Data Preparation for Data Mining

Lesson 1

Lesson 1 Overview: Introduction to Data Preparation for Data Mining

-
  - Motivation and Importance
  - Data Preparation as a Part of the Data Mining Process
  - Exploring the Problem Space
  - Exploring the Solution Space
  - The Nature of the World and How It Impacts Data Preparation

Motivation

- A variety of resources on databases and data warehouses
- A variety of resources on data mining tools and algorithms
- Gap between identifying the data and building models

Importance

- “Garbage in, garbage out”
- A crucial step of the DM process
- Need to know what to do with the “dirty data” – after it is collected and before it is mined

Data Preparation

- Could be the difference between a successful data mining project and a failure
- Could take 60-80% of the whole data mining effort

Definition

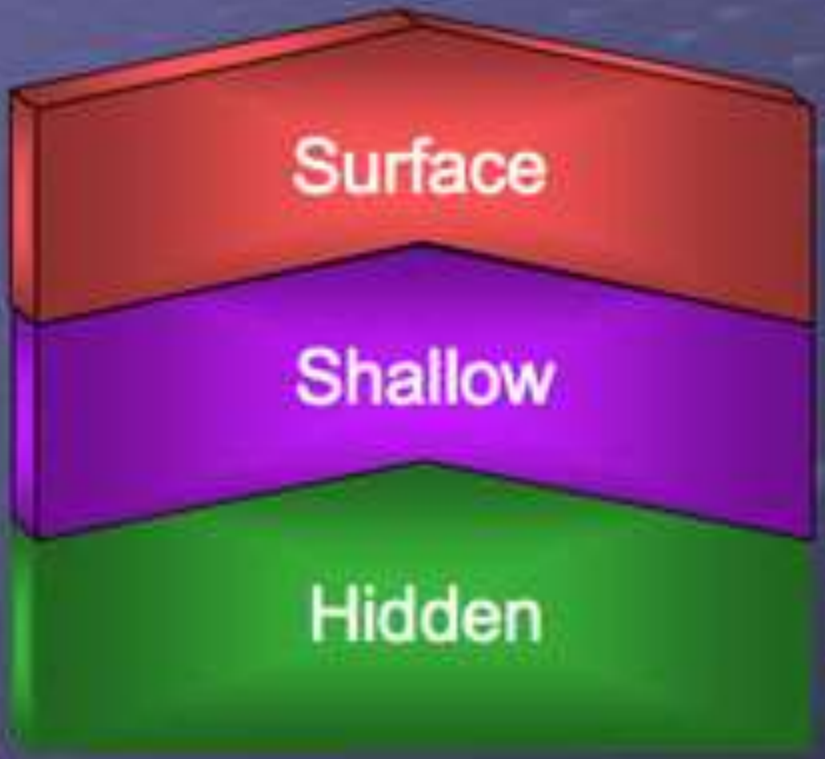
- Data Preparation is a process of cleaning, filtering and organizing the data for successful mining and modeling, by solving or avoiding problems in the data, and presenting the data to the modeling schema in the optimal way.

Good Data Preparation Practice

- When data is properly prepared and surveyed:
  - high quality modeling results are more likely
  - the quality of models produces will depend mostly on the content of the data, not so much on the modeler's expertise level.

Data Mining Refresher

Top-Down Methodology



Bottom-Up Methodology

Analytical Tools

SQL for simple queries and reporting

Statistics & OLAP for summaries and analysis

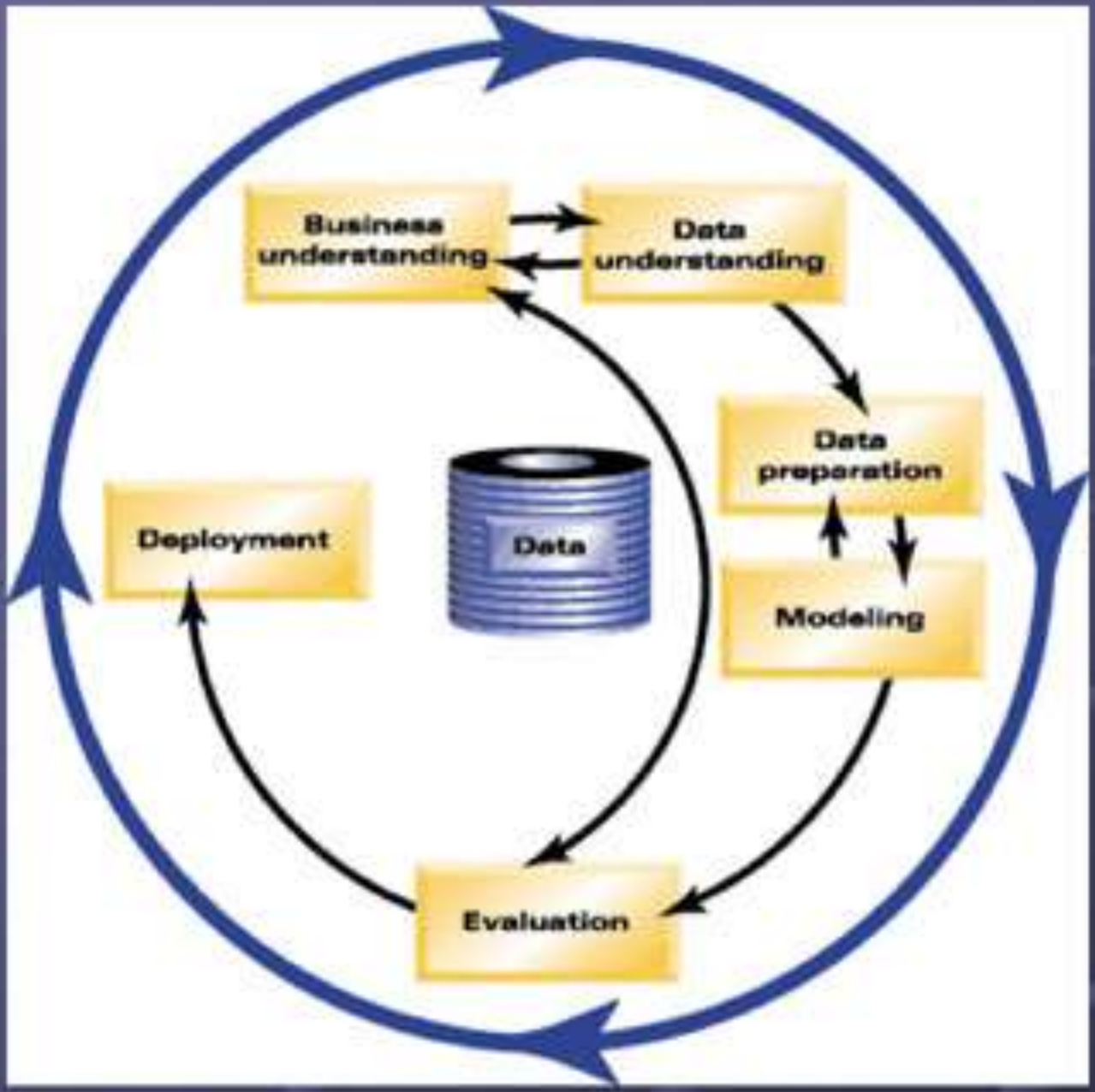
Data Mining for knowledge discovery

The Process of Data Mining

- Basic steps:
  - Data Acquisition
  - Data Preparation
  - Modeling
  - Evaluation
  - Feedback

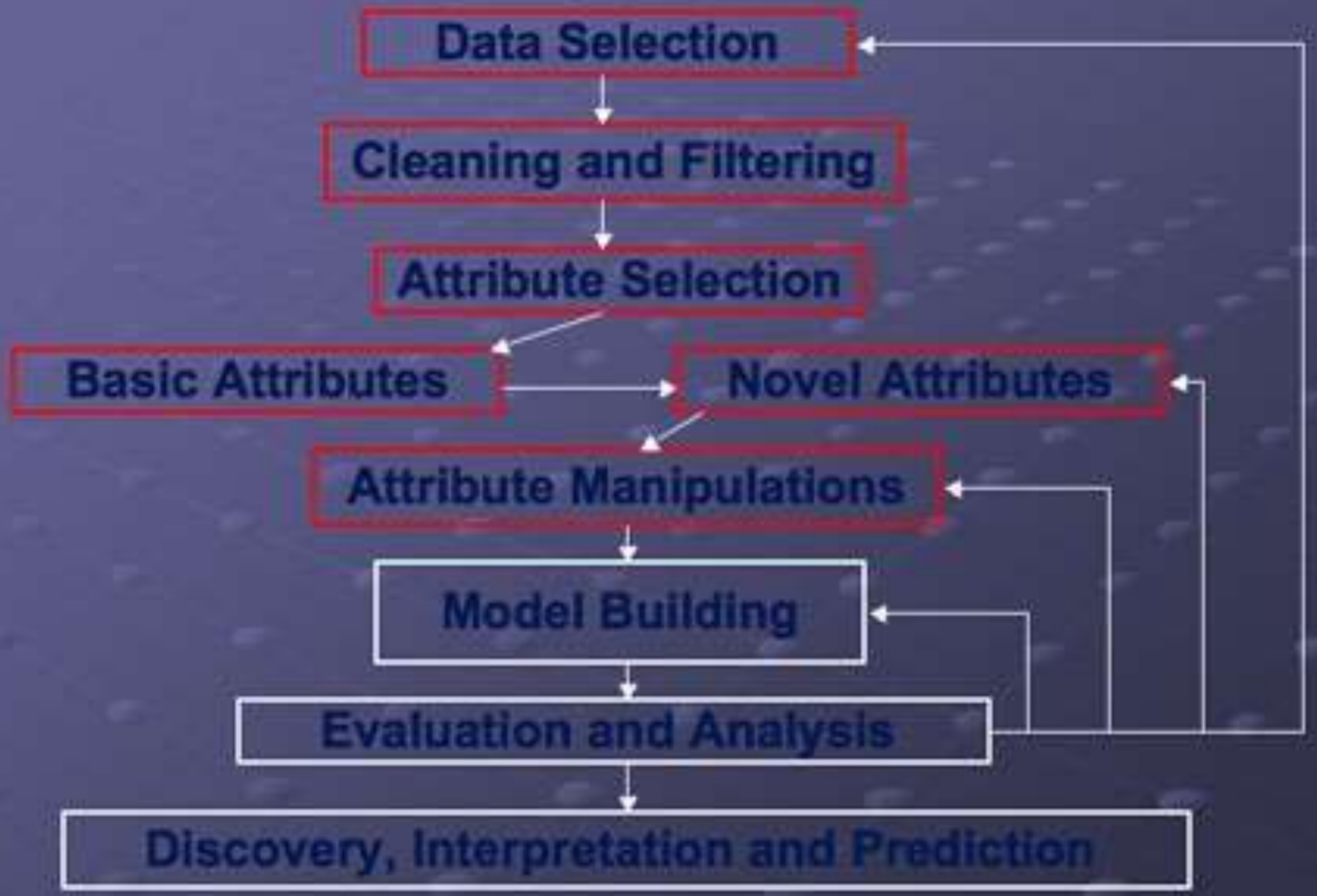
CRISP-DM Methodology

- Cross Industry Standard Process for Data Mining
  - http://www.crisp-dm.org/
- Six Phases:
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment


Business understanding
Data understanding
Data preparation
Modeling
Evaluation
Deployment
Data

The Details of the DM Process

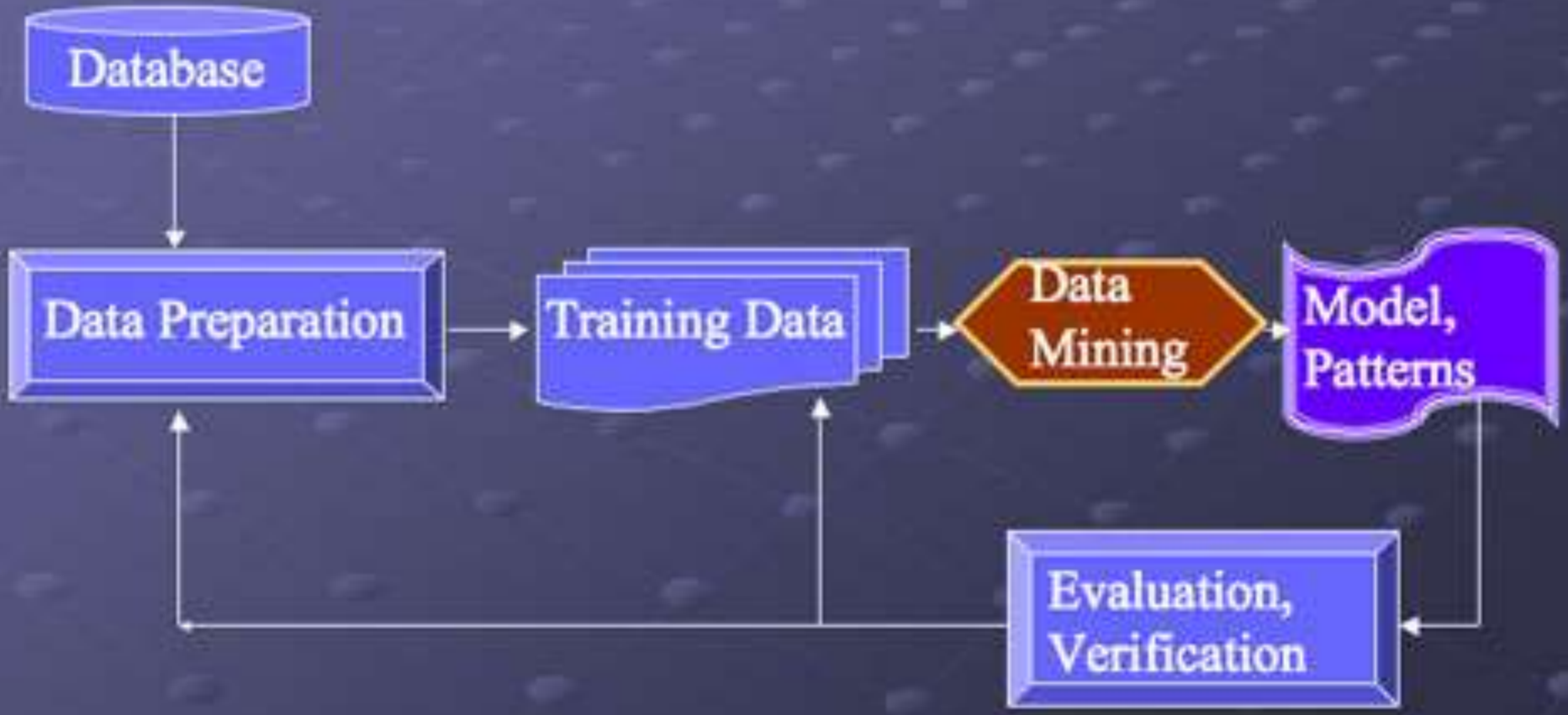
Data Preparation in the DM Process



The Data Mining Process

-
- Iterative Nature
- Exploratory Process
- Highly tailored to the dataset
- Need for Fine-Tuning
- Need for Model Revision from time to time

KDD Process


Database
Data Preparation
Training Data
Data Mining
Model, Patterns
Evaluation, Verification

The “Dark Side” of Data

- Missing values (null = empty or something else)
- Dirty data (erroneous zip codes, etc.)
- Inconsistent values (different revisions)
- Duplicate values
- GIGO (Garbage In, Garbage Out)

Prerequisites

- Data understanding:
  - Descriptors, values, ranges, labels
- Data history
- Exploring the Problem Space
- Exploring the Solution Space
- Implementation Method Specification
- Familiarity with the Nature of the World

Exploring the Problem Space

- A crucial starting point
- Avoids any possible misconceptions and unrealistic expectations from DM project

- A MUST: *identify the right problem to solve*

Identifying the Right Problem

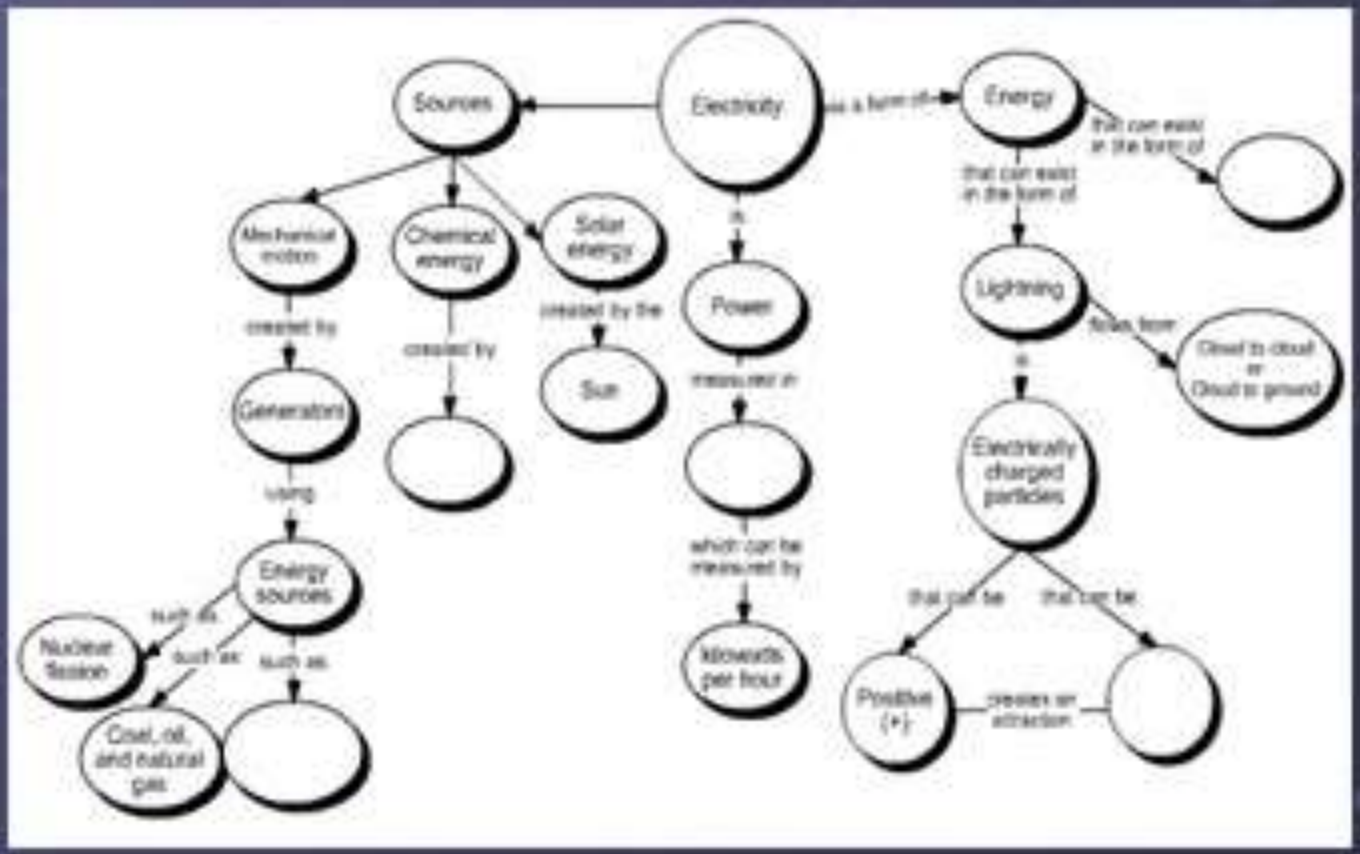
- Example: predicting the churn or stopping it? (unemployed customers over 80 of a telecommunication company had a most regrettable tendency to churn – they died and no incentive program had much impact on them!)

Problem Definition Methodology

- Start by defining problems in a precise way, no general statements
- Example: identifying failure rates on the manufacturing assembly
  - What constitutes a failure
  - How is a failure detected and measured
  - Why are some failure rates seen as a problem
  - What components need to be looked at (equipment, personnel etc)

Problem Definition Methodology

- Use tools such as:
  - Cognitive maps



Problem Definition Methodology

- Use tools such as:
  - Ambiguity resolution (when there are alternative interpretations, any assumptions are explicated)
  - Problem matrix (uses intuitive ranking to create the final rank)

Problem	Importance	Difficulty	Yield	Final Rank
A	5	3	2	3.72
…				

Results

- The best mix of precisely defined problems to solve
- The problems are ranked appropriately
- Next: what would the solution look like?

Example

- Problems: a range of problem concerning fraudulent activity in branch offices
- What output is desired?
  - What are the driving factors of the fraud?
  - Where is the easiest point in the system to detect it?
  - The most cost-sensitive measures to stop it?
  - Which activity patterns are most likely fraudulent?

Exploring the Solution Space

- List of precise problems and proposed solutions, and their associated ranking
- Proposed solutions should be:
  - Precise
  - Complete
  - Real-world
  - “Implementable”

Back to the Example

- “A computer model capable of modeling one million transactions per minute, scoring each with a value estimating the probability that this is fraudulent activity, routing any transaction above a specific threshold to an operator for manual intervention”

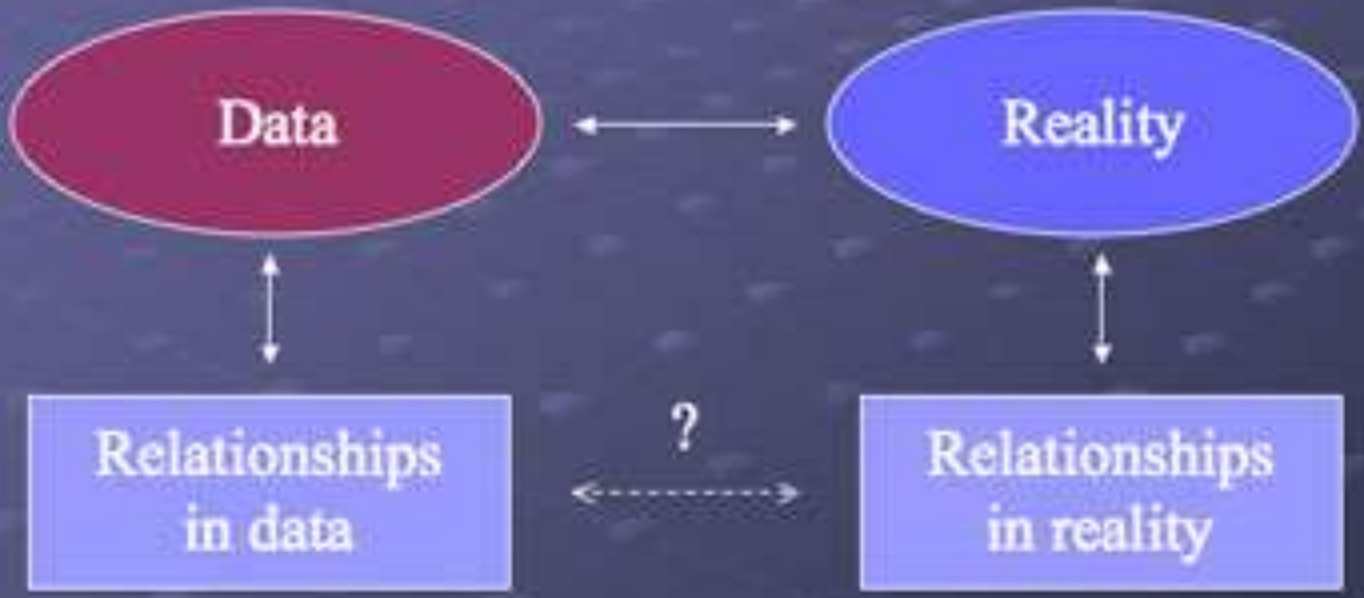
Another Example

- A company wanted improved response to their mailed catalogs
- Discovering what they really wanted was harder than creating the model!
- How was response to be measured? How was the result to be used? No clear deliverables. No precise problem/solution definition.
- Their true goal: optimize the value per page of the catalog.

Specifying the Implementation Method

- Defines how the solution will actually be applied in practice
- Specify the six "w's":
  - who, how, what, when and where (why is already covered in the problem specification)
- "Buy-in" needed from both "problem owners" and "problem holders" (those who control the resources that allow the solution implementation)

Data vs. World



Basic assumption in data mining.

Measuring the World

- World usually perceived as objects
- Objects are associated with properties and relations with other objects
  - a car has wheels, seats, color, length etc.
- Measurement freezes the world at a validating feature
  - Timestamp usually is the validating feature

Errors of Measurement

- Noise (precision) vs. bias (calibration)
- Environmental errors
  - due to the nature of interaction between variables
  - give important information to miners
- Sensitivity to changing conditions
  - bank account balance vs. income
  - estimating limits essential in modeling
- *Distortion* a better word for laymen

Types of Measurements

- Measurements differ in their nature and the amount of information they give
- Scalar vs. Nonscalar
- Qualitative vs. Quantitative

Types of Measurements

- Nominal scale
  - Gives unique names to objects
  - No other information deducible
  - Example: names of people

Types of Measurements

- Nominal scale
- Categorical scale
  - Names categories of objects
  - Although maybe numerical, not ordered
  - ZIP codes, cost centers

Types of Measurements

- Nominal scale
- Categorical scale
- Ordinal scale
  - Measured values can be ordered naturally
  - Transitivity: $(A > B) \& (B > C) \Rightarrow (A > C)$
  - “blind” tasting of wines

Types of Measurements

- Nominal scale
- Categorical scale
- Ordinal scale
- Interval scale
  - the scale has a means to indicate the distance that separates measured values
  - temperature

Types of Measurements

- Nominal scale
- Categorical scale
- Ordinal scale
- Interval scale
- Ratio scale
  - measurement values can be used to determine a meaningful ratio between them
  - bank account balance

Types of Measurements

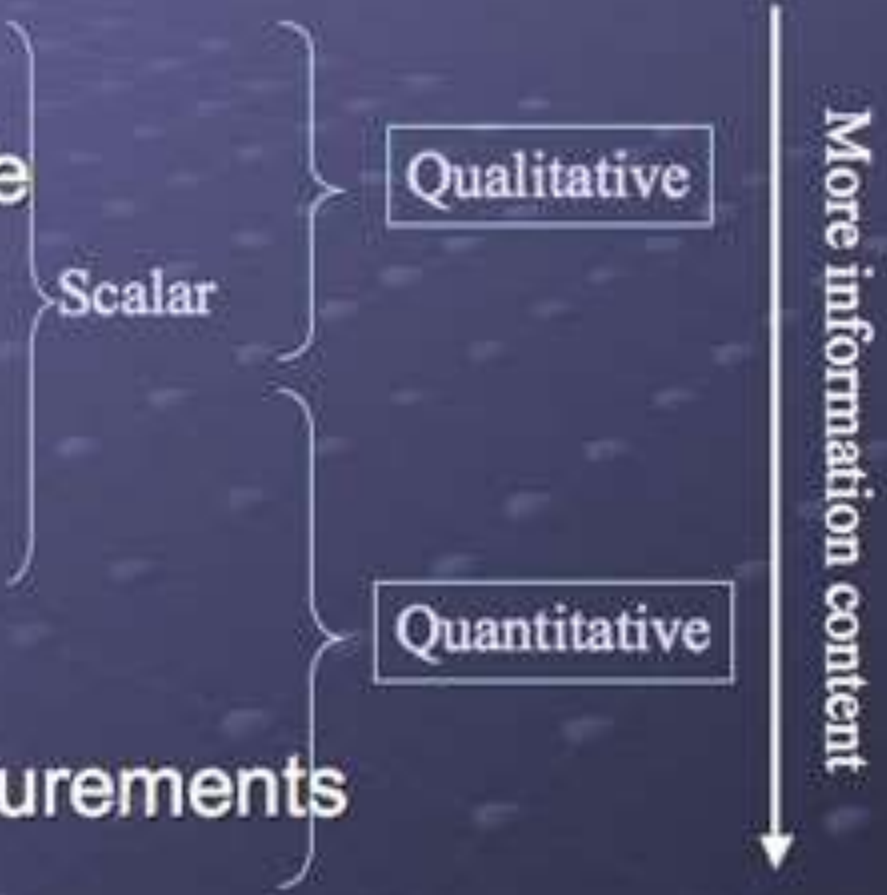
- Nominal scale
- Categorical scale
- Ordinal scale
- Interval scale
- Ratio scale

- Nonscalar measurements
  - vector: a collection of scalars
  - nautical velocity

Types of Measurements

- Nominal scale
- Categorical scale
- Ordinal scale
- Interval scale
- Ratio scale

- Nonscalar measurements



Continua of Variable Attributes

- The qualitative-quantitative continuum
- The discrete-continuous continuum

Continua of Variable Attributes

- The qualitative-quantitative continuum
- The discrete-continuous continuum
  - single-valued variables = constants
    - days in week, inches in a foot

Continua of Variable Attributes

- The qualitative-quantitative continuum
- The discrete-continuous continuum
  - single-valued variables = constants
  - two-valued variables
    - gender: male/female
    - empty and missing values
    - binary variables: “0 or 1”, “true or false”, “active or inactive”, “healthy or diseased”, “fraudulent or non-fraudulent”

Continua of Variable Attributes

- The qualitative-quantitative continuum
- The discrete-continuous continuum
  - single-valued variables = constants
  - two-valued variables
  - other discrete variables
    - difference between discrete and continuous?
    - Is bank account balance discrete or continuous?
    - Salary groups: salary variable becomes discrete?

Continua of Variable Attributes

- The qualitative-quantitative continuum
- The discrete-continuous continuum
  - single-valued variables = constants
  - two-valued variables
  - other discrete variables
  - continuous variables

Data representation

- Data set: a collection of measurements for several variables
- Superstructure of the data set: underlying assumptions and choices

Dealing with variables

- Variables as objects
  - try to figure out the features of each variable:
    - Values, min, max, range etc.
  - gain insight into the behavior of variables
  - know how they were created and/or modified

Dealing with variables

- Variables as objects
- Removing variables
  - entirely empty or constant variables can be discarded
  - Decide what to do with duplicates, if any
  - Find and deal with incorrect values
  - beware of scarcity

Dealing with variables

- Variables as objects
- Removing variables
- Sparcity
  - only a few non-empty values available, but these are significant
  - sparse data problematic for mining tools
  - dimensionality reduction may help

Dealing with variables

- Variables as objects
- Removing variables
- Sparcity
- Monotonicity
  - increasing without bound
  - datestamps, invoice numbers
  - new values never been in the training set

Dealing with variables

- Variables as objects
- Removing variables
- Sparcity
- Monotonicity
- Increasing dimensionality
  - ZIP to latitude and longitude

Dealing with variables

- Variables as objects
- Removing variables
- Sparcity
- Monotonicity
- Increasing dimensionality
- Outliers
  - values completely out of range

Dealing with variables

- Variables as objects
- Removing variables
- Sparcity
- Monotonicity
- Increasing dimensionality
- Outliers
- Numerating categorical variables
  - natural ordering must be retained!
  - Day, half-day, half-month, month

Dealing with variables

- Variables as objects
- Removing variables
- Sparcity
- Monotonicity
- Increasing dimensionality
- Outliers
- Numerating categorical variables
- Anachronisms

Building mineable data sets

- Make things as easy for the tool as possible!
  - Expert knowledge of the algorithms needed
- Exposing the information content
  - if you know how to deduce a feature, do it yourself and don't make the tool find it out
  - to save time and reduce noise
  - i.e. include relevant domain knowledge
    - Understanding of the data, problem and solutions space needed

Building mineable data sets

- Make things as easy for the tool as possible!
- Exposing the information content
- Getting enough data
  - Do the observed values cover the whole range of data?
  - Combinatorial explosion of features
    - Is a lesser certainty enough? Makes problems tractable.

Building mineable data sets

- Make things as easy for the tool as possible!
- Exposing the information content
- Getting enough data
- Missing and empty values
  - to fill in or to discard?

Building mineable data sets

- Make things as easy for the tool as possible!
- Exposing the information content
- Getting enough data
- Missing and empty values
  - to fill in or to discard?
- Shape of the data set