Data Mining and Machine Learning

Lecture 1

Dr. Edgar Acuña
Department of Mathematical Sciences
University of Puerto Rico - Mayaguez
http://academic.uprm.edu/eacuna

E-mail: edgar.acuna@upr.edu , eacunaf@gmail.com

Objectives

- Understand the basic concepts to carry out data mining and knowledge discovery in databases.
- Implement the most well known data mining algorithms on real world datasets

Course Content

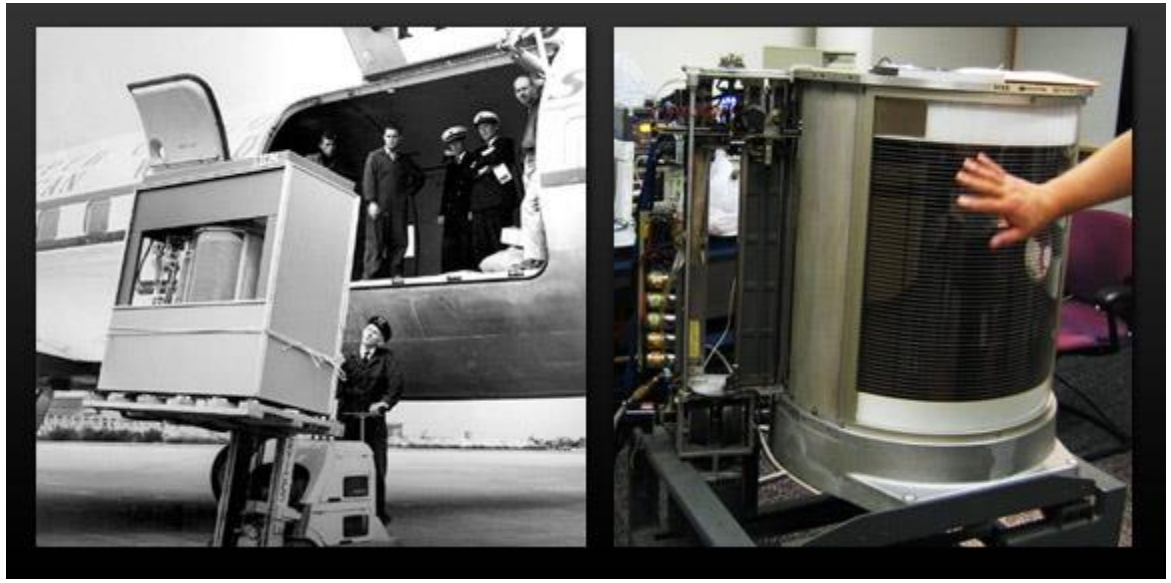
I. Introduction
II. Data Preprocessing
III. Visualization
IV. Supervised Classification
V. Clustering
VI. Novelty Detection (Outliers)
VII. Association Rules
VII. Recommendation System

Introduction

The mechanisms for automatic recollection of data and the development of databases technology has made possible that a large amount of data can be available in databases, data warehouses and other information repositories.

Nowdays, there is the need to convert this data in knowledge and information.

The first hard drive, 1956



IBM 350, had the size of two refrigerators and about 3.75MB of storage. It weighted over a ton. Approximately price of 50,000 dollars. Today, Seagate sells a 2TB hard drive it weights only .33 pounds. It costs around $100.

Size of datasets (in bytes)

Description	Size	Storage Media
Very small	10^2	Piece of paper
Small	10^4	Several sheets of paper
Medium	10^6 (megabyte)	Floppy Disk
Large	10^9(gigabite)	USB/Hard Disk
Massive	10^{12}(Terabyte)	Hard disk/USB
Super-massive	10^{15}(Petabyte)	File of distributed data
Exabyte(10^{18}), Zettabytes(10^{21}), Yottabytes(10^{24})		

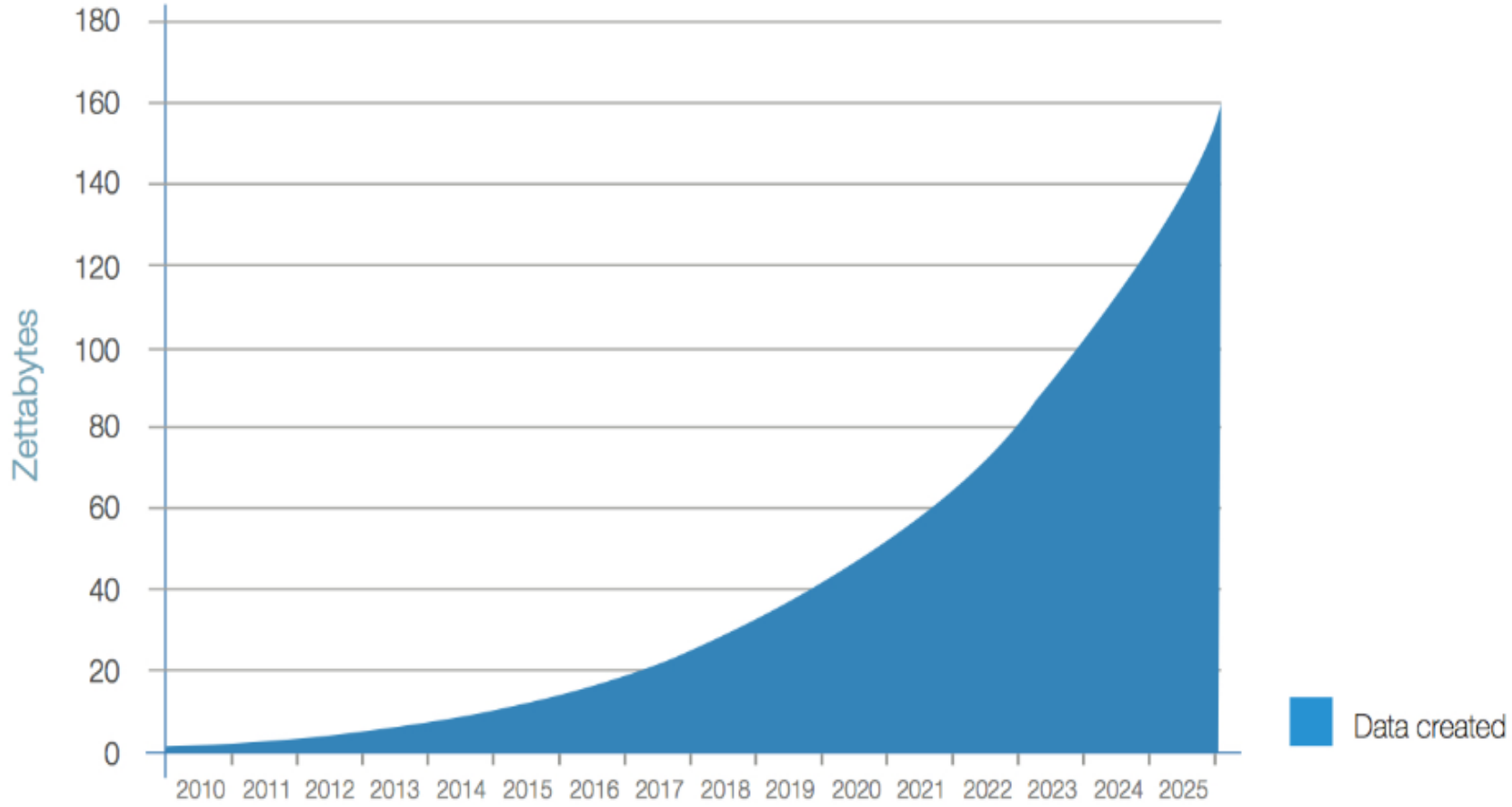
Data Age 2025:

The Evolution of Data to Life-Critical

Don't Focus on Big Data; Focus on the Data That's Big

David Reinsel John Gantz John Rydning | April 2017

An IDC White Paper, Sponsored by 


180
160
140
120
100
80
60
40
20
0
Zettabytes
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025
Data created

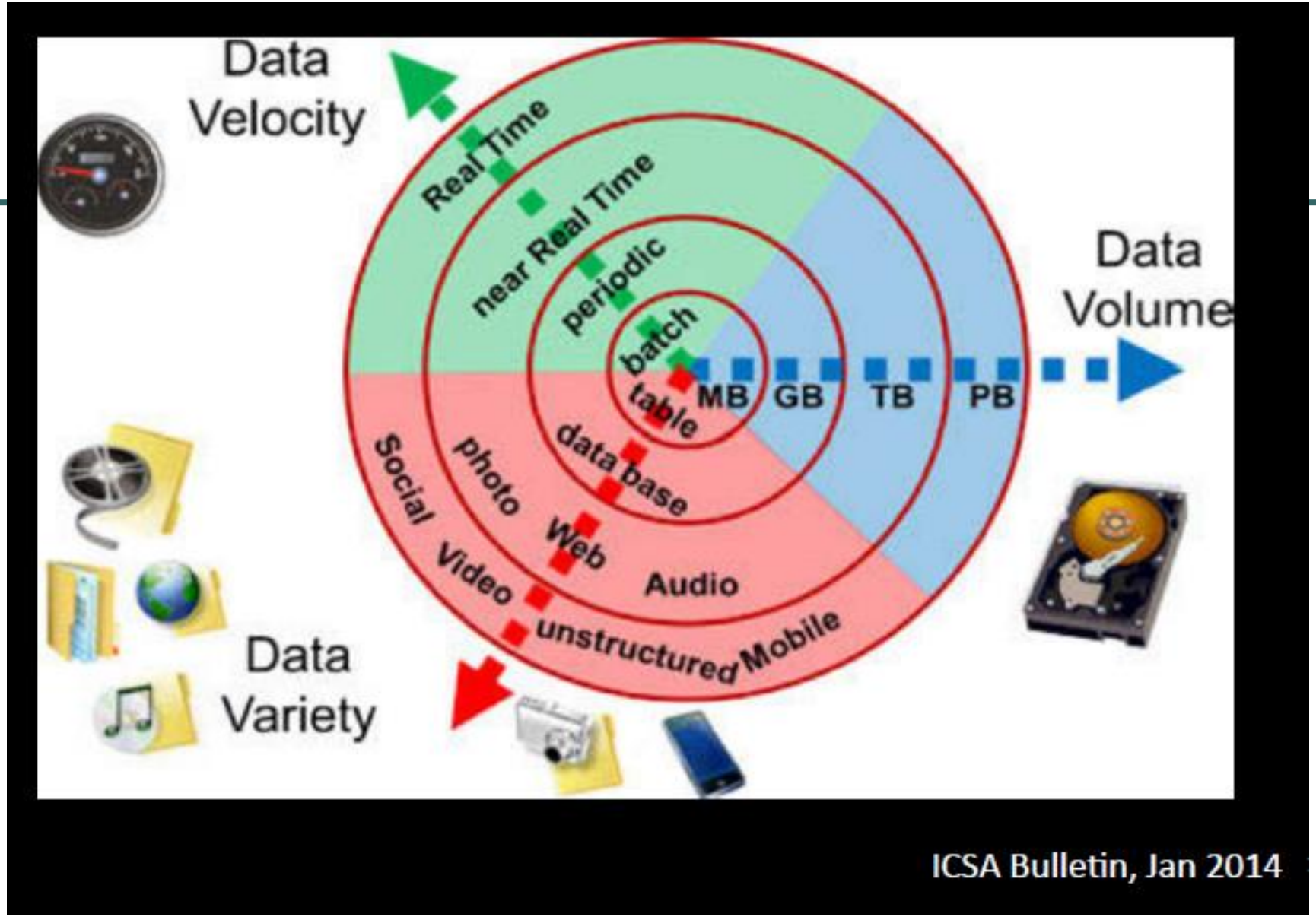
Examples of very large datasets

- Until 2016, the flight delay dataset was approximately 25 Gigabytes
- Amazon.com 45 TB of information from 60 million customers
- In 2010, the ATT call database was 323 Terabytes
- Until 2016, Google searches in more than 130 trillion pages, which represents more than 390 Petabytes
- The Large Hadron Collider (LCH) telescope stores about 600 Petabytes of sensor data per year.
- El 2010, Walmart handled 2.5 Petabytes of transactions.
- In 2014, it was built the NSA's Data Center at Utah. It storages up to 5 zettabytes (5,000 exabytes).

Data Revolution

Year	Digital	Analog	Amount
2000	25%	75%	2 Exabytes
2007	93%	7%	300 Exaby
2013	98%	2%	1,200 Exaby

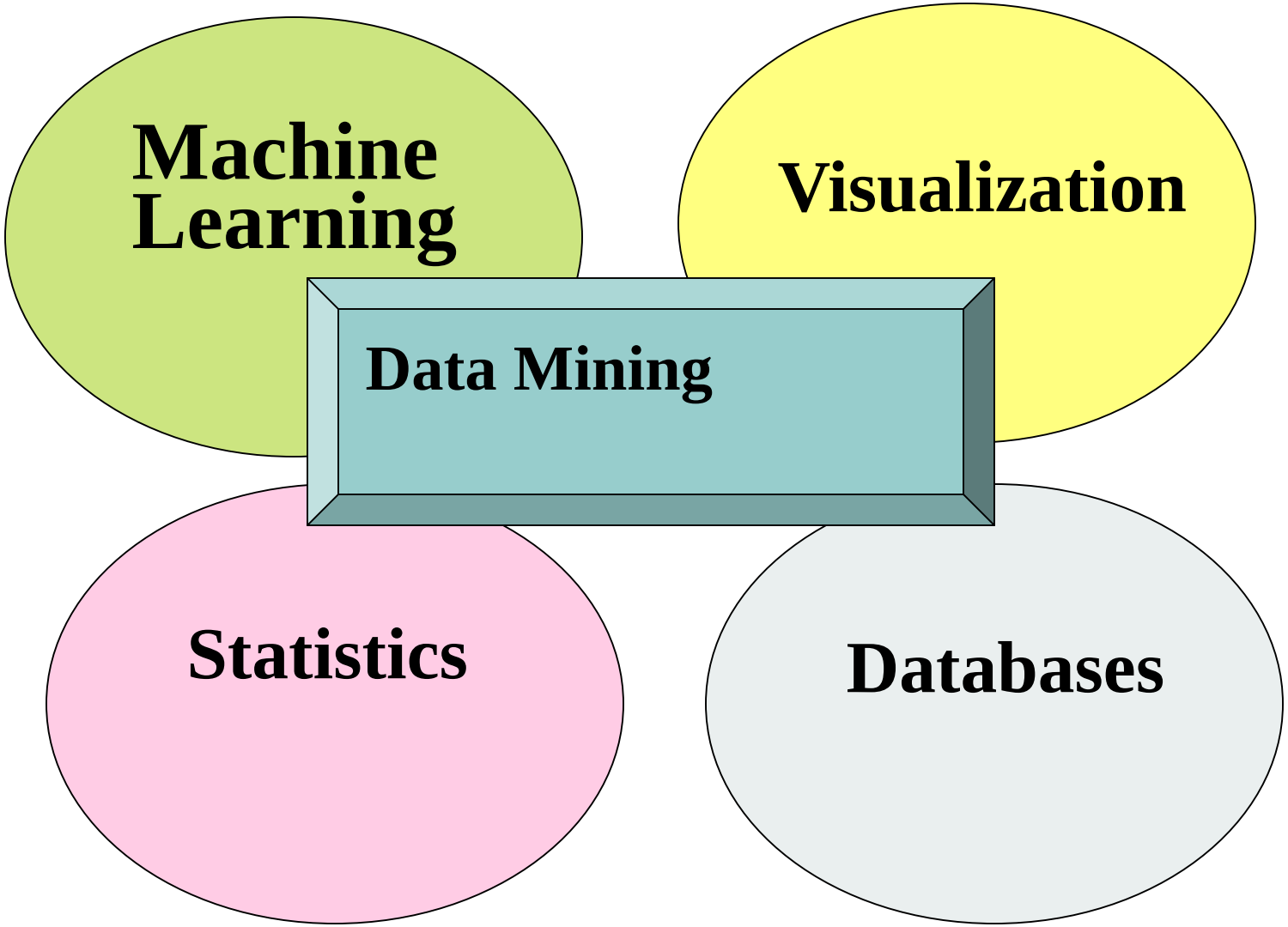
Source: Viktor Mayer-Schönberger and Kenneth Cukier: Big Data: A Revolution that will Transform how We Live, Work and Think 2013)


Data Velocity
Real Time
near Real Time
periodic
batch
table
data base
photo
Web
Video
Social
unstructured
Audio
Mobile
Data Volume
MB
GB
TB
PB
Data Variety
ICSA Bulletin, Jan 2014

What is Data Mining?

- It is the process of extracting valid knowledge/information from a very large dataset. The knowledge is given as patterns and rules that are non-trivial, previously unknown, understandable and with a high potential to be useful.
- Other names: Knowledge discovery in databases (KDD), Intelligent Data Analysis, Business Intelligence.
- The first paper in Data Mining: Agrawal et al. Mining Association rules, ACM SIGMOD 1993.

Areas related to Data Mining



Statistics, Machine Learning

- Statistics (~30% de DM)
  - Based on theory. Assume distributional properties of the features being considered.
  - Focused in testing of hypothesis, parameter estimation and model estimation (learning process).
  - Efficient strategies for data recollection are considered.
  - Model estimation
- Machine learning (~30 % de DM)
  - Part of Artificial Intelligence. Machine Learning is equivalent to a statiscal model.
  - More heuristic than Statistics
  - Focused in improvement of the performance of a classifier based on prior experiences
  - It also considers the length of the learning process.
  - Includes: Neural Networks, decision trees , Genetic algorithms.

Visualization, databases

- Relational Databases (~20% de DM)
  - A relational database is a set of tables containing data of a predetermined category. Each table contains one or more columns which represents some attributes. Each row of the table contains information of the categories defined in the columns.
  - Introduced by E. F. Codd, IBM in 1970.
  - The most used interface between the user and the relational database is SQL(structured query language).
  - A relational database can be easily enlarged.
- Visualization (~10 % de DM)
  - The dataset structure is explored in a visual form.
  - It can be used in either pre or post processing step of the Knowledge discovery process.
- Other Areas (~ 10%): Pattern recognition, expert systems, High Performance Computing.

Software

- **Free:**
- R (cran.r-project.org). Statistical oriented (48.5% users, May 2018)
- Python (python.org 65.6% users)
- Rapidminer (rapidminer.com). (52.7% users)
- **Comercials:** Microsoft SQL (39.6%), (Excel (39.1%), KNIME (12.3%) , SAS Enterprise Miner (4.3%), IBM Watson(3.1%).

What Data Mining is not…

- Search for a number in a phone book
- Look for a definition in Google
- Generate salary histograms by age groups
- Make a query in SQL and read the response of the query

What Data Mining really is …

- Find groups of people suffering from the same disease
- Determine if a person with certain characteristics could apply to a bank loan.
- Detect intruders (anomalous cases) in a system.
- Determine if a bank customer with certain characteristics could commit fraud.
- Recommend products to a customer, based on its online shopping history.
- Determine the characteristics of a client who leaves the subscription to a service.

Data Mining Applications

Science: Astronomy, Bioinformatics (Genomics, Proteonomics, Metabolomics), drug discovery.

Business: Marketing, credit risk, Security and Fraud detection,

Govermment: detection of tax cheaters, anti-terrorism.

Text Mining:

Discover distinct groups of potential buyers according to a user text based profile. Draw information from different written sources (e-mails).

Web mining: Identifying groups of competitors web pages. Recomemder systems(Netflix, Amazon, Ebay)

Types of tasks in Data Mining

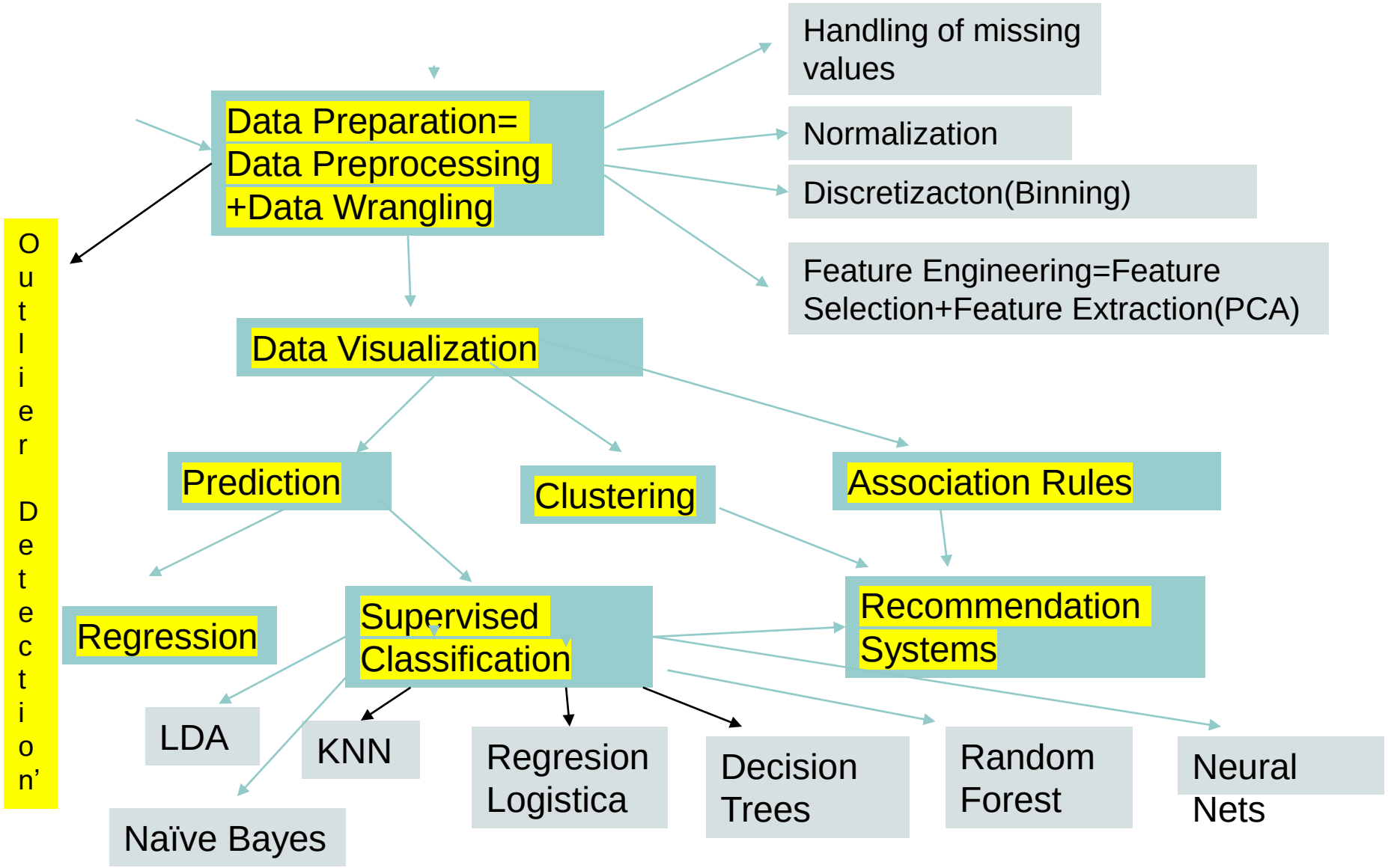
- Descriptive: General properties of the database are determined. The most important features of the databases are discovered.
- Predictive: The collected data is used to train a model for making future predictions. Never is 100% accurate and the most important matter is the performance of the model when is applied to future data.

Tasks in Data Mining

- Regression (Predictive)
- Classification (Predictive)
- Unsupervised Classification – Clustering (descriptive)
- Association Rules (descriptive)
- Outlier Detection (descriptive)
- Visualization (descriptive)
- Recommendation Systems (Predictive)
- Sentiment Analysis (Descriptive/Predictive)

Data Mining /Machine Learning

Flowchart



Regression

- The value of a continuous response variable is predicted based on the values of other variables (predictors), assuming that there is a functional relation among them.
- Statistical models, decision trees, neural networks can be used.
- Examples: car sales of dealers based on the experience of the sellers, advertisement, type of cars, etc.

Regression[2]

- Linear Regression $Y=b_o+b_1X_1+\ldots..b_pX_p$
- Non-Linear Regression $Y=g(X_1,\ldots,X_p)$, where g is a non-linear function. For example, $g(X_1,\ldots X_p)=X_1\ldots X_p e^{X1+\ldots Xp}$
- Non-Parametric Regression $Y=g(X_1,\ldots,X_p)$, where g is estimated using the available data.

Supervised Classification

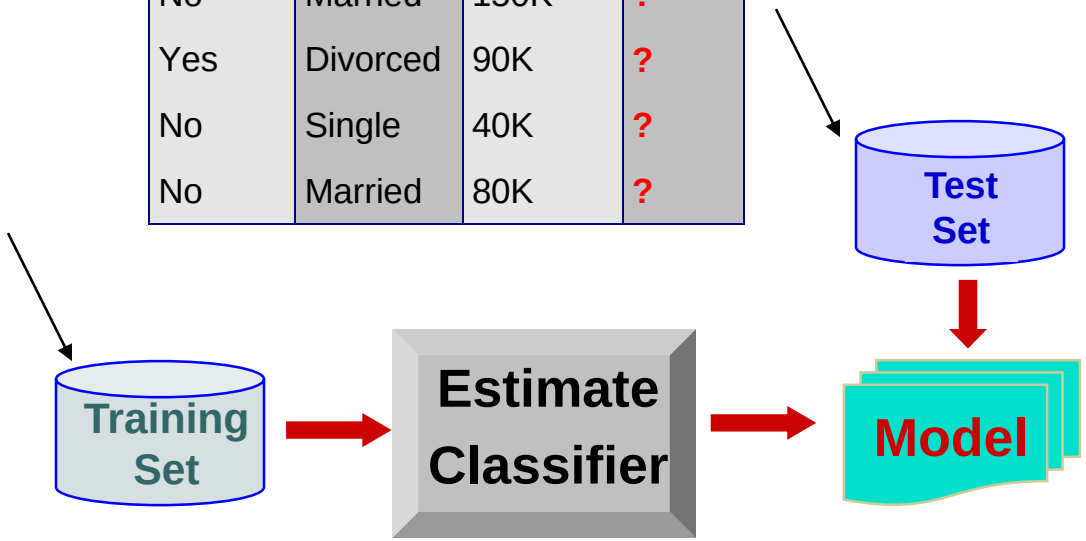
- Given a set of records, called the training set (each record contains a set of attributes and usually the last one is the class), a model for the attribute class as a function of the others attributes is constructed. The model is called the classifier.
- Goal: Assign records previously unseen (test set) to a class as accurately as possible
- Usually a given data set is divided in a training set (70%) and a test set (30%). The first data set is used to construct the model and the second one is used to validate. The precision of the model is determined in the test data set.

Classification Example

categorical categorical Continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Supervised Classification[2]

- The Supervised Classification can be considered as a decision process and the decision rule is called a classifier .
- Some Classifiers: Linear Discriminant Analysis (LDA), Logistic Regression, K-Nearest Neighbors, density estimators, Naïve Bayes, Decision Trees, Neural Networks, random forest, support vector machines.

Unsupervised Classification (Clustering)

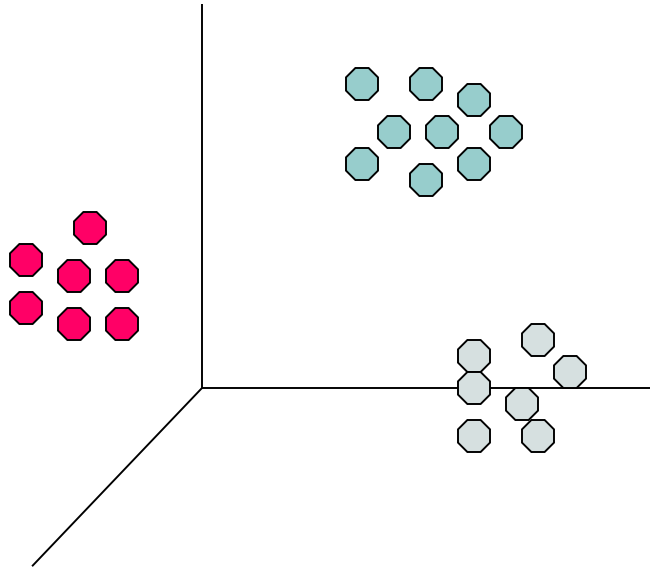
- Find out groups of objects (clusters) such as the objects within the same clustering are quite similar among them whereas objects in distinct groups are not similar.
- A similarity measure is needed to establish whether two objects belong to the same cluster or to distinct cluster.
- Examples of similarity measure: Euclidean distance, Manhattan distance, correlation, Grower distance, hamming distance, etc.
- Problems: Choice of the similarity measure, choice of the number of clusters, cluster validation.

Clustering[2]

- Tri-dimensional clustering based on Euclidean distance

The Intracluster distances are minimized

The Intercluster distances are maximized



Outlier Detection

- The objects that behave different or that are inconsistent with the majority of the data are called outliers.
- Outliers can be affected by a measurement or execution error . They can represent some kind of fraudulent activity.
- The goal of outlier detection is to find out the instances that do not have a normal behavior.

Outlier Detection [2]

- Application: Credit card fraud detection, Network intrusion

Association Rules

- Given a set of records each of which contain some number of items from a given collection. The goal is to find out 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

Rules Discoverd:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rules[2]

- The rules (X->Y) must satisfy a minimum support and confidence set up by the user. X is called the antecedent and Y is called the consequent.
- Support=(# records containing X and Y)/(# records)
- Confidence=(# records containing X and Y/(# de records containing X)

Example: The first rule has support .6 and the second rule has support .4

The confidence of rule 1 is .75 and for the rule 2 is .67

Applications: Marketing and sales promotion.

Recommendation Systems

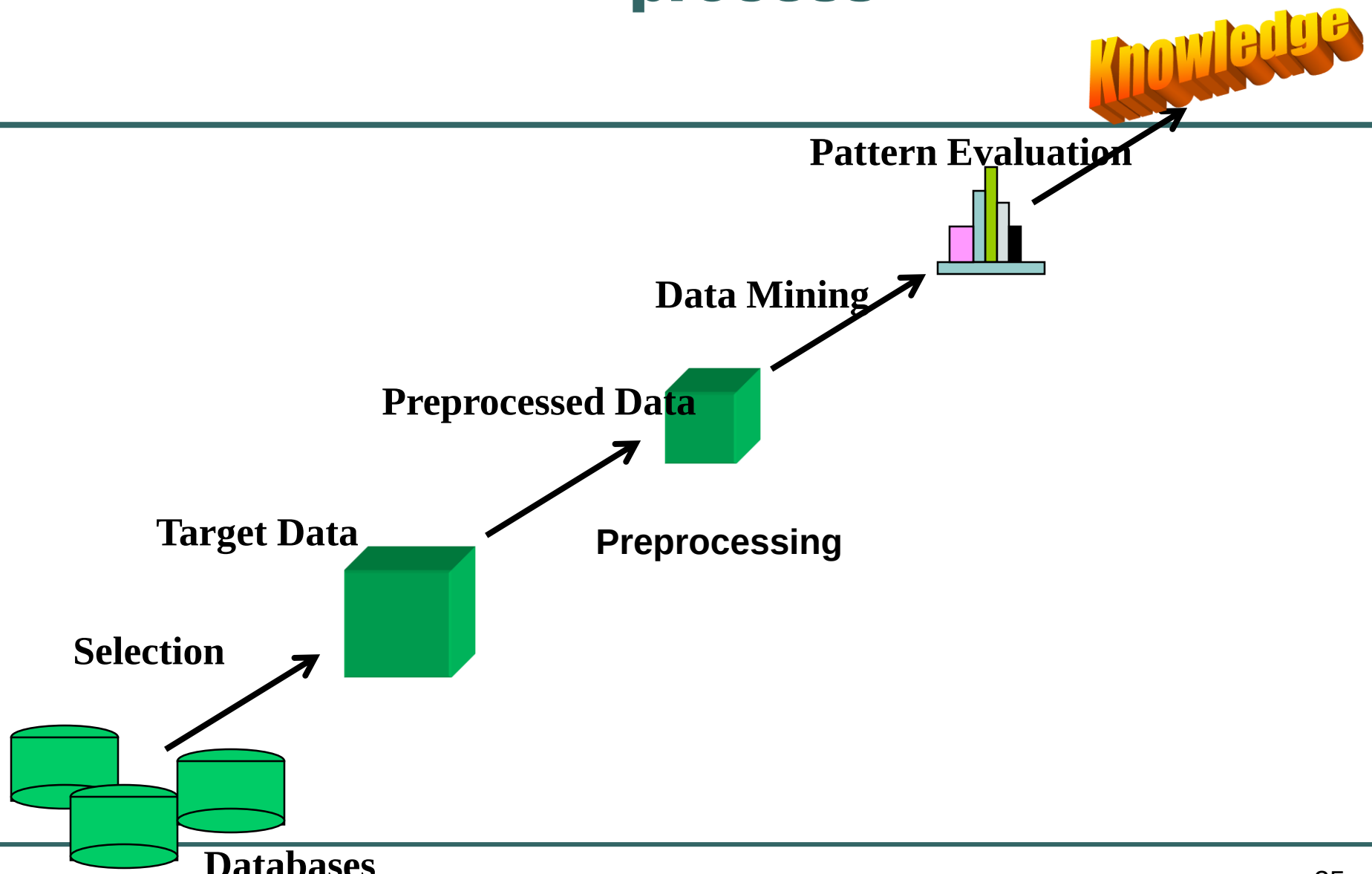
Based on Popularity: Items most viewed/bought are recommended to users.

Based on content: Recommend items similar to those liked by the user in the past.

Based on classification Models: Features of the user/item are used to build a model to predict if a new user would buy or not a certain product.

Based on collaborative filtering using either knn or Matrix Factorization.
In the first approach, one needs to find k similar users to a given users and it recommends items that those k neighbors enjoyed.
In the second approach assuming that we know items' ratings given by several users the algorithm recommends a rating to a item that he has not bought yet.

Data Mining as one step of the KDD process



Steps of a KDD Process

- Comprehend the KDD process, it's background and objectives.
- Determine a target data set.
- Data cleaning and pre-processing (it may require between 60-80% of the total process)
- Data reduction and transformation. Identify important variables and reduce dimensionality.
- Choose your task: Summarization, Classification, Regression, Association, Clustering.
- Choose the data mining algorithm to be used.\
- Look for interesting patterns
- Pattern Evaluation and knowledge representation.

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data.
- Quality of Data
- Privacy Data
- Streaming Data

Google Trends for DM/Big Data/ML