[COGS 9] Discussion
Reading 2, Data, Pandas

Reading Quiz 2 due on 3rd Feb (Fri)

Logistics

- For assignments and 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!)
- The course load from next week onwards will be high. There’s a deadline every week (sometimes even 2 in a week!)
- Start early, do often! They are not difficult

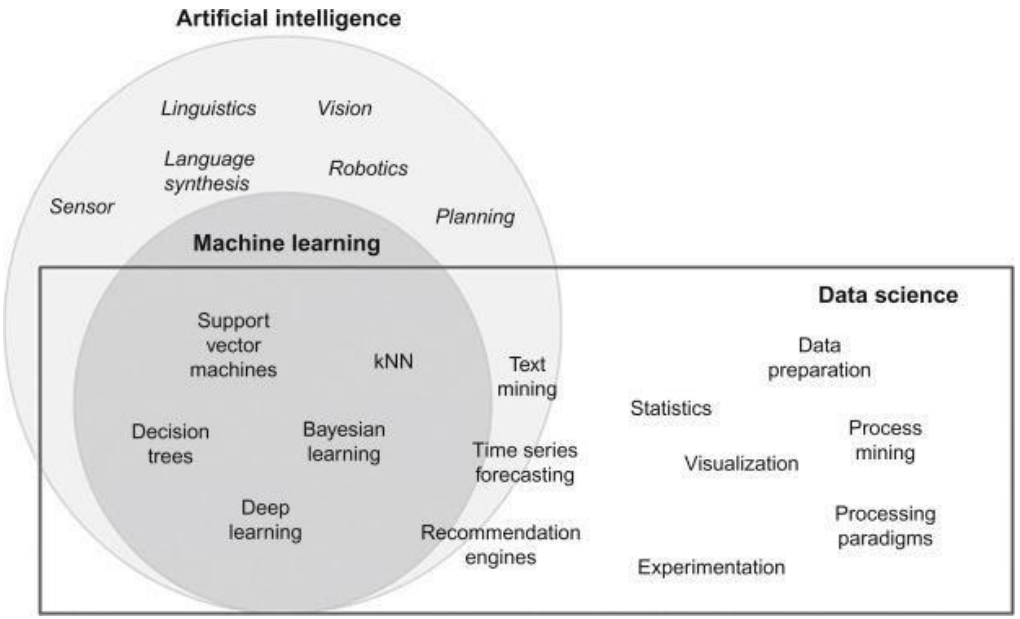
Final Project Part 1

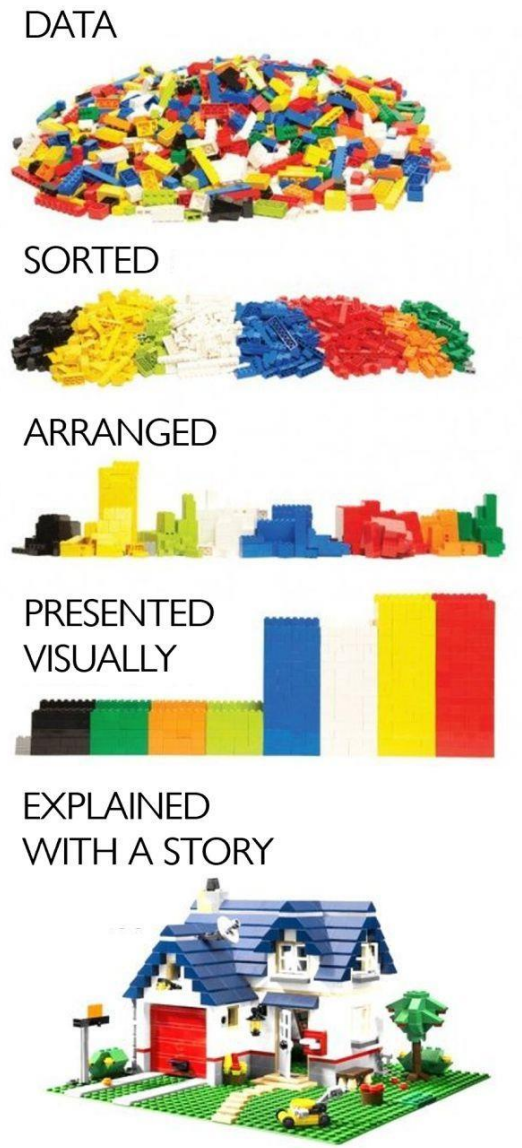
- Make sure you have a group. Check here and search by email
- Go through lecture 3 for ideas on how to create a data science question
- A list of example datasets have been provided here
- Ethical considerations (20 pts) – Go through Lecture 2 and Reading 2
- We will dedicate Week 6 discussion to provide feedback on your project

Fun projects

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

https://projects.fivethirtyeight.com/





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