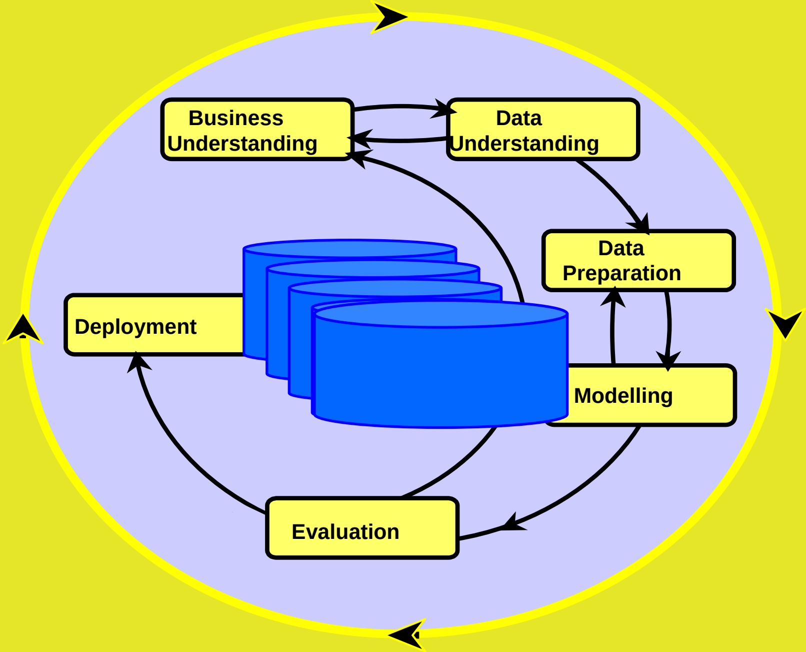
Statistic Methods in Data Mining



Data Mining Process
(Part 3)

Professor Dr. Gholamreza Nakhaeizadeh

Short review of the last lecture

Data Pre-Processing

Observation and attribute reduction

- **Sampling**
  - Representative sample
  - Random sampling
  - Sampling with and without replacement
  - Systematic sampling
  - Stratified sampling

- **Attribute reduction**
  - Supervised and unsupervised leaning
  - Why attribute reduction ? What are the benefits ?
  - Curse of dimensionality
  - Attribute Reduction: Generating new attributes, selection of attributes subsets
  - First elementary steps: Consideration of background knowledge, screening
  - Attribute ranking according to importance
  - supervised and unsupervised ranking
  - Embedded approaches, Wrapper
  - PCA

Data Mining Process

CRISP-DM: Data Preparation

Data Cleaning

Dealing with :

Missing Values

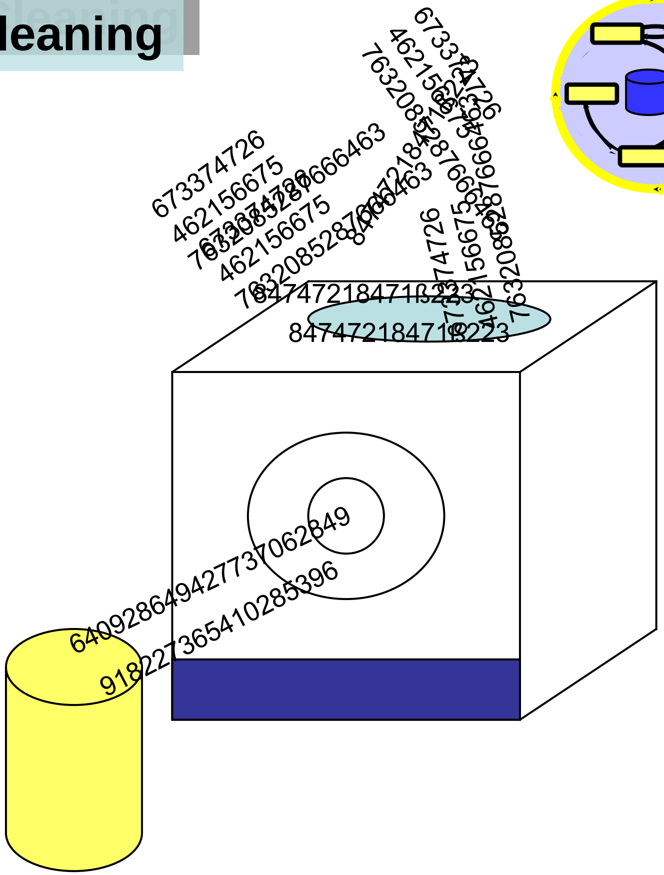
- Ignore the observation
- Ignore the attribute
- Using the attribute mean
- Predict the missing value
  - Decision tree
  - Regression
  - ……..

Inaccurate data

- Using Background Knowledge (Rules)

Duplicates

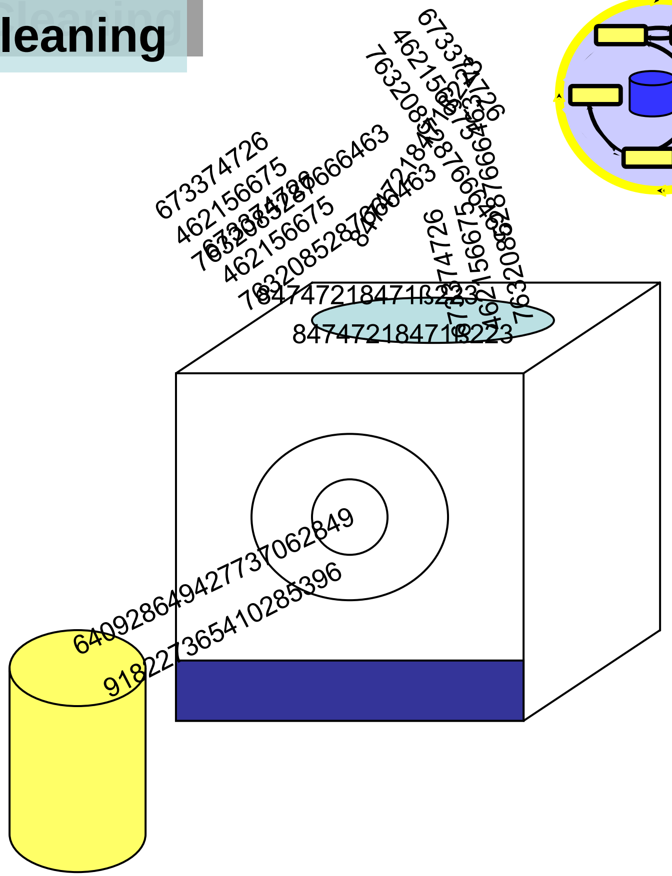
- Straße , Strasse, Str. Robert X, Bob X
- Professor, Prof. Dr.



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

Data Mining Process

CRISP-DM: Data Preparation

Data Transformation

- **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

Data Mining Process

CRISP-DM: Data Preparation

Data Transformation

Example: monthly sales instead daily sales

	Day1	…..	Day30
Company 1	S_{11}	…..	S_{130}
Company 2	S_{21}	…..	S_{230}
…………………		……	……
Company n	S_{n1}	…..	S_{230}

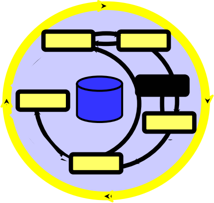
	Month1
Company 1	M_1
Company 2	M_2
…………………	……..
Company n	M_n

$$M_i = \sum_{j=1}^{30} S_{ij}$$

Data Mining Process

CRISP-DM: Data Preparation

Data Transformation

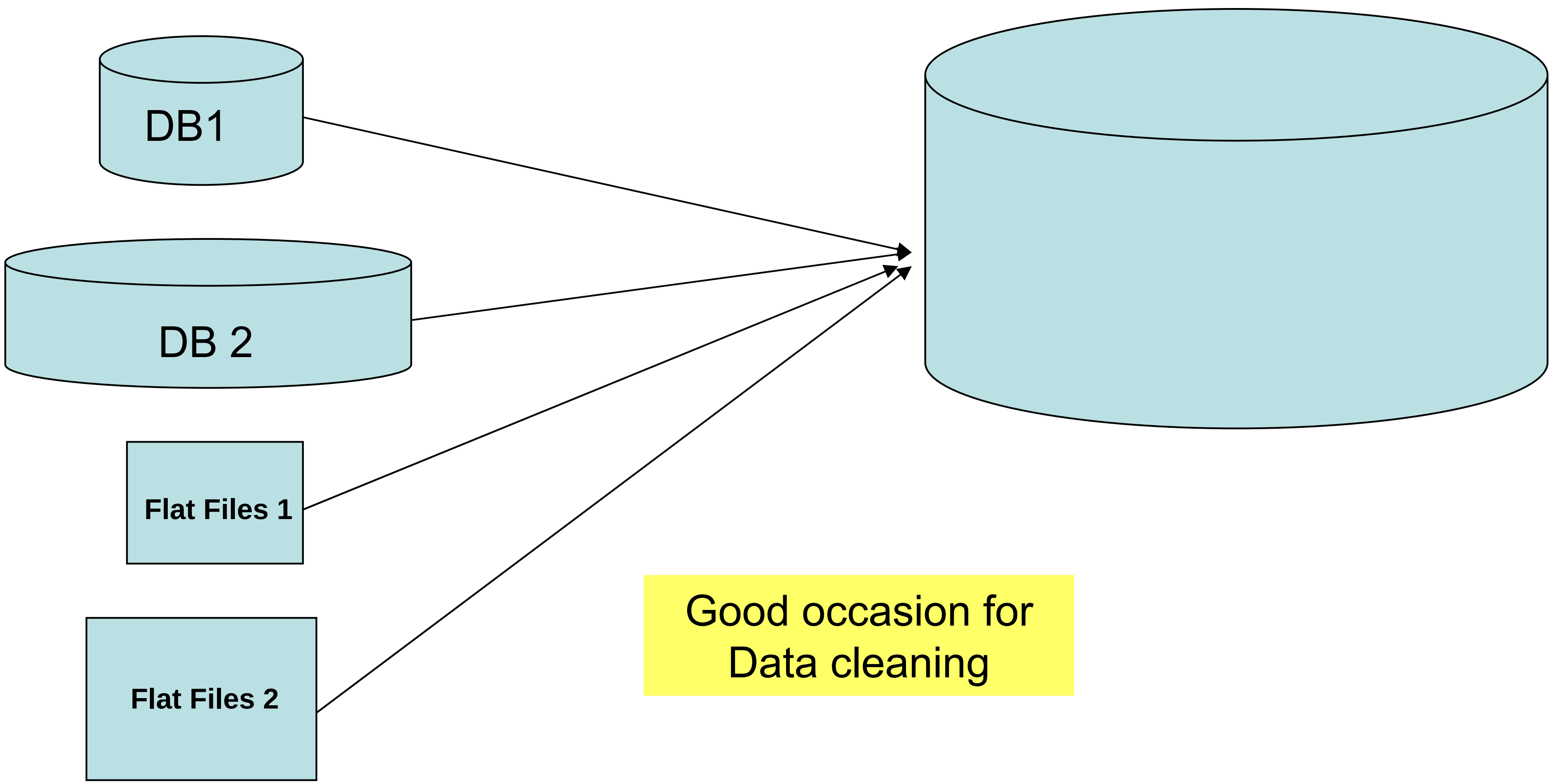


- **Binarization of categorical attributes**
- **Discretization of Continuous-Valued attributes**
- **Normalization of attributes value**
  - Values between 1 and 0

Data Mining Process

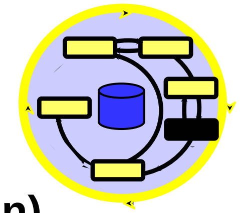
CRISP-DM: Data Preparation

Data integration



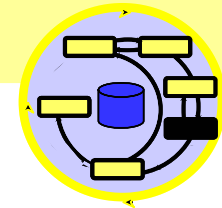
Data Mining Process

CRISP-DM: Modeling



1. **Task Identification**
   - Classification
   - Prediction
   - Clustering
   - …
2. **Determining the DM-algorithms**
   - Decision Trees
   - Neural Networks
   - Association Rules
   - …
3. **Choosing the evaluation function and evaluation method**
   - Sum squared Errors
   - Accuracy rates
   - Loss function
   - Cross-Validation
   - …..
4. **Choosing the search (optimization) method**
   - Analytical methods
   - Greedy search
   - Gradient descent
   - …
5. **Choosing the data management method**
   - Not always necessary

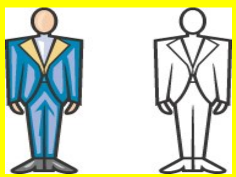
Data Mining Process



CRISP-DM: Modeling

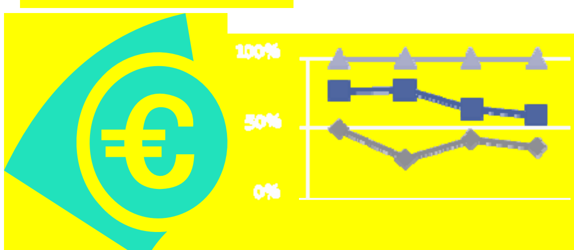
Task Identification

Classification
Credit Scoring
- **Good customer**
- **Bad customer**

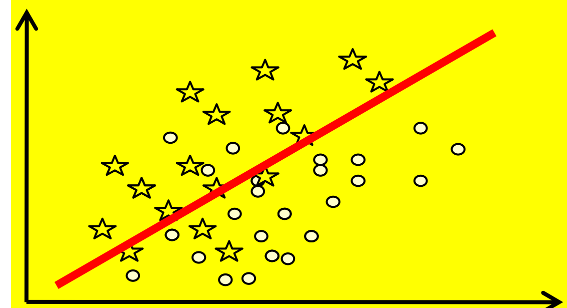


Concept Description
Customers Loyalty :
- **Age**
- **Income**
- **Education**
- ….

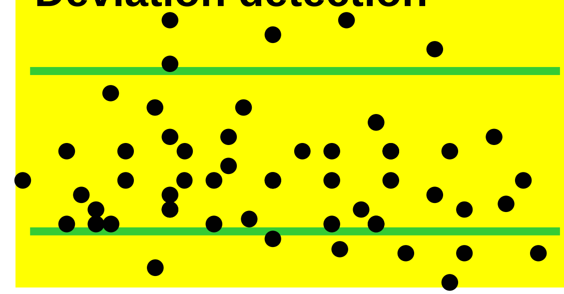
Prediction



Clustering



Deviation detection



Dependency Analysis

A and B → C

Sequence Pattern

Data Mining Process

CRISP-DM: Modeling | Task Identification | Classification

Examples:

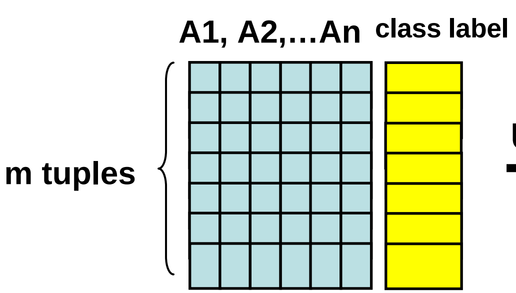
Credit Scoring
- Good customer
- Bad customer

Quality Management
- device defect
- device not defect
- perhaps defect

Marketing
- Customer buys
- Customer doesn't buy

Univ. entrance exam
- successful
- unsuccessful

Suppose that a tuple X is represented by n Attributes A1, A2,…, An and is assigned to another predefined attribute called target variable or class label attribute. For m tuples we will have a matrix representation like:



Using this data →

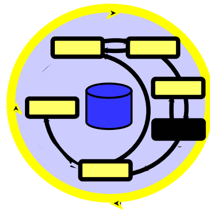
Goal: Build a classifier that can predict the class label of a new tuple only by using its attribute values

Data Mining Process

CRISP-DM: Modeling

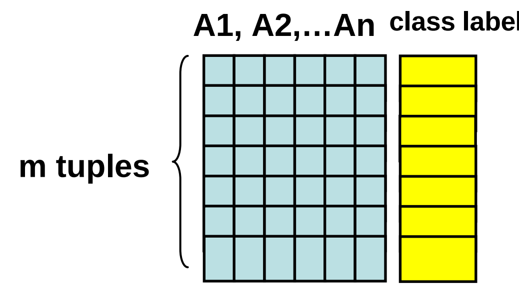
Task Identification

Classification



Simple fictive example; Credit Rating in a Bank

	Income >2000	Car	Gender	Credit Rating
Customer 1	no	yes	F	bad
Customer 2	no statement	no	F	bad
Customer 3	no statement	yes	M	good
Customer 4	no	yes	M	bad
Customer 5	yes	yes	M	good
Customer 6	yes	yes	F	good
Customer 7	no statement	yes	F	good
Customer 8	yes	no	F	good
Customer 9	no statement	no	M	bad
Customer 10	no	no	F	bad

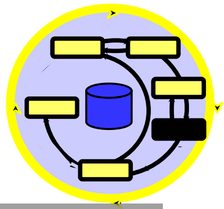


In classification, the class label (target variable) is a nominal variable

Income=3000 , car=yes, gender=female → Credit rating ?

Data Mining Process

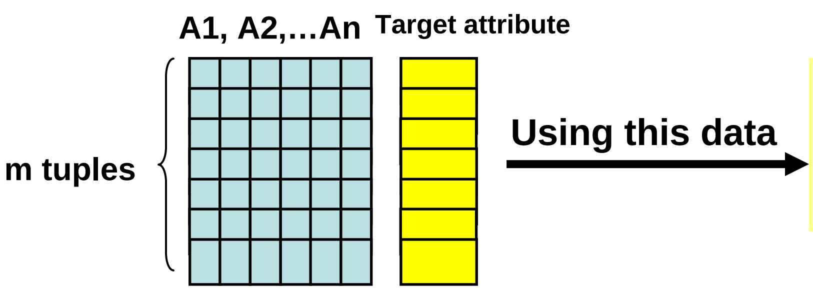
CRISP-DM: Modeling | Task Identification | Prediction



Examples:

Monthly Sales	**Hourly Exchange Rate**	**Average Daily Temperature**
2000	1.3918	23.4
2560	1.3917	25.6
1947	1.3914	24.6
……	……….	…

Suppose that a tuple X is represented by n Attributes A1, A2,…, An and is assigned to another predefined attribute called target attribute. For m tuples we will have a matrix representation like:



Goal: Build a Predictor that can predict the target value of a new tuple only by using its attribute values

Data Mining Process

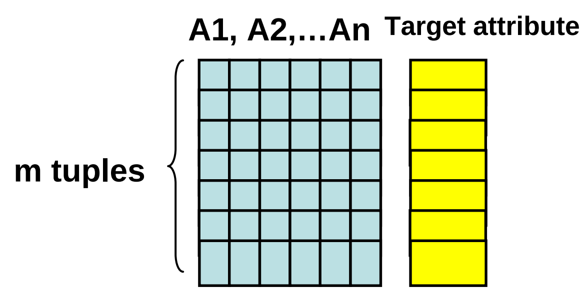
CRISP-DM: Modeling

Task Identification

Prediction

Simple fictive example; Prediction of annual income

ID	Income in three years ago	Education	Age	Income
1A	24552	High School	32	27026
2A	88282	BSc	52	93725
3B	82902	PhD	41	82356
4A	39838	High School	56	36828
5C	53542	PhD	32	62542
6M	63826	MS	28	64882
7D	82783	MA	43	89025
8A	72886	High School	33	74925
9Q	21383	BA	37	62572
1R	63552	BA	41	66427
1T	62522	High School	25	63552
1E	65254	PhD	56	67252



In prediction, the target variable is a continuous-valued variable

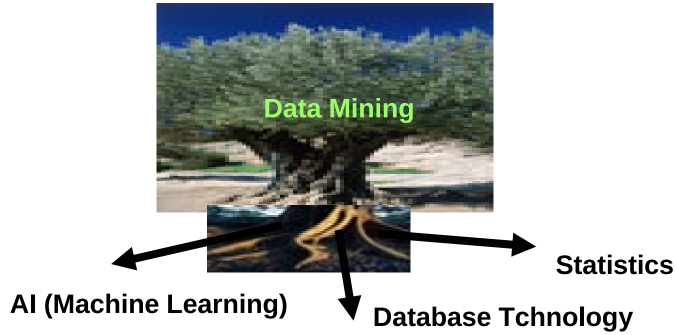
Income in three years ago=60000 , education=BA, Age=35

Annual income ?

Data Mining Process

CRISP-DM: Modeling

Determining the DM-algorithms



Data Mining Algorithms

Machine Learning

- Rule Based Induction
- Decision Trees
- Neural Networks
- …….

Statistics

- Discriminant Analysis
- Cluster Analysis
- Regression Analysis
- Logistic Regression Analysis
- …….

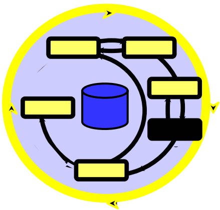
Database Technology

- Association Rules
- Sequence Mining
- ….

Data Mining Process

CRISP-DM: Modeling

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
  - …………..

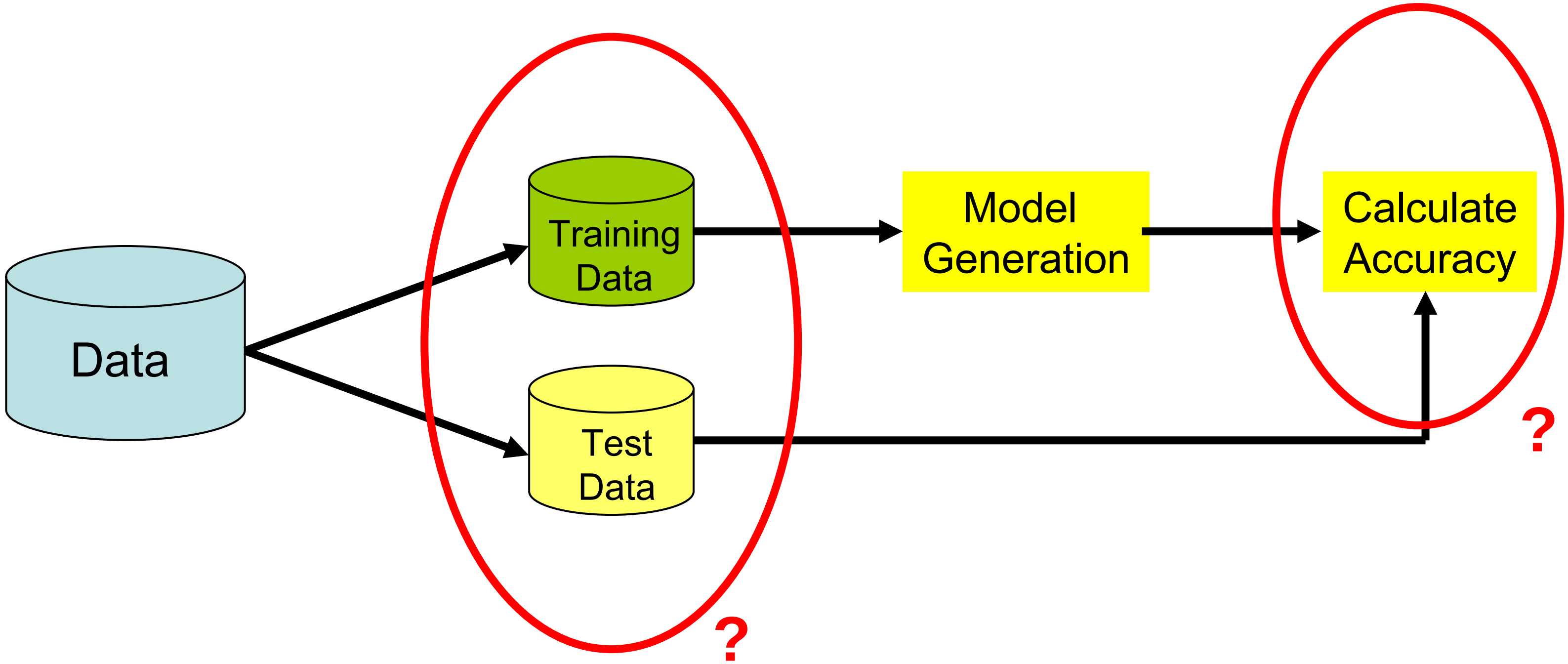
To evaluate the results produced by the model we need :
- Evaluation methods
- Evaluation functions

Data Mining Process

CRISP-DM: Modeling

Choosing evaluation function and evaluation method

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

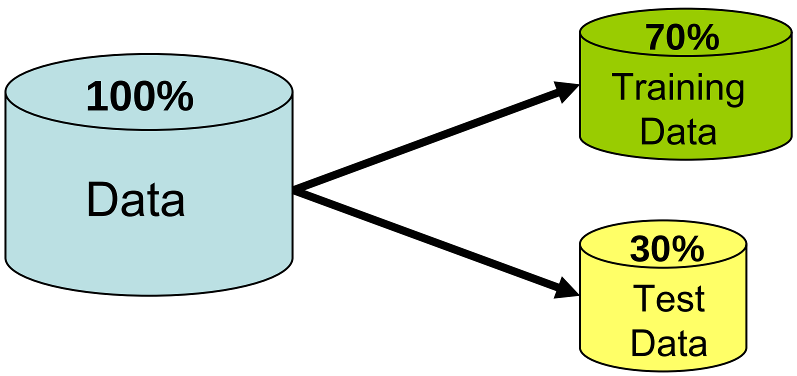


Data Mining Process

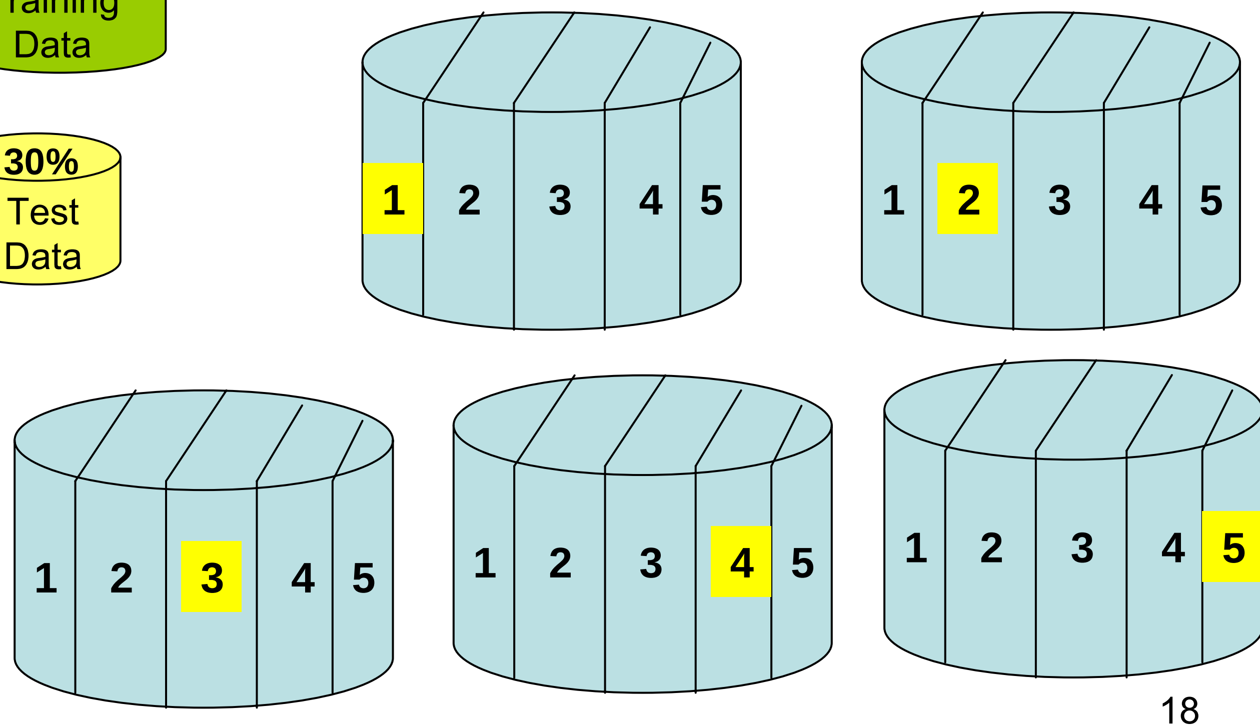
CRISP-DM: Modeling

Choosing evaluation function and evaluation method

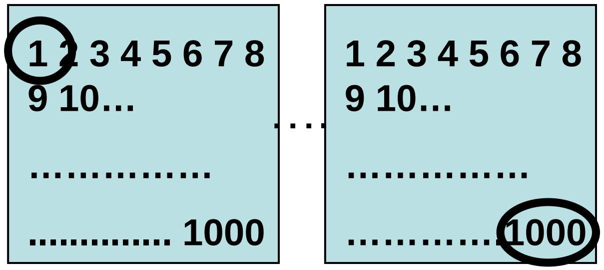
Choosing a certain portion of data for training and test
Example: 70% training, 30% test



• Cross Validation



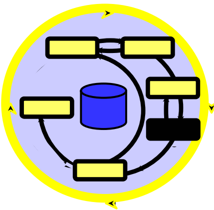
•Leave-one-out



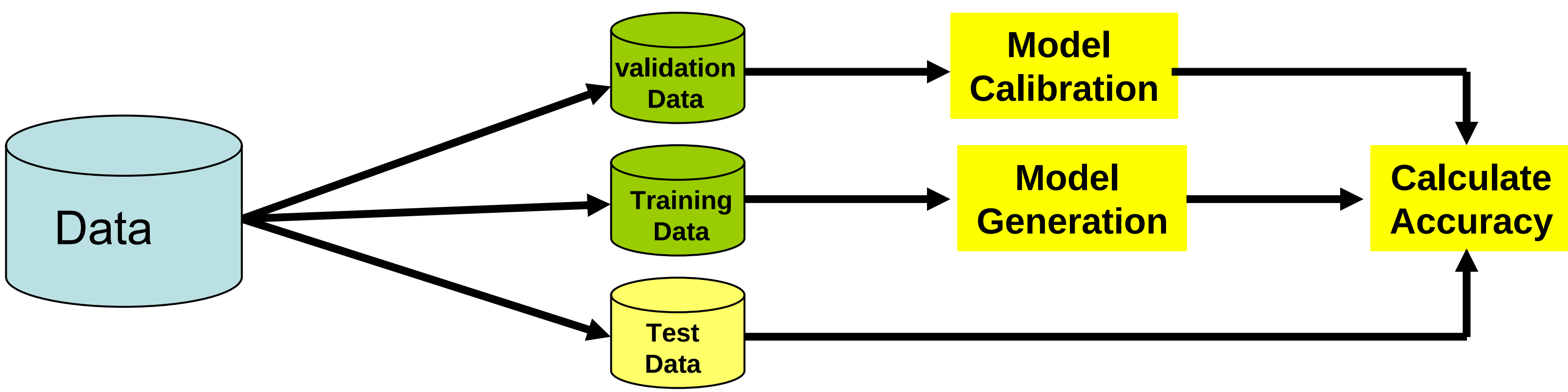
Data Mining Process

CRISP-DM: Modeling

Choosing evaluation function and evaluation method



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



Data Mining Process

CRISP-DM: Modeling

Choosing evaluation function and evaluation method

Ripley (1996), (p.354)

Training set:

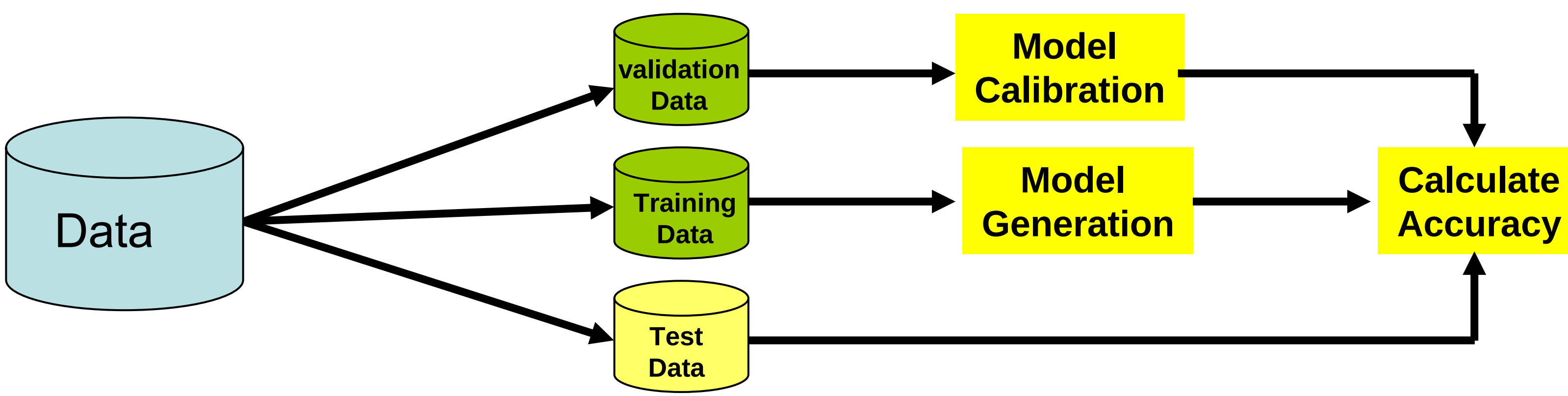
A set of examples used for learning, that is to fit the parameters [i.e., weights] of the classifier.

Validation set:

A set of examples used to tune the parameters [i.e., architecture, not weights] of a classifier, for example to choose the number of hidden units in a neural network.

Test set:

A set of examples used only to assess the performance [generalization] of a fully-specified classifier.



Data Mining Process

CRISP-DM: Modeling

Choosing the evaluation function and evaluation method

Classification

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

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

AR= 1- ER

Confusion Matrix

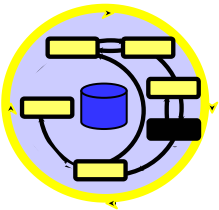
		Model	
		Class 1	Class 2
Actual	Class 1	N1	M1
	Class 2	M2	N2

- 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 and evaluation method

Example

	Income >2000	Car	Gender	Credit Rating Actual	Credit Rating predicted
Customer 1	no	yes	F	bad	bad
Customer 2	no statement	no	F	bad	bad
Customer 3	no statement	yes	M	good	bad
Customer 4	no	yes	M	bad	bad
Customer 5	yes	yes	M	good	bad
Customer 6	yes	yes	F	good	good
Customer 7	no statement	yes	F	good	good
Customer 8	yes	no	F	good	bad
Customer 9	no statement	no	M	bad	good
Customer 10	no	no	F	bad	bad

Confusion Matrix

		Model: good	Model: bad
Actual	good	2	3
Actual	bad	1	4

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

Data Mining Process

CRISP-DM: Modeling

Classification

Choosing the evaluation function and evaluation method

Class 1= positive
Class2 = negative

Confusion Matrix

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

Confusion Matrix

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

Data Mining Process

CRISP-DM: Modeling

Choosing the evaluation function and evaluation method

Prediction

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

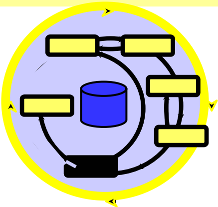
$$RAE = \frac{\sum_{i=1}^{n} |Yi - Y'i|}{\sum_{i=1}^{n} |Yi - \bar{Y}|} \qquad RSE = \frac{\sum_{i=1}^{n} (Yi - Y'i)^2}{\sum_{i=1}^{n} (Yi - \bar{Y})^2}$$

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

Data Mining Process

CRISP-DM: Evaluation

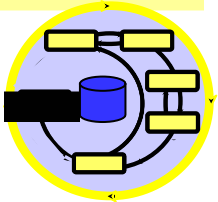


http://www.crisp-dm.org/CRISPwP-0800.pdf

- **Evaluate results**
- **Review process**
- **Determine next steps**

Data Mining Process

CRISP-DM: Deployment



http://www.crisp-dm.org/CRISPwP-0800.pdf

- **Plan deployment**
- **Plan monitoring and maintenance**
- **Produce final report**
- **Review project**

Data Mining Process

Another Example: SAS data Mining Process SEMMA

Sample **the data by creating one or more data tables. The sample should be large enough to contain the significant information, yet small enough to process**

Explore **the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas**

Modify **the data by creating, selecting, and transforming the variables to focus the model selection process**

Model **the data by using the analytical tools to search for a combination of the data that reliably predicts a desired outcome**

Assess the data by evaluating the usefulness and reliability of the findings from the data mining process

Source: http://support.sas.com/documentation/onlinedoc/miner/getstarted.pdf

Data Mining Algorithms

Data Mining algorithms

Machine Learning

- Rule Based Induction
- Decision Trees
- Neural Networks
- Conceptional clustering
- …….

Statistics

- Discriminant Analysis
- Cluster Analysis
- Regression Analysis
- Logistic Regression Analysis
- …….

Database Technology

- Association Rules
- ….