A/B TESTING INCORRECTLY

(AND SOME JOB SEARCH ADVICE)

September 16th

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 Industry-Academic Partnership*
- 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

5

Motivation & Goals

- Currently building an AB-testing system

- 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)
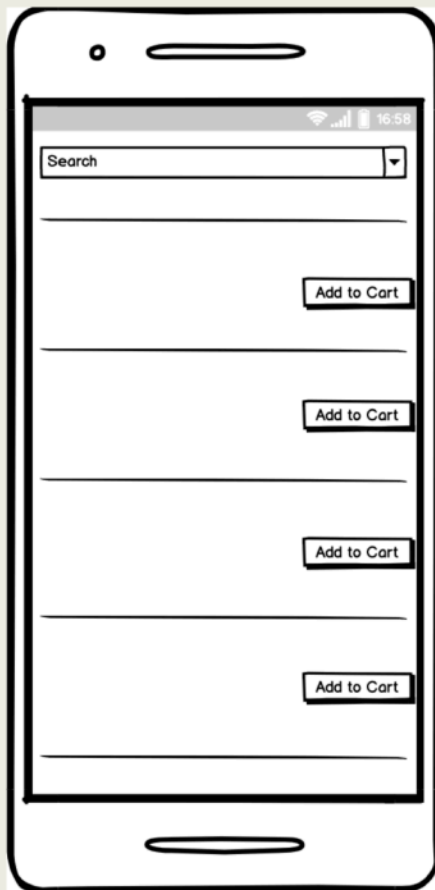
A/B Testing: What?

- **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
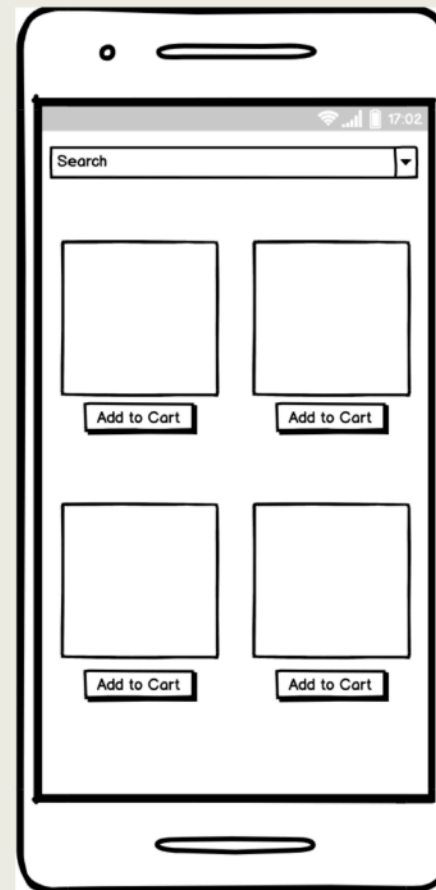
A/B Testing: What?
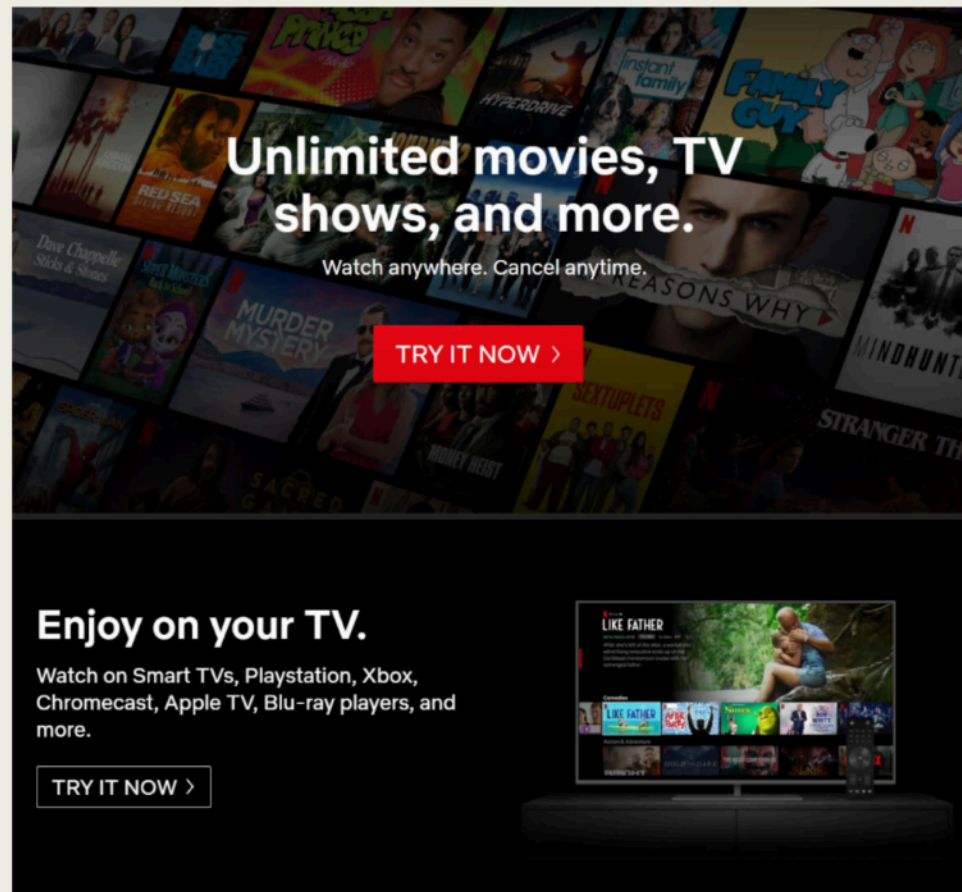

A

B
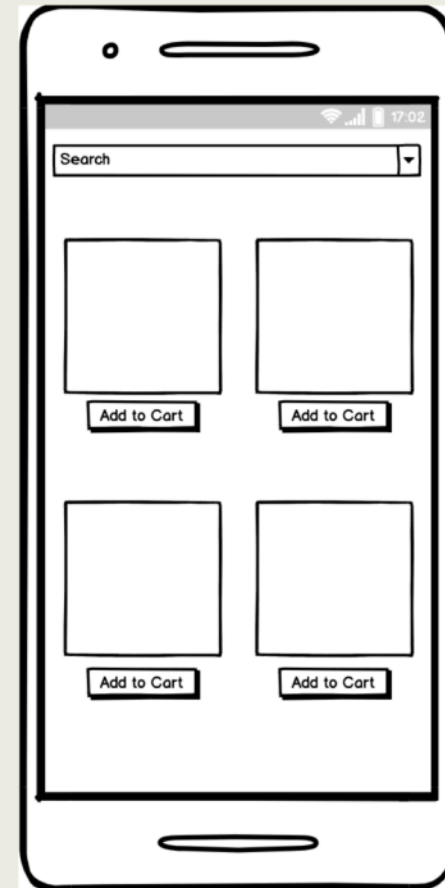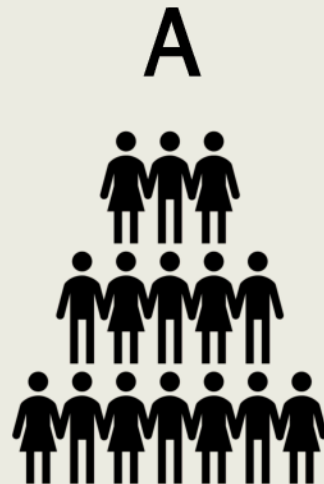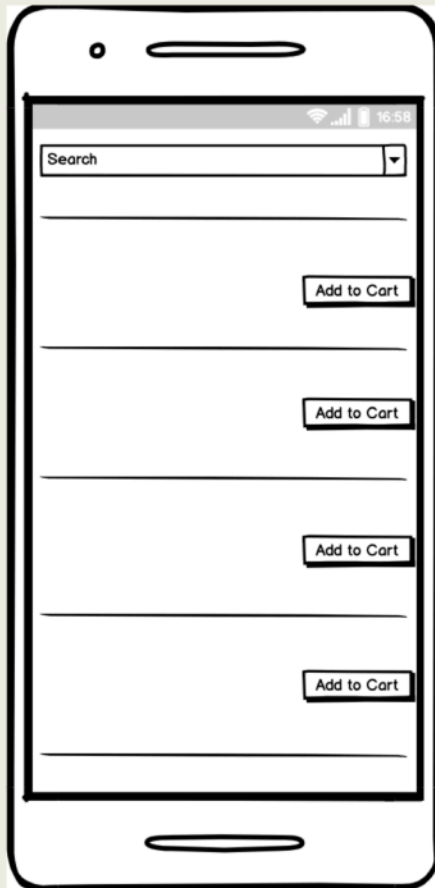
A/B Testing: What?

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?

A/B Testing: What?

A/B Testing: What?

A/B Testing: What?

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 Dat...

Glassdoor ★★

Apply Now

Requirements:

- 3+ years of quantit...
- Track record in buil...
  presenting actional...
- No less than exper...
- Experienced with c...
  Learning or regress...
- Thorough knowledg...
  and analyzing A/B t...
- Strong visualizatio...
- 1+ years of ment...

This Role:
...is role you... drive...
marketplace and y... seer...
...tatistic... nd analytic... ...
...ply for a job? Wha...
...ntion? How should we...
a... nt of a... ...ations to...
ho... ... we solve thos...
predict whether prospect...
subscription, a customer...

Decision Science at Glas...
The Decision Science team...

Data Scientist, Product analytics

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

Apply On Company Site ♡

You have:

- Strong quantitative background with Product... alytics ex... ience... roven
  track record in solving business proble... s th... gh fact bas... and
  scientific analytics
- Excellent technical skills - ... ding th... ability... query databa... s (SQ...
  Hive etc.), lev... ing... ython... ... te com... x data, and ... sing
  Tableau (c... er visuali... on tools)... eliver ins... ...t...
- ... under... ...ding of... ...tistics, exp... ...ental design, and causal
  inference
- The ability to... ... slate an... al insights into clear recommendations and
  eff... ...ly deliv... t to techni... nd non-technical stakeholders
- ...orked in cross-... ...tional and cross-cultural teams, and are able to
  ...ommunicate tech... ... ally intricate concepts/results in business terms
- ... ssio... 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

---

Dat...

Turo...

About...

You ar...
proble...
ability...
busine...
to prop...

You ha...
enjoy...
excite...

About Yo...
- 4+ yea...
- 2+ yea...

...quantitative... analysi...
...usiness... ...duct

...e results of complex

...e OR building
...tion, or exploration in a

...ns and methods

...g, consulting, or

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

...ss 400+ cities...
...and deepening...
...rketing. You wi...
...t, including qua...
...tonomy gives...

...gs together people,
...effectively work
...of thousands of
...m global Fortune 100

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

EVERYWHERE

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

- **Step 5:** Collect the data
$$\{x_1, x_2, \dots, x_{n_C}\} \text{ and } \{y_1, y_2, \dots, y_{n_T}\}$$

- **Step 6:** Estimate the metric of interest in each condition: $\hat{\theta}_C$ and $\hat{\theta}_T$

A/B Testing: How?

- **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?

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

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	30%	10%
Control	150	30%	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:

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

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 $\overline{X} - \overline{Y}$

So where's the problem?

Problem:

■ So what we're actually estimating in practice is

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

not

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

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

So where's the problem?

Problem:

When

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

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:

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

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

Between Treatment Groups

Group "A" Group "B"

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 unless we know it's value it's difficult to conduct a test.

- 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

Group "A"

Group "B"

What about correlation within groups?

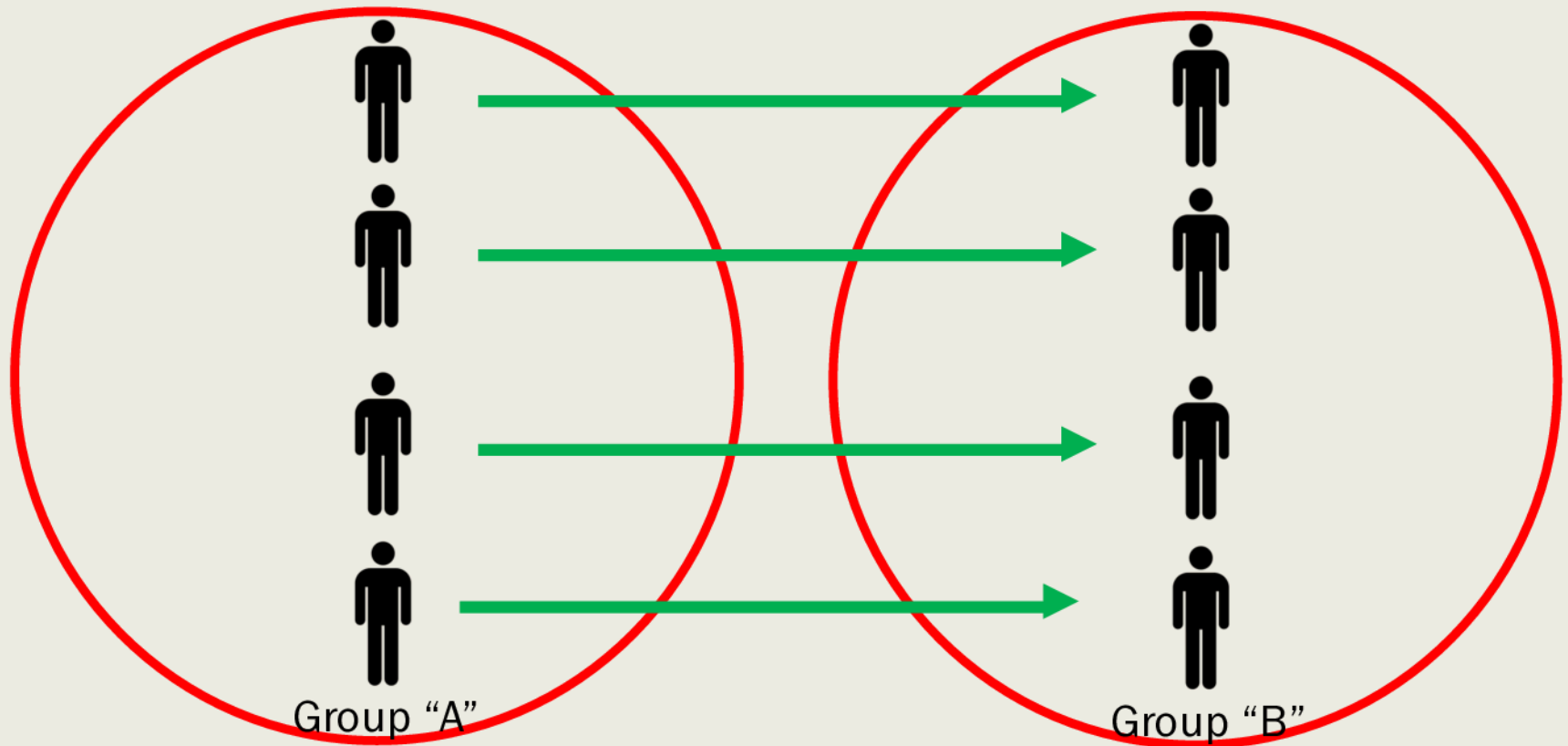- Once again, standard T-test of significance
- Users from one treatment group effect the outcome of the other treatment group
- Specifically:

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

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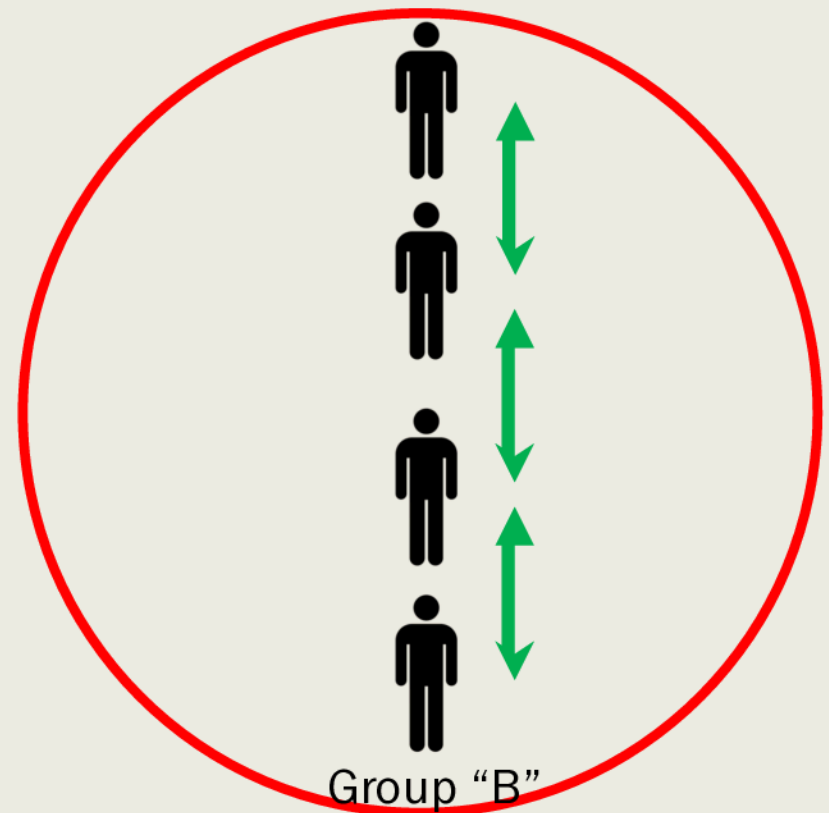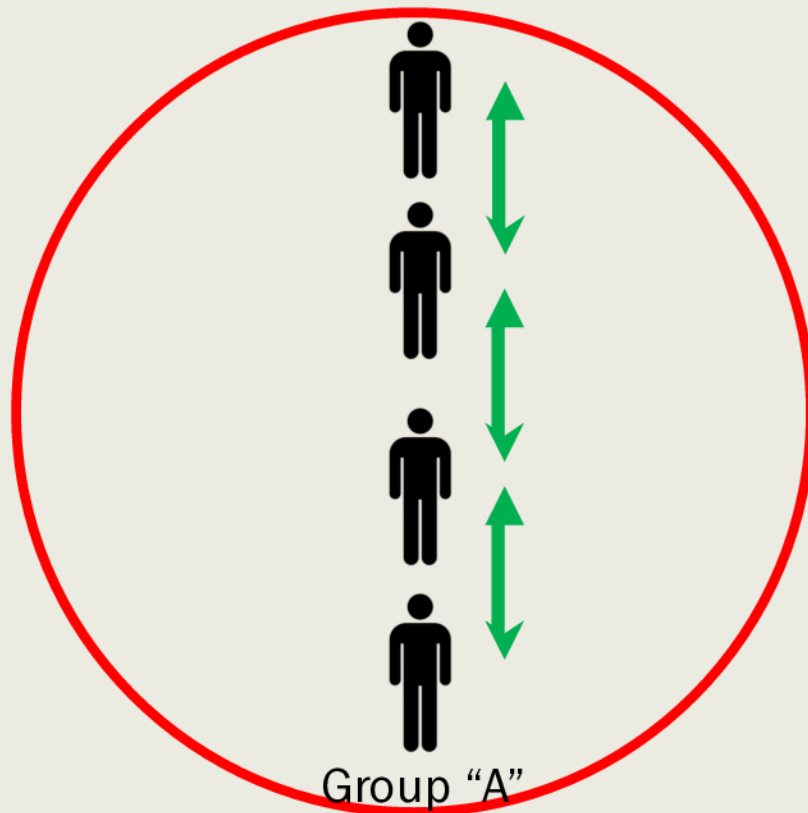
Within Treatment Groups

- Our T Statistic:

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

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

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

- 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
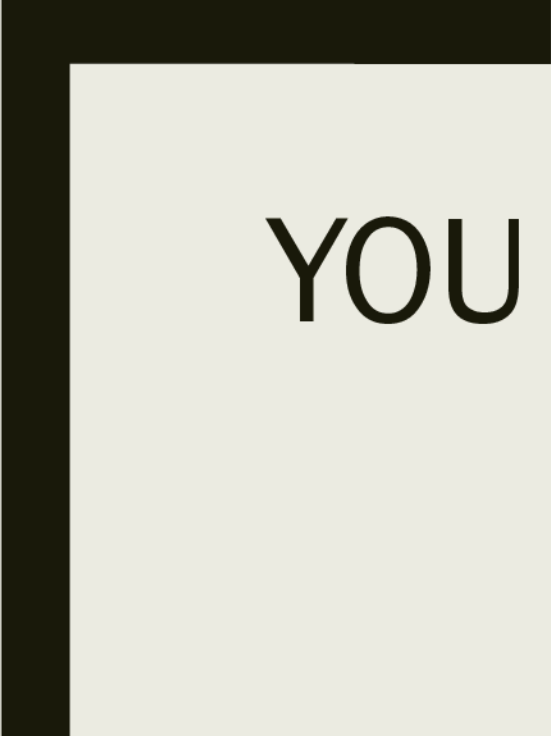
CONCLUSION

Conclusion

- While a lot of experimentation is well-known, the details of implementation are difficult

- Concepts like peeking, estimating lift and interference are "solved" academically

- Putting that solution into practice is incredibly difficult which is why understanding them = getting good jobs :)

YOU AND YOUR JOB SEARCH

JOB SEARCHES ARE TERRIBLE
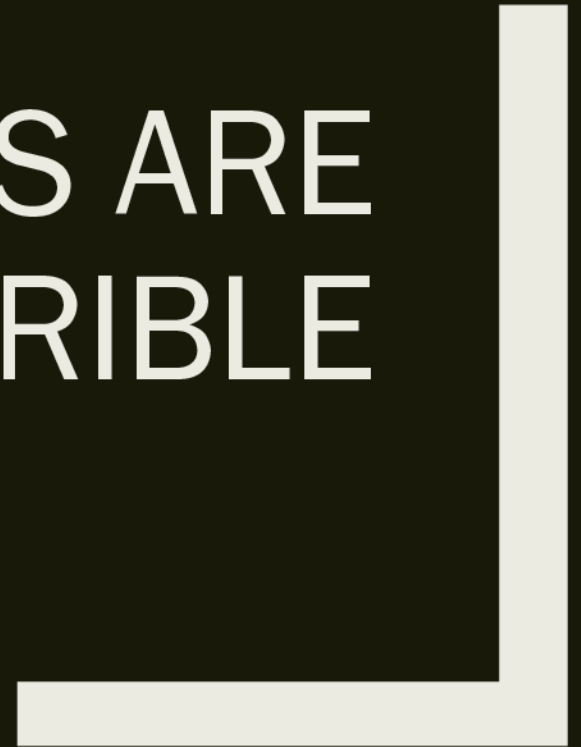
Why is getting a job difficult?

- Getting a job is a <u>bilateral</u> "asymmetric information <u>problem</u>"
- What's that mean?
  - *<u>Bilateral</u>: two-parties (you and the company)*
  - *<u>Asymmetric information</u>: each of you know something, but you can't easily and credibly disclose it*
    - "I'm a hard worker"
  - *<u>Problem</u>*
    - Yea.. It is a problem
- In this environment: Matches are difficult and sub-optimal

What do you (job seeker) see?

- You apply to hundreds of companies
  - *Your friend applies to 1 job and gets it*
- You don't hear back
- Why not? A few options:
  - *Underqualified*
  - *Bad Resume*
  - *Bad Cover Letter*
  - *Bad LinkedIn/Online Presence*
  - *Bad IRL Presence*
  - *Bad Luck*
- How do you know which one is you?
  - *Hint: You won't*

What does the company see?

- A job posting is created which contains a bunch of information about the job

- The job posting has requirements (some firmer than others)

- Applications come in, hundreds at a time

- All are nearly indistinguishable: *Everyone is a talented, hard-working, unique and exceptionally smart person.*

- Phone interview tons of people. Most fail there. Eventually (hopefully) one get hired.

- For companies this is incredibly frustrating.

End up with a dirty, imperfect match

- Because of asymmetric information:
  - *Many people apply to jobs they have no shot at*
  - *Companies filter out good candidates all the time*
- The final match is not optimal and nearly random.
- For you, this feels **terrible**. All this energy, effort and work for nothing.
- Moral:

DON'T LET THE JOB SEARCH GET YOU DOWN. IT ISN'T A MIRROR.
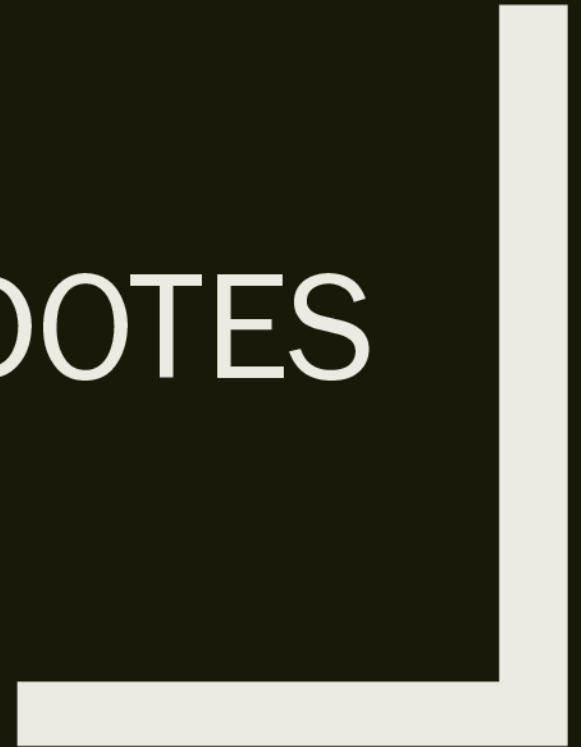
Anecdote #1: Hiring a Data Analyst at Sega

- I put up a job posting for one week.
- The job posting said:
  - *Cover letter required*
  - *SQL knowledge preferred*
  - *Lots and lots of other things*
- 500 applications in 5 days
- Solution?

Anecdote #2

- Senior job search, but still pretty common
- Will they hire from MSU?
- How many jobs are *like this*, but don't say it?

We are recruiting strong experienced engineers to build our next-gen Big Data predictive analytics platform & products with a focus on B2B marketing/sales. In particular, we're recruiting a strong Devops engineer to join a talented team with background from top schools (such as Harvard, MIT, Stanford, Tsinghua, UC Berkeley, U Penn, etc) & top software/tech companies & successful startups (such as Amazon, Apple, Google, Salesforce, ServiceNow, VMware, Zynga, DemandTec, ProfitLogic, etc).

Required qualifications:

Anecdote #3: USF's MSDS Program

- Entry level data science positions
- Statistics:
  - *Average number applied: ~85*
  - *High: ~400*
  - *Low: 3*
- Zero correlation with GPA

Anecdote Take-aways

- You aren't special (to the company)
- You are at disadvantage (not MIT)
- It can take a ton of time, energy and effort with no feedback

Being organized (really) helps...

- There are only **five** things you can do to get a job:
  1. *Improve your docs (Resume, LinkedIn, Cover Letter)*
  2. *Network*
  3. *Interview Prep*
  4. *Market yourself*
  5. *Apply to jobs*

- Having a plan & being efficient = less energy and effort

1. Optimize your actions

- Everyone (including me) tends to only work on some of those 5 things.

- "Diminishing Marginal Returns"

- If you are a introvert-grammar fiend (like me), your docs are probably already OK, but you've spent no time networking.

- Make sure that you are doing all 5 activities.

2. Be Organized / Always Be Learnin'

- Every time you apply / interview / get rejected:
  - *Document it.*

- What application paths are working for you? What types of companies are responding?

- Every question you get – did you answer it "the best"?

- What could you have done better?

3. Own your image/narrative

- Everything on your resume / LinkedIn:
  - *Why did you do it?*
  - *What did you do?*
  - *Why did you leave?*

- Have Intentionality for everything.

- DO NOT BE DESPERATE.
  - *"Actively looking for a job"*
  - This => TRASH IT.

4. Logistics of the interview

- In person: be ready for anything.
  - *Dress*
  - *Food, Water, Coffee (Bring your own)*
  - *Copies of your resume*
- Zoom: be ready for anything.
  - *Dress*
  - *Food, Water, Coffee*
  - *Battery*
  - *Quiet Place & Background*
  - *Headphones*

THE END!