

Course Reminders

- **A1** - due Friday (11:59 PM)
- **Week 2 quiz** - due Friday (11:59 PM)
- final project group by the end of the week; Project proposal (due *next* Friday - week 4)
- A2 now available on datahub (due Fri - week 5)
- Guest Lecture Dates:
 - Tu 2/4: Agustin Lebron
 - Th 2/13: Andy White

Data & Data Science Questions

Shannon E. Ellis, Ph.D
UC San Diego



Department of Cognitive Science
sellis@ucsd.edu

Data Structures Review

Structured data

- can be stored in database SQL
- tables with rows and columns
- requires a relational key
- 5-10% of all data

Semi-structured data

- doesn't reside in a relational database
- has organizational properties (easier to analyze)
- CSV, XML, JSON

Unstructured

- non-tabular data
- 80% of the world's data
- images, text, audio, videos

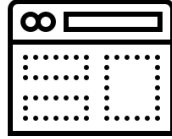
Unstructured Data

Some datasets record information about the state of the world, but in a more heterogeneous way. Perhaps it is a large text corpus with images and links like Wikipedia, or the complicated mix of notes and test results appearing in personal medical records.

Unstructured Data Types



Text files and
documents



Websites and
applications



Sensor
data



Image
files



Audio
files



Video
files



Email
data



Social media
data



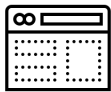
Positive:
70%

Negative:
20%

Neutral:
10%



Text:
Sentiment Analysis



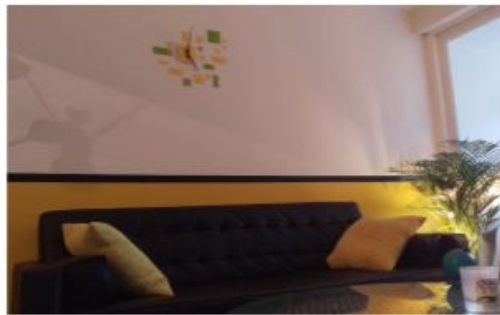
PYTHON

BEAUTIFULSOUP WEB SCRAPING





Bedroom Or Not?



“The left two photos were correctly predicted as bedrooms; The right two photos were correctly predicted NOT as bedrooms.”

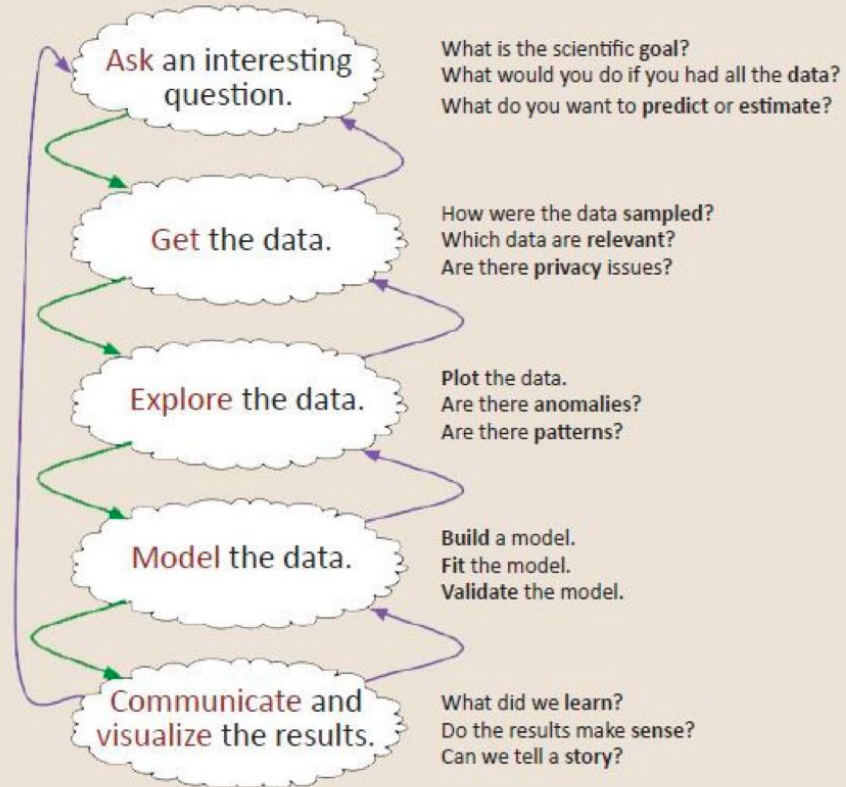
Formulating Data Science Questions

When you and your group sit down to figure out what you're going to do for your final project in this class, you'll have to formulate a strong question - one that is specific, can be answered with data, and makes clear what exactly is being measured.

Nature of a data scientist

- data-driven.
- care about answers. They analyze data to discover something about how the world works.
- care about whether the results make sense, because they care about what the answers mean.
- are comfortable with the idea that data have errors.
- know nothing is ever completely true or false in science, while everything is either true or false in computer science or mathematics.

The Data Science Process



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course <http://www.cs109.org/>.

If I had an hour to solve a problem and my life depended on it, I would use the first 55 minutes determining the proper question to ask, for once I know the proper question, I could solve the problem in less than five minutes. —Einstein

Data Science questions should...

- Be specific
- Be answerable with data
- Specify what's being measured



What makes a question a
good question?

Imagine you're passionate about public transportation...

What are some things you may care about? That a city may care about? That the people living in the city may care about?

- Efficiency (system, riders)
- Accessibility (citizens aware of how to use; actually in 'appropriate' locations)
- User-friendly; convenience
- Size; span (how far network is going)
- Rush hours (utilization) - working well at this time? Right number of buses/lines?

- Profit? (making money; not losing money)
- Cost for users (how much money will it cost people to get where you need to go; transfers)
- Accuracy? Time estimates correct?
- How quickly from point A to point B?
- Weather?
- Maintenance
- Neighborhood disruption
- Safety (riders, pedestrians, other cars, drivers)
- Compensation (drivers, workers)
- Equity & accessibility
- emissions/climate effects

A - have some ideas

B - totally stuck

C - confused about what's going on

Brainstorm: Public Trans. DS Questions

A - have some ideas

B - totally stuck

C - confused about what's going on



1. Cost per route vs. volume of customers served : Are there any routes that are not needed and/or would be better absorbed into/combined with another route? (efficiency; route)
2. (safety) What are the crime rates that affect the pedestrians or riders?
3. How does compensation affect the safety of the ride (based on how the driver performed)?
4. (Car reduction) - What areas would receive the largest reduction in cars if there were a line/bus/route there?
- 5.

Our equity/accessibility data science question:



A - have a specific, data-centric ?
B - stuck
C - confused about what's going on

What is a fair price for the transit system?

What is a fair price for different groups (adults vs. students)

Fair? How much does the transportation cost relative to the amount a person makes (income/salary)?

What affects price for the transit system? (putting system in place/maintenance/newer vs. older; how often route is used/rider density; location of transit system - labor; how much funding does the company get; distance traveled; cost to power buses/trains/etc.; what is it running on?; should/is government subsidize cost)

Given current costs, who is the system serving fairly? Who is being under/overserved?

What data would you imagine you'd have or could get?

- A** - have a specific, data-centric ?
- B** - stuck
- C** - confused about what's going on

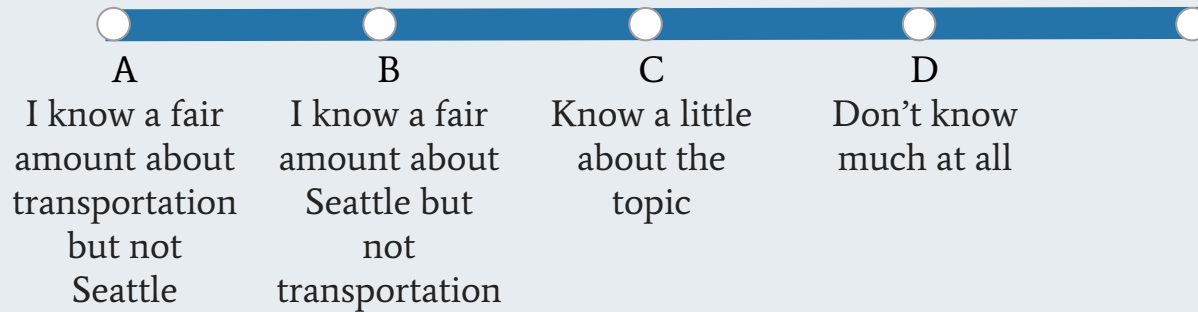
What is a fair price for the transit system? OR Given current costs, who is the system serving fairly? Who is being under/overserved?

- Average household income within proximity to each stop
- Historic fares of bus/routes in region of interest
- Number of complaints/improvements from customers
- Stop density vs population density
- Government spending on public transportation
- How much would it take to build a similar system? How much do those systems charge?
- Riders demographics (students; age; etc.)
- Average amount of rides per person (ridership; location) relative to income
- "Commutes" - common commutes; affecting fair relative to commute



Background Knowledge

How much do you know about transportation in Seattle?



Background: What do you know about public transportation (in Seattle)?

- They have ferries (shuttle cars and people)
- Has underground transportation
- What do we know about fairness/accessibility/equity already?
-

A - have some ideas

B - totally stuck

C - confused about what's going on



Imagine now you work for Seattle's DOT



ORCA = One Regional Card for All

Course Reminders

- A1 due Friday (11:59 PM)
- Project Proposals due next Friday (11:59 PM)
 - **By this Friday:** http://bit.ly/groups_wi20
 - One entry per group
 - Will create a GH repo for each group
 - Will give access to group members
 - Please type GH usernames into form correctly

What is a fair price for the transit system?

Given current costs, who is the system serving fairly?

Who is being under/overserved?

What ethical considerations should a project like this consider?

- Consent from people whose information is being used/data that they consent to be used (bias in who consents?)
- Privacy in individuals location information!!!!
- Private information contained in the data ; not include in the dataset (usage, storage, sharing/making public)
- When communicating, keep people anonymous
- Data aren't biased on their own
- Socioeconomic status - affordable for everyone; check that your analysis isn't biased (check analysis; unintended consequences)
- Deployment: what are you looking for/checking to make sure not benefitting some people/groups more than others?
- If you find an issue, plan to fix?

ORCA LIFT

Get where you need to go. Pay a lot less to get there.

Now there's a more affordable way to get to work, school, shopping, day care or anywhere else you need to go. It's ORCA LIFT, a new, reduced transit fare that can help you get more out of your public transportation system.

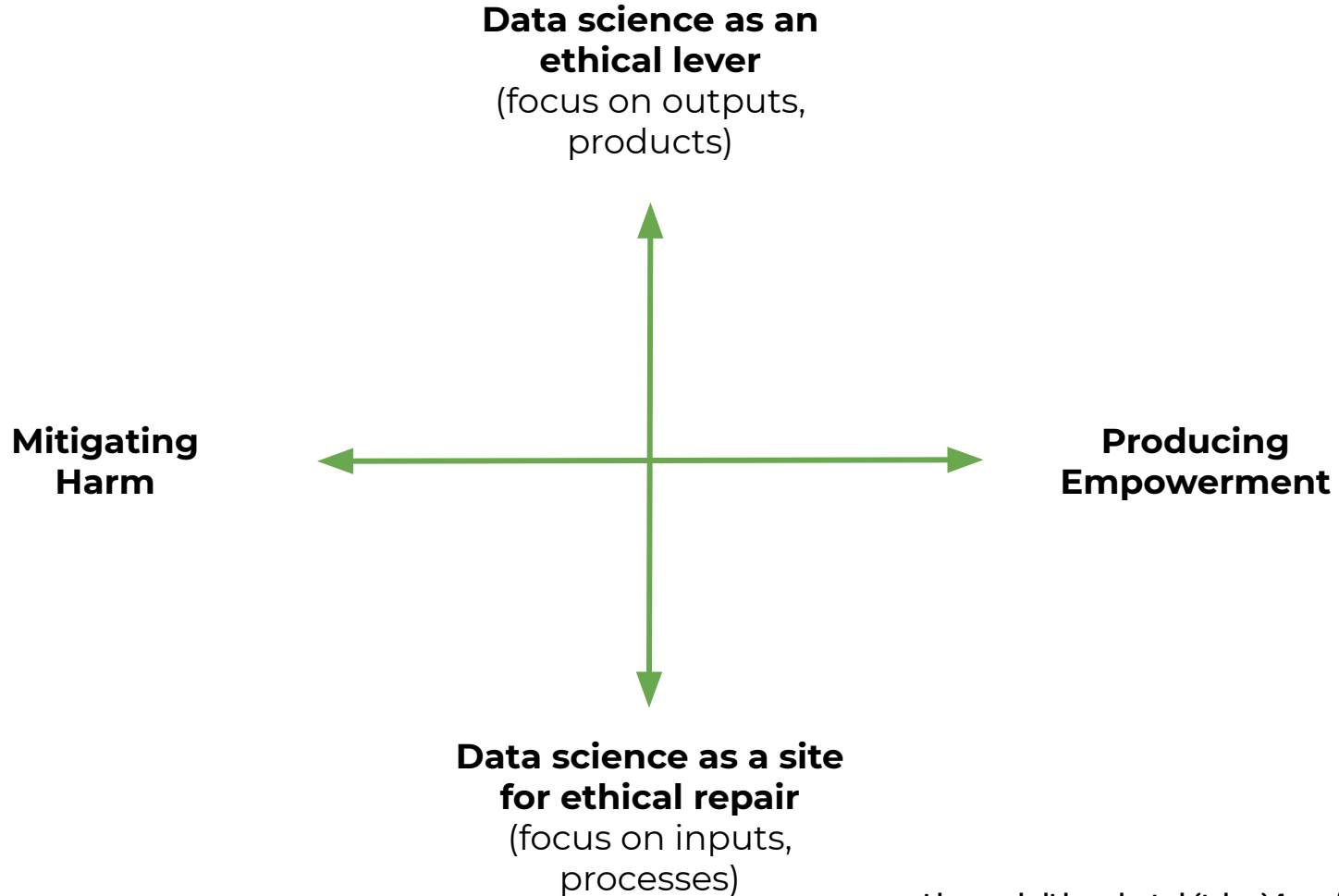
See if you qualify

Enrollment Locations

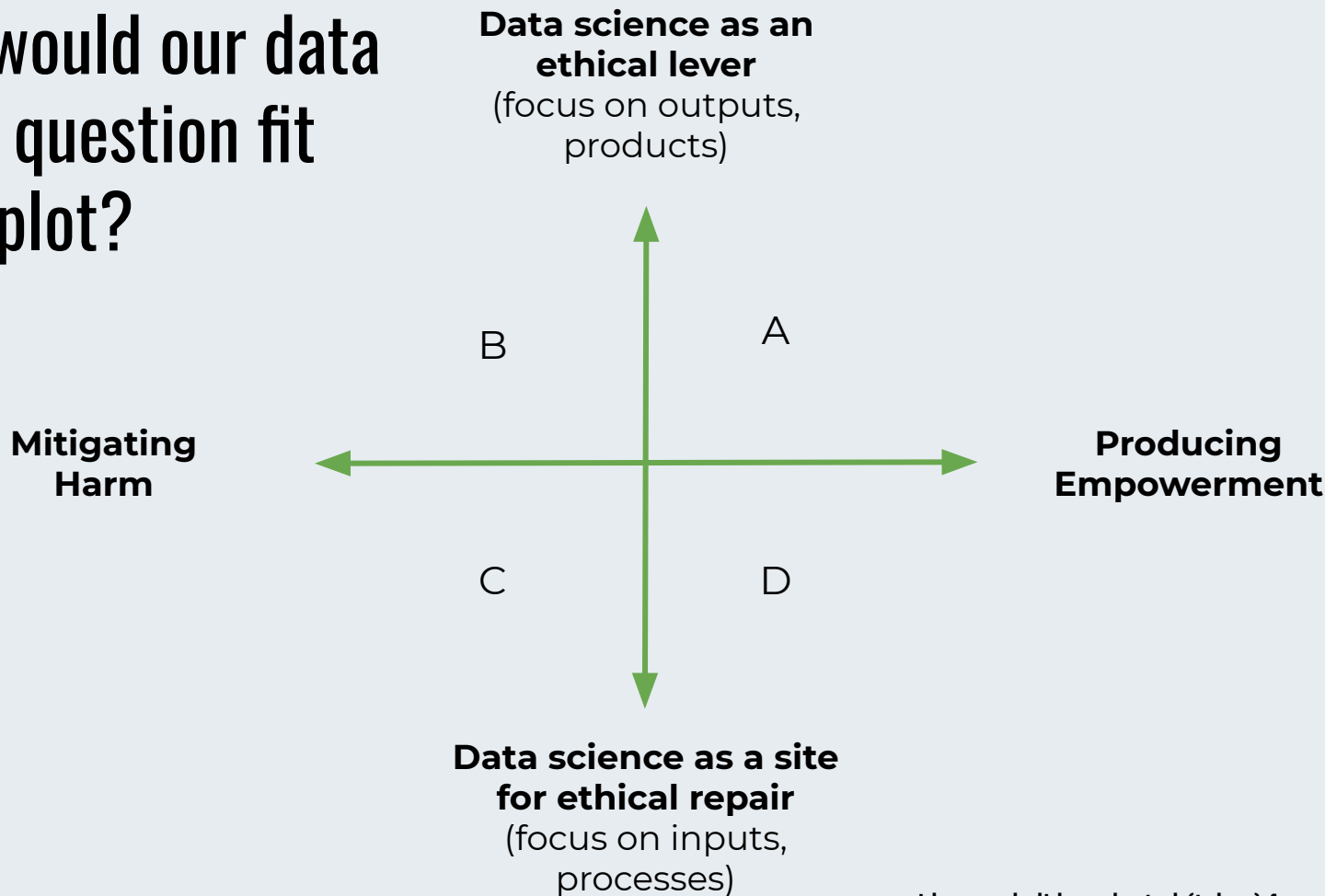


“ the evidence will more than likely show that there are differences between different types of ORCA users. From an analytical point of view, of course they're different. That's why they use different passes. For us, that wasn't particularly interesting. It's like, this is what we already knew and doing that would give us evidence to back it up. ”

- “Jamie,” DSSG fellow



Where would our data science question fit on this plot?



**Data science as an
ethical lever**

(focus on outputs,
products)

**Mitigating
Harm**

**Producing
Empowerment**

**Data Science
as Ethical
Convention**

An approach that seeks to develop rigorous methods, design principles, and professional norms that will make the practice of data science less harmful

**Data science as a site
for ethical repair**

(focus on inputs,
processes)

Bias in the system:

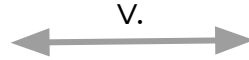
Data science as an ethical lever
to bring about change

Producing empowerment

Bias in the system:

Data science as an ethical lever
to bring about change

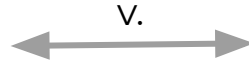
Producing empowerment



Bias in the data:

Data science as a site for ethical
repair

Mitigate harm



use the results of data science to identify and expose harms in the world

Data science as an ethical lever
(focus on outputs, products)

Data Science as Ethical Interrogation

Mitigating Harm

Data Science as Ethical Convention

Producing Empowerment

Data science as a site for ethical repair
(focus on inputs, processes)

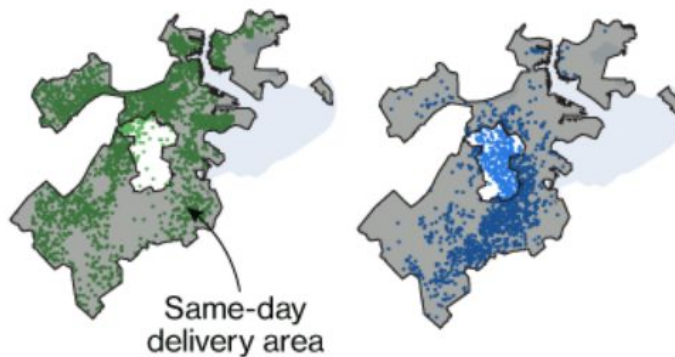


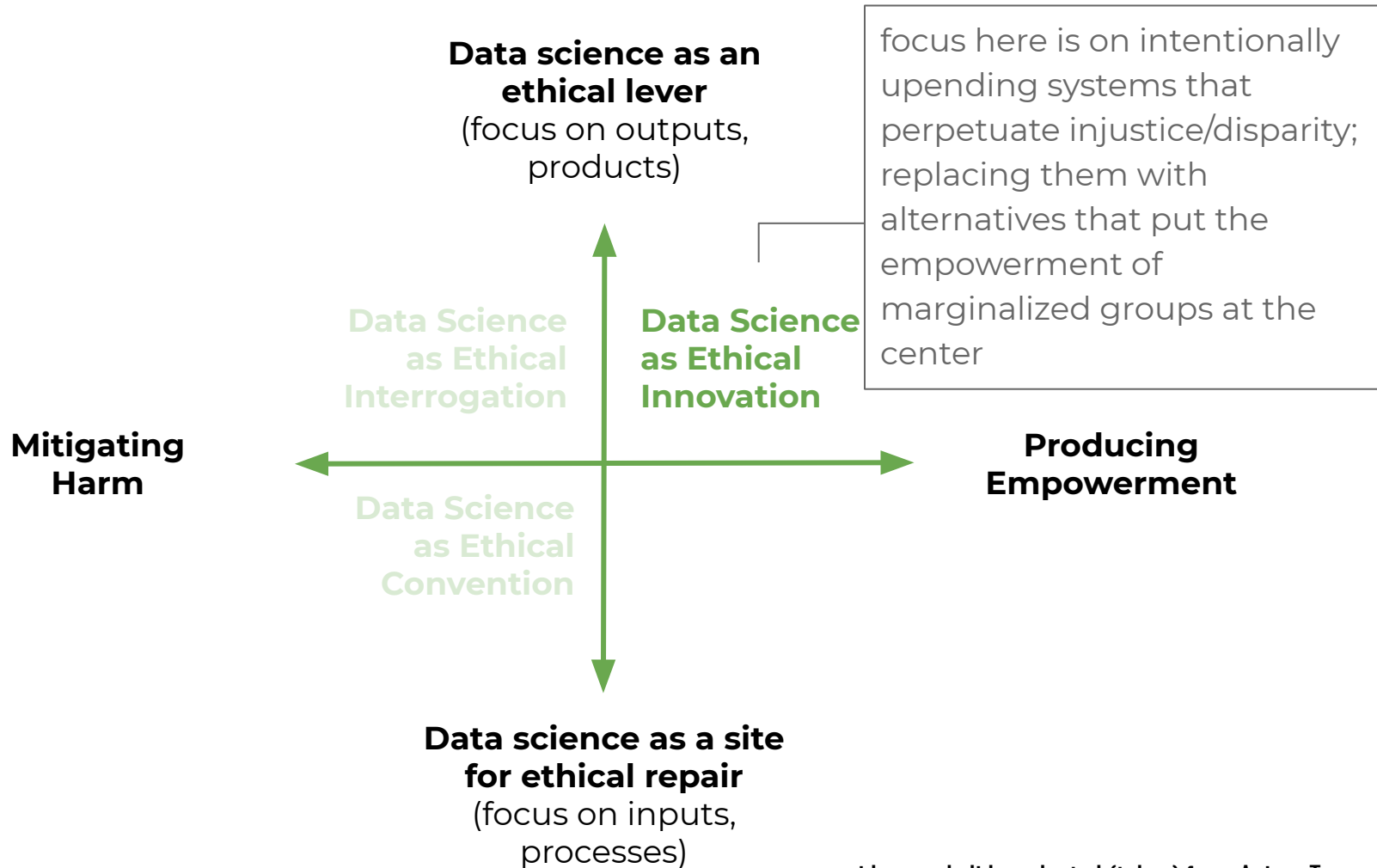
Amazon Doesn't Consider the Race of Its Customers. Should It?

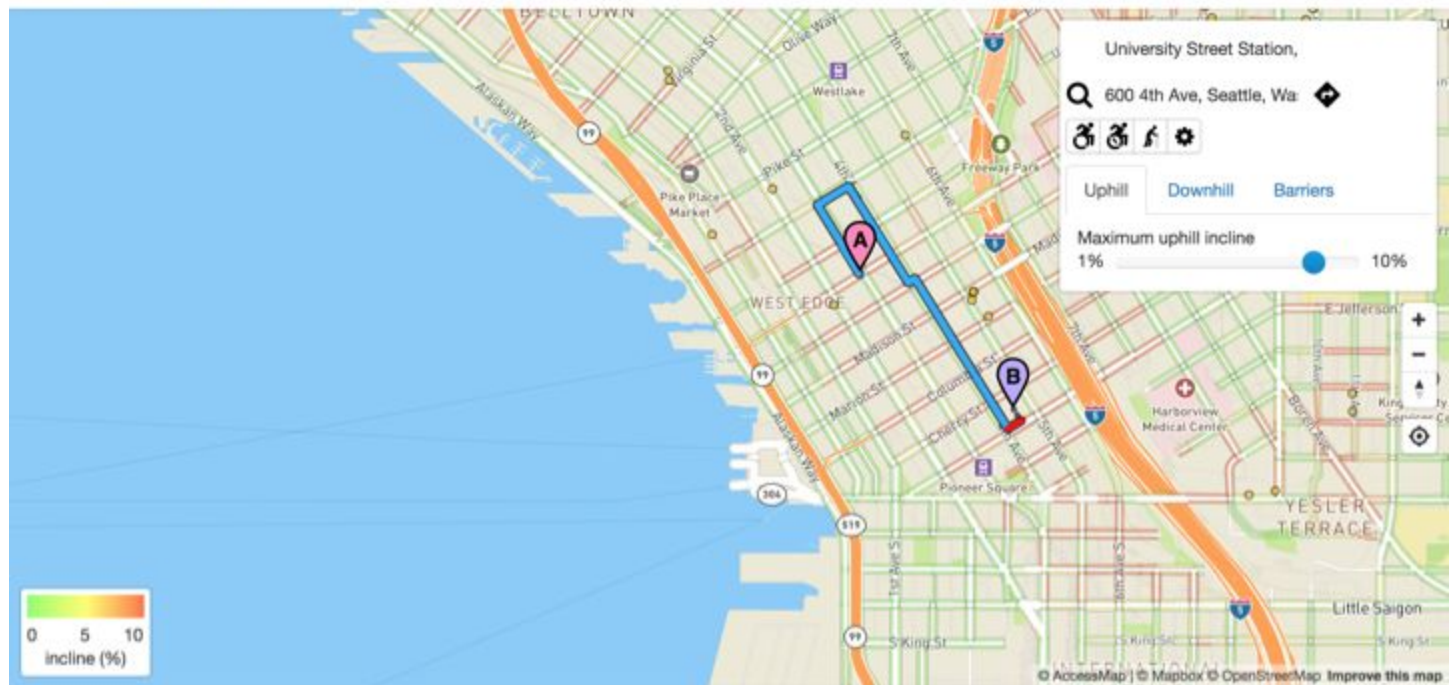
Three ZIP codes in the center of Boston, including the Roxbury neighborhood, are excluded from same-day coverage.

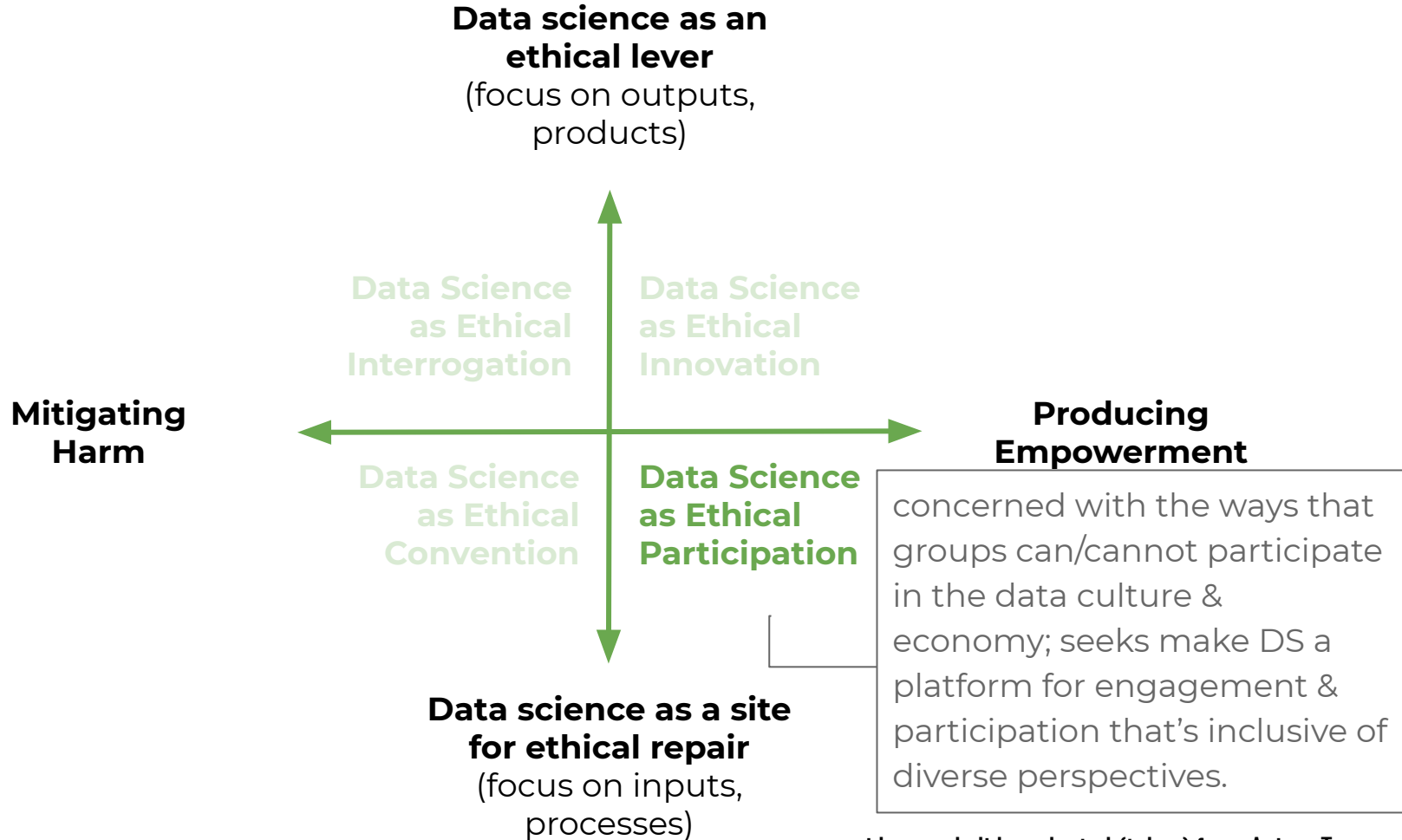
White residents

Black residents









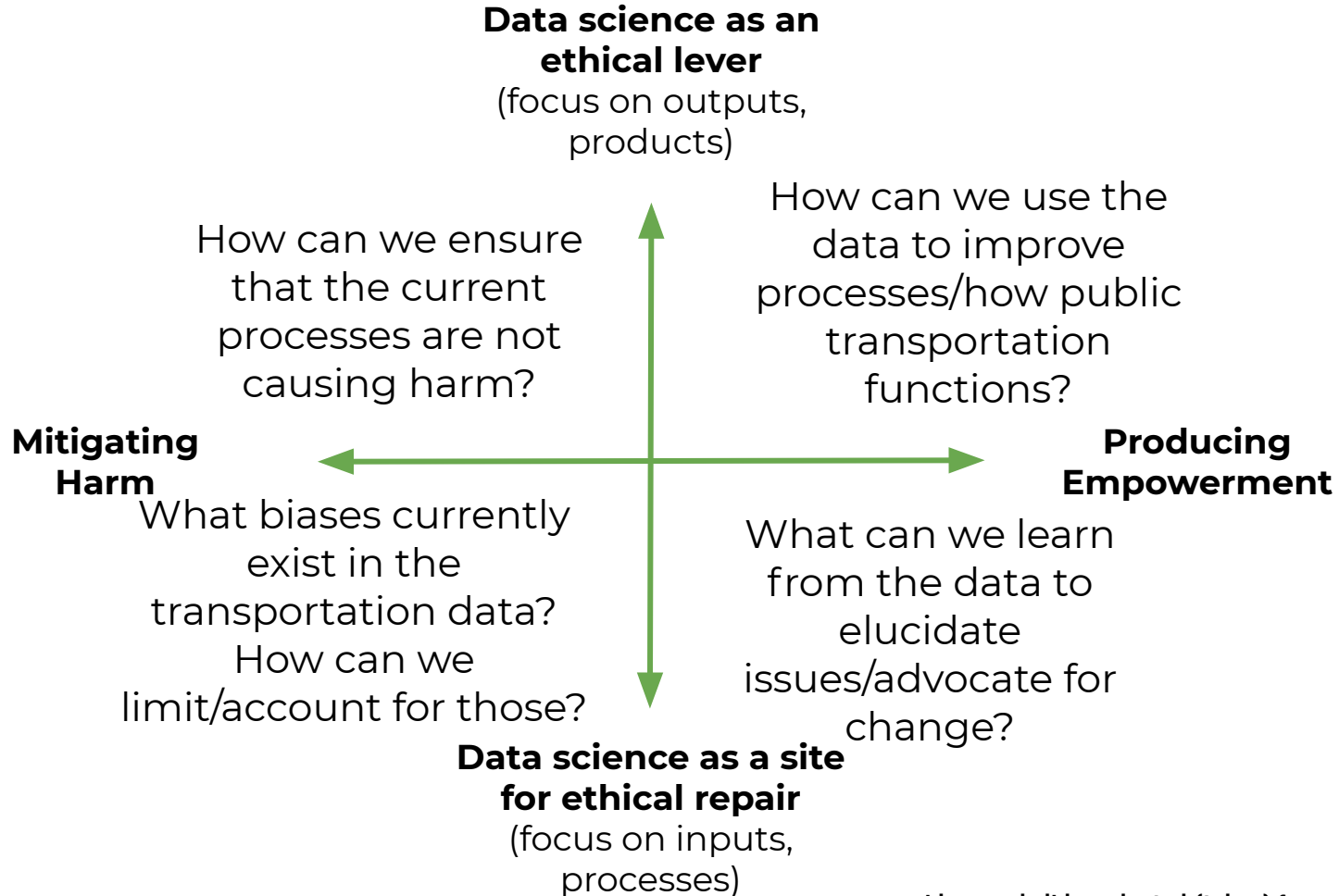
ideas and slides adapted (taken) from Anissa Tanweer

[Home](#)[About](#)[Conference '19](#)[Donate](#)[Action](#)[Press](#)

Data for Black Lives

Data as protest. Data as accountability. Data as collective action.

Sign up for more information.



Iteration: DS Questions (Unnecessary routes)

1. Are there unnecessary routes?
2. Are there unnecessary public transportation routes in San Diego?
3. Are there unnecessary public transportation routes and does that differ by type of route (bus, trolley, etc.) in San Diego?
4. Do certain public transportation routes cost more than other routes and does that differ by type of route (bus, trolley, etc.) in San Diego?
5. Unnecessary routes: Do certain public transportation routes in San Diego cost more than other routes and does that differ by type of route (bus, trolley, etc.) relative to the volume of customers served?
 - a. Does this differ by neighborhood?
 - b. Does the route cost less because it is less maintained or because it is more efficient?
 - c. Is the cost subsidized by the government or companies?
 - d. Would the less utilized routes be more utilized if in a different location?
 - i.whole line of thought to figure out how you would measure this then required...

Iteration: DS Questions (Safety)

1. What's the effect of crime on public transportation?
2. What's the effect of crime on public transportation in San Diego?
3. In San Diego, what are the crime rates that affect the pedestrians? riders?
4. In San Diego, does crime rate affect the number of pedestrians on the streets? The number of riders who use public transportation?
5. In San Diego, does the type of prominent crime and/or the crime rate affect the number of pedestrians on the streets? The number of riders who use public transportation?
6. In San Diego, does the type of prominent crime and/or the crime rate affect the *rate* of pedestrians on the streets? The *rate* of riders who use public transportation?
 - a. What changes would we recommend and to which areas to reduce crime, increase the number of people on the streets, and increase the number of people who use public transportation?

Iteration: DS Questions (Compensation)

1. Do the best drivers get paid the most?
2. Do the best public transportation drivers get paid the most?
3. Do the best public transportation drivers get paid the most in San Diego?
4. How are public transportation drivers compensated in San Diego?
5. Are public transportation drivers in San Diego compensated based on performance (as measured by timeliness, safety, number of complaints, and customer rating)?
 - a. If there is an effect, how much of this is driven by seniority (how long a driver has been in their position)
 - b. Does the routes an individual drives affect compensation?
 - c. Looking at public transportation in other cities, what would we suggest to incentivize safer driving?

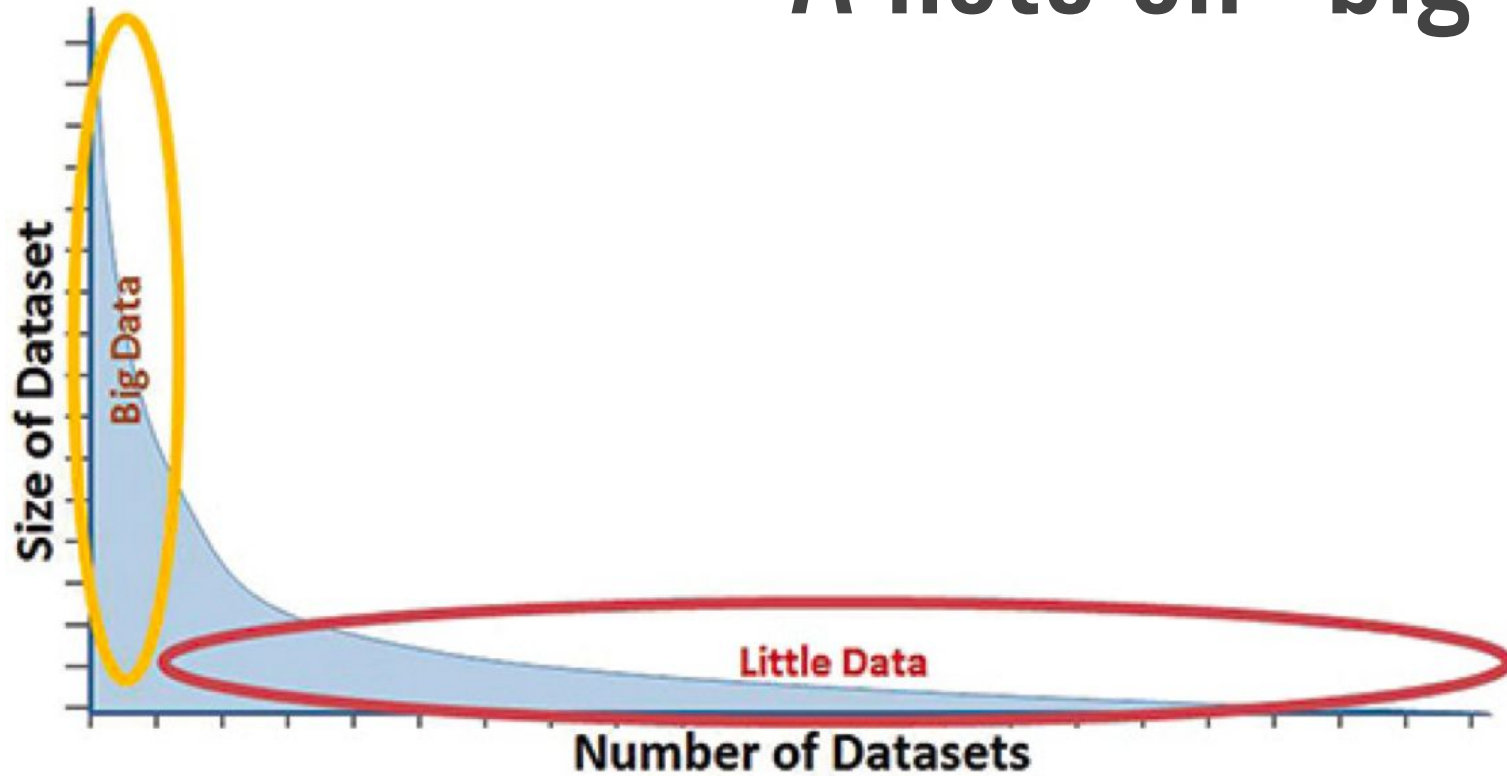
Iteration: DS Questions (Car Reduction)

1. What would reduce the number of cars on the road?
2. What would reduce the number of cars on the road in San Diego?
3. Parallel Approaches:
 - a. Traffic Portion:
 - i. Measuring the largest difference in estimate for travel during peak travel times relative to off-peak, which areas of the city have the worst traffic? What features affect these transportation patterns? What is the size of each feature's effect?
 - ii. Considering historical effectiveness of public transportation in San Diego, what areas would receive the largest reduction in cars if there were a line/bus/route there?
 - b. Car Reduction:
 - i. What has reduced the number of cars in other cities? Considering the estimate of cars taken off the road, how does each effective strategy compare to one another? (Most effective? Least effective?)
 - ii. In comparison to other cities and their similarity to San Diego, which strategy is likely to be most effective in San Diego? Which are to be avoided?

Finding Data

*Once you formulate a question, you'll need data. Data are *everywhere*. It's up to you to find the data that can best answer your question of interest. Often, this will involve multiple datasets. Typically, no one dataset will be perfect. And, often, even the combination datasets will not be perfect, but they will be good enough.*

A note on “big data”



Types of data: Big vs. Little

- There are difficulties in working with large data sets.
 - The analysis cycle time slows as data size grows (slow to iterate)
 - Large data sets are complex to visualize
- Simple models do not require massive data to fit or evaluate

Big Data Approach? Small Data Approach?

What are current voter preferences about the democratic presidential campaign pool?

Which approach is more accurate?

Take away: The right data set is the one most directly relevant to the tasks at hand, not necessarily the biggest one.

The best projects start with a question NOT the dataset.

The most boring projects are dataset-first.

Once you figure out the question, the links on GitHub may be helpful:

https://github.com/COGS108/Projects/blob/master/FinalProject_Guidelines.md#dataset-resource-list


**Where to look for and get
data for your projects?**

- Home
- What is Data?
- Frequently Used Statistics
- Frequently Used Data
- Find Data by Topic ▾
- Data APIs
- Text Data ▾
- Statistical Analysis Software ▾
- Data Visualization ▾

Library Data Services

- Data Services
- Data & GIS Lab
- GIS @ UCSD
- Finding Data & Statistics
- Research Data Curation

Featured UC San Diego Collections

- UC San Diego Dataverse 

Miscellaneous datasets purchased for the UC San Diego community. Includes:

 - Data on Terrorist Suspects (DOTS)
 - Field (California) Poll, 1956 - [Latest Release]
 - International Country Risk Guide: Table 3B: Political Risk Points by Component
 - International Terrorism: Attributes of Terrorist Events (ITERATE), 1968 - [most recent]
 - Latin American Public Opinion Project (LAPOP) 1978-2003

tics/artculture

Finding Data & Statistics

Welcome to the UC San Diego Library's guide

Data repositories and datasets linked here are our **specific guides**, many of which include sections are also available. If you need help finding particular out to a librarian for assistance.

Data Spotlight




A one-stop-shop interface for accessing statistics collected by **United Nations agencies**. Search across 32 databases (60 million records!) by topic, country, or region.

Not finding what you need? Check the website of the specific UN agency for additional statistics.

Off-Campus Access & Wireless

Many of the resources listed on this guide have **Off-campus access**, as well as **Diego Library**.

- Art and Culture
- Country Statistics & Data
- Crime
- Data Science
- Economic & Financial Data
- Economics: Datastream software
- Education
- Environment & Energy
- Food & Beverage
- Government Spending & Infrastructure
- Health/Health Care & Mortality
- Labor, Employment, Wages
- Latin American Public Opinion Project (LAPOP)
- Latinobarómetro
- Marketing
- Migration & Immigration
- People: Census guide 
- People: Children, Families, Aging
- People: Demographics & Population (general)
- People: Gender Studies & Women
- People: Race & Ethnicity
- People: Religion
- Political Science
- Political Science: Worldwide Elections Guide
- Public Opinion, Social Attitudes and Values
- San Diego & California
- Social Media

Consultations

Winter Quarter 2019

on consultation hours

esdays, 11am-12pm

Data & GIS Lab

who can provide

anie Labou

science Librarian

Barsh

an for Economics and

ss

se Sklar

an for Political Science,

Society, and International

ment Information

. Smith

an for US Government

ation, Urban Studies &

ng, Environmental Policy,

San Diego Government

When the data aren't ready and waiting for you...

- APIs
- Web Scraping
- Collecting your own data

You'll likely need more than one dataset to complete your final project.

You can see previous COGS 108 Projects

<https://github.com/COGS108/FinalProjects-Sp17>

<https://github.com/COGS108/FinalProjects-Wi18>

<https://github.com/COGS108/FinalProjects-Sp19>