Compact Course in Data Mining

Data Mining Process (CRISP-DM)

Professor Dr. Gholamreza Nakhaeizadeh

Data Mining Process

CRISP-DM :

- Provides an overview of the life cycle of a data mining project

- Consists of six phases

- was partially funded by the European Commission

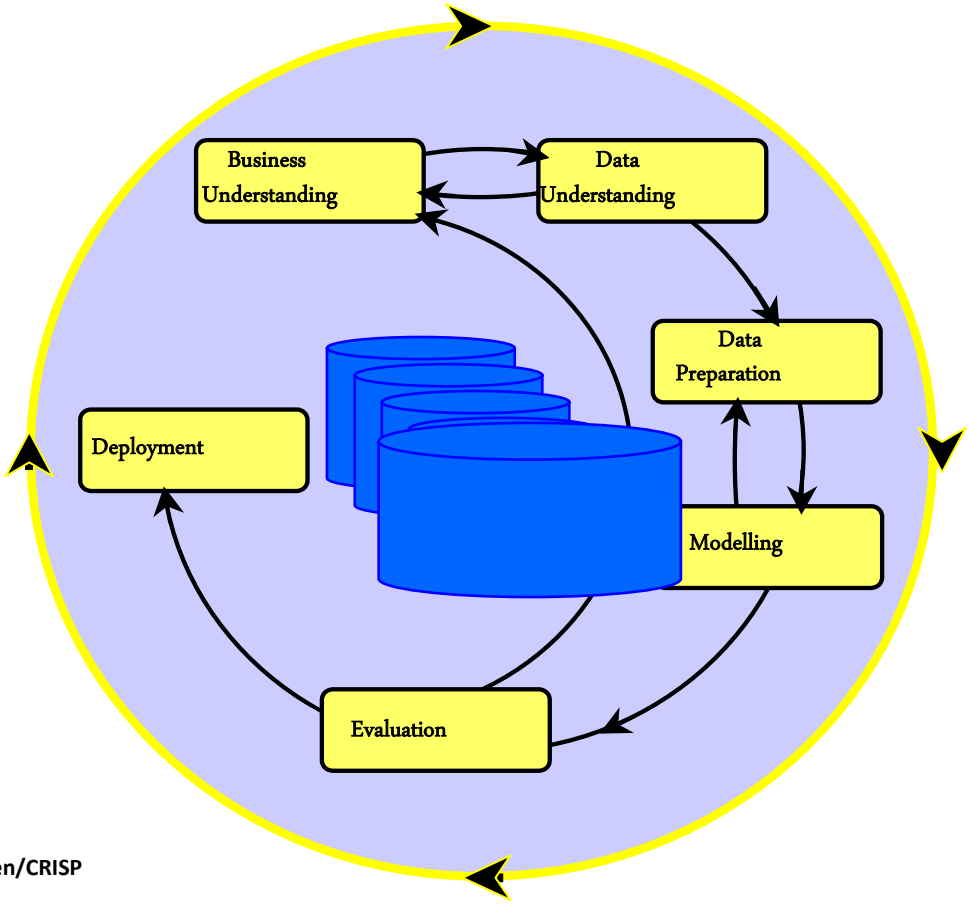
Project Partner:

Teradata a division of NCR

DAIMLERCHRYSLER

SPSS

van de mensen van OHRA

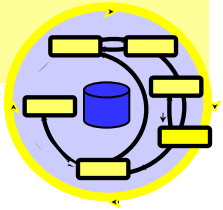


- CRISP-DM Process Model is described in:

ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf

ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/de/CRISP-DM.pdf

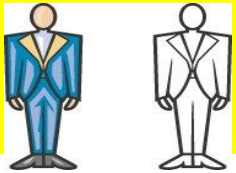
CRISP-DM ,Task Identification



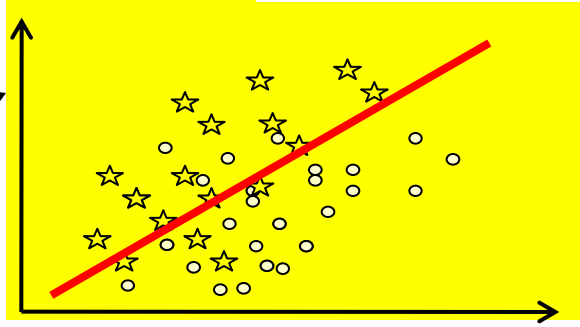
Classification

Credit Scoring

- Good customer
- Bad customer



Clustering



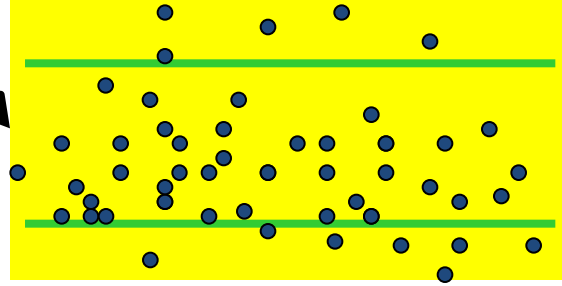
Business Challenge

Concept Description

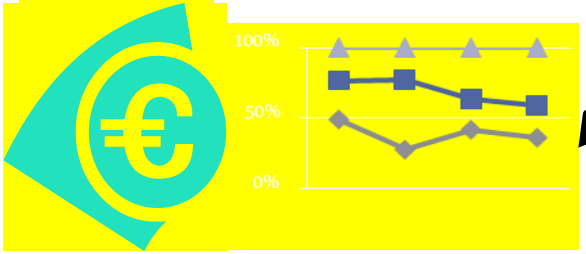
Customers Loyalty :

- Age
- Income
- Education
- ....

Deviation detection



Prediction

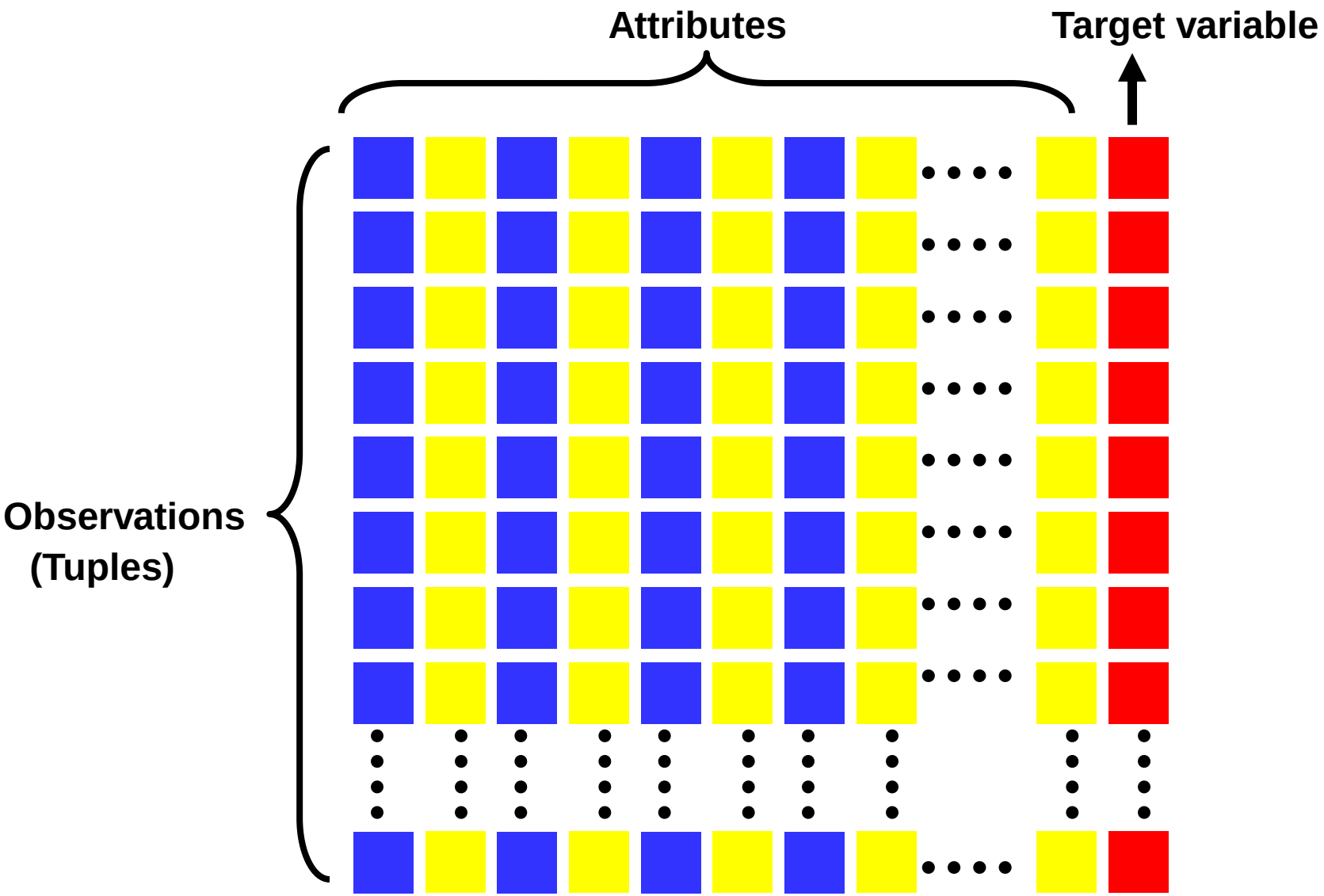


Dependency Analysis

A and B → C

Sequence pattern

Supervised and unsupervised learning



Supervised Learning

Nr.	A1	A2	A3……	An	T
1	a11	a12	a13	a1n	t1
2	a21	a22	a23	a2n	t2
3	a31	a32	a33	a3n	t3
.					
.					
.	..	….	….	….	….
.					
.					
.					
m	am1	am2	am3	amn	tm

Examples for Supervised Learning : Classification, Prediction

Unsupervised Learning

Nr.	A1	A2	A3……	An
1	a11	a12	a13	a1n
2	a21	a22	a23	a2n
3	a31	a32	a33	a3n
.				
.				
.	..	….	….	….	….
.				
.				
.				
m	am1	am2	am3	amn

Example for Unsupervised Learning: Clustering, Association Rules

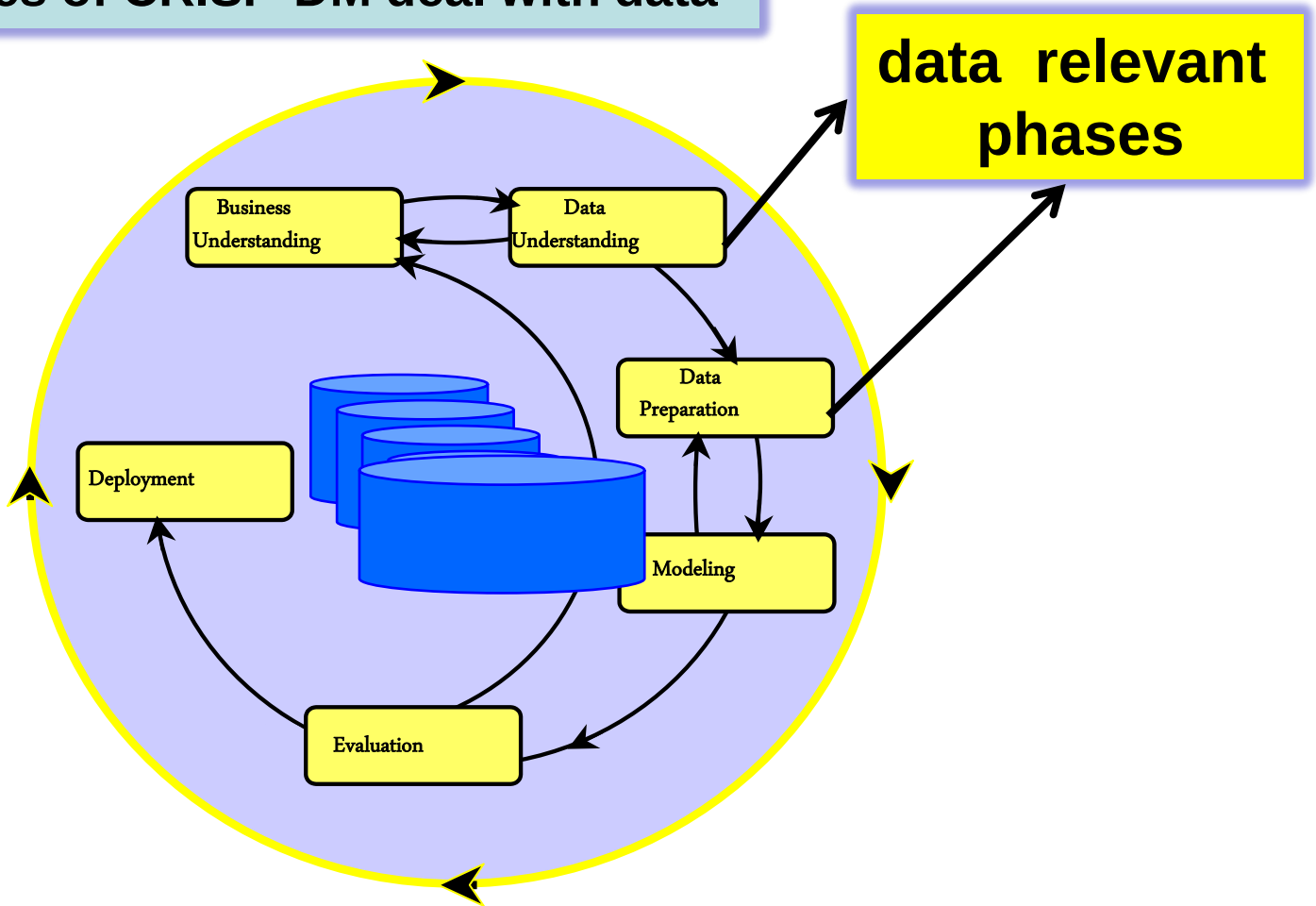
Data Mining Algorithms

- **Supervised Learning**
  - Tree Family
    - Decision Trees
    - Regression Trees
    - Model Trees
  - Artificial Neural Networks
  - K-Nearest Neighbors
  - Naïve Bayes
  - Linear Regression
  - ……..

- **Unsupervised Learning**
  - Clustering
    - K-Means
    - K-Medoids
    - ……..
  - Association Mining
    - Association Rules
    - Sequence Mining

Data Mining Process, Role of Data in Data Mining

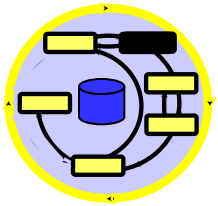
- No Data , no Data Mining
- Two phases of CRISP-DM deal with data



Data Mining Process

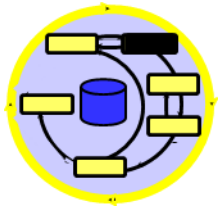
CRISP-DM: Data Understanding

General aspects



- Collect initial data
- Describe data
- Explore data
- Verify data quality

Data Understanding

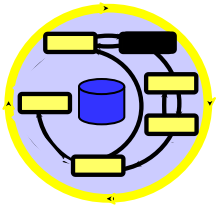


- What are the data we need ?
- Where are the data we need?
- Determining type, structure and quality of data
- Descriptive data summarization

Data Mining Process

CRISP-DM: Data Understanding

Collecting initial data



what are the needed data ?

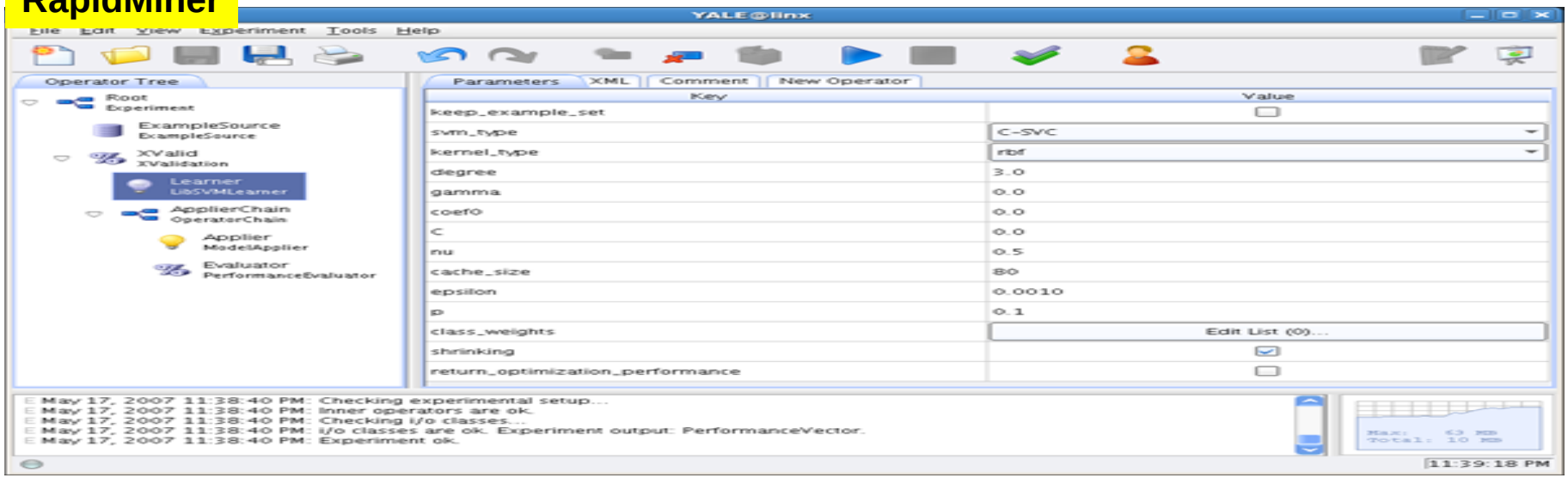
- **where are the data ?**
- Flat Files
- Databases
- Heterogeneous Databases
- Connected autonomous databases
- Legacy Databases

– Inherited from languages, platforms, and techniques earlier than current technology

- Data warehouse

Practical Work: data sources (main page)

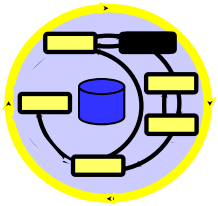
RapidMiner



Data Mining Process

CRISP-DM: Data Understanding

Collecting initial data

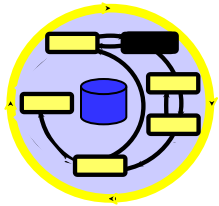


what are the needed data ?

- where are the data ?
- Can the data be accessed effectively and efficiently ?
  - How big is the needed storage ?
  - How long does it take to access the data ?
- Is there any restriction in collecting the data ?
  - privacy issues,
  - too expensive data,
  - too expensive collecting process,..
- ……

Data Mining Process

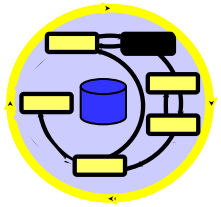
CRISP-DM: Data Understanding



Examples of data sources (1)

UCI KDD Database Repository for large datasets used machine learning and knowledge discovery research.
UCI Machine Learning Repository.
Delve, Data for Evaluating Learning in Valid Experiments
FEDSTATS, a comprehensive source of US statistics and more
FIMI repository for frequent itemset mining, implementations and datasets.
Financial Data Finder at OSU, a large catalog of financial data sets
GeneSifter Data Center, access to microarray datasets through the GeneSifter microarray data analysis system.
GEO (GEO Gene Expression Omnibus), a gene expression/molecular abundance repository supporting MIAME
compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.
Grain Market Research, financial data including stocks, futures, etc.
Investor Links, includes financial data

Source: http://www.kdnuggets.com/datasets/



Examples of data sources (2)

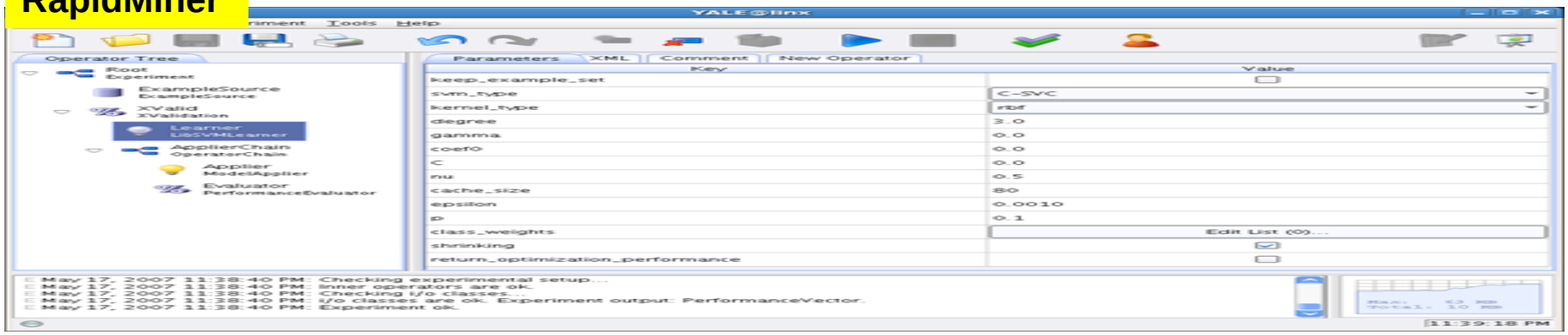
http://people.stern.nyu.edu/padamopo/blog/2014-02-20-data-sources.html

Microsoft's TerraServer, aerial photographs and satellite images you can view and purchase.
MIT Cancer Genomics gene expression datasets and publications, from MIT Whitehead Center for Genome Research.
National Government Statistical Web Sites, data, reports, statistical yearbooks, press releases, and more from about 70
web sites, including countries from Africa, Europe, Asia, and Latin America.
National Space Science Data Center (NSSDC), NASA data sets from planetary exploration, space and solar
physics, life sciences, astrophysics, and more.
PubGene(TM) Gene Database and Tools, genomic-related publications database
SMD: Stanford Microarray Database, stores raw and normalized data from microarray experiments.
SourceForge.net Research Data, includes historic and status statistics on approximately 100,000 projects and
over 1 million registered users' activities at the project management web site.
STATOO Datasets part 1 and part 2
UCR Time Series Data Mining Archive, offering datasets, papers, links, and code.
United States Census Bureau.

Source: http://www.kdnuggets.com/datasets/

Practical Work (Meta Data)

RapidMiner

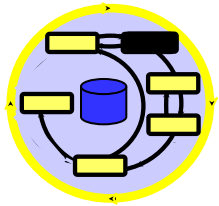


Load Churn

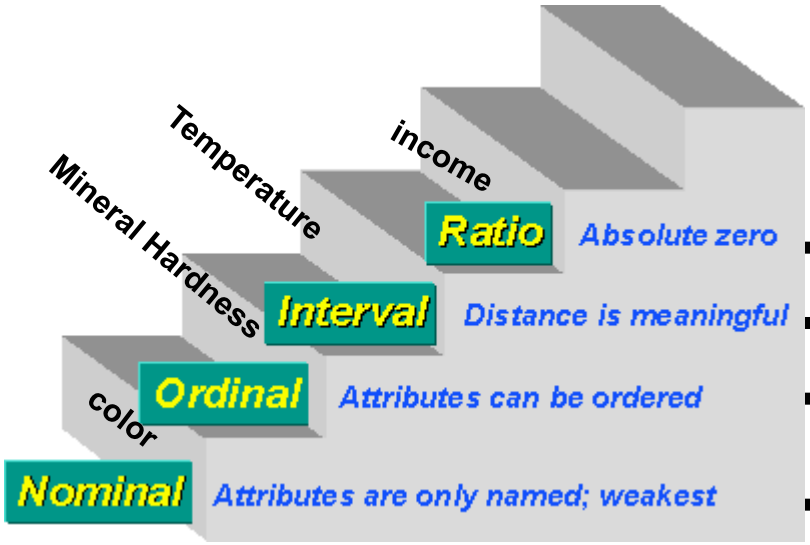
Data Mining Process

CRISP-DM: Data Understanding

Describing data



Attribute type



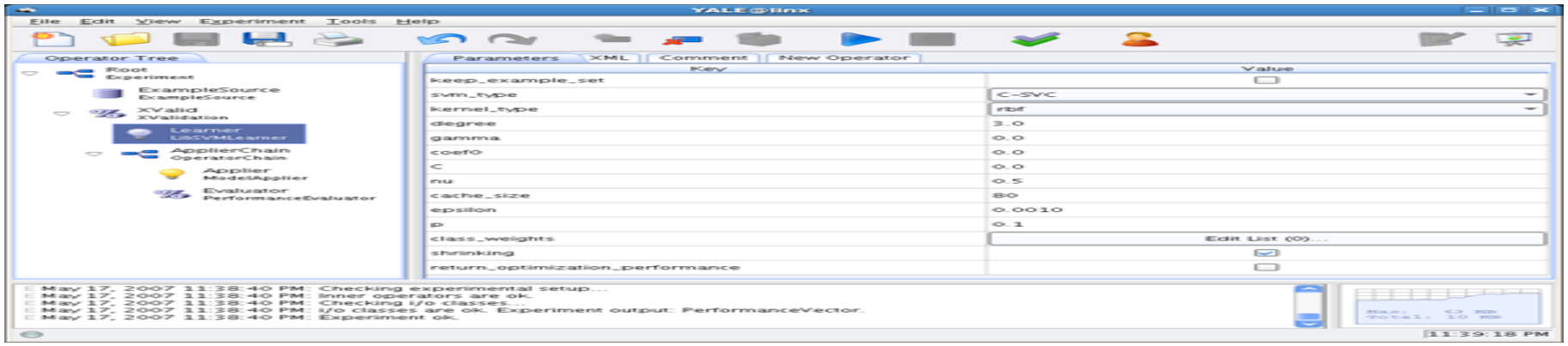
Source: http://www.socialresearchmethods.net/kb/measlevl.php

Meanigful are:

- Multiplication, devision (*, /), (-), (> , <)(= ≠)
- Difference (-), (> , <), (= ≠)
- Greater, les (> , <), (= ≠)
- Equality, inequality (= ≠)

Practical Work : Attribute Type

RapidMiner

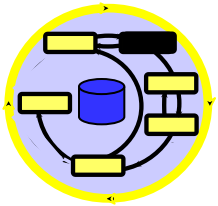


Use: Churn

Data Mining Process

CRISP-DM: Data Understanding

Describing data



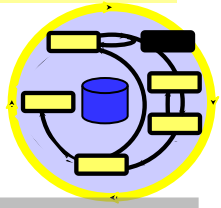
Attribute type : another classification

- **Discrete Attributes**
  - Have a finite or countable infinite set of values
  - Examples: number of children , counts
  - Often represented as integer variables
  - Special case of discrete attributes : binary attributes

- **Continuous Attributes**
  - Have real numbers as attribute values
  - Examples: Income, sales , weight

Data Mining Process

CRISP-DM: Data Understanding



Data Type

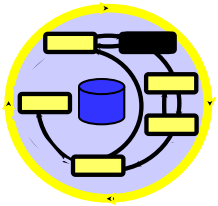
- Cross-Section data
- Time Series data
- Panel data
- Sequences
  - Postman Routes
  - Web Click Streams

- Data Streams
  - Infinite volumes
  - Dynamically Changing
  - Real time processing
- Spatial data
- Spatiotemporal data
- Text data
- web data
- Multimedia data

Data Mining Process

CRISP-DM: Data Understanding

Verifying data quality



The real world data are often "dirty", data "Cleaning" is needed

- **Is data accurate ?**
  - noisy data
  - Outliers
- **Is data complete ?**
  - missing values
- **Is data consistent ?**
  - Coding Errors
- **Is data sufficiently up-to date ?**
  - Timeliness
  - .....

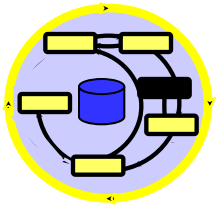
Most important

You can find more criterion describing data quality e.g. in:

http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-3-540-33172-8?cm_mmc=Google-_-Book%20Search-_-Springer-_-0

Data Mining Process

CRISP-DM: Data Preparation



- Select data
- Clean data
- Transfer data
- Integrate data

The Role of Data in Data Mining

Data preparation (preprocessing)

Data Selection

– Data reduction

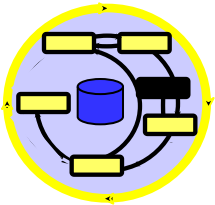
– Various sampling methods

– Attribute reduction

- Selection a subset of attributes
- Forward selection
- Backward selection
- Generating new attributes
- PCA approach

–Data quality management

- What should we do with missing values ?
- Outlier detection
- …….

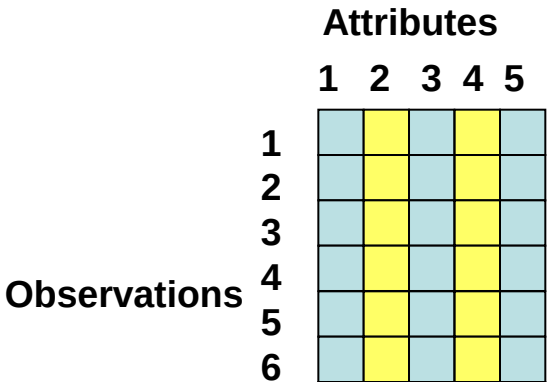


Data Selecting

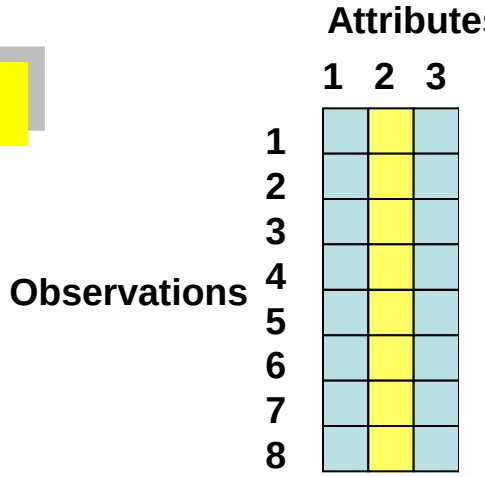
Observation Reduction

- Sampling
- Intelligent Sampling
- Learn to forget

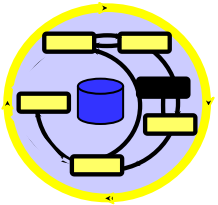
.....



Attribute Reduction



Data Selecting



Observation Reduction : Sampling

Statisticians: Sampling because *obtaining* the entire dataset (population) is too expensive or time consuming (often they *do not have* the data and start collecting)

Data Miners: Sampling because *processing* of the population is too expensive or time consuming (often they *have* the data)

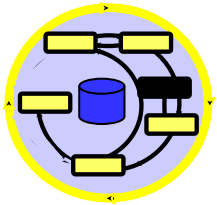
good sample ~ representative sample

↓

has nearly the same property as the population :

→

sample mean is very close to population mean
sample variance is very close to population variance
……



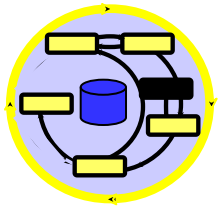
Data Selecting

Observation Reduction : Sampling

Task: Choose a sampling method that with high probability leads to a representative sample

↓

- Choosing the right sampling technique
- Choosing the right sample size



Data Selecting

Observation Reduction : Sampling technique

Random sampling: Equal and known probability of being selected for each member of the population

General aspects:

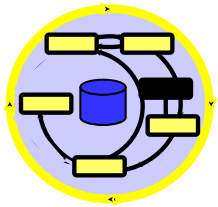
- **Sampling without replacement (s.wo.r)**
- **Sampling with replacement (s.w.r.)**

→ **During the sampling process the probability of selecting any objects remains constant**

↓

Analyzing is easier

Data Selecting



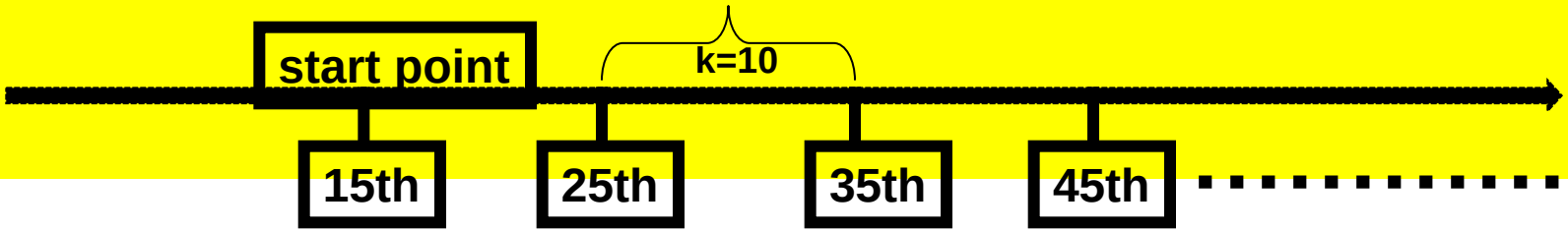
Observation Reduction : Sampling technique

Systematic Sampling (called also kth name selection method)

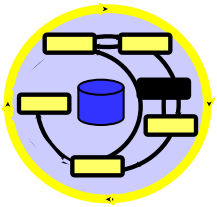
- Selection of k; k= population size / sample size (k sampling interval)
- Selection of a start point
- Selection of every kth member as sample

Example: Population size = 2000 sample size = 200

- k=10
- start point = member number 15
- then sample consists of members number 15, 25, 35, 45,…



Data Selecting



Observation Reduction : Sampling technique

Stratified Sampling

Population consists of different mutually exclusive subgroups (strata) varying considerably in size.
Examples: (120 men, 30 women), (1900 employment, 100 unemployment), (300 white, 20 black)

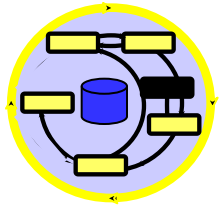
↓

Random sampling can fail to adequately represent the members with low frequency

↓

Solution: Stratified Sampling: Random sampling in each Subgroup (stratum) independently

Data Selecting



Observation Reduction : Sampling technique

Stratified Sampling Strategies

Stratified sampling strategies

1. **Number of members drawn from each subgroupa is proportional to the size of that subgroup**

2. **Equal numbers of members are drawn from each subgroup even though the gropus are of different sizes**

Example: Size of population 2000: 1900 employment, 100 unemployment
Size of needed sample: 50

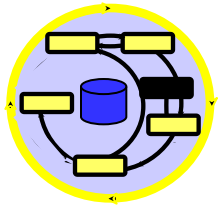
Strategy 1 : 50/2000 = 1/40 1900 * 1/40 = 47,5 100 * 1/40 = 2,5
Sample consists of 47 employment and 3 unemployment

Strategy 2 : Sample consists of 25 employment and 25 unemployment

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



Sampling technique

Bootstrap Sampling

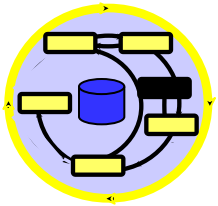
- Bootstrap is sampling with replacement from a dataset.
- Bootstrap sampling relies on its own sample as often the only resources a researcher has
- The name may come from phrase “pull up by your own bootstraps” which mean ‘rely on your own resources'

Source: http://people.revoledu.com/kardi/tutorial/Bootstrap/index.html

Data Mining Process

CRISP-DM: Data Preparation

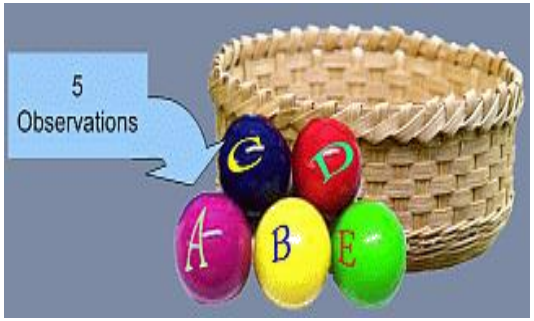
Data Selecting



Sampling technique

Bootstrap Sampling

How to do that? Suppose your sample contain only 5 observations. You label your observation into 5 balls with name A, B, C, D, E

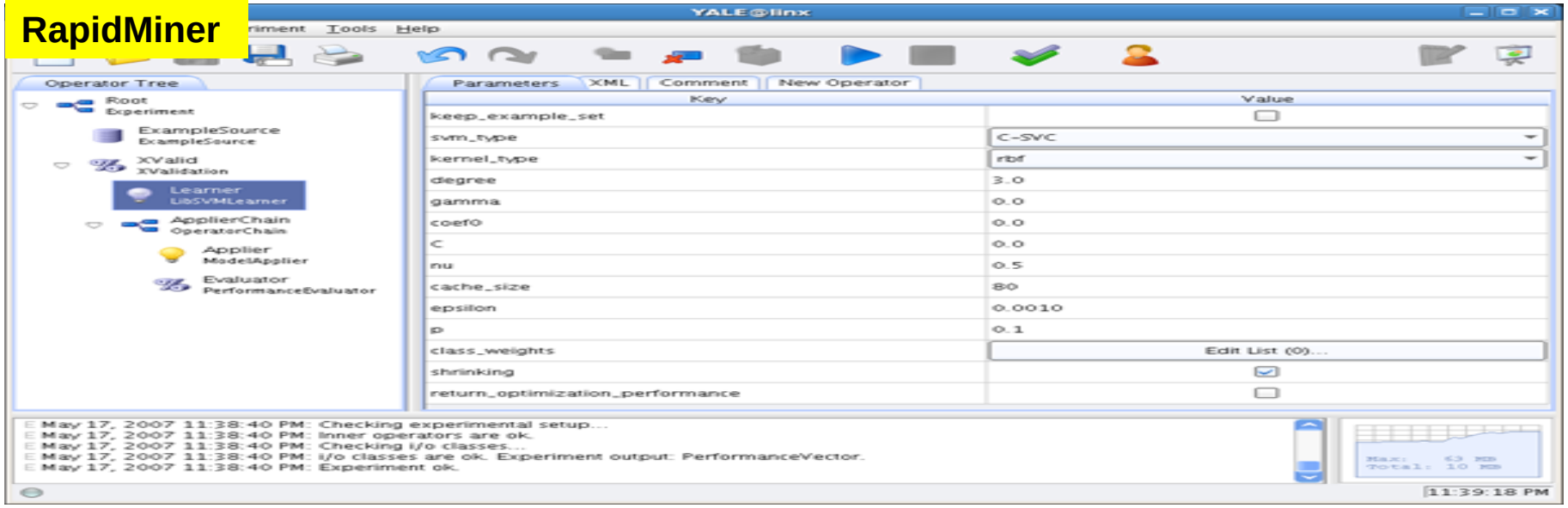


Put all the 5 balls on a basket. Then, from these 5 balls, you draw 1 ball randomly and record the name. After you record it, put back this ball in the basket. Make sure that you return the ball in the basket before making another random draw. This is sampling with replacement. Repeat the work of draw another ball randomly, record the label and put back the ball to the basket until N of time. The recorded labels are called bootstrap sampling.

Source: http://people.revoledu.com/kardi/tutorial/Bootstrap/index.html

Practical Work: Sampling

RapidMiner



Load: German Credit_Tr
Use: sample size = 100

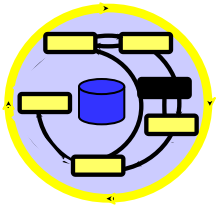
1. **Sampling in Preprocessing Data**
   - **Bootstrap**
   - **Sampling**
   - **Stratified**
2. **Example Range Filter**

→ **Analyze the class ratio Class 1/Class 2**

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



Attributes

Observations	1	2	3	4	5
1					
2					
3					
4					
5					
6					

Observation Reduction

- Sampling
- Intelligent Sampling
- Learn to forget

.....

Attribute Reduction

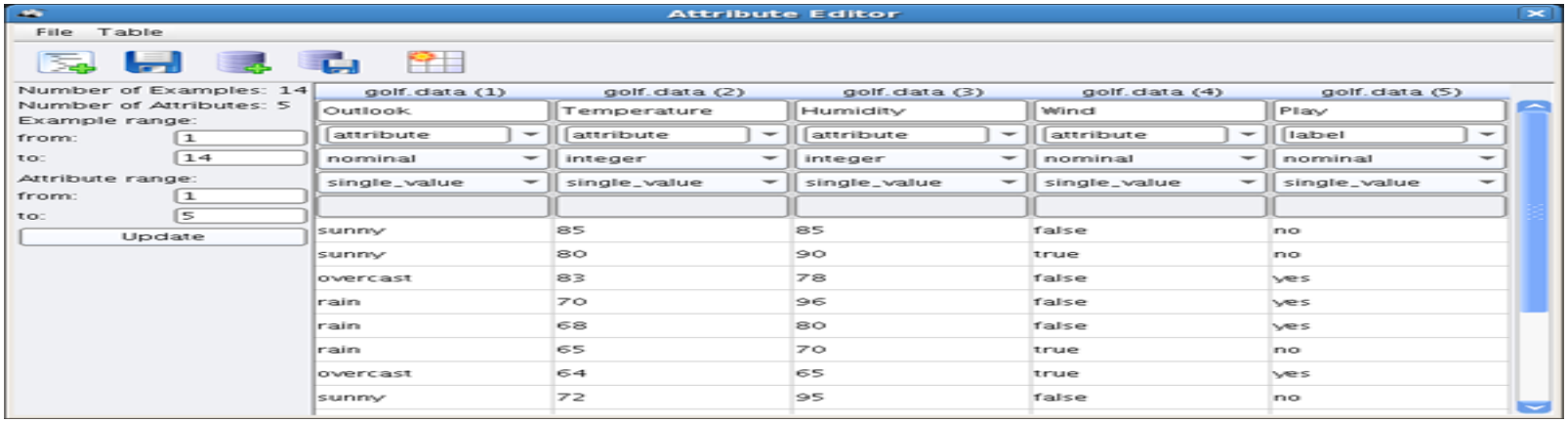
Attributes

Observations	1	2	3
1			
2			
3			
4			
5			
6			
7			
8			

Practical Work : Attribute Skipping

Using Background Knowledge: “Manual procedure”

RapidMiner

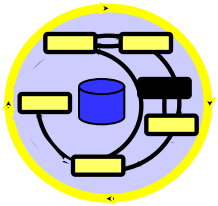


Load: churn

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



Attribute Reduction General Aspects

Data mining problems that deal with classification and prediction may involve hundreds or even thousands of attributes that can potentially be used as predictors. Example: Document classification in Text Mining: ***Bag-of-words****: >100000 attributes* , fault analysis in the automotive industry,…

↓

Problem: A lot of time and effort may be needed to decide which attribute should be included in the model

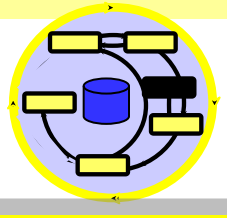
↓

Solution: In the last years Statisticians and Data Miners have developed many attribute reduction algorithms

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



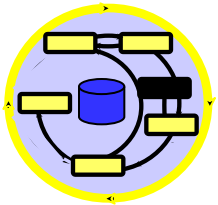
Why we need attribute Reduction ?

- to speed up learning process
- to reduce the amount of memory required
- to improve model interpretability
- to do visualization easier
- to make the datasets with many nominal attributes scalable

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



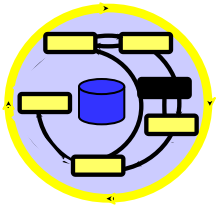
Attribute Reduction

- **creating new attributes**
  (combination of old attribute)
  attribute extraction
- **Selection a subset of old attributes**
  FSS: feature subset selection
  attribute selection
  - **Loss of information ?**
    - **no information lost if redundant and irrelevant attributes are present**

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

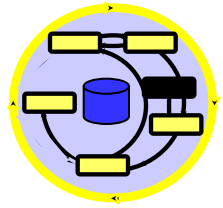


Attribute Reduction

First elementary steps

- Using common sense or domain knowledge (if available) to select a subset of attributes
  - Attribute Screening

Data Selecting



Attribute Reduction

First elementary steps

- **Attribute Screening**

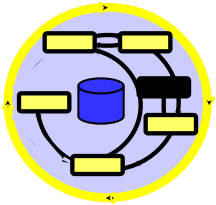
  removes problematic attributes e.g:

- **attributes with many missing values**

- **attributes with values that have too much or too little variation**

Example
Income of 100 individuals = { 20, 20, 20, 20, ……..20, 20 }

Data Selecting



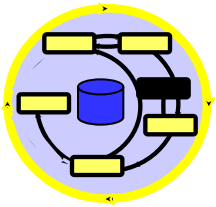
Attribute Reduction

Attribute Ranking

Determining attribute importance by criteria like:

- **Information Gain**
- **Gini-Index**
- **Pearson Chi-Square**
- **Correlation coefficient**
- **Akaike information criterion (AIC)**
- **….**

Data Selecting



Attribute Reduction

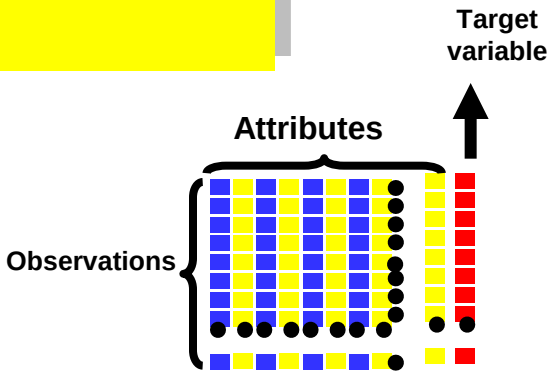
Attribute Ranking

The ranking criteria mentioned before can be used to measure the correlation between

1. **each attribute and the target variable (applicable only to Supervised Learning**
2. **between two attributes, pairwise**

Remarks

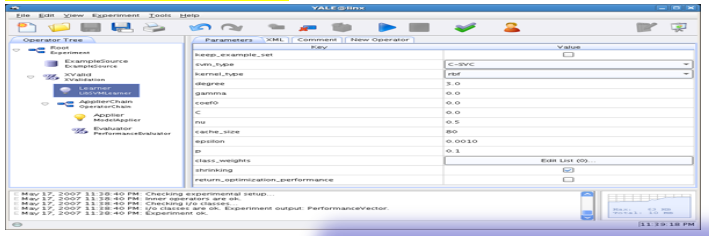
- **In the case 1, an attribute useless by itself can be useful together with others**
- **In case 2 attribute selection is independent of the target variable or , generally, independent of the data mining task**



→ **Known as Filter Approach**

Practical WorK: Data Understanding and Preprocessing: Attribute Reduction

RapidMiner



Load the Excel Dataset Weather..sameAttribute

Removing Useless Attributes:

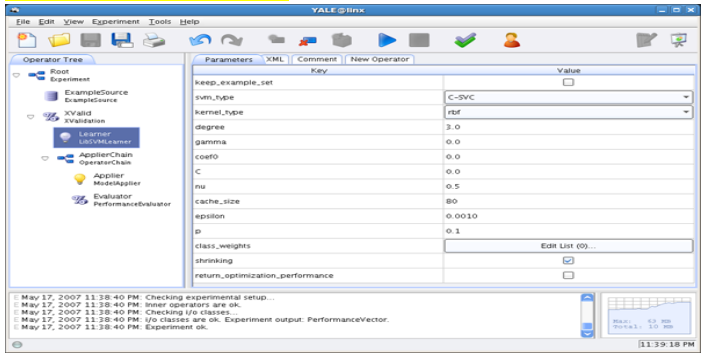
- Attributes with same value
- Attributes with many different values

Use “Remove useless attributes”

Use “Remove correlated attributes”

Practical Work: Dealing with missing values

RapidMiner

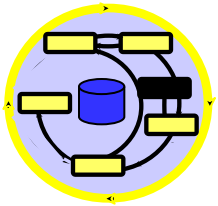


Eliminate or substitute Missing Values

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



Attribute Reduction

Embedded Methods

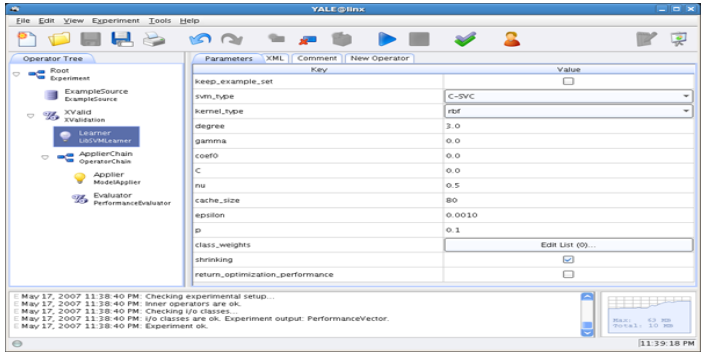
- in embedded approaches attribute selection is a part of the training process
- not all Data Mining algorithms have this built-in mechanism to perform attribute selection within the training process
- due to avoiding retraining for different attribute subsets , embedded approaches are more efficient
- Examples: Decision and Regression Trees

Remark

- in some studies, in a first step simple linear embedded systems are used for attribute selection
- later in a second step the selected attributes are used for training of a more complicated non-linear system

Practical Work: Embedded Attribute Selection

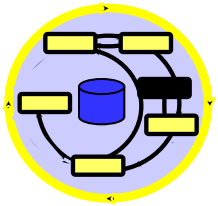
RapidMiner



Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



Attribute Reduction

Wrapper Methods

Main Idea :

- given a classification or prediction algorithm to evaluate the prediction performance of different subsets of attributes
- Select the attribute subset with the highest performance

Three Attributes

A1 , A2 , A3 →

{A1, A2, A3 }

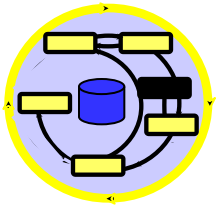
{ A1 } , { A2 } , { A3 }

{ A1, A2 } { A1, A3 }, { A2, A3 }

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



Attribute Reduction

Wrapper Methods

Main Challenges :

1. Selecting a search method to find all possible attribute subsets
2. Selecting an evaluation approach and an evaluation function to asses the prediction performance of different attribute subsets

About 1: Total search in the case of too large number of attributes needs massive amounts of computation. Greedy search like forward selection and backward elimination are more appropriate

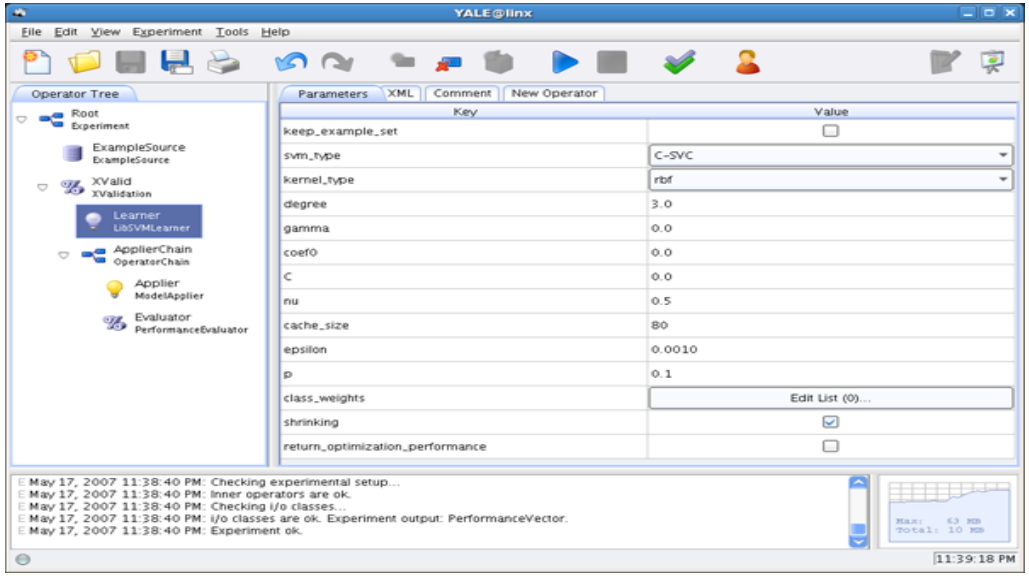
About 2 : Validation datasets or cross validation as well as evaluation functions (e.g accuracy rate or mean squared error) can be used

Practical Work: Feature Construction Preprocessing

Attribute Reduction

Wrapper Methods

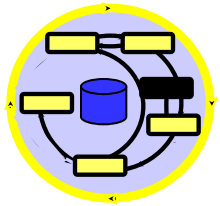
RapidMiner



Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

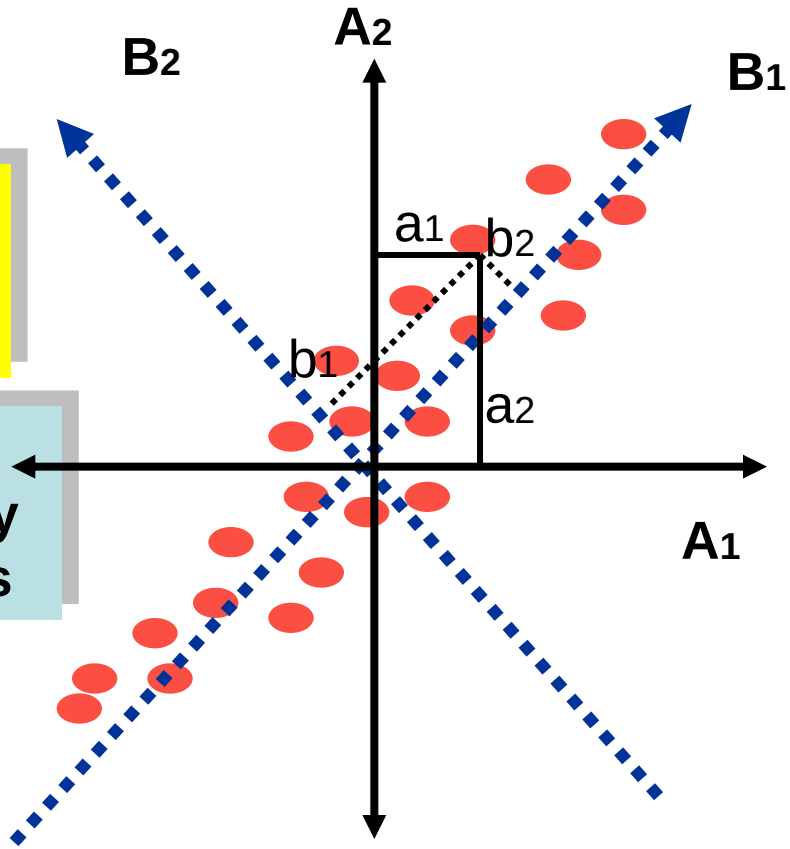


Principal Component Analysis (PCA)

Main Idea

Reducing multidimensional data sets to lower dimensions by combination of old attributes

the variance of the observations in original space should be satisfactory covered by the new created dimensions



$b_1 = p_1 a_1 + p_2 a_2$

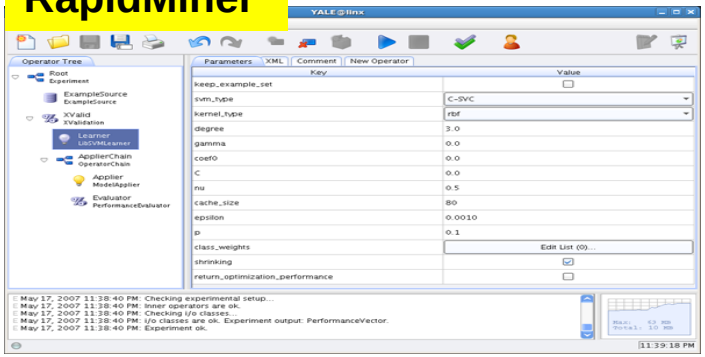
$b_2 = q_1 a_1 + q_2 a_2$

Instruments:
- **Covariance Matrix**
  - **Eigenvalues**
  - **Eigenvectors**

Interpretation ?

Practical Work: Feature Construction Preprocessing

RapidMiner



Use PCA Example of RapidMiner (in „Attributes“)

Data Mining Process

Data Cleaning

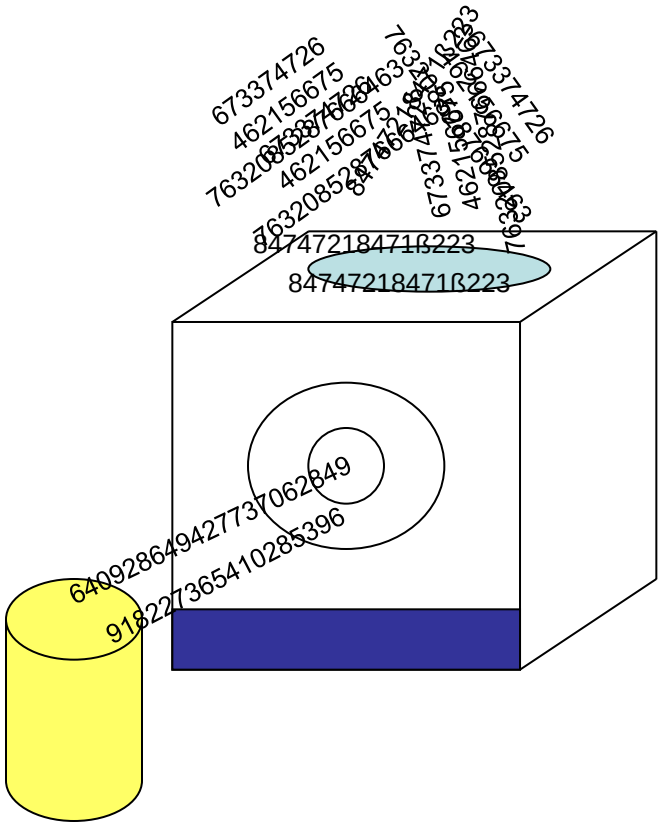
Data should has a high quality
Otherwise, garbage in => garbage out

Dealing with :

Missing Values
- Ignore the observation
- Using the attribute mean
- Predict the missing value
  - Decision tree
  - Regression
  - ……..

Inaccurate data
- Using Background Knowledge (Rules)

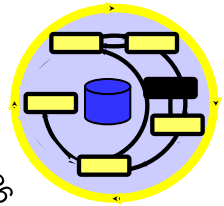
Duplicates
- Straße , Strasse, Str. Robert X, Bob X
- Professor, Prof. Dr.



Data Mining Process

CRISP-DM: Data Preparation

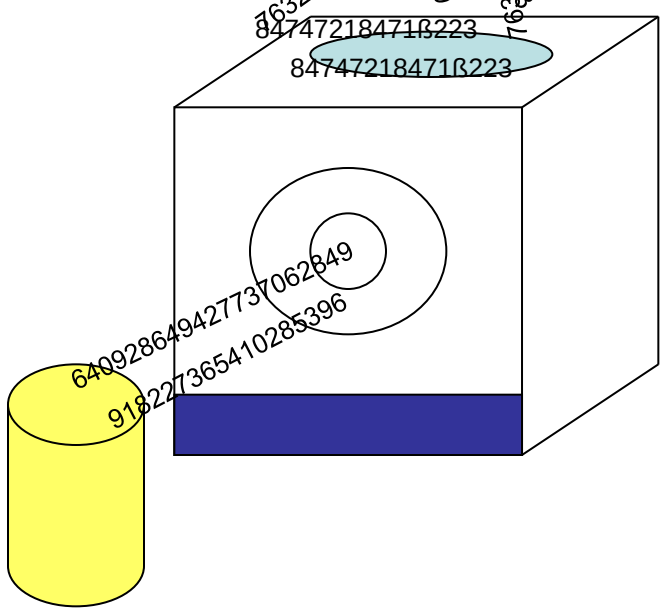
Data Cleaning



Dealing with Outliers

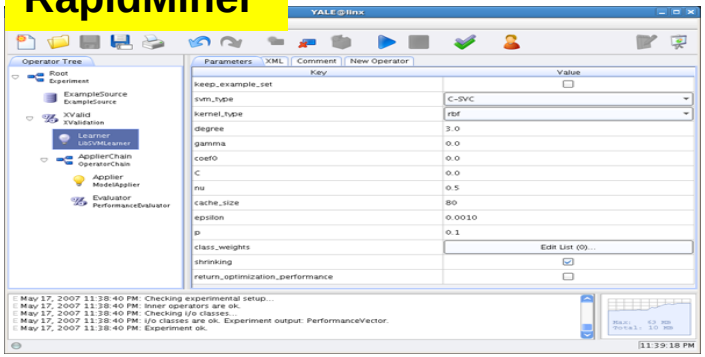
- **Outlier as noise**
- **Outlier detection as interesting finding**

- **Outliers Analysis Methods**
  - Model-based outlier detection
  - Using distance measures
  - Density-Based local Outlier Detection



Practical Work

RapidMiner



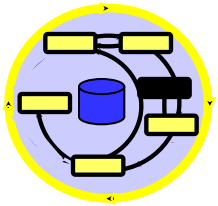
Outlier Detection: Dataset: Weather only Rain

Use : Distance Based Outlier, Parameter=3 & 3

Data Mining Process

CRISP-DM: Data Preparation

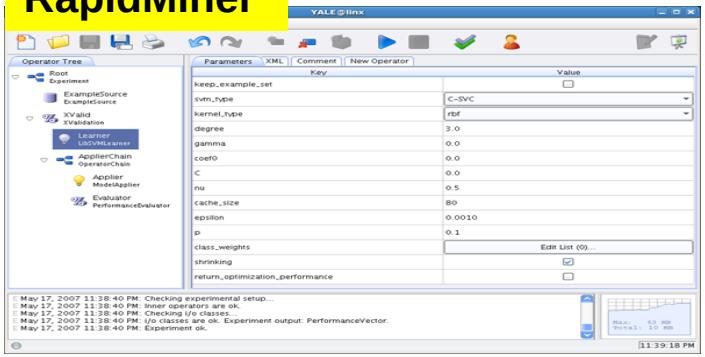
Data Transformation



- **Normalization of attributes value**
  - Values between 1 and 0
- **Adding new attributes**
  - According to new facts or background knowledge
- **Creation of new attributes using available attributes**
  - Surface area instead of length and width
- **Aggregation and Generalization of attribute values**
  - Monthly sales instead of daily sales
  - City instead of streets
- **Discretization of Continuous-Valued attributes**

Practical Work

RapidMiner



Attribute Transformation:

Transfer in [0 1] interval or Z-Trasformation

Use Sample of RM

ML-Algorithms

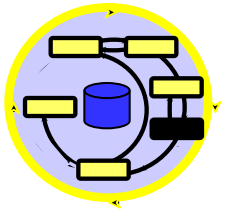
Top 10 Algorithms in Data Mining

Selected by International Conference on Data Mining (ICDM) in December 2006

The algorithms cover:		Selected algorithms :
➢ Classification and Prediction	→	C4.5, CART, KNN, Naïve Bayes
➢ Association Mining	→	Apriori
➢ Clustering	→	K-Means
➢ Statistical Learning	→	SVM, EM
➢ Bagging and Boosting	→	AdaBoost
➢ Link Mining	→	PageRank

CRISP-DM: Modeling

Model Evaluation → **Choosing evaluation function and evaluation method**



General remarks

- Results produced by the model are normally worse than the real facts

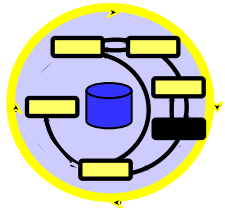
Reasons:

- Error in data
- Model Misspecification
- Structural Change
- ……..

To evaluate the results produced by the model we need :

- Evaluation functions
- Evaluation methods

Data Mining Process

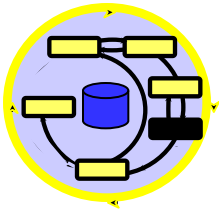


Model Evaluation

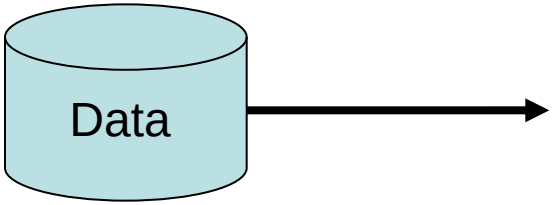
"when you have two competing theories
which make exactly the same predictions,
the one that is simpler is the better."

Based on "Occam's Razor" Rule after Willium of Occam (c. 1285 -1347).

Data Mining Process



Choosing evaluation function



Model Generation

Calculate Model Accuracy

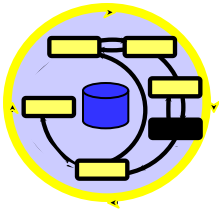
overfitting overspecialization

model evaluation based on data not used for model generation

Resubstitution Accuracy

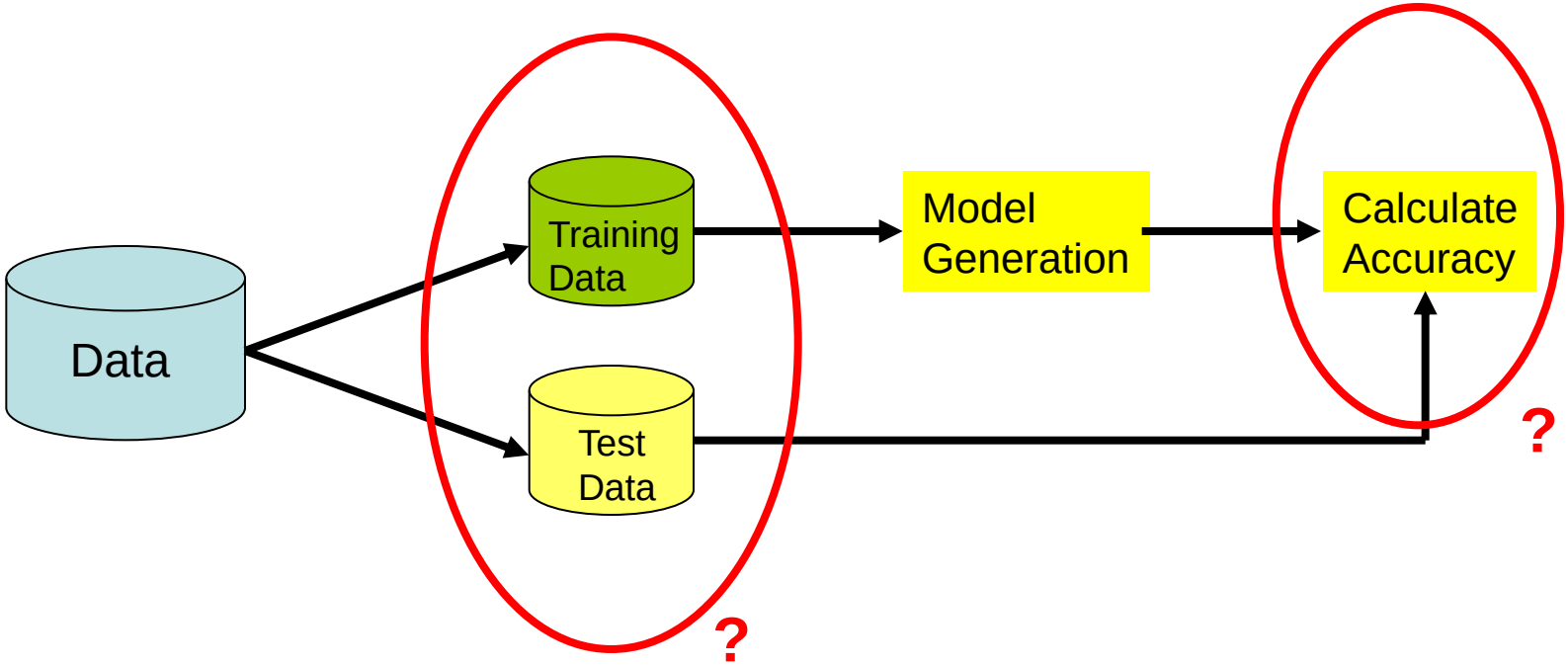
optimistic estimation of true accuracy

Data Mining Process



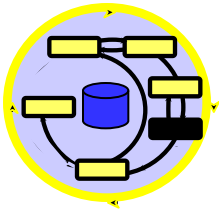
model evaluation

- **Shaping training and test data**
- **Choosing evaluation function**



Data Mining Process

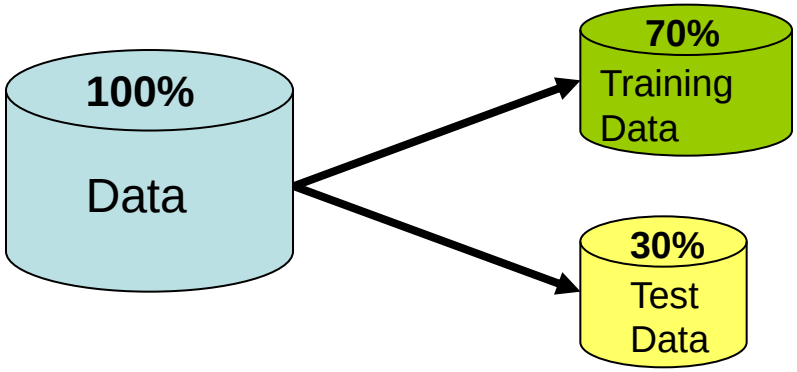
model evaluation → Choosing evaluation method



Holdout Method

Choosing randomly a certain portion of data for training and the rest for test

Example: 70% for training, 30% for testing

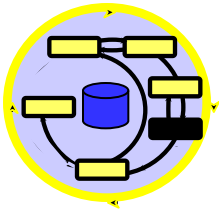


Disadvantage of Holdout Method:
Generating of the model based on a portion of the data

↓

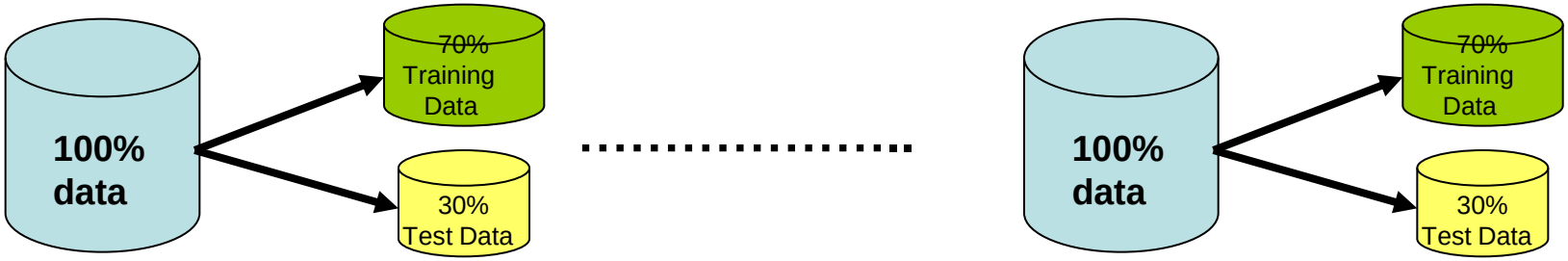
If the dataset is small than the generated model may not be as good as when the whole dataset is used for training

Data Mining Process



model evaluation → **Choosing evaluation method**

Random Subsampling

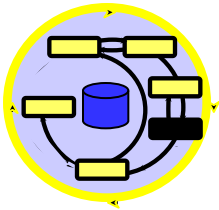


acc1 ..... **acc N**

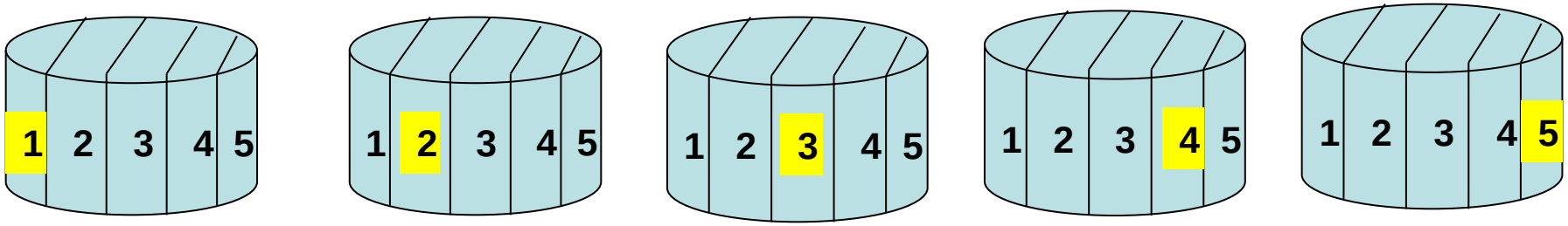
acc = Sum (acc1+ ……. + acc N) / N

Data Mining Process

model evaluation → **Choosing the evaluation method**



- **Cross Validation**



acc1 …….. acc N

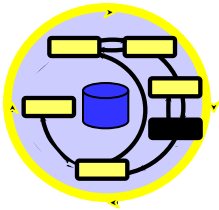
acc = Sum (acc1+ ……. + acc N) / N

In above example:
N = 5

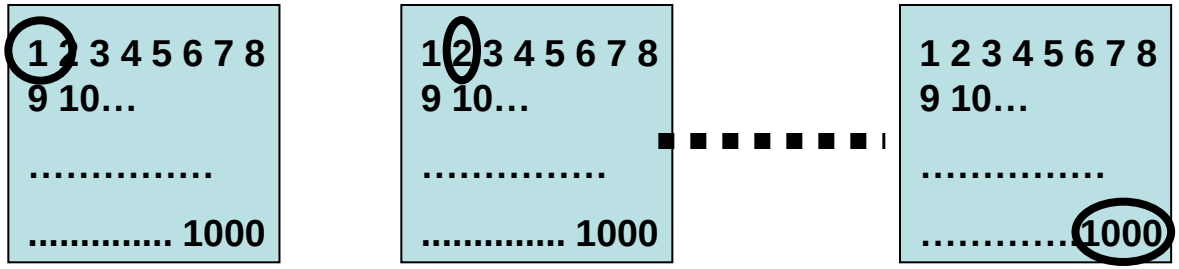
Empirically obtained: Optimum N = 10 (10-fold CV)

Data Mining Process

model evaluation → **Choosing the evaluation method**



- **Leave-one-out (LOO)**



acc1 …….. acc N

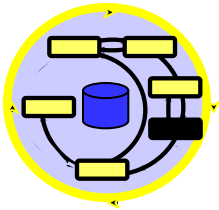
acc = Sum (acc1+ ……. + acc N) / N

In above example:
N = 1000

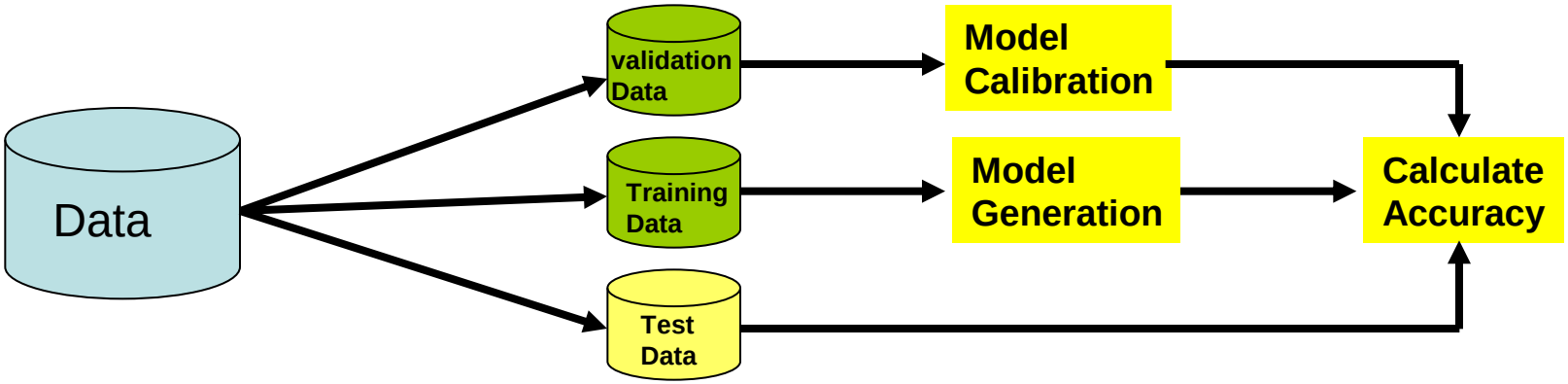
Data Mining Process

CRISP-DM: Modeling

Choosing the evaluation method



In some cases, besides the training and test datasets, a "validation dataset " is used too



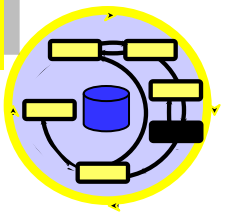
After evaluation, all the available data can be used to generate the final model

Data Mining Process

CRISP-DM: Modeling

Classification

Choosing the evaluation function



Confusion Matrix

		Actual	
		Class 1	Class 2
Model	Class 1	N1	M2
	Class 2	M1	N2

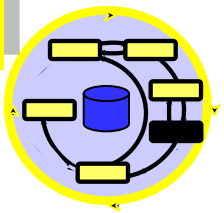
- **N1 is the number of correct classified observations of class 1**
- **N2 is the number of correct classified observations of class 2**
- **M1 is the number of incorrect classified observations (from class1 to class2)**
- **M2 is the number of incorrect classified observation (from class2 to class1)**

Data Mining Process

CRISP-DM: Modeling

Classification

Choosing the evaluation function



$$\text{Accuracy Rate} = \frac{N1+N2}{N1+N2+M1+M2}$$

$$\text{Error Rate} = \frac{M1+M2}{N1+N2+M1+M2}$$

Sometimes: Recognition Rate

(In Pattern Recognition)

Misclassification Rate

Confusion Matrix

		Actual	
		Class 1	Class 2
Model	Class 1	N1	M2
	Class 2	M1	N2

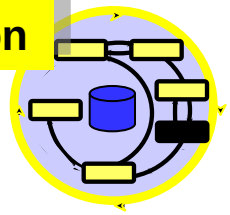
AR= 1- ER

Data Mining Process

CRISP-DM: Modeling

Classification

Choosing the evaluation function



Example

	Income	Car	Gender	Credit Rate	
				Actual	Predicted
Customer 1	low	new	F	bad	bad
Customer 2	middle	old	F	bad	bad
Customer 3	middle	new	M	good	bad
Customer 4	low	new	M	bad	bad
Customer 5	high	new	M	good	bad
Customer 6	high	new	F	good	good
Customer 7	middle	new	F	good	good
Customer 8	high	old	F	good	bad
Customer 9	middle	old	M	bad	good
Customer 10	low	old	F	bad	bad

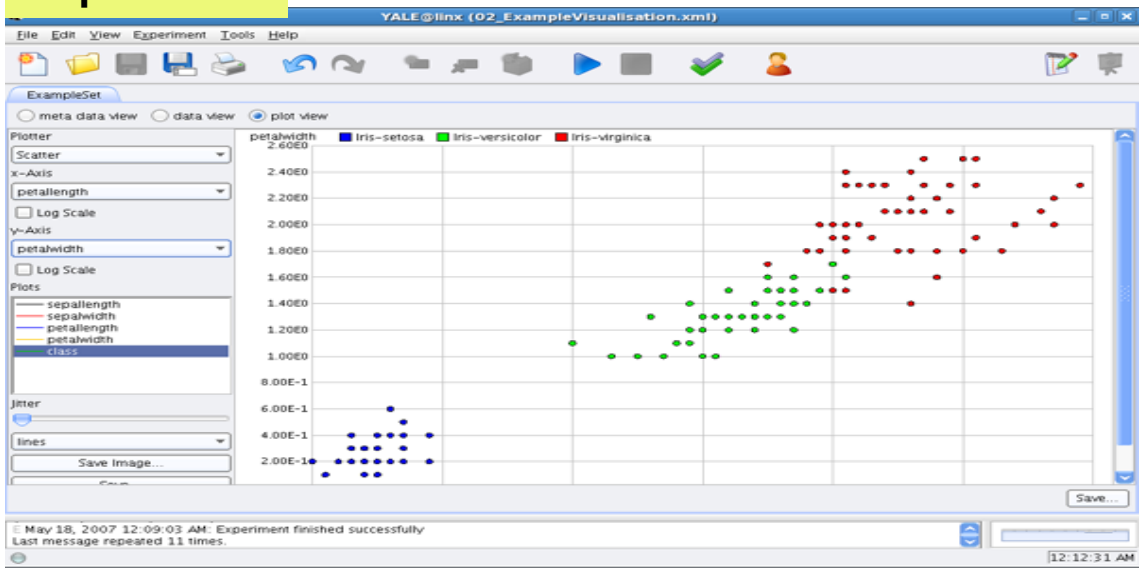
Confusion Matrix

		actual	
		good	bad
Model	good	2	1
	bad	3	4

Accuracy Rate = 6/10 = 60%
Error Rate= 40%

Practical Work Part : Model Evaluation

RapidMiner



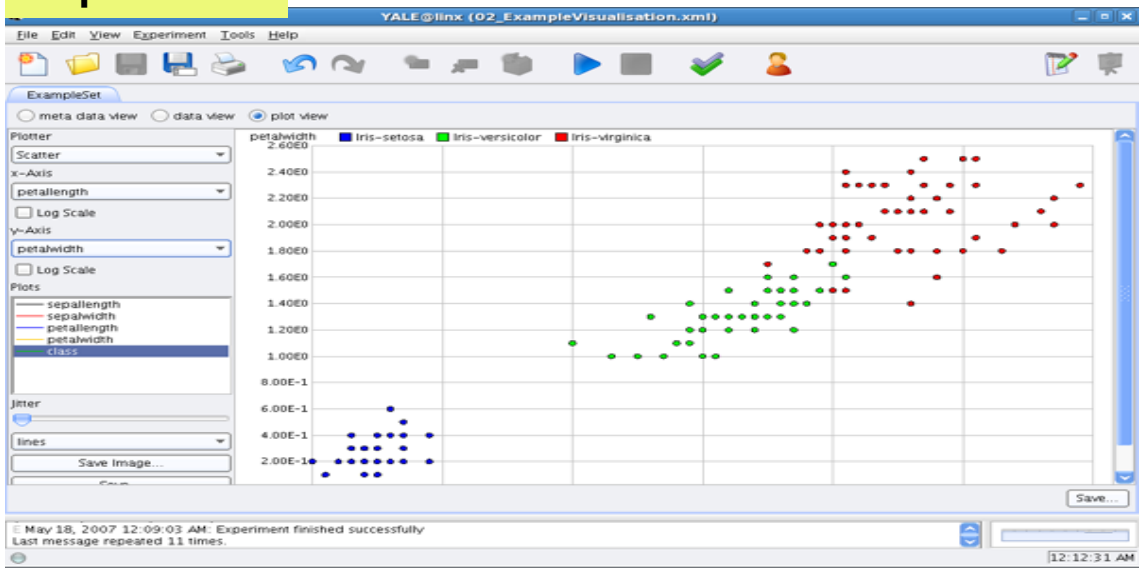
Load Churn data

Change:
Int'Plan, VmailPlan, AreaCode and Churn to "NOMINAL"

Demonstrate simple_validation

Practical Work Part : Model Evaluation

RapidMiner



Demonstrate

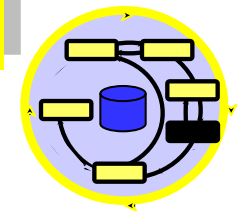
- **Cross validation**
- **Leave – One - Out**

Data Mining Process

CRISP-DM: Modeling

Classification

Choosing the evaluation function



Class 1= positive
Class2 = negative

Confusion Matrix

		Actual	
		Class 1	Class 2
Model	Class 1	N1	M2
	Class 2	M1	N2

Confusion Matrix

		Actuel	
		Class 1 positive	Class 2 negative
Model	Class 1 positive	true positive	false positive
	Class 2 negative	false negative	true negative

Model Evaluation, Disadvantage of Accuracy Rate

Choosing the evaluation function

Case1: Different misclassification costs

Confusion Matrix

Model		Actual: Positive	Actual: Negative
	Positive	10	1
	Negative	9	80

Confusion Matrix

Model		Actual 2: Positive	Actual 2: Negative
	Positive	18	9
	Negative	1	72

ACC1 = 90%

ACC2 = 90%

Cost of misclassification would be a better criteria

Model Evaluation, Disadvantage of Accuracy Rate

Case2: unbalanced classes

Suppose that number of Test Data = 100

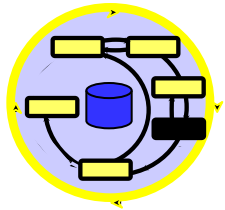
Suppose that we have two classes 1 and 2:
In Class 1 2 observation ~ 2%
In Class 2 98 observation ~98%

A case of unbalanced classes

Called Default Accuracy Rate

The Accuracy Rate of the model should be generally higher than the default Accuracy rate

Choosing the evaluation function and evaluation method



Evaluation of Prediction Models (numerical target variable)

$$MAE = 1/n \sum_{i=1}^{n} |Yi - Y'i|$$

$$MSE = 1/n \sum_{i=1}^{n} (Yi - Y'i)^2$$

$$RAE = \frac{\sum_{i=1}^{n} |Yi - Y'i|}{\sum_{i=1}^{n} |Yi - \bar{Y}|}$$

$$RSE = \frac{\sum_{i=1}^{n} (Yi - Y'i)^2}{\sum_{i=1}^{n} (Yi - \bar{Y})^2}$$

MAE = Mean Absolute Error
RAE = Relative Absolute Error
n= Number of observations in test data
Y= Actual value, Y'= Predicted value,

MSE = Mean Squared Error
RSE = Relative Squared Error
$\bar{Y}$ = mean of Ys