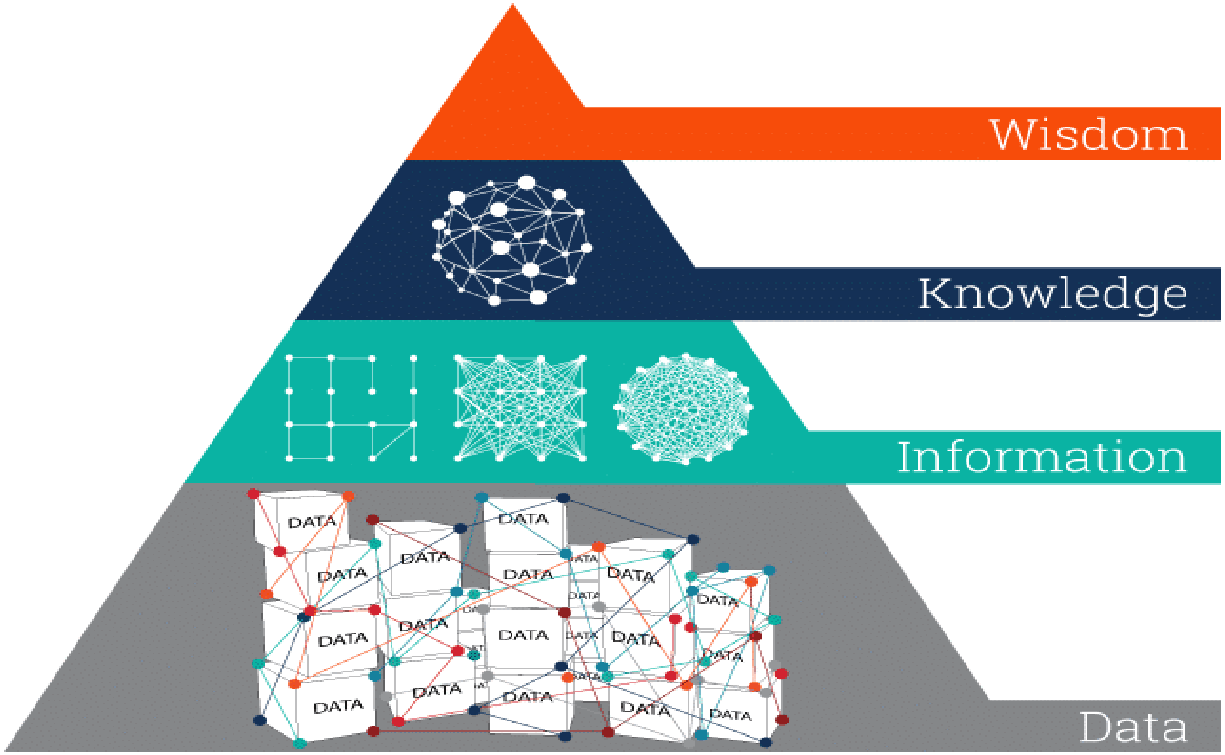
Chapter one
Introduction to Data-Analytic Thinking

Objectives

- **By the end of this chapter, you will be able:**
  - Explain the relationship between data, information, knowledge and wisdom
  - Explain and distinguish the various data types
  - Explain data mining
  - Describe the difference between database, data mining , and data warehouse

Data, Information, Knowledge, Wisdom

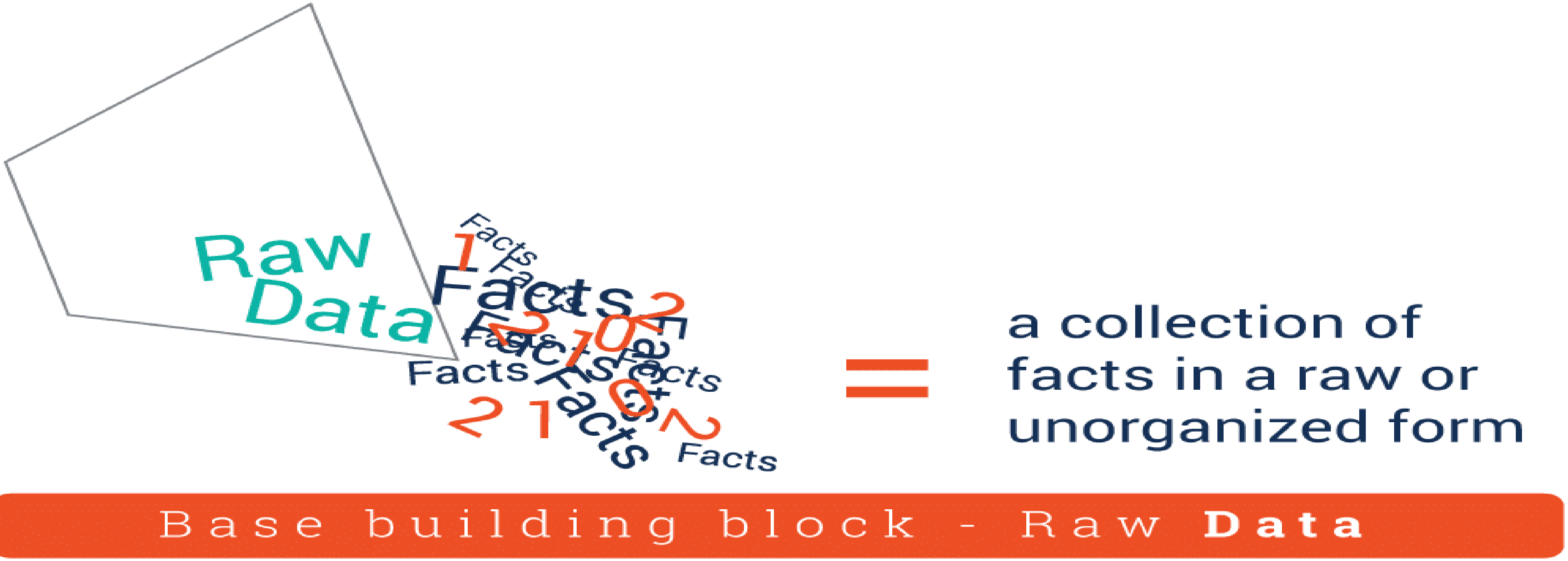
- What is Data and Information? Are they different from Knowledge? Wisdom?
  - Data!=information!=knowledge!=wisdom.
- **Known as Knowledge pyramid, wisdom hierarchy, and information hierarch**



Each step up the pyramid answers questions about and adds value to the initial data.

Data, Information, Knowledge, Wisdom

- What is Data and Information? Are they different from Knowledge? Wisdom?
  - data != information != knowledge!=wisdom
- **Data:** Unorganized and unprocessed facts; static; a set of discrete facts about events
  - No meaning attached to it as a result of which it may have multiple meaning
  - Example: what does “green123” mean?

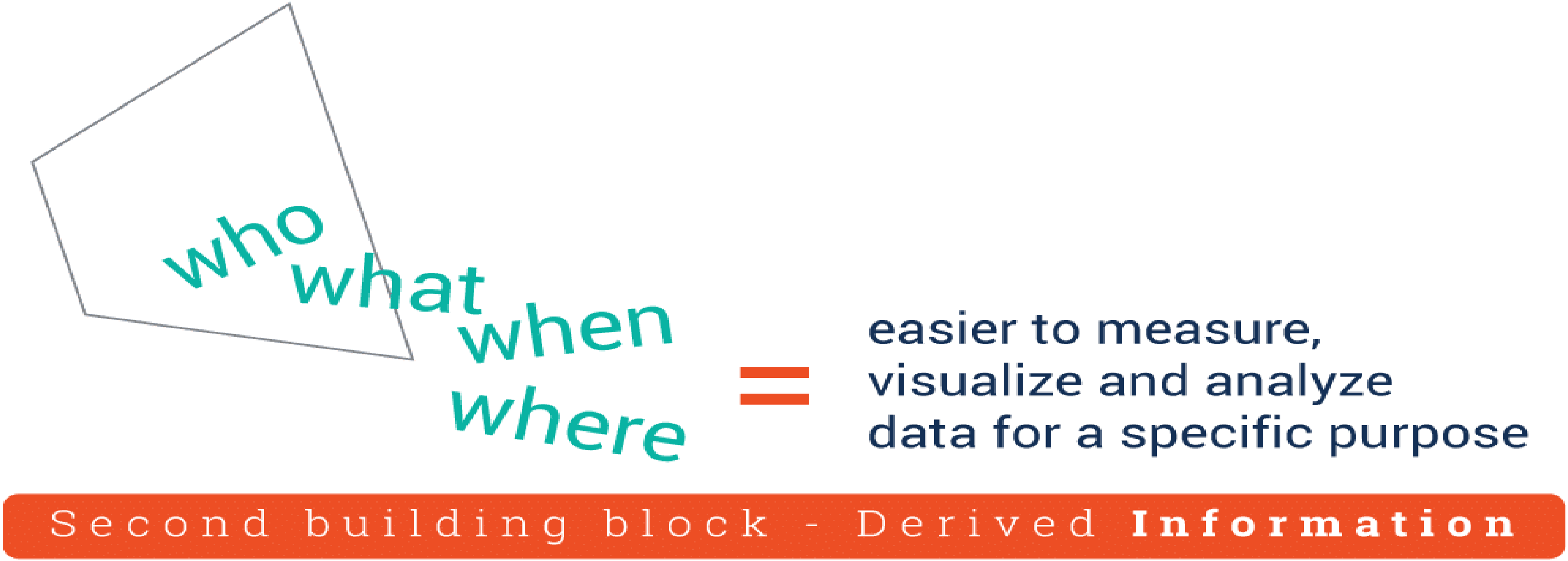


a collection of facts in a raw or unorganized form

Base building block - Raw **Data**

Data, Information, Knowledge, Wisdom

- What is Data and Information? Are they different from Knowledge? Wisdom?
  - data != information != knowledge!=wisdom
- **Information:** Aggregation of data that makes decision making easier.
  - Meaning is attached and contextualized
  - Answers questions: what, who, when, where

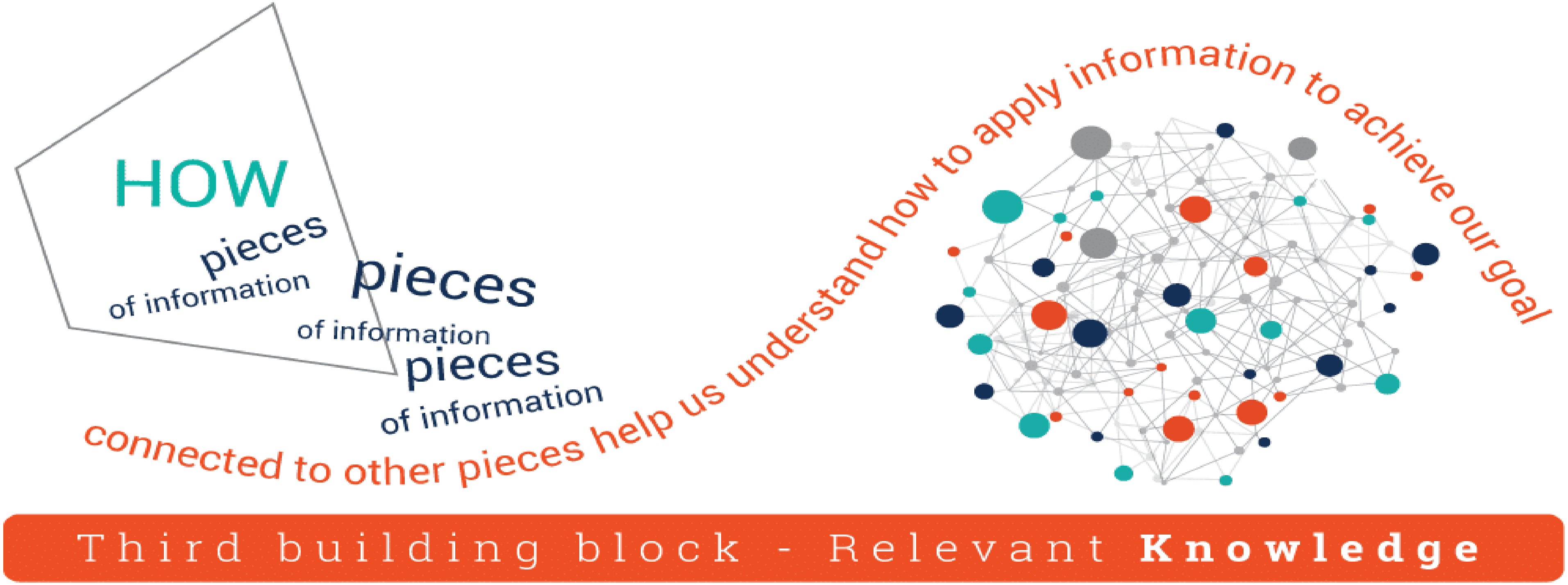


easier to measure, visualize and analyze data for a specific purpose

Second building block - Derived **Information**

Data, Information, Knowledge, Wisdom

- What is Data and Information? Are they different from Knowledge? Wisdom?
  - data != information != knowledge!=wisdom
- **Knowledge:** includes facts about the real world entities and the relationship between them. It is an Understanding gained through experience
  - Answer ‘how’ question


HOW
pieces
of information
pieces
of information
pieces
of information
connected to other pieces help us understand how to apply information to achieve our goal
Third building block - Relevant **Knowledge**

Data, Information, Knowledge, Wisdom

- What is Data and Information? Are they different from Knowledge? Wisdom?
  - data != information != knowledge!=wisdom
- **Wisdom:** Knowledge applied in action
  - Answer ‘why’ question

why do something?
what is best?
Wisdom is knowledge applied in action
The top of the DIKW hierarchy - Guiding **Wisdom**

Data, Information, Knowledge, Wisdom

- What is Data and Information? Are they different from Knowledge? Wisdom?
  - data != information != knowledge!=wisdom
- **Wisdom:** Knowledge applied in action
  - Answer 'why' question

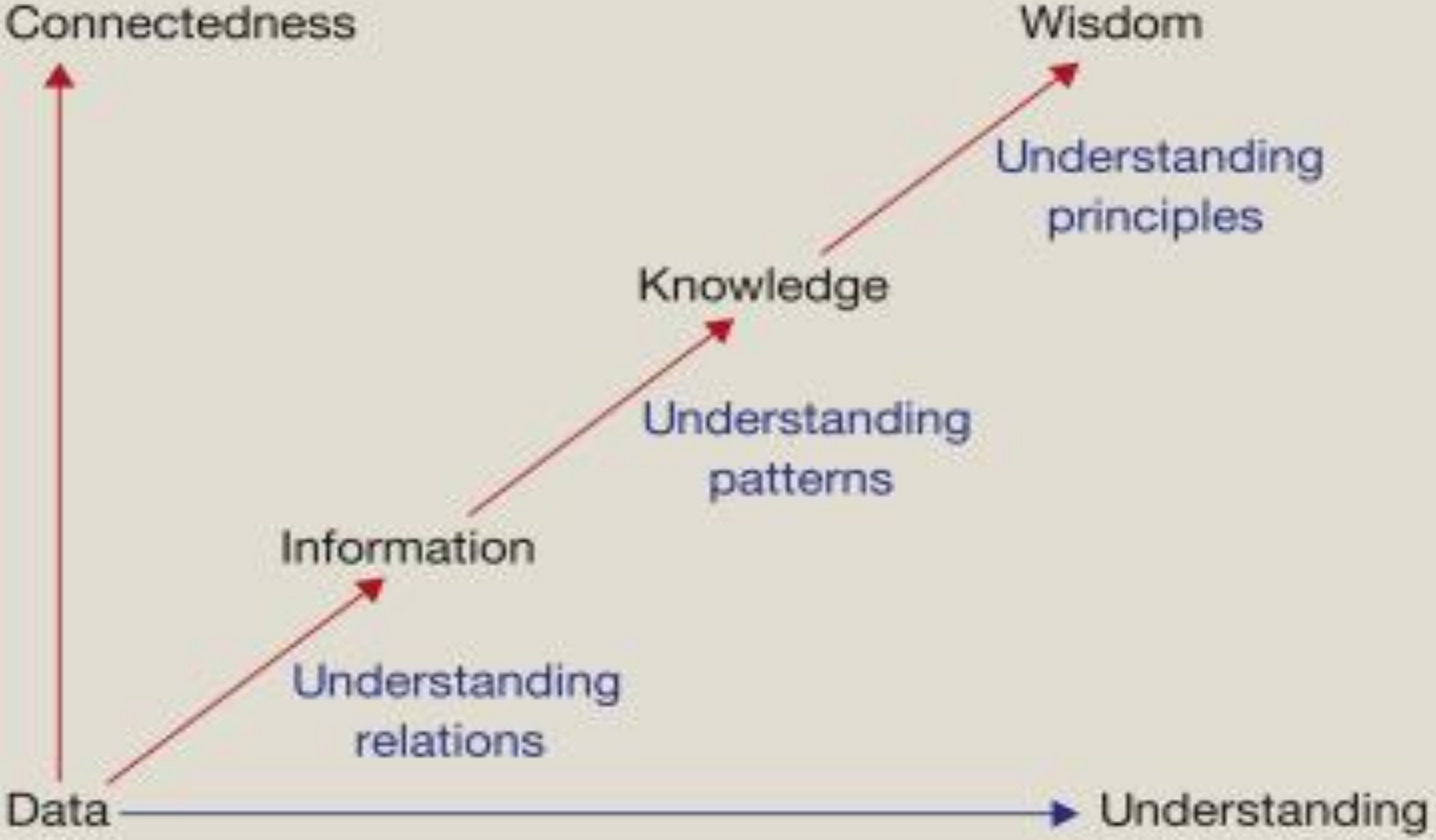
why do something?
what is best?
Wisdom is knowledge applied in action
The top of the DIKW hierarchy - Guiding **Wisdom**

Data, Information, Knowledge, Wisdom

- What is Data and Information? Are they different from Knowledge? Wisdom?
  - data != information != knowledge!=wisdom
- **Wisdom:** Knowledge applied in action
  - Answer ‘why’ question

why do something?
what is best?
Wisdom is knowledge applied in action
The top of the DIKW hierarchy - Guiding **Wisdom**

The transition from data, to information, to knowledge, to wisdom

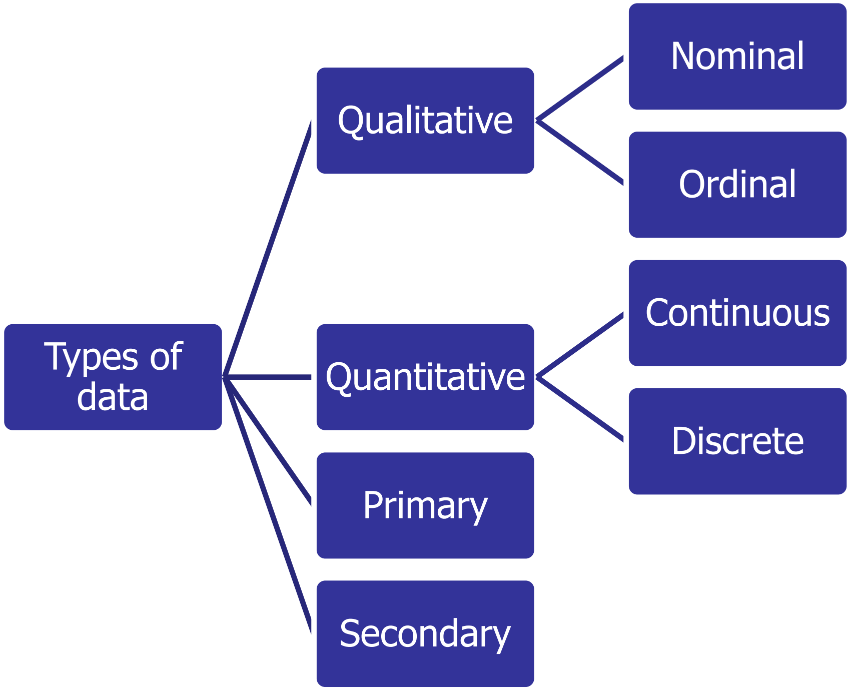


Discussion

Create a group that consists of 2 -3 students and identify a scenario for discussing and elaborating the difference between data, information, knowledge, and wisdom. Upon completion elect one student for presenting the scenario.

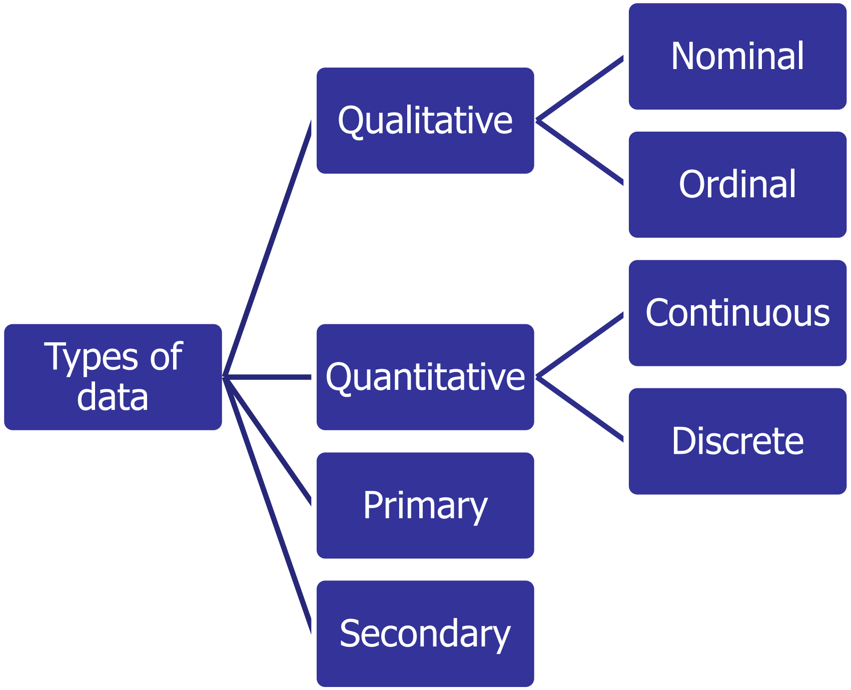
Classification of Data

- Types of data
  - Data can be classified based on two criterion (I) Nature of data (II) Source of data



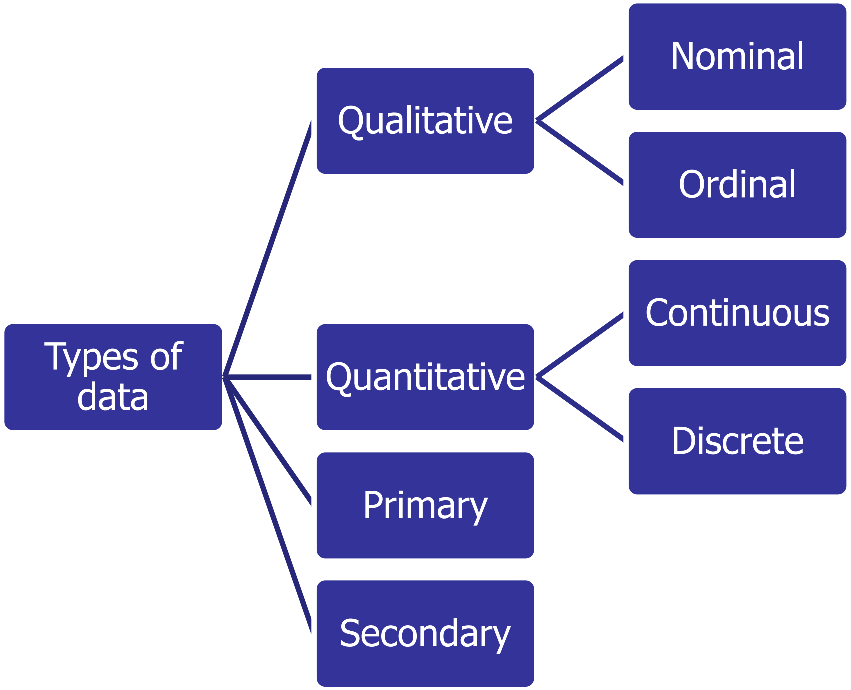
Classification of Data

- **Secondary Data**:
  - Data gathered and recorded by someone else prior to and for a purpose of other than the current project.
  - Is data that has been collected for another purpose
  - It involves less cost, time and effort.
  - Data collected from a source that has already published in any form
  - Is data that is being reused. Usually in different context.
- Source of Secondary Data
  - Government Organizations
  - Research journals and research organizations
  - Newspapers and magazines
  - Teaching and research organizations
  - Internet



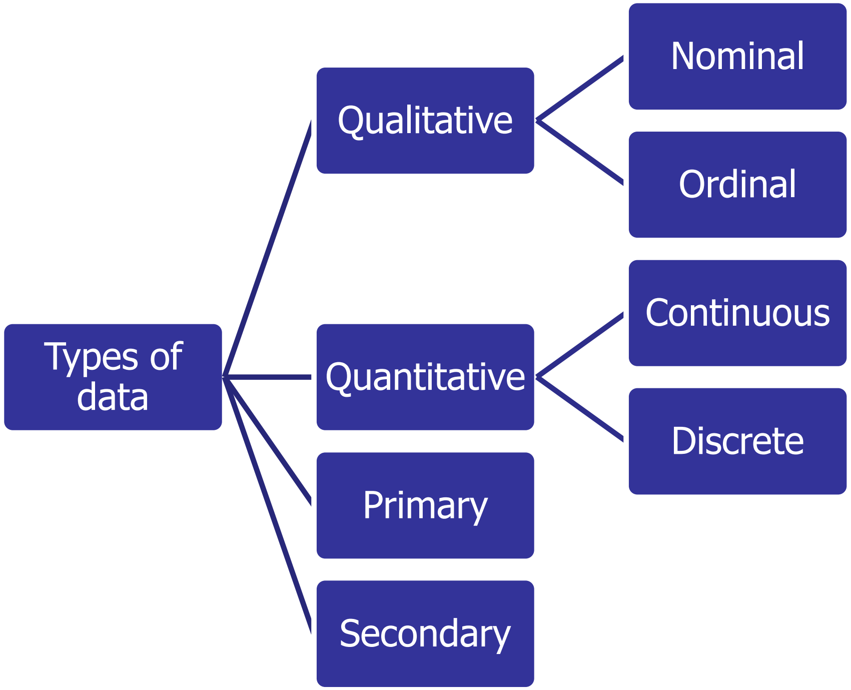
Classification of Data

- **Care to be taken before taking the secondary data**
  - The reliability of the data
  - The competency of the individual who collected the data
  - The suitability of the data for the particular study
  - The quantity of the data
  - The time of data collection
  - The original method adopted for the collection of data



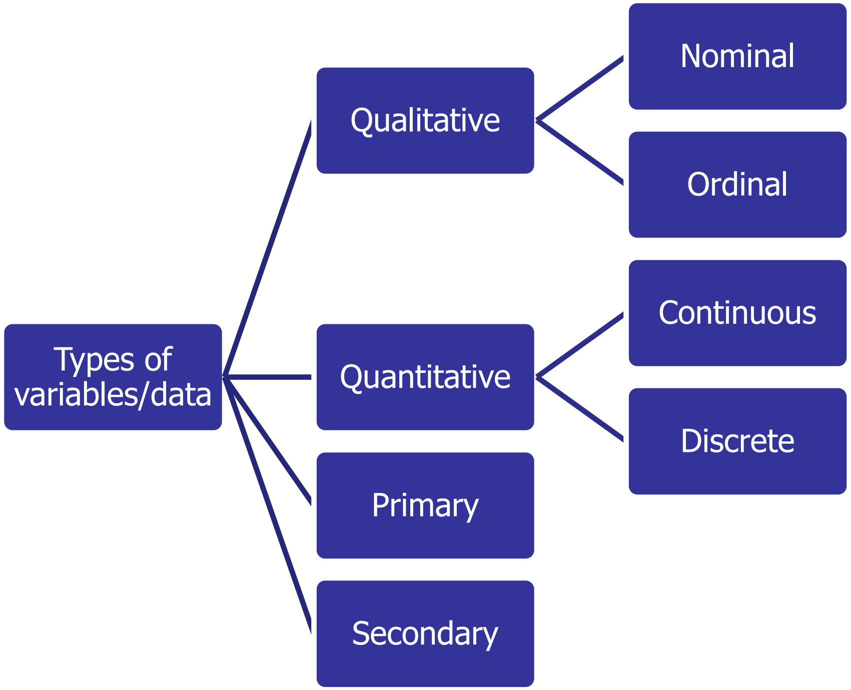
Classification of Data

- **Primary Data**:
  - Means original data that has been collected especially for the purpose in mind
  - The data which are collected from the field under the control and supervision of an investigator
  - This type of data are generally afresh and collected for the first time
  - It is useful for current studies as well as for future studies
  - Has not been published yet and is more reliable, changed or altered by human being , therefore its validity is greater that secondary data.



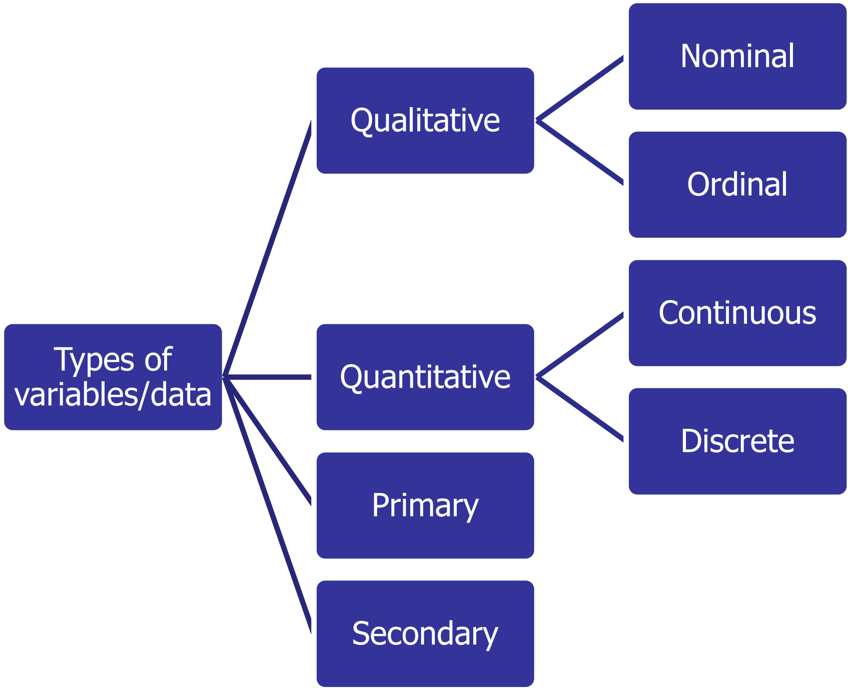
Classification of Data

- **Quantitative Variables :**
  - Are the measurable or countable variable
  - They are called better numerical data because they give the numerical data
  - Are always numbers
  - Example plant height, fruit weight , etc
  - Discrete variable
    - Also called as discontinuous data/variables
    - The value which the variable can assume are limited to whole numbers only (0,1,2,3..)
    - Example number of brothers,
    - There will be gaps between the successive values of the variable.
  - Continuous Variables
    - Are those variable that can take any value within a certain range
    - There are no gaps between the successive values of the variables



Classification of Data

- **Qualitative Variables/data :**
  - Are un-measurable variables
  - The are also called as non-numerical since they give qualitative data
  - Generally described by words or letters
  - Example: hair color, shape of leaves
  - Nominal variables
    - Have distinct levels that have no inherent ordering
    - Example gender (Male, Female)
  - Ordinal Variables
    - Have levels that follow distinct ordering

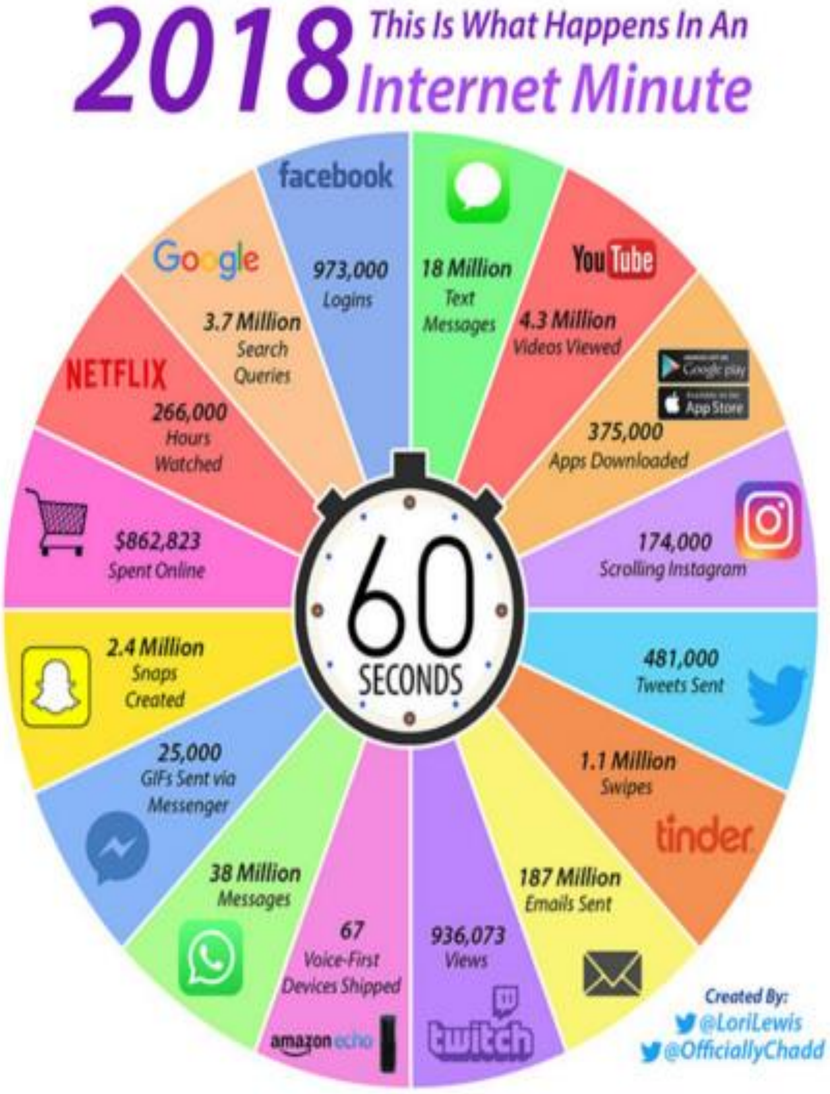


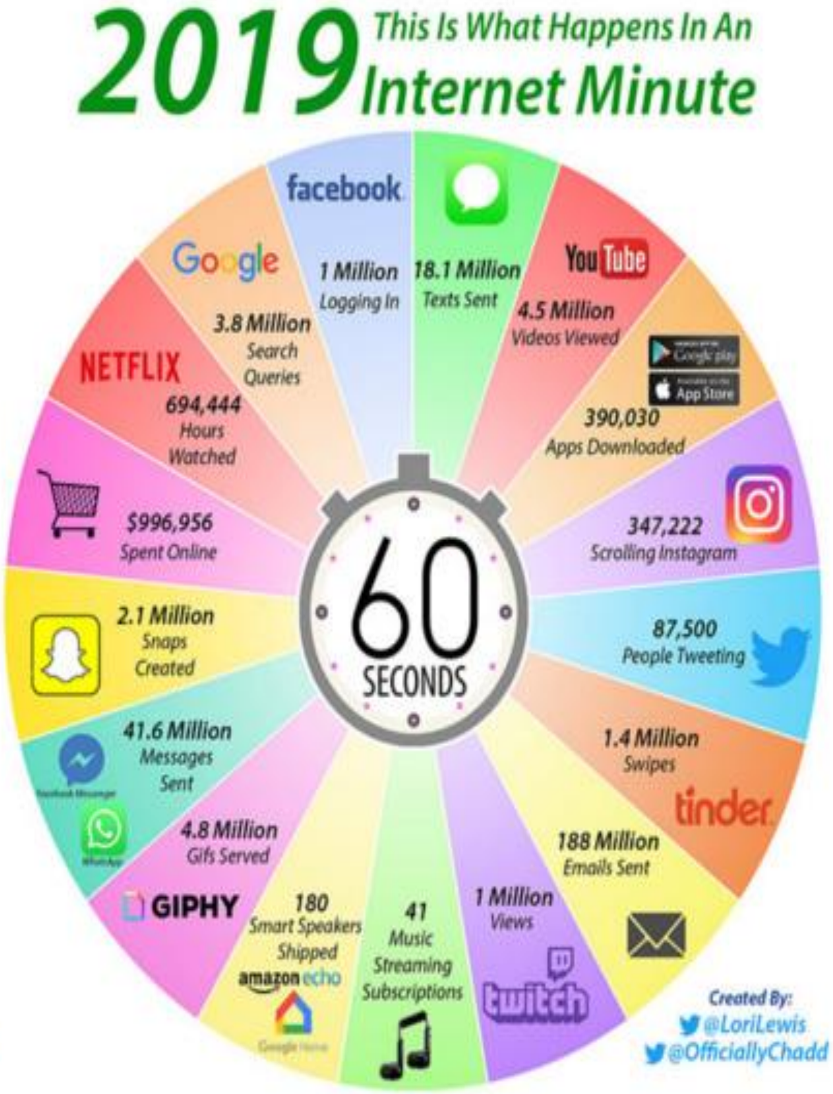
Exercise

- Identify the following variables as either quantitative or qualitative. Indicate whether the quantitative data are continuous or discrete and Indicate whether the qualitative data are nominal or ordinal.

  1. The scores of 40 students on a big data and management test.
  2. The blood types of 10 teachers in ITSC
  3. Number of times person checks their e-mail per day.
  4. Daily temperature (in degrees Fahrenheit) for last August.
  5. Weight (in grams) of tomatoes at a grocery store.
  6. Hair color of women on a high school tennis team.
  7. The final grades (A, B, C, D, and F) for students in this class.
  8. The ratings of a movie ranging from “poor” to “good” to “excellent”.
  9. The type of car you drive
  10. Number of correct answers on a quiz
  11. The distance it is from your home to here
  12. The number of courses you take per year.
  13. Peoples’ attitudes toward the government
  14. The speed of a car in miles per hour.
  15. A person's height.
  16. The color of your house.
  17. Outcome of tossing a coin.

Data/Information Overload





Data/Information Overload

- Data is being produced (generated & collected) at alarming rate because of:
  - The computerization of business & scientific transactions
  - Advances in data collection tools, ranging from scanned texts & image platforms to satellite remote sensing systems
  - Popular use of WWW as a global information system

- With the phenomenal rate of growth of data, users expect more sophisticated useful and valuable information
  - A marketing manager is no longer satisfied with a simple listing of marketing contacts, but wants detailed information about customers **past purchasing behavior and prediction of future purchases**

Too much data & too little knowledge

- There is a need to extract knowledge (useful information) from the massive data.
  - The competitive pressures are strong, which needs useful information for prediction
- Facing too enormous volumes of data, human analysts with no special tools can no longer make sense.
  - mining and management of big data can automate the process of finding patterns & relationships in raw data and the results can be utilized for decision support. That is why data mining is used, especially in science and business areas.
- If we know how to reveal valuable knowledge hidden in raw data, data might be one of our most valuable assets.
  - Data mining is the tool that involves retrospective analysis to extract diamonds of knowledge **from historical data & predict outcome of the future.**

What is data mining?

- Data Mining is a technology that uses **various techniques** to discover **hidden knowledge** from **heterogeneous** and **distributed historical data** stored in **large databases, warehouses** and **other massive information repositories** so to find patterns in data that are:
  - valid: not only represent current state, but also hold on new data with some certainty
  - novel: sound and relevant
  - useful: should be possible to act on the item or problem (useful to solve business problem)
  - understandable: humans should be able to interpret the pattern

Why DM Now?

- Four main reasons why DM now?

The competitive pressure is very strong

- How to gain competitive advantage?
- How to control the volatile market?
- How to satisfy customers need?
- How to manage the high turnover rate of professionals?

Why DM Now: Massive data collection

- **Massive data collection**: large databases (data warehouses) are growing at unprecedented rates to manage the explosive growth in stored data.

- Examples of massive data sets
  - – The current NASA Earth observation satellites generate a terabyte (i.e. 10^9 bytes) of data every day.
  - – Google: Order of 10 billion Web pages indexed
    - 100’s of millions of site visitors per day
  - – MEDLINE text database: 17 million published articles
  - – Retail transaction data: EBay, Amazon, Wal-Mart: order of 100 million transactions per day
    - Visa, MasterCard: similar or larger numbers

Why DM Now: Powerful computers

- Powerful computers: The computing power is increased and is also affordable
  - The need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology.
- Technological Driving Factors
  - Larger, cheaper memory (in hundred GBs, not in MBs)
    - Moore’s law for magnetic disk density
      “capacity doubles every 18 months”
    - Storage cost per byte falling rapidly
  - Faster, cheaper processors (in GHz, not in MHz)
    - the CRAY of 15 years ago is now on your desk
  - Success of Relational Databases and the World Wide Web
    - everybody is a “data owner”

Why DM Now: DM algorithms

- Commercial products (for data mining) are available
  - Data mining algorithms have been matured & there are reliable tools that consistently outperform older statistical methods.
  - New ideas in machine learning/statistics
    - Boosting, SVMs, decision trees, Bayes, text models, etc
  - Existence of around 20-30 mining tool vendors

Example: Why Data Mining

- Fraud detection/Network intrusion detection
  - Which types of transactions are likely to be fraudulent, given the demographics and transactional history of a particular customer?
- Customer relationship management:
  - Which of my customers are likely to be the most loyal, and which are most likely to leave for a competitor?
- Credit ratings:
  - Given a database of 100,000 names, which persons are the least likely to default on their credit cards?
- Targeted marketing:
  - Identify likely responders to sales promotions

Data Mining helps extract such useful information

Database Processing vs. Data Mining Processing

	Database	Data mining	Comments
Query	Well defined Structured Query Language	- Poorly defined - No precise query language	The data miner might not know what he exactly wants to see
Data	Operational data	Non-Operational data	The data have been cleansed and modified to better support the mining process
Output	Precise and Subset of database	Not a subset of database	The output is some hidden useful patterns & knowledge in the database

Query Examples

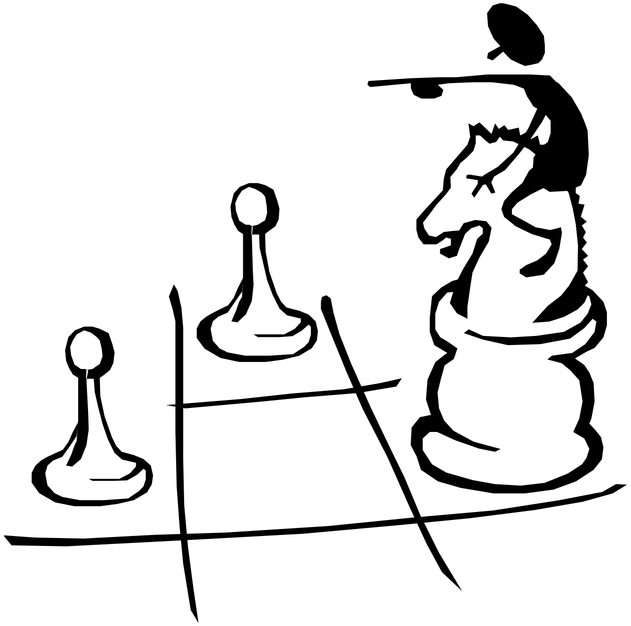
- Database
  - Find all credit applicants with first name ‘Alex’.
  - Identify customers who have purchased more than Birr 10,000 in the last month.
  - Find all customers who have purchased **Bread**

- Data Mining
  - Find all credit applicants who have no credit risks. (classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with **Bread**. (association rules)

Data Mining works with Data Warehouse

- Data Warehouse provides the Enterprise with a memory

- Data Mining provides the Enterprise with intelligence



Exercise

- Discuss (shortly) whether or not each of the following activities is a data mining task.
    1. Dividing the customers of a company according to their profitability.
    2. Computing the total sales of a company.
    3. Sorting a student database based on student identifications numbers
    4. Predicting the outcomes of tossing a pair of dice.
    5. Predicting the future stock price of a company using historical record
    6. Monitoring the heart rate of a patient for abnormalities
    7. Monitoring seismic waves for earthquake activities
    8. Extracting the frequencies of a sound wave

Data Warehouse

- Data warehouse
  - A data warehouse is a collection of different relational database management system responsible for the collection and storage of data to support management decision making and problem solving.
  - It enables managers and other business professionals to undertake Big data mining, online analytical processing, market research and decision support.
  - Current evolution of Decision Support Systems (DSSs)
- Data mart
  - A subset of a data warehouse for small and medium-size businesses or departments within larger companies

Data warehousing

- Data warehouse is an integrated, subject-oriented, time-variant, non-volatile database that provides support for decision making.

- ***Integrated*** → centralized, consolidated database that integrates data derived from the entire organization.
  - Consolidates data from multiple & diverse sources with diverse formats.
  - Helps managers to better understand the company’s operations.

- ***Subject-Oriented*** → Data warehouse contains data organized by topics.
  - E.g. Sales, marketing, finance, etc.

Data warehousing

- ***Time variant*** **→** In contrast to the operational database that focus on current transactions, the data warehouse represent the flow of data through time.
  - – Data warehouse contains data that reflect what happened last week, last month, past five years, and so on.

✓ **Non volatile** → Once data enter the data warehouse, they are never removed. Because the data in the warehouse represent the company's entire history.

✓ Because data is added all the time, warehouse is growing.

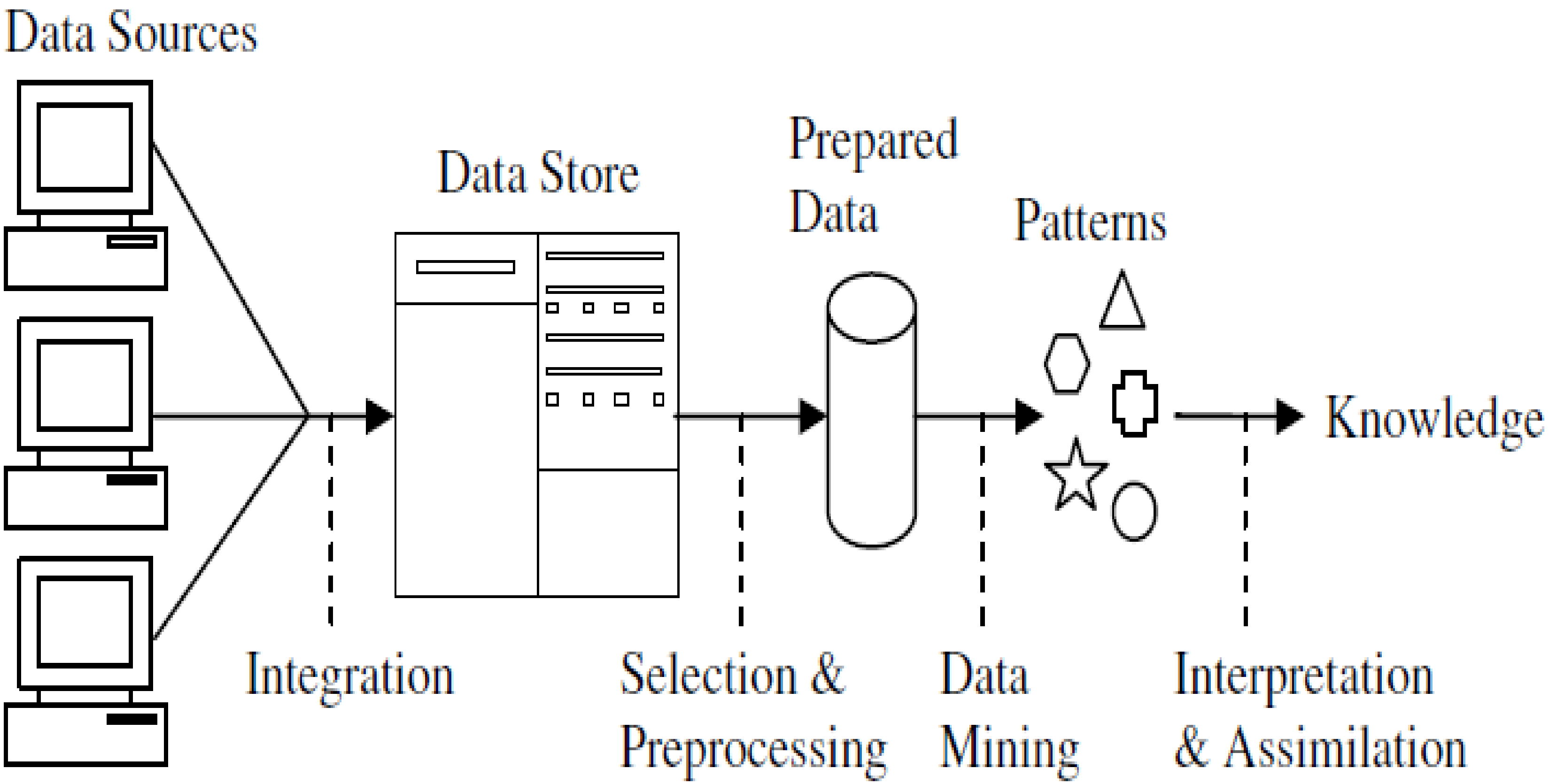
Database & data warehouse: Differences

- The data warehouse and operational environments are separated. Data warehouse receives its data from operational databases.
  - Data warehouse environment is characterized by read-only transactions to very large data sets.
  - Operational environment is characterized by numerous update transactions to a few data entities at a time.
  - Data warehouse contains historical data over a long time horizon.
- Ultimately Information is created from data warehouses. Such Information becomes the basis for ***rational decision making***.
- The data found in data warehouse is analyzed to discover previously ***unknown data characteristics, relationships, dependencies, or trends.***

Business Intelligence

- BI takes advantage of data mining and data warehousing to help organizations gather their information in a **timelier and in a more valuable manner**
- BI keeps the organization:
  - informed about the market trends,
  - alerts to new market potentials,
  - helps to determine how competitors are doing
- Business intelligence is information about a company's ***past performance that is used to help predict the company's future performance.***
  - It can reveal emerging trends from which the company might profit.
- Without such information and knowledge the organization may suffer false growth or setbacks

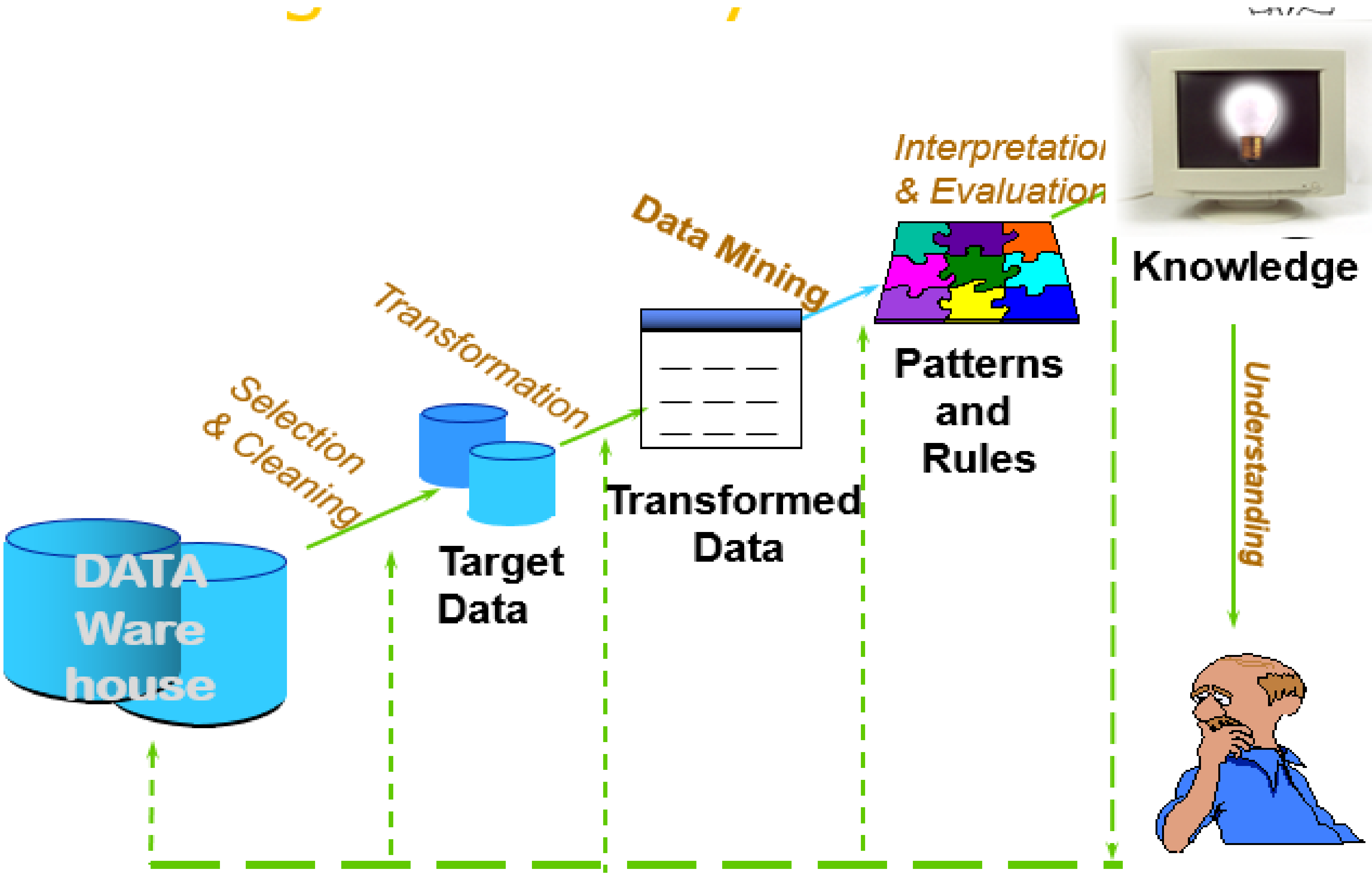
Data Warehouse as part of Data Mining



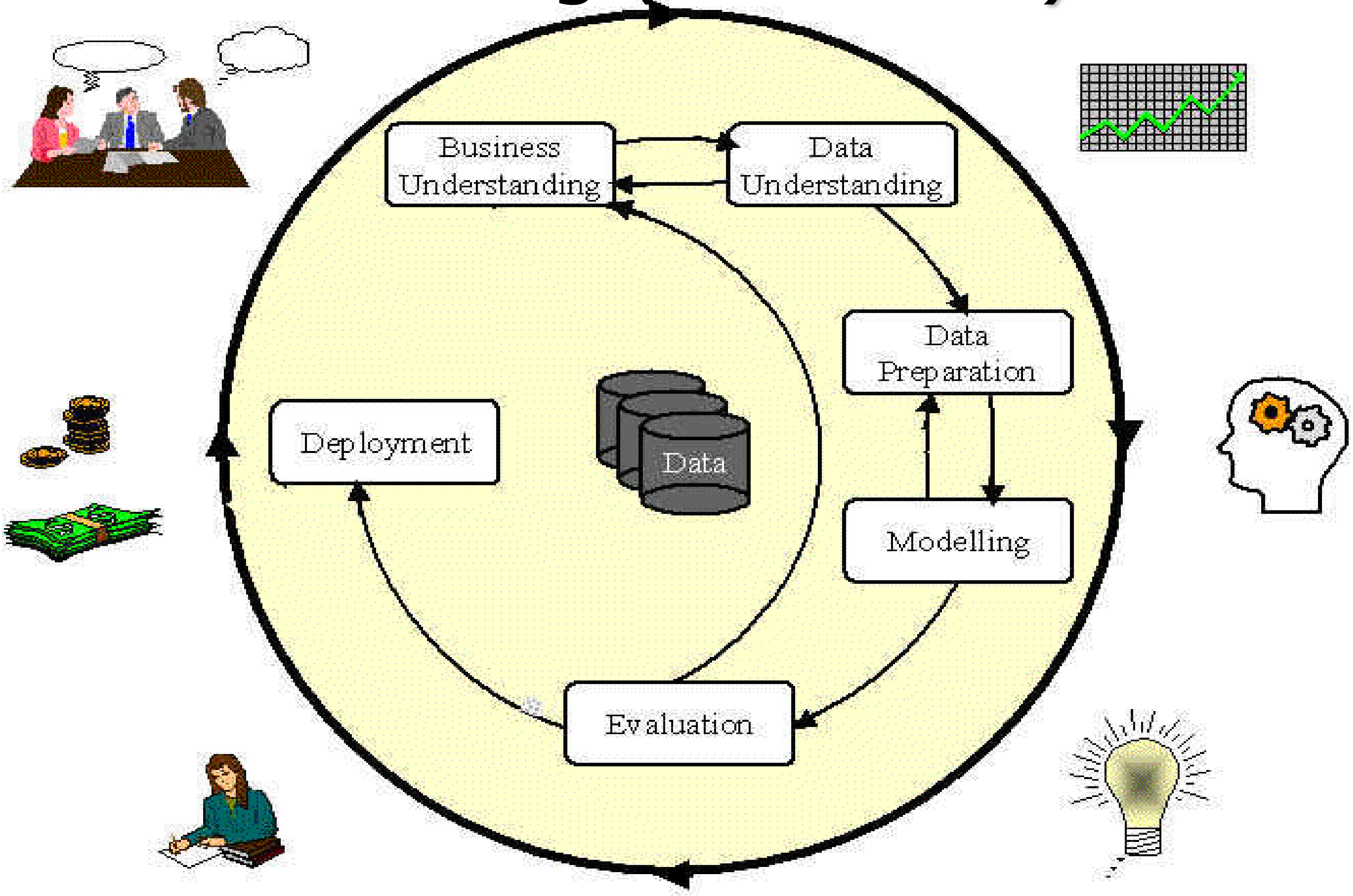
Knowledge Discovery in Databases(KDD)

- KDD is often used as a synonym for Data Mining.
  - Some author define KDD as the whole process involving: data selection → data pre-processing: cleaning → data transformation → mining → result evaluation → visualization
  - Data Mining, on the other hand, refer to the modeling step using the various techniques to extract useful information/pattern from the data.
- KDD is the process of finding useful information and patterns in data
- DM is the use of algorithms to extract hidden patterns & knowledge in data

The KDD process



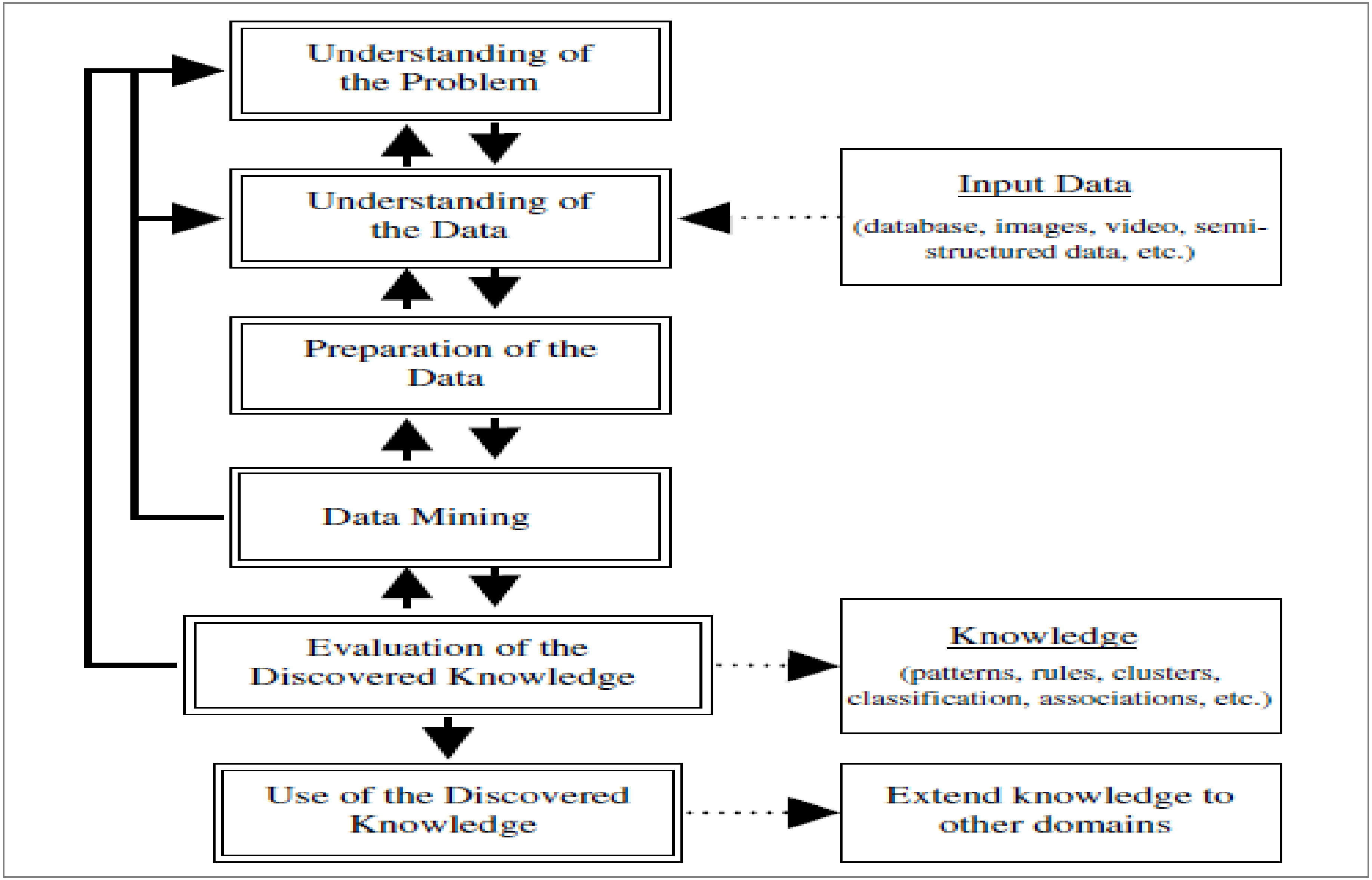
CRoss Industry Standard Process for Data Mining (CRISP-DM)



Phases and Tasks

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria*	**Collect Initial Data** *Initial Data Collection Report*	*Data Set Description* **Select Data** *Rationale for Inclusion / Exclusion*	**Select Modeling Technique** *Modeling Assumptions*	**Evaluate DM Results** *Assess Results w.r.t. Business Success criteria* *Approved Models*	**Plan Deployment** *Deployment Plan*
Assess Situation *Inventory of Resources* *Requirements, Assumptions, & Constraints, Risks & Contingencies* *Terminology ; Costs and Benefits*	**Describe Data** *Data Description Report*	**Clean Data** *Data Cleaning Report* **Construct Data:** *Derived Attributes* *Generated Records*	**Generate Test Design** *Test Design*	**Review Process** *Review of Process*	**Plan Monitoring & Maintenance** *Monitoring and Maintenance Plan*
Determine Data Mining Goal *Data Mining Success Criteria*	**Explore Data** *Data Exploration Report*	**Integrate Data** *Merged Data*	**Build Model** *Parameter Settings;* *Model construction*	**Determine Next Steps** *List of Possible Actions* *Decision*	**Produce Final Report** *Final Presentation*
Produce Project Plan *Initial Asessment of Tools and Techniques*	**Verify Data Quality** *Data Quality Report*	**Format Data** *Reformatted Data*	**Assess Model** *Model Assessment* *Revised Parameter Settings*		**Review Project** *Experience* *Documentation*

Hybrid Knowledge Discovery Process



Understanding of the Problem

Understanding of the Data

Preparation of the Data

Data Mining

Evaluation of the Discovered Knowledge

Use of the Discovered Knowledge

Input Data

(database, images, video, semi-structured data, etc.)

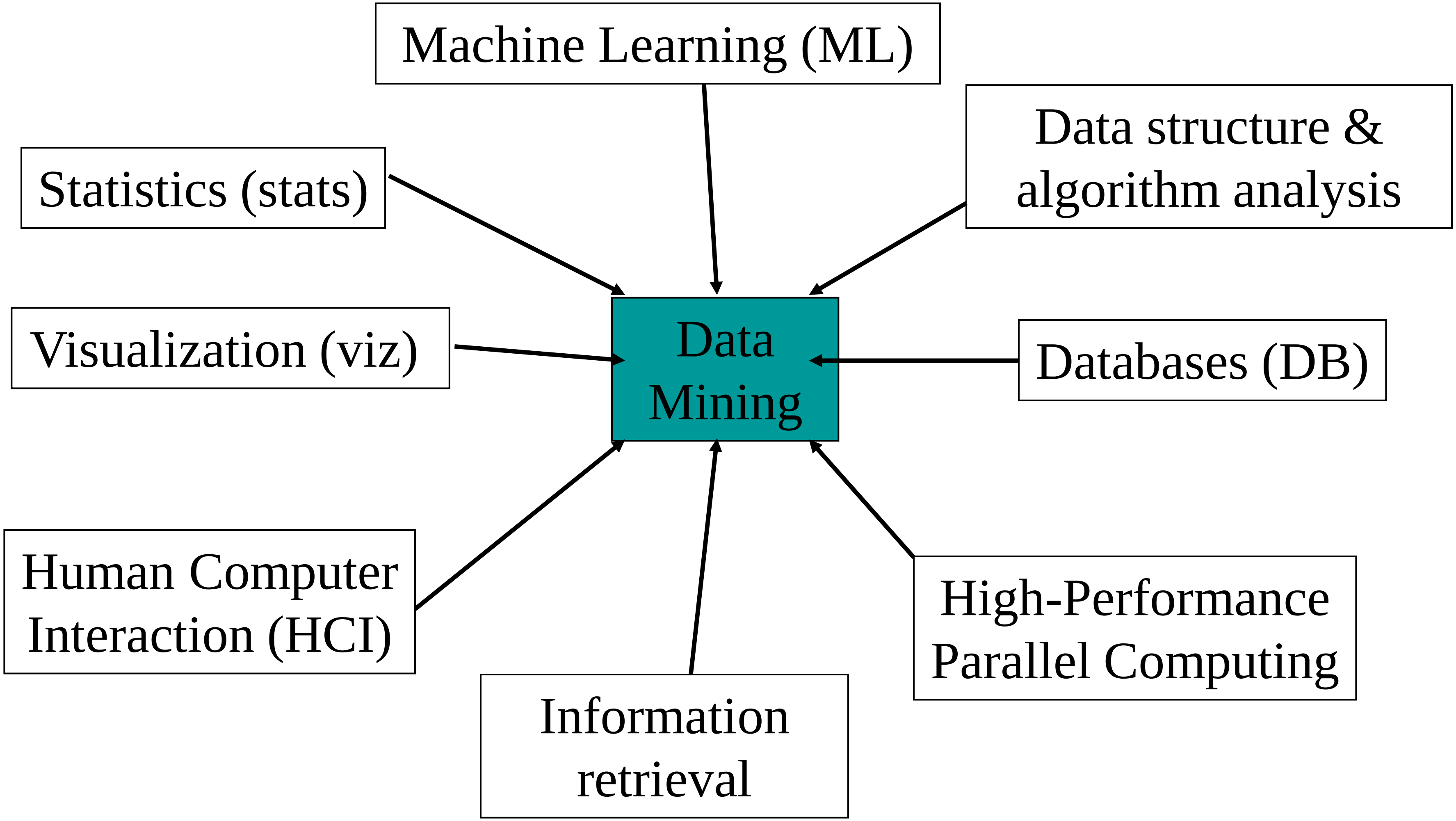
Knowledge

(patterns, rules, clusters, classification, associations, etc.)

Extend knowledge to other domains

DM: Intersection of Many Fields

- Data mining overlaps with machine learning, statistics, artificial intelligence, databases, visualization



Data Mining Metrics

How to measure the effectiveness or **usefulness** of data mining approach?

- Return on Investment (ROI)
  - From an overall business or usefulness perspective a measure such as ROI is used
  - ROI compares costs of DM techniques against savings or benefits from its use
- Accuracy in classification
  - Analyze true positive and false positive to calculate recall, precision of the system
  - Measure percentage of correct classification
- Space/Time complexity
  - Running time: how fast the algorithm runs
  - Storage or memory space requirement

Data Mining implementation issues

- **Scalability**
  - Applicability of data mining techniques to perform well with massive real world data sets
  - Techniques should also work regardless of the amount of available main memory
- **Real World Data**
  - Real world data are noisy and have many missing attribute values. Algorithms should be able to work even in the presence of these problems
- **Updates**
  - Database can not be assumed to be static. The data is frequently changing.
  - However, many data mining algorithms work with static data sets. This requires that the algorithm be completely rerun any time the database changes.

Data Mining implementation issues

- **High dimensionality**:
  - A conventional database schema may be composed of many different attributes. The problem here is that all attributes may not be needed to solve a given DM problem.
  - The use of unnecessary attributes may increase the overall complexity and decrease the efficiency of an algorithms.
  - The solution is dimensionality reduction (reduce the number of attributes). But, determining which attributes are not needed is a tough task!
- **Overfitting**
  - The size and representativeness of the dataset determines whether the model associated with a given database states fits to also future database states.
  - Overfitting occurs when the model does not fit to the future states which is caused by the use of small size and unbalanced training database.

Data Mining implementation issues

- **Ease of Use of the DM tool**
  - Since data mining problems are often not precisely stated, interfaces may be needed with both domain and technical experts
  - Although some techniques may work well, they may not be accepted by users if they are difficult to use or understand

- **Application**
  - Determining the intended use for the information obtained from the DM tool is a challenge.
  - Indeed, how business executives can effectively use the output is sometimes considered the most difficult part. ***Because the results are of a type that have not previously been known***.
  - Business practices may have to be modified to determine how to effectively use the information uncovered