GENERAL ASSEMBLY

INTRODUCTION TO A/B TESTING

Matt Brems, Data Science Immersive (massive h/t Joseph Nelson)

AGENDA

- What is A/B Testing?
- A/B Testing Design
- Case Studies

WHAT IS A/B TESTING?

- Some of you have likely seen or completed A/B tests before. What are they?

WHAT IS A/B TESTING?

- Some of you have likely seen or completed A/B tests before. What are they?

- A/B Testing is a term for a randomized experiment with two variants, A and B. These tests consist of test design, data collection, and data analysis stages.

WHAT IS A/B TESTING?

- A common use of A/B testing is to audition proposed changes to a website. Once the variants are designed, data is collected by assigning users to 'test' and 'control' groups, which will dictate the version of the site they will be served.

WHAT IS A/B TESTING?

- When designing an A/B test, it’s very important to make only one change between the two variants.

- Why?

WHAT IS A/B TESTING?

- When designing an A/B test, it’s very important to make only one change between the two variants.

- Why?

- Changing the number of images on a page
- Changing the font on a page
- Adding or removing single elements from a page
- Altering the text on a button
- Re-organizing a pages content

WHAT IS A/B TESTING?

- When analyzing the effect of an A/B test, we want to know what metrics we seek to compare.

- Consider an e-commerce site. The main effect in e-commerce is the flow of the user through the conversion funnel! Once users land on the site, test to see if the variant has any effect on how many products they **view**, how many products are **added to cart**, **changes in cart abandonment rates**, changes in **conversion rates**, **order volume**, **average order value**, etc.

A/B TEST DESIGN

- 1. What element(s) will be changed?
- 2. Who will be a part of the test group?
- 3. How long will the test run?
- 4. Why is this test truly necessary?

A/B TEST DESIGN

- 1. What element(s) will be changed?
- While working with a (or as a) project manager, you will likely have little say in what elements are changed for a test. Keep in mind that to prevent false correlations in the data, the **smallest changes possible** will likely have the most meaningful results.
- 2. Who will be a part of the test group?
- 3. How long will the test run?
- 4. Why is this test truly necessary?

A/B TEST DESIGN

- 1. What element(s) will be changed?
- 2. Who will be a part of the test group?
- Will you be splitting the incoming traffic 50/50 between variants, or can you get away with serving the variant under test to a smaller group? Also, will the test split change? (We'll discuss one strategy for assigning test groups in a minute)
- 3. How long will the test run?
- 4. Why is this test truly necessary?

A/B TEST DESIGN

- 1. What element(s) will be changed?
- 2. Who will be a part of the test group?
- 3. How long will the test run?
- This is a very important question to ask. If the test doesn't run **long enough**, your data won't be useful. If it runs **too long**, that can impact business needs. Ensure that you have enough data to capture across multiple periods, or seasons, but not too much data that your result will be heavily affected by trend.
- 4. Why is this test truly necessary?

A/B TEST DESIGN

- 1. What element(s) will be changed?
- 2. Who will be a part of the test group?
- 3. How long will the test run?
- 4. Why is this test truly necessary?
- A/B testing is a gamble. If the business result of the test is less valuable than the possible negative effects on churn or conversion rate, then it might be worth re-evaluating your design.

MULTI-ARM BANDIT TESTING

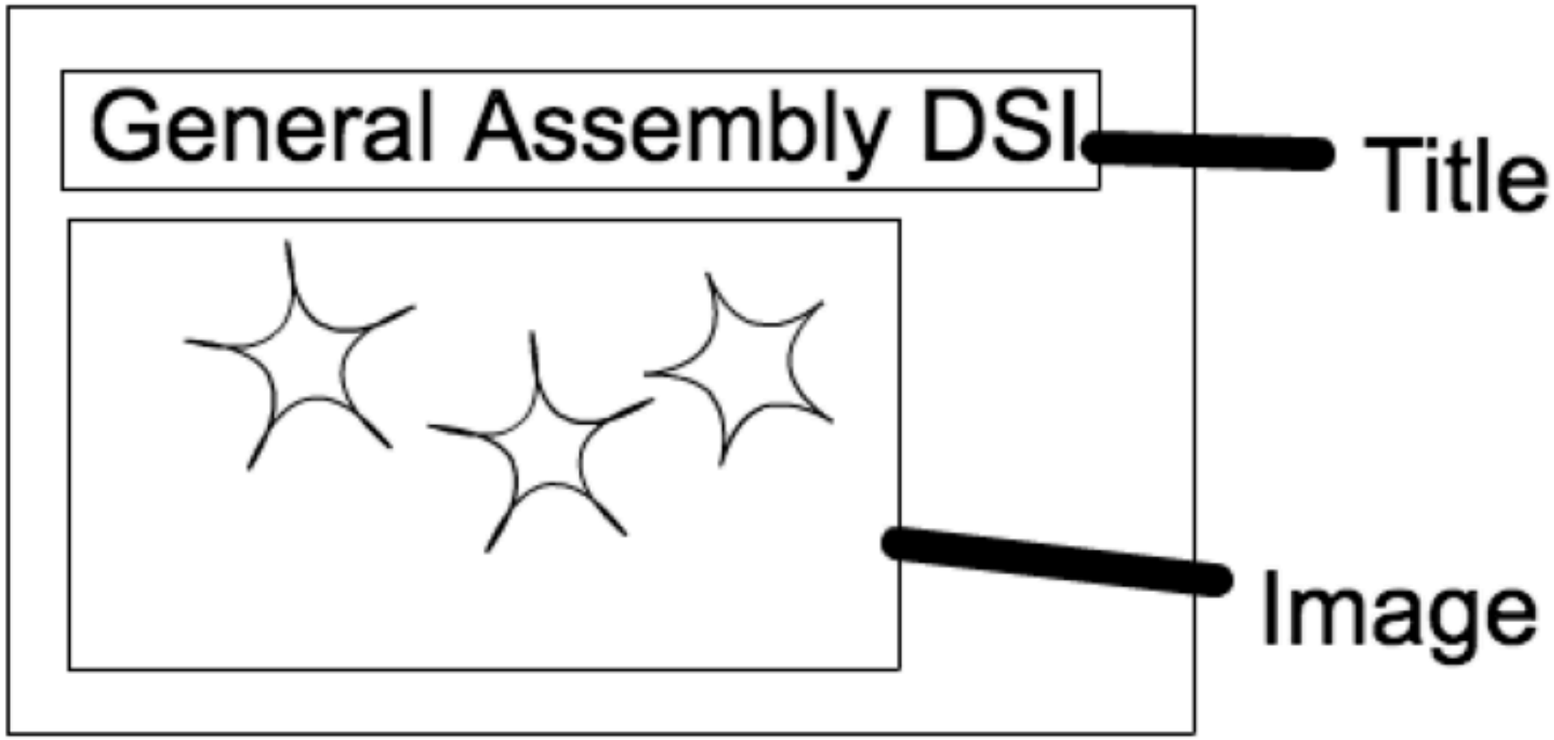
- Multi-arm bandit testing is an innovative way to split traffic (rather than simply 50/50) developed by Google (more on this in Resources).
- There are two phases:
- **Exploration Phase:** During the first ~10% of the test, traffic is split 50/50. This phase picks a short-term 'winner', and a short-term 'loser'.
- **Exploitation Phase**: For the remainder of the test, shift the majority of traffic to the higher performing variant. Continue to adjust traffic as performance increases/decreases.

- Pros/Cons?

MULTI-ARM BANDIT TESTING

- Pros/Cons?
- In practice, Multi-Arm Bandit testing does a fairly good job of optimizing conversion rates. The downside to this method, however, is increased difficulty in evaluation of results. Simply picking a 'winner' variant is not always the best strategy, especially since the 'loser' variant often gets so little traffic that it can be hard to validate the statistical significance of the lift.

FACTORIAL DESIGNS



Variant	Image	Title
Control	Stars	General Assembly DSI
Test 1	Flowers	General Assembly DSI
Test 2	Stars	Data Science Immersive
Test 3	Flowers	Data Science Immersive

- This is called a two-way factorial design.

0PTIMUS EXAMPLE

- In political consulting, suppose I wanted to test three different commercials to see if any of them had a significant effect.

- What might I do?

HOW DO WE COMPARE GROUPS?

- Identify statistically significant differences. (Contrasts, factorial designs, t-tests, χ^2 tests.)

HOW DO WE COMPARE GROUPS?

- Identify statistically significant differences. (Contrasts, factorial designs, t-tests, χ^2 tests.)

- Compare posterior distributions!

A NOTE ABOUT SAMPLING…

- What might we want to keep in mind about assigning individuals to treatment groups?

CASE STUDY ONE: OBAMA FUNDRAISING

- https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

CASE STUDY TWO: SOCIAL SHARING

- https://vwo.com/blog/amd-3600-social-sharing-increase/
- In the interest of time, you'll check this out on your own.