Two Things…

3625952

Pre-semester Survey

- Take a minute to fill out the pre-semester survey **now** if you haven't already!
  - We need these responses to assign discussion sections.

Spotify Playlist

- We created this collaborative Spotify playlist to be played before class. Share your favorite songs with the class!

CC BY NC SA

⚠️ Reminder to start the Zoom recording!

3625952

LECTURE 1

Course Overview

An overview of data science, Data 100/200, and the data science lifecycle.

Data 100/Data 200, Spring 2025 @ UC Berkeley

Narges Norouzi and Josh Grossman

CC BY NC SA

Two Things…

3625952

Pre-semester Survey

- Take a minute to fill out the pre-semester survey **now** if you haven't already!
  - We need these responses to assign discussion sections.

Spotify Playlist

- We created this collaborative Spotify playlist to be played before class. Share your favorite songs with the class!

CC BY NC SA

3625952

slido

Join at slido.com
#3625952

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

CC BY NC SA

slido

3625952

What emoji best describes your mood today?

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

CC BY NC SA

Slido Q&A

3625952

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

This is the best place to clarify course content.

CC BY NC SA

3625952

Roadmap

Lecture 01, Data 100 Spring 2025

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

CC BY NC SA

Intro - Narges Norouzi

3625952

Pronunciation of my name

- **Ph.D.** from **University of Toronto** (2017) **in CE** from Artificial Perception Lab.

- **Research:** ML with applications in biomedical and educational domain
  - Bioelectronic closed loop wound healing
  - Intelligent AI tutoring as classroom assistant
  - Find out more at: https://nargesnorouzi.me/

- **Teaching Professor** at **UC Santa Cruz** (2017-2022).
- **Teaching Professor** at **UC Berkeley** (2022-now).
- **Fifth time teaching this class!**
- This semester, I'm:
  - Leading a research group focused on AI education.
  - Teaching two research seminar courses as well as the DS honors course in addition to Data100/200.

CC BY NC SA

Intro - Josh Grossman

3625952

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

Other courses: *Stat 131A (fa24), MS&E 125 and 298 @ Stanford*

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 20-minute chat!

CC BY NC SA

3625952

What is Data Science?

Lecture 01, Data 100 Spring 2025

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

CC BY NC SA

What Is Data 100?

3625952

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

?

CC BY NC SA

3625952

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)

CC BY NC SA

3625952

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?

CC BY NC SA

Initial Research Question

3625952

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.

CC BY NC SA

Calculating Audit Rates Across Race

3625952

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
Home address (number and street). If you have a P.O. box, see instructions.: 789 Tuxedo Drive
Apt. no.
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?

CC BY NC SA

Calculating Audit Rates Across Race

3625952

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?

CC BY NC SA

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

3625952

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)

CC BY NC SA

Inferring Race from Hometown

3625952

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.

CC BY NC SA

Inferring Race from First Name, Last Name, and Neighborhood

3625952

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.

CC BY NC SA

Visualizing the Data

3625952

CC BY NC SA

Black Taxpayers 3-5x More Likely to be Audited

3625952

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.

CC BY NC SA

Visualizing the Data, with Income

3625952

CC BY NC SA

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

3625952

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.

CC BY NC SA

Identifying a Potential Root Cause

3625952

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.

CC BY NC SA

Identifying a Potential Root Cause

3625952

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

Income underreporting error
Up to $5,000 recovered

Tax credit error
Up to $500 recovered

CC BY NC SA

Identifying a Potential Root Cause

3625952

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)!

CC BY NC SA

New York Times, September 2023

3625952

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)!

CC BY NC SA

What Is Data 100?

3625952

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

?

CC BY NC SA

What Is Data Science?

3625952

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)

CC BY NC SA

Transitioning From Data 8 to Data 100

3625952

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**).

CC BY NC SA

What Do Industry Data Scientists Do All Day?

3625952

The tasks that data scientists say they work on regularly.

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

CC BY NC SA

3625952

What Will You Learn in This Class?

Lecture 01, Data 100 Spring 2025

- Intros
- What is data science?
- **What will you learn in this class?**
- Course overview
  - Lots of important details
- Data Science Lifecycle
- Demo

CC BY NC SA

What Is Data 100?

3625952

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

BY NC SA

Course Goals

3625952

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**.

CC BY NC SA

Tentative List of Topics to be Covered in Data 100

3625952

- 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
- Data science in the physical world
- Logistic Regression
- Clustering
- PCA

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

CC BY NC SA

3625952

Course Overview

Lecture 01, Data 100 Spring 2025

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

CC BY NC SA

3625952

Prerequisites

BY NC SA

Prerequisites

3625952

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.

CC BY NC SA

3625952

Staff

BY NC SA

Leads

3625952

Head TA

Sam Bobick

Shreya Gupta

Meenakshi Mittal

Xiaorui Liu

Rachel Xin

Dan Nguyen

Rose Niousha

Sarah Kiefer

Sammie Smith

Malavikha Sudarshan

Leads run one aspect of the course (examples include: infrastructure, student support, logistics, etc.). Contact info: ds100.org/sp25/staff/.

CC BY NC SA

UCS2s

3625952

Emrie Loh

Dun-Ming Huang

Conan Smallwood

Azalea Bailey

James Geronimo

Sarika Pasumarthy

Jake Pastoria

UCS2s teach discussions and assist in a wide variety of tasks (examples include: developing content, exam prep, etc.). Contact info: ds100.org/sp25/staff/.

CC BY NC SA

UCS1s

3625952

Cristina Prieto

Kelly Hu

Skyla Ma

Steven Luo

William Lee

Willy Guan

Wolfgang Edholm

Kathryn Sun

Ella Hammond

Anshul Jambula

Terry Kim

Sardaana Eginova

Noah Abji

Neil Chen

Dhruv Pendharkar

Vicky Huang

UCS1s help review and grade our assignments, answer lecture and Ed questions, host office hours, and get trained for future semesters! Contact info: ds100.org/sp25/staff/.

CC BY NC SA

UCS1s

3625952

Collin Duong

Rohan Bijukumar

Alex Huang

Kevin Hu

Henry Hong

Haichao Li

Gisella Chan

Tyler Pham

Aiden Wang

Simon Kang

Surya Ramkumar

Satleen Gill

Alex Proshkin

UCS1s help review and grade our assignments, answer lecture and Ed questions, host office hours, and get trained for future semesters! Contact info: ds100.org/sp25/staff/.

CC BY NC SA

3625952

Course Logistics
Content and Workflow

BY NC SA

Syllabus Walkthrough

3625952

ds100.org/sp25/syllabus/

CC BY NC SA

3625952

Course Websites / Platforms

BY NC SA

Online Platforms

3625952

Course website (ds100.org/sp25/)

- Where all lectures, assignments, and discussions are posted.

DataHub (data100.datahub.berkeley.edu)

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

Ed (https://edstem.org/us/courses/74189/discussion)

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

Gradescope (https://www.gradescope.com/courses/949942)

- Where all assignments are submitted, and where all of your grades in this course will live.

Lecture Notes (https://ds100.org/course-notes/)

- A summary of the highlights of topics covered in each lecture.

Textbook (www.textbook.ds100.org)

- Supplemental reading (not 100% synchronized with the class schedule).

CC BY NC SA

Programming Environment for our Course: JupyterLab

3625952

File Edit View Run Kernel Tabs Settings Help

Files | Running | Commands | Cell Tools | Tabs

notebooks > transit-zurich

Name	Last Modified
transit.ipynb	2 minutes ago
passenger.csv	2 hours ago
routes.json	2 hours ago
stops.json	2 hours ago

transit.ipynb — Code — Python 3

We plot the number of passengers at the Rosengartenstrasse stop.

In [93]:
```
load = df[df.stopNameShort=='ROSE'].passengerLoadStop
sns.distplot(load, kde=False)
plt.axvline(load.median())
plt.title('Passenger Load at Rosengartenstrasse stop')
plt.xlabel('Number of passengers');plt.ylabel('Frequency');
```

Compare the median load at this stop with the medians of all stops.

In [94]:
```
sns.distplot(df.groupby('stopNameShort')
             .passengerLoadStop.median(), kde=False)
plt.axvline(load.median())
plt.title('Passenger load medians across all stops');
plt.xlabel('Median passenger load')
plt.ylabel('Frequency');
```

routes.json

stops.json | routes.json

```
▾ 564: {} 3 keys
    type: "Feature"
  ▾ properties: {} 4 keys
      stopId: 2749
      stopNumber: 2104
      stopNameShort: "ROSE"
      stopName: "Zürich, Rosengartenstrasse"
  ▸ geometry: {} 2 keys
```

passenger.csv

Delimiter: ,

stopSequer	stopId	stopNameShort	stopName
5	2104	ROSE	Zürich, Rosengartenstra
6	564	BUCH	Zürich, Bucheggplatz
7	2017	RADI	Zürich, Radiostudio
8	498	BIRD	Zürich, Birchdörfli
9	1705	NEUA	Zürich, Neuaffoltern
10	1000	GLAU	Zürich, Glaubtenstrasse
11	767	EINF	Zürich, Einfangstrasse

CC BY NC SA

Learning Advanced JupyterLab

3625952

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:

- A 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.

CC BY NC SA

3625952

Data Science Lifecycle

Lecture 01, Data 100 Spring 2025

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

CC BY NC SA

Data Science Lifecycle

3625952

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

CC BY NC SA

Data Science Lifecycle

3625952

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

Note the two distinct entry points!

CC BY NC SA

1. Question/Problem Formulation

3625952

- 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?

CC BY NC SA

2. Data Acquisition and Cleaning

3625952

- 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?

CC BY NC SA

3. Exploratory Data Analysis & Visualization

3625952

- 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?

CC BY NC SA

4. Prediction and Inference

3625952

- 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?

CC BY NC SA

slido

3625952

Why do you think the data science lifecycle is iterative?

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

CC BY NC SA

3625952

Demo: The Data Science Lifecycle

Lecture 01, Data 100 Spring 2025

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

Available on the course website:
https://ds100.org/sp25/lecture/lec01

CC BY NC SA

Demo Slides

Ask a Question: Who are you?

3625952

CC BY NC SA

Demo Slides

3625952

CC BY NC SA

Demo Slides

CC BY NC SA

3625952

Let's understand what our data tells us, and let's clean the data while we're at it.

Demo Slides

EDA and Visualization

3625952

Population: Data 100 students, Spring 2025

Some sub-questions:

✅ How many students are in the class?
✅ What are your majors?
✅ What year are you?
4. What is the ratio between grad & undergrad students in Data 100, and its comparison with campus distribution over time?

CC BY NC SA

Demo Slides

Again, but for Campus Data

3625952

What does each row/column represent?

CC BY NC SA

Demo Slides

What's the Point of This Demo?

3625952

We make assumptions in data science

- Is the data representative:
  - Of the question being asked
  - Of the world and its implications
- Beliefs/backgrounds of data collectors
- Beliefs/backgrounds of data analysts
- Beliefs/backgrounds of the population

Data Science does not and cannot live in a theoretical vacuum. **Data Science is a human-centered technical practice.**

CC BY NC SA

Two Things…

3625952

Pre-semester Survey

- Take a minute to fill out the pre-semester survey **now** if you haven't already!
  - We need these responses to assign discussion sections.

Spotify Playlist

- We created this collaborative Spotify playlist to be played before class. Share your favorite songs with the class!

CC BY NC SA

3625952

See you soon!

BY NC SA

3625952

LECTURE 1

Course Overview

Content credit: Acknowledgments

CC BY NC SA