YaleNUSCollege

Welcome to YSC2239!

Hengnan (Henry) Hu

Spring 2023

Today’s class

- ❑Course organization
- ❑Introduction: What is Data Science?
- ❑Jupyter notebook for Python
- ❑Demo: Huckleberry Fin,
  Little Women

Course staff

- **Instructor:**
  - *Hengnan Hu*
    - **Email:** henry.hu@nus.edu.sg
    - **Office Hours:** By appointment over Zoom

- **Lectures**:
  - Time: Monday, Thurs 7pm – 8:30pm
  - Location: Y-LT1
  - Remotely (need to check with IT and update over Canvas)

Course staff (Peer tutors)

•John Jacob Go	
Email:	jacobgo@u.yale-nus.edu.sg

•Nihal Zuhayar Parash Miaji	
Email:	nihalzuhayar@u.yale-nus.edu.sg

Weekly drop-in sessions

- Weekly drop-in sessions will begin on **Week 2**

- Detailed schedule and venue: TBA

Social distancing policy

- We suggest the following social distancing policy:
- 1. Wear masks in lectures, drop-in sessions and
- 2. Keep appropriate distance between you and other classmates, peer tutors and professors
- 3. Maintain good personal hygiene at all times

Course contents

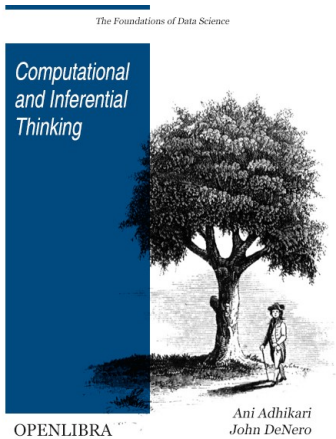
- Introduction of Python for data science: tables, data types, charts, histograms, functions, groups, joins, iteration
- Statistics: chance, sampling, models, distributions, A/B testing, confidence intervals, central limit theorem, correlation, p-values
- Machine learning: linear regression, multiple linear regression, regression diagnostics, feature engineering, logistic regression, classification, clustering, decision tree

Course resources

- We will follow the lecture slides and demo of the instructor.

- I strongly encourage bringing your laptop to lecture and play around with the demo yourself in real-time for best learning experience.

Course resources



Text for the first-half:

- Title: Computational and Inferential Thinking: The Foundations of Data Science
- Authors: Ani Adhikari and John DeNero
- Link: https://inferentialthinking.com/chapters/intro.html

- We will use Canvas for communication and posting of material

Assessment Scheme

COURSE ASSESSMENT BREAKDOWN

Attendance and participation 5%
Assignments 20%
Labs 15%
Project 8%
Peer evaluation 2%
Midterm 20%
Final 30%

Assessment Scheme

Attendance: QR code at the beginning of each lecture + 4 in-class quizzes

Assignments: weekly in the first half of the course. First assignment will be posted on Week 2, due on Friday 23:59.

Labs: weekly. First lab will be posted on Week 1, due on Wednesday 23:59.

Midterm exam: March 2nd (Thursday) 7pm – 8:30pm, lecture-time, in-person. Please block this timeslot. No make-up exams will be given for midterm exam.

Final exam: April 28th (Friday) 3pm – 5pm, please take note this timeslot. No make-up exams will be given for final exam.

Project

The course culminates in a final group project in which students are expected to create appropriate data science models using the methods covered in class for data analysis. Each group composes of 2-3 students, except in special circumstances approved by the instructor.

Further details on the projects will be provided as the due dates approach.

Late work policy

- Checkout the syllabus on penalty for late work policy

Other policies and resources

- Checkout the syllabus on penalty for academic integrity policy, intellectual property and privacy, class climate, etc.

Today’s class

☑Course organization

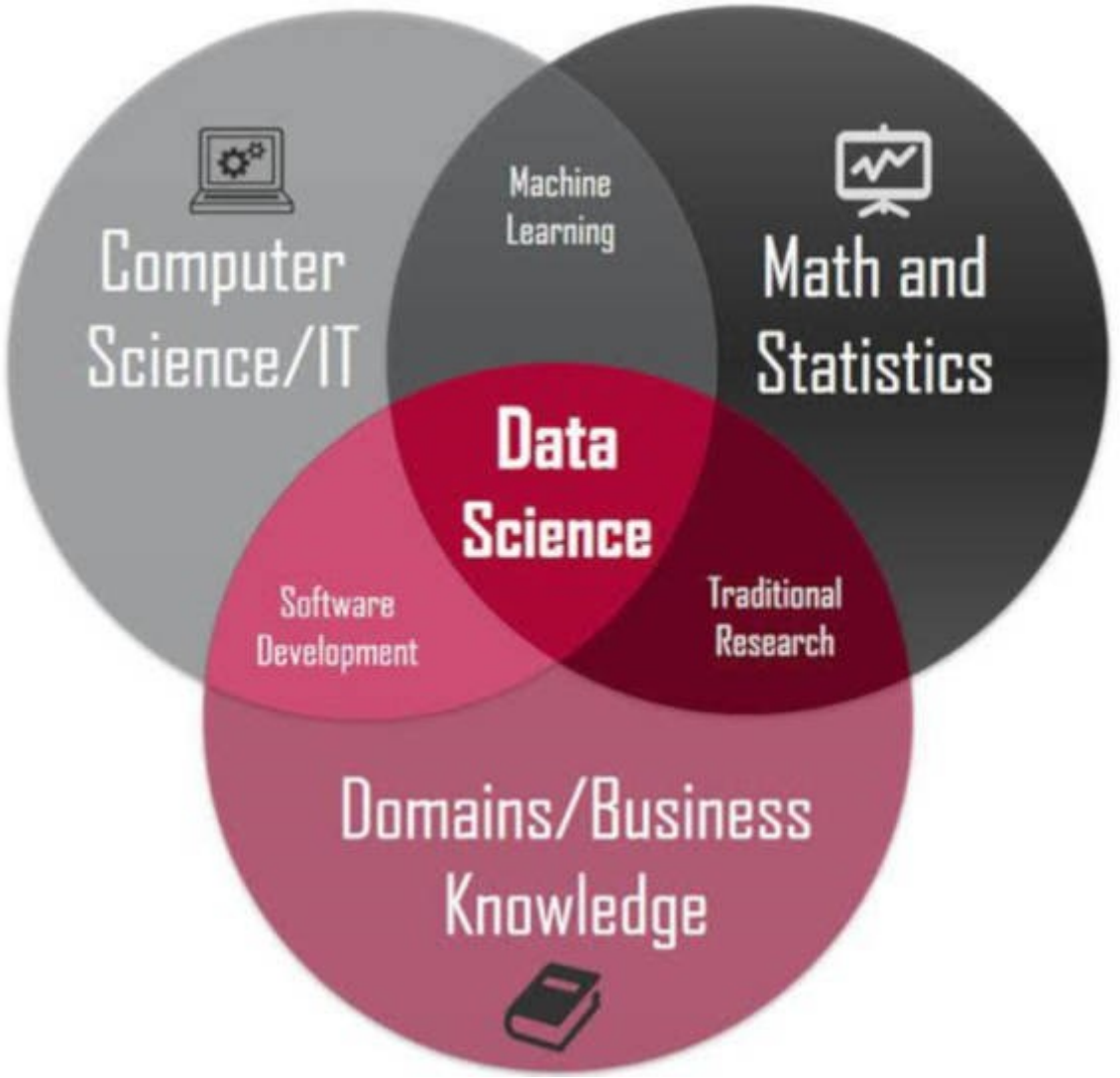
❑Introduction: What is Data Science?

❑Jupyter notebook for Python

❑Demo: Huckleberry Fin,
Little Women

Data Science

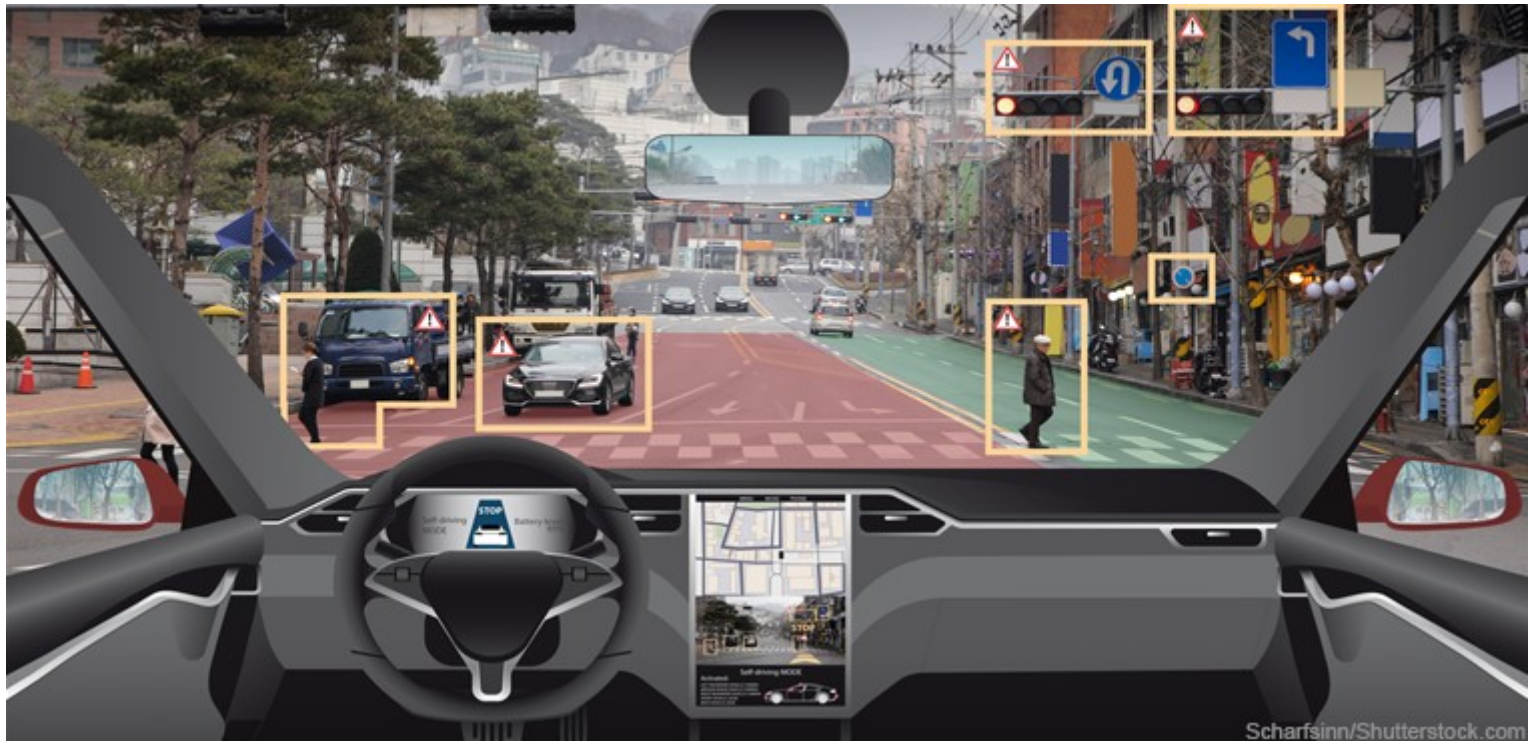
Reading: Book Chapter 1



Data is the new oil

- We now live in a time of big data
- Ability to leverage these data to make intelligent decisions and to help designing faster and better algorithms are very important
- A wide range of fields that rely on data-driven decision making, e.g. Biology, Economics, self-driving cars etc..

Applications of data science: computer vision and self-driving cars



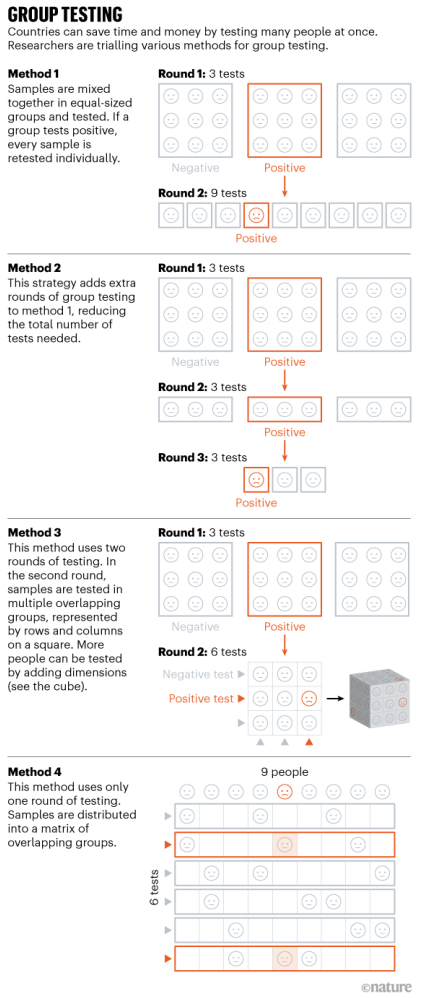
Applications of data science: accelerate scientific discovery in protein folding

- Google Deepmind has developed AlphaFold that can accurately predict 3D models of protein structures and is accelerating research in nearly every field of biology:
- https://www.deepmind.com/research/highlighted-research/alphafold

Applications of data science: combat COVID-19

- Group testing

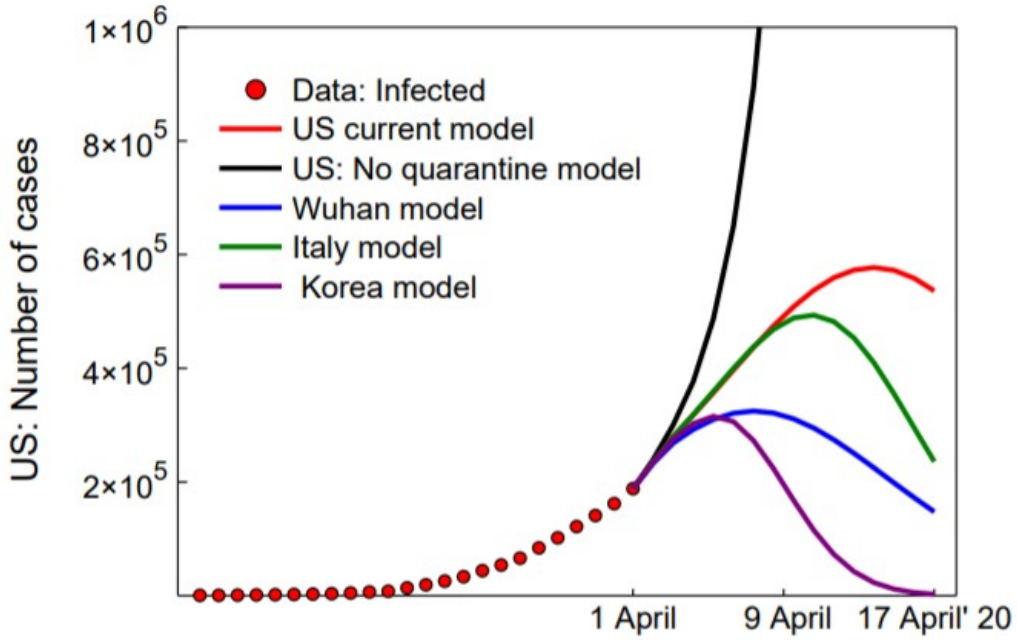
- https://www.nature.com/articles/d41586-020-02053-6



Applications of data science: combat COVID-19

- Forecasting models

- https://projects.fivethirtyeight.com/covid-forecasts/



Today’s class

☑Course organization

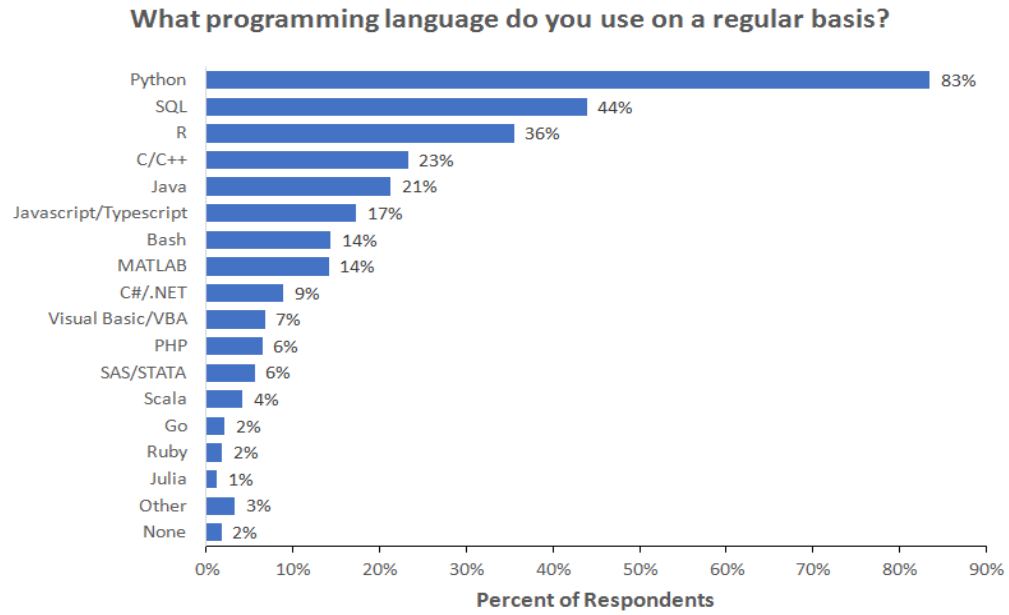
☑Introduction: What is Data Science?

❑Jupyter notebook for Python

❑Demo: Huckleberry Fin,
Little Women

Python

- Python is a very popular computing language for data science and software development



Note: Data are from the 2018 Kaggle Machine Learning and Data Science Survey. You can learn more about the study here: http://www.kaggle.com/kaggle/kaggle-survey-2018. A total of 18827 respondents answered the question.

BUSINESS BROADWAY
DATA SCIENCE | CUSTOMER ANALYTICS | MACHINE LEARNING

Jupyter notebook

- In this course, we will be using Jupyter notebook with Python **3** for all of the Labs, Assignments and Exams!

- Install the latest version: https://www.anaconda.com/distribution/

- Checkout the installation guide on Canvas

- Please let the instructors or peer tutors know if you have any issue on installation

datascience package in Python

- We shall need the datascience package in Python for the first half of the course

- Checkout the installation guide on Canvas

Python packages that we will learn in this course

- Datascience
- Pandas
- Scikit-learn
- NumPy
- Matplotlib

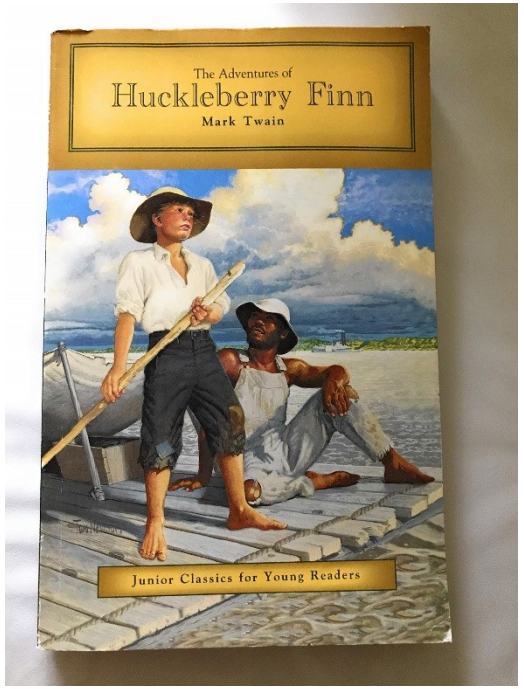
Today’s class

☑Course organization

☑Introduction: What is Data Science?

☑Jupyter notebook for Python

❑Demo: Huckleberry Fin,
Little Women



To do

- Install Jupyter notebook and the datascience package
- Lab 0 (ungraded)