You can download the slides from this link

NYCU NLP

Introduction of the Course

顏安孜

azyen@nycu.edu.tw

Online Classes or In-Person Classes?

- In-person and synchronous online options are provided
- In-person classes:
  - EC115
- Synchronous online classes
  - The link to the online class: https://meet.google.com/wpz-haxx-npd
  - Please use your NYCU email
  - Please turn off your microphone during the class
  - If you have any questions, please leave a message
- If you feel symptoms of COVID-19 or any discomfort, please join the online class via Google Meet
- If it is necessary to call off in-person classes, I will inform you via email

- 這學期基本上以實體與同步線上課程進行，同學可自由選擇修課方式，如當週須改僅線上課程方式進行，會另行通知
- 若自己或近期曾接觸過的人有相關症狀，就不必勉強參與實體課程，可選擇加入線上課程

Synchronous and Asynchronous classes

- Two synchronous classes every Tuesday (according to the calendar)
  - In-person and online classes (13:20~15:10)
- One asynchronous class (almost per week, depend on actual progress)
  - Videos
  - I will upload the videos to Google Drive. The link will be provided on the E3 platform.

Course Description

- Association Analysis
- Machine Learning and Deep Learning for Data Mining
  - Linear regression, Logistic regression, Neural Network, etc.
- Generative AI for Data Mining Applications

The schedule may differ from the weekly course schedule shown on the website before.

Homework and Final Project

- 3 homework in total　3次個人作業，各項作業規定於公告作業時說明
  - You are asked to complete the homework independently
  - We will hold a Kaggle competition for each homework
  - (Other details will be announced)
- 1 final project and 1 group presentation　1個期末專題與1個分組報告
  - The topic of the final project will be announced around mid-term
  - Work in groups (2 to 4 students in a group)
  - Please fill out the excel table with the link below if you find the group members
  - Excel table: https://docs.google.com/spreadsheets/d/1becHl6MSGsR43NFRXSD0miowj8yW47V90xM4LZO2iMs/edit?usp=sharing

Evaluation

- Homework (**Individual**): 60% of your semester grade
  - Rating based on your Kaggle ranking and report
  - There are three assignments this semester
- Final Project (**Group**): 30% of your semester grade
  - Rating based on your Kaggle ranking and report
  - There are one final project this semester
- We will set a baseline for all assignments and the final project.
- Detailed requirements and evaluation for assignments and the final project will be announced
- Presentation (**Group**): 10% of your semester grade
  - Each group is asked to select 3 papers on data mining topics and prepare slides for a 20-minute presentation **in English**
  - 5% rated by peer assessment, and 5% rated by teaching assistants
- Late submissions are not accepted for all assignments and final projects
- I reserve the right to the evaluation policy and final grade adjustment
- Plagiarism or cheating on assignments or projects ➔ Semester grade will be 0

評量方式

- Homework (Individual): **3**次作業佔學期成績 **60%**
  - Kaggle 比賽排名、需繳交電子檔報告
  - 評分細則於每次作業公告時說明
- Final Project (Group): 期末專題佔學期成績 **30%**
  - Kaggle 比賽排名、需繳交電子檔報告
  - 評分細則於公告時說明
- Homework 和 Final Project 都會設定 baseline
- Presentation (Group): 分組報告佔學期成績 **10%**
  - 各組閱讀 DM 相關議題的3篇論文，準備20分鐘左右的線上英文報告
  - 其中5%各組互評，5%助教評分
- 作業和期末專題皆不接受遲交、補交
- 遇到任何問題，請盡快向老師和助教反應
- 老師保有調整評量方式、加分、最終調分之權利，若**作業或專題有抄襲或作弊的情形，學期成績 0 分**

Homework vs. Final Project

- Q: What is the difference between homework and the final project?
- A: For the homework, I will ask you to implement the methods taught in the classes.

  For the final project, the method is not limited, so I will ask everyone to analyze and discuss the results of different methods.

Report Requirements

- Please describe your ideas and implementation in detail within the specified number of pages, discuss the experimental results, analyze and discuss based on the experimental results
- Please discuss the characteristics of the task and experimental results as much as possible, instead of describing the process of tuning parameters or the challenge of writing programs
- Example:
  - Q: Please describe your methods
  - Lower-scoring answer:
  - A: I used BERT. And I ensembled all models with different parameters so as to improve the performance.
  - Better answer:
    - A: I implemented BERT, considering that the data in this assignment has the nature of @#$%^, I #$%^&*. Finally, as shown in Table 1, the results of &^%$# are better. Finally, I used ensemble learning. So that when I decided on the classification result, I could refer to the decisions of different models. The experimental results show that the ensemble learning method improves the performance of predicting the classes which just have a few training samples.

報告內容規定

- 請在規定頁數範圍內詳述自己的想法與做法、討論實驗結果、根據實驗結果進行分析與討論
- 盡可能對任務的特色與相關實驗結果進行討論，減少單純調參數或如何寫程式的說明
- 範例:
  - Q: 請描述你的方法
  - 可能得到較低分的答案:
    - A: 我用 BERT，為了改善效能，最後 ensemble 所有不同參數的 model。
  - 較好的答案:
    - A: 我實作 BERT，考量到這個作業中的資料有@#$%^的性質，我將 BERT #$%^&* 之後，如表1顯示，在 &^%$# 的面向得到更好的結果，所以最後我使用 ensemble learning，使得最後再決定預測結果時，可以參考不同模型的決策，最後的實驗結果證實這個做法可以在數量較少的類別上得到更好的分類結果。

Topics for Group Presentation

NYCU NLP

- Major Conferences:

1. KDD - Knowledge Discovery and Data Mining
    - https://kdd.org/kdd2022/toc.html
    - Find papers you are interested in from Research Track Full Papers and ADS Track Papers

2. CIKM - International Conference on Information and Knowledge Management
    - https://www.cikm2022.org/papers-posters

3. WSDM - Web Search and Data Mining
    - https://www.wsdm-conference.org/2023/program/accepted-papers

Calendar

- 3/19(Tue) HW1 Announcement
- 4/9(Tue) Midterm (No Class)
- 4/16(Tue) HW1 Deadline & HW2 Announcement & Deadline of filing the group members
- 4/30(Tue) Final Project Announcement
- 5/14(Tue) HW2 Deadline & HW3 Announcement (No Class – Break for WWW 2024)
- 5/21(Tue) Deadline of submitting your research topic (for group presentation)
- 5/28(Tue) Group Presentation (Asynchronous & No Class)
- 6/4(Tue) Final Exam (No Class)
- 6/11 (Tue) & 6/18(Tue) No Class
- 6/21(Fri) Final Project Deadline & HW3 Deadline

How can I enroll this course?

- If you would like to apply for course enrollment:
  - Please fill the form with your information (e.g., department, student ID, email, and contact number), and I will send the list to the department office
  - If there are too many students applying to this course, I will randomly select students from the list
- The link and QR code to the form:
  - https://forms.gle/VwoKYmtZvx7Yoocc6
- I will announce the decision by the end of today's class

Teaching Assistant

- 何冠儀 joy861106.cs11@nycu.edu.tw
- 徐唯凌 weiling.hsu.cs11@nycu.edu.tw
- 劉宇承 liu2022113.cs11@nycu.edu.tw
- 陳柏瑋 h7a4n1k.cs12@nycu.edu.tw

- Office hours:
  - To be announced

What is Data Mining?

Data Mining

- aka knowledge discovery in data (KDD)
- combines statistics and artificial intelligence to analyze large datasets to discover useful information
- Data Warehousing

Data Warehousing & Data Mining

- Data mining is defined as the analytical process of identifying patterns and trends within data, whereas data warehousing refers to the systematic collection of data from various sources
- Data mining involves a deep dive into the collected data to extract trends, patterns, and insights that are not immediately visible
- Data mining's ability to forecast future trends based on historical data positions it as a valuable tool for predictive analytics

https://chat.openai.com

Four steps of data mining

- Set objectives
  - define a business problem
- Data preparation
  - clean and remove any noise in our data (e.g., duplicates, missing values, and outliers)
- Model building and pattern mining
  - investigate any interesting data relationships
  - apply machine learning techniques
- Evaluation of results and implementation of knowledge

Machine learning applications

- Supervised Learning
  - Risk evaluation
  - Forecast sales
- Unsupervised Learning
  - Recommendation
  - Anomaly detection
- Reinforcement Learning
  - Self driving cars
  - Gaming
  - Online advertising

Data mining techniques

- Association rules
- Regression
- Classification
- Clustering

Association Rule Discovery

- Given a set of records each of which contain some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Source: Introduction to data mining

Regression

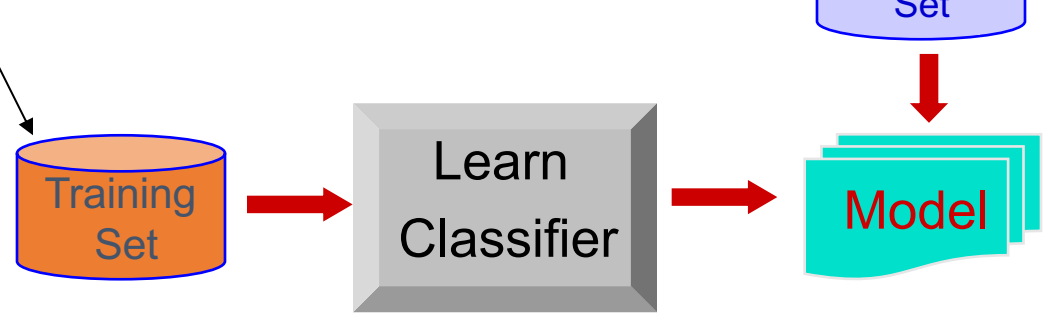
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Modeling the relationship between patient characteristics and health outcomes, such as mortality rates and disease progression.
  - Time series prediction of stock market indices.

Source: Introduction to data mining

Classification

Tid	Employed	Level of Education	# years at present address	Credit Worthy
	categorical	categorical	quantitative	class
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

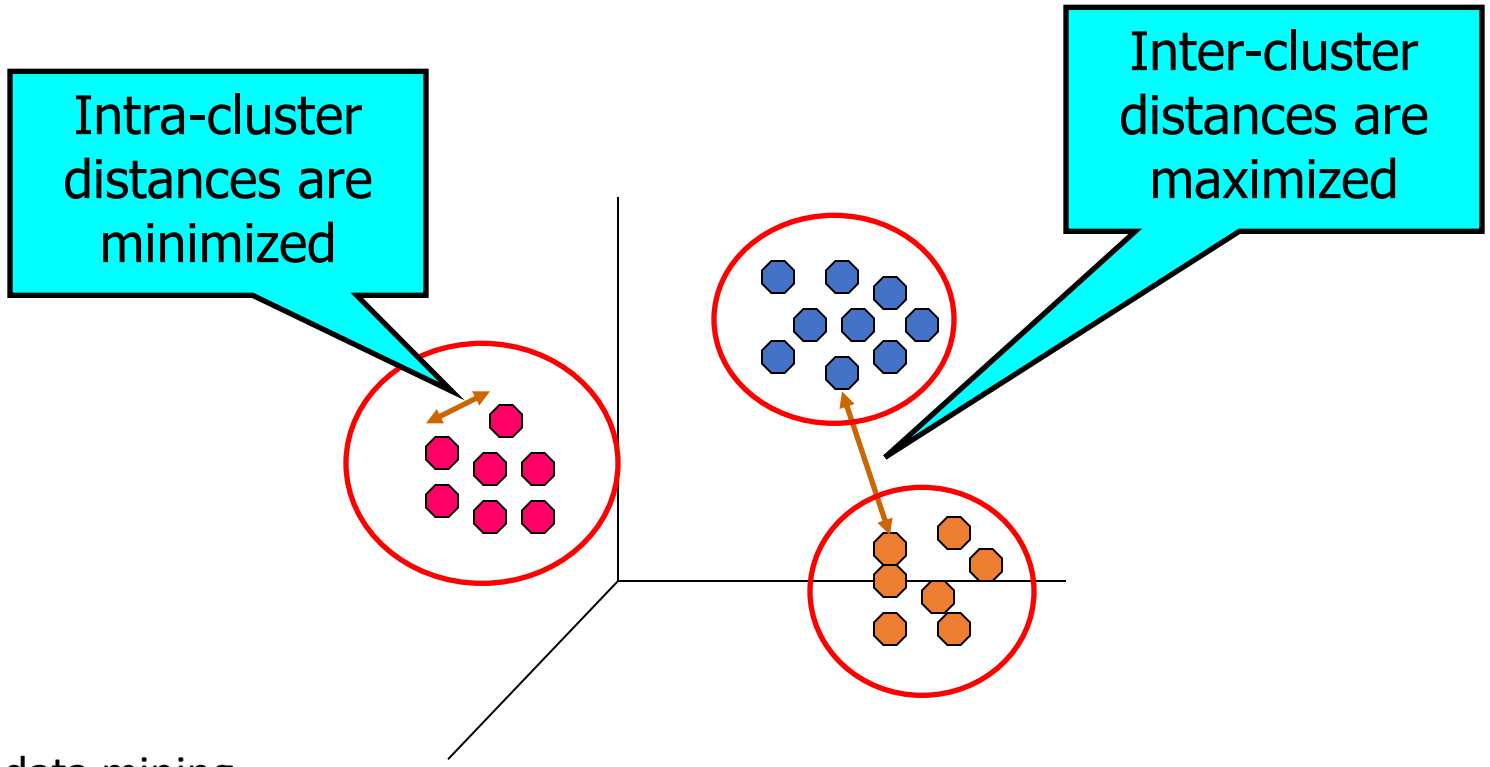
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



Source: Introduction to data mining

Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



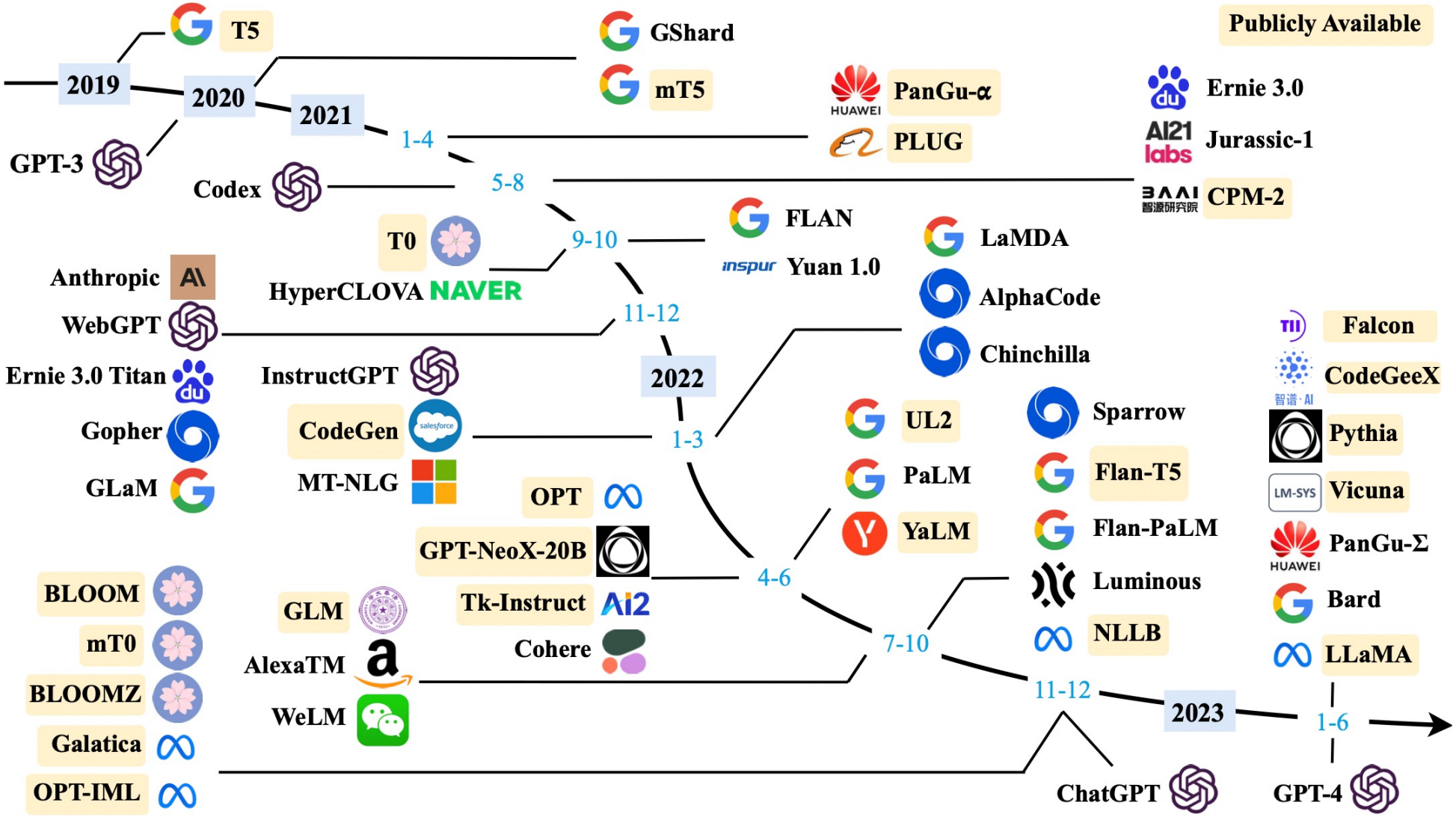
Source: Introduction to data mining

Motivating Challenges

- Scalability
  - big data
- High Dimensionality
  - hundreds or thousands of attributes in dataset
- Heterogeneous and Complex Data
  - complex objects: temporal and spatial autocorrelation, graph connectivity, etc.
  - need for techniques that can handle attributes of different types
- Data Ownership and Distribution
  - data security and privacy issues
- Non-traditional Analysis

Source: Introduction to data mining

Large Language Models



https://arxiv.org/pdf/2303.18223.pdf

NYCU NLP

The decision of who can enroll this course

Reference

- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. Introduction to data mining. Pearson Education India, 2016.
- https://www-users.cse.umn.edu/~kumar001/dmbook/slides/chap1_intro.pdf