SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Computing and Information Systems

Week 1 - Introduction

Instructor: Zhiwu HUANG

School of Computing and Information Systems

Singapore Management University

Email: zwhuang@smu.edu.sg

Courtesy: **Zhaoxia WANG**

Outline

- Definition of Data Mining (What)
- Motivation of Data Mining (Why)
- Background & Application of Data Mining
  - Data Mining & Knowledge Discovery in Databases (KDD)
  - Data Mining & Other Domains
  - Data Mining & Business Intelligence
- Data Mining Tasks (How)
- Classification of Data Mining Systems
- Summary
- Lab Section
- Appendix*

What is Data Mining?

Discovering interesting patterns from large data

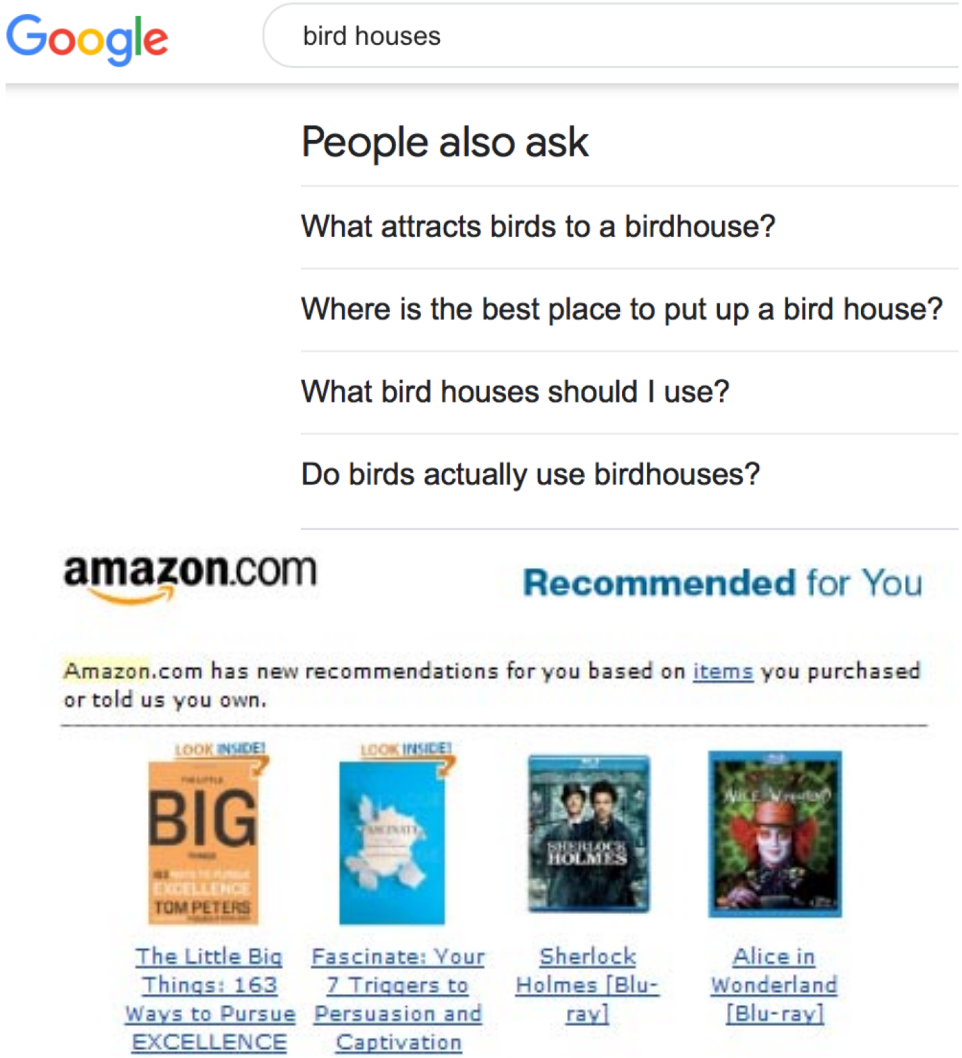
Source: https://adexchanger.com/comic-strip/adexchanger-audience-data-mining/

What is Data Mining?

- Data mining is the process of analysing large amounts of data in order to discover patterns or information, which are:
  - **implicit**
  - **previously unknown**
  - **potentially useful**
- It is typically performed on databases, which store data in a structured format.
- By "mining" large amounts of data, hidden information can be discovered and used for other purposes.

https://techterms.com/definition/data_mining
https://www.sas.com/en_sg/insights/analytics/data-mining.html

Data Mining Examples



- **implicit**
- **previously unknown**
- **potentially useful**

Data Mining Examples

- Consider the example of SMU student grade database
  - The chair of SCIS wanted to find out the relationship between students and instructors: whether for a particular instructor a student tends to perform well
- By using SQL query, it is not at all simple.
- To achieve this,
  - Several tables are joined together, such as Grades, Courses, Students and Instructors
  - Specific columns are extracted, such as Student_Id, Instructor_Id, grade
  - Association rule mining can be applied to obtain possible association (**useful knowledge**) between instructor and grade

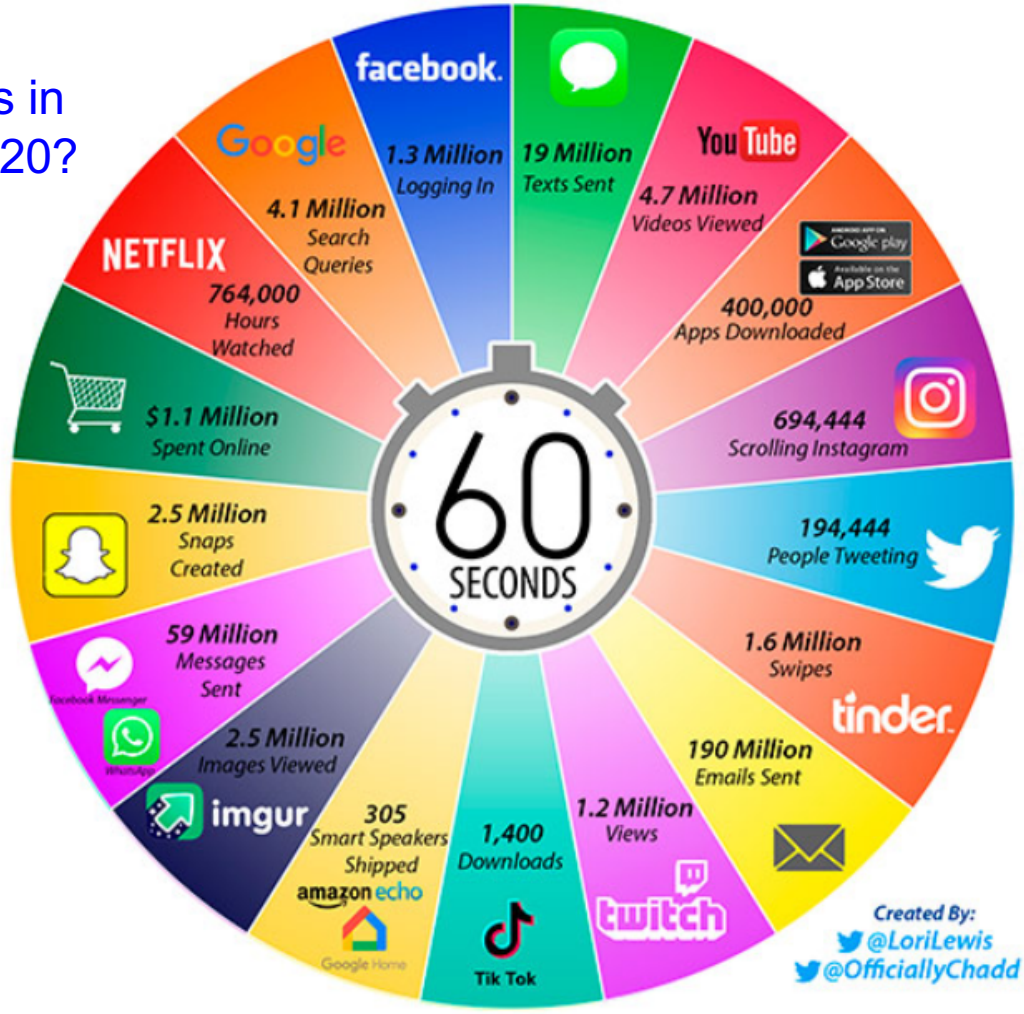
What is (Not) Data Mining?

- Example 1: Look up phone number in a phone directory
- Example 2: Look for certain last names which are more prevalent in certain regions
- Example 3: Query a Web search engine for information about “Amazon”
- Example 4: Group together similar documents returned by search engine according to their content/context (e.g. Amazon rainforest, Amazon.com, etc)

- **Implicit**
- **previously unknown**
- **potentially useful**

Why Data Mining?
"We are Drowning in Data…"

What Happens in a Minute in 2020?



Source:https://www.allaccess.com/merge/archive/31294/infographic-what-happens-in-an-internet-minute

"We are Drowning in Data…"
– From Commercial Viewpoint

- Lots of data are being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions
  - Social networks

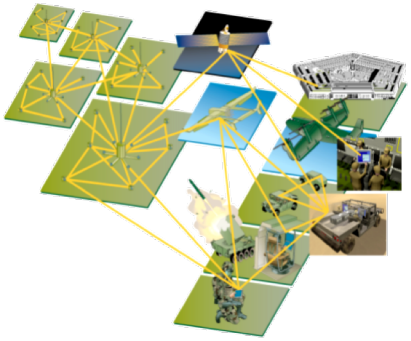
twitter

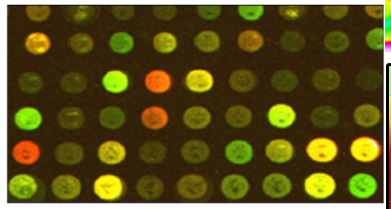
facebook

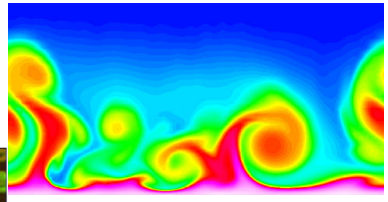
- Computers have become cheaper and more powerful
- Competitive pressure is strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

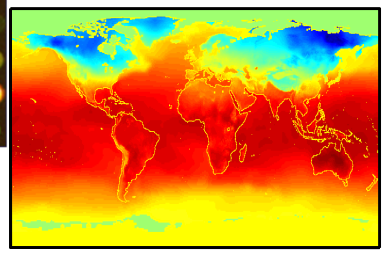
"We are Drowning in Data…"
– From Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data in biology
  - scientific simulations generating terabytes of data



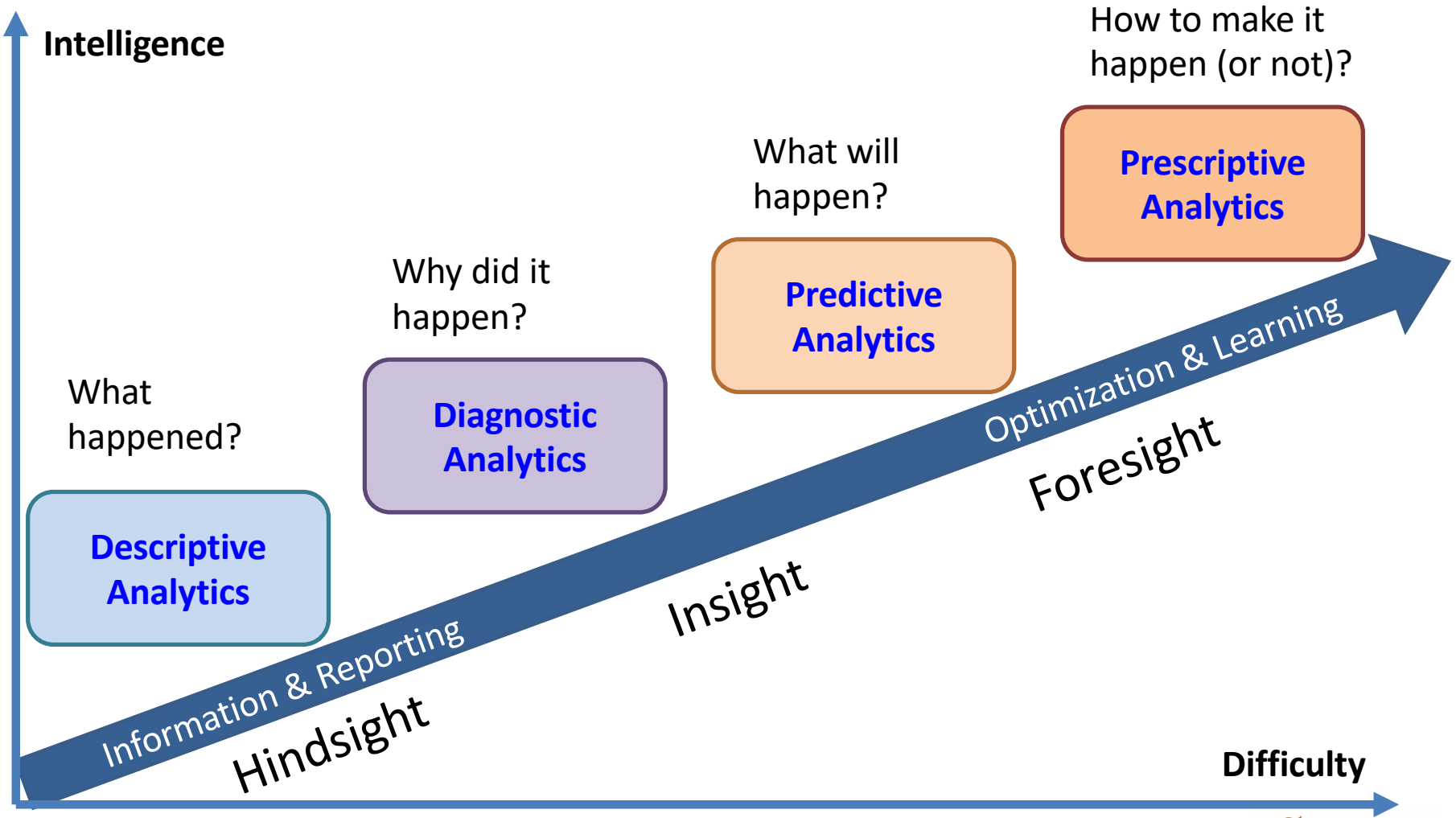






- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

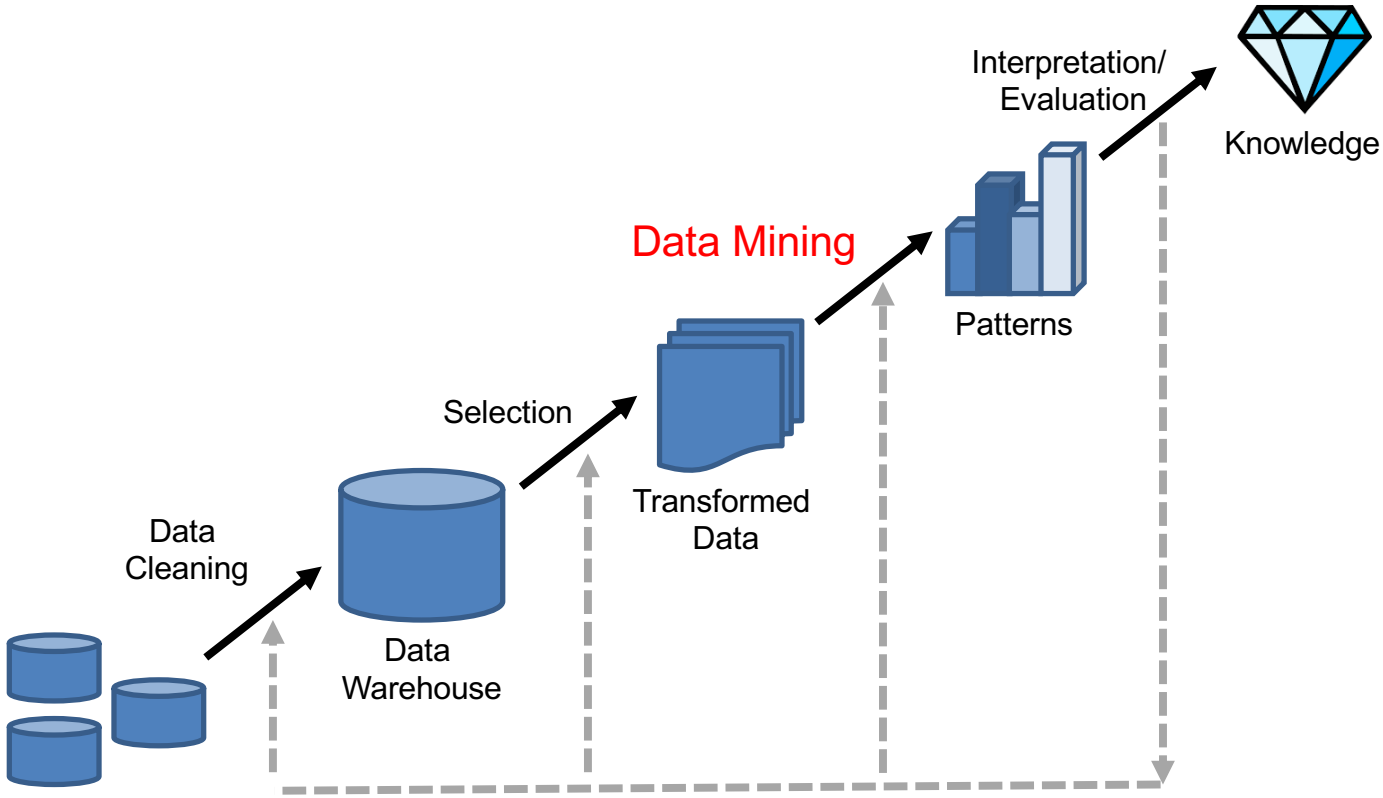
"…But Starving for Knowledge"



Courtesy: Zhaoxia Wang and et al.

Data Mining & KDD

- **Data Mining**: a **core** step of **Knowledge Discovery in Databases (KDD)** that extracts **implicit**, **previously unknown** and **potentially useful** knowledge from large amounts of data



Courtesy: Zhaoxia Wang and et al.

Major Steps in Data Mining (KDD)

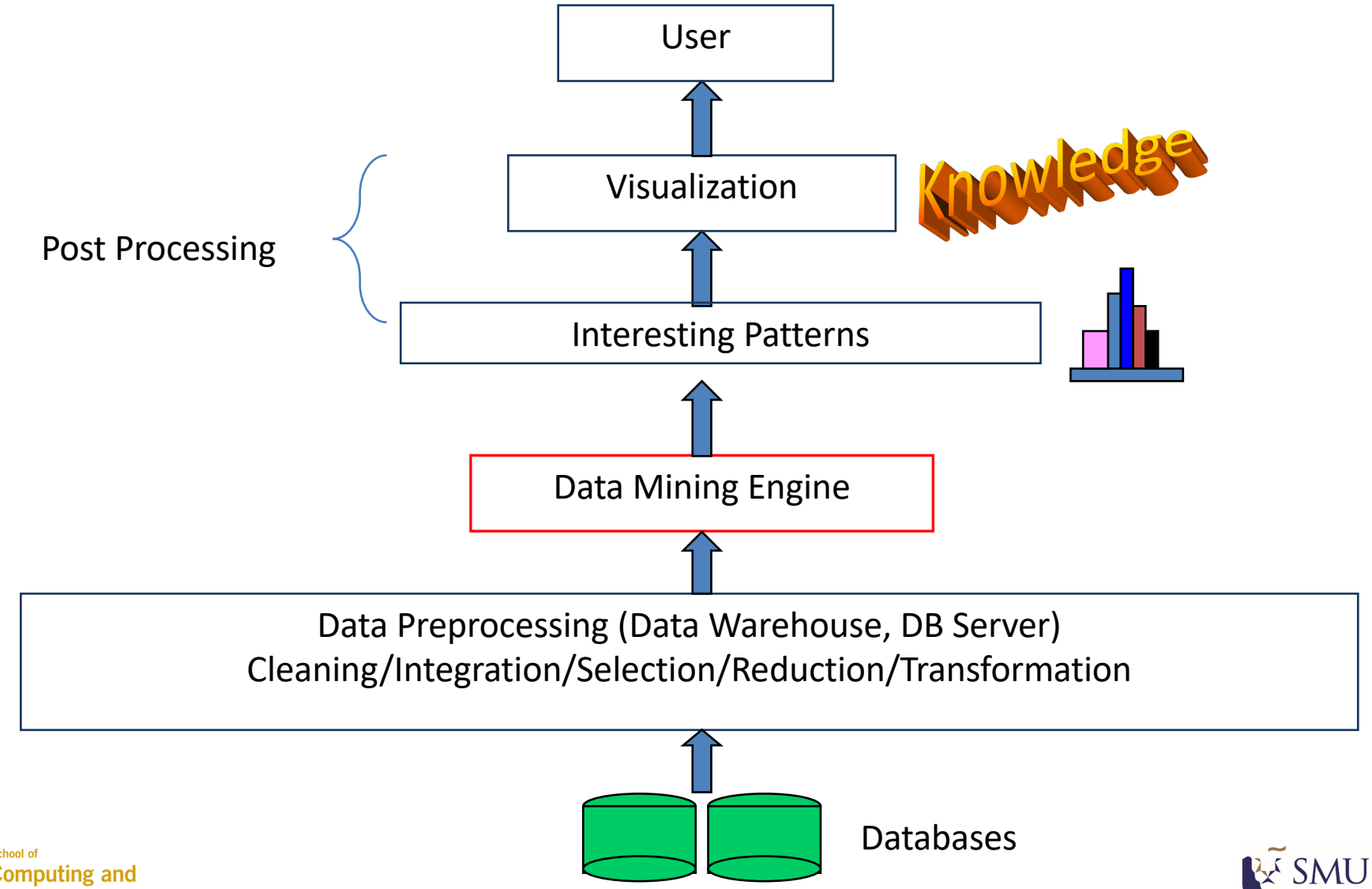
1. Data Preprocessing
   - A. Data Integration
     - Combine multiple data sources
   - B. Data Cleaning
     - Remove noise and inconsistent data
   - C. Feature Selection
     - Select task-relevant features or attributes
   - D. Data Transformation
     - Transform/consolidate selected data for further analysis

Major Steps in Data Mining (KDD)

2. Data Mining
   - Apply data mining & machine learning methods (e.g., classification/clustering) to extract patterns from data
3. Post Processing
   - Visualization
     - Present the mined patterns to users
   - Pattern Evaluation
     - Evaluate and Identify truly interesting patterns

Common Misconception: Although "Data Mining" is just one of the many steps, it is usually used to refer to the whole process of KDD

Architecture of Typical Data Mining System

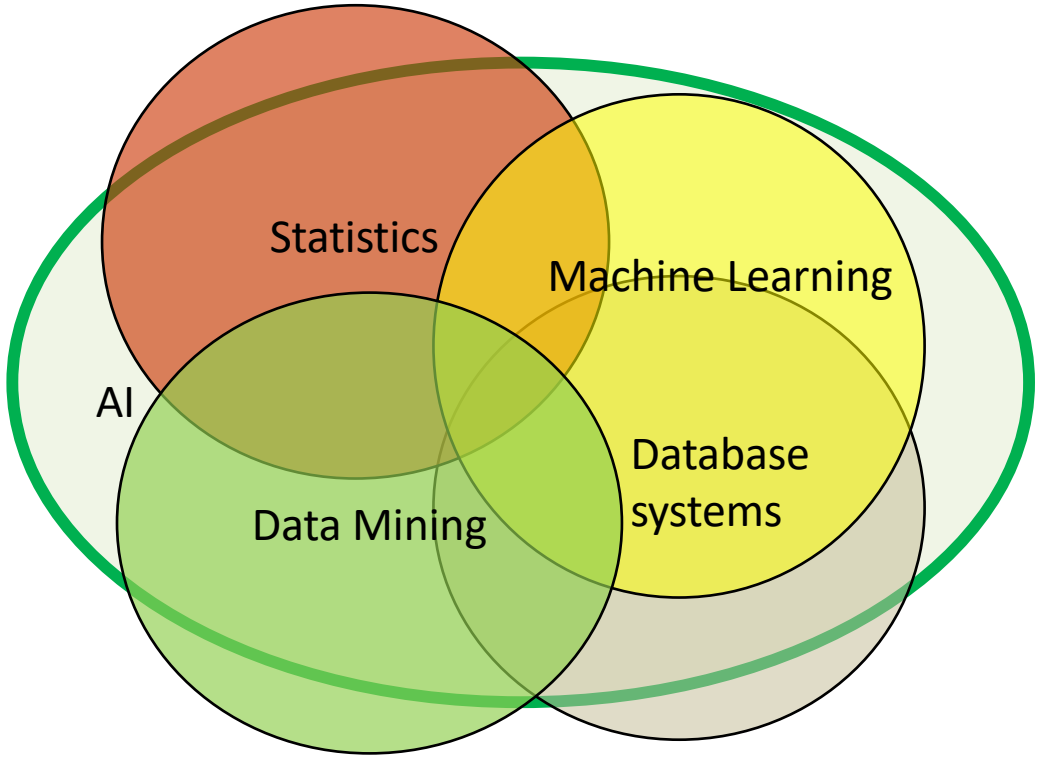


Why Not Use Classical Data Analysis?

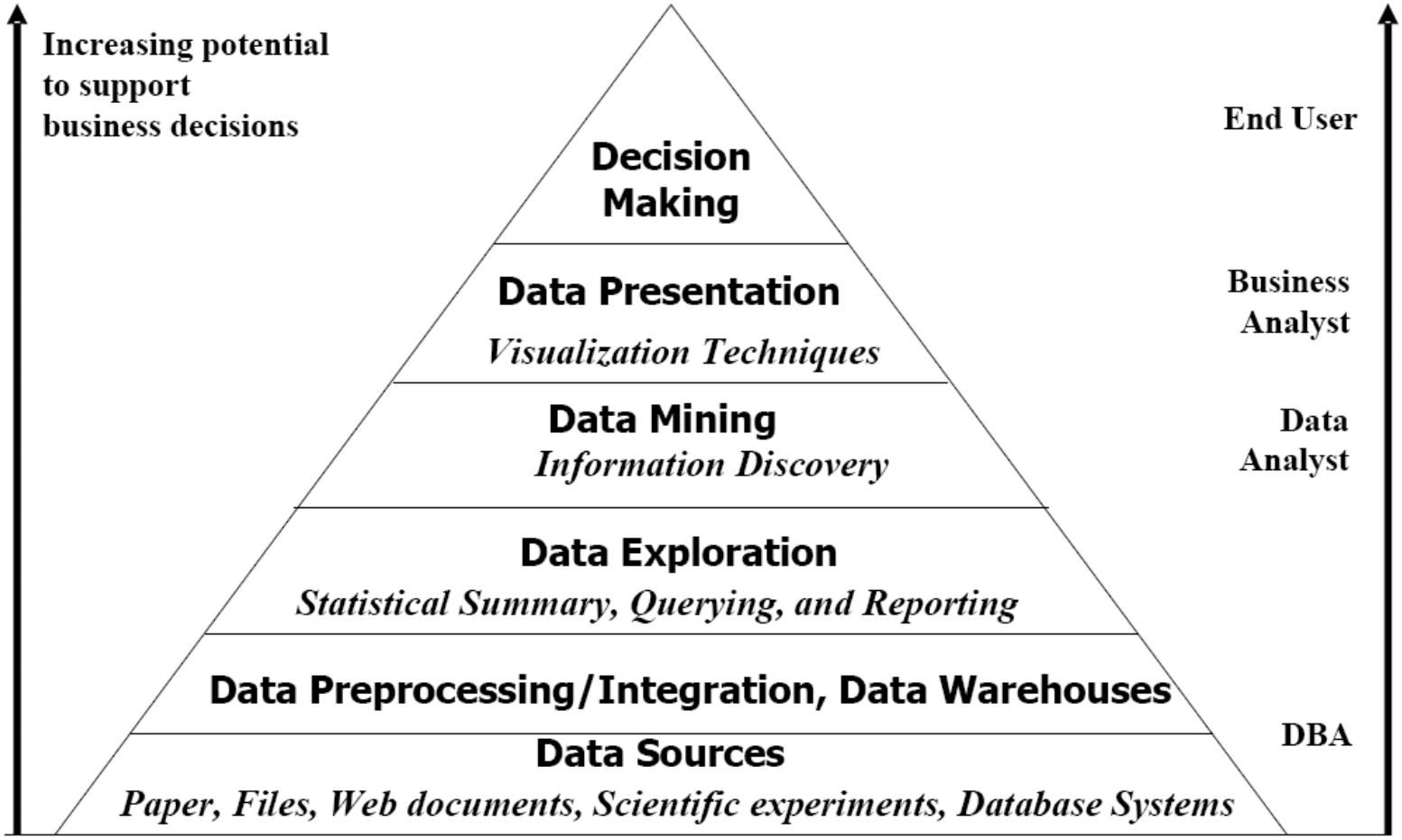
- **Tremendous amount of data**
  - Algorithms must be highly scalable to handle massive data, such as tera-bytes of data
- **High-dimensionality of data**
  - E.g., micro-array data may have tens of thousands of dimensions
- **High complexity of data**
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- **New and sophisticated applications**

Data Mining & Other Domains

- Draws ideas from statistics, machine learning, database systems.
- A key component of the emerging field of data science and data- driven discovery



Data Mining & Business Intelligence

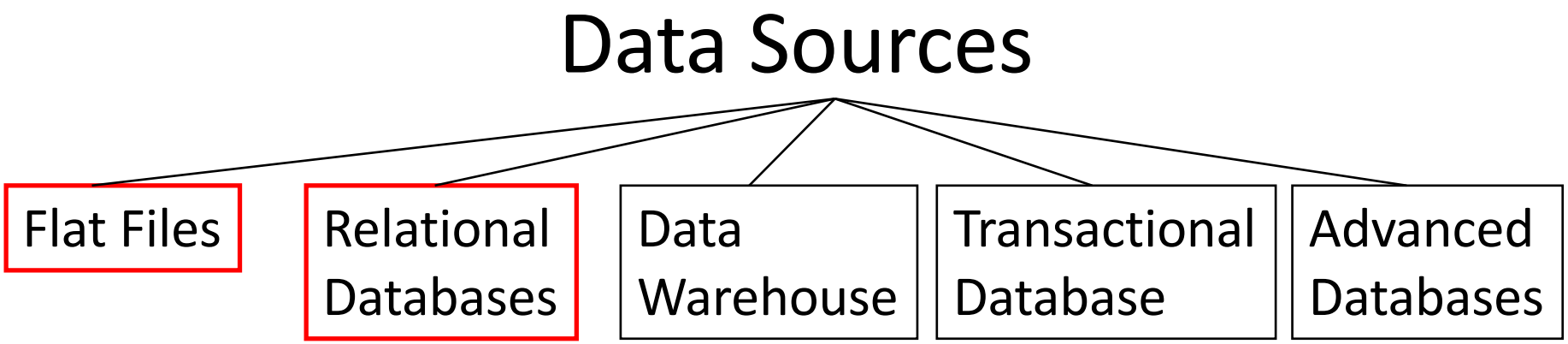


Source: https://www.unitrackind.com/mis-587.html

Multi-Dimensional View of Data Mining

- **Data** to be mined
  - Relational, data warehouse, transactional, stream, object oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW
- **Knowledge** to be mined
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques** utilized
  - Database-oriented, data warehouse, machine learning, statistics, visualization, etc.
- **Applications** adapted
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

What Kind of Data to Be Mined?



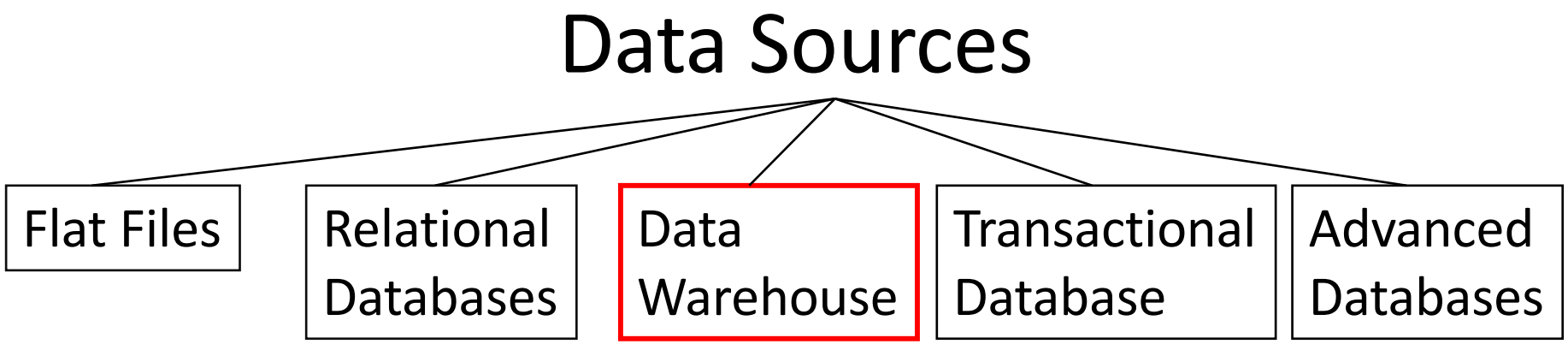
Flat Files

- Data is stored in **textual format**
- Data mining algorithms are applied to extract **useful knowledge**

Relational Databases

- Data is in **a collection of tables**, each table consists of tuples and attributes
- Database systems facilitate queries, but data mining aims to achieve **deeper** knowledge
- Usually data mining algorithms are applied to mine **useful knowledge** from a number of tables; several tables are usually joined together in such tasks

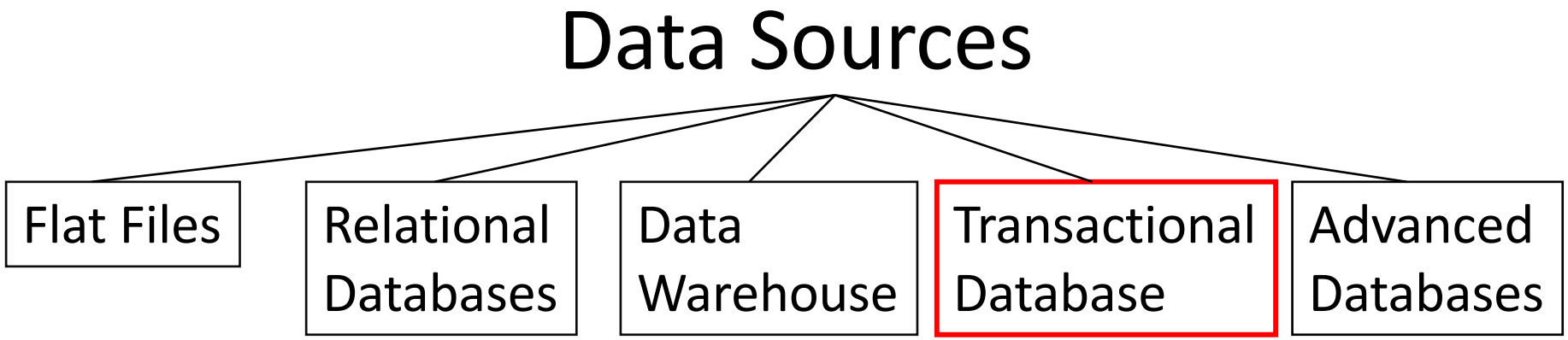
What Kind of Data to Be Mined?



Data Warehouse

- A repository of information collected from **multiple sources**
- E.g., information is extracted from historical information stored in transactional databases
- Built to support On-Line Analysis Processing (OLAP), but data mining is even deeper than OLAP

What Kind of Data to Be Mined?

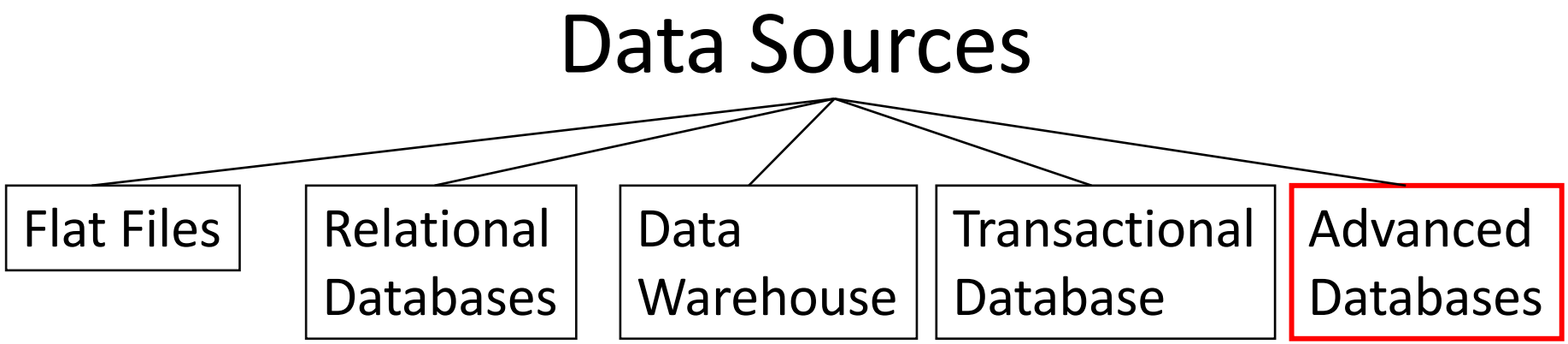


Transactional Databases

- Consists of records where **each record represents a transaction**
- A transaction typically includes a unique **transaction identity** number (trans_ID) and a list of items
- Example: Items purchased in a store
  - "Market basket analysis": one goal is to find which items are purchased together frequently

Trans_ID	List of item_Ids
T100	milk, beer, diaper
T200	milk, diaper

What Kind of Data to Be Mined?



Advanced Databases

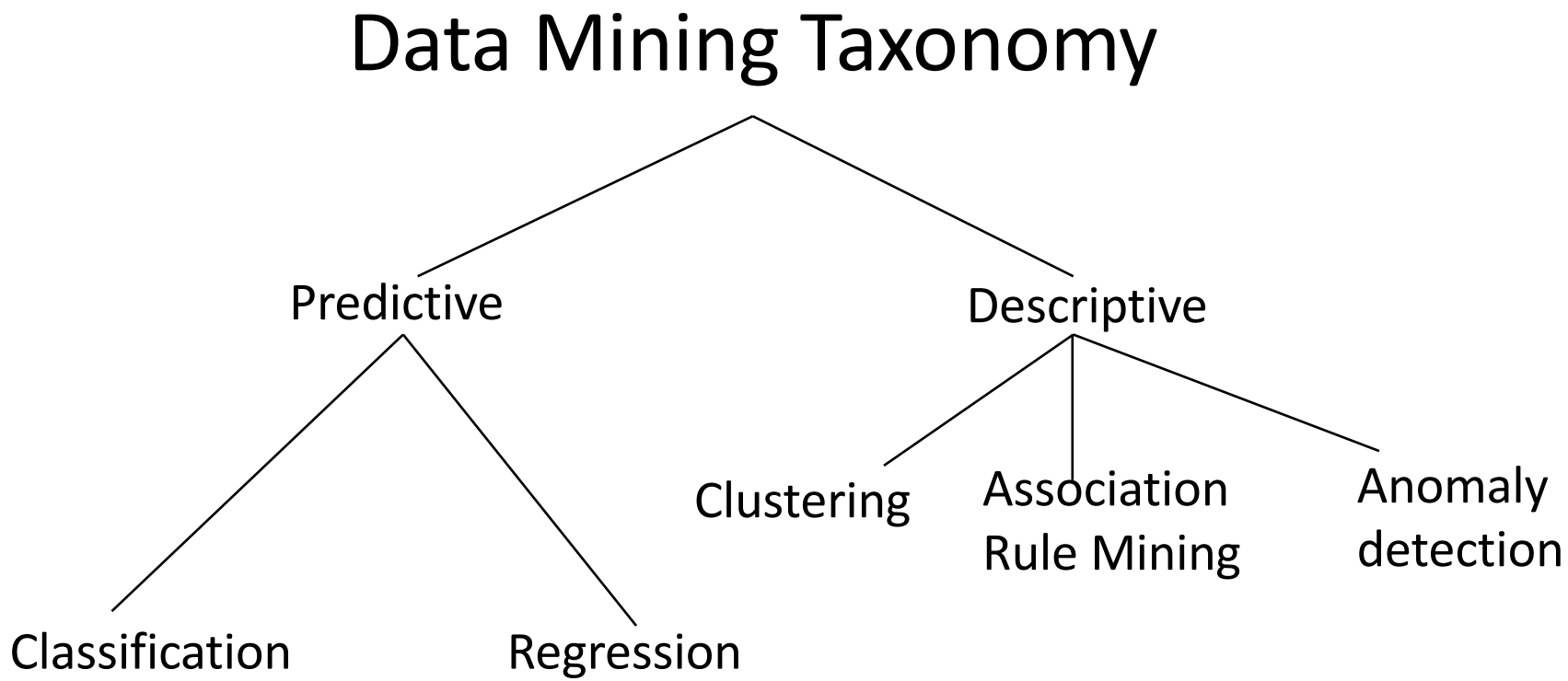
- Require **sophisticated data structures** and scalable methods
- Provide fertile ground for applying data mining
  a) Object-Relational Databases
  b) Temporal/Time-Series Databases
  c) Spatial/Spatio-Temporal Databases
  d) Text Databases
  e) Multimedia Databases
  f) Heterogeneous/Legacy Databases
  g) Data Streams
  h) WWW

Data Mining Tasks

- **Predictive** Tasks
  - Use some variables to **predict** **unknown or future values** of other variables.

- **Descriptive** Tasks
  - Find human-interpretable patterns that **describe** the data.

Data Mining Tasks



This taxonomy is based on the kinds of patterns output by data mining tasks.

Predictive Tasks: Classification/Regression

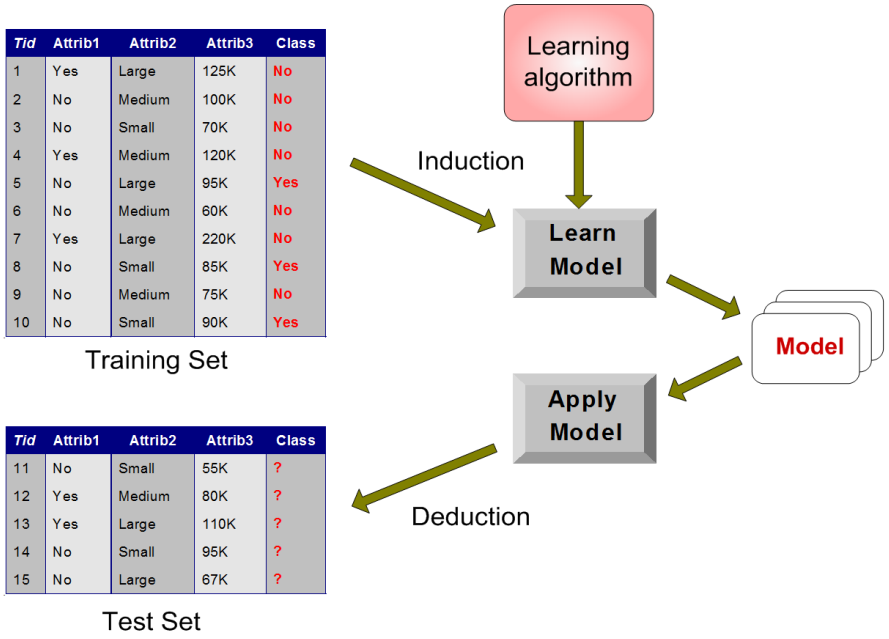
- Classification
  - Aim at learning to distinguish data instances between different classes, based on the training data with **class labels**
  - The models are generally represented as **classification rules**, **decision trees**, **neural networks**, etc.
  - The models are used to **predict** the **discrete class labels** of new (test) data without class labels comes,

- Regression
  - Aim to **map** a data item to a **continuous prediction variable**
  - Involves the learning of the function that does the mapping

Classification learns a model to predict **categorical labels,** whereas Regression learns a model to predict **continuous target.**

Predictive Task: Classification

- Given a collection of records (***training set***), each record contains a set of ***attributes***, one of the attributes is the ***class***.
- Learn a ***model*** for class attribute as a ***function*** of the values of other attributes.
- **Goal**: To ensure that **previously unseen (test data)** records should be assigned with **a class label** as **accurately** as possible.



Classification: Application

Targeted Marketing

- Goal:
  - Reduce marketing mailing cost by **targeting** a set of consumers likely to buy, e.g., a new cell-phone product.
- Approach:
  - Use the data for a similar product introduced before. We know which customers decided to buy and which decided otherwise. This {**buy**, **don't buy**} decision forms the **class** attribute.
  - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
    - Type of business, where they stay, how much they earn, etc.
  - Use this information as input attributes to learn a classification model for forecasting.

Predictive Task: Regression

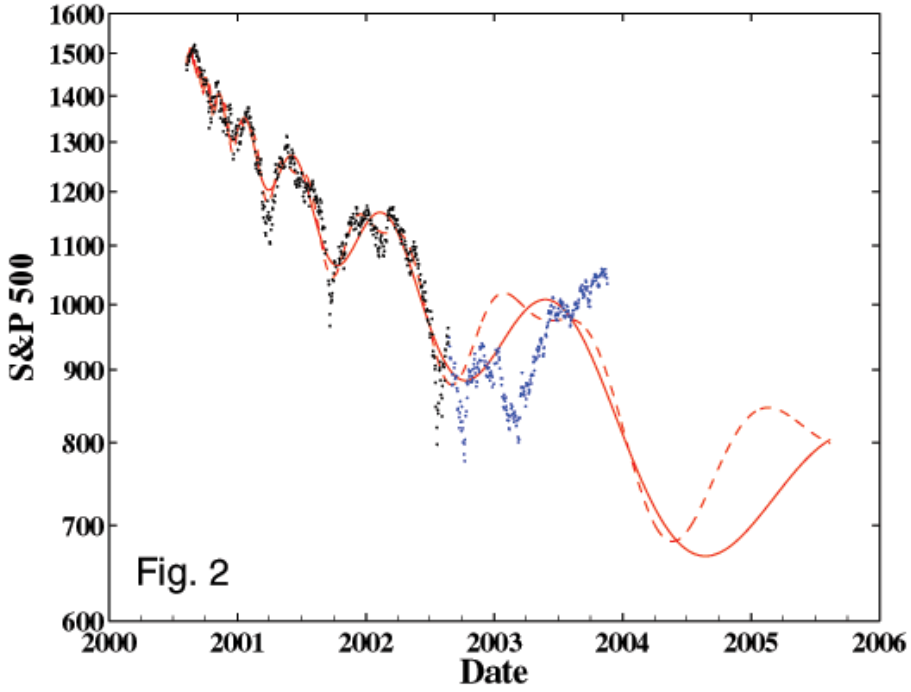
- **Goal:** Predict a value of a given **continuous valued variable** based on the values of other variables, assuming a linear or nonlinear model of **dependency**.

- Extensively studied in statistics, neural network fields.

- In statistics, regression analysis is a statistical process for estimating the relationships among variables, where the focus is on the relationship between a dependent variable and one or more independent variables.

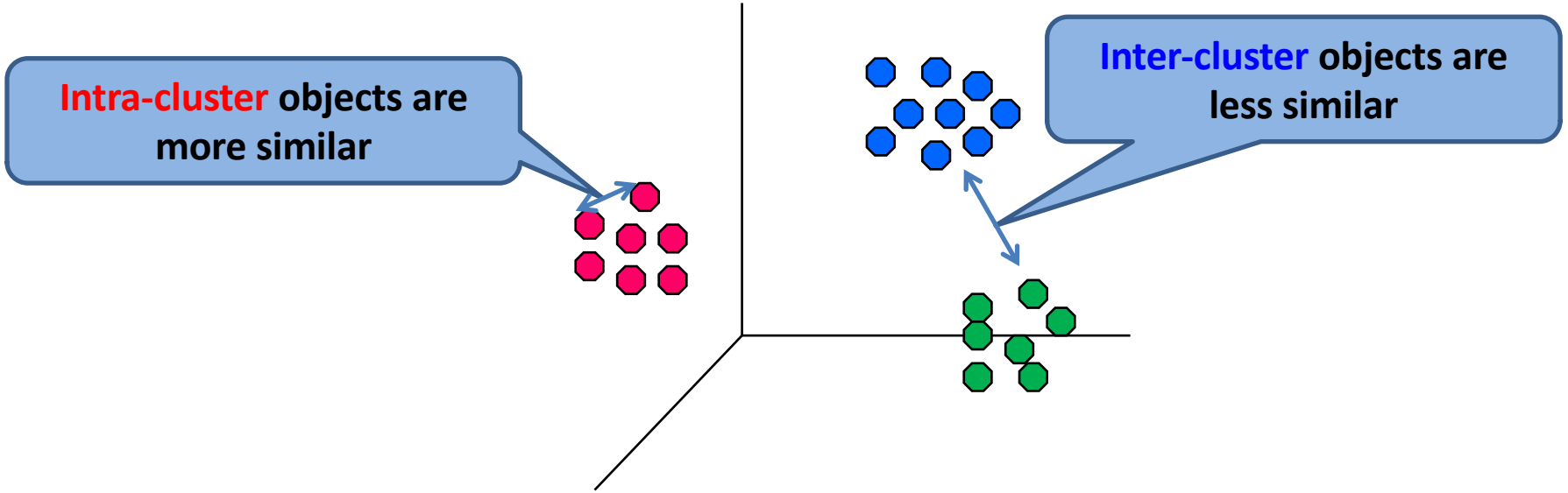
Regression - Example

- Stock Market Prediction
  - Black dots: training data
  - Red Line (continuous and dashed): Predictions
  - Blue dots: test (unseen) actual data
  - http://www.gold-eagle.com/editorials_03/sornette112403.html



Descriptive Task: Clustering

- Given a set of data points, each having a set of attributes, and some proximity measure among them, the goal of clustering is to find **clusters** such that
  - Data points in the same cluster are more similar to one another.
  - Data points in different clusters are less similar to one another



- Unlike classification and regression which analyze labeled data objects, clustering analyzes data without class labels

Clustering: Application 1

- **Market Segmentation**:
  - **Goal:**
    - To subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - **Approach:**
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- **Clustering of S&P 500 Stock Data**
  - Observe Stock Movements every day.
  - Clustering points: Stock-{**UP/DOWN**}
  - Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Descriptive Task: Association Rule Mining

- Given a set of **records** each of which contains 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

...

market transactional DB

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

{Pepsi, … } --> {Potato Chips}

Association Rule Mining: Application 1

Marketing and Sales Promotion

- Let assume the rule discovered is:

  {Pepsi, … } --> {Potato Chips}

- Potato Chips as consequent
  - Can be used to determine what should be done to boost its sales.
- Pepsi in the antecedent
  - Can be used to see which products would be affected if the store discontinues selling Pepsi.

- Implication of such rules
  - Can be used to see what products should be sold with Pepsi to promote the sale of Potato chips!

Association Rule Mining: Application 2

Supermarket shelf management.

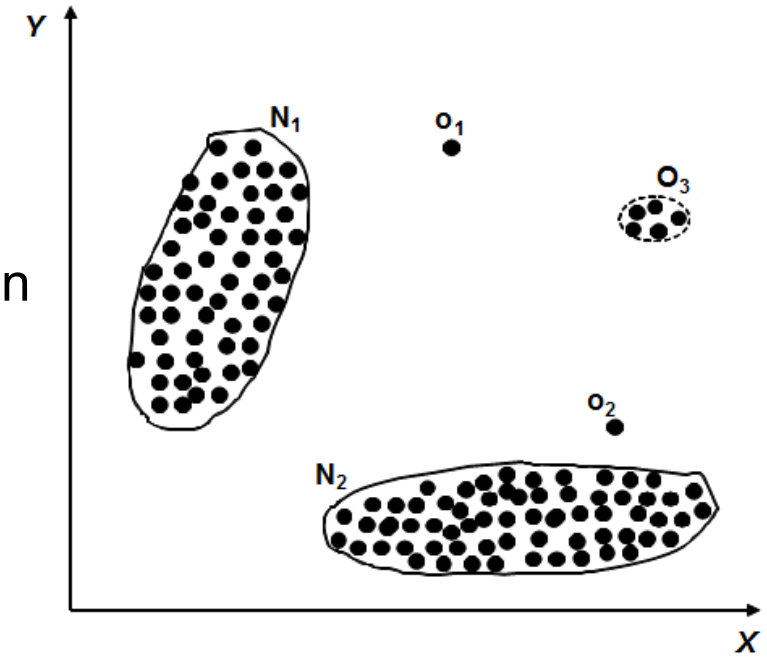
- Goal:
  - To identify items that are bought together by sufficiently many customers.
- Approach:
  - Process the point-of-sale data collected with barcode scanners to find dependencies among items.
- Here is a classical rule:

{diaper, milk} --> {beer}

  - If a customer buys diaper and milk, then he is very likely to buy beer.
  - So, don't be surprised if you find six-packs stacked next to diapers!

Descriptive Task: Outlier/Anomaly Detection

- **Outlier**: some data point does not comply with the general behavior of the data
- **Goal**: To detect significant deviations (outliers) from the normal behavior
- Although in many applications outliers are unnecessary, in some applications they are very useful - Fraud detection in credit card purchase, authentication (password), network intrusion detection

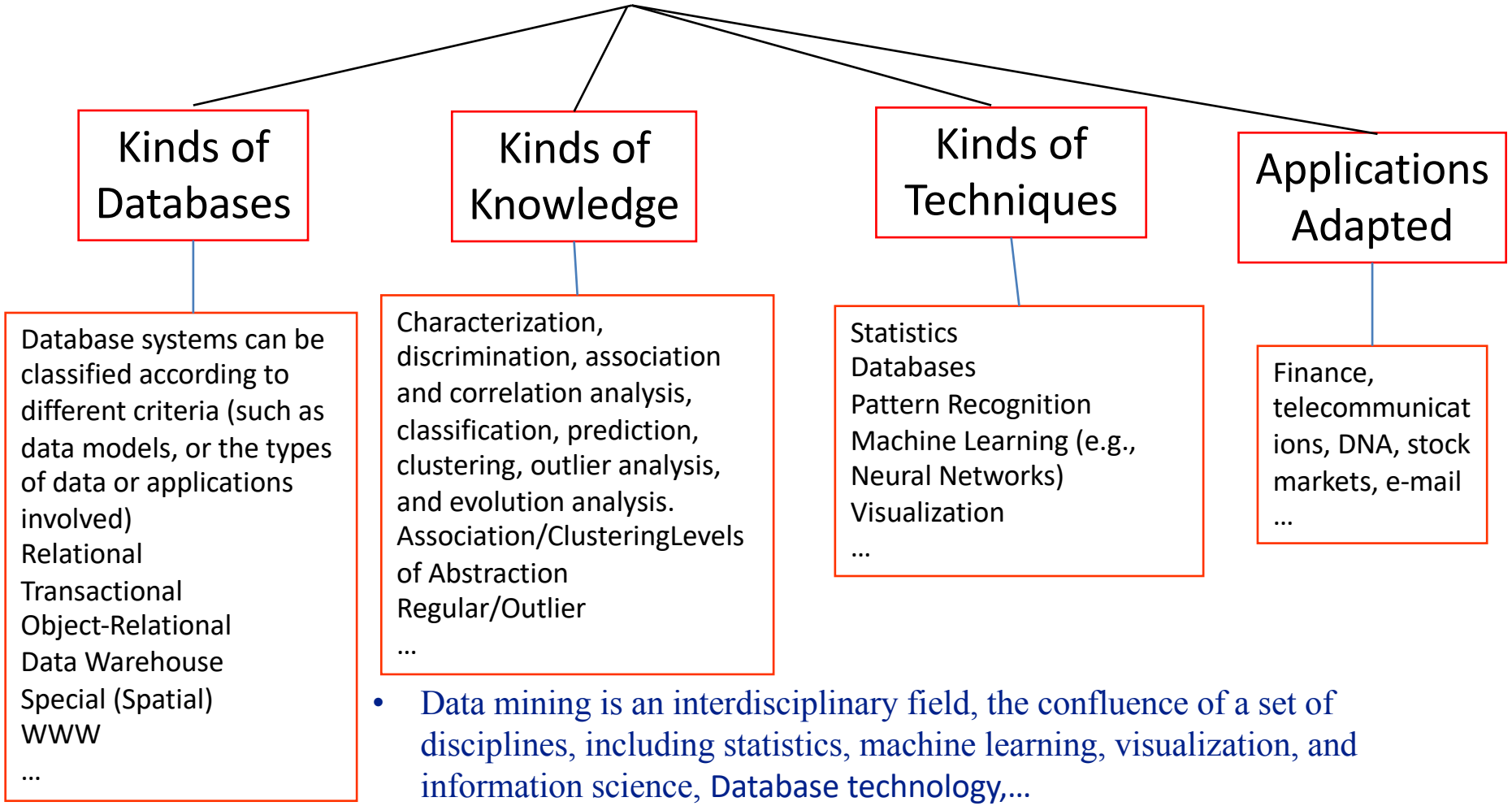


Anomaly Detection: Application

Fraud Detection

- Goal:
  - Predict fraudulent cases in credit card transactions.
- Approach
  - Use credit card transactions and the information on its account-holder as attributes.
    - When does a customer buy, what does he buy, how often he pays on time, etc.
  - Label past transactions as **fraud** or **fair** transactions. This forms the **class** attribute.
  - Learn a **model** for the class of the transactions.
  - Use this **model** to detect fraud by observing credit card transactions on an account.

Classification of Data Mining Systems



- Data mining is an interdisciplinary field, the confluence of a set of disciplines, including statistics, machine learning, visualization, and information science, Database technology,…
- Data mining systems can be categorized according to various criteria, as shown in above.

School of Computing and Information Systems

SMU SINGAPORE MANAGEMENT UNIVERSITY

Summary

- Definitions of Data Mining (What)
  - Data mining: implicit, previously unknown, and potentially useful
  - Data Mining Requirements: lots of data, and lots of computing power
- Motivation of Data mining (why)
  - We are drowning in data, starving for knowledge
  - Discovering interesting patterns from large data
  - From Commercial Viewpoint and From Scientific Viewpoint
- Background & Application of Data Mining
  - A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
  - Why Not Use Classical Data Analysis?
  - Draws ideas from statistics, machine learning, database systems
- What Kind of Data to Be Mined?
  - Flat Files, Relational Databases, Data Warehouse, Transactional Database, Advanced Databases
- Data Mining Tasks and Data Mining Taxonomy
  - Predictive: Classification, Regression
  - Descriptive: Clustering, Association Rule Mining, Outlier Detection
- Classification of Data Mining Systems
  - Kinds of Databases, Kinds of Knowledge, Kinds of Techniques, Applications Adapted

Conferences and Journals on Data Mining

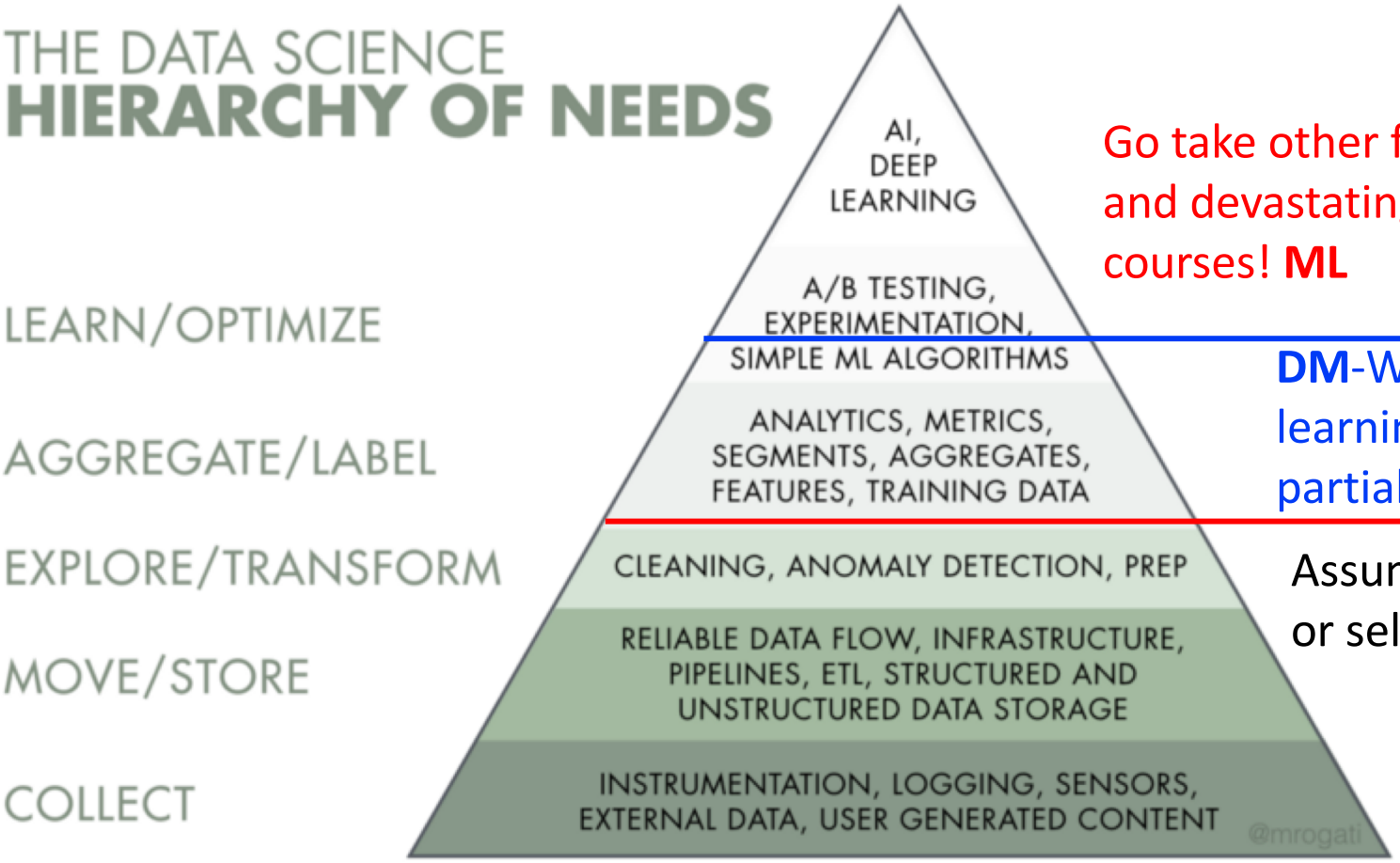
- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Other related conferences
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Lab Section
Introduction to Python

Appendix*

- More Background of Data Mining
- Example of “useful knowledge”
- More about advanced databases
- Primitives Defining a Data Mining Task
- One More Application for Clustering

What are we really teaching you?



Go take other fun and devastating courses! **ML**

DM-What you will be learning (at least partially)

Assume you know or self-learn

source: https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

Evolution: From DB to DM

- Data collection ⟹ Database creation
  - **1960s:** From primitive data collection systems to sophisticated and powerful database systems (Navigational DBMS)
- Data management ⟹ Advanced data analysis
  - **1970s:** From early hierarchical and network models to relational models online transaction processing (**OLTP**) , Structured Query Language (**SQL**) DBMS
  - **1980s:** Further research into object-oriented database systems, Internet, application-oriented, etc.
  - **1990s:** Cheaper hardware, advanced DBMS, Data Warehouse, On-Line analytical processing (**OLAP**)
  - **Late 1990s/present**: Data rich but information poor situation required powerful analytical tools, which motivates Data Mining technology

Data Mining vs. Artificial Intelligence (AI)

SMU Singapore Management University | School of Computing and Information Systems

AI

- Computer Vision
- Natural Language Processing
- Knowledge Reasoning, Planning, Causality
- Big Data and Machine Learning
- Robotics

...

↔ Data Mining

https://en.wikipedia.org/wiki/Artificial_intelligence

Supervised vs Unsupervised Learning

Unsupervised Learning

Data: x
x is data, no labels!

Goal: Learn some *hidden or underlying structure* of the data (learn the true explanatory factors, e.g. latent variables, from only the observed data)

Examples: Clustering, feature or dimensionality reduction, etc.

Supervised Learning

Data pairs: (x, y)
x is data item, y is label

Goal: Learn function to map
$x \rightarrow y$

Examples: Classification, regression, object detection, semantic segmentation, etc.

The Concepts and Algorithms Covered by DM* and ML** Courses

- Skills/ DM techniques:
  - Data Exploration & data pre-processing
  - Learning based DM/ML methods
  - Classification
    - Decision Tree
    - k-Nearest Neighbors (k-NN)
    - Ensemble methods, such as bagging, boosting
  - Clustering
    - K-means clustering
    - Hierarchical clustering
  - Association Analysis

- Skills/ ML algorithms:
  - Naïve Bayes
  - Perceptron
  - Support Vector Machine
  - Linear Regression
  - Logistic Regression
  - Neural Networks
    - Feedforward NN
    - Backpropagation Algorithm
  - Deep learning, such as RNN, LSTM
  - Unsupervised learning, such as PCA and SVD

** courses are more difficult than * courses

Advanced Databases

- **Advanced Databases**
  - Examples: ORDBMS/spatial/hypertext/multimedia/time-series/streaming/WWW
  - Require sophisticated data structures and scalable methods
  - Provide fertile ground for applying data mining

Advanced Databases

(a) Object-Relational Databases

- Simple Relational databases cannot handle sophisticated data, such as picture, sound, hypertext, etc
- In an object-relational databases
  - Each tuple is an object
  - Variables that describe an object correspond to attributes in relational DB
  - Each object has several messages, and methods that can process the messages
  - Objects that share a common set of properties is called an object class
  - Inheritance property exists between class and its subclasses
- For data mining, techniques need to be developed for handling complex object structures

Advanced Databases

(b) Temporal/Time-Series Databases

- Typically stores data that include time-related attributes
- Examples: Stock exchange, inventory control, temperature
- Data mining can find object evolution, or trend of changes of objects

(c) Spatial/Spatio-Temporal Databases

- Examples:
  - geographic (maps), very large-scale integration (VLSI), computer-aided design, medical and satellite image
- Spatial data can be represented in
  - raster format: *n*-dimensional pixel maps
  - vector format: unions or overlays of basic geometric constructs
- Applications:
  - forestry, ecology planning to provide public service info, sewage system, vehicle navigation and dispatching systems

Advanced Databases

(d) Text Databases

- Contain word description of objects
- Examples
  - product specification, error or bug reports, warning messages, summary reports, notes, newspaper articles
- Structured/Semistructrued/Unstructured
  - Structured – can be represented using relational databases
  - Semistructured – e.g., XML/HTML web page, e-mail messages
  - Unstructured – e.g., other documents
- Data mining can find
  - association between keywords, clustering of documents
  - To do this, standard data mining methods need to be integrated with information retrieval techniques, dictionaries, ontologies, etc.

Advanced Databases

(e) Multimedia Databases

- Examples
  - image, audio, video data used in applications such as picture content-based retrieval, voice-mail systems, video-on-demand, WWW, Speech recognition
- Special requirement
  - Very large size, efficient search techniques
- Data mining should be integrated with storage/retrieval system
  - to construct multimedia data cubes, pattern matching, extract multiple features

Advanced Databases

(f) Heterogeneous/Legacy Databases

- Heterogeneous database consists of a set of interconnected autonomous component databases
- A legacy database is a group of heterogeneous databases that combines different kinds of data systems
- Example:
  - Exchanging information regarding student academic performance among different schools
- Data mining can
  - first transform the data into higher, more generalized, conceptual levels, and then extract the knowledge

Advanced Databases

(g) Data Streams

- Properties of Data Streams
  - potentially infinite in size, dynamically changing, allowing only one scan, demanding fast response time
- Examples
  - telephone, network, traffic, stock exchange, Web click streams, video surveillance
- Special requirement
  - Efficient management and analysis of data streams
- Data mining can
  - detect changes in the general pattern over the stream and update the data mining model

Advanced Databases

(h) WWW

- Provides rich, worldwide, online information services
- Can be very unstructured to understand the semantics
- Data mining can help to
  - understand user access patterns
  - search web pages based on semantics
  - cluster and classify the web pages
  - identify hidden Web social networks (Web community analysis)

Features Defining a Data Mining Task

- **Task-relevant data**
  - What is the data set that I want to mine?
- **Type of knowledge to be mined**
  - What kind of knowledge do I want to mine?
- **Background knowledge**
  - What background knowledge can be useful here?
- **Pattern interestingness measurements**
  - What measures can be used to estimate pattern interestingness?
- **Visualization/presentation of discovered patterns**
  - How do I want the discovered patterns to be presented?

Clustering: Application 3

- **Document Clustering:**
  - **Goal:**
    - To find groups of documents that are similar to each other based on the important terms appearing in them.
  - **Approach:**
    - To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - **Gain/Consequence:**
    - Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.