



A/B TESTING INCORRECTLY

November 12th

Overview

- Introduction
 - *Bio*
 - *Motivation & Goals*
- What's an AB Test
- Failing to define a user: Identification
- Failing to get the right result: Peeking
- Failing to lift off: Selection Bias
- Failing to Design: Interference
- Conclusion

About me Academically

- PhD at UCLA in Management (2012)
- Masters in Economics at UC Davis (2007)
- BA in Applied Math/Statistics at UC Berkeley (2002)

About me (Professionally)

- Director of Backend Engineering & Data Science at The Meta (Kovaak)
- Assistant Professor of Data Science at USF
 - 2014-2020
- Director of Analytics at Sega
 - 2014-2015
- Director of Analytics and User Acquisition at TinyCo
 - 2011-2014
- Senior Consultant Bates White
 - 2002-2006

Some Games I've worked on

KOVAAK 2.0
M+ THE META



Currently at The Meta

- Director of Backend Engineering & Data Science at The Meta (Kovaak)
- eSports Digital Training Platform
 - *“Gym for eSports”*
 - *Provide tools to help people improve at (mainly) FPS titles*

Example Video



Talk Motivation & Goals

- Going to cover a few issues that often aren't covered when learning the statistics (at least when I was taught this)
- Hopefully you'll come away with a better understanding of some of the real world difficulties of A/B Testing.



WHAT'S AN A/B TEST



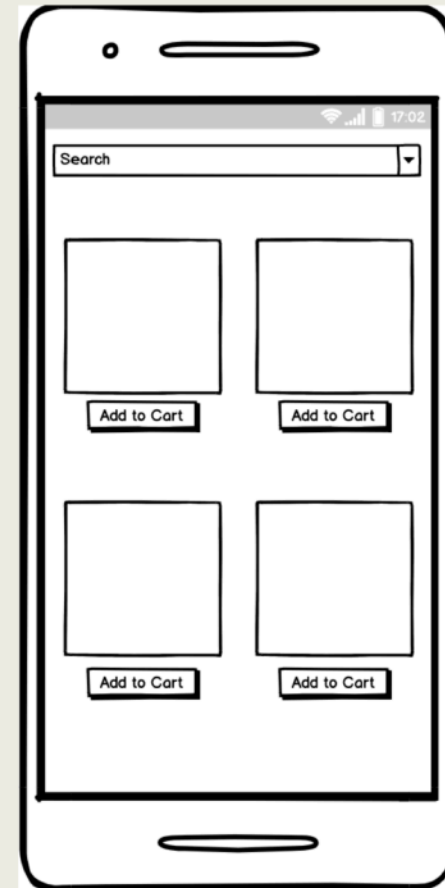
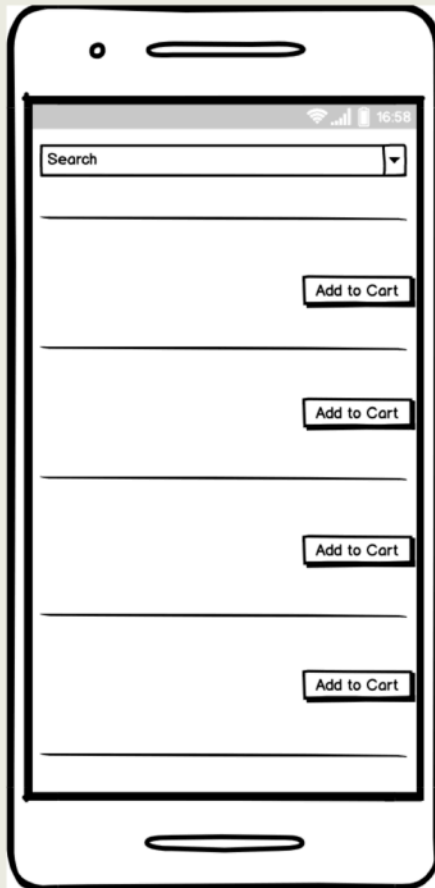
A/B Testing: What?

- An A/B test is a **controlled experiment**
- Interest lies in determining how some metric of interest (i.e., a KPI) is **causally** related to one or more factors
- Different levels of these factors define two or more **experimental conditions** (aka: variants, buckets, cells, treatments)

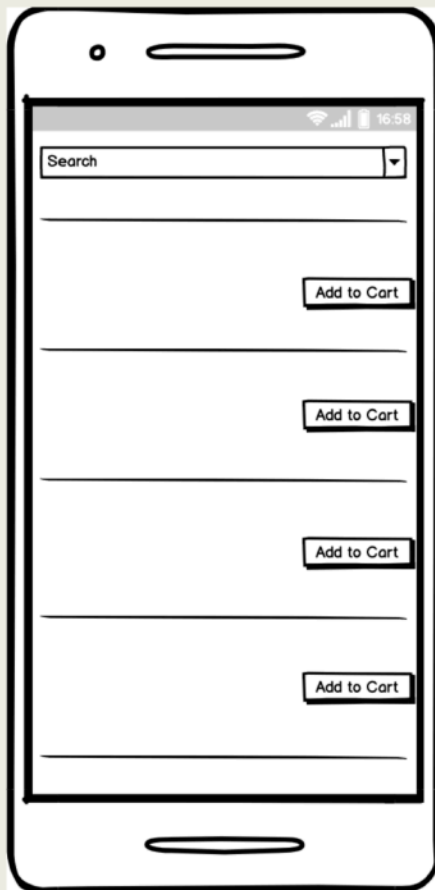
A/B Testing: What?

- Experimental **units** are **randomly** assigned to these conditions
- **Randomization principle:**
 - Random assignment ensures that users in different conditions will be homogenous and the only collective difference among them is the fact that they're in different conditions.
 - So any difference observed among the conditions should be due only to that which the experimenter is controlling

A/B Testing: What?

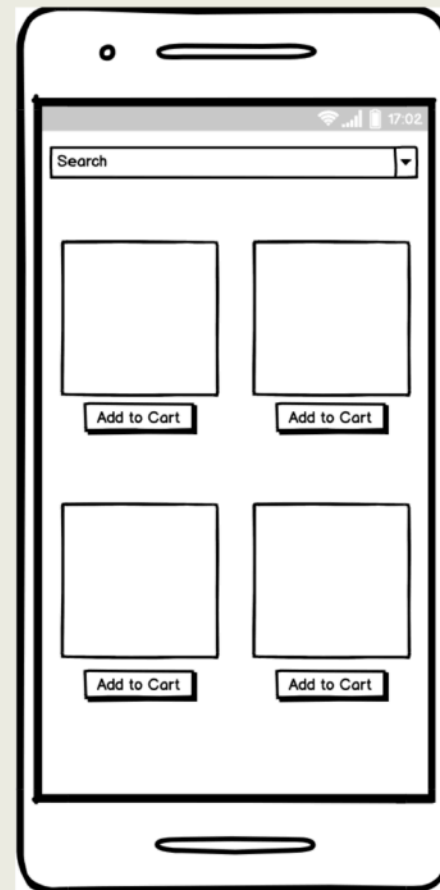


A/B Testing: What?



5%

A



2%

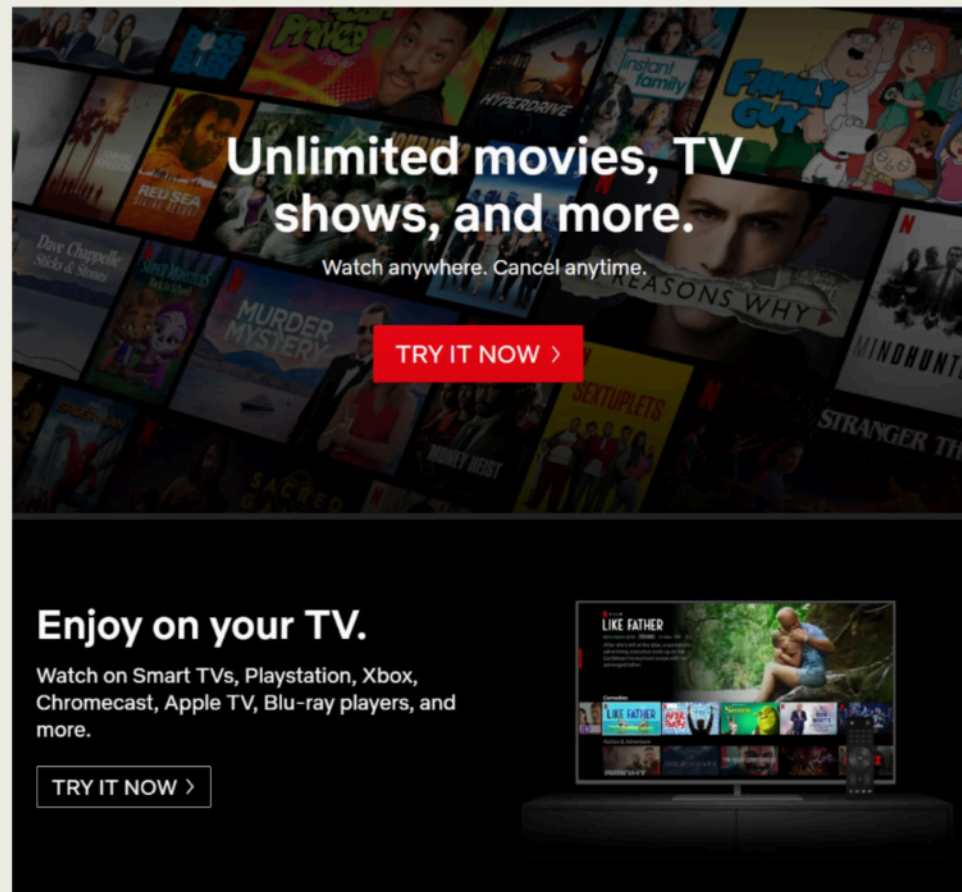
B



A/B Testing: What?

- Pick the winner:
 - The metric of interest is compared across the conditions, and the condition that optimizes the metric is declared the winner

A/B Testing: What?



Unlimited movies, TV shows, and more.

Watch anywhere. Cancel anytime.

[TRY IT NOW >](#)

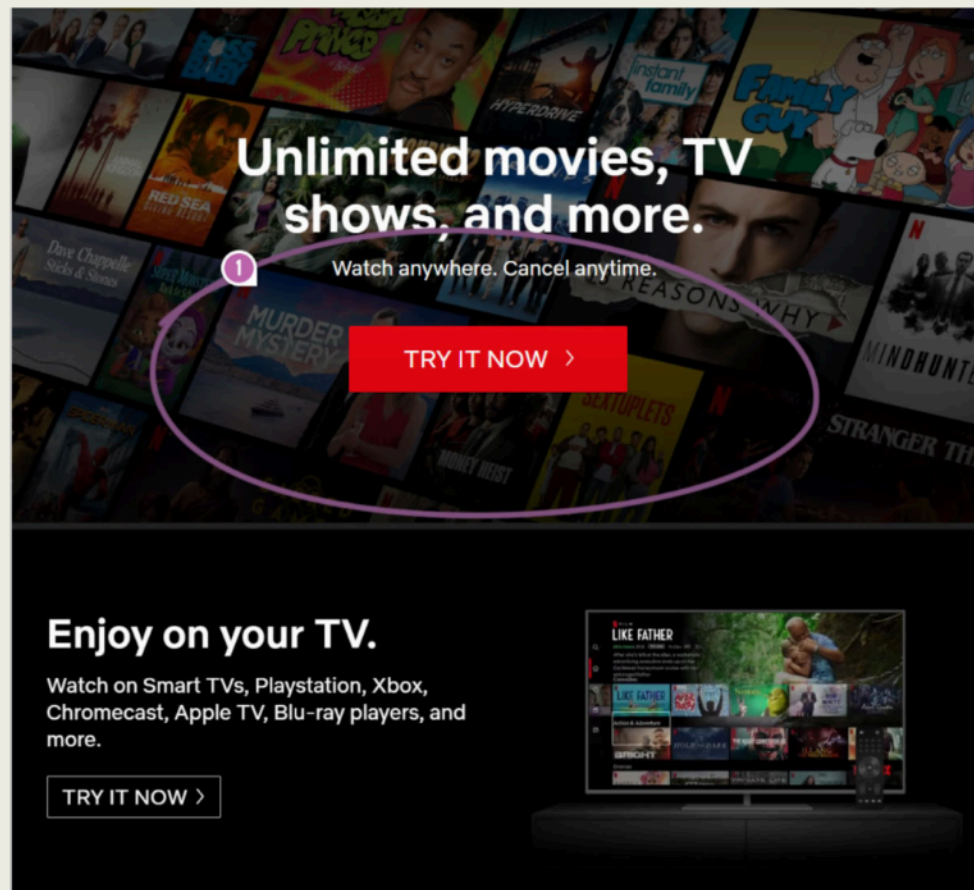
Enjoy on your TV.

Watch on Smart TVs, Playstation, Xbox, Chromecast, Apple TV, Blu-ray players, and more.

[TRY IT NOW >](#)

The advertisement features a collage of various Netflix titles including 'Prince of Persia: The Sands of Time', 'Instant Family', 'Family Guy', 'Dave Chappelle: Sticks & Stones', 'Red Sea Diving Resort', 'Murder Mystery', 'Sextuplets', 'Money Heist', 'Sacred Games', 'Like Father', 'Stranger Things', 'Mindhunter', and 'Reasons Why'. Below the collage, a television set is shown displaying the Netflix interface with the 'Like Father' title highlighted.

A/B Testing: What?



A/B Testing: What?

Unlimited movies, TV shows, and more.

Watch anywhere. Cancel anytime.

TRY IT NOW >

Enjoy on your TV.

Watch on Smart TVs, Playstation, Xbox, Chromecast, Apple TV, Blu-ray players, and more.

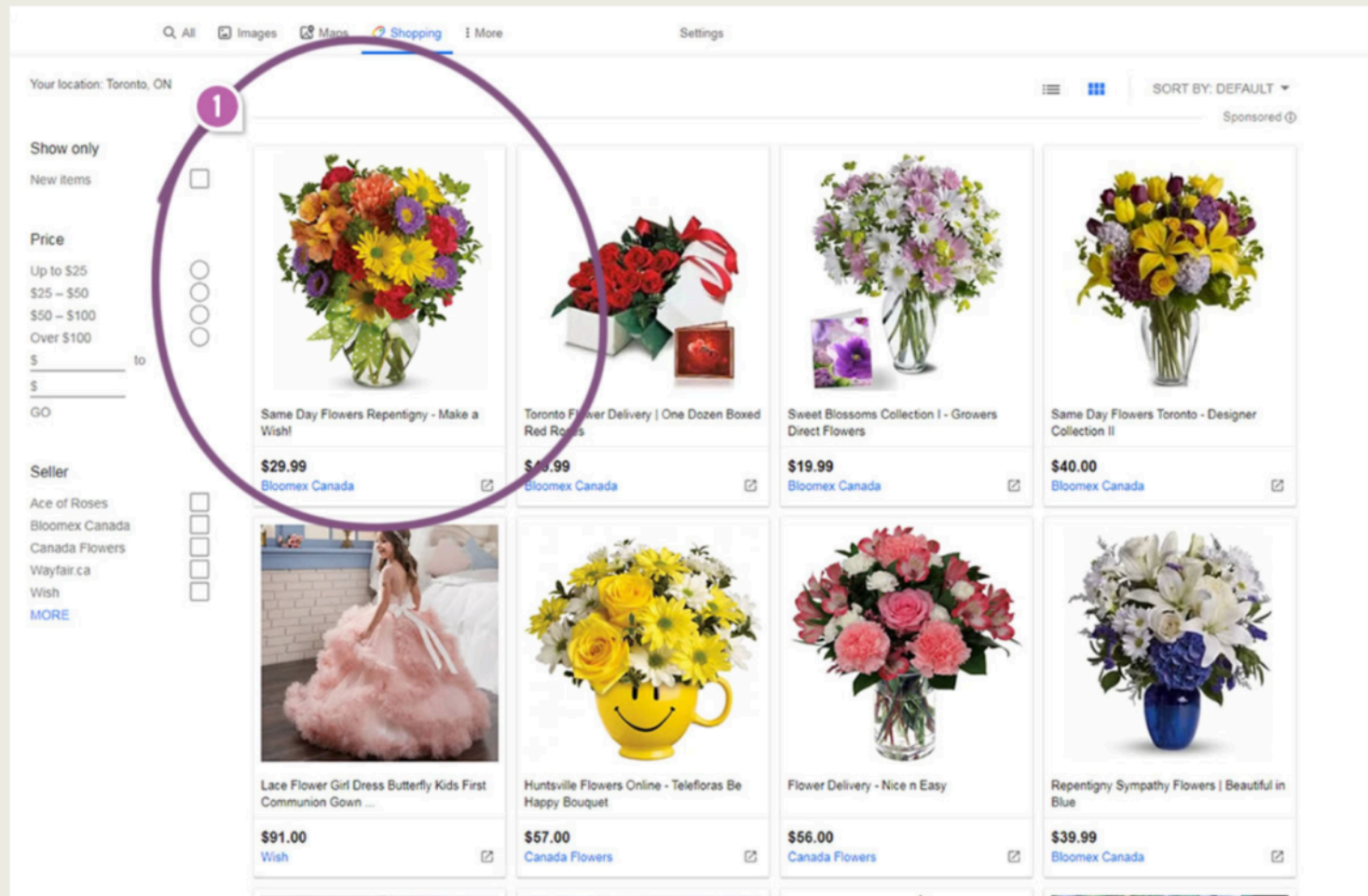
TRY IT NOW >

A/B Testing: What?

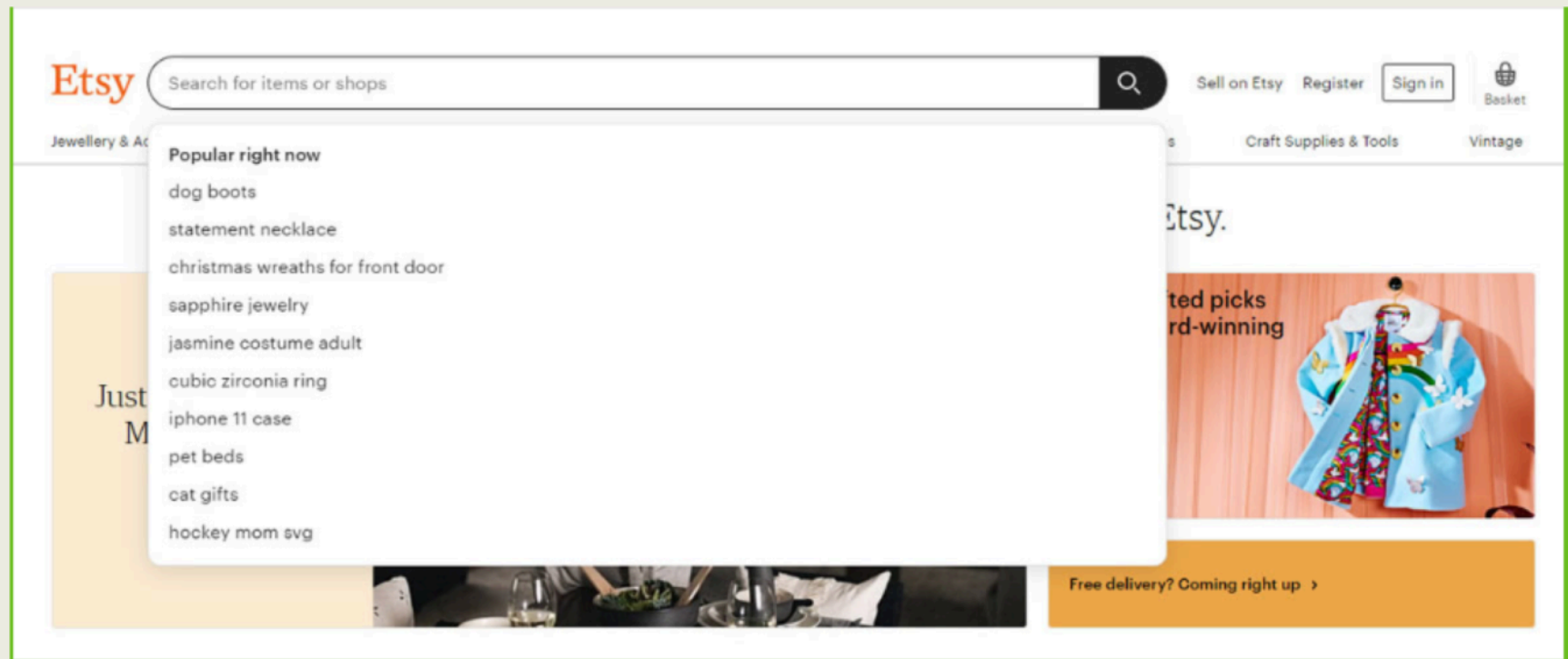
The screenshot shows a Google Shopping interface with the following elements:

- Navigation Bar:** Search icons for All, Images, Maps, Shopping, and More. A 'Settings' link is on the right.
- Location:** 'Your location: Toronto, ON'.
- Filters:**
 - Show only:** A checkbox for 'New items'.
 - Price:** Radio buttons for price ranges: 'Up to \$20', '\$20 - \$50', '\$50 - \$100', and 'Over \$100'. Input fields for custom price ranges are also present.
 - Seller:** Checkboxes for 'Ace of Roses', 'Bloomex Canada', 'Canada Flowers', 'Wayfair.ca', and 'Wish'. A 'MORE' link is at the bottom.
- Search Results:** Four flower bouquet listings are shown. The first two are circled in a purple oval:
 - Item 1:** 'Same Day Flowers Repentigny - Make a Wish!' by Bloomex Canada. Price: \$29.99. Description: 'If you close your eyes and make a wish, perhaps someone will send you this deluxe version of our Make a Wish bouquet, with a ... Bouquet'.
 - Item 2:** 'Sweet Blossoms Collection I - Growers Direct Flowers' by Bloomex Canada. Price: \$19.99. Description: 'Mauvelous for any Occassion! This sweet bouquet of fresh cut flowers features Lavender and white daisy spray chrysanthemums ... Bouquet · Daisy · Chrysanthemum'.
 - Item 3:** 'Huntsville Flowers Online - Telefloras Be Happy Bouquet' by Canada Flowers. Price: \$57.00. Description: 'Telefloras Be Happy Bouquet Huntsville. Unique and expressive arrangement of 6 lush red roses with greenery and beargrass in ... Bouquet · Rose · With Vase'.
 - Item 4:** 'Same Day Flowers Toronto - Designer Collection II' by Bloomex Canada. Price: \$40.00. Description: 'Our specially priced Designers Collection bouquets allow our professional, on-staff Floral Designers to work their magic and ... Bouquet'.

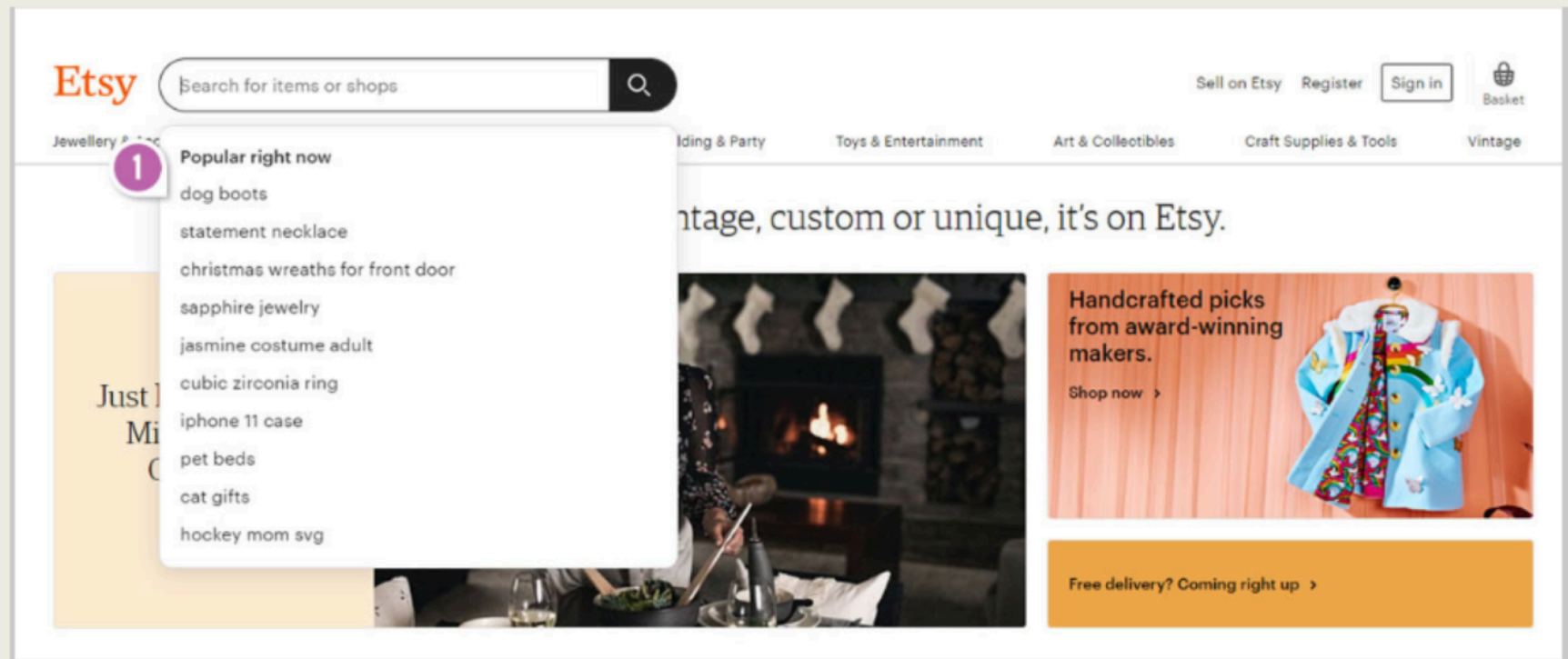
A/B Testing: What?



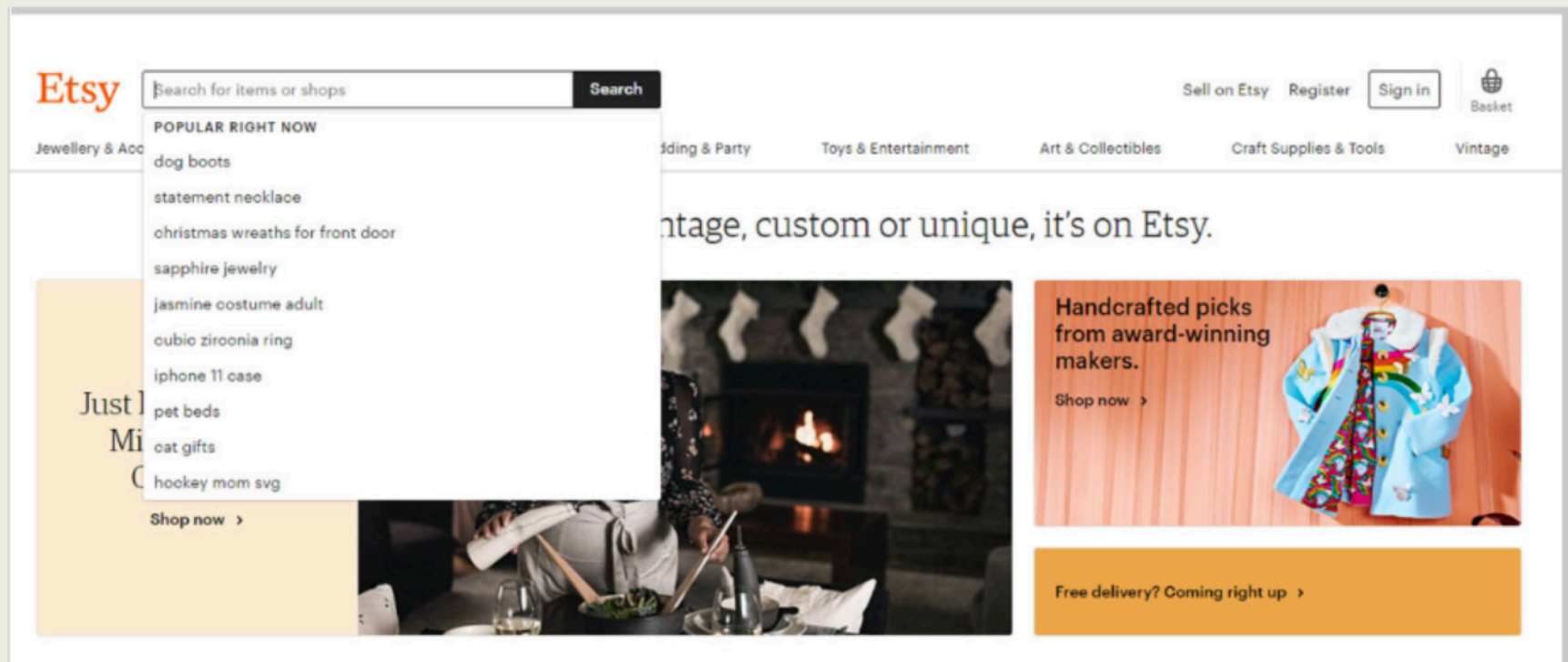
A/B Testing: What?



A/B Testing: What?



A/B Testing: What?



A/B Testing: Who?

- Large organizations such as Google, Facebook, Amazon, Microsoft are running 10,000+ experiments per year
- LinkedIn is reportedly simultaneously running 400+ experiments per day
- 1,000's of companies use tools such as Optimizely, KissMetrics, MixPanel, VWO and Split.io to run tests
 - Optimizely has around 500 employees and is reportedly worth \$500M+ ^[3]



Senior Data Analyst
Glassdoor ★★

Apply Now



Data Scientist, Product analytics

StubHub ★★★★★ 140 reviews - San Francisco, CA 94105
(Financial District area)

Apply On Company Site



Requirements:

- 3+ years of quantitative background
- Track record in building and presenting actionable insights
- No less than 2 years of experience with SQL
- Experienced with machine learning or regression analysis
- Thorough knowledge of A/B testing and analyzing A/B tests
- Strong visualization skills
- 1+ years of mentoring junior analysts

This Role:

As a Senior Data Analyst, you will drive the data strategy for StubHub's marketplace and you will see the impact of your statistical and analytical work on the business. You will be responsible for a job? What is the question? How should we answer it? What are the implications of our actions to how we solve those problems? We solve those problems by predicting whether prospect subscription, a customer's

Decision Science at Glassdoor
The Decision Science team

You have:

- Strong quantitative background with Product Analytics experience, proven track record in solving business problems through fact based and scientific analytics
- Excellent technical skills - including the ability to query databases (SQL, Hive etc.), leverage tools like Python, R, Google Analytics, and Tableau (or other visualization tools) to deliver insights
- Deep understanding of statistics, experimental design, and causal inference
- The ability to translate analytical insights into clear recommendations and effectively deliver to technical and non-technical stakeholders
- Worked in cross-functional and cross-cultural teams, and are able to communicate technically intricate concepts/results in business terms
- Passion for technology and consumer products

Why StubHub?

StubHub is the world's largest ticket marketplace, enabling fans to buy and sell tickets to tens of thousands of sports, concert, theater and other live entertainment events. StubHub reinvented the ticket resale market in 2000 and continues to lead it through innovation. The company's unique online marketplace, dedicated solely to tickets, provides all fans the choice to buy or sell their tickets in a safe, convenient and highly reliable environment. All transactions are processed and delivered by StubHub and backed by the company's FanProtect Guarantee™ processing a ticket every second today with

quantitative analysis
business product
results of complex

OR building
tion, or exploration in a
ns and methods
ing, consulting, or

the U.S. Federal E-
ities and protected
ified applicants with
icisco Fair Chance

gs together people,
effectively work
of thousands of
global Feature 100

ss 400+ cities v
and deepening e
arketing. You wi
t, including qua
onomy gives u

A/B Testing: Where?

- User acquisition funnels
- User engagement mechanics
- User retention mechanics
- Email promotions
- Website layout
- Esthetic features
- Checkout experience
- Freemium conversion
- Branding
- Ad campaigns
- Call to action language
- ML algorithms

A/B Testing: Why?

BECAUSE \$\$\$\$

A/B Testing: Why?

- A controlled experiment is the **only** way to cleanly establish causal relationships.
- It facilitates data-driven decision making...
- ...where you listen to your **customers**
 - Not your gut
 - Not your designers
 - Not the HiPPO

A/B Testing: How?

- **Step 1:** Define a business hypothesis framed in terms of the metric θ you wish to optimize
- **Step 2:** Translate the the business hypothesis into a statistical hypothesis:

$$H_0: \theta_C = \theta_T \text{ vs. } H_A: \theta_C \neq \theta_T$$

$$H_0: \theta_C \geq \theta_T \text{ vs. } H_A: \theta_C < \theta_T$$

$$H_0: \theta_C \leq \theta_T \text{ vs. } H_A: \theta_C > \theta_T$$

A/B Testing: How?

- **Step 3:** Define and produce your experimental conditions
- **Step 4:** Determine how many experimental units are required in each condition (i.e., sample size determination): n_C, n_T
 - *Minimize significance (α) or Type I Error*
 - *Maximize Power ($1-\beta$) or 1-Type II Error*
- **Step 5:** Collect the data

$$\{x_1, x_2, \dots, x_{n_C}\} \text{ and } \{y_1, y_2, \dots, y_{n_T}\}$$

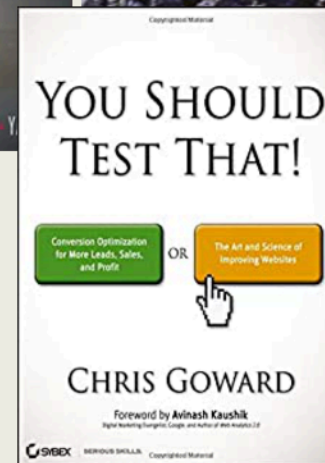
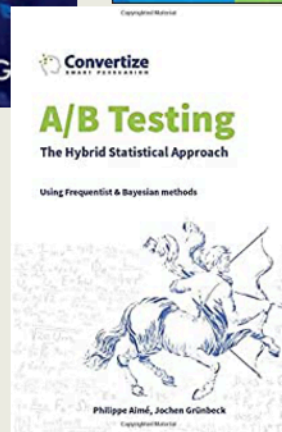
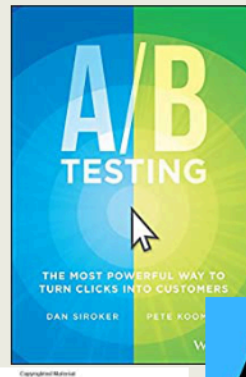
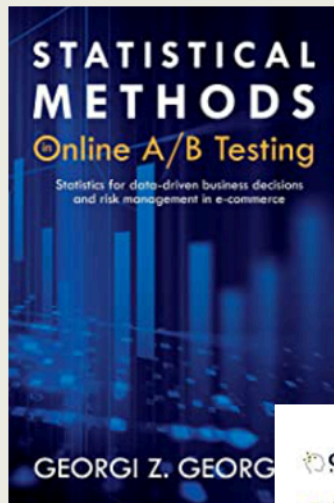
A/B Testing: How?

- **Step 6:** Estimate the metric of interest in each condition: $\hat{\theta}_C$ and $\hat{\theta}_T$
- **Step 7:** Determine whether the difference between $\hat{\theta}_C$ and $\hat{\theta}_T$ is statistically significant
 - t-test (F-test)
 - Z-test (χ^2 -test)
 - Permutation test
- But usually just a Z-test

Other Strategies

- Many more sophisticated experimental design and analysis strategies are available:
 - Factorial designs
 - Fractional factorial designs
 - Response surface designs
 - Multi-armed bandits

So many resources...





SO WHY IS THIS
HARD?

Industry Problems != Academic Interest

- Industry Problems may be sensitive:
 - *Not publishable*
- Industry Problems may be too specific:
 - *Not general enough to publish in good journals*
- Industry Problems may too hard:
 - *Not worth investing time if you can't possibly solve it.*

Industry Problems != Academic Interest

Appears in the June 2019 issue of [SIGKDD Explorations](#) Volume 21, Issue 1

<https://bit.ly/OCESummit1>

Top Challenges from the first Practical Online Controlled Experiments Summit

Somit Gupta (Microsoft)¹, Ronny Kohavi (Microsoft)², Diane Tang (Google)³, Ya Xu (LinkedIn)⁴, Reid Andersen (Airbnb), Eytan Bakshy (Facebook), Niall Cardin (Google), Sumitha Chandran (Lyft), Nanyu Chen (LinkedIn), Dominic Coey (Facebook), Mike Curtis (Google), Alex Deng (Microsoft), Weitao Duan (LinkedIn), Peter Forbes (Netflix), Brian Frasca (Microsoft), Tommy Guy (Microsoft), Guido W. Imbens (Stanford), Guillaume Saint Jacques (LinkedIn), Pranav Kantawala (Google), Ilya Katsev (Yandex), Moshe Katzwer (Uber), Mikael Konutgan (Facebook), Elena Kunakova (Yandex), Minyong Lee (Airbnb), MJ Lee (Lyft), Joseph Liu (Twitter), James McQueen (Amazon), Amir Najmi (Google), Brent Smith (Amazon), Vivek Trehan (Uber), Lukas Vermeer (Booking.com), Toby Walker (Microsoft), Jeffrey Wong (Netflix), Igor Yashkov (Yandex)

ABSTRACT

Online controlled experiments (OCEs), also known as A/B tests, have become ubiquitous in evaluating the impact of changes made to software products and services. While the concept of online

1.1 First Practical Online Controlled Experiments Summit, 2018

To understand the top practical challenges in running OCEs at scale, representatives with experience in large-scale

Given this...

- There are long-standing problems / difficulties that often can't be solved via a formula
- Operationally AB Testing requires organizational support
 - again outside of the scope of what is taught.
- Lets talk about some specifics:
 - *User Identification*
 - *Peeking*
 - *Lift Bias*
 - *Interference*



FAILING TO DEFINE A USER



User identification

- What is a user?
- Identification strategy:
 - *Software based identification*
 - Web cookie, file on the hard drive
 - *Hardware based identification*
 - Serial numbers (IDFV on mobile)
 - *Required login*
 - 3rd party (Facebook, Twitter)
 - 1st party (roll your own)

Most common

- Some Combination:

- *No login required until a threshold achieved*
- *Login “optional” but gives additional features*
- *Different accounts that may/may not be linkable*
 - Quora (FB vs. email)
 - The Meta (Steam vs. Meta Login)

Who cares?

- A “User” may experience multiple experimental treatments:
 - *Nick has no login and is assigned to treatment group A*
 - *Nick creates an account and is assigned to group treatment B*
 - *Nick uses Twitter login on desktop and FB login on mobile. Accounts not linked, one in group A and one in group B.*
 - *Nick sign up to the news letter with two different email addresses and looks for the best deals in any A/B situation*

Result

- For knife-edge conclusions, a small percentage of users being misidentified can swing the results
- Systematic misidentification (especially) can skew a test completely
 - *1% of users have two accounts, but 100% of those users choose experience treatment B.*

Solution

1. Well defined tests:
 - *Avoid obvious identification issues*
 - *Focus on subsets of users (post account creation, users who do not cross play, etc.)*
2. Rely on **organizational** momentum:
 - *Define a framework for testing which avoids these issues*
 - *Organizations tend to have momentum. Once it's done once, rely on “This is how we do it.”*



FAILING TO GET THE
RIGHT RESULTS



What is peeking?

- **Peeking** is the phenomenon whereby you regularly check the results of the experiment before it finishes
- Peeking can be a good thing!
 - *Make sure the experiment is not negatively impacting other important metrics*
 - *Verify experiment is running correctly*
- The problem arises when, as a result of peeking, you decide to end the experiment early

Example

- I set up my experiment:
 - *I need 1,000 users in both the Treatment and Control group*
- On the first day, I look at my data:

	Observations Collected	Conversion Rate
Treatment	150	10%
Control	150	5%

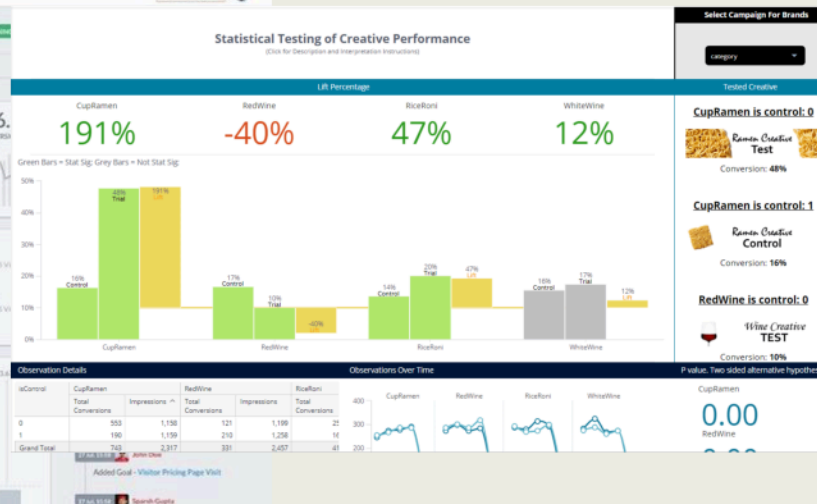
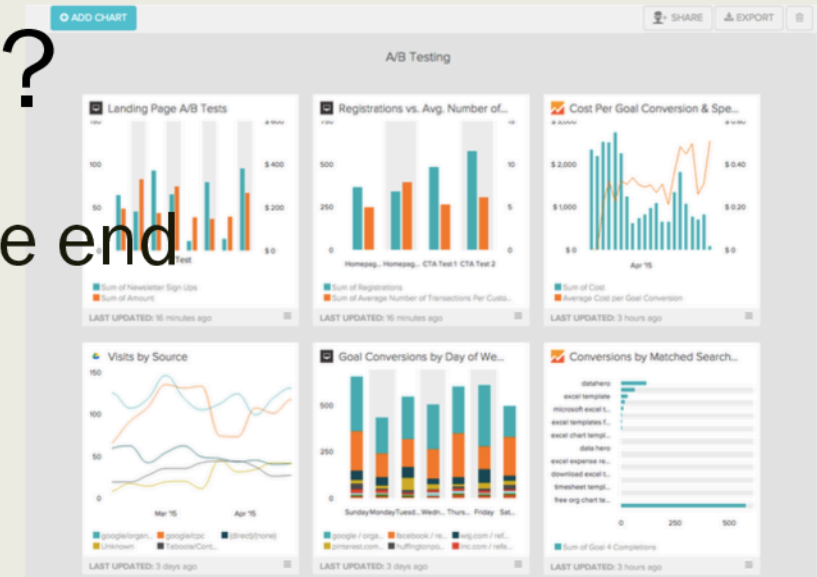
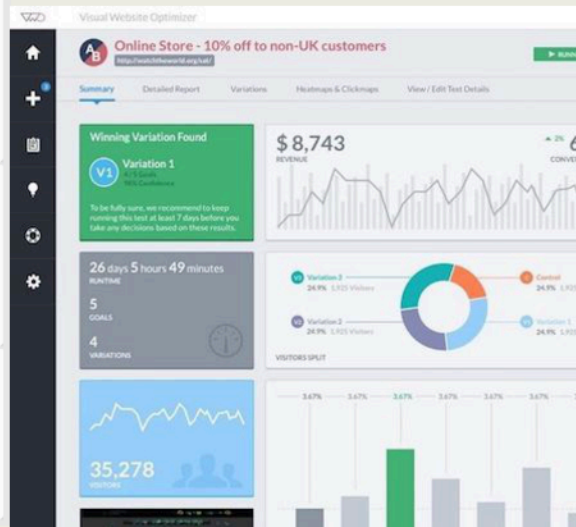
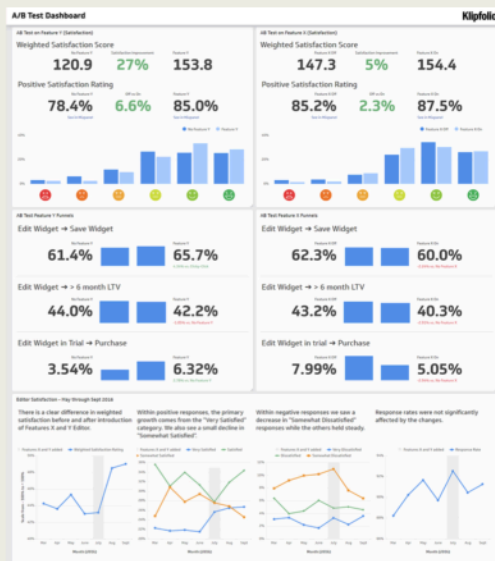
- *The conversion rate in the Treatment group is twice that of the Control. Should we stop the test?*

Why is this a problem?

- By stopping the experiment early you have not observed enough data to be confident in your conclusion
 - Just because the results suggest a winner or a significant difference at one point in time does not mean that the results won't change as more data is collected

Why is this a problem?

- Okay, so just run the test to the end
- But it's hard
- So many dashboards



Why is this a problem?

- When you stop the experiment you are rejecting the null hypothesis
- Which means you might be making a Type I error
- And by stopping the experiment early the chances you make a Type I error are **much higher** than the prespecified statistical significance (α)

Why is this a problem?

Illustrative Simulation

- $n_C = n_T = 1,000$ data points are drawn independently from the $N(0,1)$ distribution
- The observations are used to perform a Z-test of

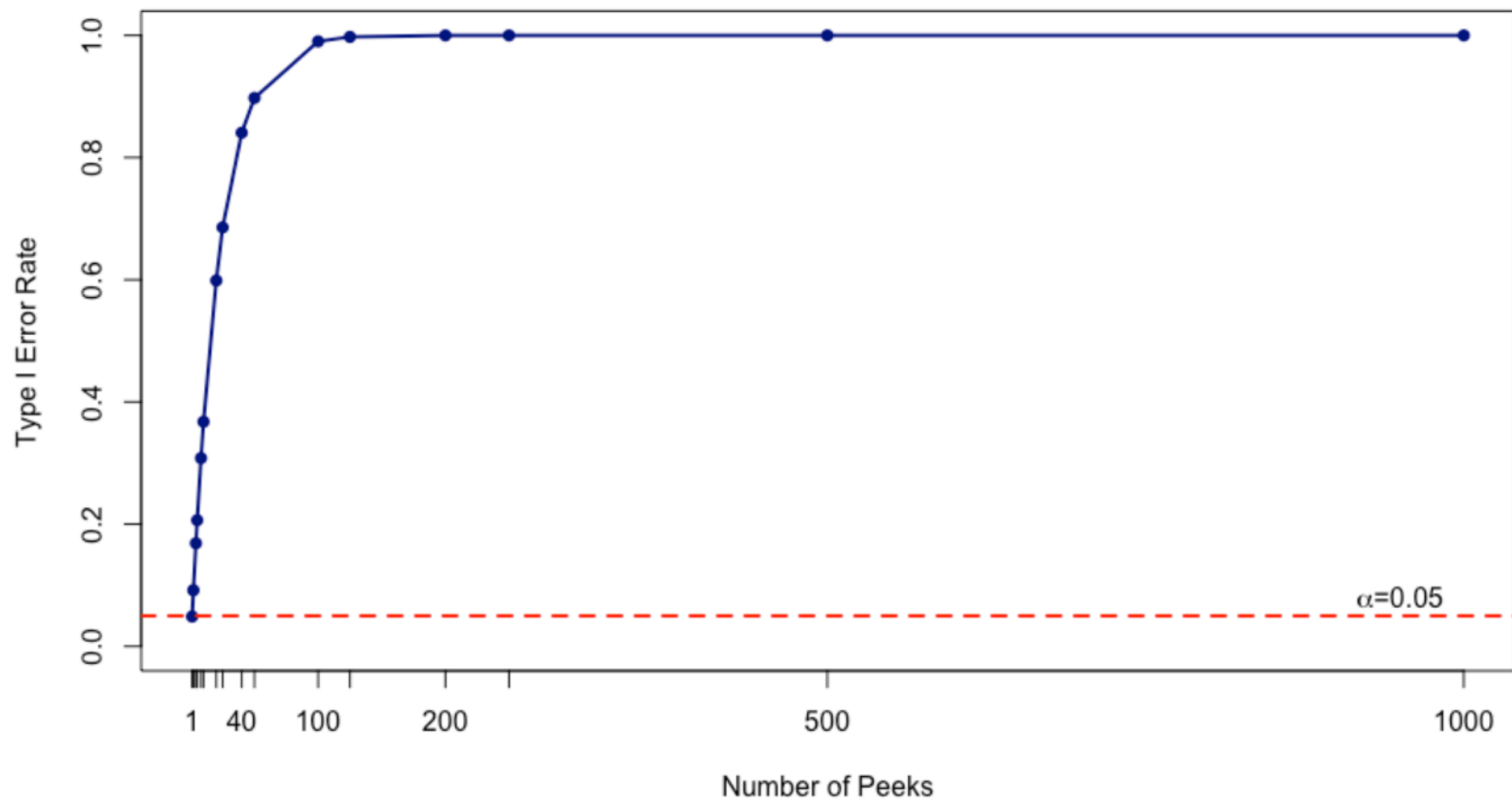
$$H_0: \theta_C \leq \theta_T \text{ vs. } H_A: \theta_C > \theta_T$$

- Because $\theta_C = \theta_T = 0$ we should not reject H_0 very often (no more than $\alpha \times 100\%$ of the time)

Why is this a problem?

- To study the consequences of peeking, we peek – **and end the experiment if a significant result is indicated** – at regular intervals
- Repeat this 10,000 times
- The Type I Error rate is the fraction of the 10,000 simulations that an experiment is ended prematurely

Why is this a problem?



What is the solution?

■ Sequential Testing

- An analysis method where the sample size is not fixed a priori
- Data are accumulated and analyzed sequentially until a stopping rule is met
- Stopping rule is based on α and β -spending functions
- Resulting lift estimates need to be bias-corrected
- More complex to implement

What is the solution?

- Avoid having non-sophisticated users end tests early
 - *Presentation layer:*
 - Modify presentation with explicit warnings
 - Hide results
- Require test to have a minimum number of units (as part of the the design)

Observations Collected		PERCENT COMPLETE	Conversion Rate
Treatment	150	15%	10%
Control	150	15%	5%



FAILING TO LIFT OFF

Hmmmm... ?

- You run a test
- Treatment effect has 5% higher revenue than control
- So you make the change, but revenue only increases by 2%
- This happens on **every** test.

	Control	Treatment	Estimated Diff	Actual Diff
Test #1	17%	12%	5%	2%
Test #2	5%	2%	3%	1.8%
Test #3	7%	3%	4%	3.2%
Test #4	9%	4.5%	4.5%	4%
Test #5	8%	6%	2%	1.4%

So what is this bias?

- Let $\delta = \theta_C - \theta_T$ be the true unknown **treatment effect** (aka: **lift**)
- This is estimated by:

$$\hat{\delta} = \hat{\theta}_C - \hat{\theta}_T = \bar{X} - \bar{Y}$$

- This is an **unbiased estimate** of lift:

$$E[\bar{X} - \bar{Y}] = \theta_C - \theta_T = \delta$$

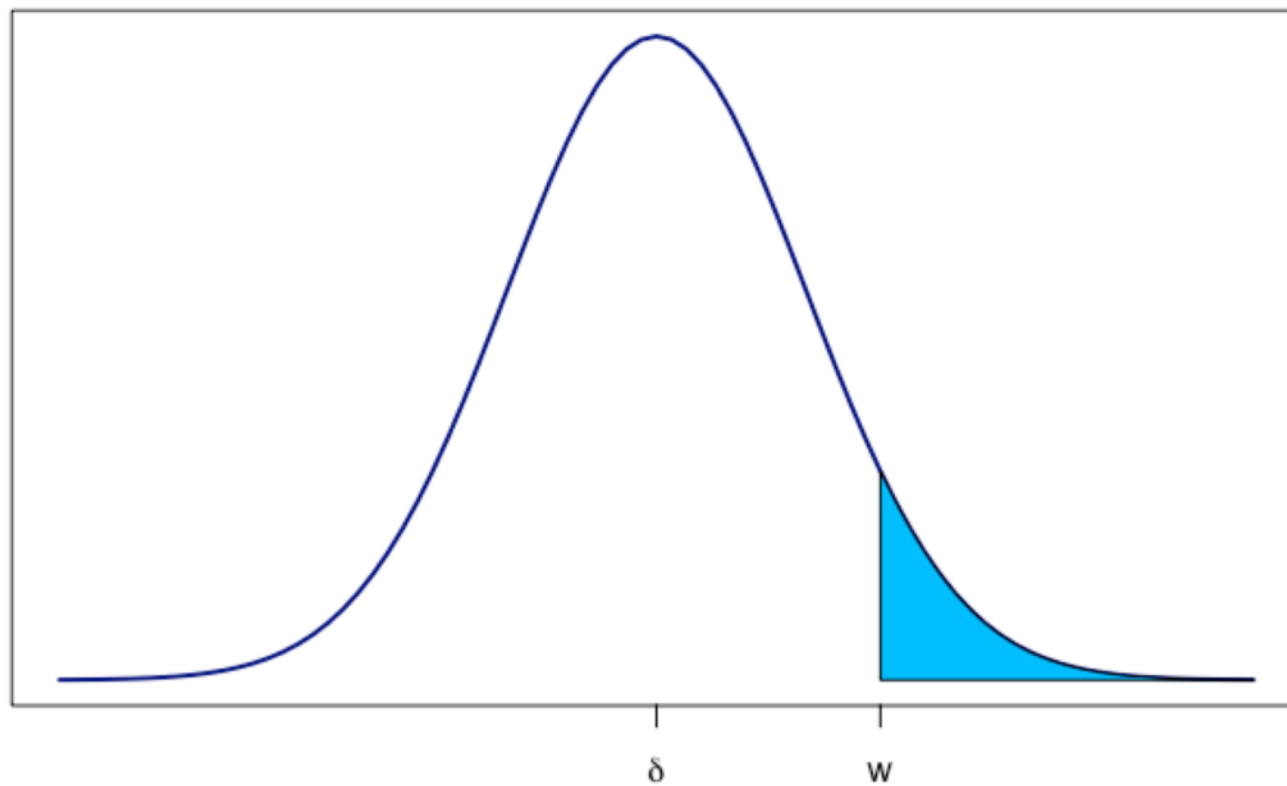
So where's the problem?

Problem:

- This isn't how we estimate lift in practice
- In practice lift is only ever estimated **if the null hypothesis is rejected**
- For illustration assume we're testing the hypothesis

$$H_0: \theta_C \leq \theta_T \text{ vs. } H_A: \theta_C > \theta_T$$

So where's the problem?



Distribution of $\bar{X} - \bar{Y}$

So where's the problem?

Problem:

- So what we're actually estimating in practice is

$$E[\bar{X} - \bar{Y} | \bar{X} - \bar{Y} \geq w]$$

not

$$E[\bar{X} - \bar{Y}]$$

Note: $w = \sigma \times z^*$ where $\sigma = SD[\bar{X} - \bar{Y}]$ and z^* is the appropriate critical value of $N(0,1)$ determined by α

So where's the problem?

Problem:

When

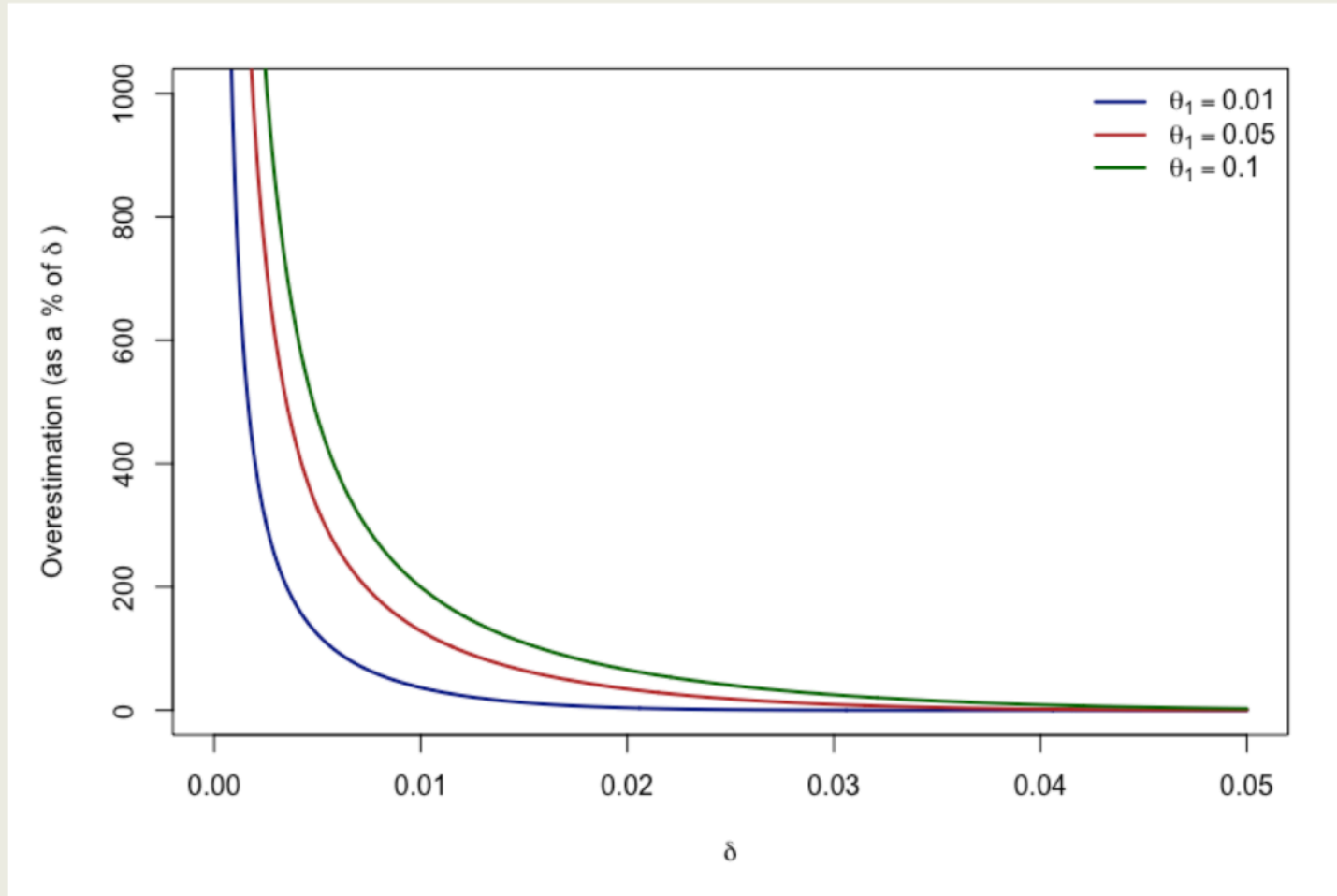
$$H_0: \theta_C \leq \theta_T \text{ vs. } H_A: \theta_C > \theta_T$$

then

$$E[\bar{X} - \bar{Y} | \bar{X} - \bar{Y} \geq w] = \delta + \sigma \frac{\phi\left(\frac{w - \delta}{\sigma}\right)}{1 - \Phi\left(\frac{w - \delta}{\sigma}\right)}$$

which is strictly **greater** than δ

How big a problem is this?



So what can we do?

- Accept that the lift estimated from your experiment is an overestimate
- Sadly, the statistics behind estimating this are difficult so can't just "undo" it
- **Presentation layer:**
 - *Add "Max Difference" or add an "Estimated" lift to the presentation.*



FAILING TO DESIGN

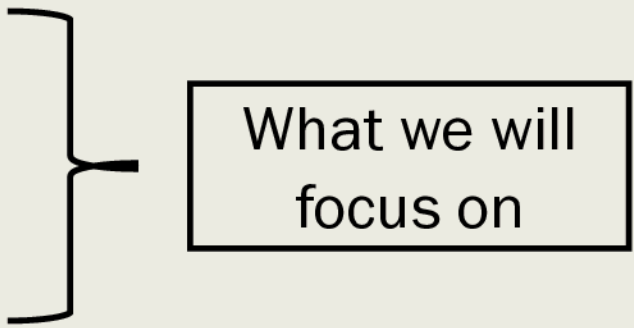
What is Interference?

- Problems of interference occur when your experimental conditions become **contaminated**
- This typically means that the **Stable Unit Treatment Value Assumption (SUTVA)** has been violated
 - SUTVA: The outcome observed on one unit should be unaffected by the treatment assignment of other units
- Your experimental conditions are no longer independent

What is Interference?

- Interference/contamination can happen for a variety of reasons:

- Unit unidentifiability (spoken about before)
- Colliding experiments (be careful)
- **Network interference**
 - **Intra contamination**
 - **Inter contamination**



What we will
focus on

Network Interference

- What if my experiment effects other users and, in turn, modifies their behavior?
- Facebook does an A/B test on “People you May Know”
 - Control group sees “as is”
 - Treatment sees “new flow”
- If treatment causes more friend requests, which then increase friend requests for control users, then my lift estimates will be incorrect
- What if my users directly communicate to each other about test conditions?
 - *This will change behavior (test/control group may be unhappy and do something negative)*

Network Interference

- Academically, this is solved by modeling as a network/graph problem
 - *Many assumptions*
 - *Specific knowledge / parameter estimates, etc.*
- “Real world”
 - *Tend to either **ignore** or design around (geo-fencing + light modeling)*

How bad is ignoring?

- Given that this “only happens a little” in my product, how much does it matter?
- At the Meta, we don’t expect this to be too much of an issue (outside of leaderboards our product does not have too much of a social element)
- Go over two models and see what happens
 - *Correlation between treatment groups*
 - *Correlation within treatment groups*

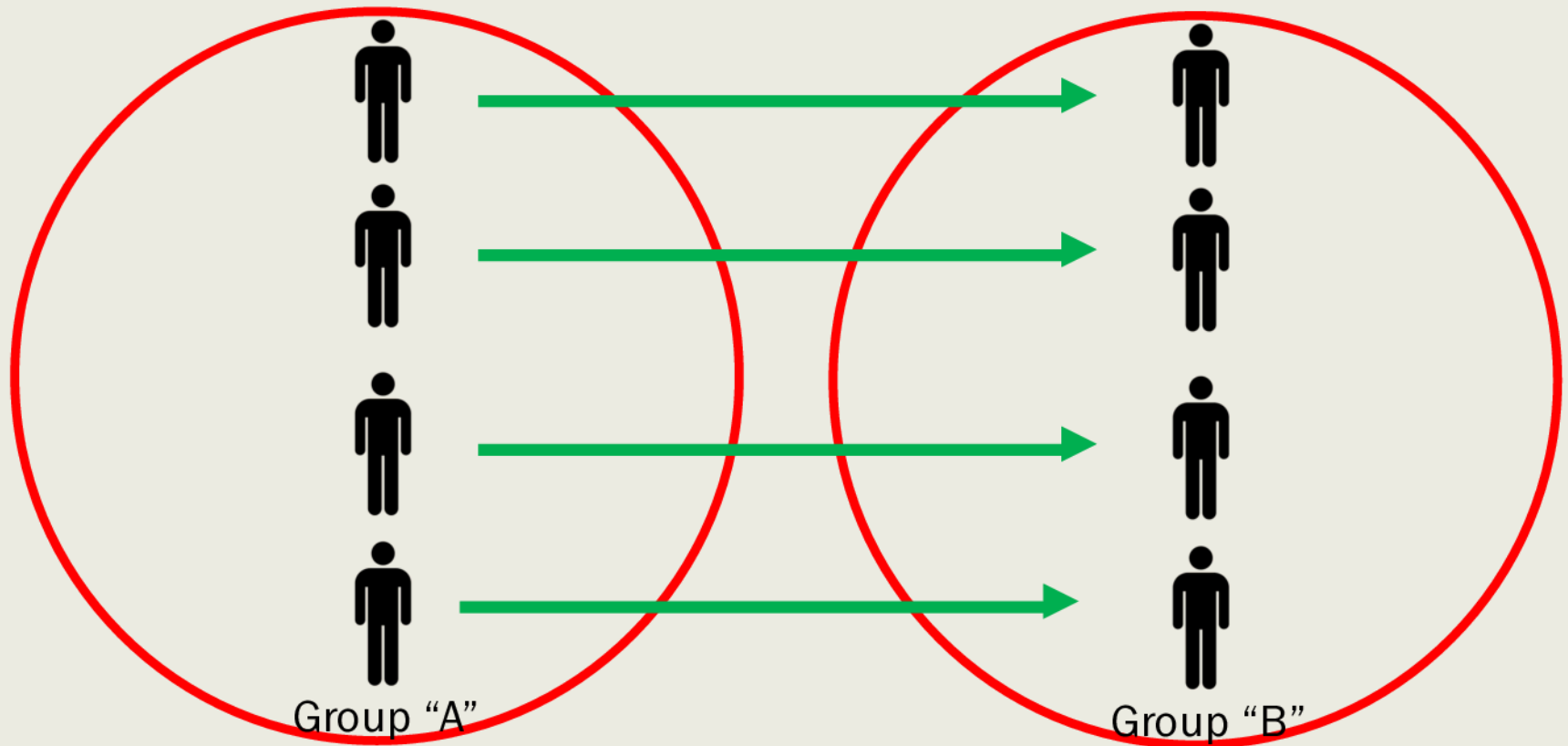
Between Treatment Groups

- Standard T-test of significance
- Users from one treatment group effect the outcome of the other treatment group
- Specifically:

$$COV(Y_{i,A}, Y_{j,B}) = \begin{cases} 0 & otherwise \\ \lambda\sigma^2 & if\ i = j \end{cases}$$

- If user #1 in treatment A does something => effects the outcome of user #1 in treatment B

Between Treatment Groups



Between Treatment Groups

- Our T Statistic:

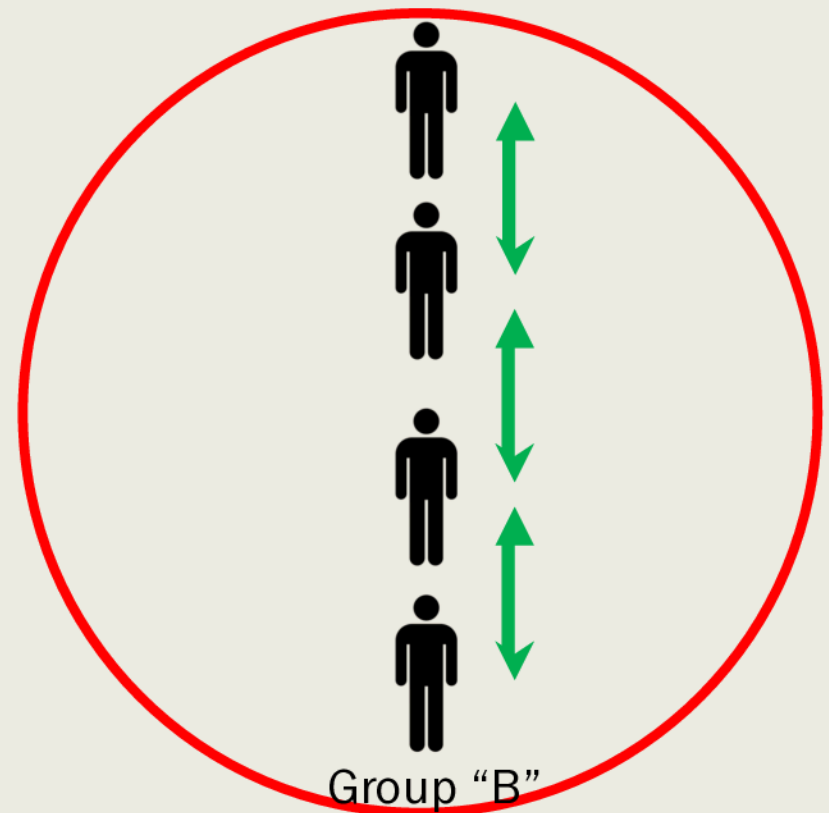
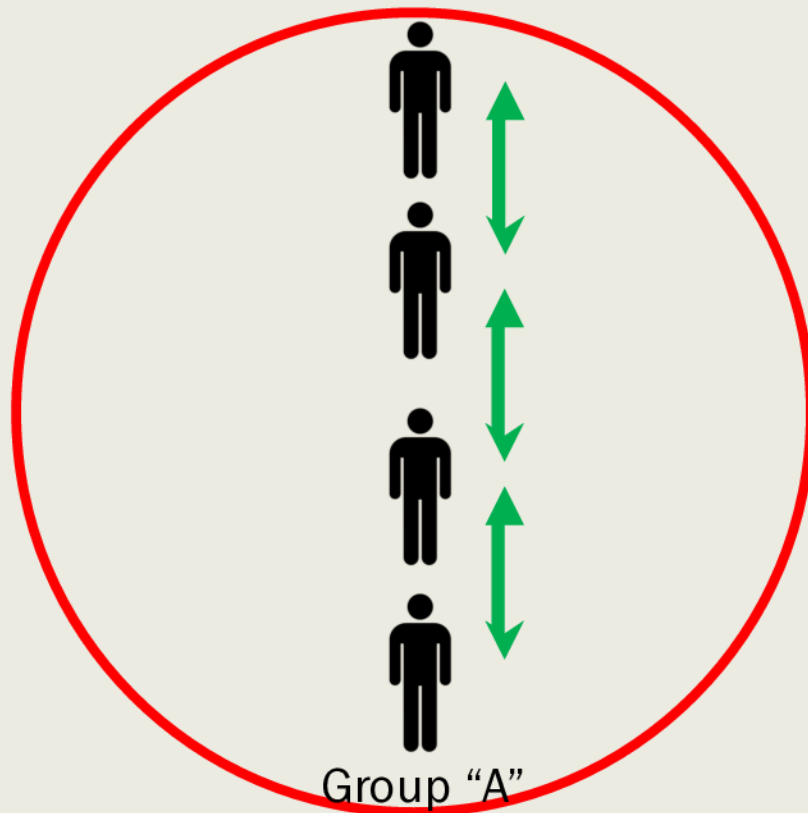
$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma \sqrt{\frac{1}{n} + \frac{1}{n}}}$$

- Without Interference: $\text{VAR}[T] = 1$
- With Interference: $\text{VAR}[T] = 1 - \lambda$

Conclusion (correlation between groups)

- Since λ can be positive or negative so we don't even know the direction of the bias!
- This is a really simple, well specified case.
- One nice thing – in this case increasing our sample size will naturally help things:
 - *While it doesn't solve the interference it does spread out our estimators (\bar{Y}_1, \bar{Y}_2) making the interference less costly.*

Between Treatment Groups



What about correlation within groups?

- Once again, standard T-test of significance
- Users from one treatment group effect the outcome of the other treatment group
- Specifically:

$$COV(Y_{i,j}, Y_{k,j}) = \begin{cases} \lambda\sigma^2 & \text{if } i \neq k \\ \sigma^2 & \text{if } i = k \end{cases}$$

- Note that $j \in (A, B)$. We assume zero correlation between groups

Within Treatment Groups

- Our T Statistic:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma \sqrt{\frac{1}{n} + \frac{1}{n}}}$$

- Without Interference: $\text{VAR}[T] = 1$
- With Interference: $\text{VAR}[T] = 1 + (n - 1) \lambda$

Conclusion (correlation between groups)

- Since λ can be positive or negative, so unless we know it's value it's difficult to conduct a test.
- This is a really simple, well specified case.
- Increasing the sample size in this case makes interference **worse**.

How to handle interference

- Academically:
 - *Network modelling*
 - *Econometric (but-for analysis)*
 - *Matched-pairs experimental design (geo-fencing)*
- All of these are difficult and costly (man power)

How to handle interference

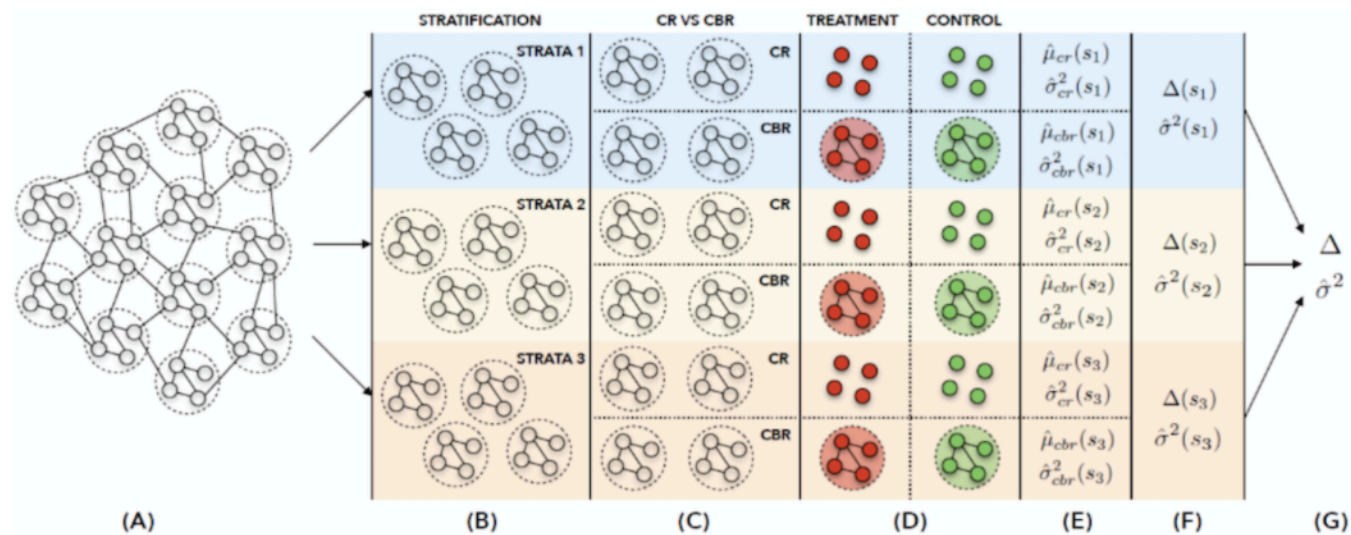


Figure 2: Illustration of the proposed experimental design for detecting network effects, using stratification to reduce variance and improve covariate balance. (A) Graph of all units and their connections; the dashed circles represent clusters. (B) Assigning clusters to strata (Section 2.4), (C) Assigning clusters within each strata to treatment arms: completely randomized (CR) and cluster-based randomized assignment (CBR). (D) Assigning units within strata to treatment buckets: treatment and control, using corresponding assignment strategy. (E) Computing the treatment effects of each treatment arm within each strata: $\hat{\mu}_{cr}(s)$ and $\hat{\mu}_{cbr}(s)$, and variance within each strata: $\hat{\sigma}_{cr}^2(s)$ and $\hat{\sigma}_{cbr}^2(s)$. (F) Computing the difference between the estimated effects using CR and CBR within each strata: $\Delta(s)$, and sum variances in each strata: $\hat{\sigma}^2(s)$. (G) Aggregating the differences across strata to compute the overall difference in differences (Δ) and the total variance ($\hat{\sigma}^2$).

How to handle interference

- So... I'm still not sure
- In the short term, try to avoid experimental situations that might make it worse:
 - *Social offers*
 - *Leaderboards functionality*
 - *Tie-in testing (adding social logins, rewarding for streaming, etc.)*
- In particular – probably (technology side), make these types of test costly to do



CONCLUSION



Conclusion

- While a lot of experimentation is well-known, the details of implementation are difficult
- “Solved” Academic Problem still exist in Industry
 - *Defining Users*
 - *Peeking*
 - *Estimating lift*
 - *Interference*
- Putting that solution into practice is incredibly difficult which is why understanding them = getting good jobs :)



THE END!



I DON'T WANT TO RISE & GRIND ANYMORE