CSSA PRESENTS

Register on
EventBrite now!
http://bit.ly/NCSC_19

UCSD COGNITIVE SCIENCE
STUDENT ASSOCIATION

COGNITIVE CROSSROADS

NATIONAL COGNITIVE SCIENCE CONFERENCE 2019

APRIL 14, 2019 |10AM-4:30PM | PC THEATER

All are welcome. Participants do *not* have to be UC San Diego students.

Course Reminders

- Survey due tonight 4/5 (11:59 PM) : http://bit.ly/cogs108_survey
- A1 - due *next* Sunday 4/14 (11:59 PM)

Data & Data Science Questions

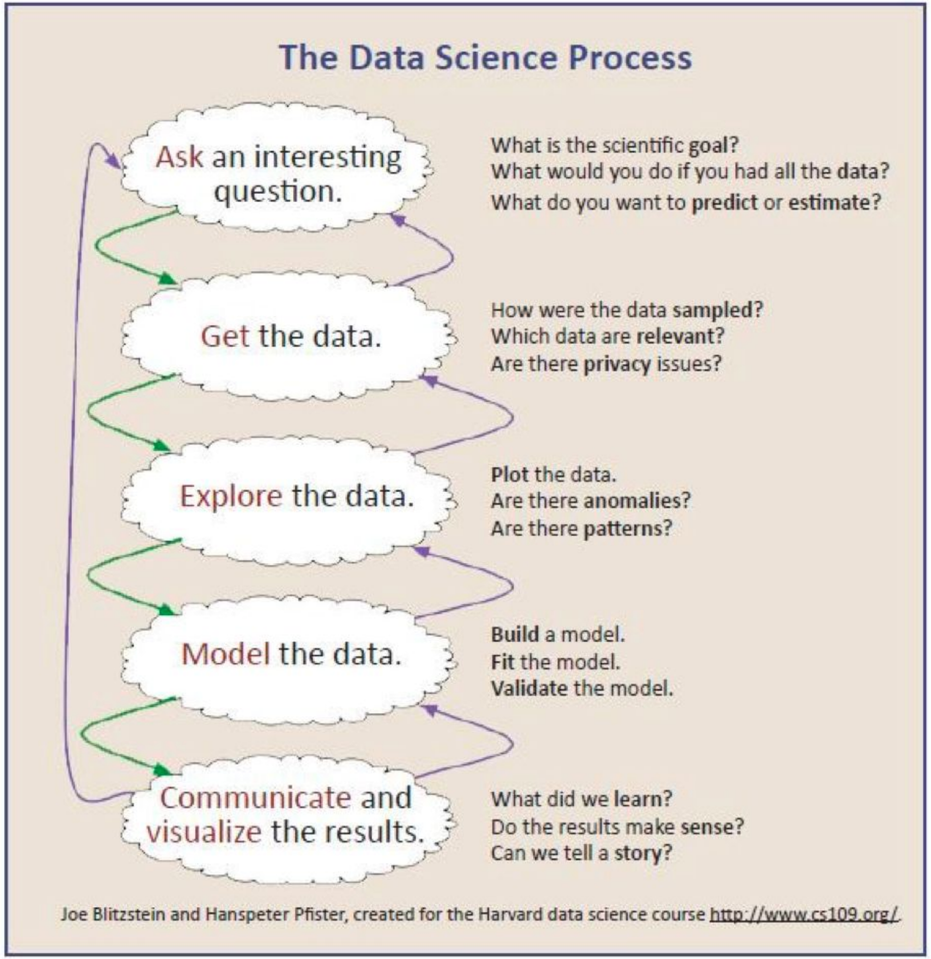
Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

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.

adapted from Chris Keown



The Data Science Process

Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

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

adapted from Chris Keown

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

adapted from Chris Keown

Data Science questions should...

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

What makes a question a good question?

Specifying what you’re going to measure is important

Examples of poor questions that leave wiggle room for useless answers:

- What can my data tell me about my business?
- What should I do?
- How can I increase my profits?

Examples of good questions where the answer is impossible to avoid:

- How many Model 3s will Tesla sell in San Diego during the third quarter?
- How many students will apply for admission to UCSD in 2019?
- How many students should UCSD admit in 2019 for a target class size of 5000?

adapted from Chris Keown

Working toward a strong data science question

Nailing down the right question: politics

Too-vague question: What impacts politics in America?

https://github.com/COGS108/FinalProjects-Wi18/blob/master/071-FinalProject.ipynb

Nailing down the right question: cause of death

Too-vague question: What gets attention in the news?

https://github.com/COGS108/FinalProjects-Wi18/blob/master/010-FinalProject.ipynb

Nailing down the right question: policing

Too-vague question: Why isn’t police response time always the same?

https://github.com/COGS108/FinalProjects-Sp17/blob/master/029-FinalProject.ipynb

Nailing down the right question: policing

Too-vague question: Why isn’t police response time always the same?

Improving: How can we improve police response time?

… Do crime levels and time of day affect response time?

… Where should police cars be stationed, accounting for crime levels and time of day, to make police response times equitable?

… Where should police cars be stationed, accounting for crime levels and time of day, to make police response times equitable throughout San Diego?

https://github.com/COGS108/FinalProjects-Sp17/blob/master/029-FinalProject.ipynb

Data Science Question

You're interested in learning more about age in US politics

Which of the following is the BEST data science question?

○ **A** How old are Congress members?

○ **B** How many people are in Congress currently?

○ **C** What is best about US politics? What is worst?

○ **D** What should I learn about US politics age and where should I learn that information?

○ **E** How has the average age of members in Congress changed over time?

How may data you’ll use in this course be structured?

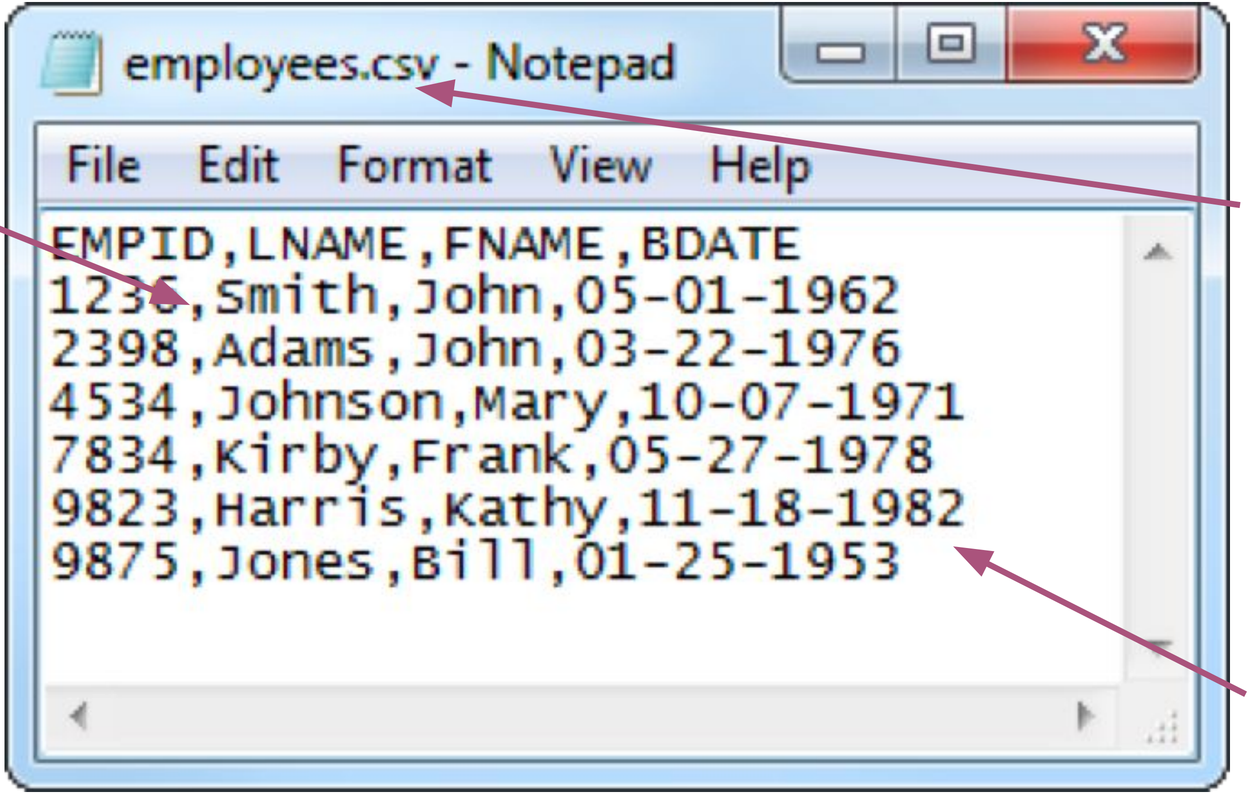
Types of data we’ll work with:

- Structured & semi-structured*
  - Spreadsheets (CSVs, .xlsx)*
  - JSON & XML*
  - relational databases (SQL)
- Unstructured
  - everything else: video, audio, images, websites, apps, text, etc.

*Data Scientists work with semi-structured data most frequently, but have to familiar with and comfortable in all types

CSVs

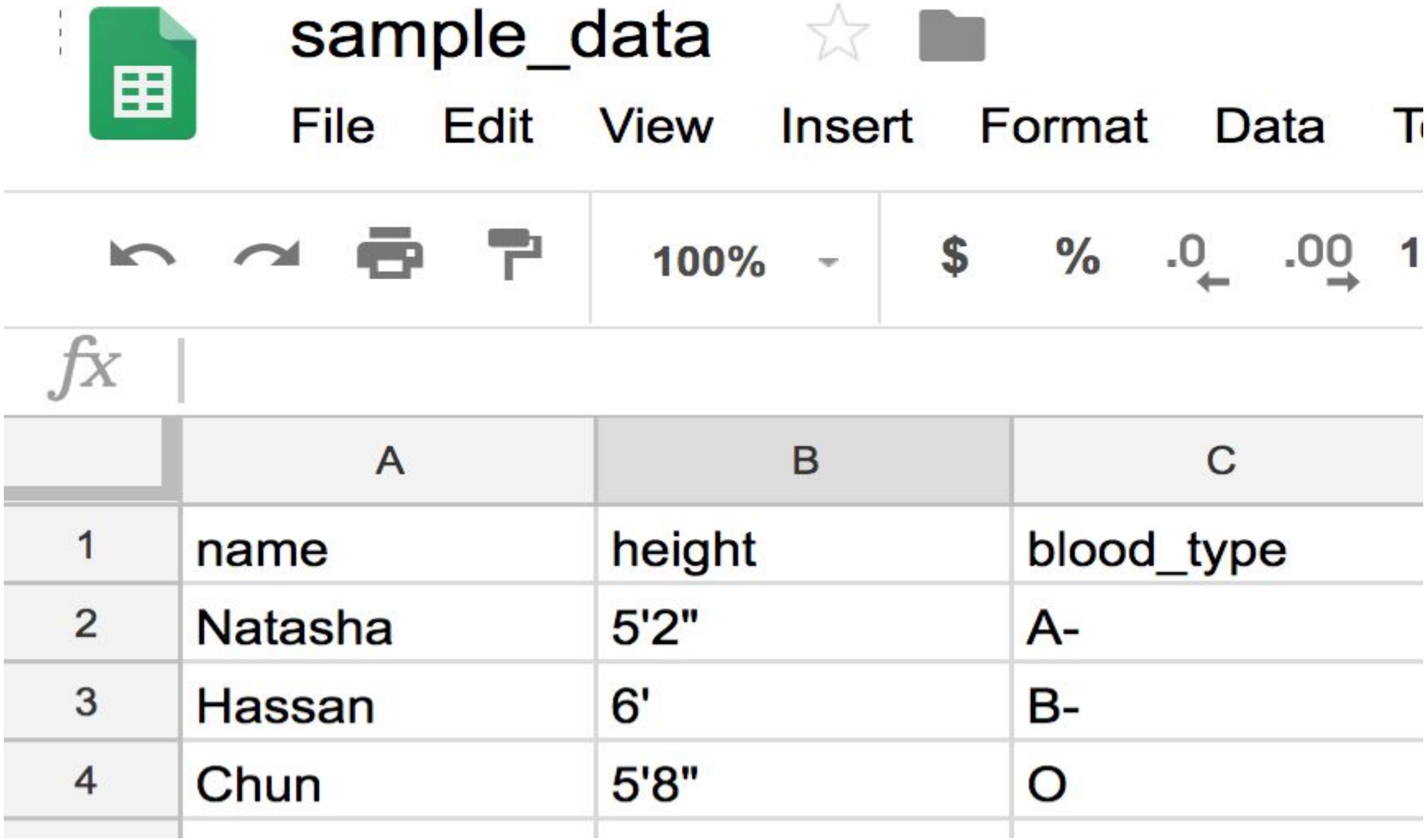
Each column separated by a comma



```
EMPID,LNAME,FNAME,BDATE
1236,Smith,John,05-01-1962
2398,Adams,John,03-22-1976
4534,Johnson,Mary,10-07-1971
7834,Kirby,Frank,05-27-1978
9823,Harris,Kathy,11-18-1982
9875,Jones,Bill,01-25-1953
```

Has the extension ".csv"

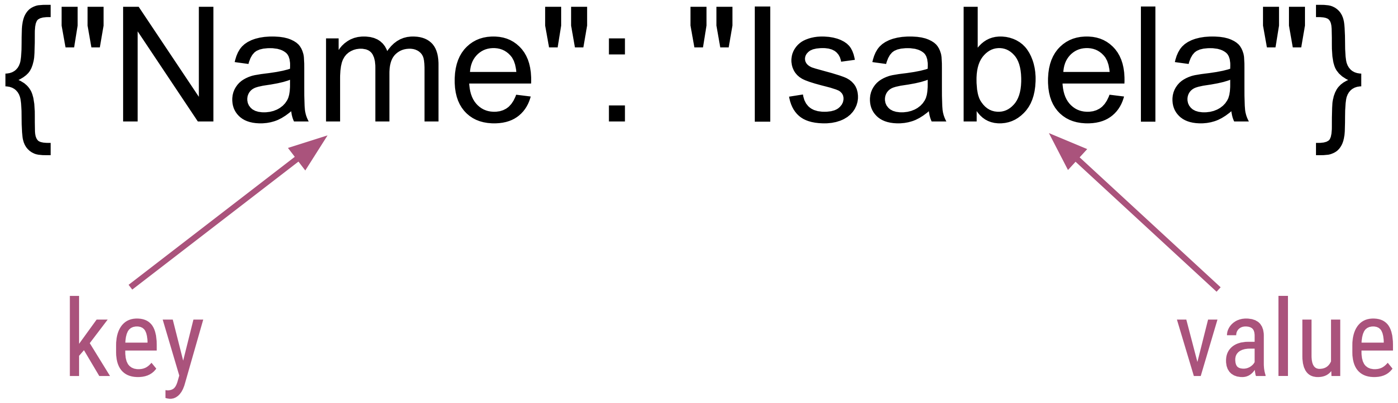
Each row is separated by a new line



	A	B	C
1	name	height	blood_type
2	Natasha	5'2"	A-
3	Hassan	6'	B-
4	Chun	5'8"	O

JSON: key-value pairs

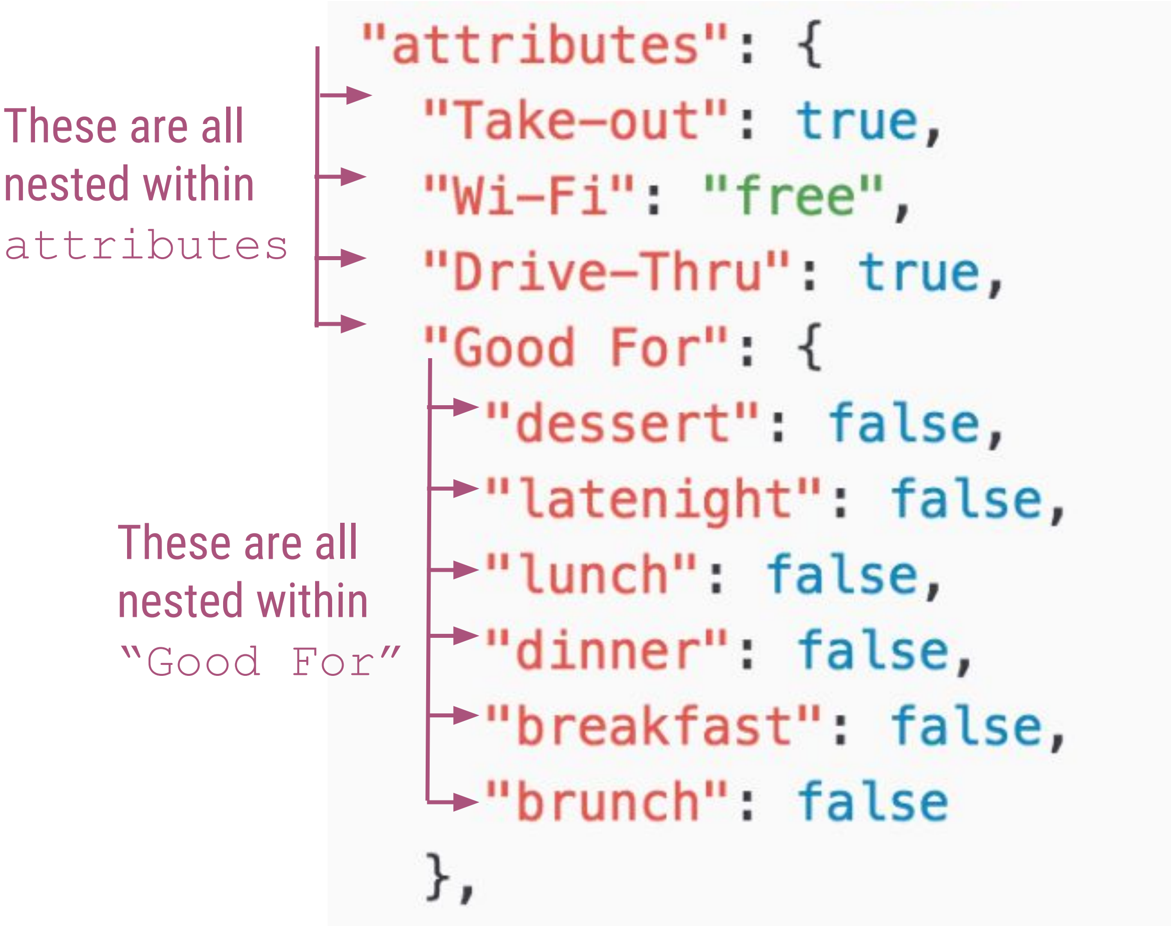
nested/hierarchical data



```
{"Name": "Isabela"}
```

key

value



These are all nested within `attributes`

These are all nested within “Good For”

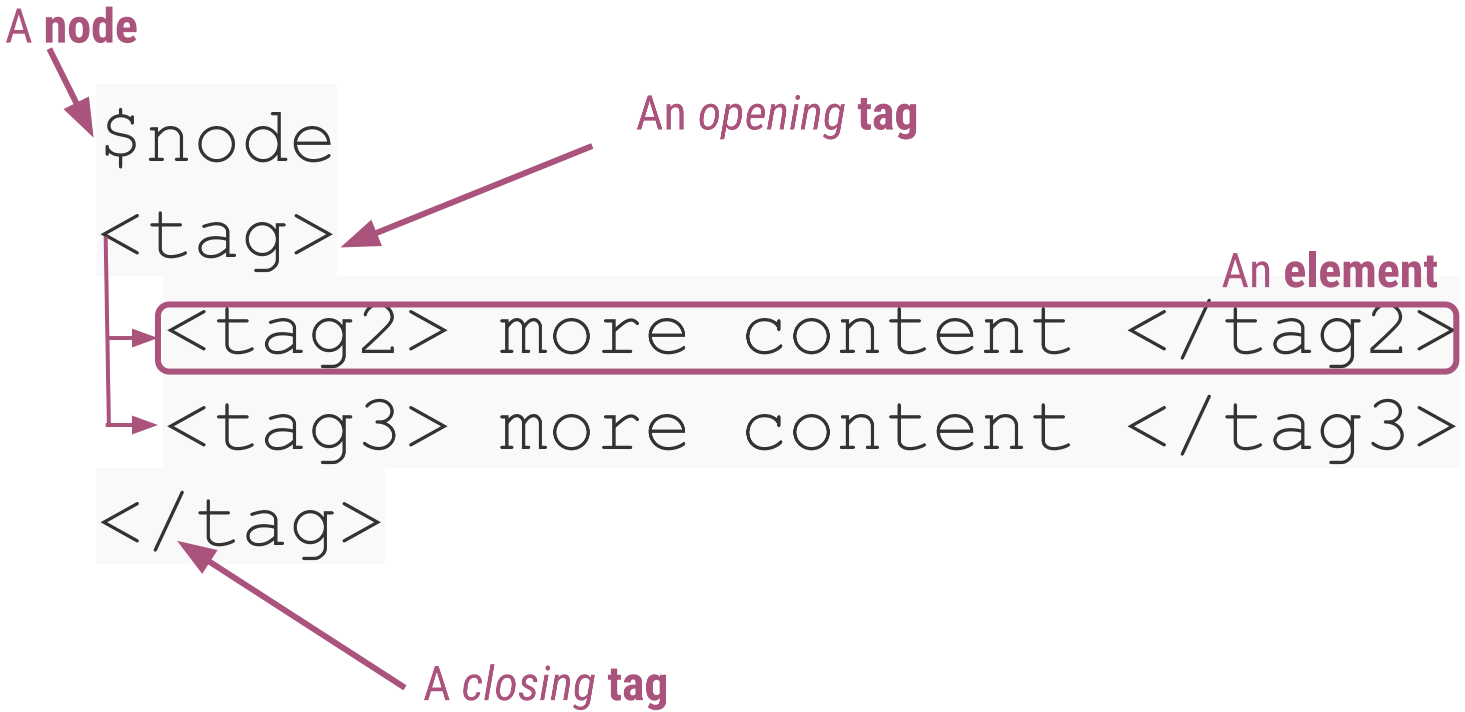
```
"attributes": {
  "Take-out": true,
  "Wi-Fi": "free",
  "Drive-Thru": true,
  "Good For": {
    "dessert": false,
    "latenight": false,
    "lunch": false,
    "dinner": false,
    "breakfast": false,
    "brunch": false
  },
```

JSON

https://blog.exploratory.io/working-with-json-data-in-very-simple-way-ad7ebcc0bb89

Extensible Markup Language (XML): nodes, tags, and elements

nested/hierarchical data



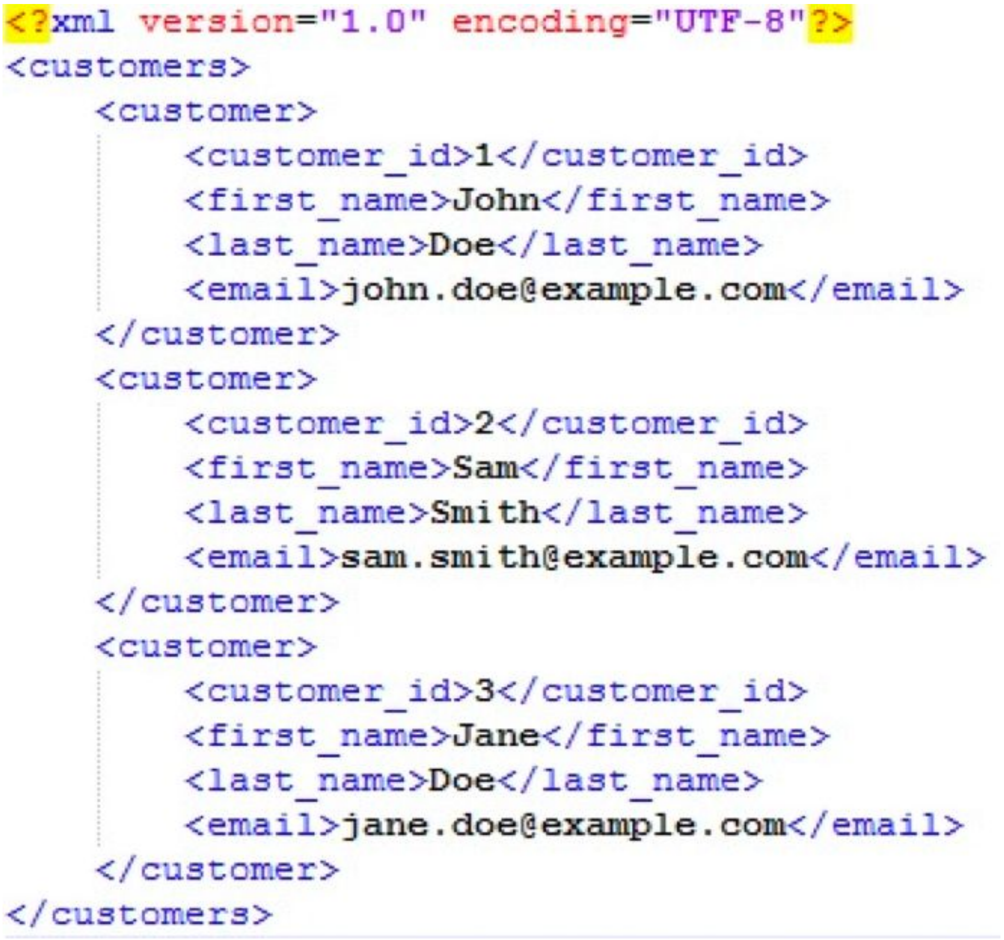
A **node**

An *opening* **tag**

An **element**

A *closing* **tag**

```
$node
<tag>
  <tag2> more content </tag2>
  <tag3> more content </tag3>
</tag>
```



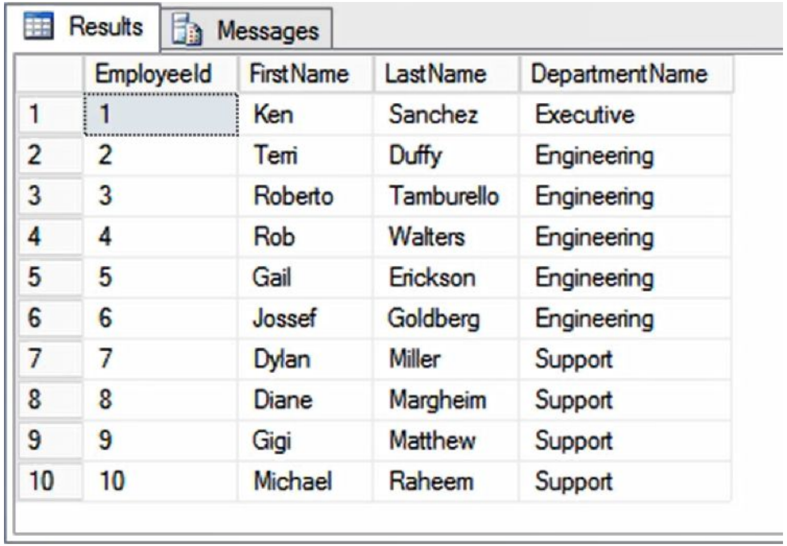
```xml
<?xml version="1.0" encoding="UTF-8"?>
<customers>
    <customer>
        <customer_id>1</customer_id>
        <first_name>John</first_name>
        <last_name>Doe</last_name>
        <email>john.doe@example.com</email>
    </customer>
    <customer>
        <customer_id>2</customer_id>
        <first_name>Sam</first_name>
        <last_name>Smith</last_name>
        <email>sam.smith@example.com</email>
    </customer>
    <customer>
        <customer_id>3</customer_id>
        <first_name>Jane</first_name>
        <last_name>Doe</last_name>
        <email>jane.doe@example.com</email>
    </customer>
</customers>
```

XML

adapted from Chris Keown

Relational Databases: A set of interdependent tables

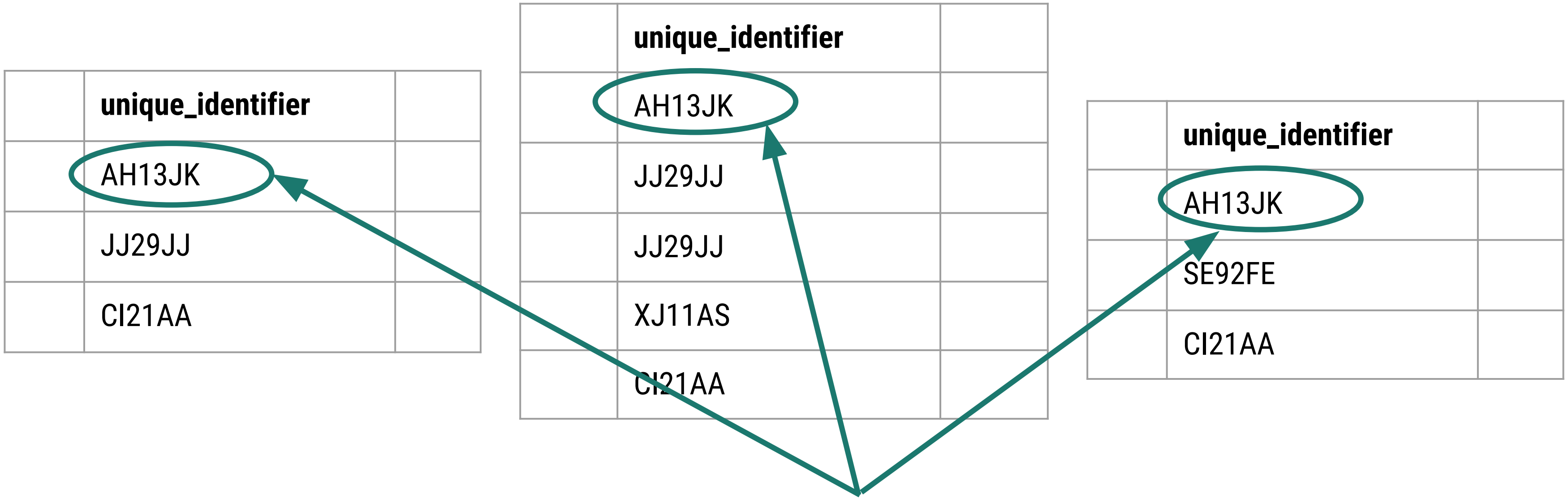
1. Efficient Data Storage
2. Avoid Ambiguity
3. Increase Data Privacy



	EmployeeId	FirstName	LastName	DepartmentName
1	1	Ken	Sanchez	Executive
2	2	Terri	Duffy	Engineering
3	3	Roberto	Tamburello	Engineering
4	4	Rob	Walters	Engineering
5	5	Gail	Erickson	Engineering
6	6	Jossef	Goldberg	Engineering
7	7	Dylan	Miller	Support
8	8	Diane	Margheim	Support
9	9	Gigi	Matthew	Support
10	10	Michael	Raheem	Support

relational database

Information is stored across tables



	unique_identifier	
	AH13JK	
	JJ29JJ	
	CI21AA	

	unique_identifier	
	AH13JK	
	JJ29JJ	
	JJ29JJ	
	XJ11AS	
	CI21AA	

	unique_identifier	
	AH13JK	
	SE92FE	
	CI21AA	

entries are *related* to one another by their unique identifier

relational database

restaurant

name	id	address	type
Taco Stand	AH13JK	1 Main St.	Mexican
Pho Place	**JJ29JJ**	192 Street Rd.	Vietnamese
Taco Stand	XJ11AS	18 W. East St.	Fusion
Pizza Heaven	CI21AA	711 K Ave.	Italian

health inspections

id	inspection_date	inspector	score
AH13JK	2018-08-21	Sheila	97
JJ29JJ	2018-03-12	D'eonte	98
JJ29JJ	2018-01-02	Monica	66
XJ11AS	2018-12-16	Mark	43
CI21AA	2018-08-21	Anh	99

rating

id	stars
AH13JK	4.9
JJ29JJ	4.8
XJ11AS	4.2
CI21AA	4.7

relational database

restaurant

name	id	address	type
Taco Stand	AH13JK	1 Main St.	Mexican
Pho Place	**JJ29JJ**	192 Street Rd.	Vietnamese
Taco Stand	XJ11AS	18 W. East St.	Fusion
Pizza Heaven	CI21AA	711 K Ave.	Italian

Two different restaurants with the same name will have different unique identifiers

health inspections

id	inspection_date	inspector	score
AH13JK	2018-08-21	Sheila	97
JJ29JJ	2018-03-12	D'eonte	98
JJ29JJ	2018-01-02	Monica	66
XJ11AS	2018-12-16	Mark	43
CI21AA	2018-08-21	Anh	99

rating

id	stars
AH13JK	4.9
JJ29JJ	4.8
XJ11AS	4.2
CI21AA	4.7

relational database

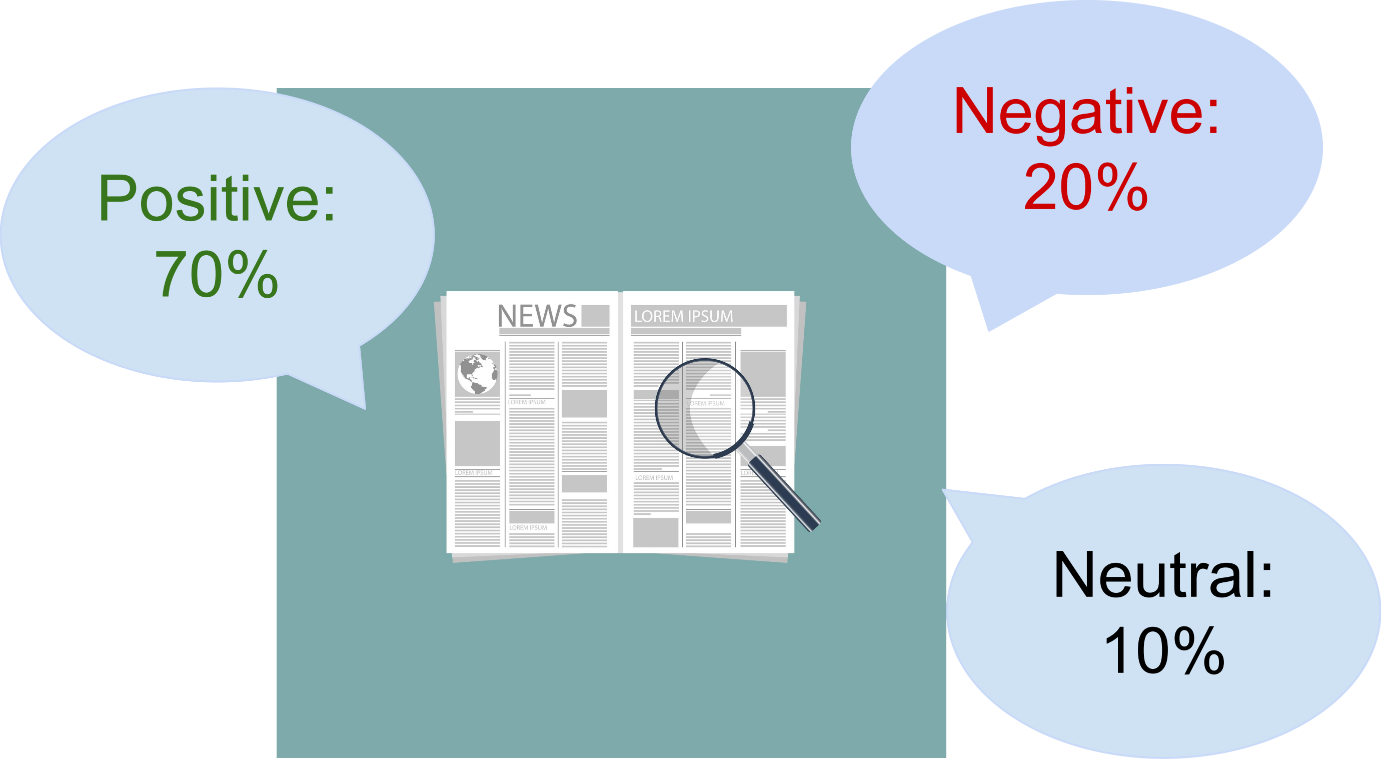
Within structured data, what information will be stored?

Variable types

- **Quantitative data** consists of numerical values, like height and weight.
- **Categorical data** consists of labels describing the properties of the objects under investigation, like gender, hair color, and occupation
  - Categorical data doesn’t have an order to it
  - Does it make any sense to talk about the maximum or minimum hair color? What is the interpretation of my hair color minus your hair color?

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.

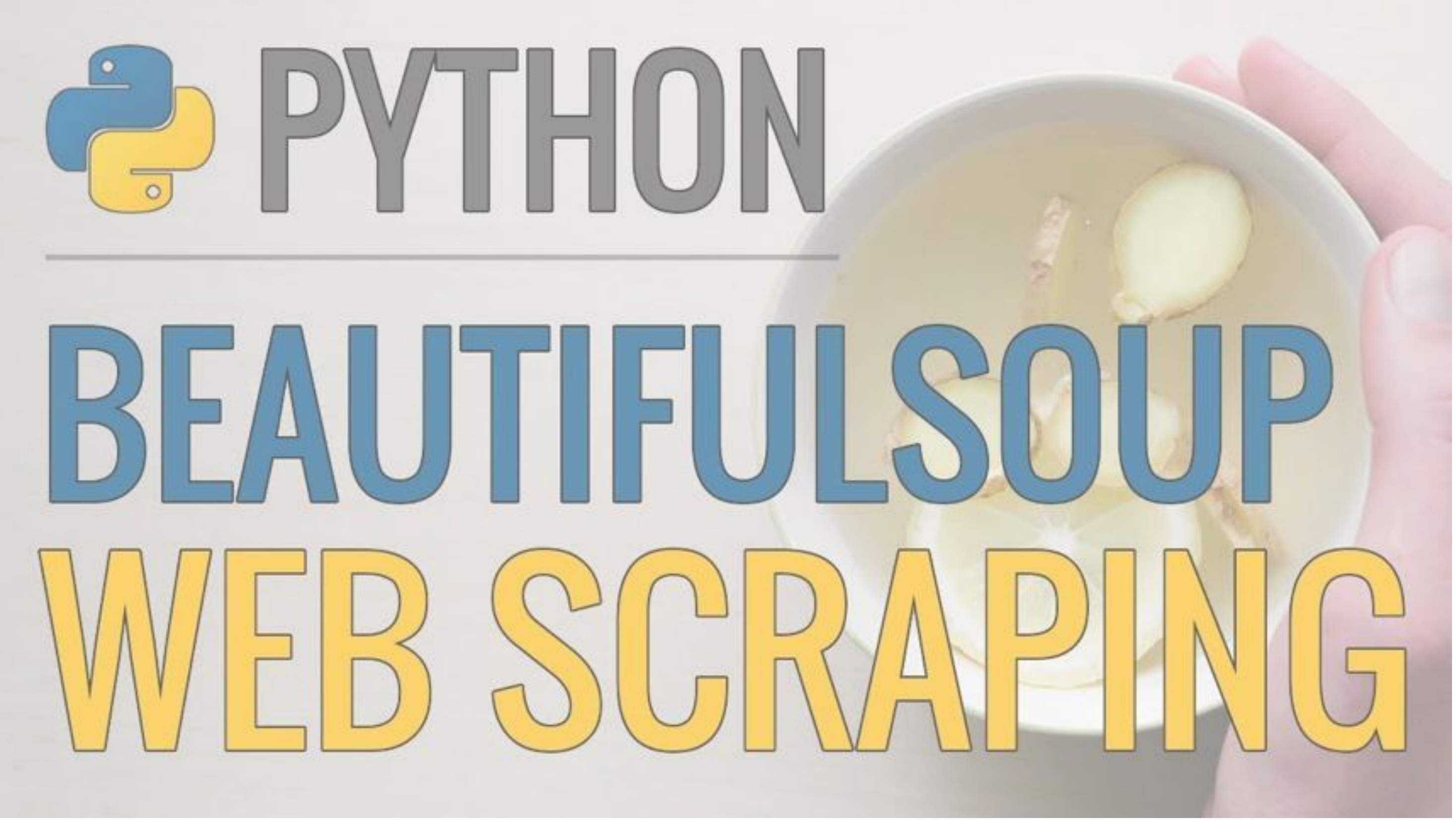


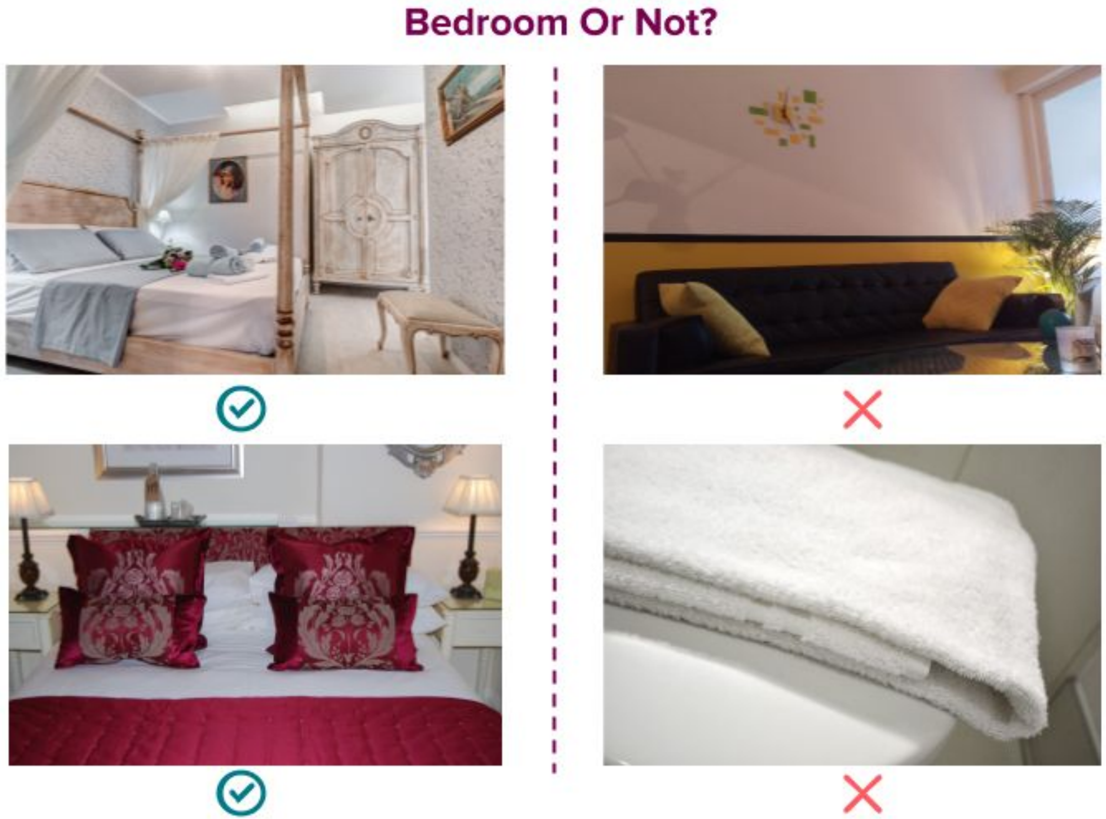
Positive: 70%

Negative: 20%

Neutral: 10%

Text:
Sentiment Analysis


PYTHON
BEAUTIFULSOUP
WEB SCRAPING



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

https://medium.com/airbnb-engineering/categorizing-listing-photos-at-airbnb-f9483f3ab7e3

Unstructured Data Types

Text files and documents

Websites and applications

Sensor data

Image files

Audio files

Video files

Email data

Social media data

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

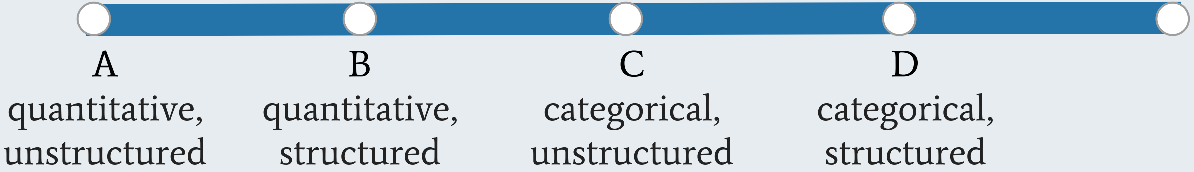
Unstructured

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

Data Sleuth I

You have information about shoe size stored in a JSON file for 1000 people.

Which of the following best describes these data?



A
quantitative,
unstructured

B
quantitative,
structured

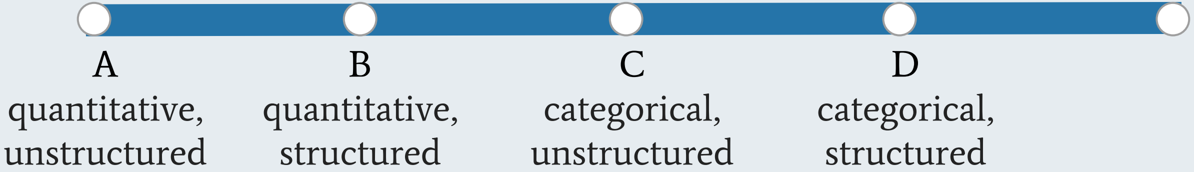
C
categorical,
unstructured

D
categorical,
structured

Data Sleuth II

You have information about everyone in the class' favorite ice cream flavor displayed on a website.

Which of the following best describes these data?



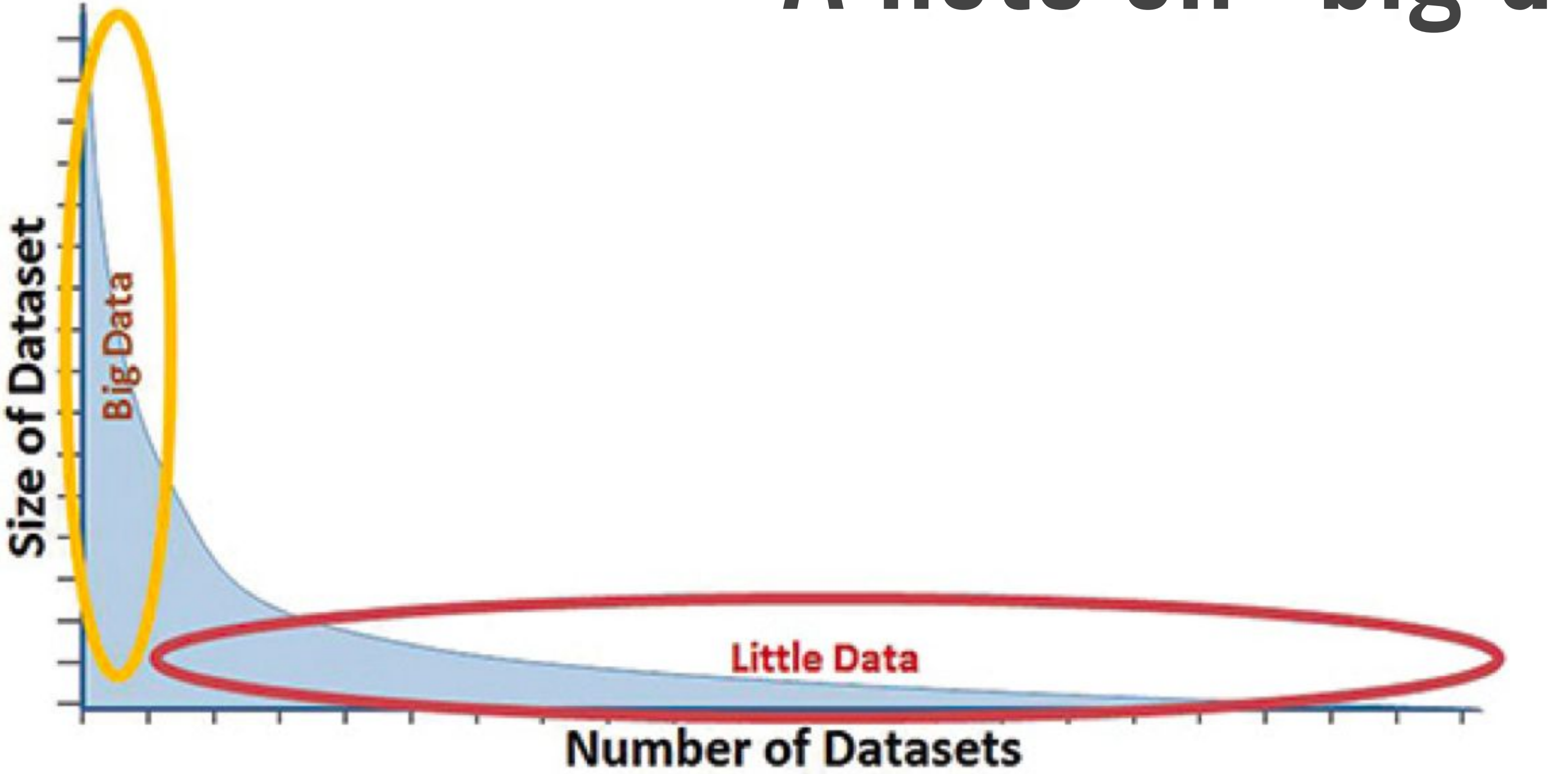
A
quantitative, unstructured

B
quantitative, structured

C
categorical, unstructured

D
categorical, structured

A note on “big data”



adapted from Chris Keown

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 demographic 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.

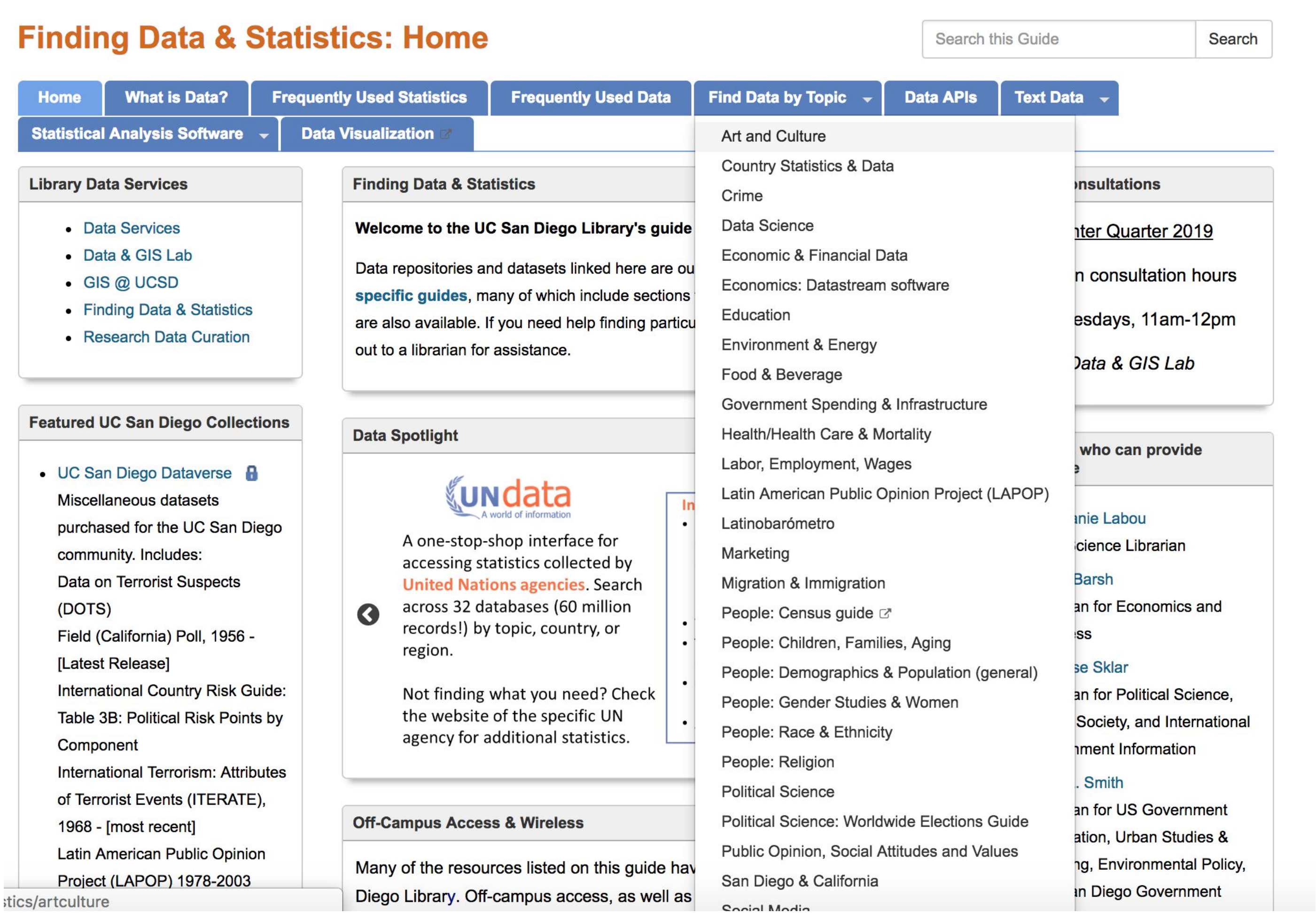
The best projects start with a question NOT the dataset. The most boring projects are dataset-first.

Where to look for and get data for your projects?

Available Datasets

- US Census Data
- data.gov
- Awesome Public Datasets
- Data Is Plural
- Datasets | Deep Learning
- Stanford | Social Science Data Collection
- Open Climate Data
- Eviction Lab (email required)
- Data and Story Library

There are more dataset sources listed on the FinalProject_Guidelines document on GitHub


Search this Guide
Search
Home
What is Data?
Frequently Used Statistics
Frequently Used Data
Find Data by Topic
Data APIs
Text Data
Statistical Analysis Software
Data Visualization
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
UNdata
A world of information
stics/artculture

Finding Data & Statistics: Home

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

Finding Data & Statistics

Welcome to the UC San Diego Library's guide

Data repositories and datasets linked here are ou **specific guides**, many of which include sections are also available. If you need help finding particu 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.

In

Off-Campus Access & Wireless

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

nsultations

nter Quarter 2019

n consultation hours

esdays, 11am-12pm

Data & GIS Lab

who can provide e

nie Labou
cience Librarian

Barsh
an for Economics and ss

se Sklar
an for Political Science, Society, and International nment Information

. Smith
an for US Government ation, Urban Studies & ng, Environmental Policy, n Diego Government

https://ucsd.libguides.com/data-statistics/home

When the data aren’t ready and waiting for you

- APIs
- Web Scraping
- Collecting your own data