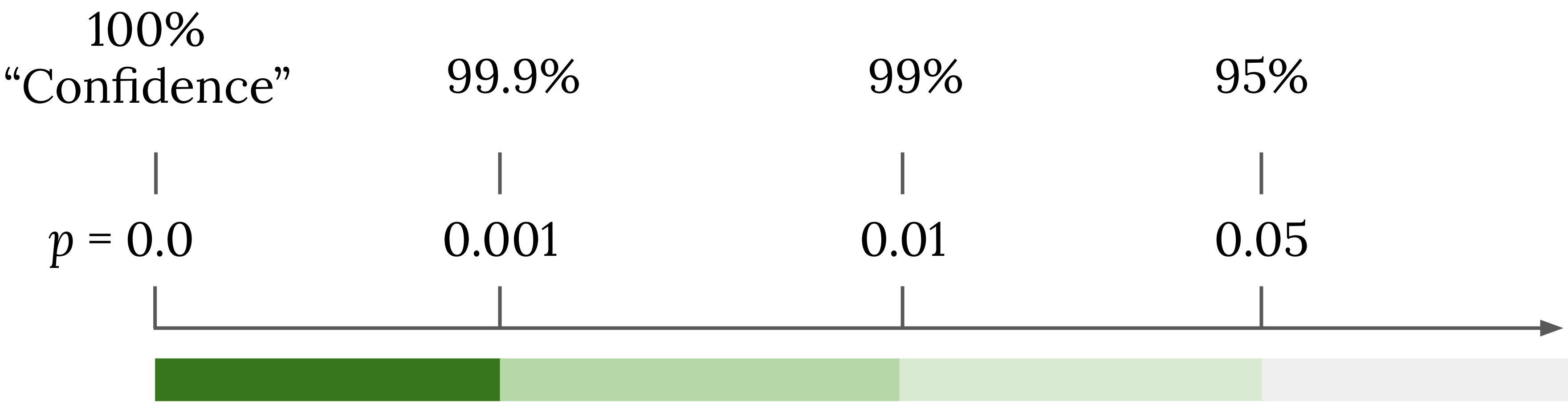


p-values for humans (rules of thumb)

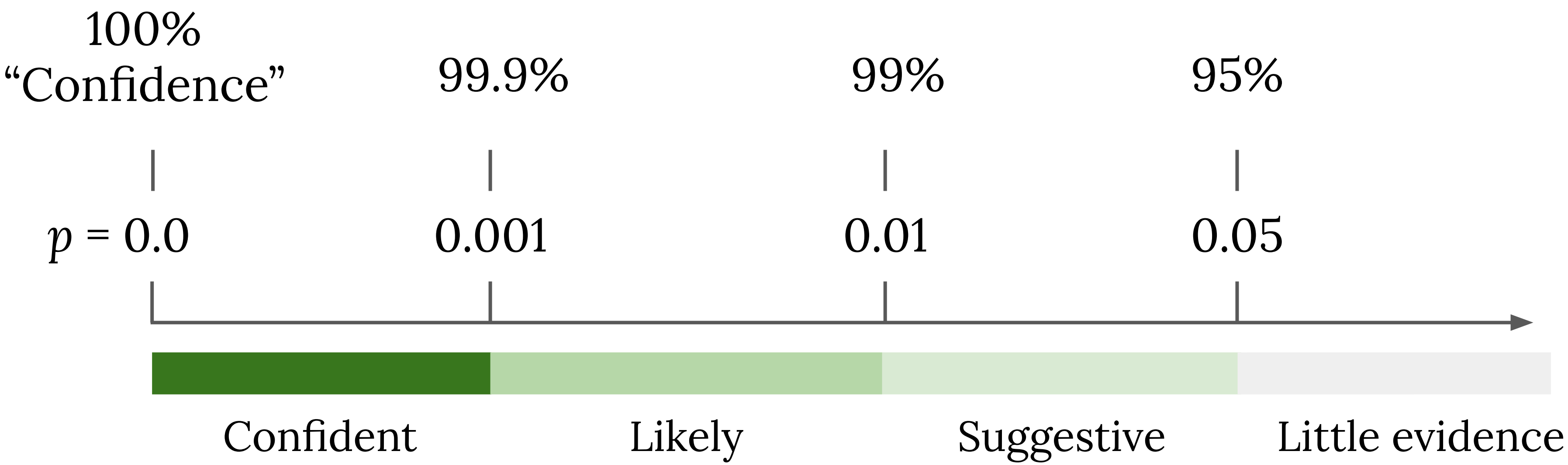


- The most common rule of thumb is to say a $p<0.05$ is “statistically significant”
- There is nothing magic about $p=0.05$! (or “95% confidence”)

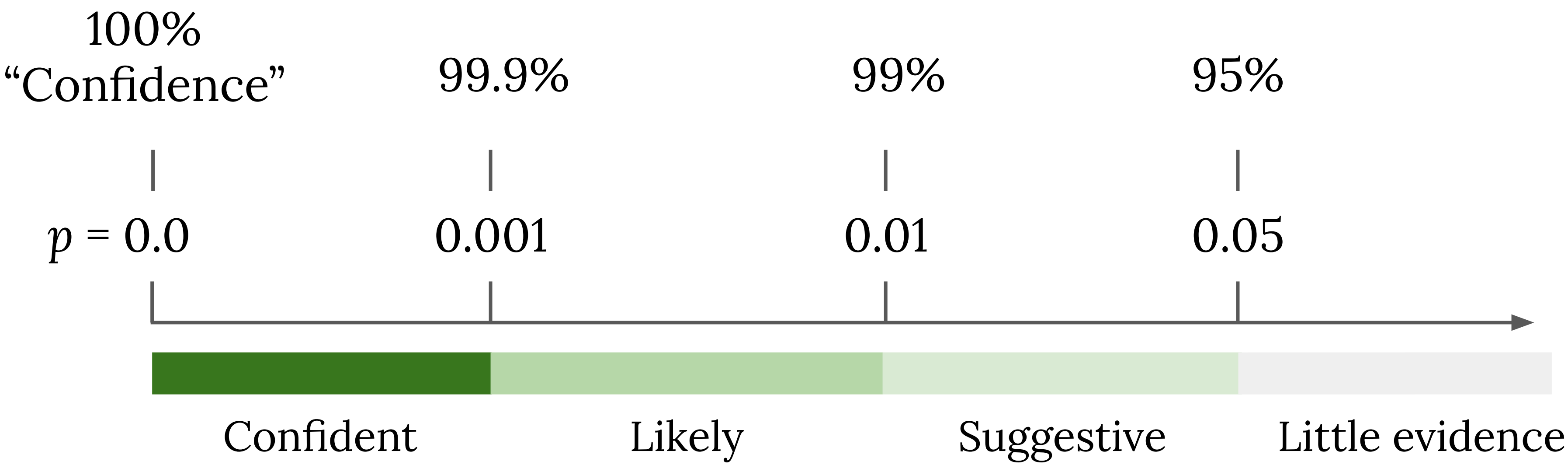
p-values for humans (rules of thumb)



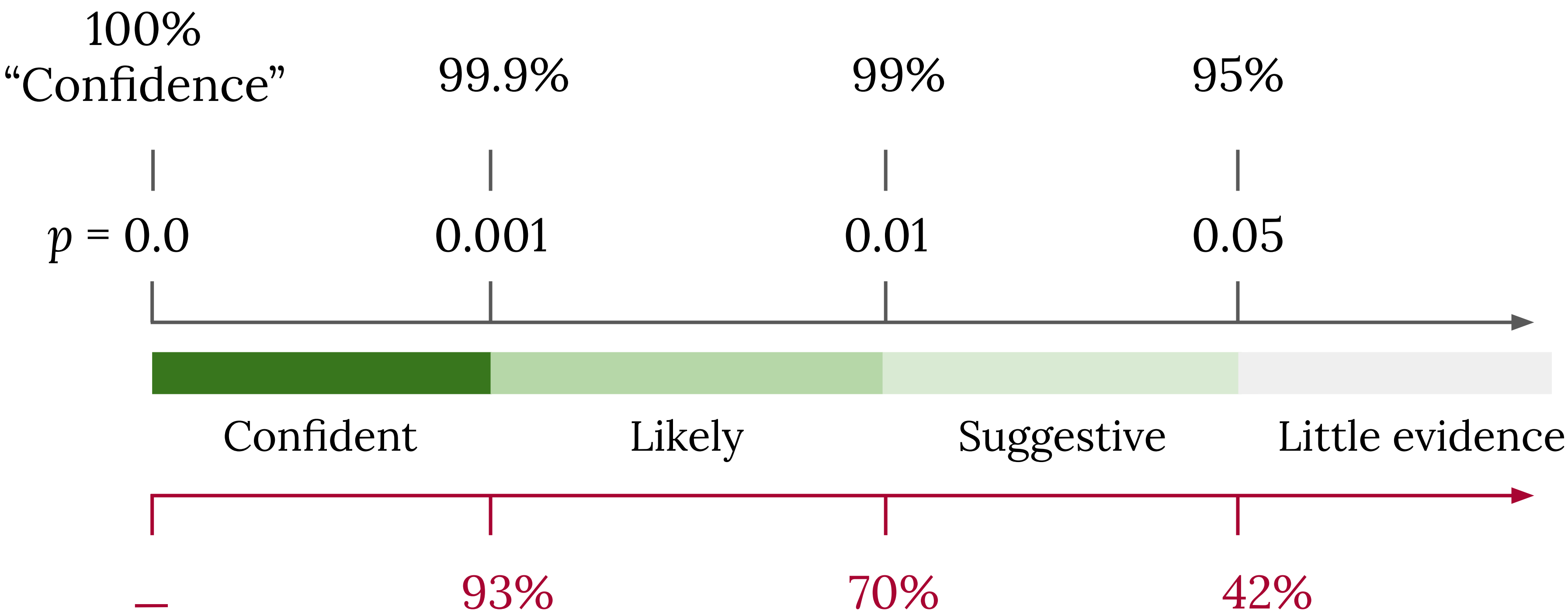
p-values for humans (rules of thumb)



p-values for humans (rules of thumb)

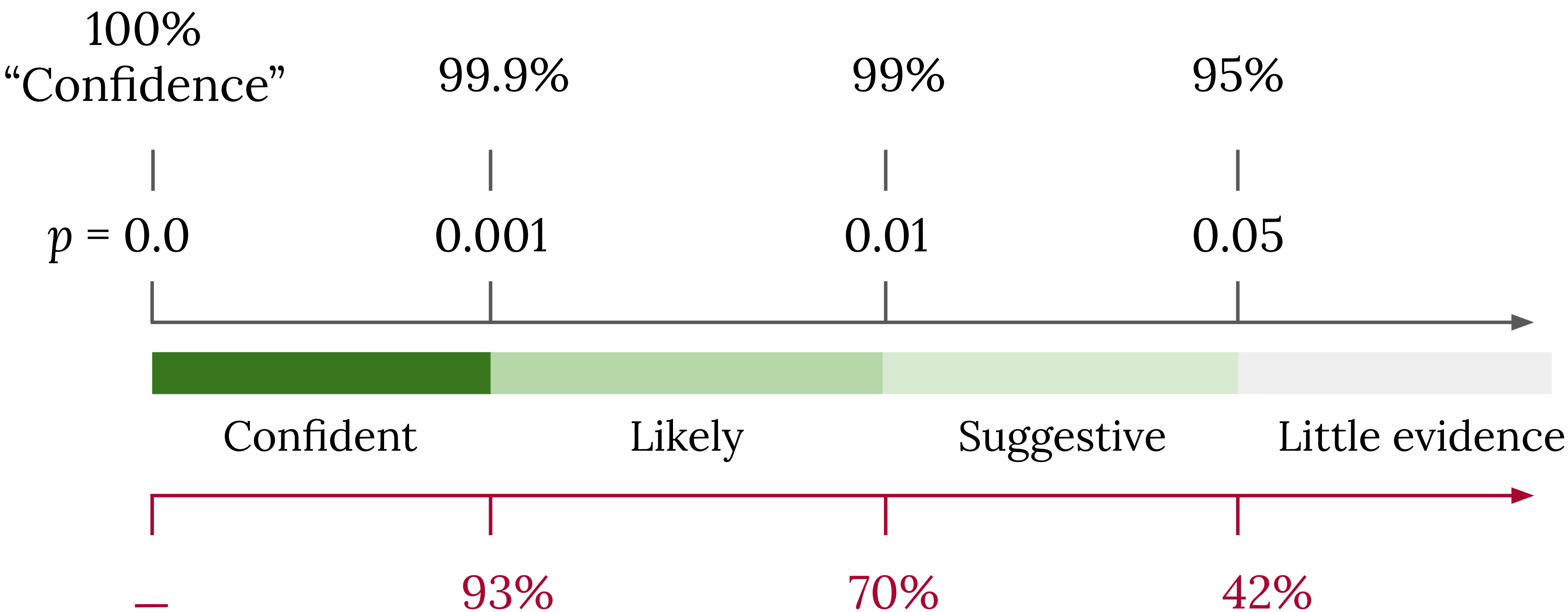


p-values for humans (rules of thumb)



My research suggests that the true probability of observing a non-zero effect at the given *p*-value levels is much, much lower than naive "confidence" levels

p-values for humans (rules of thumb)



My research suggests that the true probability of observing a non-zero effect at the given p-value levels is much, much lower than naive “confidence” levels

Evaluating the results from an A/B test

Sample Dashboard (simulated data)

	Variant	Sessions	Conversion	Conversion Rate	Lift over baseline	p-value
	A	4912	127	2.59%	--	--
	B	4866	78	1.60%	-0.98	**0.02***

- To conclude this example:
  - It appears quite likely that the "A" variant (i.e., orange button) has a higher conversion rate than the "B" variant (green button)
  - Decision: Keep orange button

Testing Paradigms for Business Decisions

The importance of...

Understanding and Defining the Goal of A/B Tests

Statistics in the real world

- There's a fundamental trade-off in statistics:

Precision ⟷ **Speed**

Larger sample sizes — **Smaller sample sizes**

Statistics in the real world

- There's a fundamental trade-off in statistics:

Precision ⟷ **Speed**

Larger sample sizes ⟷ **Smaller sample sizes**

- It's useful to think about the goals of an experiment as falling into one of two paradigms:

Hypothesis Testing

Metric Optimization

Hypothesis Testing

- You come to the table with a set of predetermined hypotheses

- Primary concerns:

  - Trying to learn something fundamental about your customer

  - To measure and quantify the difference between arms **with precision**

  - The correct choice is made between A & B (making a mistake has external costs)

Metric Optimization

- The primary goal is to maximize a particular metric (e.g., conversion rate, revenue) over a fixed period of time

- You care less about:
  - making the best decision 100% of the time
  - exactly why or how things work

Metric Optimization

- The primary goal is to maximize a particular metric (e.g., conversion rate, revenue) over a fixed period of time

- You care less about:
  - making the best decision 100% of the time
  - exactly why or how things work

Fixed period of time

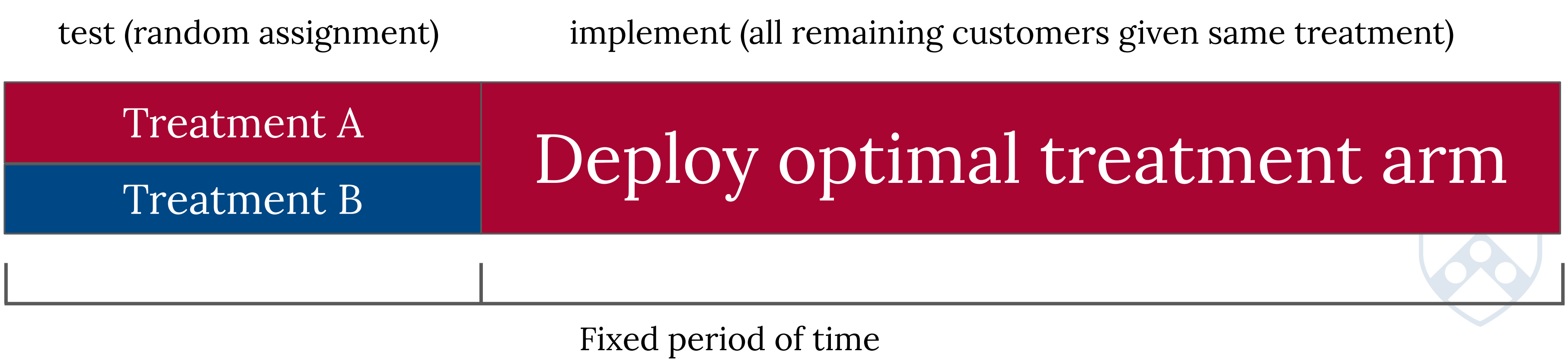
Metric Optimization

- The primary goal is to maximize a particular metric (e.g., conversion rate, revenue) over a fixed period of time

- You care less about:
  - making the best decision 100% of the time
  - exactly why or how things work

Metric Optimization

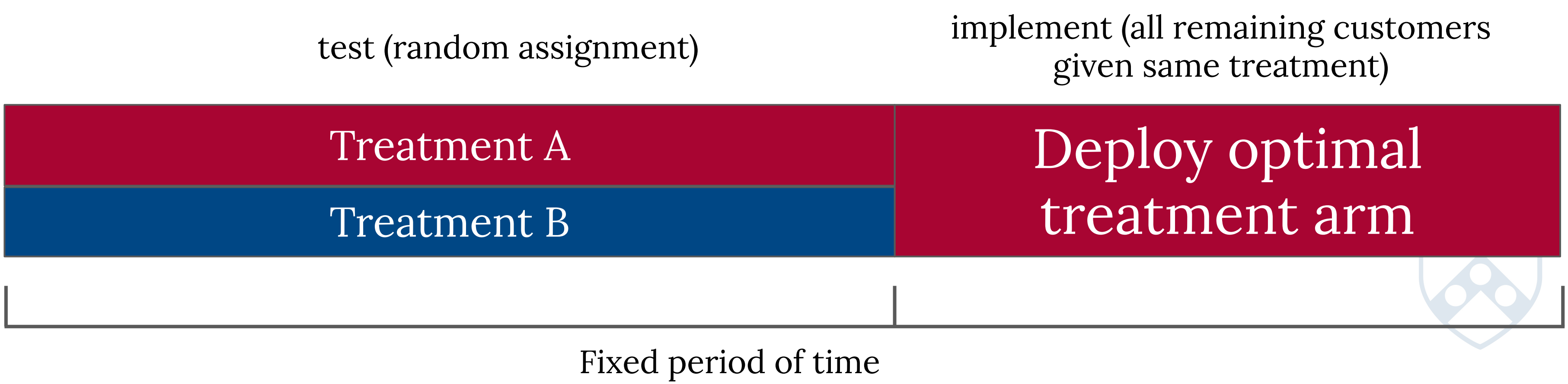
- The primary goal is to maximize a particular metric (e.g., conversion rate, revenue) over a fixed period of time
- You care less about:
  - making the best decision 100% of the time
  - exactly why or how things work



Metric Optimization

- The primary goal is to maximize a particular metric (e.g., conversion rate, revenue) over a fixed period of time

- You care less about:
  - making the best decision 100% of the time
  - exactly why or how things work



Which paradigm is “correct”?

- Neither; both have valid use-cases and they aren’t even necessarily mutually exclusive

Which paradigm is "correct"?

- Neither; both have valid use-cases and they aren't even necessarily mutually exclusive

- However:

    - Sample sizes needed for very precise experiments are much larger than many people realize

    - "Optimization" paradigm more closely matches most scenarios I've encountered in A/B testing

Sample size example using classical "significance" and "power" levels

Suppose website conversion rate is 5%...

- To detect a
  - **0.5%** absolute difference (~10% relative difference)
- You need: **90,000 observations**

- To detect a
  - **0.1%** absolute difference (2% relative difference)
- You need: **1 million+ observations**

In my research at medium-to-large e-commerce firms,
half of all A/B tests have effect sizes smaller than 0.1% (in absolute terms)

Note on sample size calculations

- I highly encourage you to play around with a sample size calculator:
  e.g., https://www.evanmiller.org/ab-testing/sample-size.html
- Can be very valuable for setting sample sizes ahead of time when in the "hypothesis testing" paradigm
  - i.e., can give you principled reasons for knowing when to stop an experiment
- This will help you develop intuition about the magnitude of effect sizes that you can expect to detect at your company's scale

Why classical notions of “significance” may be irrelevant for many A/B tests

- Classical “statistical significance” are based on “false positive control” guarantees
    - “False positive”: You conclude there is a true difference between A & B, when in reality there is no difference
    - 5% significance level = 5% of results will be false positive

Why classical notions of “significance” may be irrelevant for many A/B tests

- Classical “statistical significance” are based on “false positive control” guarantees
  - “False positive”: You conclude there is a true difference between A & B, when in reality there is no difference
  - 5% significance level = 5% of results will be false positive
- This is very valuable when precision is important and false positives are costly…
  - but is this really the main thing you care about when making business decisions?

Why classical notions of "significance" may be irrelevant for many A/B tests

- For many business decisions, "false positives" are not that costly
  - Often by the time some variation can be tested in an experiment, most of the design/development work is already done

Why classical notions of "significance" may be irrelevant for many A/B tests

- For many business decisions, "false positives" are not that costly
    - Often by the time some variation can be tested in an experiment, most of the design/development work is already done
- If there is no difference between A & B, and the cost to implement both is negligible, it really doesn't matter if you make a "wrong" decision
- Precision is less important → Metric optimization paradigm can be more useful
    - Smaller sample sizes with less "significance" can be okay

Hypothesis Testing

“precision mindset”

Metric Optimization

“risk mindset”

Hypothesis Testing

"precision mindset"

Metric Optimization

"risk mindset"

- Precision matters
- False positives are costly

Hypothesis Testing

“precision mindset”

- Precision matters
- False positives are costly

Metric Optimization

“risk mindset”

- Precision is “nice to have”, but maximizing profits is the primary goal
- False positives are less costly

Key insight #1 for using A/B testing within a “metric optimization” framework:

Key insight #1 for using A/B testing within a “metric optimization” framework:

- If there is a big difference between variations A & B, it will be obvious!
  - You don’t need millions of observations
- If there is a small difference between variations A & B, it is not costly to make the wrong decision
  - *“If I couldn’t detect an effect after 1 month, it’s too small to stress about.”*

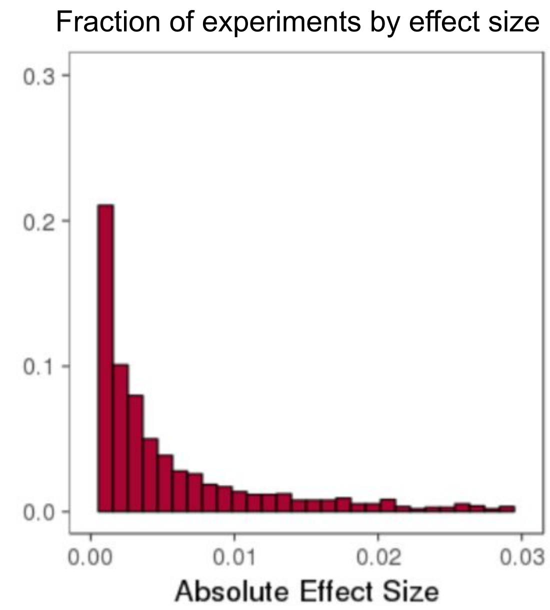
Key insight #1 for using A/B testing within a “metric optimization” framework:

- If there is a big difference between variations A & B, it will be obvious!
  - You don’t need millions of observations
- If there is a small difference between variations A & B, it is not costly to make the wrong decision
  - *“If I couldn’t detect an effect after 1 month, it’s too small to stress about.”*
- **With smaller samples, you won’t get every decision correct, but you will get the big ones**

Key insight #2 for using A/B testing within a “metric optimization” framework:

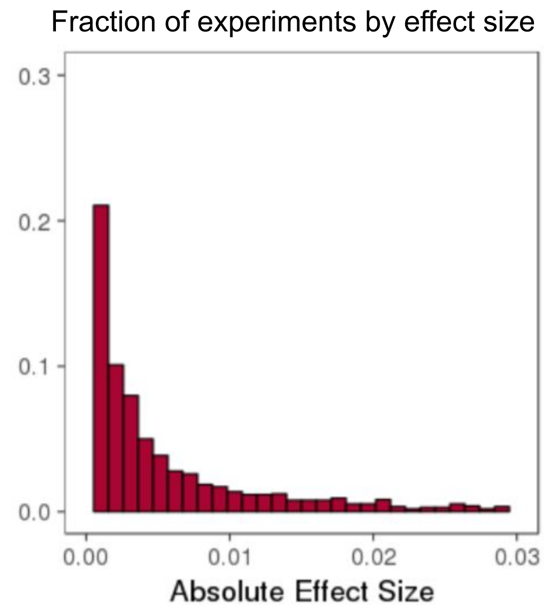
Key insight #2 for using A/B testing within a “metric optimization” framework:

- A/B test results follow the “Pareto principle”:
  - 80% of gains will be found in 20% of tests
  - Distribution of effect sizes →



Key insight #2 for using A/B testing within a "metric optimization" framework:

- A/B test results follow the "Pareto principle":
  - 80% of gains will be found in 20% of tests
  - Distribution of effect sizes →



- **Getting the most out of A/B testing consists of finding the few "big wins", rather than expecting gains from every attempt**
  - More shots on goal → More chances of scoring big

Upshot of both insights:

- **You will get more value by running MORE experiments with SMALLER sample sizes** compared to running fewer experiments with larger sample sizes

- Subject of recent research by Wharton professors:

Test & Roll: Profit-Maximizing A/B Tests

Elea McDonnell Feit
LeBow College of Business
Drexel University
eleafeit@gmail.com

Ron Berman
The Wharton School
University of Pennsylvania
ronber@wharton.upenn.edu

May 21, 2019

A/B Testing with Fat Tails[*]

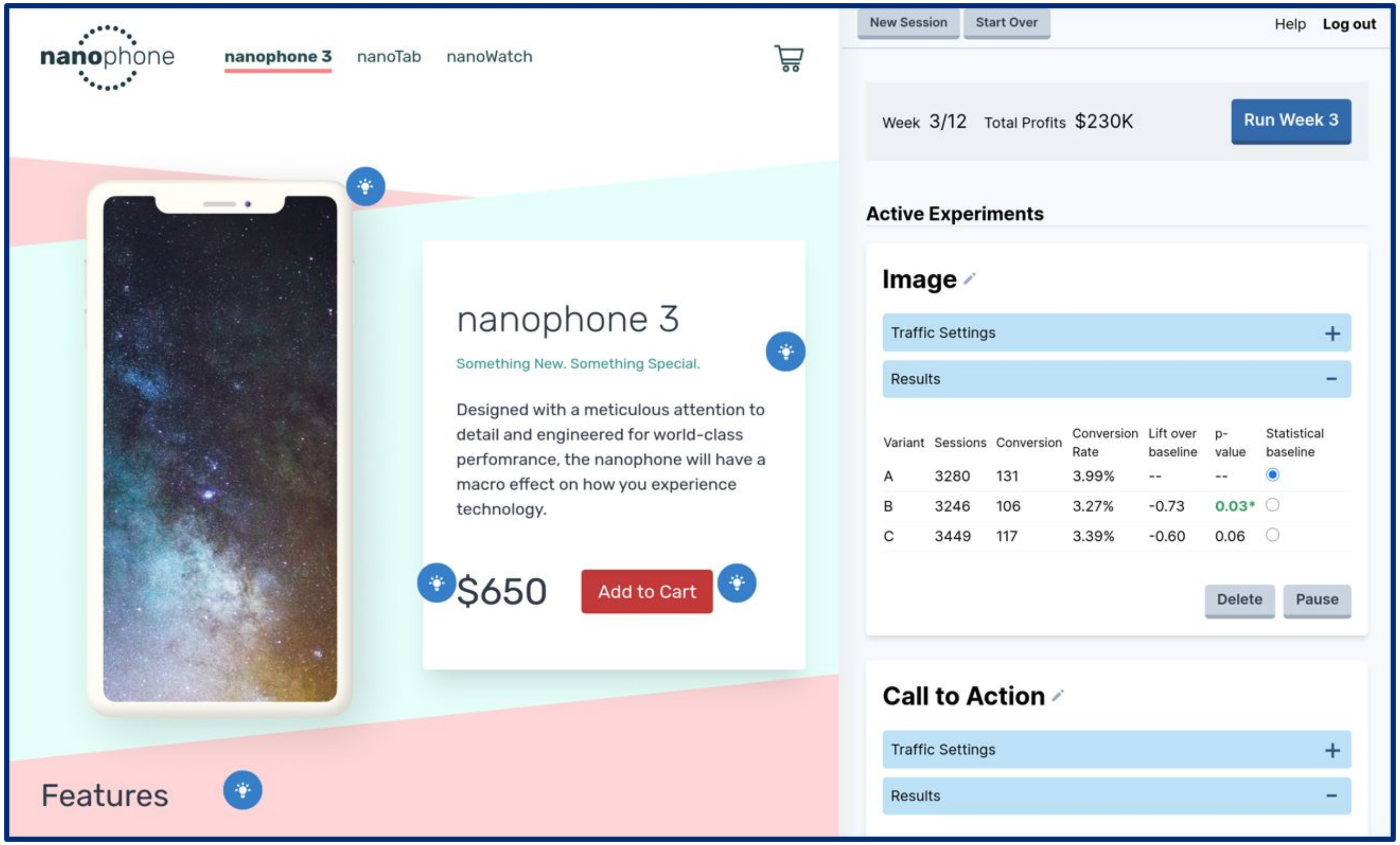
Eduardo M. Azevedo[†] Alex Deng[‡] José Luis Montiel Olea[§]
Justin Rao[¶] E. Glen Weyl[‖]

First version: April 30, 2018
This version: August 9, 2019

Simulation Exercise

- I've helped develop an interactive tool designed to:
    - Give you a hands-on feel of what it looks and feels like to run an e-commerce A/B test
    - Allow you to experience & internalize key principles of using A/B testing for decision making (covered in this session)
- We are making continuous improvements, so input/feedback is welcome

- I will give a brief demo of how to use the tool



Logistics

- I’ll be breaking you out into smaller rooms to form teams
    - 1st Stage: Practice mode (20 min)
        - → Familiarize yourself with the interface; discuss strategies for maximizing score with group
    - 2nd Stage: Competition Mode (15 min)
        - → Groups will compete by playing the same version of the game
    - Debrief (15 min)
        - → I’ll asking highest-scoring team(s) to describe their strategy

Practice Mode! (20min)

- Spend 5-10 minutes playing the game on your own to familiarize yourself with interface

- Think carefully about the objective of the game and how you can maximize your total profits at the end of the 12 week period

- Spend 5-10 minutes discussing your insights with your group

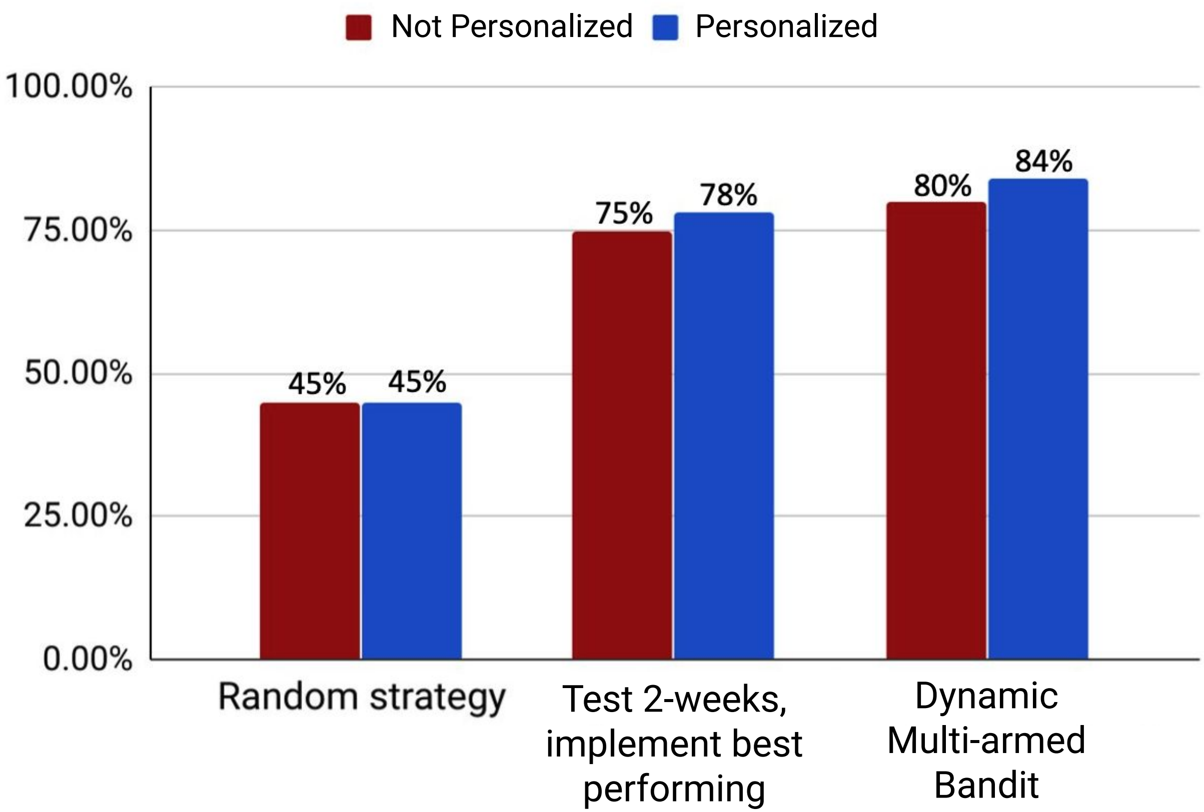
- Select ONE (1) person to act as your group's avatar

I'll reconvene whole session before moving to competition

Competition Mode (15-20min)

- You’ve had a chance to practice; now one member from each group will play in a “competition mode”
- One member from each group will click the competition link (shared in chat)
  - When in break-out room, share screen with your group and walk through the simulation
- Once finished, we’ll reconvene once more to compare scores & debrief

How do different strategies compare on average?



Dynamic "AI" based strategies only achieve marginal gains above a simple "explore first" strategies

Summary of key takeaways:

- If you really want precision, demand very small p-values and large sample sizes

- However, precision is costly and, in many situations, imprecision may not be that bad

- If you care about "Metric Optimization", adopt a risk mindset and lower your standards for precision:
    - Run more experiments, more quickly
    - Most gains come from finding the rare interventions with big effects; not precisely measuring typical interventions with small effects

Future of A/B Testing

- A/B testing + Machine Learning = Much more sophisticated personalization
  - e.g., Moving from targeting customers based on 2 variables (Location, Device) to 50 variables
  - Recent advances in ML make this easy/automatable in principled ways

- Testing platforms will move away from rules of thumb for decision making (e.g., p=0.05) and toward "Bayesian" paradigms based on data