Two Things…

8705986

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!

8705986

LECTURE 1

Course Overview

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

Data 100/Data 200, Fall 2024 @ UC Berkeley

Narges Norouzi and Joseph E. Gonzalez

slido

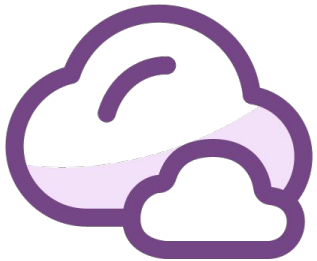
8705986

Join at slido.com
#8705986

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

slido

8705986



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

8705986

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

This is the best place to clarify course content.

8705986

Roadmap

Lecture 01, Data 100 Fall 2024

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

Intro - Narges Norouzi

8705986

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

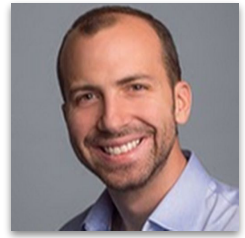
Pronunciation of my name

- **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).
- Fourth time teaching this class!
  - Previously taught graduate and undergraduate Artificial Intelligence, Machine Learning, Deep Learning, and Algorithms courses.
- This semester, I'm:
  - Leading a research group focused on AI education.
  - Teaching two research seminar courses in addition to Data100/200.
  - Starting several granted projects.

Intros - Joseph E. Gonzalez

8705986



BS in 2006 @ **Caltech** → **PhD in 2012** @ **Carnegie Mellon University**

- **Thesis**: *Parallel and distributed inference in probabilistic graphical models*

Associate Prof. in EECS: co-director of **RISE**, **Sky Lab** and **LM-Sys Org** also in **BAIR**

Research "Focus" in AI Systems

- Training and Inference Systems (e.g., **Clipper**, **vLLM**, **Alpa**)
- Developing new LLMs (e.g., **Gorilla LLM**, **MemGPT**, and **Vicuña**)
- Evaluating LLMs (e.g., **Chatbot Arena** and **LLM-as-a-judge**)
- Visual Language Reasoning (**Lisa's** Projects)
- Robotics and Autonomous Driving Systems (**ERDOS**)
- Large-Scale Data Processing Systems (e.g., **Apache Spark**, **GraphLab**, **Ray**)

Entrepreneurial Activities

- Co-founder at **Turi** and **RunLLM** and advisor at **Genmo.ai** and **MemGPT.ai**

Teaching Focus: ***Data100*** *(v0, v1, v2, …,* Sp24, Fa24) and **Grad. AI-Systems** (FA24)

Life Focus: I have two little kids and most of my free time is spent with family.

8705986

What is Data Science?

Lecture 01, Data 100 Fall 2024

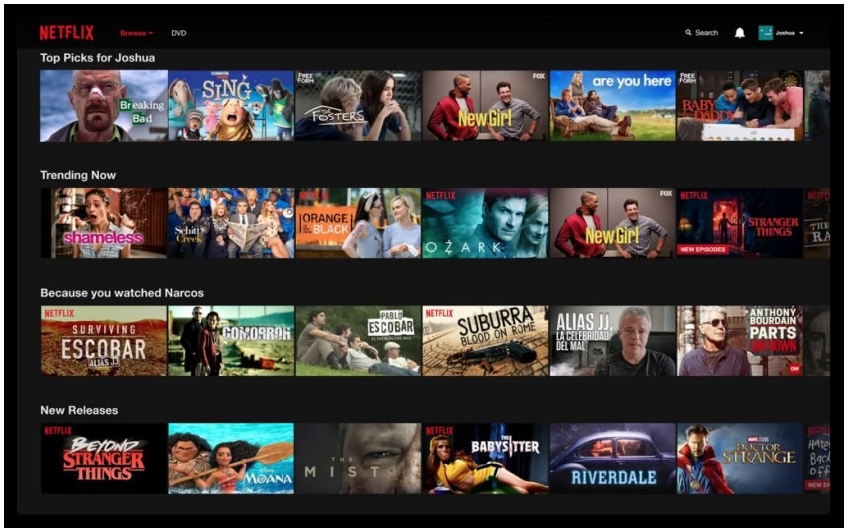
- Intros
- **What is data science?**
- What will you learn in this class?
- Course overview
- Data Science Lifecycle
- Demo

8705986

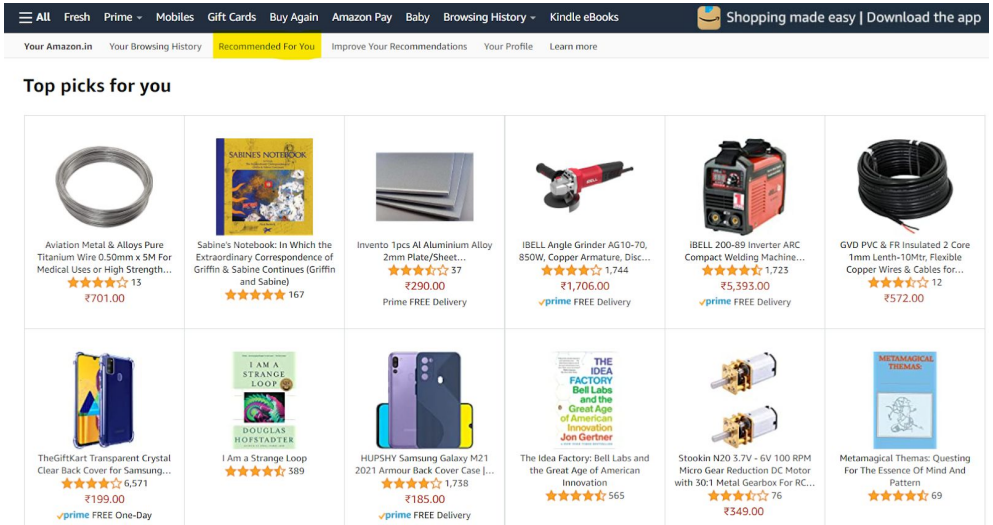
Some examples of Data Science?

Recommendation Systems

8705986



amazon



8705986

DJ Patil calls data scientists ‘a new kind of first responder’

By Rachel Leven | April 13, 2023

https://www.youtube.com/watch?v=LiHMrn2AHpw

DJ Patil spoke to UC Berkeley data science students on April 10. (Photo/

On March 14, 2020, the United States was on the brink of a pandemic. Covid-19 had killed at least 60 people and two cruise ships with ill passengers were set to dock in San Francisco. That’s when DJ Patil received a call: How can data help California combat this?

So the former White House chief data scientist put together a plan. His team acquired hospital and community data, developed surveys, models, dashboards and data catalogs, and used those tools and insights to inform public officials across the state and the country.

“All of this came together in this effort to really take on Covid,” said Patil, who is now a general partner at GreatPoint Ventures, at an April 10 UC Berkeley

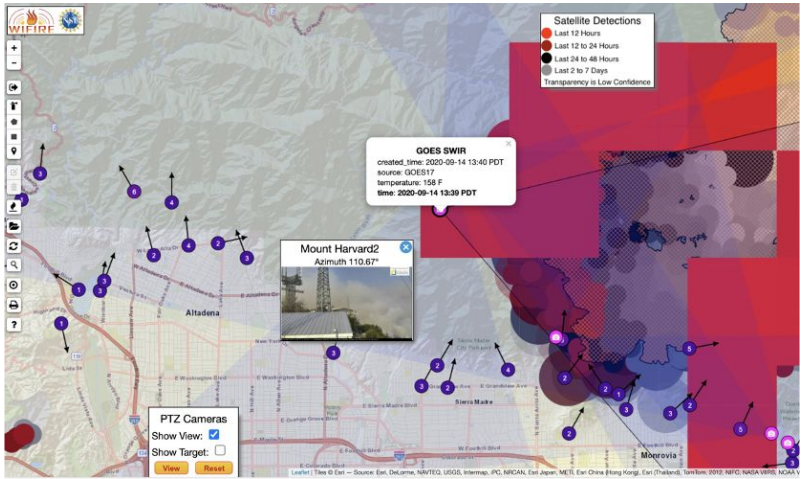
First U.S. Chief Data Scientist
(Obama Adm.)

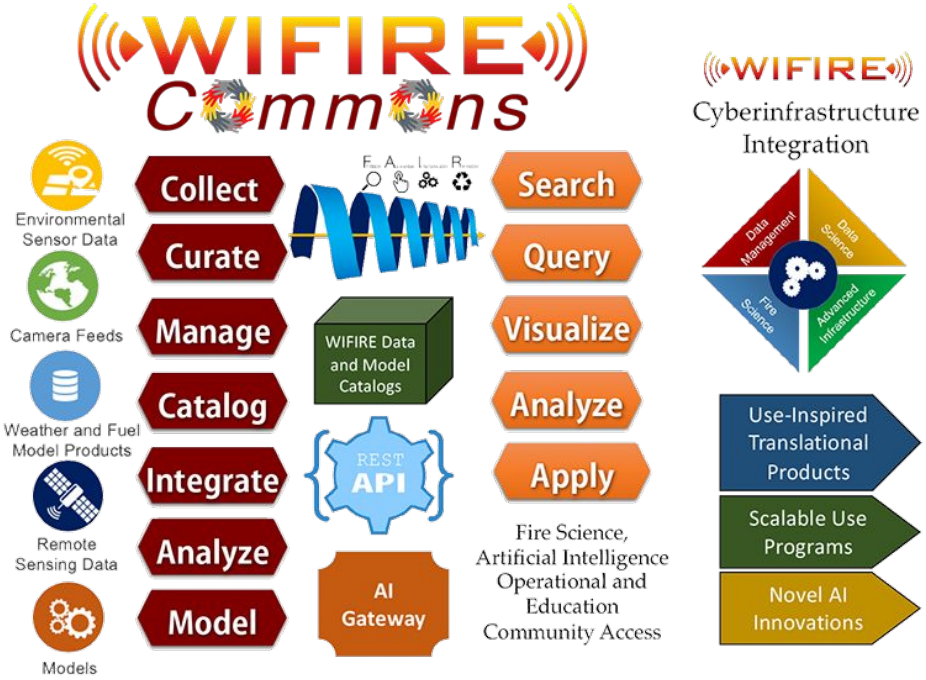
https://www.sciencefriday.com/segments/an-exit-interview-with-u-s-chief-data-scientist-dj-patil/

WIFIRE (UCSD) - Wildfire modeling and management

8705986

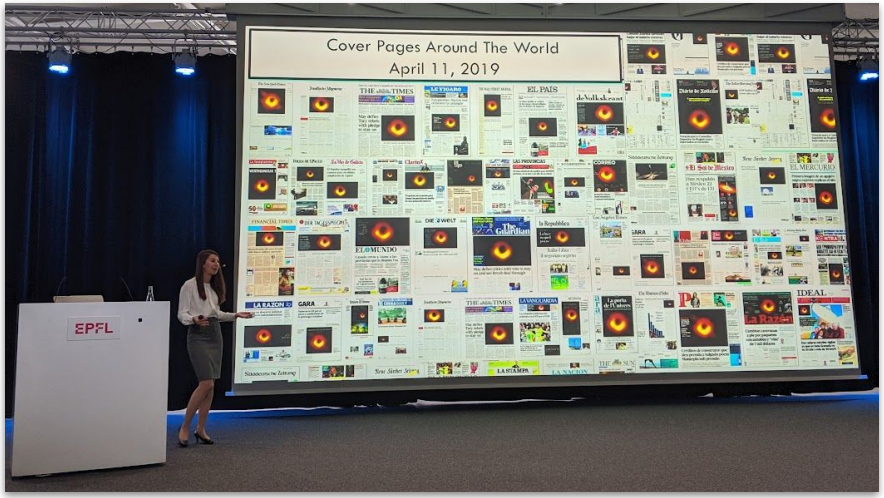
https://wifire.ucsd.edu





2019: First Image of a Black Hole

8705986



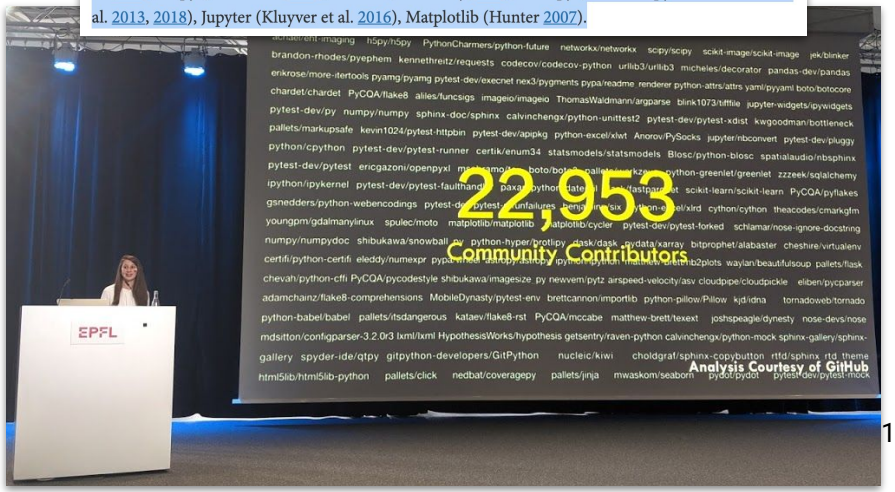
THE ASTROPHYSICAL JOURNAL LETTERS

First M87 Event Horizon Telescope Results. III. Data Processing and Calibration

The Event Horizon Telescope Collaboration, Kazunori Akiyama[1,2,3,4], Antxon Alberdi[5], Walter Alef[6], Keiichi Asada[7], Rebecca Azulay[8,9,6], Anne-Kathrin Baczko[6], David Ball[10], Mislav Baloković[4,11], John Barrett[2] +Show full author list

Published 2019 April 10 • © 2019. The American Astronomical Society.
The Astrophysical Journal Letters, Volume 875, Number 1

Software: DiFX (Deller et al. 2011), CALC, PolConvert (Martí-Vidal et al. 2016), HOPS (Whitney et al. 2004), CASA (McMullin et al. 2007), AIPS (Greisen 2003), ParselTongue (Kettenis et al. 2006), GNU Parallel (Tange 2011), GILDAS, eht-imaging (Chael et al. 2016, 2018), Numpy (van der Walt et al. 2011), Scipy (Jones et al. 2001), Pandas (McKinney 2010), Astropy (The Astropy Collaboration et al. 2013, 2018), Jupyter (Kluyver et al. 2016), Matplotlib (Hunter 2007).



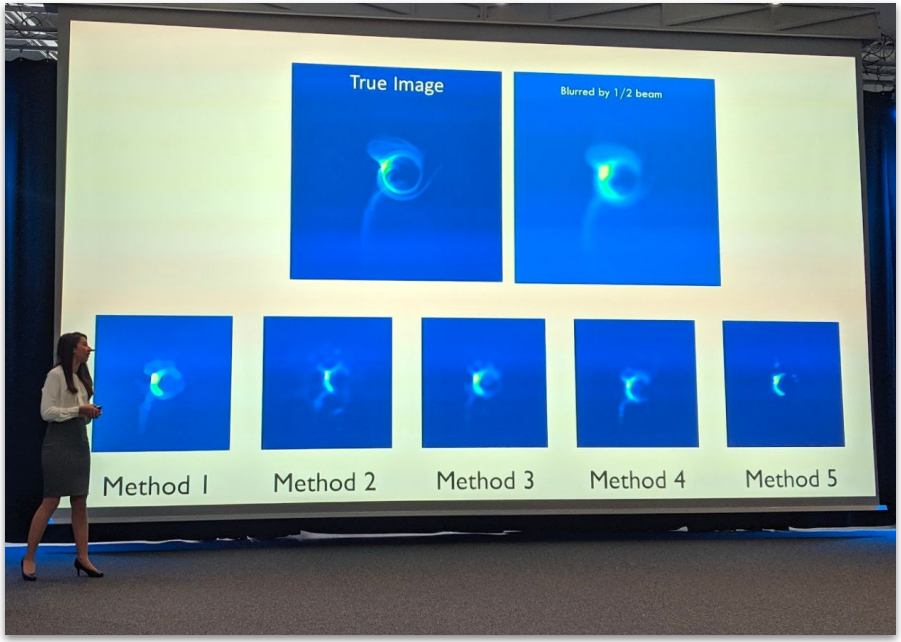
Katie Bouman
MIT/Caltech

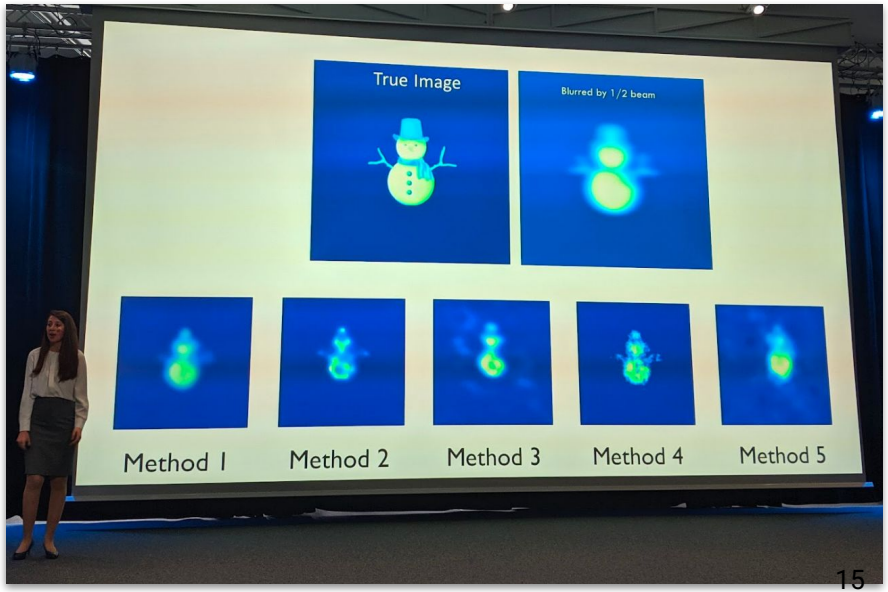
Event Horizon Telescope

Talk Video: https://youtu.be/TSgpliktkwc

2019: First Image of a Black Hole

8705986





8705986

Why Data Science Matters?

Data Science Enhances Critical Thinking

8705986

The world is complicated! Decisions are hard.

Data is used everywhere to answer hard questions and make tough decisions:

- Science
- Medicine
- Social science
- Engineering
- Sports

Claims about data come up in discussing almost any important issue:

- A common line these days: **"the data says"** … but does it really?
- It is usually not easy to tell what the data "says".
- **Empower yourself** to participate in the arguments that shape your life and your society.

8705986

Technology Trends

- 2020s — **AI…?**
- 2010s — Data Industry
  - Collect and sell information
- 2000s — Internet Industry
  - Online retailers and services
- 1990s — Software Industry
  - Sold computer software
- 1980s — Hardware Industry
  - Sold computers

The Darker Side of Data Science?

8705986

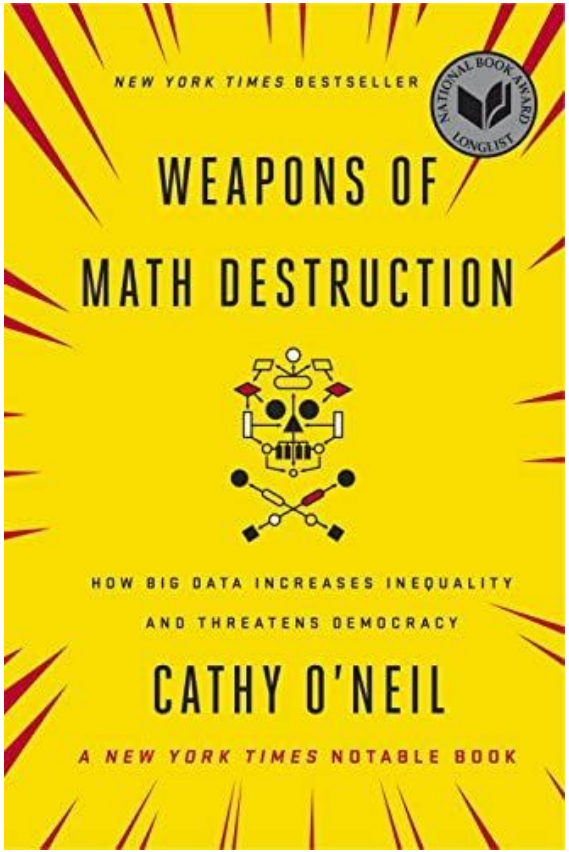
Obscuring complex decisions:

- Mortgage-backed securities → market crash
- Teaching scores & job advancement

Reinforcing historical trends and biases:

- Hiring based on previous hiring data
- Recidivism and racially biased sentencing
- Social media, news, and politics

We will discuss the ethics of data science throughout the class!



NPR author interview
with Cathy O'Neil

But…I Am Optimistic!

8705986

Knowledge is empowering.

Data science offers **immense potential** to address challenging problems facing society.

The future is in your hands, and I believe:

You will use your knowledge for good.

…I am thrilled to teach Data 100 :-)

Data Science Enhances Critical Thinking

8705986

Data science is a fundamentally **human-centered field that facilitates decision-making** by quantitatively balancing tradeoffs.

- To quantify things **reliably** we must:
  - **Find** relevant data;
  - Recognize its **limitations**;
  - Ask the right **questions**;
  - Make reasonable **assumptions**;
  - Conduct an appropriate **analysis**; and
  - **Synthesize and explain** our insights.
- Apply **critical thinking and skepticism** at every step; and
- Consider how our decisions **affect others**.

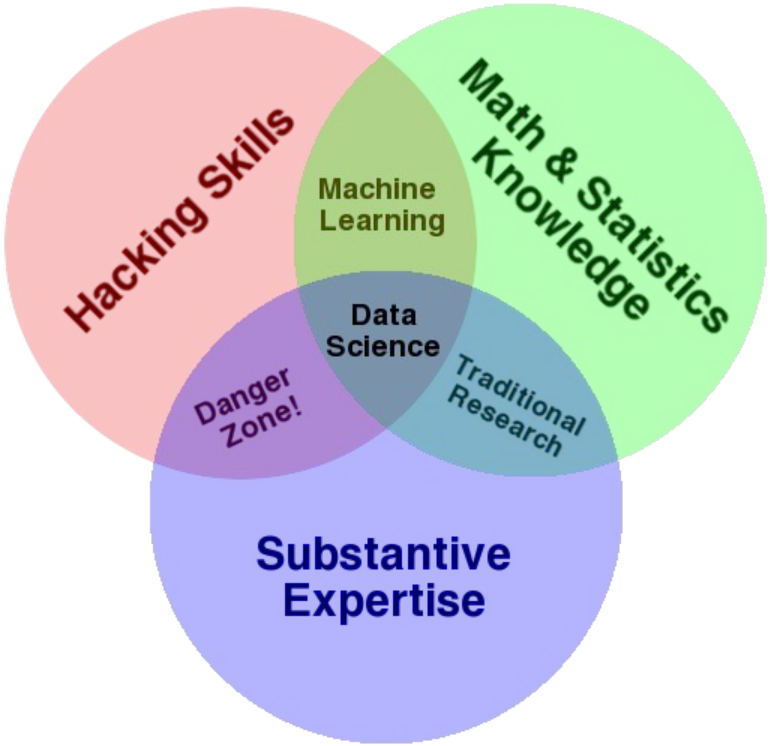
After this course, you should be able to take data and produce useful insights on the world's most challenging and ambiguous problems.

8705986

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

Data Science Venn Diagram

8705986



by Drew Conway in 2010 (link)

Data Science Is a Fundamentally Interdisciplinary Field

8705986

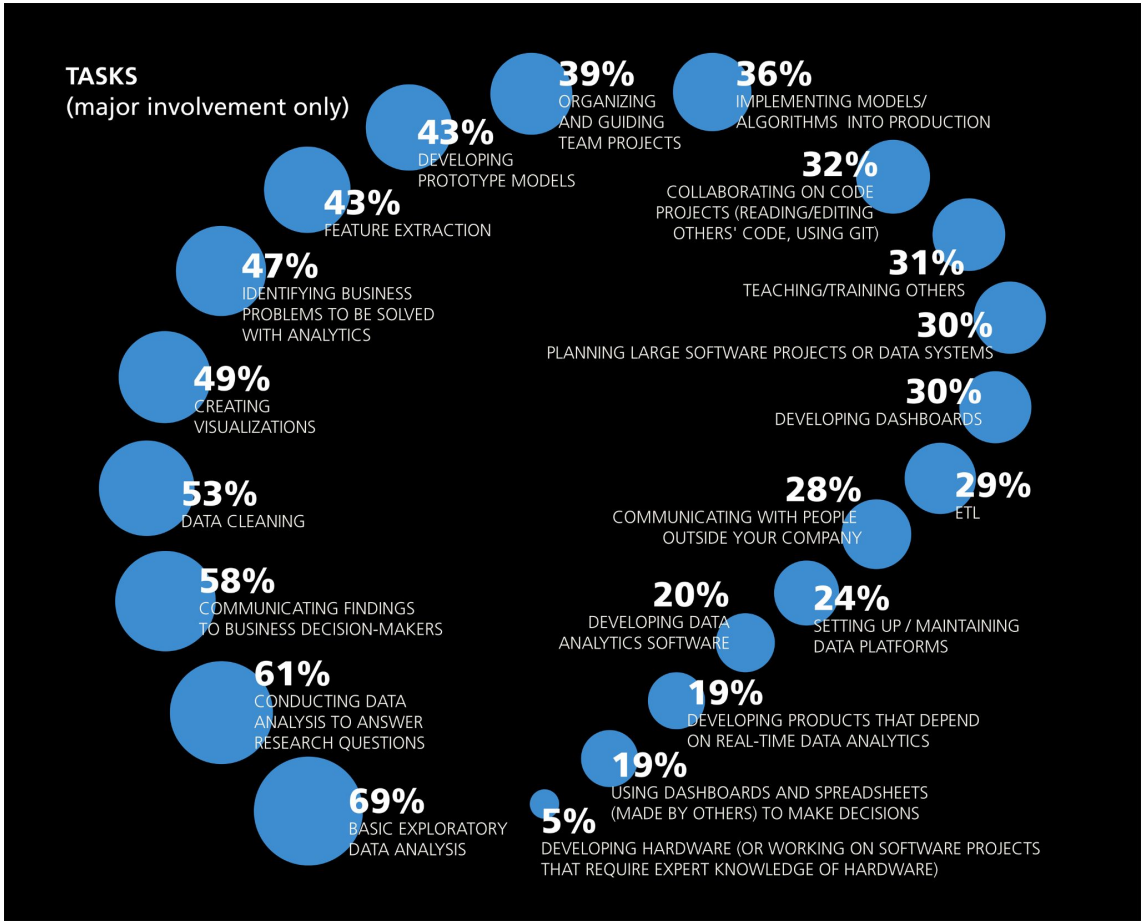
Joey Gonzalez

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

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

Data Science in Industry

8705986



The major tasks that data scientists say they work on regularly.

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

Data Science Requires Engineering and Scientific Insight

8705986

Good data analysis is *not*:

- Simple application of a statistics recipe.
- Simple application of software.

There are many **tools** out there for data science, but they are merely tools.

- **They don't do any of the important thinking!**

"The purpose of computing is insight, not numbers."

R. Hamming. *Numerical Methods for Scientists and Engineers (1962).*

Example Questions in Data Science

8705986

Some (broad) questions we might try to answer with data science:

Business

- What show should we recommend to our user to watch?
- In which markets should we focus our advertising campaign?
- Where should we put docking ports for our bikes?

Government

- What areas of the world are at higher risks for climate change impact in 10 years? 20?
- Is the use of the COMPAS algorithm for prison sentencing fair?
- Do immigrants from poor countries have a positive or negative impact on the economy?

Life

- What is an original (but not too original) name for my daughter?
- What should we eat to avoid dying early of heart disease?
- Should I send my kids to daycare?

8705986

What Will You Learn in This Class?

Lecture 01, Data 100 Fall 2024

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

8705986

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

Course Goals

8705986

Prepare

Prepare students for advanced Berkeley courses in **data management**, **machine learning**, and **statistics**, by providing the necessary foundation and context.

Enable

Enable students to start careers as data scientists by providing experience working with **real-world data, tools, and techniques**.

Empower

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

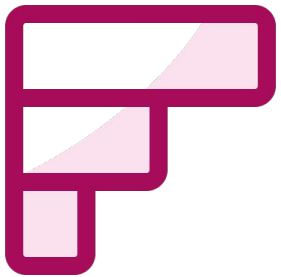
Tentative List of Topics to be Covered in Data 100

8705986

- 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

slido

8705986



Which of these topics excites you the most? Please rank.

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

Prerequisites

8705986

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.

8705986

Course Overview

Lecture 01, Data 100 Fall 2024

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

8705986

Staff

Leads

8705986

Head TAs

Angela Feng

Shreya Gupta

Sarah Kiefer

Aneesh Durai

Malavikha Sudarshan

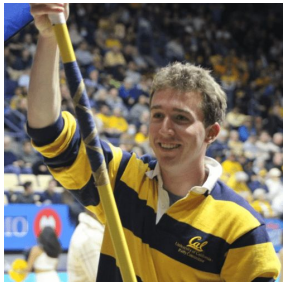
Abby O'Neill

Dan Nguyen

Rose Niousha

Brandon Huang

Prabhleen Kaur



Sam Bobick

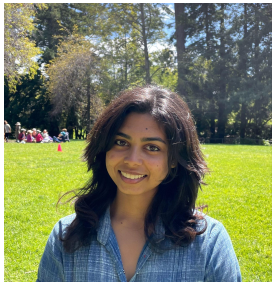
Jessica Lin

Leads run one aspect of the course (examples include: infrastructure, student support, logistics, etc.). Contact info: ds100.org/fa24/staff/.

UCS2s

8705986

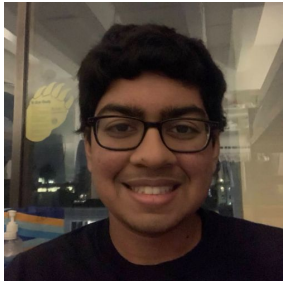
Claire Ding



Minoli Daigavane

Xiaorui Liu

Rayna Bhattacharyya



Nikhil Reddy

Sarika Pasumarthy

UCS2s teach discussions and assist in a wide variety of tasks (examples include: developing content, exam prep, etc.). Contact info: ds100.org/fa24/staff/.

UCS1s

8705986

Meenakshi Mittal

Sammie Smith

Anshul Jambula

Victor Shi

James Geronimo

Willy Guan

Vicky Huang

Brie Zhou

Tyler Zhao

Rohan Bijukumar

Qijun Li

Julianna Lee

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/fa24/staff/.

UCS1s

8705986

Jesse Yao

Nehal Sindhu

Gisella Chan

Yewen Xu

Jake Pastoria

Vladyslav Shevkunov

Ella Hammond

Rishi Khare

Tyler Pham

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/fa24/staff/.

8705986

Course Logistics
Content and Workflow

Syllabus Walkthrough

8705986

ds100.org/fa24/syllabus/

8705986

Course Websites / Platforms

Online Platforms

8705986

Course website (ds100.org/fa24/)

- 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/62812)

- 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/827978)

- 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).

Programming Environment for our Course: JupyterLab

8705986

File Edit View Run Kernel Tabs Settings Help

Files | Running | Commands | Cell Tools | Tabs

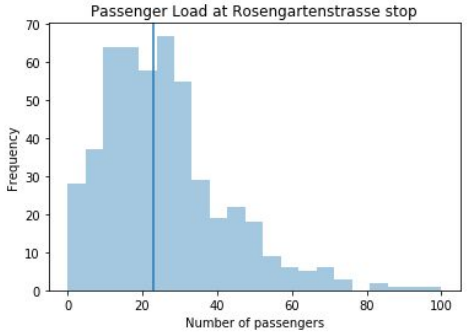
> notebooks > transit-zurich

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

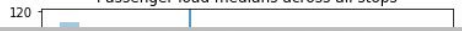
We plot the number of passengers at the Rosengartenstrasse stop.

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

Learning Advanced JupyterLab

8705986

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.

8705986

Data Science Lifecycle

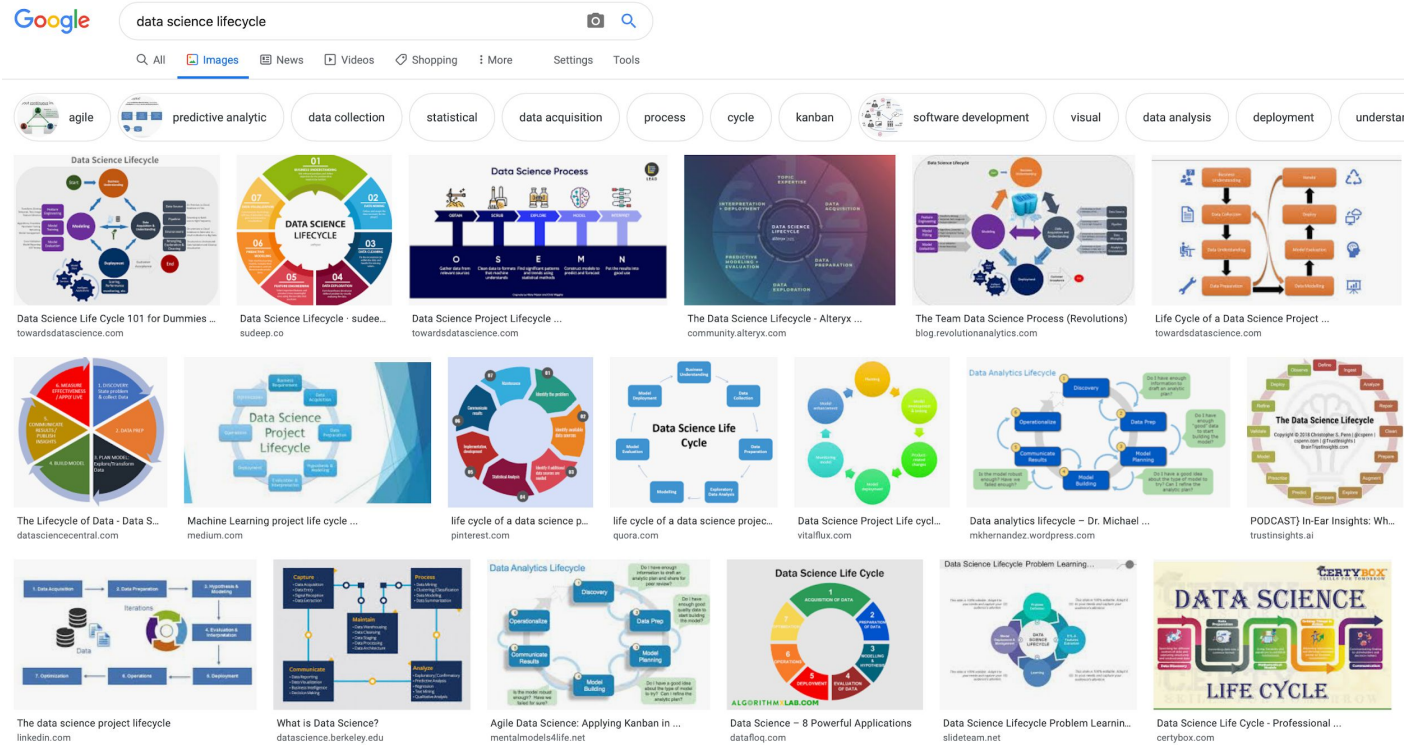
Lecture 01, Data 100 Fall 2024

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

Data Science Lifecycle

8705986

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.

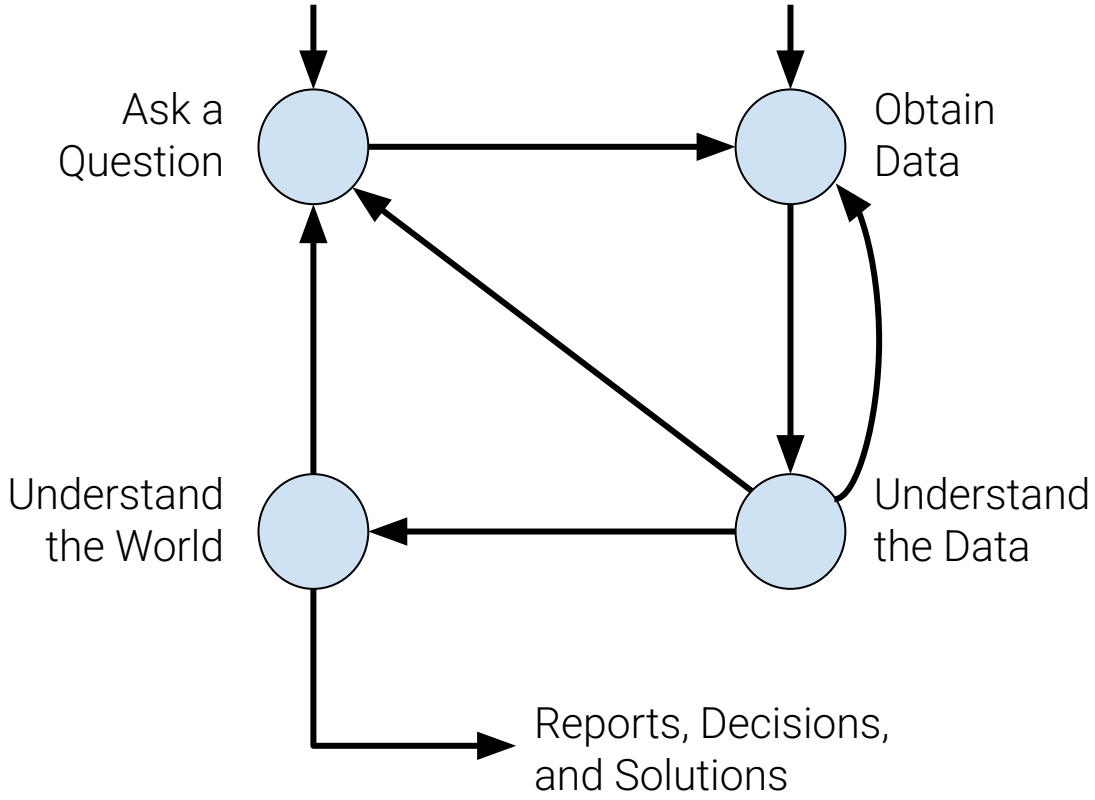


Data Science Lifecycle

8705986

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

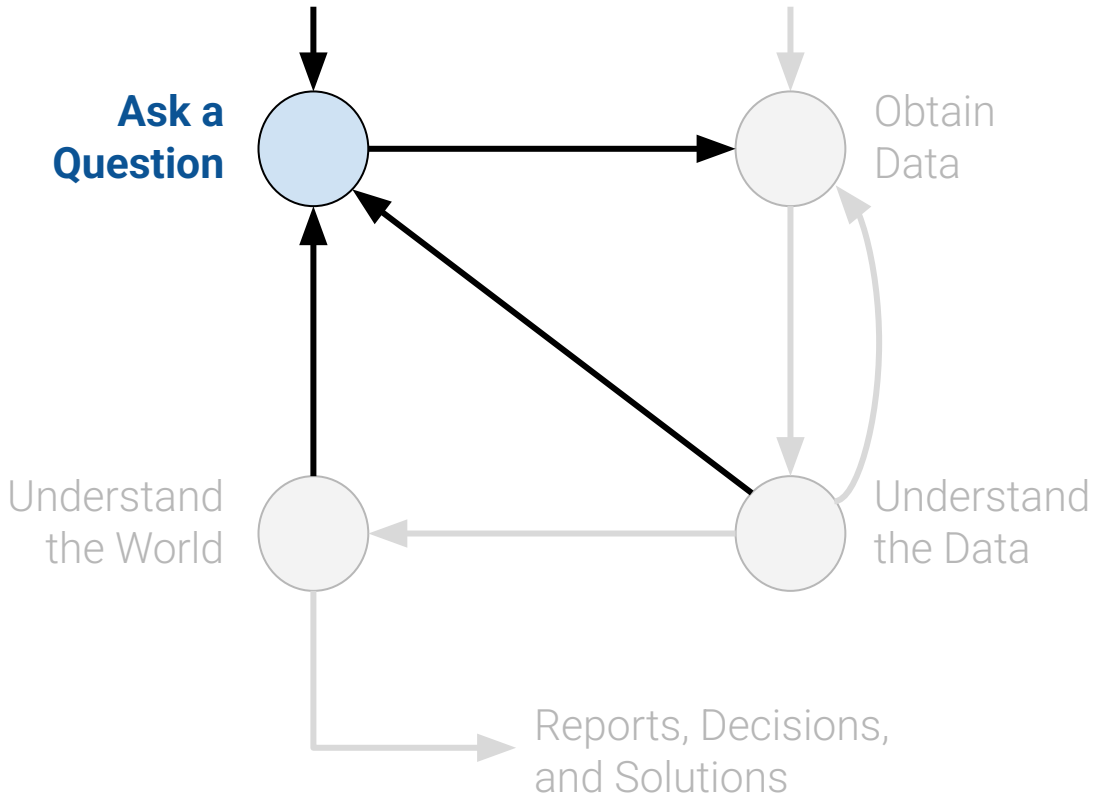
Note the two distinct entry points!



1. Question/Problem Formulation

8705986

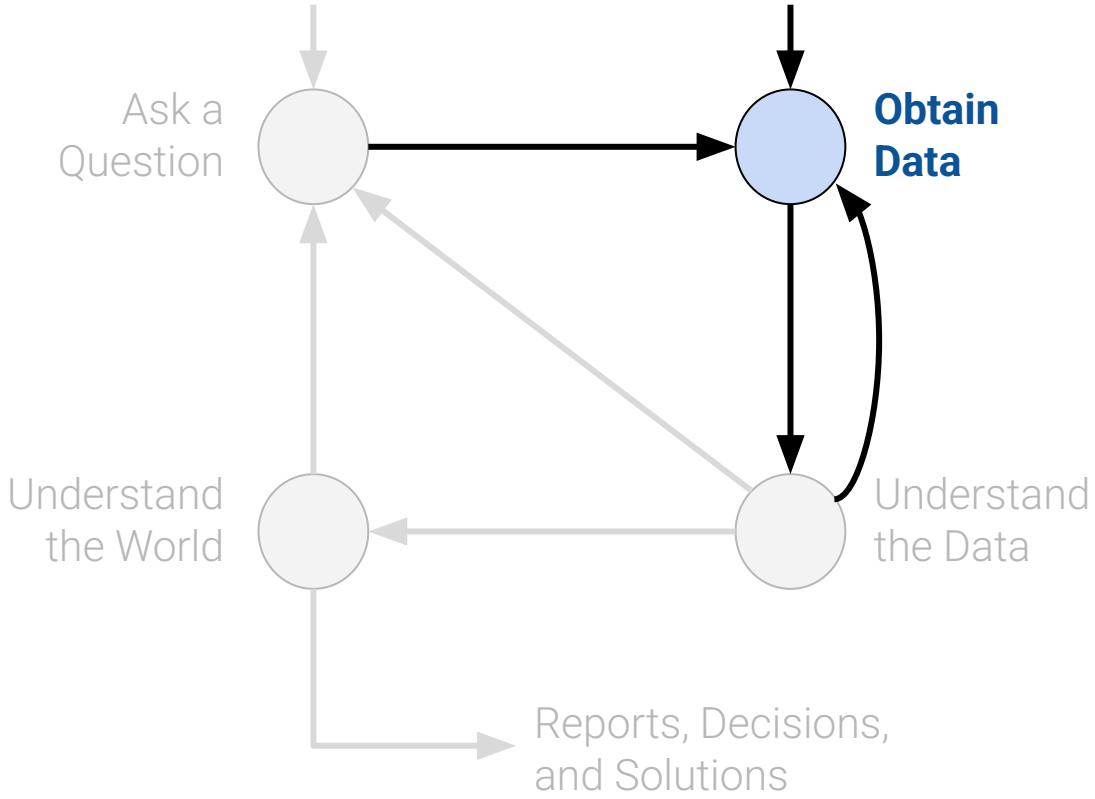
- What do we want to know?
- What problems are we trying to solve?
- What hypotheses do we want to test?
- What are our metrics for success?



2. Data Acquisition and Cleaning

8705986

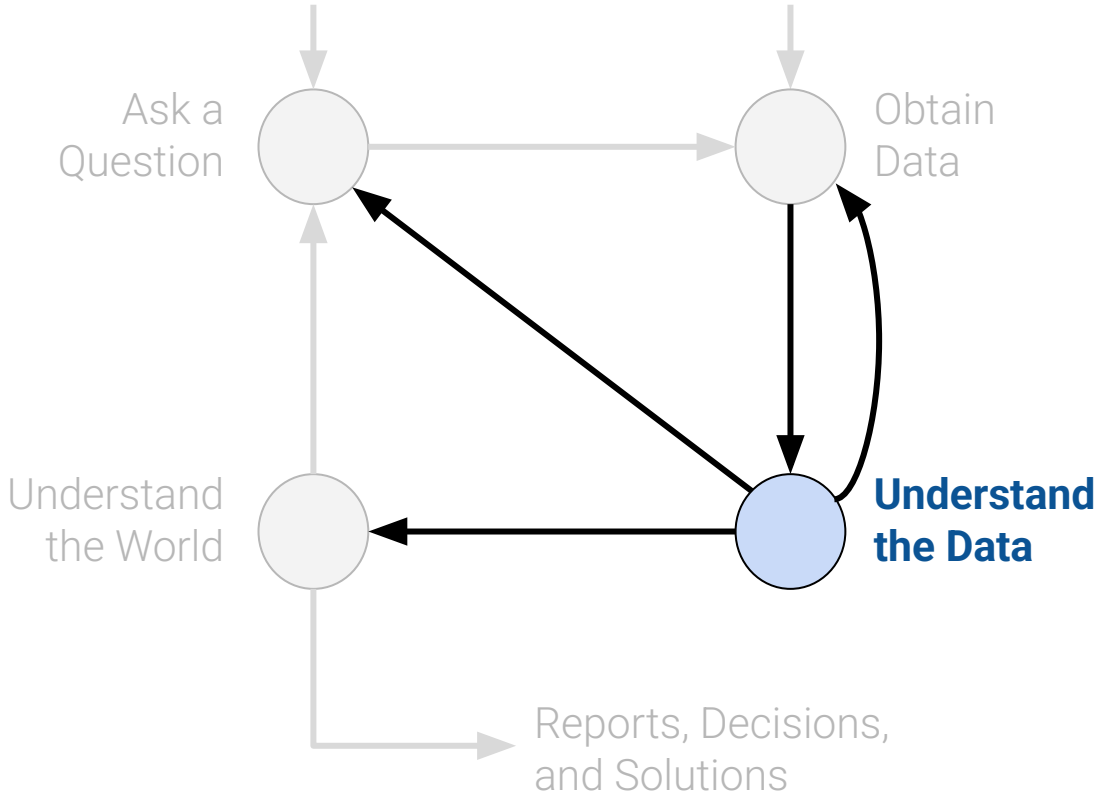
- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?



3. Exploratory Data Analysis & Visualization

8705986

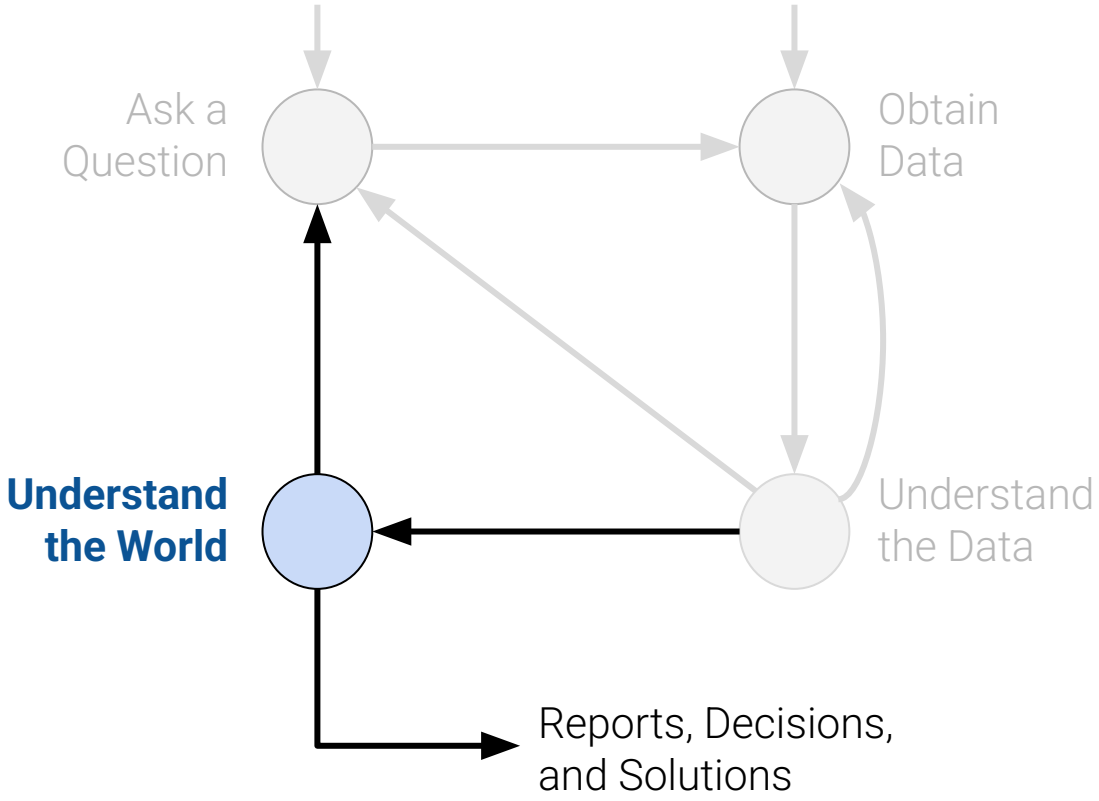
- How is our data organized and what does it contain?
- Do we already have relevant data?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?



4. Prediction and Inference

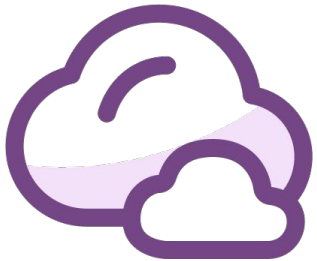
8705986

- 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?



slido

8705986



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.

8705986

Demo: The Data Science Lifecycle

Lecture 01, Data 100 Fall 2024

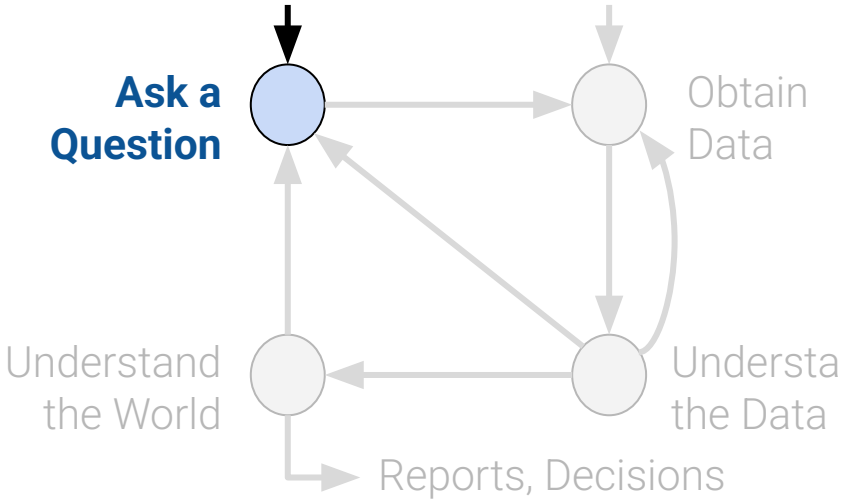
- Intros
- What is data science?
- What will you learn in this class?
- Course overview
- Data Science Lifecycle
- **Demo**

Available on the course website:
https://ds100.org/fa24/lecture/lec01

Demo Slides

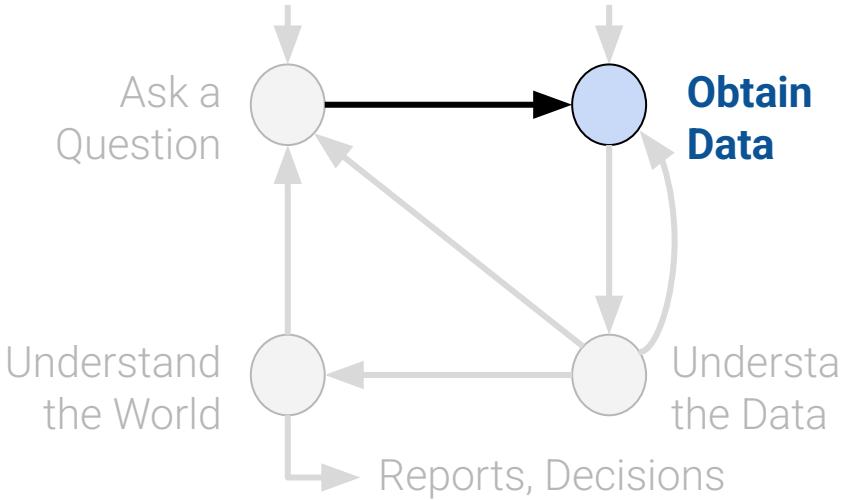
Ask a Question: Who are you?

8705986



8705986

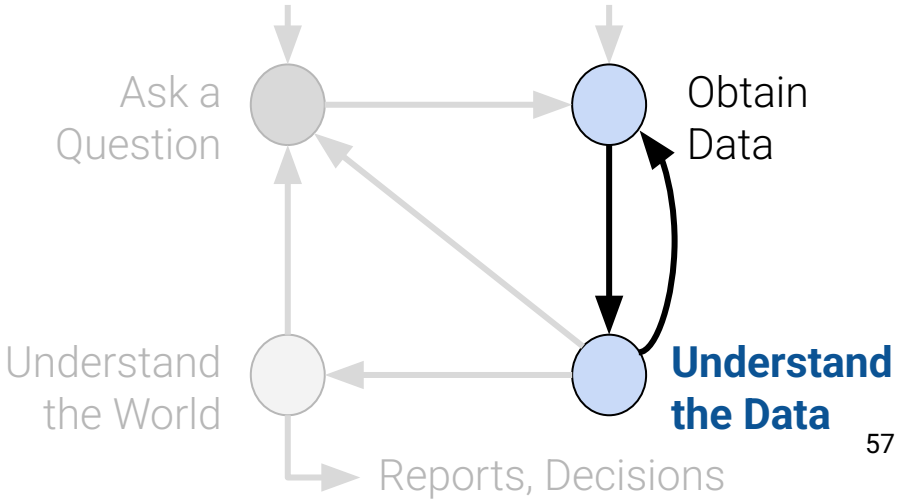
Demo Slides



Demo Slides

8705986

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



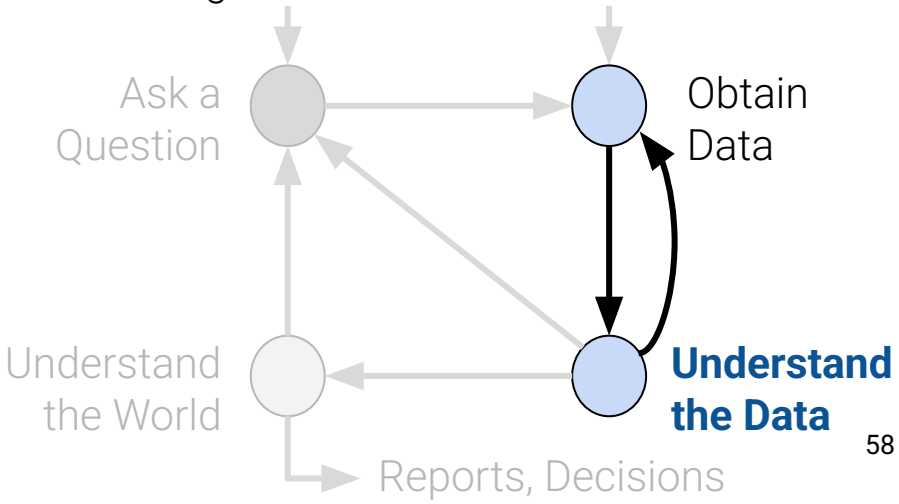
Demo Slides

8705986

Population: Data 100 students, Fall 2024

Some sub-questions:

✅ How many students are in the class?
✅ What are your majors?
✅ What year are you?
4. How did your major enrollment trend change over time?

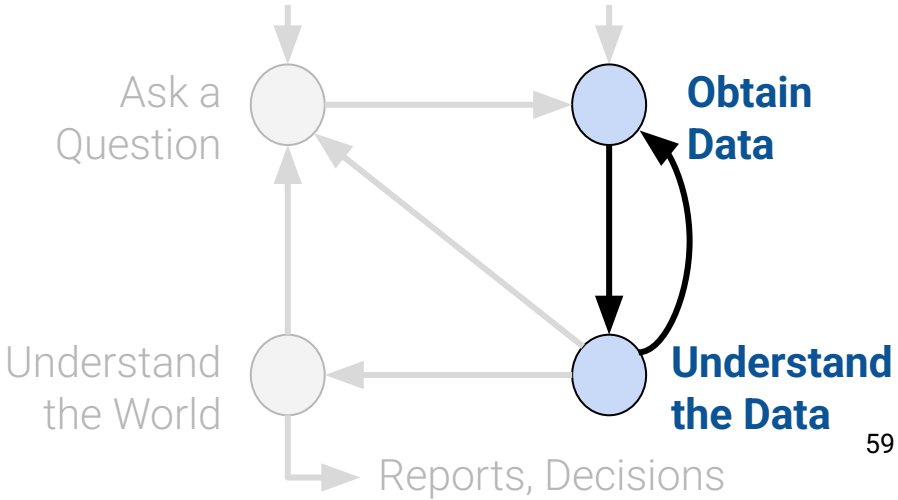


Demo Slides

Again, but for Campus Data

8705986

What does each row/column represent?



Demo Slides

What's the Point of This Demo?

8705986

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.**

Two Things…

8705986

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!

8705986

See you soon!

8705986

LECTURE 1

Course Overview

Content credit: Acknowledgments