S4 Internal and External Validity

Sanity Checks, SRM, A/A tests

Shan Huang, HKU

Review: Internal vs. External Validity

- Internal Validity: the correctness of the experiment results *without* attempting to generalize to other populations or time periods.
- External Validity: the extent to which the results of a controlled experiment can be generalized to other populations (e.g., other countries, other platforms) or time periods (e.g., will the 2% revenue increase continue for a long time or diminish).

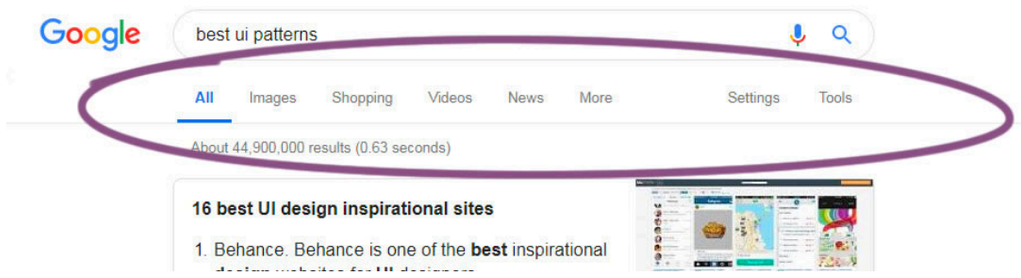
Internal vs. External Validity

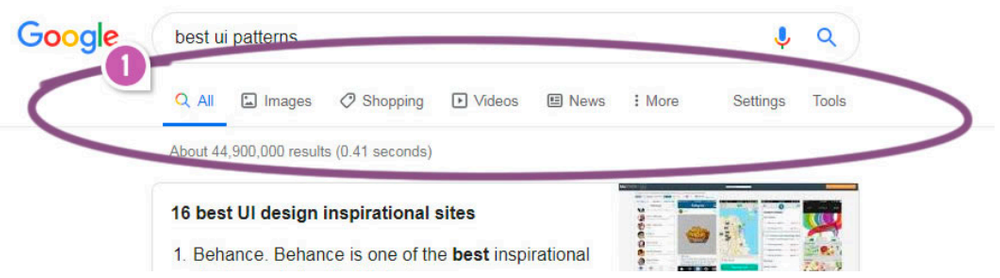
Treatment: Blue background
Control: Green background

- Internal Validity: The treatment effects are **100%** caused by the change of background colour. We need to guarantee the internal validity.
- External Validity: The treatment effects can be applied to other websites, user groups, and time periods. We don't need perfect EV, which really depends on our needs.

Internal Validity

- The degree of confidence to establish a trustworthy *cause-and-effect* relationship between a treatment and an outcome.
- The degree of confidence to attribute the difference between treatment and control to the treatment.
- **Example:** Google run an experiment to test whether including icon in addition to the labels would increase users' clicks on the navigation buttons.





- You should be confident about that the difference, $\Delta = CTR_1 - CTR_0$, is *100% caused* by the treatment - including the icons.

Can you attribute Δ (δ) to the Treatment?

1. If the users in control group are older than those in the treatment group (different users' characteristics)
2. If the users in control group are more active on Google than those in the treatment group (different users' behaviors)
3. If the users in control group see a different background colour from those in the treatment group (different "Treatments")
4. If the users in the control group sometimes use their friends' accounts, which are assigned to the treatment group.

Control and Treatment Groups should be indifferent in all the other aspects except for receiving different treatments.

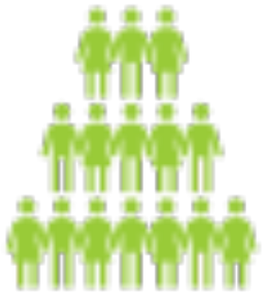
In this way, we can safely attribute Δ to the Treatment.

Sanity Checks

- Sanity checks are to make sure the experiment (e.g., randomization) was run properly (mainly for internal validity).
- *Guardrail Metrics* are critical metrics that are designed to alert experimenters about a violated assumption.
- When a Treatment effect unexpectedly moves a guardrail metric,
    - You may want to debug, or stop an experiment.

Sanity Checks

- Trust-related guardrail metrics
    - Samples to be sized according to the design (Sample Ratio Mismatch -*SRM*)
    - The variables, which are not impacted by the treatment, should be indifferent across variants.
        - User demographics (age, gender, city)
        - Variables measured by the data before the experiment.
            - Users' historical behaviors

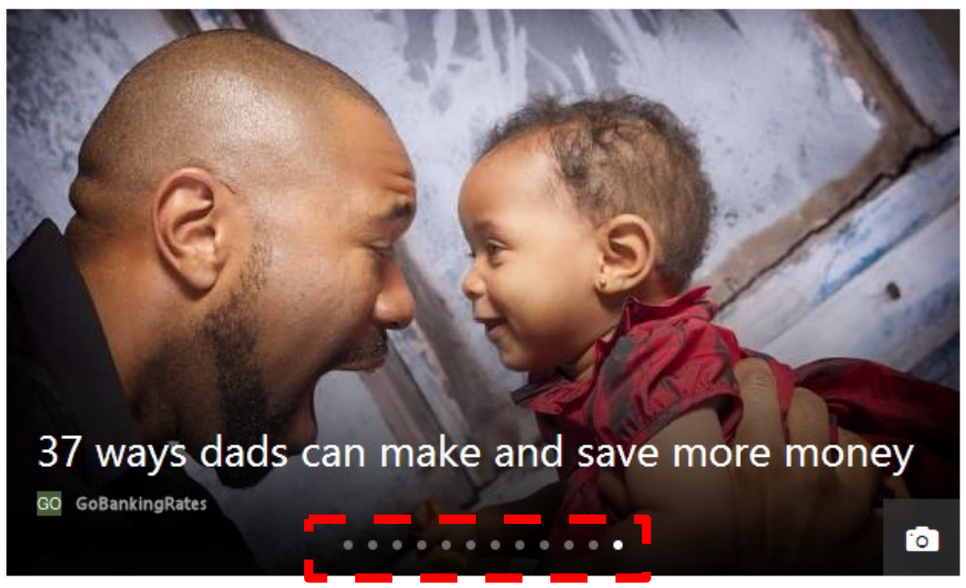


Challenge: Sample Ratio Mismatch

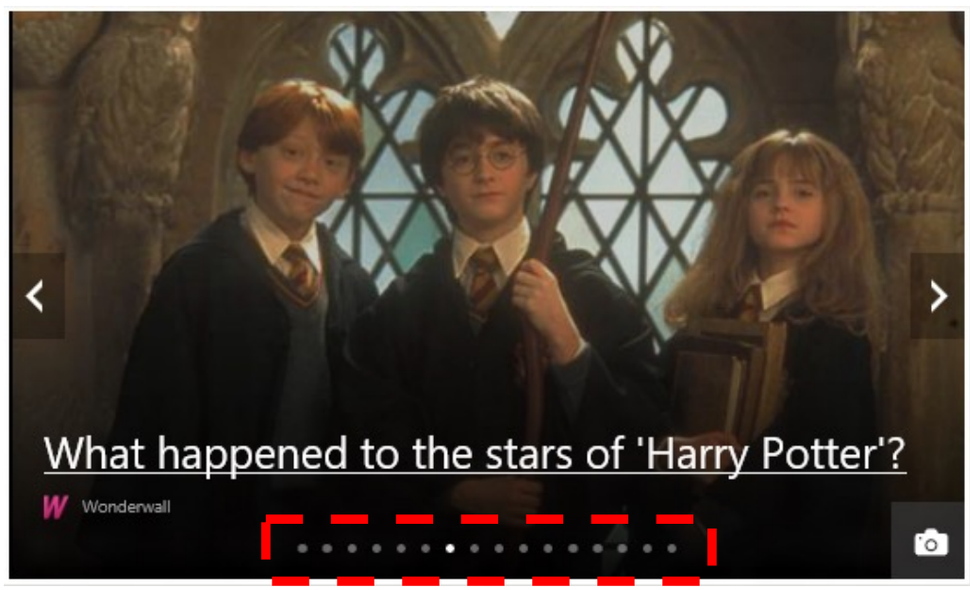
- It's easy to lose data on users. For example, if you expect to experiment on 10%/10% you should have equal numbers of users in Treatment and Control.
- Or if you do a 20/80% traffic split for the treatment and control, you find the sample size ratio is not 1/4.
    - Unfortunately, it happens.
- This indicates something wrong with your experiment.

Real Examples

- The user engagement decreased because of the treatment. Surprising!
- There is a SMR in the experiment.
- The ratio is 0.992 instead of 1 (expected ratio).
- Bots were removed from the experiment.
    - Incorrectly classified the most heavy users as bots and removed them.



Control: 12 Slides



Treatment: 16 Slides

Sample Ratio Mismatch (SRM)

- Sample Ratio = #Control/#Treatment
  - If the randomization is correct *(before and after the data cleaning)*, the observed sample ratio should be statistically indifferent to the expected sample ratio.
- Otherwise, there must be something *wrong* with the randomization or data collection procedure.

Sample Ratio Mismatch (SRM)

- The SRM should be included for every experiment to ensure the internal validity.
- The SRM should be the first/easiest thing to check after the experiment.
- SRM happens much more than people expected.
- If you find SRM happens, you should stop analyzing the treatment effects and *debug*. Otherwise, the analysis will be heavily biased.
- Facebook, Linkedin, Google and many other companies emphasize the importance of SRM tests.

SRM Tests

- Chi-square Homogeneity Tests
- t-tests (z-tests)

Chi-square Homogeneity Tests

- H0: the proportion of Control Group identical to the expected proportion (e.g., 50% for equal split). Similarly, for the other groups.
- H1: At least one of the null hypothesis statements is false.
- Chi-square Statistics:
  - $\chi^2 = \sum (O_i - E_i)^2 / E_i$
  - O_i = observed sample size of group (variant) i
  - E_i = expected sample size of group (variant) i
  - Degree of freedom = k(#groups) - 1

Chi-squared Test for SRM

- For example, we allocate 50% traffic to control and treatment.
- If p-value is very low (we use < 0.05), then you very likely commit a sample-ratio mismatch and the experiment results are questionable.

Scenario 1

	Proportio	Observed	Expected
Control	0.5	445,000	447,500
Treatment	0.5	450,000	447,500
SUM	1	895,000	895,000
Chi-square: p-value	1.25592E-07	Rejected the Null	

895000*
50%

SRM: t-tests

- Consider the random assignment process a Bernoulli trial with $p = 0.5$ for example.
- Null Hypothesis: $p = 0.5$.

$$\sigma^2 = p(1-p) = 0.25$$

$$t = \frac{\Delta}{se}$$

$$se = \frac{\sigma}{\sqrt{n}}$$

	Proportion	Observed N	Observed p
Control	0.5	445,000	0.497206704
Treatment	0.5	450,000	
SUM	1	895,000	
t: p-value	1.25621E-07	Rejected the Null	

Class Exercise

- Random Assignment with p = 0.2
- 80% Treatment/20% Control
- Please use t-test to conclude SRM for the following experiment

	Observed N	Expected N	Expected p
Count in Control	20,041	20,040	0.2
Count in Treatment	80,159	80,160	0.8
Sample Size	100,200	100,200	1

Scenario 2

- You allocate 10% to Treatment 1 , 10% to Treatment 2, and 80% to Control Group.
- We usually do not put too many users into Treatment Groups.

	proportion	Observed	Expected
Treatment 1	0.1	10,800	10737
Treatment 2	0.1	10,570	10737
Control	0.8	86,000	85896
SUM	1	107,370	107,370
Chi-square: p value	0.21298759	ACCEPT NULL	

Scenario 3

- If you find the p value is small but not too small, you may conduct more randomization checks (you will learn soon).

	Proportion	Observaed	Expected
Control	0.1	11,000	10757
Treatment	0.1	10,570	10757
Outside	0.8	86,000	86056
SUM	1	107,570	107570
Null Hypothesis	There is no difference between observed and expected group sizes.		
Chi-squared	0.012422	Reject or Accept the Null ?	

SRM Causes

Browse Redirections

- A very common mechanism for A/B tests is to redirect the Treatment to another page.
- Treatment groups often suffer an extra redirect and delay user experience, causing loss of units in the Treatment groups.
- Users may send the redirect link to their friends, causing interferences.
- Robots handle redirects differently.
- Avoid redirects in A/B tests or at least make sure the redirect costs are the same across groups.

SRM Causes

Unequal Drop of the Data Across Groups

- You always need to clean data after the experiment, e.g., missing attributes for some users.
- However, the users dropped from different groups may not be equal
    - e.g., the data was dropped more from the control group than treatment group.

SRM Causes

Triggering based on attributes impacted by experiment

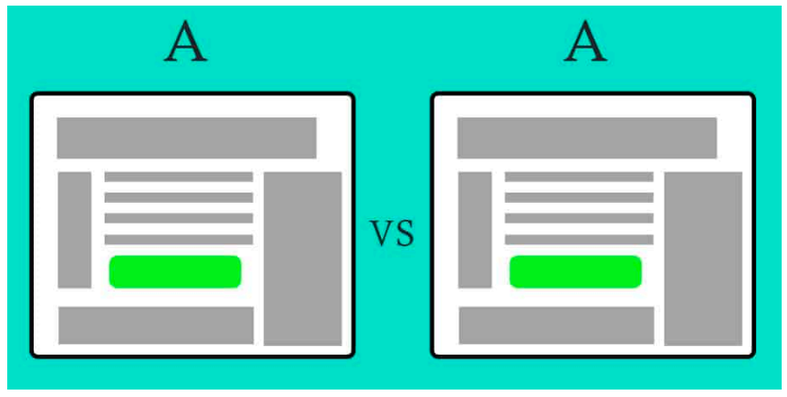
- Suppose you are running a campaign on inactive users based on an activeness attribute.
- However, the treatment can make the inactive user more active and not an inactive user anymore.
- The trigger condition should be uncorrelated with the treatment.
    - The state of the dormant attribute should be fixed before the experiment.

SRM Causes

Residual Effects

- Experiments are sometimes restarted after fixing a bug.
- However, if the bug is serious enough, users in the bugged group tend to abandon the treatment, causing SRM.
- When the experiment is visible to users, there is a desire not to re-randomize the users.

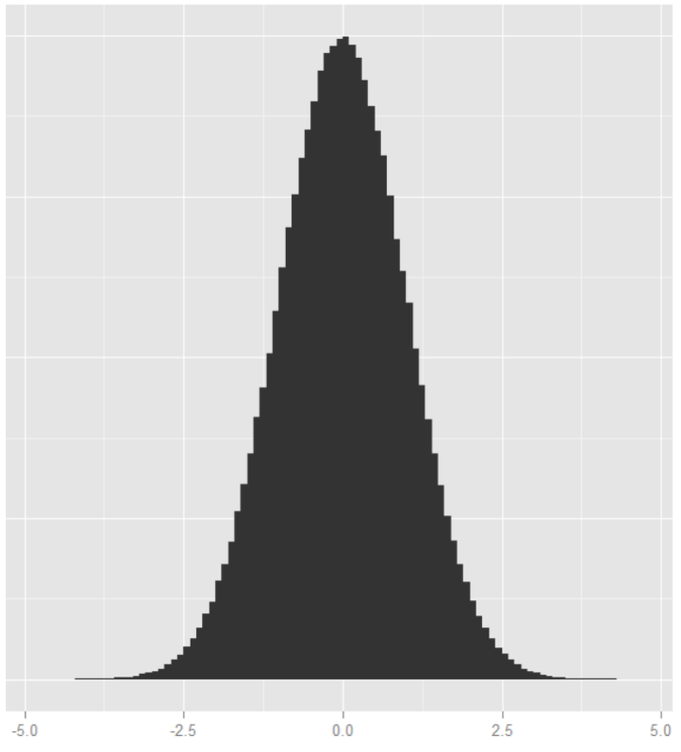
A/A Tests



- Randomly split the users into two groups as in a regular A/B test but make B identical to A (A/A).
    - There is no Treatment in A/A tests.
    - The *goal* of an A/A test is to find *no difference* between your control and treatment groups.
    - A/A test is highly useful to establish trust in establishing your experimentation platform.
    - AA - AB tests or AAB tests

Challenge: How random is random?

- Causality will only be established if the split is random. If it is not random, there is any hidden bias and can affect your results.
- If we do 1 million random assignments of the same sample, what is the distribution of Δ?



Observed differences between two groups
1 million random assignments

A/A Tests

1. Checking the accuracy of an A/B Testing platform.
    - Check random assignments
    - Interferences between overlapping experiments
    - Check data collection (logging system)
2. A sanity check before a new experiment
    - Randomization
    - Normality assumption, standard errors estimation.
3. Setting a baseline conversion rate for future A/B tests
    - Estimate the metrics for control group
4. Deciding a minimum sample size
    - How many units, and how long for the experiment

Threats to A/A Tests

- The threats to SRM
- The threats to Randomization
- The distribution is skewed
- Outliers

How to Decide the Sample Size

- To achieve 80% statistical power and 5% significance level; the variants are assumed to be of equal size
    - The number of sample in each variant is:

$$n = 16\sigma^2/\delta^2$$

σ^2 = population variance

We can estimate it by A/A tests

δ = practically significant effect size (the amount of change you want to detect between control and treatment .)

We can estimate a baseline using A/A tests. For example, baseline metric = a, we want a 10% increase. δ = 10%*a

Class Exercise - How to Run A/A Tests

- Randomization Check
    - Use data sample_data_aatest.csv
    - Multiple Testing (Consider Type I error)
    - Visualize the data

```
# Mutiple Testing
def multi_cm(x):
    x0 = df[df['expid'] == 0][x]
    x1 = df[df['expid'] == 1][x]
    x2 = df[df['expid'] == 2][x]
    cm01 = sms.CompareMeans(sms.DescrStatsW(x0), sms.DescrStatsW(x1))
    cm02 = sms.CompareMeans(sms.DescrStatsW(x0), sms.DescrStatsW(x2))
    cm12 = sms.CompareMeans(sms.DescrStatsW(x1), sms.DescrStatsW(x2))
    print(x)
    print(cm01.tconfint_diff(alpha=0.05, alternative='two-sided', usevar='pooled'))
    print(cm02.tconfint_diff(alpha=0.05, alternative='two-sided', usevar='pooled'))
    print(cm12.tconfint_diff(alpha=0.05, alternative='two-sided', usevar='pooled'))
    return

for feature in df.columns[2:]:
    multi_cm(feature
```

Class Exercise

- Estimating baseline metrics and their variances, σ
- $\delta = 5\% * Metric_0$
- Metrics are Click, Like, Like Count, Comment Count.
- Find δ for different metrics.

```
like0=np.mean(df.like)
var_like0=np.var(df.like)
```

```
click0=np.mean(df.click)
var_click0=np.var(df.click)
```

```
sns_like_cnt0=np.mean(df.sns_like_cnt)
var_sns_like_cnt0=np.var(df.sns_like_cnt)
```

```
sns_comment_cnt0=np.mean(df.sns_comment_cnt)
var_sns_comment_cnt0=np.var(df.sns_comment_cnt)
```

Class Exercise

- Power Analysis - Choose sample size for different metrics.
- $n = 16\sigma^2/\delta^2$

Power Analysis

click

```
delta = 0.05*click0
print(delta)
```

```
0.0014185
```

Sample Size

```
size = 16*var_click0/(delta**2)
```

```
print(size)
```

```
219190.4124070348
```

Other Threats to Internal Validity

- Violation of SUTVA
- Survivorship Bias

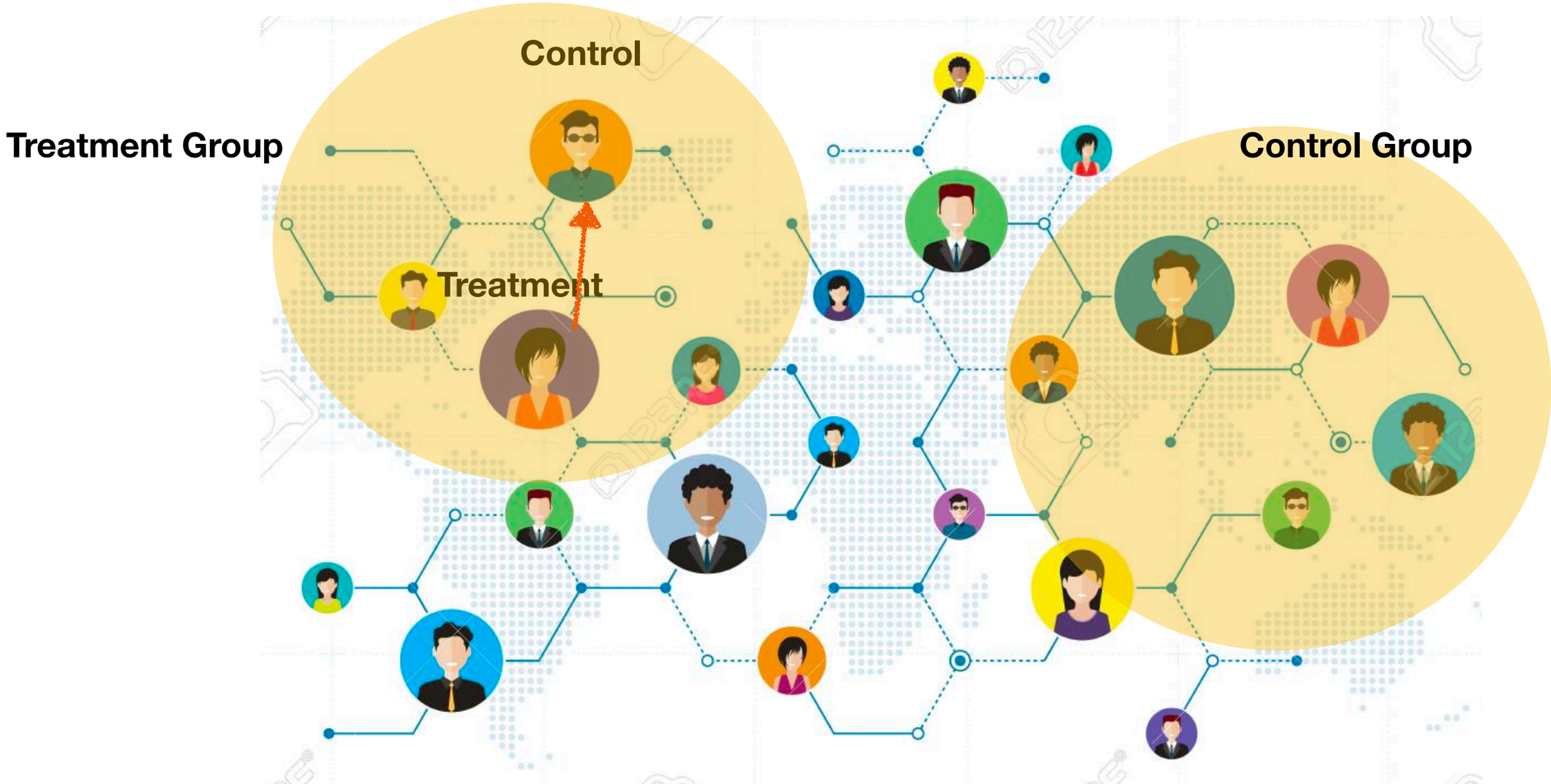
Violations of SUTVA

- **SUTVA:** Stable Unit Treatment Value Assumption *(Imbens and Rubin 2015)*
    - Experiment units (e.g., users) do not interfere with one another.
        - Their treatments are not impacted by others' treatments.
- The response of a particular unit depends only on the treatment to which he himself was assigned, not the treatments of others around him.
- However, violations of SUTVA often happen.

Violations of SUTVA

SUTVA can be violated in (not limited to) the following settings:

A. Social networks, where a feature might spillover to a users' friends.



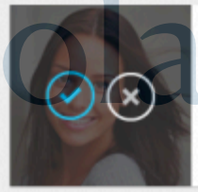
Search for people, jobs, companies and more... Advanced

Home Profile Contacts Jobs Interests Business Solutions | Upgrade

Interested in Mobile? - Real User Monitoring For Your Mobile App w/ New Relic! Free Sign Up & Shirt

People You May Know

Pending invitations



Carrie Westlake
Content editor at Axius Communications
San Francisco Bay Area
73

Gordon Lambert
Graduate student at New University
Greater New York City
43

Paul McCarthy
Product Strategist and Entrepreneur
San Francisco Bay Area
10

See more

Samuel Wang 57
Product manager at Atlas
San Francisco Bay Area

Christine Vitale 55
Senior data scientist at KMD Corporation
San Francisco Bay Area

Stella Cole 60
Global customer support at Idea
San Francisco Bay Area

Paul Thompson 78
Research assistant at Stanford University
San Francisco Bay Area

Justin Liu 80
Account Manager at Citi
San Francisco Bay Area

Lilian White 13
Head of Design/UX
San Francisco Bay Area

Sharon Lucas 63
External Communications at IBG International
San Francisco Bay Area

Jesse Vo 20
Global customer support at Eastern Technology
Washington D.C. Metro Area

Louis Hills 45
Product Marketing Manager at Nova
San Francisco Bay Area

Joanna Park 90
Investment analyst at Paradise Management
Greater Boston Area

Heather Criscuolo 33
Designer
San Francisco Bay Area

Andrew Iriondo 24
User Researcher at Good Technologies
Greater New York City

Joseph Randall 14
Software engineer at Evergreen
Greater New York City

Dan Cleveland 23
Quant analyst at Partners
San Francisco Bay Area

Irene Baoler 9
Investment analyst at Paradise Management
Greater Boston Area

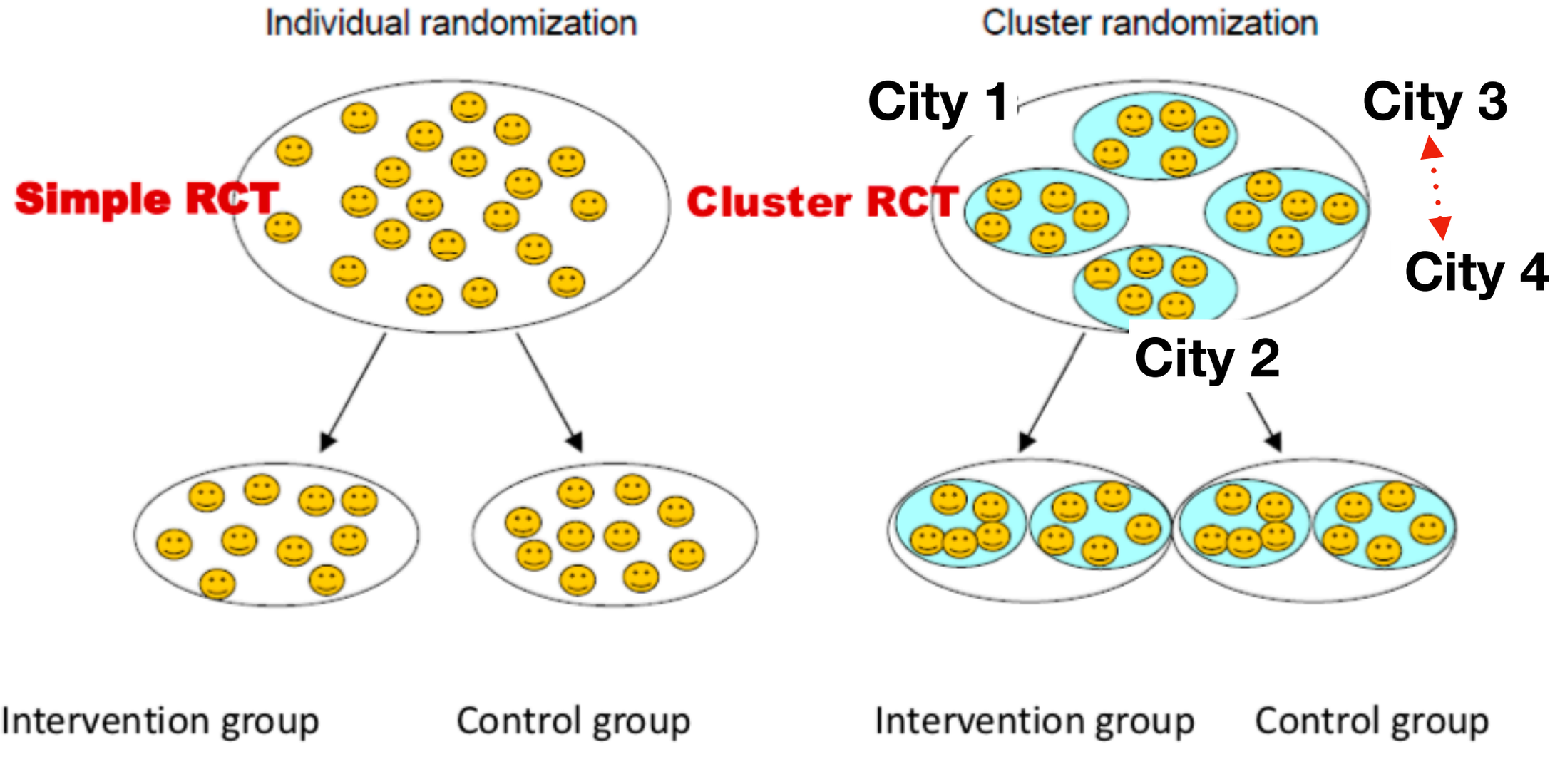
Violations of SUTVA

- Test “People You May Know” function @ Linkedin on # Users’ connections
- A is a friend of B but A is Treated but B is not
- If A receives recommendations of friends (e.g., B), A may connect with B.
- Once B receives the notification, B may actively discover more people to connect with.
- The treatment spills from A to B (in some ways).
- Will the treatment effects under/over estimated?
- Solution: Randomization at Clusters

Violations of SUTVA

SUTVA can be violated in (not limited to) the following settings:

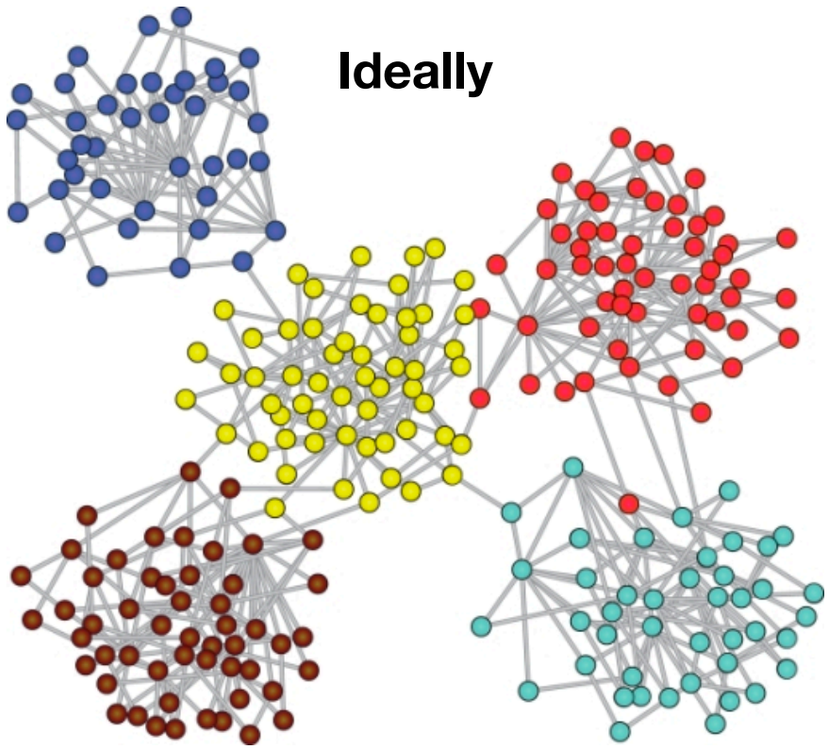
A. Social networks, where a feature might spillover to a users' friends. Solution: Change the Randomization Unit

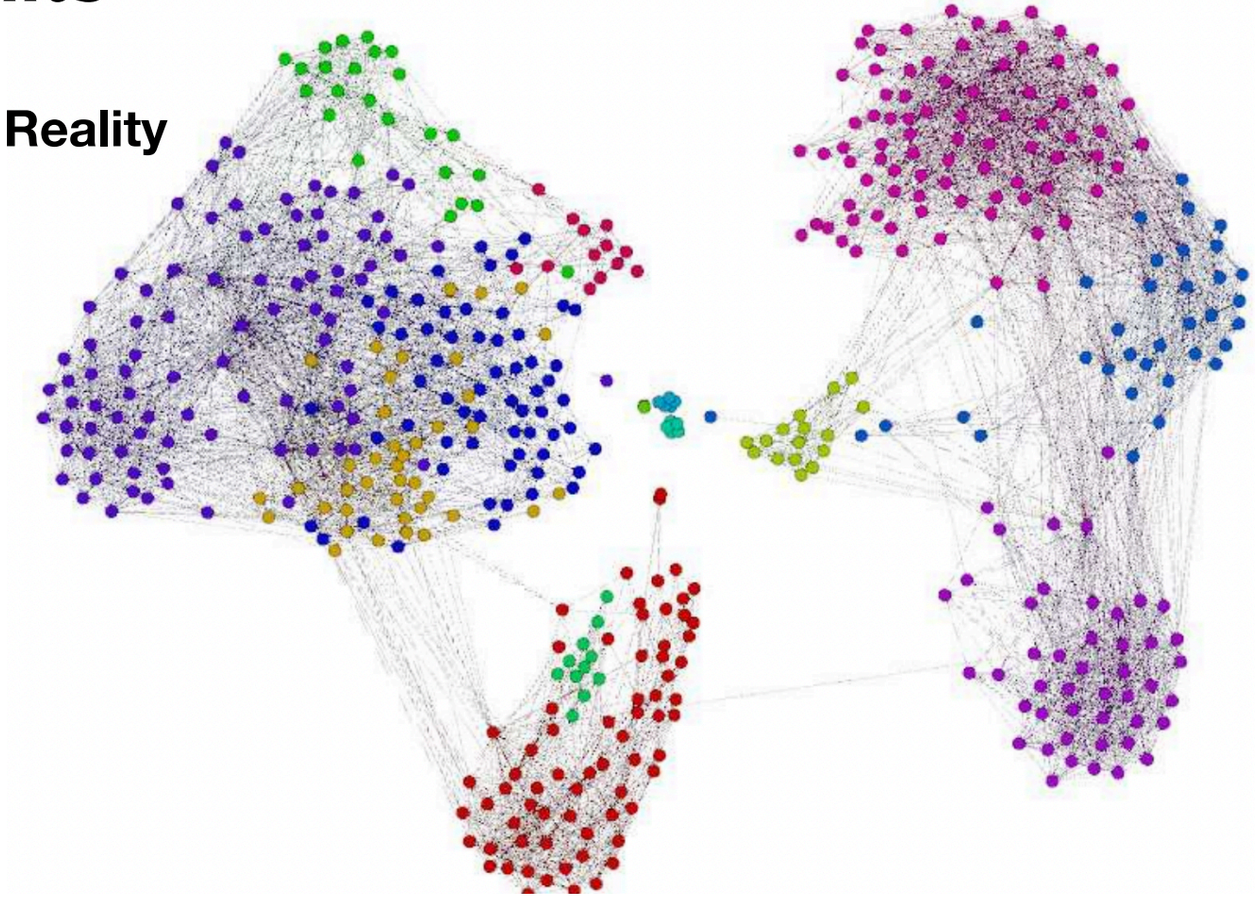


Violations of SUTVA

Solution - Change the Randomization Unit

- Cluster (community) detection
- Randomize at cluster units





Gargi, Ullas, et al. "Large-scale community detection on youtube for topic discovery and exploration." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5. No. 1. 2011.

Violations of SUTVA

SUTVA can be violated in (not limited to) the following settings:

A. Social networks, where a feature might spillover to users' friends.

B. Skype: the treatment can be spilled from senders to receivers.

C. Two-sided Platforms

D. Shared resources: a new feature can slow down the whole system affecting both treatment and control.

Survivorship Bias

- Only some of the subjects in the treatment group adopt the treatments and are treated.
- **Example:** WeChat aims to test whether larger font in the Moments can increase the user engagement. They allocate 10% traffic to each group.

Control

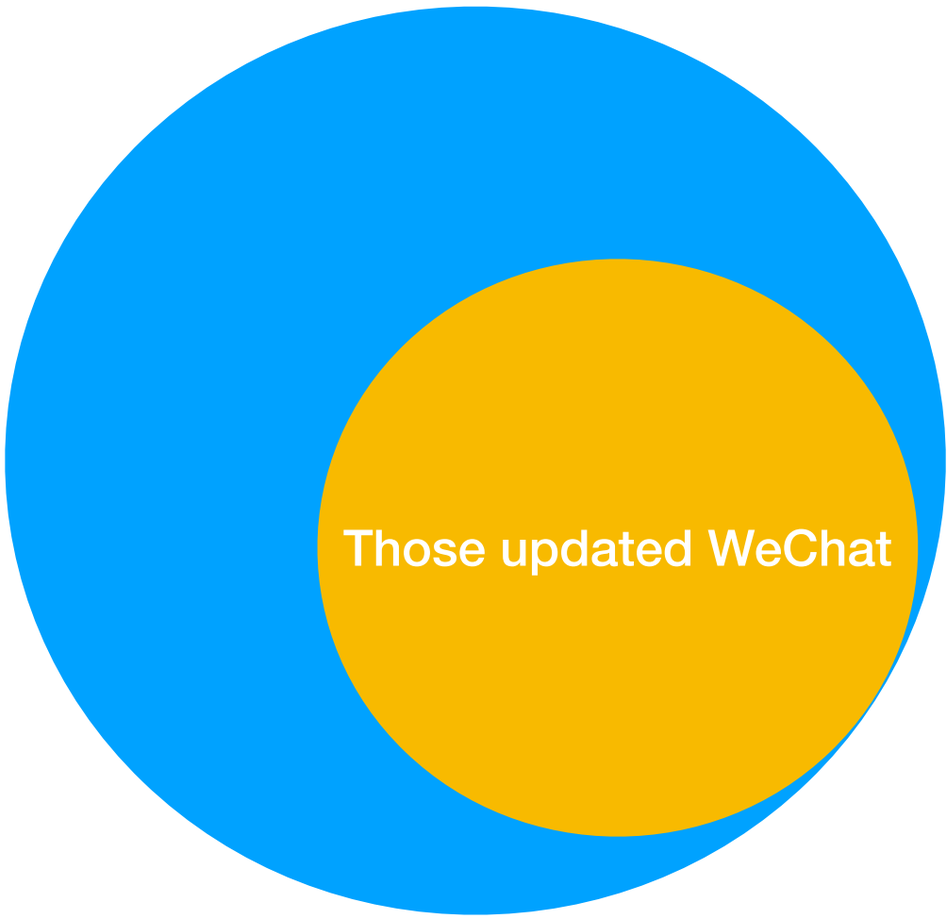
Treatment

- However, users need to update WeChat to the latest version to get treated.
- Some users never update their WeChat during experiment.

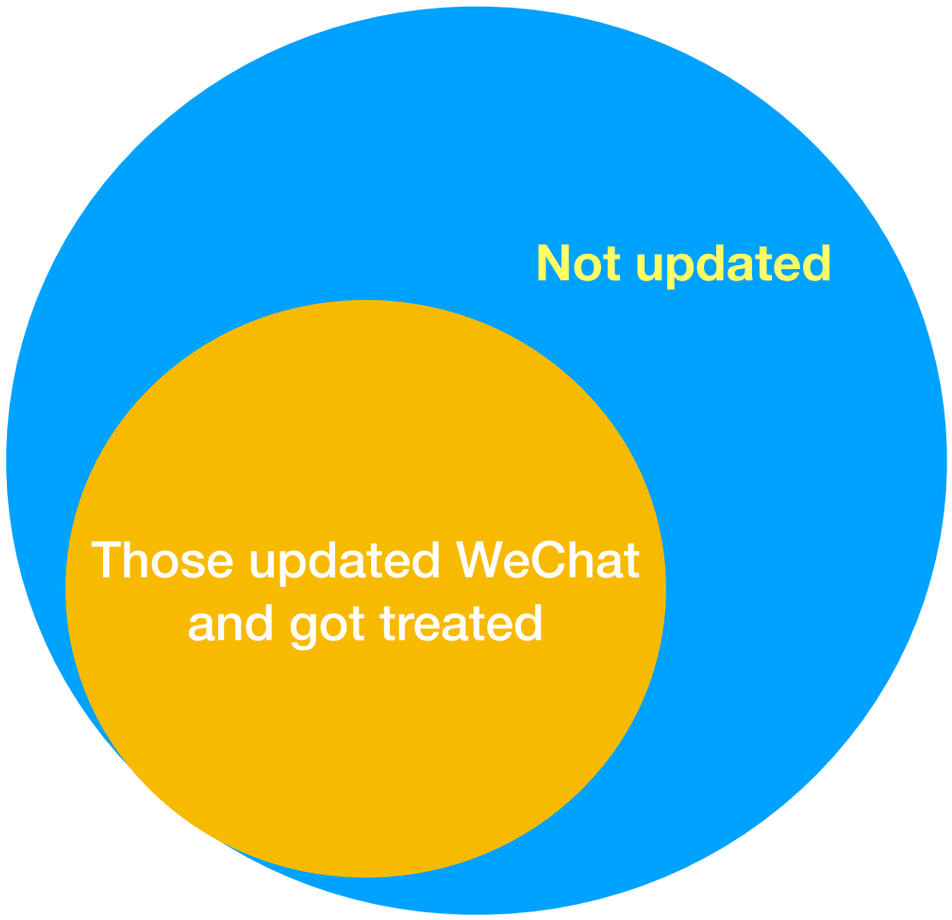
Survivorship Bias

- Will you compare the treatment and control directly?
- Will you compare the users who indeed received the treatment and the control group?
- Are there better ways?

Control Group



Treatment Group



Intention-to-Treat

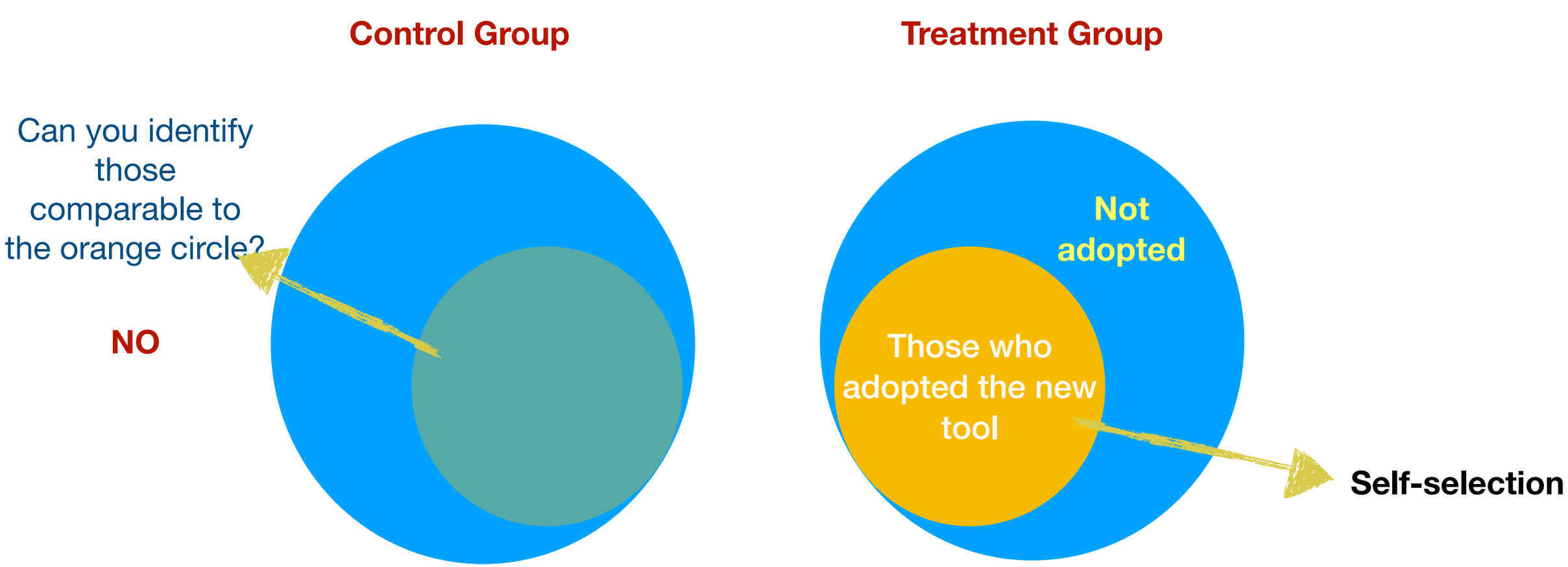
- Non-random attrition from the variants.
    - The users who update WeChat tend to be more active than those who do not.
- ***Solutions:***
    - Check whether the attrition is random. If it is not:

A. If the condition of attrition can be identified, you can analyze only the survivors i*n both treatment and control groups*.
    - You can only analyze the users who update WeChat in *control and treatment*. However, you should be aware that the treatment effects can only apply to those who update WeChat (e.g., more active users).

B. If the condition of attrition can not be identified.
    - We can measure the treatment effects based on *the offer of the new feature*, intention-to-treat, which is the Treatment of Offering the New Feature.

Intention-to-Treat

- Taobao provided a better tool for the sellers, and wanted to test whether it could improve the sales. *However, some of the sellers never adopted the new tool during the experiment.*



Intention-to-Treat

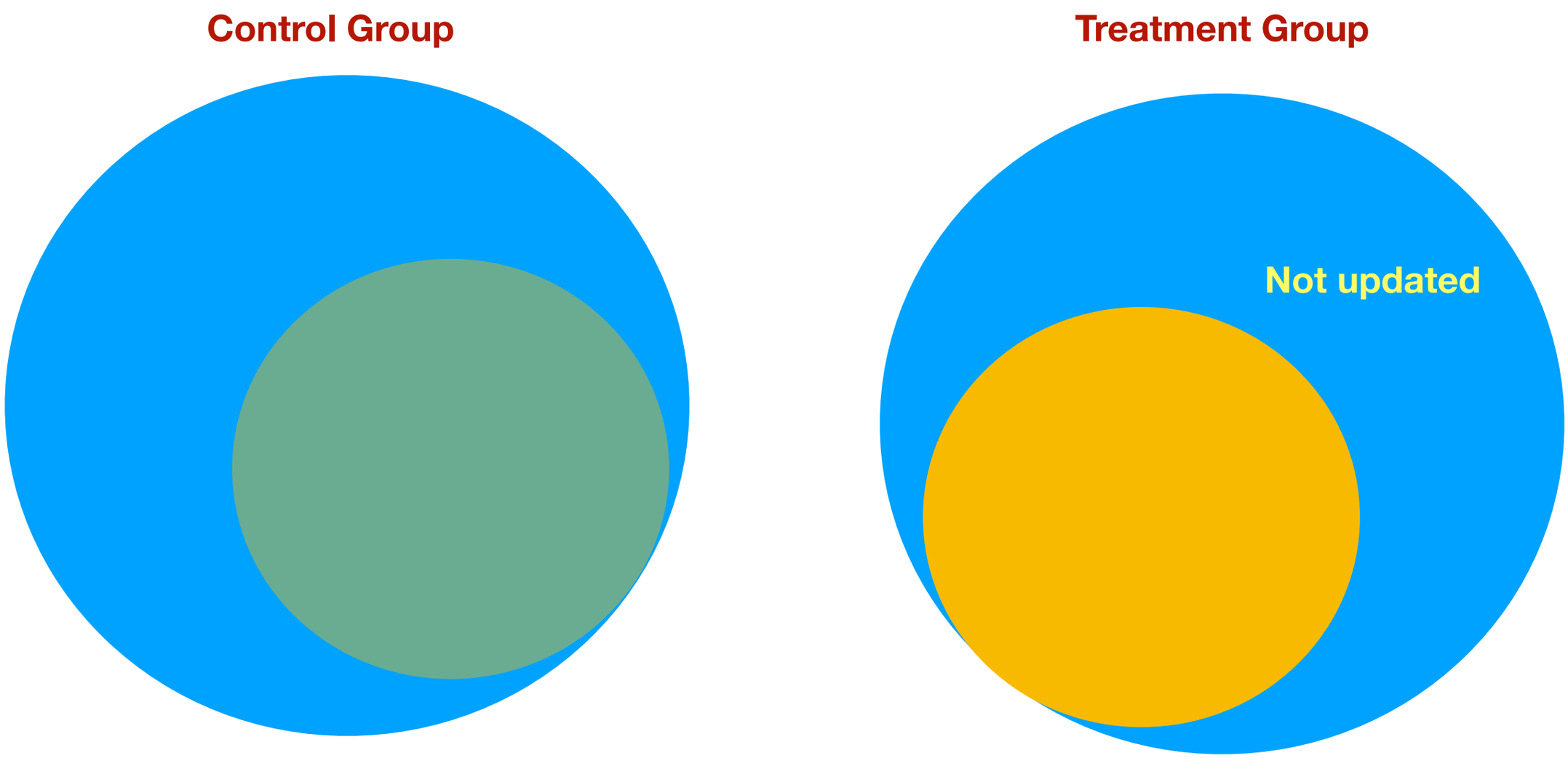
- Non-random attrition from the variants.
    - The users who update WeChat tend to be more active than those who do not.
- ***Solutions:***
    - Check whether the attrition is random. If it is not:

A. If the condition of attrition can be identified, you can analyze only the survivors i*n both treatment and control groups*.
    - You can only analyze the users who update WeChat in *control and treatment*. However, you should be aware that the treatment effects can only apply to those who update WeChat (e.g., more active users).

B. If the condition of attrition can not be identified.
    - We can measure the treatment effects based on *the offer of the new feature,* intention-to-treat, which is the Treatment of Offering the New Feature.

Survivorship Bias

- We need a larger Sample. WHY?



Heterogenous Treatment Effects

- Same treatment may affect different subgroups (individuals, time-periods, etc.) differently
- Conditional Average Treatment Effect (CATE)

$$\delta(x) = E(Y_i(1) - Y_i(0)) \,|\, X_i = x)$$

  - e.g., the new features works better for young people/ in India/ during holidays.

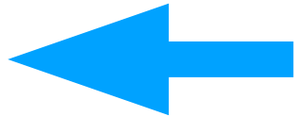
- Stratify the data and estimate ATE within each strata.
- The stratification factors should be independent of Treatment. WHY?
- $\Delta(x) = m_1(x) - m_0(x)$

Heterogenous Treatment Effects

- Treatment: A feature to improve user engagement
- Stratification Factor: Active Users (visit > 3 times in the last week) vs. Inactive Users based on the data *after* the experiment
- OEC: # visits
- This will cause a SRM: *WHY?*
    - Sample size of the treatment group increases(decreases) for the active (inactive) group because of the treatment effects.

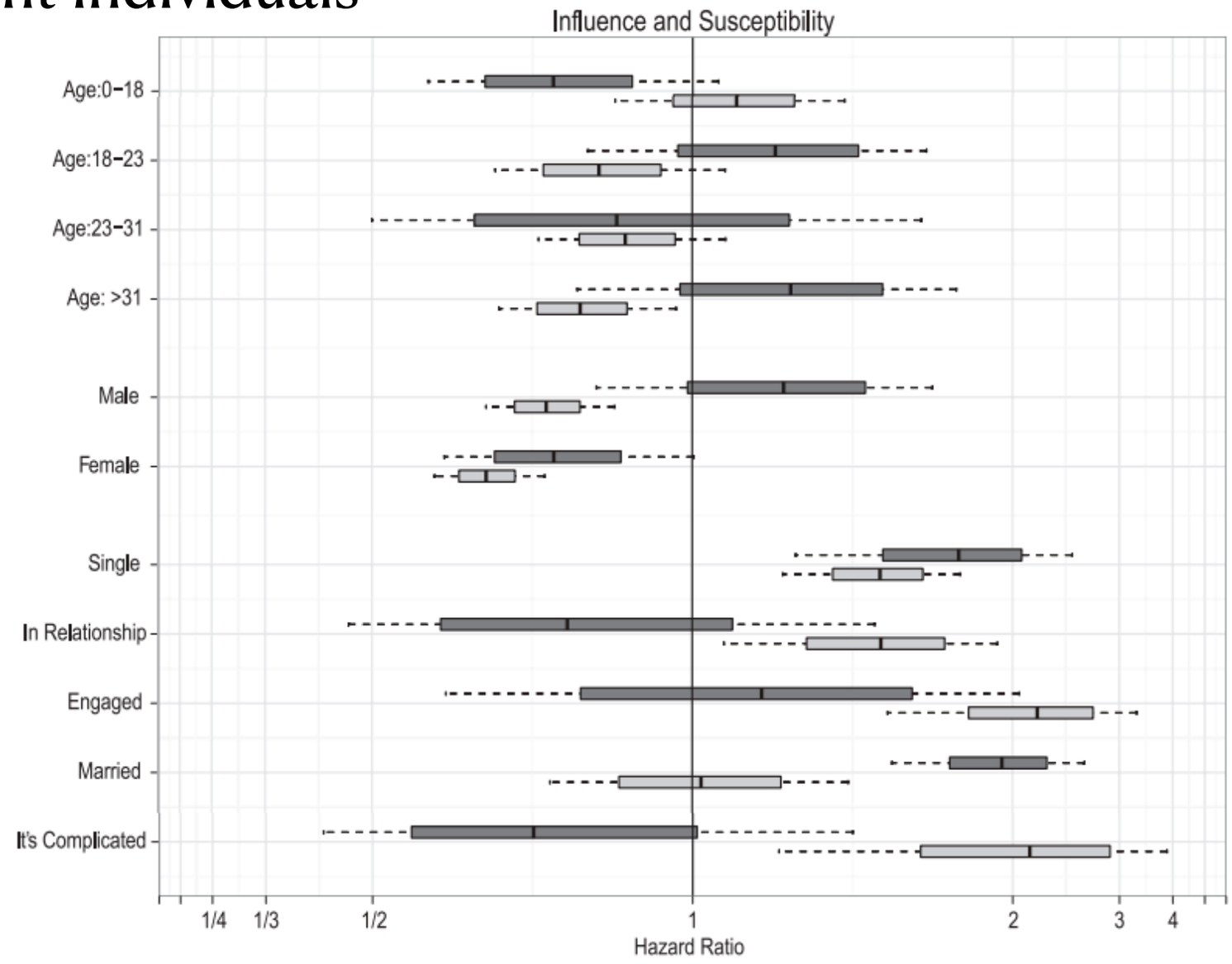
X<=2 are inactive users

Treatment	#Vists (Active)	#Vists (Inactive)
0	3	1
0	4	1
0	5	2
0	3	1
0	4	1
1	6	1
1	4	2
1	7	2
1	5	2–>4
1	4	1–>3



Heterogenous Treatment Effects

- Treatment Effects: effects of social cues on clicking ads across different individuals



Aral, Sinan, and Dylan Walker. "Identifying influential and susceptible members of social networks." *Science* 337.6092 (2012): 337-341.

Heterogenous Treatment Effects

- Average Treatment Effects are not informative enough with big/rich data.
    - CATE can be large even if ATE is not significant.
- Big data allows us to find heterogenous treatment effects.
    - Large N
    - Across many different segments
- What is the business value of detecting heterogenous effects?
    - Find the user segments/time periods most/more impacted by the Treatment.
    - Implement different strategies for different users.
    - Design a new product to serve a specific user group.

Compare Two Treatment Effects

- Rule of Thumb:
  - If the two CIs don't overlap, we can safely conclude that they are statistically significantly different.
  - Otherwise, we need to run additional tests to compare them
- What user groups with statistically significantly different treatment effects?
  - Age 0-18 < Age 18-23?
  - Male > Female?
  - Married > Complicated?

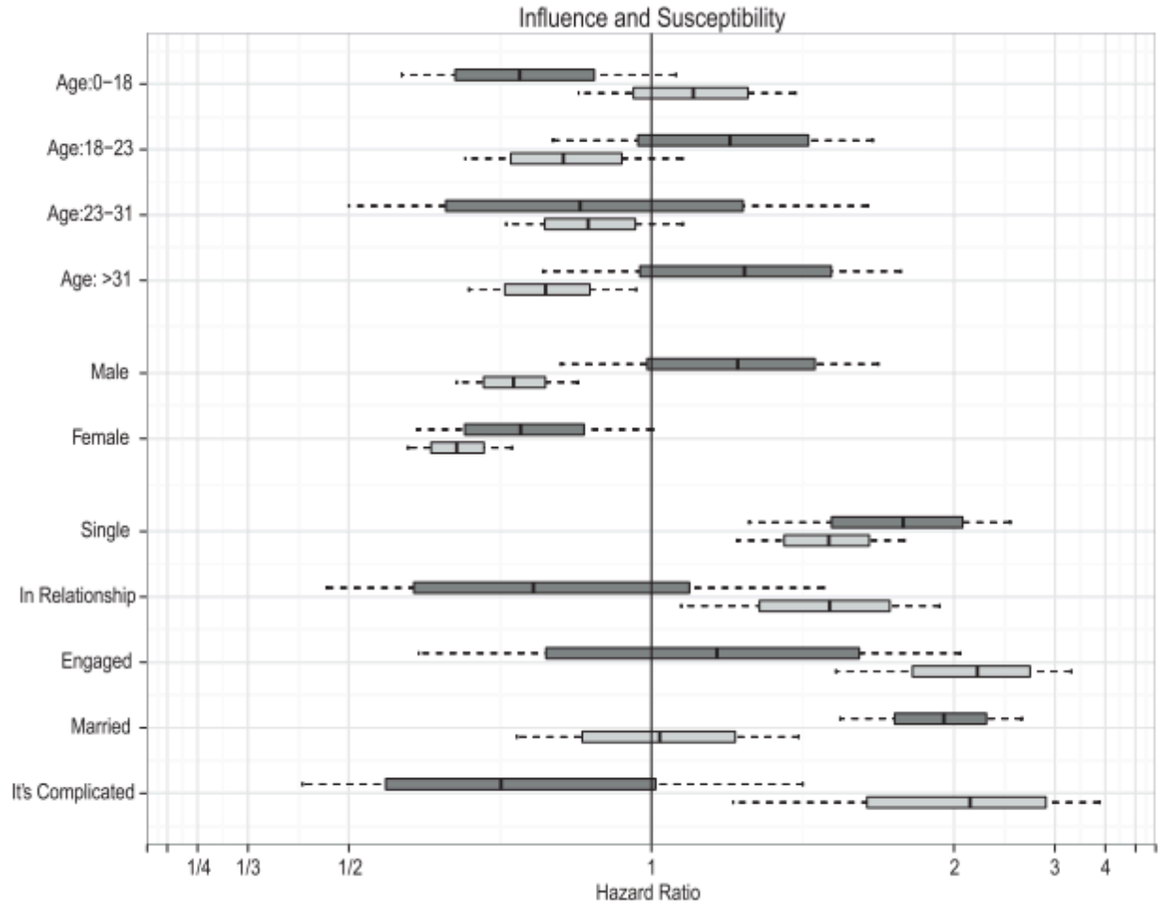


Fig. 1. Effects of age, gender, and relationship status on influence and susceptibility. Influence (dark gray) and susceptibility to influence (light gray) are shown with SEs (boxes) and 95% confidence intervals (whiskers). The figure displays hazard ratios (HRs) representing the percent increase (HR > 1) or decrease (HR < 1) in adoption hazards associated with each attribute. Age is binned by quartiles. Each attribute is shown as a pair of estimates, one reflecting influence (dark gray) and the other susceptibility (light gray). Personal relationship status reflects the status of an individual's current romantic relationship and is specified on Facebook as Single, In a Relationship, Engaged, Married, or It's Complicated. Estimates are shown relative to the baseline case for each attribute, which is the average for individuals who do not display that attribute in their online profile.

Compare Two Treatment Effects with t-test

- Ho: $\delta = \delta_1 - \delta_2 = 0$
- $\text{se}(\Delta) = \sqrt{se(\Delta_1)^2 + se(\Delta_2)^2}$
  - Two unit segments are independent
- $t = \dfrac{\Delta}{se(\Delta)}$

Compare Two Treatment Effects with Regressions

- Analyze Treatment Effects with OLS (Ordinary Least Squares):

$$y_i = \beta_0 + \beta_1 \,.\, T_i + \epsilon_i, \quad \epsilon_i \text{ is IID}$$

$$T_i = \{0,1\}, \text{Control or Treatment}$$

y_i is the OEC (or other metrics)

$$\delta = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

$$\Delta = \hat{\beta}_0 + \hat{\beta}_1 - \hat{\beta}_0 = \hat{\beta}_1$$

- The simplest version of Regression
- The results should be consistent with comparing the means with t-tests (z-tests, bootstrap)
- OLS cannot be used to analyze other statistics, e.g., median, quantiles, lifts.

Interaction Effects with Regressions

- $y_i = \beta_0 + \beta_1 \,.\, T_i + \beta_2 \,.\, G_i + \gamma \,.\, T_i \,.\, G_i + \epsilon_i\,,\ \epsilon_i$ is IID

$$\gamma = E(\delta \,|\, G_i = 1) - E(\delta \,|\, G_i = 0) = \delta_1 - \delta_2$$

$$\gamma = (E(y_i \,|\, T_i = 1, G_i = 1) - E(y_i \,|\, T_i = 0, G_i = 1)) - (E(y_i \,|\, T_i = 1, G_i = 0) - E(y_i \,|\, T_i = 0, G_i = 0))$$

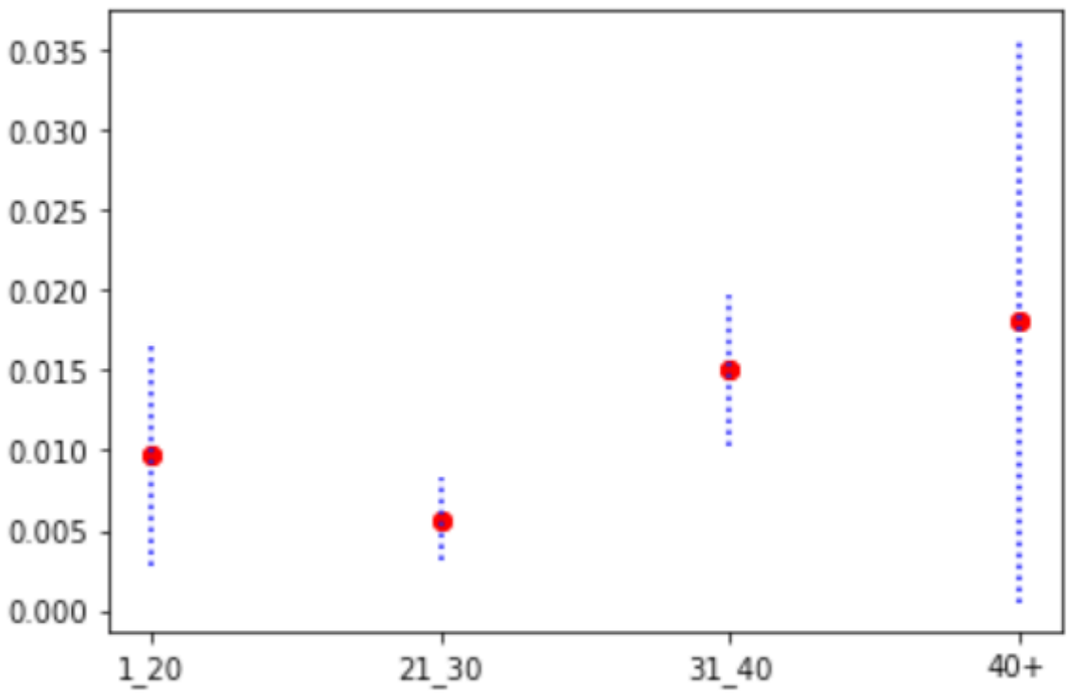
$$= \beta_0 + \beta_1 + \beta_2 + \gamma - \beta_0 - \beta_2 - \beta_0 - \beta_1 + \beta_0$$

$$\Delta = \Delta_1 - \Delta_2 = \hat{\gamma}$$

- If the regression results show that $\hat{\gamma}$ is statistically significant, we can conclude δ_1, δ_2 are statistically significantly different.
- We call $\hat{\gamma}$, the estimator of interaction effects between T_i and G_i.
- Interaction Effects:
    - How the Treatment Effects are impacted by Gender (G).

Class Exercise

- Find the Treatment Effects for different genders and age groups with t-test and OLS.
- Compare CATE between different user segments with OLS.
- Visualize the heterogenous treatment effects
- Use the data sample_data_segments.csv



Class Exercise

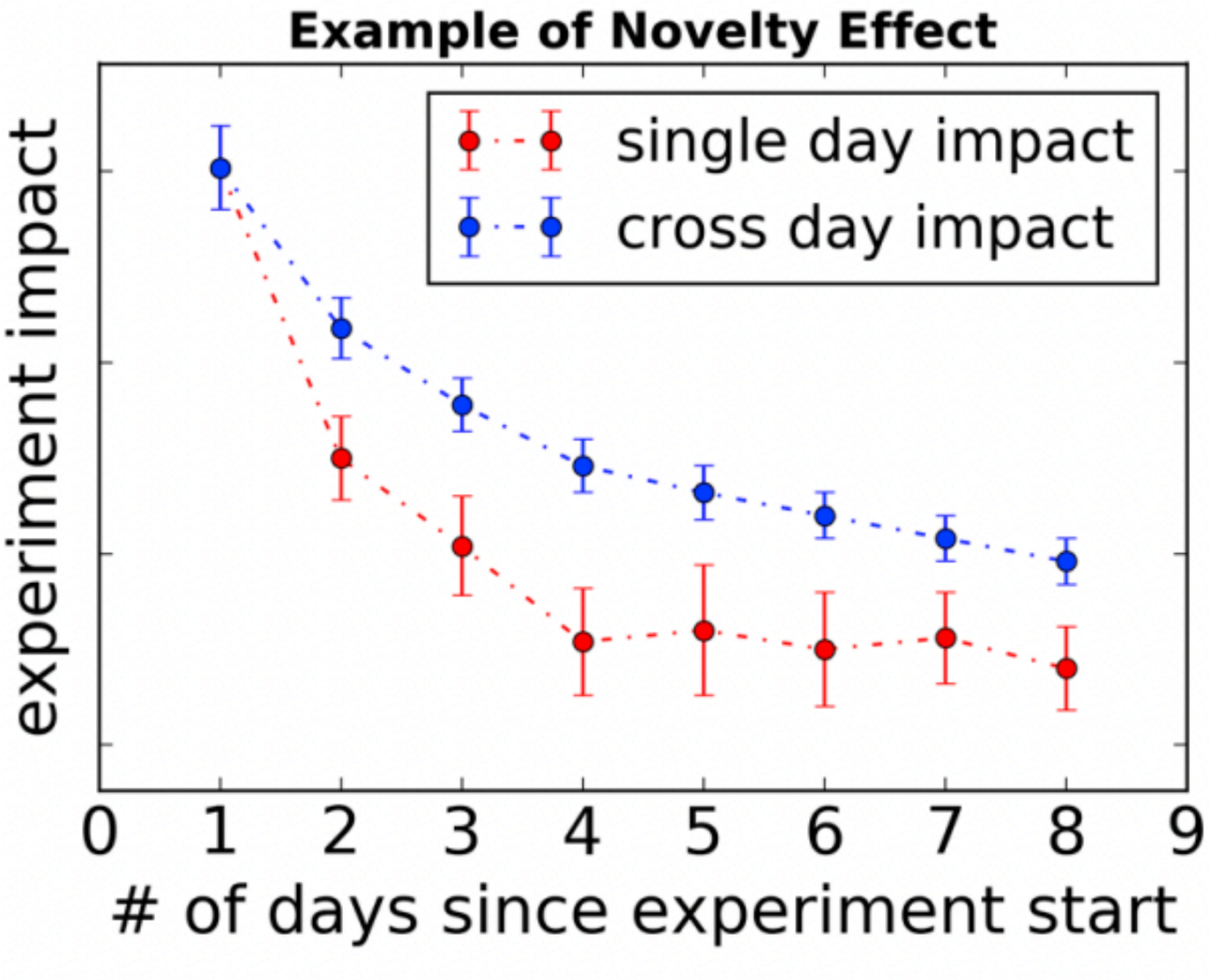
- Python Code for OLS

```
import statsmodels.formula.api as smf

mod = smf.ols(formula='click ~ expid', data=df)
res = mod.fit()
print(res.summary())

mod_gender = smf.ols(formula='click ~ gender + expid + gender*expid',
data=df)
res_gender = mod_gender.fit()
print(res_gender.summary())

** mod_age = smf.ols(formula='click ~ age + expid + age*expid',
data=df)
```

Novelty Effects

- The standard assumption of experiments assumes the Treatment Effects are constant over time.
- *However, it is not always the case!*
- New features often initially attract users' attention, which however may decline over time.
    - If users don't find the feature useful, repeat usage will be small.
- The treatment effects may appear to perform well at first but will quickly decline over time.

Detect Novelty Effects



- The decline of single-day effects clearly indicate novelty effects.
- Time-dependent effects due to users' interests change.

Which effects will you reference to make decisions?

Detect Novelty Effects

- It indicates the experiment needs to run longer to get a stable impact estimate.
- It offers unique insights on how users interact with the new feature being experimented on.
- Take the users who arrived in the first few days and plot the treatment effects for them over time. WHY?
  - The early arrivers may be different from the later ones.
  - This further controls individual characteristics.
    - The effects' change may be caused by user characteristics
  - This also introduces selection bias. WHY?

Primacy Effects

- When a change is introduced, users may need time to adopt.
- Users are primed in the old feature and get used to the way it works.
- The treatment effects may be increasing overtime.
- It often happens, when we introduce a new feature.
- The experiment needs to be longer.
- Keep optimistic for your product during the early stage!