Data & Information Systems Management

Business Intelligence, Data Mining & Data Warehousing

LECTURE 4 –

OUTLINE

1. What is Business Intelligence

2. What is a Data warehouse

3. What is Data Mining & why is it important

4. The Knowledge, Data and discovery process

What is Business Intelligence ?

- What is BI?

What is Business Intelligence (BI) ?

▶ "Business intelligence, or BI, is an umbrella term that refers to a variety of software applications used to analyze an organization's raw data. BI as a discipline is made up of several related activities, including data mining, online analytical processing, querying and reporting". (And Data Warehousing) – CIO.COM

What is Business Intelligence (Gartner) ?

"*Business intelligence (BI) is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance*". - Gartner

Who are the market leaders with BI tools ? 2022

Figure 1: Magic Quadrant for Analytics and Business Intelligence Platforms

Who are the market leaders with BI tools ? 2023

7 Terms used in Business Intelligence ?

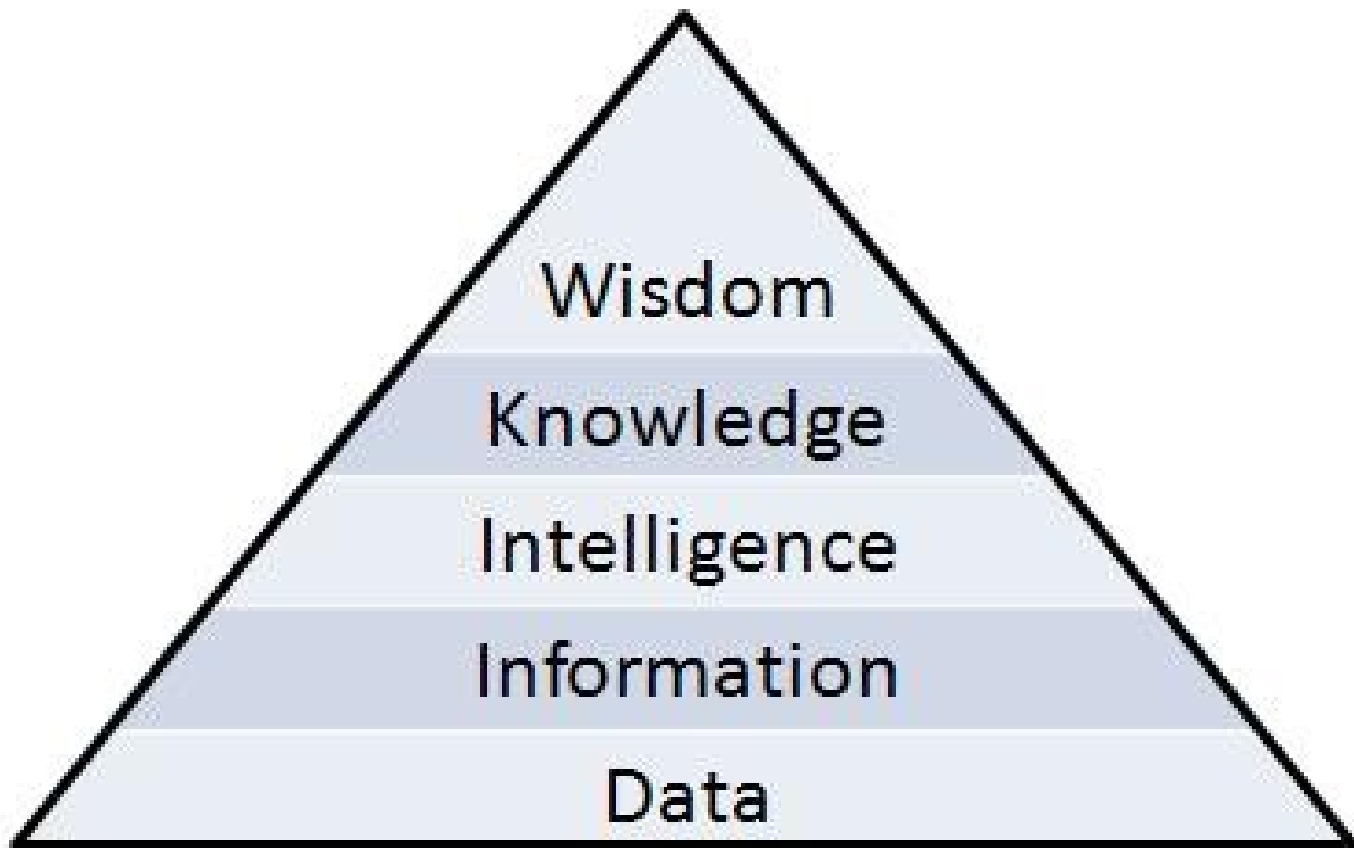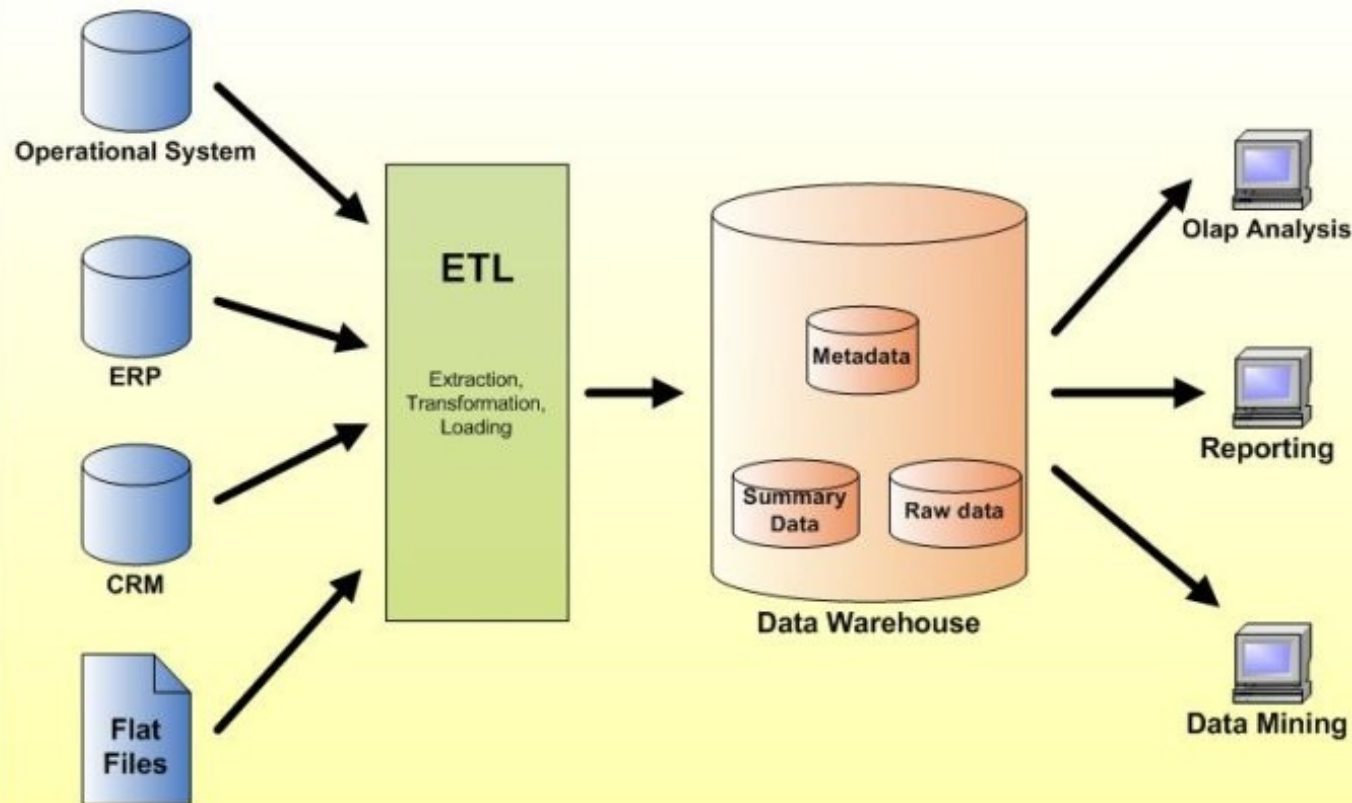- [7 Terms used in Business Intelligence](#)

Data warehousing ?

- [Data Warehousing - An Overview](#)

The Information hierarchy ?

Data Warehouse architecture ?

OLAP is an acronym for Online Analytical Processing.
OLAP performs multidimensional analysis of business data and provides the capability for complex calculations, trend analysis, and sophisticated data

Is Data Mining important?

"The process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques". Gartner

Data Mining and Big Data

Data Mining is <u>NOT</u> about Big Data, but it is part of Big Data, Big Data is a marketing term which encompasses a broader range of issues and technologies – note Google & CERN have big data
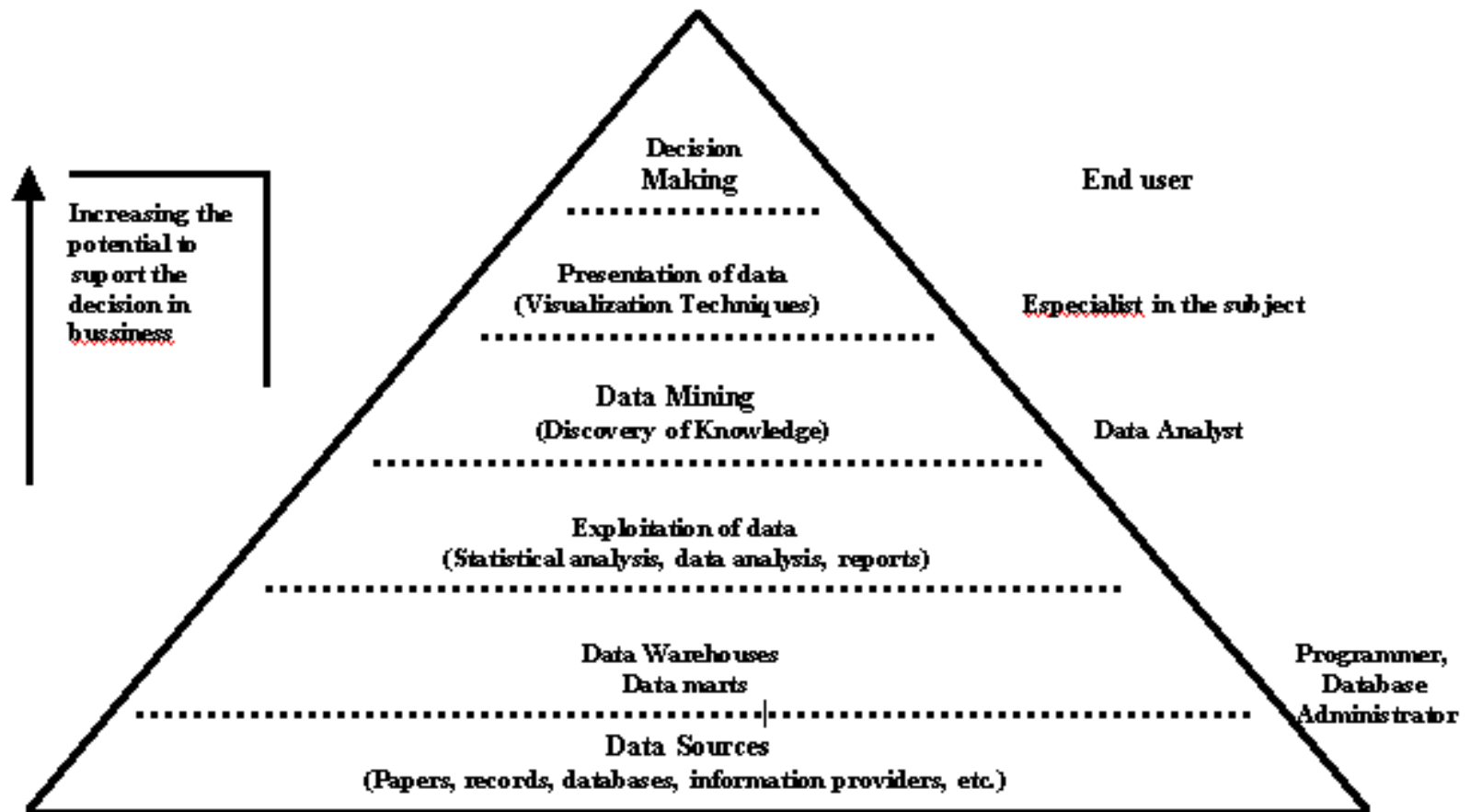
Is Data Mining Important ?

Most in demand tech jobs USA 2023

Data Mining ?

Decision Making — End user

Presentation of data (Visualization Techniques) — Especialist in the subject

Data Mining (Discovery of Knowledge) — Data Analyst

Exploitation of data (Statistical analysis, data analysis, reports)

Data Warehouses Data marts — Programmer, Database Administrator

Data Sources (Papers, records, databases, information providers, etc.)

Increasing the potential to suport the decision in bussiness
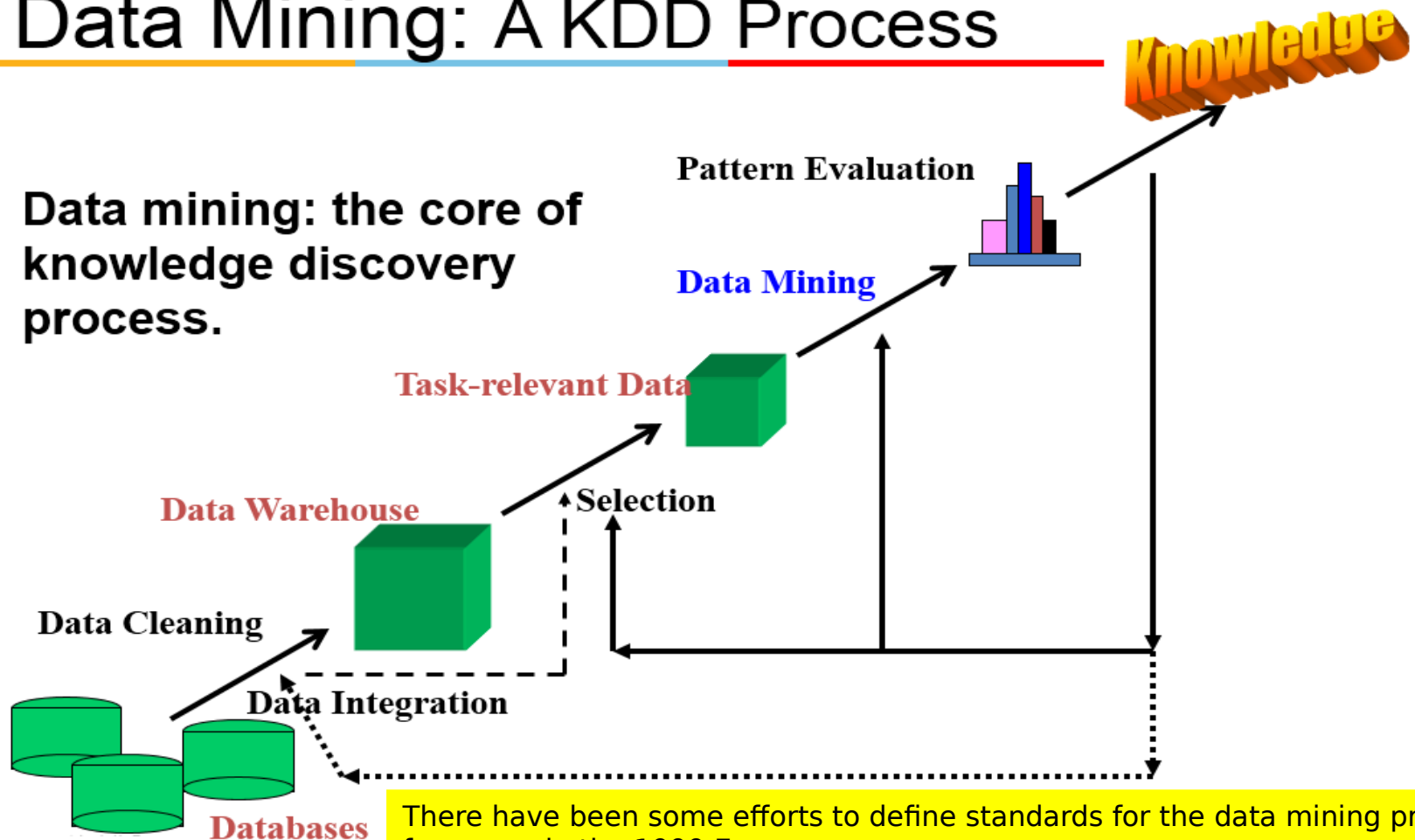
What data is being captured ?

- Class/group discussion – select some industries

- In Groups identify:-

  - What data is being captured now
  - What data will be captured in the future
  - Where can profit be made from Data/Information/Knowledge

Knowledge Discovery in Databases (KDD)

Data Mining: A KDD Process

Knowledge

Data mining: the core of knowledge discovery process.

Pattern Evaluation

Data Mining

Task-relevant Data

Data Warehouse

Selection

Data Cleaning

Data Integration

Databases

There have been some efforts to define standards for the data mining process, for example the 1999 European Cross Industry Standard Process for Data Mining (CRISP)

CRISP(CRoss Industry Standard Process for Data Mining (CRISP-DM))
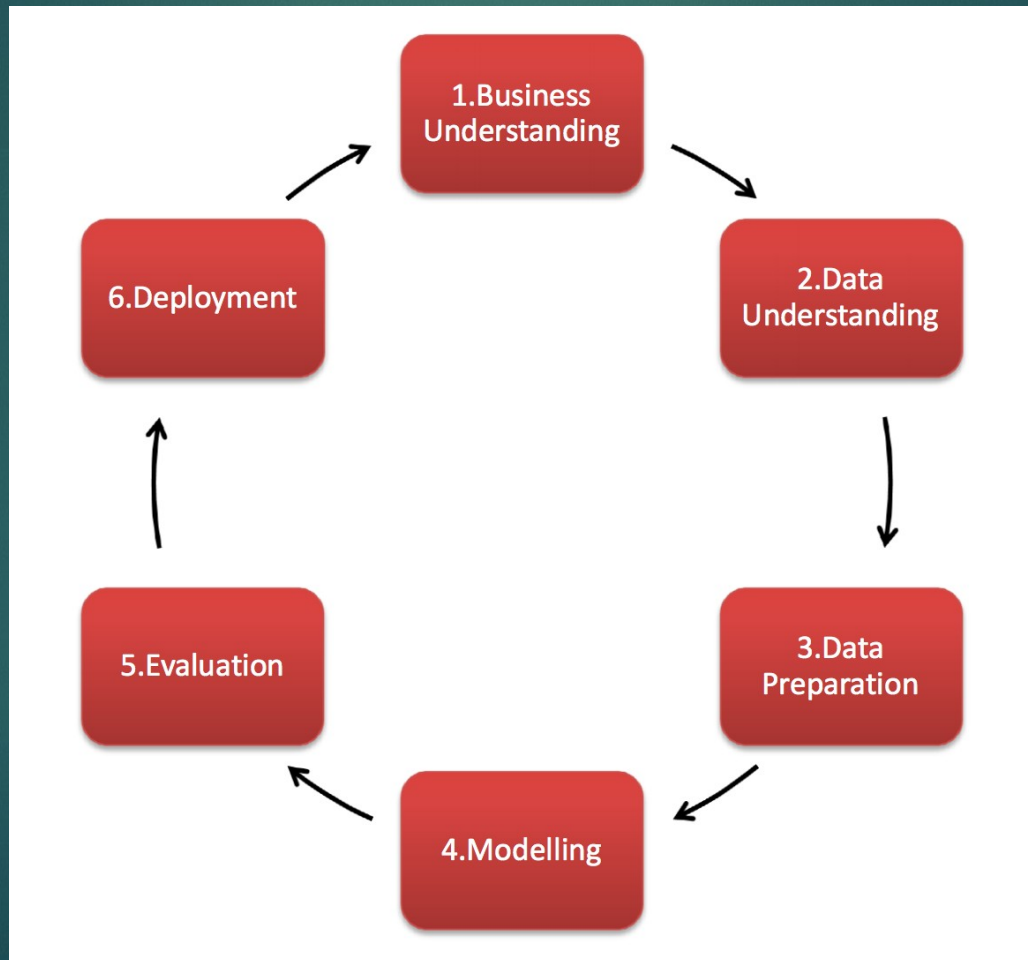
The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) is a process model that serves as the base for a data science process. It has six sequential phases:

1.Business understanding – What does the business need?
2.Data understanding – What data do we have / need? Is it clean?
3.Data preparation – How do we organize the data for modeling?
4.Modeling – What modeling techniques should we apply?
5.Evaluation – Which model best meets the business objectives?
6.Deployment – How do stakeholders access the results?

Published in 1999 to standardize data mining processes across industries, it has since become the most common methodology for data mining, analytics, and data science projects.
Data science teams that combine a loose implementation of CRISP-DM with overarching team-based agile project management approaches will likely see the best results.

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

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

KDnuggets Polls

What main methodology are you using for data mining?
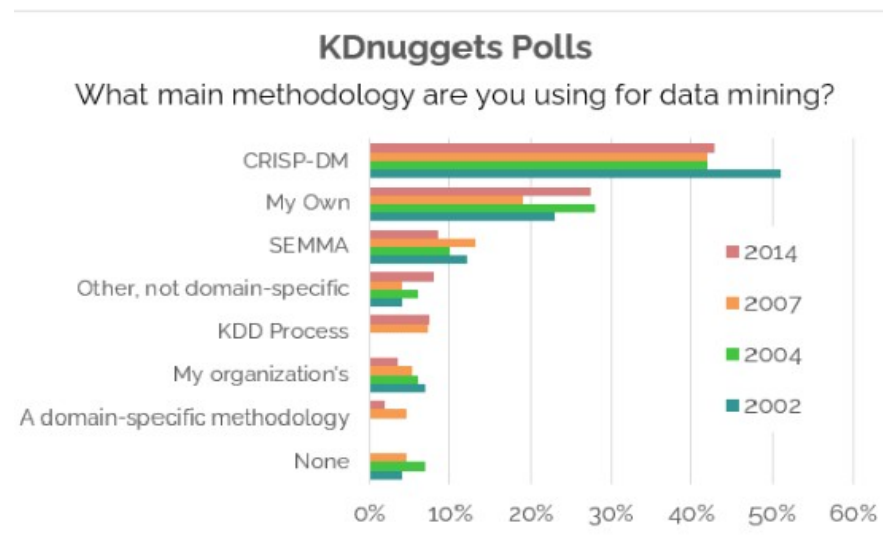
CRISP-DM was the popular methodology in each poll spanning the 12 years.

Data Mining preprocessing

► Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

The six common classes of Data Mining ?

▶ Data mining involves six common classes of tasks:

1. Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
2. Association rule learning (Dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
3. Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
4. Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
5. Regression – attempts to find a function which models the data with the least error.
6. Summarization – providing a more compact representation of the data set, including visualization and report generation.