COMP5121
Data Mining and Data Warehousing Applications

Week 1: Introduction to Data Mining and Data Warehousing

Dr. Fengmei Jin

- Email: fengmei.jin@polyu.edu.hk
- Office: PQ747 (+852 3400 3327)
- Consultation Hours: 10am-12pm every Thursday

Outline

- Why Data Mining?
- What Is Data Mining?
- How to Mine Data? – *a multidimensional view of data mining*
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - Which Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- Summary

The Evolution of Information Technology

Data collection and database creation
(1960s and earlier)

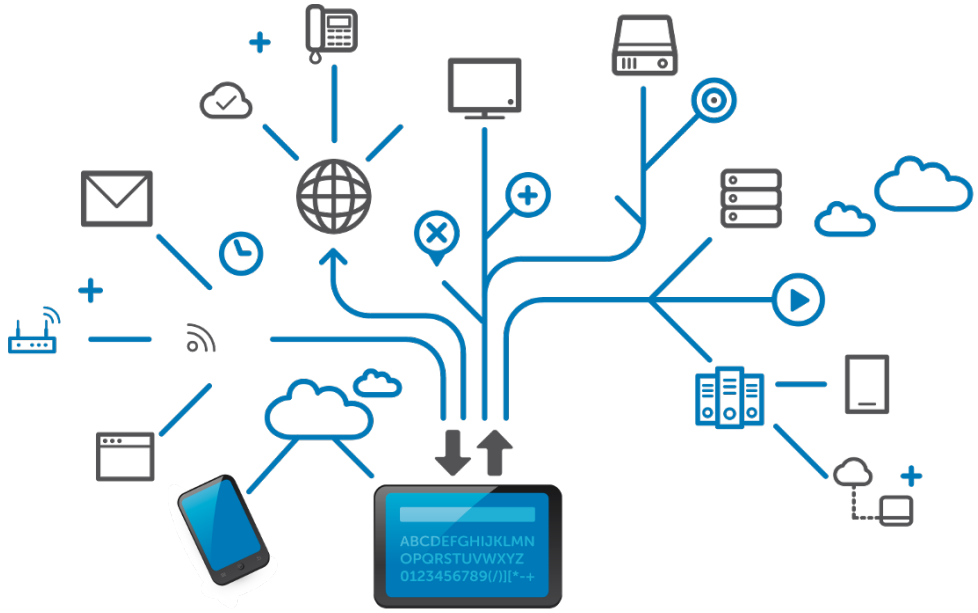
Database management systems
(DBMS, 1970s – early 1980s)
- Relational DBMS
- Indexing and accessing methods
- Query language, processing and optimization …

Advanced database systems
(mid-1980s to present)
- Advanced data models
- Managing complex data
- Cloud computing and parallel data preprocessing …

Advanced data analysis
(late-1980s to present)
- Data warehouse and OLAP
- Data mining and knowledge discovery
- Mining complex types of data
- …

Why Data Mining?

- We are living in the **data** age.
  - The explosive growth of data:
    - from *terabytes* (1 TB ≈ 1,000 GB) to *petabytes* (1 PB ≈ 1,000 TB)
  - The extensive data collection, availability, and storage:
    - *Business*: e-commerce, stocks, …
    - *Telecommunication networks*: call detail records, web search, …
    - *Social media*: news, blogs, photos, …

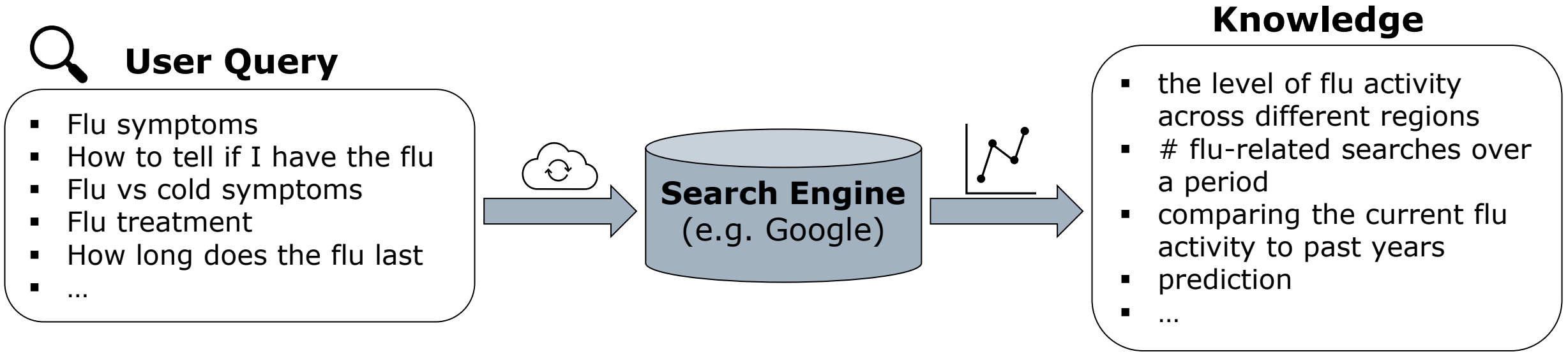


However, a data rich but information poor situation!

Why Data Mining?

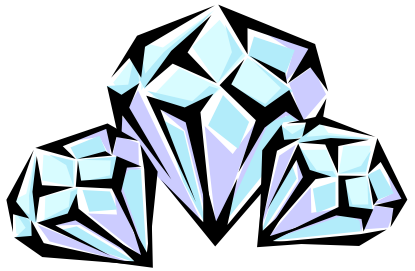
"Necessity is the Mother of Invention"

- We are drowning in data, but starving for knowledge!
- **Solution**: data warehousing and data mining
  - To turn a large collection of data into knowledge and move from the data age toward the information age



What Is Data Mining?

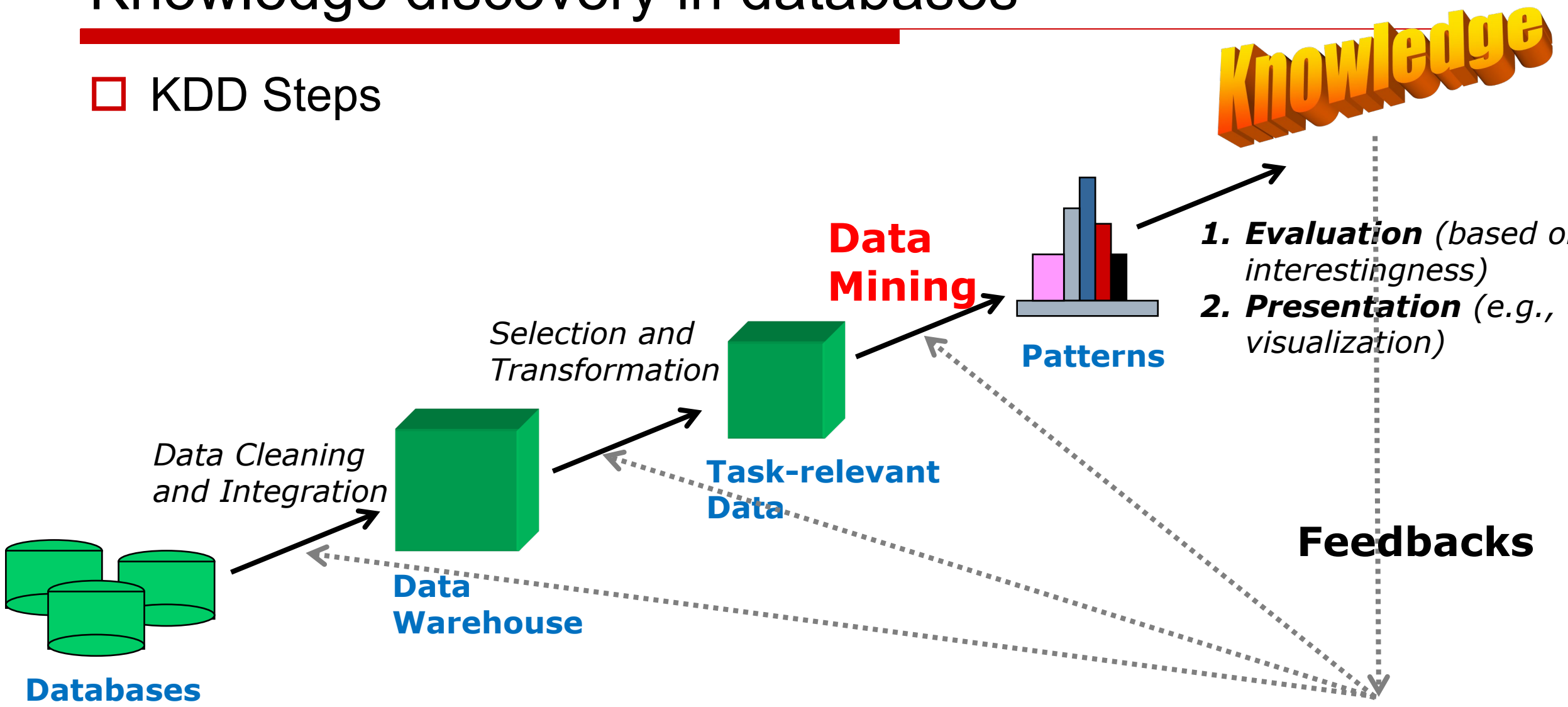
- Alternative names
  - Knowledge mining → *not emphasis on mining from data*
  - Knowledge mining from data → *too long*
  - Knowledge extraction
  - Data/pattern analysis
  - …



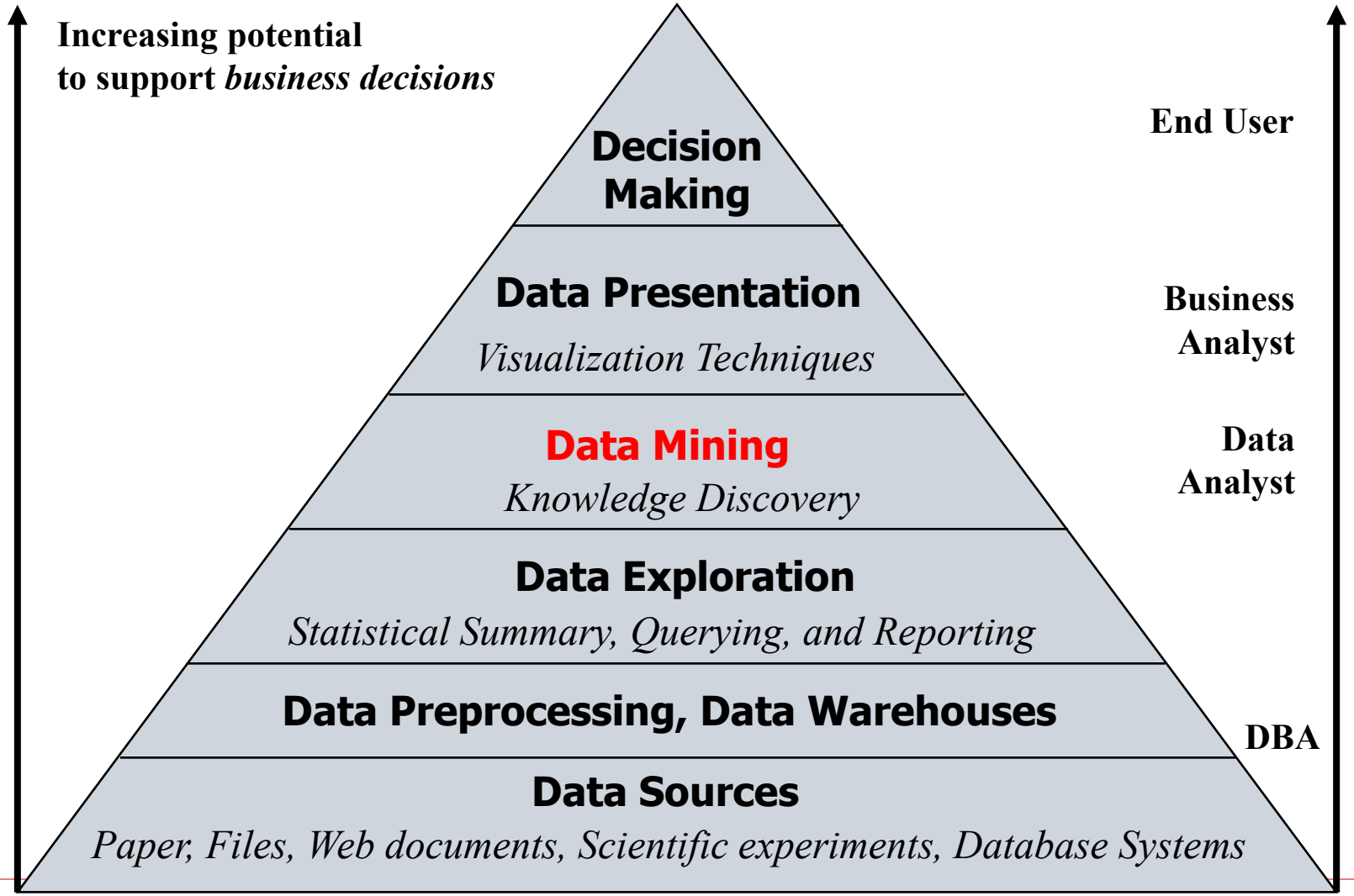
- Data mining as the core of knowledge discovery process
  - To extract interesting, non-trivial, implicit, previously unknown, and potentially useful information from huge amount of data
    - e.g., around 60% of the customers buy diapers also buy beer

Knowledge discovery in databases

- KDD Steps



Data Mining in Business Intelligence



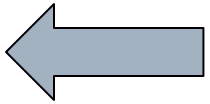
data, knowledge/pattern, technologies, applications, …

A MULTIDIMENSIONAL VIEW OF DATA MINING

What Kinds of Data Can Be Mined?

- Database data (Relational)
  - A collection of tables; each consists of a set of attributes and stores a set of tuples. Each tuple represents an object identified by a unique key and described by values.
  - **What to mine:** searching for trends or data patterns

e.g., by analyzing customer data, we could predict the credit risk of new customers based on their income, age, and previous credit info.



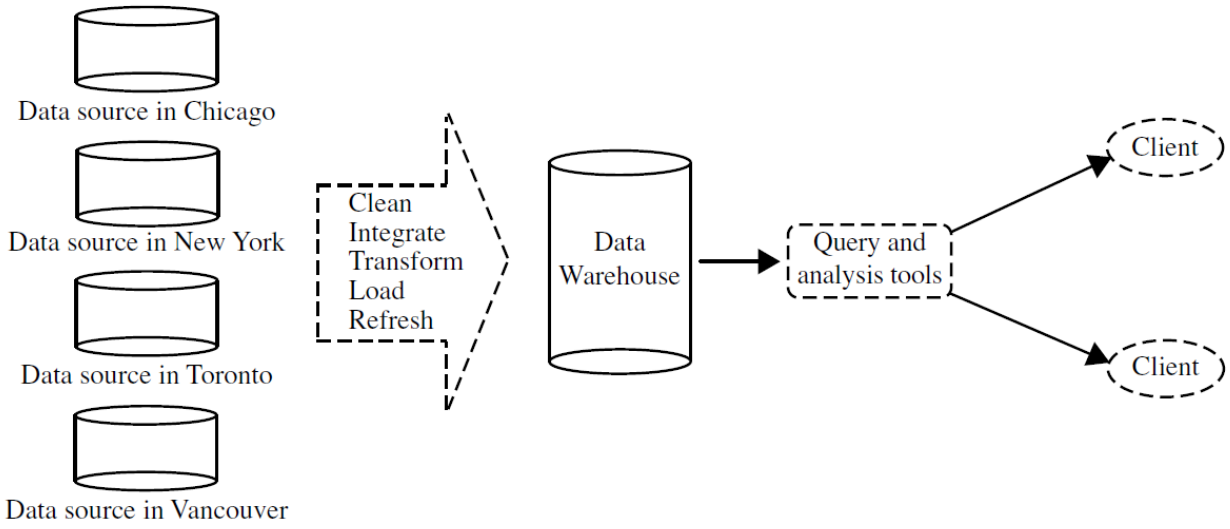
customer	(*cust_ID, name, address, age, occupation, annual_income, credit_information, category, . . .*)
item	(*item_ID, brand, category, type, price, place_made, supplier, cost, . . .*)
employee	(*empl_ID, name, category, group, salary, commission, . . .*)
branch	(*branch_ID, name, address, . . .*)
purchases	(*trans_ID, cust_ID, empl_ID, date, time, method_paid, amount*)
items_sold	(*trans_ID, item_ID, qty*)
works_at	(*empl_ID, branch_ID*)

Relational schema for a relational database, *AllElectronics*.

What Kinds of Data Can Be Mined?

- Data warehouse
  - A repository of information collected from multiple sources, but stored under a unified schema, and often residing at a single site.
  - The data are typically summarized – a multidimensional view of data and the precomputation/fast access of historical data.

e.g., rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item type for each store, or summarized to a higher level, for each sales region.



Typical framework of a data warehouse for *AllElectronics*.

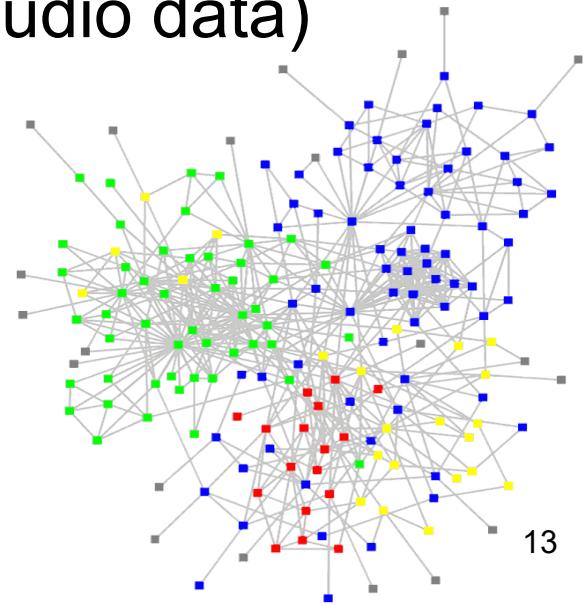
What Kinds of Data Can Be Mined?

- Transactional data
  - One record per transaction (change, event, interaction, etc.)
    - e.g., a customer's purchase, a flight booking, or a click on a web page
  - **What to mine:** market basket data analysis, e.g., *"Which items sold well together?"* → bundle groups of items for boosting sales
    - Data mining on transactional data can do so by mining frequent itemsets (patterns), i.e., sets of items that are frequently sold together.

trans_ID	*list_of_item_IDs*
T100	{computer, printer}
T101	{computer, printer, keyboard}
…	…

What Kinds of Data Can Be Mined?

- Other kinds of data
  - data streams (e.g., video surveillance and sensor data)
  - time-related or sequence data (e.g., stock data, time-series data)
  - graph or networked data (e.g., social networks)
  - spatial data (e.g., maps)
  - multimedia data (e.g., text, image, video, and audio data)
  - the Web data
  - …



What Kinds of Patterns Can Be Mined?

- A classification of data mining functionalities:
  - Descriptive mining tasks *characterize* properties of the data in a target data set.
    - Data characterization and discrimination
    - Mining frequent patterns, associations, and correlations
    - Cluster analysis
    - Outlier analysis
    - …
  - Predictive mining tasks *perform induction* on the current data in order to make predictions.
    - Classification and regression …

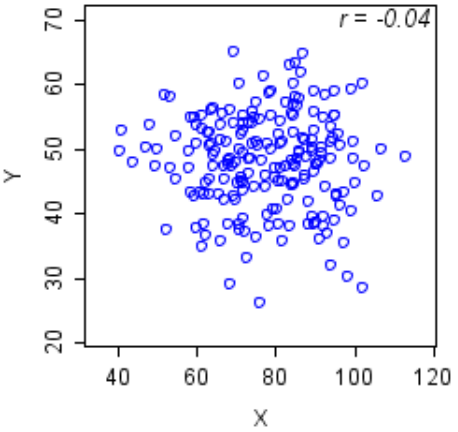
(1) Data Characterization and Discrimination

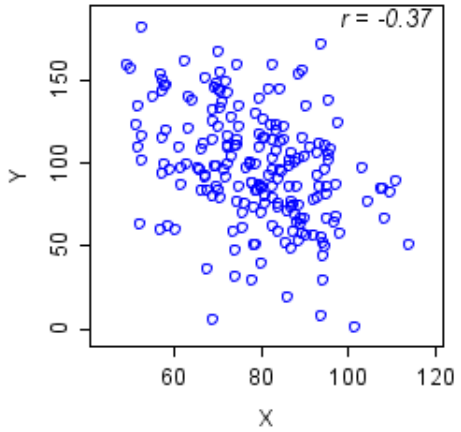
- Data entries can be associated with classes or concepts, e.g.,
  - classes of items for sale include *computers* and *printers*
  - concepts of customers include *bigSpenders* and *budgetSpenders*

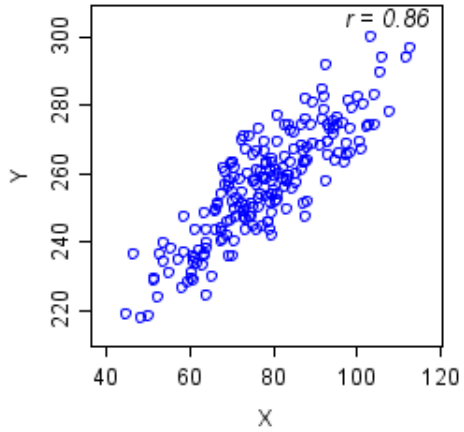
- Such class/concept descriptions can be derived using:
  - Data characterization: by summarizing the data of the class under study (often called the *target class*) in general terms
  - Data discrimination: by comparing the target class with one or a set of comparative classes (often called the *contrasting classes*)

(2) Mining Frequent Patterns, Associations, and Correlations

- Frequent itemset mining: *very fundamental*
  - To find a set of items that often appear together in transactions
  - More advanced: frequent subsequences or substructures
- Mining frequent patterns → the discovery of interesting associations and correlations within data
  - Association: "IF-THEN" rules with frequent co-occurrence in a dataset
  - Correlation: statistical relationship between two random variables







A typical association rule:

- Computer → Printer [*support* = 1%, *confidence* = 50%]

Q: Are strongly associated items also strongly correlated?

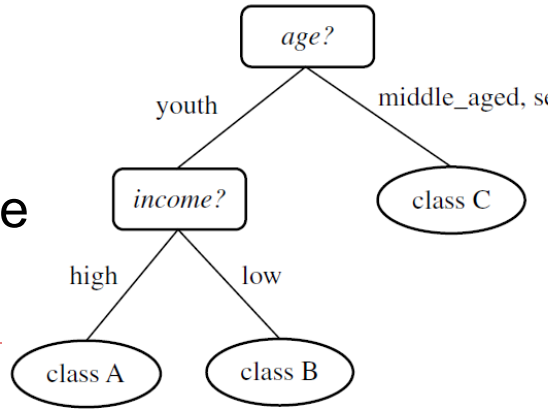
(3) Classification and Regression for Predictive Analysis

- **Classification** and label prediction – *categorical/discrete* labels
  - Build models (functions) based on a set of training data
  - Describe and distinguish classes or concepts for future prediction
  - Predict label-unknown objects

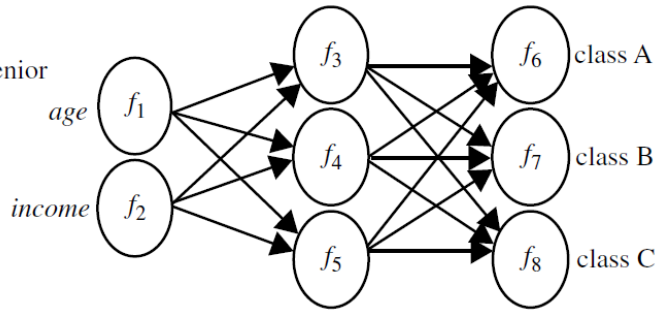
$$age(X, \text{“youth”}) \; AND \; income(X, \text{“high”}) \rightarrow class(X, \text{“A”})$$
$$age(X, \text{“youth”}) \; AND \; income(X, \text{“low”}) \rightarrow class(X, \text{“B”})$$
$$age(X, \text{“middle\_aged”}) \rightarrow class(X, \text{“C”})$$
$$age(X, \text{“senior”}) \rightarrow class(X, \text{“C”})$$

Classification rules (i.e., IF-THEN)

- **Regression** and numeric prediction
  - Model continuous-valued functions
  - Predict missing or unavailable numerical data values
    - e.g., the Hang Seng Index, stock price



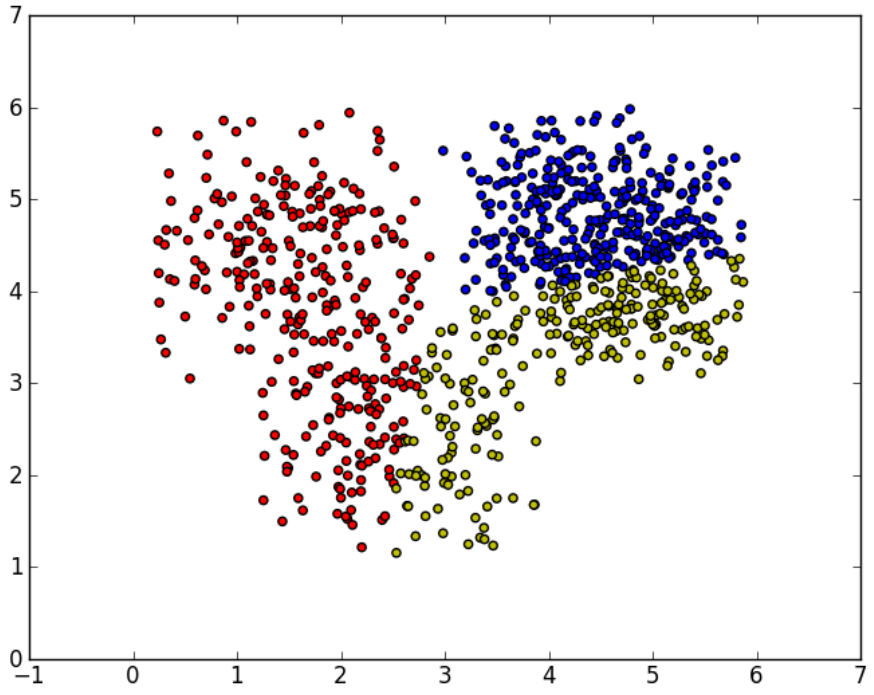
Decision trees



Neural networks

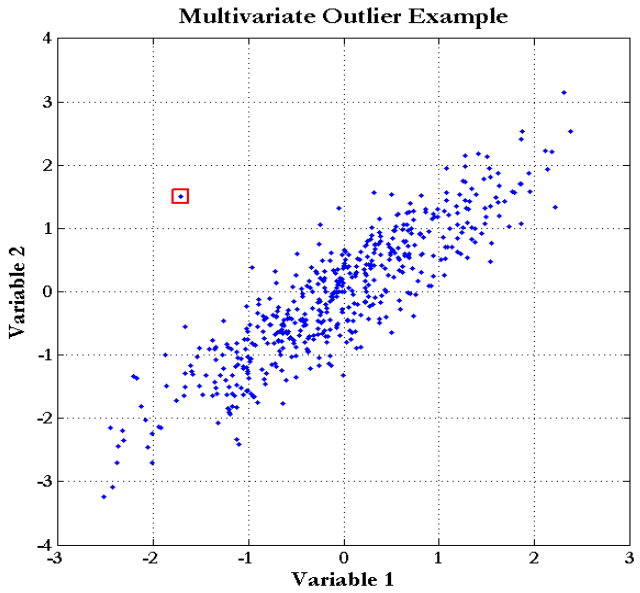
(4) Cluster Analysis

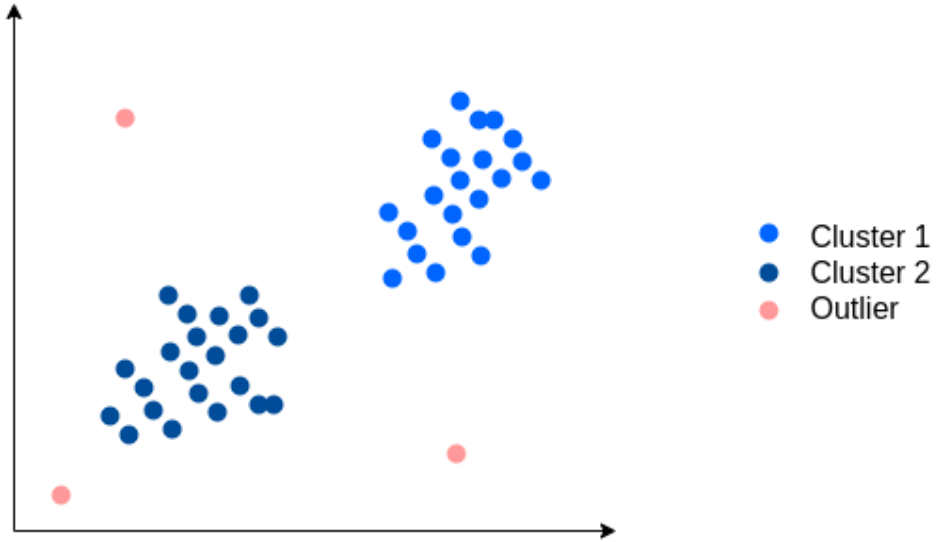
- **Class labels is unknown**: group data objects to form new classes, e.g., categorize web pages to define topics
- **Principle**: maximize the intra-cluster similarity & minimize the inter-cluster similarity
  - By doing so, objects within a cluster have high similarity, but are rather dissimilar to objects in other clusters.
  - Each cluster so formed can be viewed as *a class of objects*.



(5) Outlier Analysis

- **Outlier / Anomaly**: a data object that does not comply with the general behavior of the data, typically *noise.*
  - *Exception*? Sometimes, rare events can be more interesting than the more regularly occurring ones, e.g., fraud detection, hackers.
- **Methods**: by regression or clustering analysis, …





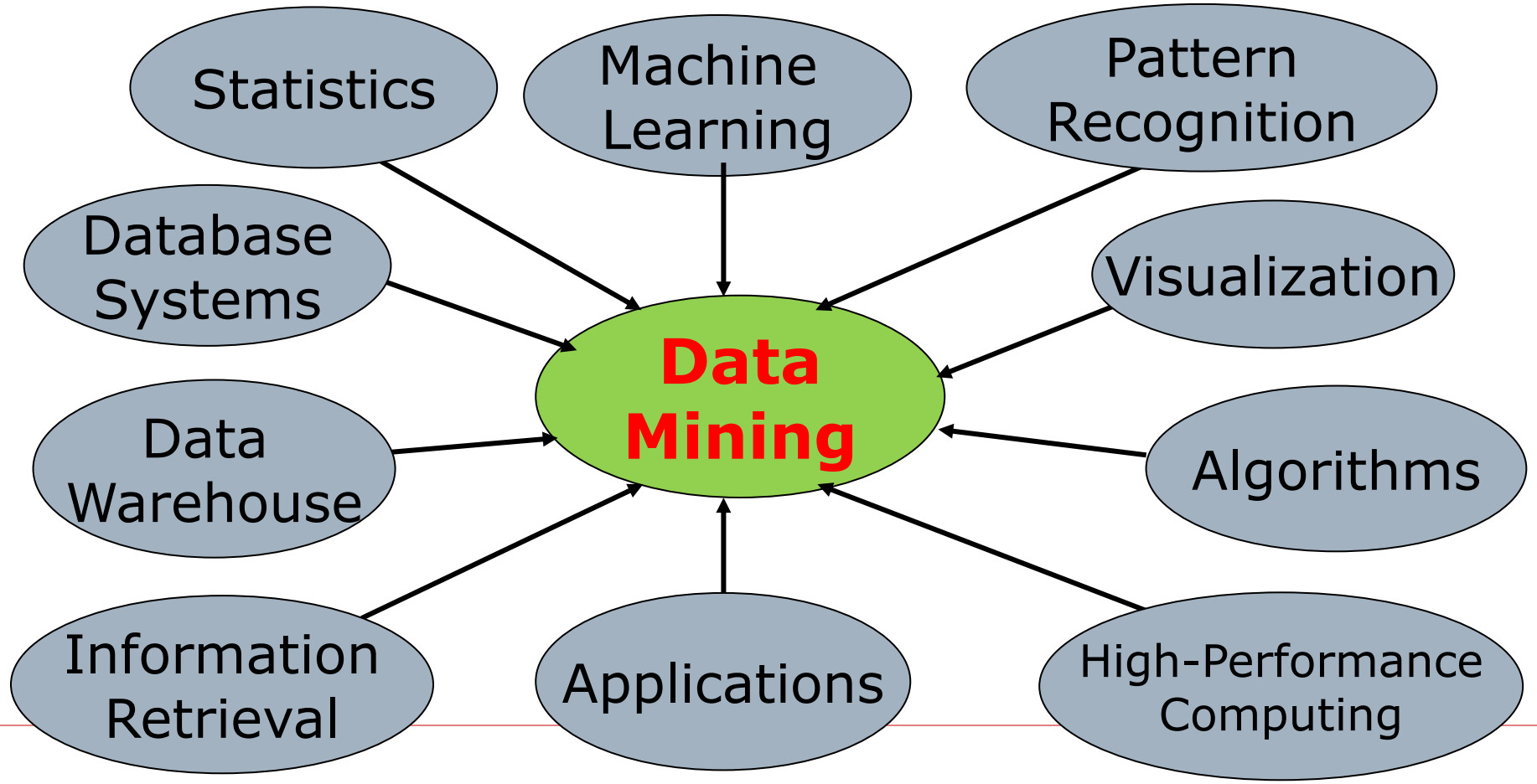
Are All Patterns Interesting?

- **No!**
  - One can mine tremendous amount of "patterns"
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative enough, may be transient, …
- A pattern is interesting if it is:
  - 1) easily understood by humans; 2) valid on new or test data with some degree of certainty; 3) potentially useful; and 4) novel
- An interesting pattern represents knowledge.
  - **Interestingness measures**: e.g., support, confidence, …, each associated with a user-controlled threshold

$$support(X \Rightarrow Y) = P(X \cup Y),$$
$$confidence(X \Rightarrow Y) = P(Y|X).$$

Which Technologies Are Used?

- **Data mining:** highly application-driven and interdisciplinary



What Kinds of Applications Are Targeted?

- Business Intelligence (BI)
  - To understand the commercial context of an organization, e.g., their customers, the market, supply and resources, competitors
    - *Target marketing*: find clusters of customers who share similar features
  - To offer historical/current/predictive views of business operations
    - *Cross-market analysis*: association/correlation between products
    - *Sales prediction, customer profiling, recommendation, …*

What Kinds of Applications Are Targeted?

- Web Search Engines: search for information on the Web

  - The returned list consists of web pages, images, etc.
    - How pages should be ranked?
    - Which ads should be added?
    - How the search results can be personalized?
  - Web page analysis: classification, clustering, ranking
    - By mining the Web access logs in different ways, we could discover user preference and behavior, analyze Web marketing, improve Web site organization, etc.
- Text Mining, e.g., emails
  - SPAM filtering, email decluttering by topics, etc.

Major Issues in Data Mining – *Mining Methodology*

- Mining various and new kinds of knowledge
  - Data mining is not a one-size-fits-all field. As new kinds of data and questions arise, data mining must evolve, leading to the invention of new techniques.
- Mining knowledge in *multi-dimensional* space (across multiple attributes)
  - To provide a flexible way to analyze data and reveal diverse insights at various levels of detail, from high-level summaries down to fine-grained details
- Data mining: an *interdisciplinary* effort
  - e.g., text data → DM + IR + NLP; bug analysis → DM + software engineering
- Boosting the power of discovery in an *interconnected* environment
- Handling noise, uncertainty, or incompleteness of data
- Pattern evaluation and *pattern-* or *constraint-guided* mining
  - What makes a pattern interesting may vary from user to user.

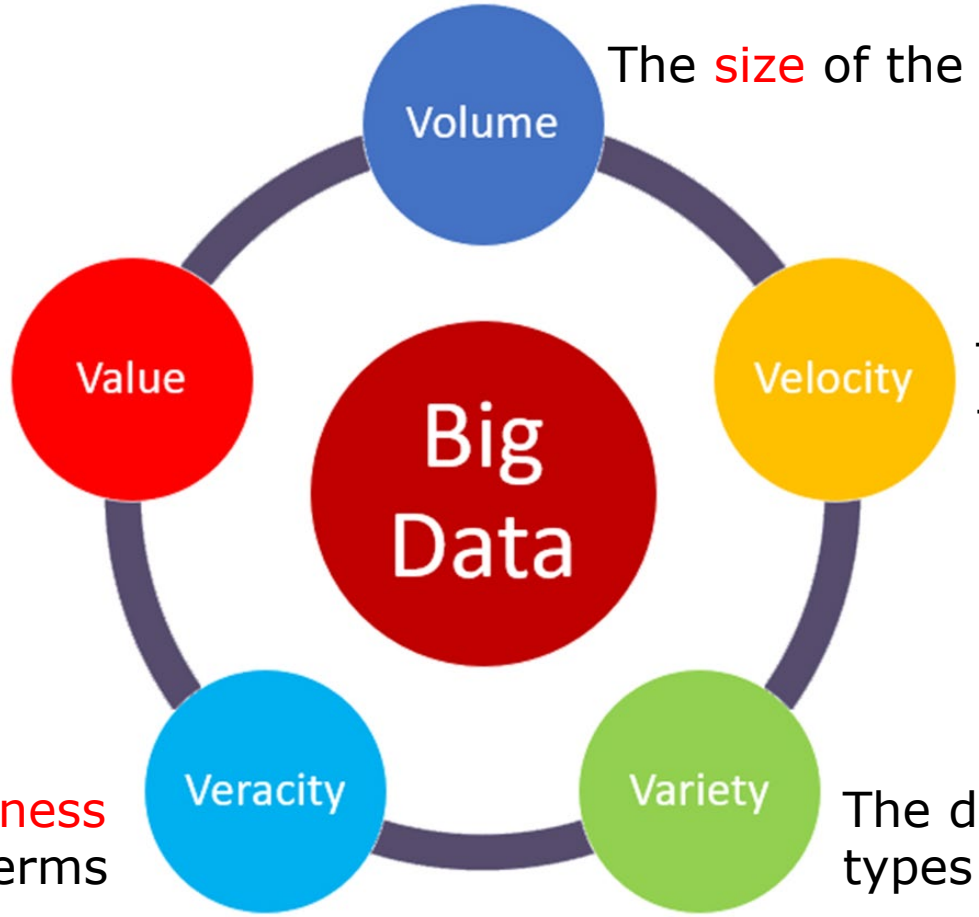
Major Issues in Data Mining – *User Interaction*

- Interactive mining
  - Build flexible and user-friendly interface that allows users to engage with the data and the mining system in an exploratory manner
  - e.g., a summary of sales data → sales by region/product/quarter
- Incorporation of background knowledge
  - Domain-specific pattern evaluation
  - Guide the search toward interesting patterns
- Ad-hoc data mining and data mining query languages
  - SQL-like query languages support ad-hoc queries.
  - Need high-level flexible data mining query languages for more ad-hoc mining tasks.
- Presentation and visualization of data mining results
  - The discovered knowledge shall be easily understood and directly usable by humans.

Major Issues in Data Mining

- *Efficiency and Scalability*
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- *Diversity of data types*
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- *Data mining and society*
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

The 5 Vs of Big Data



The size of the data

The speed at which the data is generated

The different types of data

The trustworthiness of the data in terms of accuracy

Just having big data is of no use unless we can turn it into **value**!

Summary

- *"Necessity is the mother of invention"* – **big data → data mining**
- Data mining: the process of discovering interesting patterns from massive amounts of data
  - *KDD Process*: data cleaning/integration, data selection/transformation, **pattern mining/evaluation**, knowledge representation
  - *Tasks*: characterization/discrimination, frequent patterns, association, correlation, classification, regression, clustering, outlier detection, etc.
  - *Interestingness*: certainty, novel, potentially useful, easily understood by human → guide the discovery process in turn
- Data warehouse: a repository for multiple-source data stored under a unified schema and typically summarized
  - with multidimensional data analysis capabilities for decision making

Email: fengmei.jin@polyu.edu.hk

Office: PQ747

THANK YOU!