

Introduction to Data Mining

Man-Kwan Shan

Dept. of Computer Science

National Chengchi Univ.

mkshan@nccu.edu.tw



國立政治大學

外系生修資科系資料採掘

政治大學 · 8月18日 09:31

如題

請問商院仔修資科所沈錕坤老師開的資料採掘

先備課程只有py和線代

去修這門課會被勸退嗎？



國立政治大學 資訊科學系



0



卡，想知道詳細的課程內容和甜涼度

B3 · 8月18日 回覆



國立政治大學 資訊科學研究所



4



老師教的東西滿多的課後要花一點時間複習（我都沒在上課也沒在複習）

期中期末考也不好寫（也可能是我太混）

不過最後還是過了感覺老師不會太刁難（我真的很混）

B4 · 8月18日 回覆

加簽

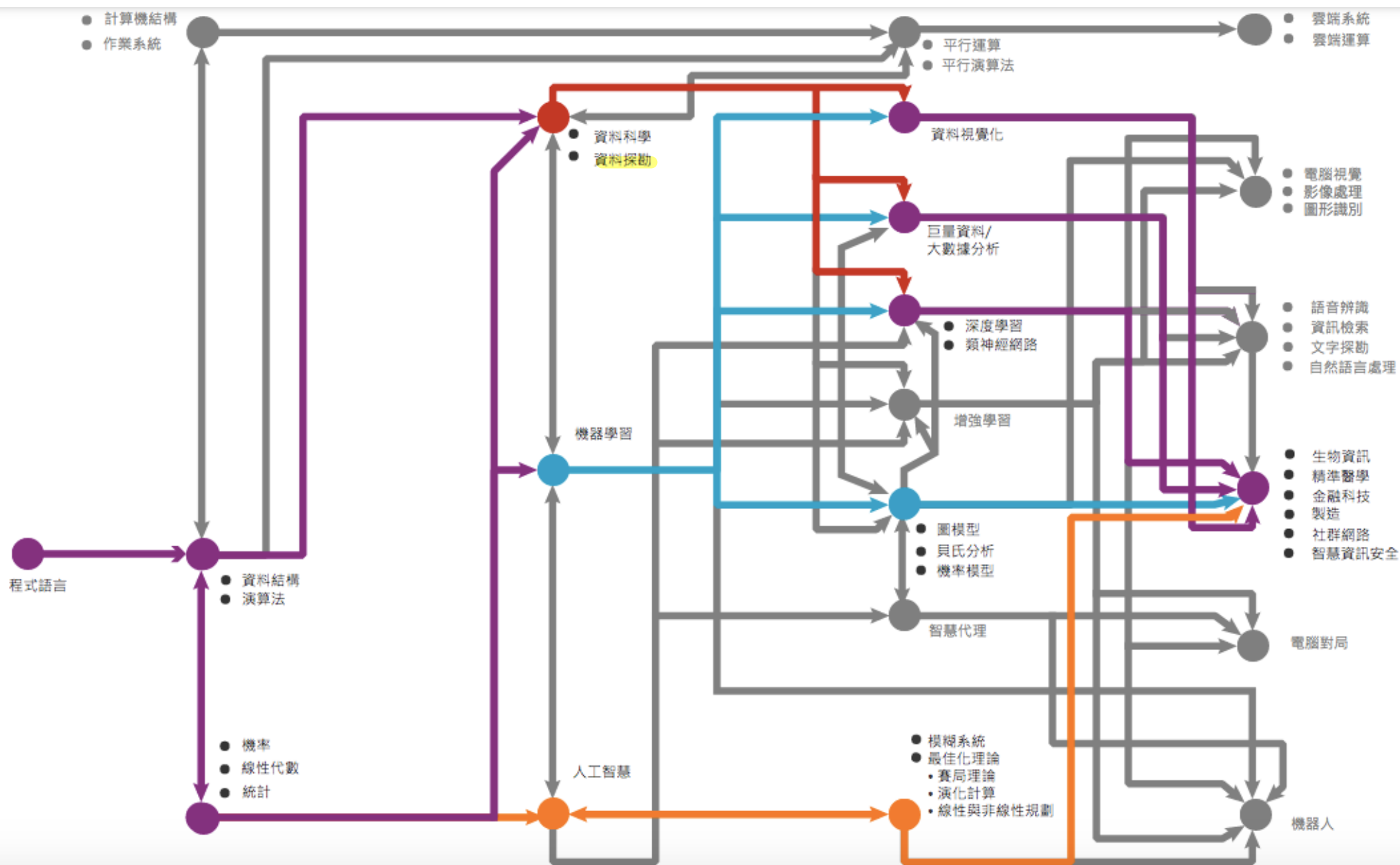
- 由遞補名單中，依照下列順序依序加簽
資科碩智慧計算組 > 資科碩一般組 > 資科博 > 資科碩跨領域組
> 資科大四 > 人智應學程 > 資訊學院其他系所 > 外院
- 修課同學必須具備資料結構、演算法、資料庫系統、
程式設計(python)的先備能力
- 9/13(五) 12:00 前，遞補名單中想加簽的同學，請附上姓名、學號、
系級、學校email、先備能力證明，email至
mkshan@nccu.edu.tw
- 9/15(日) 23:59 前，我將會email 告訴同學遞補結果。
- 可以加簽的同學，我會在9/19(四)前把加簽名單送交資科系辦公室
手動幫同學加簽(同學不需繳交加簽單)。

Agenda

- 課程相關規定
- Introduction to Data Mining
 - What
 - Why
 - How
 - Data
- Large Language Model vs. Data Mining

Course Information

- **Instructors:** 沈錕坤 Man-Kwan Shan
 - mkshan@nccu.edu.tw
 - data mining, database, multimedia, social network, deep learning
- **Office hours:** By appointment
- **Prerequisites**
 - Linear Algebra
 - Data Structures, Algorithms
 - Databases, SQL
 - Python



Goals

- 學習目標：培養同學資料探勘演算法的基礎能力
- 學習成效：
 - 資料探勘的專業知識
 - 資料探勘演算法的設計思維
 - 解決資料探勘的問題

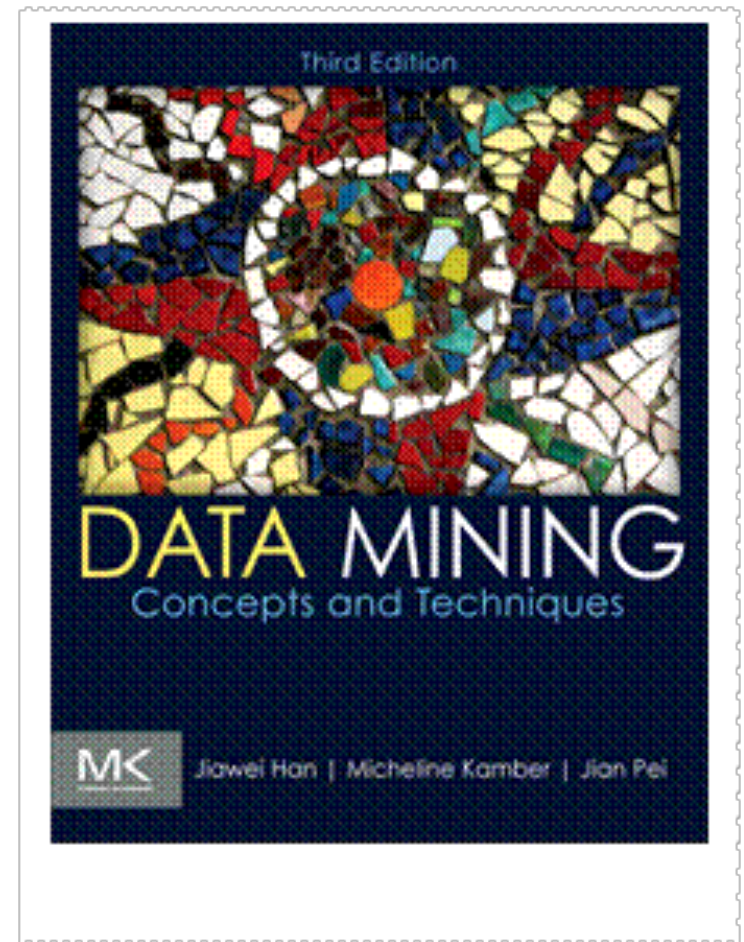
* 本課程設計主要針對沒有修讀過資科系的

- 資料科學導論
- 大數據分析
- 機器學習
- 網路搜索與探勘

的同學而設計。修讀過上述課程的同學，選修前請審慎考量。

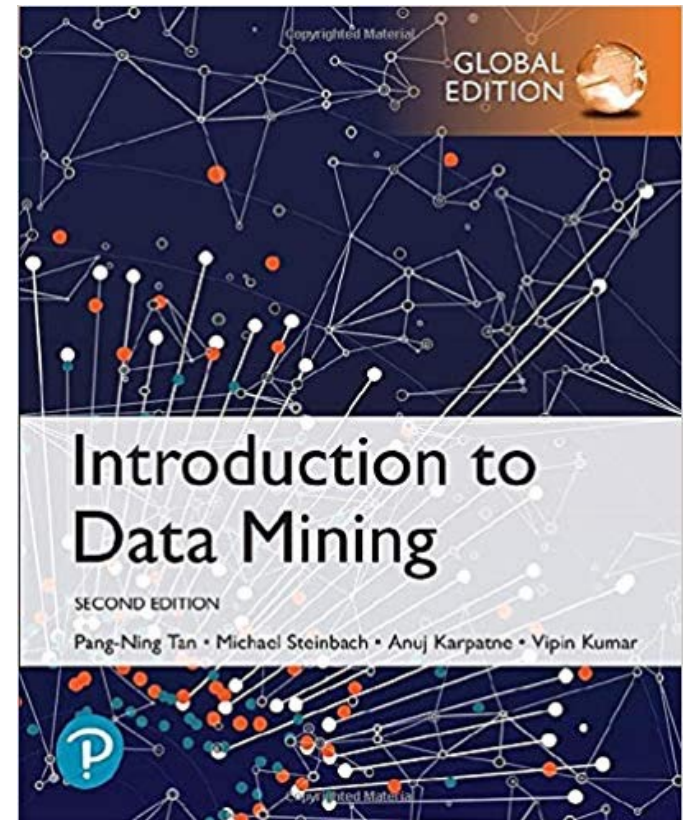
Text Book-1

- Data Mining: Concepts and Techniques, J. Han, M. Kamber, and J. Pei, 3rd. Ed., Morgan Kaufmann , 2011.



Text Book-2

- Introduction to Data Mining,
Pang-Ning Tan, Michael
Steinbach, and Vipin Kumar,
Addison Wesley, 2020.



Topics

1. Introduction to Data Mining
2. Data Preprocessing
3. Frequent Pattern Mining
4. Clustering
5. Classification
6. Dimensionality Reduction
7. Recommendation
8. Data Warehouse
9. Social Network Mining (optional)
10. System Design (optional)
11. Deep Learning (optional)
12. Large Language Model (optional)

Grading (Tentative)

- 期中考、期末考(optional) 25~30 %
- 五次作業 15~25% (python scikit-learn)
 - Data Preprocessing
 - Frequent itemset mining
 - Clustering
 - Classification
 - Recommendation
- 學期專題 25~30%

Data Mining Conferences

- ACM KDD (Knowledge Discovery from Data)
- IEEE ICDM (Data Mining)
- SDM (SIAM Data Mining)
- ACM CIKM (Conference on Information & Knowledge Management)
- ACM WSDM (Web Search and Data Mining)
- PAKDD (Pacific–Asia)
- PKDD (Principle...)
- ACM SIGMOD (Management of Data)
- IEEE ICDE (Data Engineering)
- VLDB (Very Large Database)
- WWW Conference (World Wide Web Conference)
- ACM SIGIR (Information Retrieval)

Data Mining Journals

- Data Mining & Knowledge Discovery (DMKD)
- IEEE Transactions on Knowledge & Data Engineering (IEEE TKDE)
- ACM Transactions on Knowledge Discovery from Data (TKDD)
- ACM Transactions on Database Systems (TODS)
- ACM Transactions on Information Systems (TOIS)
- ACM Transactions on The Web (TWEB)
- ACM Transactions on Internet Technology (TOIT)
- ACM Transactions on Intelligent Systems and Technique (TIST)
- IEEE Intelligent Systems
- AI Magazine
- Data and Knowledge Engineering
- Expert Systems with Applications
- Knowledge Information systems
- Information Science
- ...

Introduction to Data Mining

為什麼會有 Data Mining ?



Motivation

What motivated Data Mining ?

We are data rich but information poor



Technologies to Support Data Mining

- Technology to **collect** data
 - barcode reader for basket data
 - sensors: location aware sensors for air quality PM 2.5
 - video surveillance
 - search engines for search behaviors
 - social media...
- Technology to **store** data
 - e.g.: RAID, Data Center, Cloud
- Technology to **manage** data
 - e.g.: Database Management Systems

什麼是 Data Mining ?



What is Data Mining ?

- Nontrivial process of extraction of
 - valid (with some degree of certainty)
 - novel (surprising, previously unknown)
 - potential useful
 - understandable

patterns from large collection of data

Patterns

- 30% of customers
who purchased breads
also purchased milks
- 30% of female customers
who purchased skin care products,
also purchased men's shirts.

Related Terms

- Data Mining, Knowledge Discovery in Databases (KDD)
- Intelligent Data Analysis
- Statistics 統計
- Data Science 資料科學
- Big Data 大數據
- Machine Learning 機器學習
- Deep Learning 深度學習

Related Terms

- Data Mining, Knowledge Discovery in Databases (KDD)
 - Pattern mining
- Statistics 統計
 - Sampling
- Data Science 資料科學
 - Data Mining + Visualization + Interpretation for Decision Support
- Big Data 大數據
 - Data Mining + Cloud Computing
- Machine Learning, Deep Learning 機器學習, 深度學習
 - For Prediction, Generation

4 Types of Data Analytics

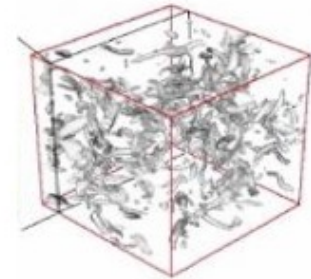
- **Descriptive Analytics:** tells you what happened in the past.
- **Diagnostic Analytics:** helps you understand why something happened in the past.
- **Predictive Analytics:** predicts what is most likely to happen in the future.
- **Prescriptive Analytics:** recommends actions you can take to affect those outcomes.

Evolving Science

- Thousand years ago:
science was empirical
describing natural phenomena
- Last few hundred years:
theoretical branch
using models, generalizations
- Last few decades:
a computational branch
simulating complex phenomena
- Today:
data exploration (eScience)
*synthesizing theory, experiment and
computation with advanced data
management and statistics
→ new algorithms!*

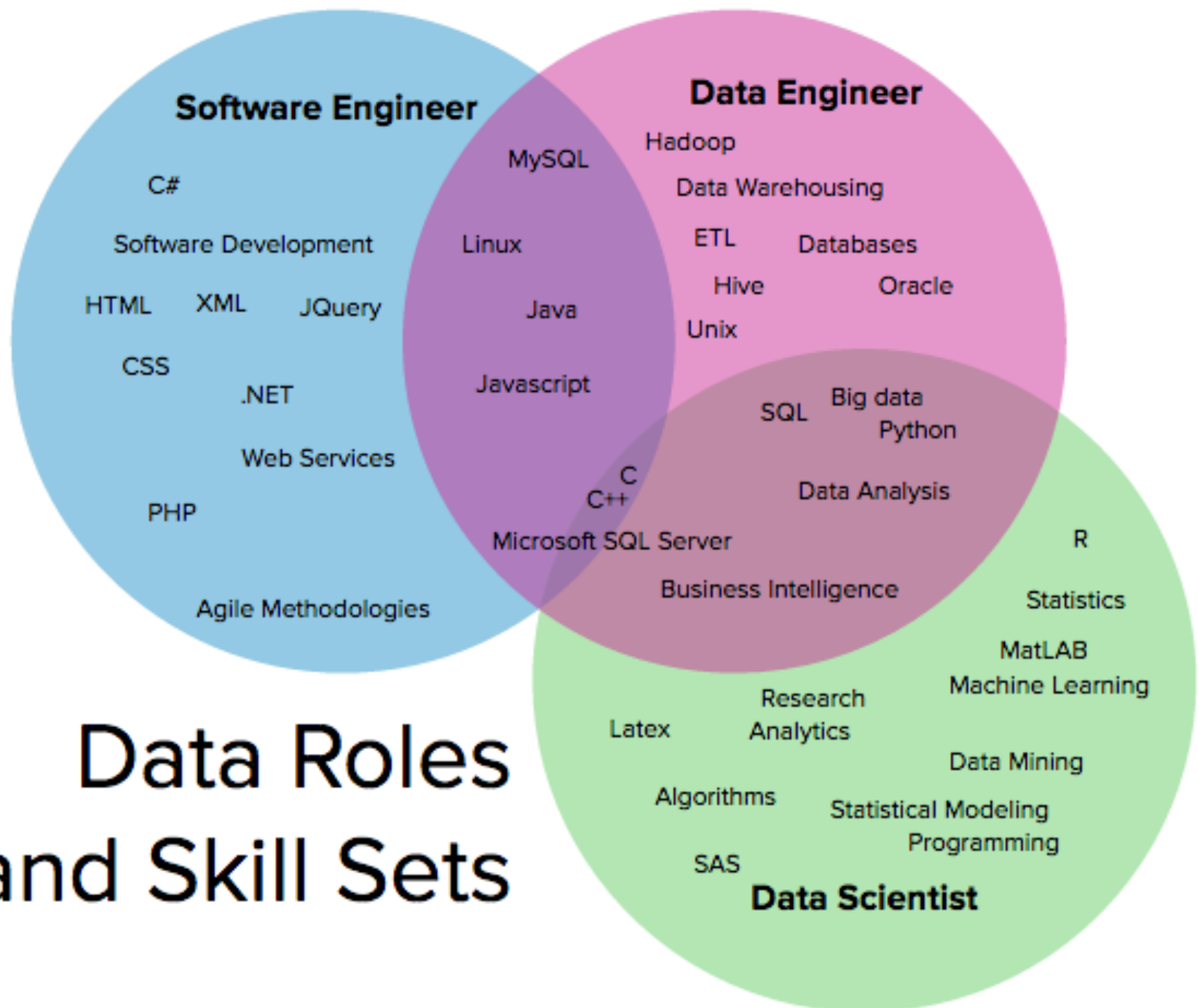


$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K \frac{c^2}{a^2}$$



* Jim Gray: The Fourth Paradigm: Data-Intensive Scientific Discovery

Data Roles and Skill Sets



Data Mining的步驟為何？

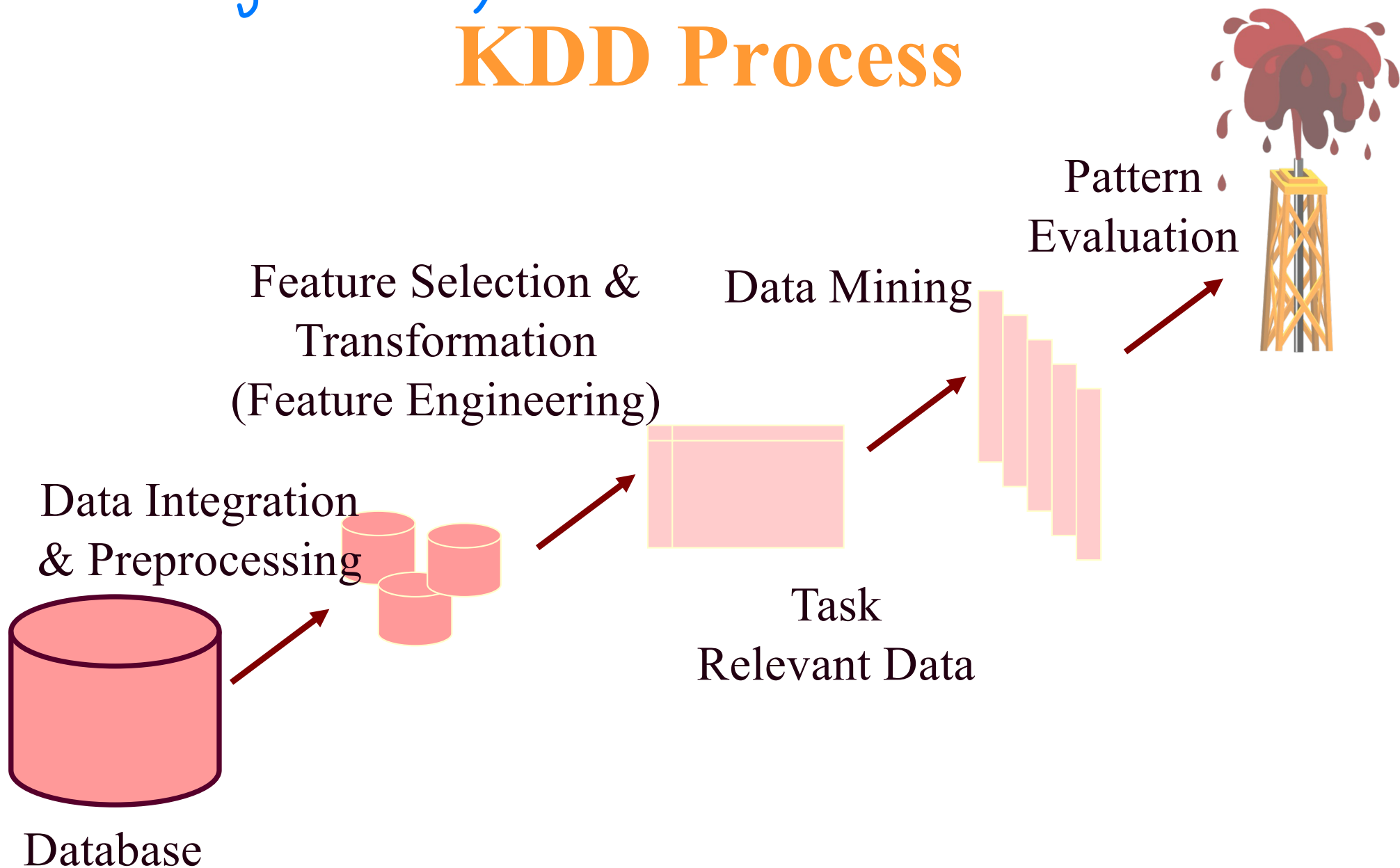


Specification of Data Mining Tasks

1. Data sets
2. Mining task specification: kinds of knowledge to be mined
3. Background knowledge
4. Interestingness measurement: confidence, significance,, thresholds, concept levels

Knowledge Discovery in Databases

KDD Process

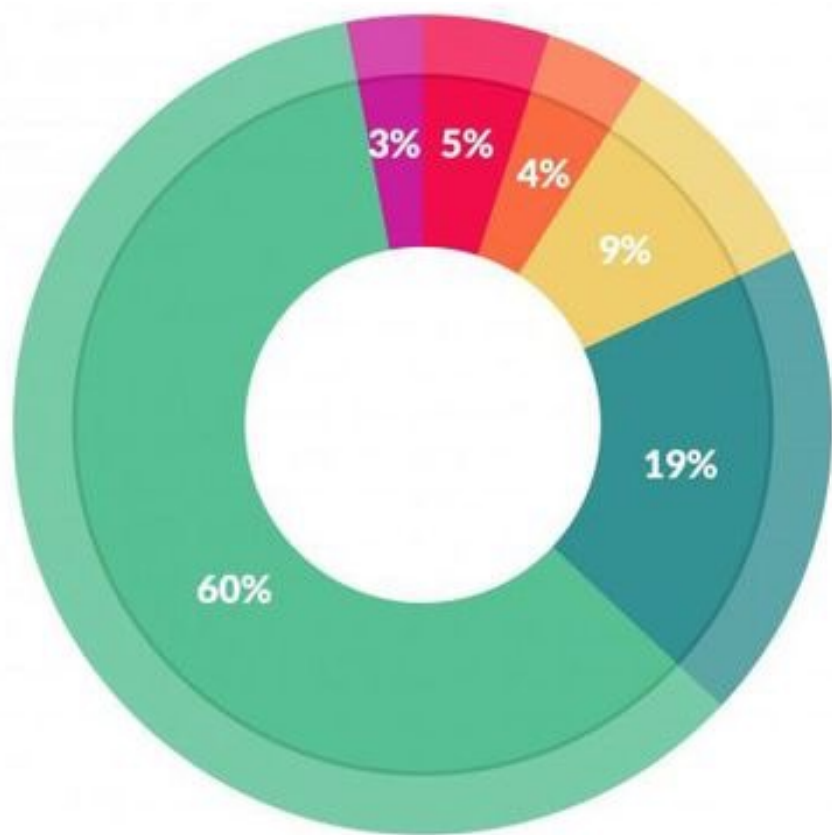


Steps of KDD Process

- Learning the application domain (prior knowledge & goals of applications)
- Gathering & integrating data
- Cleaning & preprocessing data
- Reducing & projecting data (feature selection, dimensionality reduction)
- Choosing functions of data mining
- Choosing the mining algorithm
- Data mining
- Evaluation results
- Interpretation (visualization, alteration, removing redundant patterns)
- Use of discovered knowledge

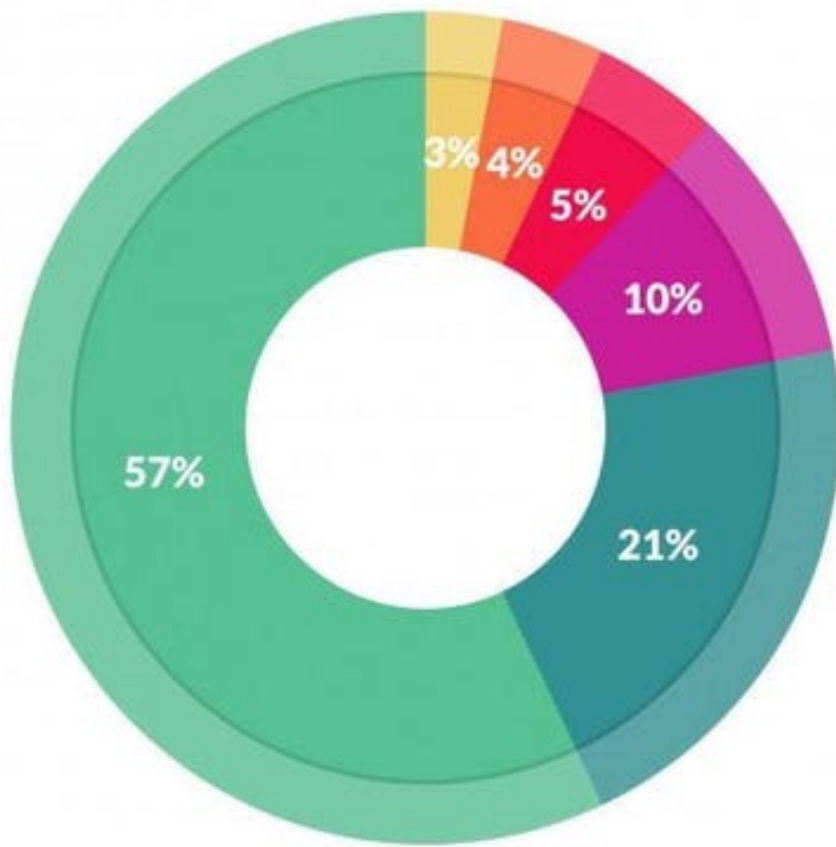
Data Mining這些步驟中，
哪個步驟最花時間？最無趣？





What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

Data Mining技術有哪幾種？



Classification of Data Mining Techniques

- What kinds of **databases** to work on (input)
- What kinds of **knowledge** to be mined (output)
- What kinds of **techniques** to be utilized (processing)

What Kinds of Databases to
Work on ?

Databases to Work on

- Relational data
- Transactional data
- Spatial data
- Temporal data
- Mobility data
- Textual data
- Multimedia (Text, Image, Video, Audio, Animation)
- Graph (Social network, Chemical compound)
- ...

Relational Data

- Relational data
 - A set of records
 - Each record consists of a fixed number of attribute values (structured data)

People			
Record ID	Age	Married	NumCars
100	23	No	1
200	25	Yes	1
300	29	No	0
400	34	Yes	2
500	38	yes	2

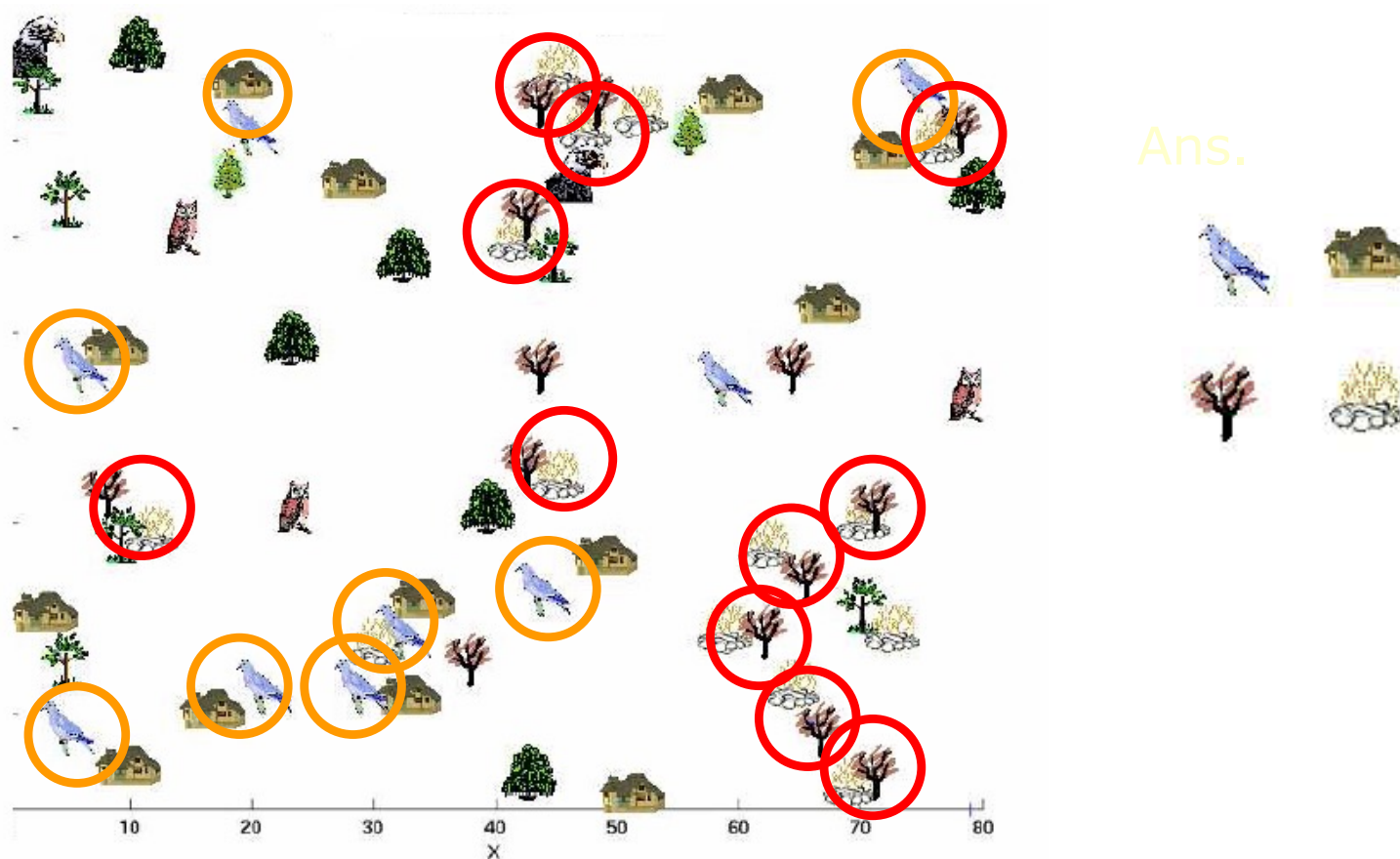
Transactional Data

- Transactional data
 - A set of transactions
 - Each transaction is an itemset

TID	Items
100	A, C, D
200	B, C, E
300	A, B, C, E
400	B, E

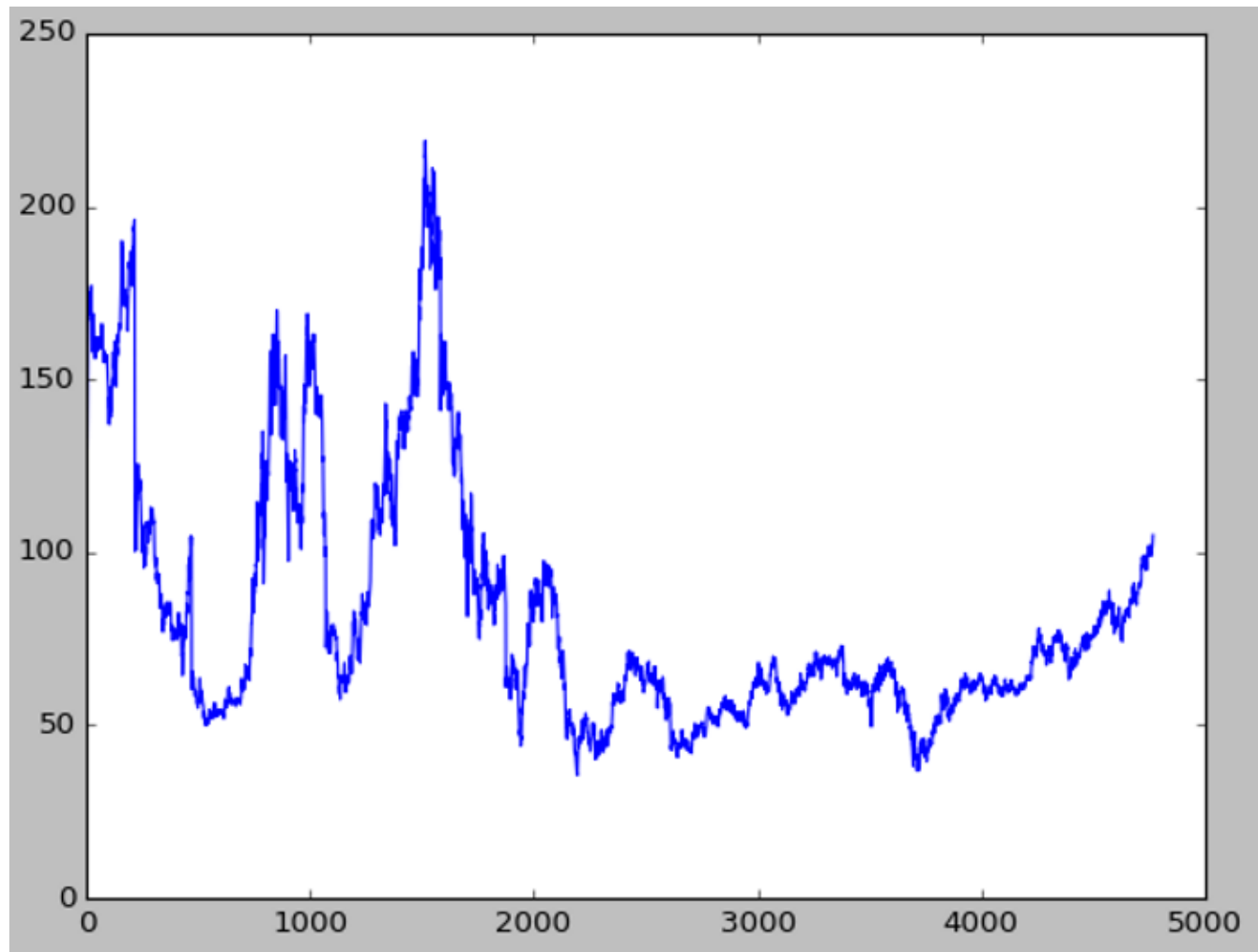
Spatial Data

- Spatial data: location (e.g. geographic) information
- Spatial patterns: e.g. co-location patterns

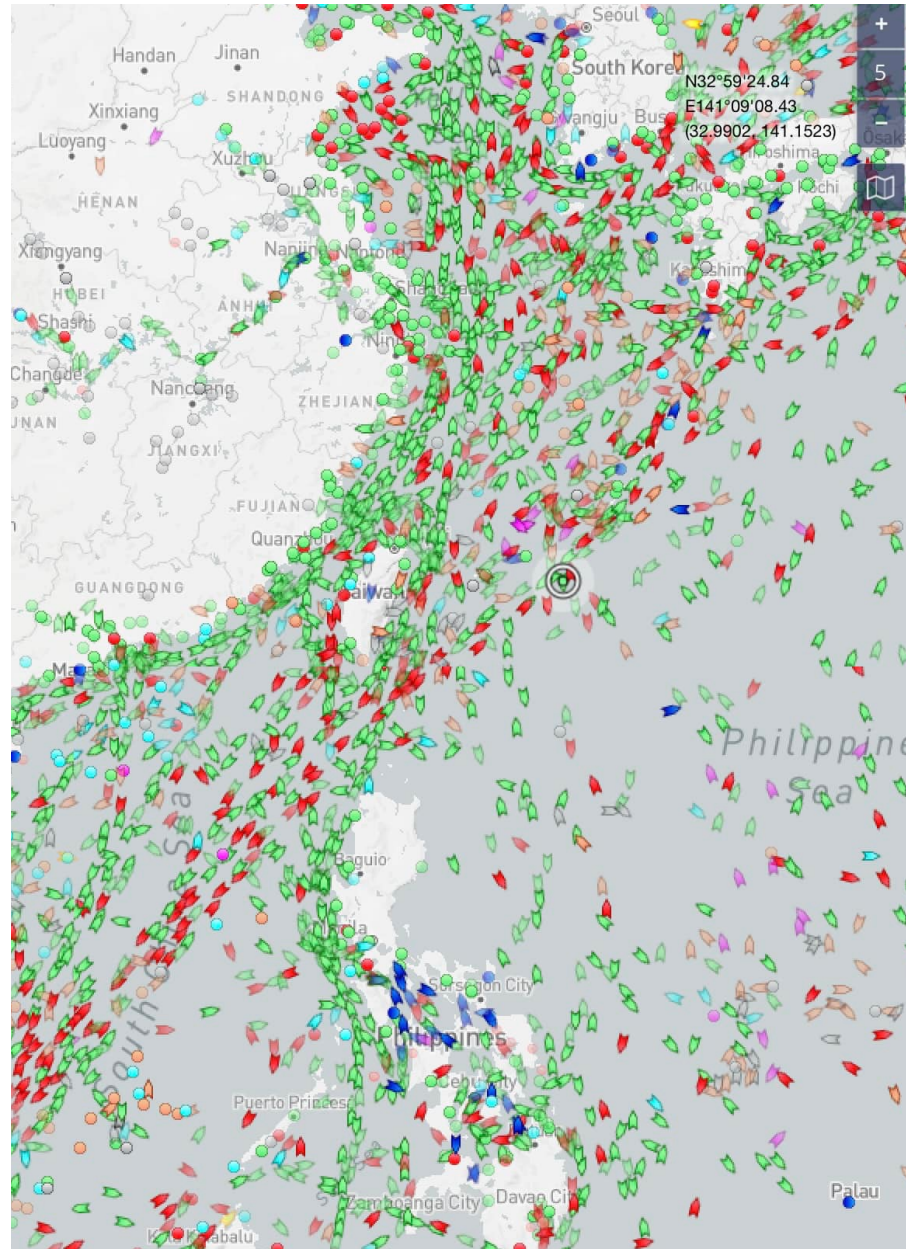


Temporal Data

- Time series data



Trajectory Data



Biological Sequence Data

- The patterns with 2 point mutations:

cctgatagacgctatctggctatccaggtacttaggtcctctgtgcgaatctatgcgtttccaacc
at

agtactggtgtacatttgatccatacgtacaccggcaacctgaaacaaacgctcagaaccagaagt
gc

aaacgttagtgcaccctctttcttcgtggctctggccaacgagggctgatgtataagacgaaaatt
tt

agcctccgatgtaagtcatactgtaactattacctgccaccctattacatcttacgtccatata
ca

ctgttatacaacgcgctcatggcgggtaagcgtcttggtcgtcgtacgctcgatcgttaccgtacg
gc

acgtacgt

Text

Id: Abc123 on 5-1-2008 “/
bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, ...

...

Text

- Text (document)
 - Unstructured data, not structured data
 - Text retrieval (Information Retrieval)
- Text Mining
 - Opinion mining, sentiment mining
 - Document classification
 - ...

Music

- Motif discovery: variation of repeating pattern



Exact repeat



Interval repeat



Sequence



Contrary motion

1 1 1 -2 1, -2 -1 -1 -1 2 -1 2

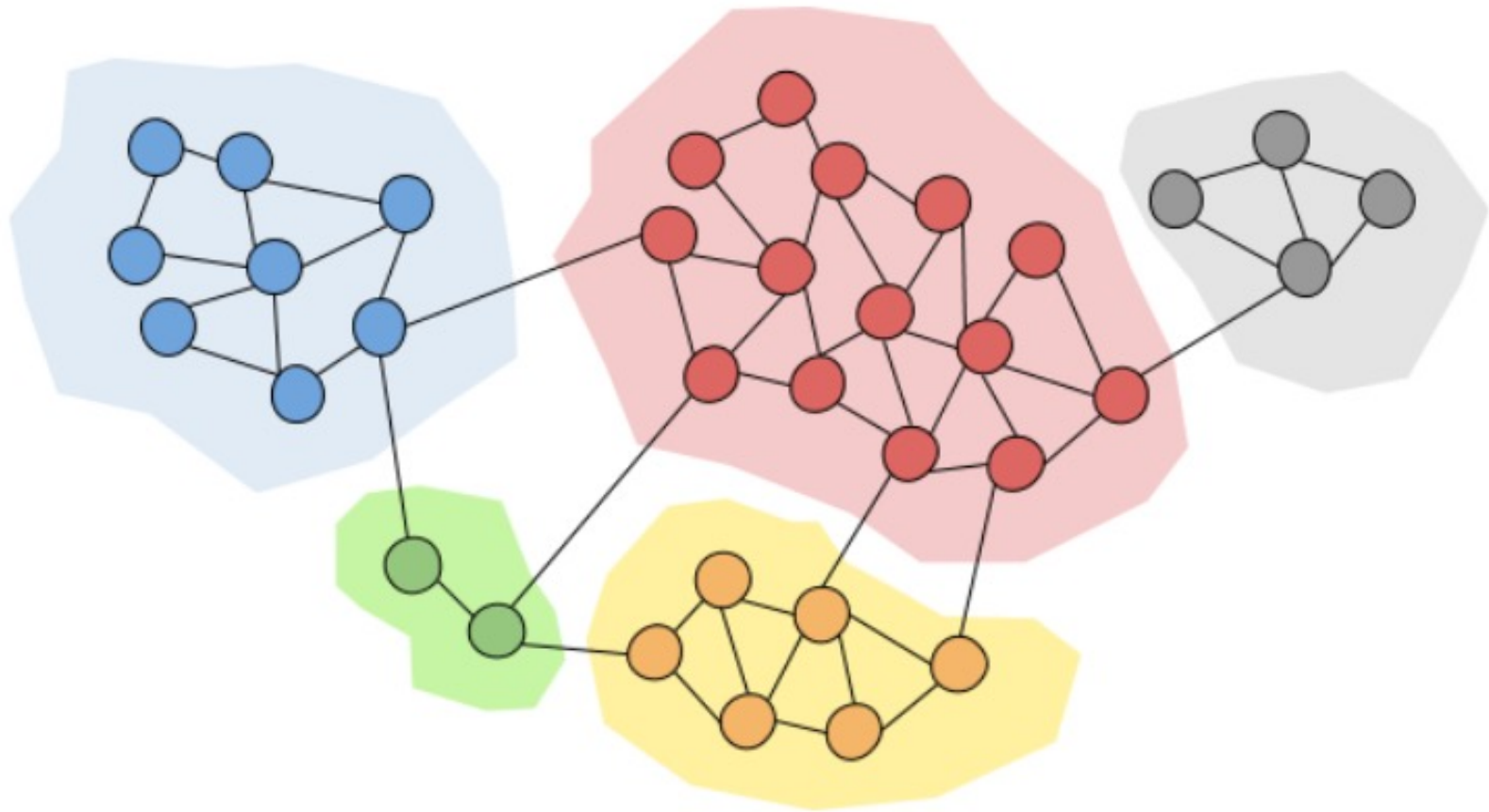


Retrograde



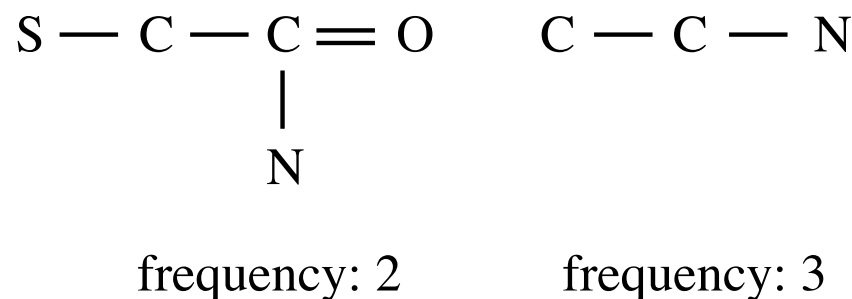
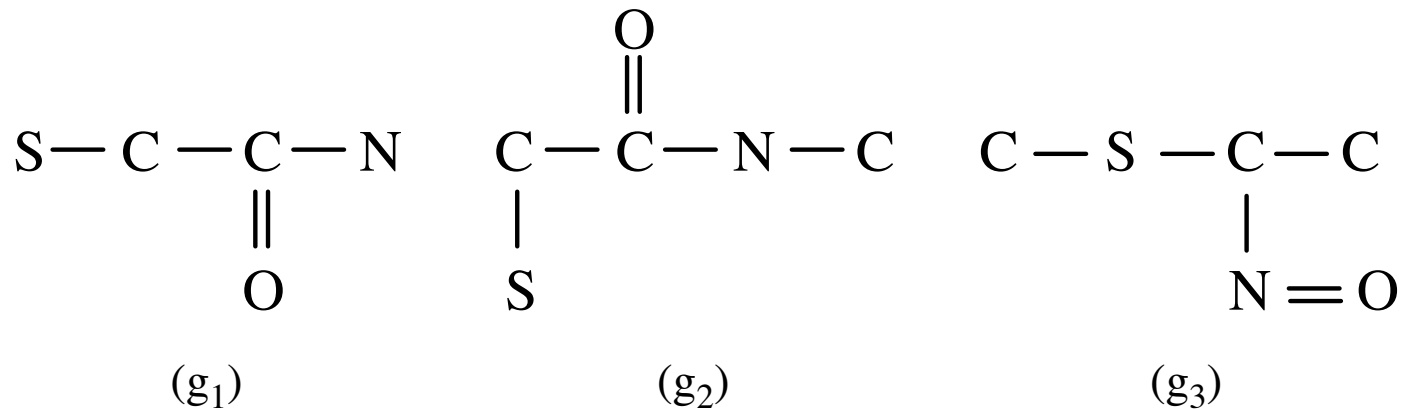
Augmentation or Diminution

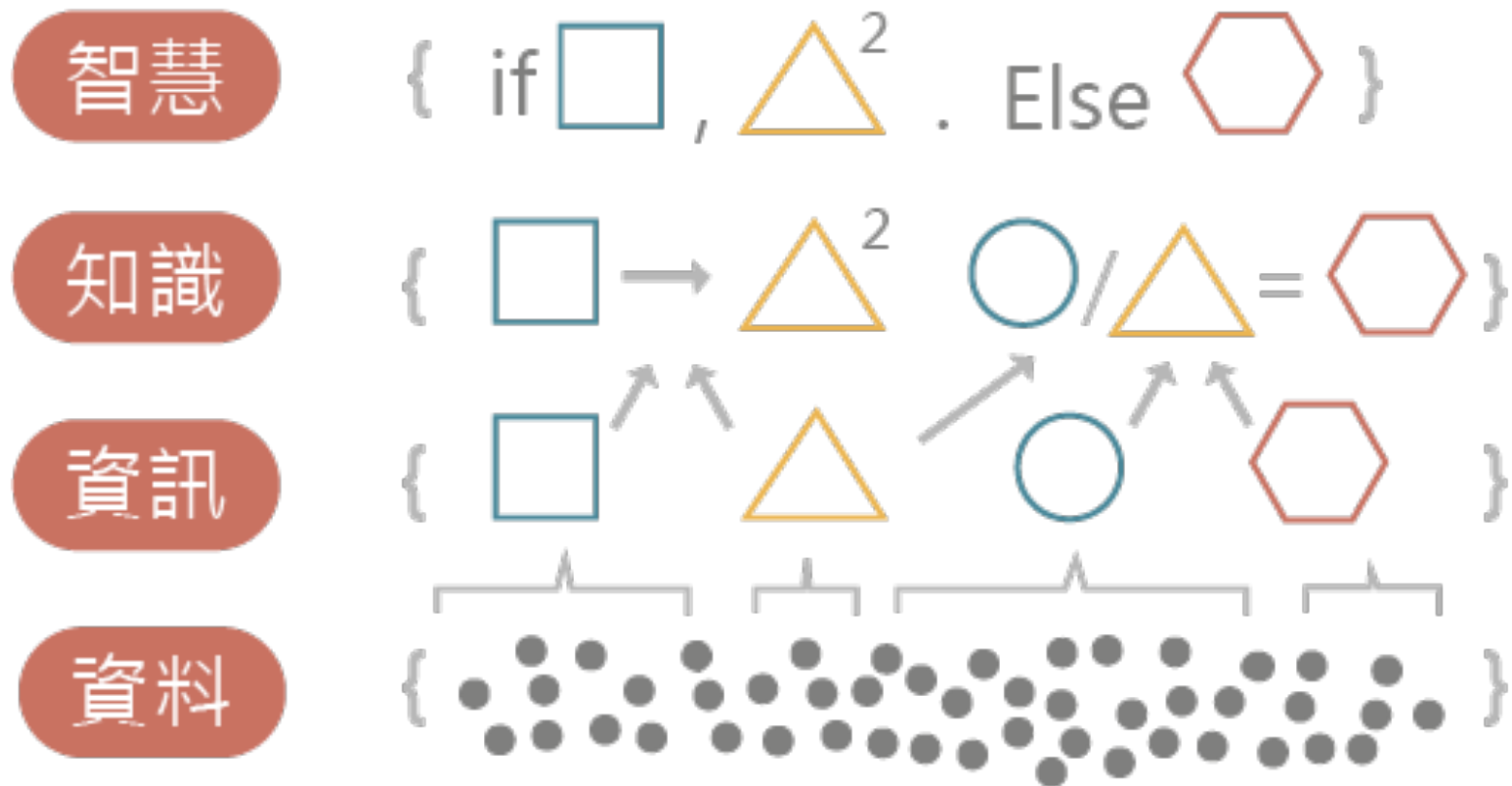
Social Network



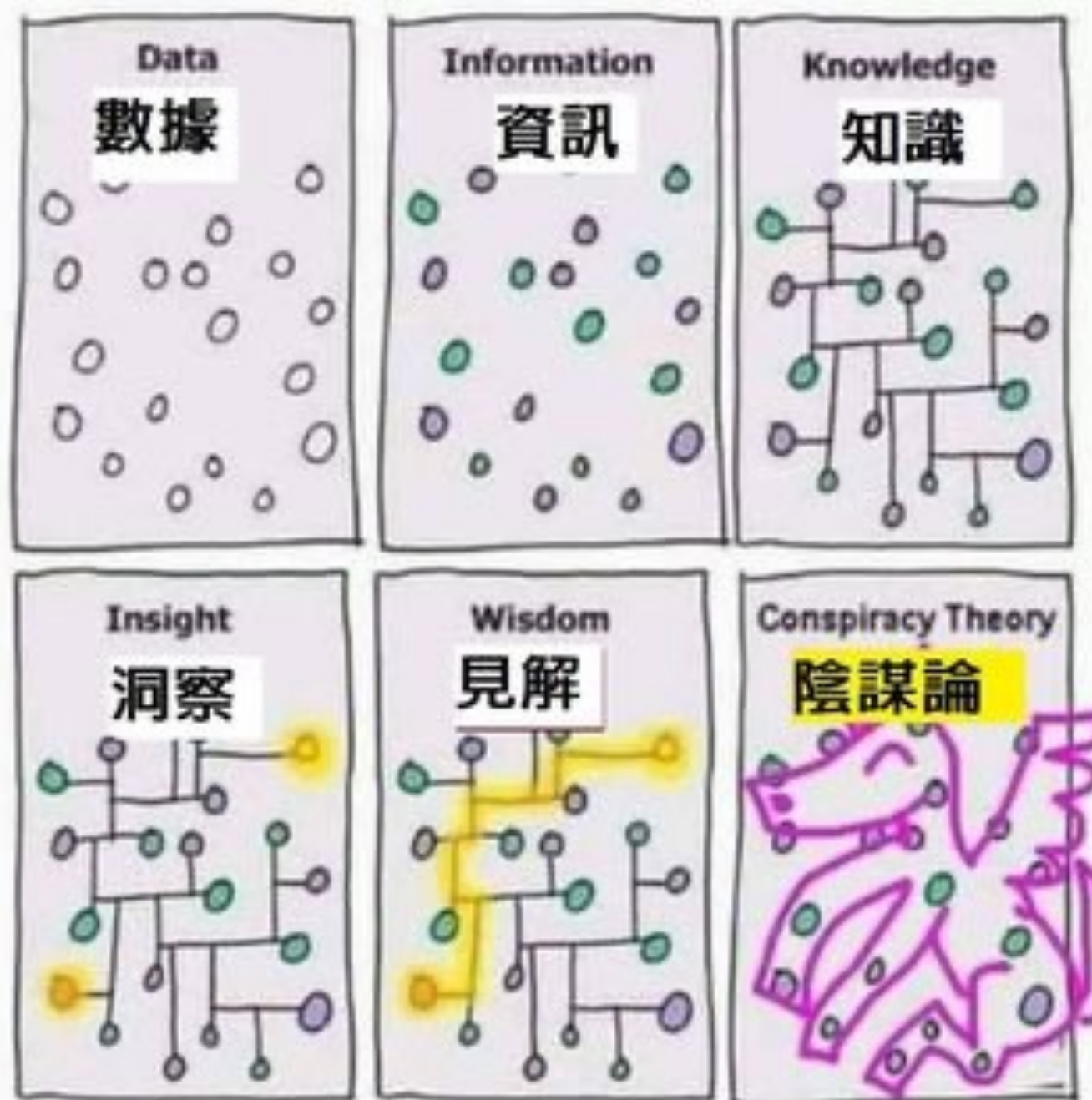
Frequent Graph Mining

- Frequent graph
 - sub-graph that appears frequently over a set of graphs





* Scott E. Page, The Model Thinker



Large Language Model

- Is LLM skilled at data mining ? (Claude)
 - LLMs (Large Language Models) like myself are not typically used directly for data mining in the traditional sense.
 - Data mining usually involves analyzing structured datasets to discover patterns and extract useful information.
 - LLMs, on the other hand, are primarily designed to process and generate human-like text.

Large Language Model (cont.)

- However, LLMs can be valuable tools in certain aspects related to data analysis and information extraction:
- 1. Natural Language Processing (NLP): LLMs excel at understanding and generating human language, which can be useful for analyzing **unstructured text data**.
- 2. Information Extraction: LLMs can extract relevant information from large volumes of text, which could be considered a form of data mining from **unstructured data**.
- 3. Question Answering: LLMs can quickly provide answers to questions based on their training data, which can be useful for querying large amounts of information.

Large Language Model (cont.)

- However, LLMs can be valuable tools in certain aspects related to data analysis and information extraction:
4. Text Summarization: LLMs can summarize large documents, potentially helping to identify key points in text data.
 5. Pattern Recognition in Text: While not as precise as traditional data mining algorithms, LLMs can identify trends or patterns in textual data.
 6. Assisting in Data Analysis: LLMs can help interpret results, generate hypotheses, or suggest analytical approaches, though they shouldn't replace rigorous statistical analysis.

Large Language Model (cont.)

- It's important to note that for structured data analysis & traditional data mining tasks, specialized tools and techniques (like statistical analysis software, data mining algorithms, or machine learning models specifically designed for structured data) are generally more appropriate and efficient.

Summary

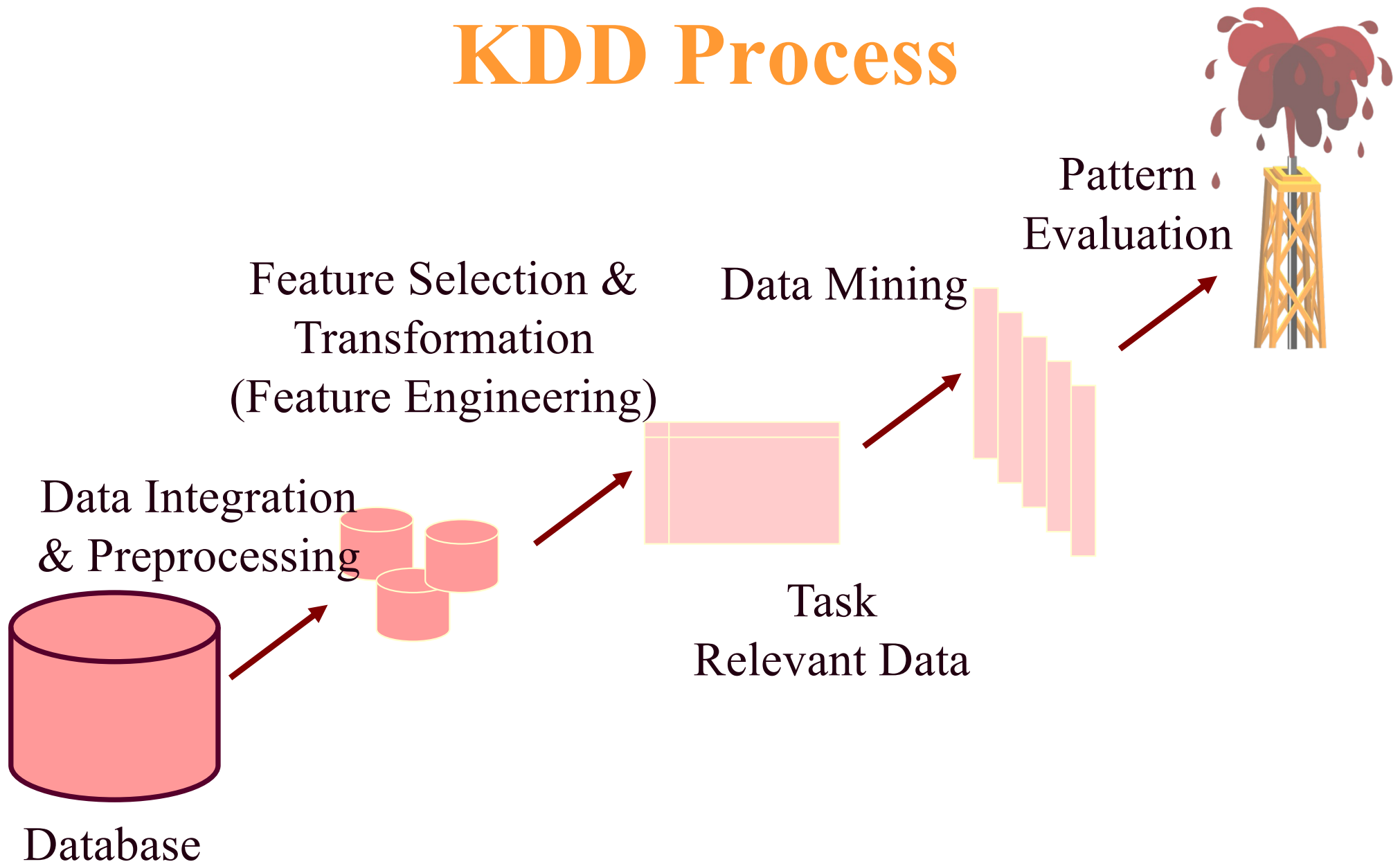
Knowledge Discovery from Databases

- Nontrivial process of extraction of
 - valid (with some degree of certainty)
 - novel (surprising, previously unknown)
 - potential useful
 - understandablepatterns from large collection of data

4 Types of Data Analytics

- **Descriptive Analytics:** tells you what happened in the past.
- **Diagnostic Analytics:** helps you understand why something happened in the past.
- **Predictive Analytics:** predicts what is most likely to happen in the future.
- **Prescriptive Analytics:** recommends actions you can take to affect those outcomes.

KDD Process



Databases to Work on

- Relational data
- Transactional data
- Spatial data
- Temporal data
- Mobility data
- Textual data
- Multimedia (Text, Image, Video, Audio, Animation)
- Graph (Social network, Chemical compound)
- ...