### Data preparation

### Data Quality

* Problems

− garbage in, garbage out

* + DQ problems cost hundreds of billion $$$ each year.
  + Resolving data quality problems is often the biggest

effort in a data mining study

* 75% of 599 companies had economic losses because of data quality problems ( ‘Price Waterhouse Coopers’ survey in New York 2001)
* If the underlying data is not accurate, any relations based on this data will be misleading.
* The cost of bad or “dirty” data exceeds $600

billion annually. (insidearm.com 2008)

* Wrong price data in retail databases along cost US consumers $2.5 billion annually. (DW Review Magazine 2000)

#### Dell in Taiwan (PC World.com 2009)

* 19-inch LCD monitors for NT$500 ($15.26) each
* This price had been available for 8 hours (11pm

on 25 June to 7am on 26 June 2009)

* 26,000 people placed orders for 140,000 monitors

#### Amazon UK (computerweekly.com2009)

* Hewlett-Packard Pocket PCs for £7 each (retail

price £290)

* This price had been available only for just under 1 hour (19 March 2009)
* Some shoppers thought to have placed orders for 50 or 60 of the devices at a time.

#### Other Examples

* Kodak advertised it's digital cameras for £100 (worth £329). (honoured)
* Argos (a catalogue giant in UK) advertised a type of TVs for £3 (worth £300). (1999)
* William Hill “refuse to pay” after a student wins

£1,000 from 50p bet on Roger Federer at French

Open 2015. (The Independent, 04 June 2015)

* Data Quality – How to improve data quality
* Data Cleaning – is required whenever there is any level of doubt existing.
* Data Cleaning is a costly and time consuming task, esp. in VLDBs
* Questions:
  + How to represent the wide variety of dirty data?
  + How to select and apply appropriate algorithms?
  + How to reduce the time used for data cleaning?
  + How to improve the degree of automation when performing data cleaning?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **2000** | **1** | **60** | **l** | **1** |
| 2015 | 4 | 3o | 6 | 3 |

* Can we interpret the data?
  + What do the fields mean?
  + What is the key? The measures?
* Data glitches
  + Typos, multiple formats, missing / default values
* Metadata and domain expertise
  + Column two is Road Type. What does the 1 mean?
  + Column five is the number of accidents. Is it only 3 cases or 3k cases, or other unit?
* A multi-dimensional concept. It’s dimensions:
  + Accuracy – conformity of the recorded value with the actual value
  + Timeliness – the recorded value is not out of date
  + Completeness – all values for a certain variable are recorded
  + Consistency – the representation of data is uniform in

all cases

* + Uniqueness – the property that there is only one record for each unique entity

***Error*** – Any one of the dimensions is violated

Regarding both Data and Meta-Data

* Lack of validation routines
* Valid, but not correct
* Mismatched syntax, formats and structures.
* Unexpected changes in source systems
* Lack of referential integrity checks
* Poor system design
* Data conversion errors
* Data Integration errors

Aim: detect, correct errors and inconsistencies from

data in order to ensure data quality

* Identification of error types
* Selection of Algorithms
* Selection of Methods/Approaches
* Outliers – are observations that are unexpectedly

different from the majority in the sample.

* Duplicates – are two or more tuples representing

the same entity

* Misspelling
* More types -- 33 +

*Better understanding helps*

### Error Detection

|  |  |
| --- | --- |
| **Error Data Type** | **Algorithm** |
| **EDT1** | **ALG11, ALG12, ALG13, …** |
| **EDT2** | **ALG21, ALG22, ALG23, …** |
| **EDT3** | **ALG31, ALG32, ALG33, …** |
| **…** | **…** |
| **EDTn** | **ALGn1, ALGn2, ALGn3, …** |

**Detection Methods**

* Statistical Methods
* Data Transformation
* Integrity Constraint enforcement
* Duplicate Elimination
* Data Mining Techniques – Association rules

***Time Consuming***

### Outlier

* A value or data object that is far away or very

different from all or most of other data a university student aged 65

* Single attribute
* Multidimensional data
* Methods
* Visualisation techniques
* Clustering
* Distance-based
* Projection-based

### Missing Values

* Ignorance/Deletion – simply remove records with at least one missing value
* Imputation – replace the missing value by an estimated value, such as the most popular one/mean, or a value determined by other attributes
* Leave it as missing, but some methods/tools can’t deal with missing values

### Duplicate Detection

* To Identify records in the same or different

databases that refer to the same real-world entity.

* It is also called *record linkage* or *record matching*
* Field Matching Techniques – deal with single

record

* Character-based Similarity Methods Smiht  Smith
* Token-Based Similarity methods

John Smith vs Smith, John

* Duplicate Records – deal with multiple records

### Select Data

* Key – Only select data that is relevant for the

given problem

* Record selection – create a subsample of the data and use it for the analysis, instead of the whole dataset
* Reasons for record selection
* Timeliness – in a changing world to use recent data
* Representativeness – Sometimes the population of interest is different from the population available in the database
* Rare events – predicting that the event will not occur
* How large does a subsample have to be?

# Prepare Data in Practice

* Different software takes different format of data as input
* Data format accepted by Weka
  + Files with the following format:

WEKA’s ARFF format, having a .arff extension

CSV format, a .csv extension

C4.5 format, having a .data and .names extension

serialized Instances objects, having a .bsi extension

– Database -- Reads data from a database. (Note that to make this work you might have to edit the file in weka/experiment/DatabaseUtils.props.)

# Weka’s arff format

* A dataset has to start with a declaration of its name:

@relation name

* followed by a list of all the attributes in the dataset

(including the class attribute)

@attribute attribute\_name specification

* If an attribute is nominal, specification contains a list

of the possible attribute values in curly brackets:

@attribute nominal\_attribute {first\_value, second\_value, …}

* If an attribute is numeric, the keyword real is used

@attribute numeric\_attribute real

* actual data is introduced by a statement

@data

### weather data in arrf (part)

@relation weather

@attribute outlook {sunny, overcast, rainy}

@attribute temperature real

@attribute humidity real

@attribute windy {TRUE, FALSE}

@attribute play {yes, no}

@data sunny,85,85,FALSE,no sunny,80,90,TRUE,no overcast,83,86,FALSE,yes rainy,70,96,FALSE,yes

**References**

* F. Hoppner and F. Klawonn

Guide to Intelligent Data Analysis, Spriner-Verlag, 2010

– Section 4.2, 6.1.3, 6.2

* Online materials about Google-Refine
* A. Elmagarmid, P. Ipeirotis, and V. VeryKios, “Duplicate Record Detection: A Survey,” IEEE Trans. Knowl.Data Eng., Vol.19, No.1, pp. 1-16, 2007