INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

Chapter 3

Knowledge discovery from web data

PROFESSOR

Juan D. Velásquez

Outline

1. The KDD Process
2. Data sources and cleaning
3. Data consolidation and information repositories
4. Data mining
5. Tools for mining data
6. Using data mining to extract knowledge
7. Validation of the extracted knowledge
8. Mining the web

Section 3.1

- The KDD Process

Introduction

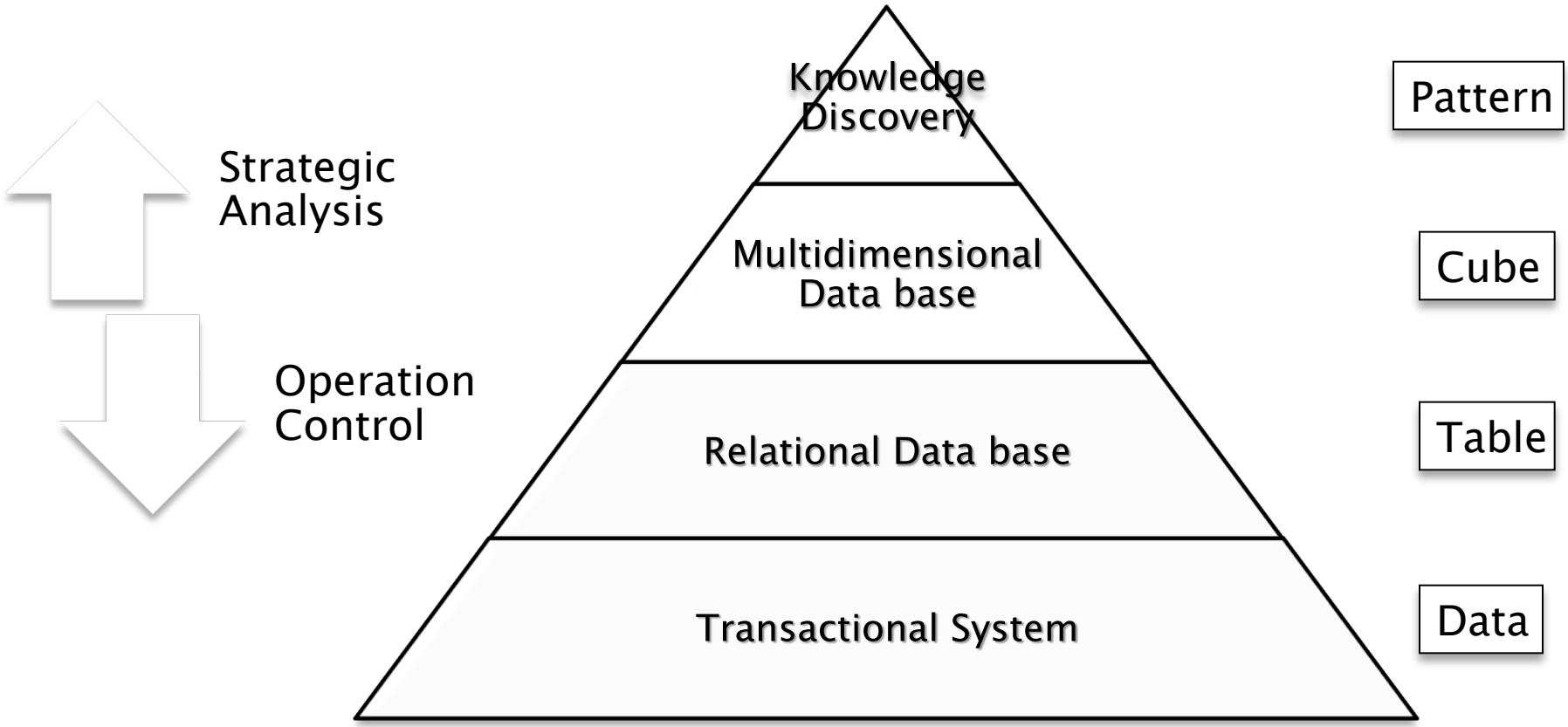
- The data resume for **decision making** is a **traditional report of the statistics**.

- Today **Information** is a **valuable resource** that have important implication in the **productivity of the company**. It administration is called **knowledge discovery**.
  - This process usually imply a **large amount of data**.

Introduction (2)

- **The Goal**
  - Find knowledge valid, useful, relevant and new over a phenomena or activity.
- A visual representation of the result in order for an easier interpretation.
- The usability must allow a flexible, dynamic and collaborative process.
- Scalability and efficiency are important requirement.

Knowledge Pyramid



Considerations

- Scientific Method:
  - Hypothesis-Experiment-Knowledge
- Knowledge Discovery:
  - Data-Hypothesis-Knowledge
- Expert support
- Space-Time dimension.
- Data Quality, Homogeneity

Asking the right question:
Who is the right person?

Query. 1 | Query. 2 | Query. 3 | Query. 4 | Query. 5

Asking the right question (2)

- Have first a **clear objective**
- **Understanding the goals** of the process.
- Question oriented to generate knowledge.
- **We don't want another statistical report.**

Asking the right question (3)

- Evaluating Point Of Sale
  - Which **POS** has **better/worst production level**?
  - Which **products** are **best selling** on some POS?
  - Which **promotions** has been with the **best impact** in sales.
- Evaluating promotion
  - Which **effect over the total sales** had the last year **marketing campaign**?
  - Has the **promotion** an **effective cost**?

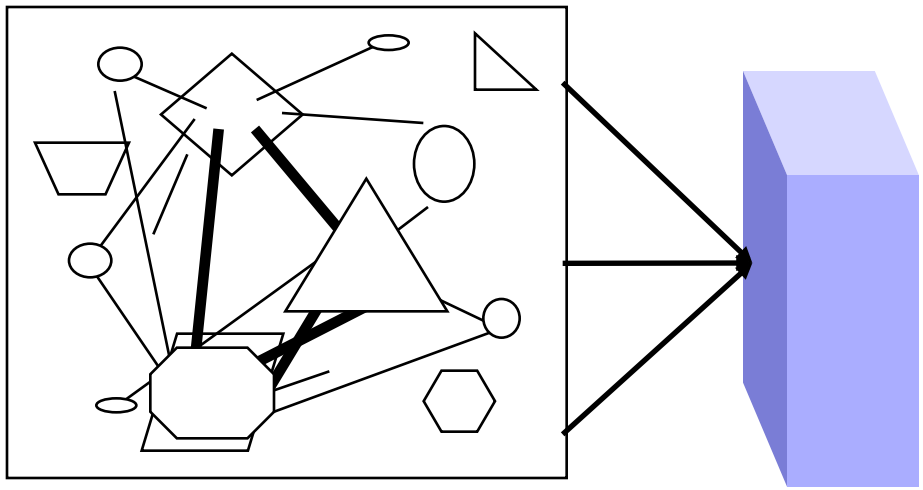
Asking the right question (4)

- Sales and tendencies
  - Which **product** has **top selling score** and which ones doesn't moves?
  - How much **utility** are perceived by **each product**?
  - How affected are the **pattern of purchases** by a **change in the price**?
  - How affected are the **popularity of a product** in relation to **the time**?
  - Which **special characteristics** has the **client** that **buy this product**?
  - Which are the **best selling brand**?

Asking the right question (5)

- The client preferences and sales
  - How many **man bought** this product?
  - How many **married women bought** in this store?
  - What is the **impact of the education** on the **family sales pattern**?
  - What is the impact the **change on the utility** by product on the **sales of men v/s women**?
- Inventory
  - Which warehouse had the **best usage ranking on special period**?
  - What is the **behaviour of the inventory** level of the warehouses?
  - What are the **most valuable warehouse**?

Data Selection



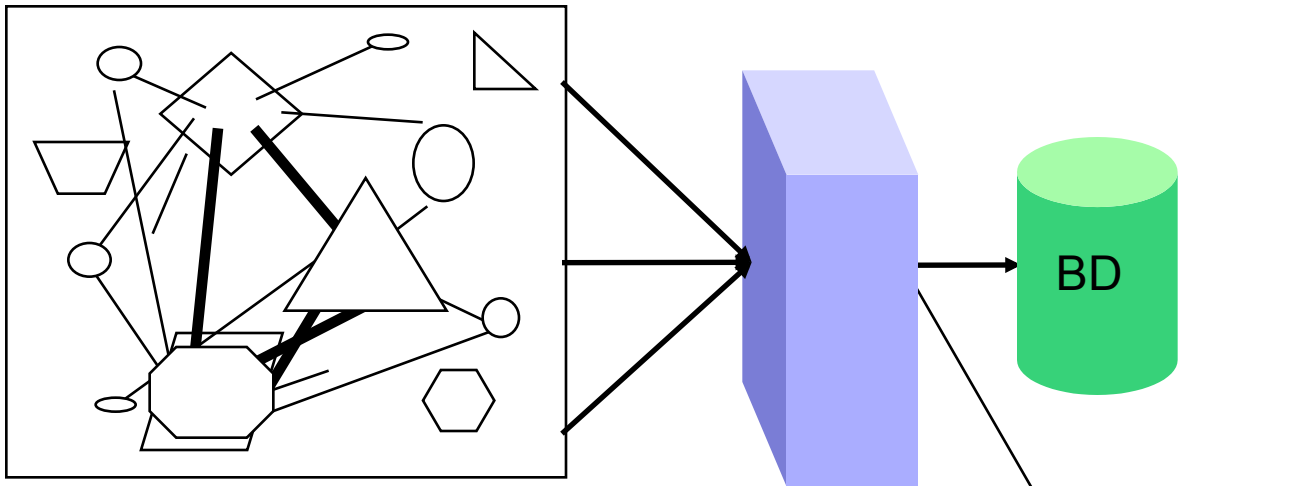
Operational Data

Data Selection

Data Selection (2)

- **All data in a Data Warehouse.**
  - Huge amount of data, more time to process.
- **Data Marts, segmenting the study to an operational sector.**
  - More conventional amount of data, but less time to process.
  - Oriented to a more global strategy for the business.

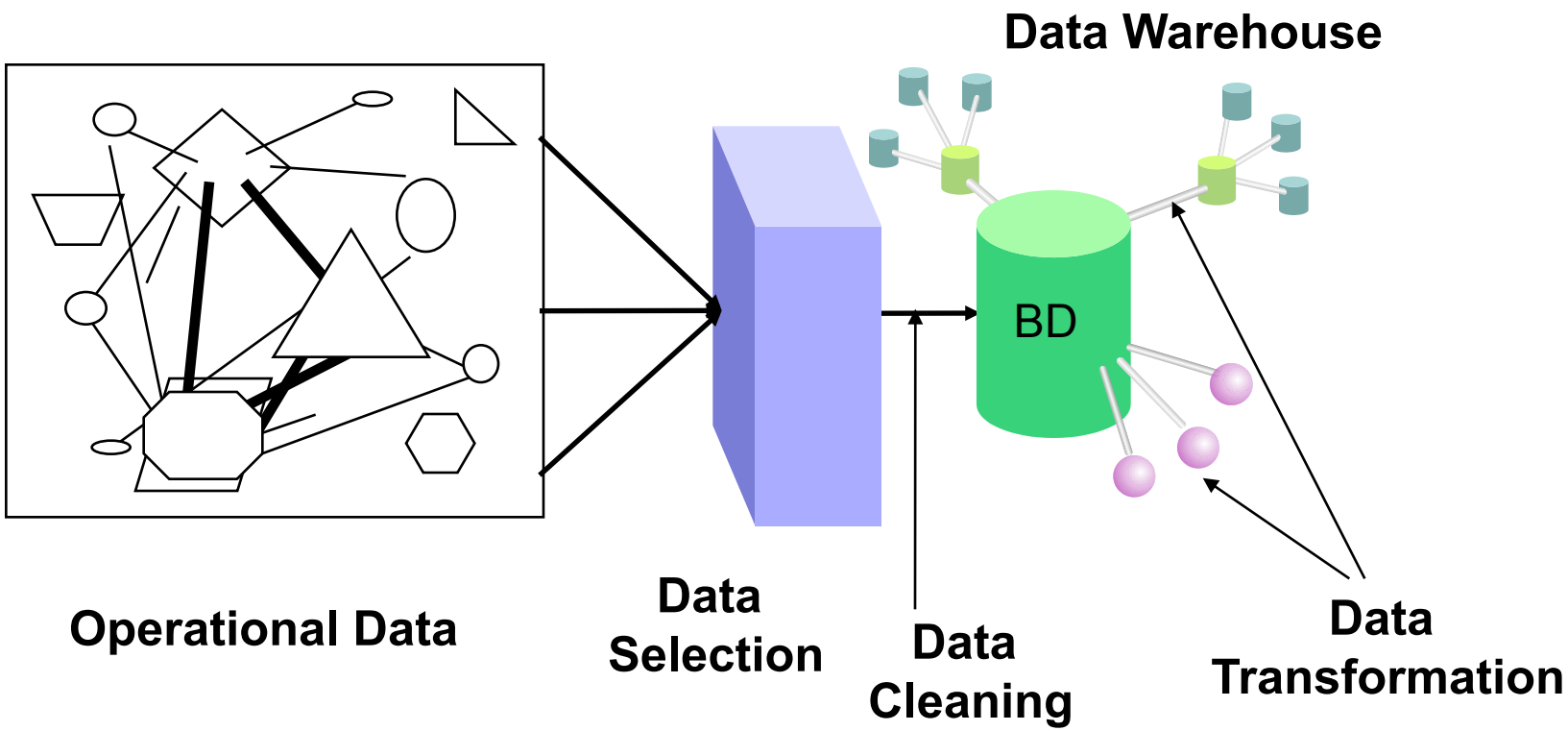
Preprocessing and data cleaning



Preprocessing and data cleaning (2)

- ✓**Missing** Values
- ✓**Dynamical** data
- ✓**Distributed** and **big data** bases.
- ✓**Noise**

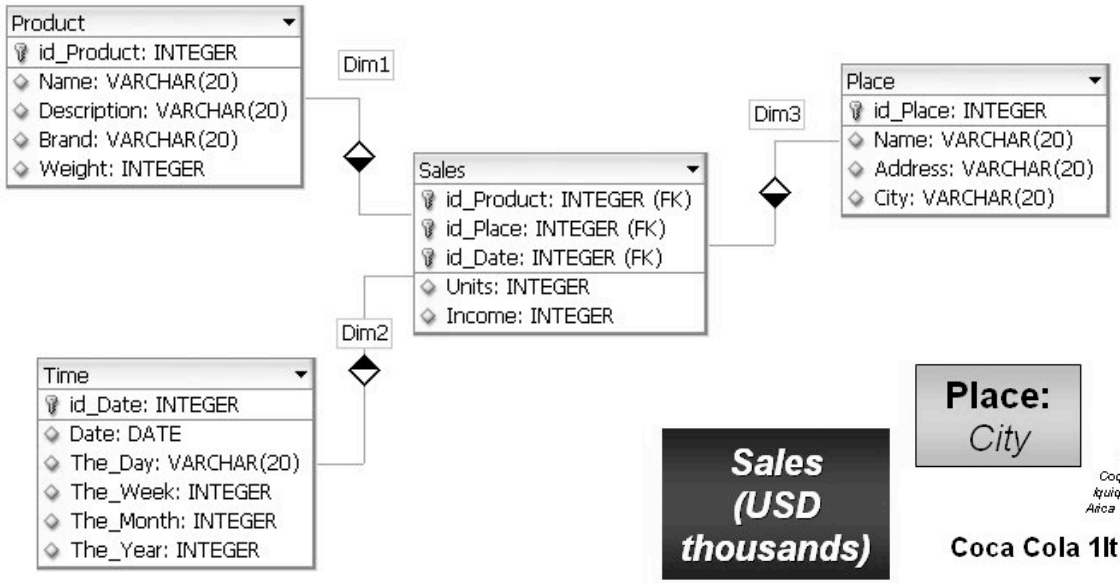
Data transforming

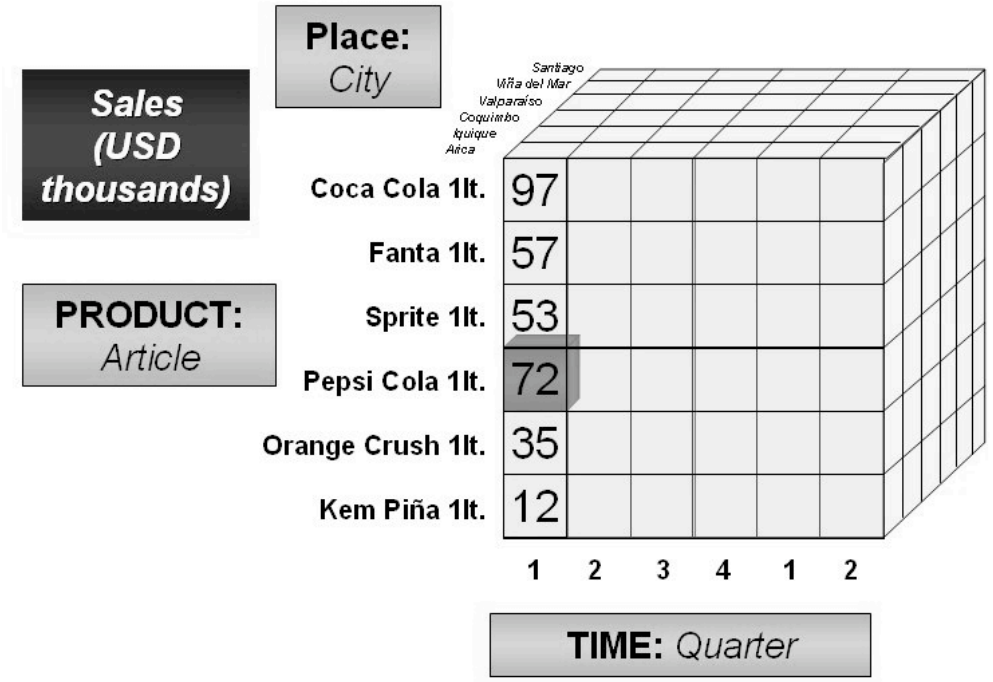


Data transforming (2)

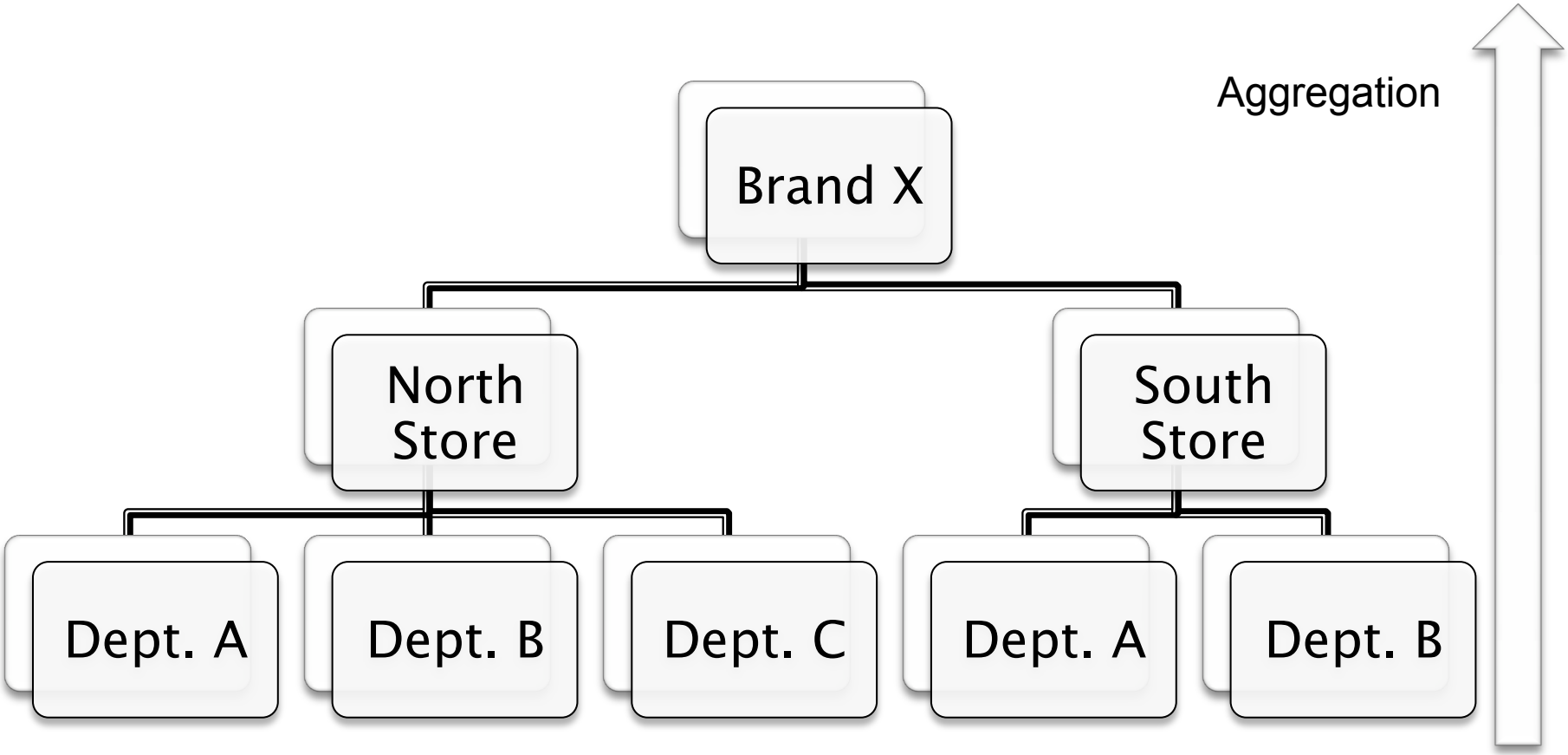
- Dimensions (OLAP)
- Hierarchies
- Variables
- Rotations
- Segmentation
- Drill-Down

Dimension

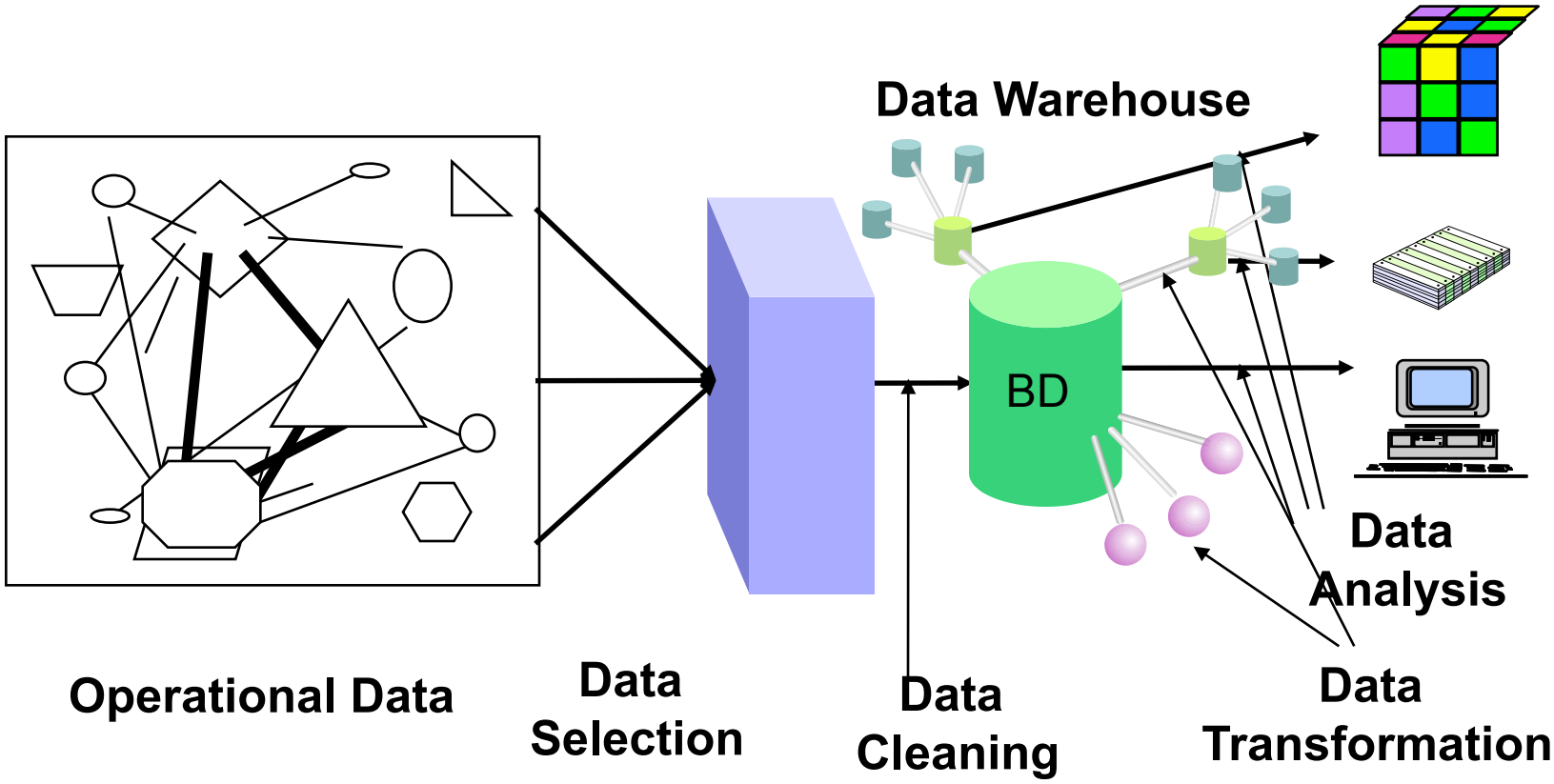

Product
id_Product: INTEGER
Name: VARCHAR(20)
Description: VARCHAR(20)
Brand: VARCHAR(20)
Weight: INTEGER
Dim1
Sales
id_Product: INTEGER (FK)
id_Place: INTEGER (FK)
id_Date: INTEGER (FK)
Units: INTEGER
Income: INTEGER
Dim3
Place
id_Place: INTEGER
Name: VARCHAR(20)
Address: VARCHAR(20)
City: VARCHAR(20)
Dim2
Time
id_Date: INTEGER
Date: DATE
The_Day: VARCHAR(20)
The_Week: INTEGER
The_Month: INTEGER
The_Year: INTEGER


Sales (USD thousands)
Place: City
Santiago
Viña del Mar
Valparaíso
Coquimbo
Iquique
Arica
PRODUCT: Article
Coca Cola 1lt.
Fanta 1lt.
Sprite 1lt.
Pepsi Cola 1lt.
Orange Crush 1lt.
Kem Piña 1lt.
97
57
53
72
35
12
1 2 3 4 1 2
TIME: Quarter

Hierarchy on the business



Data Mining


Data Warehouse
BD
Data Analysis
Operational Data
Data Selection
Data Cleaning
Data Transformation

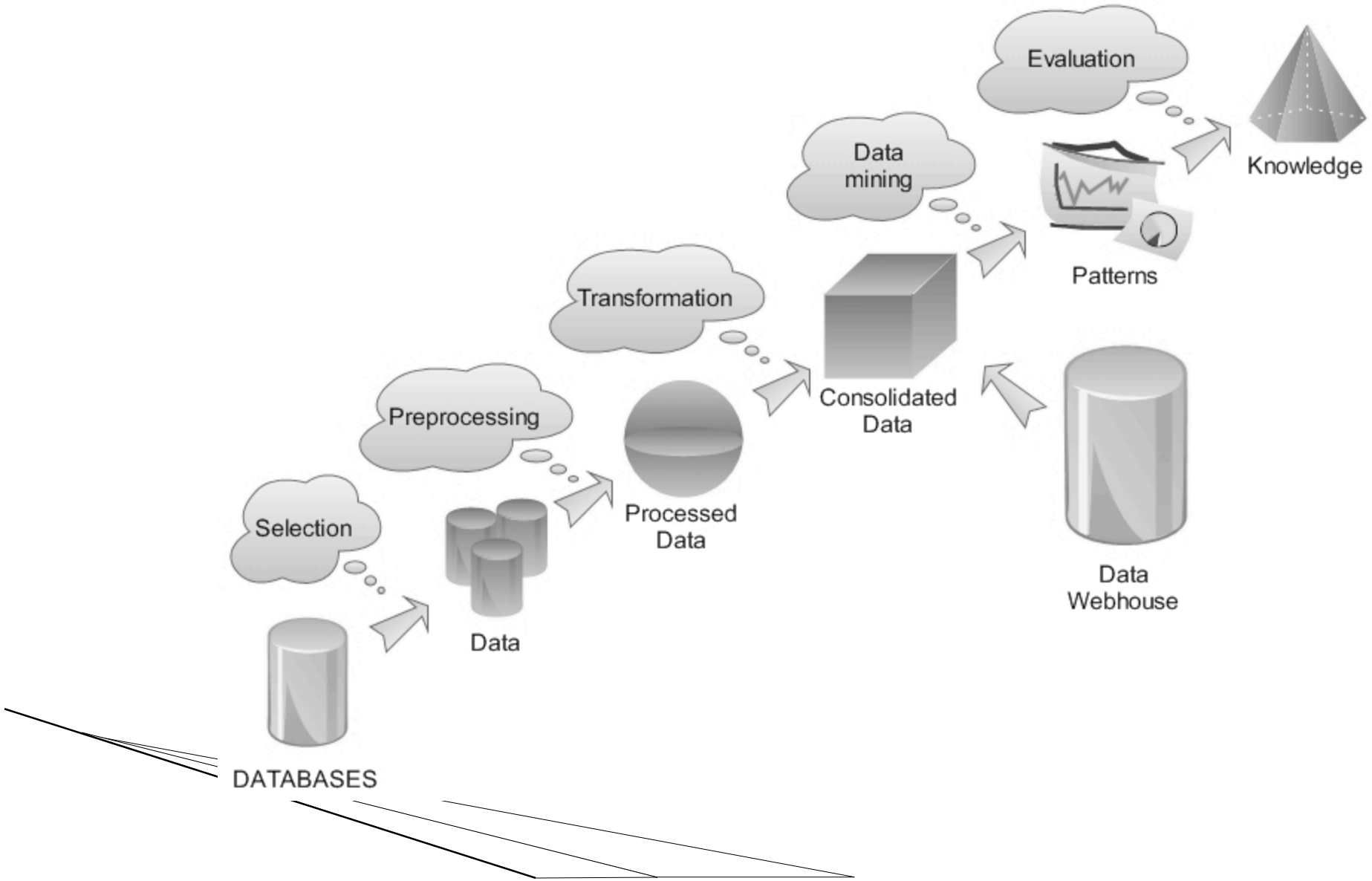
Data Mining (2)

- Classification
- Clustering
- Regression
- Dependency modelling
- Tendencies and similarities of them
- Relations

Interpreting and verifying the results

- **Summarized critical factors**,
  - observing its **impact** on the business and **try to explain** them.
- The **Expert** identify the **knowledge.**
- **Store the knowledge generated**, **reuse** it on a future KDD process.

The KDD Process



Conclusion

- **Tactical Value** of large amount of data.
- **Analysis capacities**,
  - finding useful knowledge v/s Cost and time.
- The **new knowledge MUST BE validated** by the **new data** in order to plan the business.
- The **expert must validate** each step in the process.
- **The correct interpretation** of the result will generate the **knowledge**
- **The new discovery must be stored** in a structure that allows **reuse in others KDD.**

Section 3.2

- Data source and cleaning

Data Source and Cleanning

- The **identification of real data sources** are **important step** in the KDD process.
- **Irrelevant data** (noise) often leads to **analytic errors.**
- Different Data Format introduce a **cost of interpretation/transformation.**
  - **Metadata** allows us to **standardize** the data.

Most frequent data problem

- **Data consistency**: Operational system are constructed on base of the **direct business requirements.** That means any other requirement on them (like KDD) have been never implemented and tested.
  - That imply **inconsistency.**

- **Data Manipulation Errors**: usually occur when **testing is avoided.** Example: Client with name "Batman" that remains from the development process.

- **Irrelevant Data**: Some data **need to be filtered** because is not part of the analysis

The Web Data

- **The problem**
  - **Garbage-in, Garbage-out**
- **Web Data**
  - Highly Variable in **Type**
  - Highly Variable in **Format**
- HTML includes:
  - Tags
  - Text
  - Multimedia
- **Logs**: have an standard but the information that we want (sessions) is **not explicit.**
- **Web Sites** Change over time and **usually nobody track these changes.**

Web Data in the KDD process

- We require:
  - Pages transformed to **Feature Vectors**
    - (*to be explain in further chapter*)
  - **Clean individual Session** from users.
  - The **Web site graph**
- The **web data cleaning** and **pre-processing** activities should store the result in an **information repository** for further **data mining process.**

Section 3.3

- Data consolidation and information repositories

Data consolidation and information repositories

- Information **architecture complexity** depends on **computing background resources.**
- Usually:
  - **Heterogeneous operational systems** not designed to work together.
  - Not a **single data storage convention**
  - Data **without explicit stored correlations**
    - (like: owner-car-repair)
- In order to restore the suitable integrity we need to perform:
  - Integration Task
  - Consolidation Task

Data Staging Area
DSA

- Temporary Location for data.
  - Used for performing the intermediate transformation.
  - Usually in all data has already an standard data type.
- **Data is prepared for more complex transformations**

XOR

X1	X2	Y
1	1	0
1	0	1
0	1	1
0	0	0

The ETL Process:
Extraction, Transformation & Loading

- **Extraction**
  - Data are **extracted and stored in the DSA.**
  - **Some data standardization** are performed and **cleaning.**
- **Transformation**
  - More complex transform are made
  - For example: **sessionization.**
- **Loading**
  - The **clean (integrity, consistency) data are stored** in the historic **final repository**
  - **For example**: webhouse and user sessions.

The Data Warehouse: The Web House

- Data transforms resume:
  - Cleaning: Avoiding null values.
  - Pre-processing: Data type standardization
  - Integration
  - Consolidation
- Store Historic information:
  - Time is an important dimension.
- The business memory of
  - web client behavior on the site
  - semantic content of the site
  - hyperlink structure.
- Is the **source for data mining**

Section 3.4

Data Mining: Machine Learning

Based on the Machine Learning Lecture by Vasile Palade, Oxford University. Co-Author of the book:

J.D. Velásquez and V. Palade “Adaptive Web Site”. IOS Press, Netherland, 2008.

Traditional Motivation:
Beer and Diaper

- Case large US supermarket

  Customer purchase behaviour:

  product linked with another

  (bread -> butter, beer -> diapers)

  Market segment:

  young men married in the last three years with small children.

- Based on this information, we deduce:

  place diaper and beer on the same place on Friday afternoons.

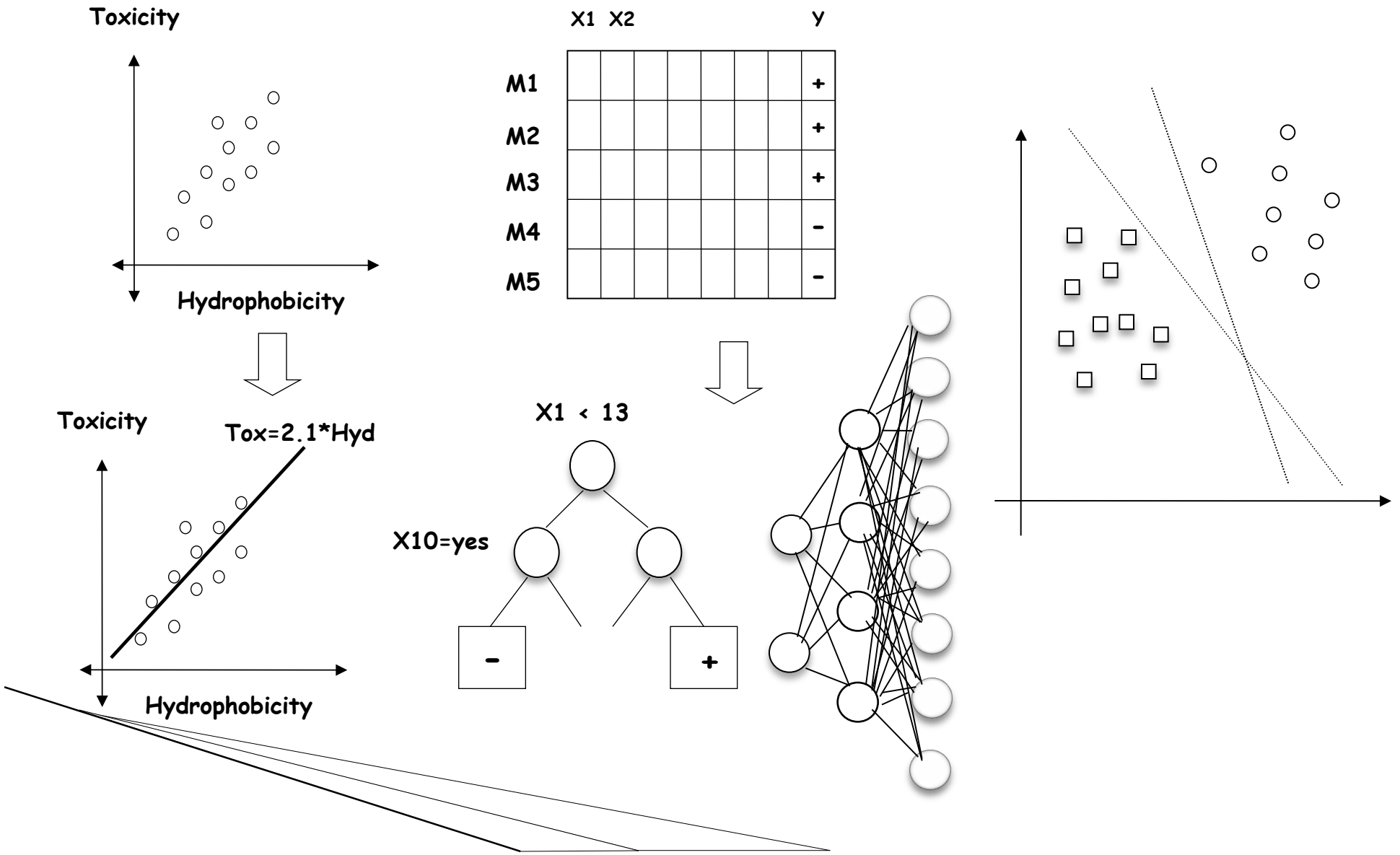
- **data mining** algorithm are **key to pattern discovery**.

- The business take the pattern and take some **profitable actions**.

What is Machine Learning (ML)?

- Shows how to build computer programs that improve their performance at some task through experience (Mitchell, 1997).
- ML paradigm can be viewed as
  - "programming by examples". Construct models from data.
- The goal of ML is to develop algorithms that do
  - the learning automatically without human intervention or assistance.
- **Useful when you have lots of data.**

Programs that Construct Models for Data



Machine Learning

- It is a **core sub-area** of **Artificial Intelligence**

  Intelligence = Learning + Reasoning

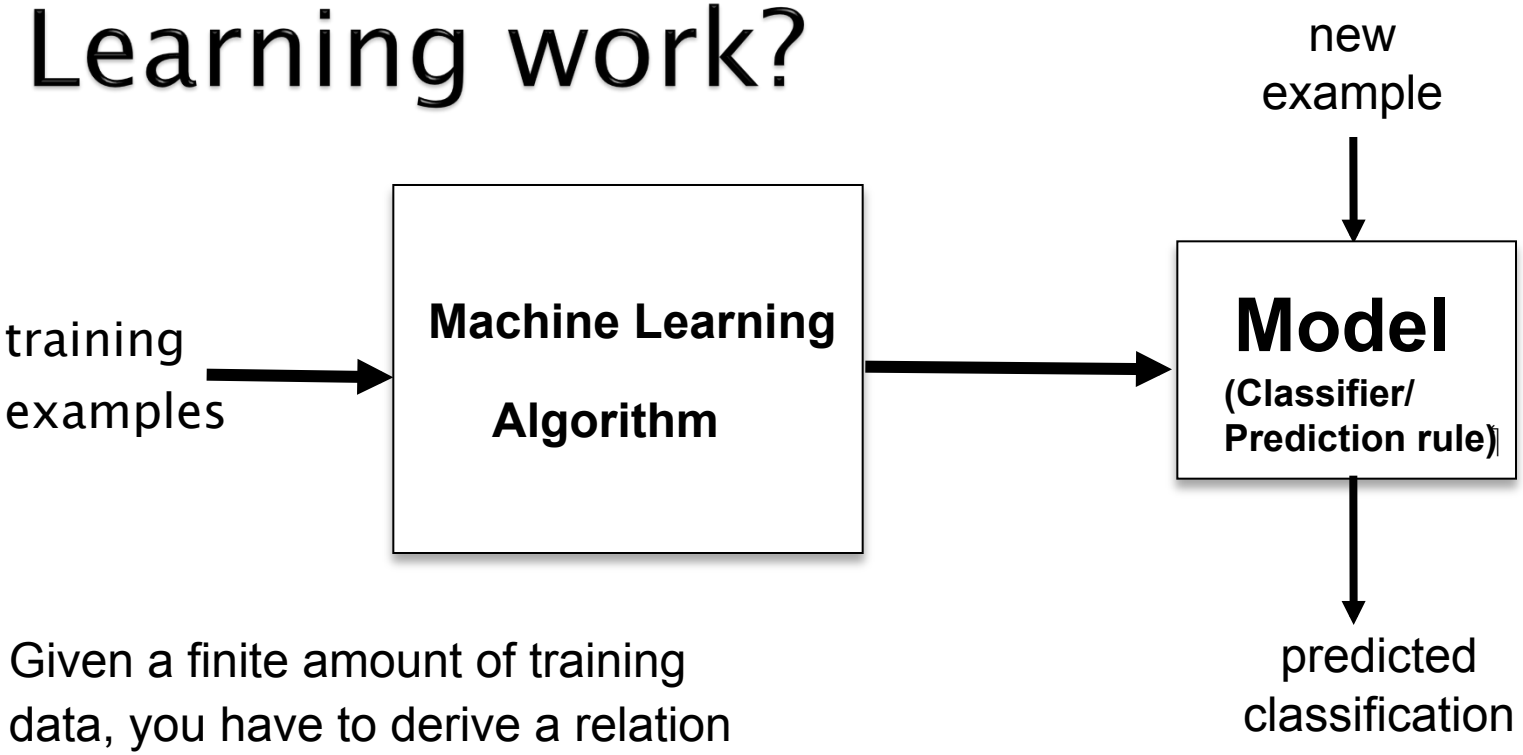
  [**Learning is at the core of intelligence**]

- It is related also to
  - Statistics, Computational Complexity Theory, Information Theory, Biology (Genetics, Neurobiology), Cognitive Science, Control Theory.

Approaches in Machine Learning

- Decision tree
- Instance/Case-based learning
- Bayesian learning
- Neural Networks
- Learning through fuzzy logic
- Genetic algorithms
- Multistrategy Learning
- Rule Induction (ILP)
- Analytical learning
- Reinforcement learning
- Support Vector Machines
- Multi-agent learning

How does the Machine Learning work?



Given a finite amount of training data, you have to derive a relation (function/model) for an infinite domain

Data set is divided into:

- Training set – used for training
- Testing set – used for validating the model

Examples

Training examples for the target concept PlayGolf

Day	Outlook	Temperature	Humidity	Wind	PlayGolf
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Cool	Normal	Weak	Yes
5	Sunny	Mild	Normal	Weak	Yes
6	Rain	Mild	High	Strong	No
7	Overcast	Hot	Normal	Weak	Yes

Predict

8	Rain	Hot	Normal	Strong	??????

Examples

Training examples for learning the 3 classes of Iris flower

	Sepal-length	Sepal-width	Petal-length	Petal-width	Class
1	6.8	3	6.3	2.3	Versicolour
2	7	3.9	2.4	1.1	Setosa
3	2	3	2.3	1.7	Verginica
4	3	3.4	1.5	1.5	Verginica
5	5.5	3.6	6.8	2.4	Versicolour
6	7.7	4.1	1.2	1.4	Setosa
7	6.3	4.3	1.6	1.2	Setosa
8	1	3.7	2.2	2	Verginica
…	…	…	…	…	…

Predict

	4.5	3.0	6.3	1.5	??????

Some Definitions …

- Examples (Instances) –
  - objects that are being classified
- Attributes (Features/Variables)
  - are used for describing examples
- Label
  - category (class) that we are trying to learn and predict.
  - During training the learning algorithm is supplied with labeled examples, while during testing, unlabeled examples are provided.
- Concept
  - mapping from examples to labels (classes)

Some Definitions (2)

- **A learning method**
  - utilizes a **set of training examples** $\{(x_1,y_1) \ldots (x_n,y_n)\}$ to approximate a function f(x). x values are usually vectors of the form $<x_{i1}, x_{i2}, \ldots, x_{im}>$ (features). The y values are usually drawn from a discrete set of classes.
- **Purpose of learning**
  - **approximate an input to output mapping**
  - approximate the function f(x) and produce a classifier or a model
- **Classification**
  - process of **assigning presented information into classes/categories** of the same type.
- **Classifier**
  - **computer-based agent that can perform classification**

Some Basic Distinction

Supervised Learning (prediction)	• Build models to predict a variable/class using data for other variables/features. There is a "teacher" who tells what is the right class of any given example in the training set (direct feedback).
Unsupervised Learning (description)	• Build models to describe a set of variables (or relations). Given a population of unclassified examples, invent reasonable concepts (clusters), and find definitions/ meanings of those concepts. No teacher exists during training (no feedback
Reinforcement Learning (Adaptation)	• Indirect feedback after many examples, an agent that evolve according to its environment (Robotic Movement).

Some Basic Distinction (2)

Feature-based Learning
(Memory Based learning)

- Supervised based, build models from data that are in the form of a single table of features and their values.
- These tables acts as memory resource for *relevant* instance found in the past for reinforcing the learning.

Relational Learning

- Build models from data that are in the form of several related tables (each table represents a relation).
- Requires data in some relational format like Relational DB.
- It's like Feature-Based but with an additional structure that the algorithm discover.

Applications of Machine Learning

- medical diagnosis
- industrial fault diagnosis
- **text categorization**
  - **(text mining/web mining)**
- **information retrieval**
- speech recognition
- natural language processing
- signal and image processing
- etc, etc
- industrial control/ automation
- **data mining**
- **business application**
  - (stock market prediction)
- bioinformatics
- intelligent tutoring systems
- virtual reality
- **decision support systems**

Data mining techniques

Association Rules

- If-then expressions

Classification

- Classes of objects

Clustering

- Finding similar objects

Association Rules

- Find **significant correlations** among a large data set.
  - An association rule is an implication of the form [X => Y]
  - Where X, Y are group of items.
- The meaning of [X => Y] is:
  - in "most of the cases" if we have the set of X items in a transaction then we will found also these others set Y of items.
- If we known the most probable rules [X=>Y]
  - then we could **configure our business** for **efficiency and profitability.**

Measuring Association Rules:
Support and confidence

- We need some measure of acceptability of the rules.
- **Support (α):**
  - The probability of occurrence of X or Y in a tra n saction. That means all the possible cases related with X or Y.
- **Confidence (β):**
  - The probability that X => Y occu r in a transaction. That mean the conditional probability P(Y|X)
- Example (Beer, Diaper):
  - 5% of all transaction show or beer or Diaper.
  - 53% of the consumer that consume beer also consume diaper.

Classification

- Sort objects in different categories or classes
- The process:
  - Learning (training):
    - Iterative process working with training data associated with **known class**.
    - The training stop where the % of correct prediction is superior to a given threshold.
  - Classification:
    - Having a trained model we can measure the real classification ability by using it on an **independent data test set.**
  - We could iterate testing the classification with a test set and training set, and changing parameters of the model in order to obtain better adjustment.

Clustering

- Is the process of grouping objects with similar characteristics.
  - We need a **similarity measure and a neighbourhood definition**.
- **A similarity measure is not a distance function!**
- In a supervised learning process
  - the classification is known a priori.
- In a UNsupervised learning process
  - the classification is NOT known a priori.

Clustering

- **Cluster**
  - A collection of similar physical/abstract objects.
    - (Similar within the same cluster, dissimilar to objects in other clusters)
- ***Clustering***
  - *process of grouping a set of objects into clusters*
    - It is an unsupervised learning method.
    - It is a common and important ML task
    - It has many applications:
      - stand-alone classification tool
      - preprocessing tool for other ML algorithms.

Clustering Approach:
Partition Clustering

1. Divide n object into k group called partitions

2. A partition $P=\{p_1,...,p_k\}$ of a set $X=\{x_1,...,x_n\}$, satisfy:
   a) $p_i \subseteq X$, $p_i \mathrel{!}= \emptyset$
   b) $X = \bigcup p_i$
   c) $p_i \cap p_j = \emptyset \quad i \mathrel{!}= j$

Clustering Approach:
Hierarchical clustering

- Same data but ...
  - there are a **tree like structure** to decide how to perform the aggregation.
- This involve more information than partition clustering
  - because we could split clusters in different types.
- The final clustering are perform by
  - the leaves of the tree structure.

Concepts in Clustering

- First we should define a
  - distance (similarity measure) between points
- A good clustering is one where:
  - high intra-cluster similarity
    - the sum of distances between objects in the same cluster are minimized,
  - low inter-cluster similarity
    - while the distances between different clusters are maximized
- Clusters can be evaluated using:
  - "internal measures" that are related to the inter/intra cluster distance
  - "external measures" that are related to how representative are the current clusters to "true" classes. This is typically highly subjective.
- ***Centroid*** – centre of the cluster (i.e. mean point of the cluster)
- ***Medoid*** – the most centrally located (or most representative) object in a cluster

Distance measures – Examples

- Euclidean distance $$d(x_i, x_j) = \sqrt{\sum_{f=1}^{nof} (x_{i,f} - x_{j,f})^2}$$
- Minkowski distance $$d(x_i, x_j) = \sqrt[p]{\sum_{f=1}^{nof} (x_{i,f} - x_{j,f})^p}$$
- Manhattan distance $$d(x_i, x_j) = \sum_{f=1}^{nof} | x_{i,f} - x_{j,f} |$$

where **nof** is the **number of features**

Inter/Intra Cluster Similarity Measures

Intra-cluster similarity measure

- Sum/Min/Max/Avg of the absolute/squared distances between:
  - All pairs of points in the cluster OR
  - Between the "centroid" and all points in the cluster OR
  - Between the "medoid" and all points in the cluster

Inter-cluster similarity measure

- Sum the (squared) distance between all pairs of clusters, where the distance between two clusters is defined as:
  - distance between their centroids/medoids
    (i.e. spherical clusters)
  - distance between the closest pair of points belonging to the clusters
    (i.e., chain shaped clusters)

Is clustering a difficult problem?

Given n points, and say we would like to cluster them into k clusters

- How many possible cluster???

Answer: Too many

→ exponential in terms of n and k

→ **we can not test exhaustively all possibilities**.

Solution: Iterative optimization algorithms

→ **Start with an initial clustering and iteratively improve it**

→ (see K-means)

Classical clustering methods

- **Partitioning methods**
  - Construct various partitions and then evaluate them by some criterion
  - k-Means (and EM), k-Medoids
- **Hierarchical methods**
  - create a hierarchical decomposition of the set of objects using some criterion
  - agglomerative, divisive
- **Model-based clustering methods**
  - a model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

Non-Classical clustering methods

- Fuzzy clustering algorithms
  - **Fuzzy C-means** is the most popular one
- Neural networks have been used for clustering
  - Self-Organizing Maps (SOMs – Kohonen, 1984)
  - Adaptive Resonance Theory (ART) networks (Carpenter & Grossberg, 1990), unsupervised architecture.
- Evolutionary algorithms based clustering
- Simulated annealing based clustering