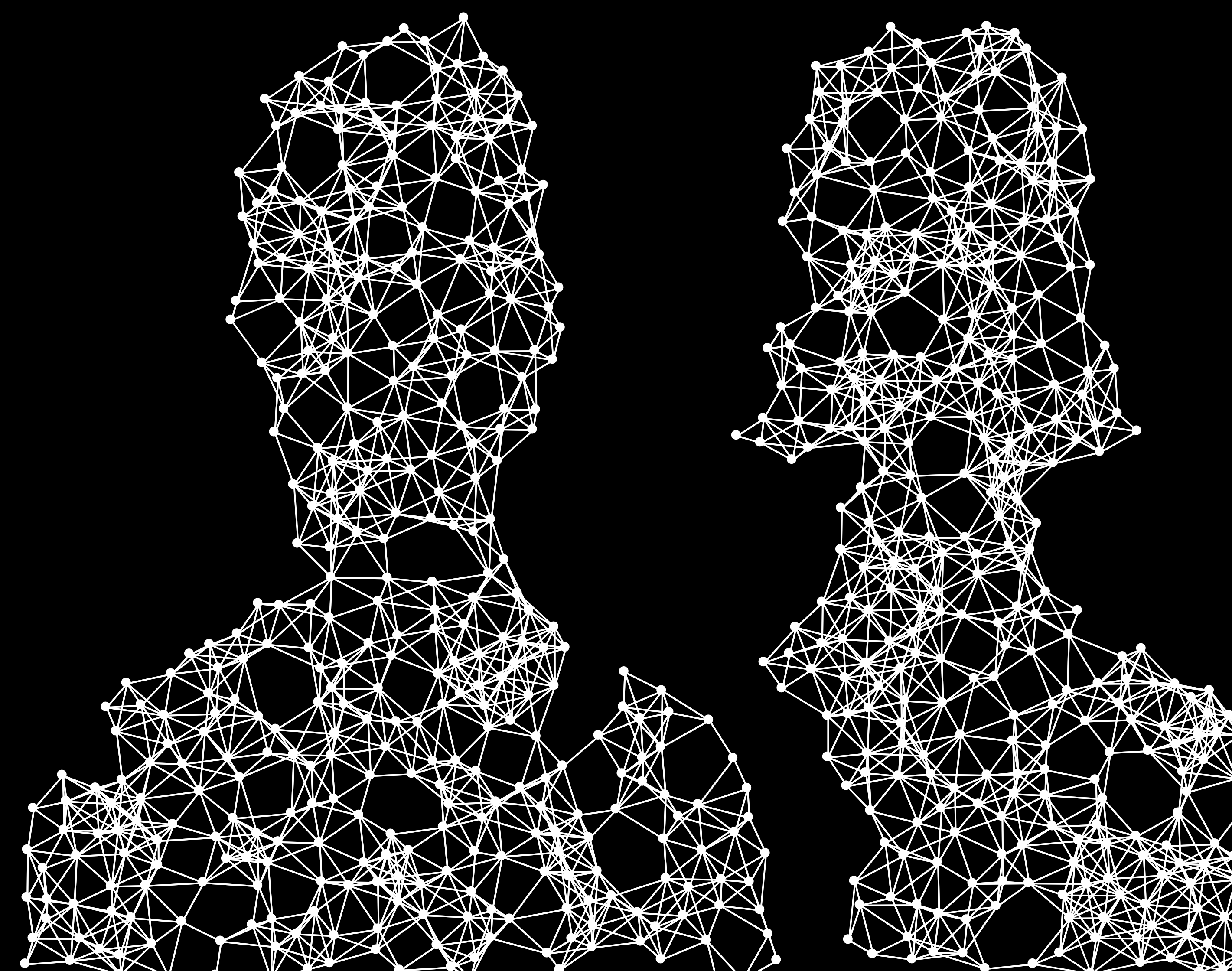
Online Experiments for Computational Social Science

ICWSM Tutorial

Eytan Bakshy
Sean J. Taylor
Facebook Core Data Science

ICWSM 2014, Ann Arbor, MI
June 1, 2014



Outline

1. Introduction and Causal Inference (30 minutes)

2. Planning Experiments (30 minutes)

   ‹30 minute break›
   Discussion + analysis exercise (15 minutes)

3. Designing and Implementing Experiments (45 minutes)

4. Analyzing Experiments (30 minutes)

Everything we assume

- Minimum requirements
  - Some basic knowledge of statistics
  - The ability to follow code
- Necessary to understand 90%
  - Intermediate knowledge of R and beginner knowledge of Python
- Necessary to understand 100%
  - Advanced knowledge of R, intermediate Python, intermediate stats, design of experiments

Don’t panic!

Don’t panic!

Buddy up!
Group learning is good for you!

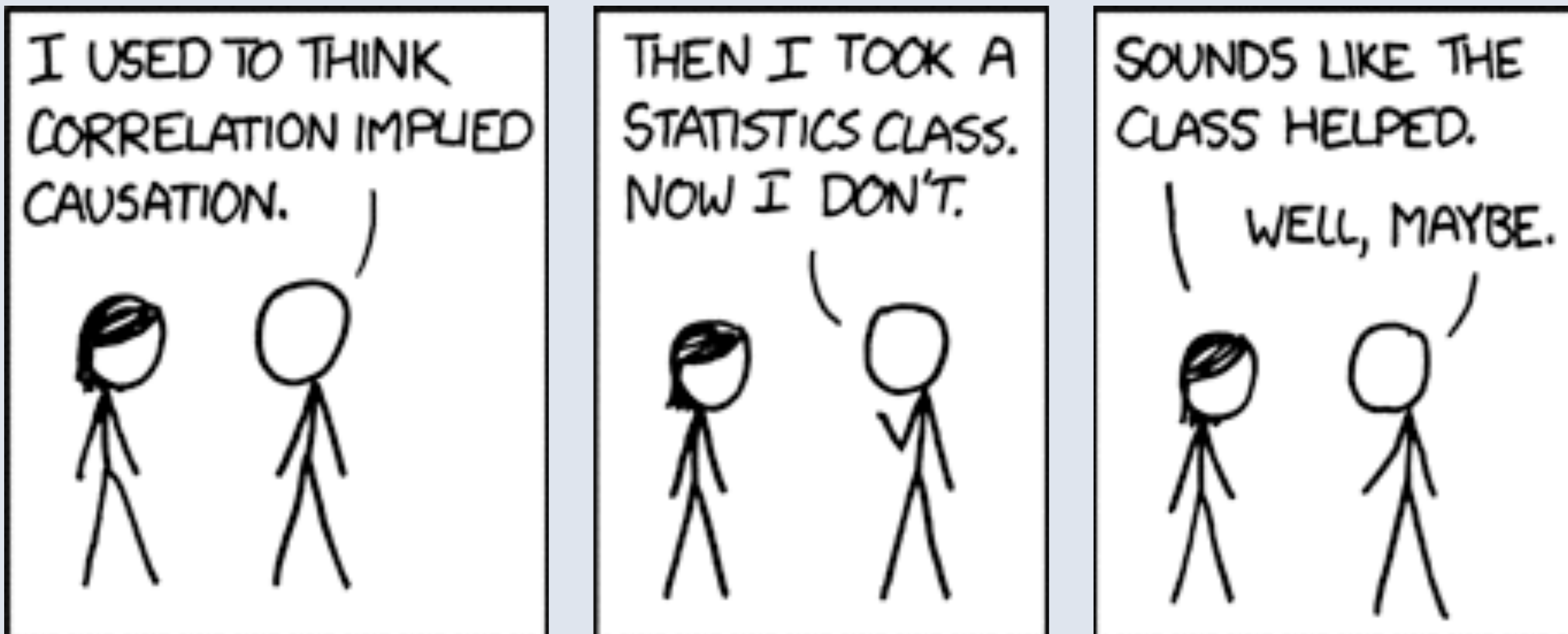
Software requirements

http://bit.ly/icwsm14_experiments

- If you are the kind of person who wants to tinker with code yourself, here are the software requirements
  - Section 1: None
  - Section 2: R
  - Part 3: IPython + PlanOut
  - Part 4: R
- Can't install the software? Buddy up

Section 1:
Introduction and Causal Inference

Obligatory


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Associations

$$(X_i, Y_i)$$

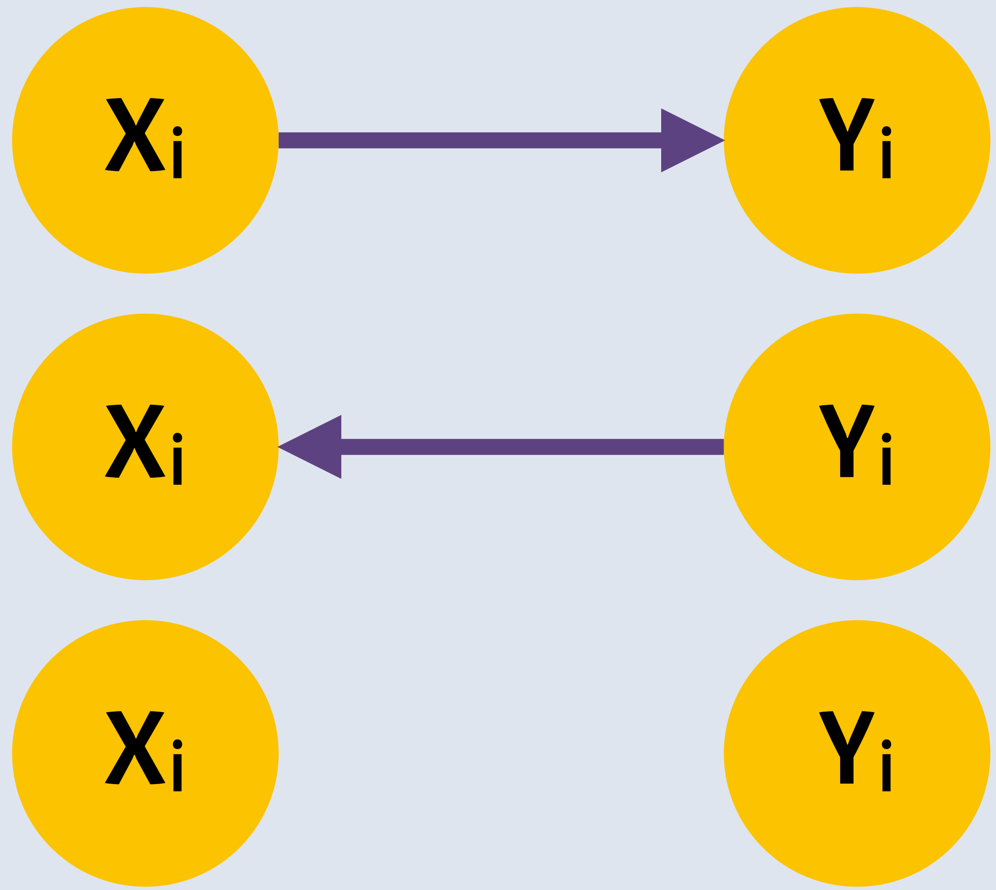
- Health: (smoking, cancer)
- Sports: (running the ball, winning games)
- Education: (completes MOOC course, gets a promotion)
- Social Media: (likes a page on Facebook, buys a product)

Possible Relationships

- $Pr(Y_i \mid X_i) = Pr(Y_i)$ (independence)
- $Pr(Y_i \mid X_i) \neq Pr(Y_i)$ (dependence)

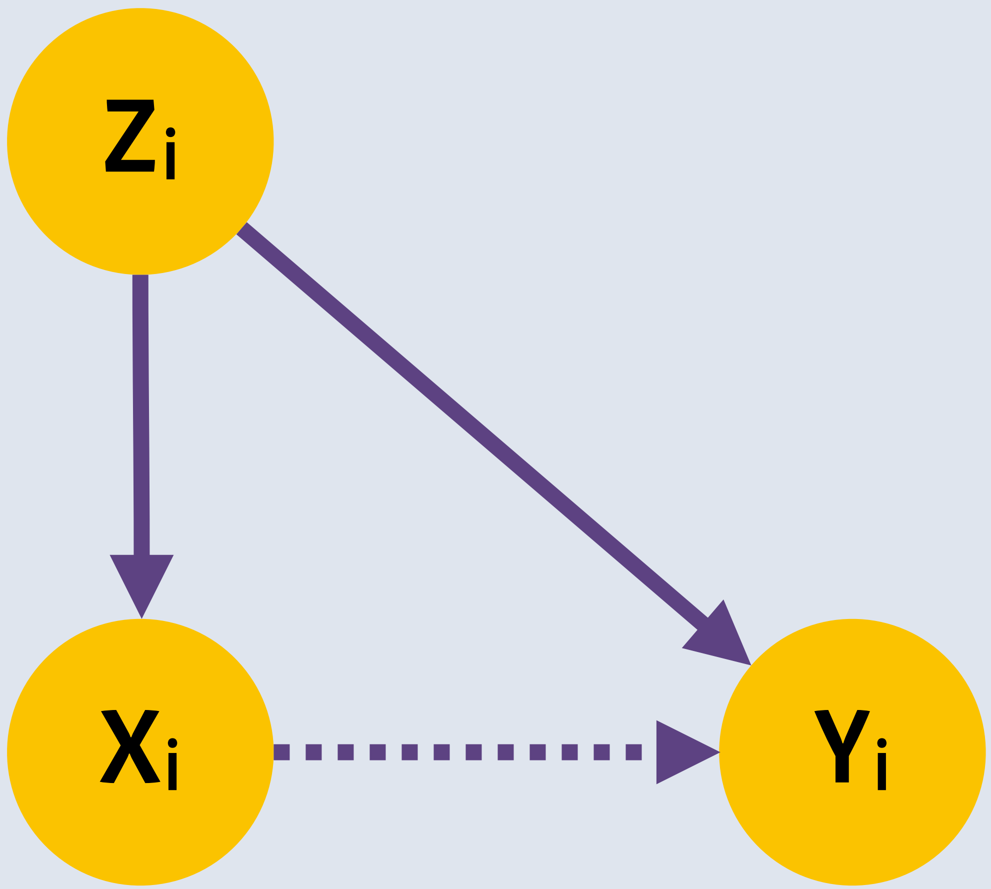
Dependence between variables is useful, but interpretation of this relationship can often be tricky.

Possible Causal Relationships



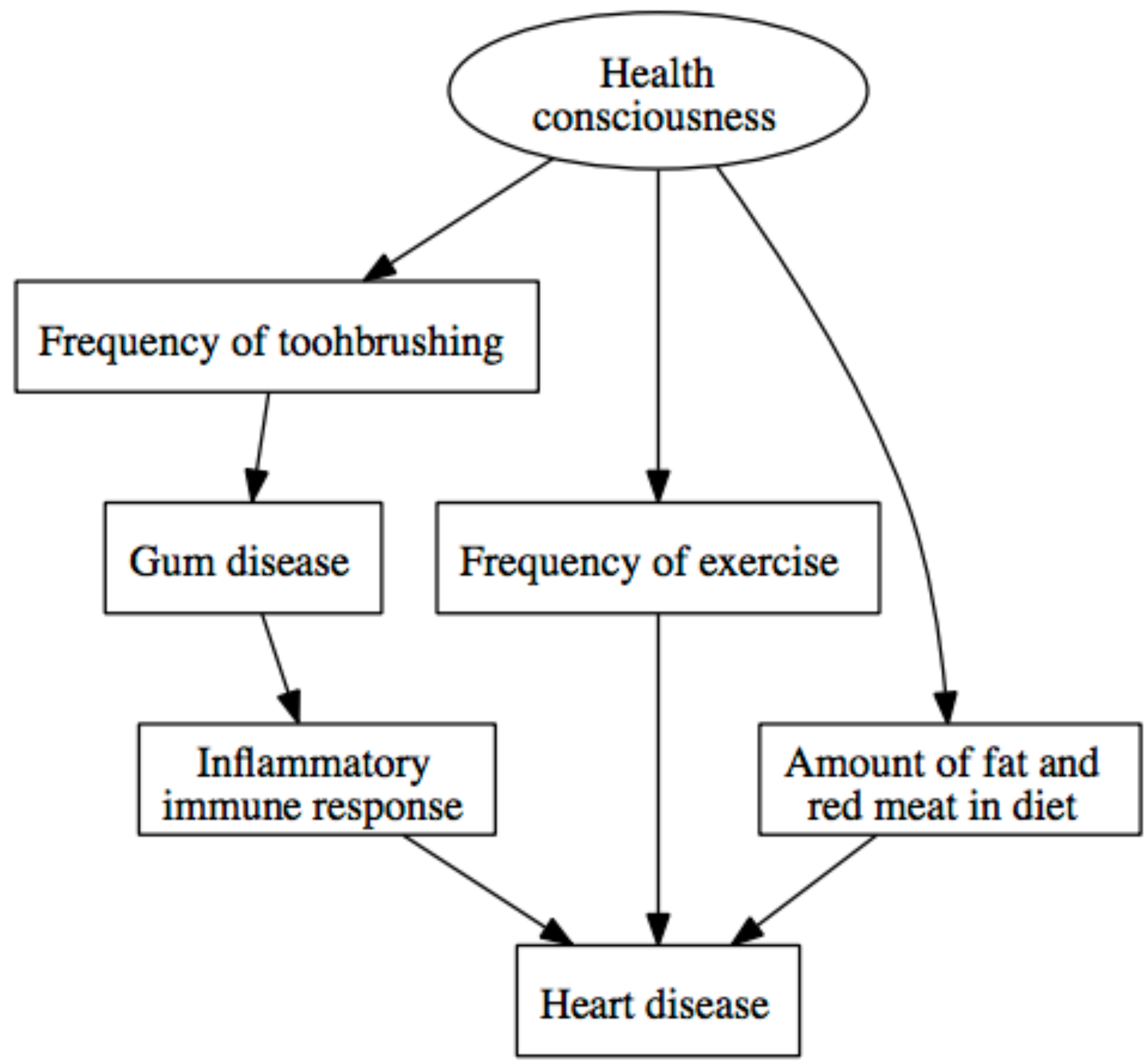
All three causal relationships are possible when there's a dependency between variables.

Correlation without Causation: Introducing Z_i



- Smoking causes cancer (genetics)
- Running the ball causes winning games (having a lead)
- Completing a MOOC causes a promotion (self-motivation)
- Liking a page on Facebook causes a person to buy a product (brand loyalty)

Bigger Example



Why Causal Inference?

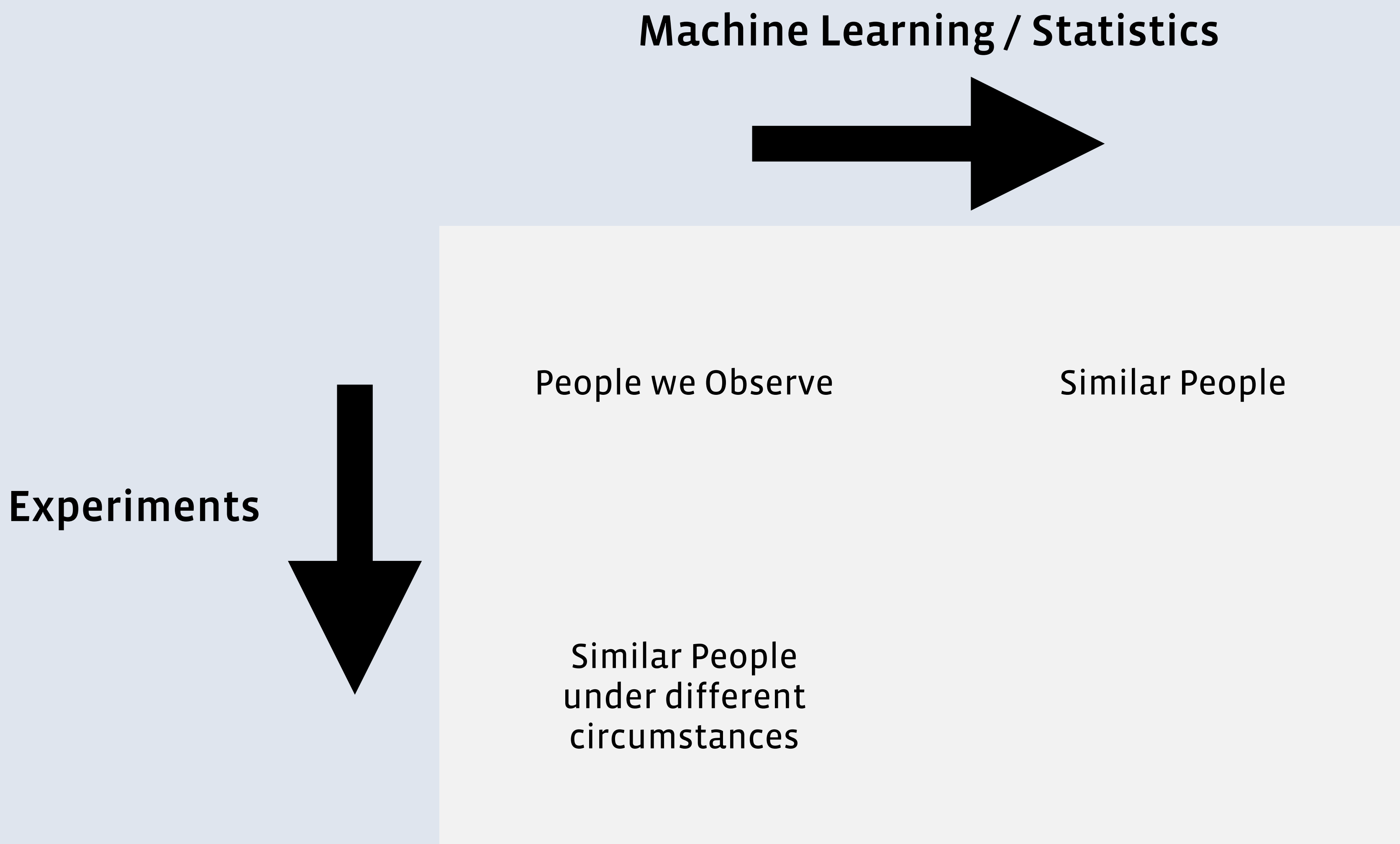
1. **Science:**
   Why did something happen?
2. **Decisions:**
   What will happen if I change something?

Causal Inference for Science

1. Associations between two variables are always more interesting when they're causal.
2. Understanding a phenomenon is different from predicting it.

Explanatory Modeling	**Predictive Modeling**
model captures causal function	model captures association
model carefully constructed from theory	models constructed from data
retrospective	forward-looking
minimize bias	minimize variance
basic science	applied science
make better data	make better features

Two Kinds of Out-of-Sample



Attribute Outcomes to Causes

non-social ad

social ad

Social Influence in Social Advertising: Evidence from Field Experiments.
Bakshy, Eckles, Yan, Rosenn. EC 2012.

Attribute Outcomes to Causes

1 liking friend
0 friends shown

1 liking friend
1 friend shown

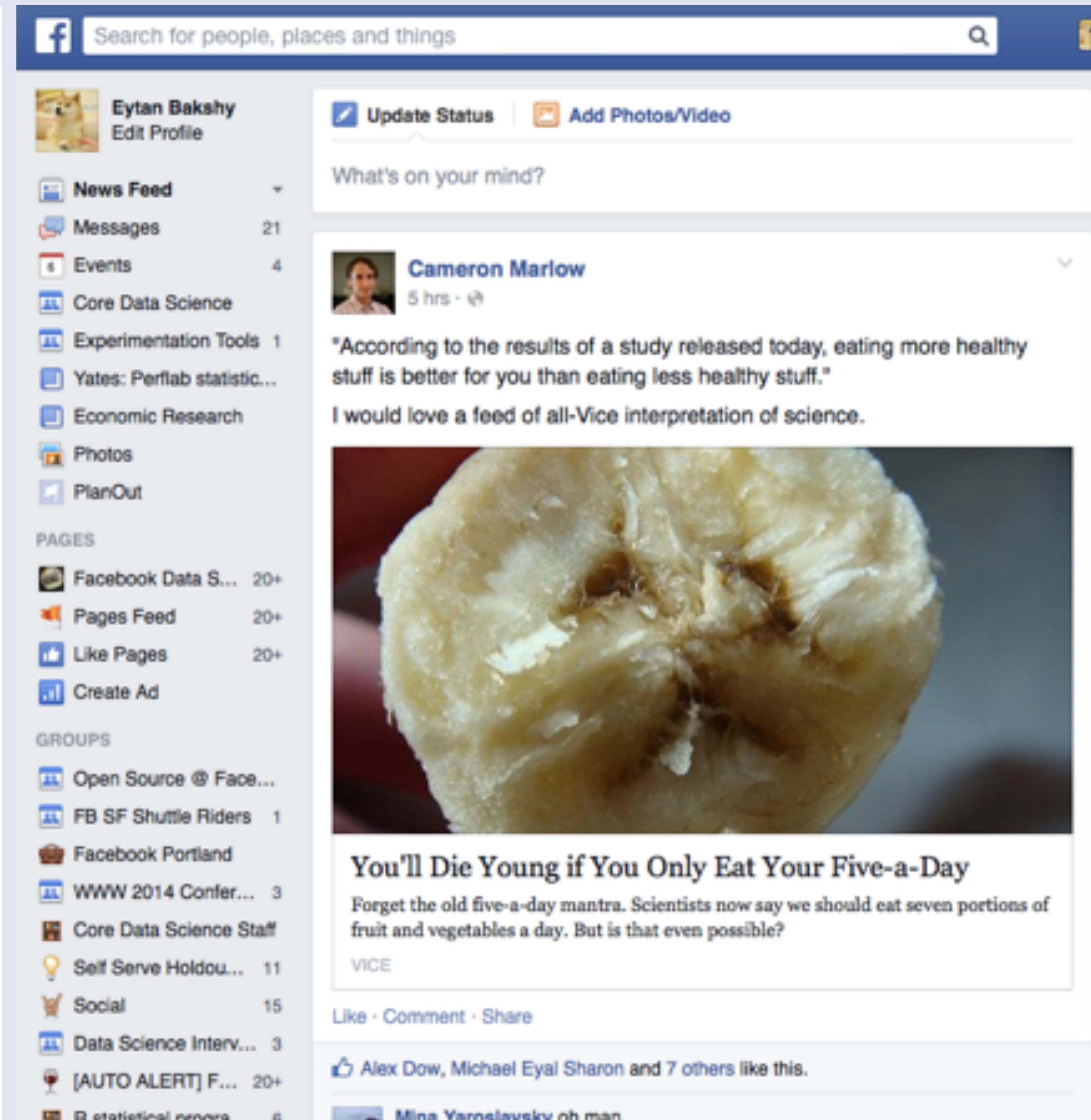
Social Influence in Social Advertising: Evidence from Field Experiments.
Bakshy, Eckles, Yan, Rosenn. EC 2012.

Causal Inference for Decisions

- Health: Quit smoking? Brush your teeth? :)
- Sports: run the ball more?
- Social media: recruit more Facebook fans for my page?
- Advertising: purchase ads?
- Education: invest time in completing a MOOC?

Test complete alternatives

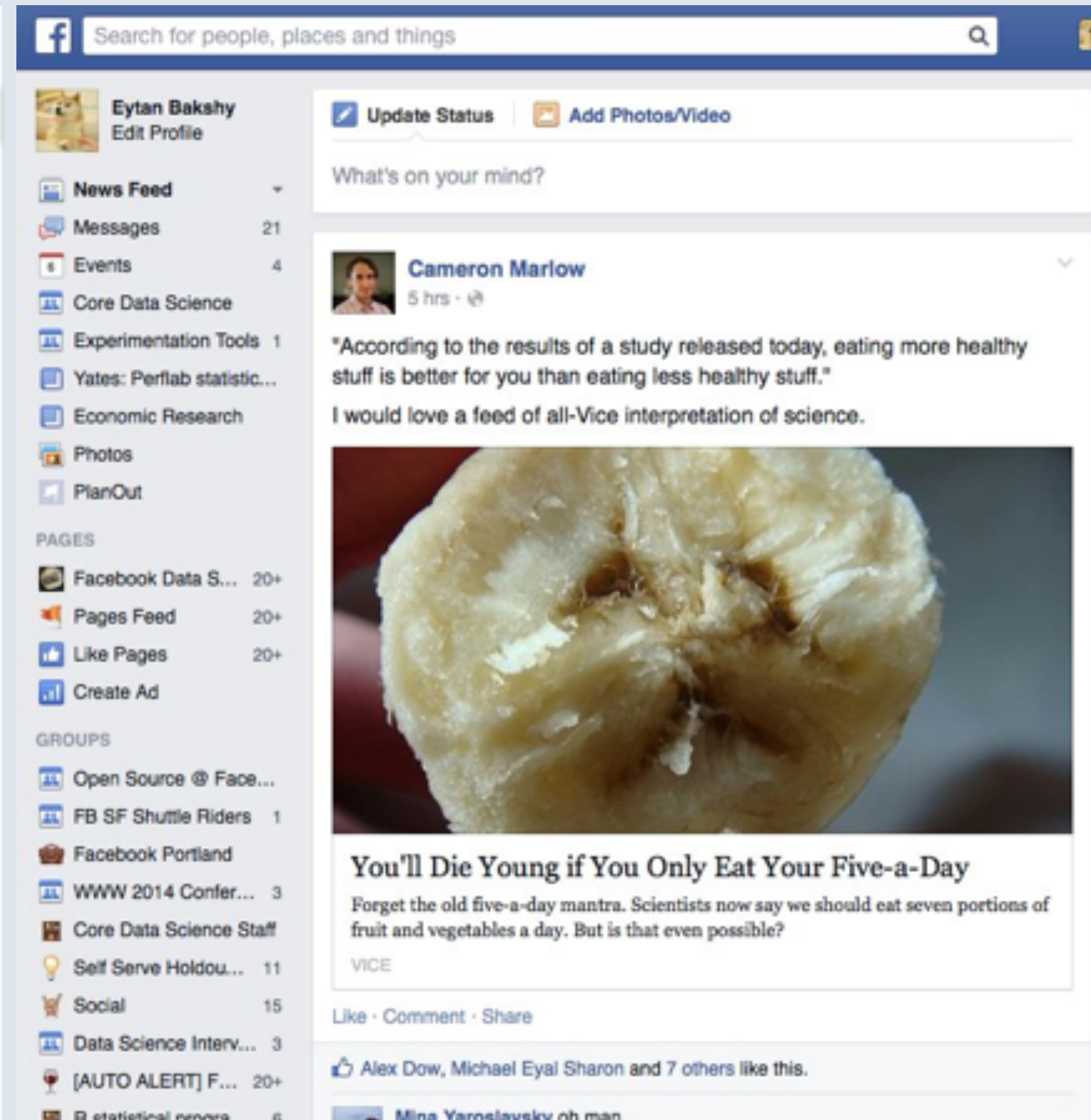
News Feed (2011)



News Feed (2014)

Explore a design space

News Feed (2011)



News Feed (2014)

A Social Science Problem

Causal inference is easy for the natural sciences:

- Manipulation is easy!
- Molecules, cells, animals, plants are exchangeable!

For people, there's always something we may not be observing perfectly.

- Latent traits provide alternative explanations that we often cannot rule out.
- Experiments are often the only way of ruling them out.

Potential Outcomes Framework

	Y	Y	D	Effect
Eytan	10	-5		15
Anna	0	5		-5
Gary	-20	5		-25
Linda	-10	10		-20
Edna	-5	0		-5
Sean	10	5		5
Mean	**-2.14**	**2.86**		**-5**

$Y_i(1)$ is the outcome under treatment
$Y_i(0)$ is the outcome under control

Fundamental Problem of Causal Inference

	Y	Y	D	Effect
Eytan	10		1	?
Anna		5	0	?
Gary		5	0	?
Linda		10	0	?
Edna		0	0	?
Sean	10		1	?
Mean				

We only ever observe a unit in either treatment or control.
Individual level effects are never defined.

Confounding

	Y	Y	D	Effect
Eytan	10		1	?
Anna		5	0	?
Gary		5	0	?
Linda		10	0	?
Edna		0	0	?
Sean	10		1	?
Mean	**10**	**5**	**0.29**	**5**

Here we assumed they selected D_i to maximize their outcome.

Randomization

	Y	Y	D	Effect
Eytan		-5	0	?
Anna	0		1	?
Gary	-20		1	?
Linda		10	0	?
Edna		0	0	?
Sean	10		1	?
Mean	**-3.33**	**1.67**	**0.5**	$\boxed{\textbf{-5}}$

With randomly assigned D_i, we get unbiased effect estimates.
The difference in means is called the **Average Treatment Effect**

The Average Treatment Effect

$$\text{ATE} = \delta = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$

- Due to the linearity of expectations, we can separate the ATE into two measurements.
- In practice, we *estimate* these expectations using the means in our treatment and control groups.

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i \in T} Y_i(1) - \frac{1}{M}\sum_{i \in C} Y_i(0)$$

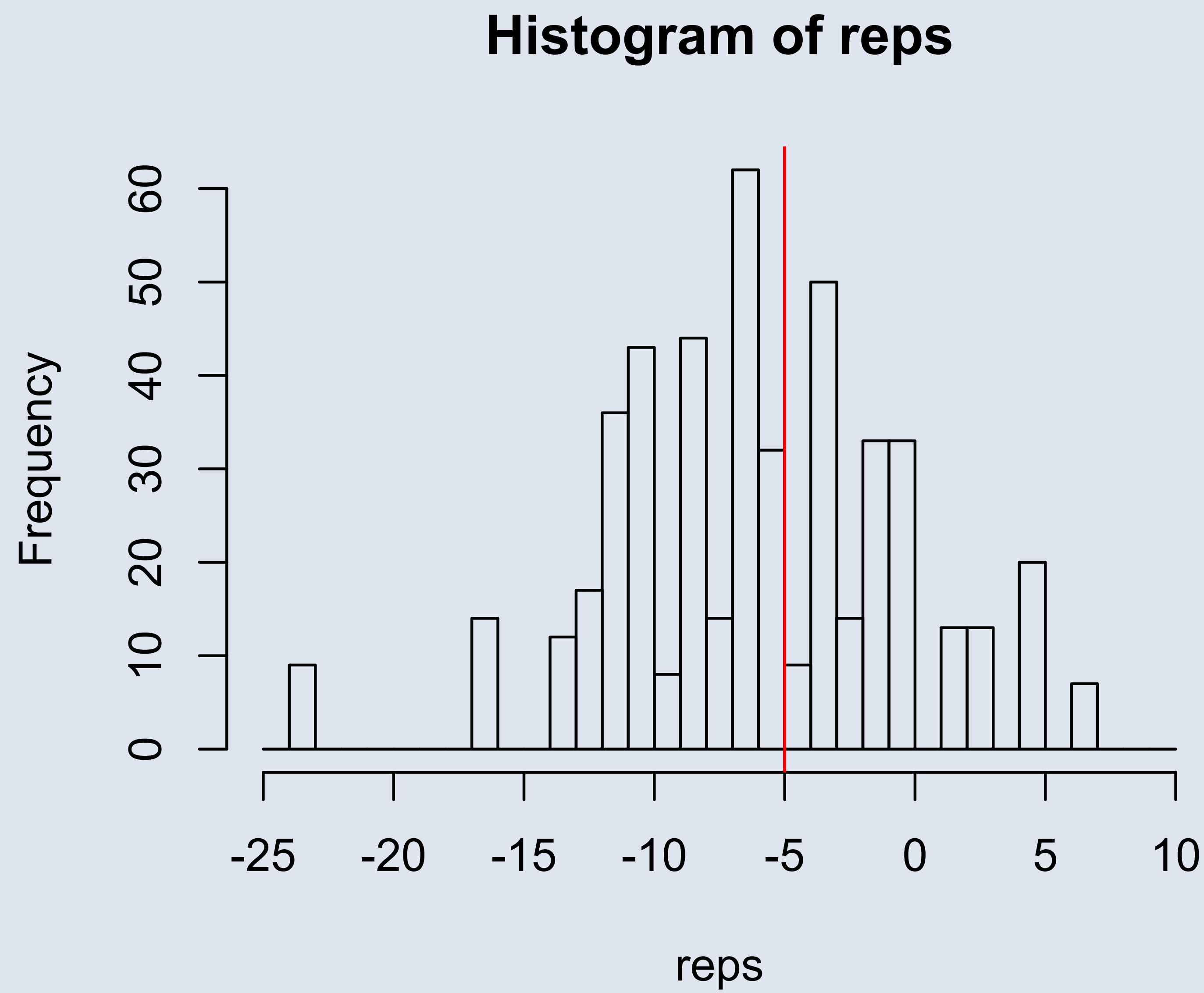
Uncertainty

- How sure are we about the ATE we measured?

Variability in ATE estimates comes from:

1. variation in random assignment
2. variation in subjects

Variation due to Random Assignment



For a given set of potential outcomes and a randomization, we may not always arrive at the right estimate, but it will not be biased.

Standard Errors

- How can we quantify uncertainty about ATEs we measure?

$$\widehat{\mathrm{SE}}(\widehat{\mathrm{ATE}}) = \sqrt{\frac{1}{n_0+n_1}}\sqrt{\frac{n_1}{n_0}\mathrm{Var}(Y_i(0)) + \frac{n_0}{n_1}\mathrm{Var}(Y_i(1))}$$

- SE decreases with √N
- SE is smaller when variances of potential outcomes are smaller
- want n_0 and n_i to be similar if the variances are the same
- want more observations for higher variance conditions

Confidence Intervals

- 95% confidence interval is 1.96 times $\widehat{\mathrm{SE}}(\widehat{\mathrm{ATE}})$

- SEs shrink with the square root of the number of observations, so to double the precision of your experiment you need four times the number of subjects.

Confidence Intervals vs p-values

- Often easy to get statistical significance with big data. Most things have effects!
- Harder to get big, <u>practically significant</u>, effects!
- CIs make uncertainty about an estimate more credible.
- Should favor reporting CIs.

Section 2:
Planning Experiments

The Routine

The routine

- **Step 1: Formulate a research hypothesis**
- Step 2: State an expected effect size
- Step 3: Design your experiment
- Step 4: Power analysis
- Step 5: Write up analysis plan
- Step 6: Collect data
- Step 7: Analyze data according to plan

- A specific hypothesis:
  - Population
  - Empirical context
  - Treatment(s)
  - Subgroups
  - Outcomes

The routine

- Step 1: Formulate a research hypothesis
- **Step 2: State an expected effect size**
- Step 3: Design your experiment
- Step 4: Power analysis
- Step 5: Write up analysis plan
- Step 6: Collect data
- Step 7: Analyze data according to plan

- Use prior literature
- Use existing data
- Collect your own observational data
- Small effects need big data

The routine

- Step 1: Formulate a research hypothesis
- Step 2: State an expected effect size
- **Step 3: Design your experiment**
- Step 4: Power analysis
- Step 5: Write up analysis plan
- Step 6: Collect data
- Step 7: Analyze data according to plan

- Follows from step 1
- Identify:
  - Constraints
  - Threats to validity
  - Ways to increase precision

The routine

- Step 1: Formulate a research hypothesis
- Step 2: State an expected effect size
- Step 3: Design your experiment
- **Step 4: Power analysis**
- Step 5: Write up analysis plan
- Step 6: Collect data
- Step 7: Analyze data according to plan

- Simulate experiment with posited effects and design
- You should feel comfortable that you'll find a clinically significant effect

The routine

- Step 1: Formulate a research hypothesis
- Step 2: State an expected effect size
- Step 3: Design your experiment
- Step 4: Power analysis
- **Step 5: Write up analysis plan**
- Step 6: Collect data
- Step 7: Analyze data according to plan

- Makes it easier to communicate your study
- Helps catch problems with plan
- Keeps you honest as a scientist

The routine

- Step 1: Formulate a research hypothesis
- Step 2: State an expected effect size
- Step 3: Design your experiment
- Step 4: Power analysis
- Step 5: Write up analysis plan
- **Step 6: Collect data**
- Step 7: Analyze data according to plan

- Collect pre-treatment data (if applicable)
- Implement
- Collect experiment data
  - Log sane data

The routine

- Step 1: Formulate a research hypothesis
- Step 2: State an expected effect size
- Step 3: Design your experiment
- Step 4: Power analysis
- Step 5: Write up analysis plan
- Step 6: Collect data
- **Step 7: Analyze data according to plan**

- Wrangle data to get it into a format that you can analyze in R
- Apply appropriate statistical procedures
- Analysis should be easy!

Planning Experiments using Simulation

open: power_part1.R

In-class Exercise

open: power_part2.R

Power analysis for peer effects study

facebook

Section 3: Designing and Implementing Experiments with PlanOut

Be a subject in a linguistic alignment study

http://icwsm.seanjtaylor.com
(3 minutes)

PlanOut

PlanOut scripts are high-level descriptions of randomized parameterizations

The PlanOut Idea

- **User experiences are parameterized by experimental assignments**
- **PlanOut scripts describe assignment procedures**
- Experiments are PlanOut scripts plus a population
- Parallel or follow-on experiments are centrally managed

Sample PlanOut script

```
button_color = uniformChoice(
 choices=["#ff0000", "#00ff00"],
 unit=userid);

button_text = uniformChoice(
 choices=["I'm voting", "I'm a voter”],
 unit=userid);
```

2x2 factorial design

Compiled PlanOut Code

```
{
  "op": "seq",
  "seq": [
    {
      "op": "set",
      "var": "button_color",
      "value": {
        "choices": {
          "op": "array",
          "values": [
            "#ff0000",
            "#00ff00"
          ]
        },
        "unit": {
          "op": "get",
          "var": "userid"
        },
        "op": "uniformChoice"
      }
    },
    {
      "op": "set",
      "var": "button_text",
      "value": {
        "choices": {
          "op": "array",
          "values": [
            "I'm voting",
            "I'm a voter"
          ]
        },
        "unit": {
          "op": "get",
          "var": "userid"
        },
        "op": "uniformChoice"
      }
    }
  ]
```

Using PlanOut

$ ipython notebook --pylab inline

open: 0-planout-intro.ipynb

Selective Exposure Experiment

- Communication literature has shown people tend to choose news sources that align with their political ideologies.
- This experiment was designed to test this hypothesis.
- News source icons were randomly applied to the sample set of stories, and presented in a random order.

Python Web Application Walkthrough

open: webapp/app.py

Extracting Data from Logs

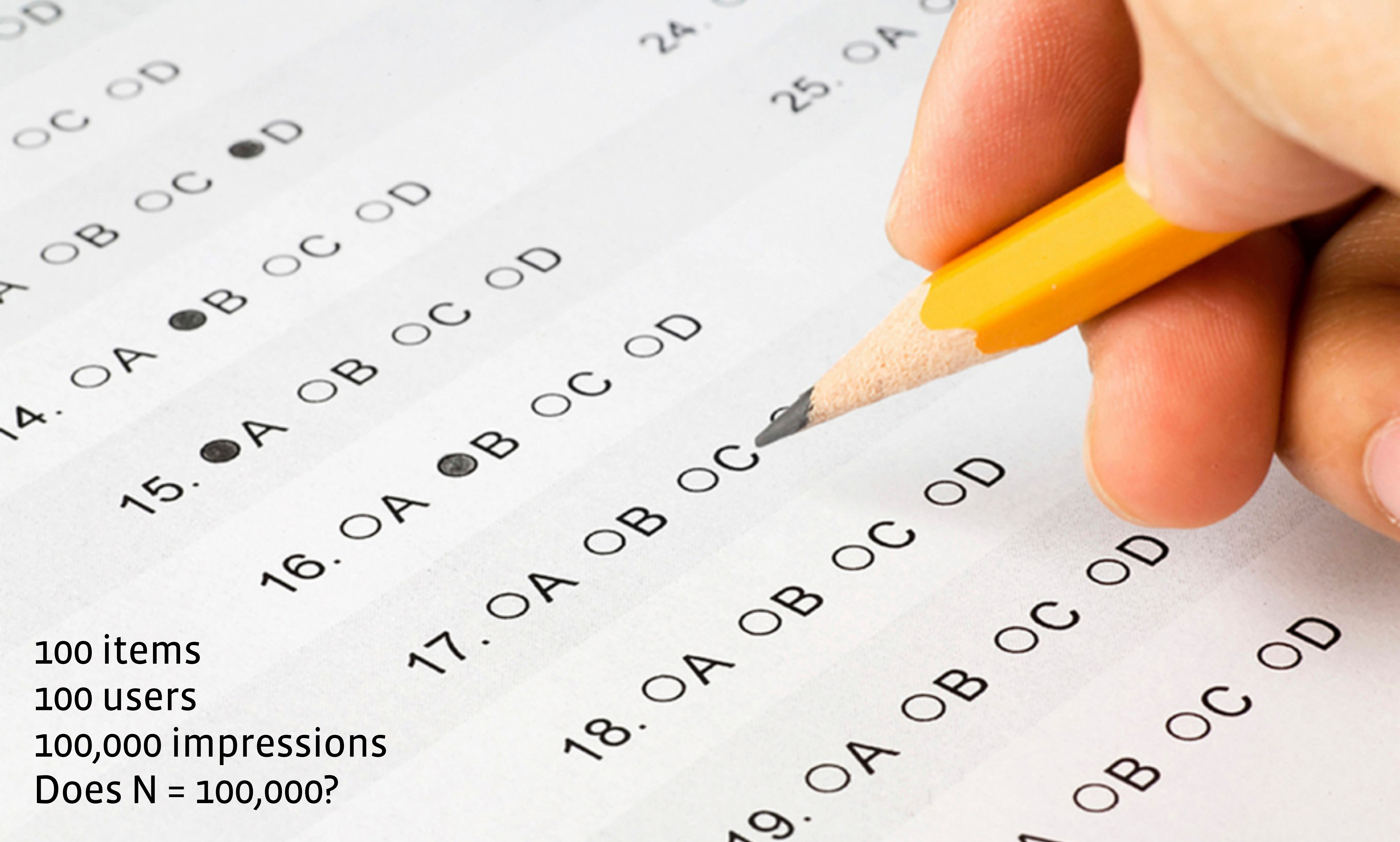
open: webapp/extract_data.py

facebook

Section 4:
Analyzing Experimental Data

Outline

1. Dependence (non- i.i.d. data)
2. The bootstrap
3. Using covariates
4. Data reduction
5. “Big Data Guide”
6. Example in R



100 items
100 users
100,000 impressions
Does N = 100,000?

http://www.tucsonsentinel.com/local/report/062712_az_students/state-natl-rankings-vary-widely-az-student-performance/

Dependence

- Most formulas for confidence intervals assume that each individual data point is independent of all the others.
- In practice, we often have repeated observations of users or content items.
- Ignoring this fact in inference will tend to make confidence intervals anti-conservative.

See Bakshy and Eckles, "*Uncertainty in Online Experiments with Dependent Data*" (KDD 2013)

The Bootstrap

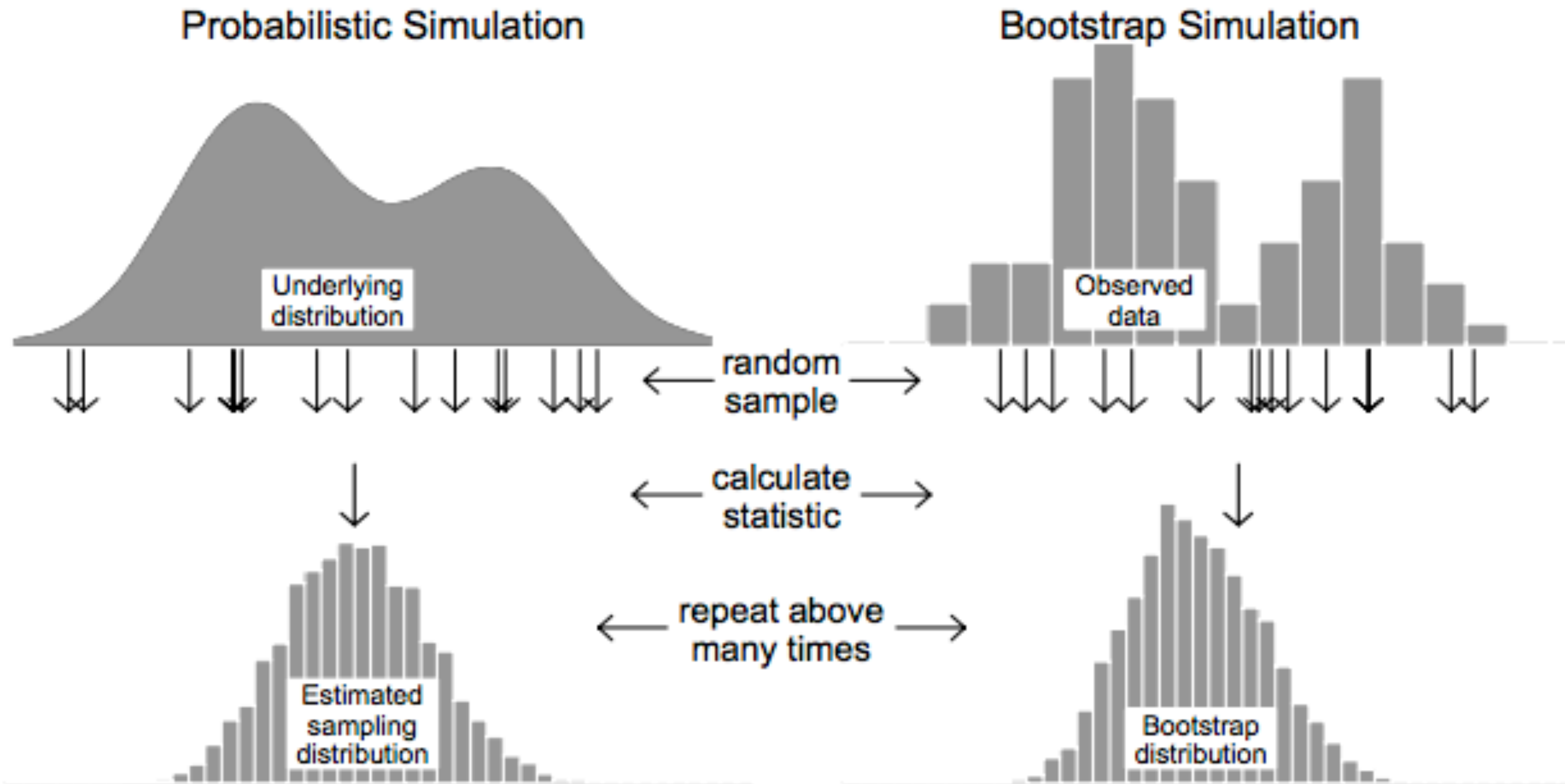
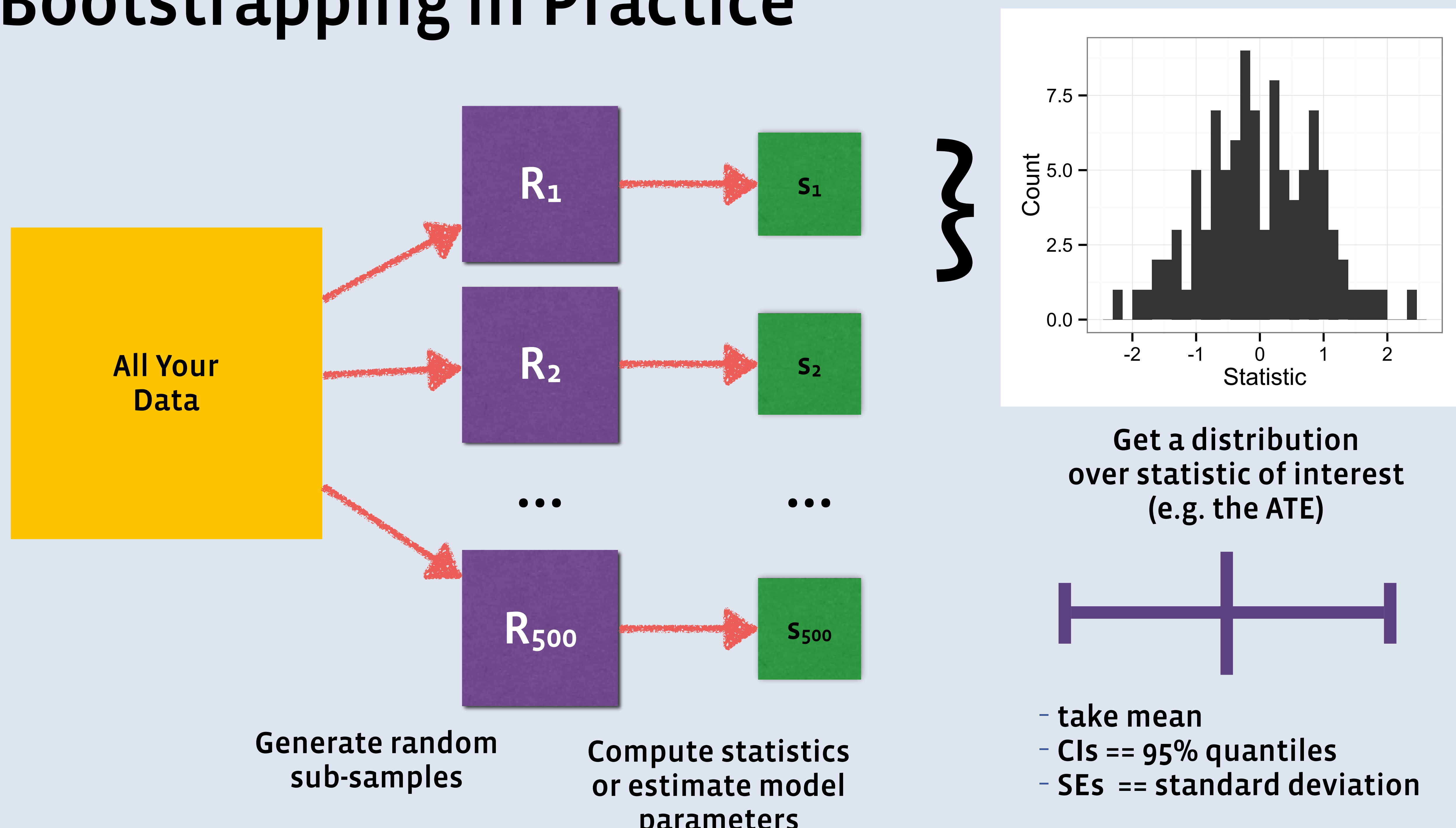


Figure 3: Diagram of probabilistic simulation and bootstrap sampling estimates of sampling distributions.

Bootstrapping in Practice



Using Covariates

- With simple random assignment, using covariates is **not necessary.**
- However, you can improve precision of ATE estimates if covariates explain a lot of variation in the potential outcomes.
- Can be added to a linear model and SEs should decrease if they are helpful.
- Should always at least report results <u>without</u> using covariates.

Data Reduction

Subject	D	Y
Evan	0	1
Ashley	0	1
Greg	1	0
Leena	1	0
Ema	0	0
Seamus	1	1

N

D	Y=1	Y=0
0	2	1
1	1	2

treatments

Data Reduction with Covariates

Subject	X	D	Y
Evan	M	0	1
Ashley	F	0	1
Greg	M	1	0
Leena	F	1	0
Ema	F	0	0
Seamus	M	1	1

N

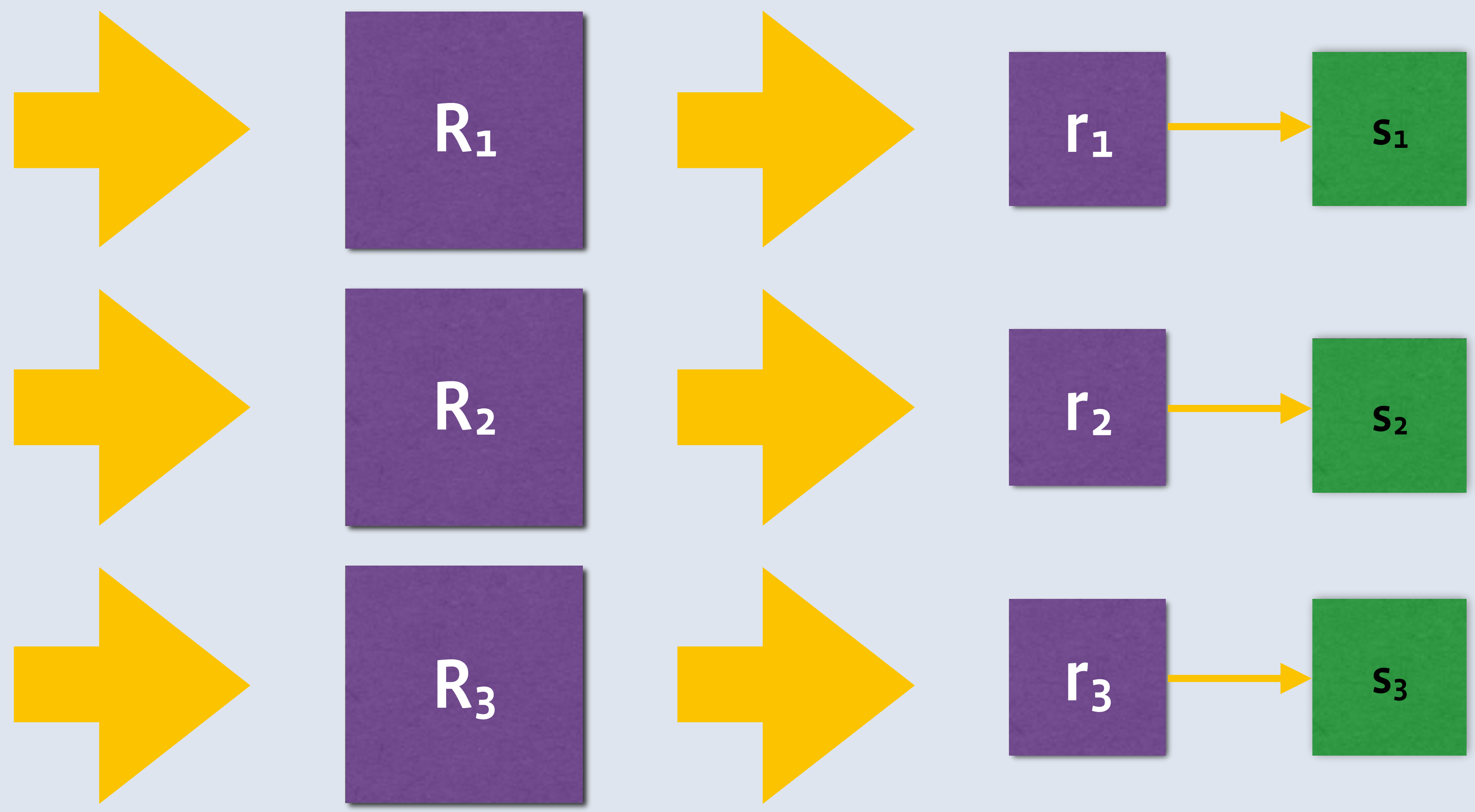
X	D	Y=1	Y=0
M	0	1	0
M	1	1	1
F	0		
F	1		

treatments X # groups

Can analyze the reduced data using a weighted linear model.

Data Reduction with Dependent Data

Subject	D	Y
Evan	1	1
Evan	1	0
Ashley	0	1
Ashley	0	1
Ashley	0	1
Greg	1	0
Leena	1	0
Leena	1	1
Ema	0	0
Seamus	1	1



Create bootstrap replicates

reduce the replicates as if they're i.i.d.

compute statistics on reduced data

Experiment Analysis (i.i.d. data)

Data Size	Fits in memory < 2M rows	Doesn't fit in memory
Using Covariates	linear regression	data reduction + weighted linear models
No Covariates	t-test	data reduction

Experiment Analysis (dependent data)

Data Size	Fits in memory < 2M rows	Doesn't fit in memory
Using Covariates	random effects models or bootstrap + linear models	bootstrap + data reduction + weighted linear models
No Covariates	random effects models or bootstrap in R	bootstrap + data reduction

Analyzing our Experimental Data

open: analyzing_experiments.R

facebook