Course Overview

LECTURE 1

An overview of data science and the data science lifecycle.

CS577, Sean Kang, Lecturer

Agenda

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

Instructor: Sean Kang

M.S., Business Intelligence Data Analytics, Carnegie Mellon

- *Focus area: Machine Learning and Data Mining*

B.S., Computer Science, Purdue University

Application Software Engineer, Software Engineering Architect, Director of Engineering, over 30+ years of commercial SW experience

- **IBM**
- **Microsoft**
- **Aion (Rule-based SW company)**
- **Various other companies**

What is data science?

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

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 J Gonzalez, UC Berkeley

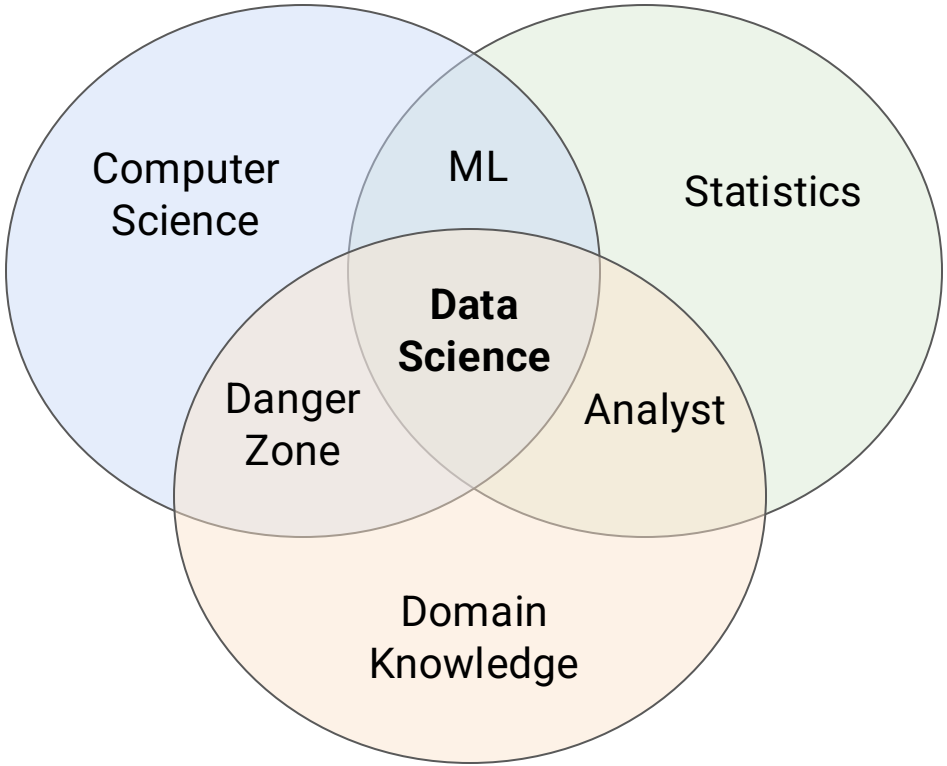
Data science is a fundamentally interdisciplinary field

Joey Gonzalez (co-creator of this course)

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

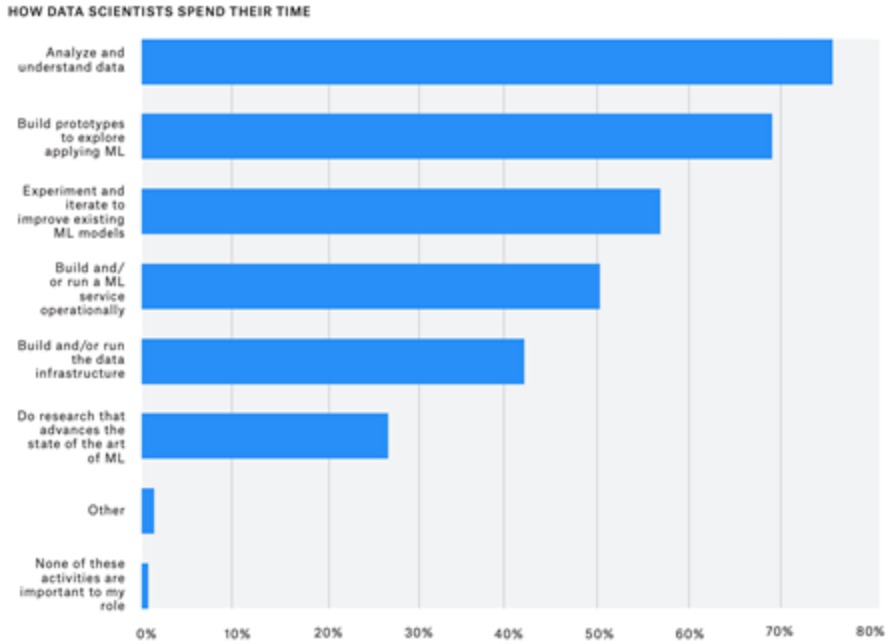
- Understand the world (science).
- Solve problems (engineering).

Data Science Venn Diagram



From Prof. J Gonzalez.

Data science in industry



The tasks that data scientists say they work on regularly.

Self-reported. Based on the results of Kaggle's 2019 Machine Learning & Data Science Survey.

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:

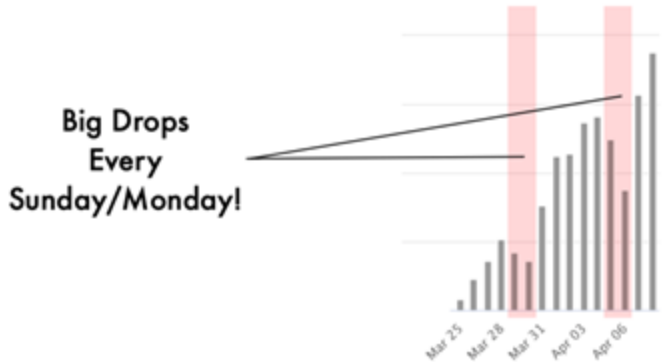
- Are the customers likely to buy our products? (result is a purchase)
- In the probability high that a COPD patient being discharged will return to the hospital for the same issue? (patient returns)
- What is the projected net income for next year? (result can be captured by the company's finance/accounting department.)
- Are there any differences in customer satisfaction in different business districts? (This can be captured in post-sales reviews or surveys)

Once we have a good question, then we need to research how to collect the data to answer the question. How do we find the correct approach or method to answer the question.

Dangers of Mis-interpreting Data – Collection was faulty

There are real-world implications of the work we do as data scientists.

Let's take a look at the daily numbers reported by the United Kingdom:

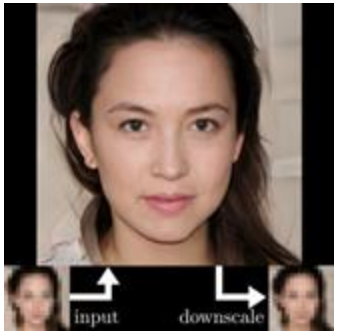


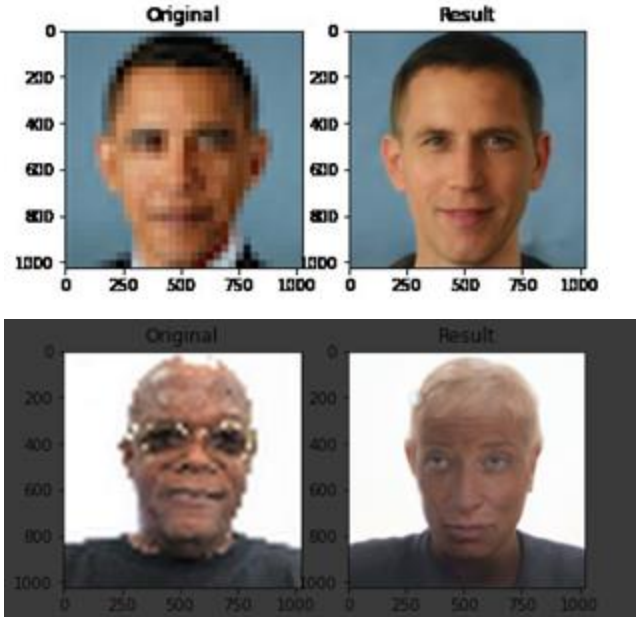
Daily Deaths due to COVID in the UK from https://www.worldometers.info/coronavirus/country/uk/

The problem is that this weekly cycle is fake. It's an artifact of how the data is collected and reported.

Unconscious bias is real – be mindful of it

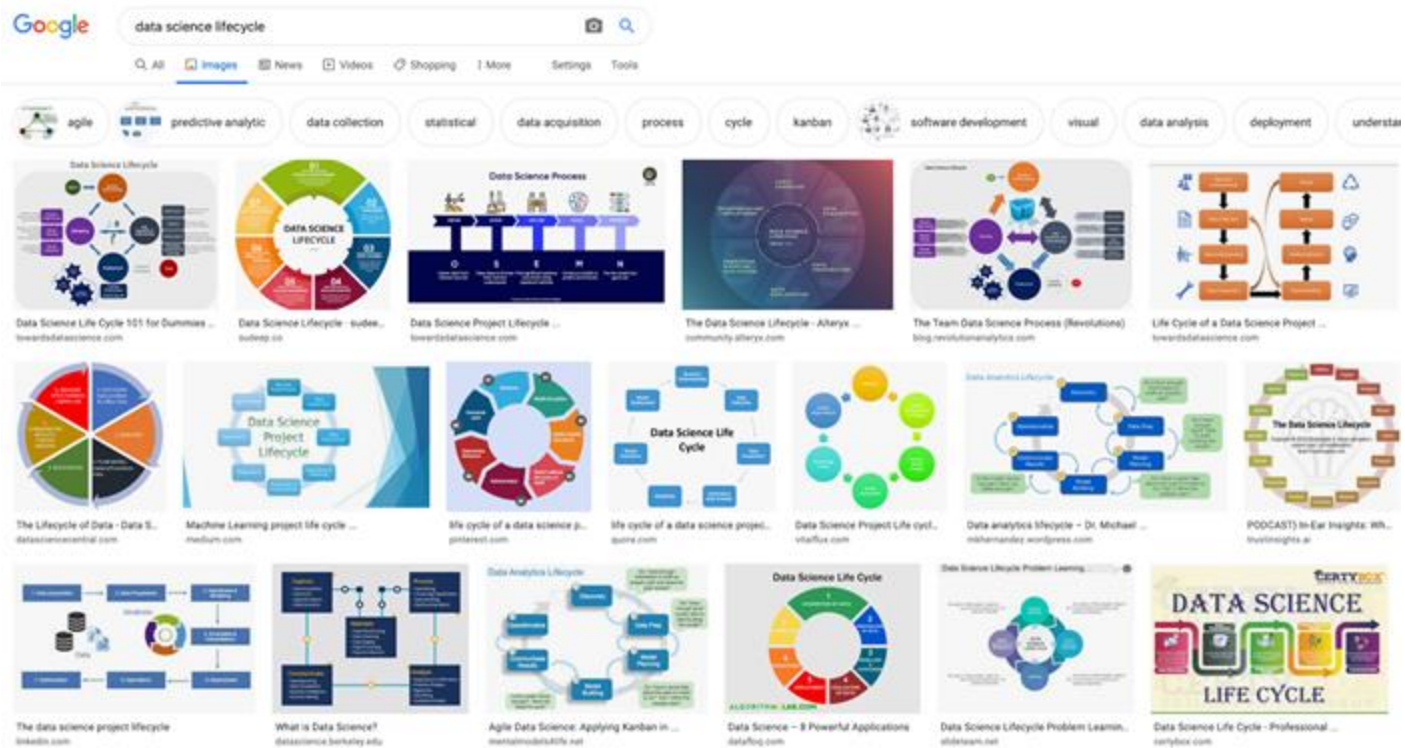
A “depixelizer” (link in description) was built that takes pixelated images and generates images that are perceptually realistic and downscale correctly.





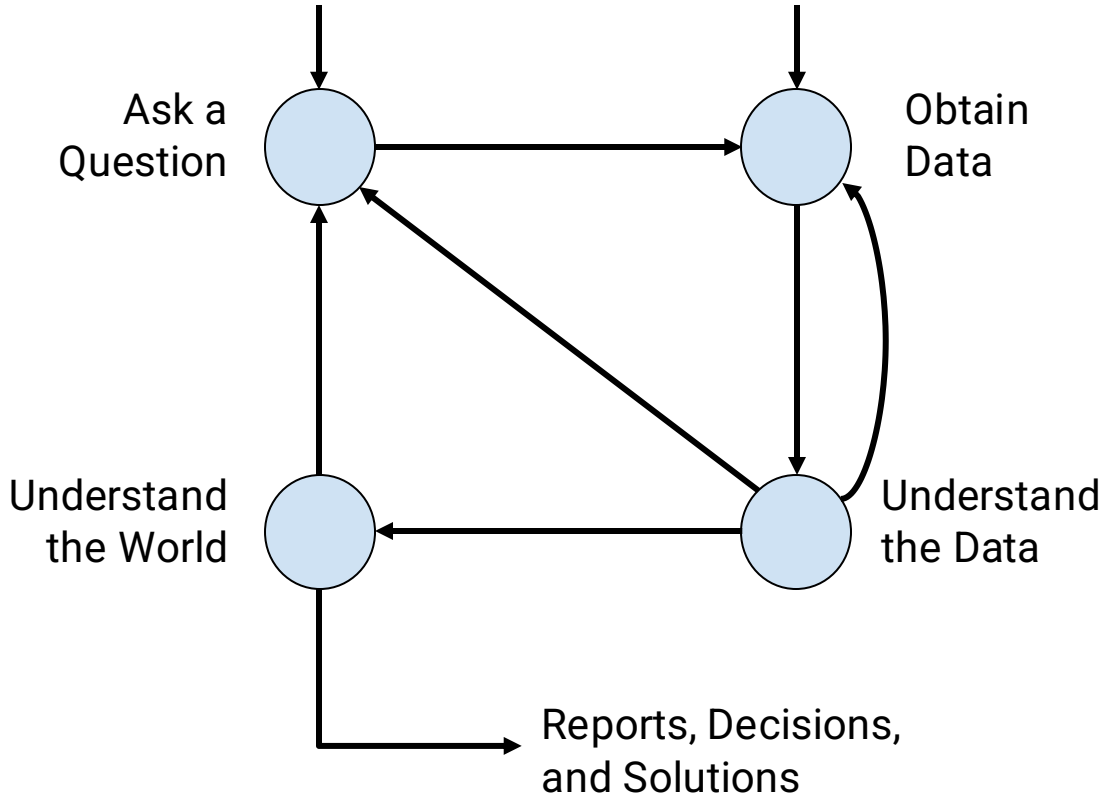
What do you notice? **Why** might this be happening?

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 lifecycle

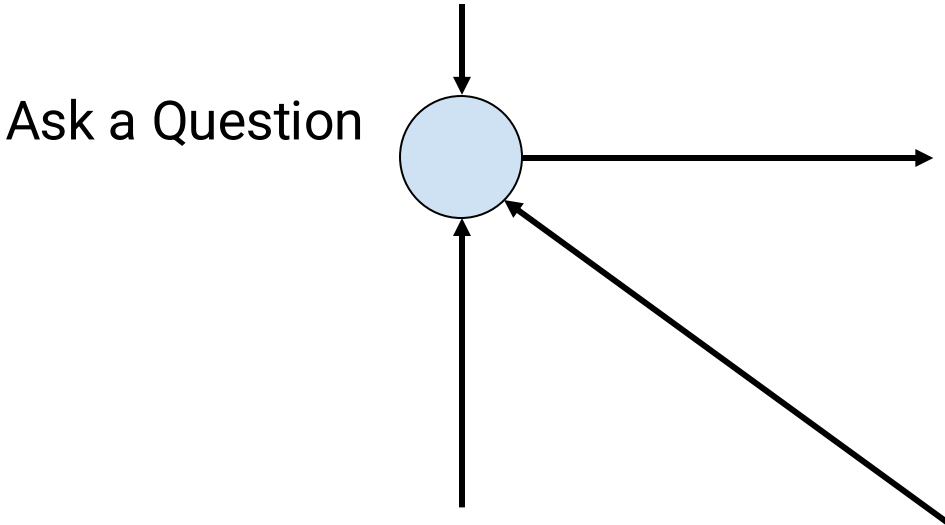


The data science lifecycle is a **high-level description** of the data science workflow.

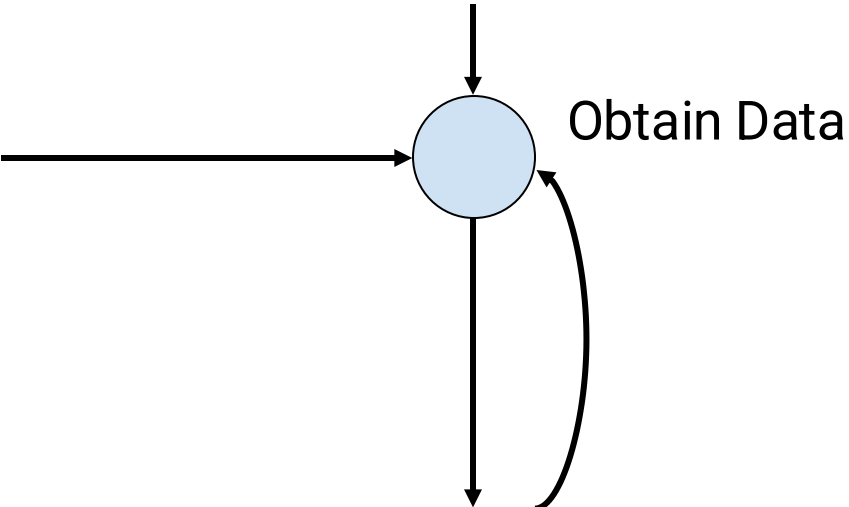
Note the two distinct entry points!

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?

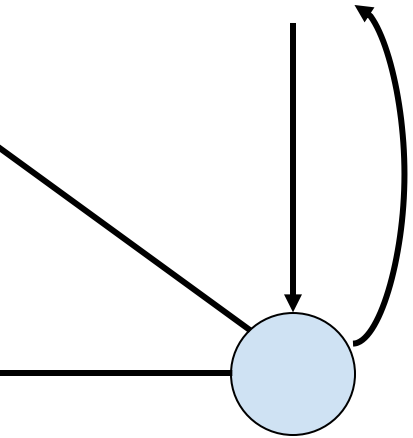


2. Data Acquisition and Cleaning



- 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

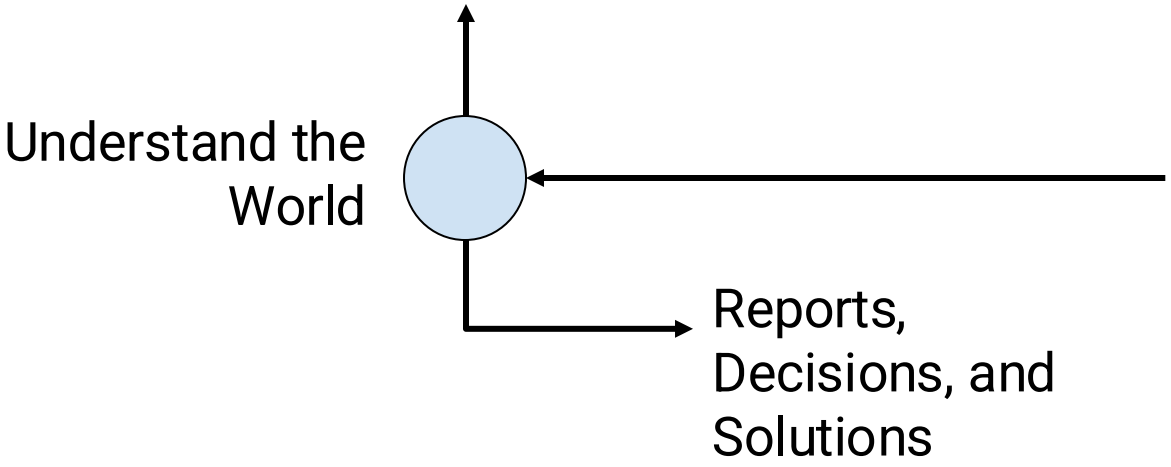


Understand the Data

- 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?



What will you learn in this class?

PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

Course goals

Prepare

Prepare students for advanced 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**.

Topics covered

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

Prerequisites

- CS 210 (formerly numbered CS 310)
- Math 254
- Stat 250
- Working knowledge of Python

Course Logistics

Lecture format

- 3-unit lecture course
- Students will spend 3 hours in lecture and between 9-12 hours outside of class per week
- This is not a do-it-yourself, work-at-your-own-speed course
- The material in this course is technical and will build on each other
- If you fall behind it will be difficult to catch up
- It is important that you do the readings on time, before attending class

Online platform: Canvas
Discussions: Attendance and/or participation in live discussion section can count towards your grade

Course Materials

Materials	Required or optional	Where and how it can be obtained
Principles and Techniques of Data Science, Sam Lau, Joey Gonzalez, and Deb Nolan	Required	https://www.textbook.ds100.org
An Introduction to Statistical Learning, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani	Optional	http://www-bcf.usc.edu/~gareth/ISL/ https://www.stat.berkeley.edu/users/rabbee/s154/ISLR_First_Printing.pdf
Data Science from Scratch, Joel Grus, O'Reilly Media, April 2015	Optional	SDSU Bookstore

Software Setup

- You will need Python3 with JupyterLabs
- There are two ways to do it.
- Easy way: Install it thru Anaconda. https://www.anaconda.com/download/success
- Harder way: Install thru scripts. https://jupyterlab.readthedocs.io/en/stable/index.html
-
- I recommend to install Python3 and JupyterLab thru Anaconda.
-
- If you want to get a bit more fancy with AI tools and Jupyter, then I also suggest ADDING Visual Studio Code to your tools. You still need to install Python3 and the JupyterLab tools from the above step.
- https://code.visualstudio.com/docs/datascience/jupyter-notebooks

Assignments

4 assignments:

- Students need to select the appropriate techniques to analysis a set of data
- Students will spend time over multiple days to complete an assignment
- Assignments are individual effort (not group)

Semester Project

- Students need to find the dataset(s) to investigate an issue
    - What can be learned from the dataset?
    - What are the issues with the dataset?
    - What techniques should they use?
- It is expected that the quality of the project report is high enough that one could submit it to a data science competition like Kaggle or used by San Diego City to justify some action
- Students need to submit project proposals on week 5 of the semester.
- Work on the project is expected to take at least a month

Exams

- There will be exams during week 9, and a final on the date scheduled by the university
- One midterm
- One final exam
- No coding in the exam

Grading

- Mid-term exam 25%
- Final exam 30%
- Assignments 30%
- Semester project 15%

Collaboration and academic dishonesty

Assignments

Data science is a collaborative activity! It is okay to discuss problems with friends.

- List their names at the top of your assignments. We provide a place to do this.
- You must write your solutions individually! Do not copy any other student's work.
- If we suspect that you have submitted plagiarized work, we will call you in for a meeting. If we then determine that plagiarism has occurred, we reserve the right to give you a negative full score (-100%) or lower on the assignments in question, along with reporting your offense to the Center of Student Conduct.

Exams

- Cheating on exams is a serious offense. We have methods of detecting cheating on exams – so **don't do it**!
- Students caught cheating on any exam will fail this course.

We are here to help!

This class will be moving very quickly. But we have plenty of resources to help you, including

- Office hours
- Several discussion sections.

We really want you to succeed in this class.

- Feel free to reach out to me with any questions or concerns you have.

Welcome to CS577!