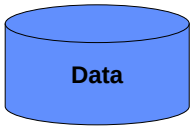
Data Mining & Machine Learning F2.4DN1/F2.9DM1

Nick Taylor
N.K.Taylor@hw.ac.uk
Room EM1.62



Data Mining - Content

- Introduction to Data Mining
  - What it is, Who does it and Why
  - Data Warehousing
  - "Virtuous Cycle of Data Mining"
- Data Mining Methods
  - Statistical Techniques
  - Machine Learning
  - Soft Computing

Data Mining - Definition & Goal

- Definition
  - Data Mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules
- Goal
  - To permit some other goal to be achieved or performance to be improved through a better understanding of the data

Data Mining - Basics

- Data Mining is the process of discovering patterns and inferring associations in raw data
- Data Mining cannot, on its own, identify cause and effect relationships
- Data Mining is a collection of powerful techniques intended to analyse large amounts of data
- There is no single Data Mining approach
- Data Mining can employ a range of techniques, either individually or in combination with each other

Data Mining - Why now?

1. Data is being generated in enormous quantities
2. Data is being collected over long periods of time
3. Data is being kept for long periods of time
4. Computing power is formidable and cheap
5. A variety of Data Mining software is available

Data Mining - History

- The approach has its roots over 40 years ago
- In the early 1960s Data Mining was called statistical analysis, and the pioneers were statistical software companies such as SPSS
- By the late 1980s these traditional techniques had been augmented by new methods such as machine induction, artificial neural networks, evolutionary computing, etc.

Data Mining – Two Major Types

- Directed (Farming)
  - Attempts to explain or categorise some particular target field such as income, medical disorder, genetic characteristic, etc.
- Undirected (Exploring)
  - Attempts to find patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes
- Compare with Supervised and Unsupervised systems in machine learning

Data Mining - Tasks

- **Classification** - Example: high risk for cancer or not
- **Estimation** - Example: household income
- **Prediction** - Example: credit card balance transfer average amount
- **Affinity Grouping** - Example: people who buy X, often also buy Y with a probability of Z
- **Clustering** - similar to classification but no predefined classes
- **Description and Profiling** – Identifying characteristics which explain behaviour - Example: “More men watch football on TV than women”

Data Warehousing

- Note that Data Mining is very generic and can be used for detecting patterns in almost any data
  - Retail data
  - Genomes
  - Climate data
  - Etc.
- Data Warehousing, on the other hand, is almost exclusively used to describe the storage of data in the commercial sector

Data Warehouse - Two Definitions

1. "A subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process"

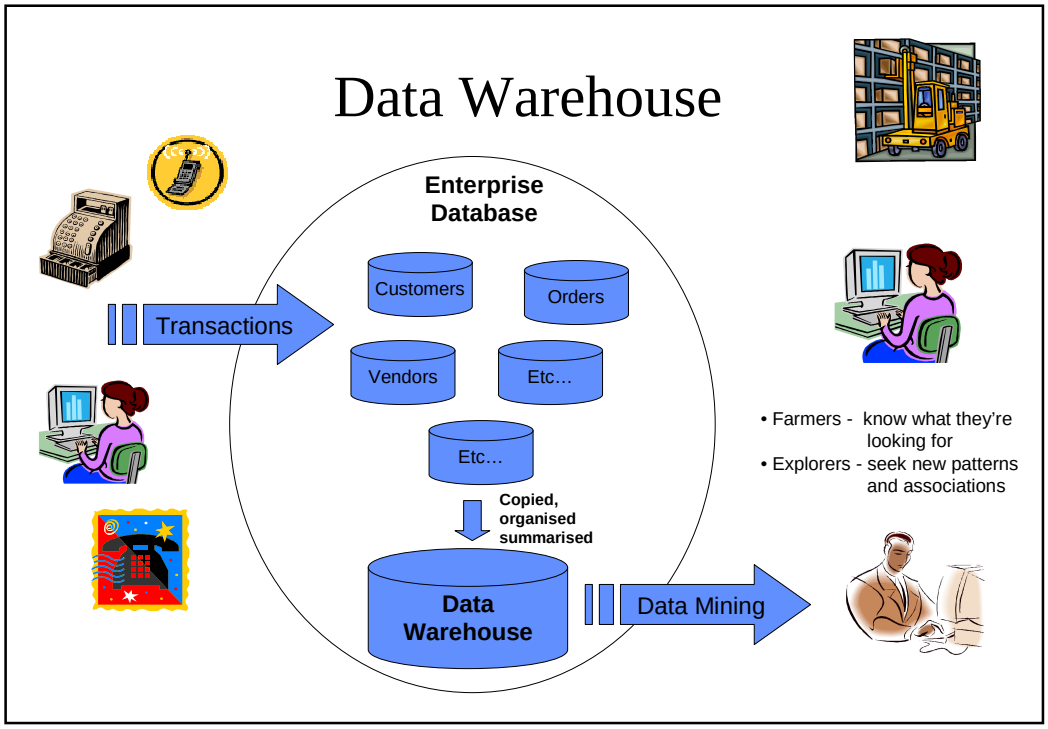
- W.H. Inmon

2. "A copy of transaction data, specifically structured for query and analysis"

- Ralph Kimball

Data Warehouse - Purpose

- For organisational learning to take place data from many sources must be gathered together over time and organised in a consistent and useful way
- Data Warehousing allows an organisation to remember its data and what it has learned about its data
- Data Mining techniques make use of the data in a Data Warehouse and subsequently add their results to it



Data Warehouse - Contents

- A Data Warehouse is a copy of transaction data specifically structured for querying, analysis and reporting
- The data will normally have been transformed when it was copied into the Data Warehouse
- The contents of a Data Warehouse, once acquired, are fixed and cannot be updated or changed later by the transaction system - but they can be added to of course

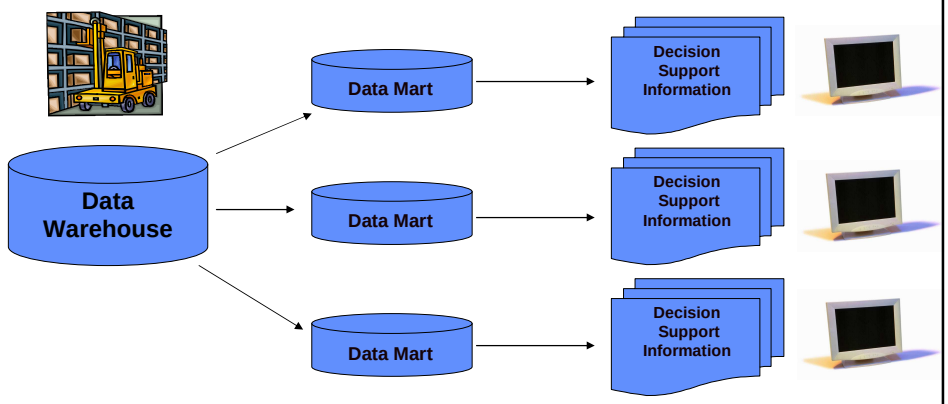
Data Marts

- A Data Mart is a smaller, more focused Data Warehouse – a mini-warehouse
- A Data Mart will normally reflect the business rules of a specific business unit within an enterprise – identifying data relevant to that unit’s acitivities

Data Warehouses and Data Marts

- Typically a Data Mart will consist of tables, of two or more dimensions, which index into the Data Warehouse
- They are designed to extract from the Data Warehouse particular data records and fields of relevance to a specific analysis
- The Data Mart will then aggregate this data into a smaller more manageable data set for Data Mining

Data Warehouse to Data Mart



Data Mining's Big Challenge

- The largest challenge that a Data Miner may face is the sheer volume of data in the Data Warehouse
- It is very important, then, that summary data also be available to get the analysis started
- The sheer volume of data may mask the important relationships in which the Data Miner is interested
- Being able to overcome the volume and interpret the data is essential to successful Data Mining

In Practice …

- Data Miners, both "farmers" and "explorers", are expected to utilise Data Warehouses to give guidance and answer a limitless variety of questions
- The value of a Data Warehouse and Data Mining lies in a new and changed appreciation of the meaning of the data
- There are limitations though - A Data Warehouse cannot correct problems with its data, although it may help to more clearly identify them