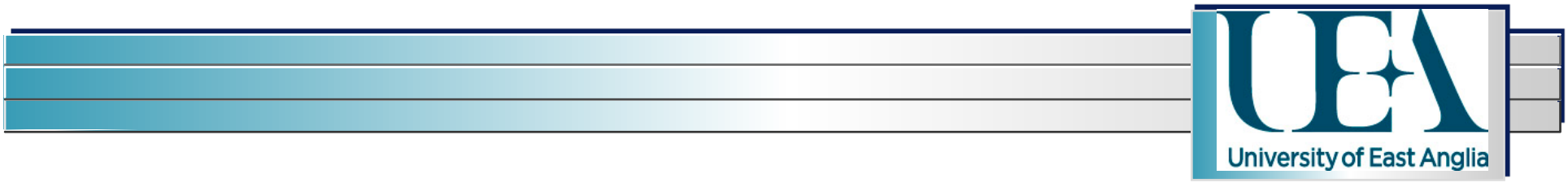
A Methodology: The KDD Roadmap



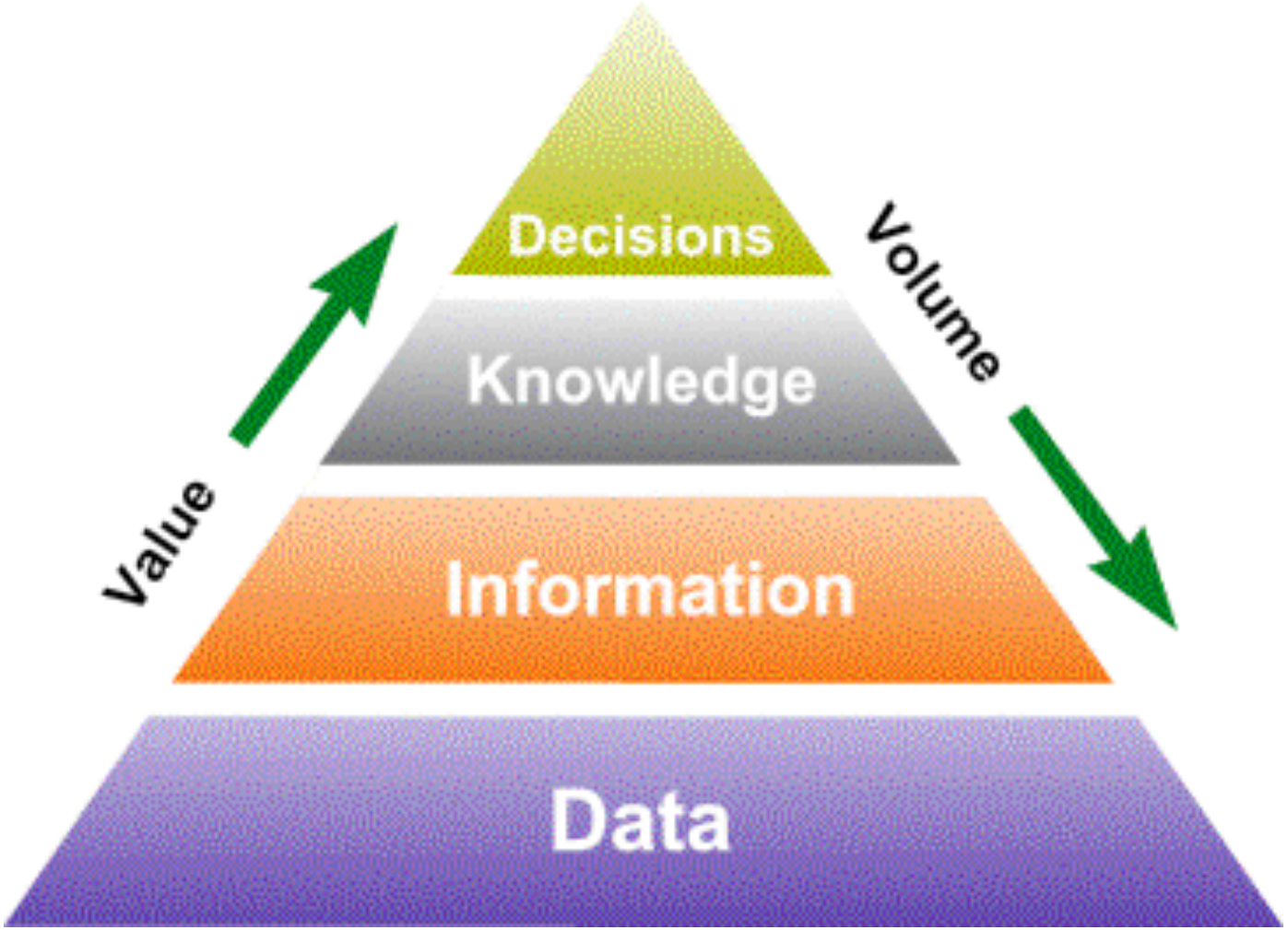
Dr Beatriz de la Iglesia

email: b.Iglesia@uea.ac.uk

Session outline

- Introduction to KDD and Data Mining
- KDD and its stages

The knowledge pyramid



Rationale for data mining

- With databases of enormous size, the user needs help to analyse the data more effectively than reporting, querying and statistics.
- Semi-automatic methods to extract useful, unknown (higher-level) knowledge in a concise format will help the user make more sense of their data.
- Emphasis on large data sets and real world ("messy") data.

Data Mining History

- Emerges in the late 1980s.
- Starts to take shape as a discipline on its own right in the mid 1990s.
- But it grows from a body or knowledge in statistics, machine learning, databases and others dating back to 1700s.
- Research discipline: *It is still in its youth!!*

Data Mining

- Ultimate aim:

> "To provide automatic tools to assist us in transforming the vast amounts of data into **useful** information and knowledge"

Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases is the automatic **non-trivial** *process* of identifying **valid**, **novel**, **potentially useful**, and **ultimately understandable** patterns in data.

Data Mining

Data Mining is *a **step*** in the KDD process.

It consists of the application of particular data mining **algorithms** to extract the information and patterns derived by the KDD process.

Big data (the current buzzword)

University of East Anglia

- What is big data? By dictionary definition, data sets that are too large and complex to manipulate or interrogate with standard methods or tools.

- Examples:

  - Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces **1 Gigabit/second** of astronomical data over a 25-day observation session.

  - Twitter firehose: Every second, on average, around 6,000 tweets are tweeted on Twitter, this is equivalent to 500 million per day

  - Amazon handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers.

Characteristics of big data

- So much data, it cannot always be all stored, i.e. analysis has to be done "on the fly", on streaming data.
- Even if data is stored, it often requires "massively parallel software running on tens, hundreds, or even thousands of servers."
- Challenges and opportunities are defined as three-dimensional, i.e.
  - increasing volume (amount of data),
  - velocity (speed of data in and out),
  - and variety (range of data types and sources**).**

Patterns or models

A **model** is a description of the original database, summarising the important characteristics of the data, that can be successfully applied to new data or to better describe the DB, e.g. a decision tree for classification.

A **pattern** is a statement that describes some relationship among a subset of the data with a degree of certainty, e.g. a frequent itemset rule.

Examples of models/patterns

- Classification rules:

  "if IntCallsLastWeek < 5 and IntCallsThisWeek > 20 then Fraud"

- Associations:

  "customers who buy items *a* and *b* also buy *c* and *d*"

- Cluster model:

  A description of three closely related groups of customers that buy from a web-site.

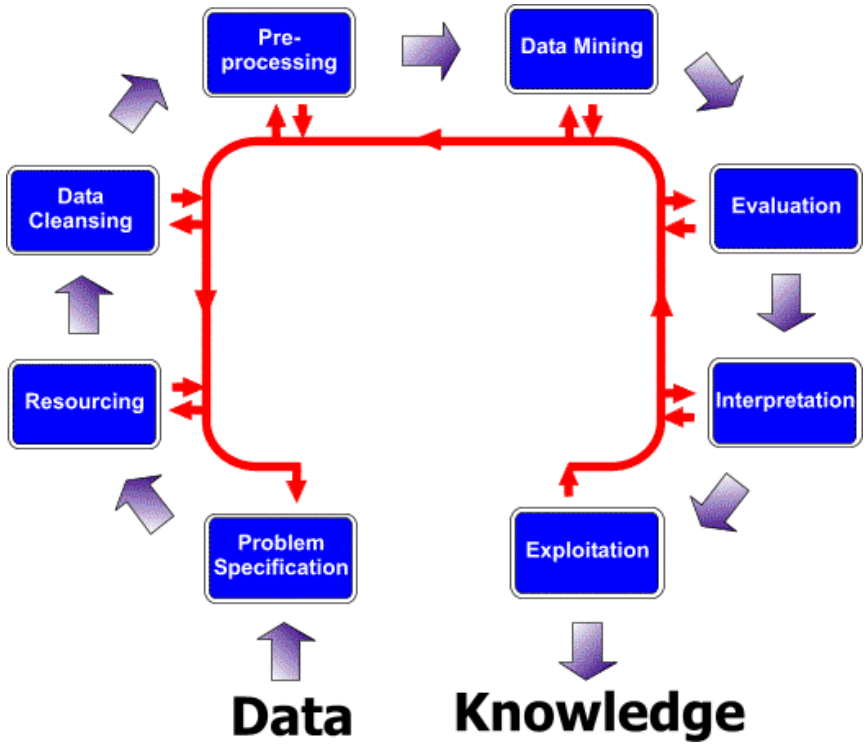
- Classification model

  A decision tree or Neural network to predict high risk customers.

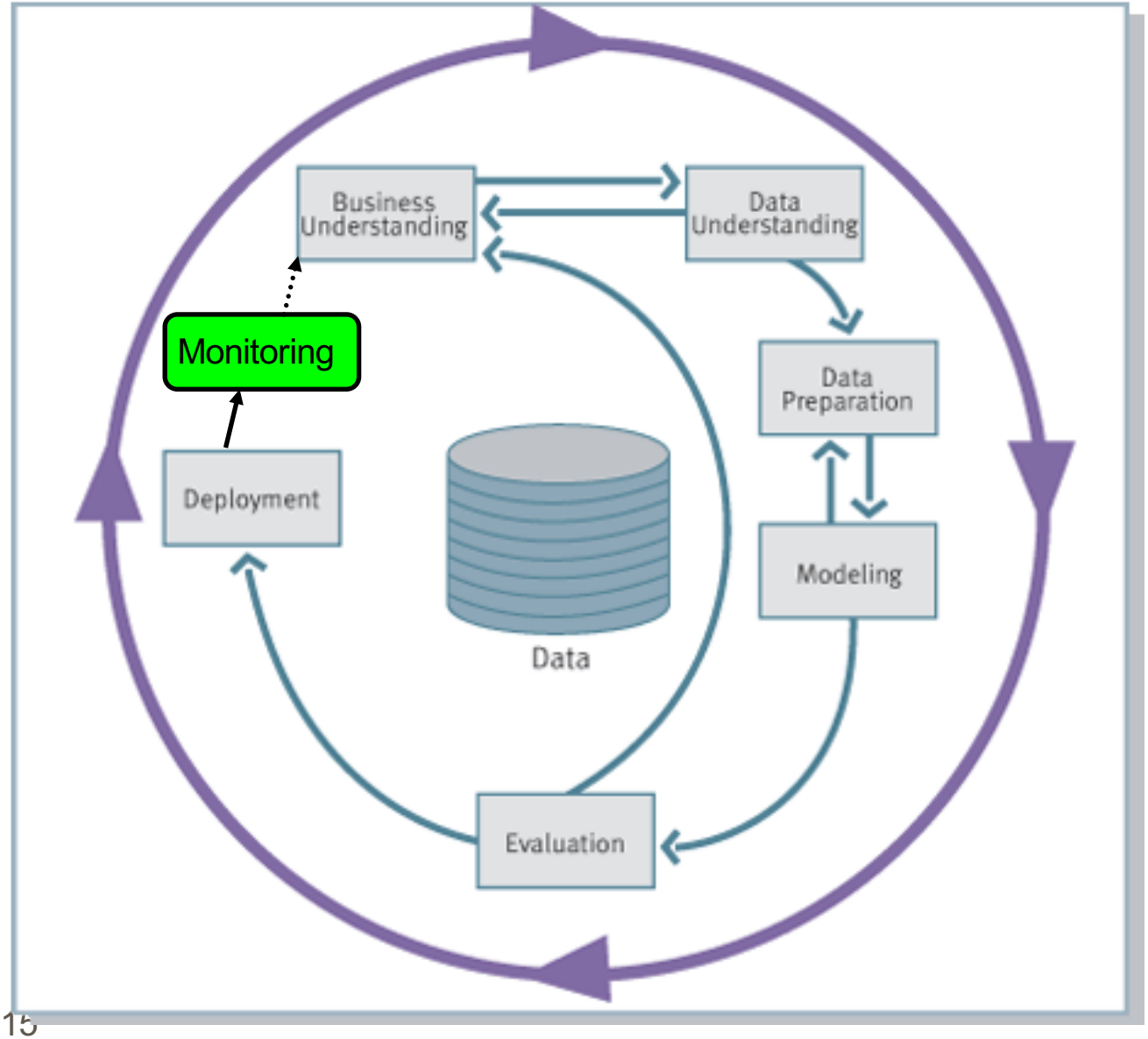
KDD: cross-roads

Outline KDD process

- KDD process will be presented in the form of a Roadmap.
- Provided rules of the road are obeyed, a variety of routes may be taken.
- Suggested routes can be formulated for particular applications.

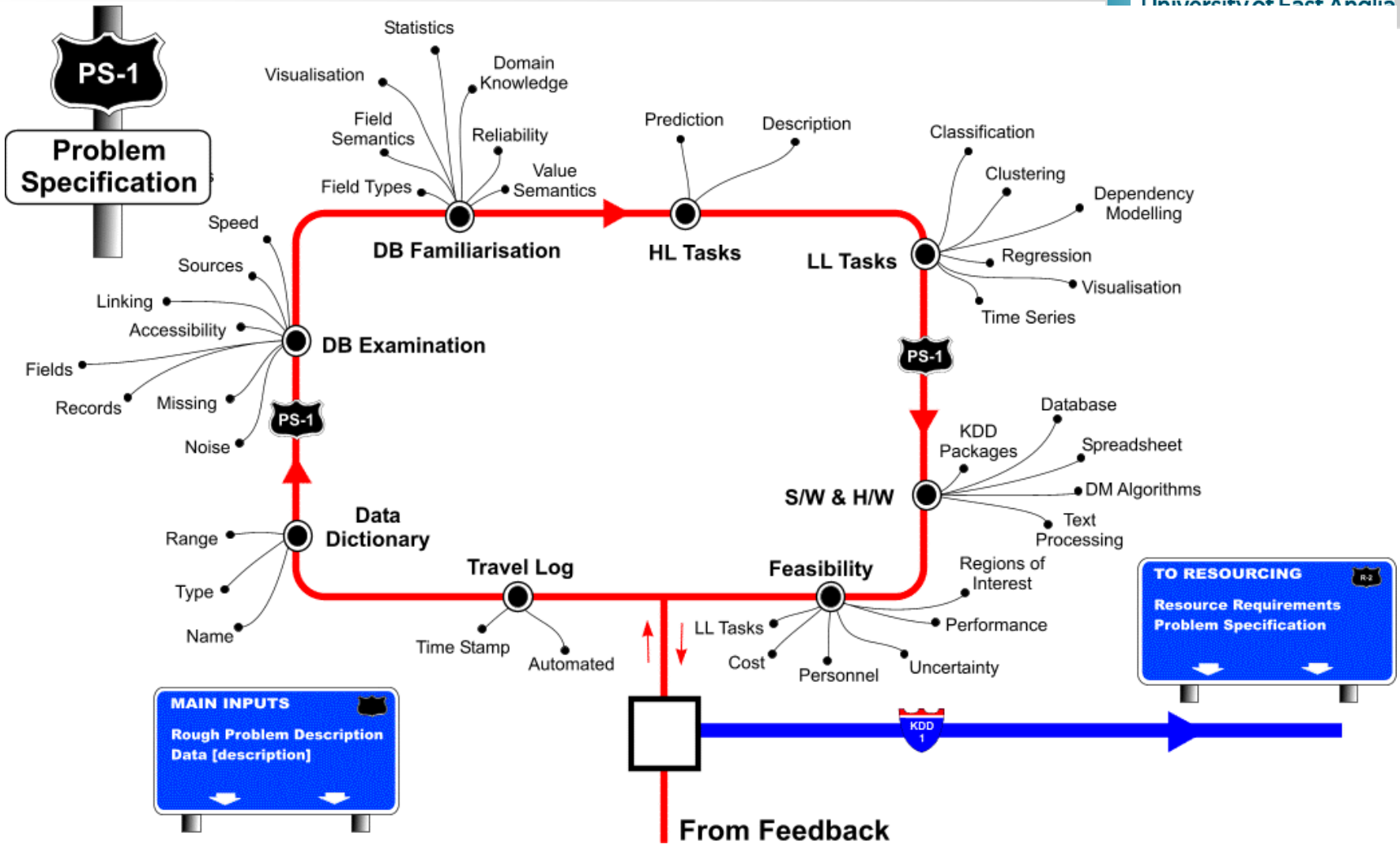


Knowledge Discovery Process flow, according to CRISP-DM



see www.crisp-dm.org for more information

Problem specification



Problem specification

- The purpose of this phase is to move from a **loosely** defined problem description to a **tightly** defined problem specification.
- Processes which are performed within this phase include preliminary database examination and familiarisation, determination of required tasks, data availability and software and hardware requirements.

Inputs to specification

- A problem description which may be loosely defined.
- Some (rough) description of the data to be mined.
- A typical enquiry - "We have some data from a marketing exercise and would like to use it to target our marketing better."

Travel log

- A **travel log** should be initiated to store details of the operations performed at each stage of the process.
- Each operation should be **time-stamped.**
- This document is updated throughout the project.
- Useful for tracking, repeating and reversing operations performed.
- Manual or automatic updates.

Case study: Customer Response

- The Customer Response (CR) database contains profile information about customers of a software company. The company stores this information about customers who have purchased their main software packages.
- The data describes various interactions between the company and the customer, historical information about the customers' accounts and what additional packages (or modules) they have purchased.
- Recently, the company sent a newsletter to these customers describing a new module that will soon be available.
- One month after the distribution of the newsletter, a count is taken to determine the response rate, i.e. those who subsequently bought the new module.

Data dictionary

- A **data dictionary** (DD) must be available for each data source.
- A data dictionary will contain attribute (field) names, types and ranges, together with information on missing values and reliability of values.
- It may also give details of the source of the data, when and where it was created, as well as the purpose for which it was gathered, if appropriate.

Preliminary DB examination

- The following are determined:
  - number of records and fields (attributes)
  - proportion of DB that is missing
  - proportion of fields with missing values
  - proportion of records with missing values
  - whether or not there is a target field (prediction)
  - the integration required to form the DB from multiple sources
  - speed with which data can be accessed (some data may be in paper format)

Customer response

- Data Dictionary - Customer Response database
  - Source - Software Company database
  - Number of instances (records) - 1012
  - Number of attributes (fields) - 11
  - Missing attribute percentage - 0%
  - Target attribute? - Response (=YES/NO)
  - Target class distribution
    - Response = Yes (11.86%)
    - Response = No (88.14%)

Customer response

Field Name	Description	Type
Satisfaction	A rating based on a recent customer survey	Num. Cont.
Software	An indicator of which main software packages have been purchased	Categorical (A,B,C,D,E,F,G)
Date	Number of months since purchase of main software package	Num. Disc.
Modules	The number of add-on modules purchased in last 3 years	Num. Disc.
VisitFromSales	The number of visits made by a sales representative in the last year	Num. Disc.
VisitFromSupport	The number of visits made by support staff in the last year	Num. Disc.
DaysConsultancyThisYear	The number of days of additional consultancy projects in the current year	Num. Disc.
DaysConsultancyLastYear	The number of days of additional consultancy projects last year	Num. Disc.
UserGroup	Whether the customer attended the last user group meeting	Categorical (Y,N)
TrainingScore	The average score gained after attending the training session that accompanies the purchase a main software package	Num. Cont.
Response	Whether the customer responded to the latest newsletter	Categorical (Y,N)

DB familiarisation

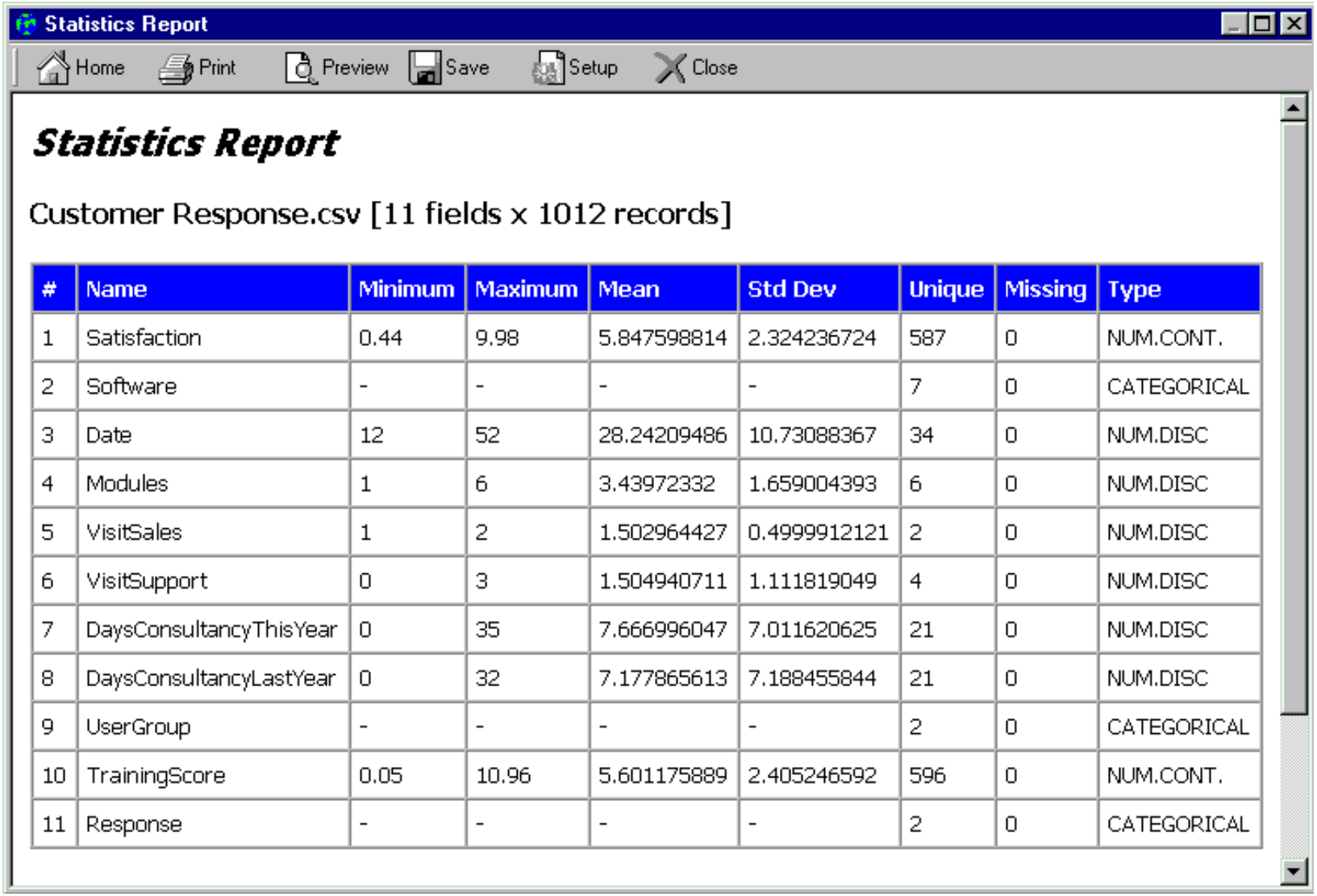
Field types:

- Numerical - discrete (number of children)
- Numerical - continuous (temperature)
- Categorical - ordinal has an implied ordering (small, medium, large)
- Categorical - nominal has no implied ordering (labels such as gender)
- Some fields may use wrong type (e.g. integer instead of categorical nominal)

DB familiarisation

- Statistics
  - A basic understanding of each field may be gained by statistical analysis (range, mean, median, standard deviation as appropriate).
  - Also, looking at distributions of records within each attribute for categorical variables.
  - Such analysis may suggest suitable data cleansing and pre-processing(see later).

Customer response

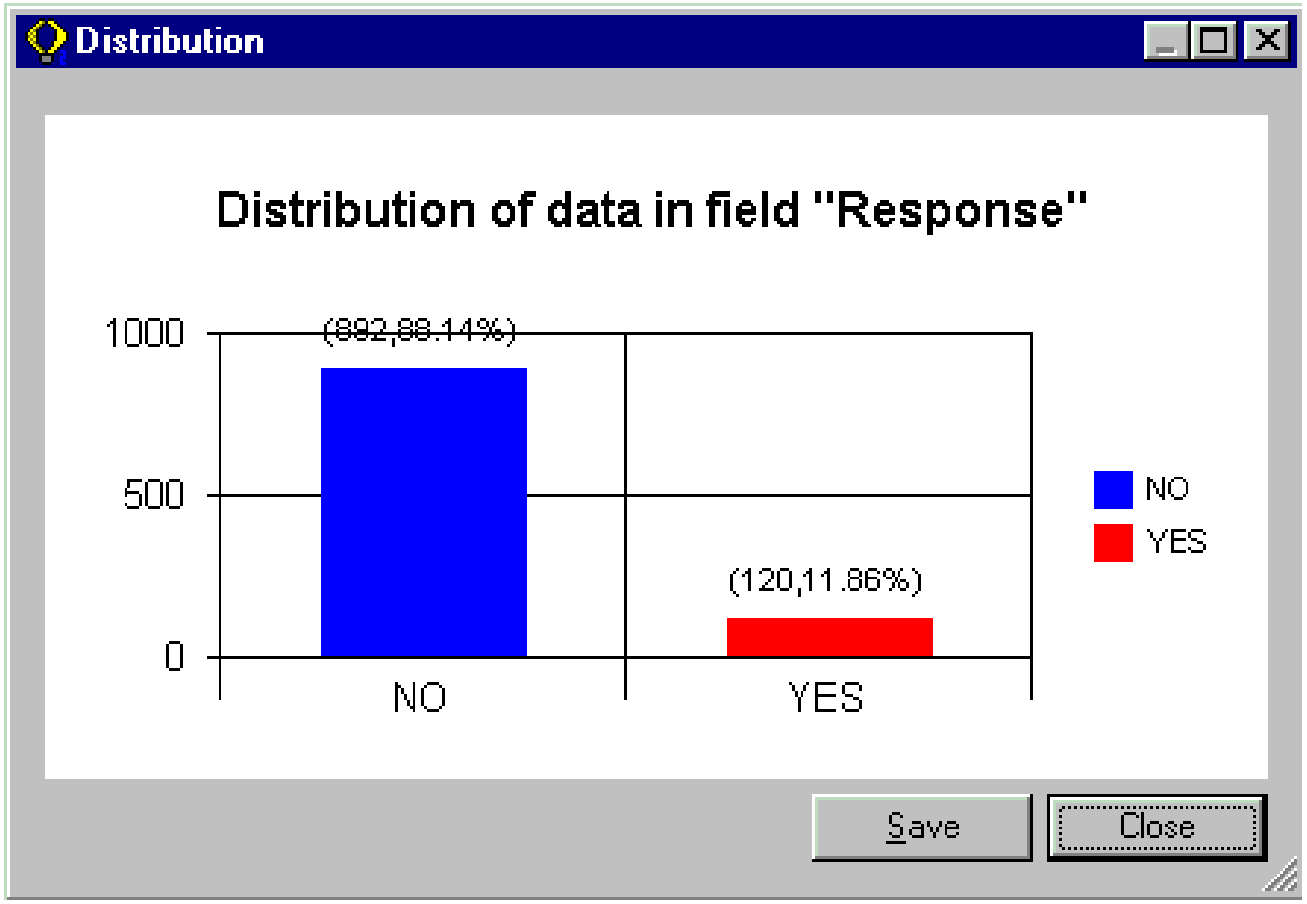


Statistics Report

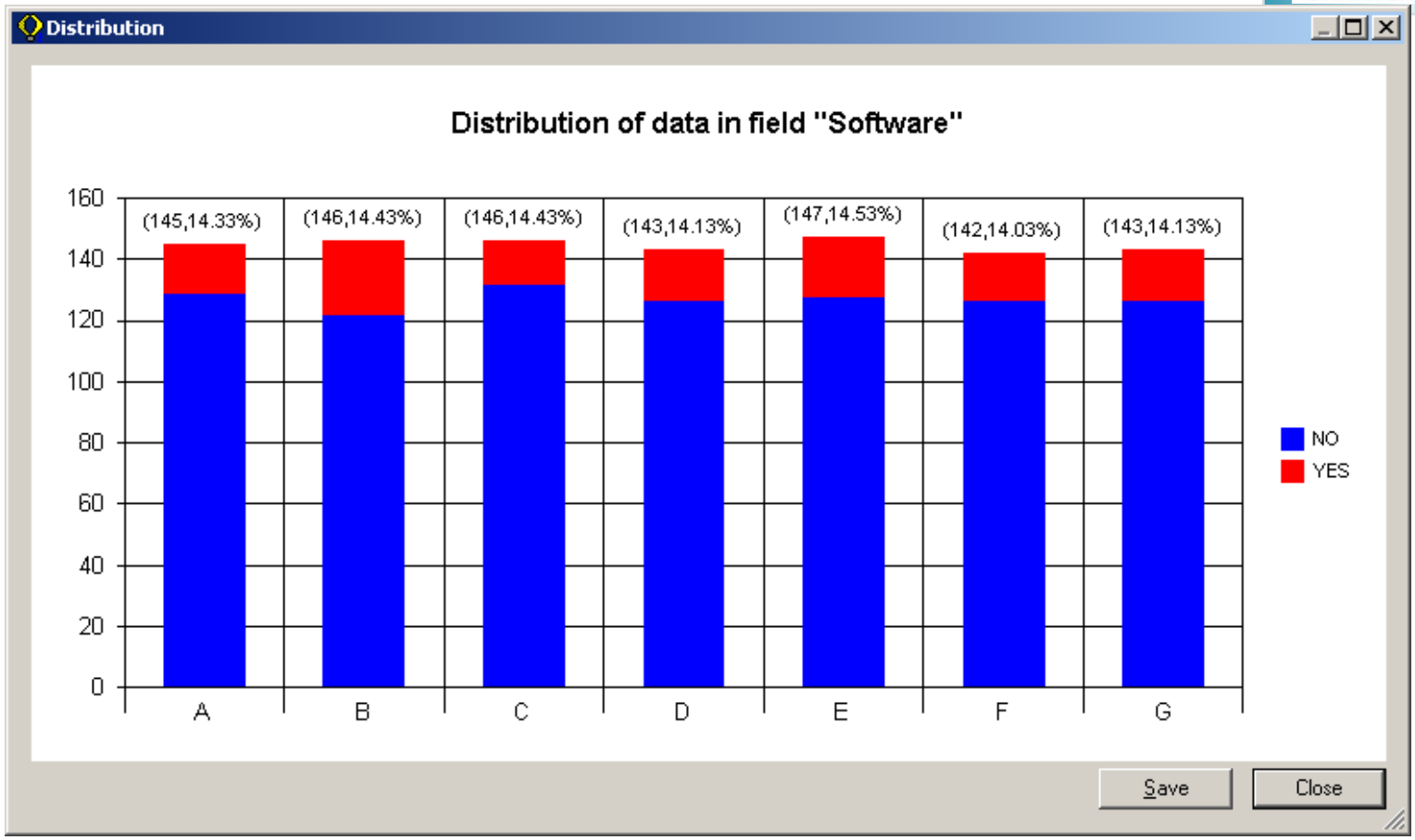
Customer Response.csv [11 fields x 1012 records]

#	Name	Minimum	Maximum	Mean	Std Dev	Unique	Missing	Type
1	Satisfaction	0.44	9.98	5.847598814	2.324236724	587	0	NUM.CONT.
2	Software	-	-	-	-	7	0	CATEGORICAL
3	Date	12	52	28.24209486	10.73088367	34	0	NUM.DISC
4	Modules	1	6	3.43972332	1.659004393	6	0	NUM.DISC
5	VisitSales	1	2	1.502964427	0.4999912121	2	0	NUM.DISC
6	VisitSupport	0	3	1.504940711	1.111819049	4	0	NUM.DISC
7	DaysConsultancyThisYear	0	35	7.666996047	7.011620625	21	0	NUM.DISC
8	DaysConsultancyLastYear	0	32	7.177865613	7.188455844	21	0	NUM.DISC
9	UserGroup	-	-	-	-	2	0	CATEGORICAL
10	TrainingScore	0.05	10.96	5.601175889	2.405246592	596	0	NUM.CONT.
11	Response	-	-	-	-	2	0	CATEGORICAL

Customer response

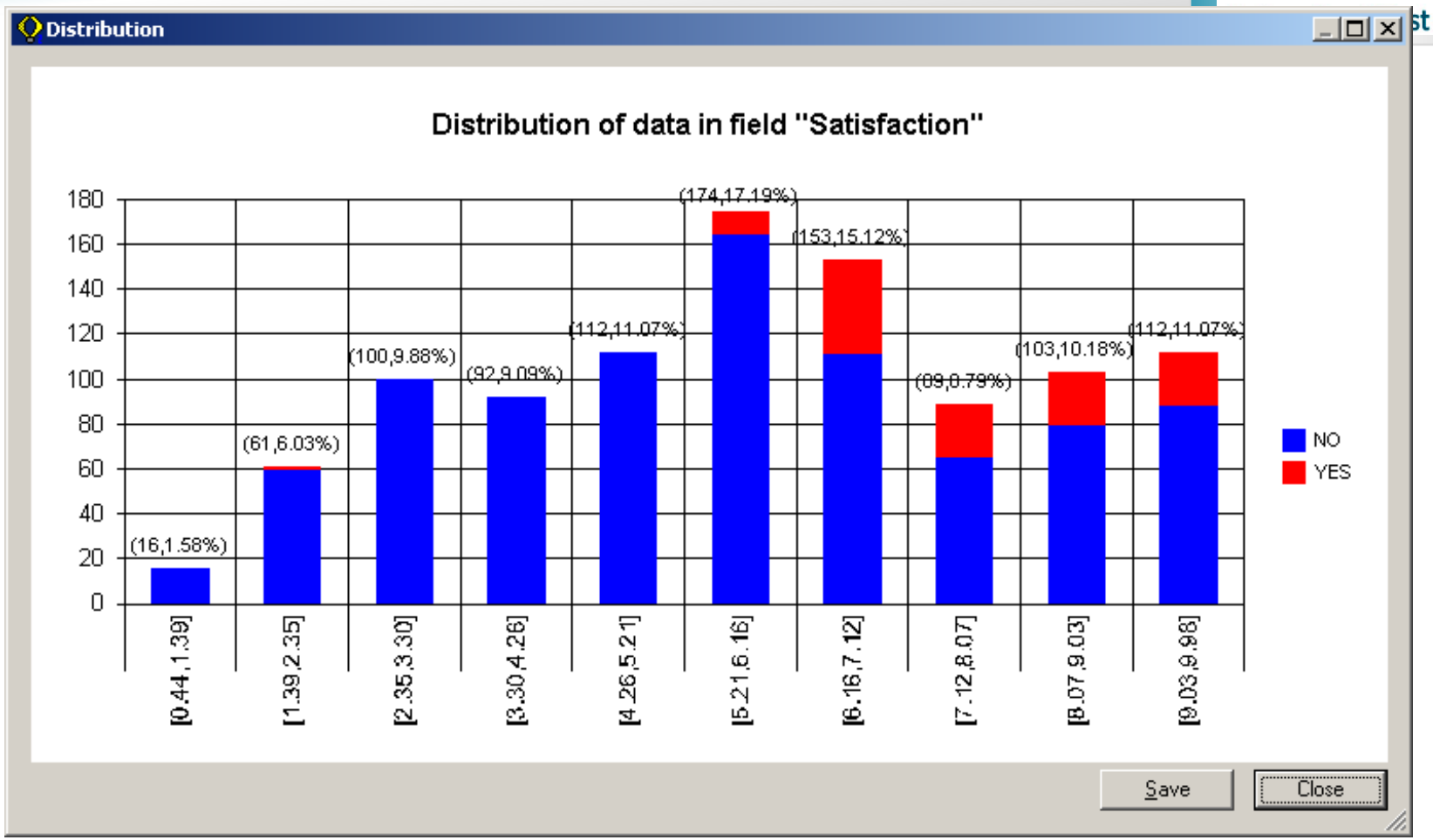


Customer response



There is very little difference in each category - so **Software** is unlikely to be a good predictor of Response.

Customer response



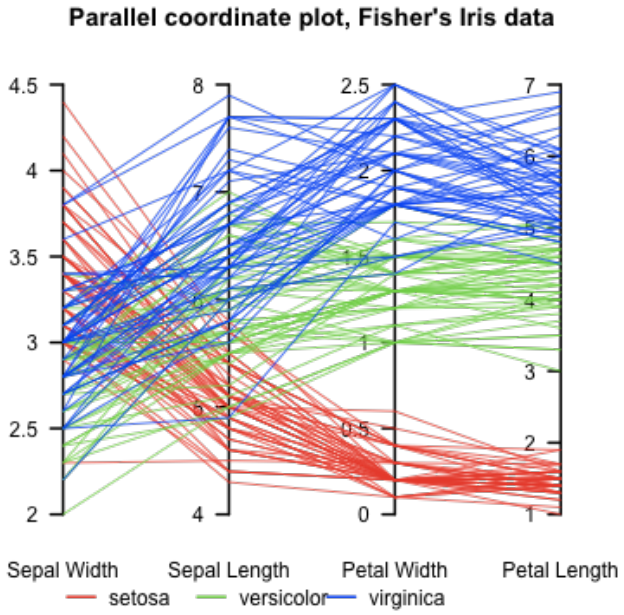
Here we see that only 1 customer with Satisfaction < 5.21 actually responded. Satisfaction is a strong predictor.

DB familiarisation

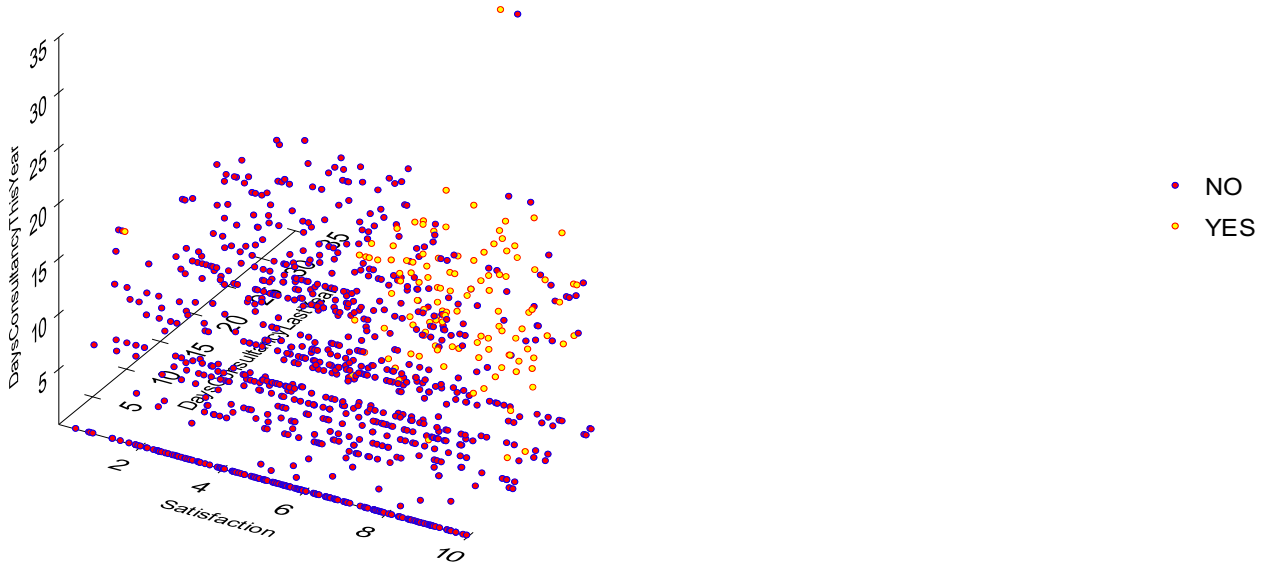
- Reliability
  - Fields (and even some values) within the database may have differing levels of reliability.
  - Knowledge of these levels can be useful in pre-processing operations and to target the more reliable data when mining begins.
  - For instance, how reliable is this **Satisfaction** field in the Customer Response database?

DB familiarisation

- Visualisation
  - Familiarisation with the database may be achieved with simple 2D or 3D plots.
  - More sophisticated visualisation may allow higher dimensionality – parallel coordinates plots many variables at once.



Customer response



x-axis is Satisfaction, y-axis is DaysConsultancyLastYear, z-axis is DaysConsultancyThisyear

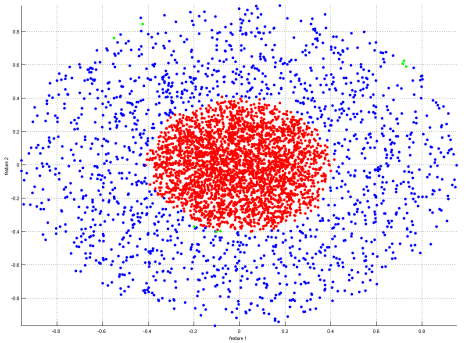
Note the cluster of Response = Yes. This is for higher values of each of the three axis variables.

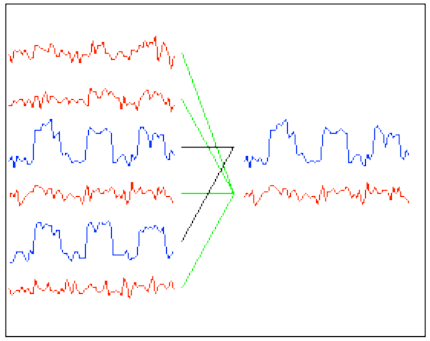
High Level Task (HLT)

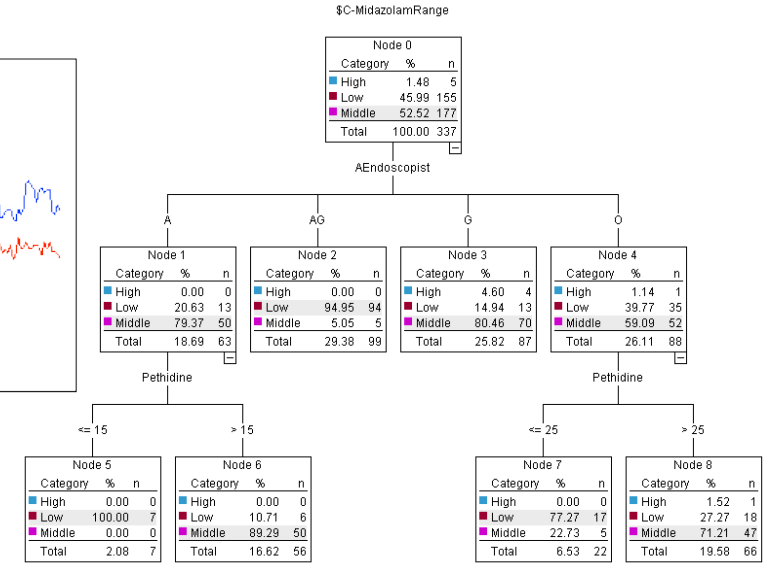
- The goal or goals of the DM project must be determined.
- There are two recognised categories of goals (or high level tasks):
  - Prediction (predict future based on historical data)
  - Description (understand your data better)

Low level task (LLT)

- Main tasks are:
  - Classification (predicting a target class)
  - Clustering (finding similarly behaving objects)
  - Association rules (describing associations in transactions)
  - Regression (predicting a target real-valued score)
  - Time series and sequence discovery
  - Text and Web mining

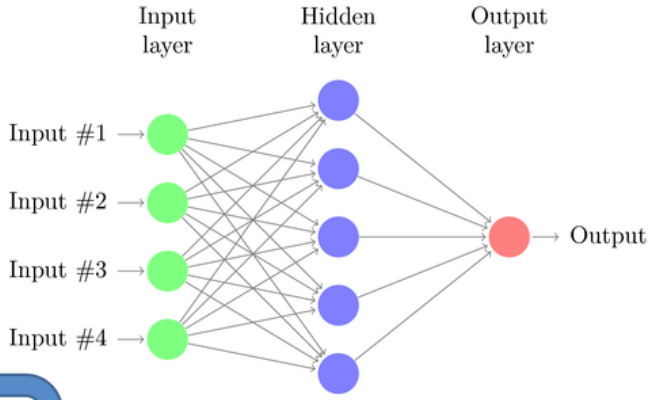


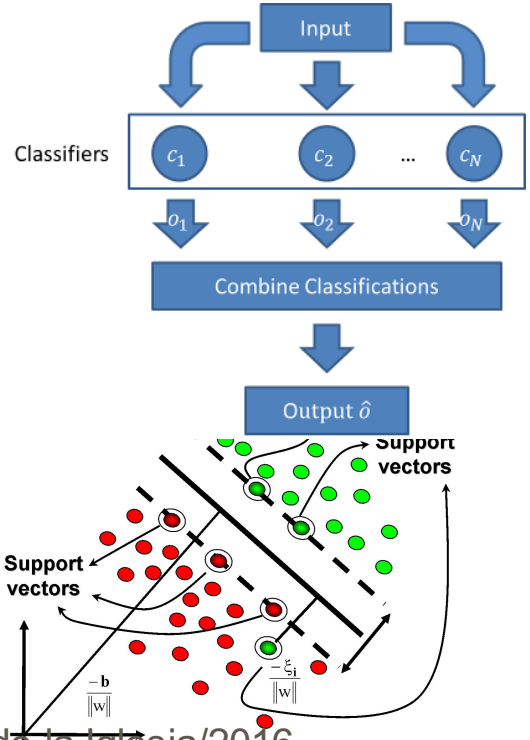


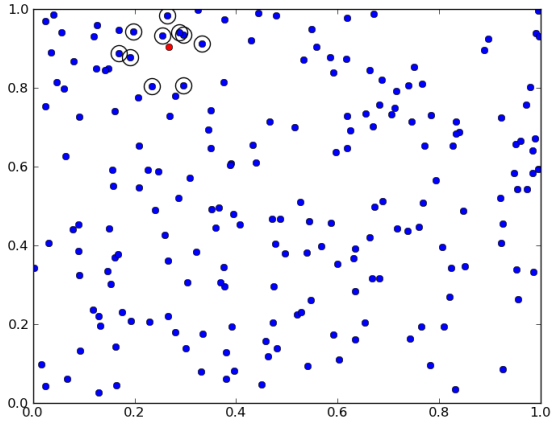


A rich toolkit: e.g. classification

- Decision trees
- Neural networks
- Nearest Neighbour algorithms
- Decision rules
- SVM
- Random Forest
- Naïve Bayes classifier
- Ensembles
- …







Software & hardware

- The decision made with respect to the low-level tasks leads to decisions about the procurement of the necessary software to support the KDD process.
- The hardware requirements follow on from the software decisions and other factors such as the database(s) used.

Feasibility

- The feasibility of the project is then determined and the detailed problem specification is then produced.
- The project may not be feasible for many reasons (too much missing or unreliable data).
- Other issues are: available hard disc space, memory, database access speed, processor speed, personnel, cost.
- Only when these issues have been fully discussed, can a decision be made as to the feasibility of the entire project.

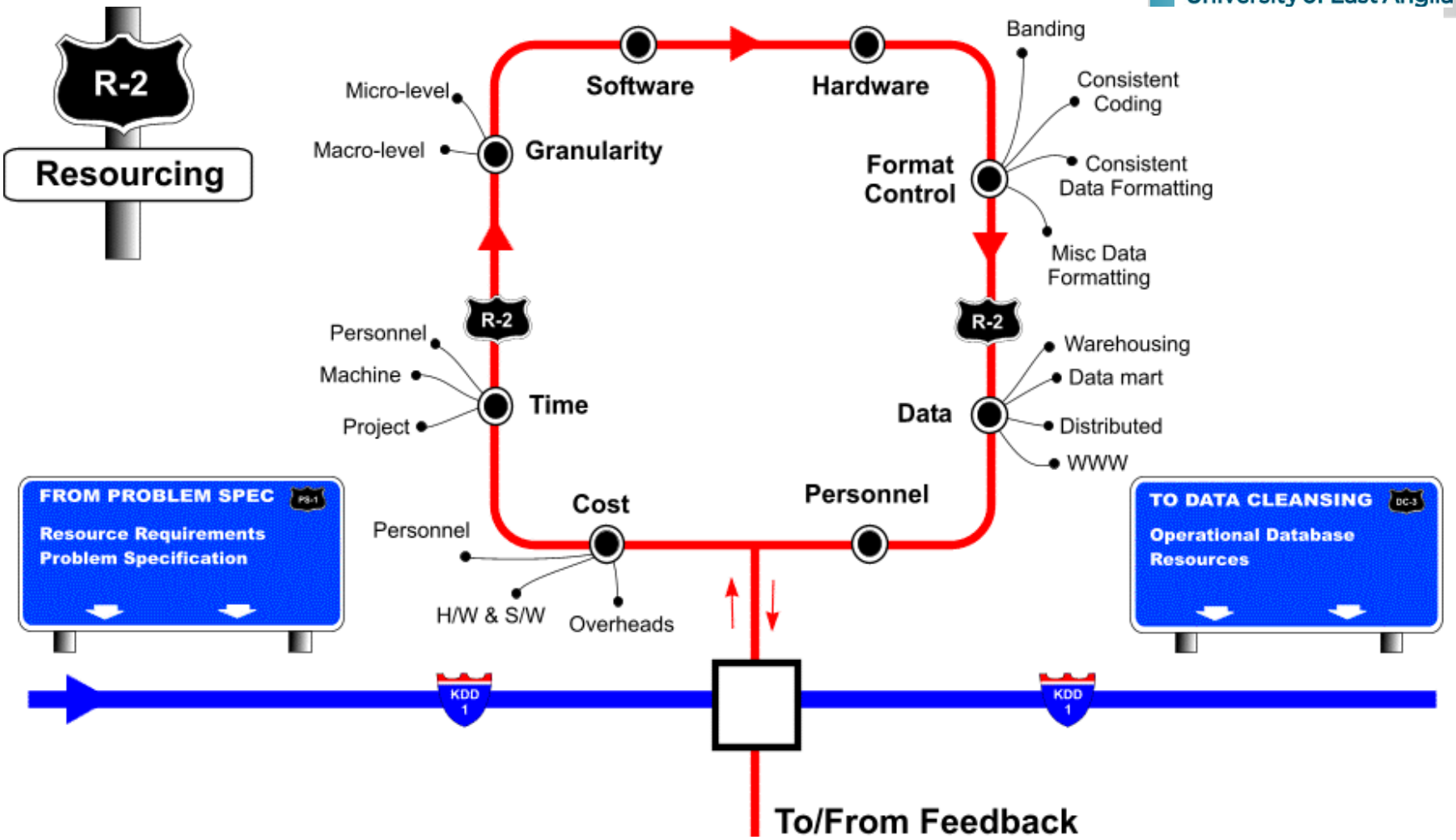
Outputs of specification

- The output is a Problem Specification containing:
  - A list of resource requirements
  - HLTs (prediction or description)
  - LLTs (classification, clustering, etc.)
  - Data dictionary
  - Feasibility report
  - Travel log (updated to record the above information).

Discussion time

- Prediction or description? What task?
- A estate agency has accumulated a large number of property sale records. The properties can be flats, semi-detached or detached or mansion houses.
  - The agency wants to investigate what kinds of customers are likely to purchase which types of property.
  - The agency wants to create a model for marketing to customers which enables them to guess the average spend per customer given other customers' characteristics.
  - The agency is going to cross sell products to high net-worth customers.

Resourcing


R-2
Resourcing
Software
Hardware
Banding
Consistent Coding
Consistent Data Formatting
Misc Data Formatting
Micro-level
Macro-level
Granularity
Format Control
R-2
R-2
Personnel
Machine
Project
Time
Data
Warehousing
Data mart
Distributed
WWW
FROM PROBLEM SPEC
Resource Requirements
Problem Specification
Cost
Personnel
Personnel
H/W & S/W
Overheads
TO DATA CLEANSING
Operational Database
Resources
KDD 1
KDD 1
To/From Feedback

Resourcing

- The main activity in this stage is to gather all resources necessary for the project
- The data for the project must be obtained and transformed into the **operational database.**
- The operational database is a unique source of data in a minable form, showing a standard format of fields and consistency of values.

Gathering the data

- The main resource to any DM project is the data
- Gathering the data may be simple and straight forward or a complex process depending on the sources of data
  - data warehouses and data marts
  - distributed databases
  - web data
  - paper data!

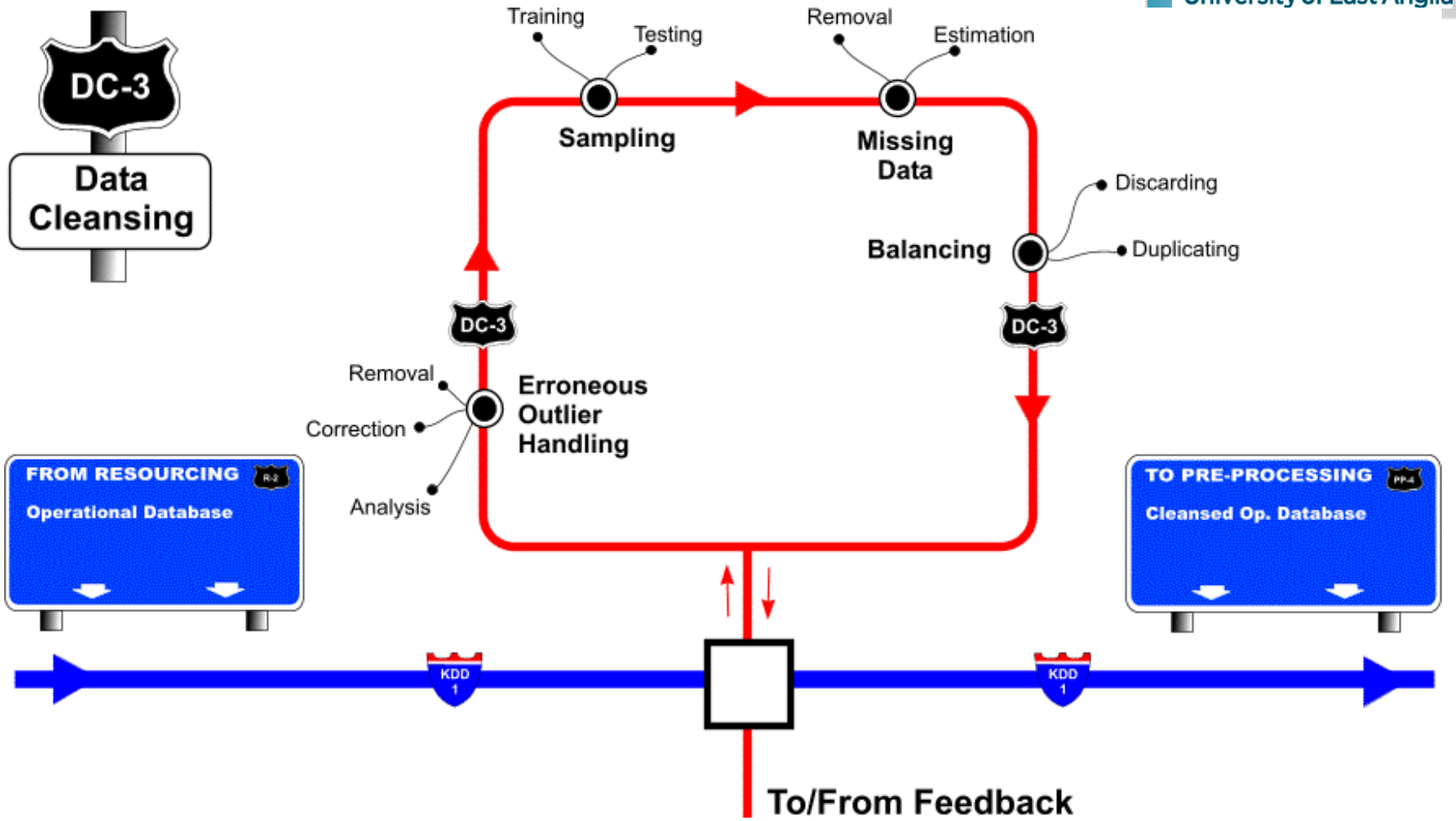
Linking and format control

- Data from different sources needs to be integrated to and standard form.
- Linkage may be required (sometimes probabilistic methods used).
- Format may be different in data from different sources:
  - Source 1: age (1,2,3, …, 99) integers
  - Source 2: age (1-5, 6-10, …, 65-90)
  - May require transformations to achieve consistent formatting.

Outputs of resourcing

- An operational database, usually separate from actual database (copy or virtual).
- The various resources required, or at least a schedule of when they need to be acquired.
- Project management documentation.

Data Cleansing



Data cleansing

- The aim of this stage is to prepare the data for subsequent stages that involve learning.
- Operations performed here include the removal of errors, dealing with missing values, balancing (if necessary) and many others.
- There is no learning in this stage.
- The main objective is to **improve the quality** of the data.

Data cleansing operations

- Outlier handling
  - Dealing with observations (cases) which appear to be inconsistent with the remainder of the data set.

- Sampling
  - Partitioning the data into a train/test set
  - Alternative methods to check error of the model on new data.
  - Used for data reduction.

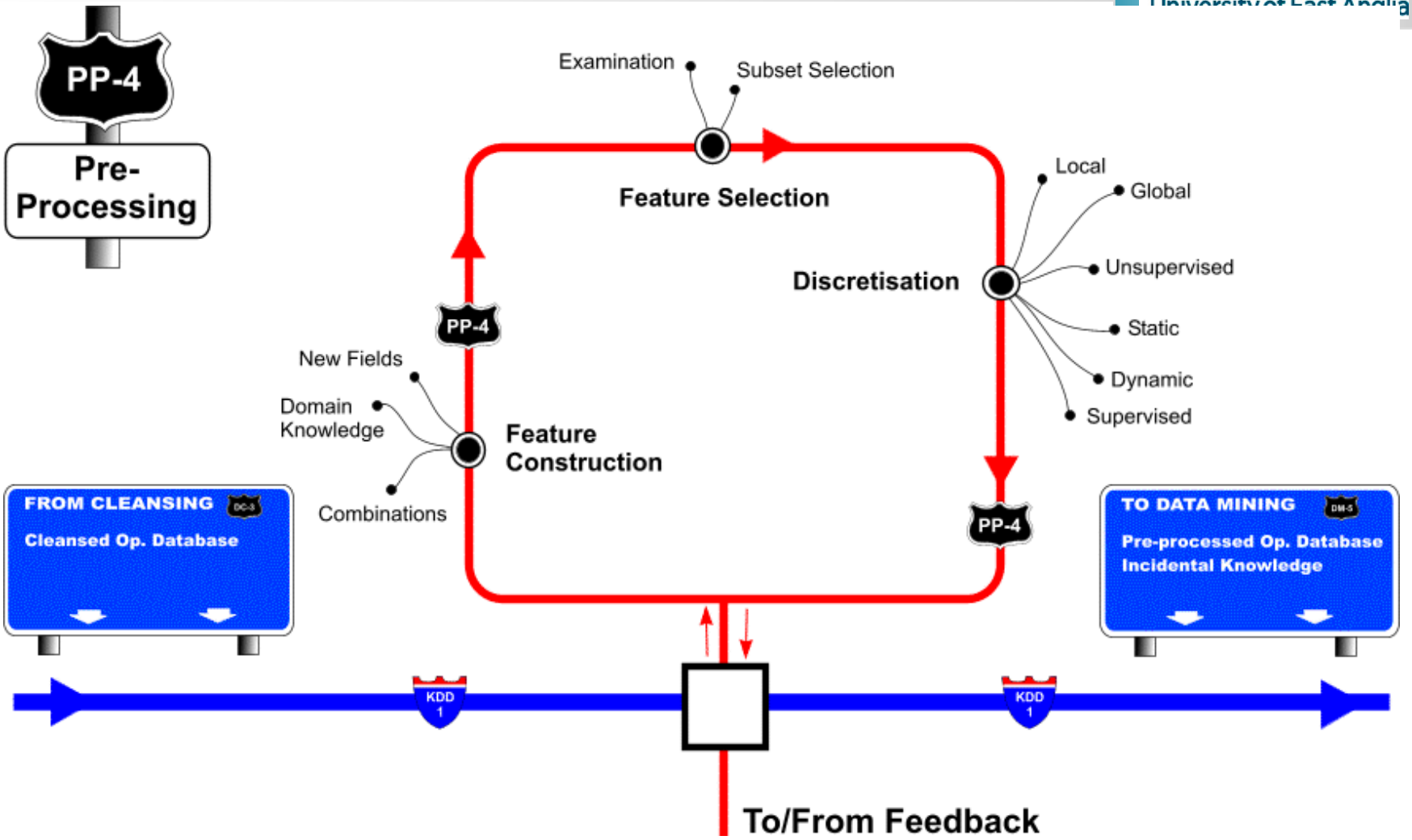
Data cleansing operations

- Missing data handling
  - removing or estimating missing values in the data
- Database balancing
  - correcting imbalances in the target field
- Others as appropriate

Outputs of cleansing

- All the data sets that may be generated in this phase should be cleansed operational data sets.
- There may be
  - train and test datasets
  - balanced datasets
- They have the following characteristics:
  - high quality data
  - free from erroneous outliers
  - missing data may have been dealt with

Pre-processing

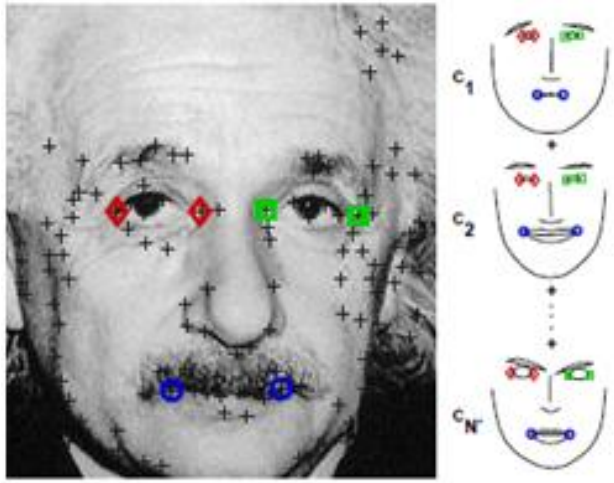


Pre-processing

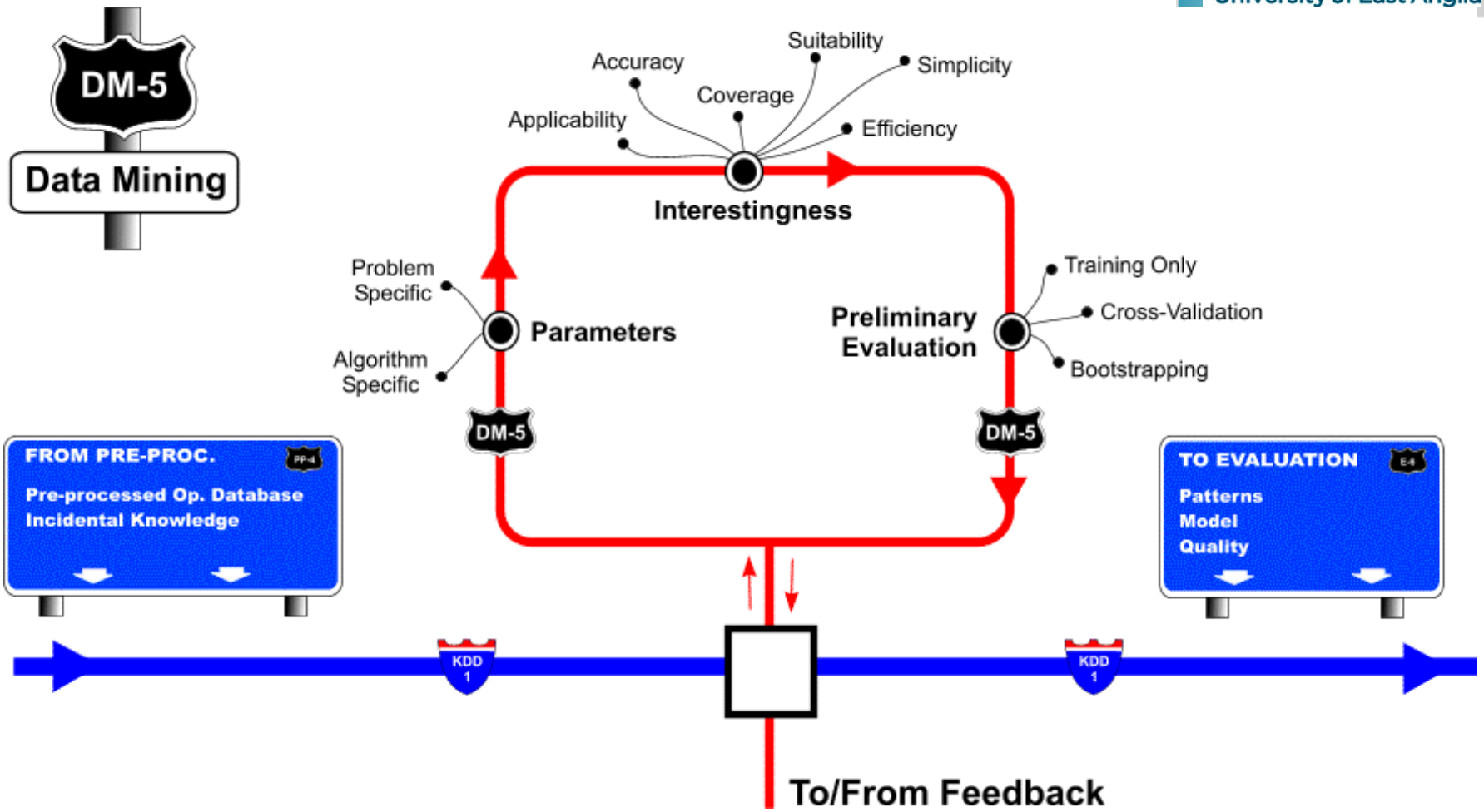
- Pre-processing the data is an essential stage of any data mining project.
- The techniques applied here may yield much more effective models for prediction and description.
- In some applications, pre-processing is sufficient to gather knowledge enough for decisions to be made.

Pre-processing techniques

- Feature construction
  - creating new features that are highly predictive
- Feature selection (Data reduction)
  - reducing the number of features to a powerful subset
- Discretisation (Data reduction)
  - reducing the number of values within a feature



Data Mining


DM-5
Data Mining
Accuracy
Suitability
Simplicity
Coverage
Applicability
Efficiency
Interestingness
Problem Specific
Algorithm Specific
Parameters
Preliminary Evaluation
Training Only
Cross-Validation
Bootstrapping
DM-5
DM-5
FROM PRE-PROC.
Pre-processed Op. Database
Incidental Knowledge
TO EVALUATION
Patterns
Model
Quality
KDD 1
KDD 1
To/From Feedback

DM and model selection

- Data mining is concerned with the **selection** of a model or the extraction of patterns from data
- Many models can fit the same data
- DM is concerned with the improvement (**optimisation**) of that model to obtain the best prediction or description.

Model selection

- Some of the model selection is done during the Problem Specification stage, when deciding on Low-level tasks, for example, classification.
- As there are many different classification models (artificial neural network, statistical models, rule induction, etc), the particular model(s) chosen will have a bearing upon software resources.

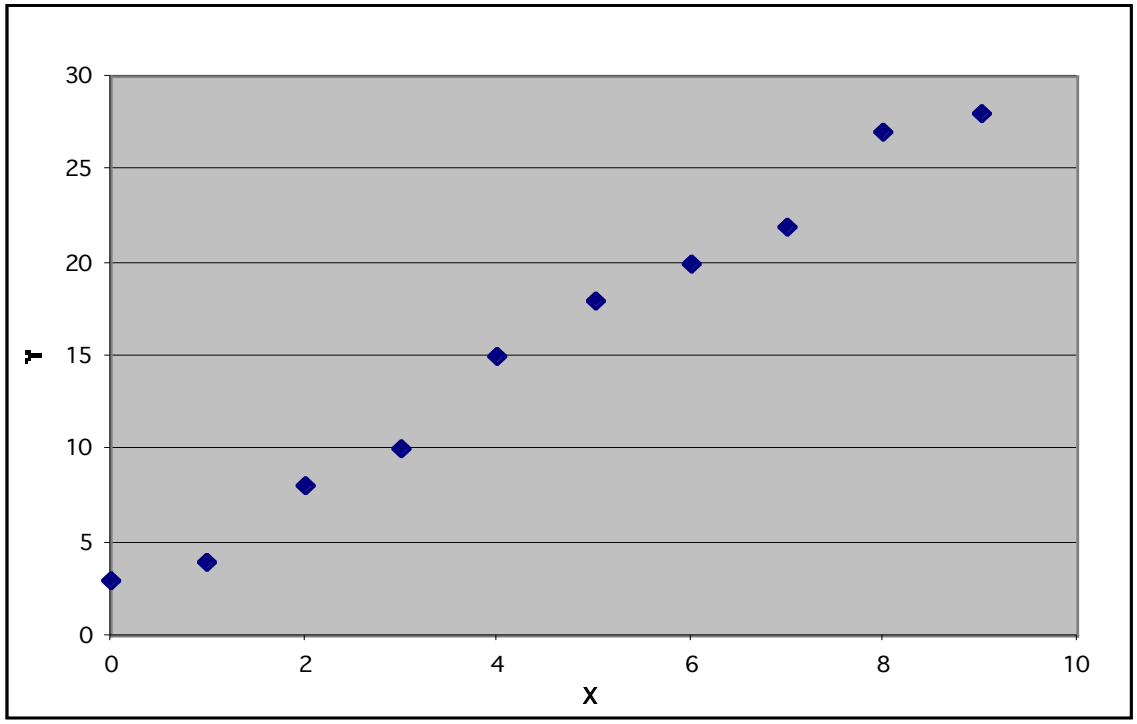
Model selection

- The key aspect is that a model should be selected
  - not to just model the data (train set)
  - but to model the (real-world) process that is generating this data.
- Only then can we be sure that the model will generalise to other samples from the same (real-world) process.

Model selection

- This is best illustrated using a diagram.
- Consider a database with two fields, X and Y, with values shown in the following scatter plot:

There appears to be a linear relationship between the values of X and Y (with noise).



Model selection

- We can safely assume that the real-world process that **generated** this data is linear:

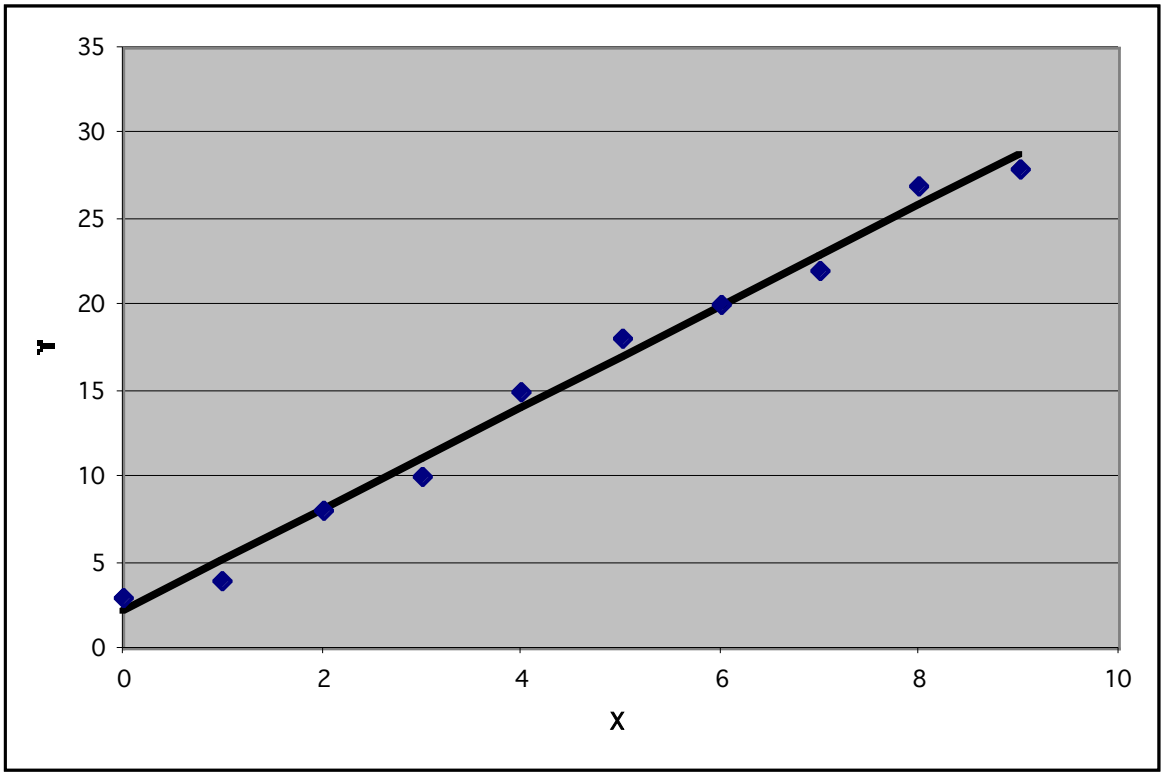
$$Y=aX+b$$

- Thus the **model** we have selected to represent the process is a linear one.
- Here, a and b are parameters to be chosen so that the line is a ‘best fit’ to the data.
- This is the **optimisation** part.

Model selection

- In this simple case, we use a least squares regression process to determine **best** values of a and b.

In fact
a=2.96
and
b=2.15



Model selection

- In data mining, the process is no different to the one just described.
- A model is selected, based upon LLTs and objectives of the KDD project.
- It is during the DM algorithm stage, when the algorithm (model) is applied to the train data set, that the best parameters for the model are determined (optimisation).
- Examples of models are neural nets, rules, induction trees.

Data mining

- Hence, data mining consists of the **application** of the **selected algorithms** and subsequent **parameter** optimisation to obtain best possible model.
- Initial estimation of model fit is assessed on the training data, the test data must not be compromised during parameter optimisation.
- A validation set may be used to assess fit of model to new data.

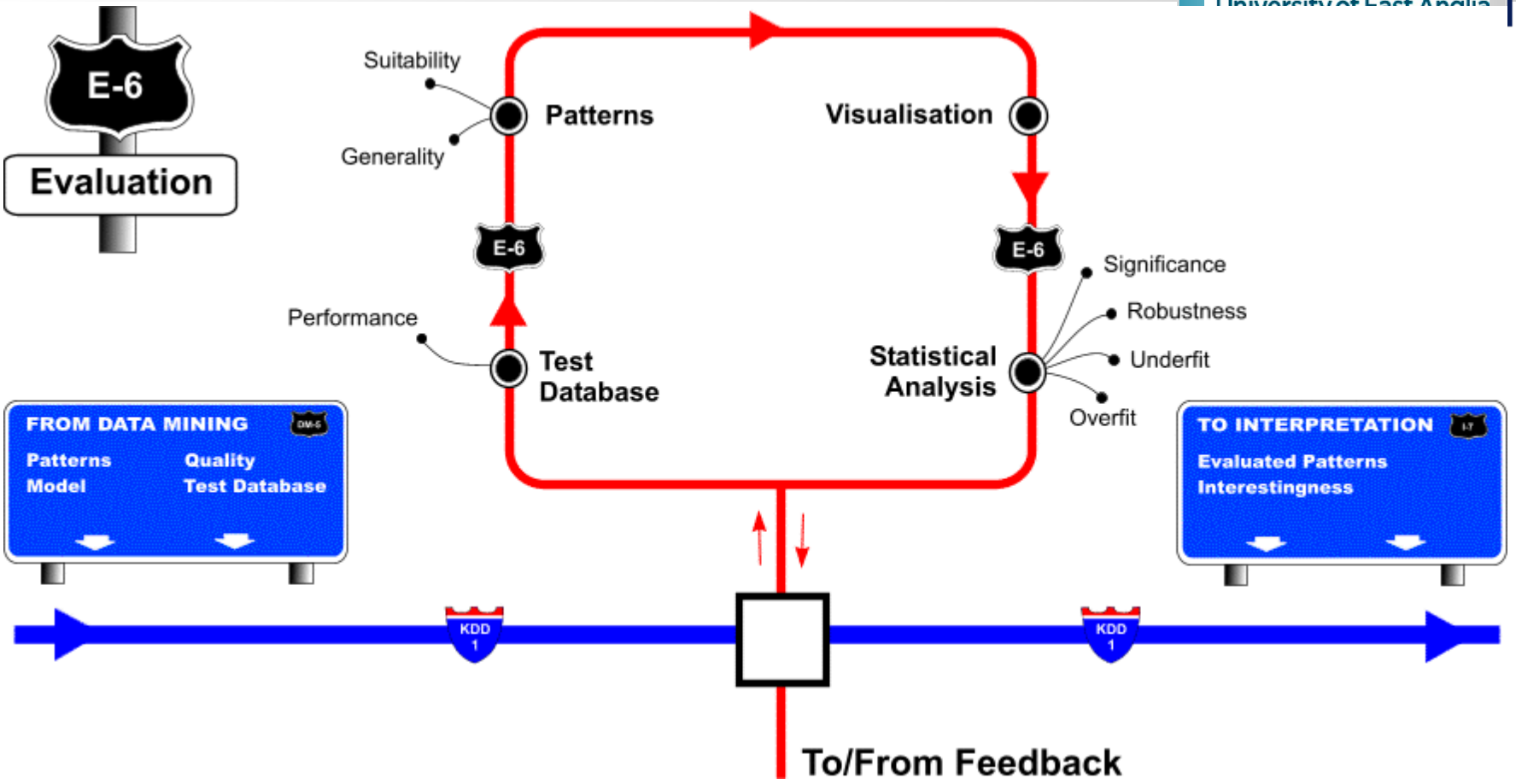
Data mining

- Alternative methods to check model fit include
  - cross validation
  - bootstrapping
- The criteria to be measured may change from model to model, but involves the notion of **interest** of the knowledge discovered.
- The criteria should be decided early in the project, at the time of model selection.

Interest

- Accuracy
- Coverage
- Simplicity
- Novelty
- Surprise
- Efficiency
- Many more...

Evaluation


E-6
Evaluation
Suitability
Generality
Patterns
Visualisation
E-6
E-6
Performance
Test Database
Statistical Analysis
Significance
Robustness
Underfit
Overfit
FROM DATA MINING
Patterns
Model
Quality
Test Database
TO INTERPRETATION
Evaluated Patterns
Interestingness
KDD 1
KDD 1
To/From Feedback

Evaluation

- This is the phase in which we use the **test set** to evaluate performance.
- Visualisation techniques may also be used to further evaluate the model.
- Statistical tests can be applied to test if model/pattern is significant.
- A robust model should have no signs of **over-fitting** and should be resilient to small variations in conditions.

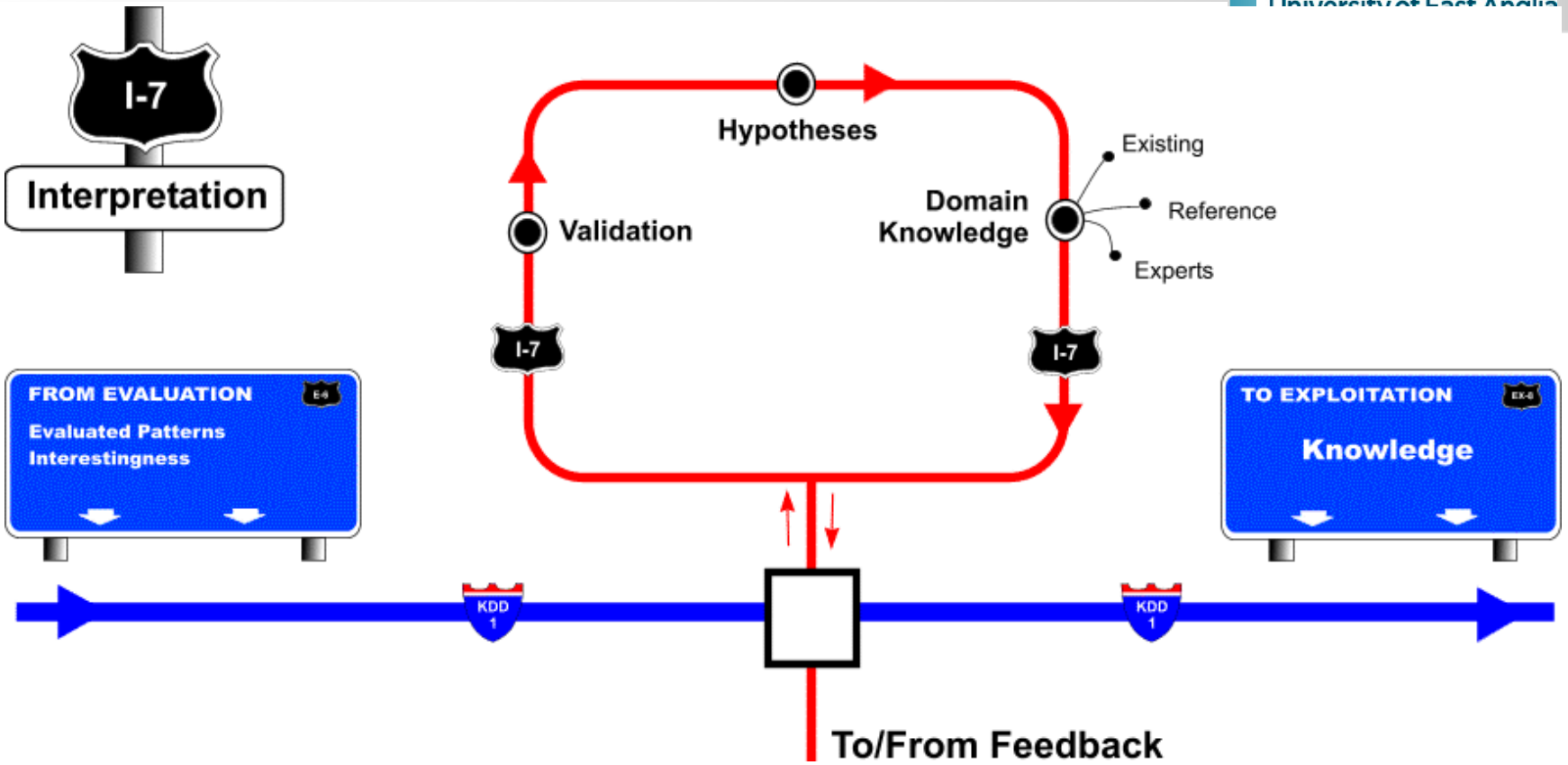
Model overfitting

- A model that **overfits** the data is one that represents closely the training data (the data it was build from) but has no generalisation powers to new data.
- For example,

IF patient-id = 12345 THEN diagnosis= X

will be 100% accurate on the training data, but will have made no useful generalisations.

Interpretation



Producing knowledge

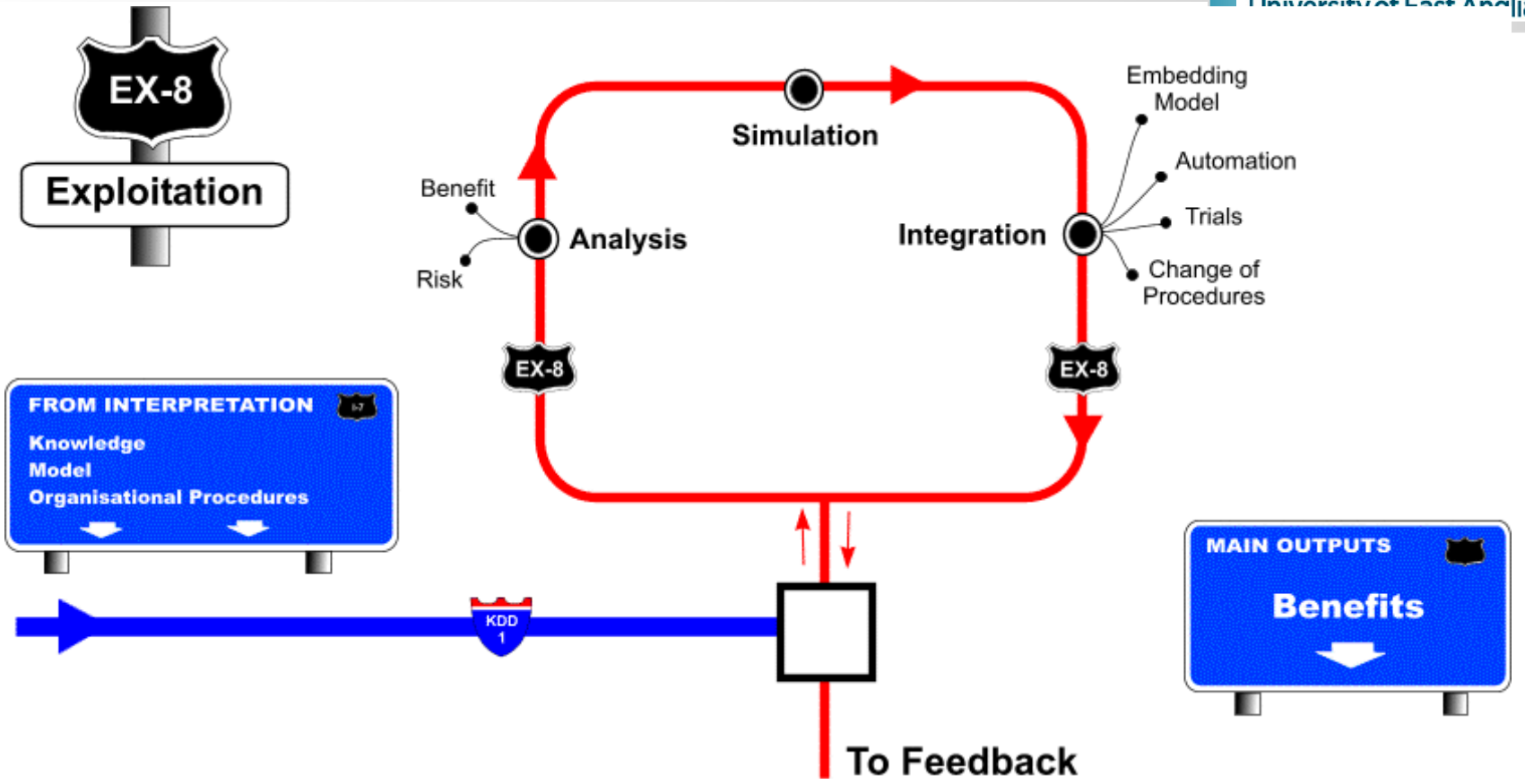
- At this last stage, the domain experts need to assess the patterns/models obtained against their own background knowledge
- The models may be
  - Invalid (flukes in the data?)

    IF male and … THEN post-natal depression=YES
  - Valid but already known facts (uninteresting)

    IF young driver and fast car THEN high-risk= YES
  - Valid and previously unknown = **Knowledge**

IF young driver and fast car and car age>25 THEN high-risk= NO

Exploitation



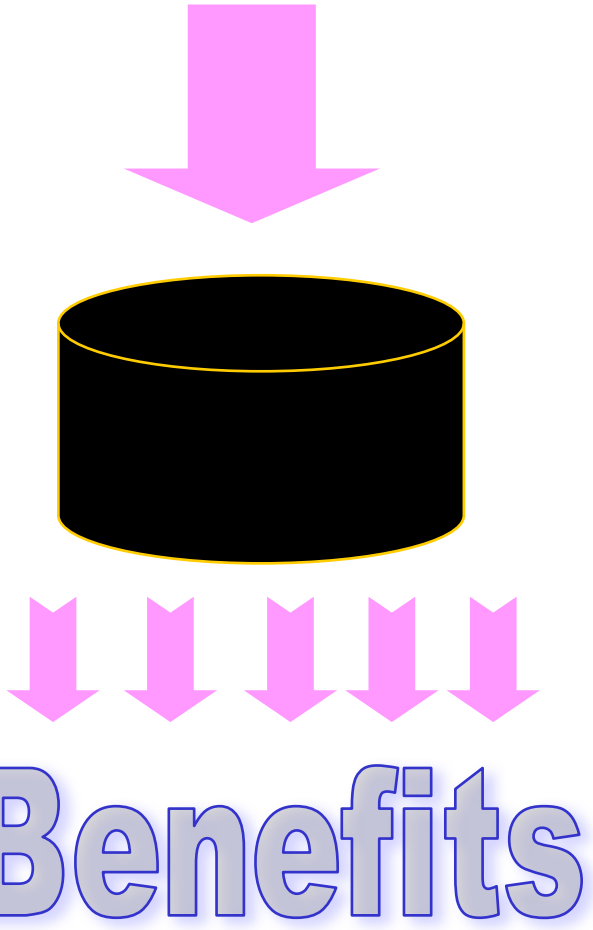
Making it happen

- This is the **most important** phase of a commercial project.
- Knowledge obtained must be used and translated into **benefits**.
- May need to present cost-benefit and risk analysis for the integration of new knowledge into existing procedures.
- May require simulations, if change of procedure involves high risk.

Incremental learning

- If the underlying data changes we can either
  - adjust the model for the new data (incremental learning)
  - build a new model from scratch.
- Most KDD systems do not support very well incremental learning.
- May need to automate the KDD process for the particular application so it can often be repeated easily.

Data



Discussion time

Sex	Country	Age	Buy?
M	France	25	Yes
M	England	21	Yes
F	France	23	Yes
F	England	34	Yes
U	France	30	No
M	Germany	107	No
M	Germany	20	No
F	Germany	18	No
F	?	34	No
M	France	55	No

Freitas and Lavington (1998) Data Mining with EAs, CEC99.

- How do we begin to analyse this data?

Learning outcomes

- What is DM and KDD?
- What are the phases of a KDD project?
- What operations take place in each phase?
- What are the ingredients for success in a KDD project?

Sources of information

- KDD Roadmap:

J. C. W. Debuse, B. de la Iglesia, C. M. Howard and V. J. Rayward-Smith (2000). Building the KDD roadmap: A methodology for knowledge discovery. In R. Roy (Ed.), Industrial Knowledge Management. Springer-Verlag, London.

- Alternative CRISP-DM methodology:

P. Chapman, J. Clinton, J.H. Hejlesen et al. (1998). The current CRISP-DM process model for data mining [online]. Available: http://www.crisp-dm.org.

Books

Hongbo Du, Data Mining Techniques and Applications, Cengage Learning, 2010.

Pang-Ning Tan, Michael Steinbach and Vipin Kumar. *Introduction to Data Mining*, Addition Wesley, Boston, 2006.

Margaret H. Dunham, *Data Mining - Introductory and Advanced Topics*, Prentice Hall, New Jersey, USA,2003.

Sholom M. Weiss and Nitin Indurkhya, *Predictive Data Mining: a practical guide*, Morgan Kaufmann Publishers Inc., San Francisco, California, 1998.

Doryan Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann Publishers Inc., San Francisco, 1999.

Books

Ian H. Witten and Eibe Frank, *Data Mining*, Morgan Kaufmann Publishers Inc., San Francisco, 2000.

Jiawei Jan and Micheline Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers Inc., San Francisco, 2001.

Richard J. Roiger and Michael W. Geatz, *Data Mining: A tutorial-based primer*, Pearson Education LTD, 2003.

W. Klosgen and J. Zytkow, *Handbook of Data Mining and Knowledge Discovery*, Oxford University Press, 2002.

Conferences

- ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
- IEEE International Conference in Data Mining
- SIAM International Conference on Data Mining
- DAWAK: International Conference on Data Warehousing and Knowledge Discovery
- PAKDD: Pacific-Asia Conference on Knowledge Discovery and Data Mining
- PKDD: European Conference on Principles and Practice of KDD

More conferences

- VLDB: International Conference on Very Large Databases
- ICML: International Conference on Machine Learning
- UAI: Conference on Uncertainty in Artificial Intelligence
- IJCAI: International Joint Conference on Artificial Intelligence

Journals

- Data Mining and Knowledge Discovery (Kluwer)
- IEEE Transactions on Knowledge and Data Engineering
- Intelligent Data Analysis (Elsevier)
- Journal of Intelligent Information Systems (Kluwer)
- Machine Learning (Kluwer)
- Applied Intelligence (Kluwer)

Web sources

- http://www.kdnuggets.com/
  - This page is probably one of the most informative and up-to-date pages for KDD. Many links here.
- http://www.acm.org/sigkdd/
  - ACM Special Interest Group on Knowledge Discovery in Data and Data Mining