Causal Inference in Data Science

@jonathandinu
jondinu@gmail.com
http://jonathanjonathanjonathan.com

Carnegie Mellon University

@jonathandinu

Getting the Materials

Code: https://github.com/hopelessoptimism/causality-for-the-uninitiated

Data: http://insideairbnb.com/get-the-data.html

What Causality?

*My headache is **now** gone because I took **two aspirins** an **hour ago** instead of just a **glass of water**.*

- Rubin Causal Model

Conditional Counterfactual

*If an **hour ago** I had taken **two aspirins** instead of just a **glass of water**, my headache would **now** be gone.*

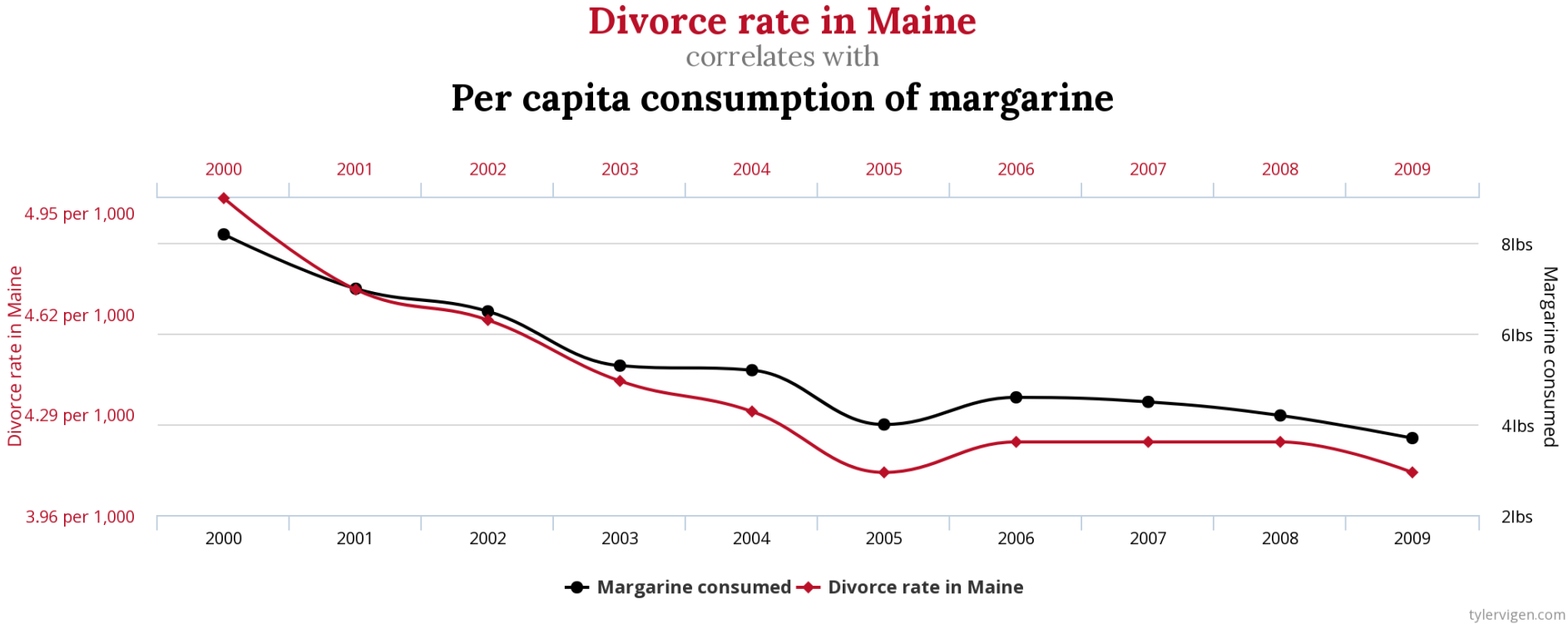
- Rubin Causal Model

What Causality?

*the causal effect of one treatment, **E**, over another, **C**, for a particular unit and an interval of time from **t1** to **t2** is the difference between what would have happened at time **t2** if the unit had been exposed to **E** initiated at **t1** and what would have happened at **t2** if the unit had been exposed to **C** initiated at **t1**.*

- Rubin Causal Model

Why Causality?



source: Tyler Vigen (https://www.tylervigen.com/spurious-correlations)

Who Causality?

source: Judea Pearl (http://bayes.cs.ucla.edu/jp_home.html)

Randomized Controlled Trials

The “Gold Standard”

Questions

- Do US eligible voters favor Hillary Clinton over Bernie Sanders?
- Is drug **A** effective at treating a medical condition?
- Has the Facebook News Feed significantly increased the amount of time users spend on the platform?
- Is Uber making NYC rush hour traffic worse?
- Has an unregulated AirBnB marketplace led to an increasing rate of evictions in San Francisco?

At the end of the day we want to reason in an informed way to make decisions

(hypothetical) Decisions

- Hillary Clinton should be the presumptive Democratic nominee
- Approve drug **A** for consumer use
- Roll out the Facebook News Feed interface to all users permanently
- Cap the number of Uber vehicles (and other services) in NYC
- Restrict private short term housing rentals in San Francisco to 75 nights per year

Case Study

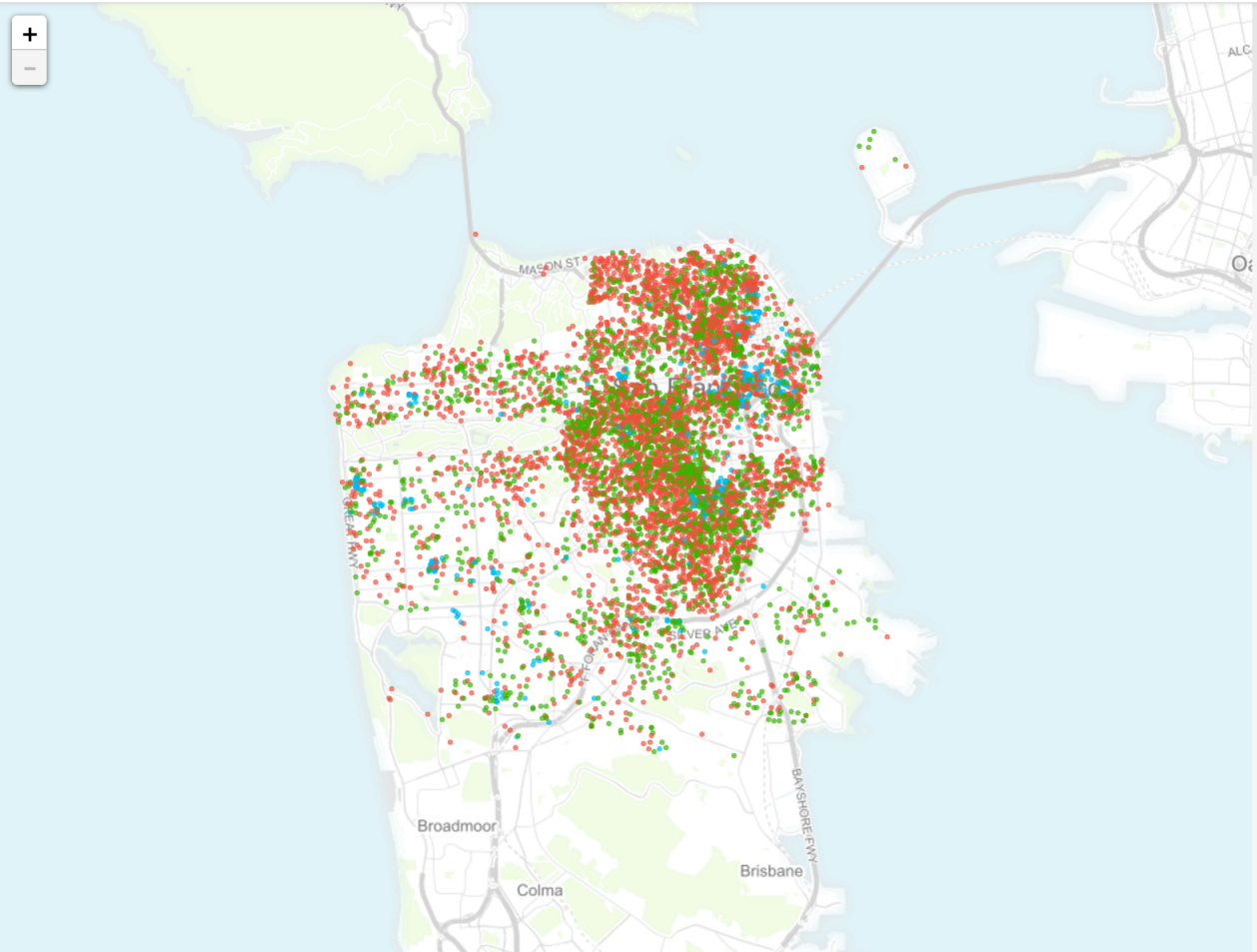
Inside Airbnb

insideairbnb.com/san-francisco/

Inside Airbnb
Adding data to the debate

About Behind Get the Data

Like Share 1.5K Tweet



San Francisco

Filter by: San Francisco

7,029
out of **7,029** listings (100%)

About Airbnb in San Francisco

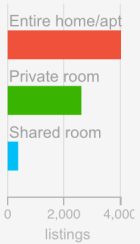
How is Airbnb really being used in and affecting your neighborhoods?

Room Type

☐ Only entire homes/apartments

Airbnb hosts can list entire homes/apartments, private or shared rooms.

Depending on the room type, availability, and activity, an airbnb listing could be more like a hotel, disruptive for neighbors, taking away housing, and illegal.



57.4%
entire homes/apartments

$221
price/night

4,033 (57.4%)
entire home/apartments

2,624 (37.3%)
private rooms

372 (5.3%)
shared rooms

Activity

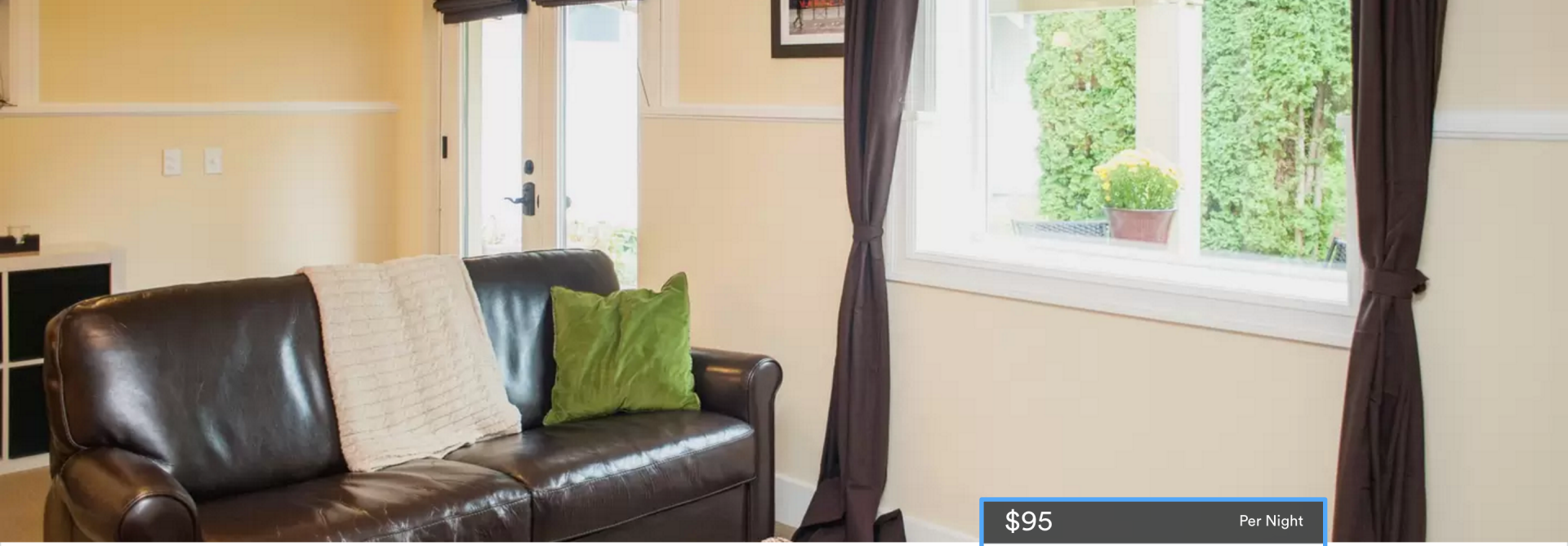
☐ Only **recent** and **frequently** booked

Airbnb guests may leave a review after their stay, and these can be used as an indicator of airbnb activity.

The minimum stay, price and number of reviews have been used to estimate the **occupancy rate**,

138
estimated nights/year

1.5
reviews/listing/month



Josh And Nicole

Sweet and Cozy Madrona Apartment

Seattle, WA, United States ★★★★½ (73)

Entire home/apt | 2 Guests | 1 Bedroom | 1 Bed

$95 Per Night

Check in: mm/dd/yyyy

Check out: mm/dd/yyyy

Guests: 1

Request to Book

What guests liked about this place

Residential | Cozy | Near restaurants and shops | Well-equipped kitchen | Private | Quiet

♡ Save to Wish List

256 travelers saved this place

$95 Per Night

Check in mm/dd/yyyy | Check out mm/dd/yyyy | Guests 1

Request to Book

About this listing

We have a recently remodeled sweet, clean, cheery, MIL apartment on the first floor of our 1901 Craftsman home in the historic Madrona neighborhood. We are on the bus line, blocks from cafes, restaurants and lovely Lake Washington.

Contact Host

Business Travel — This listing has essential business travel amenities. Learn More

The Space

Accommodates: **2**
Bathrooms: **1**
Bed type: **Real Bed**
Bedrooms: **1**
Beds: **1**
House Rules

Check In: **Anytime after 4PM**
Check Out: **11AM**
Property type: **Apartment**
Room type: **Entire home/apt**

Amenities

Kitchen
TV
Internet
Essentials
+ More

Prices

Extra people: **No Charge**
Cleaning Fee: **$40**
Security Deposit: **$150**

Weekly discount: **17%**
Monthly discount: **0%**
Cancellation: **Moderate**

Description

The Space

The is a MIL apartment on the first floor of our 1901 Craftsman home. It has been recently remodeled with a full kitchen, heated floors in the bathroom and kitchen, a cozy gas stove, a queen sleep number bed, a propane grill, patio, TV, cable and wifi. The apartment is ideally suited to fit one or two people but a small family may be comfortable if the children sleep on the small sofas. Please let us know if you need extra bedding for a child.

Save to Wish List

256 travelers saved this place

Email | Messenger | ··· More

Report this listing

$95 Per Night

Check in mm/dd/yyyy Check out mm/dd/yyyy Guests 1

Request to Book

73 Reviews

Search reviews

Summary

Accuracy	★★★★★	Location	★★★★★
Communication	★★★★★	Check In	★★★★★
Cleanliness	★★★★★	Value	★★★★★

Lucas

My wife, my little pug and I felt like home. The location is beautiful and really close to downtown. The house was perfect to spend a couple of days. I really recommend it ;-) Lucas

May 2016 Helpful

Tracy

Josh met us when we arrived. The apartment is in a lovely residential neighbourhood with great views and interesting architecture, both old and new. The place was comfy and private. We were very comfortable. There is a cute outdoor area. We didn't get a chance to use the BBQ!

April 2016 Helpful

Stephanie

We had a wonderful stay at this place! Clean, cozy, and in a great location. Would stay again!

March 2016 Helpful

Lorenz & Barbara

Josh and Nicole: We loved you place. Wonderful neighborhood and very comfy little flat- perfect base for our Seattle Symphony "side-by-side". Brigid's cello teacher, Olivia, lives near you apparently. We very much hope to return- Emily has a piano recital in June. We will be in touch! You were very grascious and helpful hosts! Thanks so much! (URL HIDDEN)Lefty & Barbara Schultz

Save to Wish List

256 travelers saved this place

Email Messenger ··· More

Report this listing

Exploratory Question

What variables affect the price of a listing?

Questions

AirBnB

- When listings are recommended, are there more reservations on the AirBnB platform?
- Do listings in neighborhoods with more restaurants and shops have more bookings?

The Public

- Do cities with AirBnB have higher rents (than those without)?
- Since AirBnB has been operating in NYC, have hotel rates fallen?

Hypothesis Statement

“Superhosts” have more expensive listings

Hypothesis Statement

“Superhosts” have more expensive listings
(implicit assumption: since they can attract more guests)

Confounding Variables

- Do they have the same volume of guests?
- Is is an equivalent price per # of guest or sq. footage? Do they just happen to have bigger places?
- What else is correlated with price?
    - (correlation can be useful is showing causation!)

Hypothesis Statement

More on this later…

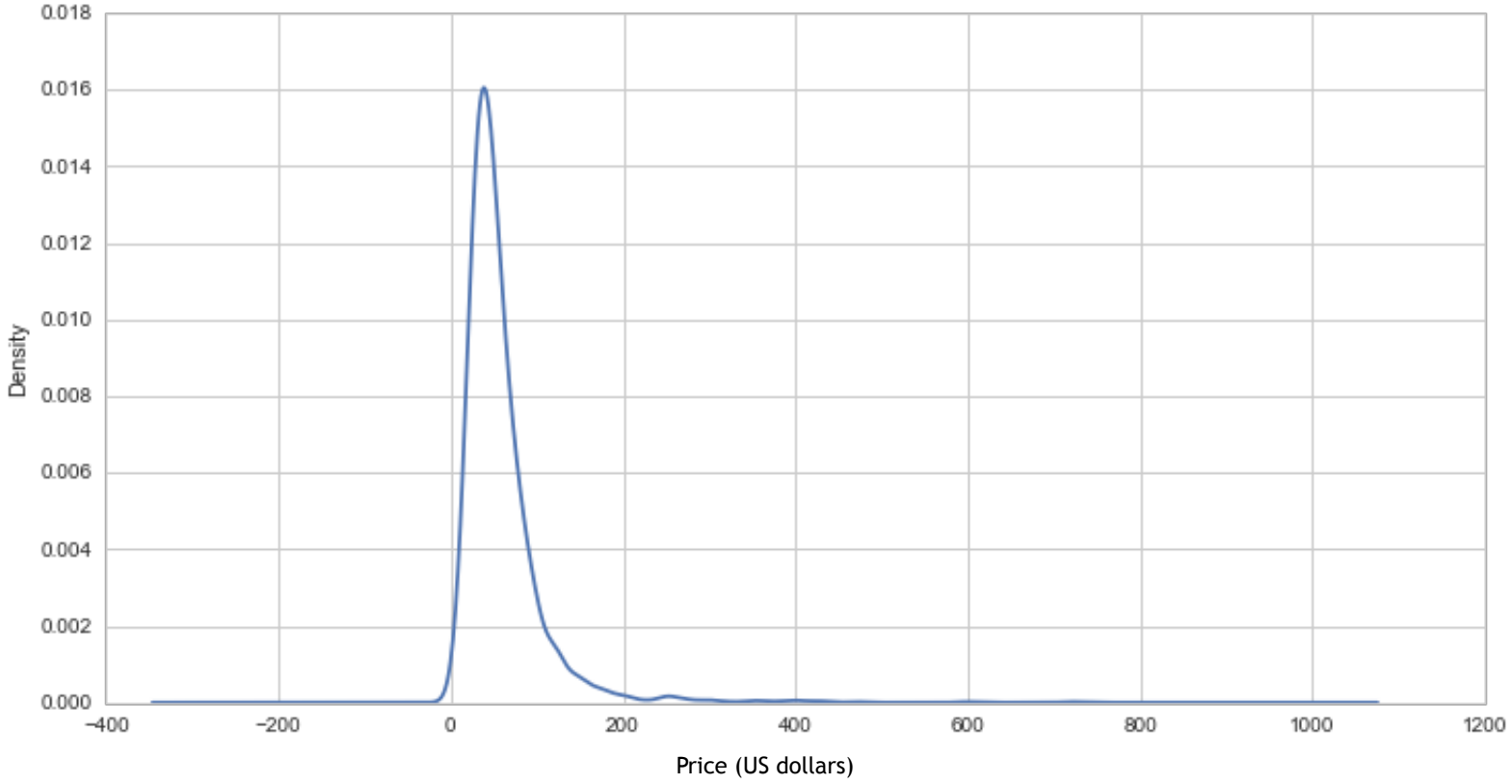
Question for Inference

What variables are correlated with the price of a listing?

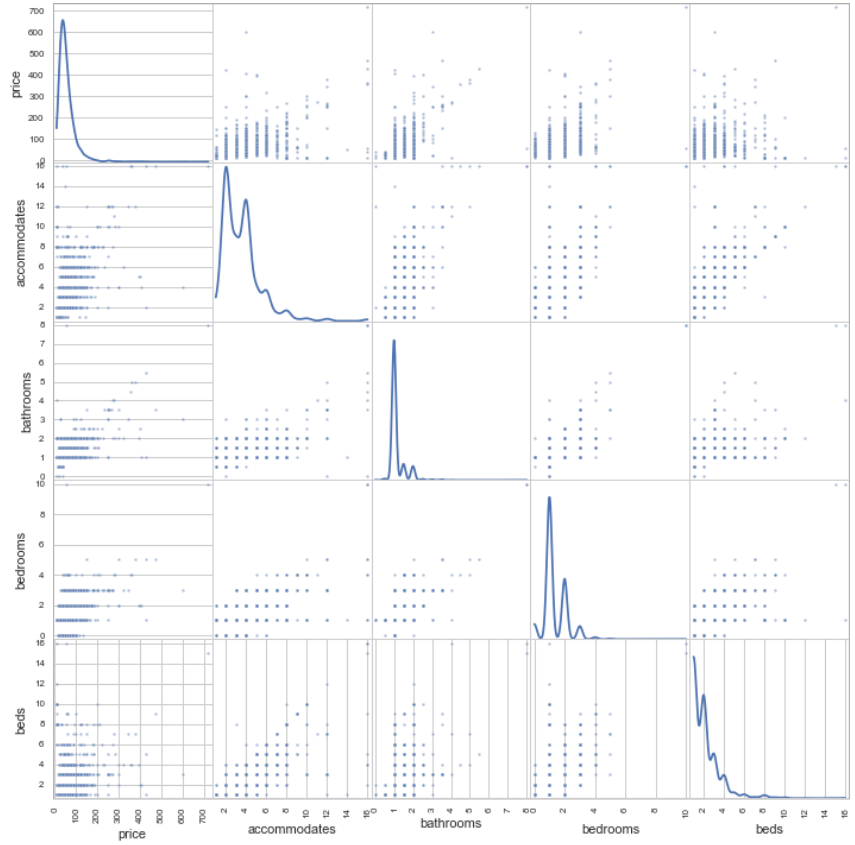
Correlation

Determining relationship between two variables

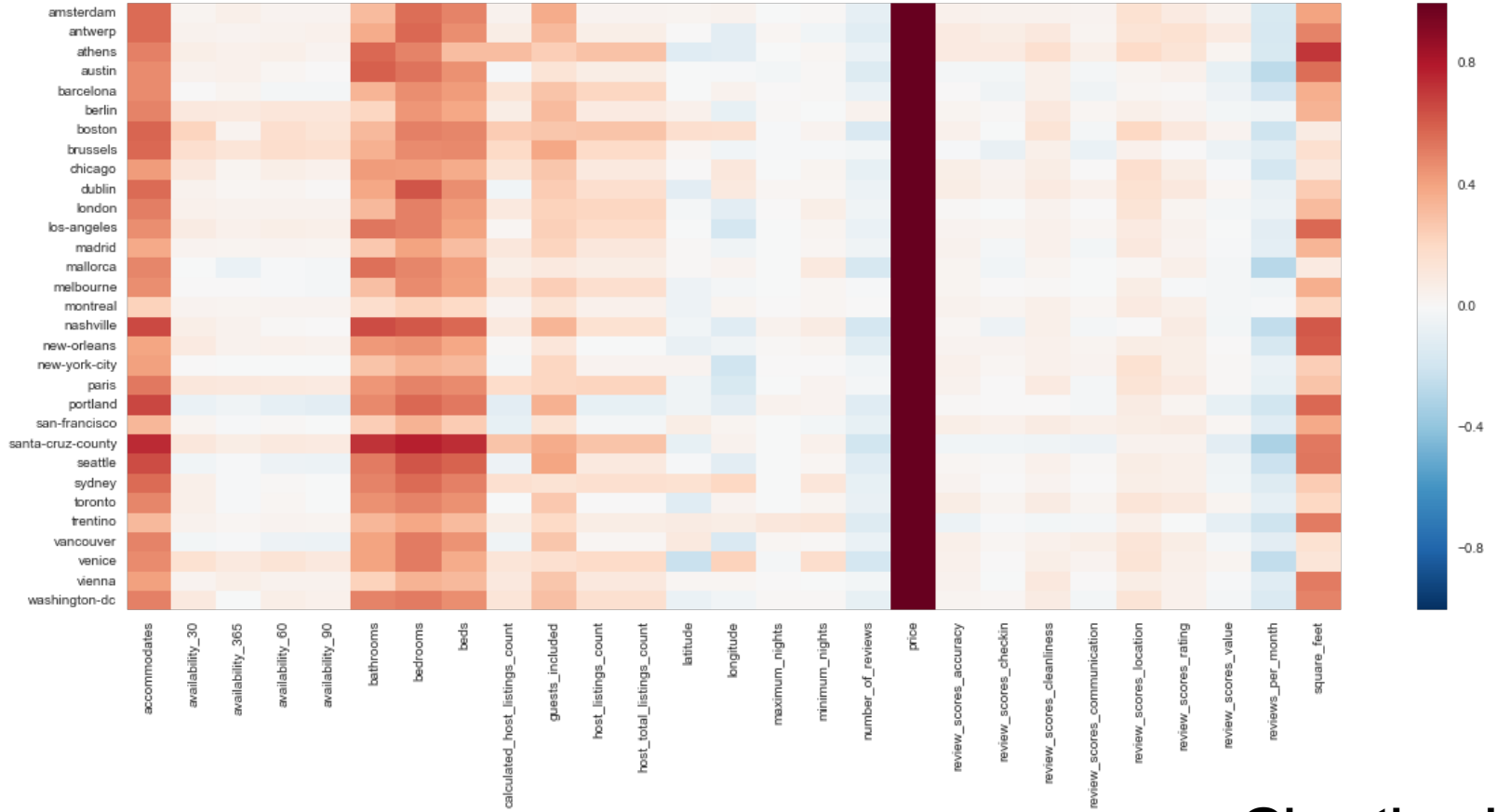
Price Distribution



Correlation: Visually



Correlation: Analytically

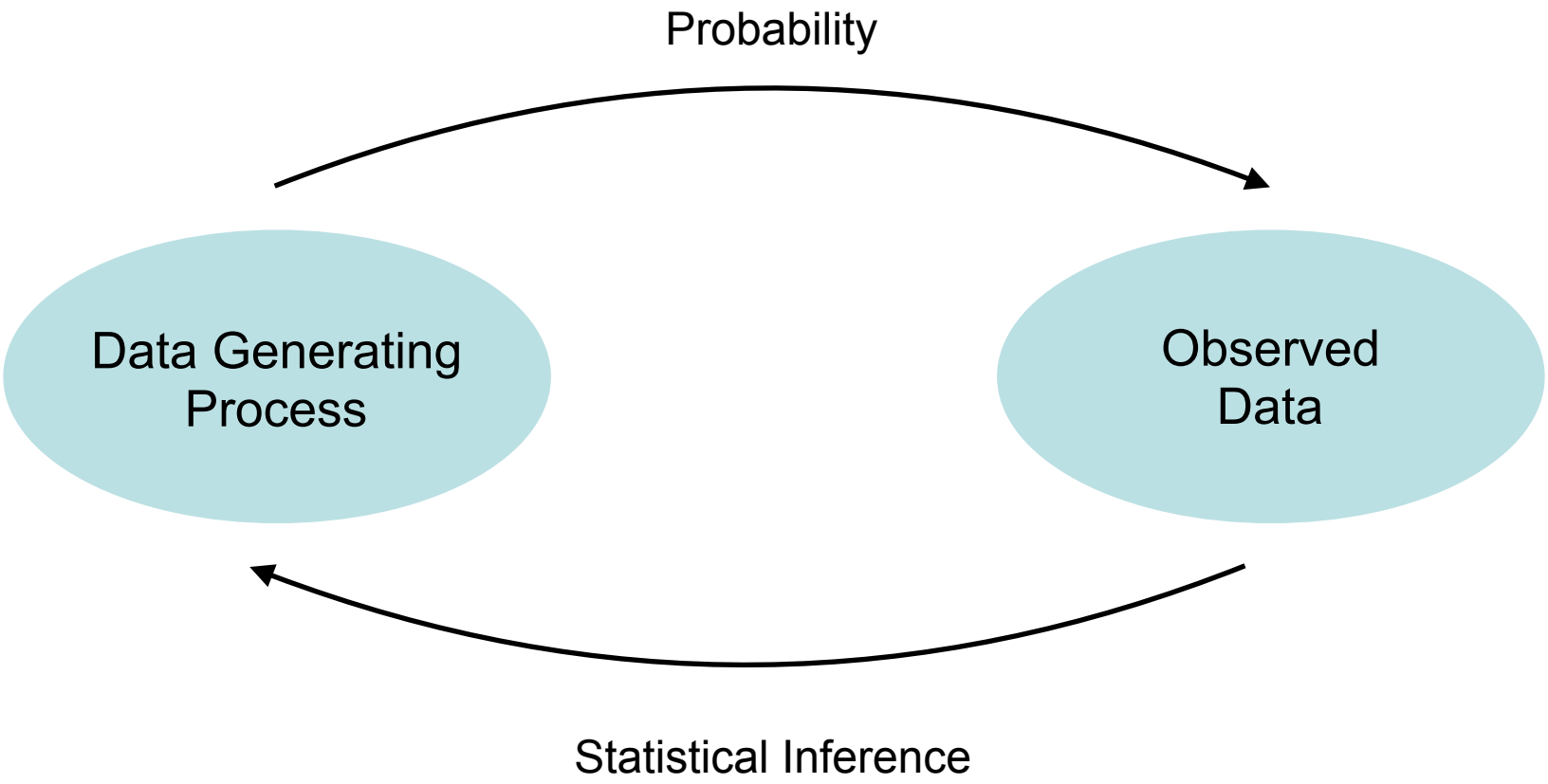


@jonathandinu

Q&A (5 min)

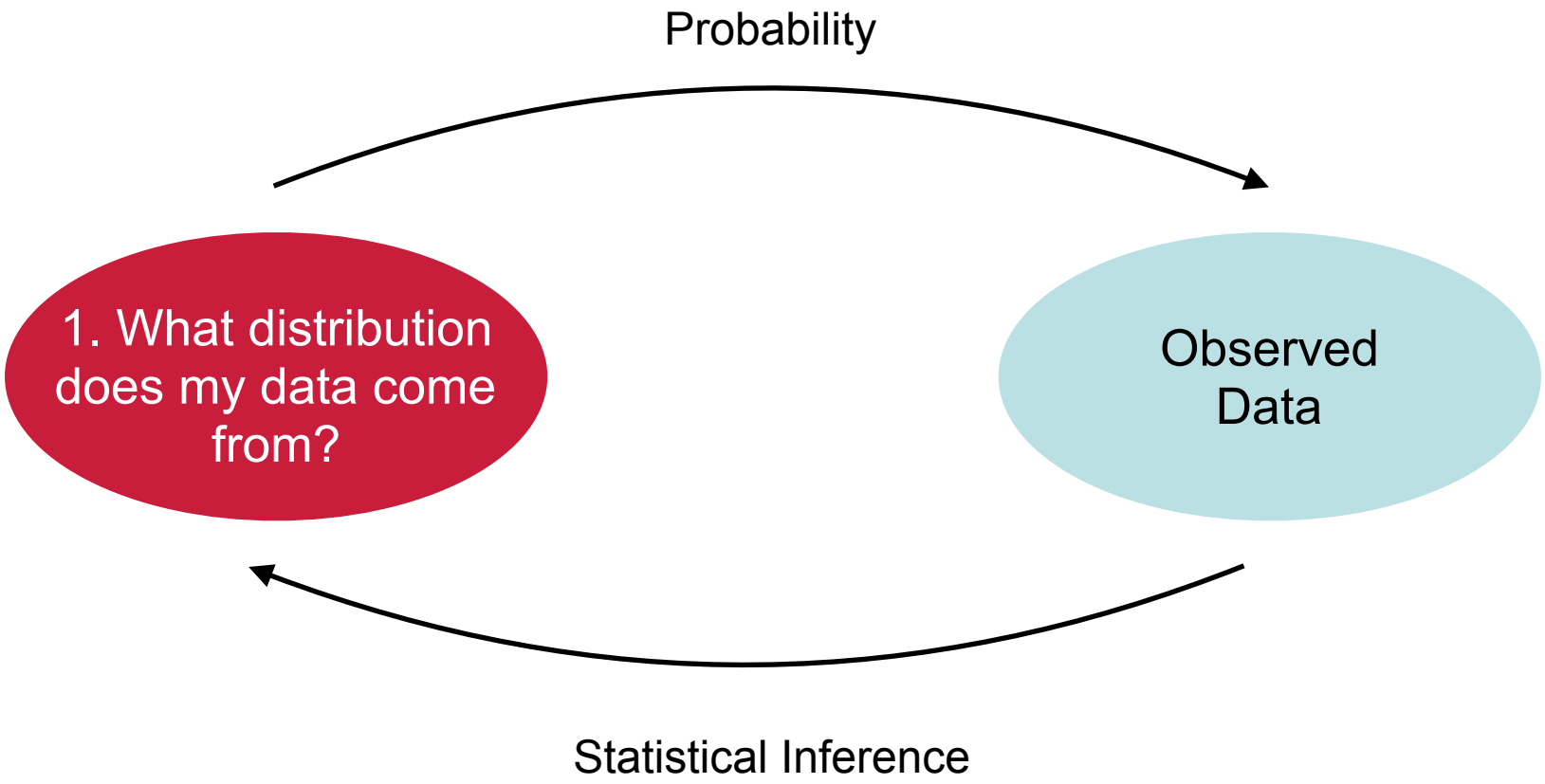
Hypothesis Testing

The Parametric Framework



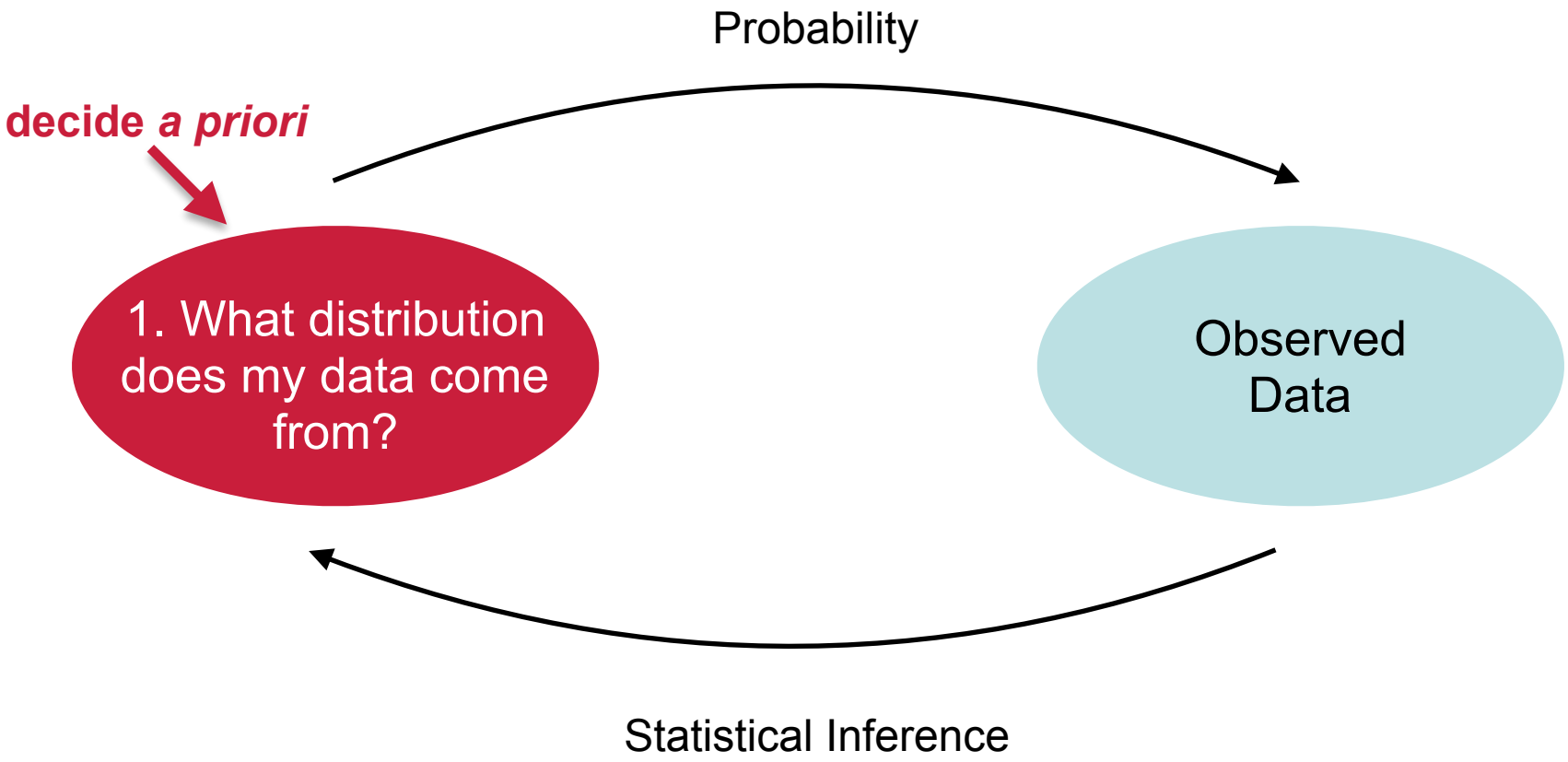
source: All of Statistics, Larry Wasserman

The Parametric Framework



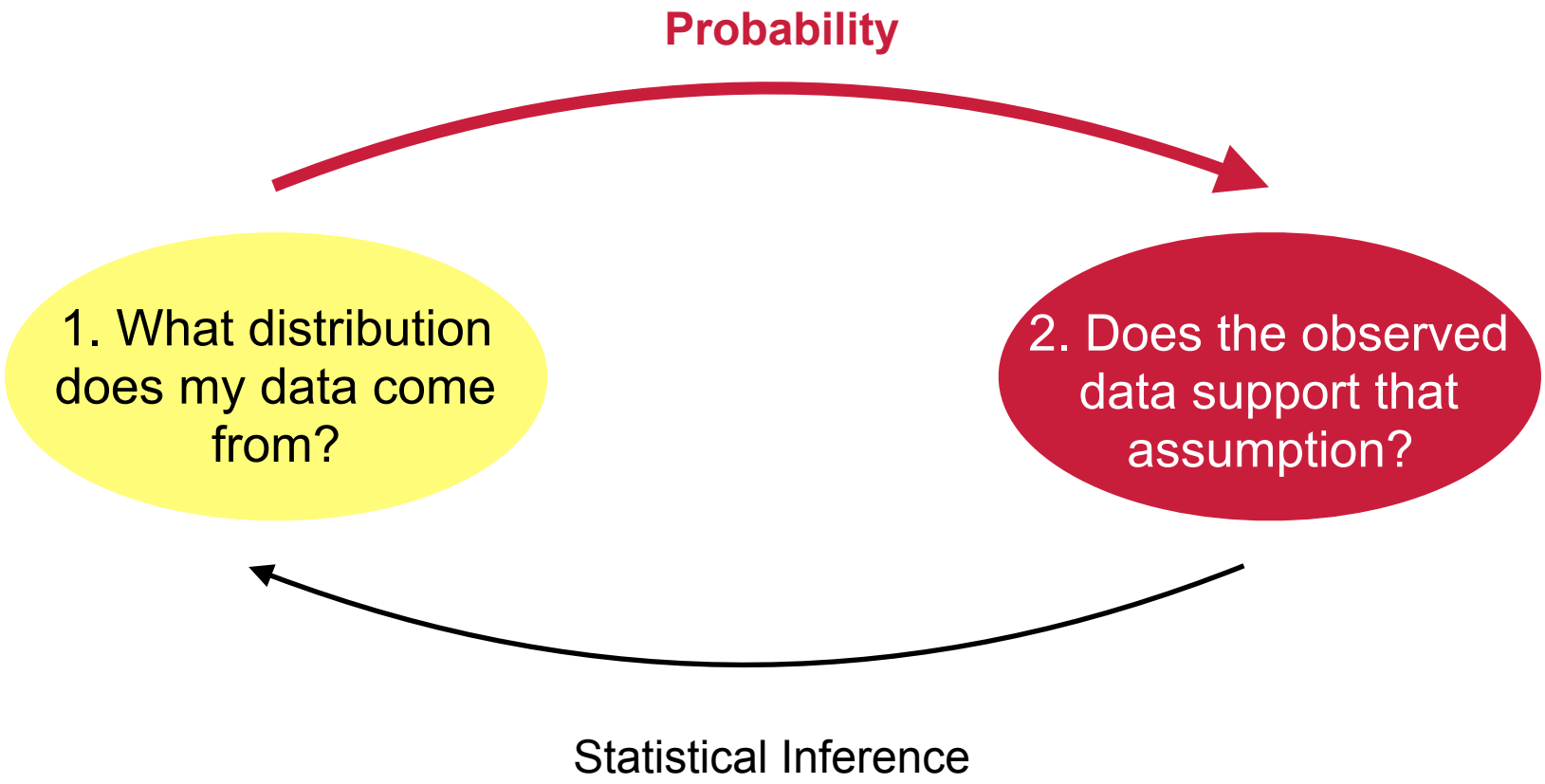
source: All of Statistics, Larry Wasserman

The Parametric Framework



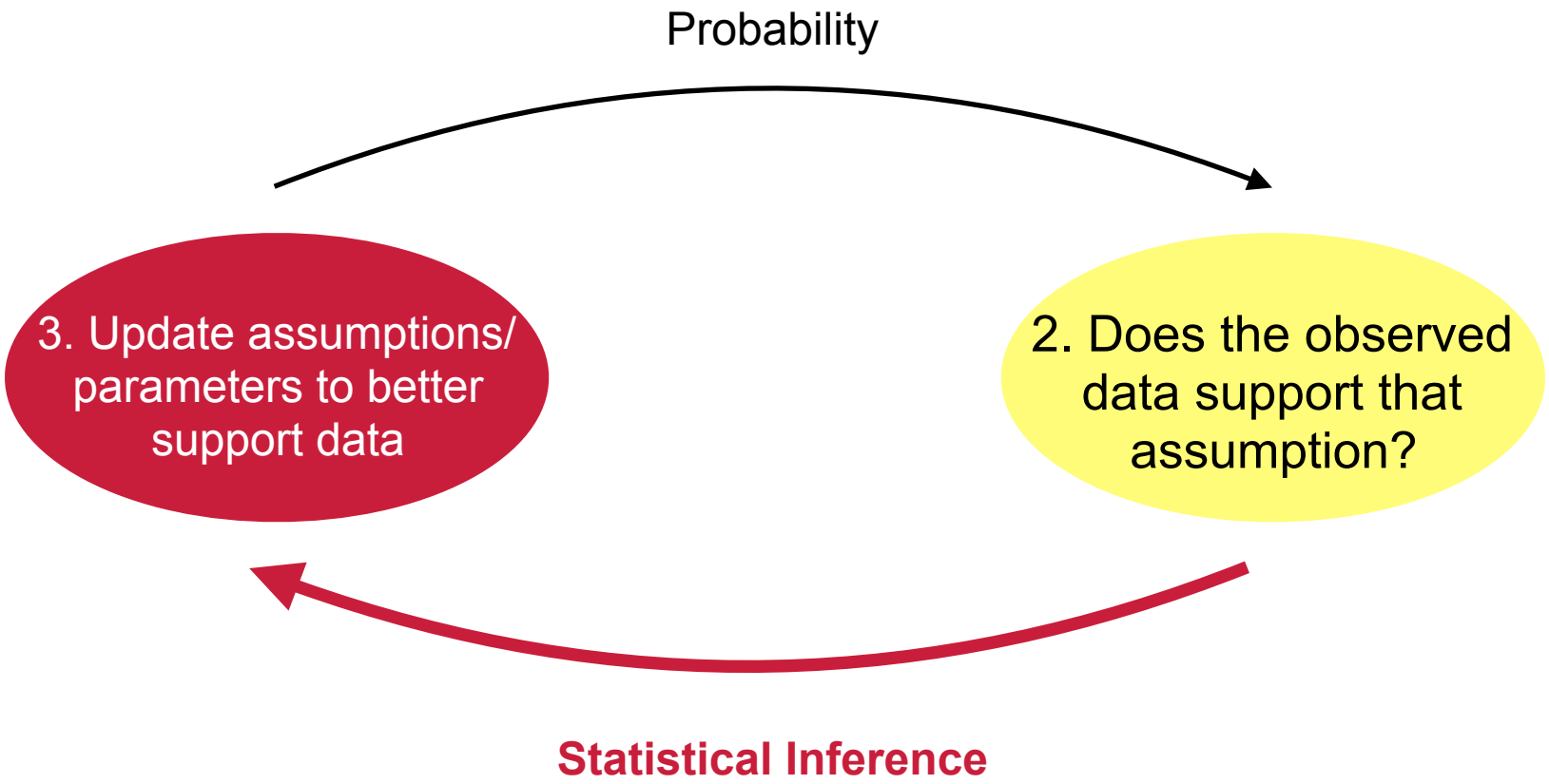
source: All of Statistics, Larry Wasserman

The Parametric Framework



source: All of Statistics, Larry Wasserman

The Parametric Framework



source: All of Statistics, Larry Wasserman

Hypothesis Testing Terminology

Null Hypothesis

A default position that there is **no relationship** between two measured phenomena (or no association among groups)

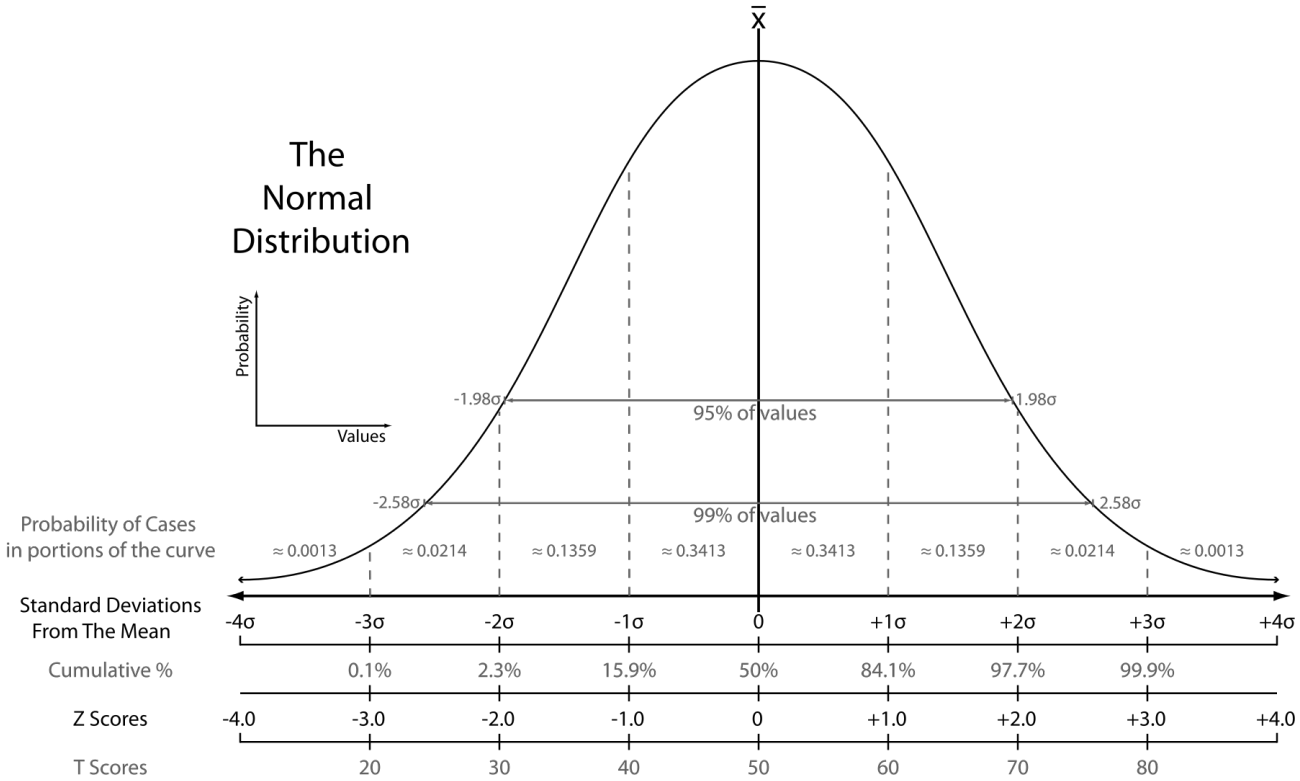
Z-score

Number of standard deviations away from the mean of the **sampling distribution**

p-value

the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true

The Normal Distribution



source: Wikipedia

Settings

Solve for? (•) Power ( ) Alpha ( ) n ( ) d

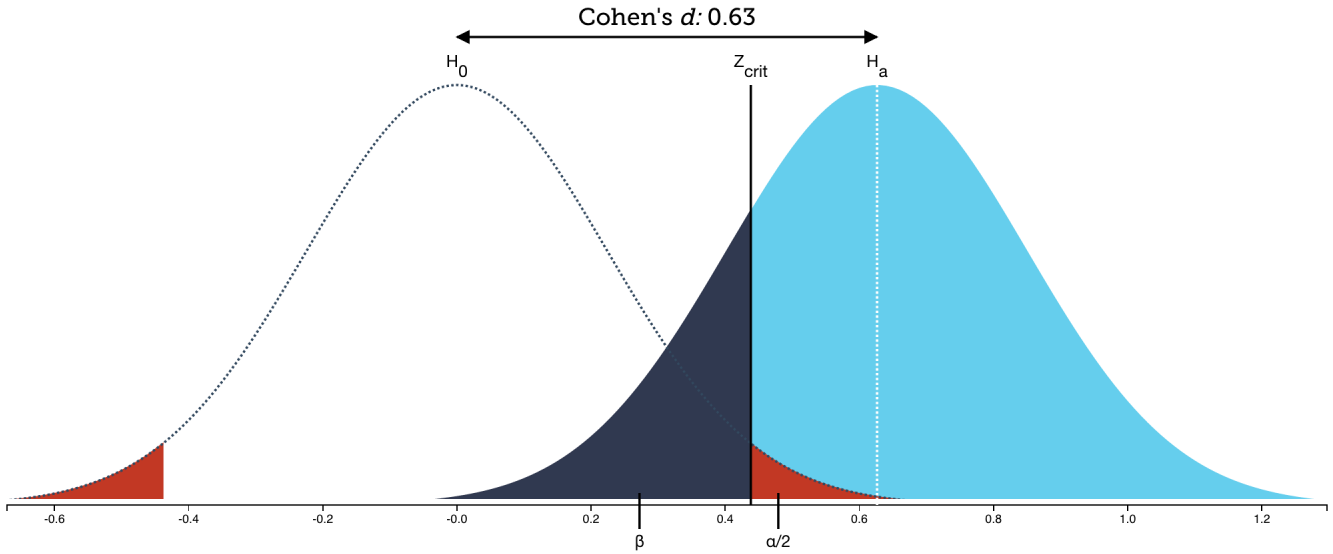
Significance level (α = 0.05)

Sample size (n = 20)

Effect size (d = 0.63)

One-tailed | Two-tailed

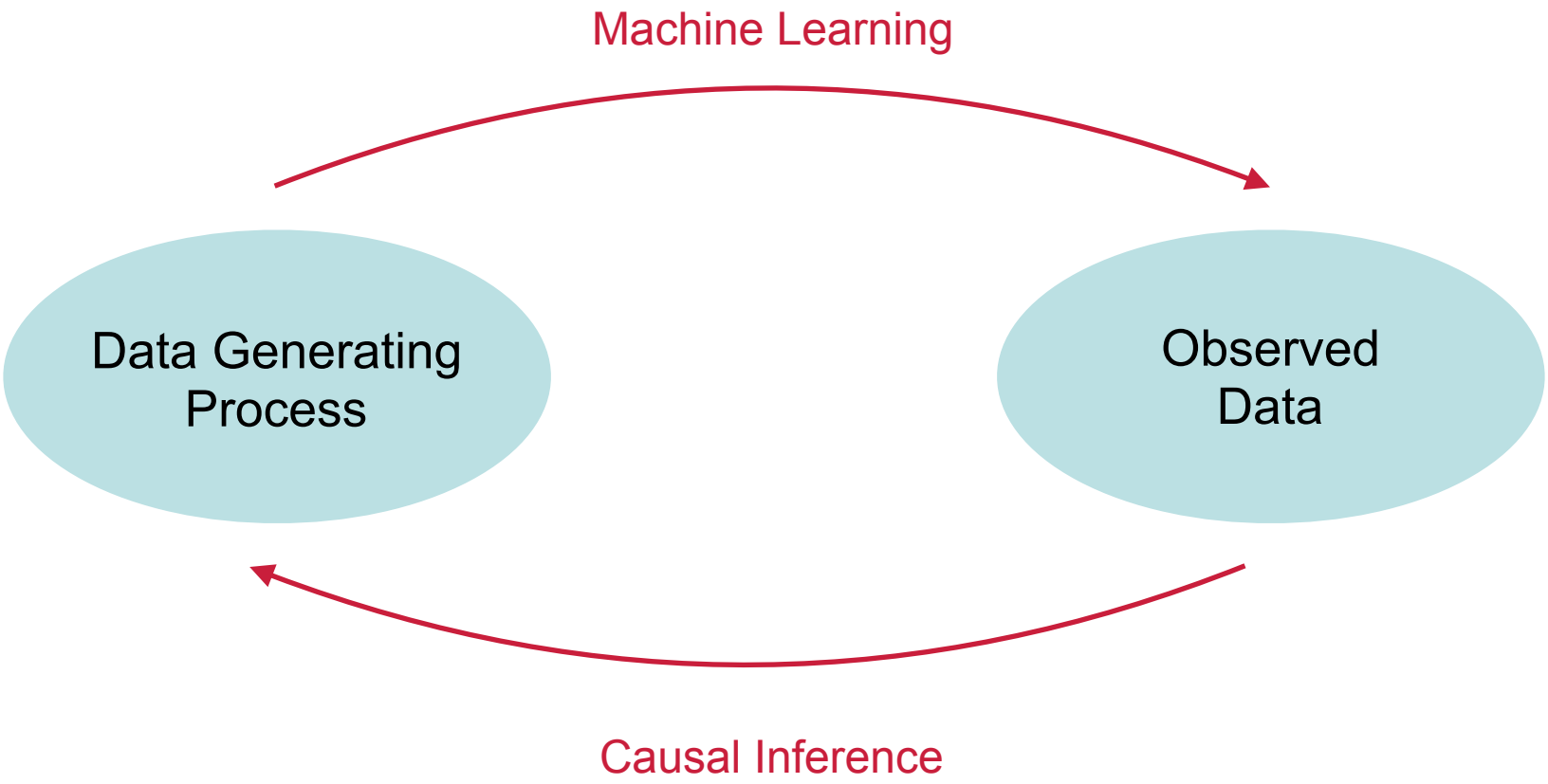
Reset zoom



5 %	20 %	80 %	20
Type I error	Type II error	Power	Sample size

Observational vs Experimental

Machine Learning vs Causal Inference



Statistics

Descriptive

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (correlation and covariance)

Inferential

- Estimation (MoM, MLE, MAP)
- Confidence Intervals
- Hypothesis Testing
- Bootstrap Methods
- Regression

Predictive

- Modeling (generative distributions)
- Machine Learning
  - Linear Regression
  - Logistic Regression
  - K Nearest Neighbors

Statistics

Understand

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (correlation and covariance)

Decide

- Estimation (MoM, MLE, MAP)
- Confidence Intervals
- Hypothesis Testing
- Bootstrap Methods
- Regression

Forecast

- Modeling (generative distributions)
- Machine Learning
  - Linear Regression
  - Logistic Regression
  - K Nearest Neighbors

Statistics

Past

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (correlation and covariance)

Present

- Estimation (MoM, MLE, MAP)
- Confidence Intervals
- Hypothesis Testing
- Bootstrap Methods
- Regression

Future

- Modeling (generative distributions)
- Machine Learning
  - Linear Regression
  - Logistic Regression
  - K Nearest Neighbors

Statistics

Descriptive

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (correlation and covariance)

Inferential

- Estimation (MoM, MLE, MAP)
- **Confidence Intervals**
- **Hypothesis Testing**
- Bootstrap Methods
- Regression

Predictive

- Modeling (generative distributions)
- Machine Learning
  - Linear Regression
  - Logistic Regression
  - K Nearest Neighbors

Statistics

Descriptive

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (correlation and covariance)

Inferential

- Estimation (MoM, MLE, MAP)
- **Confidence Intervals**
- **Hypothesis Testing**
- Bootstrap Methods
- Regression

Predictive

- Modeling (generative distributions)
- Machine Learning
  - Linear Regression
  - Logistic Regression
  - K Nearest Neighbors

Optimal… if conditions and assumptions are valid

Statistics

Descriptive

- Central Tendency (**mean**, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (**correlation** and covariance)

Inferential

- Estimation (MoM, MLE, MAP)
- Confidence Intervals
- Hypothesis Testing
- Bootstrap Methods
- Regression

Predictive

- Modeling (generative distributions)
- Machine Learning
  - Linear Regression
  - Logistic Regression
  - K Nearest Neighbors

Many make (erroneous) causal claims from descriptive measures…

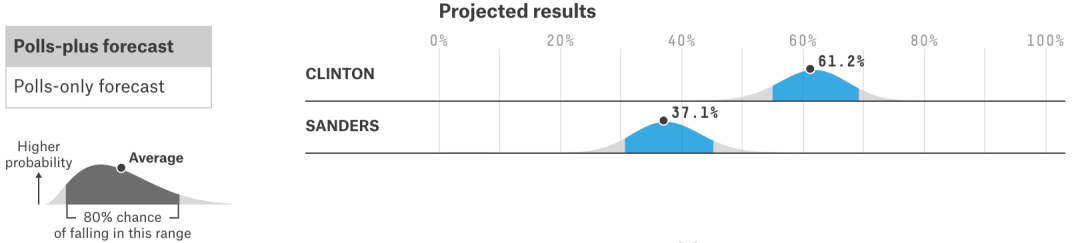
Break (10 min)

Enter Causal Inference

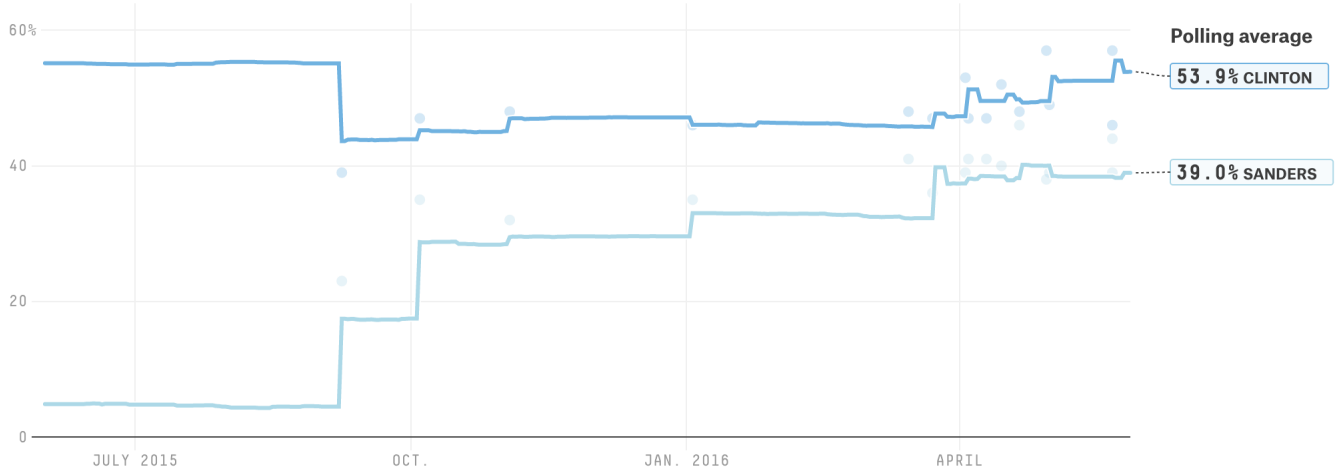
When Causal Inference is Necessary

- Randomized assignment is { unethical | costly | impossible }
- Observational data and post-hoc analysis
- Counterfactual (what-if) questions
- Common in Economics, Public Health, Political Science, …

Reasoning in an Uncertain World

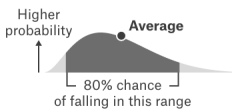


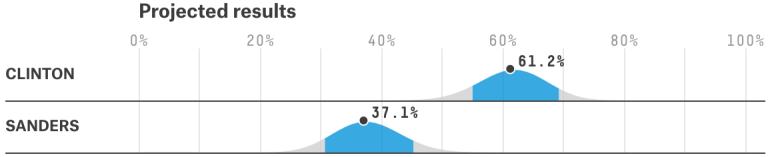
Not all **polls** ● are created equal, so our forecasts are calculated based on weighted **polling averages** 〰.



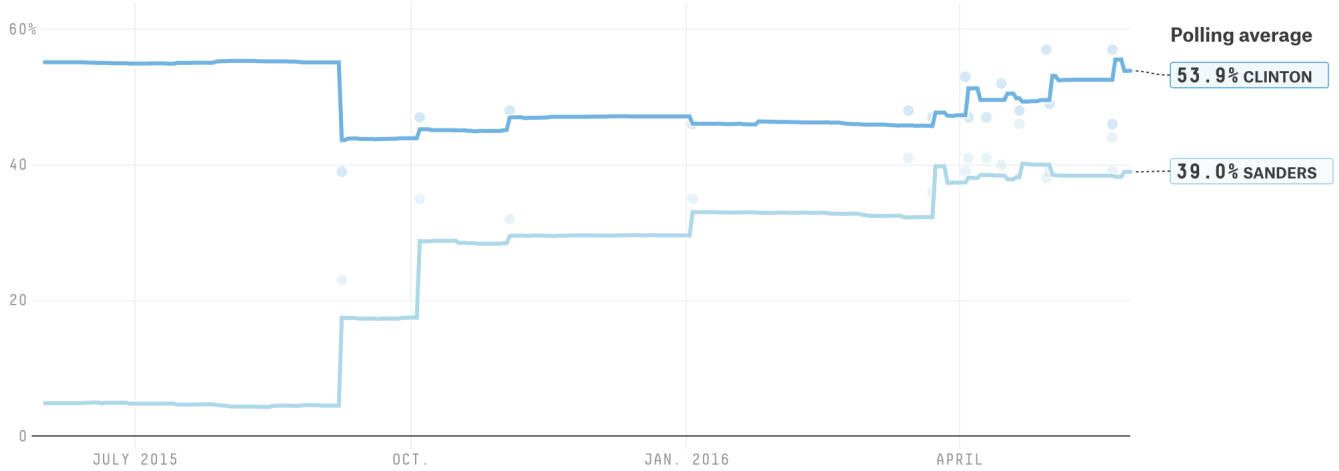
Polls-plus forecast

Polls-only forecast





Not all **polls** ● are created equal, so our forecasts are calculated based on weighted **polling averages** 〰.



● = NEW A = ALL ADULTS RV = REGISTERED VOTERS LV = LIKELY VOTERS

	POLLSTER	SAMPLE	WEIGHT	LEADER	CLINTON	SANDERS
● MAY 19-22	SurveyUSA	803 LV	1.48	Clinton +18	57%	39%
● MAY 13-22	PPIC	552 LV	0.34	Clinton +2	46%	44%
APR. 27-30	SurveyUSA	826 LV	0.05	Clinton +19	57%	38%
APR. 28-MAY 1	Sextant Strategies	1,617 RV	0.04	Clinton +10	49%	39%
APR. 18-21	Fox News	623 LV	0.01	Clinton +2	48%	46%
APR. 13-15	YouGov	1,123 LV	0.01	Clinton +12	52%	40%
APR. 7-10	Gravis Marketing	846 LV	0.00	Clinton +6	47%	41%
MAR. 24-APR. 4	Field Poll	584 LV	0.00	Clinton +6	47%	41%
MAR. 30-APR. 3	SurveyUSA	767 LV	0.00	Clinton +14	53%	39%
MAR. 16-23	USC Dornsife	363 LV	0.00	Clinton +11	47%	36%

Show more polls ▾

*Leader or runner-up is not in the race.

If you can't find a contest in the dropdown menu above, it's because there hasn't been enough polling in that state yet. We'll add new polling averages and forecasts as soon as the data is available. Notice any bugs or missing polls? Send us an email.

By Jay Boice, Aaron Bycoffe, Harry Enten, Ritchie King, Dhrumil Mehta, Andrei Scheinkman and Nate Silver.

● = NEW A = ALL ADULTS RV = REGISTERED VOTERS LV = LIKELY VOTERS

	POLLSTER	SAMPLE	WEIGHT	LEADER	CLINTON	SANDERS
● MAY 19-22	SurveyUSA	803 LV	1.48	Clinton +18	57%	39%
● MAY 13-22	PPIC	552 LV	0.34	Clinton +2	46%	44%
APR. 27-30	SurveyUSA	826 LV	0.05	Clinton +19	57%	38%
APR. 28-MAY 1	Sextant Strategies	1,617 RV	0.04	Clinton +10	49%	39%
APR. 18-21	Fox News	623 LV	0.01	Clinton +2	48%	46%
APR. 13-15	YouGov	1,123 LV	0.01	Clinton +12	52%	40%
APR. 7-10	Gravis Marketing	846 LV	0.00	Clinton +6	47%	41%
MAR. 24-APR. 4	Field Poll	584 LV	0.00	Clinton +6	47%	41%
MAR. 30-APR. 3	SurveyUSA	767 LV	0.00	Clinton +14	53%	39%
MAR. 16-23	USC Dornsife	363 LV	0.00	Clinton +11	47%	36%

Show more polls ▾

*Leader or runner-up is not in the race.

If you can't find a contest in the dropdown menu above, it's because there hasn't been enough polling in that state yet. We'll add new polling averages and forecasts as soon as the data is available. Notice any bugs or missing polls? Send us an email.

By Jay Boice, Aaron Bycoffe, Harry Enten, Ritchie King, Dhrumil Mehta, Andrei Scheinkman and Nate Silver.

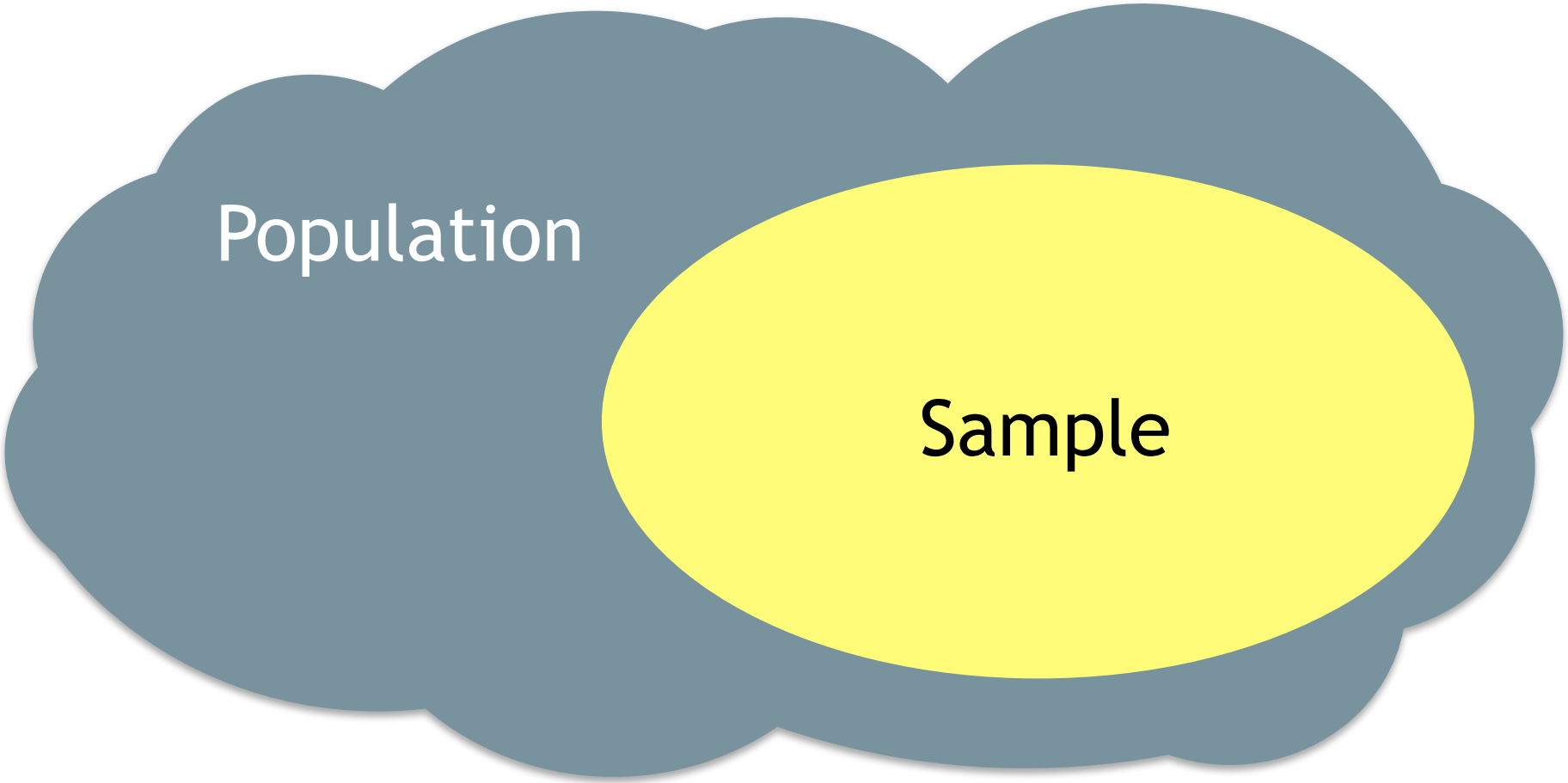
Conditional Counterfactual

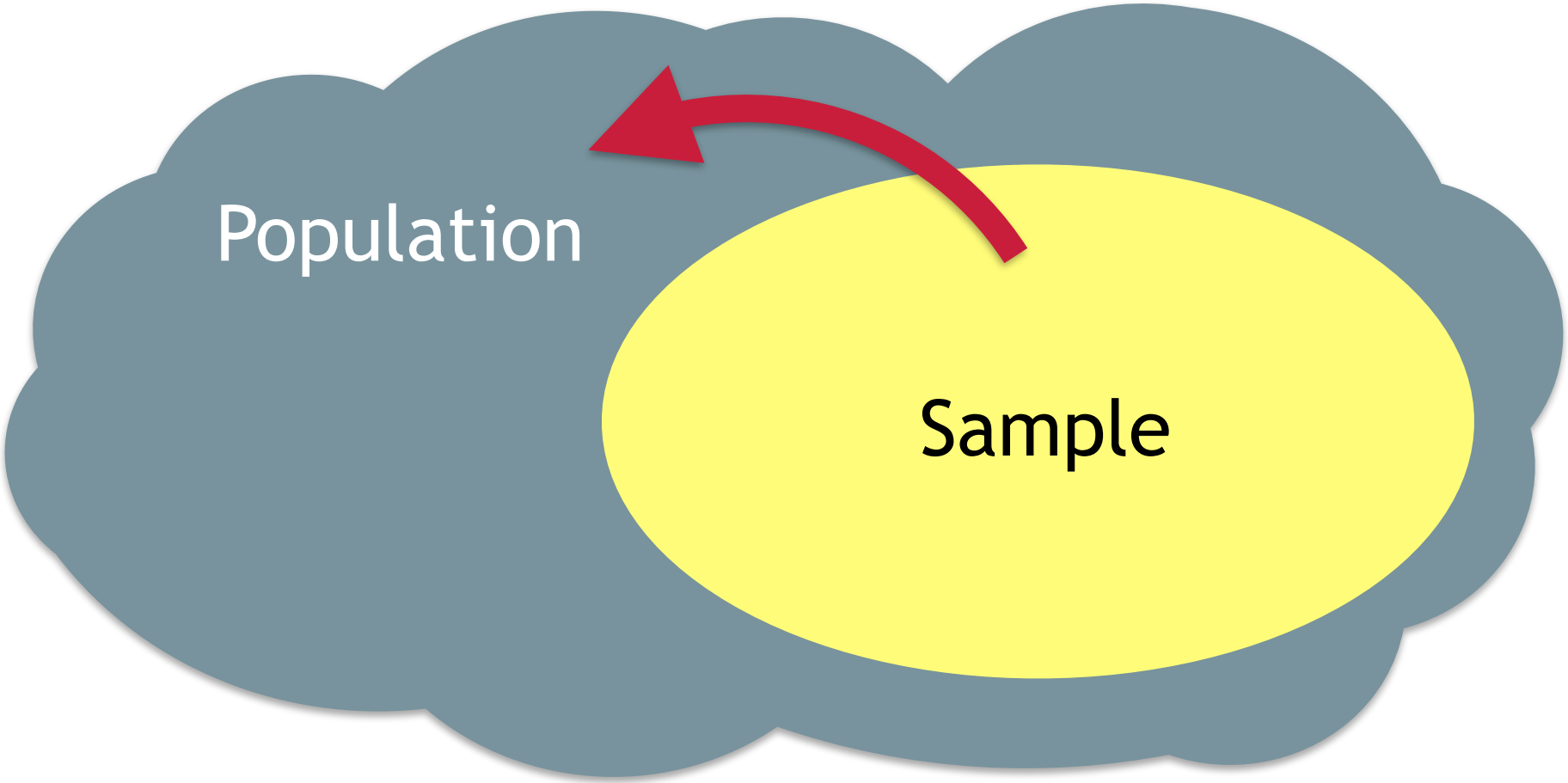
*If Candidate A had supported **Medicare for All** instead of **for-profit hospitals** during **the primary**, would they be the nominee for the **general election**.*

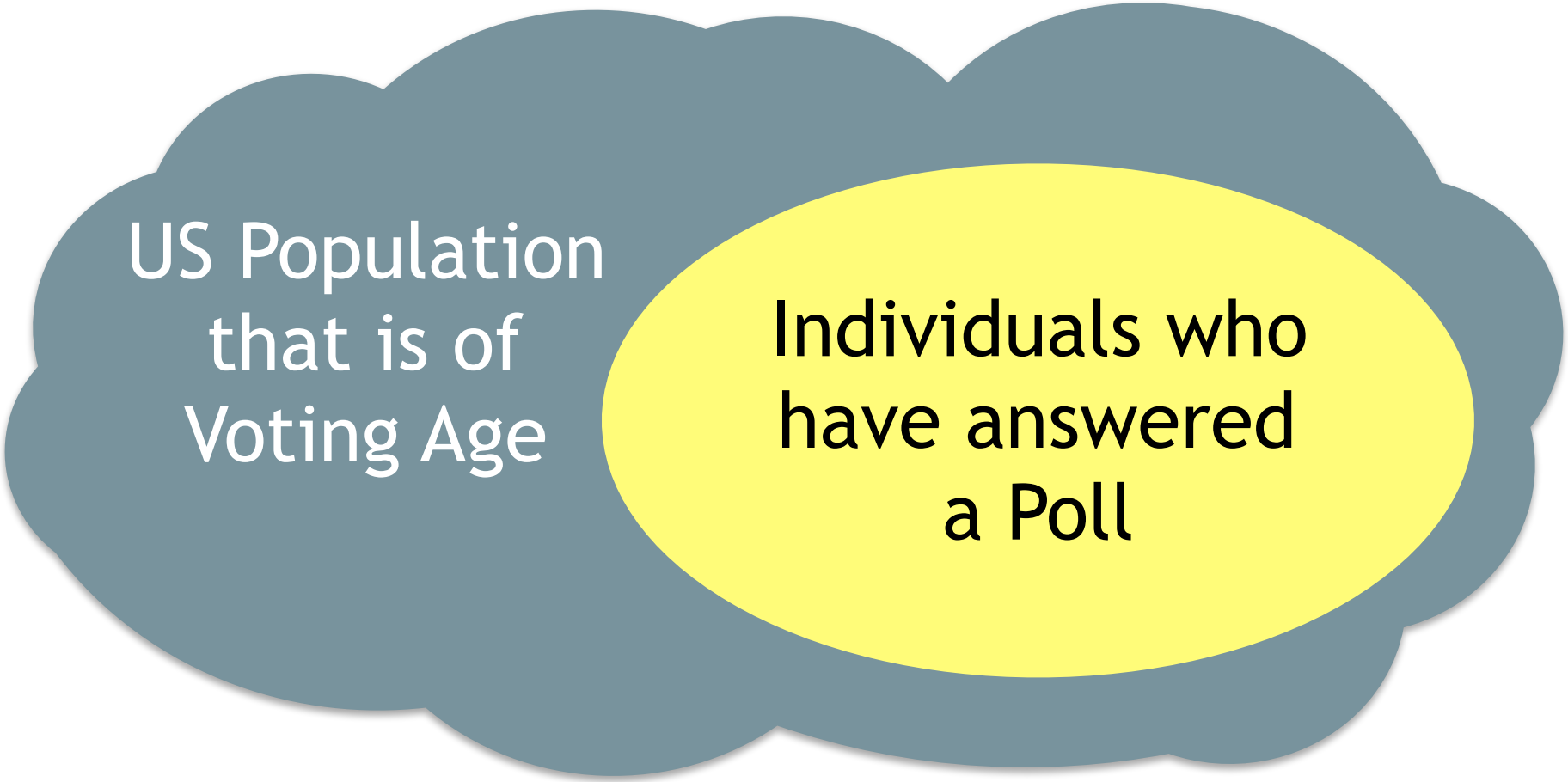
- Rubin Causal Model

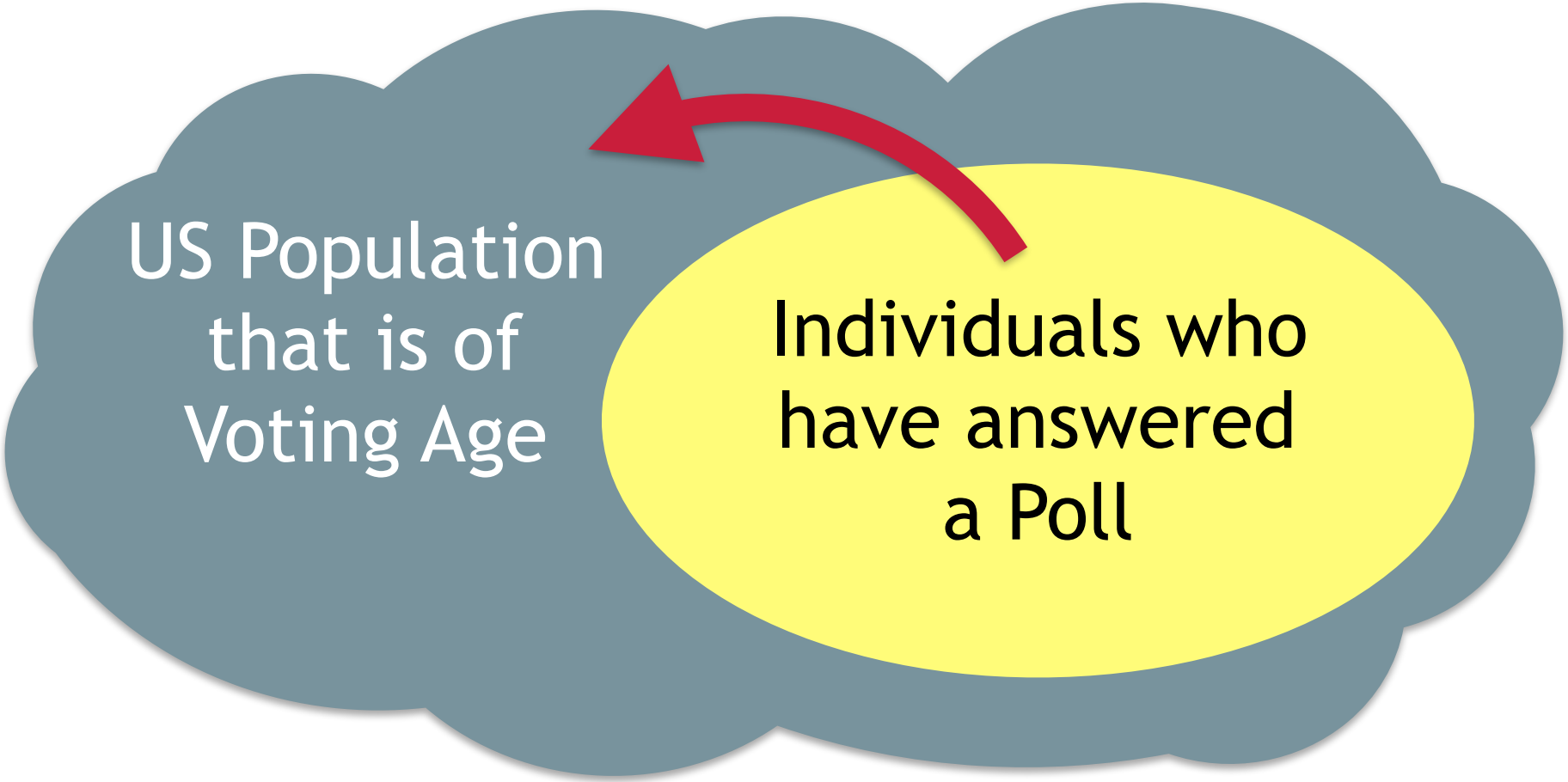
Random assignment impossible

Can’t hold multiple simultaneous elections….


Population
Sample


Population
Sample


US Population that is of Voting Age
Individuals who have answered a Poll


US Population that is of Voting Age
Individuals who have answered a Poll

Questions

AirBnB

- When listings are recommended, are there more reservations on the AirBnB platform?
- Do listings in neighborhoods with more restaurants and shops have more bookings?

The Public

- Do cities with AirBnB have higher rents (than those without)?
- Since AirBnB has been operating in NYC, have hotel rates fallen?

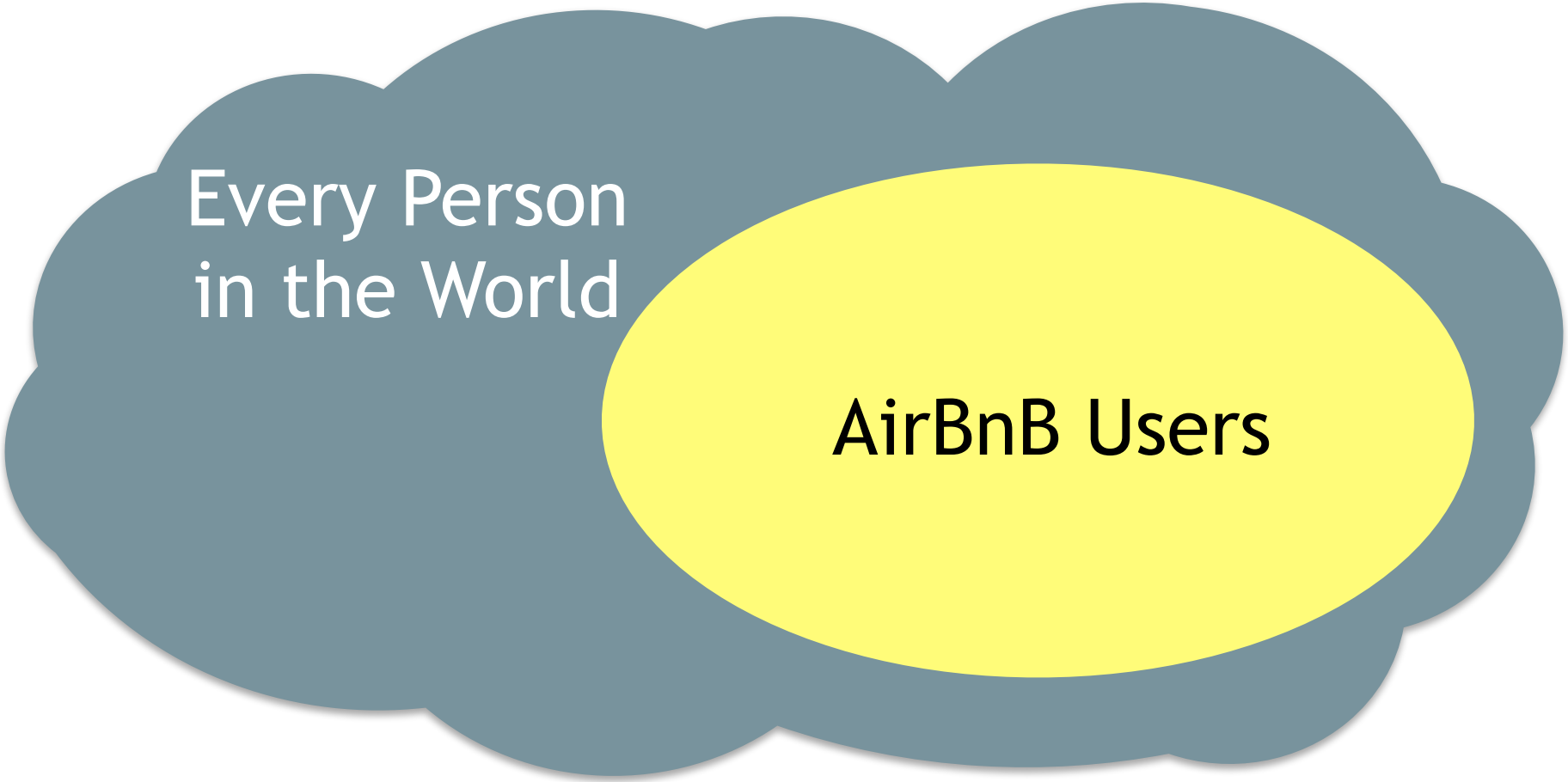
Causal Questions

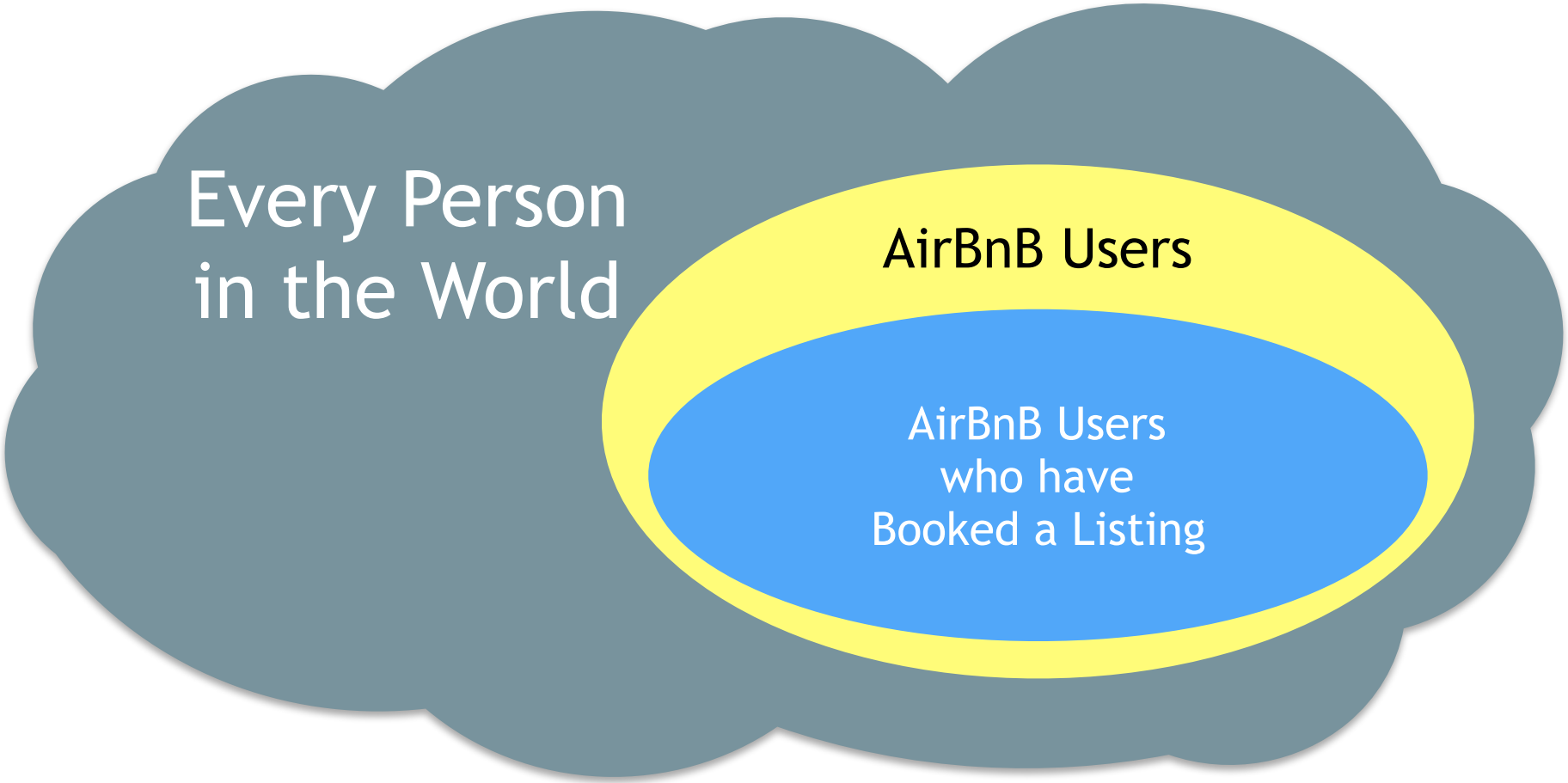
AirBnB

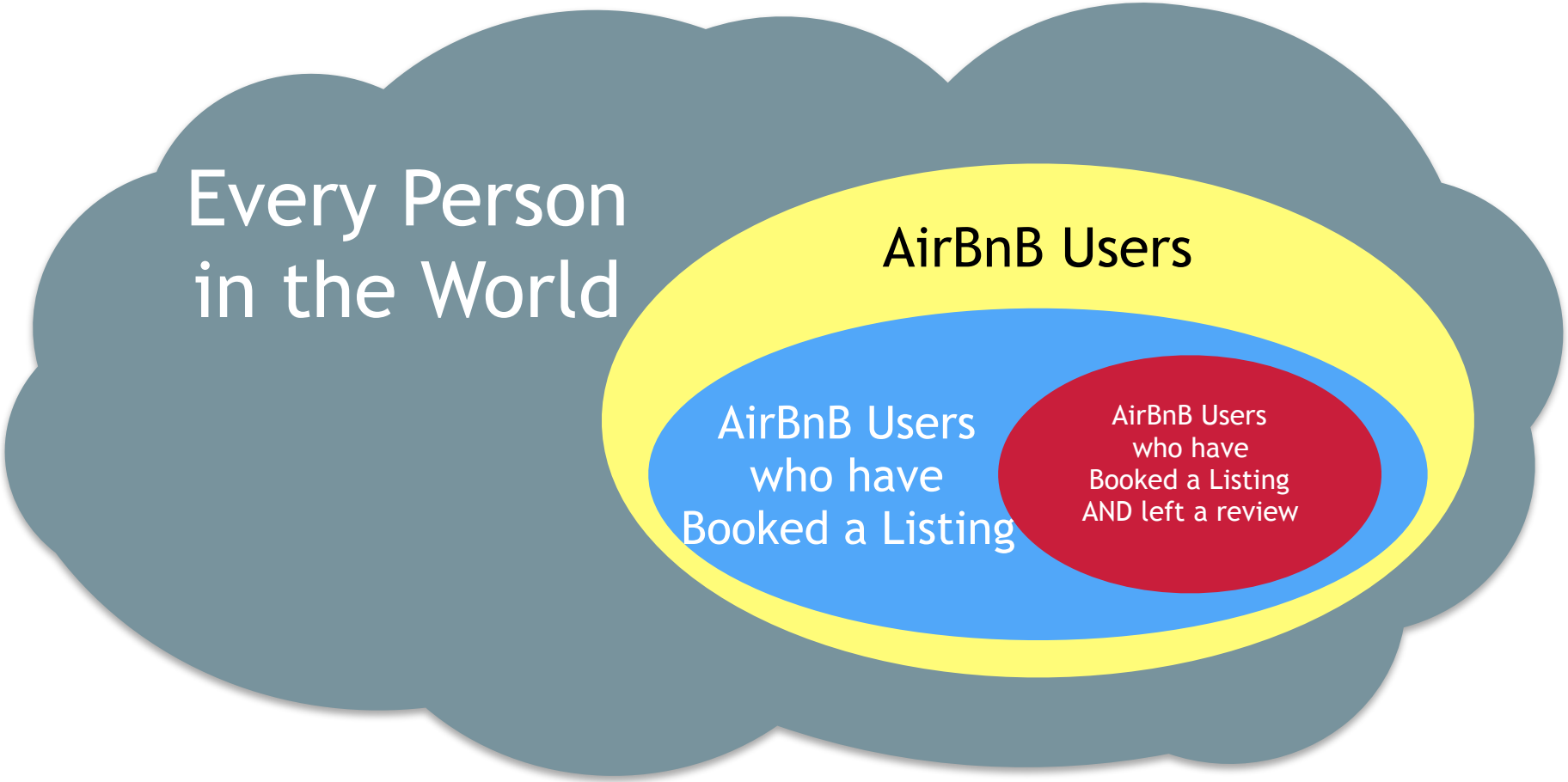
- Does recommending listings **make** users book more reservations?
- Do neighborhoods with more restaurants and shops **attract** more visitors?

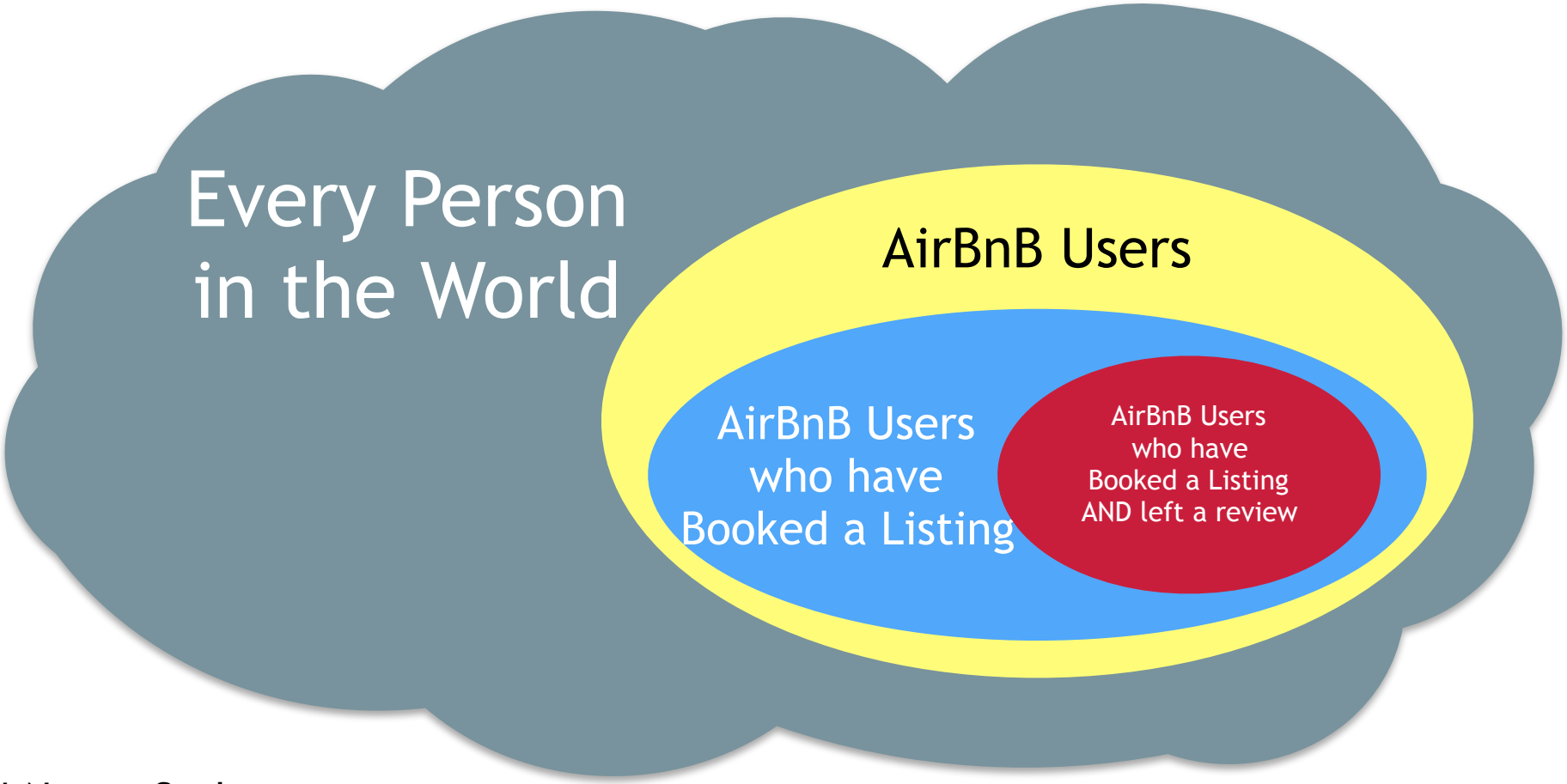
The Public

- Does AirBnB **increase** rents in cities?
- Has AirBnB **caused** hotel rates in NYC to fall?


Every Person
in the World
AirBnB Users


Every Person
in the World
AirBnB Users
AirBnB Users
who have
Booked a Listing


Every Person
in the World
AirBnB Users
AirBnB Users
who have
Booked a Listing
AirBnB Users
who have
Booked a Listing
AND left a review


Every Person
in the World
AirBnB Users
AirBnB Users
who have
Booked a Listing
AirBnB Users
who have
Booked a Listing
AND left a review

* Not to Scale

Uncertainty

The Components of Causality

Causal Model

- Encode assumptions/knowledge about problem structure

Causal Identification

- Is estimation possible, given unlimited observations

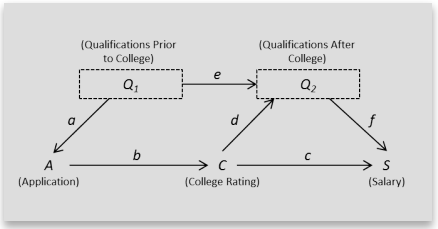
Causal Estimation

- Statically infer the value of causal estimand from observations

Causal Workflow

Domain Knowledge and Assumptions

Observed Data



Inference | Estimation

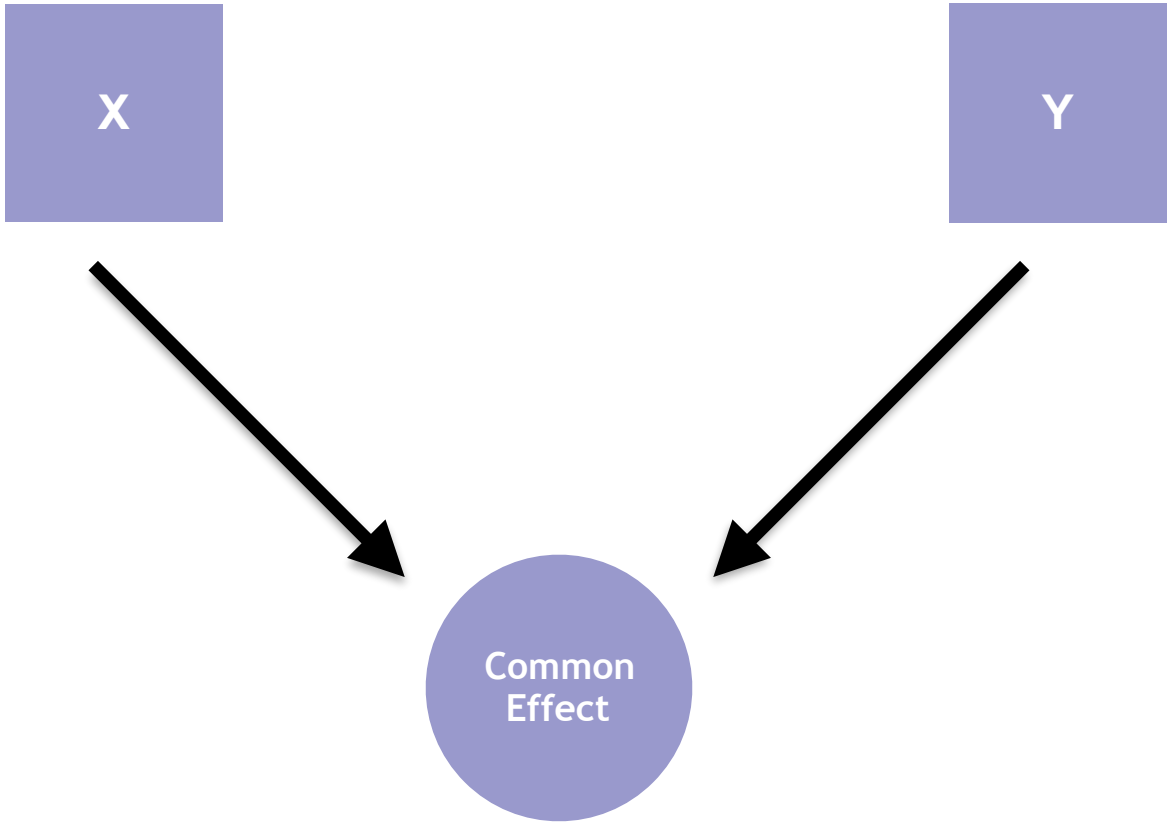
Criticize and Evaluate

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models.

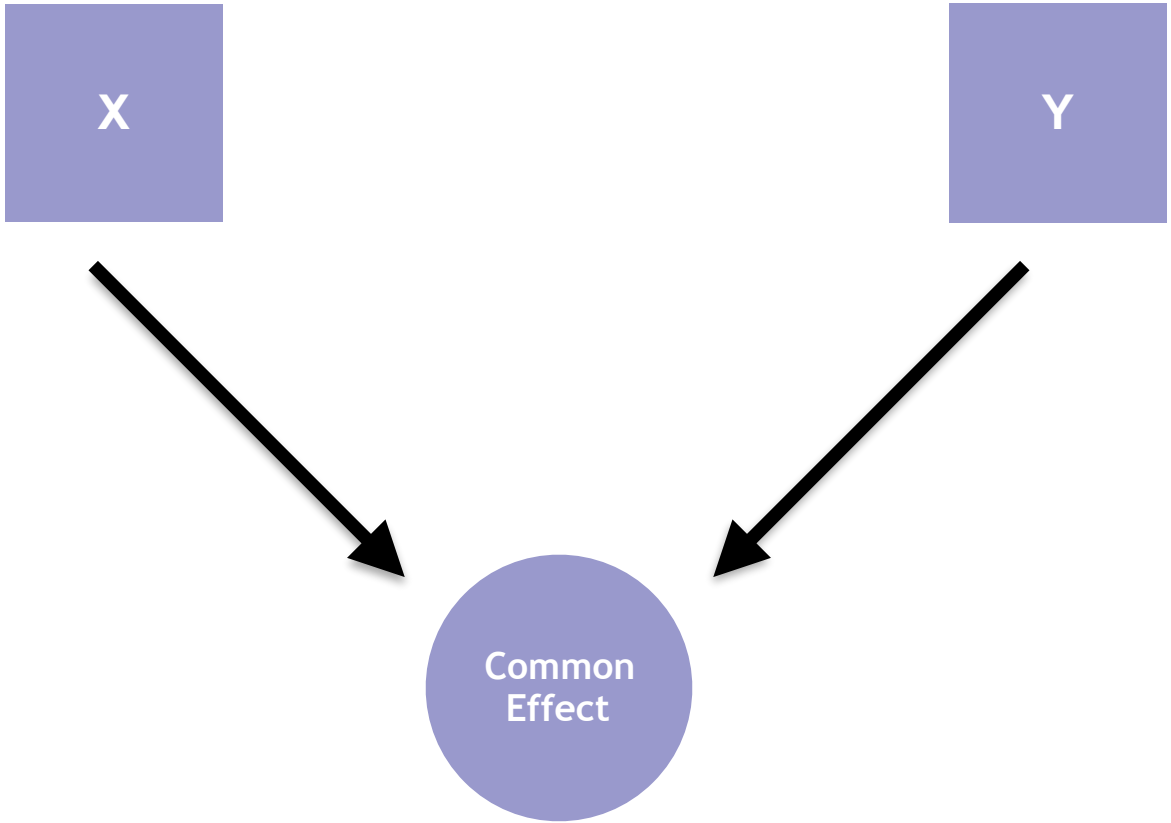
Live Code

Enter Causal Inference

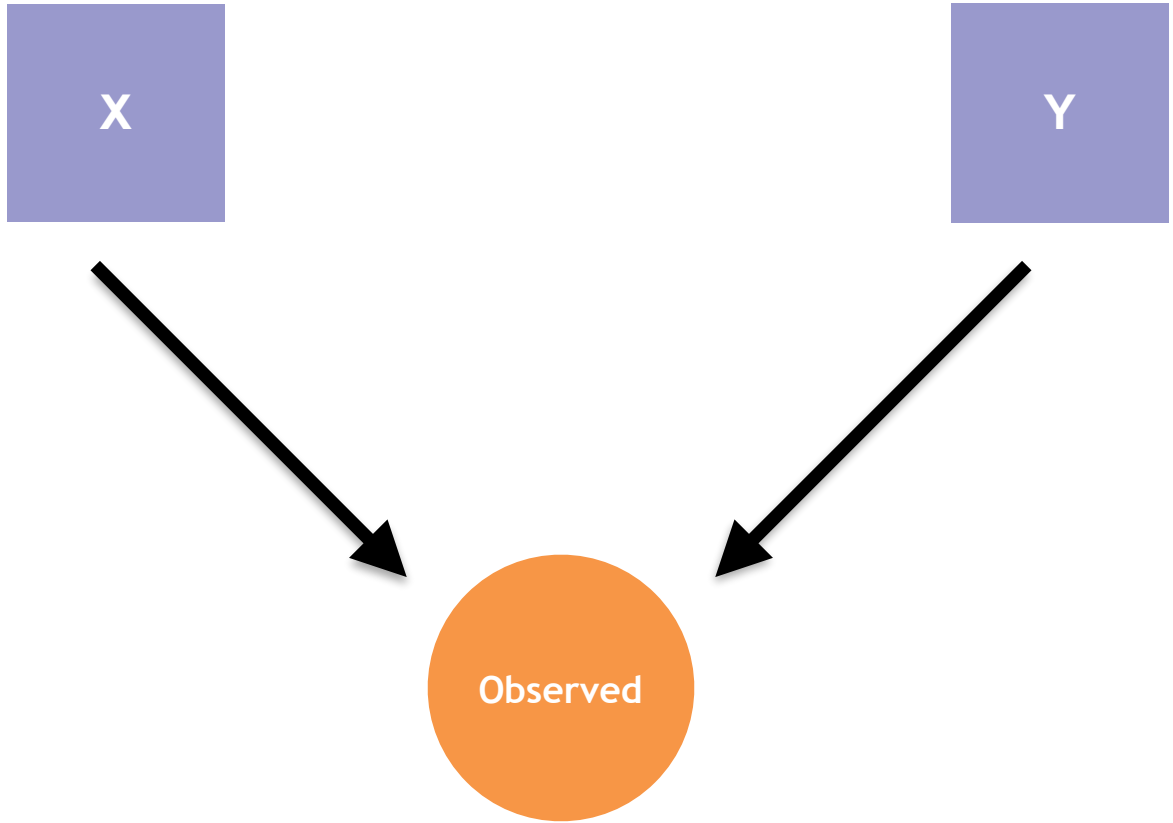
Graphical Causal Models



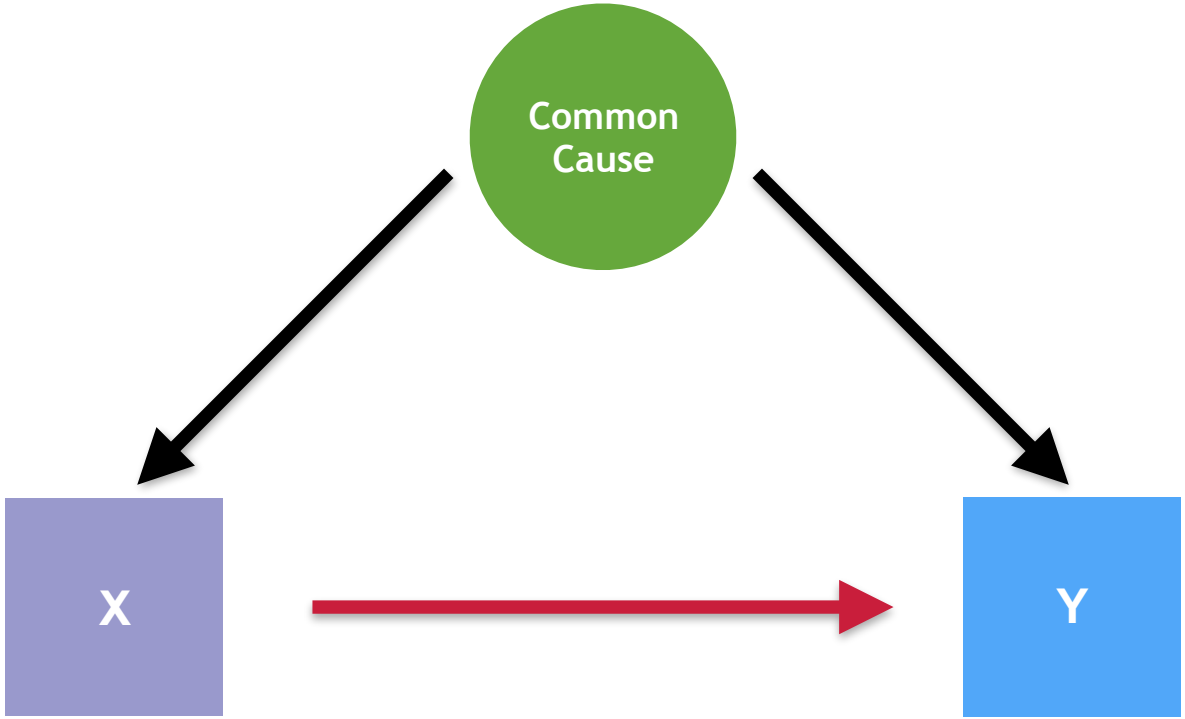
Independence



Dependence



Confounding



Identification

Back-door Criterion

- Control **covariates** that block indirect paths from X to Y

Front-door Criterion

- Find set of **mediator** variables in path from X to Y

Instrument Variables

- Use **instrument** which only affects Y by influencing X

Q&A (5 min)

Conclusion and Next Steps

Programming and Python

- Learn Enough Command Line to Be Dangerous
- Learn Enough Git to Be Dangerous
- Official Tutorial
- Think Python (online book)
- Composing Programs (Python and CS concepts)
- Google's Python Class (videos + exercises)
- Software Carpentry (videos + exercises + in-person)
- HGtP: Learning Python (compilation)

Resources in the Repository

Books

- Causality: Models, Reasoning, and Inference
- Causal Inference for Statistics, Social, and Biomedical Sciences
- Counterfactuals and Causal Inference Methods and Principles for Social Research
- Advanced Data Analysis from an Elementary Point of View
  - Chapter 18: Graphical Models
  - Chapter 19: Graphical Causal Models
  - Chapter 20: Identifying Causal Effects from Observations
  - Chapter 21: Estimating Causal Effects from Observations
  - Chapter 22: Discovering Causal Structure from Observations
- Graphical & Latent Variable Modeling

Courses

- Applied Causality (Columbia)
- Intermediate Statistics (CMU): Causal Inference
- Causal Inference and Learning (UIC)
- KDD Tutorial on Causal Inference and Counterfactual Reasoning

References

- Causal Inference Animated Plots
- Causal Data Science blog posts
- Causal inference in statistics: An overview
- The Seven Tools of Causal Inference, with Reflections on Machine Learning

Data Science Fundamentals LiveLessons

Part 2: Data Wrangling and Databases

Part 2: Machine Learning and Statistical Analysis

@jonathandinu

Thank You!

Materials: code and slides

Data: http://insideairbnb.com/get-the-data.html

@jonathandinu
jondinu@gmail.com
http://jonathanjonathanjonathan.com