Data

What is Data?

- **Data** – Information from a source or sources that can be useful for insights and decision making after being processed in various ways.
- **Data Science** – Interdisciplinary field that uses scientific methods, processes, and systems to extract knowledge and insights from data across various domains.

Where is Data?

- Medical
- Automotive
- Weather
- Social Media
- Written work
- Sensors
- Everywhere

Why is Data?

- Why do we care about data?
  - Understanding trends
  - Predicting future events
  - Decision making
  - Drawing a conclusion of a population based on a smaller subset of the group, which we call a sample
    - Population – Set of every item or individual that is of interest for a particular question
    - Sample – Subset of the population of interest
- Data analytics is the application of data science to a domain specific problem
  - Often involves summarizing findings from data in a report
    - Trends and insights
  - We will not worry too much about separating the terms data science and data analytics
- Data can be used for:
  - Diagnoses of diseases
  - Detecting hate speech
  - Facial recognition
  - Recommending movies / shows / products on Amazon
  - Detecting fraudulent purchases on a credit card

Data Science & Data Science Lifecycle

Data Science

- Data science combines computer science, statistics, and expertise from various domains where the data originated.

Data Science Lifecycle

- Iterative process
  - Data Preparation
  - Data Analysis
  - Data Storytelling

Analogies with making a salad:

- Data Preparation
  - Collecting, cleaning, transforming, integrating, and managing data to structure it for effective analysis, storytelling, and decision-making
  - Getting the ingredients for the salad
- Data Analysis
  - Examining prepared data to uncover valuable insights and guide decision making
  - Creating the salad from the ingredients
- Data Storytelling
  - Communication of data insights through summaries, visualizations, and narratives
  - Presenting the salad in a nice way for customers

Netflix Example

- What is a goal that Netflix might have?
  - Improve customer experience
- What data might they collect?
  - What movies / shows are watched
  - When movies / shows are watched
  - Binge-watching or not
- What can Netflix do with an analysis of this data?
  - Use analysis and visualization of data to steer development of future features / services

Data Science Tools

- Your tasked with calculating the mean average of all the reviews of a product on Amazon.
  - Are you going to manually do this?
- We can use tools to make our job of calculating the mean average of all the reviews easier
- Tools can assist with all stages of the data science lifecycle
  - Code (python, Java, R)
  - Spreadsheet (Excel)
- Cloud – System of computer servers accessible over the internet, allowing us to deal with large amounts of data
  - NASA used cloud services to stream images and videos from Curiosity's Mars landing

Data Form / Appearance

- An easy way to view data is to organize it in a table (tabular form)

`first_name`	`age`	`gpa`	`favorite-pet`	`voted`
Ash	22	2.3	cat	1
Brian	43	3.4	cat	1
Carole	36	2.1	dog	1
Doug	78	2.4	bird	0

- Each row in the table is called an instance / record / sample / object
- Variables / attributes / features / fields are:
  - `first_name`, `age`, `gpa`, `favorite_pet`, `voted`
- A variable / attribute's value is the value (quantitative / qualitative) given to a single instance's variable (ex. 43 for age)
- Dataset – collection of data

Structured & Unstructured Data

- Structured Data – Information organized in a standardized format for efficient access
  - Tables
  - Databases
- Unstructured Data – Information that lacks predefined organization

  - Text
  - Images
  - Videos

- Semi-structured Data – There is some imposed format, but not completely. Sometime looks like a structured object of unstructured data.
  - Tags / keys to assist
  - Emails
  - JSON
  - HTML