slido

2868377

Join at slido.com
#2868377

ⓘ Click **Present with Slido** or install our Chrome extension to display joining instructions for participants while presenting.

Pre-semester Survey

2868377

Take a minute to fill out the pre-semester survey now if you haven't already!

We need these responses to assign discussion sections.

⚠️ Reminder to start the Zoom recording!

2868377

LECTURE 1

Course Overview

An overview of data science, Data 100, and the data science lifecycle.

Data 100, Summer 2025 @ UC Berkeley

Josh Grossman and Michael Xiao

Pre-semester Survey

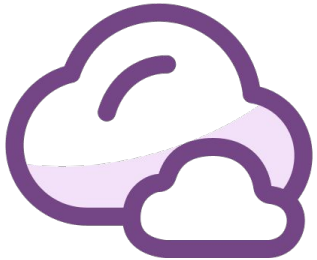
2868377

Take a minute to fill out the pre-semester survey now if you haven't already!

We need these responses to assign discussion sections.

slido

2868377

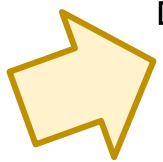


What emoji best describes your mood today?

Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

Slido Q&A

2868377



Use the QR code on any slide to post questions to slido.

This is the best place to clarify course content.

Read the syllabus to learn more about the Data 100 attendance policy.

We'll have snacks starting with the next lecture 🍬

2868377

Roadmap

Lecture 01, Data 100 Summer 2025

- **Intros**
- What is data science?
- What will you learn in this class?
- Course overview
- Data Science Lifecycle

Intro - Josh Grossman

2868377

Assistant Teaching Professor in **Statistics + Data Science** (started July 2024)

BA 2016 @ **Harvard** in Neurobiology (you don't have to get it right the first time!)

Ph.D. 2024 @ **Stanford** in Computational Social Science

Research: Intersection of public policy and data science, broadly

- See jdgrossman.com for more info

Courses: *Data 100 (sp25, su25, fa25), Stat 131A (fa24, sp26)*

Between college and grad school:

- 2 years as a product manager @ IXL Learning (K–12 edtech company)

Want to talk 1-1 about grad school, career plans, life, …? Sign up for a 15-minute chat!

Intro - Michael Xiao

2868377

Lecturer at **CDSS** (2025)

BA, MA in Statistics; **BS** in Environmental Science @ **Berkeley**

PhD (incoming) in Data Science @ **UChicago**

Research: Interpretability and Uncertainty Quantification (not sure)

- Also a bunch of fun **data projects**:
  - Environmental impact of cannabis legalization in Oregon
  - Causes and prevention of california forest fires
  - Decoding and analysis of AirBnb pricing model

Courses: *Data 140 (Fa20 - Sp25); Stat 134 (Su24); Data 188 (Fa22); Data 88S (Sp25)*

- **Probability** and **Statistical Inference**

Michael and Josh Tips

2868377

You can call us "Michael" and "Josh". Say hi in public; it's OK if we don't know you yet!

Unless you have a private question for just one of us, don't use our personal emails.

- Email data100.instructors@berkeley.edu or post on Ed.

We will always have office hours (i.e., OH, help hours) after lecture in **HFAX B1** nearby.

- No prep is required for our office hours. **We want to talk to you!**
- If you talk to anyone who has taken a class with us, most will say office hours were one of the **best parts of the course**.
- Pro tip, OH is often where cool opportunities+contacts+letters of rec originate.

2868377

What is Data Science?

Lecture 01, Data 100 Summer 2025

- Intros
- **What is data science?**
- What will you learn in this class?
- Course overview
- Data Science Lifecycle

2868377

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

?

2868377

A recent+powerful case study

One example of data science driving positive change

⚠️Keep in mind: This case study discusses racial bias. Assessing discrimination is critically important, but also challenging. Please engage to the extent you are comfortable.

Adapted from research by Elzayn et al. (2025)

2868377

Black Americans Are Much More Likely to Face Tax Audits, Study Finds

A new report documents systemic discrimination in how the I.R.S. selects taxpayers to be audited, with implications for a debate on the agency's funding.

IRS : Internal Revenue Service. U.S. federal agency that collects taxes.

Tax audit : IRS suspects your taxes are incorrect. Burdensome to the IRS + taxpayers.

Discrimination : Unjustified differences in decisions+outcomes across race, gender, age, …

But, taxpayers never provide their race to the IRS. How can the IRS discriminate?

Initial Research Question

2868377

What are audit rates of Black and non-Black taxpayers?

- 👥 Human context: Why does this comparison matter? What is implied by X≠Y?
- 📊💡 Core component of data science: Calculations across groups.

X% of tax returns of **Black** taxpayers were audited.

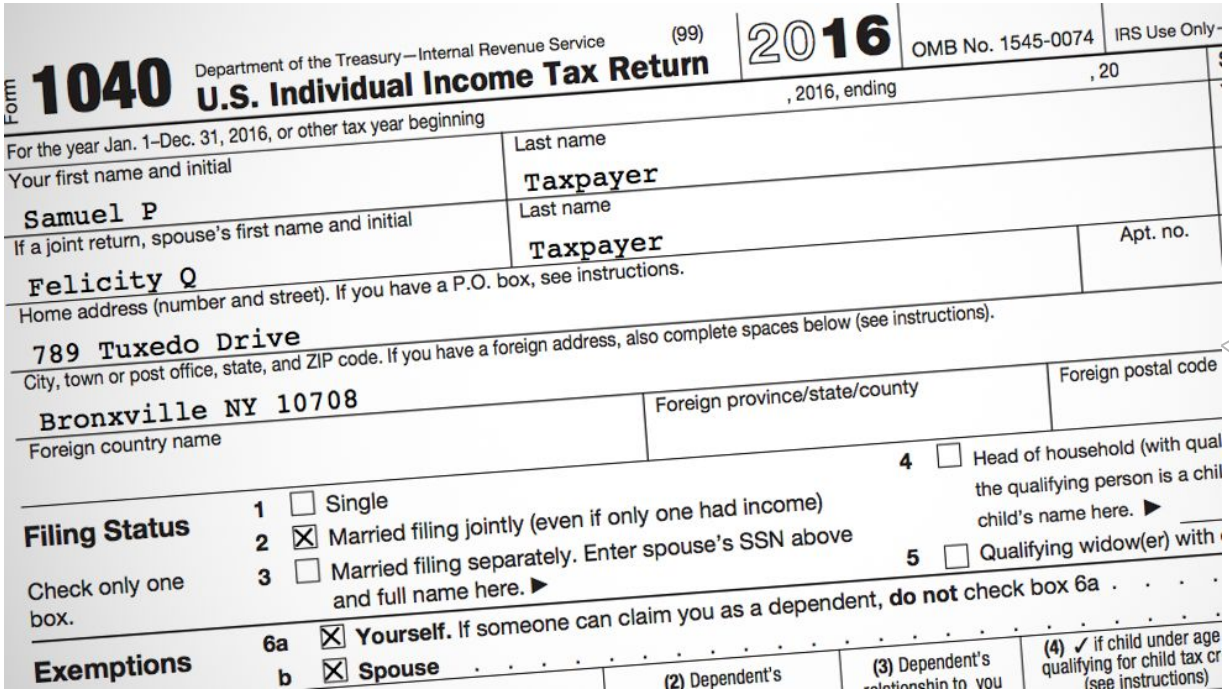
Y% of tax returns of **non-Black** taxpayers were audited.

Calculating Audit Rates Across Race

2868377

How do we know who was audited? **Data from the IRS.**

- Obtained from partnership between researchers and the IRS.
- 📊💡 Collaboration, whether across agencies or industry teams. Often messy!



Form **1040** Department of the Treasury—Internal Revenue Service (99) **U.S. Individual Income Tax Return** **2016** OMB No. 1545-0074 IRS Use Only—

For the year Jan. 1–Dec. 31, 2016, or other tax year beginning , 2016, ending , 20

Your first name and initial: Samuel P

Last name: Taxpayer

If a joint return, spouse's first name and initial: Felicity Q

Last name: Taxpayer

Apt. no.

Home address (number and street). If you have a P.O. box, see instructions.: 789 Tuxedo Drive

City, town or post office, state, and ZIP code. If you have a foreign address, also complete spaces below (see instructions).: Bronxville NY 10708

Foreign country name

Foreign province/state/county

Foreign postal code

Filing Status Check only one box.

1 ☐ Single

2 ☒ Married filing jointly (even if only one had income)

3 ☐ Married filing separately. Enter spouse's SSN above and full name here. ▶

4 ☐ Head of household (with qual the qualifying person is a chil child's name here. ▶

5 ☐ Qualifying widow(er) with

Exemptions

6a ☒ **Yourself.** If someone can claim you as a dependent, **do not** check box 6a . . . .

b ☒ **Spouse** . . . . . (2) Dependent's (3) Dependent's relationship to you (4) ✓ if child under age qualifying for child tax cr (see instructions)

Audited?

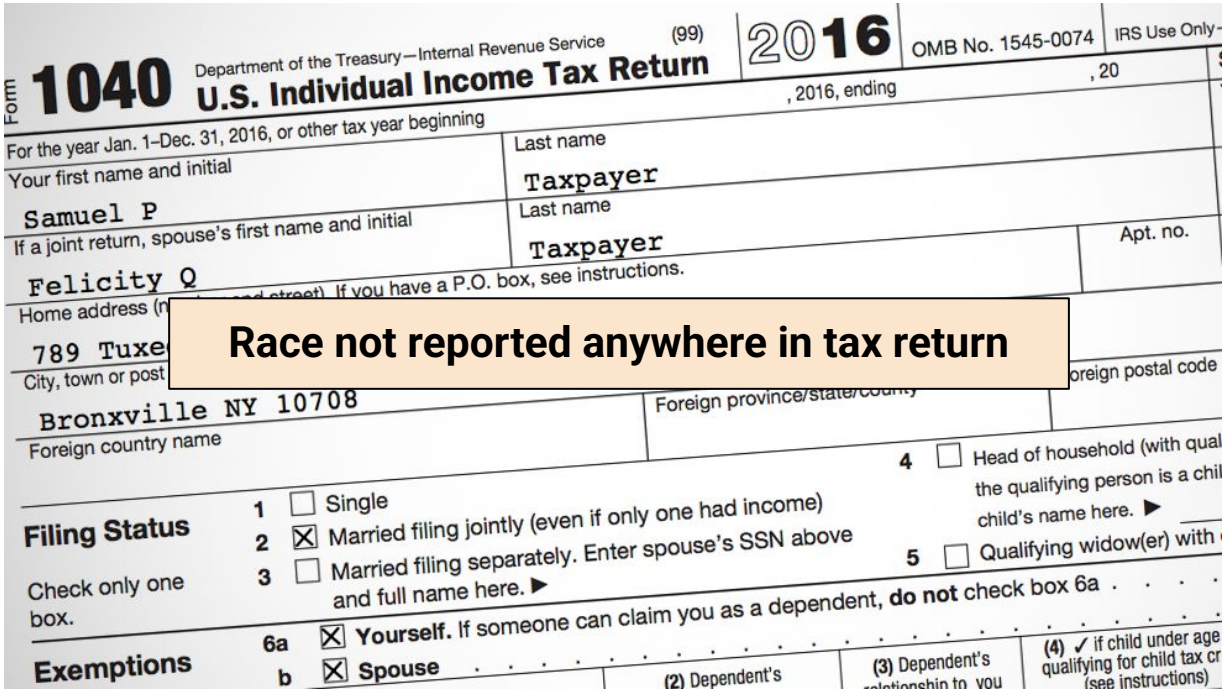
Not audited?

Calculating Audit Rates Across Race

2868377

How do we determine the race of each taxpayer? **We can't.**

- 📊💡 Making do with imperfect knowledge. Being explicit about assumptions.
- Could we make an informed **prediction** of taxpayer race, and then use that prediction?



First Names, Last Names, Hometowns, and Race/Ethnicity

2868377

The prevalence of hometowns, first names, and last names differs across race+ethnicity.

- 👥 Why might these differences exist? Are they "bad"?

Decennial Census | Universe: Total population | 2020: DEC Demographic a...

Label	Berkeley city, California
Total:	124,321
Hispanic or Latino	17,018
Not Hispanic or Latino:	107,303
Population of one race:	98,234
White alone	62,450
Black or African American al...	9,495
American Indian and Alaska ...	226
Asian alone	24,701

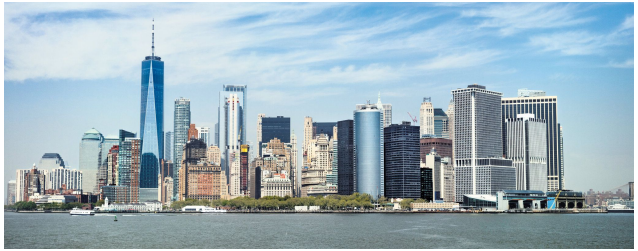
Source: 2020 U.S. Census

Race/Ethnicity	Most Popular Baby Name
Asian, non-Hispanic	Sophia
Black, non-Hispanic	Madison
Hispanic	Isabella
White, non-Hispanic	Olivia

Source: NYC Health (2013)

Inferring Race from Hometown

2868377



New York City (NYC) residents
2020 Census: **16%** Asian-identifying

San Francisco (SF) residents
2020 Census: **34%** Asian-identifying

If I were to **randomly** choose one NYC resident and one SF resident, there is a **higher probability** that the SF resident identifies as Asian.

- 📊💡 Using probability to reason about the world. Importance of random sampling.

Inferring Race from First Name, Last Name, and Neighborhood

2868377

Drawing on external datasets, the researchers estimated the probability that a taxpayer with a particular location+name identifies as Black or non-Black. See Bayes' rule, Naive Bayes, BIFSG

- 📊💡 Joining in external datasets to enable novel analysis.

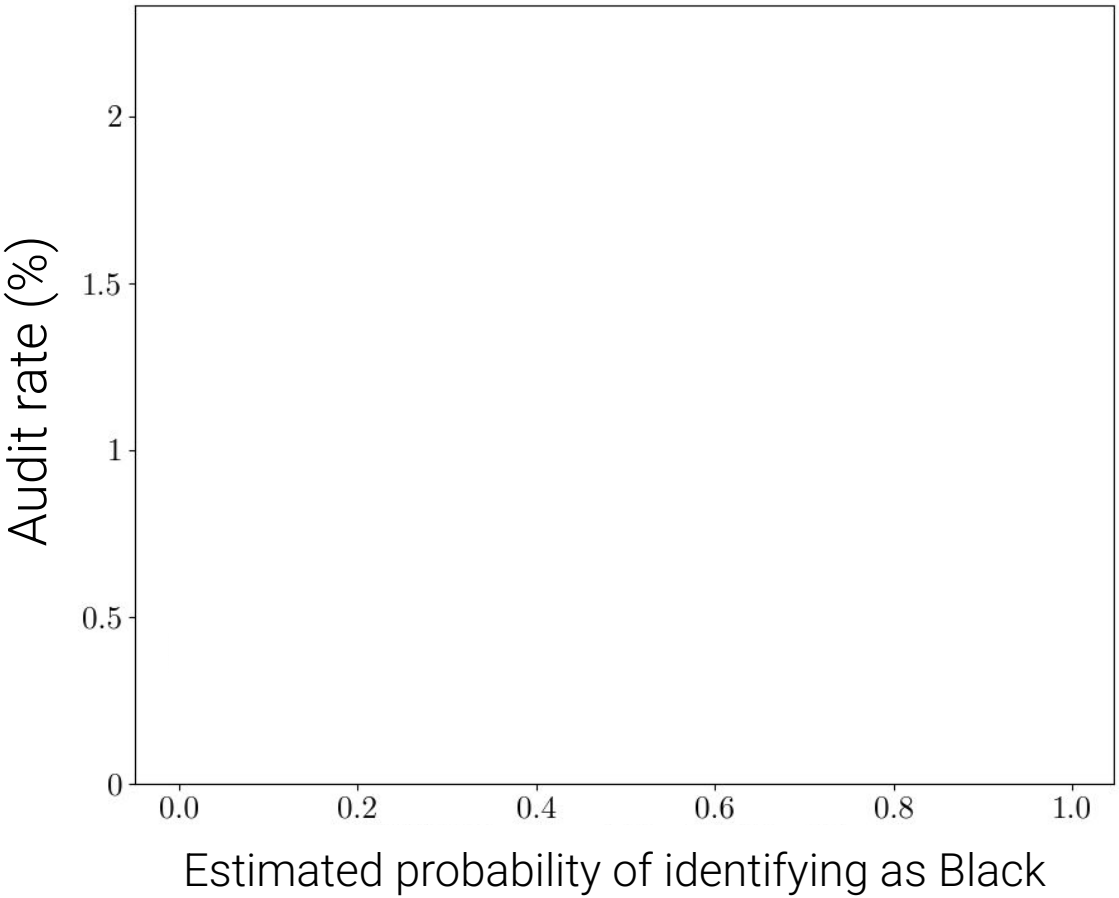
Robert Johnson from **neighborhood A** in NYC: **X%** probability of identifying as Black

Tanisha Tompkins from **neighborhood B** in SF: **Y%** probability of identifying as Black

And so on for the remaining taxpayers.

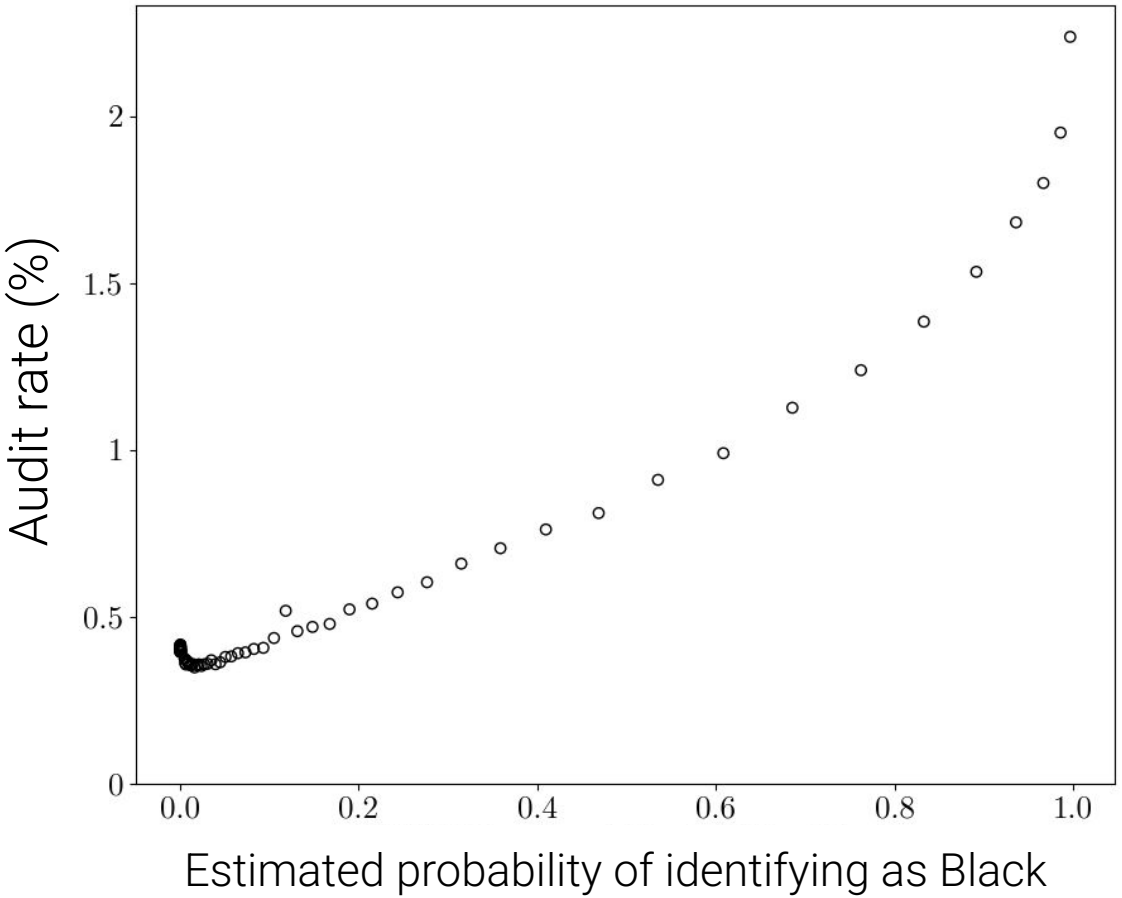
Visualizing the Data

2868377



Black Taxpayers 3-5x More Likely to be Audited

2868377



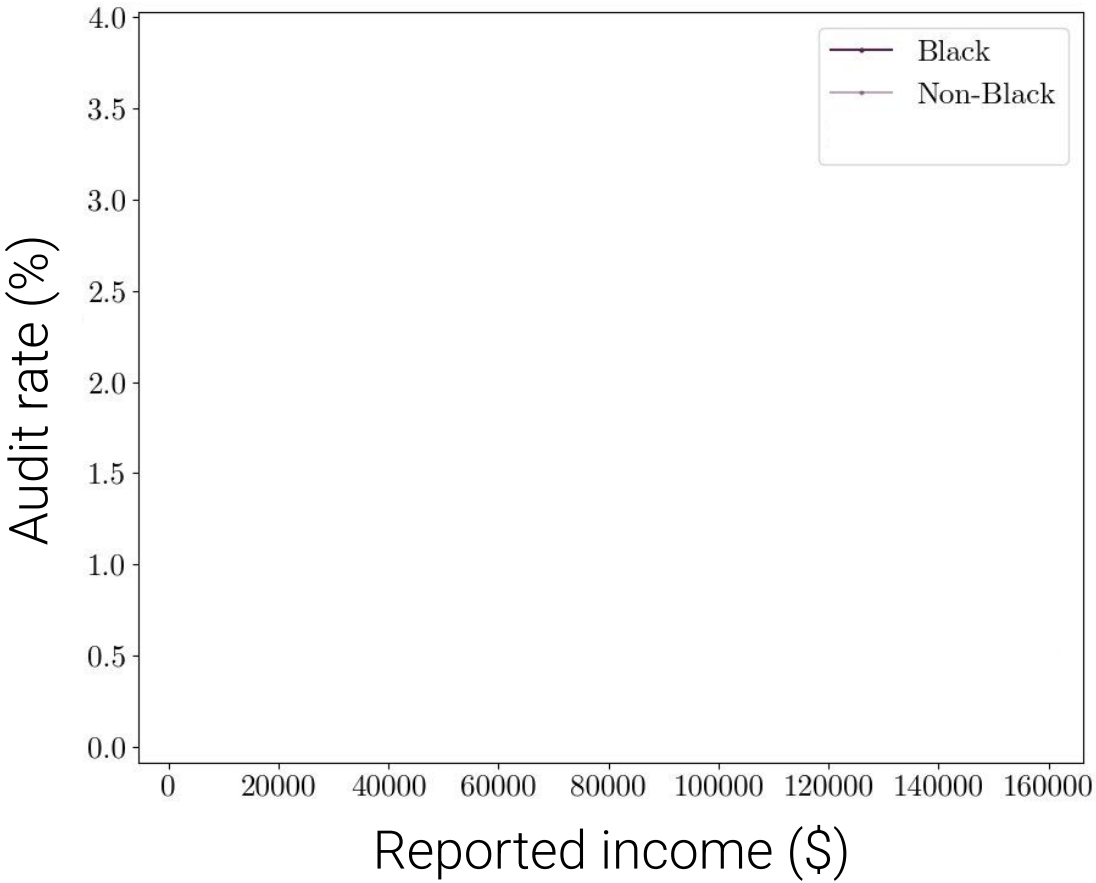
Taxpayers with a higher estimated probability of identifying as Black were more likely to be audited.

- 📊💡 Visualization is a critical form of communication

Each point is one of 100 equally-sized groups of taxpayers.

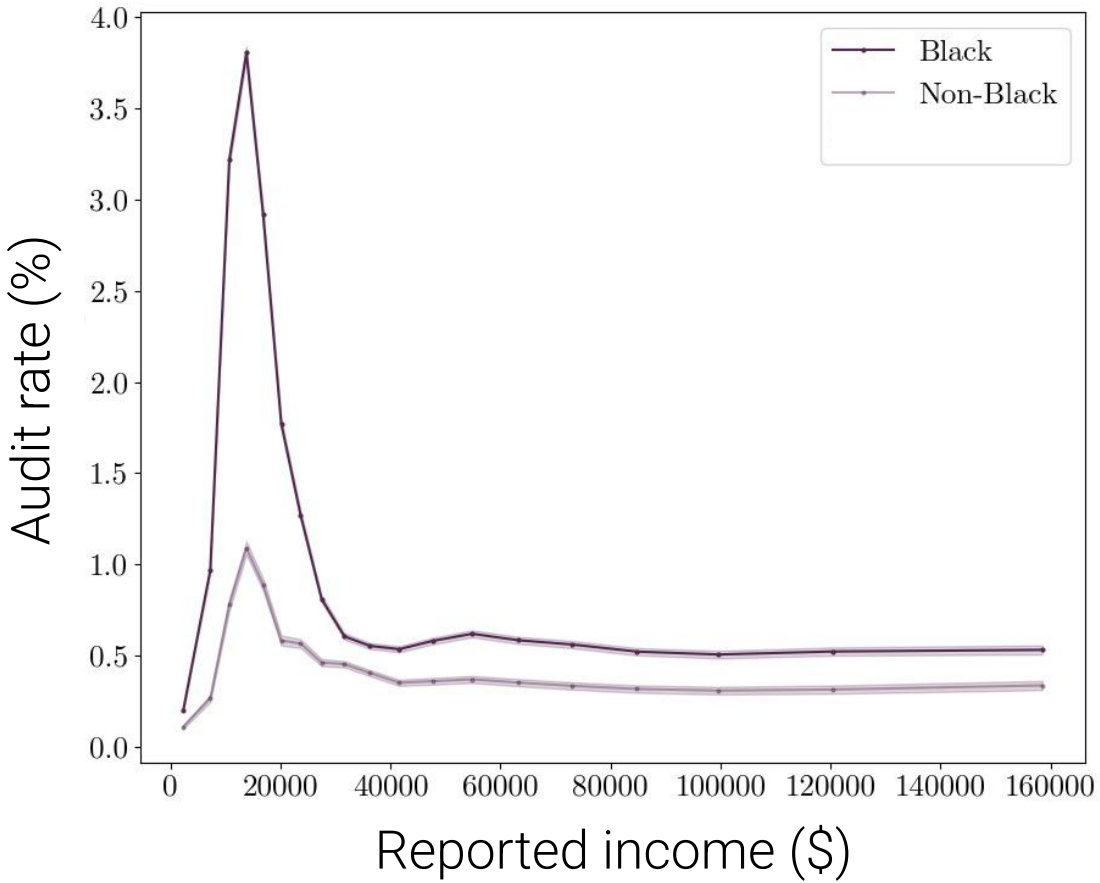
Visualizing the Data, with Income

2868377



Differences in Income Do Not Fully Explain the Gap in Audit Rates

2868377



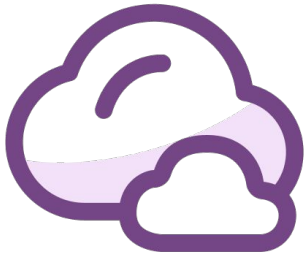
Gap in audit rates persists across income levels, esp. lower incomes.

- 📊💡"Adjusting" or "controlling" for additional variables

Remember, the IRS does not observe taxpayer race.

So, what's driving the gap?

Each point is one of 20 equally-sized groups of taxpayers.



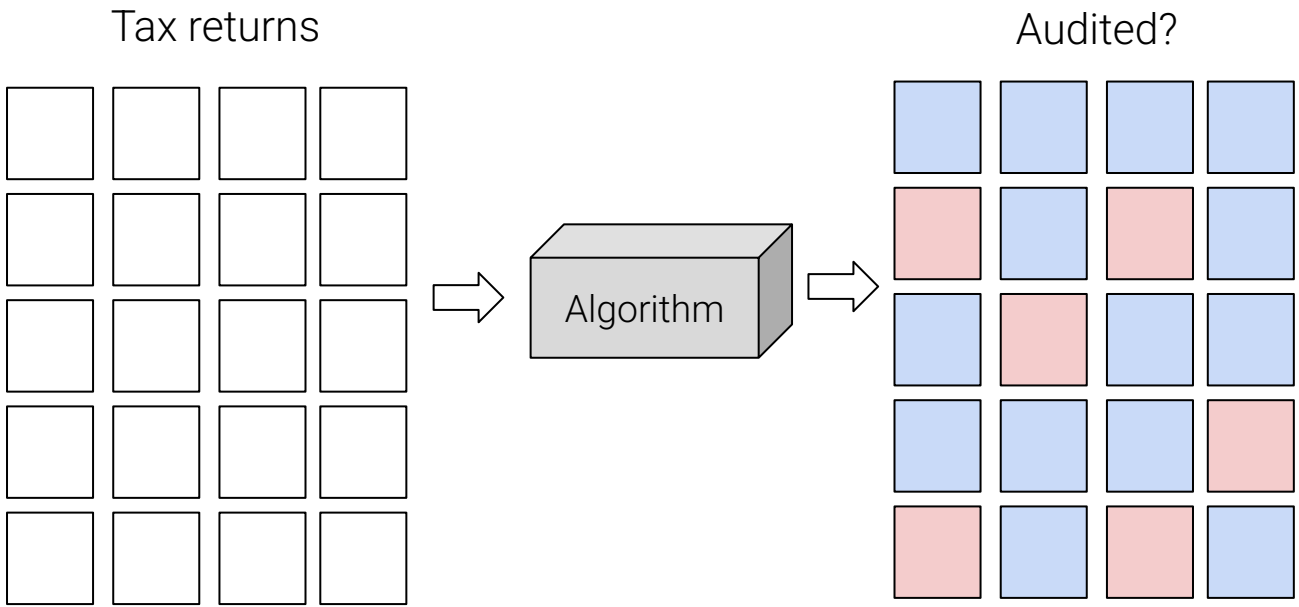
What do you think could be responsible for the patterns in this plot? Remember, the IRS does not observe race or ethnicity of taxpayers.

Identifying a Potential Root Cause

2868377

As it turns out, the IRS uses an algorithm to flag errors in tax returns.

- 📊💡 Models built by data scientists can improve efficiency, but they must be scrutinized.

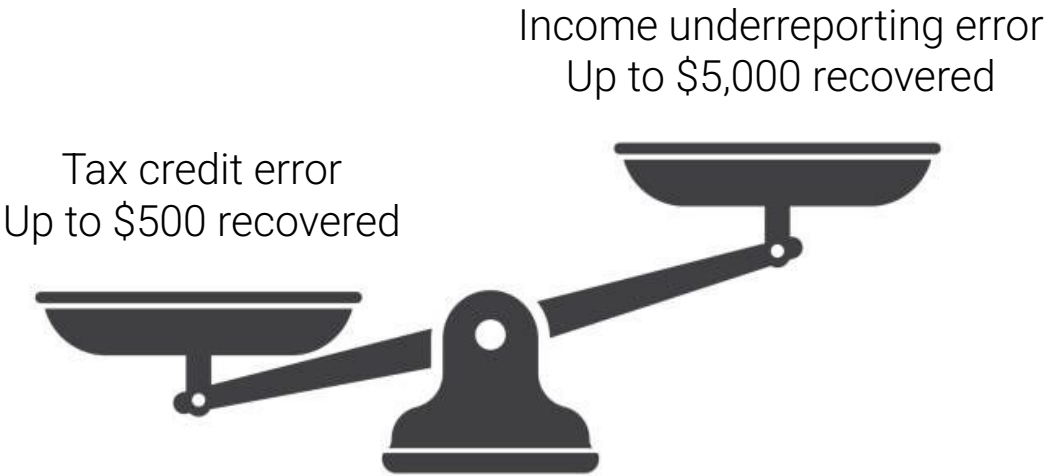


Identifying a Potential Root Cause

2868377

The algorithm seems to have prioritized catching **errors in claimed tax credits** over catching errors that, if addressed, would **recover the most money**.

- 📊💡 Thinking carefully about choosing a metric for success



Identifying a Potential Root Cause

2868377

Black taxpayers were **more likely** to file the kinds of returns targeted by the algorithm.

- Thus, audit rates of Black taxpayers were **higher**.

With better policy, IRS may have recovered **a lot more money** with **smaller racial disparities**.

There's a lot more to this story. See research paper and NYTimes article. Come discuss during office hours (i.e., help hours)!

2868377

I.R.S. Changes Audit Practice That Discriminated Against Black Taxpayers

The agency will overhaul how it scrutinizes returns that claim the earned-income tax credit, which is aimed at alleviating poverty.

You can improve the world with tools from Data 100.

There's a lot more to this story. See research paper and NYTimes article. Come discuss during office hours (i.e., help hours)!

2868377

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

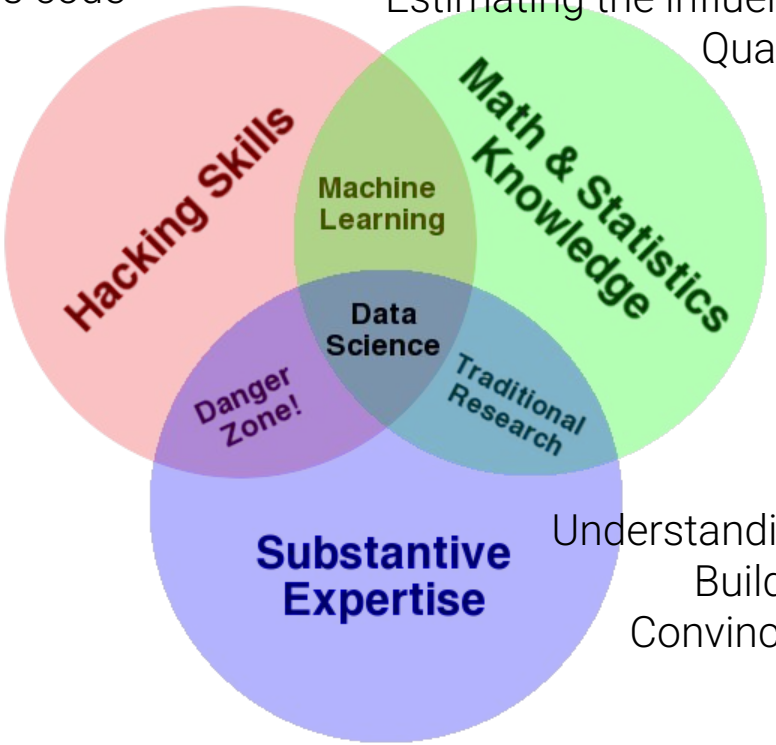
?

What Is Data Science?

2868377

Wrangling the IRS data into a usable format
Writing and debugging analysis code
Generating the visualizations
...

Designing a model for inferring race
Estimating the influence of variables on audit rates
Quantifying statistical uncertainty
...



Understanding the IRS auditing pipeline
Building relationships at the IRS
Convincing the public of the results
...

by Drew Conway in 2010 (link)

Transitioning From Data 8 to Data 100

2868377

Textbook title: Computational and Inferential Thinking

Prof. Joey Gonzalez

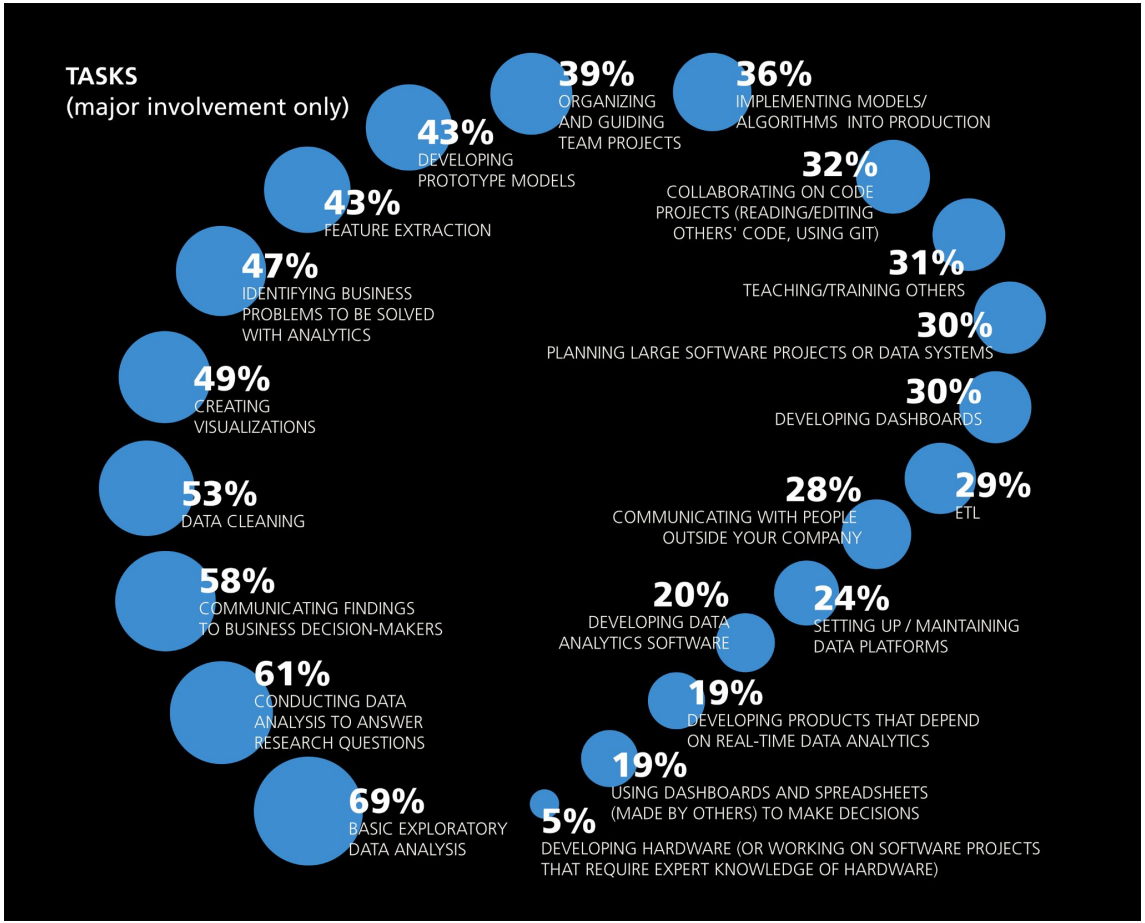
Long-time Data 100 Instructor

Data Science is the application of data-centric, computational, and inferential thinking to:

- Understand the world (**science**).
- Solve problems (**engineering**).

What Do Industry Data Scientists Do All Day?

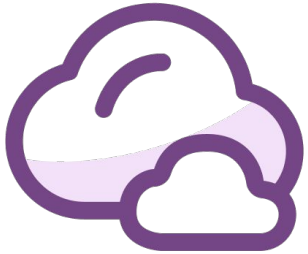
2868377



Hmm, can't a large language model (LLM) do a lot of these tasks? Will LLMs replace data scientists?

The tasks that data scientists say they work on regularly.

Self-reported. Based on the results of the 2016 Data Science Salary Survey.



What is/are the role(s) of LLMs, like Gemini and ChatGPT, in the future of data science?

Are LLMs going to replace data scientists?

2868377

Short answer: No one knows for sure. But, Michael and Josh have hope!

Step 1	Step 2	Step 3
Gather context and ask insightful questions.	Plan analyses that address your question(s).	Run your chosen analyses.
LLMs cannot deeply understand the **nuanced** desires, abilities, and personalities of complex organizations. At least for now 🙂	LLMs can provide advice as you weigh options. But, the choice of "best" analysis is often **subjective**. Experience matters.	With proper context and instructions, LLMs can help a lot here! But, you must be able to **verify** outputs. Don't do analysis by "vibes".

Much of the text on this slide is grossly oversimplified. Just a heads up!

Are LLMs going to replace data scientists?

2868377

Short answer: No one knows for sure. But, Michael and Josh have hope!

Step 1	Step 2	Step 3
Gather context and ask insightful questions.	Plan analyses that address your question(s).	Run your chosen analyses.

The secret is out: Most Data 100 homework problems can be answered by LLMs.

If LLMs write your answers, you are fully **outsourcing** Steps 1 and 2, and you **never learn to verify** the outputs of Step 3.

In other words, you are setting yourself up for future **replacement by LLMs**, regardless of your degree title/honors/grades/etc. Employers know this, too. **Beware!**

Large language models (LLMs) in Data 100

2868377

But, we encourage you to use LLMs **appropriately** in Data 100!

- Appropriate: *"What does this error message mean?"*
- Inappropriate: *"HW 3 Problem 2: Imagine you have three boxes…"* → Academic dishonesty!
- Be sure to read Data 100's academic honesty policy carefully.

In the first discussion, we will show you how to install a custom LLM for Data 100 that will help you learn more effectively.

Be sure to sign up for FREE Gemini pro before June 30th!

2868377

What Will You Learn in This Class?

Lecture 01, Data 100 Summer 2025

- Intros
- What is data science?
- **What will you learn in this class?**
- Course overview
- Data Science Lifecycle

2868377

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

Course Goals

2868377

Prepare

Prepare you for advanced courses in **data management**, **machine learning**, and **statistics**.

Enable

Enable you to start a career as a data scientist by providing experience with **real-world data, tools, and techniques**.

Empower

Empower you to apply computational and inferential thinking to address **real-world problems**.

Tentative List of Topics to be Covered in Data 100

2868377

- Pandas and NumPy
- Relational Databases & SQL
- Exploratory Data Analysis
- Regular Expressions
- Visualization
  - matplotlib
  - Seaborn
  - plotly
- Sampling
- Probability and random variables
- Model design and loss formulation
- Linear Regression
- Feature Engineering
- Regularization, Bias-Variance Tradeoff, Cross-Validation
- Gradient Descent
- Logistic Regression
- Clustering
- PCA

Remember, **these are just tools**. More importantly, you'll learn how to **reason with data**.


2868377

Course Overview

Lecture 01, Data 100 Summer 2025

- Intros
- What is data science?
- What will you learn in this class?
- **Course overview**
- Data Science Lifecycle

2868377

Prerequisites

Prerequisites

2868377

Official prerequisites for this course:

- Completion of Data 8.
- Completion of CS 61A, Data C88C, or Engineering 7.
- Co-enrollment in EE 16A or Math 54 or Stat 89A.

The prereqs are being strictly enforced! We will **not** be teaching:

- How to use Python.
- How to use Jupyter notebooks.
- Inference from Data 8.
- Linear algebra (though we will review this topic to a greater degree since linear algebra is a corequisite, not prerequisite).

Homework 1 and Lab 1 will help calibrate your background.

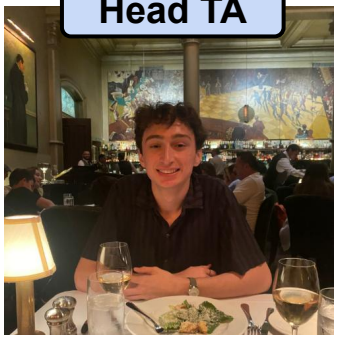
- For Homework 1, the Data 8 textbook will be helpful.

2868377

Staff

TAs

2868377



Jake Pastoria

Xiaorui Liu

Sammie Smith

Milena Novakovic

Wesley Zheng

TAs teach discussions and assist in a wide variety of tasks (examples include: developing content, exam prep, etc.). Contact info: ds100.org/su25/staff/.

Tutors

2868377

Ben Rabin

Cristina Prieto

Ella Hammond

Justin Wang

Rohan Bijukumar

Hannah Brown

Sara Eginova

Tutors help review and grade our assignments, answer lecture and Ed questions, host office hours, and get trained for future semesters! Contact info: ds100.org/su25/staff/.

2868377

Course Logistics
Content and Workflow

Syllabus Walkthrough

2868377

ds100.org/su25/syllabus/

2868377

Course Websites / Platforms

Online Platforms

2868377

Course Website (ds100.org/su25)

- All lectures, assignments, and discussions are posted here.

DataHub (data100.datahub.berkeley.edu)

- Where you will work on all assignments (links on the course website automatically take you here).

Ed (edstem.org/us/courses/80038/discussion)

- A place to ask and answer questions about assignments and concepts.
- Where all announcements are posted (exam logistics, new assignment released, etc).

Pensieve (pensieve.co/student/classes/data100_su25)

- Where all assignments are submitted, and where your assignment grades will live. Not Gradescope!

Lecture Notes (ds100.org/course-notes)

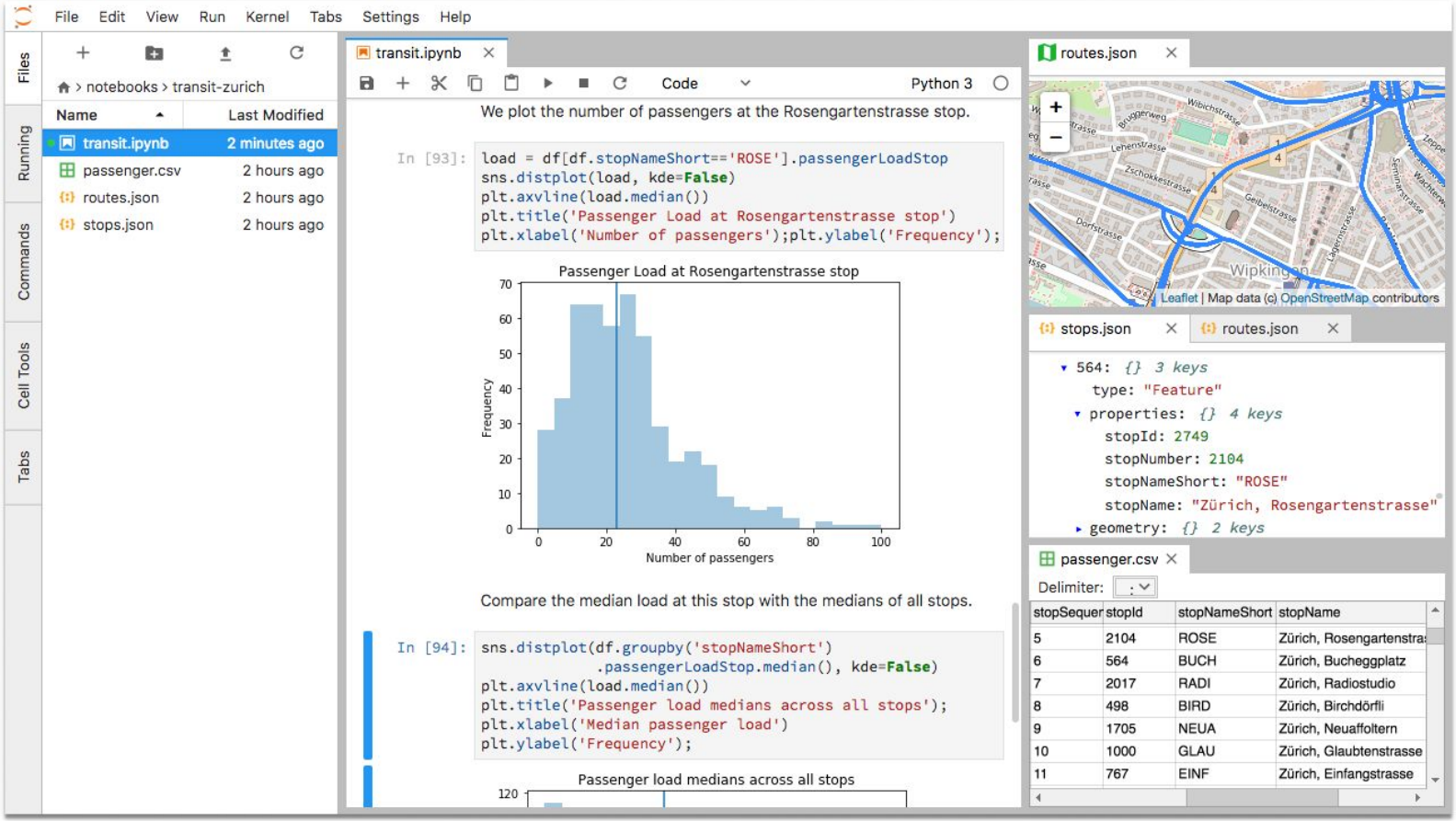
- A summary of each lecture. **Think of the course notes as the unofficial course textbook.**

Supplemental Textbook (learningds.org)

- Supplemental reading (not synchronized with the class schedule).

Programming Environment for our Course: JupyterLab

2868377



Learning Advanced JupyterLab

2868377

JupyterLab offers notebooks and more tools for data science.

We'll be accessing JupyterLab using **DataHub** (data100.datahub.berkeley.edu).

Resources for learning ~fancier~ JupyterLab functionality:

- **The quickest intro is this great 2-minute overview by Serena Bonaretti**.
  - Note: Unlike Serena's example, in our course we're using JupyterLab notebooks hosted on the internet, not on your own local computer.
- The interface overview from the official docs has more details and short, embedded videos.
- A more detailed discussion from a bio/data angle: ~45 minute video.
- Full ~3h in-depth tutorial is available from the core team.

This summer will go by fast!

2868377

The summer edition of Data 100 is **fast**.

Most weeks: Four 90-minute lectures, 2 discussions, 2 HWs, 2 labs. Nearly a full time job!

If you fall behind, don't be silent. Get in touch with course staff ASAP. We will help you! 😀

Most weeks, we will have two optional **catch-up sessions** to help you stay on track.

Also, this summer has several ~experimental~ components. We will be fair+flexible if anything doesn't go as planned!

Office hours and discussions

2868377

Discussions and TA/tutor office hours start this **Wednesday 6/25**!

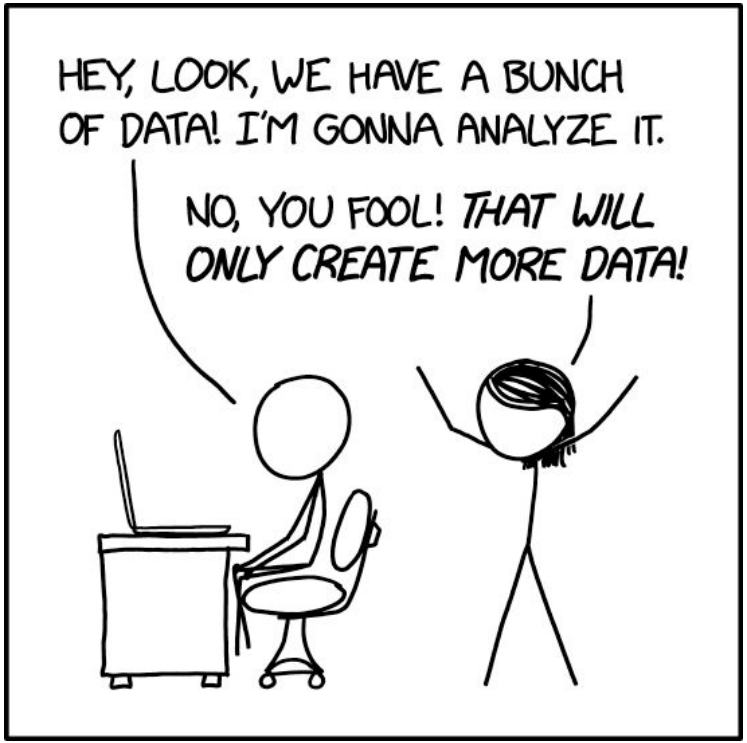
- Michael and Josh have office hours starting today immediately after lecture.

If you have questions before Wednesday, we encourage you to post on Ed.

2868377

Interlude

2-min stretch break!



xkcd.com/2582

2868377

Data Science Lifecycle

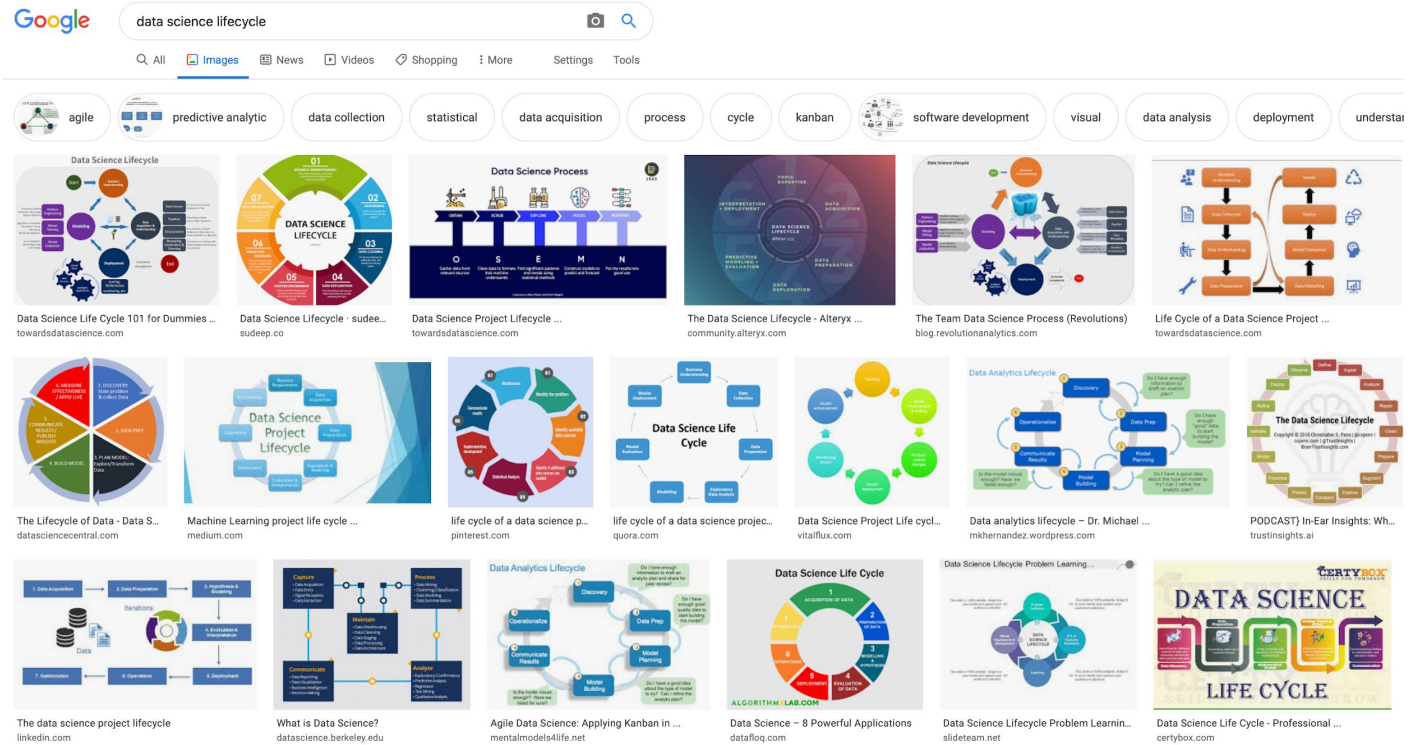
Lecture 01, Data 100 Summer 2025

- Intros
- What is data science?
- What will you learn in this class?
- Course overview
- **Data Science Lifecycle**

Data Science Lifecycle

2868377

The "data science lifecycle" you will see in the wild may be slightly different than the one we teach you, but the core ideas are all the same.

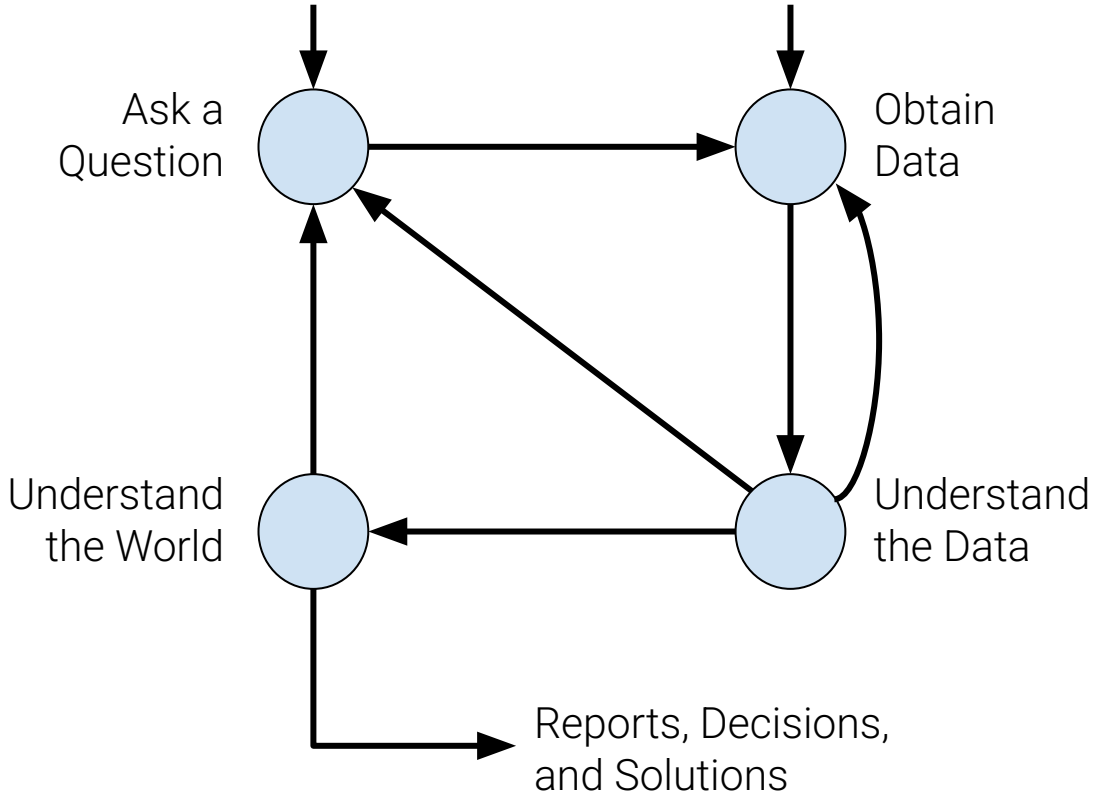


Data Science Lifecycle

2868377

The data science lifecycle is a **high-level description** of the data science workflow.

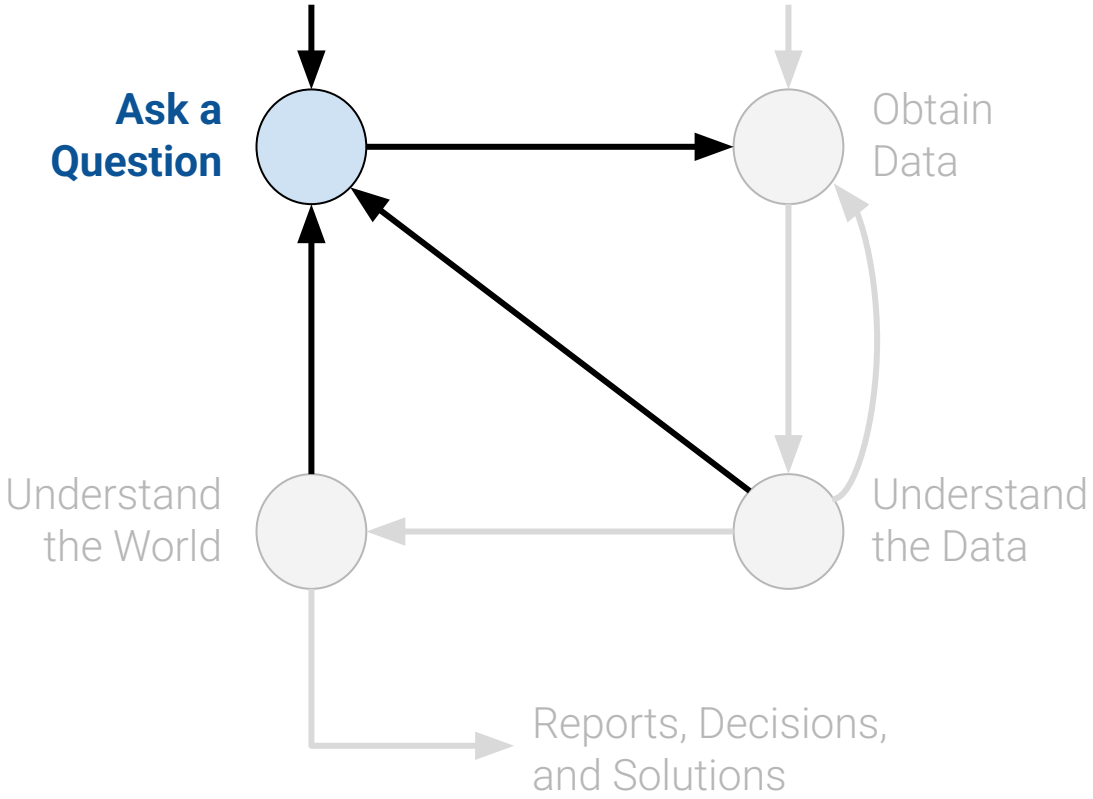
Note the two distinct entry points!



1. Question/Problem Formulation

2868377

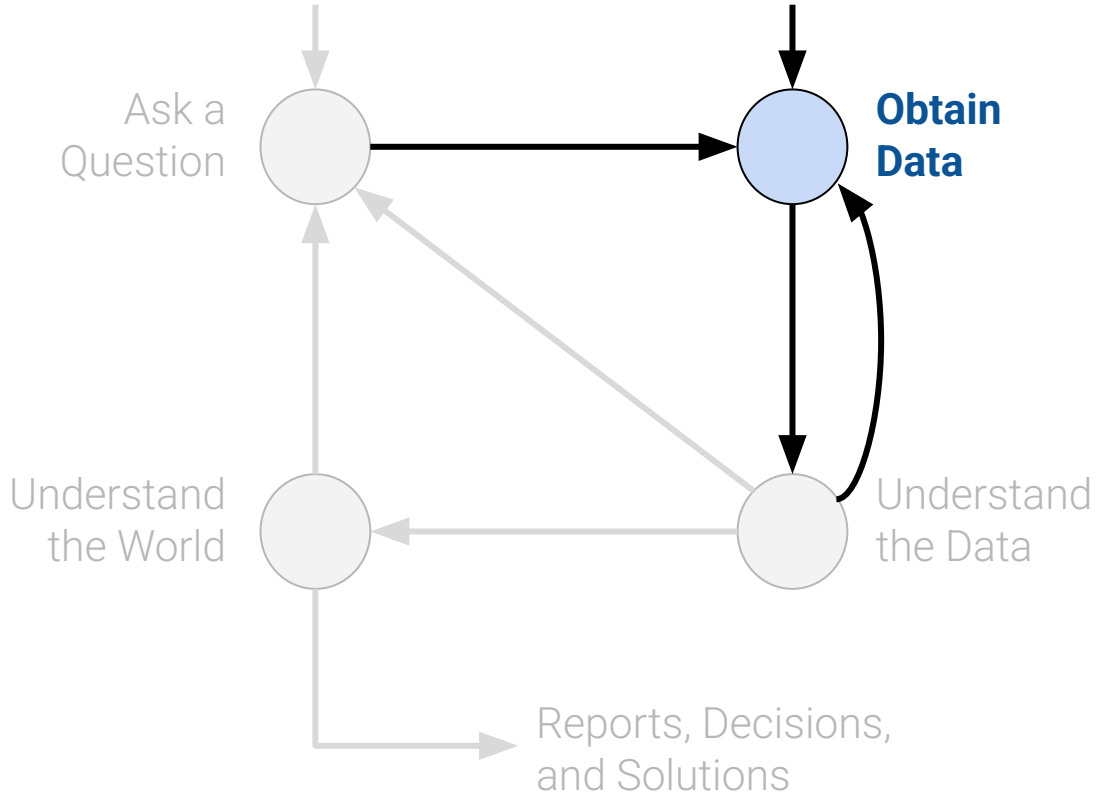
- What do we want to know?
- What problems are we trying to solve?
- What hypotheses do we want to test?
- What are our metrics for success?



2. Data Acquisition and Cleaning

2868377

- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?



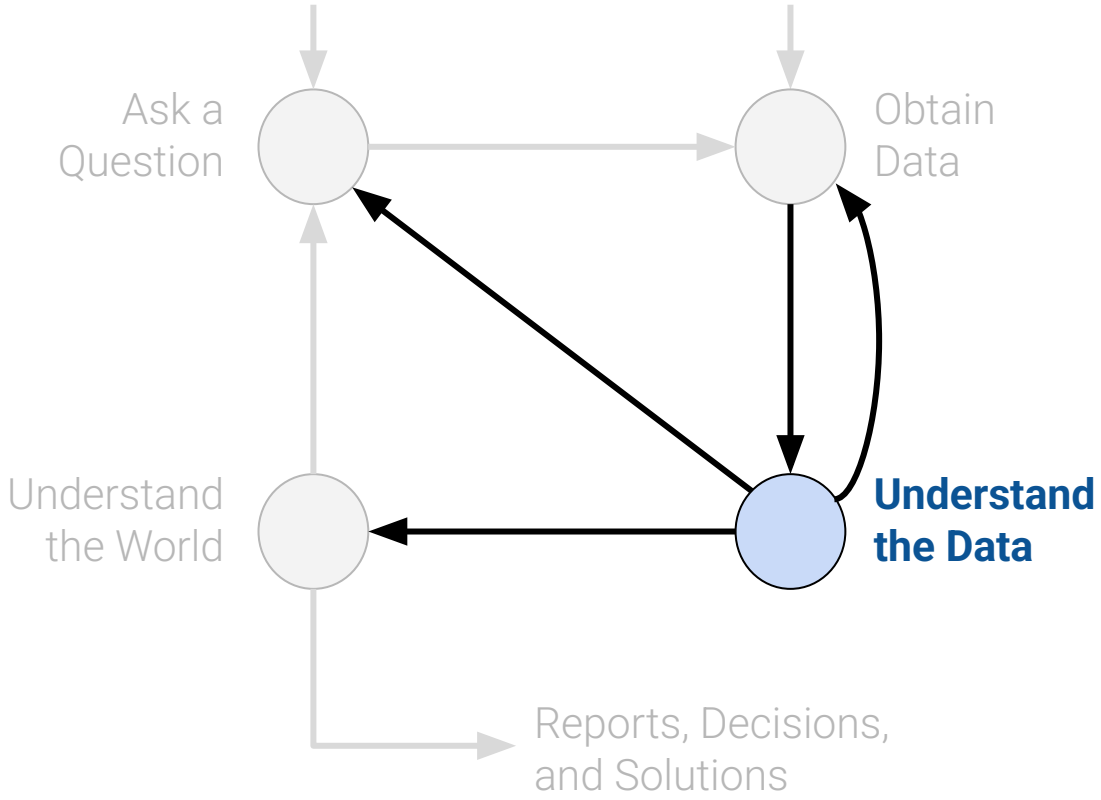
In Data 100, we hand you the data.

In real life, data can take **years** of exploration and negotiation to obtain!

3. Exploratory Data Analysis & Visualization

2868377

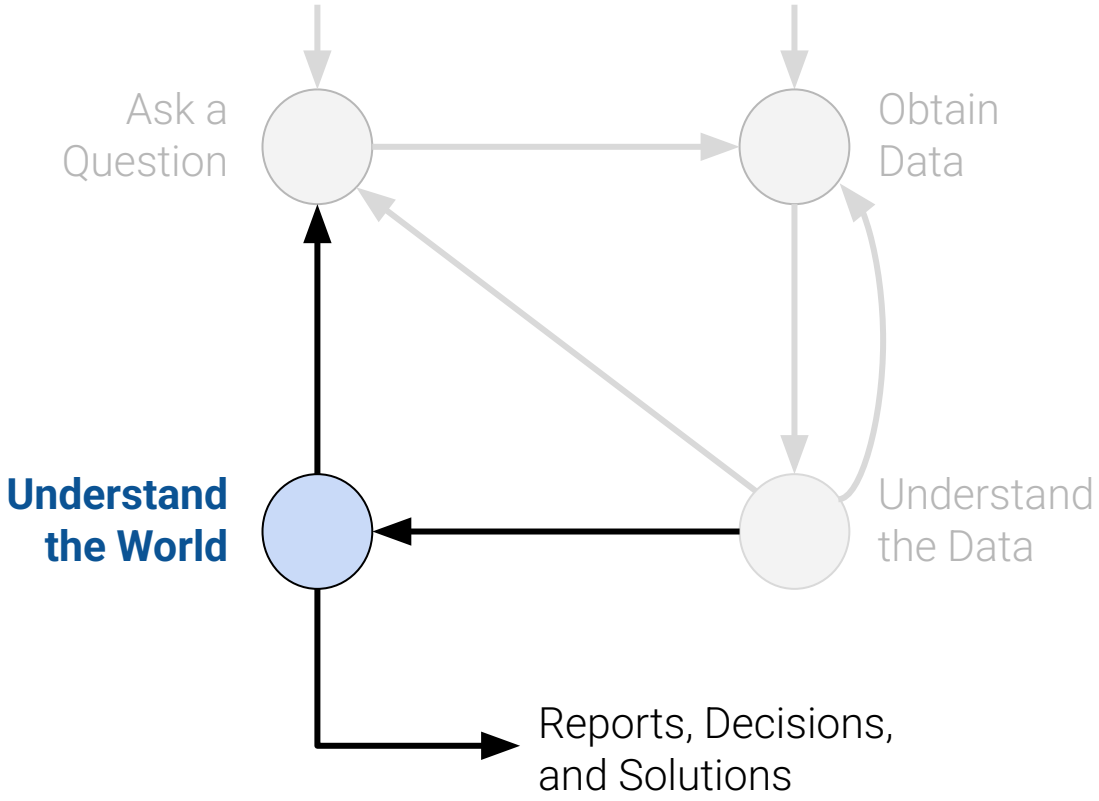
- How is our data organized and what does it contain?
- Do we already have relevant data?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?



4. Prediction and Inference

2868377

- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?



2868377

Start Pandas!
Time permitting

Lecture 01, Data 100 Summer 2025

2868377

Congratulations!!!

You **have collected** or **have been given** a box of data.

What does this "data" actually look like?
How will you work with it?

Data Scientists Love Tabular Data

2868377

"Tabular data" = data in a table.

Typically:

	Year	Candidate	Party	Popular vote	Result	%
0	1824	Andrew Jackson	Democratic-Republican	151271	loss	57.210122
1	1824	John Quincy Adams	Democratic-Republican	113142	win	42.789878
2	1828	Andrew Jackson	Democratic	642806	win	56.203927
3	1828	John Quincy Adams	National Republican	500897	loss	43.796073
4	1832	Andrew Jackson	Democratic	702735	win	54.574789
...	...	...	...	...	...	...
182	2024	Donald Trump	Republican	77303568	win	49.808629
183	2024	Kamala Harris	Democratic	75019230	loss	48.336772
184	2024	Jill Stein	Green	861155	loss	0.554864
185	2024	Robert Kennedy	Independent	756383	loss	0.487357
186	2024	Chase Oliver	Libertarian Party	650130	loss	0.418895

A **row** represents one **observation** (here, a single person running for president in a particular year).

A **column** represents some characteristic, or **feature**, of that observation (here, the political party of that person).

In Data 8, you worked with the `datascience` library using `Tables`.

In Data 100 (and beyond), we'll use an industry-standard library called `pandas`.

Introducing the Standard Python Data Science Tool: pandas

2868377

The Python Data Analysis Library

Stands for "panel data"

The Data 100 logo

a cartoon panda

DataFrames

2868377

In the "language" of pandas, we call a table a **`DataFrame`**.

We think of **`DataFrame`**s as collections of named columns, called **`Series`**.

	Year	Candidate	Party	Popular vote	Result	%
0	1824	Andrew Jackson	Democratic-Republican	151271	loss	57.210122
1	1824	John Quincy Adams	Democratic-Republican	113142	win	42.789878
2	1828	Andrew Jackson	Democratic	642806	win	56.203927
3	1828	John Quincy Adams	National Republican	500897	loss	43.796073
4	1832	Andrew Jackson	Democratic	702735	win	54.574789
...	...	...	...	...	...	...
182	2024	Donald Trump	Republican	77303568	win	49.808629
183	2024	Kamala Harris	Democratic	75019230	loss	48.336772
184	2024	Jill Stein	Green	861155	loss	0.554864
185	2024	Robert Kennedy	Independent	756383	loss	0.487357
186	2024	Chase Oliver	Libertarian Party	650130	loss	0.418895

A `DataFrame`

```
0          Andrew Jackson
1        John Quincy Adams
2          Andrew Jackson
3        John Quincy Adams
4          Andrew Jackson
              ...
182          Donald Trump
183         Kamala Harris
184            Jill Stein
185        Robert Kennedy
186          Chase Oliver
Name: Candidate, Length: 187, dtype: object
```

A `Series` named "Candidate"

Series

2868377

A **`Series`** is a 1-dimensional array-like object. It contains:

- A sequence of **`values`** of the same type.
- A sequence of data labels, called the **`index`**.

`pd` is the conventional alias for **`pandas`**

```
import pandas as pd
s = pd.Series(["welcome", "to", "data 100"])
```

```
0      welcome
1           to
2     data 100
dtype: object
```

`Index`, accessed by calling `s.index`

```
RangeIndex(start=0, stop=3, step=1)
```

`Values`, accessed by calling `s.values`

```
array(['welcome', 'to', 'data 100'], dtype=object)
```

Constructing a `Series`

- We can provide index labels for items in a **`Series`** by passing an index list.

```
s = pd.Series([-1, 10, 2], index = ["a", "b", "c"])
```

```
a    -1
b    10
c     2
dtype: int64
```

```
s.index
```

```
Index(['a', 'b', 'c'], dtype='object')
```

- A **`Series`** index can also be changed.

```
s.index = ["first", "second", "third"]
```

```
first     -1
second    10
third      2
dtype: int64
```

```
s.index
```

```
Index(['first', 'second', 'third'], dtype='object')
```

Selection in `Series`

2868377

- We can select a single value or a set of values in a **`Series`** using:
  - A single label
  - A list of labels
  - A filtering condition

```
s = pd.Series([4, -2, 0, 6], index = ["a", "b", "c", "d"])
```

```
a    4
b   -2
c    0
d    6
dtype: int64
```

Selection in `Series`

- We can select a single value or a set of values in a **`Series`** using:
  - **A single label**
  - A list of labels
  - A filtering condition

```
s = pd.Series([4, -2, 0, 6], index = ["a", "b", "c", "d"])
```

```
a    4
b   -2
c    0
d    6
dtype: int64
```

```
s["a"]
```

```
4
```

A single string label

A single value

Selection in `Series`

- We can select a single value or a set of values in a **`Series`** using:
  - A single label
  - **A list of labels**
  - A filtering condition

```
s = pd.Series([4, -2, 0, 6], index = ["a", "b", "c", "d"])
```

```
a    4
b   -2
c    0
d    6
dtype: int64
```

```
s[["a", "c"]]
```

```
a    4
c    0
dtype: int64
```

A list of string labels

A `Series`

Selection in `Series`

2868377

- We can select a single value or a set of values in a **`Series`** using:
  - A single label
  - A list of labels
  - **A filtering condition**

```
s = pd.Series([4, -2, 0, 6], index = ["a", "b", "c", "d"])
```

```
a    4
b   -2
c    0
d    6
dtype: int64
```

How to select values in the **`Series`** that satisfy a condition:

1) Apply a boolean condition to the **`Series`**, creating a **new boolean `Series`** (often called a **"boolean mask"**).
2) Index into our original **`Series`** using the boolean mask. **`pandas`** selects only the entries in the **`Series`** that satisfy the condition.

```
s > 0
```

```
a     True
b    False
c    False
d     True
dtype: bool
```

```
s[s > 0]
```

Boolean mask

```
a    4
d    6
dtype: int64
```

DataFrames of Series!

2868377

In Data 100, we primarily think of **`Series`** as columns in a **`DataFrame`**.

We can think of a **`DataFrame`** as a collection of **`Series`** that all share the same **`Index`**.

```
0       1824
1       1824
2       1828
3       1828
4       1832
       ...
182     2024
183     2024
184     2024
185     2024
186     2024
Name: Year,
```

The Series **`"Year"`**

```
0          Andrew Jackson
1       John Quincy Adams
2          Andrew Jackson
3       John Quincy Adams
4          Andrew Jackson
              ...
182          Donald Trump
183         Kamala Harris
184             Jill Stein
185         Robert Kennedy
186          Chase Oliver
Name: Candidate,
```

The Series **`"Candidate"`**

[...]

	Year	Candidate	Party	Popular vote	Result	%
0	1824	Andrew Jackson	Democratic-Republican	151271	loss	57.210122
1	1824	John Quincy Adams	Democratic-Republican	113142	win	42.789878
2	1828	Andrew Jackson	Democratic	642806	win	56.203927
3	1828	John Quincy Adams	National Republican	500897	loss	43.796073
4	1832	Andrew Jackson	Democratic	702735	win	54.574789
...	...	...	...	...	...	...
182	2024	Donald Trump	Republican	77303568	win	49.808629
183	2024	Kamala Harris	Democratic	75019230	loss	48.336772
184	2024	Jill Stein	Green	861155	loss	0.554864
185	2024	Robert Kennedy	Independent	756383	loss	0.487357
186	2024	Chase Oliver	Libertarian Party	650130	loss	0.418895

The DataFrame **`elections`**

Non-native English speaker note: The plural of "series" is "series". Sorry.

2868377

LECTURE 1

Course Overview

Content credit: Acknowledgments