



Data & Information Systems Management

Database Mining – Techniques

LECTURE 5 – DATABASE MINING – TECHNIQUES

CHAPTER OUTLINE

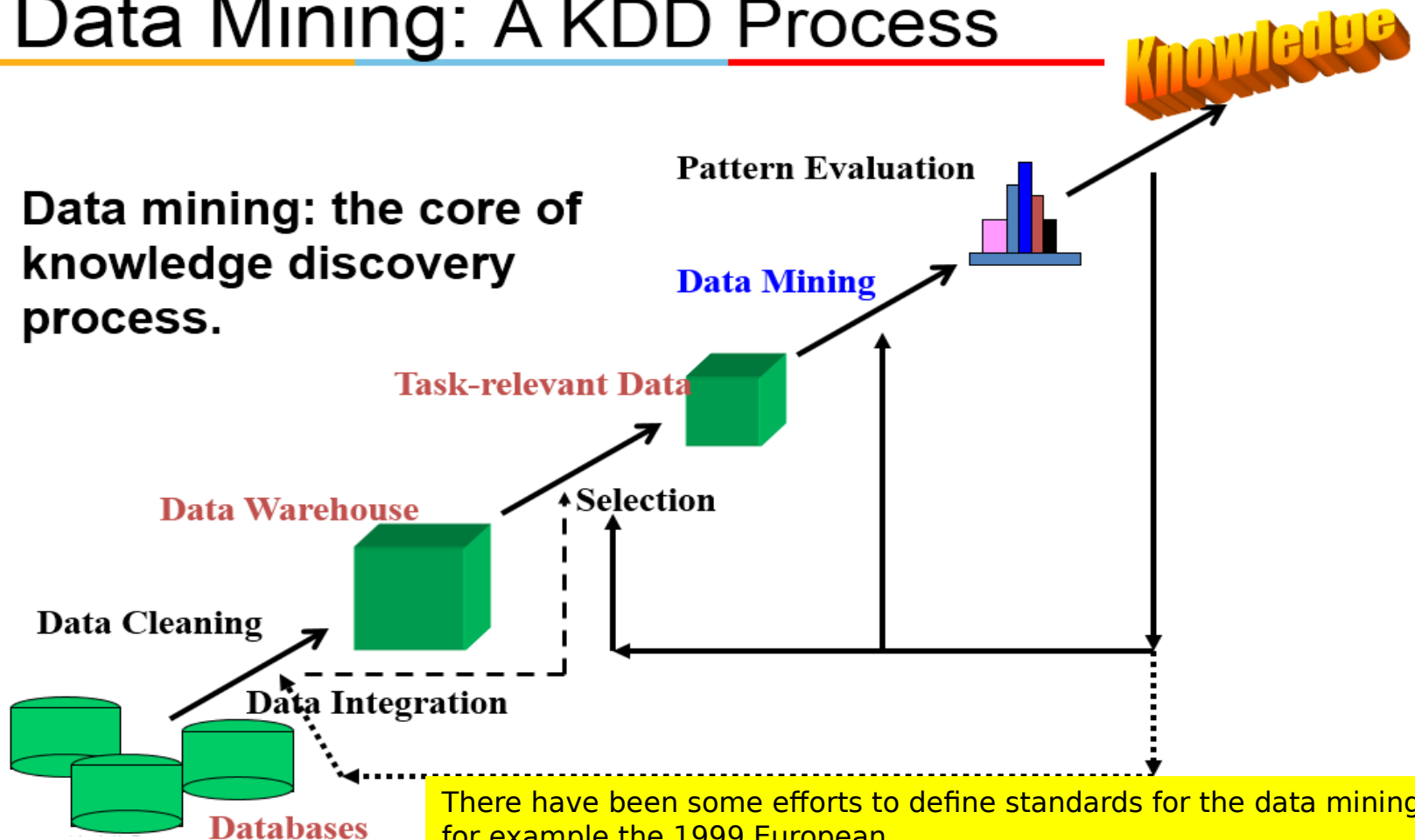
- ▶ Data Preprocessing
- ▶ Data Mining Techniques
- ▶ Visualisation



Knowledge Discovery in Databases (KDD)

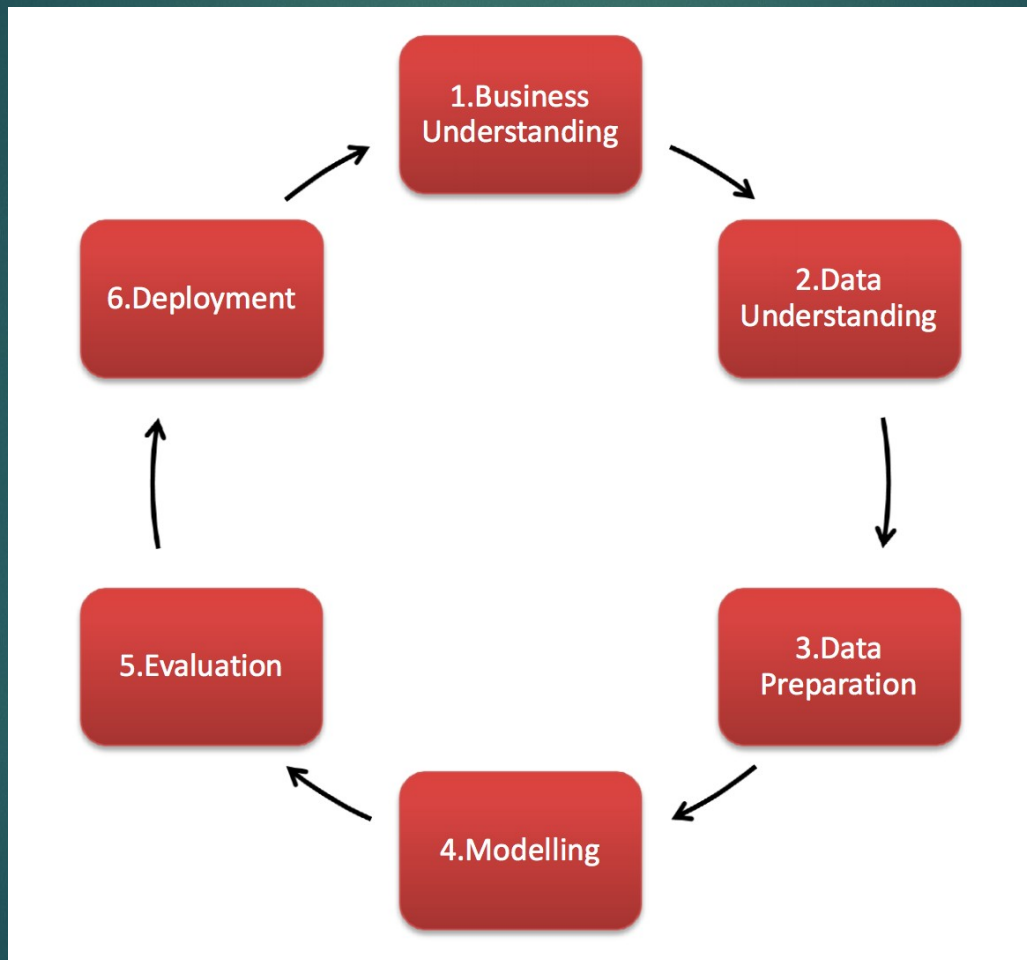
Data Mining: A KDD Process

Data mining: the core of knowledge discovery process.



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-DM 1.0) and the 2004 Java Data Mining standard (JDM 1.0).

CRISP(**C**Ross **I**ndustry **S**tandard **P**rocess for Data Mining (CRISP-DM))



Why is Data Preprocessing important

No Quality data, no quality mining results – quality decisions must be based on quality data e.g., duplicate or missing data may cause incorrect or even misleading statistics

“Data extraction, cleaning and transformation comprises the majority of the work of building a data warehouse” – Bill Inmon



Data Preprocessing

- ▶ **Data Cleaning**

- ▶ Fill in missing values, smooth noisy data, identify or remove outliers and resolve inconsistencies

- ▶ **Data integration**

- ▶ Integration of multiple databases, data cubes or files

- ▶ **Data transformation**

- ▶ Normalization and aggregation

- ▶ **Data reduction**

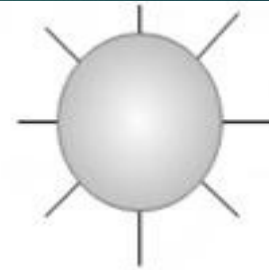
- ▶ Obtains reduced representation in volume but produces the same or similar analytical results

- ▶ **Data discretization**

- ▶ Part of data reduction but with particular importance, especially for numerical data

Data Preprocessing

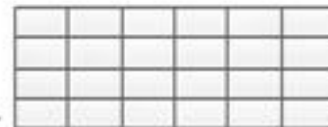
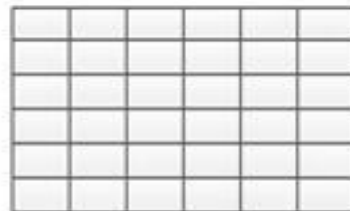
DATA CLEANING



DATA INTEGRATION



DATA REDUCTION



DATA TRANSFORMATION

-2,32,100



-0.02,0.32,1.00

Data mining...

*After we have created
good quality data, we
can then move onto
data mining - different
techniques are
available to perform
successful data
mining.....*



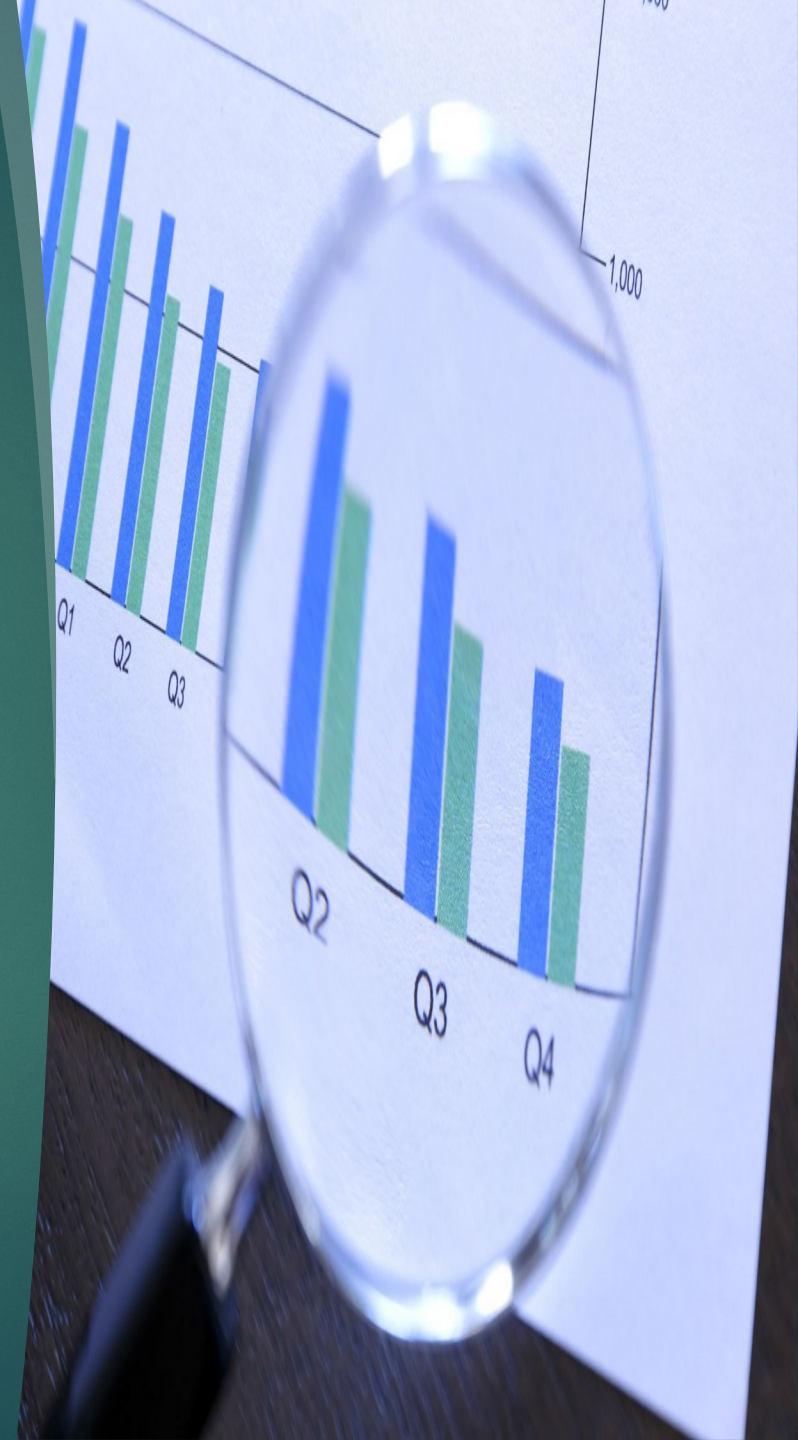
Types of Information

There are two types of information

Structured: Information which comes from transactional and database data.

Unstructured: Information which comes from files and repositories containing text, media etc..

The right combination of information, effectively exploited with data mining and predictive analytics technologies, can create analytic and predictive models of the patterns that impact business operations and can also be used to evaluate future options, risks and decisions. source Gartner.com 2020



What is predictive data mining?

Data mining is characterised by finding the general properties/attributes and relationships within a set of information.


Predictive data mining performs inference on the current information sets to create models to be used to make predictions on future information sets.

Prescriptive analytics

Prescriptive analytics intends to calculate the best way to achieve or influence the outcome — it aims to drive action. When combined with predictive analytics, prescriptive analytics naturally draws on and extends predictive insights, addressing the questions of, what should be done? or what can we do to make a given outcome happen?

Prescriptive analytics includes both rule-based approaches (incorporating known knowledge in a structured manner) and optimization techniques (traditionally used by operations research groups) that look for optimal outcomes within constraints to generate executable plans of action. Prescriptive analytics relies on techniques such as graph analysis, simulation, complex-event processing and recommendation engines

What is predictive data mining



Predictive data mining technologies provide two very important pattern-seeking functions:

Automated discovery of relationships. Data mining tools can search through databases and identify previously hidden patterns. Insight from discovering relationships using characterization, segmentation, comparisons and discrimination techniques leads to developing descriptive models. e.g., patterns in current retail sales, detecting fraud

Automated prediction of trends and behaviours. These techniques are rule and model driven and are based on classification and predictive modelling techniques. e.g., Looking at future marketing trends in data, predicting fraud

Techniques for Discovering Patterns and Relationships

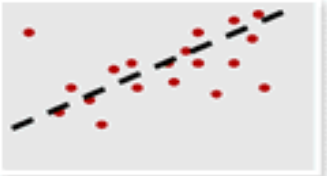




Statistics, while not exactly a part of data mining, can be essential to discovering patterns and relationships, as well as testing the validity of data mining and predictive analytic models. *Descriptive statistics* consists of organizing and summarizing information. *Inferential statistics* consists of algorithms that generalize results obtained from a sample of information to infer patterns that exist within a larger set of information (and from which the sample is obtained) and measures the reliability of the results of the inferences.

Clustering techniques are some of the most often used techniques to "make sense out of data." They group data with low "semantic" distance into the same cluster and try to maximize the semantic distance between each of the clusters. This reveals interesting views on the distribution of data, allows organizations to construct prototypes or stereotypes, and is also useful in identifying outliers (for example, insurance fraud).

Association rules provide an efficient way to find N-way correlations within large datasets. Associated items are grouped into item sets, and rules can be generated that can be used for prediction.

Techniques for Discovering Patterns and Relationships ?

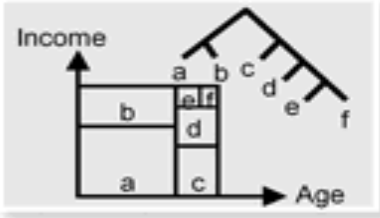
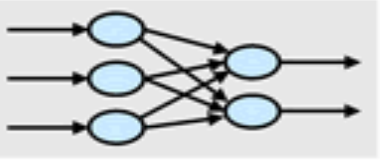
Method	Description	How Used	Comment
Statistics:  <p>Spending</p> <p>Distance</p>	<ul style="list-style-type: none"> Averages Standard deviations Central tendency Histograms Regression 	<ul style="list-style-type: none"> Scoring and segmentation Hypothesis testing Forecasting 	<ul style="list-style-type: none"> Widely used and understood Issues of interplay between variance and sample size
Clustering:  <p>Age</p> <p>Disposable Income</p>	<ul style="list-style-type: none"> Locality and connectivity within an n-dimensional space Unsupervised learning technique 	<ul style="list-style-type: none"> Segmentation: visualization 	<ul style="list-style-type: none"> Wide range of algorithms Computation intensive
Association Rule:  <p>Beer → Chips (80%)</p>	<ul style="list-style-type: none"> Identifying frequent item sets Mining association rules (correlations) Unsupervised learning technique 	<ul style="list-style-type: none"> Shopping basket analysis Cross-selling 	<ul style="list-style-type: none"> Effective but limited application

Decision trees are a form of classification shown in a tree-like structure, in which a disposable income is less than \$20,000 per year, *then* best-selling product = C.) Decision trees are useful in, for example, analyzing and treating different segments regarding their profitability by examining drivers of unprofitability, or realigning marketing or point-of-sale efforts.

Neural networks : Each input into a "neuron" has its own weight associated. A weight is a floating-point number and is adjusted in training the network. The weights in most neural nets can be both negative and positive, therefore providing excitatory or inhibitory influences to each input. As each input enters a neuron, it is multiplied by its weight. The neuron then sums all the input values, which gives us the *activation* (a floating-point number, which can be negative or positive). If the activation is greater than a threshold value, the neuron outputs a signal. If the activation is less than one, the neuron outputs zero.

Naive Bayes calculates probabilities for each possible state of the input attribute, given each state of the predictable attribute. The algorithm supports only discrete attributes and considers all the input attributes to be independent, given the prediction attribute.

Techniques for Classification and Predictions

Method	Description	How Used	Comment
Decision Trees and Rule Induction 	<ul style="list-style-type: none"> • Pairwise or multiwise splits • Generation of If ... then rules • Supervised learning 	<ul style="list-style-type: none"> • Risk analysis • Churn analysis • Customer profiling • Failure prediction 	<ul style="list-style-type: none"> • Easy to understand • Good for data exploration • Trees can become very complicated
Neural Networks 	<ul style="list-style-type: none"> • Nonlinear regression • Supervised learning 	<ul style="list-style-type: none"> • Forecasting • Scoring • Classification 	<ul style="list-style-type: none"> • Difficult to understand model and results • Significant data preprocessing
Naive Bayes <p> $P(H X) = P(X H) P(H) / P(X)$ Evaluating probability that the hypothesis H holds, given the observed data record X </p>	<ul style="list-style-type: none"> • Posterior probability • Supervised learning 	<ul style="list-style-type: none"> • Classification • Scoring 	<ul style="list-style-type: none"> • High accuracy and speed for large databases • Presumption of conditional independence

Techniques for Classification and Predictions

Information Extraction

- Extract Facts, People, Companies, Events, Relations, Phone Numbers ...

Competitive Intelligence

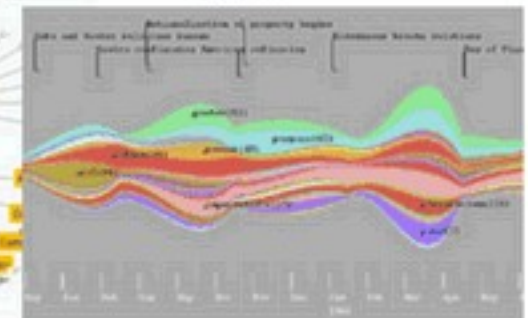
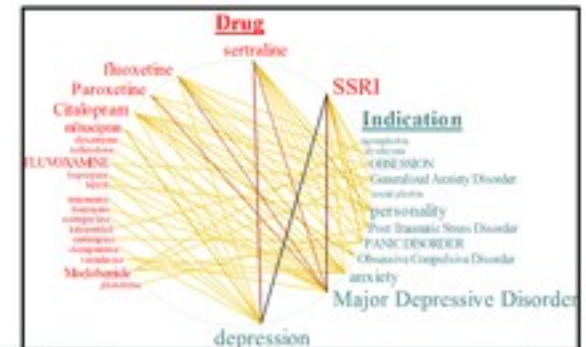
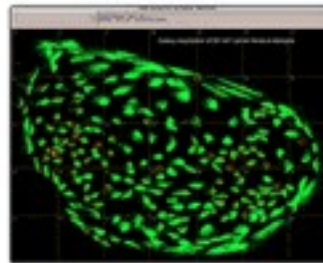


Feedback Analysis



Visualization

- Overviews/Discovery/Relationships Clustering, Trending, Graph Layout



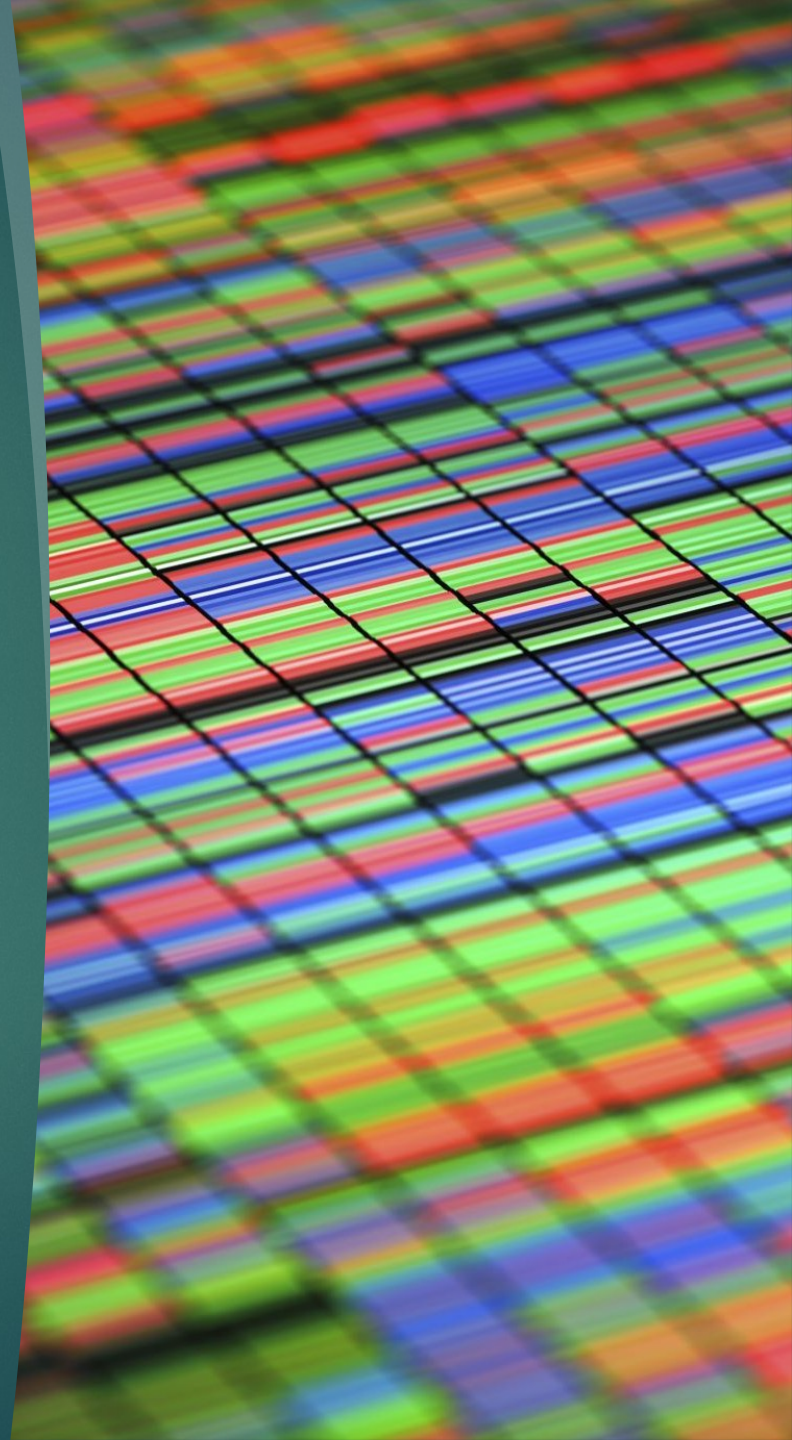
Techniques for Information Extraction and Visualization

Information extraction (IE) attempts to find and isolate facts and findings in unstructured documents. Finding facts about certain named entities (such as "Bush") is called "named entity extraction." A more advanced exercise is the extraction of relationships (the announcement of a joint venture between two companies, or a company's product announcement for a certain target market). IE technology still suffers from a lack of accuracy in many areas (false positive versus false negative classification, both of which relate directly to precision/recall statistics). The applicability of this technology is wide, ranging from shopping agents that parse salient information out of websites and product catalogs (for example, price, delivery conditions, quality, product characteristics), to competitive intelligence (such as parsing news feeds and monitoring certain events), to customer complaint analysis, text analysis in the life sciences domain and medical record analysis. This market space is emerging fast and, at the time of writing, there were more than 20 vendors.

Visualization can be a vital aid in data preparation and analysis. Visualization exploits the added dimension of graphics, as opposed to text or numbers. It can also show values that are out of range or missing values.

Commercial data-mining software & applications

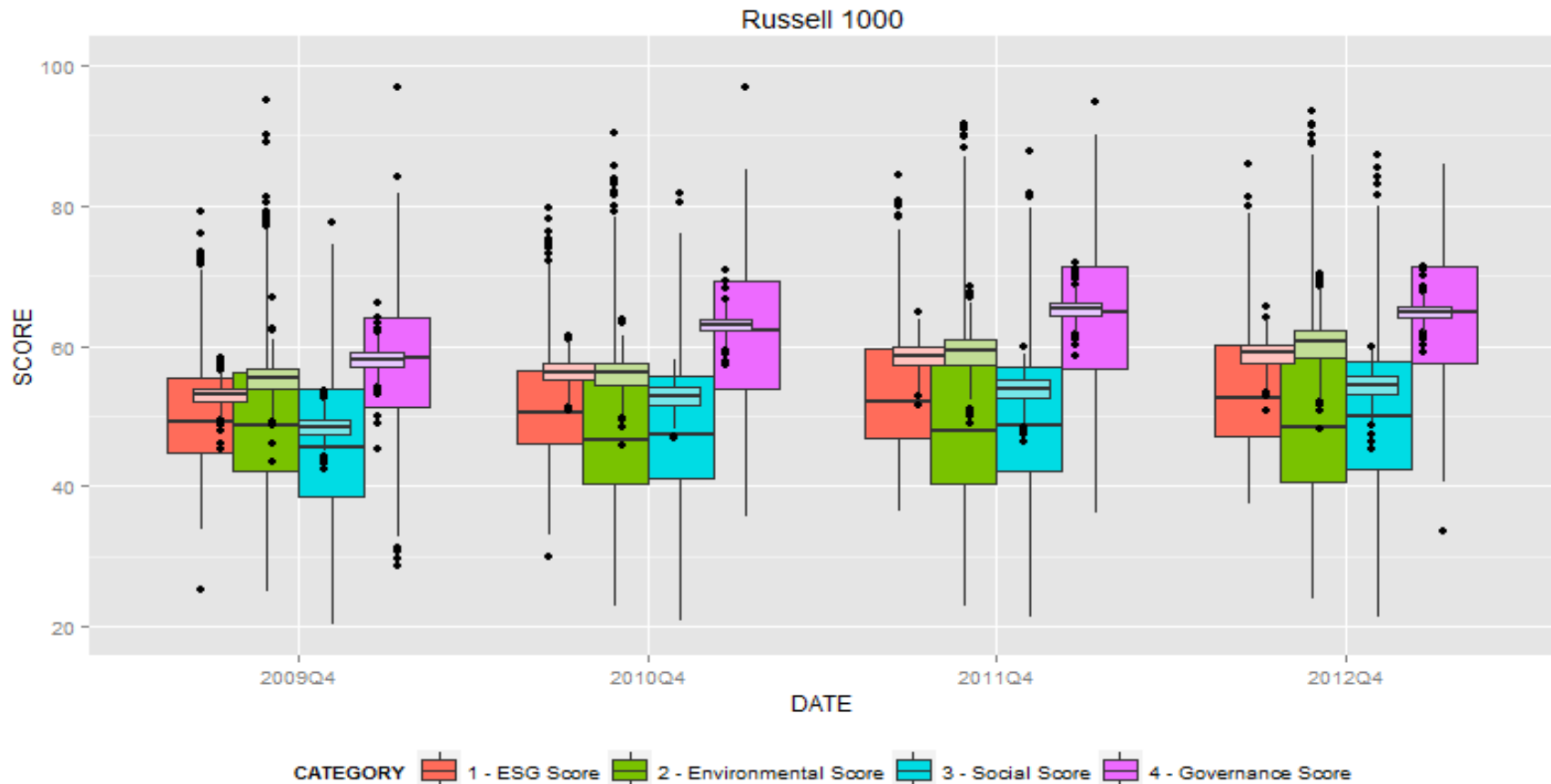
► Top Data mining software 2023





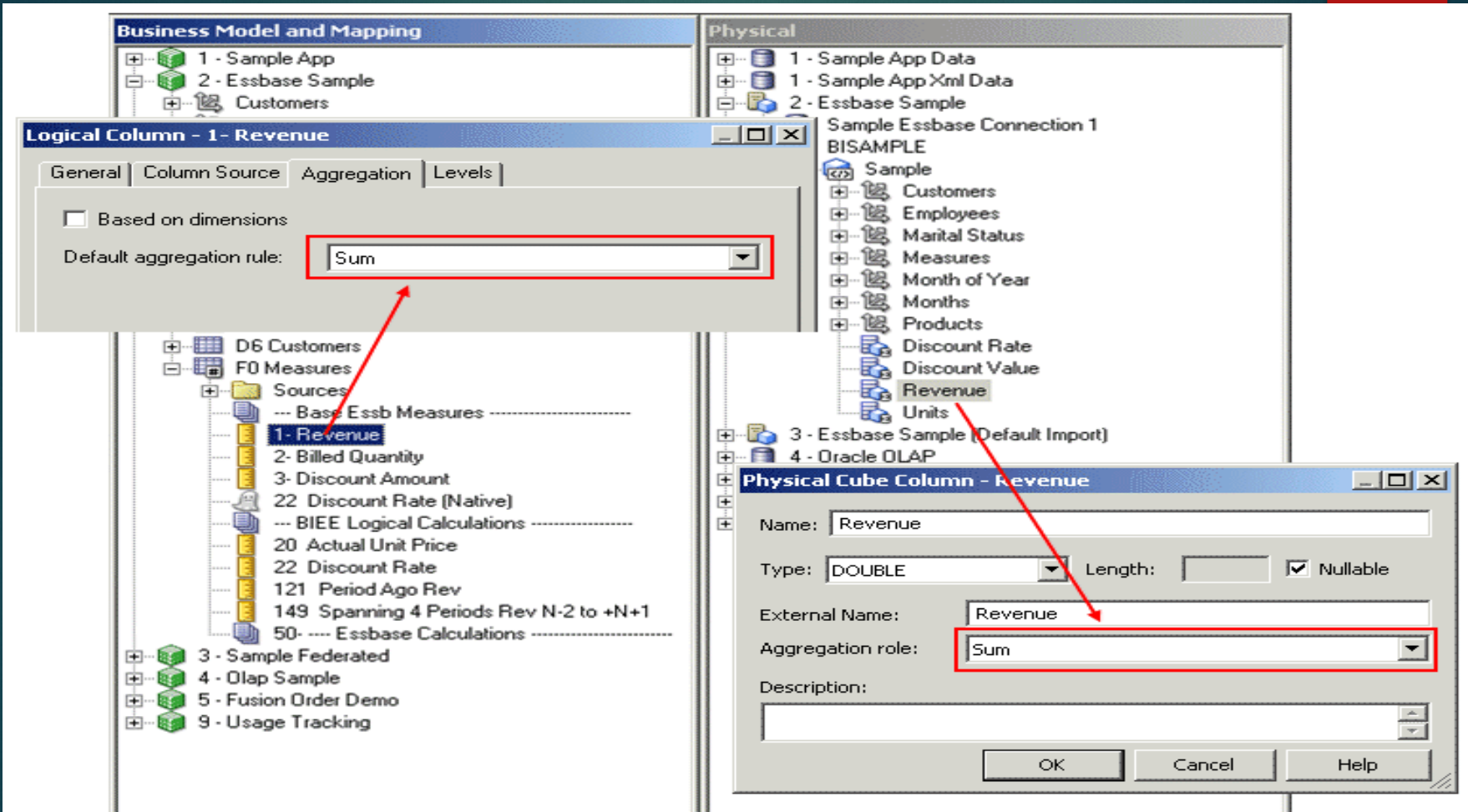
- ▶ **Visualisation:** The use of computer graphics to create visual images which aid in the understanding of complex, often massive representations of data.
- ▶ **Purpose of Visualisation:**
 - ▶ Gain an insight into an information space by mapping data into graphical primitives
 - ▶ Provide qualitative overview of large data sets
 - ▶ Search for patterns, trends, structure, irregularities, relationships among data.
 - ▶ Help find interesting regions and suitable parameters for further quantitative analysis
 - ▶ Provide a visual
- ▶ Visualisation can be part of the data mining process by presenting results of knowledge in visual forms e.g. Scatter plots (boxplots obtained from data mining in visual forms), decision tress, association rules, clusters, outliers and generalised rules

► Example of boxplots

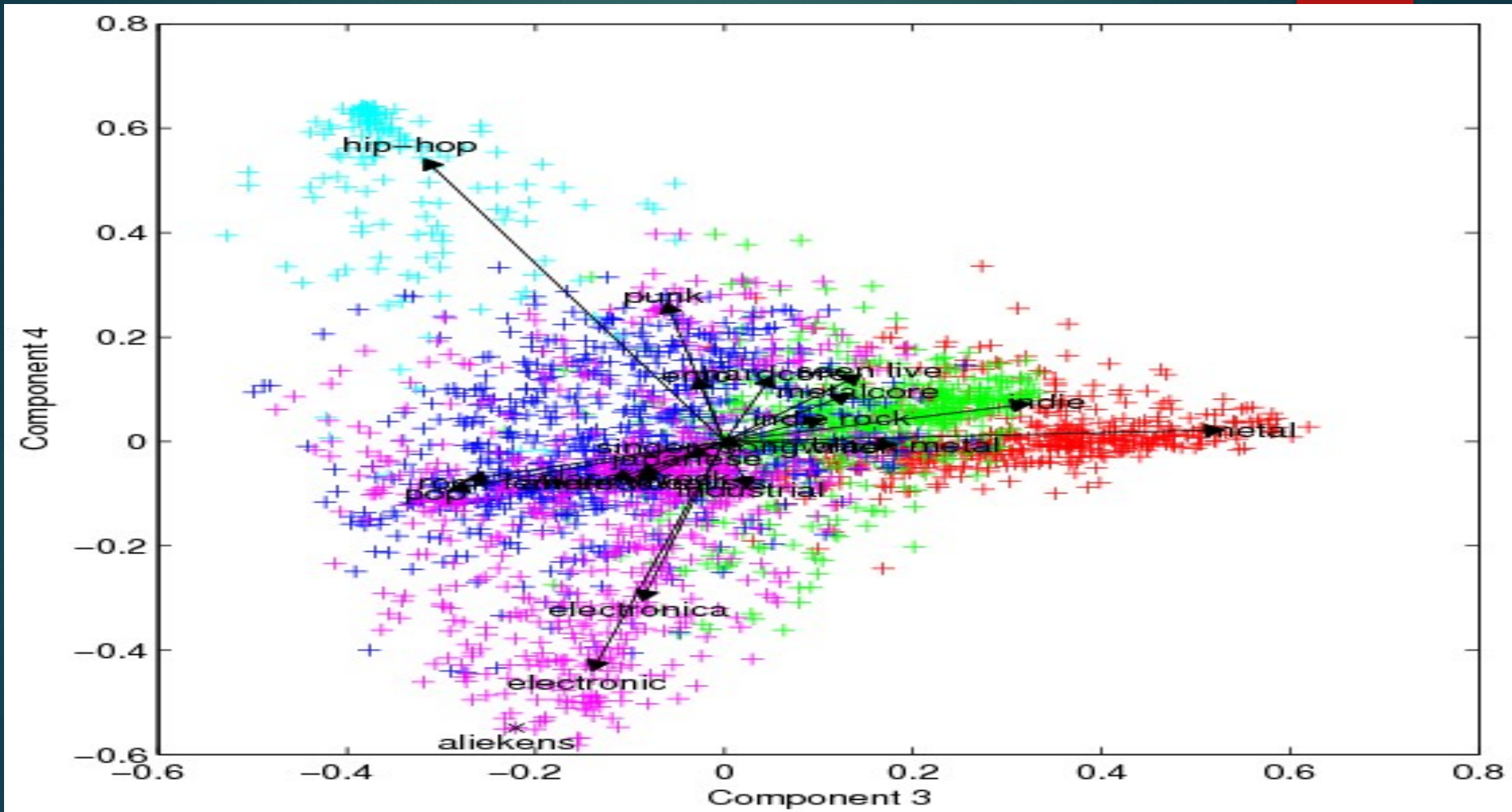


Developed in R which is an open source used for statistical computing and graphics





Produced using Oracle BI tools



The UK Government guideline for data visualisation

- ▶ [UK Government guidelines for data visualisations](#)

