CS313: Data Mining and Application

Course Overview and Introduction to Data Mining

Vo Nguyen Le Duy

University of Information Technology, VNUHCM
RIKEN, Japan

duyvnl@uit.edu.vn
http://vonguyenleduy.github.io

Fall 2025

Course Information

- Instructor: Vo Nguyen Le Duy
  - Email: duyvnl@uit.edu.vn
  - Homepage: http://vonguyenleduy.github.io
  - Office: E5.8
- You must attend every class unless you have good reasons

What We Learn in This Course

- Foundations and practical applications of data mining techniques
  - Knowledge discovery process
  - Data preprocessing
  - Data mining techniques
- Python programming for data mining tasks
- In addition, I hope you have a nice experience on having an English lecture

Expected Outcomes

- Explaining the basic concepts and terminologies in data mining
- Distinguishing the stages in knowledge discovery process
- Understanding the working mechanism of machine learning applied in data mining
- Being able to utilize softwares for exploring and analyzing data
- Developing teamwork skills and presentation skills

Course Outline

- **Part A: Lectures**
  - W1 (Week 1): Course overview and introduction to data mining
  - W2: Data pre-processing
  - W3: Regression + Lasso + Elastic Net
  - W4: Statistical test for machine learning

- **Part C: Lectures + Midterm Project Presentation**
  - W5
    - Lecture on Classification
    - Presentations by G1 (Group 1), G2 and G3
  - W6
    - Lecture on Clustering
    - Presentations by G4, G5, and G6

- W7
  - Lecture on Itemset Mining/Sequential Pattern Mining
  - Presentations by G7, G8, and G9
- W8
  - Lecture on Trajectory Data Mining
  - Presentations by G10, G11, and G12

- **Part C: NVIDIA Workshop: Applications of AI for Anomaly Detection**
  - W9: XGBoost
  - W10: AutoEncoder (AE)
  - W11: Generative Adversarial Network (GAN) + Learning Assessment

- **Part D: Final Project Presentation**
  - W12 - W15
  - 3 groups/week

Grading

- (35%) Random Assessments
- (15%) NVIDIA Certificate + Learning Assessment
- (50%) Project (Midterm + Final Presentation)
  1. Application domains: Education/Bioinformatics/Agriculture/Environmental Sustainability/Geographic
  2. Itemset Mining or Sequential Mining or Trajectory Data Mining
  3. Machine Learning
  4. Experimental Evaluation
  5. Web Demo
  6. English report (for final presentation) with 3 or 4 pages
- Note
  - Absent from 3 lectures: -1 point on total score
  - $+0.25$ point for each "good" question or answer

Group Formation

Reference Materials

- Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.
- David J. Hand, Heikki Mannila and Padhraic Smyth. *Principles of Data Mining*. The MIT Press, 2001.

Any Questions?

- Your feedback is always welcome
- Please feel free to share your concerns with the class

What Is Data Mining?

- Broad view: data mining is the process of discovering interesting patterns and knowledge from large amounts of data

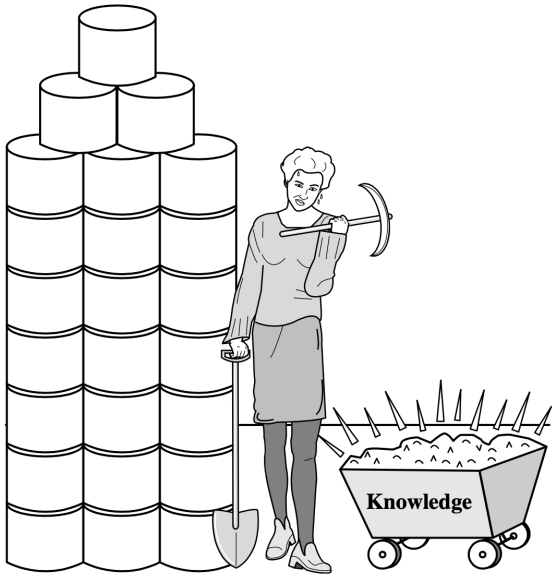


Figure 1: Data mining (Han et al. 2022)

- Data mining is also called *knowledge discovery from data (KDD)*
- Another view: data mining is a step in the process of knowledge discovery

Process of Knowledge Discovery

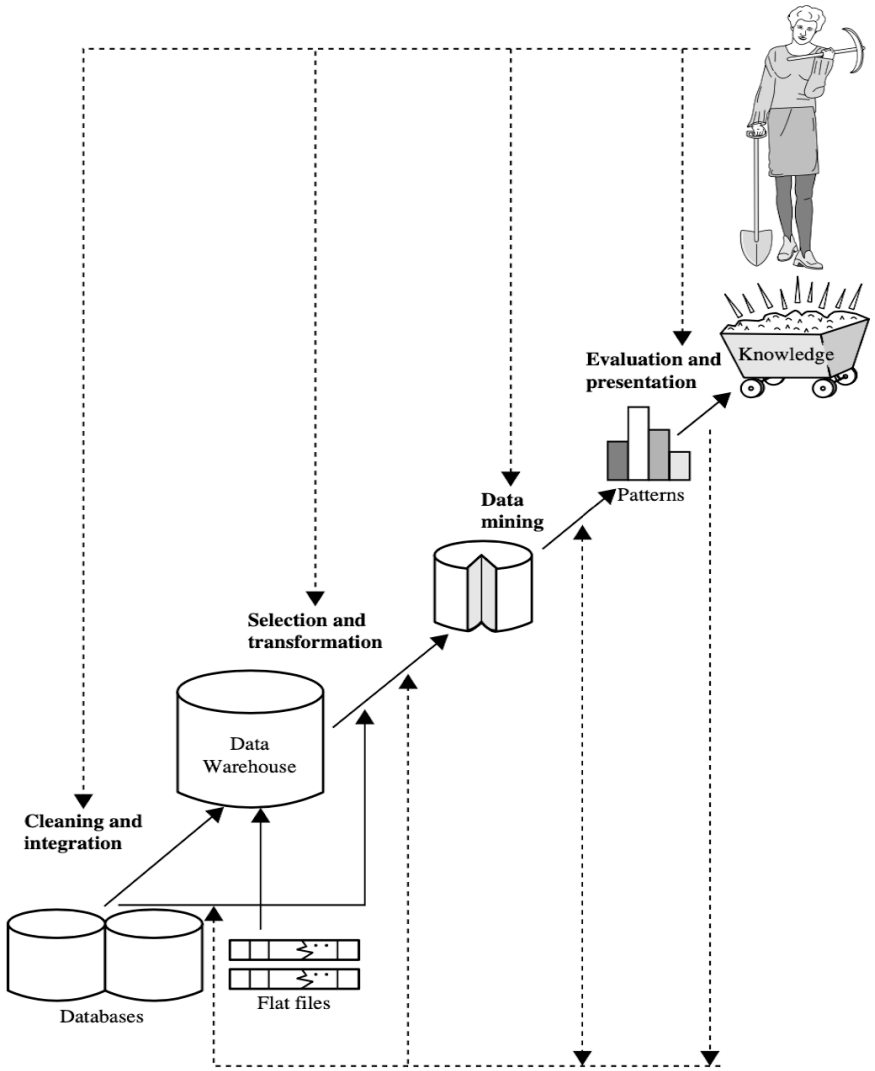


Figure 2: Process of knowledge discovery (Han et al. 2022)

Applications

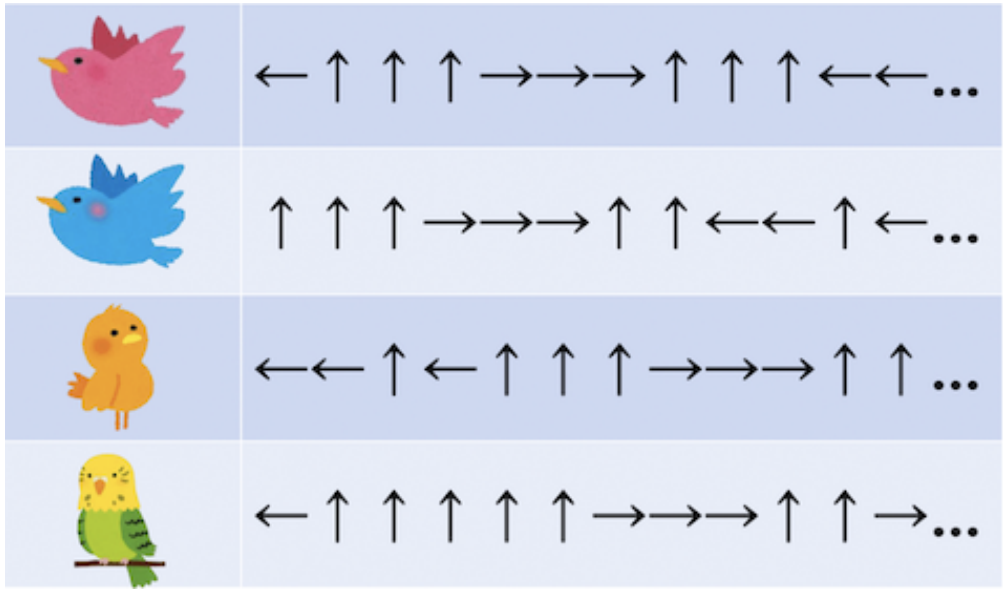
- Retail and marketing
  - Market basket analysis

Takeuchi. (2020)

  - Customer segmentation
- Finance and Banking
  - Fraud detection
  - Risk analysis

Applications

- Animal behavior analysis

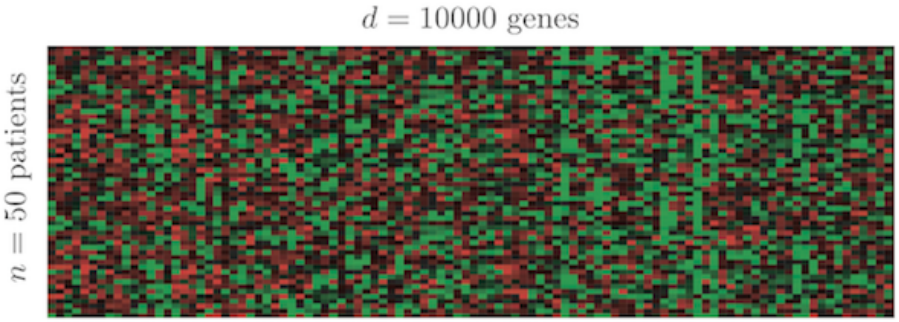


Hanada. (2023)

- Migration Patterns
- Movement Analysis

Applications

- Bioinformatics
  - Genetic data analysis



Takeuchi. (2022)

  - Drug discovery

Compounds	Character 1	Character 2	Character 3	Efficacy
	1.7	2.5	-0.4	2.7
	2.8	-1.9	1.3	-0.9
	0.6	2.2	1.7	0.2
	1.5	1.3	2.4	2.3

Hanada. (2023)

Challenges in Data Mining

- Mining methodology
- Efficiency and scalability
- Diversity of data types
- Privacy

Data Mining, Big Data, and Data Science

- Big data refers massive datasets

Data Mining, Big Data, and Data Science

- Big data refers massive datasets
  - Social media data (Facebook, Twitter, and Instagram)
  - IoT data (sensors and devices)
  - Gaming data
  - Sports data

Data Mining, Big Data, and Data Science

- Big data refers massive datasets
  - Social media data (Facebook, Twitter, and Instagram)
  - IoT data (sensors and devices)
  - Gaming data
  - Sports data
- Data science is a broad field that includes data mining

Types of Data

- Transactional data
- Sequence data
  - Time-series data
  - Biological sequence data
- Spatial and spatial-temporal data
- Graph data
- Hypertext and multi-media data
  - Text, image, video, and audio data

Transactional Data

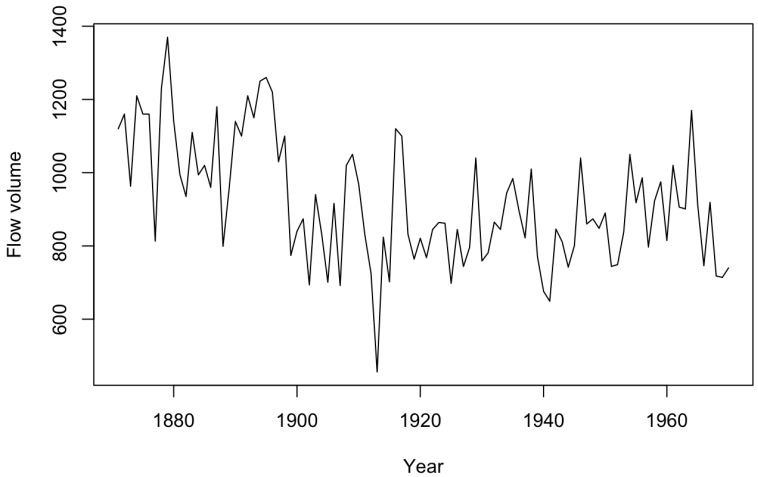
- Customer's purchase, a flight booking, or a user's clicks on a web page
- A transaction: transaction ID + list of the **items**
  - Additional tables might be included (e.g., item description)
- Example: *Which items sold well together?*

Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

Sequence Data

- Time-series data: a series of data points indexed in time order
- Examples of time-series analysis:
  - Earthquake forecasting
  - Annual river flow analysis



Annual flow volume of the Nile river

Sequence Data

- Biological sequence data

Species	Alignment of Amino Acid Sequences of β-globin
Human	1 VHLTPEEKSA VTALWGKVNV DEVGGEALGR LLVVYPWTQR FFESFGDLST
Monkey	1 VHLTPEEKNA VTTLWGKVNV DEVGGEALGR LLLVYPWTQR FFESFGDLSS
Gibbon	1 VHLTPEEKSA VTALWGKVNV DEVGGEALGR LLVVYPWTQR FFESFGDLST
Human	51 PDAVMGNPKV KAHGKKVLGA FSDGLAHLDN LKGTFATLSE LHCDKLHVDP
Monkey	51 PDAVMGNPKV KAHGKKVLGA FSDGLNHLDN LKGTFAQLSE LHCDKLHVDP
Gibbon	51 PDAVMGNPKV KAHGKKVLGA FSDGLAHLDN LKGTFAQLSE LHCDKLHVDP
Human	101 ENFRLLGNVL VCVLAHHFGK EFTPPVQAAY QKVVAGVANA LAHKYH
Monkey	101 ENFKLLGNVL VCVLAHHFGK EFTPQVQAAY QKVVAGVANA LAHKYH
Gibbon	101 ENFRLLGNVL VCVLAHHFGK EFTPQVQAAY QKVVAGVANA LAHKYH

4. What other evidence could you use to suppor your hypothesis?

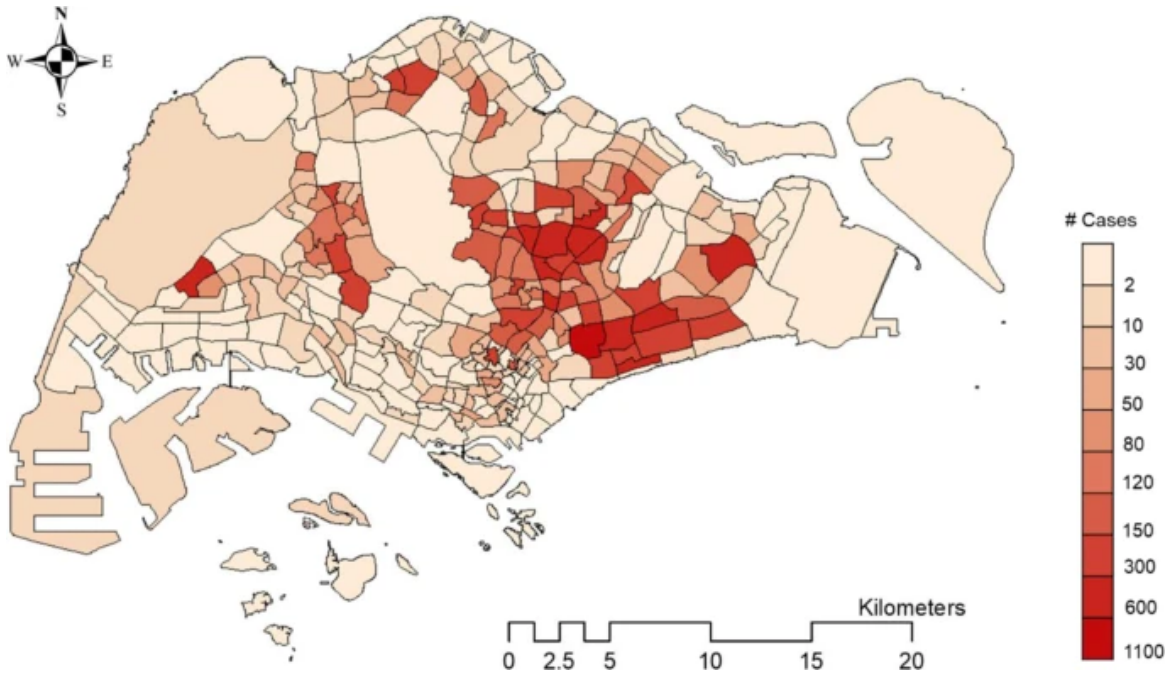
MB A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

Data from Human: http://www.ncbi.nlm.nih.gov/protein/AAA21113.1; rhesus monkey: http://www.ncbi.nlm.nih.gov/protein/122634; gibbon: http://www.ncbi.nlm.nih.gov/protein/122616

- Event Sequence Data
  - Online game
  - Soccer

Spatial Data

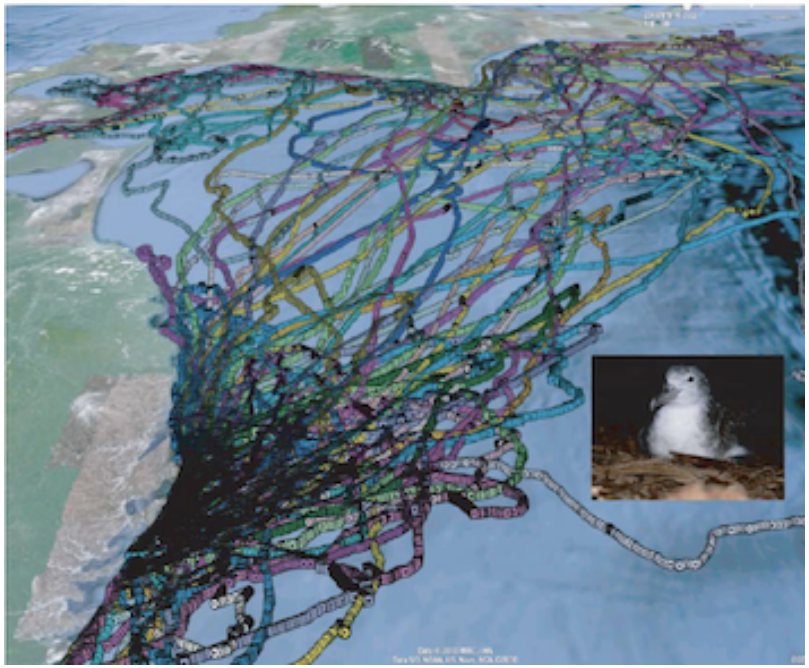
- Data that has a geographic or spatial component
- Spatial data is broader class of data whose geometry is relevant but it is not necessarily georeferenced



Dengue Outbreak in Singapore (Huang et al. 2022)

Spatial-temporal Data (Animal Trajectory)

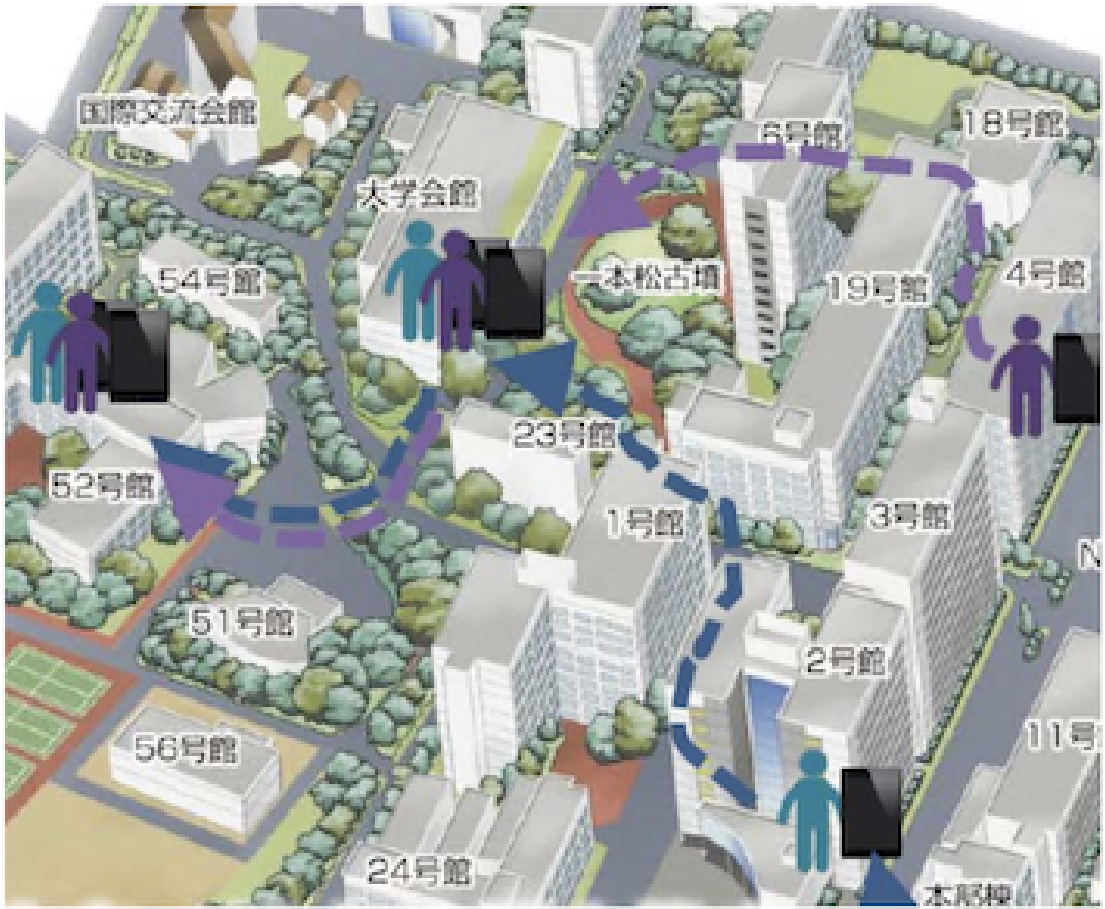
- Spatiotemporal data are data that relate to both space and time



by K. Yoda

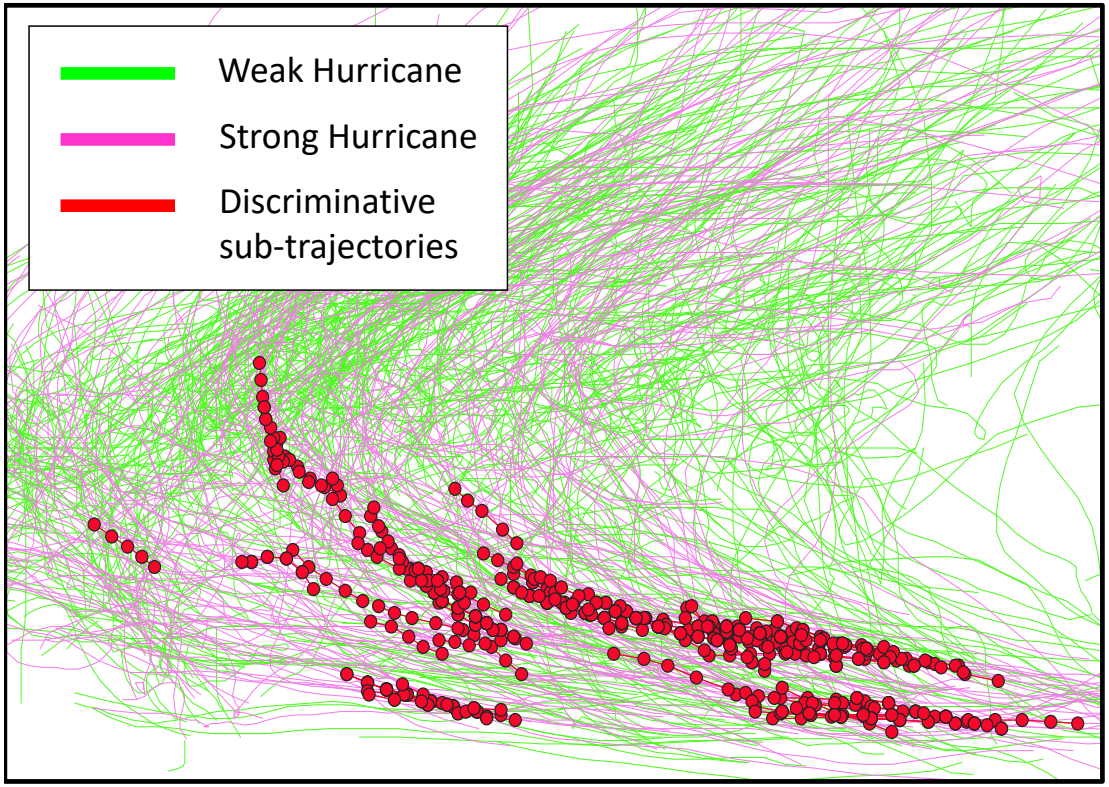
- Migration route identification
- Identification of resting areas
- Behavioral patterns:

Spatial-temporal Data (Human Trajectory)

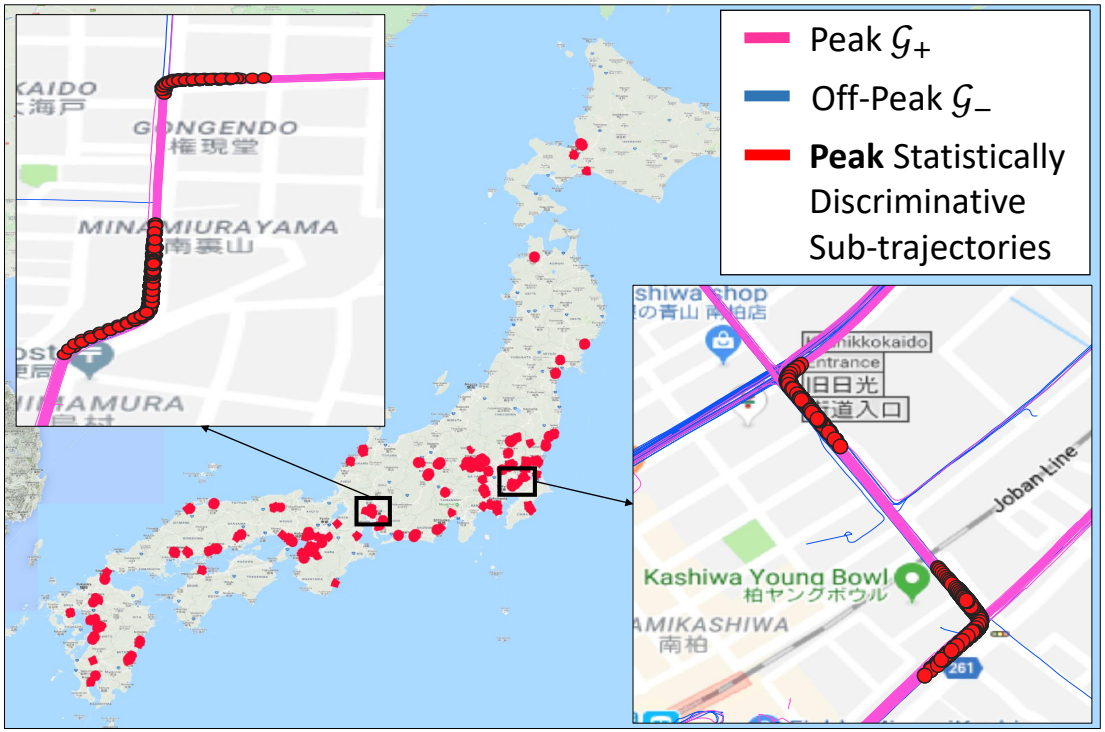


Takeuchi. (2020)

Spatial-temporal Data (Hurricane Trajectory)



Spatial-temporal Data (Japan Car Trajectory)

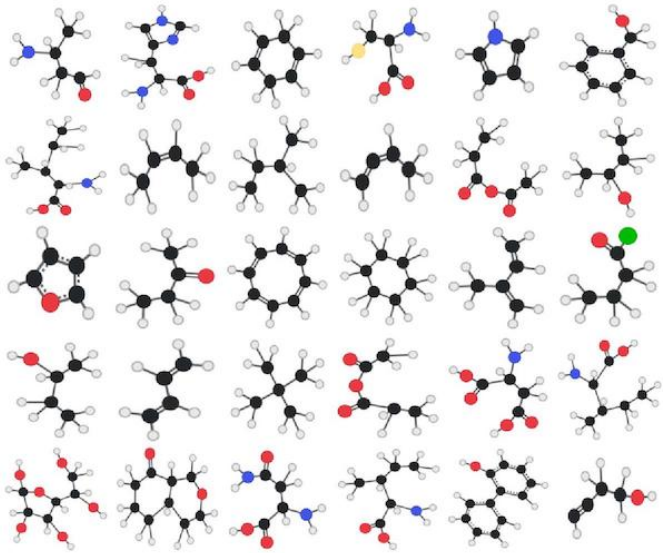


Spatial-temporal Data (Another example)

- Surgical instrument movements

Graph Data

- Graph data refers to data that is structured as a graph, which is a collection of nodes and edges.



https://github.com/chemplexity/molecules

- Biological and molecular analysis, e.g., analyzing the interactions between proteins to understand the mechanisms of cellular processes.
- Route optimization
- Social network analysis

Hypertext and Multi-media Data

- By mining text data on machine learning from the past ten years, we can identify the evolution of hot topics in the field
- Analyzing user navigation patterns to optimize website structure and layout

Data Mining Functionalities

- Characterization and Discrimination
- Mining Frequent Patterns and Associations
- Classification and Regression
- Cluster Analysis
- Outlier Analysis

Characterization and Discrimination

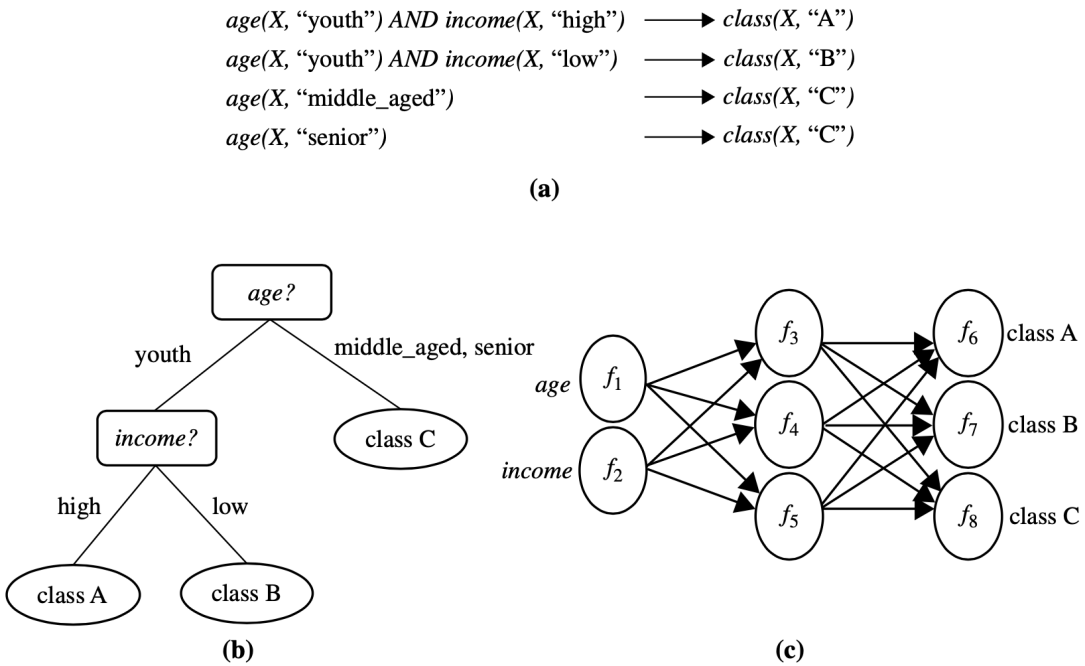
- Data characterization
  - Summarization of the general characteristics or features of a target class of data
  - The output of data characterization can be presented in various forms, e.g., pie charts, bar charts
  - Example 1: summarize the characteristics of customers who spend more than 10,000 USD a year on Shopee (40 to 50 years old, employed, and have excellent credit ratings)
  - Example 2: university entrance exam scores
- Data discrimination
  - Comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.
  - Example 1: compare two groups of customers on Shopee — regularly shop vs. rarely shop
  - Example 2: compare two groups of students
- Other examples?

Mining Frequent Patterns and Associations

- Frequent patterns: patterns that occur frequently in data
  - Frequent itemsets
  - Frequent sub-sequences
  - Frequent sub-structures
- Association analysis
  - $buys(\text{“computer”}) \Rightarrow buys(\text{“software”})$ $[support = 1\%, confidence = 50\%]$
  - $age(\text{“20..29”})$ and $income(\text{“40K..49K”}) \Rightarrow buys(\text{“laptop”})$ $[support = 2\%, confidence = 60\%]$
- Other examples?

Classification and Regression

- Classification: finding a model (or function) that describes and distinguishes data classes or concepts

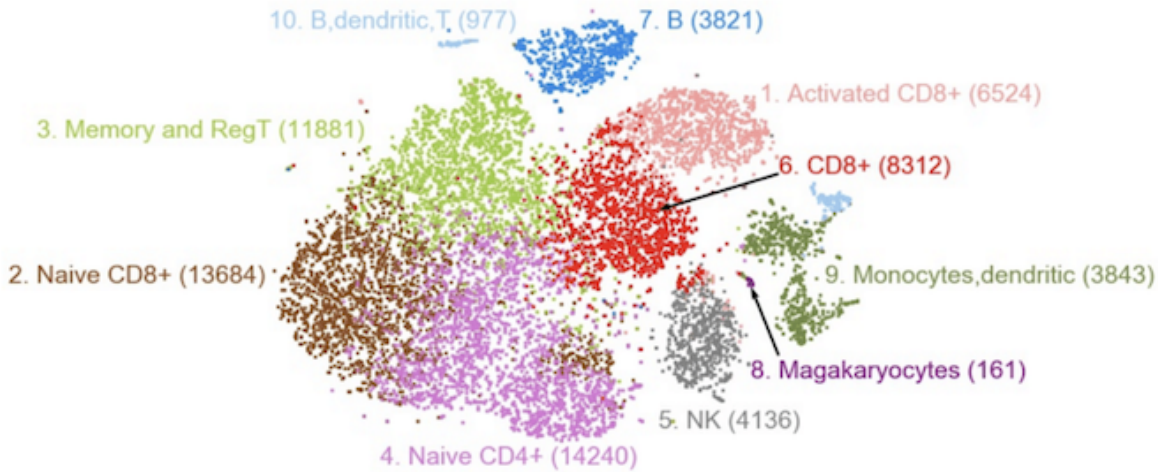


Classification model (Han et al. 2022)

- Regression: predicting missing or unavailable numerical data values
- Other examples?

Cluster Analysis

- Clustering or grouping objects based on the principle of *maximizing the intra-class similarity and minimizing the inter-class similarity*

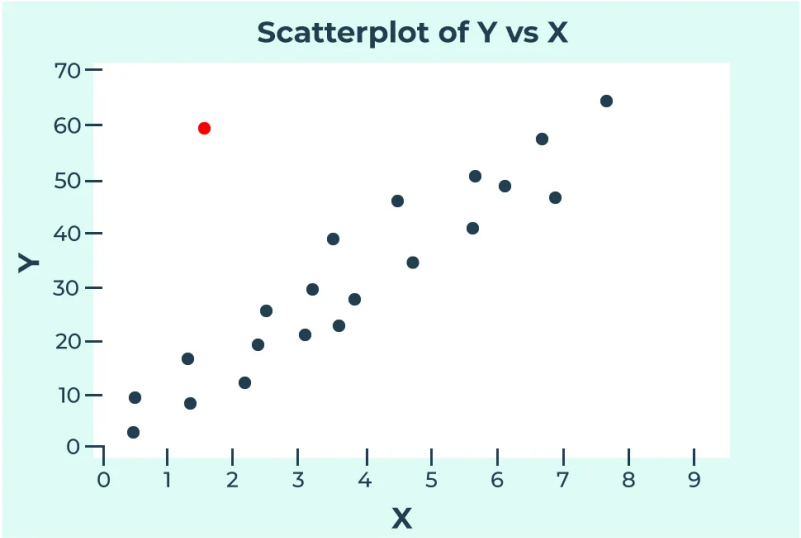


Inoue. 2019

- Other examples?

Outlier Analysis

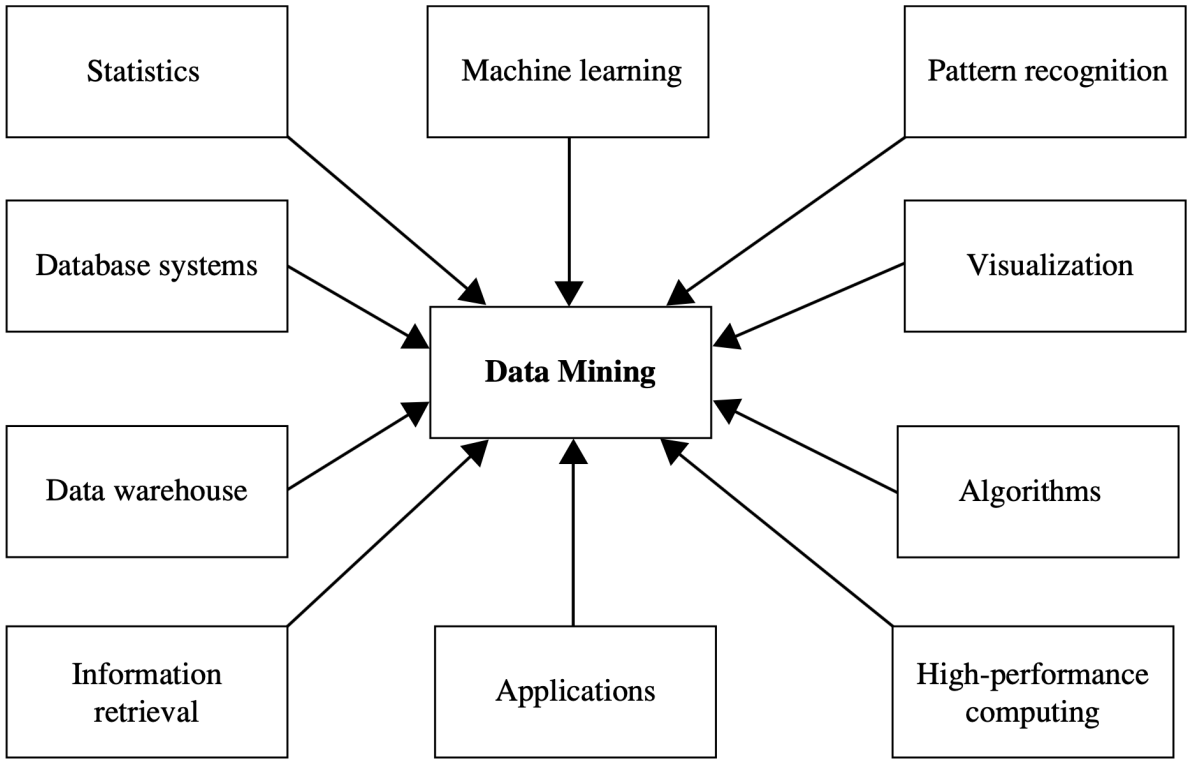
- Identifying rare and unusual observations that deviate significantly from the norm within a given dataset



https://godatadrive.com/blog/outliers-101

- Other examples?

Which Technologies Are Used?



Han et al. 2022

Summary

- Data mining is the process of discovering interesting patterns from massive amounts of data
- A pattern is *interesting* if it is valid on test data with some degree of certainty, novel, potentially useful, and easily understood by humans
- Data mining has many successful applications, such as business intelligence, bioinformatics, health informatics, and finance
- Data mining can be conducted on any kind of data as long as the data are meaningful for a target application
- Data mining functionalities include characterization and discrimination; the mining of frequent patterns, and associations; classification and regression; cluster analysis; and outlier detection
- Data mining has incorporated technologies from many other domains such as statistics, machine learning, pattern recognition.

Quiz 1

Identify the wrong step in the process of knowledge discovery

1. Data cleaning
2. Data mining
3. Data integration
4. Data analysis

Quiz 2

Which data mining technique is used to predict numeric values?

1. Prediction
2. Classification
3. Regression
4. Outlier

Quiz 3

Q & A

duyvnl@uit.edu.vn