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?

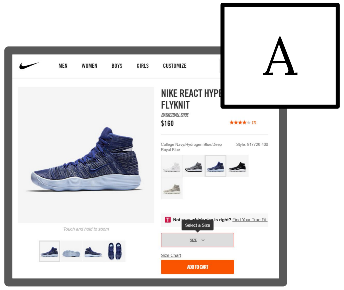
- A high-level understanding of how to use A/B testing tools to solve the **right** problem
  - Key aspects of using statistics for business decision making
  - Without getting bogged down in math

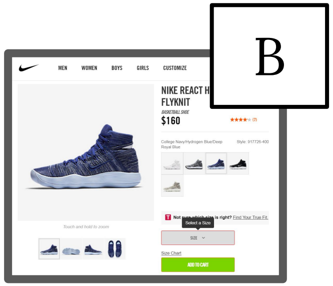
Core Concepts in A/B Testing

Definition:

A/B testing is:

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

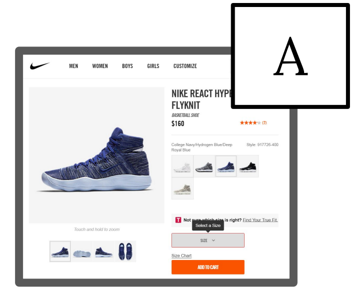


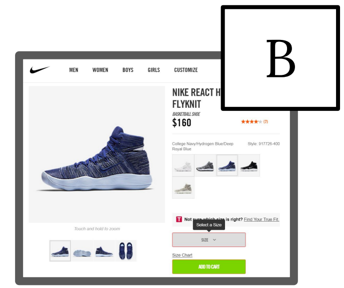


Definition:

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

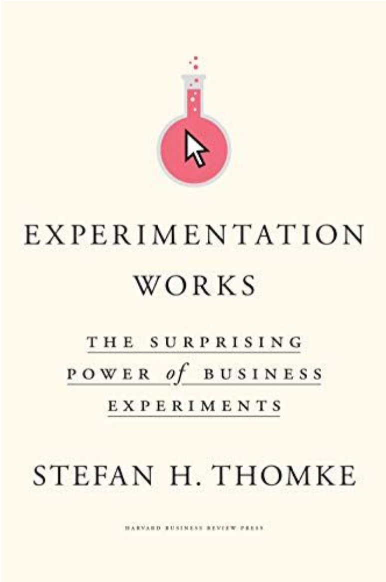
Optimizely

Google Optimize

Recommended Reading

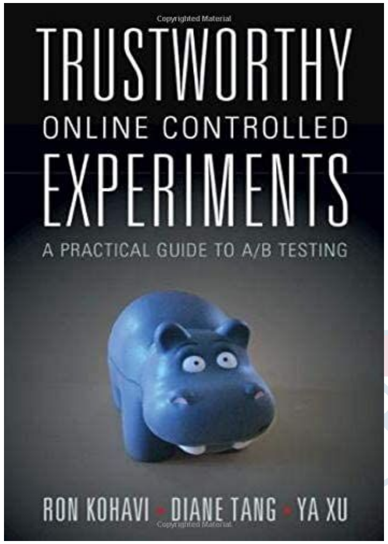
For more details on developing an experimental culture in your organization:

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:

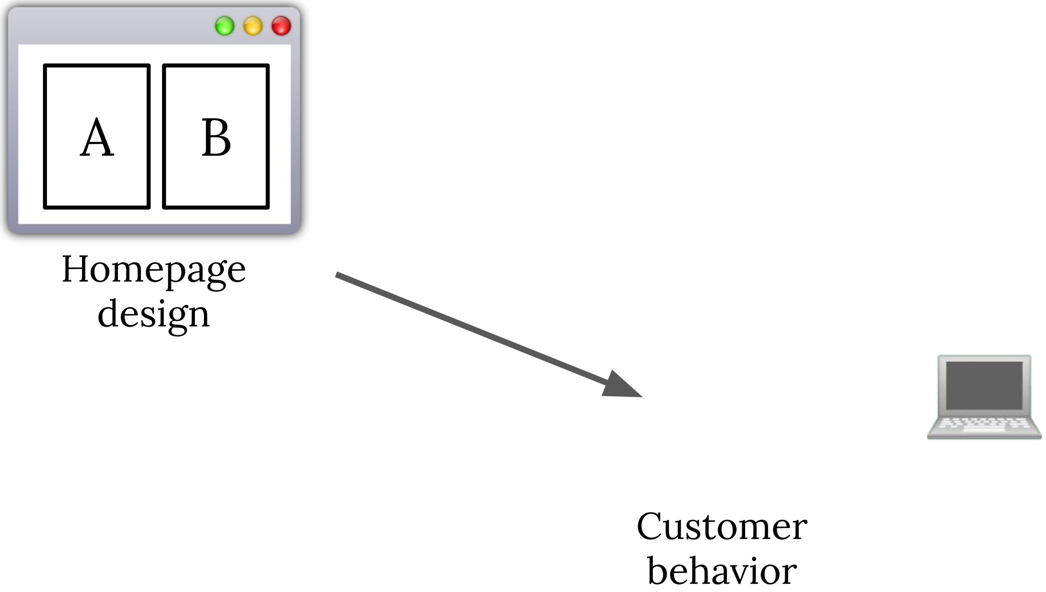
- “We redesigned our homepage last week and customer conversions increased”

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

How to tell the difference?

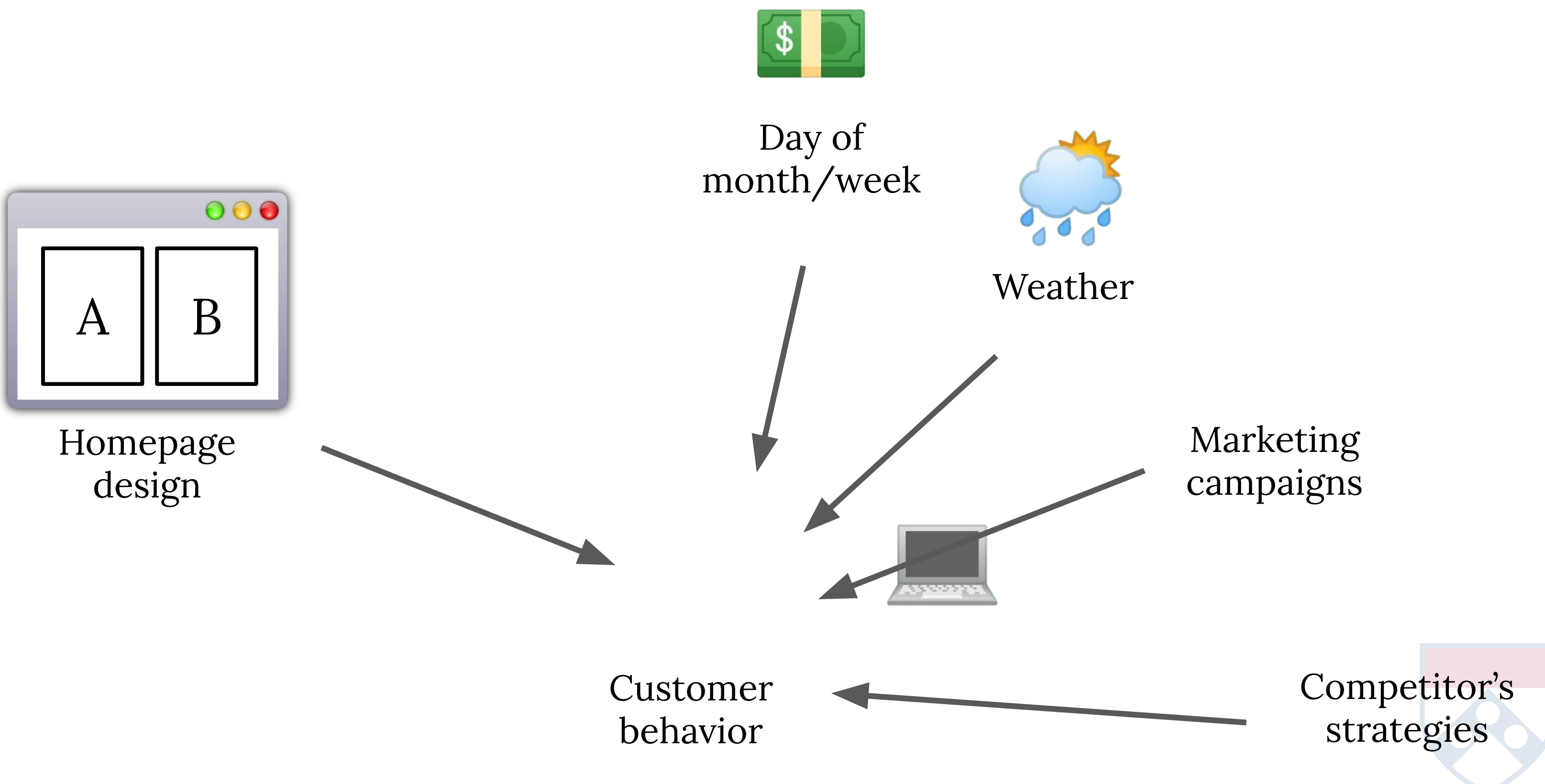
Why is this problem hard?

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

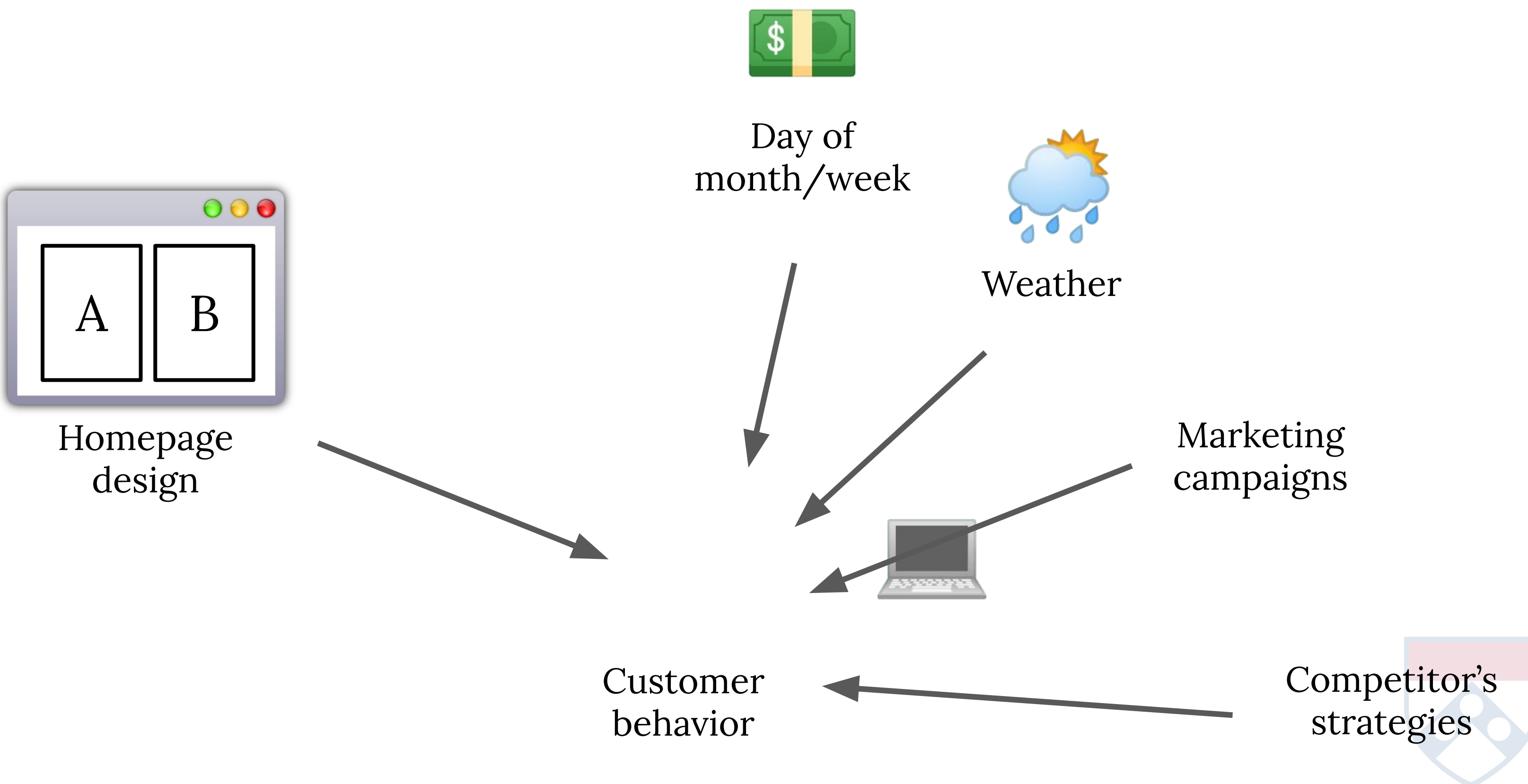


Why is this problem hard?

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

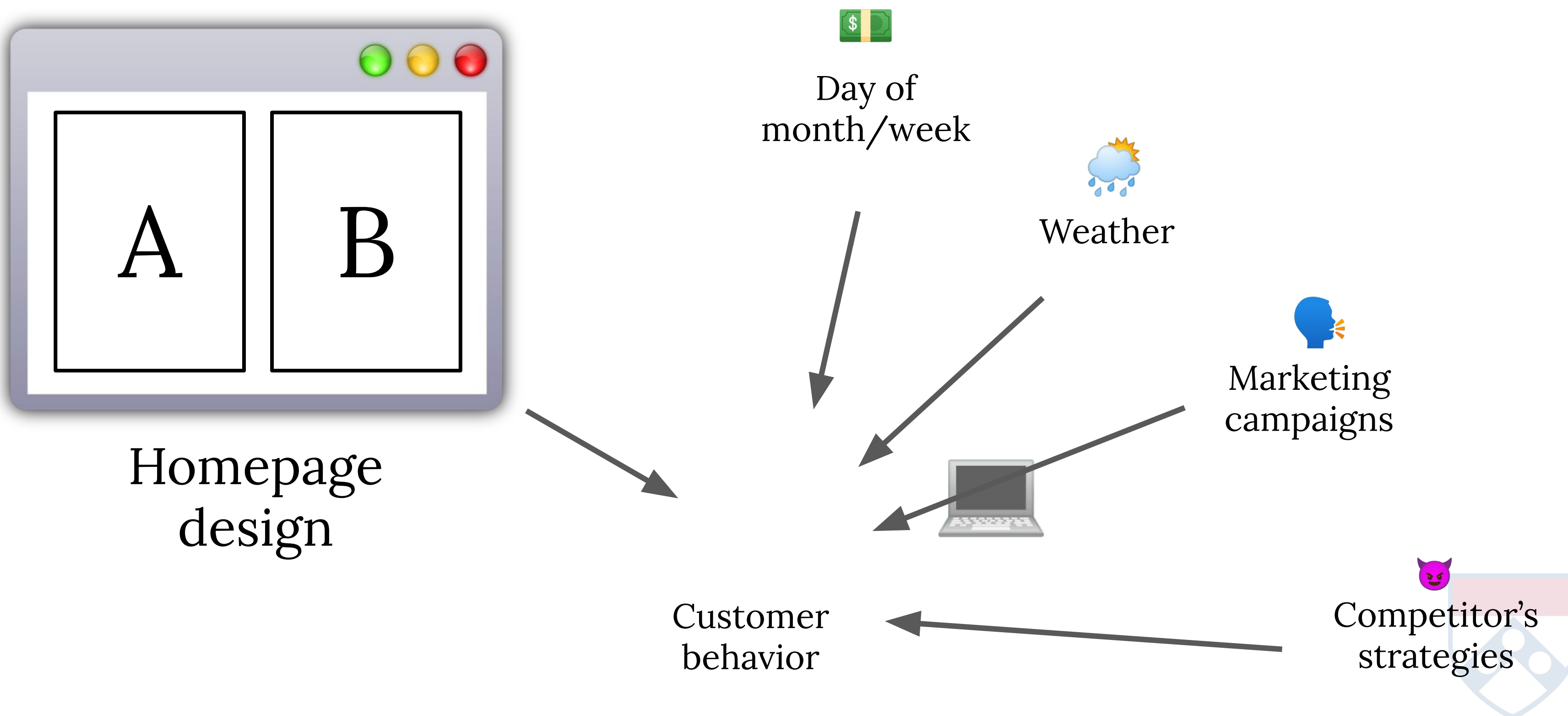


How does randomization help?



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

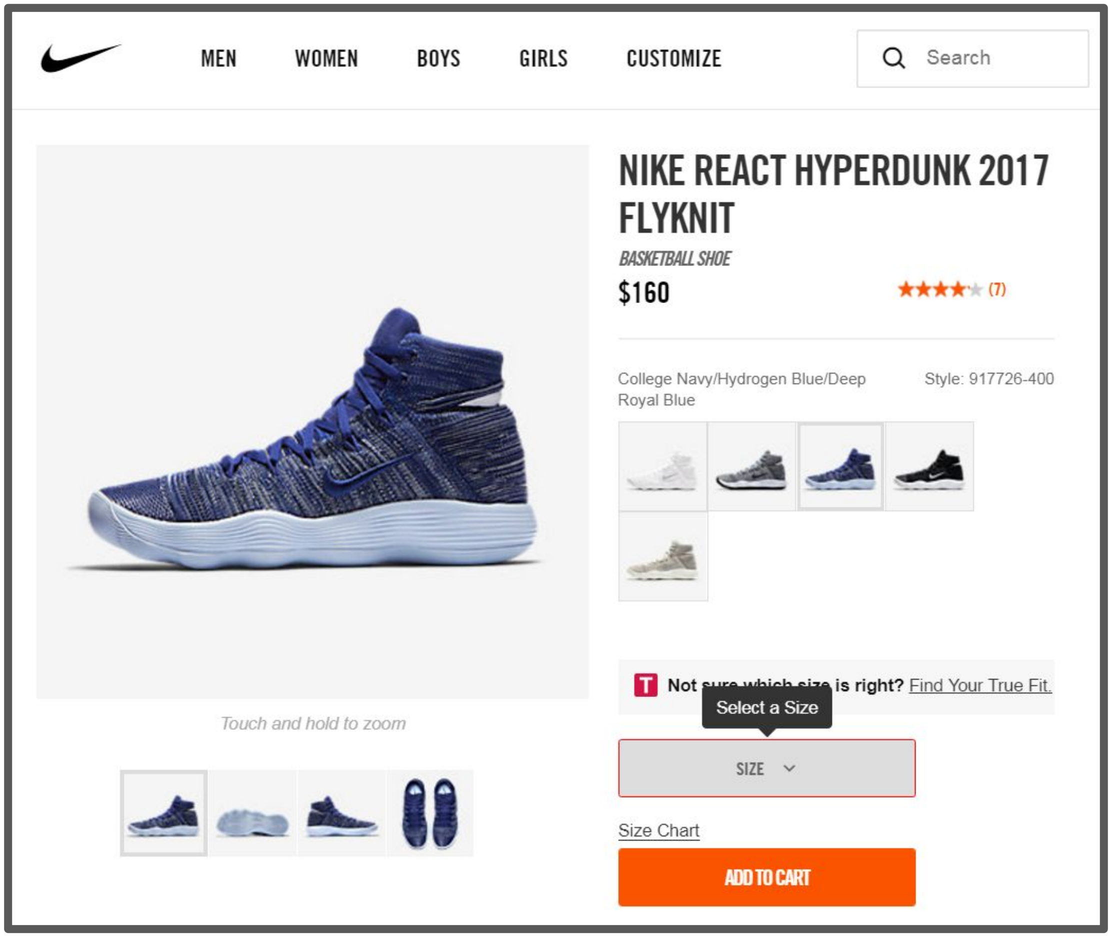
3. Define your intended sample size & stopping criteria (will revisit)

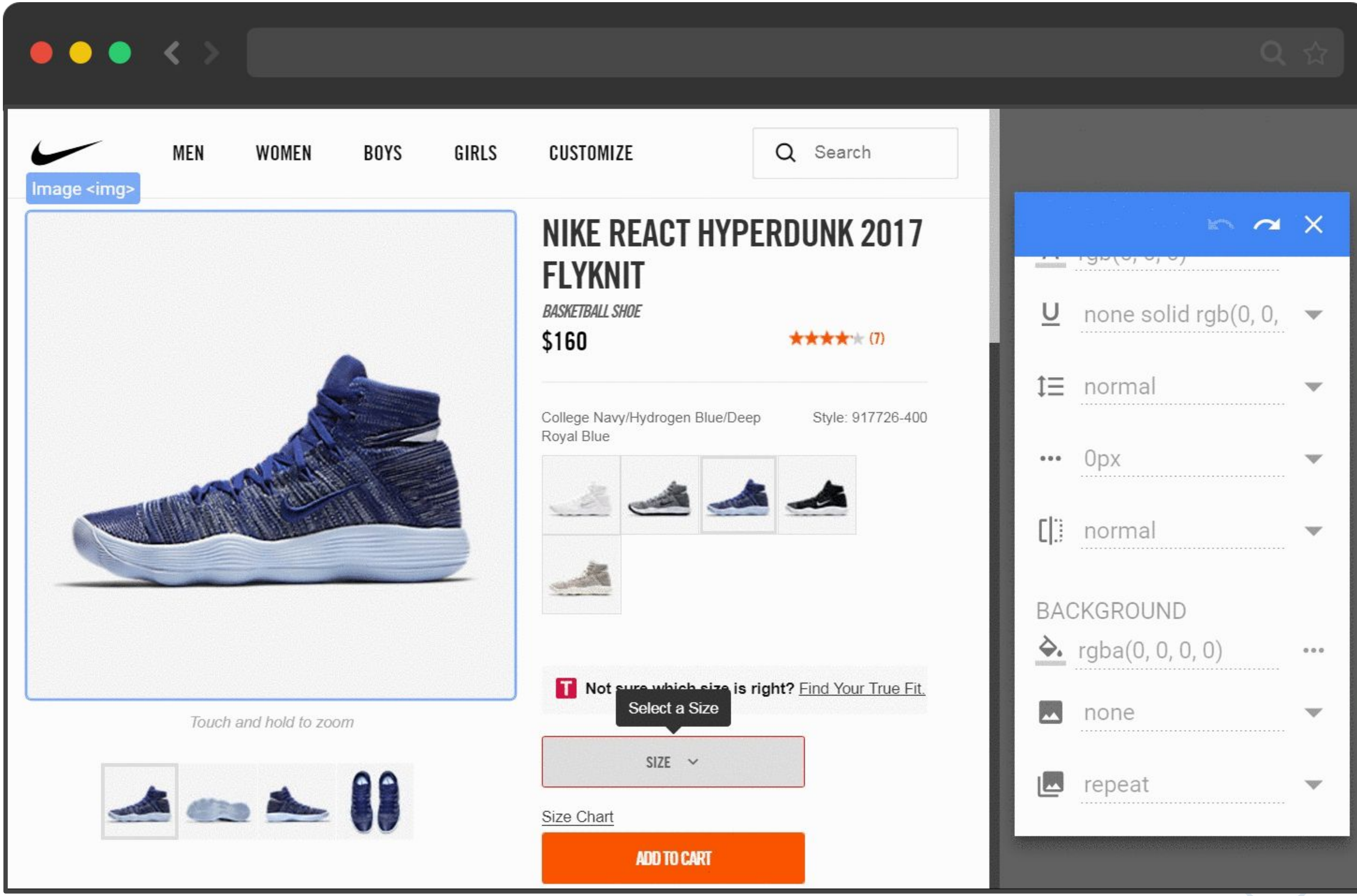
4. Run your experiment: Randomly assign customers to treatment arms

✅ 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:

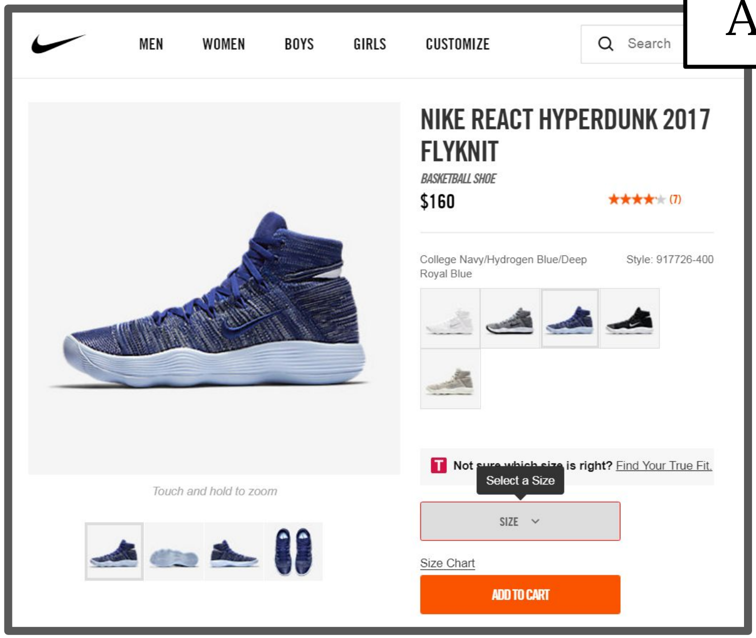




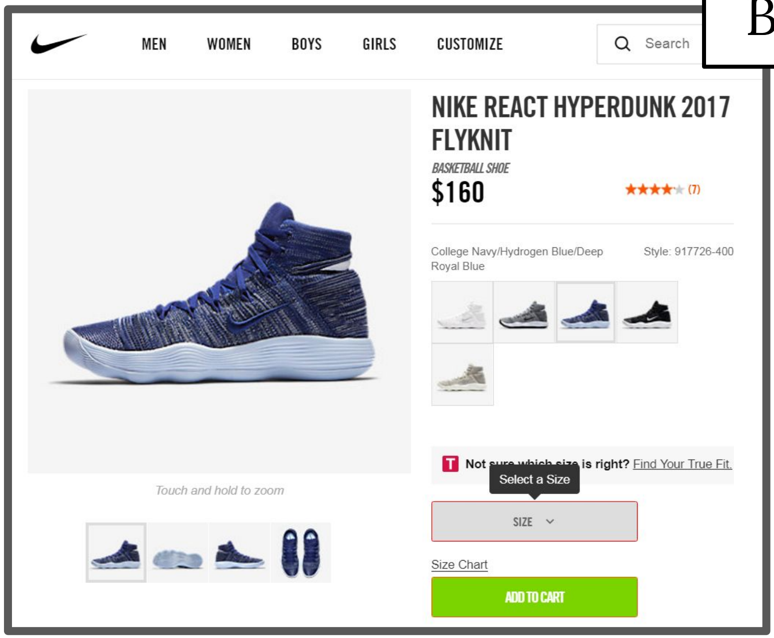
MEN
WOMEN
BOYS
GIRLS
CUSTOMIZE
Search
Image <img>
NIKE REACT HYPERDUNK 2017 FLYKNIT
BASKETBALL SHOE
$160
(7)
College Navy/Hydrogen Blue/Deep Royal Blue
Style: 917726-400
Touch and hold to zoom
Not sure which size is right? Find Your True Fit.
Select a Size
SIZE
Size Chart
ADD TO CART
none solid rgb(0,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

Hypotheses? ✅

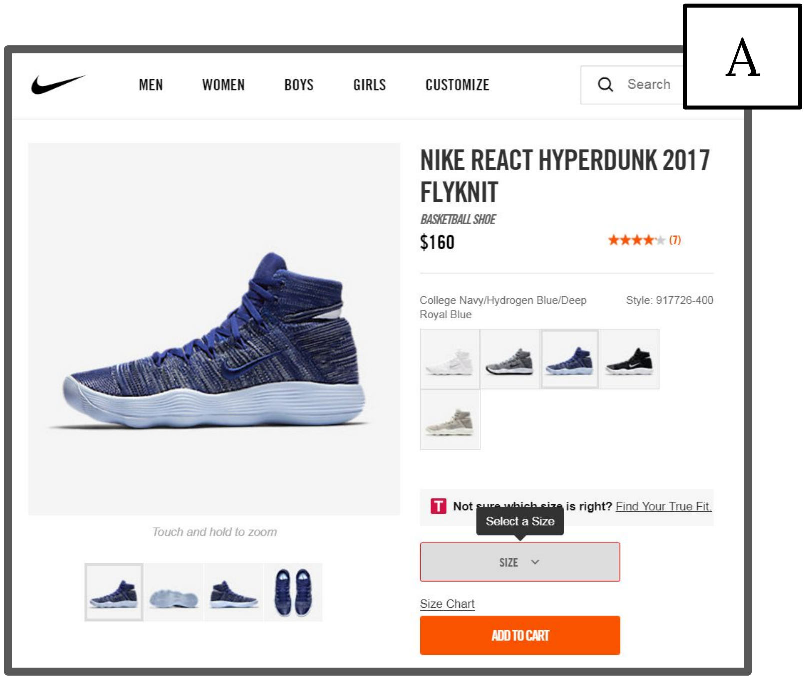
A

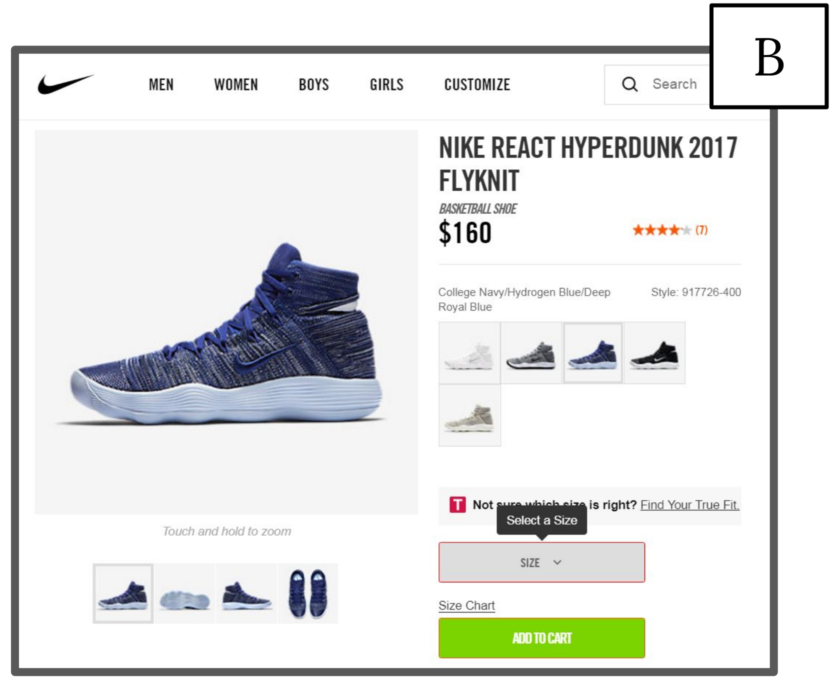


B



Hypotheses? ✅

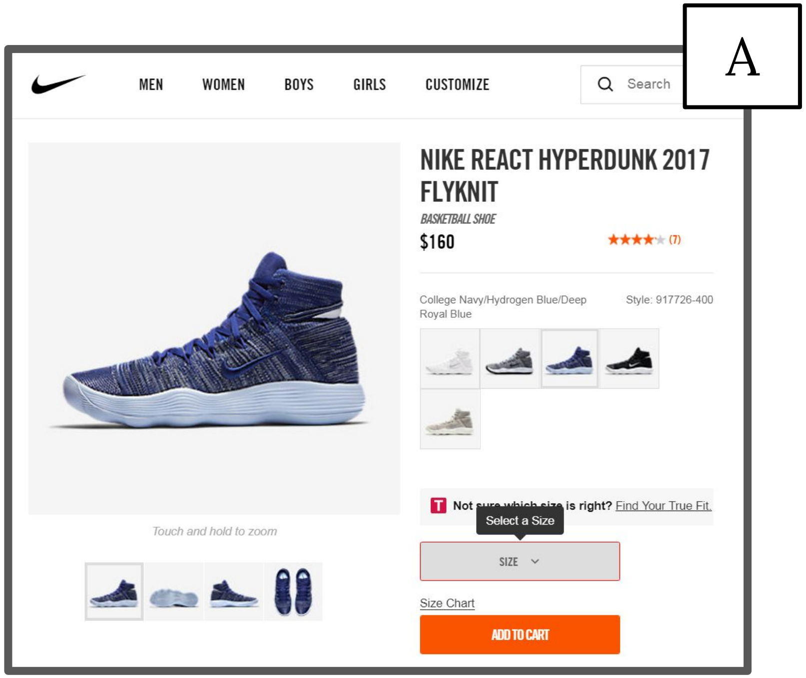


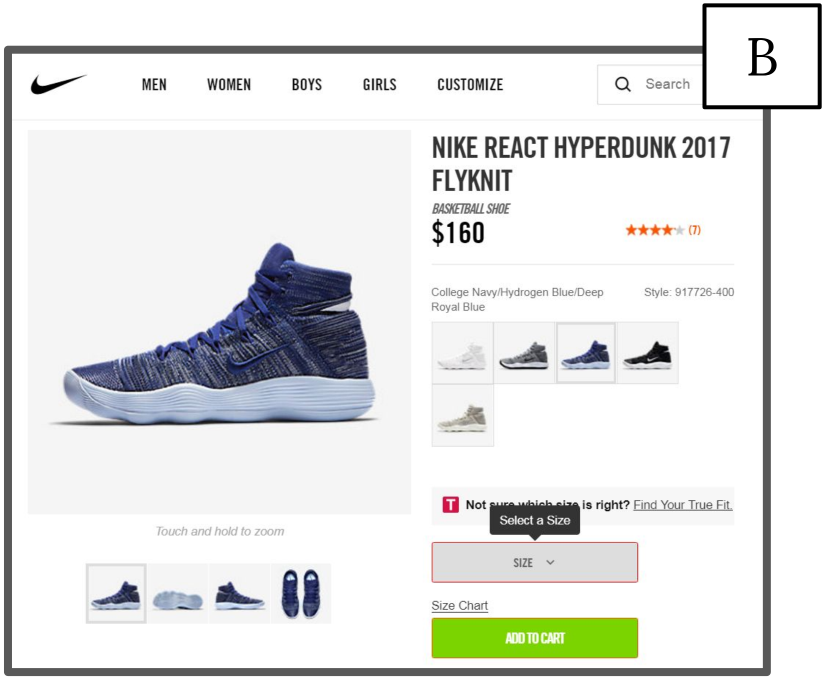


Evaluation criterion?

How long to run?

Hypotheses? ✅

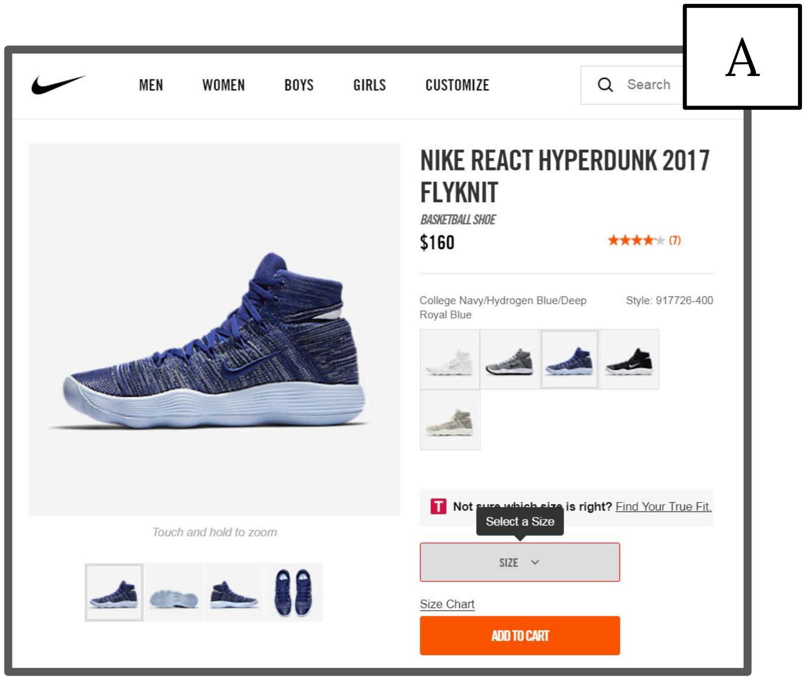


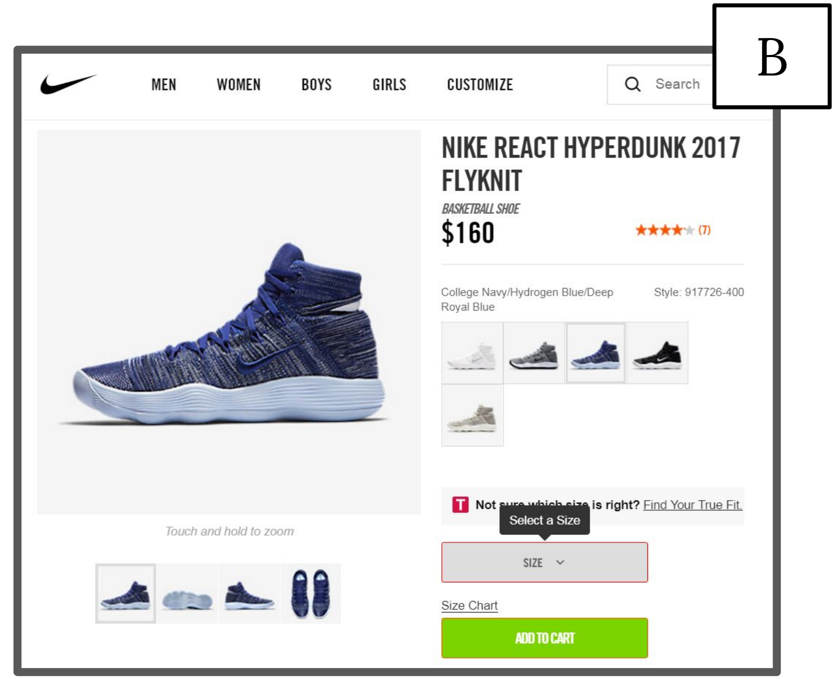


Evaluation criterion? Conversion rate ✅

How long to run?

Hypotheses? ✅



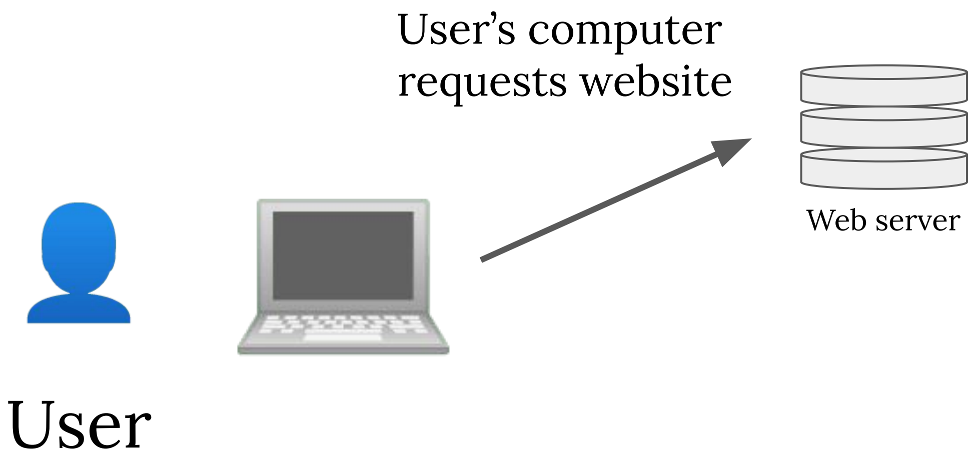


Evaluation criterion? Conversion rate ✅

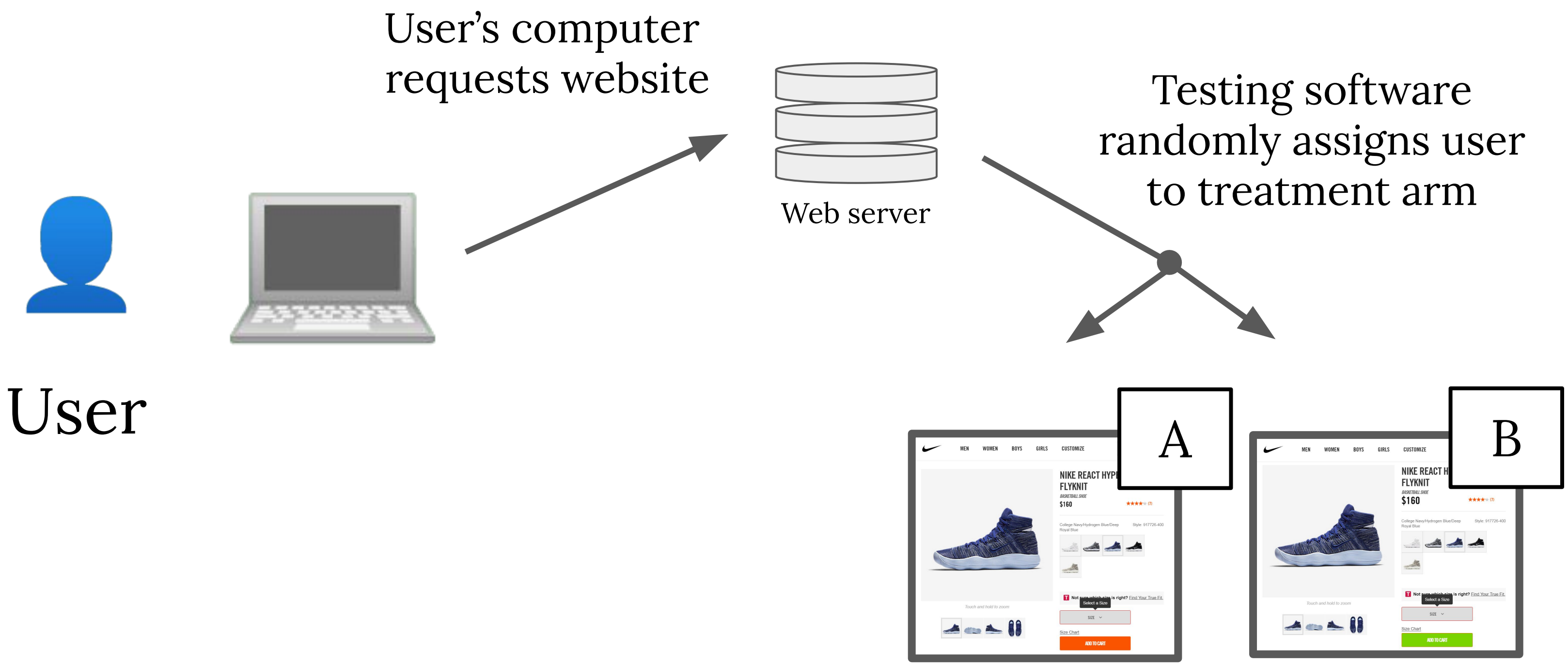
How long to run? 1 week ✅

Run experiment: A/B Test in Action

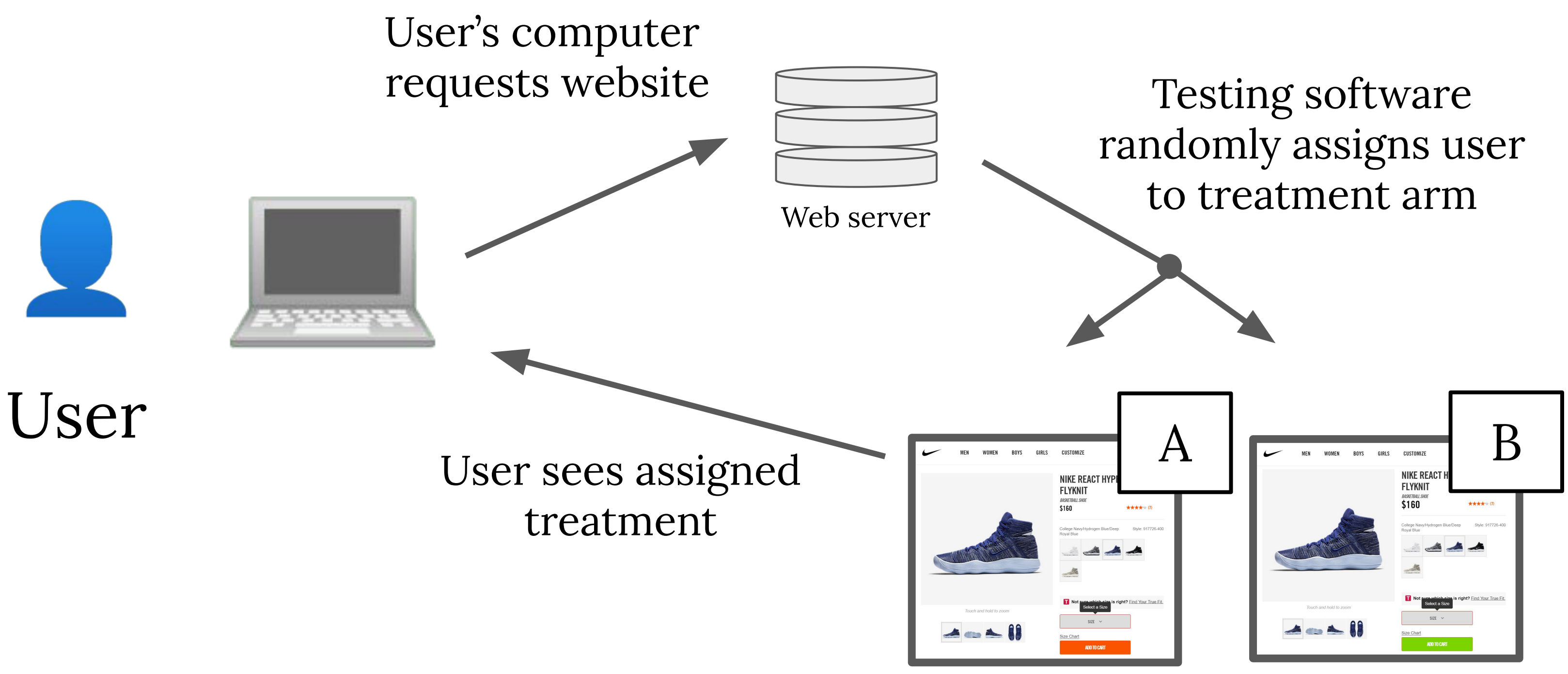
Run experiment: A/B Test in Action



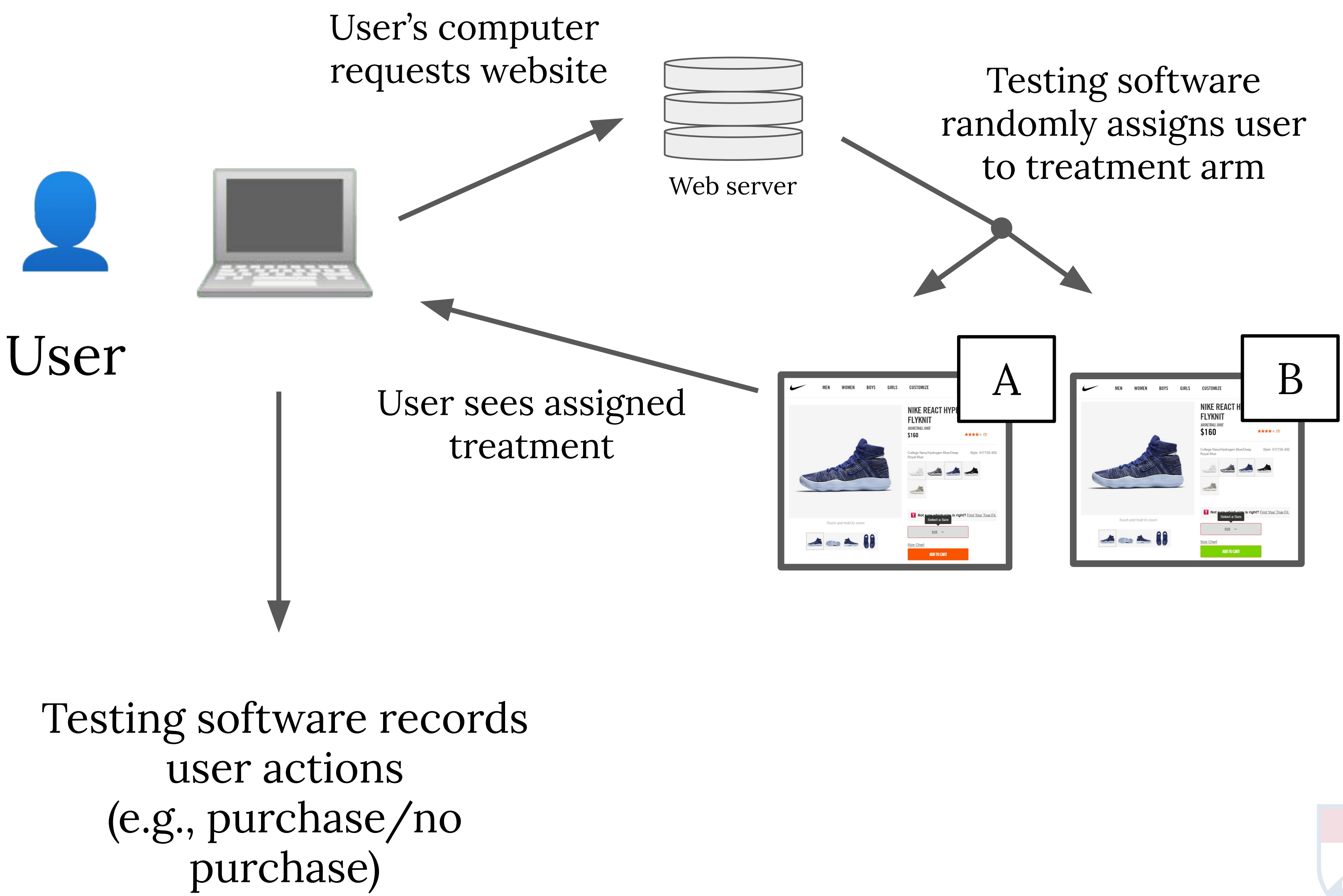
Run experiment: A/B Test in Action



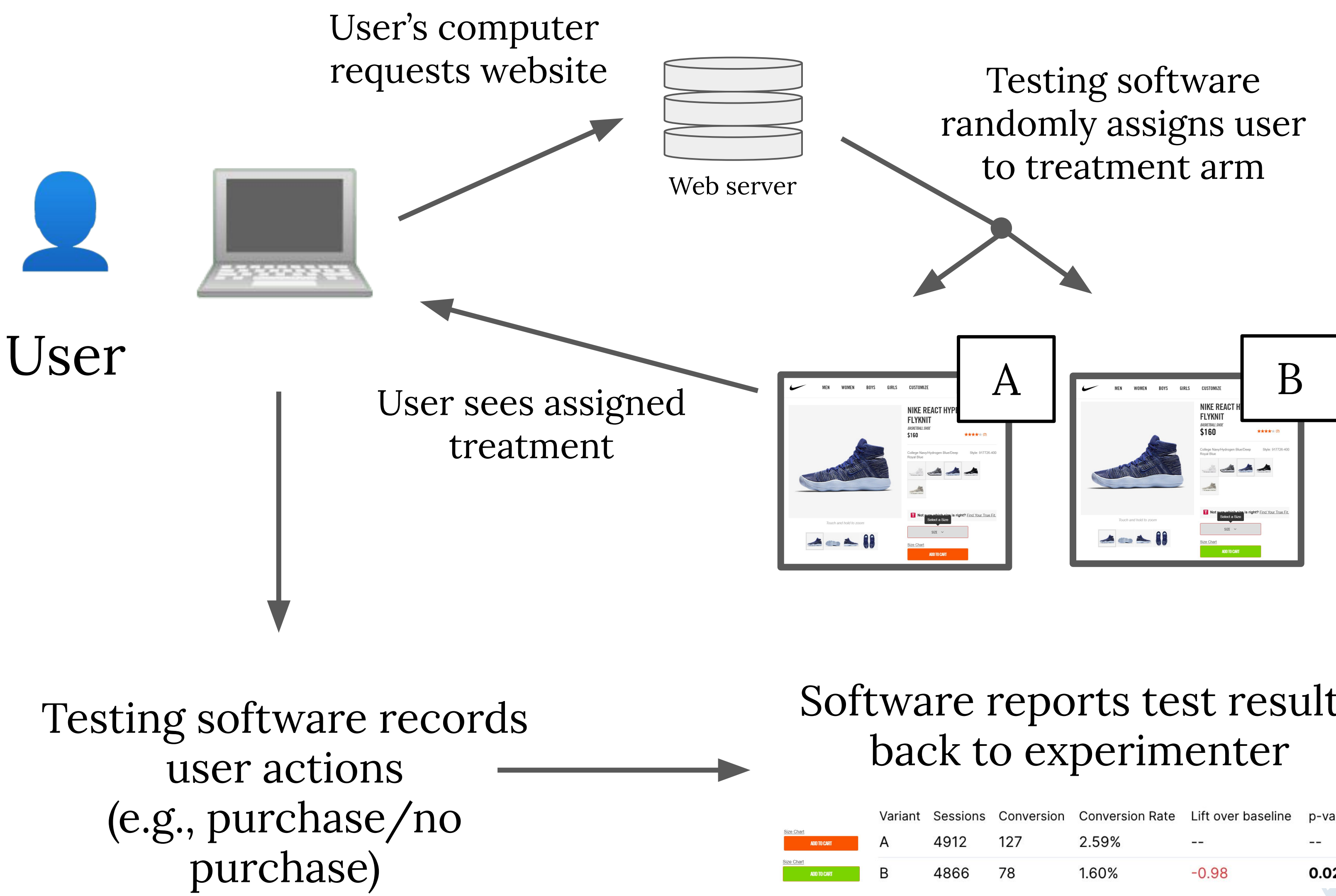
Run experiment: A/B Test in Action



Run experiment: A/B Test in Action



Run experiment: A/B Test in Action



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

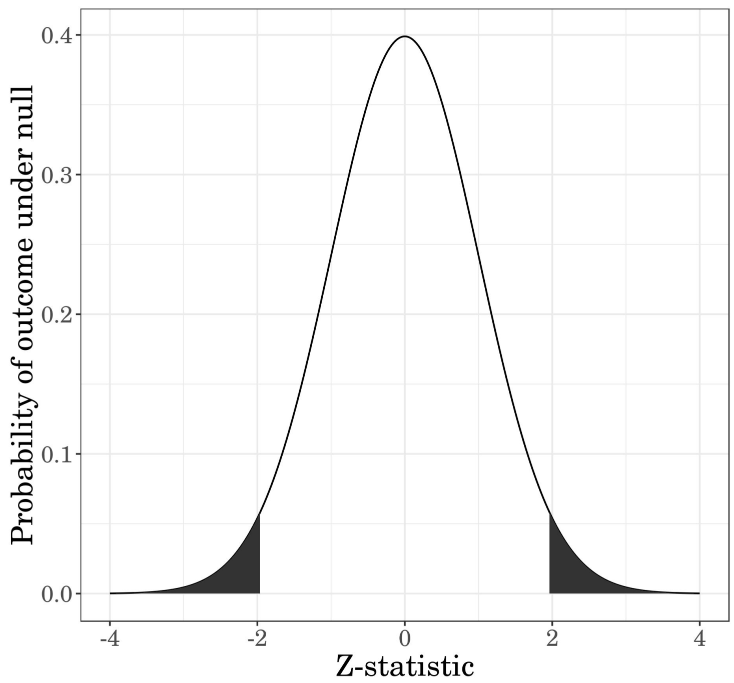
- This dashboard reports raw "p-values"
- It is common to report 1-p as "confidence" (e.g., p=0.02 implies "98% confidence")
- Practices are changing, but this is very common paradigm in statistical software

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

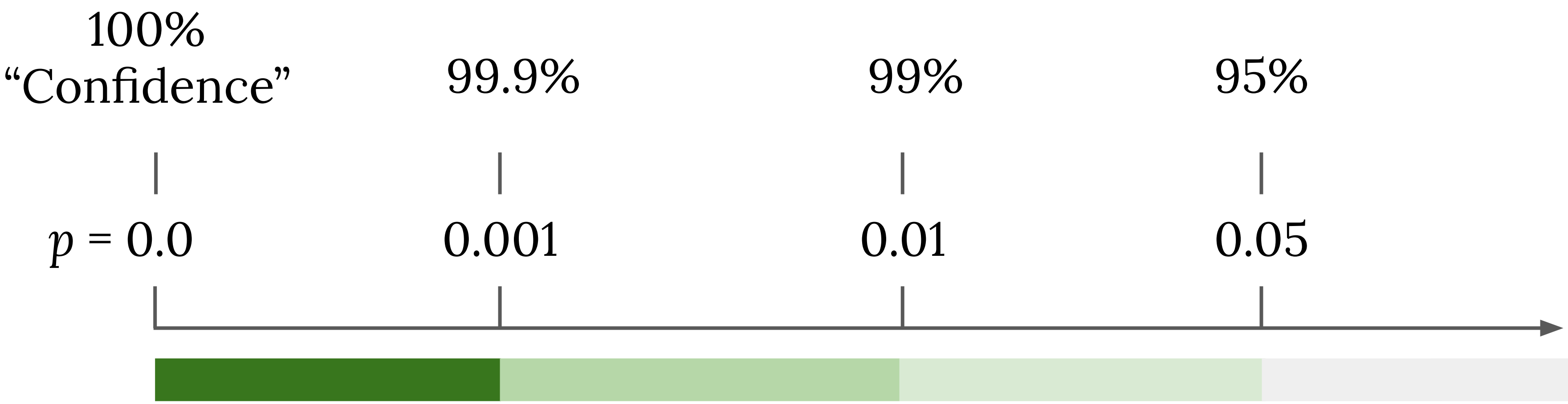
The question you want to answer:

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

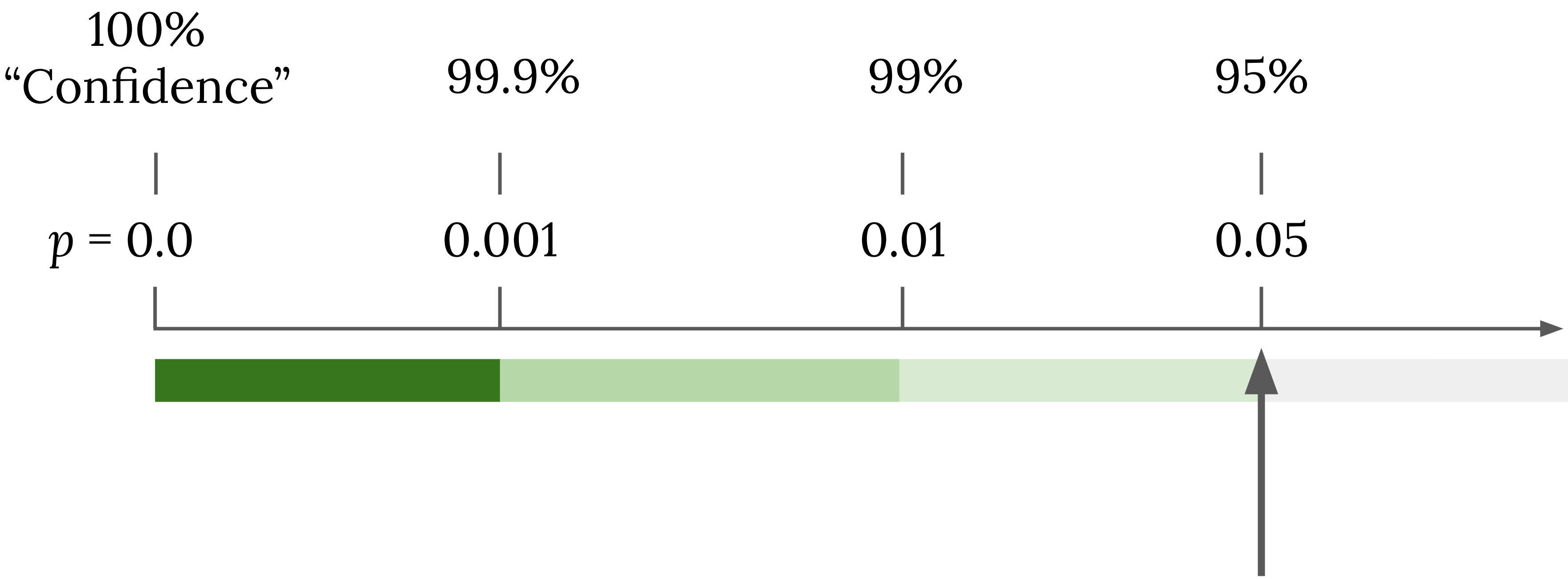
The question most A/B testing tools answer (those based on p-values or “Frequentist” statistics):

- 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)