STA 235H - Randomized Controlled Trials II

Fall 2021

McCombs School of Business, UT Austin

Some announcements

Homework 2 is due on Thursday

- Please remember to (1) submit on time or (2) fill out the extension form
  - Reach out to the instruction team if you have any issues.
- Following the analytics for the course website, less than 50% of students check out the R code:
  - Review the scripts and answer the questions there! (good practice for the midterm, too).

Last week

- Finished our chapter on causal inference framework.
- Started talking about Randomized Controlled Trials:

Assumptions: The power of randomization

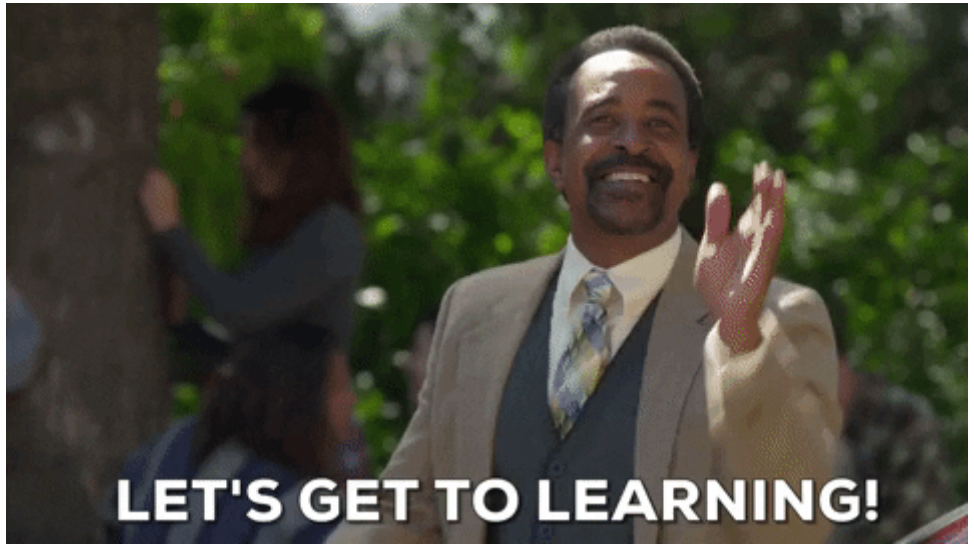
Today

- Continue with randomized controlled trials:

Design: What should we consider?

Limitations: Gold Standard?

- Comparison between observational studies and RCTs



Design: How do we randomize?

How do we randomize?

Coin flip*

E.g., in R:

```
id <- seq(1,1000)

set.seed(100)

data <- data.frame("id" = id) %>% mutate(z = sample(c(0,1), size = 1000,
                                                    replace = TRUE,
                                                    prob = c(0.5, 0.5)))
```

How do we randomize?

Coin flip*

E.g., in R:

```
id <- seq(1,1000)

set.seed(100)

data <- data.frame("id" = id) %>% mutate(z = sample(c(0,1), size = 1000,
                                                    replace = TRUE,
                                                    prob = c(0.5, 0.5)))
```

How do we randomize?

Coin flip*

E.g., in R:

```
id <- seq(1,1000)

set.seed(100)

data <- data.frame("id" = id) %>% mutate(z = sample(c(0,1), size = 1000,
                                                    replace = TRUE,
                                                    prob = c(0.5, 0.5)))
```

How do we randomize?

Coin flip*

E.g., in R:

```
id <- seq(1,1000)

set.seed(100)

data <- data.frame("id" = id) %>% mutate(z = sample(c(0,1), size = 1000,
                                                    replace = TRUE,
                                                    prob = c(0.5, 0.5)))

data %>% group_by(z) %>% count()
```

How do we randomize?

Coin flip*

E.g., in R:

```
id <- seq(1,1000)

set.seed(100)

data <- data.frame("id" = id) %>% mutate(z = sample(c(0,1), size = 1000,
                                                    replace = TRUE,
                                                    prob = c(0.5, 0.5)))

data %>% group_by(z) %>% count()
```

```
## # A tibble: 2 x 2
## # Groups:   z [2]
##       z     n
##   <dbl> <int>
## 1     0   516
## 2     1   484
```

How do we check that randomization was done correctly?

Checking for balance

```
library(modelsummary)

d <- read.csv("https://raw.githubusercontent.com/maibennett/sta235/main/exampleSite/content/Classes/Week4/2_RCT/data/covar

head (round(d,3))
```

```
##   id z     x1     x2     x3     x4     x5     x6     x7     x8     x9    x10
## 1  1 1 -0.626 -0.897 -0.962  0.217 -0.841  0.270  2.287 -0.085 -0.767  0.019
## 2  2 1  0.184  0.185 -0.293 -0.542  1.384 -0.630 -1.197  0.840 -0.816 -0.184
## 3  3 0 -0.836  1.588  0.259  0.891 -1.255  0.869 -0.694 -0.463 -0.142 -1.371
## 4  4 1  1.595 -1.130 -1.152  0.596  0.070  1.727 -0.412 -0.551 -0.278 -0.599
## 5  5 0  0.330 -0.080  0.196  1.636  1.711  0.024 -0.971  0.736  0.436  0.295
## 6  6 0 -0.820  0.132  0.030  0.689 -0.603  0.368 -0.947 -0.108 -1.187  0.390
##      x11    x12    x13    x14    x15    x16    x17    x18    x19    x20
## 1 -0.591 -1.481  0.554 -0.662  0.259  0.476 -1.015  0.926 -1.189  1.163
## 2  0.027  1.577 -0.280  1.719  1.831 -0.125 -0.080  1.823  0.389 -0.586
## 3 -1.517 -0.957  1.775  2.122 -0.340  1.096 -0.233 -1.611 -0.344  1.785
## 4 -1.363 -0.920  0.187  1.497  0.897 -1.444 -0.817 -0.285 -0.548 -1.333
## 5  1.178 -1.998  1.143 -0.036  0.488  1.148  0.772 -0.342  0.981 -0.447
## 6 -0.934 -0.272  0.416  1.232 -1.255 -0.468 -0.166  0.366 -0.237  0.570
```

Checking for balance

```
d_bal <- d %>% select(z, starts_with("x"))

datasummary_balance(~ z, data = d_bal, fmt = 2, dinm_stat
```

	0		1			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p
x1	-0.07	1.04	0.04	1.03	0.11	0.09
x2	0.06	1.03	0.06	1.00	0.00	0.98
x3	0.01	0.99	0.01	1.00	0.00	0.99
x4	-0.03	1.00	-0.04	0.94	0.00	0.96
x5	0.03	1.01	0.01	1.01	-0.02	0.77
x6	-0.08	0.99	0.03	1.02	0.12	0.07
x7	0.03	0.98	-0.03	0.98	-0.06	0.31
x8	-0.02	1.00	-0.07	1.05	-0.05	0.45
x9	0.00	0.92	0.01	1.00	0.02	0.77

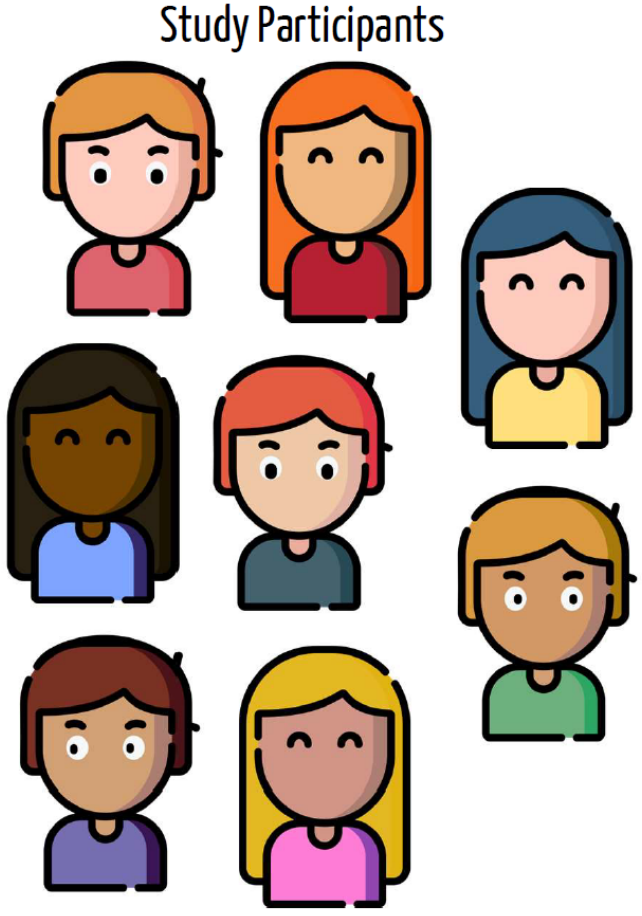
- Columns of interest:
  - Diff in Means: Difference in means between treat (col. 3) and control group (col. 1)
  - p: P-value for the difference in means (whether is statistically significant or not.)

Randomization assures balance *in expectation*

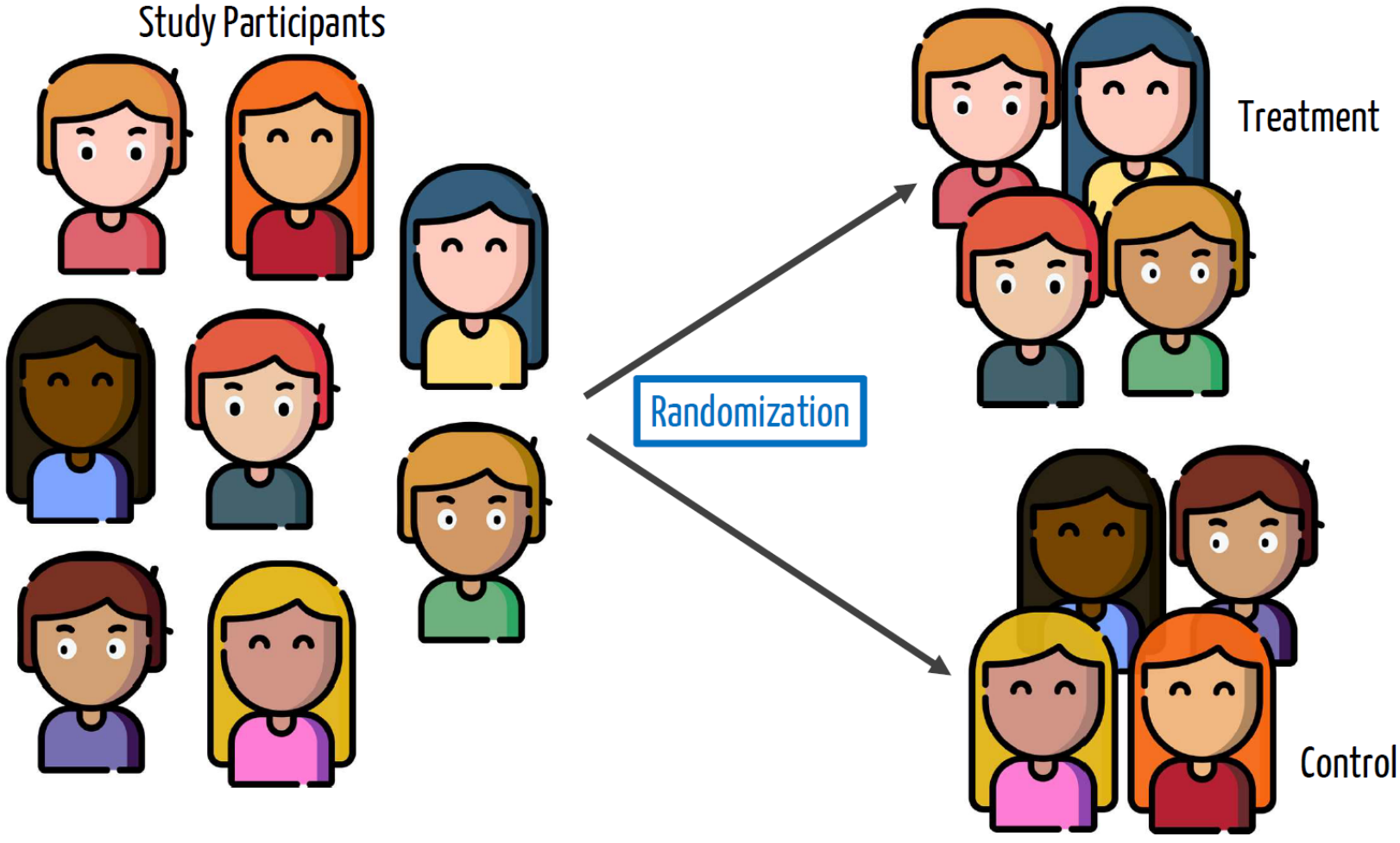
Can we ensure balance?

- In RCTs, in general, you can't ensure balance on all covariates.
- But you can **stratify**!
- Stratification means dividing your data into different stratas or groups S, based on one or more covariates.
- Then, you randomize in the same way *within strata*.

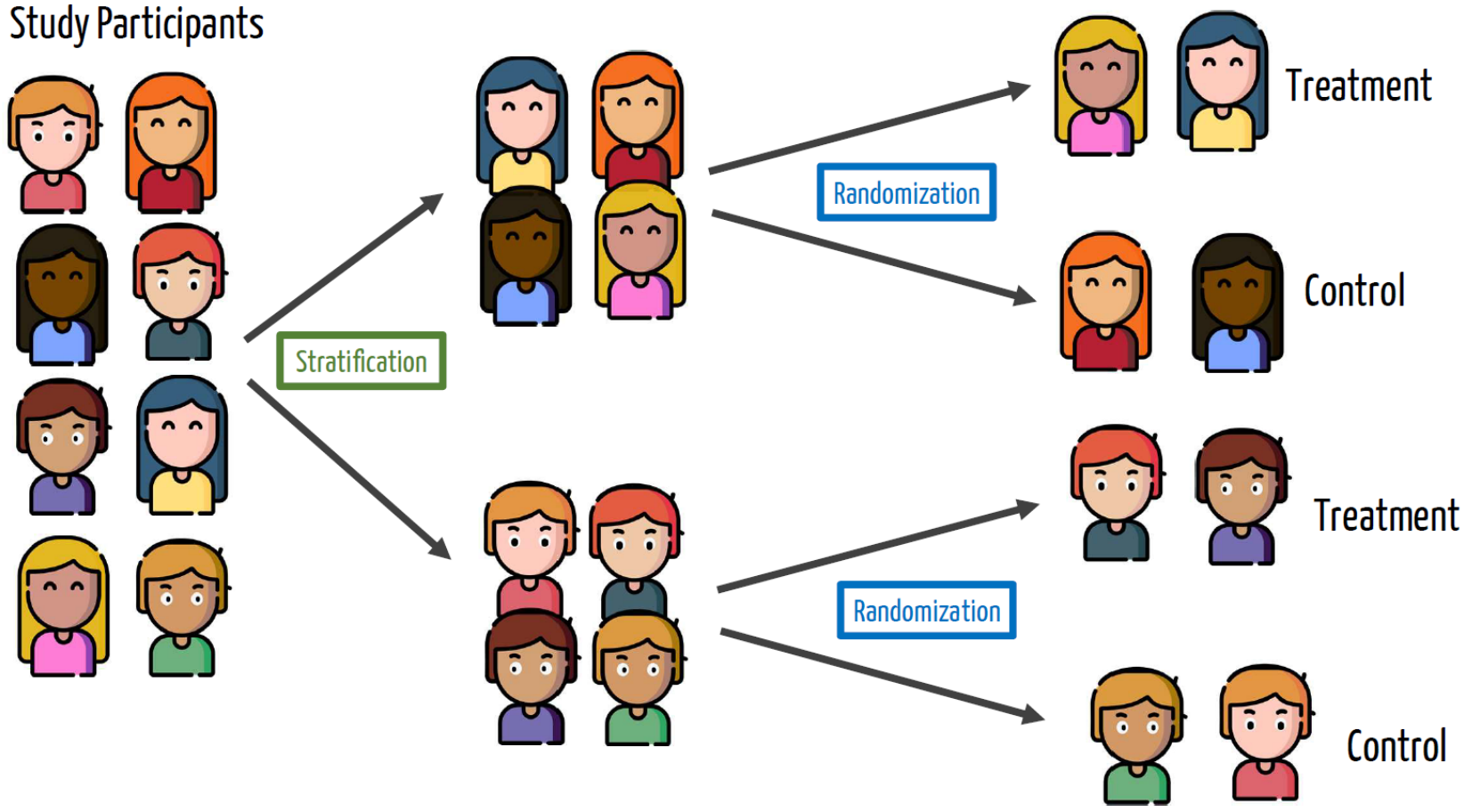
Randomization of treatment



Randomization of treatment with no strata



Randomization of treatment with strata



RCTs in Practice

Let's look at some data

- "Get out the Vote" Large-Scale Mobilization experiment (Arceneaux, Gerber, and Green, 2006)
    - "Households containing one or two registered voters where randomly assigned to treatment or control groups"
    - Treatment: GOTV phone calls
    - Stratified RCT: Two states divided into competitive and noncompetitive

Checking for balance

Balance Table by Stratum

	Non-competitive		Competitive		Non-competitive		Competitive	
	Treat	Control	Treat	Control	Treat	Control	Treat	Control
female2	0.552	0.546	0.541	0.535	0.549	0.545	0.543	0.541
fem_miss	0	0	0	0	0.026	0.025	0.022	0.021
age	52.157	51.977	50.81	50.862	55.795	55.782	53.481	53.464
newreg	0.117	0.116	0.133	0.134	0.048	0.049	0.048	0.046
persons	1.496	1.497	1.513	1.518	1.539	1.538	1.529	1.533
vote98	0.231	0.227	0.258	0.259	0.572	0.574	0.599	0.594
vote00	0.564	0.567	0.595	0.593	0.734	0.732	0.781	0.78

Estimating the effect

- Depending on the design, usually you can compare group means or fit a simple regression

$$\frac{1}{N_T}\sum_{i\in T} Y_i - \frac{1}{N_C}\sum_{i\in C} Y_i$$

$$Y_i = \beta_0 + \beta_1 Z_i + \varepsilon_i$$

How do we incorporate stratification here?

Estimating the effect

- In stratified RCTs, we need to consider the strata!
- Run a regression with *fixed effects by strata*

$$Y_i = \beta_0 + \beta_1 Z_i + \gamma_s + \varepsilon_i$$

Estimating the effect

```
library(estimatr)

d_s1 <- d_s1 %>% mutate(strata = interaction(state, competiv))

summary(lm_robust(vote02 ~ treat_real + strata, data = d_s1))
```

```
##
## Call:
## lm_robust(formula = vote02 ~ treat_real + strata, data = d_s1)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)  CI Lower CI Upper     DF
## (Intercept) 0.486984  0.0008837 551.048 0.000e+00  0.485252 0.488716 733335
## treat_real  0.000369  0.0027329   0.135 8.926e-01 -0.004987 0.005725 733335
## strata1.1   0.108725  0.0018181  59.800 0.000e+00  0.105161 0.112288 733335
## strata0.2   0.044896  0.0057059   7.868 3.602e-15  0.033712 0.056079 733335
## strata1.2   0.169422  0.0012373 136.931 0.000e+00  0.166997 0.171848 733335
##
## Multiple R-squared:  0.02537 ,    Adjusted R-squared:  0.02536
## F-statistic:  4777 on 4 and 733335 DF,  p-value: < 2.2e-16
```

Estimating the effect

- One important thing to note in the previous analysis is that **assignment to treatment $\neq$ contact**

```
d_s1 %>% count(treat_real, contact)
```

```
##   treat_real contact      n
## 1          0       0 698815
## 2          1       0  19210
## 3          1       1  15315
```

Does this affect the internal validity of the study?

When we assume...

Other assumptions

- We already talked about the ignorability assumption
- Stable Unit Treatment Value Assumption (SUTVA):
  - No interference
  - No hidden variations of treatments

SUTVA: No interference

- *"The treatment applied to one unit does not affect the outcome for other units"*
- Imagine we have two individuals, 1 and 2:
  - Z_1 and Z_2 will be the treatment assignment for 1 and 2, respectively.
  - Under SUTVA:

$$(Y_2(0), Y_2(1)) \perp Z_1$$

SUTVA: No hidden variations of treatments

- *"An individual receiving an specific treatment level cannot receive different forms of that treatment."*
- Think about the headache example from last class:
    - Individual 1 gets a new aspirin (aspirin +)
    - Individual 2 gets an old aspirin (aspirin -)
    - Individual 3 doesn't get a pill
- If we label the treatments as $Z = 0$ (doesn't get an aspirin) and $Z = 1$ (gets an aspirin), potential outcomes would change depending on whether they got *aspirin +* or *aspirin -*.

When SUTVA hits the fan

When do you think SUTVA could fail?

h/t to @paul_gp

When SUTVA hits the fan

Network effects

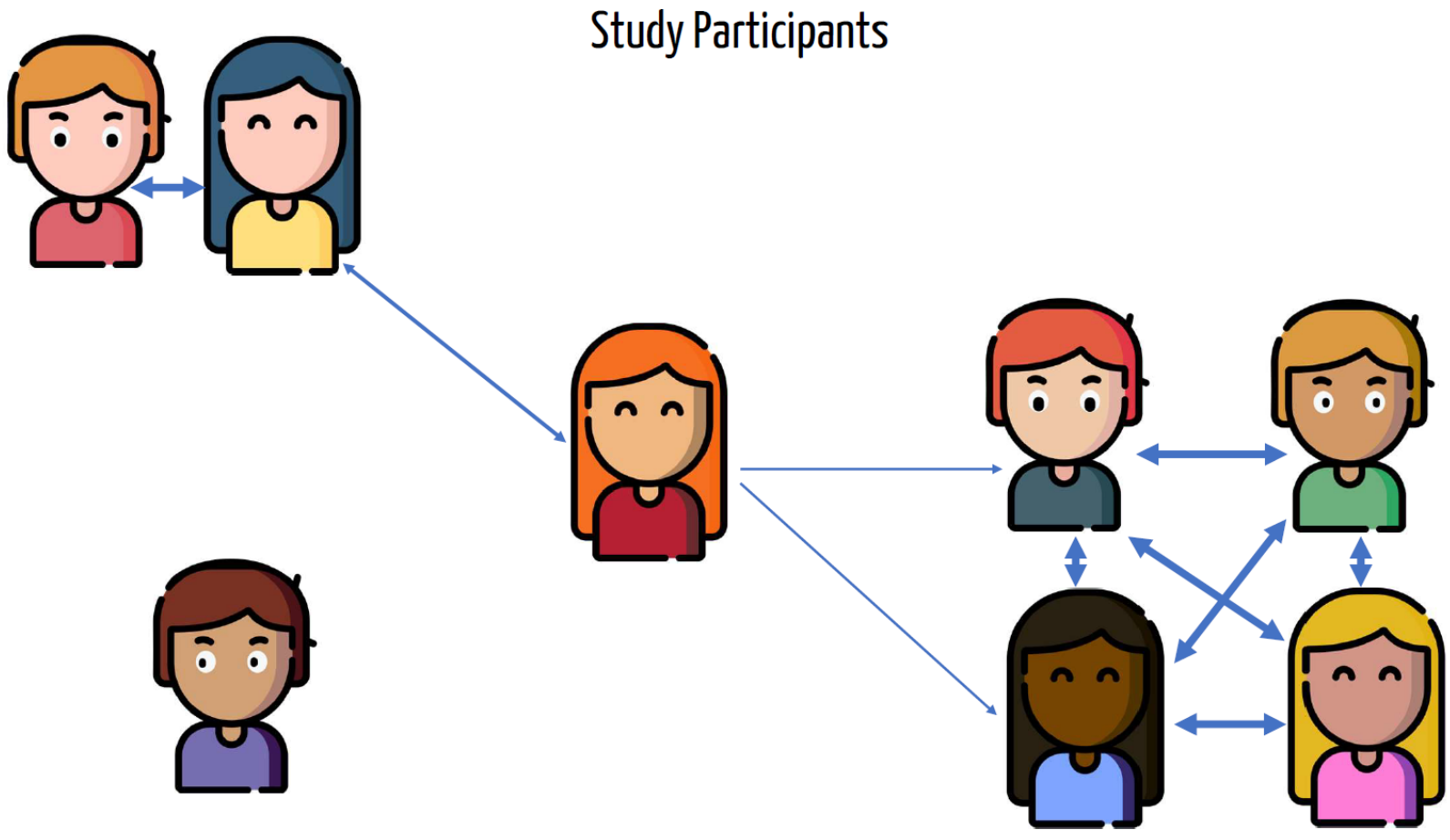
General Equilibrium Effects

Network effects

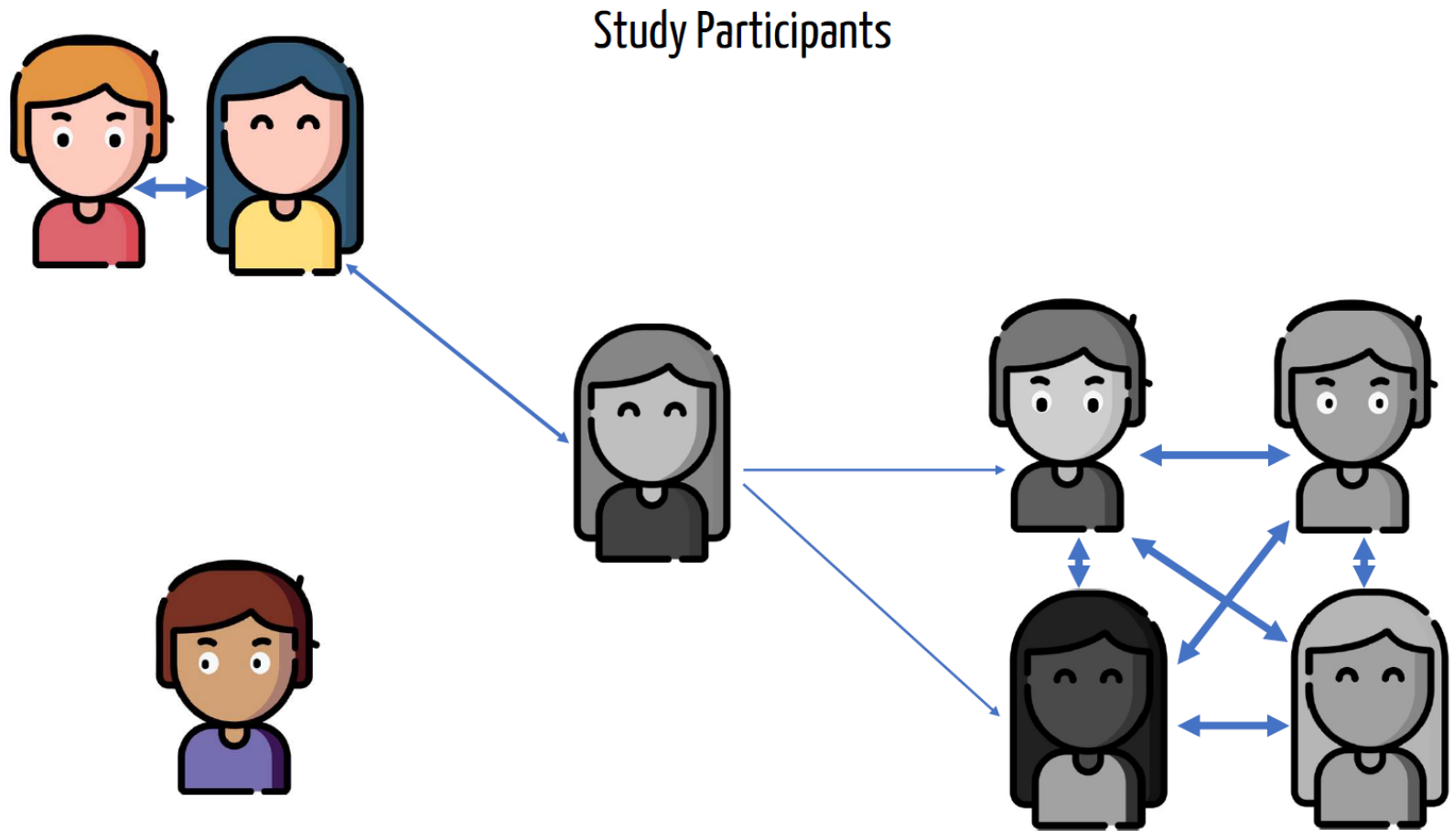
- Also referred to as spillovers
- Potential outcomes will depend on *who gets the treatment*

Let's look at an example

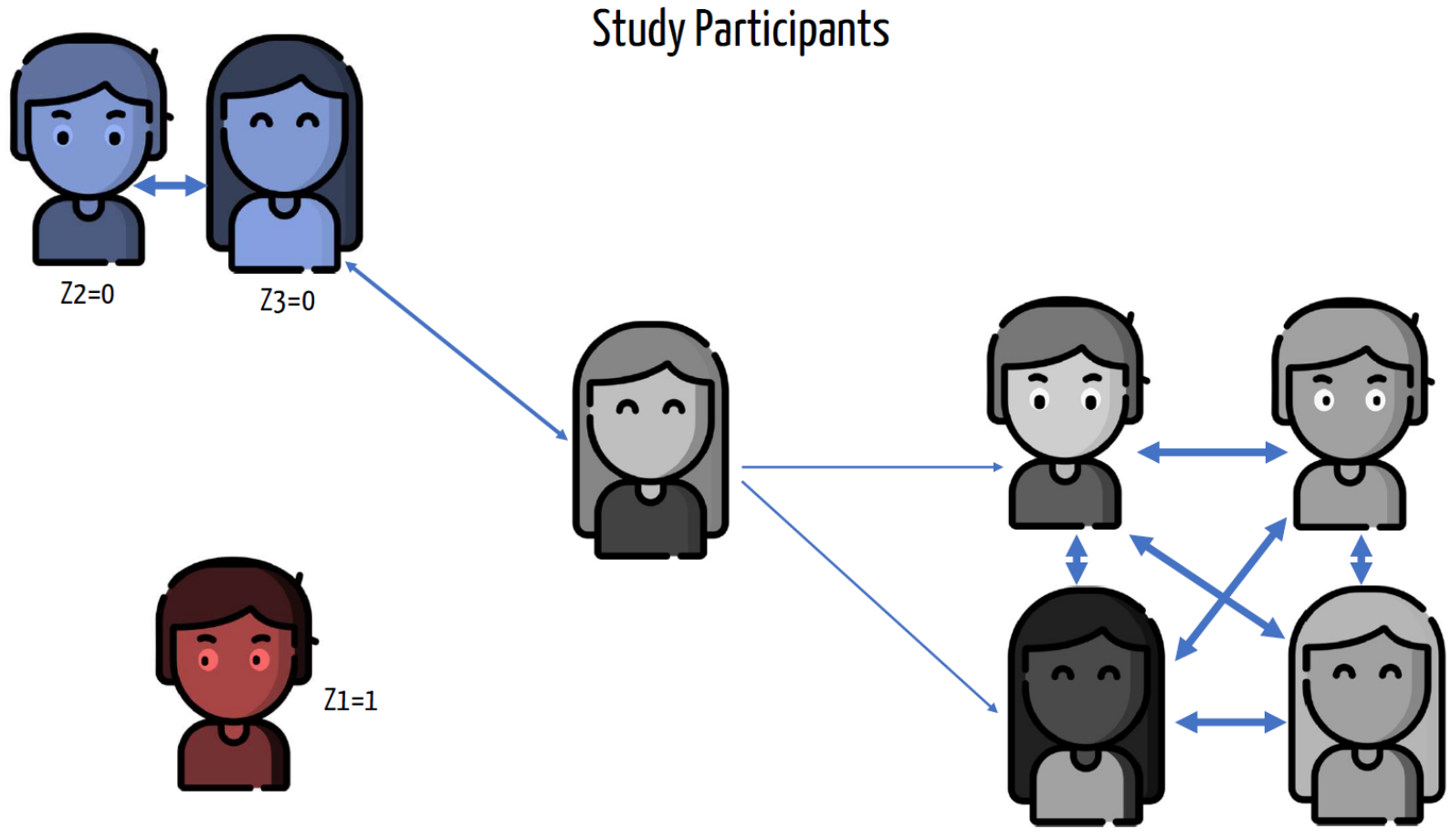
Network effects



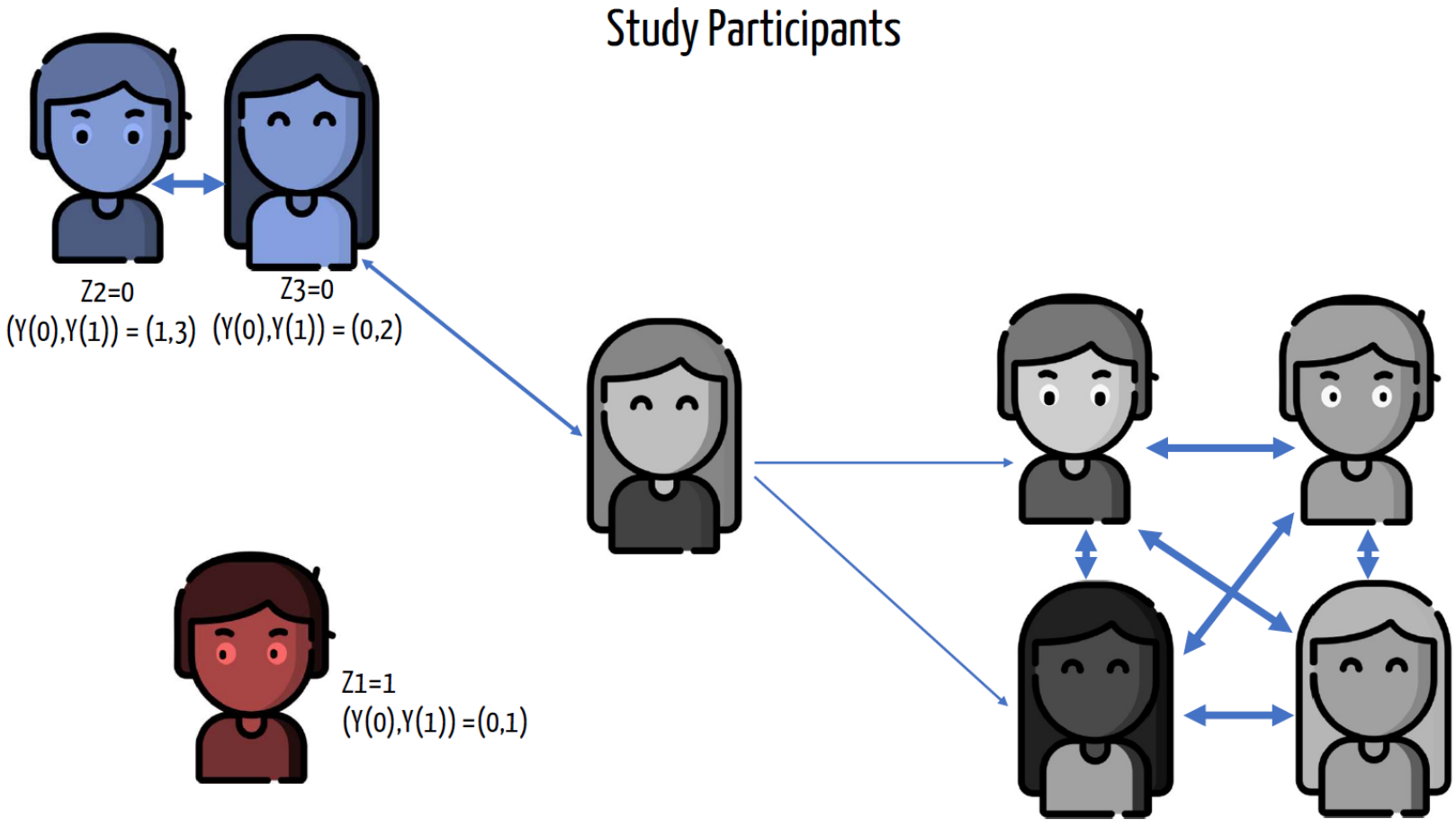
Network effects



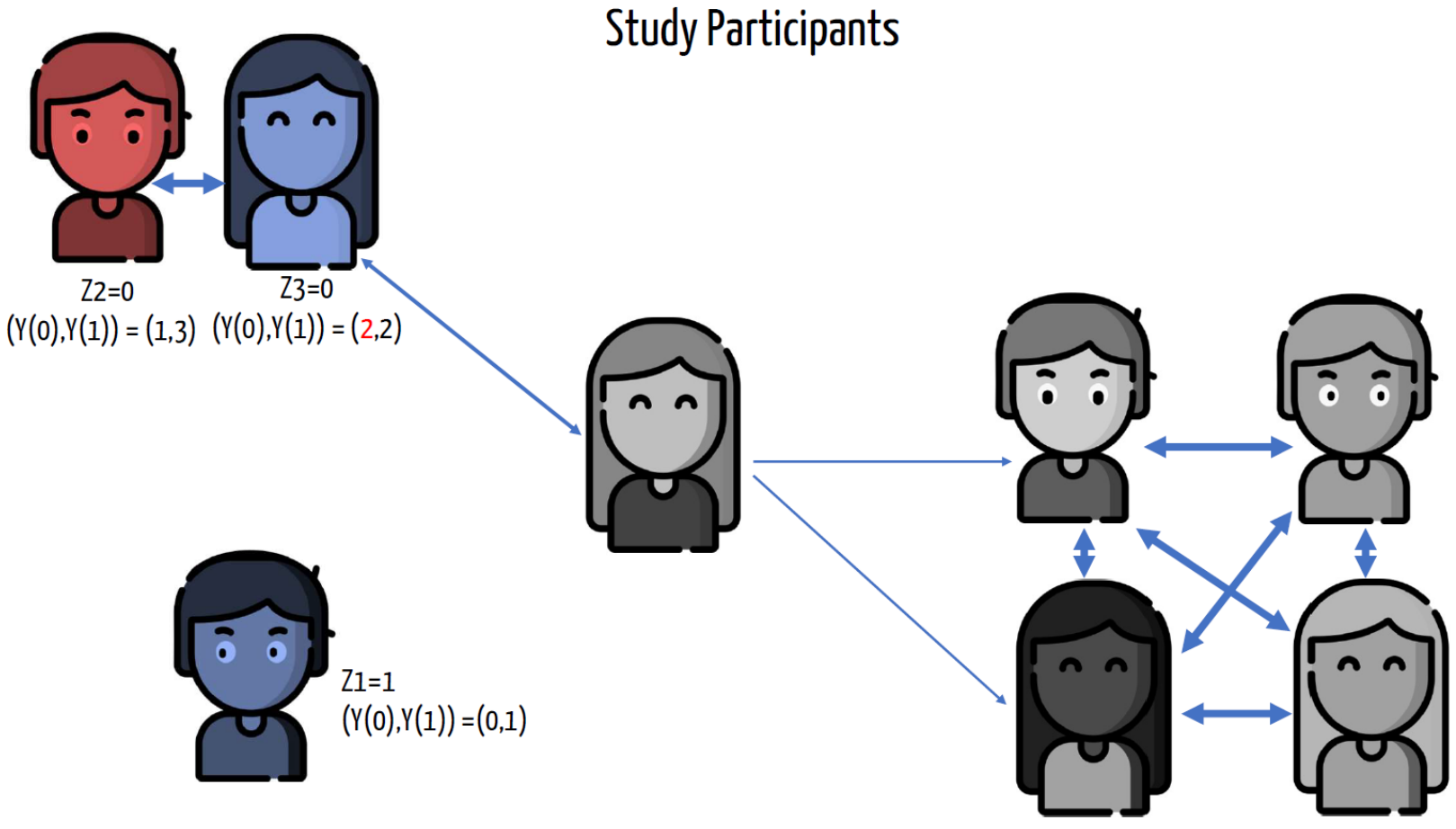
Network effects



Network effects



Network effects



Network effects

Can we do something about this?

1. Randomize at a higher level (e.g. neighborhood, school, etc. instead of at the individual level)
    - Note that you will have to cluster your standard errors
2. Model the network!

General Equilibrium Effects

- Usually arise when you scale up a program or intervention.
- Imagine you want to test the effect of providing information about employment and expected income to students to see whether it affect their choice of university and/or major.

What could happen if you offer it to everyone?

Show me the power

Statistical power

- Statistical power refers to the probability that a test correctly rejects the null hypothesis H_0, when H_0 is false.
    - It's related to sample size!

Power = 1- Pr(Type II error)

- Type I error: Probability of rejecting the null hypothesis, when the null is true (α).
- Type II error: Probability of not rejecting the null hypothesis, when the null is false (β).

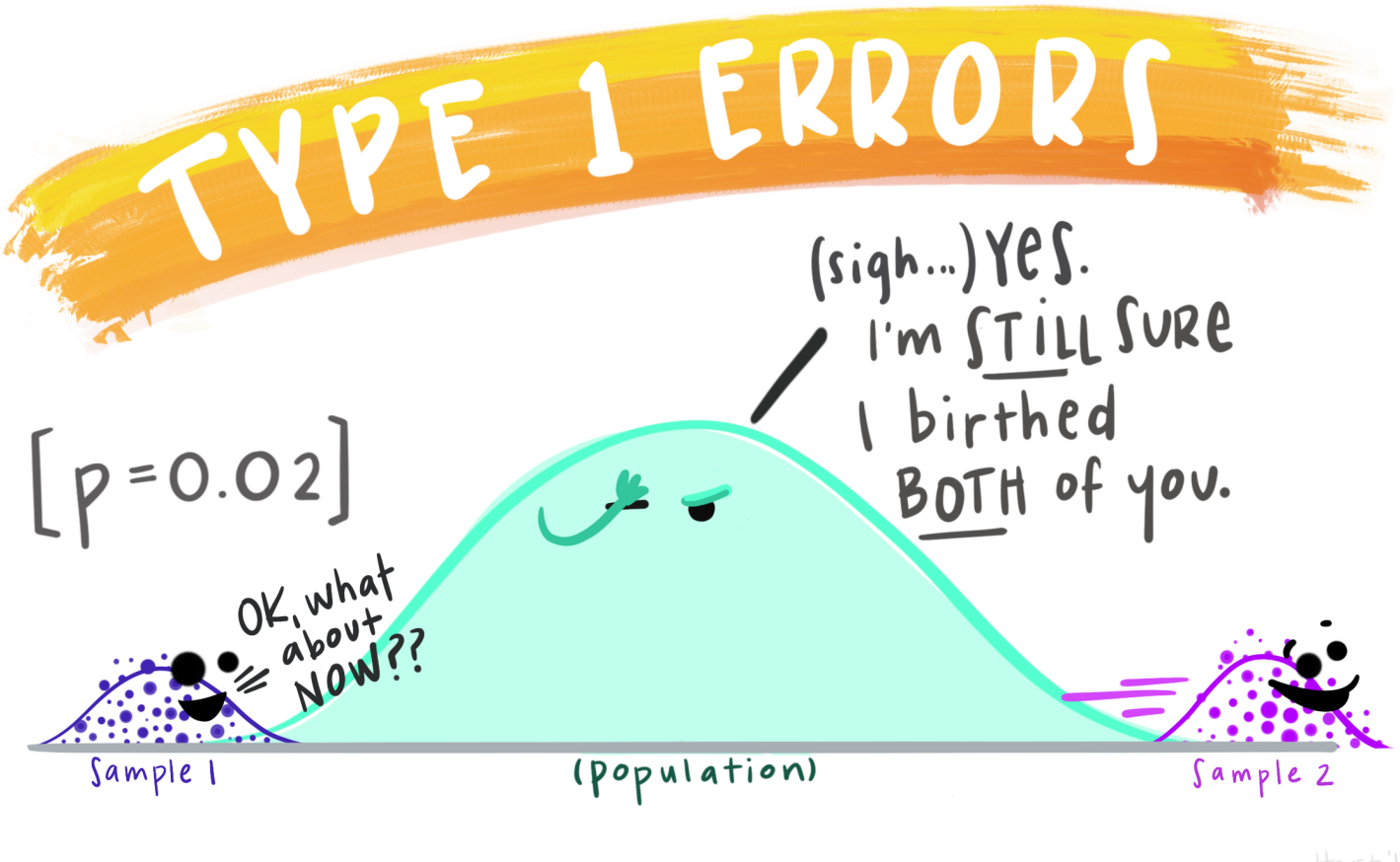
Statistical power

- Statistical power refers to the probability that a test correctly rejects the null hypothesis H_0, when H_0 is false.
    - It's related to sample size!

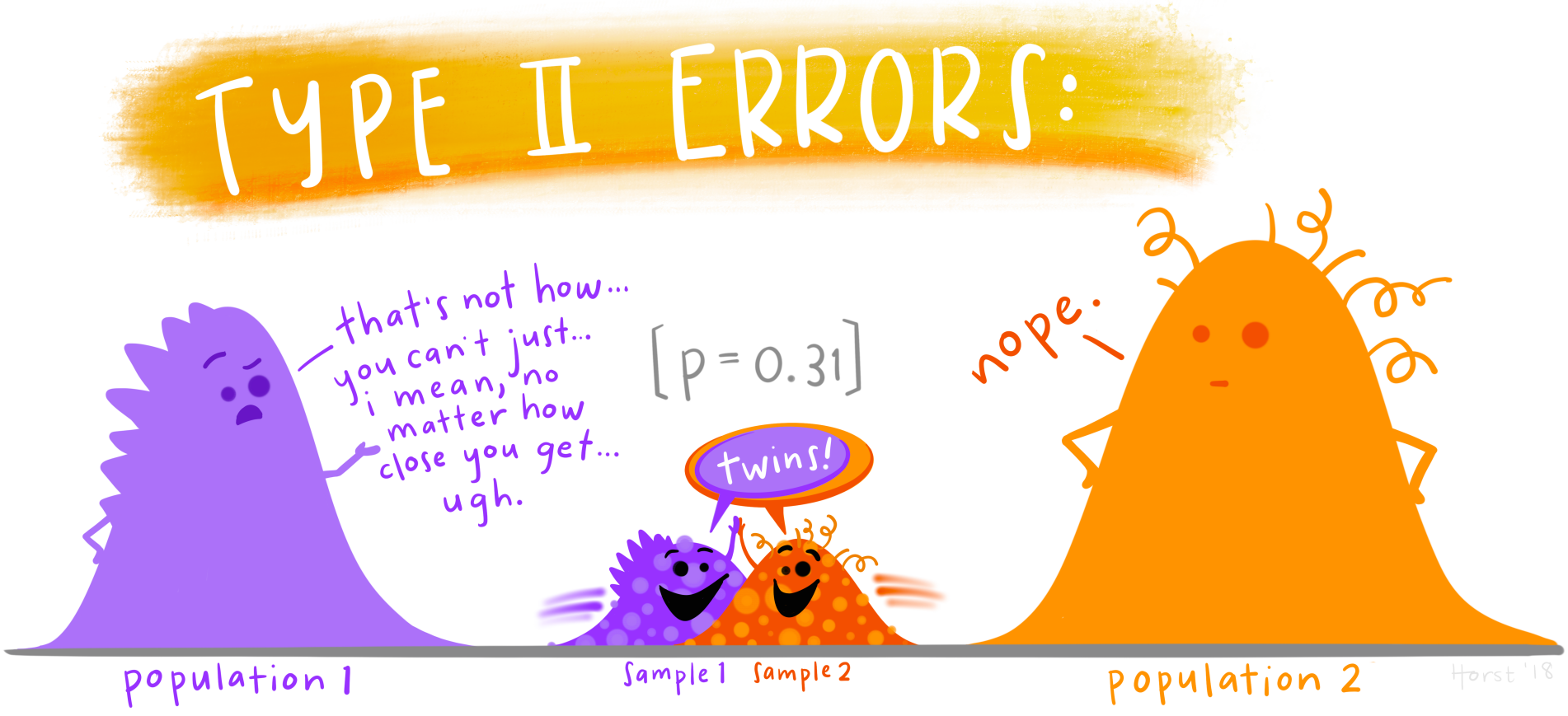
Power = 1- Pr(Type II error)

- False Positive: Probability of rejecting the null hypothesis, when the null is true (α).
- False Negative: Probability of not rejecting the null hypothesis, when the null is false (β).

Statistical power



Statistical power



Statistical power and sample size

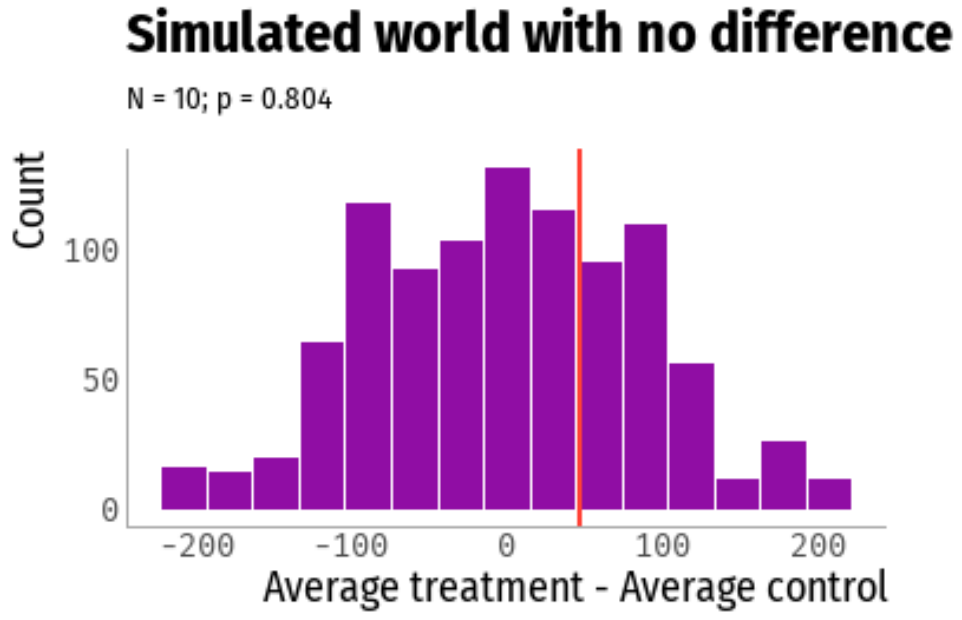
How big of a sample?

A training program causes incomes to rise by $40

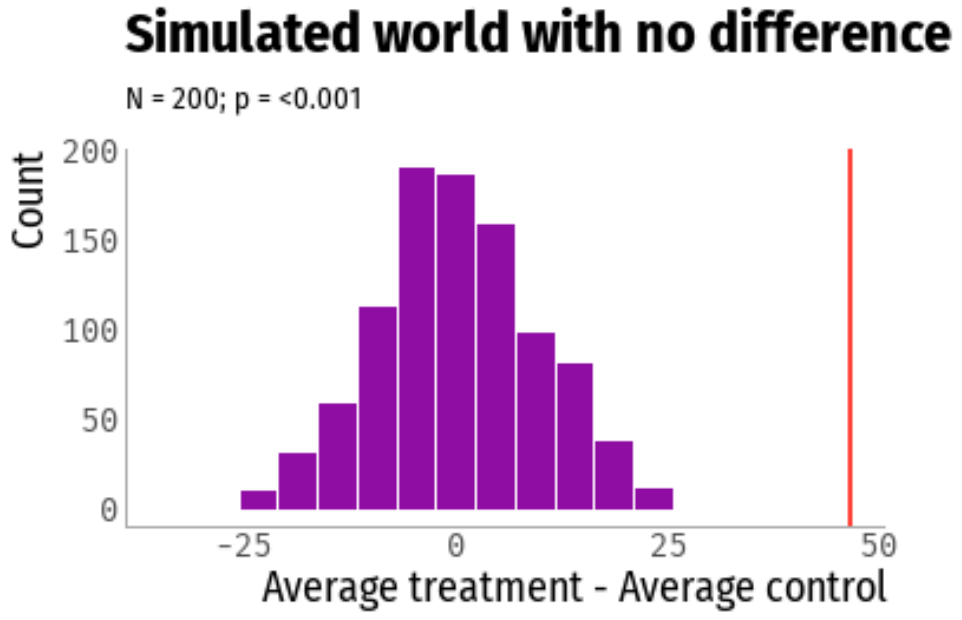
Person	Group	Before	After	Difference
295	Control	122.09	219.04	96.95
126	Treatment	205.60	199.84	-5.76
400	Control	133.25	120.40	-12.85
94	Treatment	270.11	206.56	-63.54
250	Control	344.37	212.89	-131.49
59	Treatment	312.41	268.06	-44.35

Can I detect an effect?

Enroll 10 participants

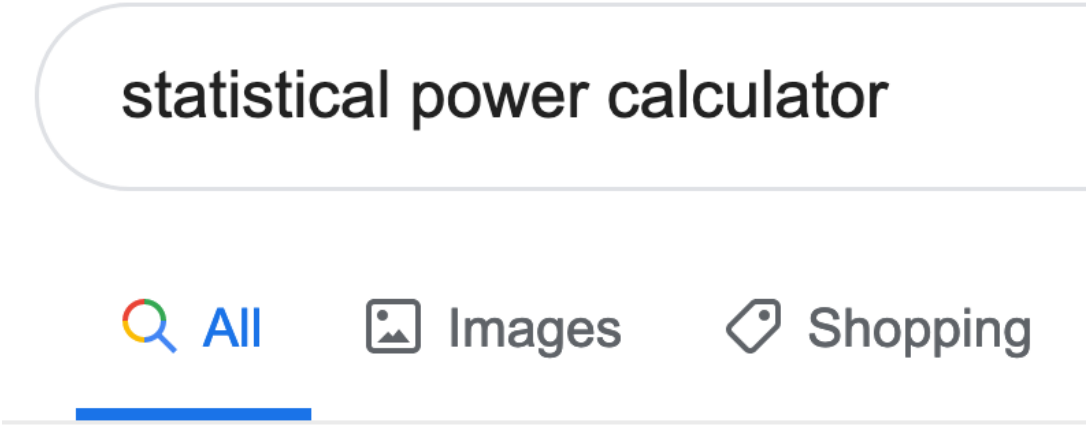


Enroll 200 participants



What's the right sample size?

Use a statistical power calculator to make sure you can potentially detect an effect



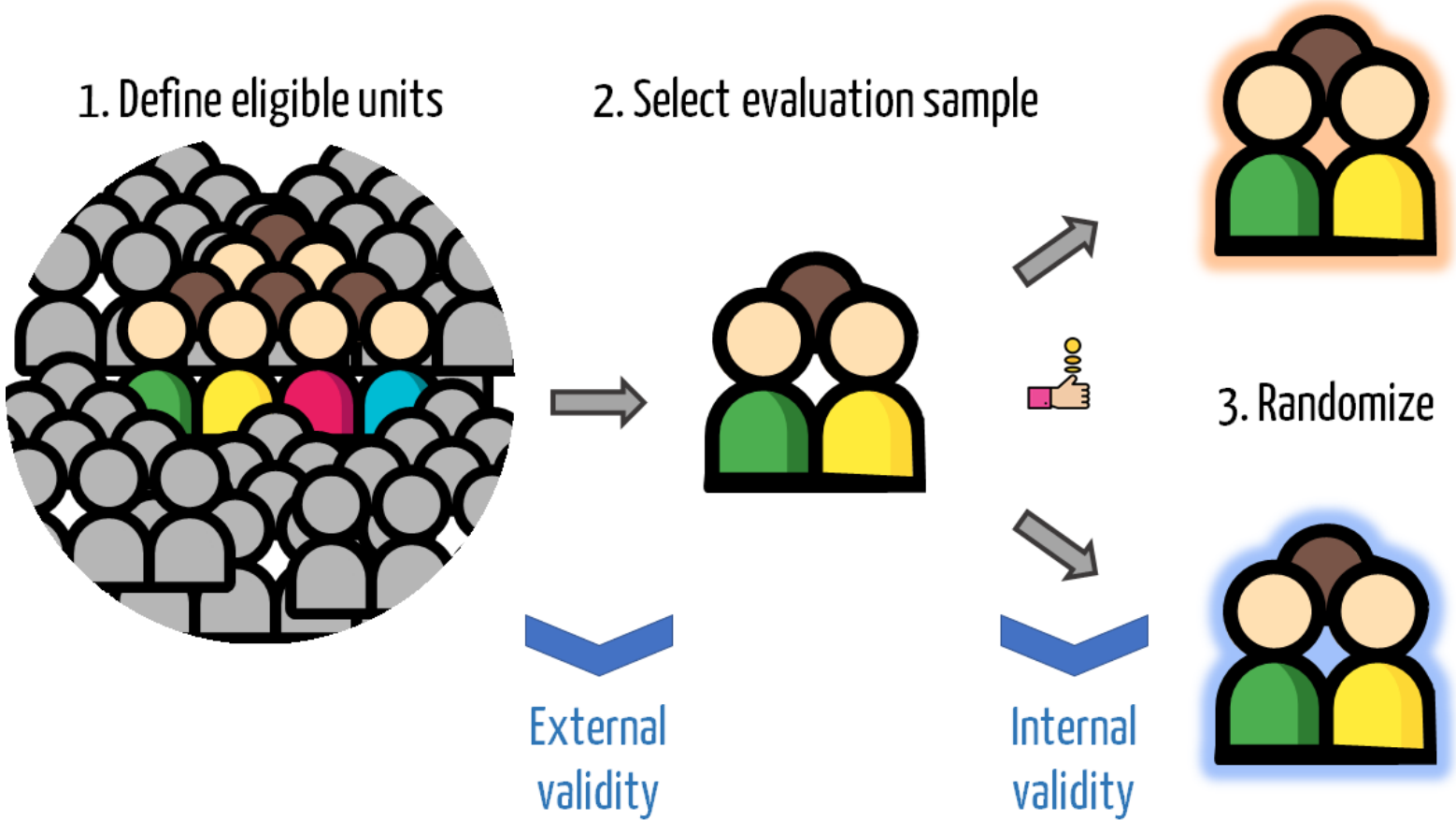
Or you can use simulations!

Limitations and Potential Problems in Randomized Controlled Trials

Generalizibility of RCTs

- External Validity vs Internal Validity:
  - External validity: "The extent to which results can be generalized to other contexts or populations."
  - Internal validity: "[T]he extent to which the observed results represent the truth in the population we are studying."

External vs Internal Validity



- Many times, RCTs use convenience samples

Noncompliance and Attrition

Attrition

Units fall out of your sample

- Can you give an example?

Noncompliance and Attrition

Attrition

Units fall out of your sample

Noncompliance

Units that where assigned to one group end up in another

- Example?

Noncompliance and Attrition

- If **attrition** is correlated with the treatment, we're in trouble.

Noncompliance and Attrition

- If **attrition** is correlated with the treatment, we're in trouble.

The ignorability assumption can break!

Noncompliance and Attrition

- If **attrition** is correlated with the treatment, we're in trouble.

The ignorability assumption can break!

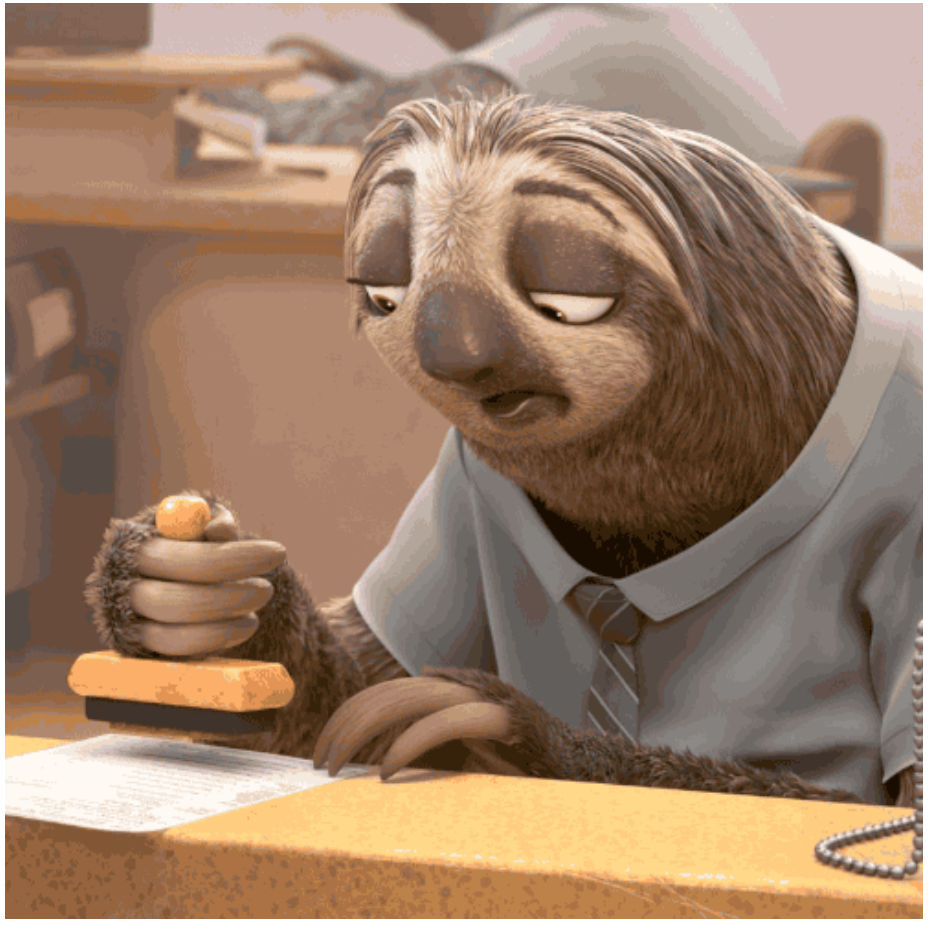
- **Noncompliance** is a problem for identifying ATE.
  - E.g. Treatment *assingment* is random, but not necessarily the *take-up* of the treatment
  - Stuck with **Intention to Treat (ITT) effect**.

Other problems to look out for

1. Hawthorne effects:
    - Effect that occurs when people behave differently because they are being watched
2. John Henry effects:
    - Effect that occurs when the control group works harder because they were not assigned to treatment.

Feasibility

- RCTs can be **expensive** and also **slow**.
- Organizations might **not be willing to randomize**.

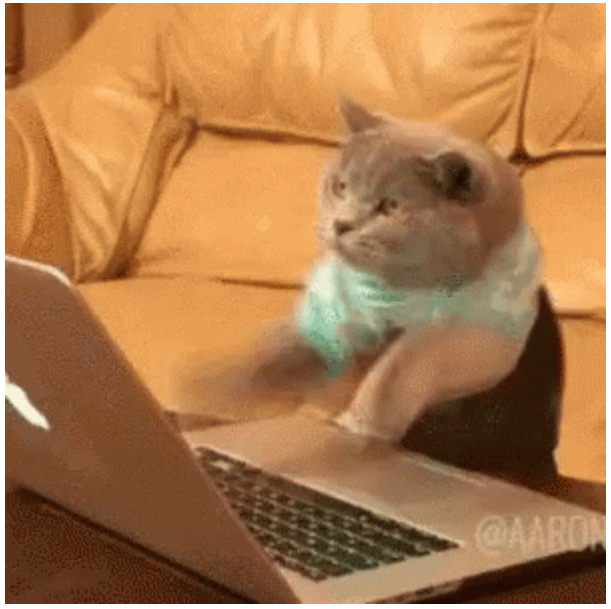


Getting around some issues

- Staggered adoption: Eventually treat everyone, but at different times.
    - You can think about this when you have panel data.
- Randomize in the bubble:
    - Not randomize everyone, but maybe those that are on the margin of receiving the treatment.

A/B Testing

A/B Testing



- A/B testing is very popular right now to test the effect of small changes in products on an outcome of interest.
- E.g. In web development, user would be exposed to different versions of the same website (only with subtle changes).
- Companies can quickly analyze the data, adapt their design, and continue testing features.

A/B Testing Jargon

- There's a lot of jargon in A/B testing that is thrown out there...

What does it all mean?

Conversions

Regret

Multi-armed bandits/ Contextual bandits

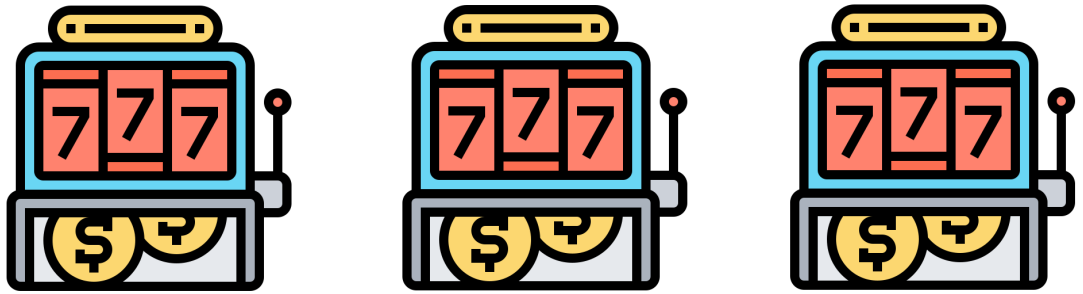
A/B Testing Jargon

- Conversions: "Cause the customer to take action"
- Regret: Loss of conversion due to a low-performing treatment.

Maximize conversions and minimize regret

A/B Testing Jargon

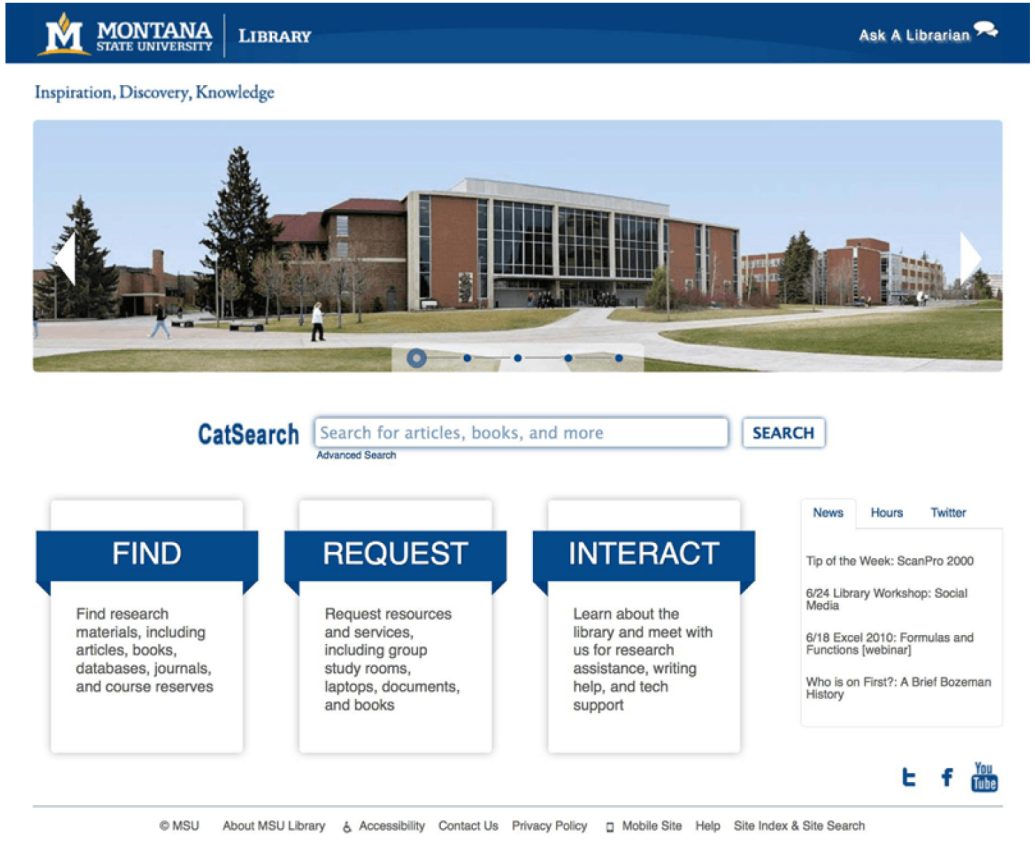
- Multi-armed bandits/ Contextual bandits
  - Name comes from machines at casinos.
  - You want to maximize total payout → Balance between exploration and exploitation



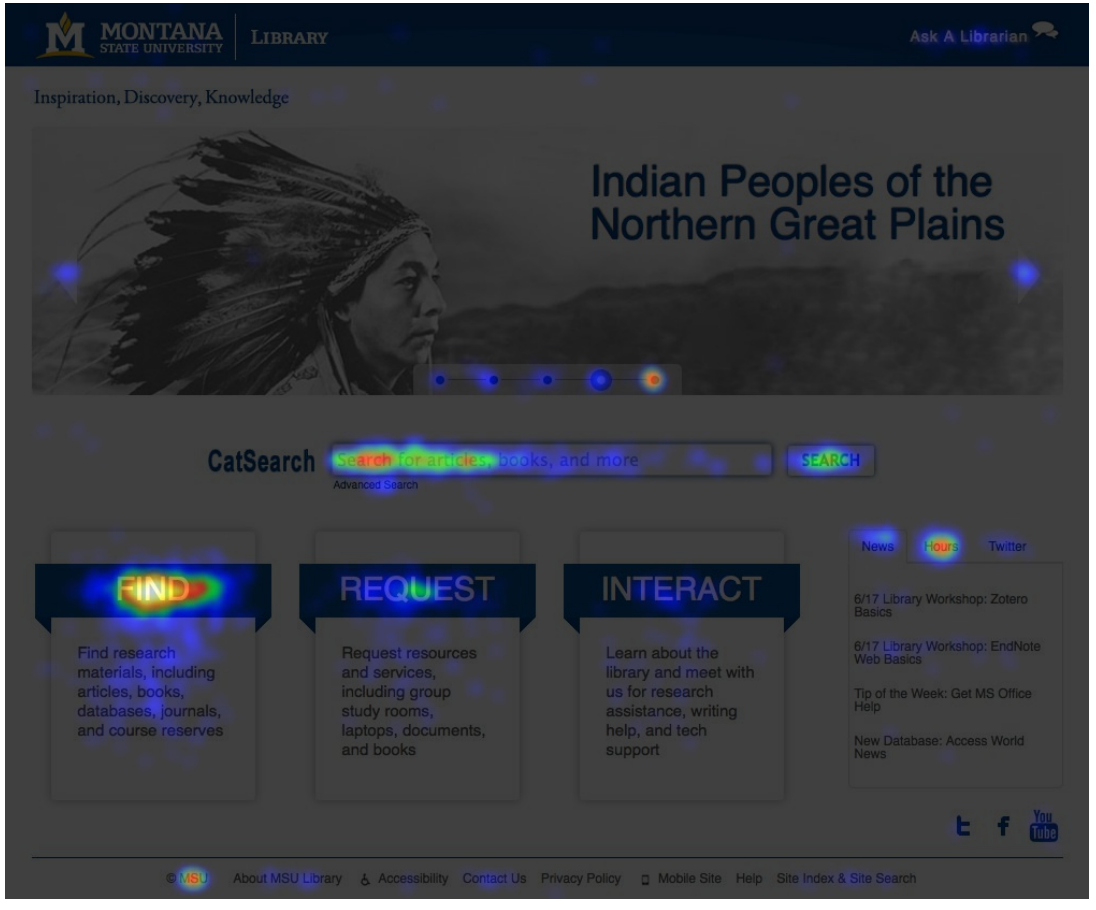
- Multi-armed bandits redirect users to the more successful versions of the treatment.
- A contextual bandit uses prior information from the user to select an action.

A/B Testing in Data: MSU Library

- Montana University Library website: Low interaction with the "Interact" Category

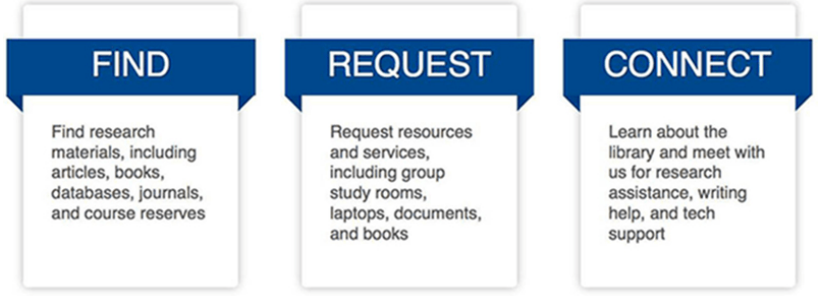


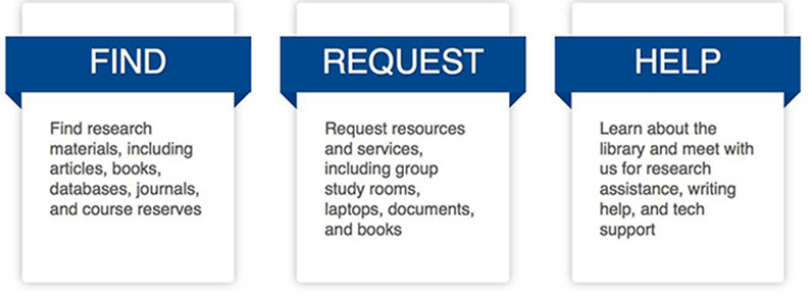
A/B Testing in Data: MSU Library

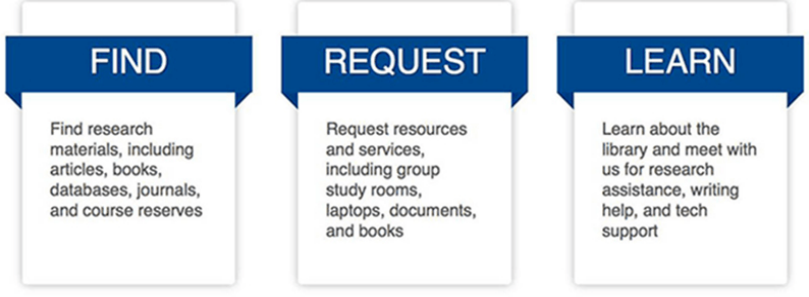


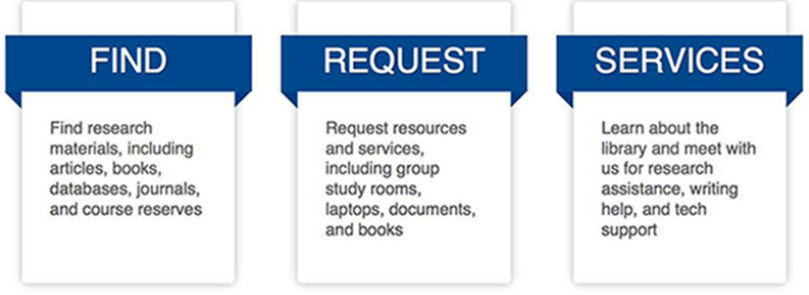
A/B Testing in Data: Different treatments

- Test four different names for that category









A/B Testing in Data: Analyze results

How would you analyze these results?

Homepage - Interact III (CrazyEgg)

May 29, 2013 - Jun 18, 2013

Stopped | Filters applied | View settings

Conversions / Page Metrics

All Sessions
37.45%

+ Add Segment

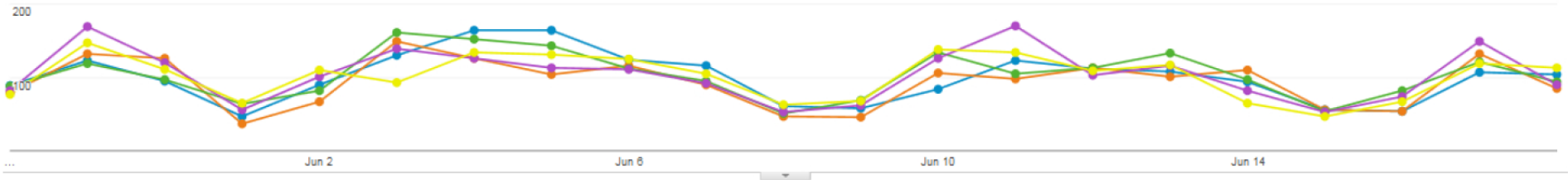
Explorer

Pageviews vs. Select a metric

Experiment stopped before winner was found

Day Week Month

Original - Interact Variation 1 - Connect Variation 2 - Learn Variation 3 - Help Variation 4 - Services



10,568 Pageviews

21 days of data

100% users included

Status:

Not enough data to identify winner

Primary Dimension: Variation

Plot Rows

Variation	Pageviews	Unique Pageviews	Avg. Time on Page	Entrances	Bounce Rate	% Exit	Page Value
Original - Interact	2,103	1,638	00:00:56	1,500	40.13%	36.38%	$0.00
Variation 3 - Help	2,187	1,690	00:01:39	1,633	50.03%	44.86%	$0.00
Variation 2 - Learn	2,166	1,693	00:01:14	1,612	46.90%	42.80%	$0.00
Variation 4 - Services	2,138	1,697	00:01:10	1,639	48.44%	44.15%	$0.00
Variation 1 - Connect	1,974	1,514	00:01:06	1,465	43.14%	39.21%	$0.00

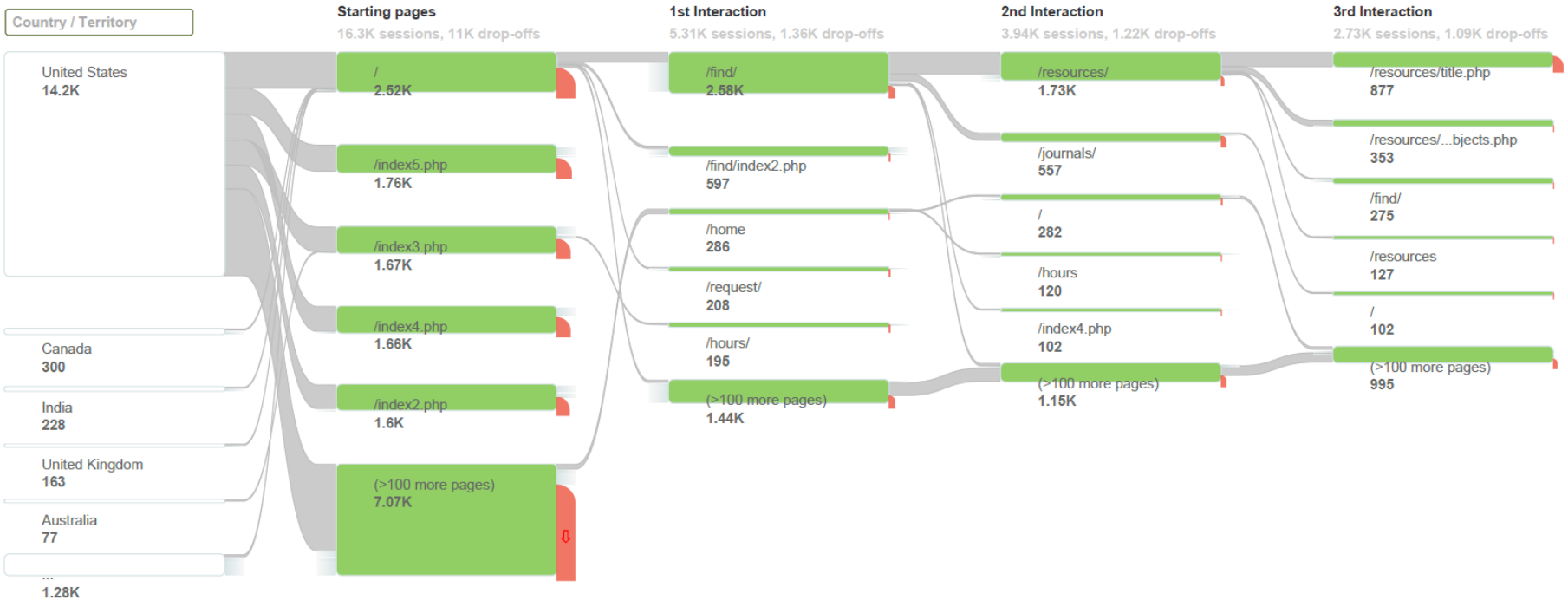
Show rows: 10 Go to: 1 1 - 5 of 5

Users Flow

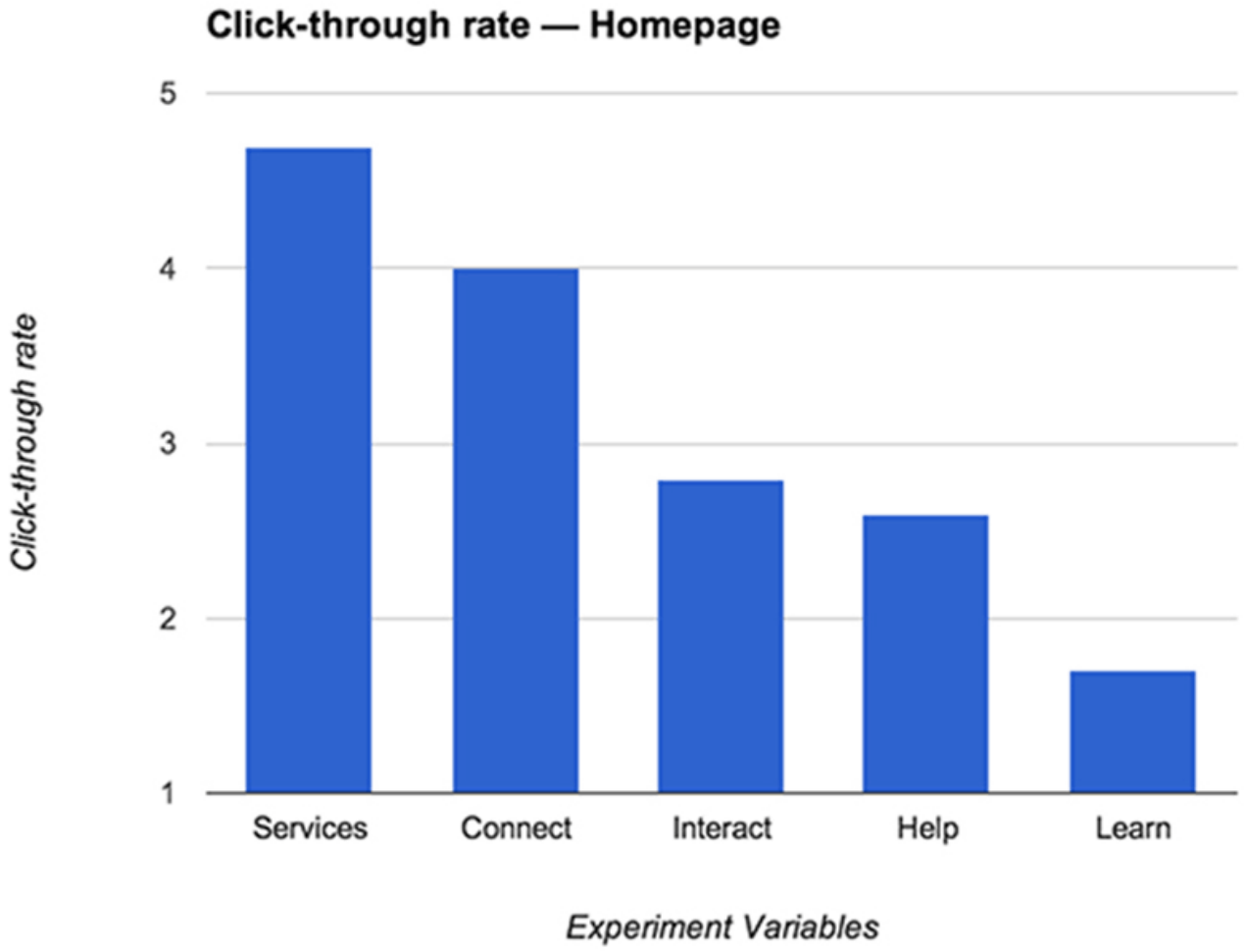
May 29, 2013 - Jun 18, 2013

All Sessions
100.00%

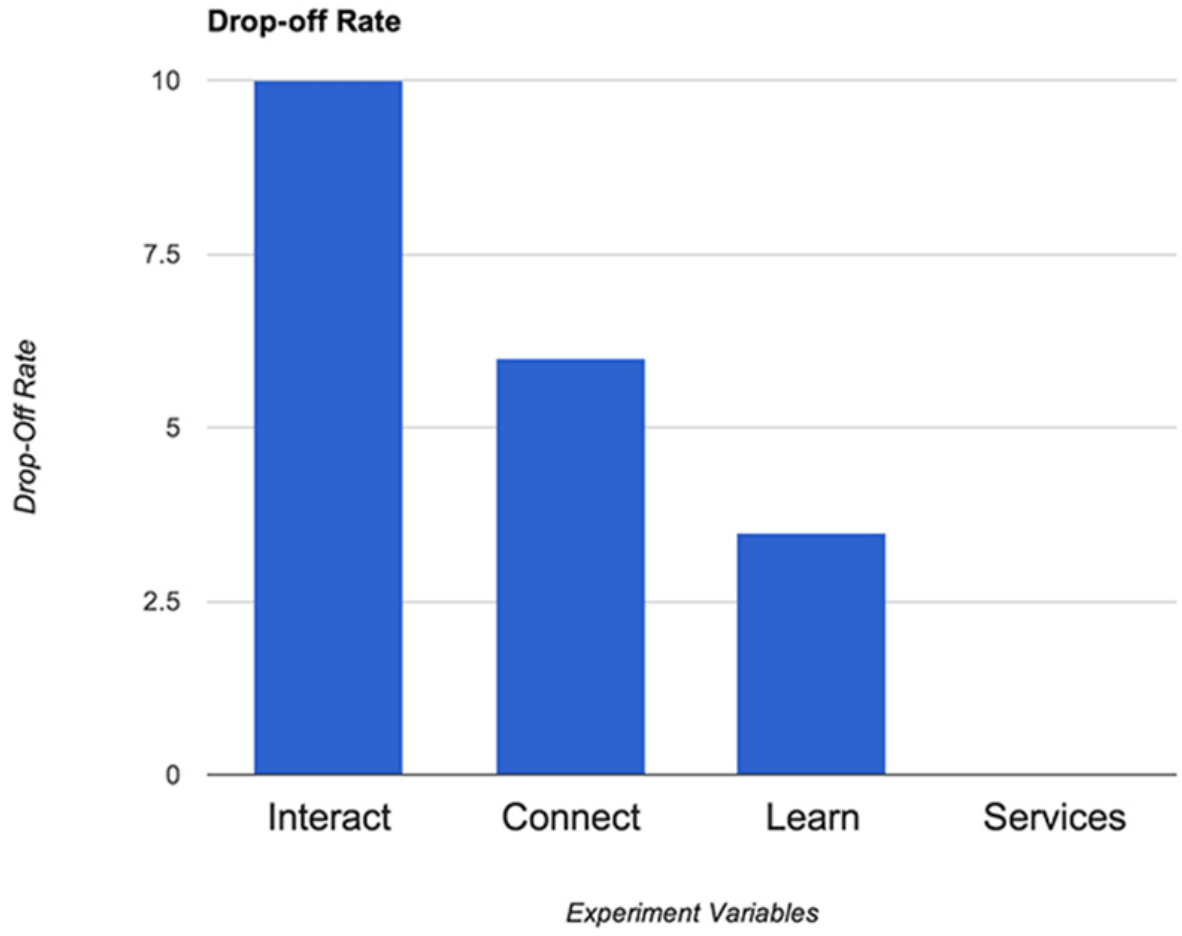
+ Add Segment



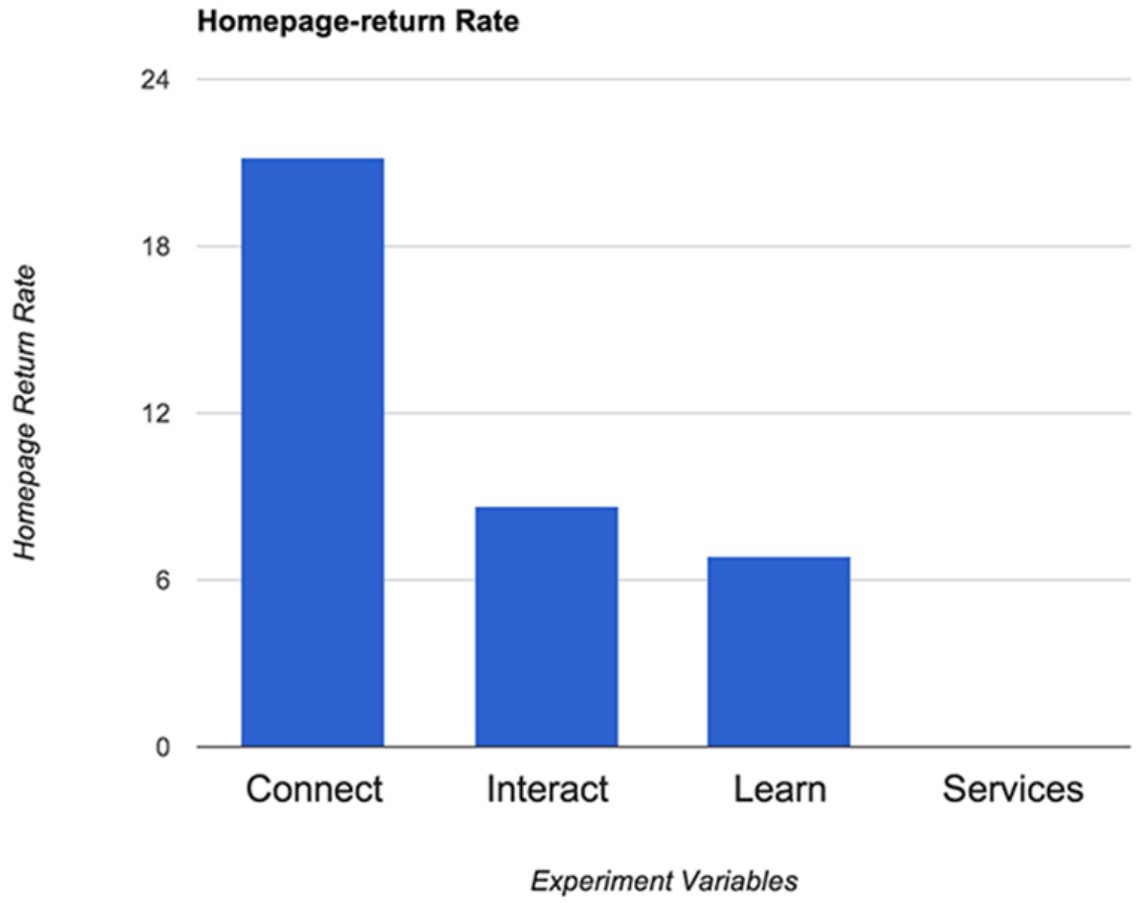
Homepage click-through



Drop-off


Drop-off Rate
10
7.5
5
2.5
0
Drop-Off Rate
Interact
Connect
Learn
Services
Experiment Variables

Homepage return



Wrapping things up

- Randomized controlled trials are great… but not for everything!
- Randomization buys us no systematic selection on observables or unobservables
    - But things can go wrong, too!

Check your assumptions and look out for potential issues!