# [COGS 9] Discussion
## Reading 2, Data, Pandas

Reading Quiz 2 due on 26th April (Fri)

# Logistics

- For the project, try and answer each question on a new page
- When submitting your work, ensure that the pages are selected correctly corresponding to each question (You can choose multiple pages!)
- Start early, do often! The assignments and project are not difficult

# Assignment 1

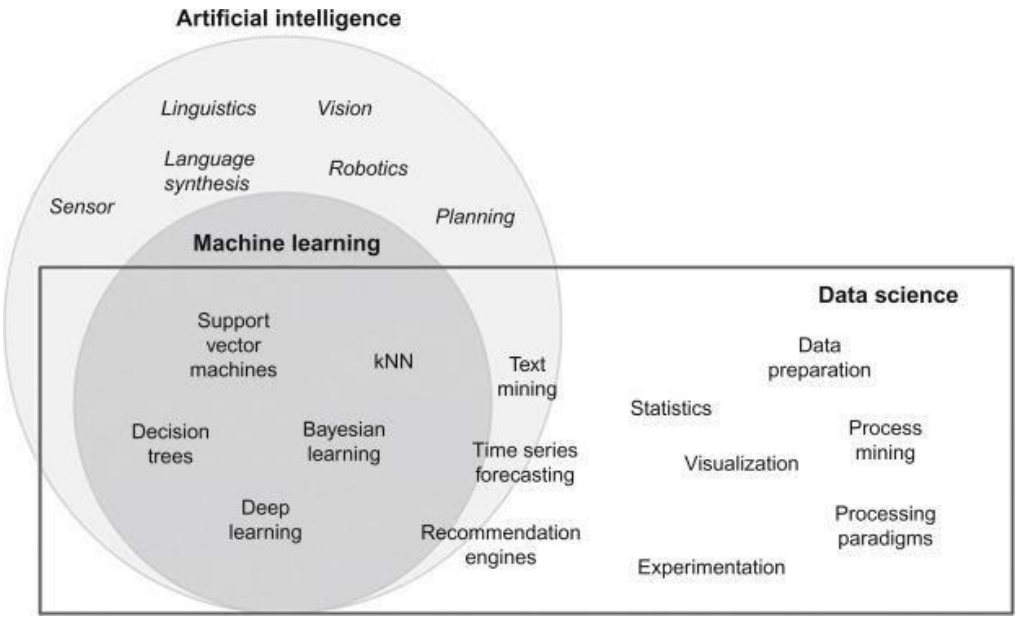
# Final Project Part 1

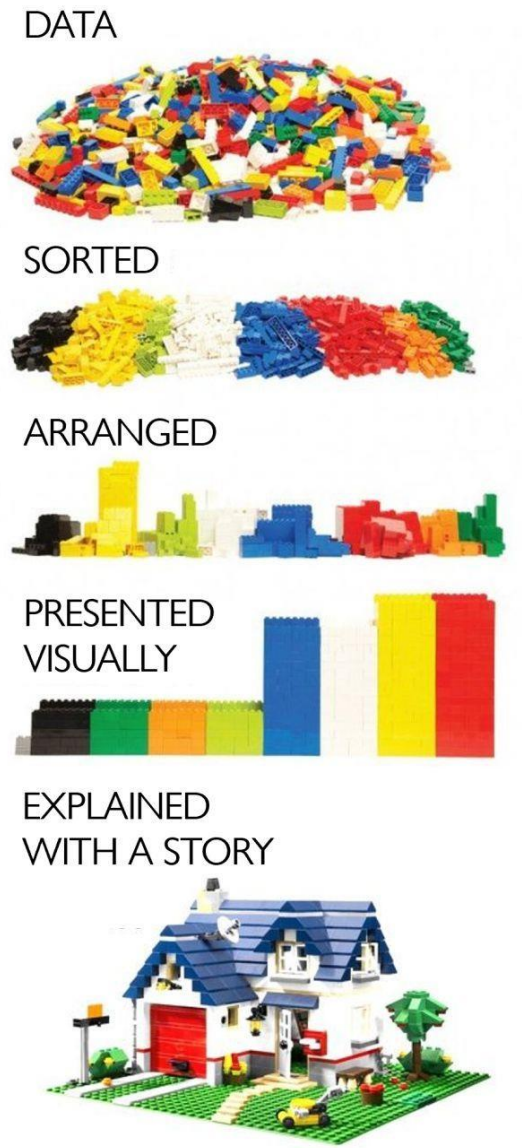
- Make sure you have a group
- Go through lecture 3 for ideas on how to create a data science question
- A list of example datasets have been provided here
- Ethical considerations (10 pts) – Go through Lecture 2 and Reading 2
- I will dedicate my OH and week 6 discussion to provide feedback on your project

# Fun projects

https://openai.com/dall-e-2/

https://projects.fivethirtyeight.com/



# Data's Day of Reckoning

Talks about the increasing role of data in our lives.

There is tension between individual privacy, public good, and corporate profits.

There is a need for responsibility in the creation and management of data and technologies.

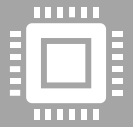
Provides examples of how data has been weaponized and used against us.

A misalignment of incentives between organizations that build and own data products and the people using those products.

The need for a day of reckoning in data science, machine learning, AI, and related technologies.

# Integrating Ethics and Security

Ethics and security are taught but they need to be integrated into the core curriculum and connect them to real-world applications.



In the case of databases, students aren’t taught to think about SQL injection attacks

The importance of continuous training and discussions on ethics and security in the workplace.

# Developing Guiding Principles

Lots of challenges of maintaining ethical principles during project development.

The role of checklists as a tool to ensure ethical considerations are addressed at every stage of a project. Deon’s tool

Need for a set of guiding principles that put good intentions into practice.

# Building ethics into data-driven culture

Ethics should be a part of the culture of an organization

Multiple members => Difficulty in getting consensus

The paper provides pointers on what can be implemented in companies

Lean/agile methodology can help with ethical issues and can allow companies to “move fast and break things”

# Regulation

How does regulation play a role in ethics and data?

Different bodies such as the GDPR which aggressively regulates data use

Even with the strictest of policies, they ultimately tend to lag when compared to the speed at which technology is moving

# Myths and Fallacies of ‘PII’

- Discusses the challenges of developing effective privacy protection technologies as the amount and variety of data collected about individuals increase exponentially.
- The concept of PII is surprisingly difficult to define and is becoming increasingly meaningless as the amount and variety of publicly available information about individuals grows.
- Gives some definitions on what constitutes PII
- Even with very strict protection, users can be indirectly identified through web browsing/purchase history, etc.

# Anonymizing PIIs

- Discusses concepts such as k-anonymity and l-diversity and how having these quasi-identifiers like zip code and gender, and removing person information are assumed to de-identify a user
- However, reidentification is still possible since the de-identified data can simply be combined with other public data sets
- The "safe harbor" provision of the HIPAA Privacy Rule argues that removal of 18 attributes is sufficient for data to be considered properly de-identified. However, as stated above, in the context of PII it is meaningless since the remaining non-removed attributes can simply be combined with a public dataset

# Getting Data

## Download button (Easy)

Find your dataset

Click download

## APIs (Medium)

Choose method

Build URL

Get Authorization/Authentication

## Web Scraping (Hard)

Configure a crawler/spider to pull required HTML pages

Look through the page for your required information

Tidy the data

# API DEMO

- Basic terminal/shell commands (ls, cd, cat, >, |, man, grep, sed, curl)
- curl -X get http://files.rcsb.org/download/10mh.pdb
- Methods: get, put, post, patch, delete (For web)
- URL: The URL from where you want information
- Authorization/Authentication: Bearer tokens, access tokens
- APIs do not necessarily mean web APIs, they can be APIs of, say, a python module as well (Calling a function)

# Web Scraping

- Require a good amount of configuration
- Scrapy is an example of a scraping framework for python
- It uses a spider that systematically goes through the websites you have configured in an ordered manner
- A spider is just a bot that gets you the required data, it is your job to parse and clean the data