Statistical Learning and Data Science

Xinwei Deng

xdeng@vt.edu

Department of Statistics

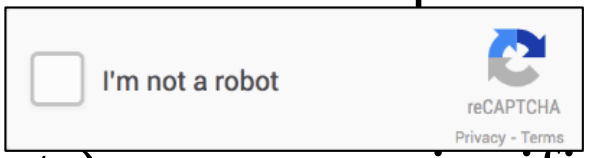
Introduction to Online Experimentation (A/B Testing)

- What is A/B testing?
- Why use A/B testing?
- Advantages/Disadvantages

Online Experimentation

- Key difference from the offline experiments (physical or computer experiments):

1. Experiment is conducted online, ramp-up automatically to a large scale. But experiments can be shut down automatically.
2. Massive data logged: tens of terabytes per day, thousands of attributes for every page shown.
    - Data quality challenges, big data challenges
    - New metrics created quickly (e.g., every week)
3. False positives issues and multiple testing.

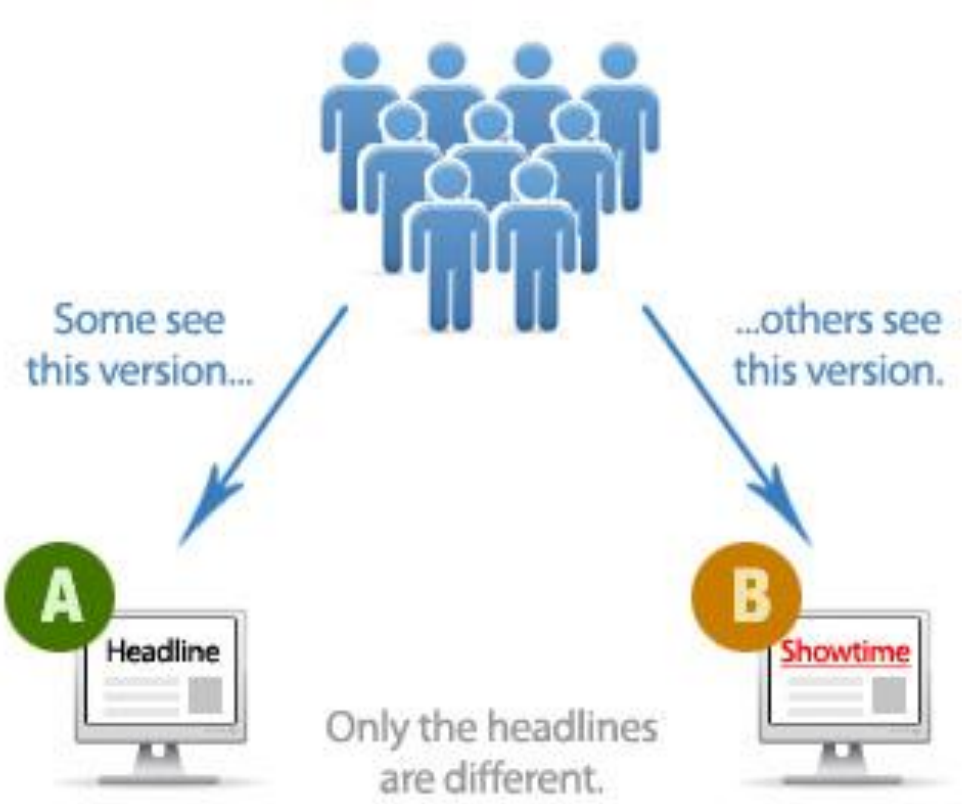


4. Artifacts (e.g., *Bots*) can cause significant skews.

What and Why:
A/B Testing

- **A/B Testing for Online Experimentation:** Test different ideas (or webpage designs) against each other in the real online world.
- Why A/B Testing?
  - No subjective guessing
  - Provides accurate and timely answers
  - Allows to rapidly iterate on ideas
  - Establish the causal relationships
- *Goal*: Find the one that statistically performs better.

What is A/B Testing



A/B testing, sometimes is also called as split testing, bucket testing, and multivariate testing.

What Is It?

- Controlled experiments
  - Web site
  - Authentic users
- Split users into two groups:
  - Control
  - Test variant
- Cookies can ensure an individual gets consistent version.
- A metric (the *overall evaluation criteria*). For example:
  - Click-through rate
  - Purchases
  - Conversation Rate

What is A/B Testing

- A/B testing is a way to test changes to your web page against the current design and determine which ones produce a more positive result.

- It has made many websites to become more successful.

- From “I visited the Amazon homepage” to “I visited an Amazon homepage.

What is A/B testing: Example 1

Version A **Version B**

Picture links

Empty Space

Different Graphics

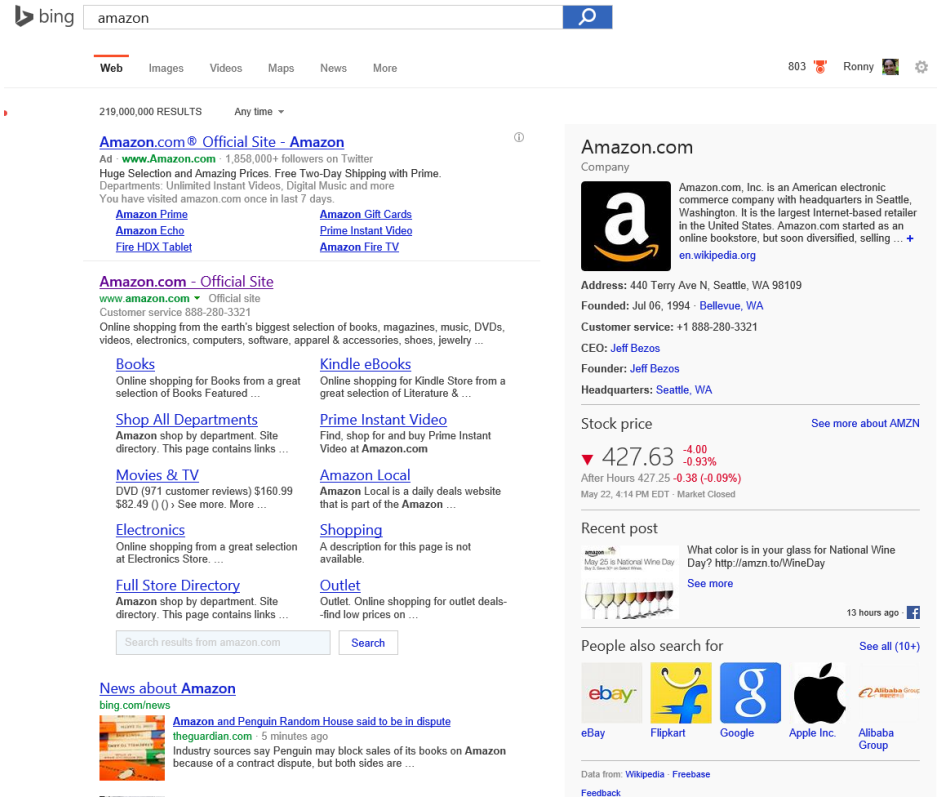
More info on page

24%

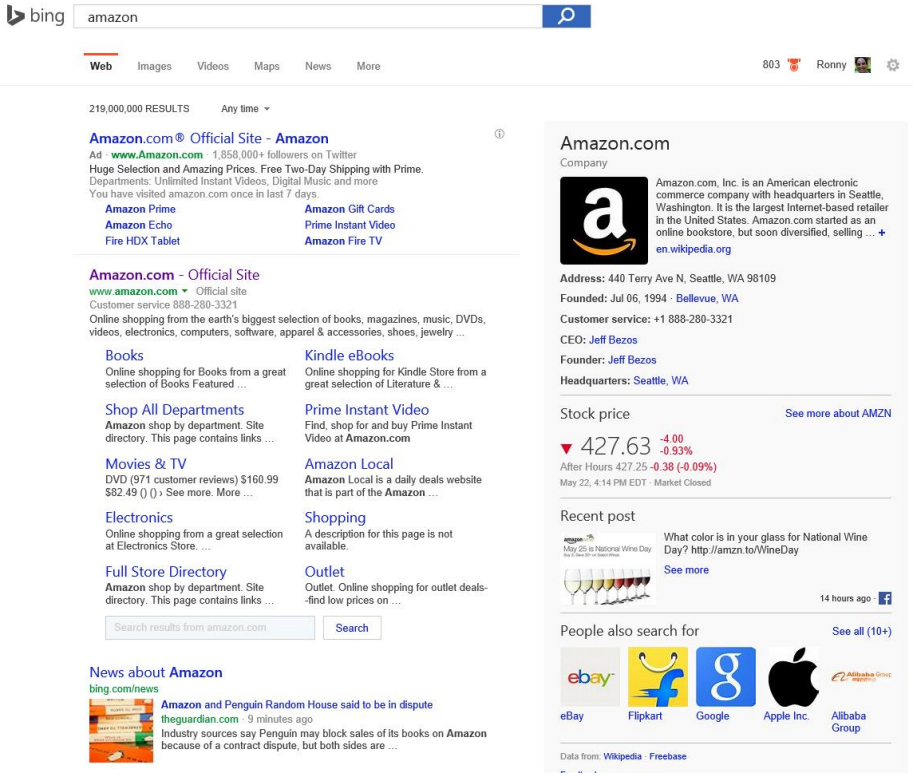
42%

Example 2: Underlining Links

- Does underlining increase or decrease click-through-rate?



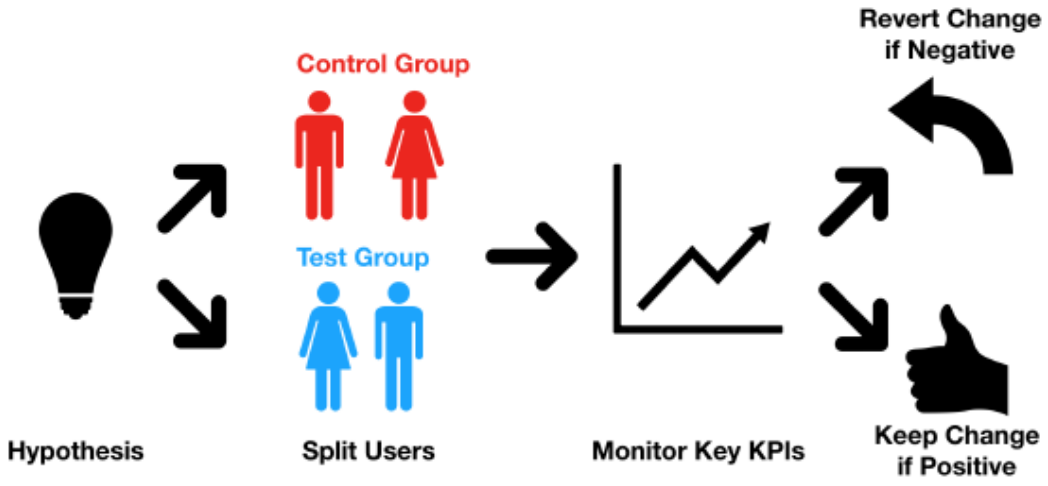
A



B

A General A/B Testing Process

- Step 1: Create Objective:
  - Example: How to increase click through rates for my homepage?
- Step 2: Develop Hypothesis:
  - Example: Underlined links will outperform my current links.



A General A/B Testing Process (Con't)

- Step 3: Estimate time to run for A/B test:
  - E.g.: A/B Test Duration Calculator.
- Step 4: Run the test:
  - **Randomly** assign users to two different groups
  - Group 1 to the current product webpage (version A)
  - Group 2 to a product webpage that tests the hypothesis (Version B).
- Review results, conduct analysis, and pick the one with better performance according to a set of *key performance indicators (*KPI).

The Importance of Randomness

- Randomization: randomly assign treatment to experiment unit.
  - isolate the impact of the change made (version A vs version B)
  - reduce the potential impact of confounding variables
- Randomization makes the analysis more towards to causal inference.
- Using an deterministic assignment criteria may introduce confounders.

Why Use A/B testing?

- Easy to use
  - The optimization platforms often are available. For example, https://www.optimizely.com/ab-testing/.
- Not as risky as changing the webpage features without testing it.
- Gives flexibility to test out new ideas since it can easily return to the old web page if new is unsuccessful.

Advantages of A/B Testing

- Using A/B testing tools, it is easy to resize, retype, delete, and make changes to the website.

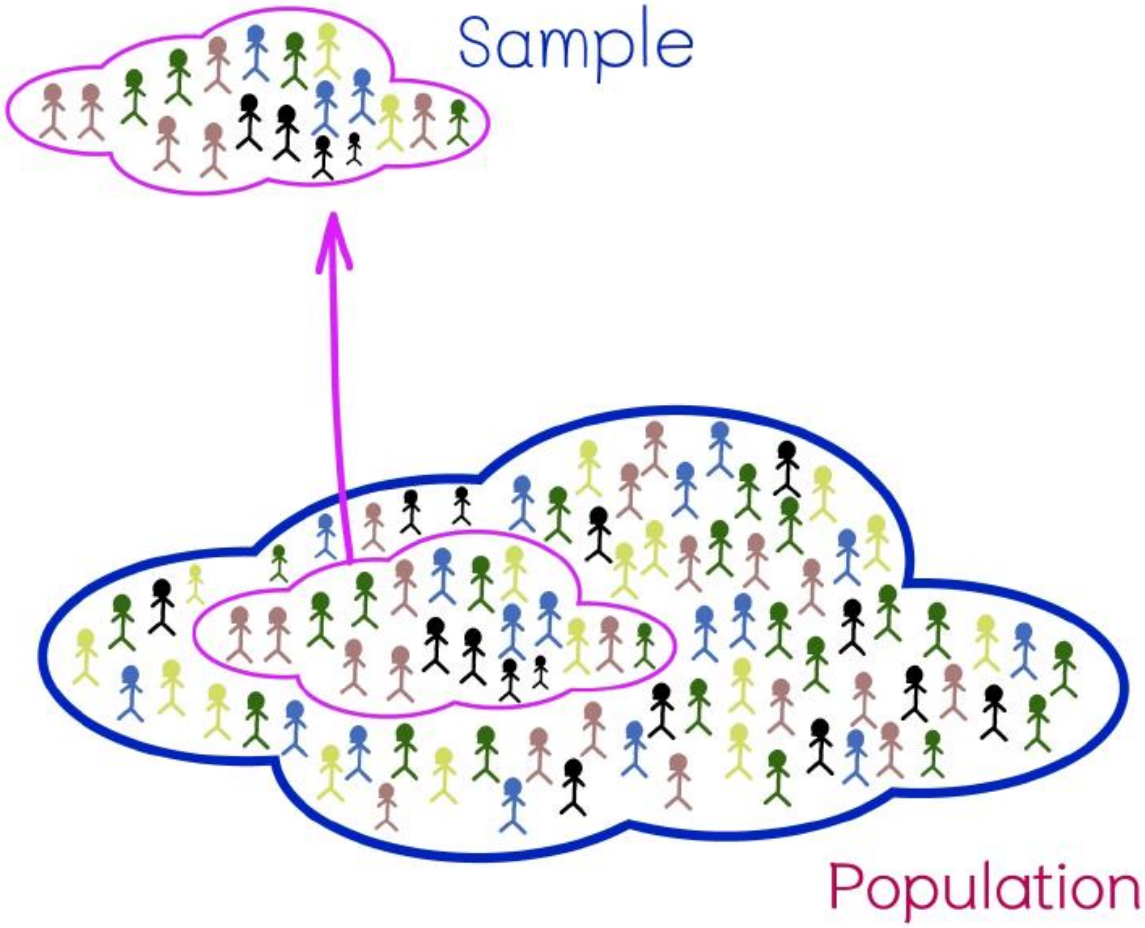
- One can easily create two similar webpages with minor differences and cause users that visit the site to randomly get redirected to either the original page (A) or the modified page (B).

- The tools track user behavior and delivers the results so that one can see which webpage is more successful.

Challenges of A/B Testing

- Network effects among users.
- Challenging to segment the users into groups.
- Difficult to untangle the impact of multiple A/B testings.
- Can has multiple response varaibles with different types.

Sample Size Calculation in A/B Testing



Null Hypothesis

- Hypothesis that control (version A) and treatment (version B) have the same impact on the response.
  - E.g.: Updated the product webpage does not improve conversion rate
  - Any observed difference is due to randomness
- Rejecting the Null Hypothesis
  - Determine their is a difference between the treatment and control.
  - Statistically significant result

Types of Error and Confidence Level

- **Confidence Level:** probability of *not* making Type I Error.
- The higher the value of confidence level is, the larger test sample is needed
- Common values: 0.90 and 0.95.

Null Hypothesis		Null Hypothesis	
		TRUE	FALSE
	Not Reject	Correct	Type II Error
	Reject	Type I Error	Correct

Statistical Power

- **Statistical Power:** Probability of finding a statistically significant result when the Null Hypothesis is false.
- Sample size increases ←→ Power increases
- Confidence level increases ←→ Power decreases

Summary

- A/B testing can be used to study user's actual behaviors.
- Treat web as experimental platform
- Ease of testing
- A/B testing can be considered as an example of
  - Large scale empirical study of Think-Aloud.

Remarks and Comments

VirginiaTech

Thank you!