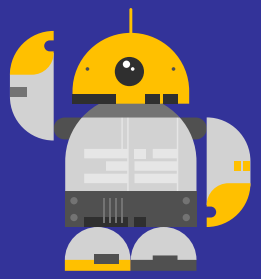


Database Systems

Lecture 2

Dr. T. Akilan

takilan@lakeheadu.ca

Database

- A database system, also called a **database management system** (DBMS)
- Consists of a collection of **interrelated data**
- Provides mechanisms
  - for defining database **structures** and data **storage**;
  - for **specifying** and **managing** concurrent, shared, or distributed data access;
  - for ensuring **consistency** and **security** of the information stored despite system crashes or attempts at unauthorized access.

Database Cont.

- A **relational database** is a collection of **tables**, each of which is assigned a **unique name**.
- Table:
  - consists of a set of **attributes** (columns or fields)
  - stores a large set of **tuples** (records or rows).
- Tuple represents an object identified by a unique key and described by a set of attribute values.

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*)

E.g., Relational schema for a relational database, *AllElectronics*.

Image: Data Mining: Concepts and Techniques

Data Warehouses

- A **repository of information** collected from multiple sources, stored under a unified schema, and usually residing at a single site.
  - It is constructed via a process of **data cleaning, data integration, data transformation, data loading**, and **periodic data refreshing**.

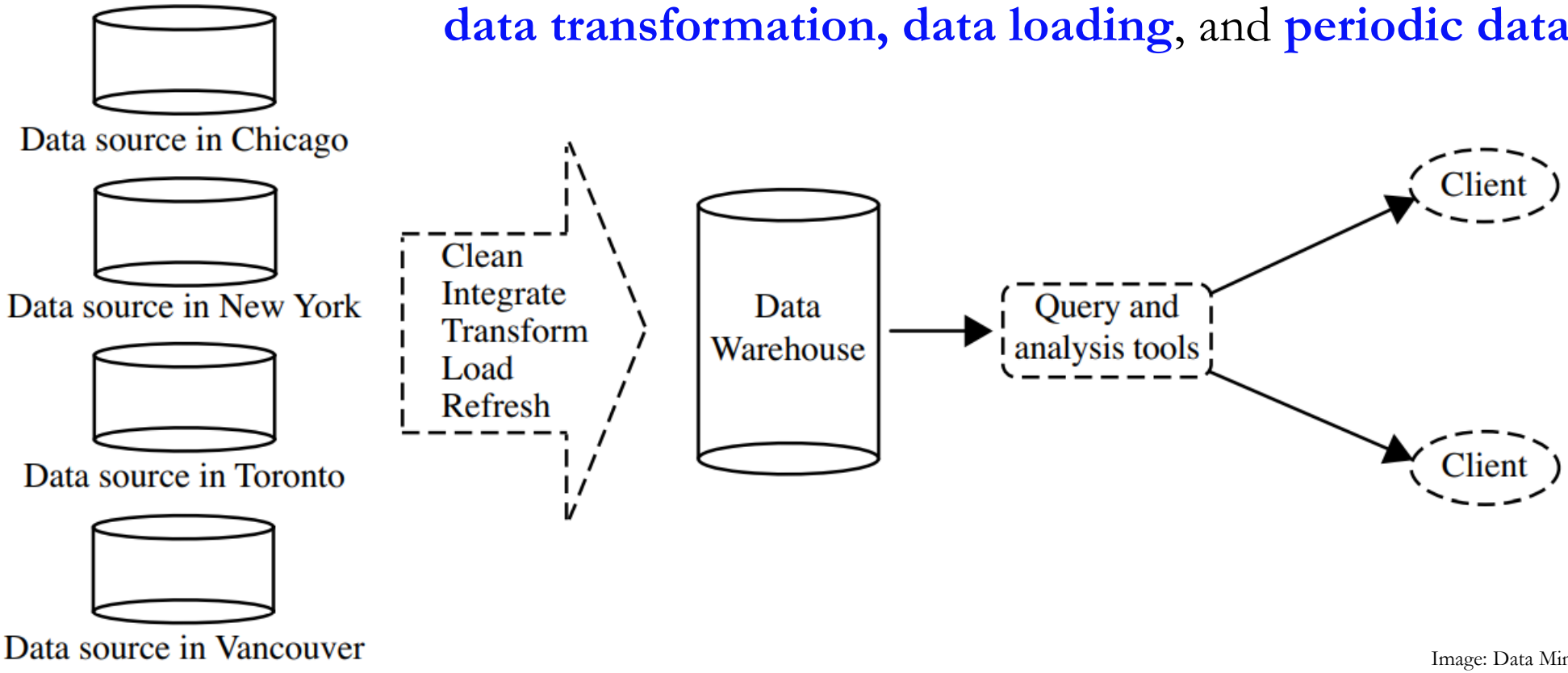


Image: Data Mining: Concepts and Techniques

Data Warehouses Cont.

- **Multidimension**:
  - It is a multidimensional **data structure**, called a **data cube.**
  - **E**ach dimension corresponds to an attribute or a set of attributes in the schema
  - Each **cell** stores the value of some aggregate measure such as count-`sum(sales_amount)`.
  - A data cube provides a **multidimensional view** of data and allows the precomputation and fast access of summarized data.

Data Warehouses Cont.

- E.g., the figure shows a multidimensional data cube, commonly used for data warehousing:
  - (a) shows summarized data for *AllElectronics*
  - (b) shows summarized data resulting from drill-down and roll-up operations on the cube in (a)

Note: for improved readability, only some of the cube cell values are shown in the figure.

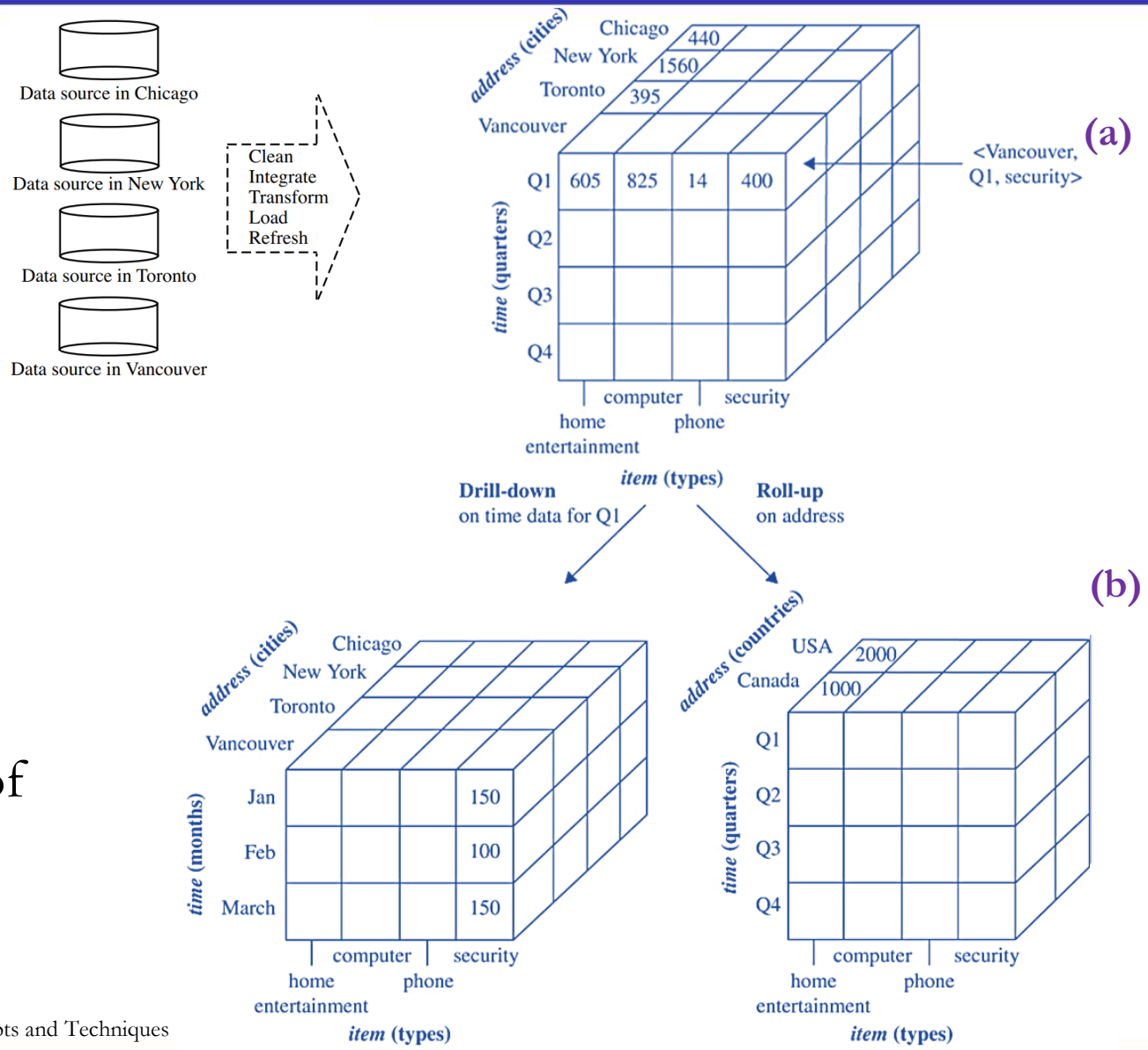


Image: Data Mining: Concepts and Techniques

Database vs Data Warehouse

Parameter	Database	Data Warehouse
Purpose	To perform fundamental operations, like transactional data recording	To analyze historical data, like predicting business trend
Processing Method	Online Transactional Processing (OLTP)	Online Analytical Processing (OLAP)
Availability	Data is available real-time	Data is refreshed from source systems as and when needed
Data Type	Current - Data stored/captured in the Database is up to date	Historical (analytical) data. May not be up to date
Query Type	Simple transaction queries	Complex queries are used for analysis purpose
Abstract Level	Detailed data is stored in a database	It stores highly summarized data

Data Sources for Knowledge Extraction

- What kind of data can we use?
  - Relational database, data warehouse, transactional database
  - Heterogeneous databases and legacy databases
  - Data streams and sensor data
  - Data sequences: temporal data, sequences of actions, DNA
  - Structured data, graphs, social networks and multi-linked data
  - Spatial data and spatiotemporal data
  - Multimedia databases
  - Text databases
  - And of course, the World-Wide Web (WWW)

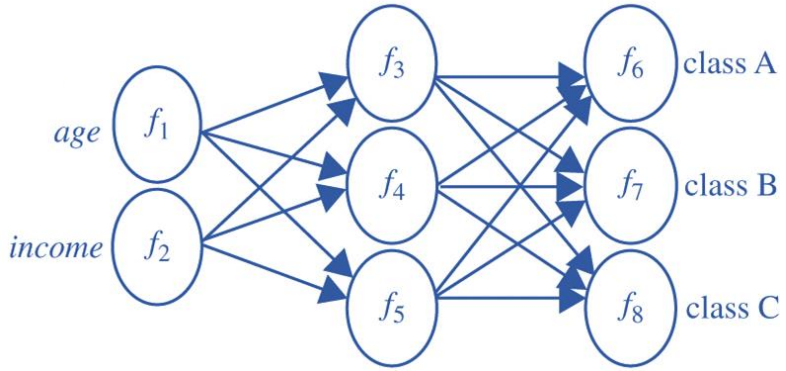
Knowledge Extraction

- What kind of knowledge (**patterns**) can we get?
  - **Frequent patterns** (or frequent itemsets)
    - ✓ What items are frequently purchased together?
  - **Association rules**
    - ✓ What items frequently lead customers to purchase a second item?

What kind of knowledge (patterns) can we get? Cont.

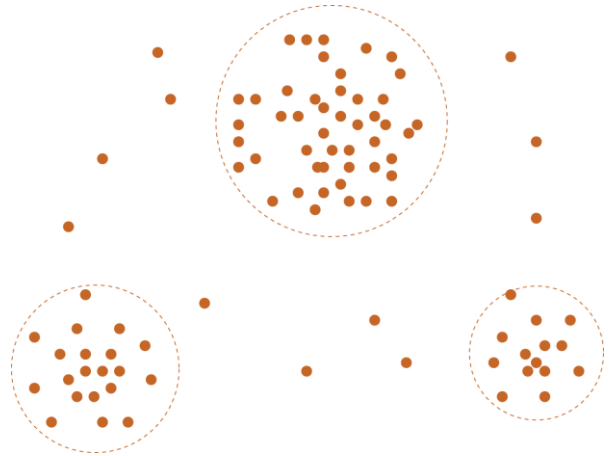
Image: Data Mining: Concepts and Techniques

- **Clustering**
  - Group data to form **new categories by maximizing intra-class similarity** & **minimizing interclass similarity**
  - Grouping houses into different social neighborhoods



A classification model is represented in the form of a NN

- **Classification** (label prediction)
  - A large family of problems, "what type of thing is this?"
  - Find attributes common to items in one class and missing from items in other classes
  - Text classification based on keyword occurrences



A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

What kind of knowledge (patterns) can we get? Cont.

- **Regression** (numeric prediction)
  - Models continuous-valued functions.
  - Predicts missing or unavailable numerical data values rather than (discrete) class labels.
  - Encompasses the identification of distribution trends based on the available data

- **Sequential patterns**
  - From sequential data, trends, time-series
  - Discovering DNA sequences common to specific populations, sequences of commands in program hacking

  - Streams:
    - ✓Potentially-infinite time sequences to analyze in real-time (video, for instance)

Data Mining Challenges

- Heterogeneous data
  - When you build your own database and populate it yourself, all the data is ok
  - But when you're dealing with a data warehouse...
    - ✓Different databases with different fields
    - ✓Legacy databases with outdated information
    - ✓Noise and missing information
    - ✓User-submitted information of questionable quality
- Efficient processing
  - Your algorithm can extract knowledge from an encyclopaedia article in 15 seconds!
  - There are 4M articles in Wikipedia
  - It will take 2 years to finish...

Data Mining Challenges Cont.

- Outliers:
  - A piece of data that is very unlike everything else around it
  - Including it causes large differences in the average statistics, but excluding it requires special exception rules
    - ✓ How much cash do you have in your pockets?
    - ✓ 10-people random sample: $10, $10, $10, $10, $10, $10, $10, $10, $10, $1,000,000
    - ✓ Average amount people have = $100,009?

- Outlier analysis:
  - It may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account.

Data Mining Challenges Cont.

- High dimensionality
  - Most complete data warehouse have a lot of information (dimensions) about each item
    - ✓ A shopping database can have very detailed data about purchases
    - ✓ "the customer bought 2% milk on special"
  - Can cause us to discover patterns that are too specific to be useful, or to miss more general patterns
  - Dimensionality reduction: abstracting away some details
    - ✓ 2% milk on special = 2% milk = milk = dairy products = groceries?
    - ✓ i.e., from multidimensional space to lower dimensional space

Data Mining Challenges Cont.

- **Handling uncertainty, noise, or incompleteness of data:**
  - Data often contain noise, errors, exceptions, or uncertainty, or are incomplete.
  - Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns.
  - Data cleaning, data preprocessing, outlier detection and removal, and uncertainty reasoning are examples of techniques that need to be integrated with the data mining process.

- **Background knowledge**
  - Some patterns are obvious from the data for us, because of our background knowledge
    - ✓ We know a text document mentioning "wall street" is probably about finance
    - ✓ Because we know the New York Stock Exchange (NYSE) - a major financial institution is on Wall Street
  - How to include such knowledge into a database system?

Data Mining Challenges Cont.

- **Evaluating** the knowledge
  - One can mine tremendous amount of patterns
  - How to know which ones are good or bad?
  - Different evaluation metrics for different patterns and applications
    - ✓ Predictive, coverage, statistical measures (precision, recall, accuracy, f-measure), computational complexity, etc.

Applications

- **Basket data analysis and targeted marketing**
  - Given a database of customers with demographic information, location, and past purchase behaviour
  - Determine the profile of the most profitable customers
  - Tailor advertisement campaigns to attract and retain these customers

- **Fraud detection**
  - Automobile Insurance Bureau of Massachusetts had a database of insurance claims, including over 60 attributes such as claimant, type of accident, type of injury/treatment, and expert opinion of real vs. fraud
  - Dimension Reduction methods used to obtain weighted variables, then identified subsets of characteristics strongly correlated with fraud

Applications

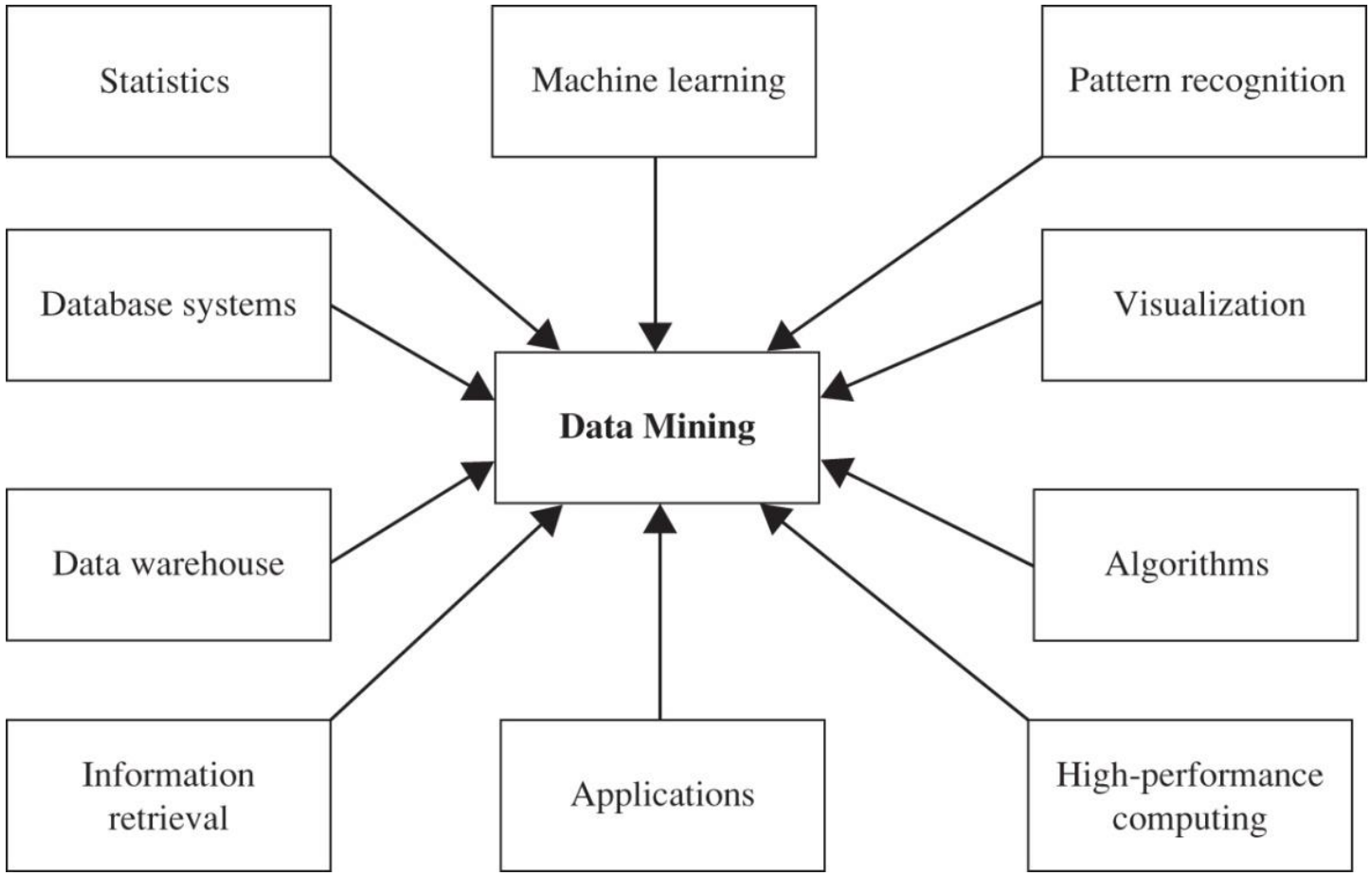
- **Web page analysis:**
  - Page ranking, for example, Google search engine results
  - Recommender systems (Amazon)
  - Clicks-to-Customers
    - ✓ 50% of Dell's customers order their computer through the web, but 0.5% of visitors of Dell's web page become customers
    - ✓ Dell has navigation history of visitors through their site
    - ✓ Cluster customers through their click sequences, and design web pages to maximize the number of customers
- **Biological and medical data analysis:**
  - classification, cluster analysis, biological sequence analysis, biological network analysis
- Engineering research and development (Watson)

Summary

- Data (collected and generated) and individual databases are being consolidated into massive data warehouses
  - Getting knowledge from this massive amount of data is a challenge

- Data mining: Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns from huge amount of data
  - Different types of data
  - Different patterns of interest
  - Different applications
  - Different challenges

Summary Cont.



Data mining adopts techniques from many domains

Image: Data Mining: Concepts and Techniques

Summary Cont.

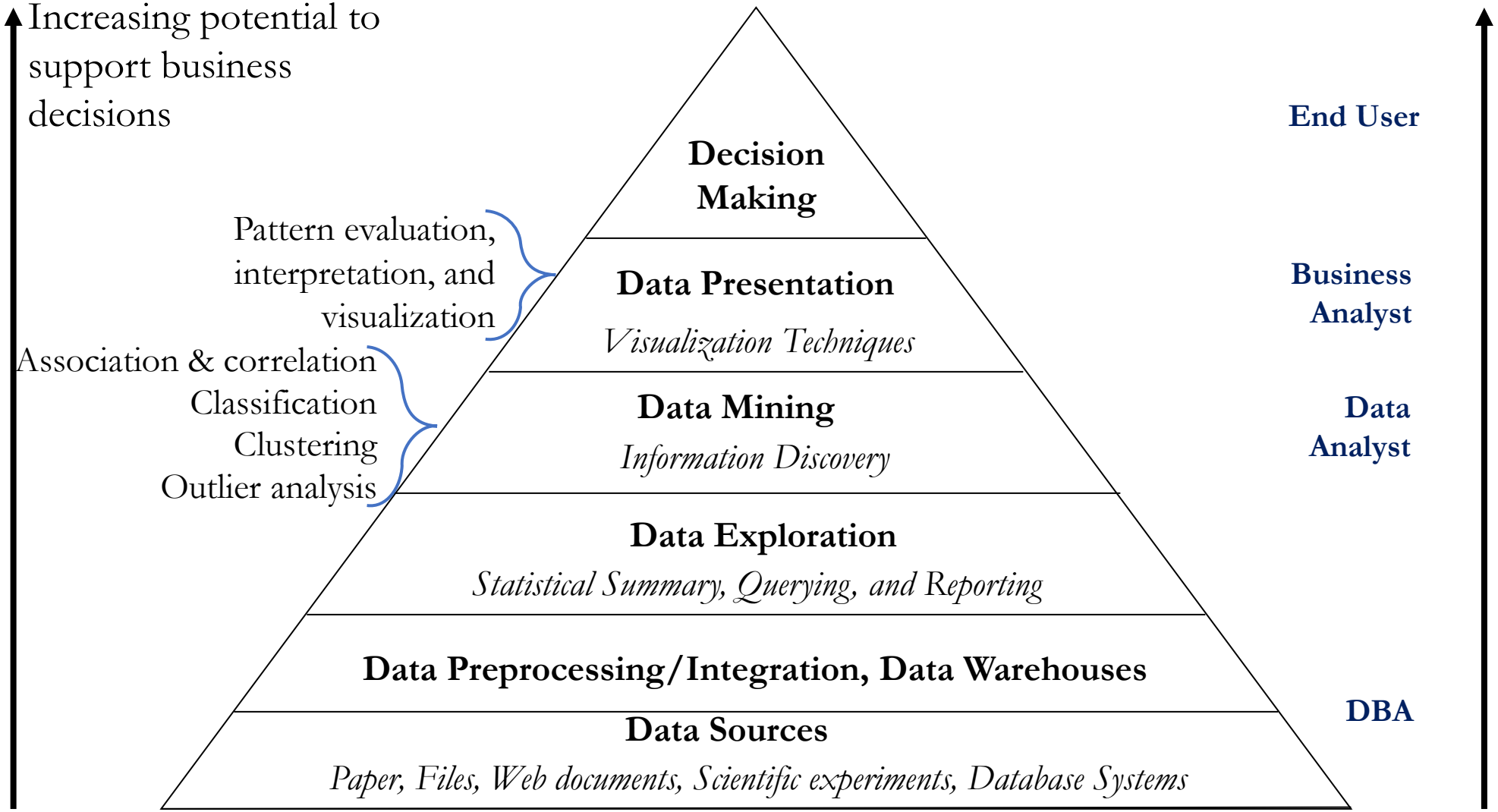


Image: Data Mining: Concepts and Techniques