Studying Online Behavior at Scale (SOC 412)

Week 2 Lecture 1

Sherrerd Hall 306

J. Nathan Matias

@natematias

civilservant.io

jmatias@princeton.edu

Department of Psychology

PRINCETON UNIVERSITY

CITP

mit media lab

What we will cover today

Discuss today's readings

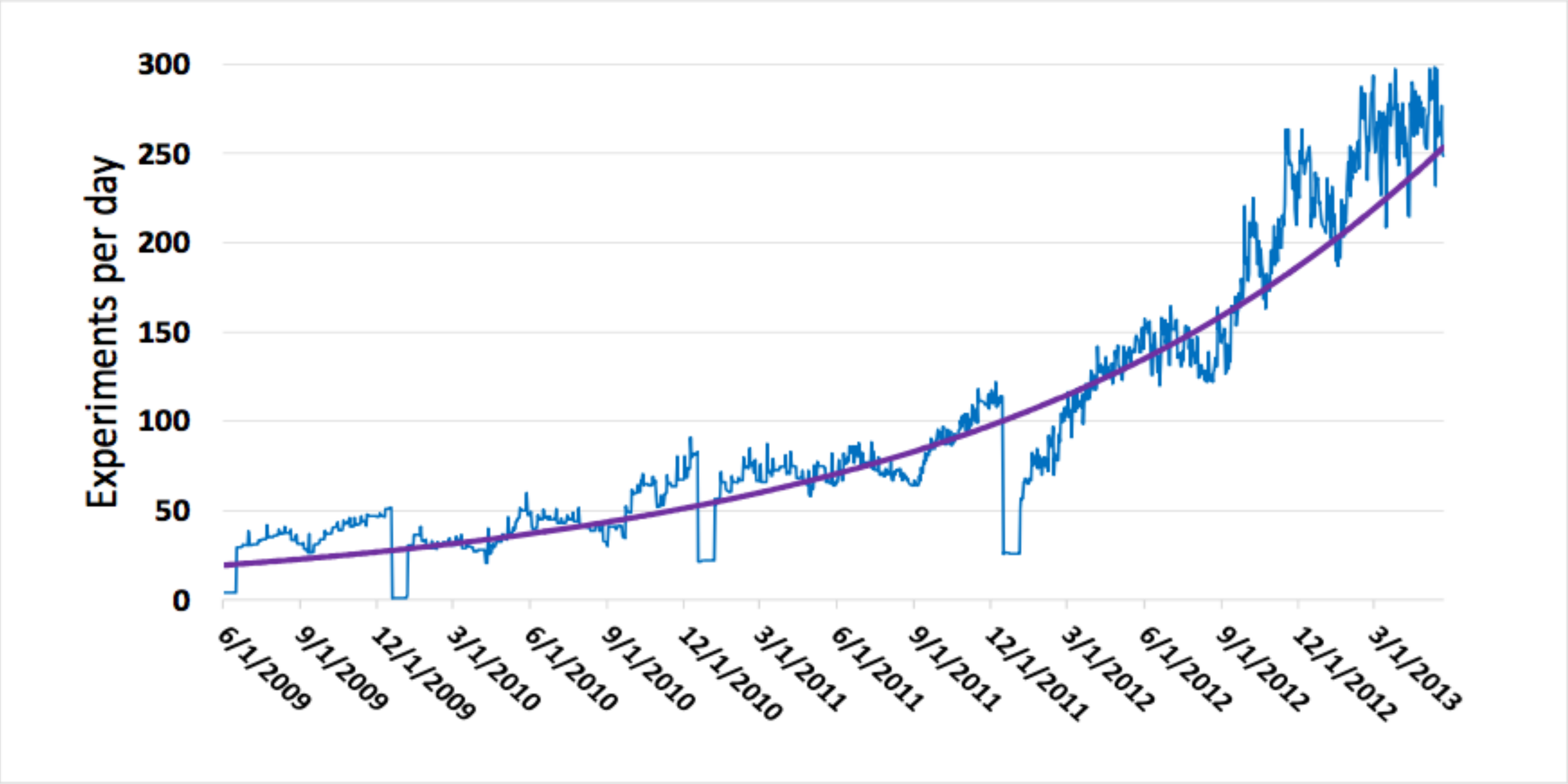
Finish last week's slides

Discuss this week's assignment

Upcoming assignments

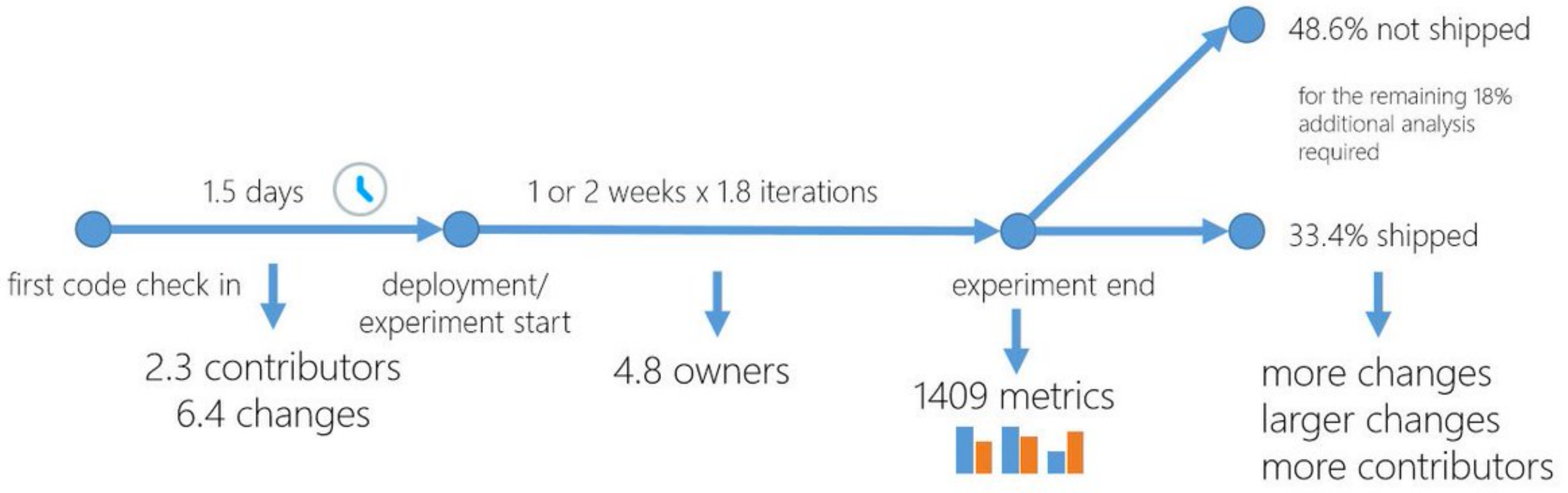
Class projects

A Rhythm for readings, discussion, presentations



Experiments Per Day on bing.com

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013, August). **Online controlled experiments at large scale**. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1168-1176). ACM.



Kevic, K., Murphy, B., Williams, L., & Beckmann, J. (2017, May). **Characterizing experimentation in continuous deployment: a case study on bing.** In Proceedings of the 39th International Conference on Software Engineering: Software Engineering in Practice Track (pp. 123-132). IEEE Press.

bing

“ the number of concurrent experiments running in ERF has grown from a few dozen (in 2014) to **about 500 concurrent experiments** [May 2017]

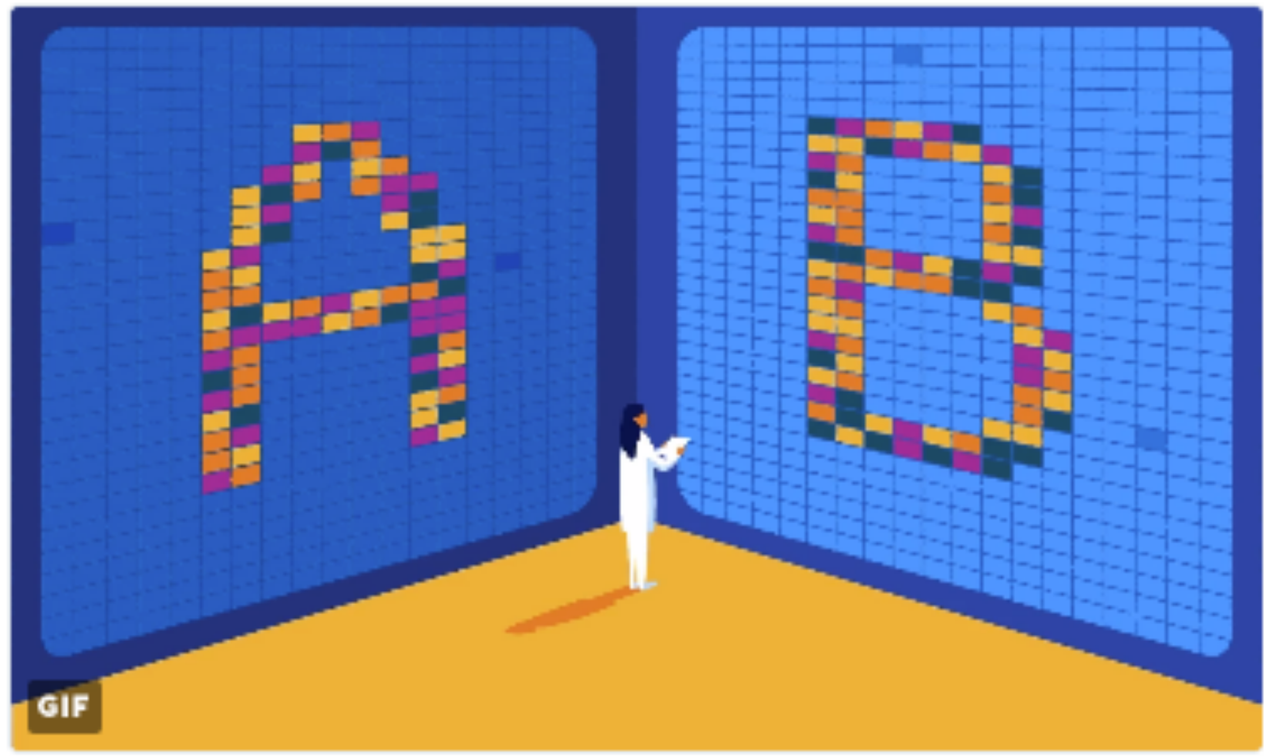
Today we compute ~2500 distinct metrics per day and roughly **50k distinct experiment/ metric combinations**.

Parks, Jonathan. Scaling Airbnb’s Experimentation Platform. Airbnb Engineering & Data Science. May 10, 2017.

airbnb

MailChimp
@MailChimp
Follow

We looked at nearly 500K A/B testing campaigns to determine how long a test needs to run to give you accurate results. Here's what we found: expi.co/01iJRL



1:00 AM - 9 Feb 2018

19 Retweets 50 Likes

19 50

Jonathan Parks Follow
May 10, 2017 · 8 min read

Scaling Airbnb's Experimentation Platform

Simon Jackson
Data scientist in the Experiment Tool team at Booking.com, Ph.D. in cognitive psychology, R guy.
Jan 22 · 10 min read

How Booking.com increases the power of online experiments with CUPED

Simon Jackson | Data Scientist at Booking.com

Data-supported decisions rule the roost at Booking.com. All product teams are empowered to do controlled experiments (A/B testing) and test any changes they make to the website (Kaufman, Pitchforth, & Vermeer, 2017). Such experiments expose some users to the existing website (base) while others see a new variant, and we statistically test the observed difference.

The Unofficial Google Data Science Blog

HOME ABOUT THIS BLOG

Designing A/B tests in a collaboration network

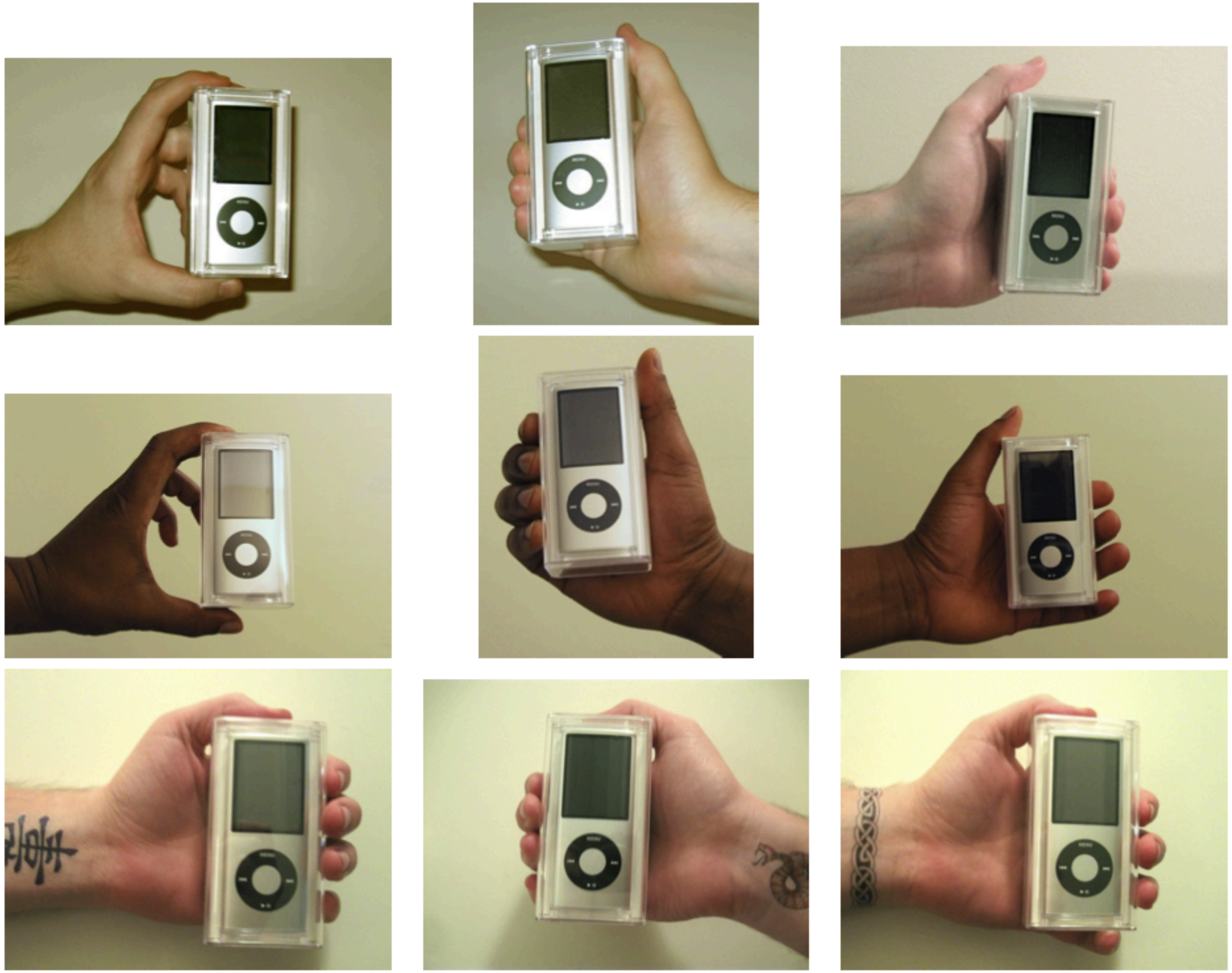
January 16, 2018

BY SANGHO YOON

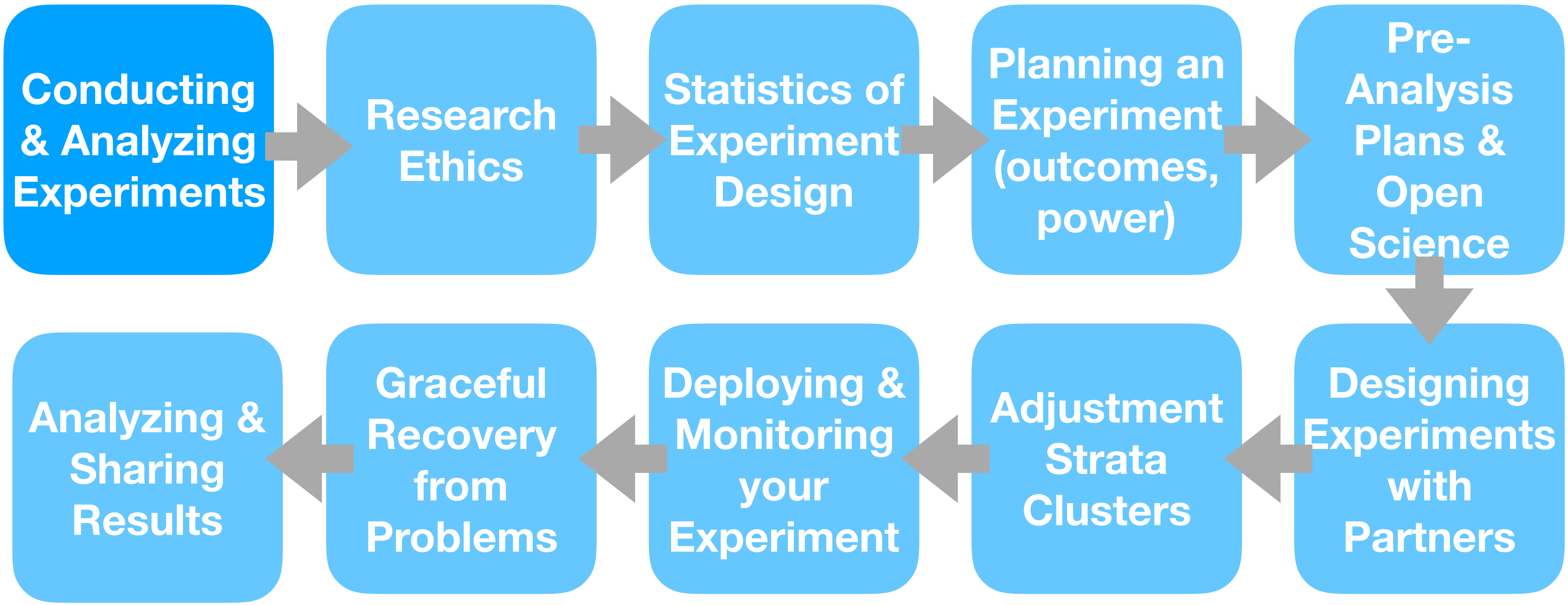
In this article, we discuss an approach to the design of experiments in a network. In particular, we describe a method to prevent potential contamination (or inconsistent treatment exposure) of samples due to network effects. We present data from Google Cloud Platform (GCP) as an example of how we use A/B testing when users are connected. Our methodology can be extended to other areas where the network is observed and when avoiding contamination is of primary concern in experiment design. We first describe the unique challenges in designing experiments on developers working on GCP. We then use simulation to show how proper selection of the randomization unit can avoid estimation bias. This simulation is based on the actual user network of GCP.

Experimentation on networks

A/B testing is a standard method of measuring the effect of changes by randomizing samples into different treatment groups. Randomization is essential to A/B testing because it removes selection bias as well as the potential for confounding factors in assessing treatment effects.



Doleac, J. L., & Stein, L. C. (2013). ***The visible hand: Race and online market outcomes.*** The Economic Journal, 123(572), F469-F492.


Conducting & Analyzing Experiments
Research Ethics
Statistics of Experiment Design
Planning an Experiment (outcomes, power)
Pre-Analysis Plans & Open Science
Designing Experiments with Partners
Adjustment Strata Clusters
Deploying & Monitoring your Experiment
Graceful Recovery from Problems
Analyzing & Sharing Results

What we will cover today

Discuss today's readings

Finish last week's slides

Discuss this week's assignment

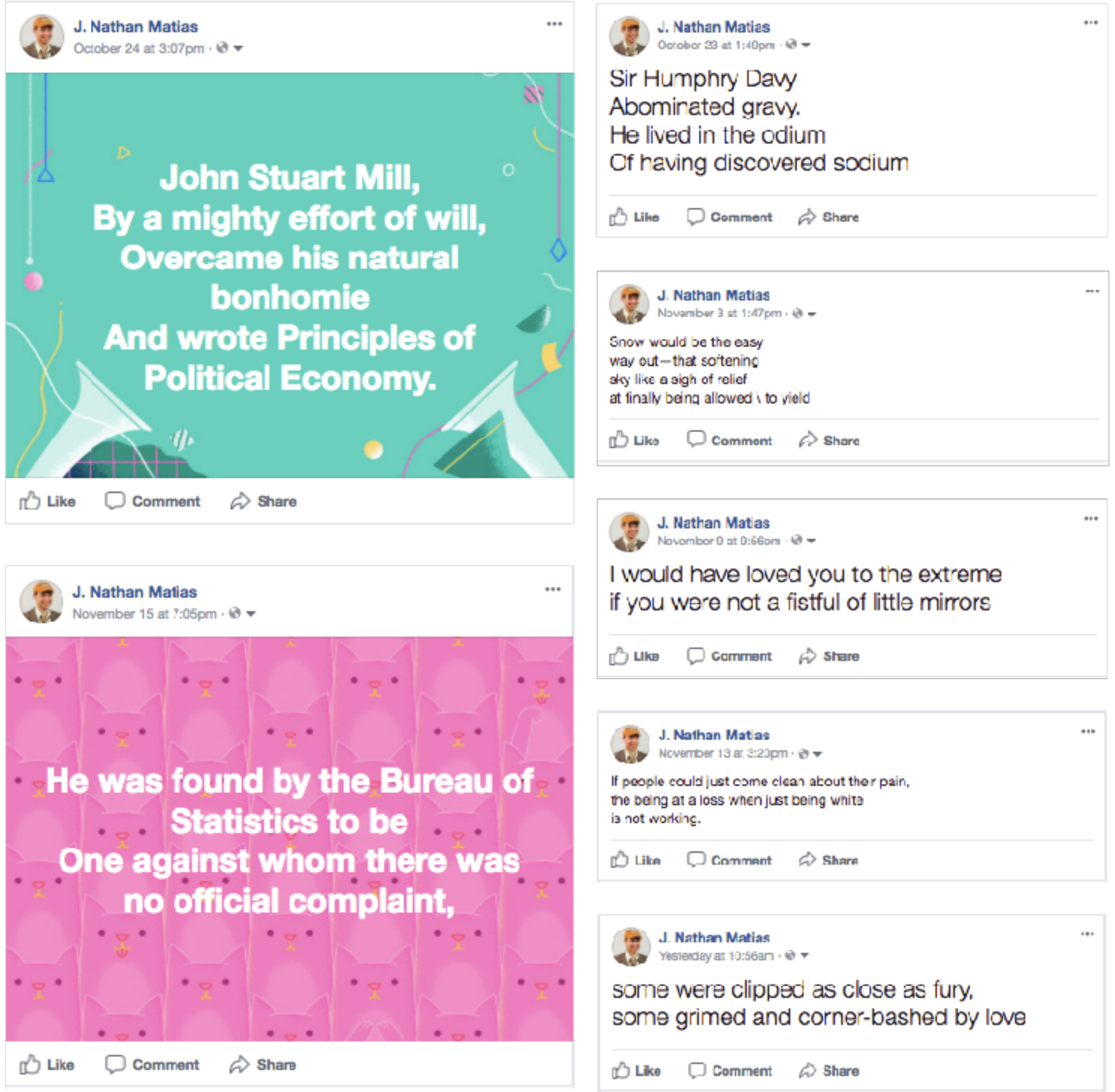
Upcoming assignments

Class projects

A Rhythm for readings, discussion, presentations

Parts of an Experiment

1. Does A **Colored Background** **Increase Facebook Interactions** on average?
2. If so, **by how much** on average?



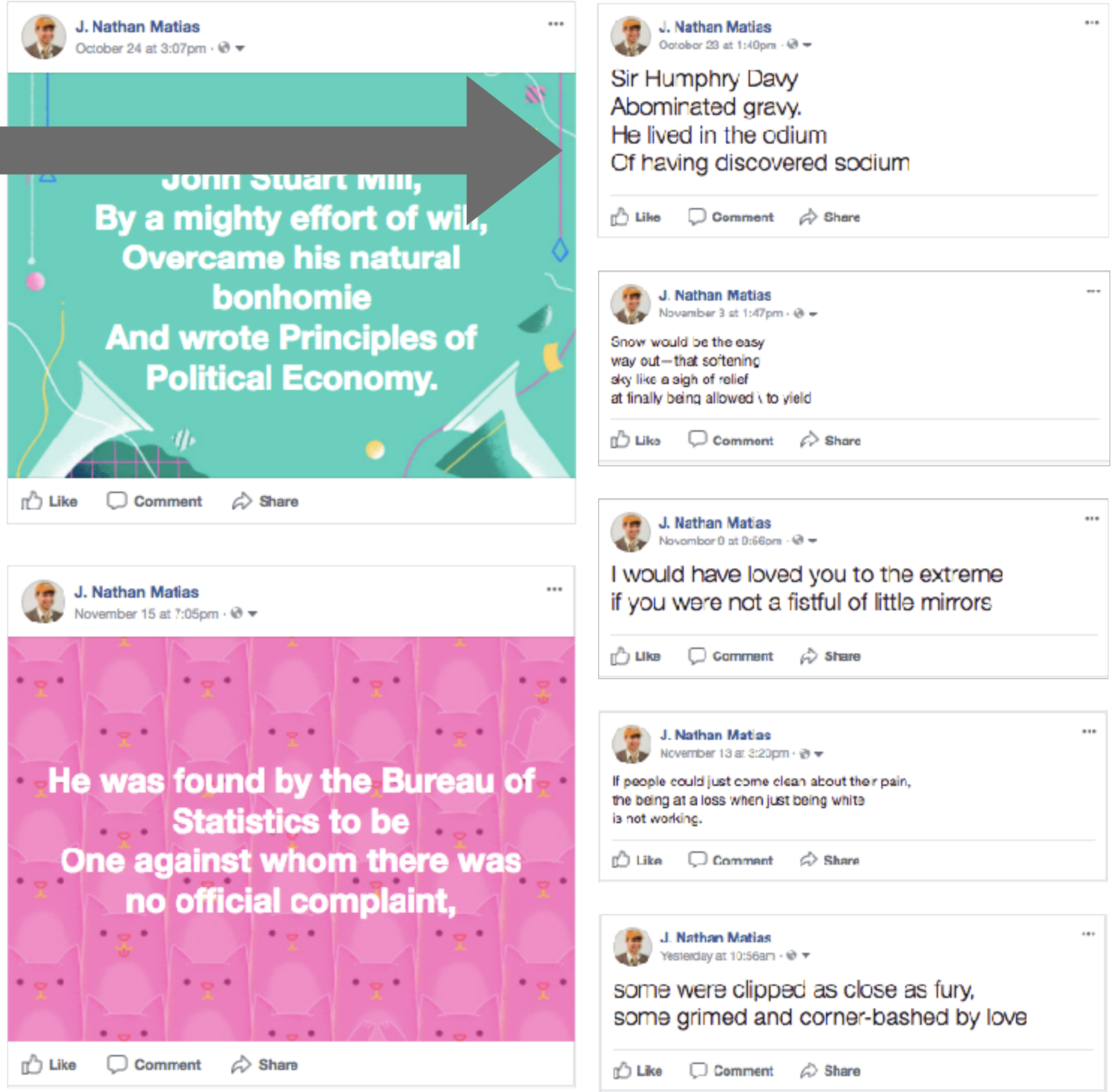
Parts of an Experiment

Background Color

?

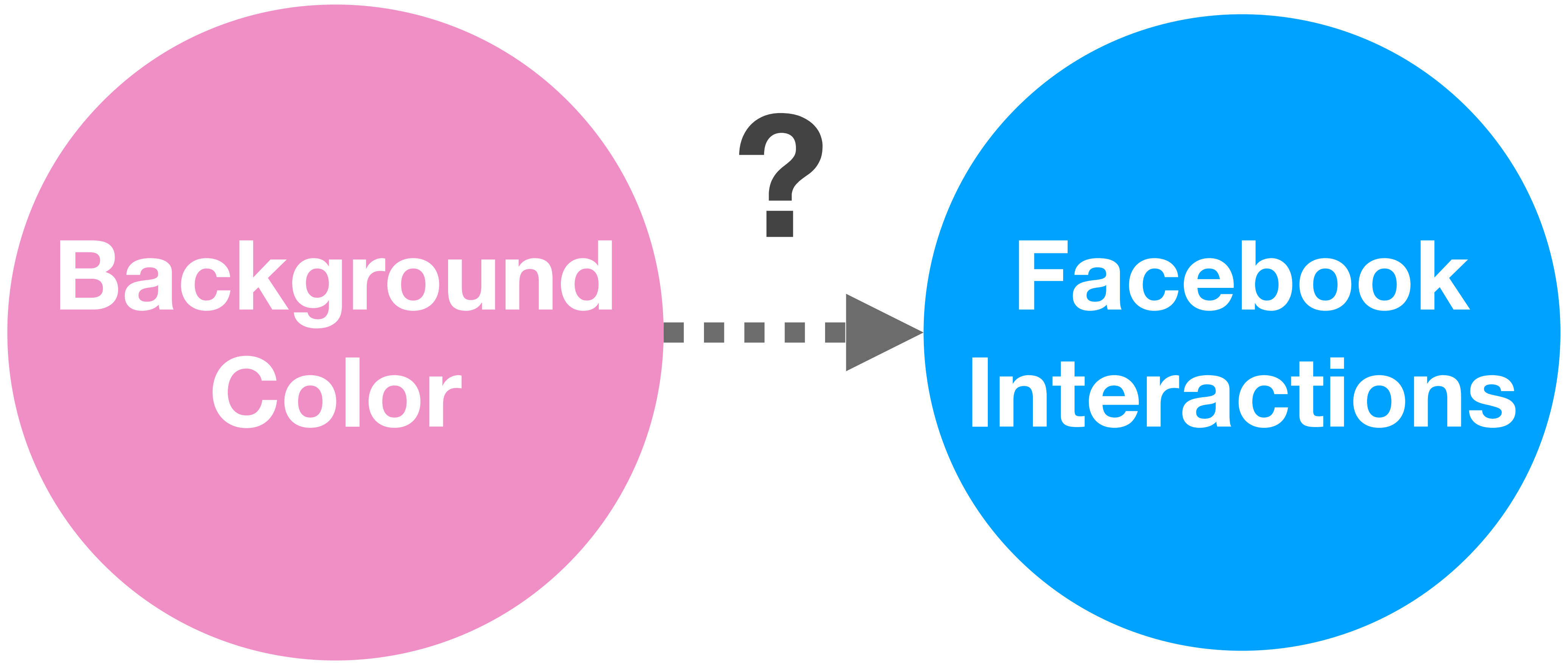
Comparison
(control)
(placebo)

Intervention
(treatment)



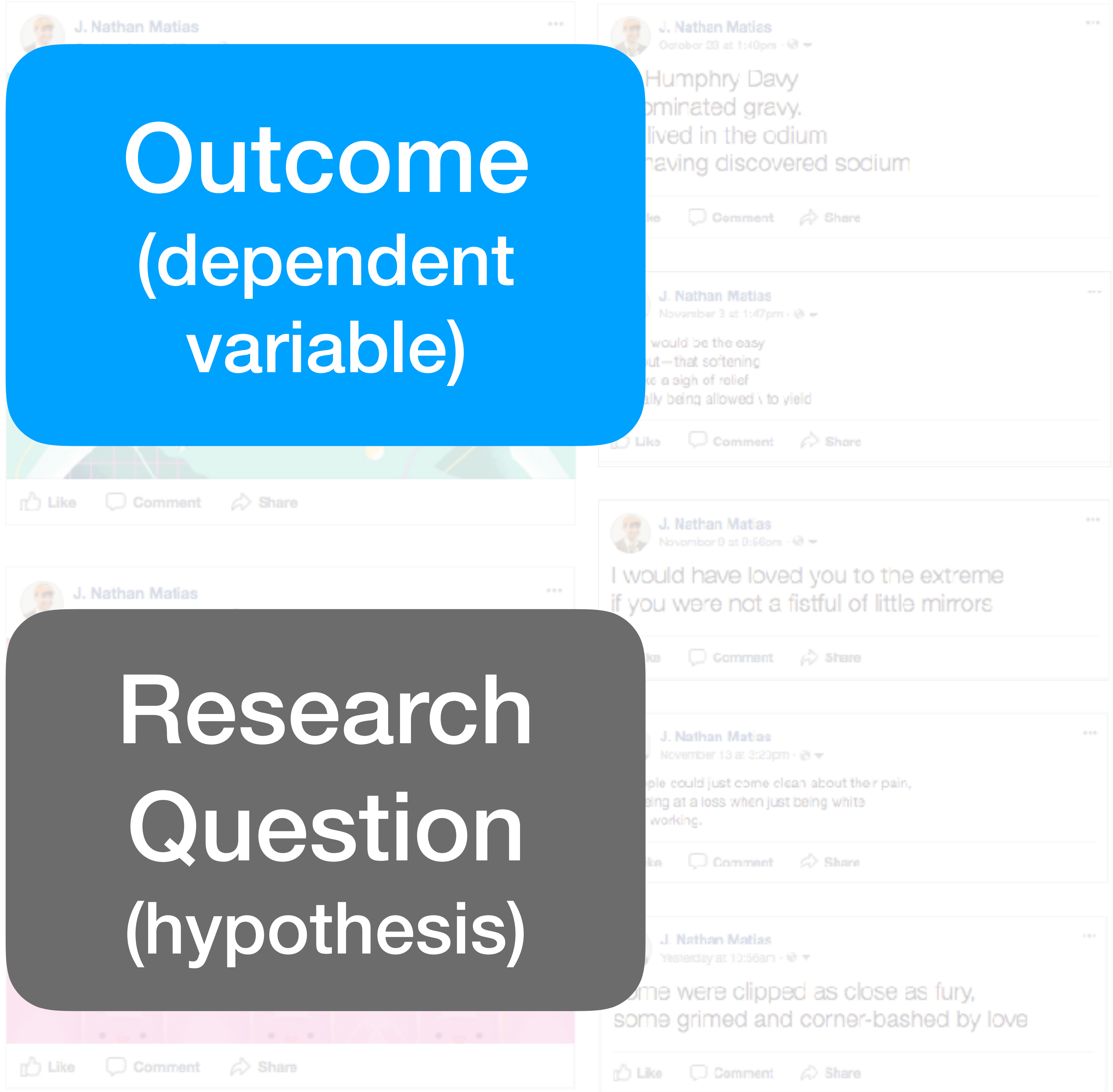
1. Does A Colored Background Increase Facebook Interactions on average?
2. If so, by how

Parts of an Experiment



Outcome (dependent variable)

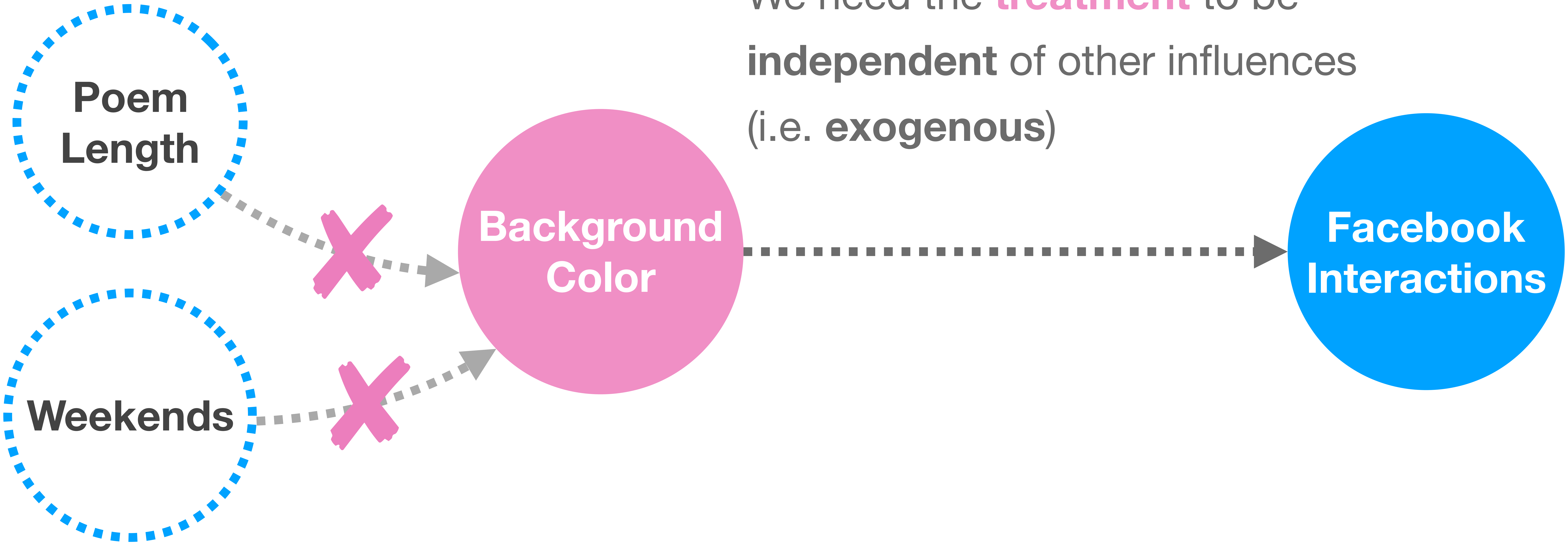
Research Question (hypothesis)



1. Does A Colored Background **Increase** Facebook Interactions on average?
2. If so, **by how much** on average?

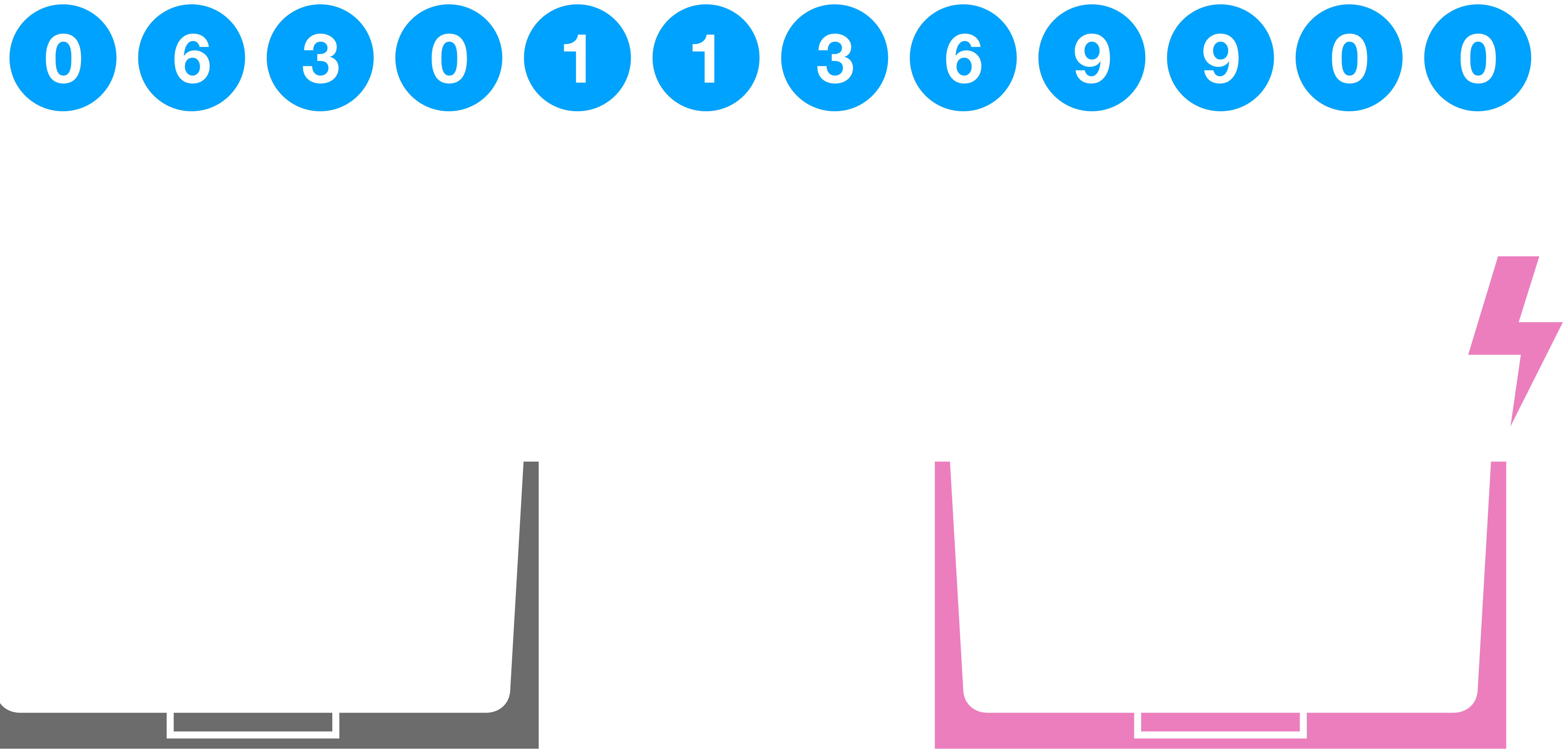
Independence of the Treatment

We need the treatment to be **independent** of other influences (i.e. **exogenous**)



Understanding Randomization via Sampling

Now imagine that posting the colored background has an effect

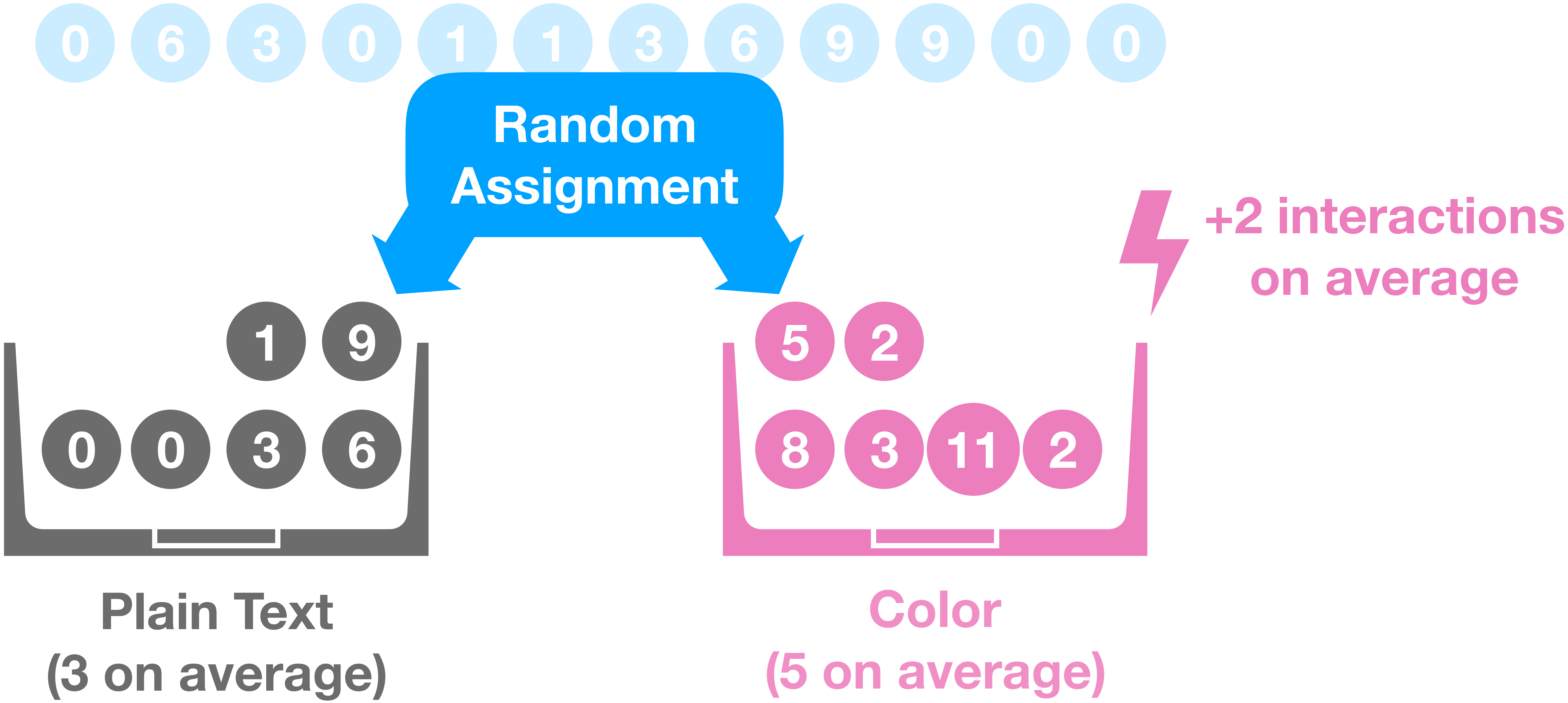


+2 interactions on average

Plain Text

Understanding Randomization via Sampling

Now imagine that posting the colored background has an effect

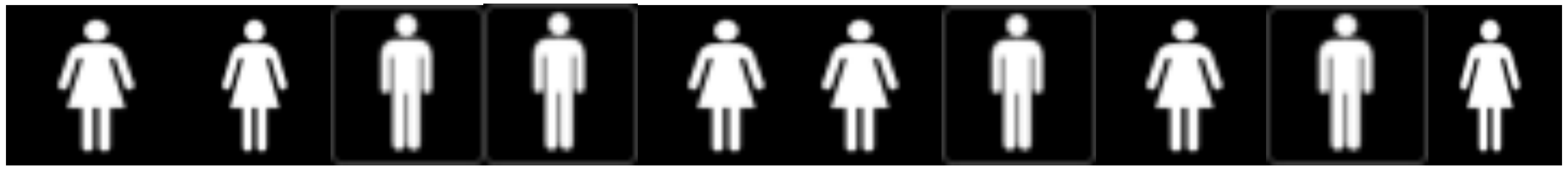


Common Methods of Random Assignment

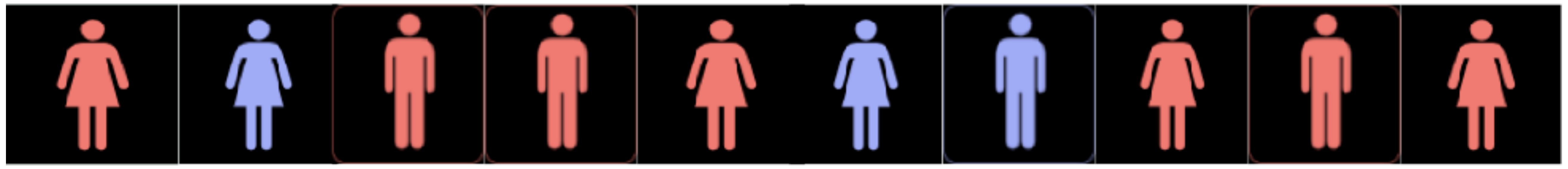
- **Simple randomization** (coin flips)
  - Problem: it's hard to get equal groups
- **Complete** (equal groups)
  - Example: sorted lists
- **Clustered**: (by group)
  - Randomized students by randomizing schools
- **Blocked**: (within groups)

Simple Randomization

Sample:



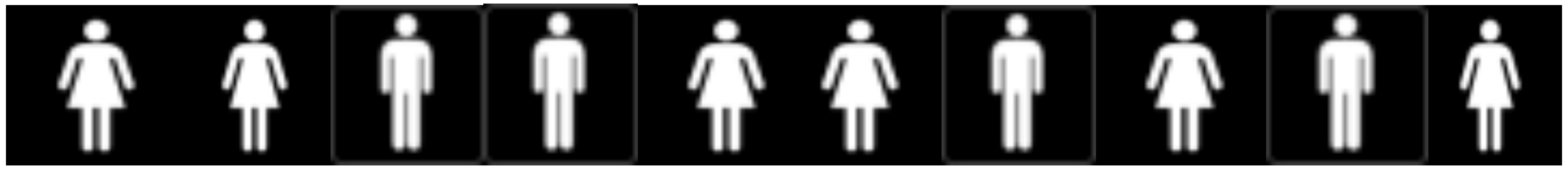
Iteration 1:



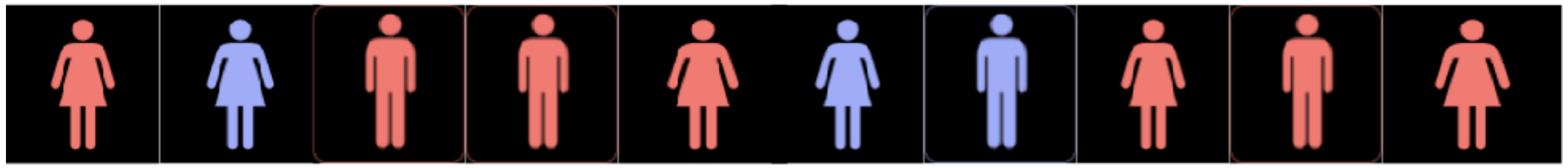
Simple Randomization

May not allocate the expected number of participants
to treatment and control

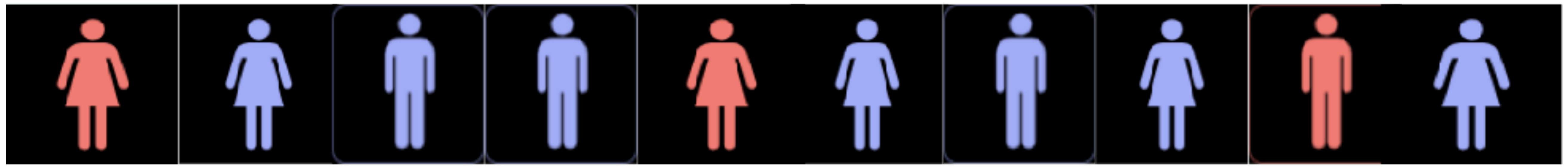
Sample:



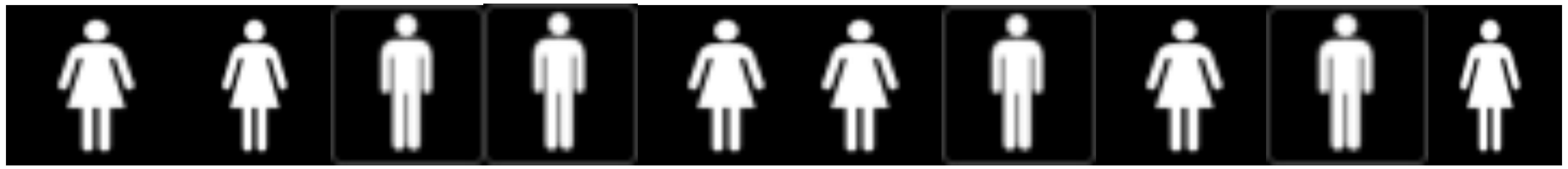
Iteration 1:

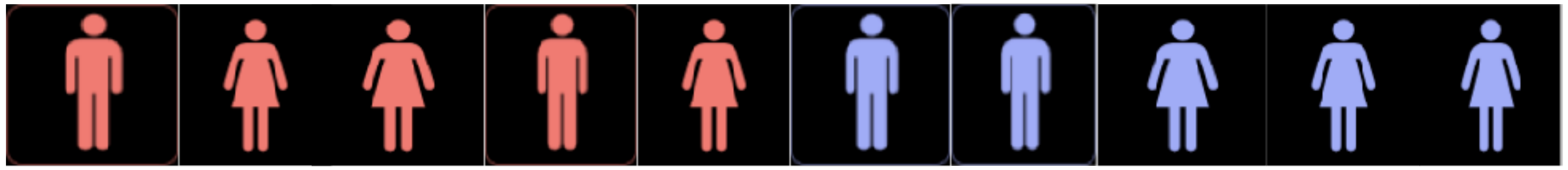


Iteration 2:



Complete Randomization

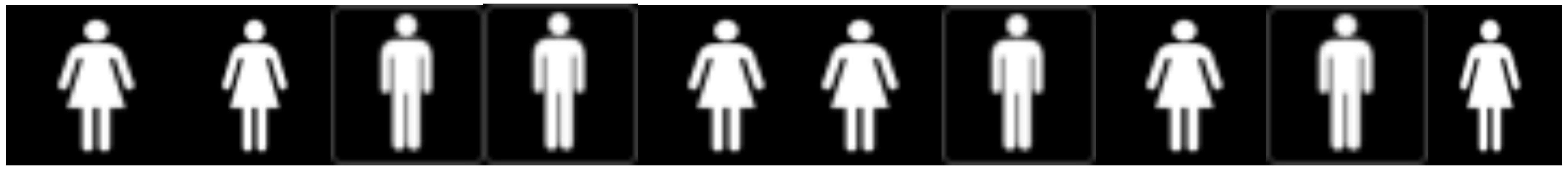
Sample: 

Iteration 1: 

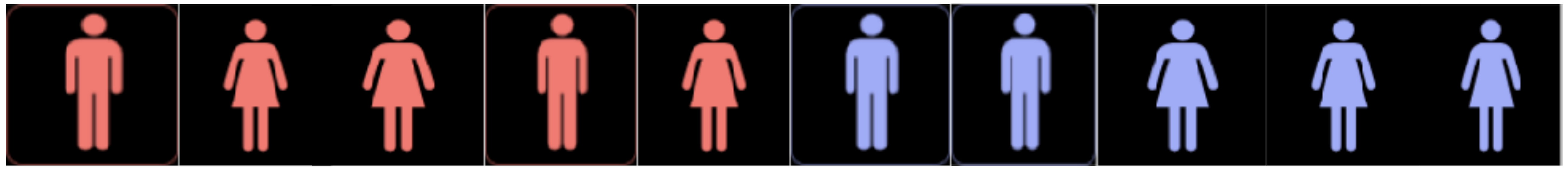
Complete Randomization

Always allocates the expected number of participants
to treatment and control

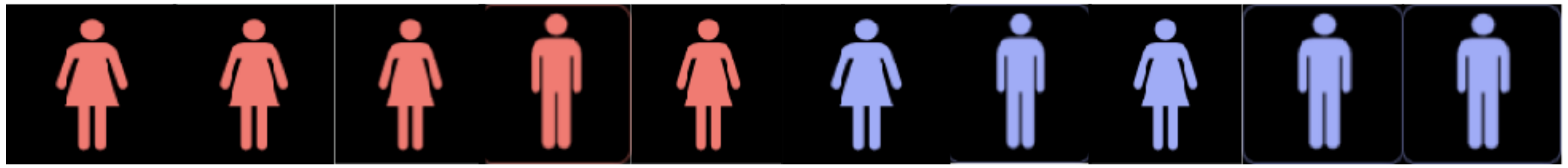
Sample:



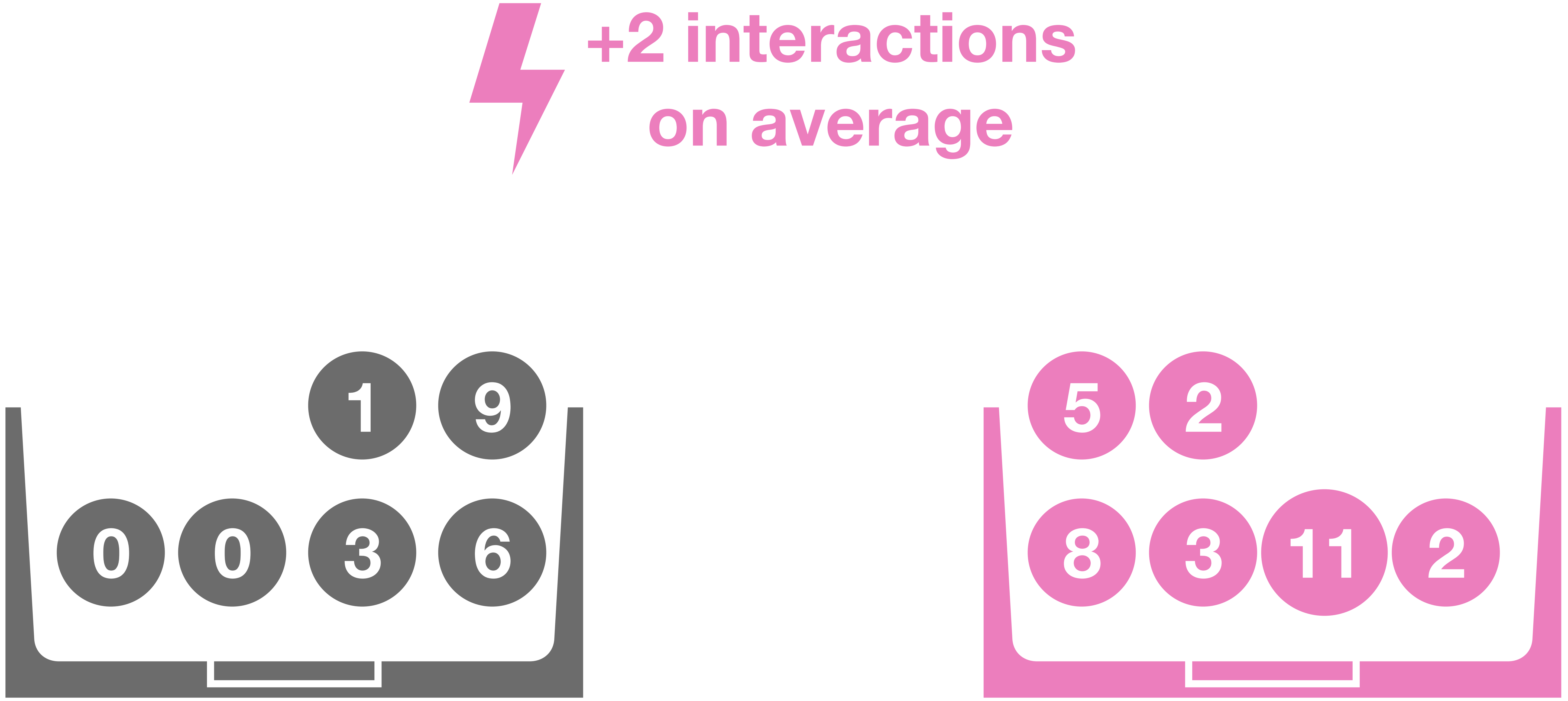
Iteration 1:



Iteration 2:



Average Treatment Effect (ATE)



Plain Text
(3 on average)

Color
(5 on average)

Average Treatment Effect (ATE)

$$Y = \alpha + \beta_1 X + \epsilon$$

$$Interactions = \alpha + \beta_1 Background + \epsilon$$

lm(interactions ~ condition, data=poems)

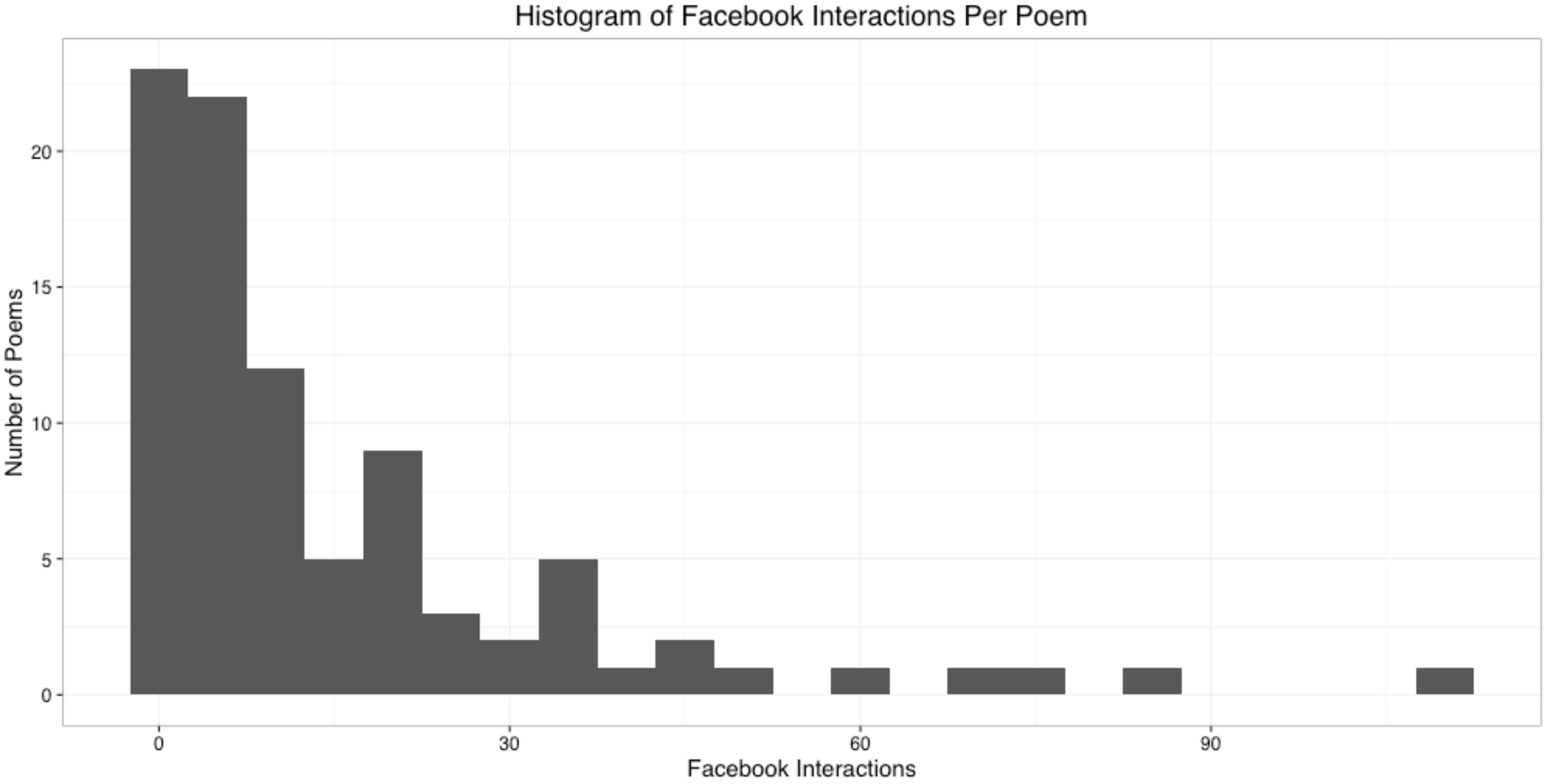
Example at github.com/natematias/SOC412/tree/master/1-facebook-poem

Average Treatment Effect (ATE)

```
Call:
lm(formula = interactions ~ condition, data = poems)

Residuals:
    Min       1Q   Median       3Q      Max
-17.578 -11.578  -8.578   3.922  99.289

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      11.711      3.008   3.893 0.000192 ***
conditionColor    7.867      4.254   1.849 0.067785 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.18 on 88 degrees of freedom
Multiple R-squared:  0.03741,    Adjusted R-squared:  0.02647
F-statistic:  3.42 on 1 and 88 DF,  p-value: 0.06778
```

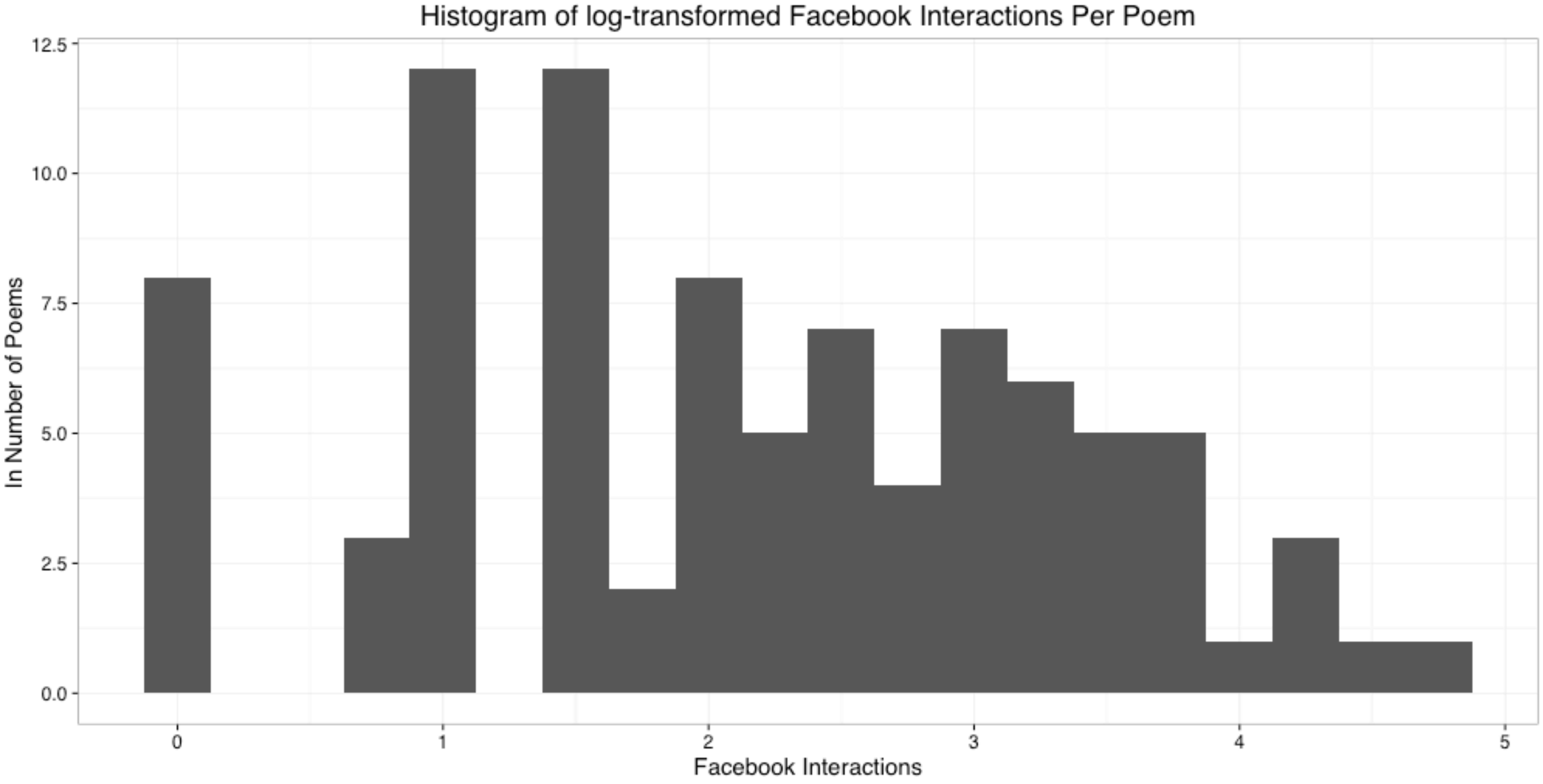
Example at github.com/natematias/SOC412/tree/master/1-facebook-poem

Average Treatment Effect (ATE)



Example at github.com/natematias/SOC412/tree/master/1-facebook-poem

Average Treatment Effect (ATE)



Example at github.com/natematias/SOC412/tree/master/1-facebook-poem

Average Treatment Effect (ATE)

(log-transformed dependent variable)

$$\ln(Interactions + 1) = \alpha + \beta_1 Background + \epsilon$$

lm(*log1p*(interactions) ~ condition, data=poems)

Example at github.com/natematias/SOC412/tree/master/1-facebook-poem

Average Treatment Effect (ATE)

(log-transformed dependent variable)

```
Call:
lm(formula = log1p(interactions) ~ condition, data = poems)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6856 -0.5869 -0.1596  0.7456  3.0329

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.6856     0.1635  10.309  < 2e-16 ***
conditionColor   0.9589     0.2312   4.147 7.74e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.097 on 88 degrees of freedom
Multiple R-squared:  0.1635, Adjusted R-squared:  0.154
F-statistic:  17.2 on 1 and 88 DF,  p-value: 7.737e-05
```

Example at github.com/natematias/SOC412/tree/master/1-facebook-poem

Average Treatment Effect (ATE)

(log-transformed dependent variable)

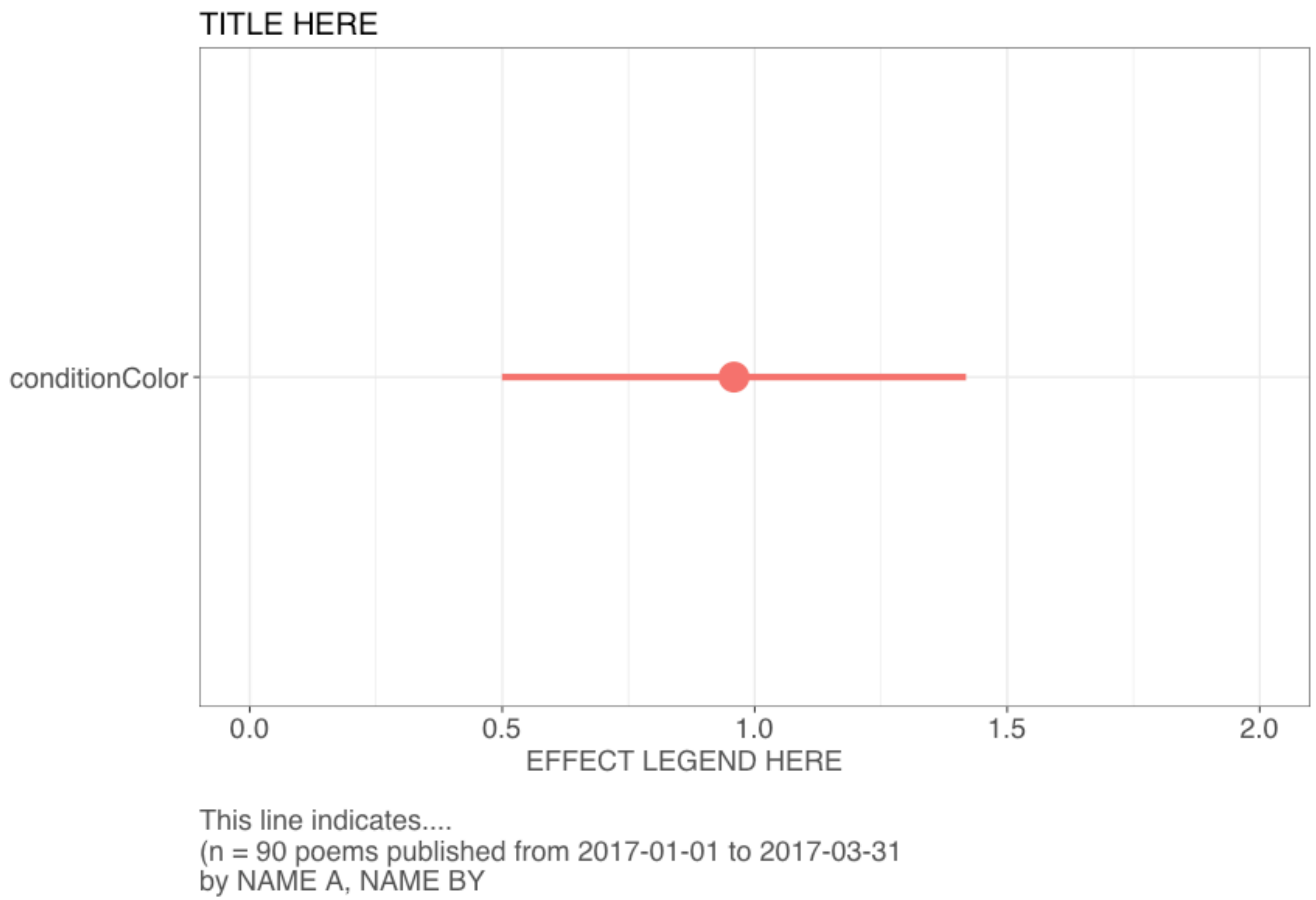
	Linear	Log-Transformed
Color	**7.87**	**0.96 *** **
	(4.25)	(0.23)
(Intercept)	11.71 ***	1.69 ***
	(3.01)	(0.16)
R^2	**0.04**	**0.16**
Num. obs.	90	90
RMSE	20.18	1.10

Linear models estimating log-transformed likes, comments, and shares per poem

Example at github.com/natematias/SOC412/tree/master/1-facebook-poem

Average Treatment Effect (ATE)

(log-transformed dependent variable)



Example at github.com/natematias/SOC412/tree/master/1-facebook-poem

Average Treatment Effect (ATE)

APPLIES TO POPULATIONS **NOT TO INDIVIDUALS**

+2 interactions
on average

Plain Text
(3 on average)

Color
(5 on average)

Assumptions of ATE

- **Random assignment** of participants to treatment
  - implies that receiving the treatment is statistically independent of participants' potential outcomes
- **Non-interference**: a participant's potential outcomes reflect only whether they receive the treatment themselves (not by others receiving it)

Assumptions of ATE

- **Excludability** a participant's potential outcomes respond only to the defined treatment, not other extraneous factors that may be correlated with treatment
  - importance of defining the treatment precisely and maintaining symmetry between treatment and control groups (e.g. through blinding)

Conspicuously Absent Assumptions

- Random sampling of subjects from a larger population is not a core assumption
- The issue of “external validity” is a separate question that relates to the issue of whether the results obtained from a given experiment apply to other subjects, treatments, contexts, and outcomes
- For now, we aim only to estimate the ATE in our subject pool

Upcoming Assignments

Why I am asking you to participate in an experiment:

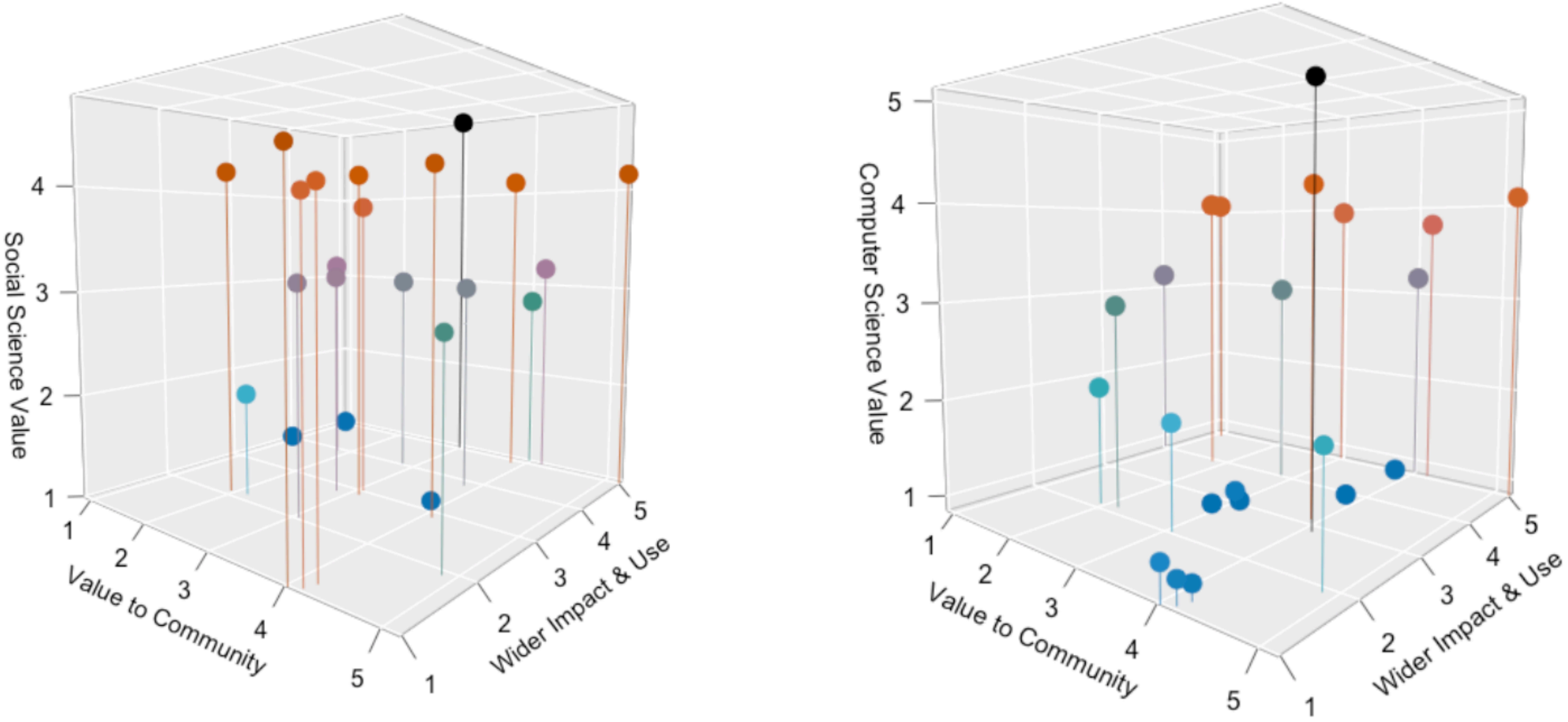
If you're going to be asking other people to participate in experiments, you need to at least be willing to think about research ethics in light of your own experience and the people you know.

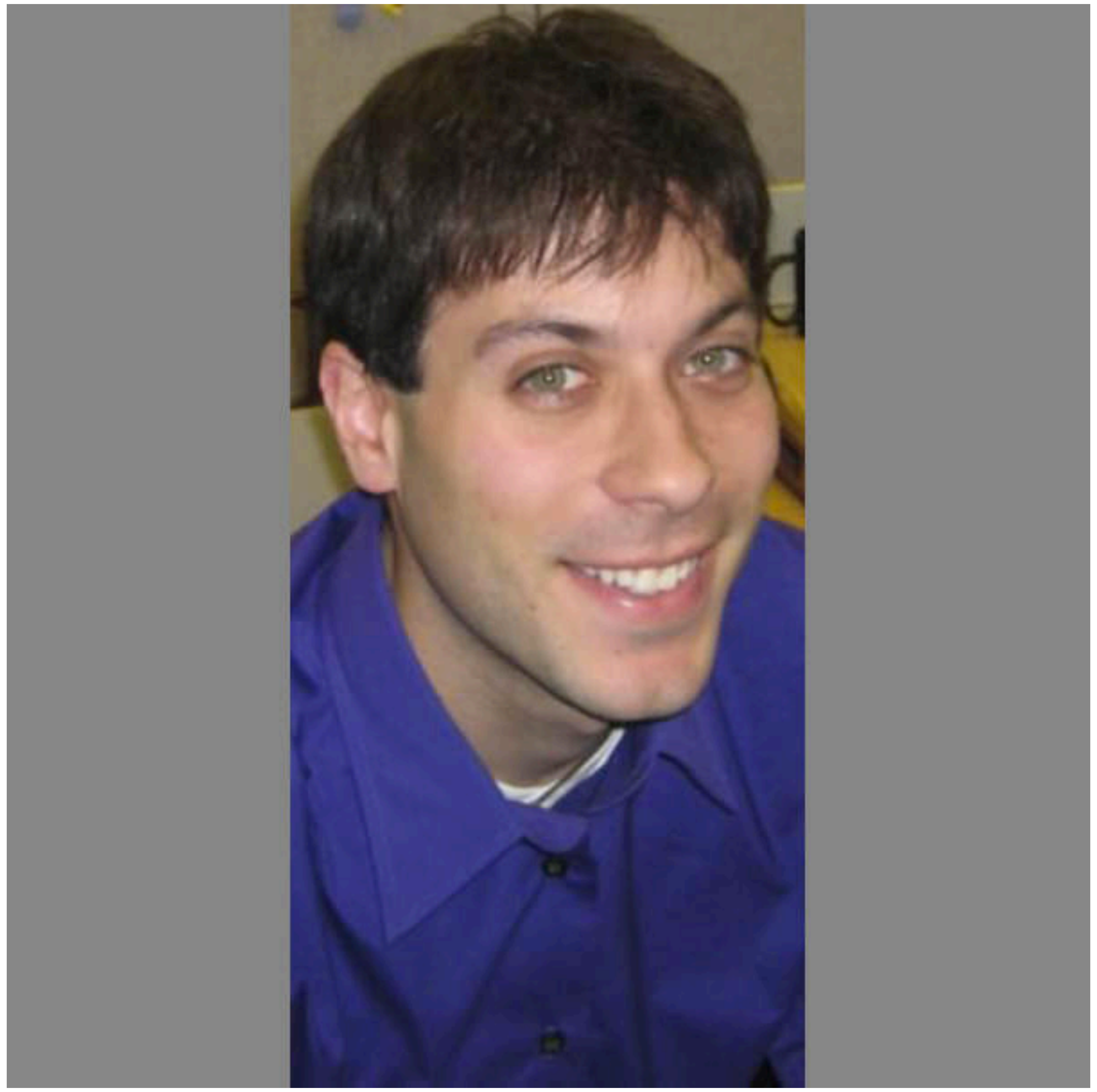
Upcoming Assignments

- Facebook Color Experiment
  - https://github.com/natematias/SOC412/tree/master/2-facebook-color
- Cornhole Experiment
  - https://github.com/natematias/SOC412/tree/master/2-cornhole-challenge

Week 2: Studying Online Behavior at Scale (SOC412)

Research proposed at the CivilServant Research Summit could **Transform Communities**, **Benefit Society**, & **Grow Knowledge** in the **Social Sciences & Computing**





Eric Pennington

Research Manager

Class Projects

- Supporting people who **post about sexual assault online**
- **Preventing harassment** by posting rules in online communities
- **Improving newcomer retention** with a guestbook
- **Reduce the influence of harmful norms** on people’s behavior
- Improving moderator accuracy with training

Weekly Rhythm

- **Post to Piazza with one observation for the upcoming discussion by Friday at 5pm**
- **Over the weekend, post at least one response to someone else's observation**
- **Submit lab assignment by Monday at 5pm**
- Office Hours Mon/Wed 10-11 (or Tues if necessary)

https://meetme.so/natematias-soc412

Components of leading a good class discussion

- Provide a summary of the material (5 minutes)
- Pose some questions
- Support the conversation

Your group should sign to lead two sessions:

http://bit.ly/2EkggoH

Behavioral Science & Public Policy Center

Thursday, February 22

Behavioral Policy Speaker Series

Kahneman-Treisman Center for Behavioral Science & Public Policy

Personalized Policies

Jens Ludwig, University of Chicago

4:30 p.m.

Robertson Hall, Bowl 002

https://behavioralpolicy.princeton.edu/S18Ludwig

https://behavioralpolicy.princeton.edu/behavioral-policy-interest-list

References to Know