Class 9: Experiments

BUS 696

Prof. Jonathan Hersh

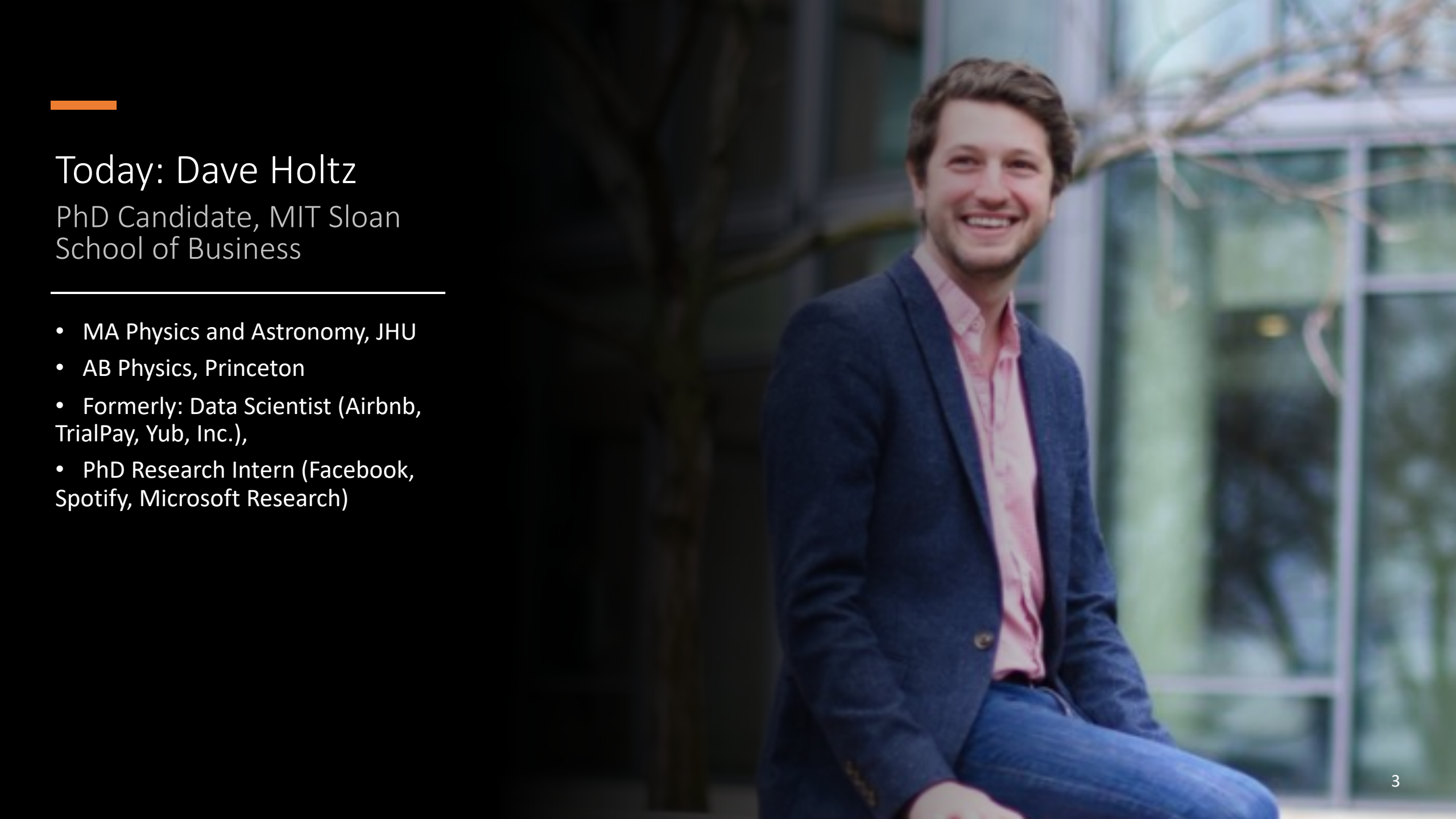
Class 9: Announcements

1. Midterms hopefully graded by next week
2. Pset 4 (Case study improving worker safety with ML) due Nov 18. Will post later this week.
3. Final Projects Description
4. Today: Interview Dave Holtz
5. Next week: read Uber Case study

Today: Dave Holtz

PhD Candidate, MIT Sloan School of Business

- MA Physics and Astronomy, JHU
- AB Physics, Princeton
- Formerly: Data Scientist (Airbnb, TrialPay, Yub, Inc.),
- PhD Research Intern (Facebook, Spotify, Microsoft Research)

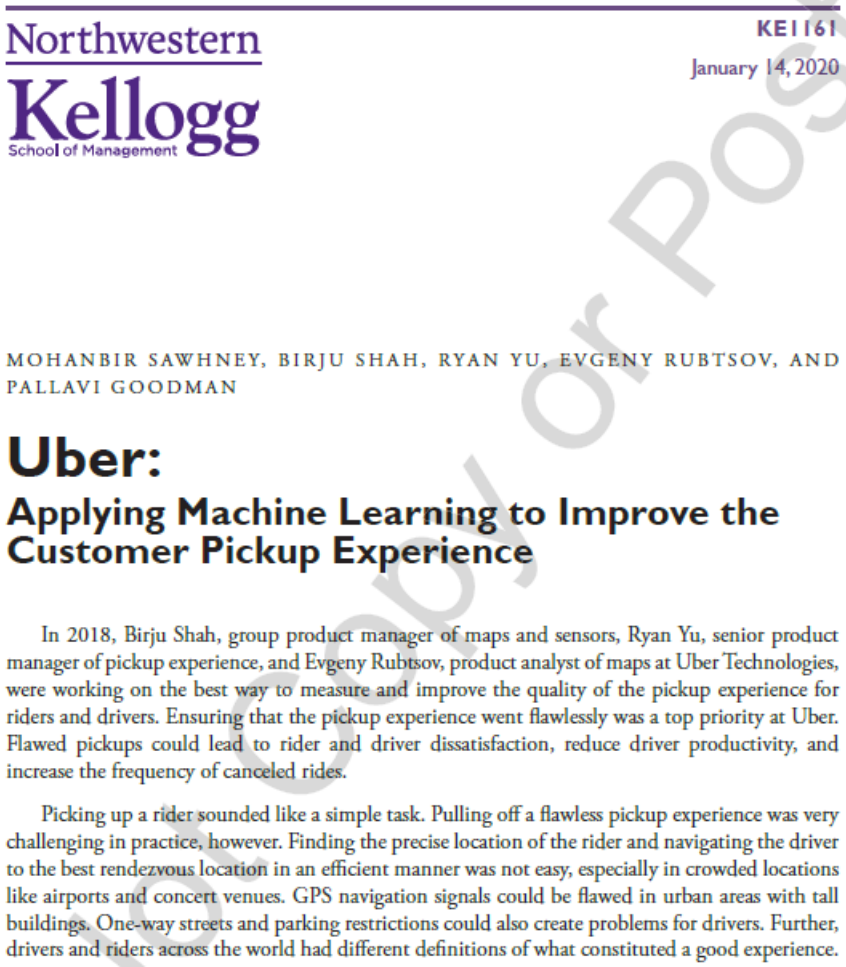


Next Week: Ana Rocca, PhD
Formerly: Head of Central and US&C Research and Insights, Uber

- PhD, Economics UC Berkeley
- AB Economics, University of Chicago

- Will discuss case study “Applying Machine Learning to Improve the Customer Pickup Experience.”

Case Study for Next Week



Northwestern Kellogg School of Management

KE1161

January 14, 2020

MOHANBIR SAWHNEY, BIRJU SHAH, RYAN YU, EVGENY RUBTSOV, AND PALLAVI GOODMAN

Uber:

Applying Machine Learning to Improve the Customer Pickup Experience

In 2018, Birju Shah, group product manager of maps and sensors, Ryan Yu, senior product manager of pickup experience, and Evgeny Rubtsov, product analyst of maps at Uber Technologies, were working on the best way to measure and improve the quality of the pickup experience for riders and drivers. Ensuring that the pickup experience went flawlessly was a top priority at Uber. Flawed pickups could lead to rider and driver dissatisfaction, reduce driver productivity, and increase the frequency of canceled rides.

Picking up a rider sounded like a simple task. Pulling off a flawless pickup experience was very challenging in practice, however. Finding the precise location of the rider and navigating the driver to the best rendezvous location in an efficient manner was not easy, especially in crowded locations like airports and concert venues. GPS navigation signals could be flawed in urban areas with tall buildings. One-way streets and parking restrictions could also create problems for drivers. Further, drivers and riders across the world had different definitions of what constituted a good experience.

- Chapman has purchased this case study for you!
- Please use the following link to download the case study and read for next week to prepare for interview

https://www.thecasecentre.org/educators/courses?id=1356705&pdid=171935&opid=855816

Data Analytics Association Guest Speaker

Annie Wang

- JD Candidate, Yale

- Formerly:
- Director of Data Science, Warren for President
- Director of Research and Analytics, Analyst Institute
- Civis Analytics, Senior Applied Data Scientist

- **Tuesday, Nov 10 @ 7pm**

Class 9: Outline

1. Interview with Dave Holtz
2. A/B Testing for Experiments
3. Potential Outcome Framework
4. A/B Testing in R
5. Lab on A/B Testing (Time Permitting)

Final Project Overview

- **Your final project should take a real-world data set, and estimate a series of predictive models against this dataset.**
- You should identify a business use-case for the prediction, and estimate at least three predictive models we covered in this class against the dataset.

November 18th – Due: students must upload to Canvas a one-page outline of their project

a) identify a dataset you will use
b) the outcome you are trying to predict, and what variables you will use to predict it
c) motivation to your project -- as in the business or practical management use case of such a prediction
d) the names of the students who will be part of your group (up to three per group)

Where to Find Datasets?

- Kaggle: https://www.kaggle.com/datasets
- Kaggle: https://www.kaggle.com/annavictoria/ml-friendly-public-datasets
- FiveThirtyEight https://data.fivethirtyeight.com/
- TidyTuesday: https://github.com/rfordatascience/tidytuesday
- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.php

How to Find Dataset and Join Groups?

- Think about the topic/industry you want to cover in your final project
  - Finance?
  - Sports Analytics?
  - Entertainment?
  - Marketing Analytics?
  - Personnel Analytics?
  - Other?

- Spend 5-10 minutes in a breakout rooms
- (Upgrade to latest zoom and can self-join different rooms based on topics)
- Find a group of up to three students
- Brainstorm ideas for datasets (see links at last slide)

Other Final Project Deadlines

December 2nd – Due: students must upload to Blackboard a compiled Rmarkdown document (and code and dataset that generates it) that includes example summary statistics from the dataset to be studied. This should include a summary table of means, max, mins and standard deviations; data transformations performed for feature engineering, as well as at least five plots revealing interesting patterns to be studied. These must be motivated by the analysis – they cannot just be random plots.

December 9th – Due: a 15-20 minute in-class presentation. Your presentation should start with the business use and motivation for the prediction task. Sell the class on the idea for why we must implement a given machine learning prediction solution. Then describe in details the data cleaning, feature creation/engineering, modeling stages. Finally, discuss the model performance and consider the adoption decision of the particular machine learning solution.

December 16th: Final project code, replication files, dataset, and revisions to presentation (if necessary).

Class 9: Outline

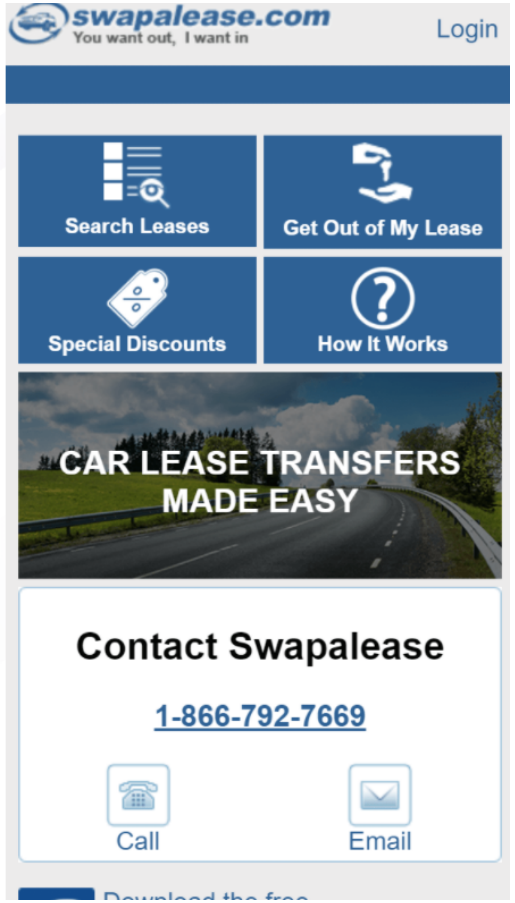
1. Interview with Dave Holtz
2. A/B Testing for Experiments
3. Potential Outcome Framework
4. A/B Testing in R
5. A/B Lab (Time Permitting)

What is A/B Testing

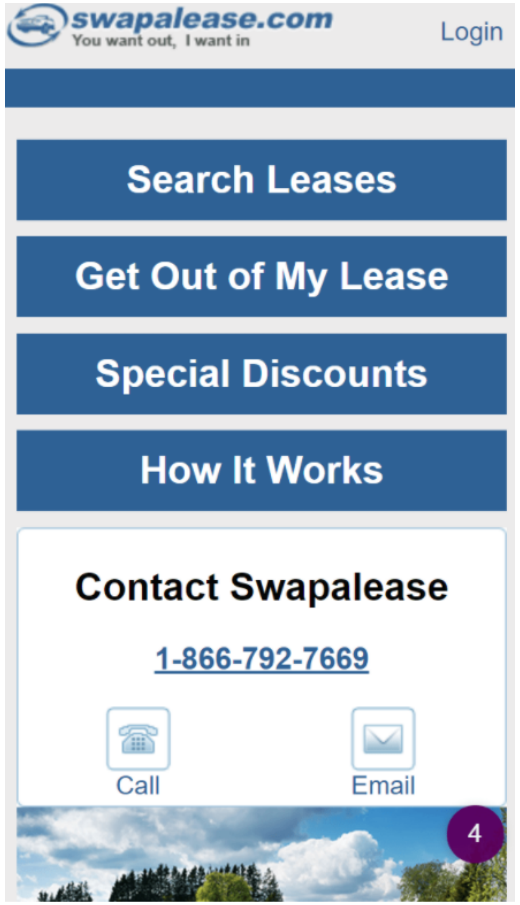
- An A/B test is a comparison of two versions of the same asset (e.g. website or email) that you expose randomly to some set of your audience
- Comparing the results on the performance metrics of interest (customer conversions, revenue, time spent on site) reveals which version best meets our goals.

Example A/B Test: Mobile App

Which do you prefer?



Version A



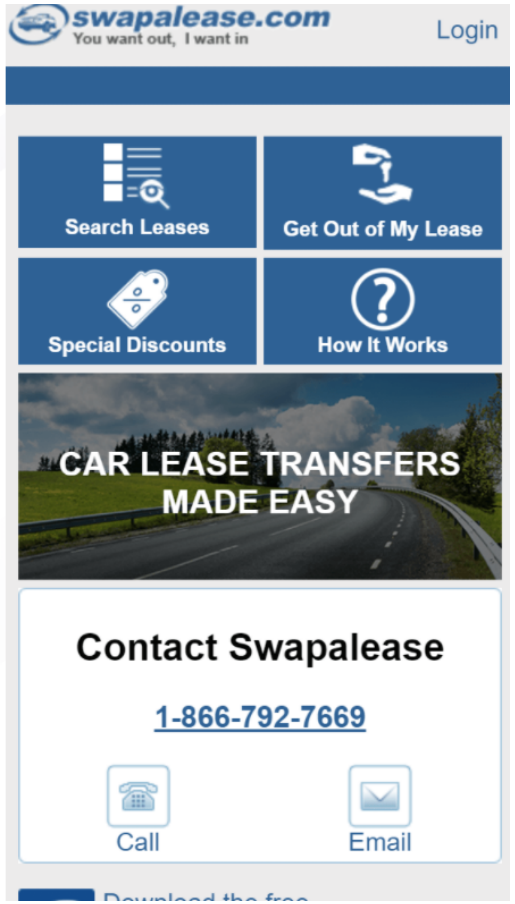
Version B

From: https://www.crazyegg.com/blog/ab-testing-examples/

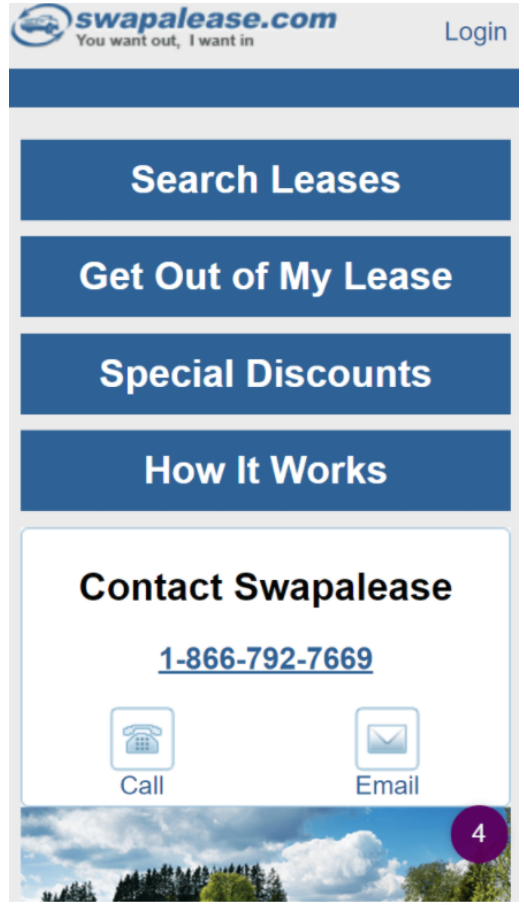
Example A/B Test: Mobile App Swapalease.com

Which do you prefer?

- Version A winner in sample of N = 4,406 mobile users.
- Version A led to a 30% increase in lead submissions



Version A



Version B

Example A/B Test: Electronic Arts Landing Page

Which do you prefer?

Version A

Version B

Example A/B Test: Electronic Arts Landing Page

Which do you prefer?

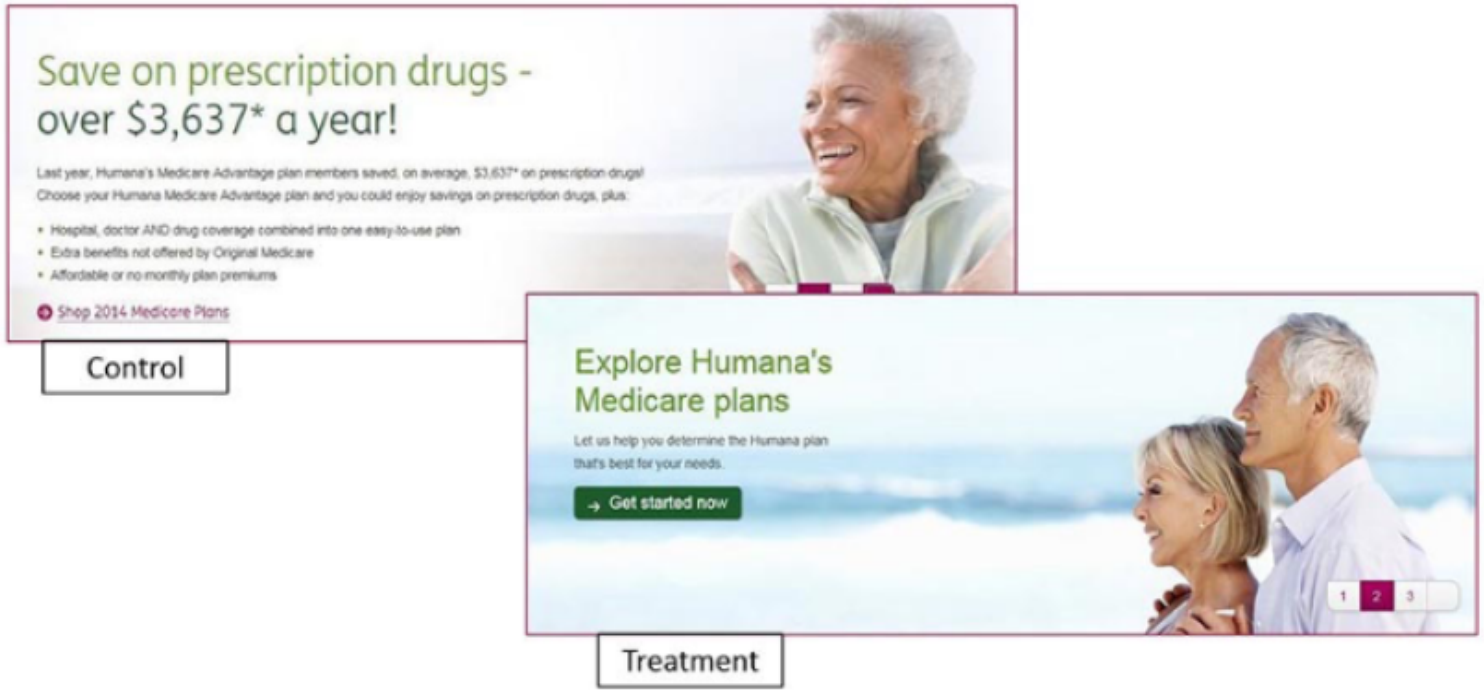
- Version B => 40% more orders

Version A

Version B

Example A/B Test: Humana Website Banner

Which had a better click through rate (CTR)?

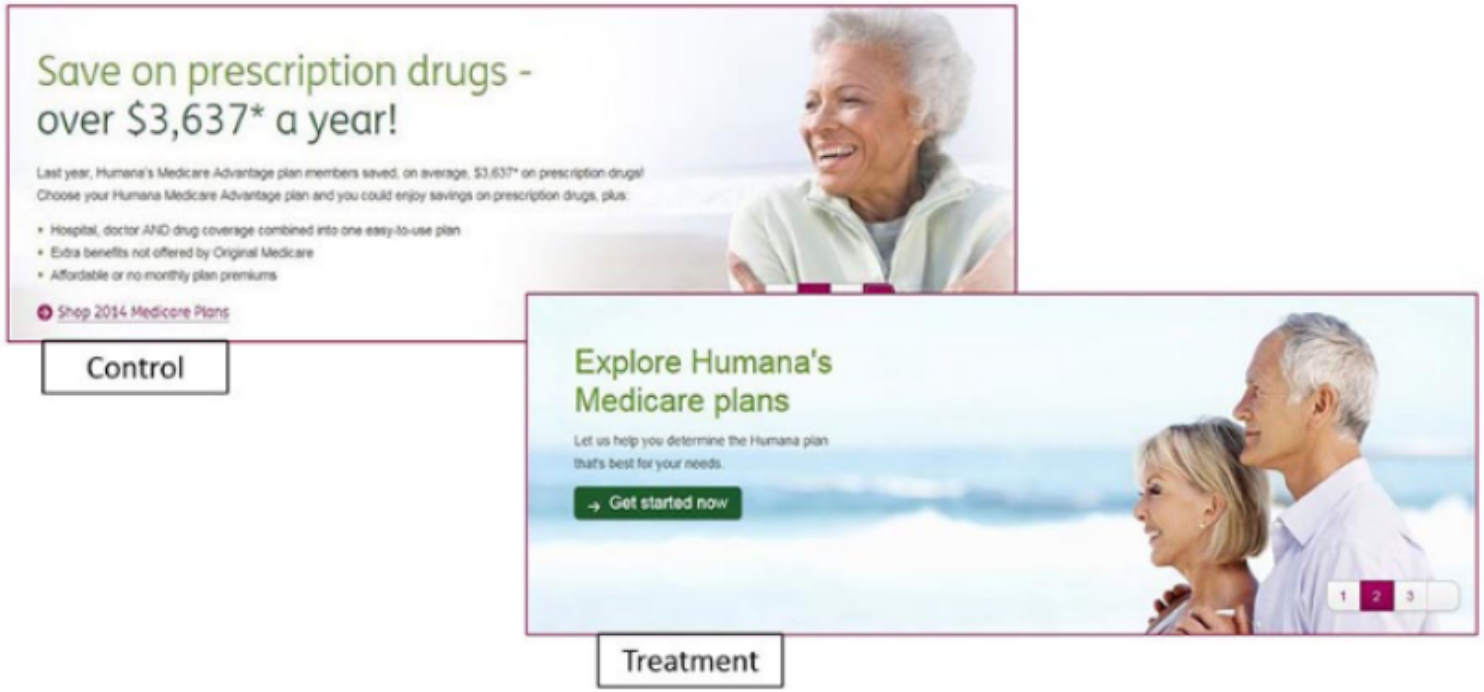


Control **Treatment**

Example A/B Test: Humana Website Banner

Which had a better click through rate (CTR)?

- Treatment: 433% increase in CTR!



Control **Treatment**

Potential Outcome Framework

$$\underbrace{ATE}_{avg\ treatent\ effect} = \underbrace{E[y|d=1]}_{avg\ for\ treated} - \underbrace{E[y|d=0]}_{avg\ for\ untreated}$$

$$\widehat{ATE} = \frac{1}{n_1}\sum_{d=1} y_i - \frac{1}{n_0}\sum_{d=0} y_i$$

$$\widehat{ATE} = \bar{y}_1 - \bar{y}_0$$

- y: outcome
- d: = 1 (treated), 0 (untreated)
- $E[\quad]$: average or expectation

Estimating ATE Via Regression

$$\underbrace{E[y|x,d]}_{avg\ treatent\ effect} = \underbrace{\alpha_d}_{=1\ if\ treated} + \epsilon$$

- It turns out an equivalent way of estimating treatment effect is via regression
- We regress the outcome (y) against a treatment indicator (α_d)
- The coefficient on α_d is our estimated treatment effect
- We estimate y continuous with OLS (linear regression) and y binary with logistic regression

ATE with Regression Adjustments

$$\underbrace{E[y|x,d]}_{avg\ treatent\ effect} = \underbrace{\alpha_d}_{=1\ if\ treated} - \underbrace{X'\beta_d}_{\boldsymbol{coefficients\ for\ group\ d}}$$

- Suppose we are concerned that treated groups are systematically different from untreated groups
- We can include observational controls that increase efficiency of estimator
- Standard errors and p-values are usually correct!
- Can “cluster” at some level (day, geography) to adjust standard errors for specific effects

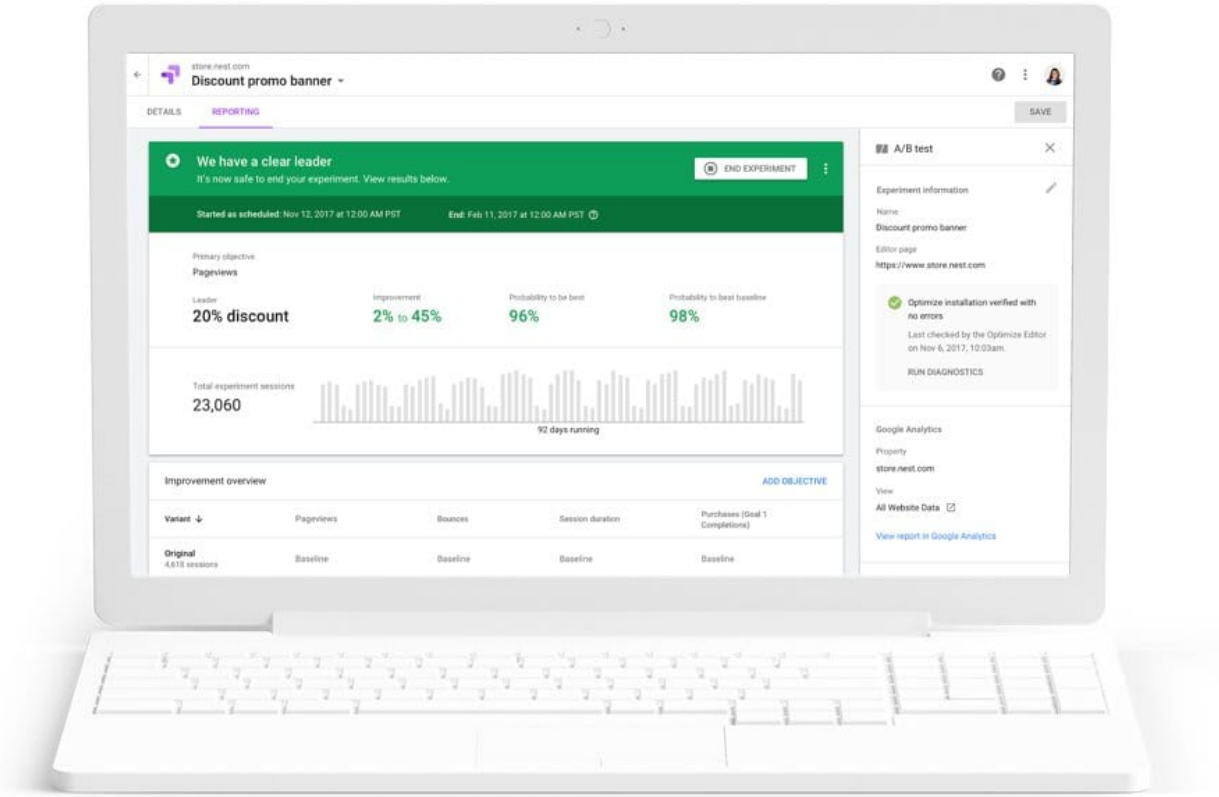
Many A/B Platforms Available

Convertize
SMART PERSUASION

Google Optimize

Optimizely

AB Tasty



A/B Treatment Effect Calculation in R

```
#-------------------------------------------------------------
# Load data and packages
#-------------------------------------------------------------
library('tidyverse')
library('lubridate')

AB <- read.csv(here::here("datasets","ab_data.csv"))

head(AB)
```

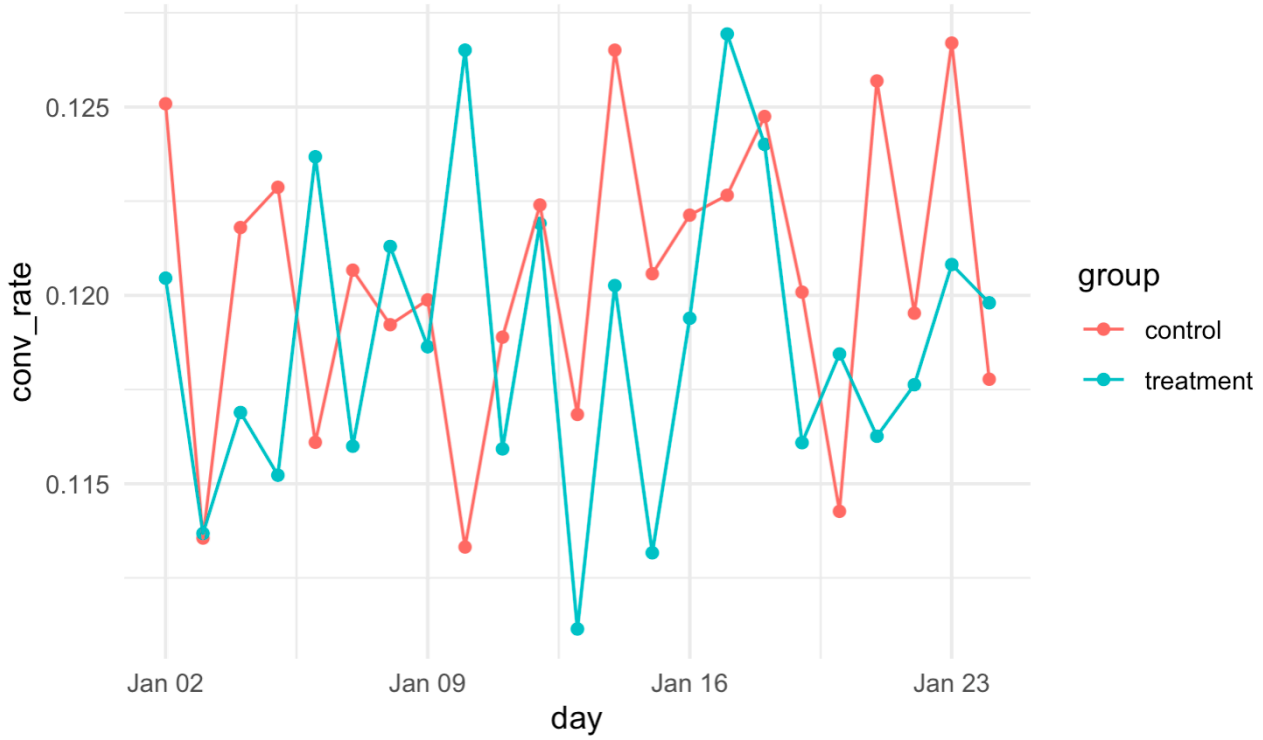
```
> AB <- read.csv(here::here("datasets","ab_data.csv"))
> head(AB)
  user_id                   timestamp     group landing_page converted
1  851104 2017-01-21 22:11:48.556739   control     old_page         0
2  804228 2017-01-12 08:01:45.159739   control     old_page         0
3  661590 2017-01-11 16:55:06.154213 treatment     new_page         0
4  853541 2017-01-08 18:28:03.143765 treatment     new_page         0
5  864975 2017-01-21 01:52:26.210827   control     old_page         1
6  936923 2017-01-10 15:20:49.083499   control     old_page         0
>
```

- Example AB dataset tests impact of new vs old landing page on website
- Dataset has user ID, timestamp (when visited site), group (treatment or control), and whether user converted to sale (1) or not(0).

Graphing and Viewing Conversion Rate

```
# group by user_id, arrange by timestamp, and remove
# duplicate IDs
# create a day date variable from the raw
#  timestamp information
AB_clean <- AB %>% group_by(user_id) %>%
  arrange(timestamp) %>%
  filter(!duplicated(user_id)) %>%
  mutate(day = as.Date(timestamp))

# and group by date and control/treatment indicator
AB_sum <- AB_clean %>%
  group_by(day, group) %>%
  summarize(conv_rate = mean(converted))
```



Estimating Logistic Model

$$y_i = logistic(d_i) + \epsilon_i$$

```
#-----------------------------------------------
# Estimate model to identify treatment effect
#-----------------------------------------------
AB_clean <- AB_clean %>%
  mutate(treated = factor(group,
         levels = c("control","treatment")))

logit_mod <- glm(converted ~ treated,
                 data = AB_clean,
                 family = "binomial")

summary(logit_mod)
```

```
Call:
glm(formula = converted ~ treated, family = "binomial", data = AB_clean)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5066  -0.5066  -0.5030  -0.5030   2.0640

Coefficients:
                  Estimate Std. Error  z value Pr(>|z|)
(Intercept)      -1.988479   0.008061 -246.679   <2e-16 ***
treatedtreatment -0.015066   0.011433   -1.318    0.188
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 212810  on 290583  degrees of freedom
Residual deviance: 212808  on 290582  degrees of freedom
AIC: 212812
```

Estimating Logistic Model With Regression Adjustment

```
#-------------------------------------------------------
# Cluster standard errors around day
#  and control for covariates (day effect)
#-------------------------------------------------------
library('miceadds')
library('sandwich')
logit_cluster <- glm.cluster(converted ~
                             treated + factor(day),
                       cluster = "day",
                       data = AB_clean)

summary(logit_cluster)
```

```
> summary(logit_cluster)
                              Estimate   Std. Error      t value  Pr(>|t|)
(Intercept)              0.1235666378 5.783863e-04    213.64033 0.0000000
treatedtreatment        -0.0015873542 1.157774e-03     -1.37104 0.1703624
factor(day)2017-01-03   -0.0091519244 1.712111e-06  -5345.40439 0.0000000
factor(day)2017-01-04   -0.0034235089 7.660123e-07  -4469.26122 0.0000000
factor(day)2017-01-05   -0.0037368772 2.714732e-06  -1376.51808 0.0000000
factor(day)2017-01-06   -0.0028324400 6.948823e-06   -407.61434 0.0000000
factor(day)2017-01-07   -0.0044433384 1.631996e-06  -2722.64117 0.0000000
factor(day)2017-01-08   -0.0025103416 1.574190e-06  -1594.68770 0.0000000
factor(day)2017-01-09   -0.0035148371 2.398258e-07 -14655.79164 0.0000000
factor(day)2017-01-10   -0.0028462845 1.450919e-06  -1961.71206 0.0000000
factor(day)2017-01-11   -0.0053639115 2.767538e-07 -19381.53058 0.0000000
factor(day)2017-01-12   -0.0006138833 4.895336e-06   -125.40167 0.0000000
factor(day)2017-01-13   -0.0087725379 1.759904e-06  -4984.66960 0.0000000
factor(day)2017-01-14    0.0006031631 2.659382e-06    226.80574 0.0000000
factor(day)2017-01-15   -0.0058686549 6.236839e-06   -940.96623 0.0000000
factor(day)2017-01-16   -0.0020096171 2.464059e-06   -815.57199 0.0000000
factor(day)2017-01-17    0.0020138495 1.624173e-06   1239.92322 0.0000000
factor(day)2017-01-18    0.0016087557 5.083899e-06    316.44133 0.0000000
factor(day)2017-01-19   -0.0046815959 1.365150e-06  -3429.36390 0.0000000
factor(day)2017-01-20   -0.0063813356 7.564410e-06   -843.59993 0.0000000
factor(day)2017-01-21   -0.0017381753 8.148768e-06   -213.30527 0.0000000
factor(day)2017-01-22   -0.0041958080 3.640622e-06  -1152.49752 0.0000000
factor(day)2017-01-23    0.0009983169 3.145816e-06    317.34751 0.0000000
factor(day)2017-01-24   -0.0040069530 6.430382e-06   -623.12828 0.0000000
```

- Appears to be no significant difference between new and old landing page
- ATE = 0

New A/B Data: Mobile App Cookie Cats

- The mobile game app is testing out a new procedure to increase paid conversions
- After letting players play for free for some rounds, customers face a “gate” or pause screen that lets them make an in app purchase
- This gate was at level 30, but they are testing whether having the gate at level 40 increases player usage and retention.

```r
#-------------------------------------------------------------
# Read in Cookie Cats Data
#-------------------------------------------------------------

cats <- read.csv(here::here("datasets","cookie_cats.csv"))
head(cats)
# userid - unique player ID
# version - whether the player was put in the control group
# (gate_30 - a gate at level 30)
#  or the test group (gate_40 - a gate at level 40).
# sum_game rounds - the number of game rounds played by the player
#    during the first week after installation
# retention_1 - did the player come back and play 1 day after installing?
# retention_7 - did the player come back and play 7 days after installing?

# 1. Arrange by userid and remove any duplicate IDs that exist
#    Use the mutate function to create a "treated" variable = 1
#    if player was in the treated category where the gate was set to level 40
```

Lab Time! (Time Permitting)

```
# 1. Arrange by userid and remove any duplicate IDs that exist
#    Use the mutate function to create a "treated" variable = 1
#    if player was in the treated category where the gate was set to level 40

cats_clean <- cats %>%
  arrange(userid) %>%
  filter(!duplicated(userid)) %>%
  mutate(treated = if_else(version == "gate_40",1,0),
         ret_1 = if_else(retention_1 == "TRUE",1,0),
         ret_7 = if_else(retention_7 == "TRUE",1,0))

# 2. Estimate the treatment impact of the gate 40
#    intervention on day 1 retention

# 3. Estimate the treatment impact of the gate 40
#    intervention on day 7 retention

# 4. Estimate the treatment impact of the gate 40
#    intervention on game rounds played

# 5. What do you conclude? Should they adopt the treatment?
#    Why or why not?
```