WHAT DOES THE DATA TELL US TO DO?
WE ONLY HAVE BAD DATA ON THIS.
@SCOTTADAMSSAYS
DILBERT.COM
DOES THE BAD DATA SUGGEST WE SHOULD DO WHAT WE WANTED TO DO ANYWAY?
WELL, YES.
4-3-18 ©2018 Scott Adams, Inc./Dist. by Andrews McMeel
THAT'S CALLED "GOOD DATA."

LECTURE 1

Course Overview

An overview of data science, CSCI 3022, and the data science lifecycle.

CSCI 3022 @ CU Boulder

Maribeth Oscamou

Content credit: Lisa Yan, Josh Hug, Suraj Rampure, Allen Shen, Joey Gonzalez, and Sam Lau

Roadmap

Lecture 01, CSCI 3022

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

Intros

In Groups of 3-4   INTRODUCE YOURSELF :

- NAME, YEAR, MAJOR, HOMETOWN
- HOBBIES/INTERESTS
- SOME RANDOM FUN FACT ABOUT YOU

Getting To Know You:

I'd like to get a chance to be introduced to each of you!

1. **Please sign-up for a 15 min. timeslot (link on Piazza) to meet with me on Zoom during the first couple weeks to briefly introduce yourself and meet a few other classmates.**
2. **Please fill out our Getting to Know you Survey (link in HW 1) by Thursday**

iClicker Q&A

Join our iClicker class:

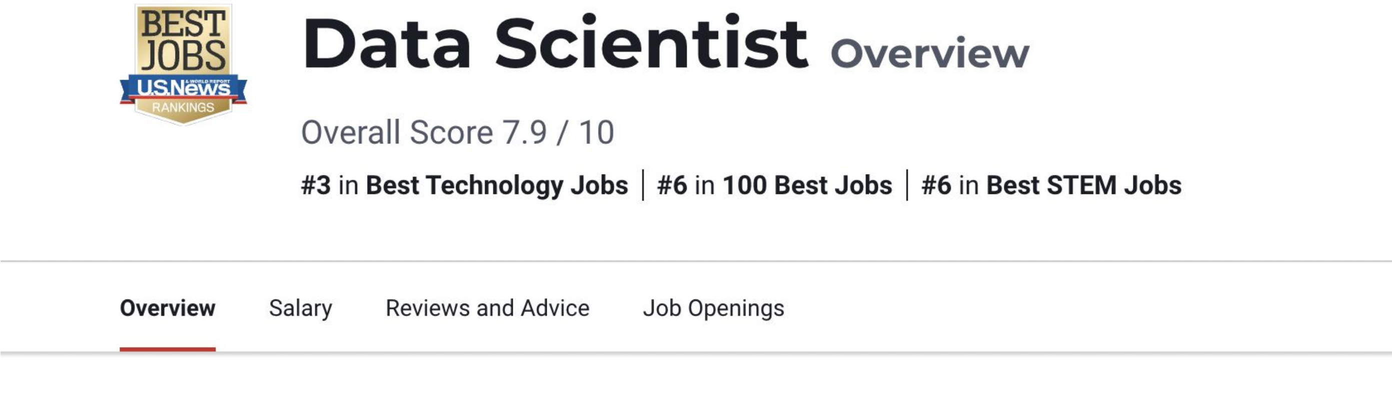
1:25pm Section: https://join.iClicker.com/LWSA

3:35pm Section: https://join.iClicker.com/IIMP

What is Data Science?

Lecture 01, CSCI 3022

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



BEST JOBS U.S.News & WORLD REPORT RANKINGS

Data Scientist Overview

Overall Score 7.9 / 10

#3 in **Best Technology Jobs** | **#6** in **100 Best Jobs** | **#6** in **Best STEM Jobs**

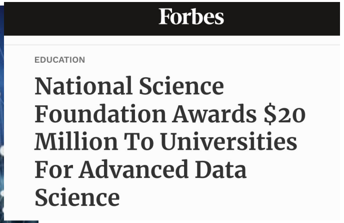
What is a Data Scientist?

Data scientists use technology to glean insights from large amounts of data they collect. It's a field that requires statistics, quantitative reasoning and computer programming skills. On top of all that, you need to be a good communicator so you can report your research findings and explain how they address a larger question you're trying to answer.

DATA SCIENCE | DATA MANAGEMENT

Tracking the Rise of Data Science

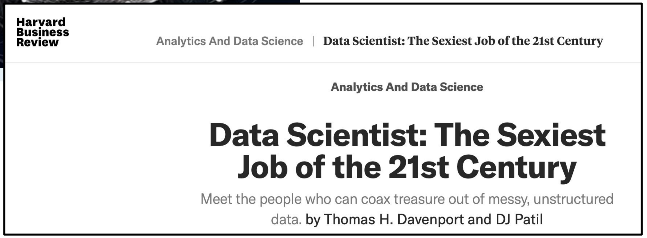
By Paul Mah on August 10, 2022



Forbes

EDUCATION

National Science Foundation Awards $20 Million To Universities For Advanced Data Science



Harvard Business Review

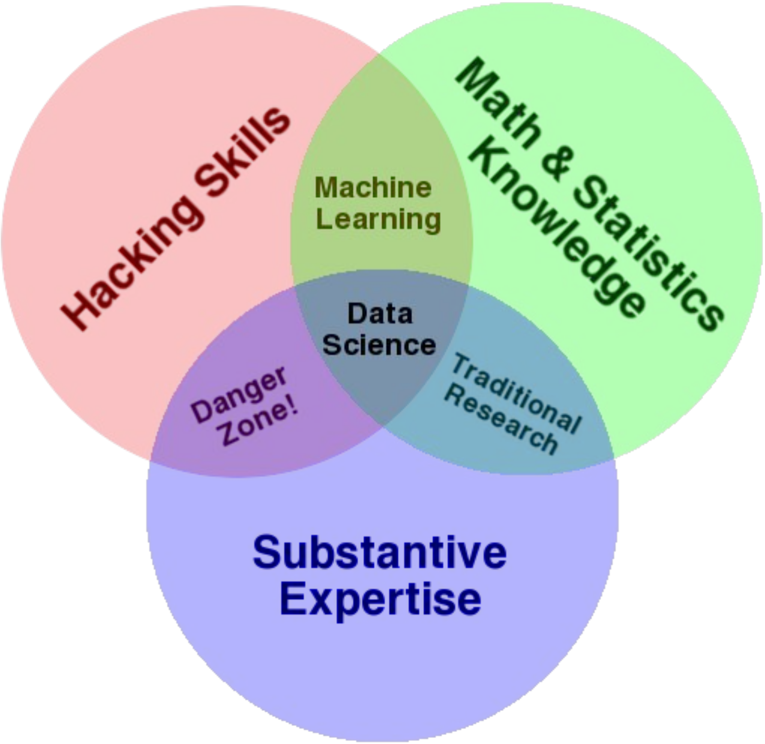
Analytics And Data Science | Data Scientist: The Sexiest Job of the 21st Century

Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

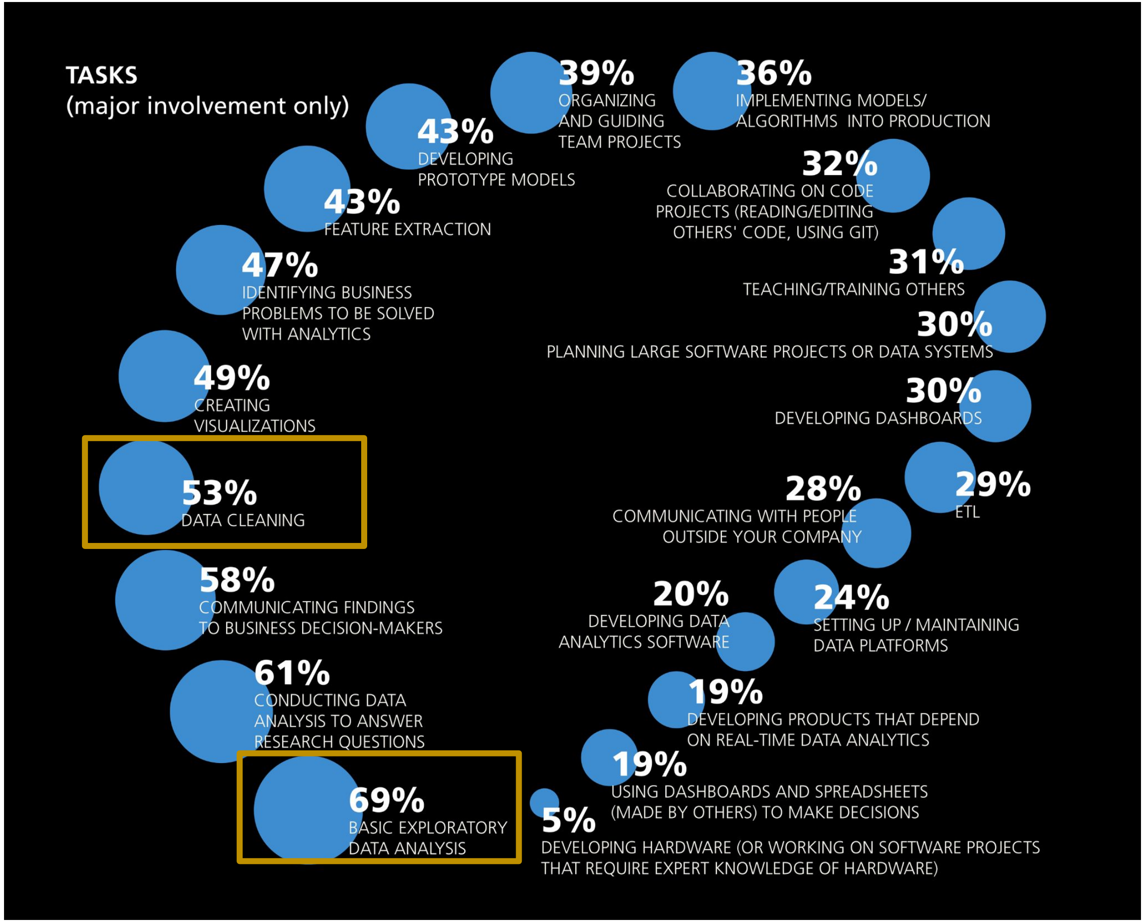
Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

Data Science Venn Diagram



by Drew Conway in 2010 ([link]())

Data science in industry



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

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

Why Data Science Matters

Data science enhances critical thinking

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:

- 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

Data is changing the world

Technology Trends

2020s ● ?

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

From Joey Gonzalez.

The Darker Side of Data Science?

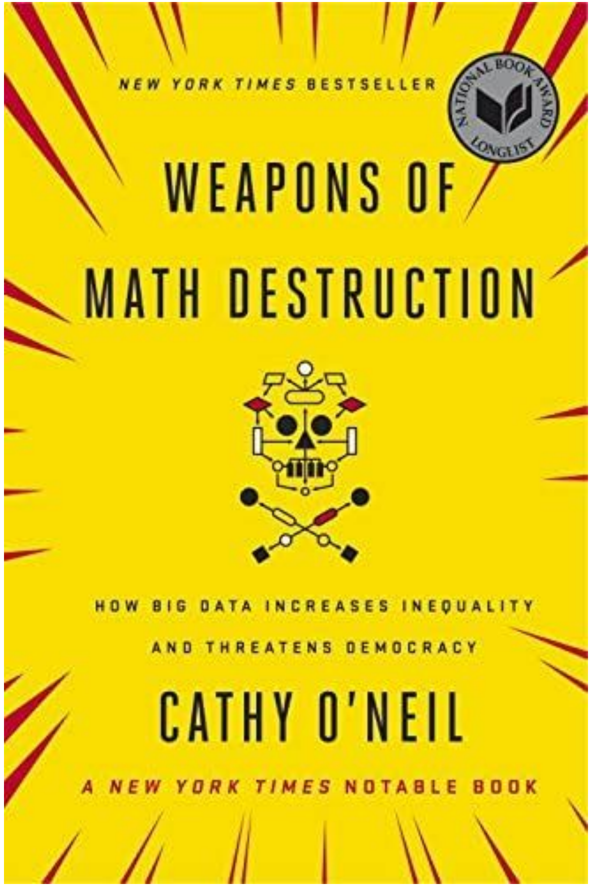
Obscuring complex decisions:

- 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!


NEW YORK TIMES BESTSELLER
NATIONAL BOOK AWARD LONGLIST
WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL
A NEW YORK TIMES NOTABLE BOOK

NPR author interview
with Cathy O'Neil

But…I am optimistic!

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 CSCI 3022 :-)

Data science enhances critical thinking

The world is complicated! Decisions are hard.

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**.

My Primary Goal for You in This Course

The world is complicated! Decisions are hard.

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.

What is CSCI 3022

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

Data science requires engineering and scientific insight

Good data analysis is not:

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

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

- What show should we recommend to our user to watch?
- In which markets should we focus our advertising campaign?
- Is the use of the COMPAS algorithm for prison sentencing fair?
- Should I send my kids to daycare?
- Is the world getting better or worse?
- What areas of the world are at higher risks for climate change impact in 10 years? 20?
- Where should we put docking ports for our bikes?
- What should we eat to avoid dying early of heart disease?
- Do immigrants from poor countries have a positive or negative impact on the economy?

What will you learn in this class?

Lecture 01, CSCI 3022

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

What is CSCI 3022?

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

Course goals

Prepare

Prepare students for advanced courses in **machine learning**, **artificial intelligence,** and **advance data science**, 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 CSCI 3022

- Pandas and NumPy
- Exploratory Data Analysis
- Visualization
  - matplotlib
  - Seaborn
  - plotly
- Sampling
- Probability and random variables
- Model design and loss formulation
- Linear Regression
- Feature Engineering
- Regularization, Bias-Variance Tradeoff, Cross-Validation
- Data science in the physical world
- Causality
- Logistic Regression

Prerequisites

Official prerequisites for this course:

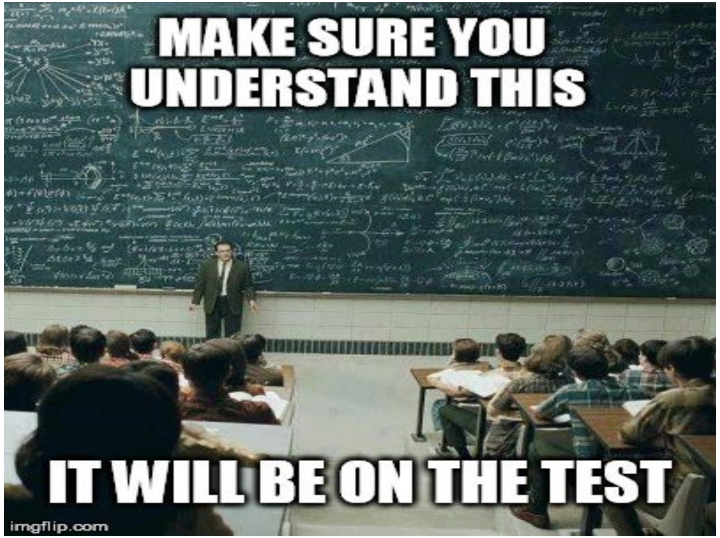
- Completion of Calculus 2 (C- or better)
- Completion of Discrete Structures (C- or better)
- Completion of Data Structures (C- or better)

Homework 1 and 2 and Lab Notebook 1 will help calibrate your background.

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

Course Overview

Lecture 01, CSCI 3022

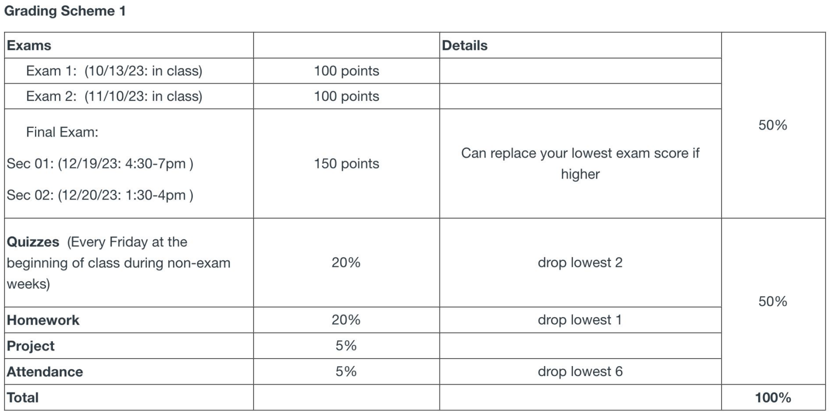


Course Logistics

Content and workflow

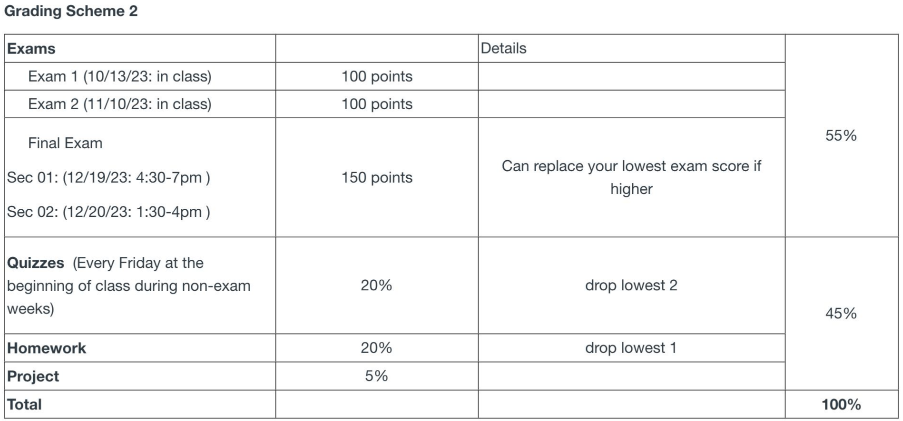
Course Logistics: Grading

At the end of the semester, your grade will be determined by the Grading Scheme below that leads to your highest grade :



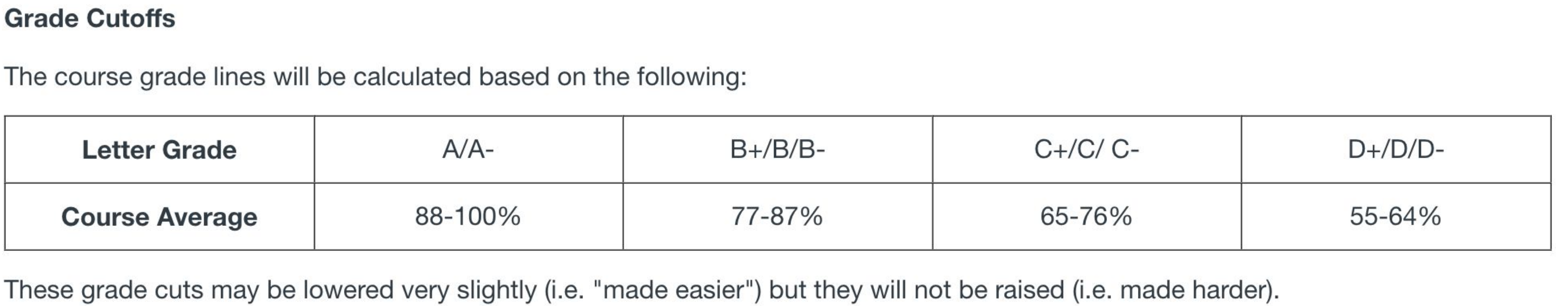
Grading Scheme 1

Exams		**Details**	
Exam 1: (10/13/23: in class)	100 points		50%
Exam 2: (11/10/23: in class)	100 points		
Final Exam: Sec 01: (12/19/23: 4:30-7pm) Sec 02: (12/20/23: 1:30-4pm)	150 points	Can replace your lowest exam score if higher	
Quizzes (Every Friday at the beginning of class during non-exam weeks)	20%	drop lowest 2	50%
Homework	20%	drop lowest 1	
Project	5%		
Attendance	5%	drop lowest 6	
Total			**100%**



Grading Scheme 2

Exams		Details	
Exam 1 (10/13/23: in class)	100 points		55%
Exam 2 (11/10/23: in class)	100 points		
Final Exam Sec 01: (12/19/23: 4:30-7pm) Sec 02: (12/20/23: 1:30-4pm)	150 points	Can replace your lowest exam score if higher	
Quizzes (Every Friday at the beginning of class during non-exam weeks)	20%	drop lowest 2	45%
Homework	20%	drop lowest 1	
Project	5%		
Total			**100%**



Grade Cutoffs

The course grade lines will be calculated based on the following:

Letter Grade	A/A-	B+/B/B-	C+/C/ C-	D+/D/D-
Course Average	88-100%	77-87%	65-76%	55-64%

These grade cuts may be lowered very slightly (i.e. "made easier") but they will not be raised (i.e. made harder).

Course Logistics: Your Week At A Glance

Mon	Tues	Wed	Thurs	Fri
Attend & Participate in Class	(Optional): Attend Notebook Discussion with our TA (5pm-6pm Zoom)	Attend & Participate in Class	HW Due 11:59pm via Gradescope	In Class Quiz (beginning of class) Attend & Participate in Class
Previous week quiz grades posted		Previous week HW grades posted		Next week HW released Next week Discussion Notebook released

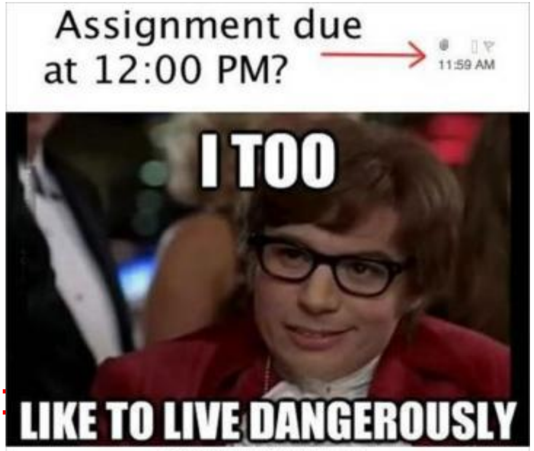
But what if…

sick/quarantined/emergency/traveling/*insert your situation here*

and miss a lecture, quiz, exam or assignment?

There are no make-up, early or online exams OR quizzes.

- Your lowest 1 midterm exam grade will be replaced by your final exam grade (if your final exam grade is higher)
- Your lowest 2 quiz scores will be dropped
- Your lowest 6 attendance scores will be dropped
- Your lowest 1 HW score will be dropped
- You are allowed up to 4 Written HW extensions for full credit
  - Two 24-hr extensions
  - Two 6-day extensions for full credit
- Course capture video lectures and lecture slides are posted Canvas



THESE APPLY TO ALL STUDENTS AND TO ANY AND ALL SITUATIO

No Other Allowances Will Be Made

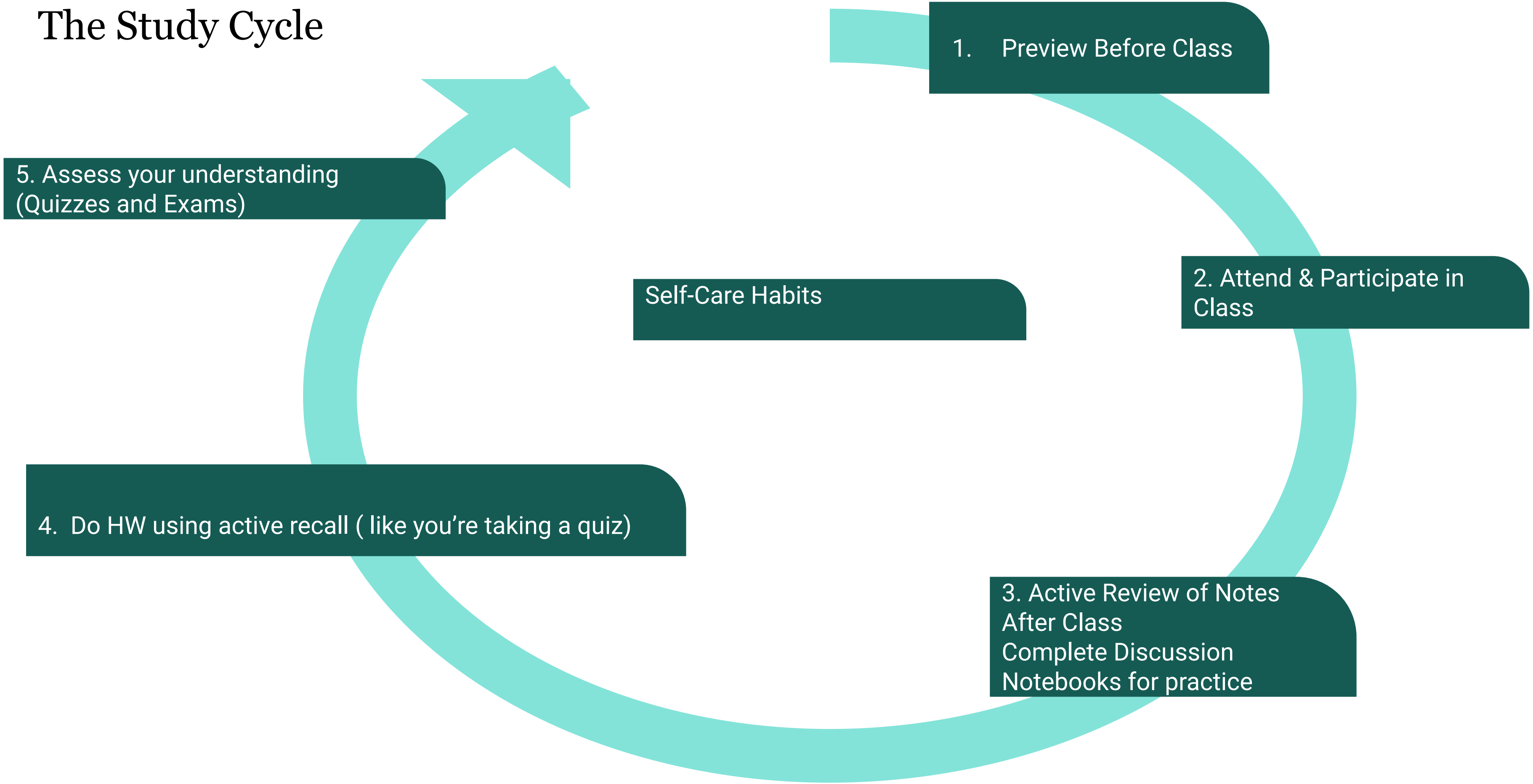
Course Expectations

- Allocate approximately **6-7 hours per week *outside of class*** to study and do work for this class

- Space out your practice
  - A little each day is much more effective for learning

What I do when a teacher says "this cannot be done the night before"
🟦 Adhere the warning and start early
🟥 Take it as a personal challenge

The Study Cycle



1. Preview Before Class

2. Attend & Participate in Class

3. Active Review of Notes After Class
Complete Discussion Notebooks for practice

4. Do HW using active recall (like you’re taking a quiz)

5. Assess your understanding (Quizzes and Exams)

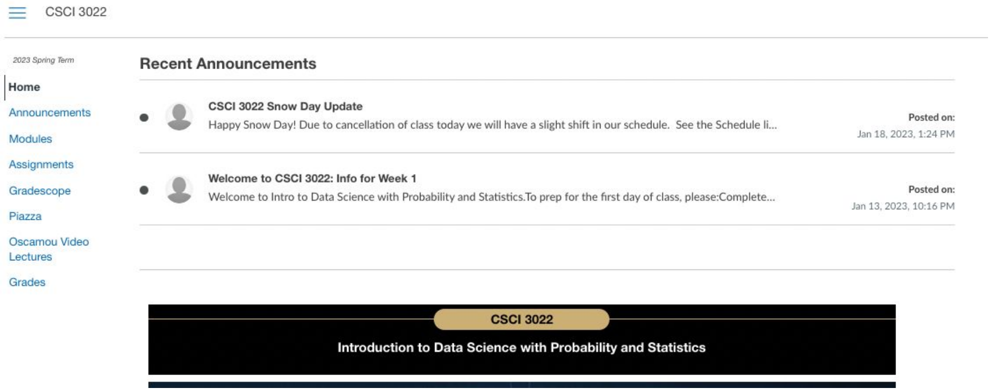
Self-Care Habits

Course Logistics: Attendance

Attendance and active participation is expected in this course.

Course lectures are recorded via course capture for you to review as you study – but it’s much better to attend.

You are responsible for getting notes from a classmate if you are absent.



Recent Announcements

CSCI 3022 Snow Day Update
Happy Snow Day! Due to cancellation of class today we will have a slight shift in our schedule. See the Schedule li...
Posted on:
Jan 18, 2023, 1:24 PM

Welcome to CSCI 3022: Info for Week 1
Welcome to Intro to Data Science with Probability and Statistics.To prep for the first day of class, please:Complete...
Posted on:
Jan 13, 2023, 10:16 PM

CSCI 3022
Introduction to Data Science with Probability and Statistics

Resources: We are Here to Help!

Piazza– Post and/or respond to any questions here!

Office Hours– Schedule is on Canvas & Piazza. You can attend ANY office hours offered by TA, Course Manager, Course Assistants and myself

Discussion Notebook Sessions With our Course TA - Tuesdays 5pm-6pm via Zoom (link is on Office Hour Page in Canvas)

Communicating With Me:

Please post on PIAZZA:

- General course questions
- Questions about HW/notes
- Answers to in-class polls
- Responses to other classmates' questions

******Piazza is checked periodically by Course Staff during business hours (8am-5pm; M-F)***

Please DO NOT notify me if you are going to miss class or miss a quiz or miss a midterm.
The course grading policies address ANY and ALL situations, regardless of the circumstances.

Please come to my OFFICE HOURS with:

- Questions about HW/notes
- Questions about exams
- To discuss your performance on an exam or in the class

Please EMAIL me with:

- Disability Accommodations
- University-related sports activity schedule conflicts
- If you have an emergency/illness the day of the final exam
  - (note: ALL students must take the final exam; however in an emergency some students may qualify for an Incomplete)

Accommodations

For quiz/exam accommodations: Please email me your **accommodations letter** from the Office of Disability Services **within the first week**.

- For **University-related sports activities**: Please email me about potential conflicts **within the first week of class**

Course Logistics – Academic Integrity

See the Course Collaboration and Academic Honesty Policy on Canvas for full details and specific examples.

ANY instance in violation of the Course Collaboration & Academic Honesty Policy will result in an F in the course and a report to Honor Code.
There are no 2nd chances.

What is Something You are Wondering??

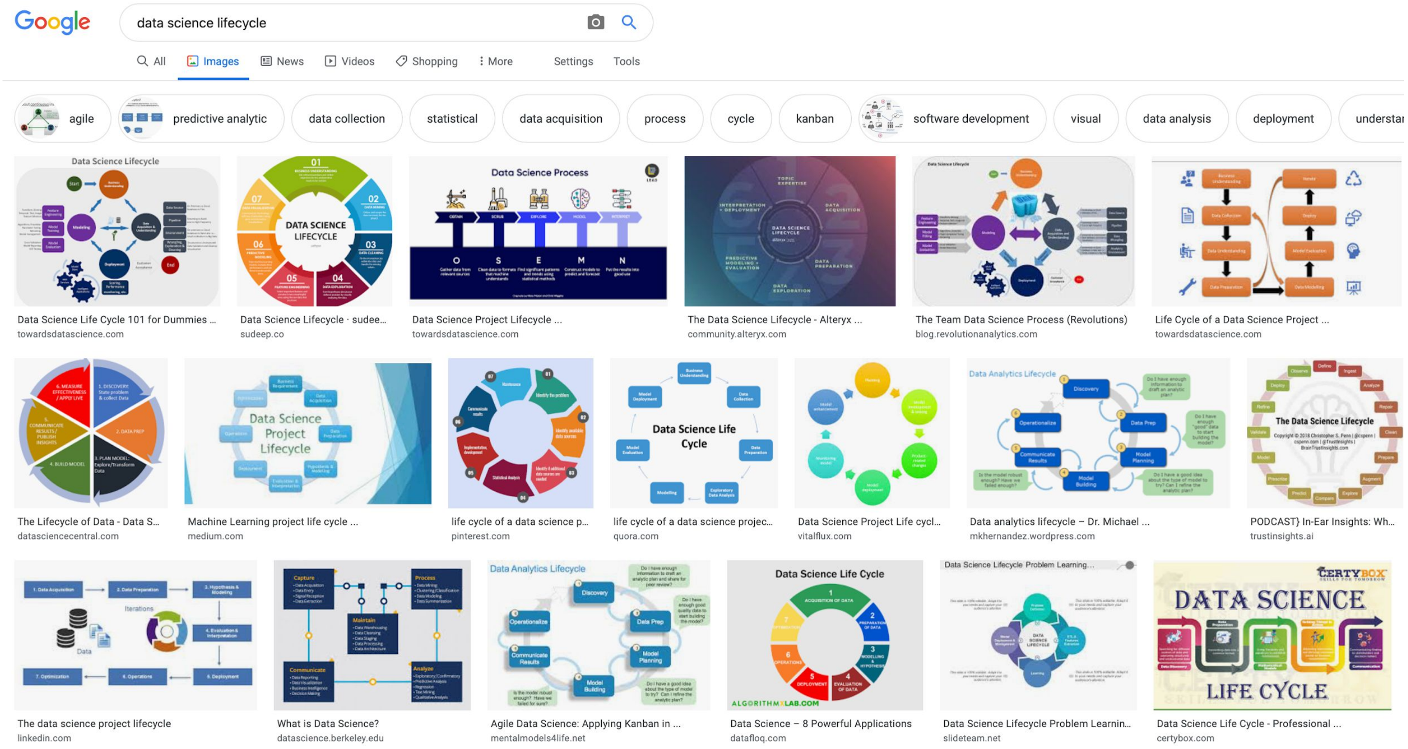
Data Science Lifecycle

Lecture 01, CSCI 3022

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

Data science lifecycle

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 Life Cycle 101 for Dummies ...
towardsdatascience.com

Data Science Lifecycle · sudee...
sudeep.co

Data Science Project Lifecycle ...
towardsdatascience.com

The Data Science Lifecycle - Alteryx ...
community.alteryx.com

The Team Data Science Process (Revolutions)
blog.revolutionanalytics.com

Life Cycle of a Data Science Project ...
towardsdatascience.com

The Lifecycle of Data - Data S...
datasciencecentral.com

Machine Learning project life cycle ...
medium.com

life cycle of a data science p...
pinterest.com

life cycle of a data science projec...
quora.com

Data Science Project Life cycl...
vitalflux.com

Data analytics lifecycle – Dr. Michael ...
mkhernandez.wordpress.com

PODCAST} In-Ear Insights: Wh...
trustinsights.ai

The data science project lifecycle
linkedin.com

What is Data Science?
datascience.berkeley.edu

Agile Data Science: Applying Kanban in ...
mentalmodels4life.net

Data Science – 8 Powerful Applications
datafloq.com

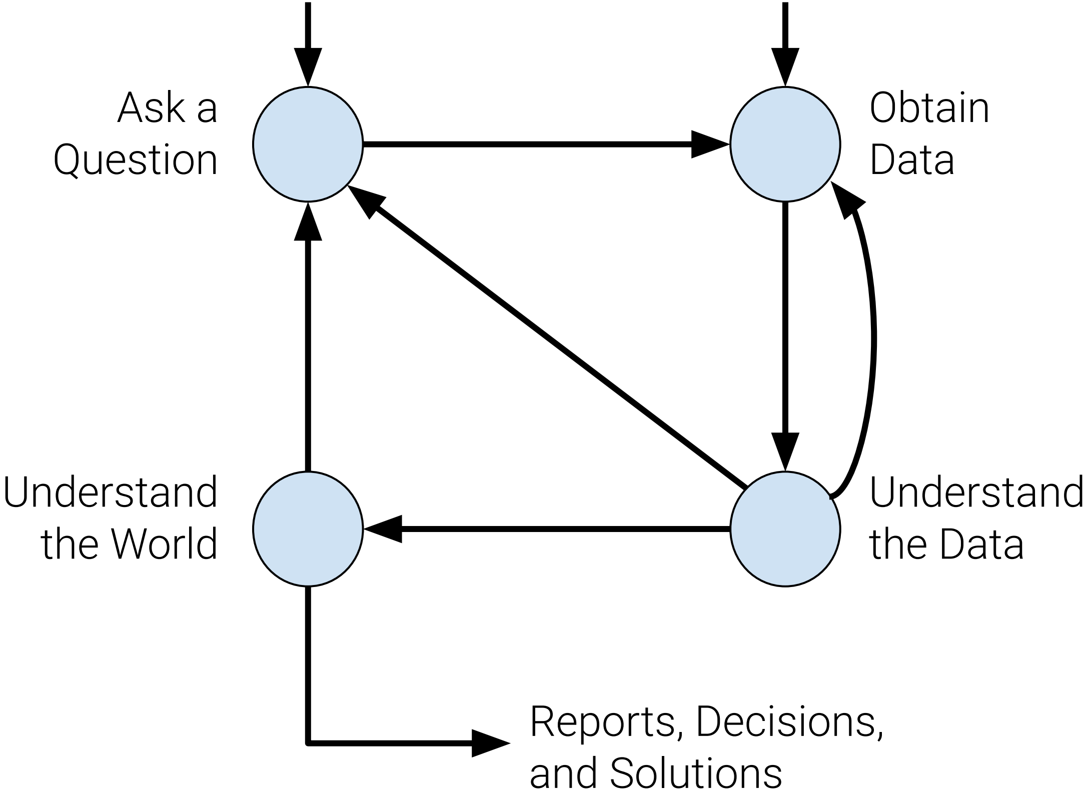
Data Science Lifecycle Problem Learnin...
slideteam.net

Data Science Life Cycle - Professional ...
certybox.com

Data science lifecycle

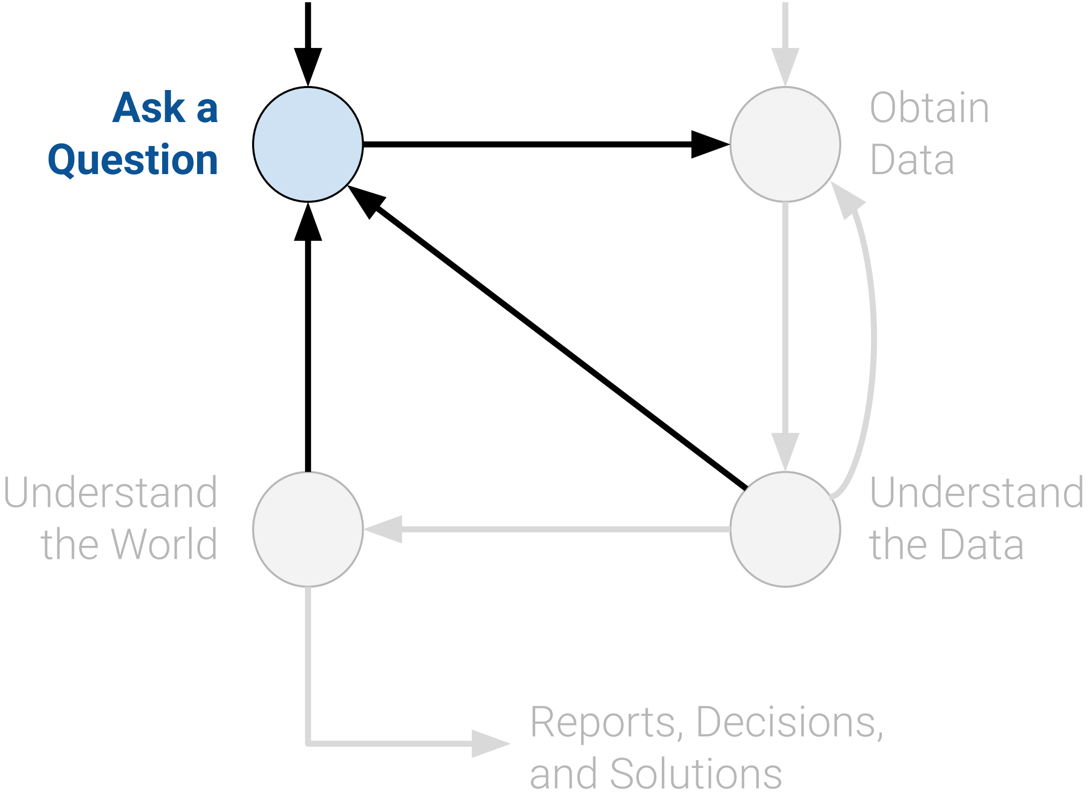
The data science lifecycle is a **high-level description** of the data science workflow.

Note the two distinct entry points!



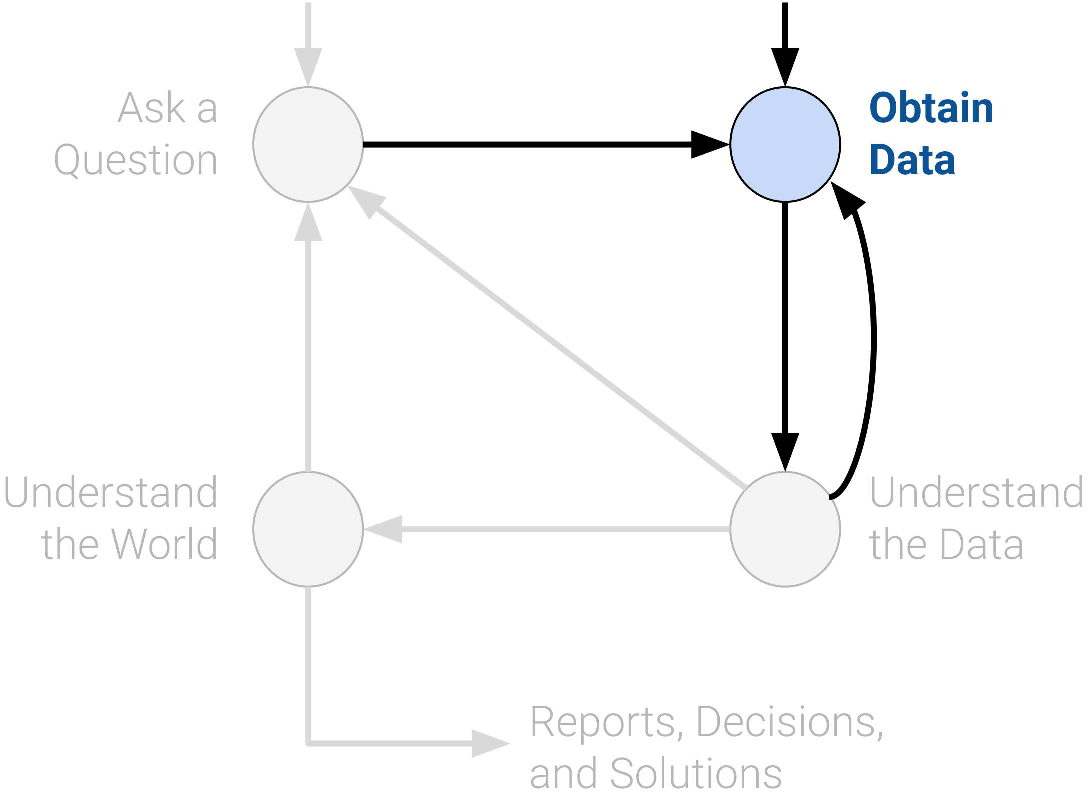
1. Question/Problem Formulation

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



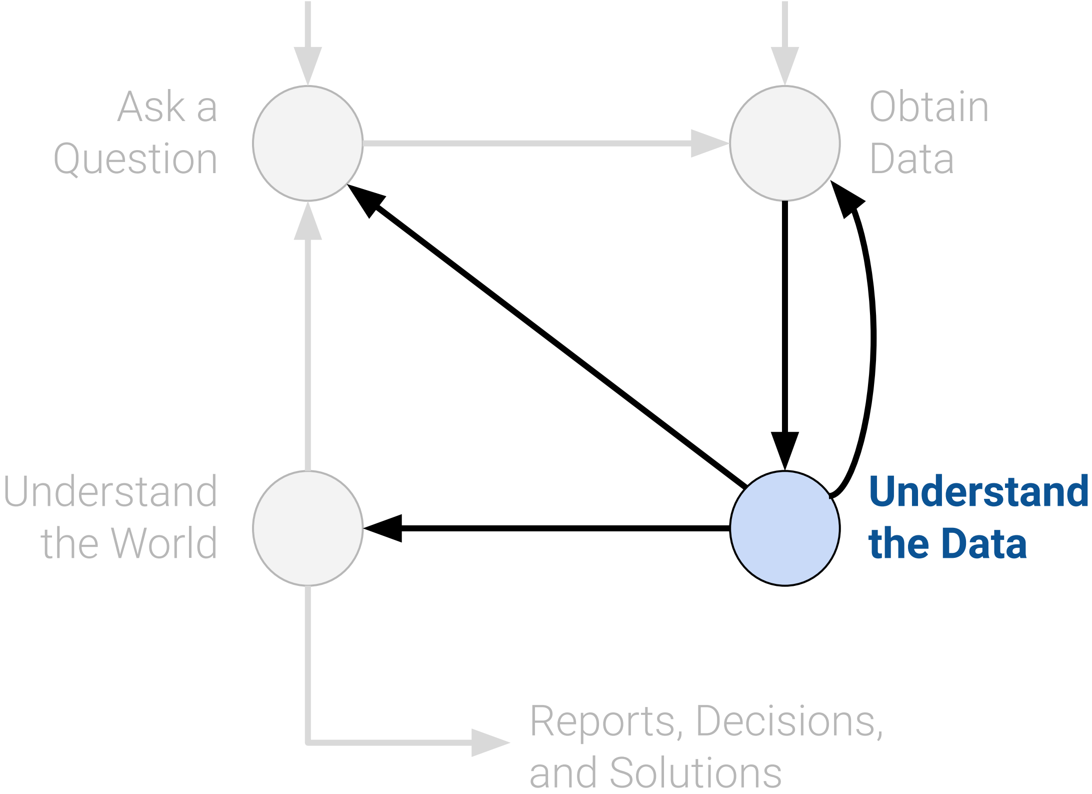
2. Data Acquisition and Cleaning

- 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?



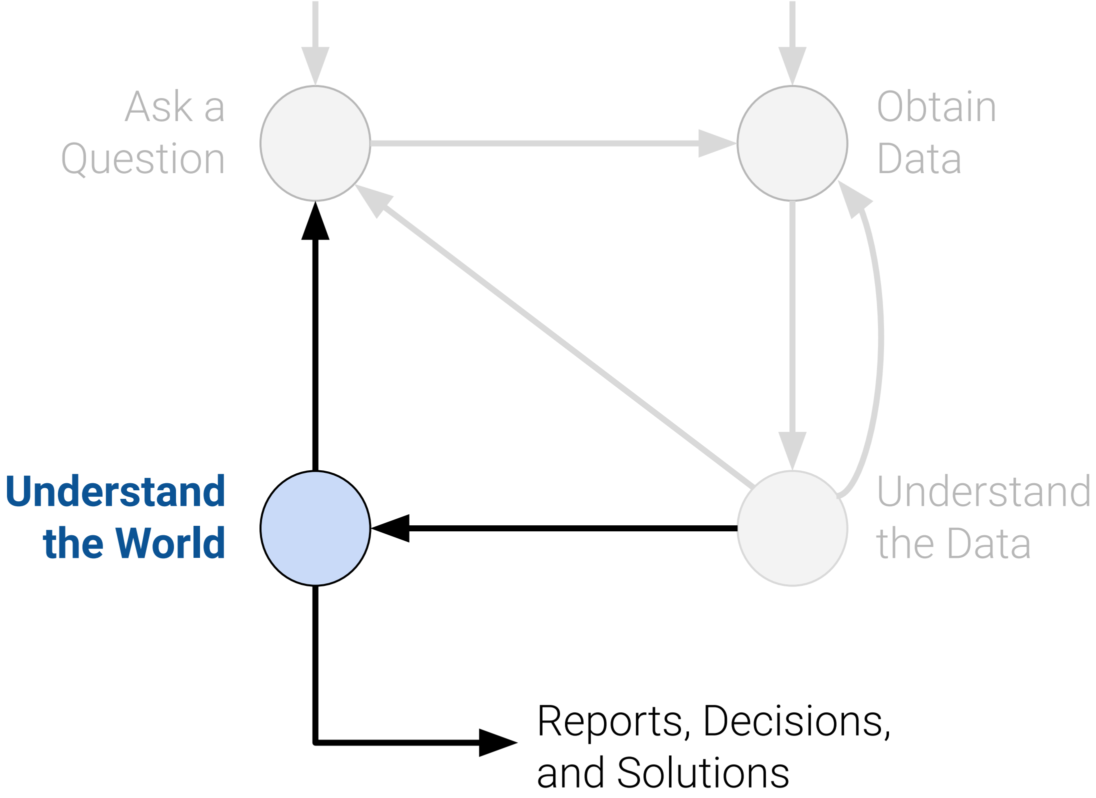
3. Exploratory Data Analysis & Visualization

- 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

- 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?



Course Websites / Platforms

Online platforms

Canvas

- Where all course information including lectures, assignments, announcements and grades are posted

CSCI 3022 JuptyerHub (https://coding.csel.io/)

- Where you will work on all assignments (links on Canvas assignments automatically take you here).

iClicker (https://student.iclicker.com/#/login**)**

- Where you answer polls during class

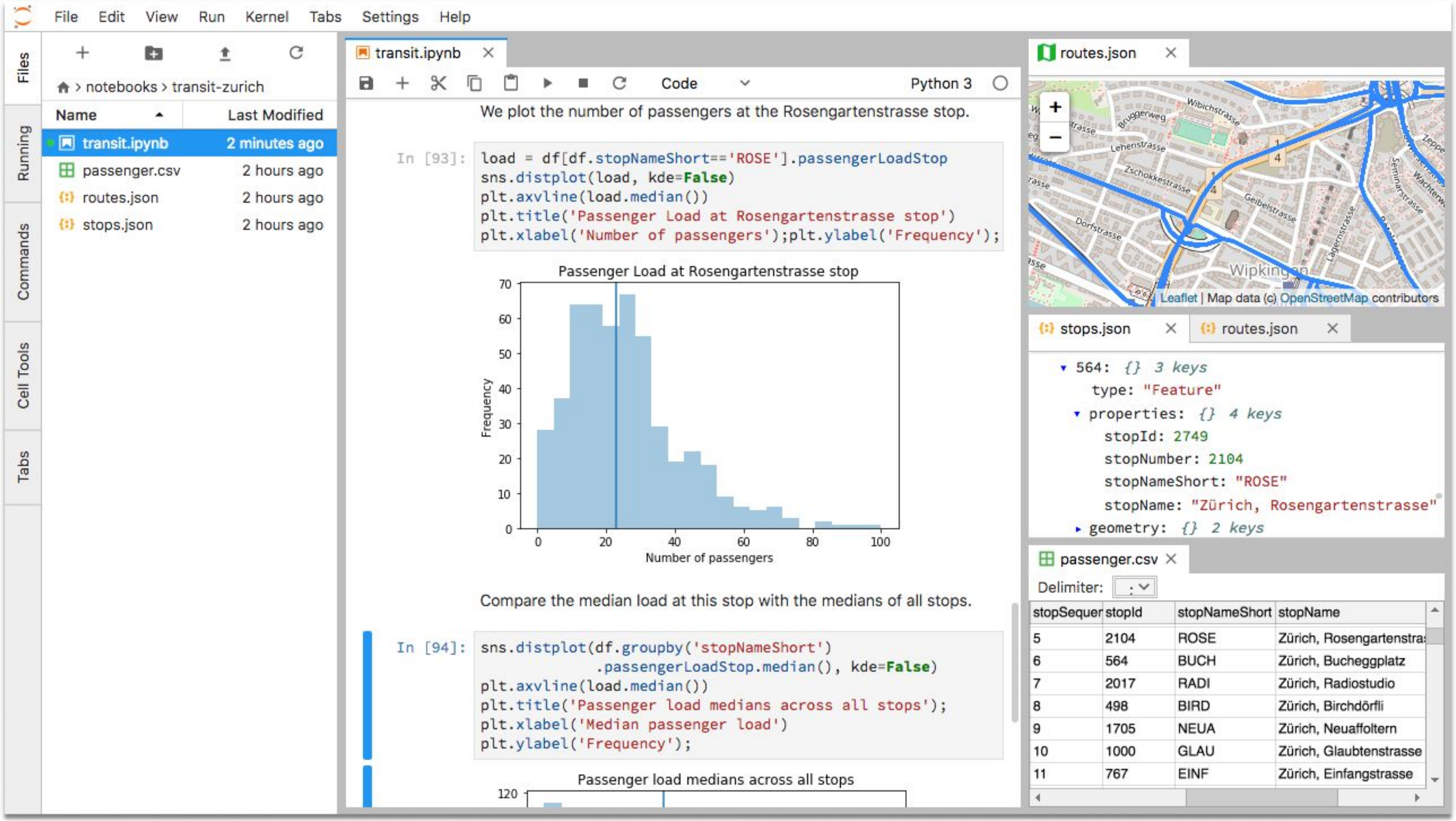
Piazza (linked from Canvas)

- Where discussions are posted

Gradescope (linked from Canvas)

- Where all assignments are submitted

Programming Environment for our Course: JupyterLab



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.

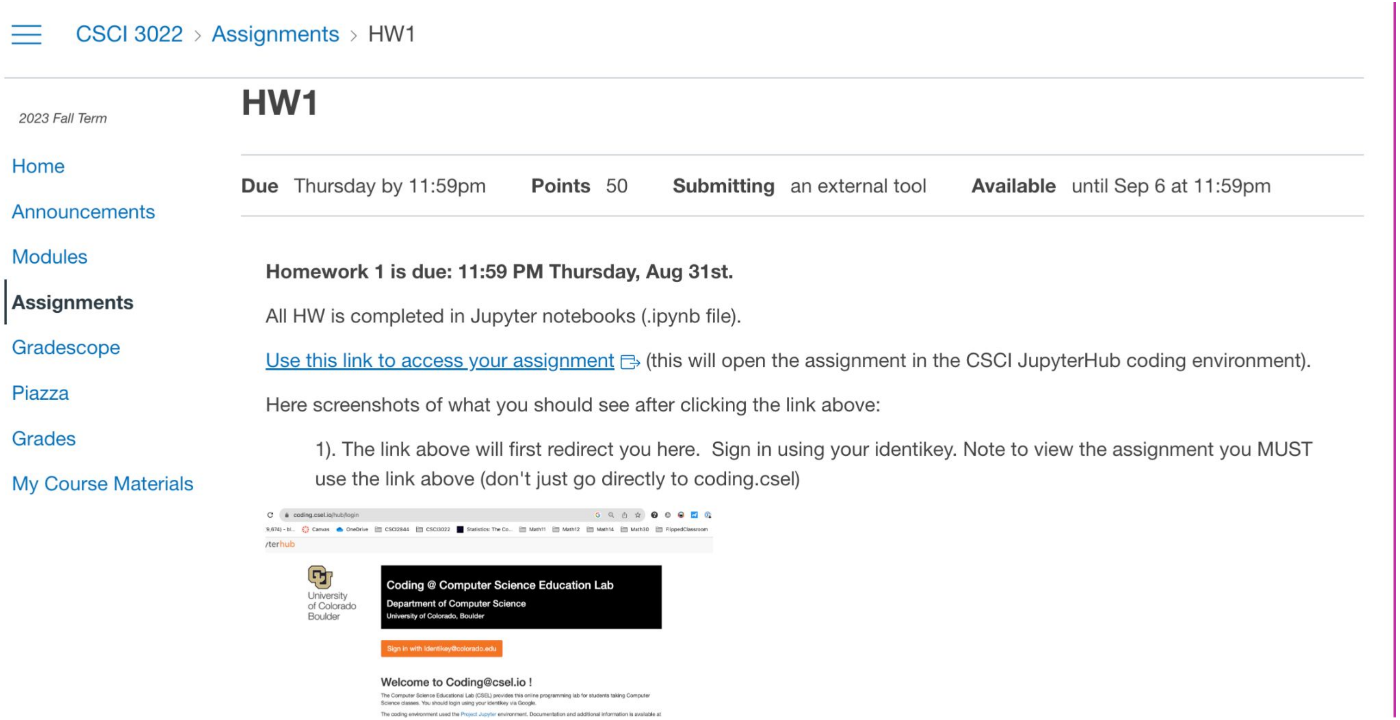
```
In [94]: sns.distplot(df.groupby('stopNameShort')
                         .passengerLoadStop.median(), kde=False)
         plt.axvline(load.median())
         plt.title('Passenger load medians across all stops');
         plt.xlabel('Median passenger load')
         plt.ylabel('Frequency');
```

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

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

Accessing HW 1

https://canvas.colorado.edu/courses/95692/assignments/1774751



CSCI 3022 › Assignments › HW1
2023 Fall Term
Home
Announcements
Modules
Assignments
Gradescope
Piazza
Grades
My Course Materials
coding.csel.io/hub/login
9,674) - bl... Canvas OneDrive CSCI2844 CSCI3022 Statistics: The Co... Math11 Math12 Math14 Math30 FlippedClassroom
/terhub
University of Colorado Boulder
Sign in with identikey@colorado.edu

HW1

Due Thursday by 11:59pm **Points** 50 **Submitting** an external tool **Available** until Sep 6 at 11:59pm

Homework 1 is due: 11:59 PM Thursday, Aug 31st.

All HW is completed in Jupyter notebooks (.ipynb file).

Use this link to access your assignment (this will open the assignment in the CSCI JupyterHub coding environment).

Here screenshots of what you should see after clicking the link above:

1). The link above will first redirect you here. Sign in using your identikey. Note to view the assignment you MUST use the link above (don't just go directly to coding.csel)

Coding @ Computer Science Education Lab

Department of Computer Science

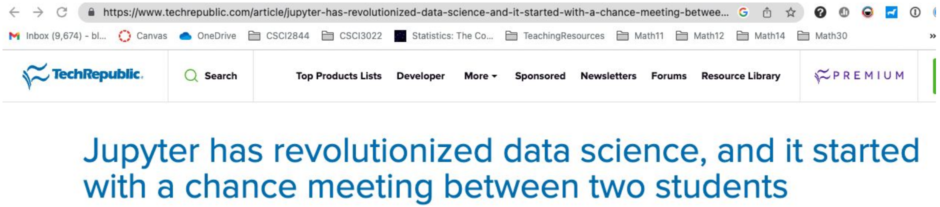
University of Colorado, Boulder

Welcome to Coding@csel.io !

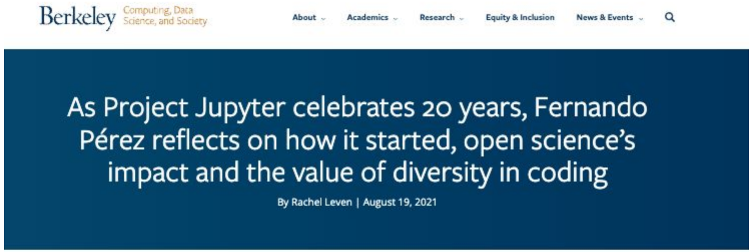
The Computer Science Educational Lab (CSEL) provides this online programming lab for students taking Computer Science classes. You should login using your identikey via Google.

The coding environment used the Project Jupyter environment. Documentation and additional information is available at

Practice With Jupyter: Notebook 01

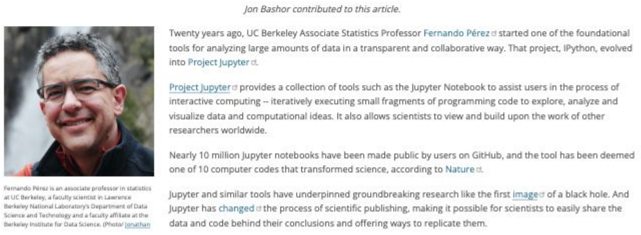


Jupyter has revolutionized data science, and it started with a chance meeting between two students



As Project Jupyter celebrates 20 years, Fernando Pérez reflects on how it started, open science's impact and the value of diversity in coding

By Rachel Leven | August 19, 2021



Jon Bashor contributed to this article.

Twenty years ago, UC Berkeley Associate Statistics Professor Fernando Pérez started one of the foundational tools for analyzing large amounts of data in a transparent and collaborative way. That project, IPython, evolved into Project Jupyter.

Project Jupyter provides a collection of tools such as the Jupyter Notebook to assist users in the process of interactive computing -- iteratively executing small fragments of programming code to explore, analyze and visualize data and computational ideas. It also allows scientists to view and build upon the work of other researchers worldwide.

Nearly 10 million Jupyter notebooks have been made public by users on GitHub, and the tool has been deemed one of 10 computer codes that transformed science, according to Nature.

Jupyter and similar tools have underpinned groundbreaking research like the first image of a black hole. And Jupyter has changed the process of scientific publishing, making it possible for scientists to easily share the data and code behind their conclusions and offering ways to replicate them.

Fernando Pérez is an associate professor in statistics at UC Berkeley, a faculty scientist in Lawrence Berkeley National Laboratory's Department of Data Science and Technology and a faculty affiliate at the Berkeley Institute for Data Science. (Photo/ Jonathan

Shout out to Fernando Perez & Brian Granger for developing Juptyer notebooks!

Fernando created iPython while a graduate student at CU in 2001 and co-founded its successor, Project Jupyter with Brian Granger.

The Jupyter team collaborates openly to create the next generation of tools for human-driven computational exploration, data analysis, scientific insight and education.

Notebook 1

Lecture 01, CSCI 3022

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

Available on Canvas -> Modules

Learning Advanced JupyterLab

JupyterLab offers notebooks and more tools for data science.

We'll be accessing JupyterLab using CSCI's CSEL **JupyterHub** (https://coding.csel.io/)

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.

See you soon!