Data Mining and Analysis
Analiza i eksploracja danych

Lecturer: JERZY STEFANOWSKI
Institute of Computing Science
Poznań University of Technology

Software Engineering – Master Course
Computer Science, PUT, revised 2015 / 2016

Background literature [Polish translations]

- Translations:
  - Larose D., Odkrywanie wiedzy z danych. Wprowadzanie do eksploracji danych, PWN, 2006.
  - Larose D., Metody i modele eksploracji danych, PWN 2008.
  - Hand D., Mannila H., Smyth P. Eksploracja danych, WNT, 2005 (Principles of Data Mining, MIT Press, 2001).

- Polskie książki – Polish language books
  - Tadeusz Morzy, Eksploracja danych. Metody i algorytmy. PWN 2013!!!!!
  - Koronacki J., Ćwik J., Statystyczne systemy uczące się, WNT 2005 (kolejne wydanie w drodze).
  - Krawiec K, Stefanowski J., Uczenie maszynowe i sieci neuronowe, Wyd. PP, 2003.

Background literature (English)

- Han Jiawei and Kamber M. Data mining: Concepts and techniques, Morgan Kaufmann, 2001 (1 ed.), there is 2d
- Hand D., Mannila H., Smyth P. Principles of Data Mining, MIT Press, 2001.
- Kononenko I., Kukar M., Machine Learning and Data Mining: Introduction to Priniciples and Algorithms. Horwood Pub, 2007.
- Maimon O., Rokach L., The data mining and knowledge discovery Handbook, Springer 2005.
- Witten I., Eibe Frank, Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 1999.

Lecture 1 a.

Data Mining: Introduction

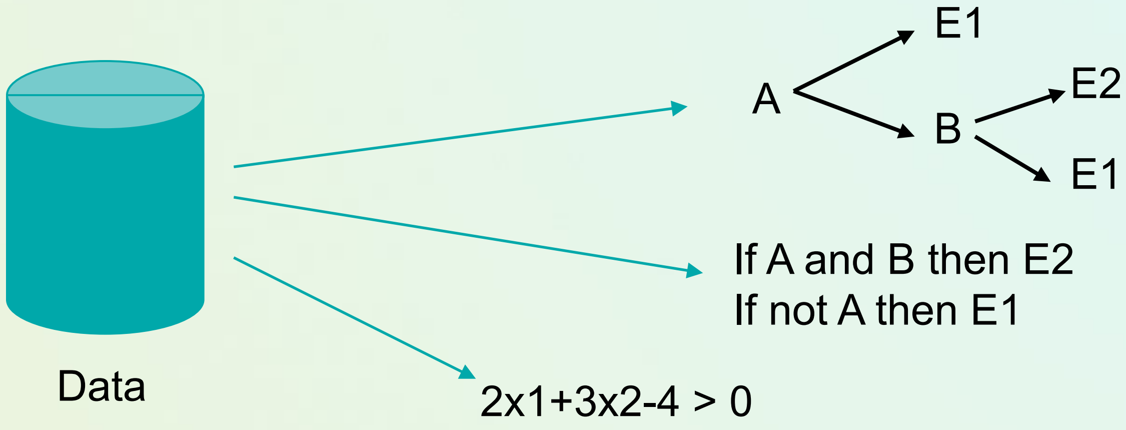
Motivations - data explosion problem

- Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories.
- More data is generated:
  - Bank, telecom, other business transactions ...
  - Scientific data: astronomy, biology, etc
  - Web, text, and e-commerce
- Very little data will ever be looked at by a human!
- <u>We are drowning in data, but starving for knowledge!</u>

Data Flood and Answers

- Data mining?
  - Extraction of useful information patterns from data
  - More than typical data analysis, machine learning or classical decision support!
- Knowledge Discovery is **NEEDED** to make sense and use of data.

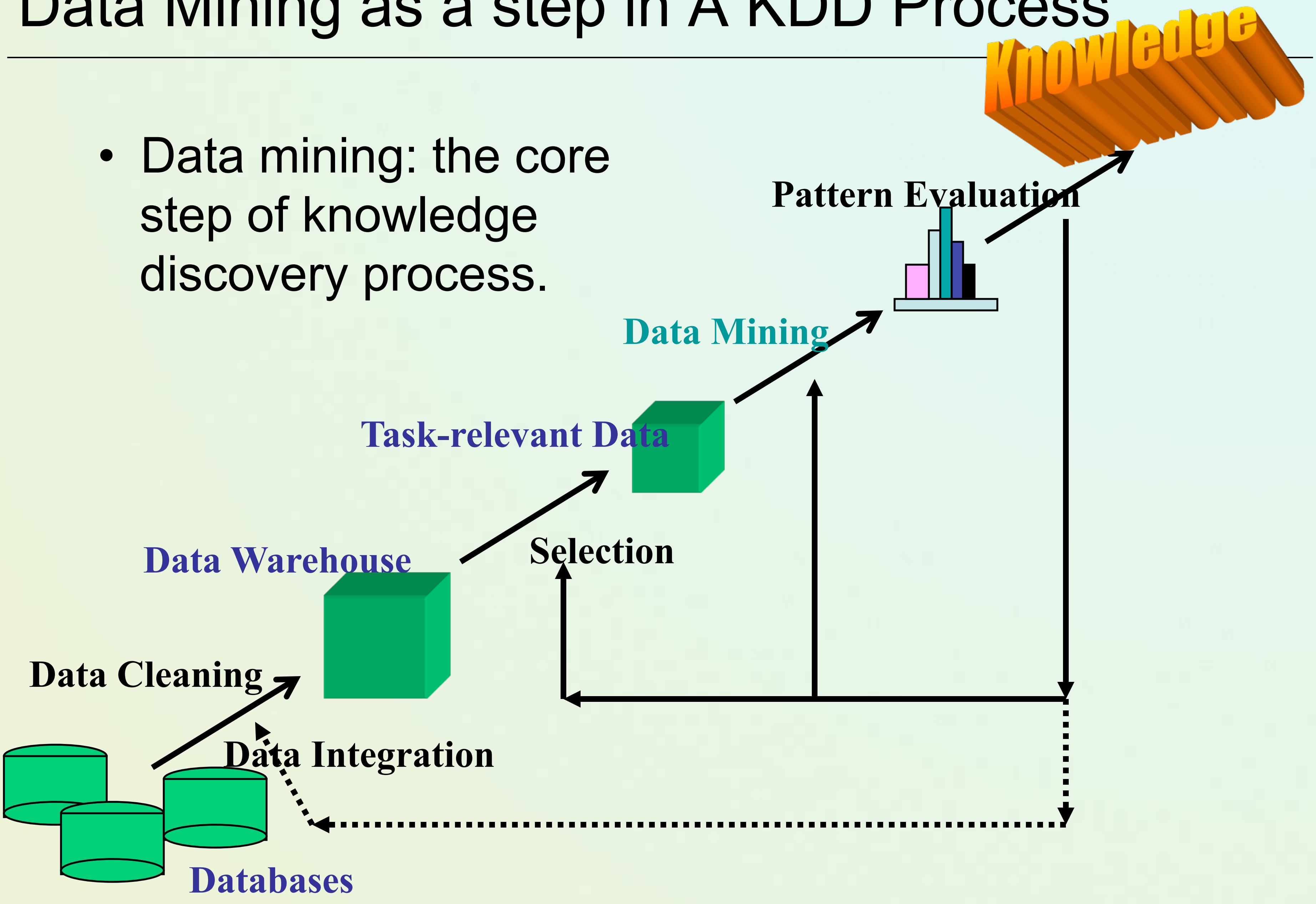
Data mining: what is it?



- Data mining is
  - Extraction of useful patterns from data sources, e.g., databases, texts, web, images.
- Patterns (knowledge representation) must be:
  - Valid, novel, potentially useful, understandable to the users.

Data Mining as a step in A KDD Process

- Data mining: the core step of knowledge discovery process.



Data Mining: On What Kind of Data?

- **Attribute-value tables (standard form / data table)**
- Multi-relational data / first order predicate calculus
- Structured data (graphs, workflows, ontologies, …)
- Sequence data bases
- Other more complex data repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Data streams
  - Text databases and multimedia databases
  - WWW resources
  - …

Flat files

- Actually the most common data source for data mining, especially at the research level.
- Simple data files in text or binary format with a structure known by the data mining algorithm to be applied.
- The data in these files can be transactions, time-series data, scientific measurements, etc.
- Big data – efficiency of access and management.

Instance	f 1	…	f k	Y
x1	V 1,1	…	V 1,k	V1,k+1
…	…	…	…	…
x i	V i,1	…	V i,k	Vi,k+1
…	…	…	…	…
xn	V n,1	…	V n,k	Vn,k+1

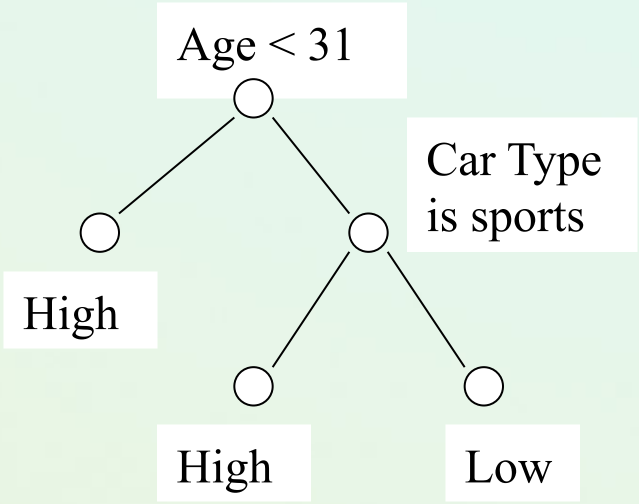
Types of attributes

- The most common distinction comes from measurement scale and statistics:
  - Nominal (also binary)
  - Ordinal
  - Interval-scaled
  - Ratio-scaled.
- Other names:
  - Categorical vs. numeric/continuous ones.
- Other types:
  - Criteria (preference-ordered), hierarchical, …

Decision trees

- Typical approach to the classification task.

Age	Car Type	Risk
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low



Numeric prediction – regression function

- Example: 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
…							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

- Linear regression function

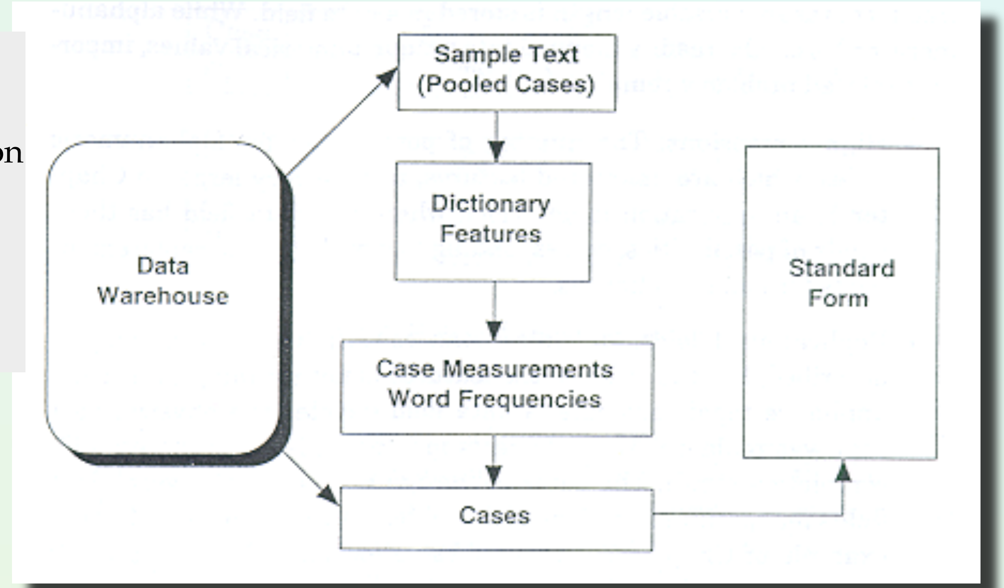
```
PRP =  -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
       + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

Transforming text documents into a standard form

- Transformation into Vector Representation

The d=7 **documents**:

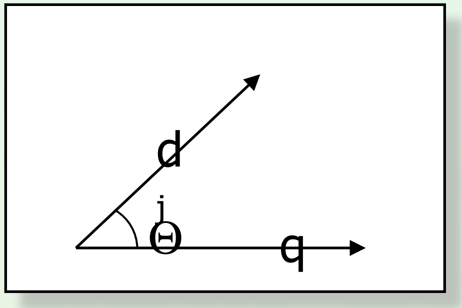
D1: Large Scale Singular Value Computations
D2: Software for the Sparse Singular Value Decomposition
D3: Introduction to Modern Information Retrieval
D4: Linear Algebra for Intelligent Information Retrieval
D5: Matrix Computations
D6: Singular Value Analysis of Cryptograms
D7: Automatic Information Organization



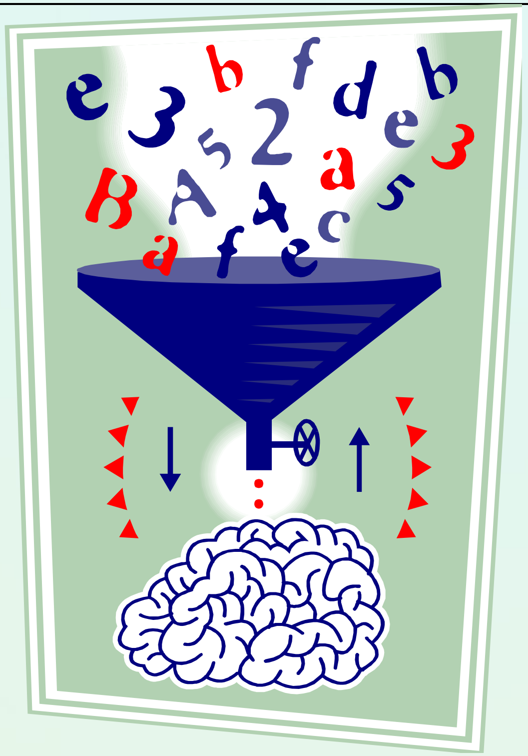
The t=5 **terms**:

T1: Information
T2: Singular
T3: Value
T4: Computations
T5: Retrieval

$$A = \begin{pmatrix} 0.00 & 0.00 & 0.71 & 0.71 & 0.00 & 0.00 & 1.00 \\ 0.58 & 0.71 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 \\ 0.58 & 0.71 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 \\ 0.58 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.71 & 0.71 & 0.00 & 0.00 & 0.00 \end{pmatrix}$$



Data Preparation for Knowledge Discovery



A crucial issue: The majority of time / effort is put there.

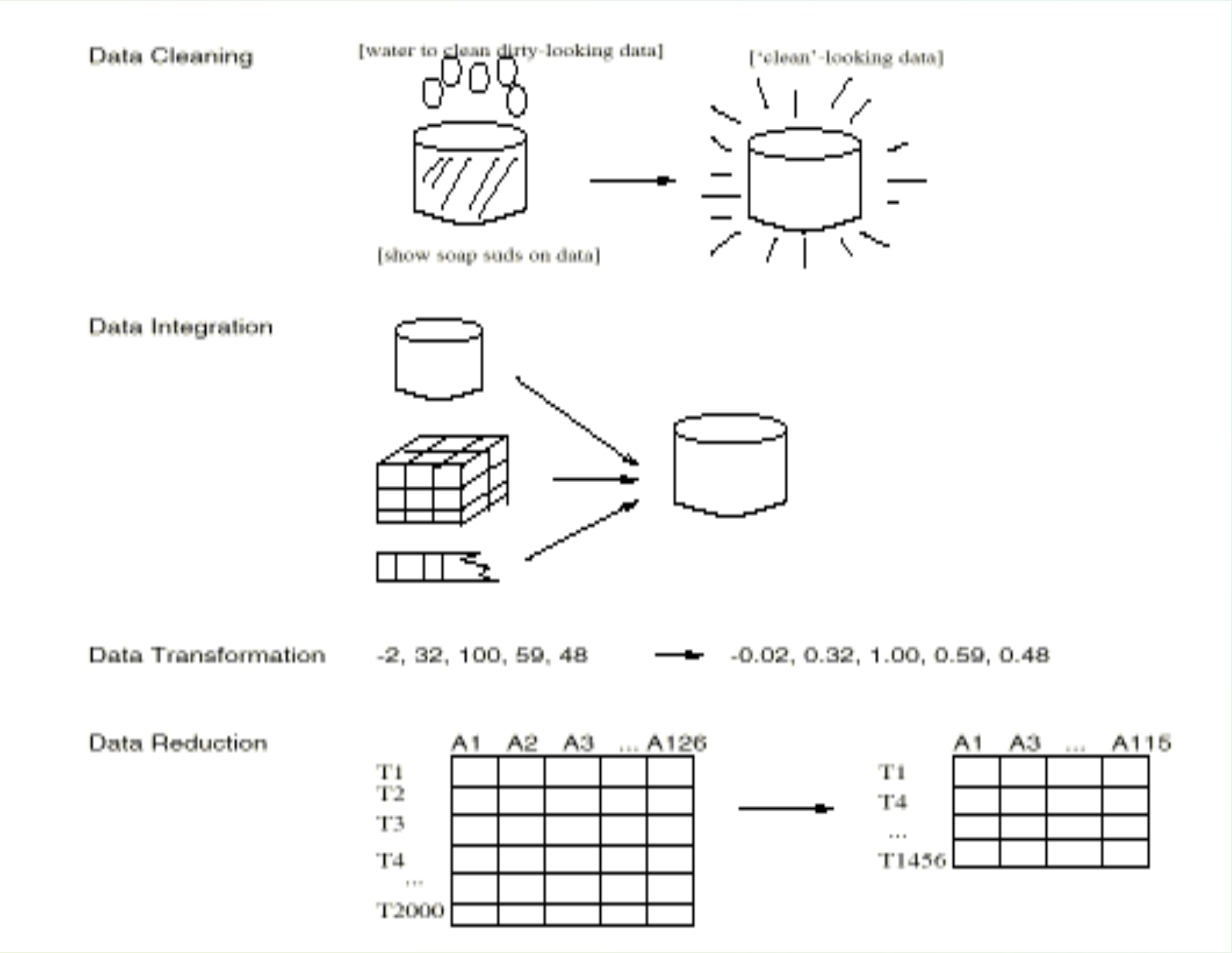
Data Understanding: Quantity

- Number of instances (records)
  - *Rule of thumb: 5,000 or more desired*
  - if less, results are less reliable; use special methods (boostrap sampling, …)
- Number of attributes (fields)
  - *Rule of thumb: for each field (attribute) find 10 or more instances*
  - If more fields, use feature reduction and selection
- Number of targets
  - *Rule of thumb: >100 for each class*
  - if very unbalanced, use stratified sampling or specific preprocessing (SMOTE, NCR, etc.)

Why Data Preprocessing?

- Data in the real world is „dirty” …
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - inconsistent: containing discrepancies (disagreements) in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., discrepancy between duplicate records

Basic forms of data preprocessing



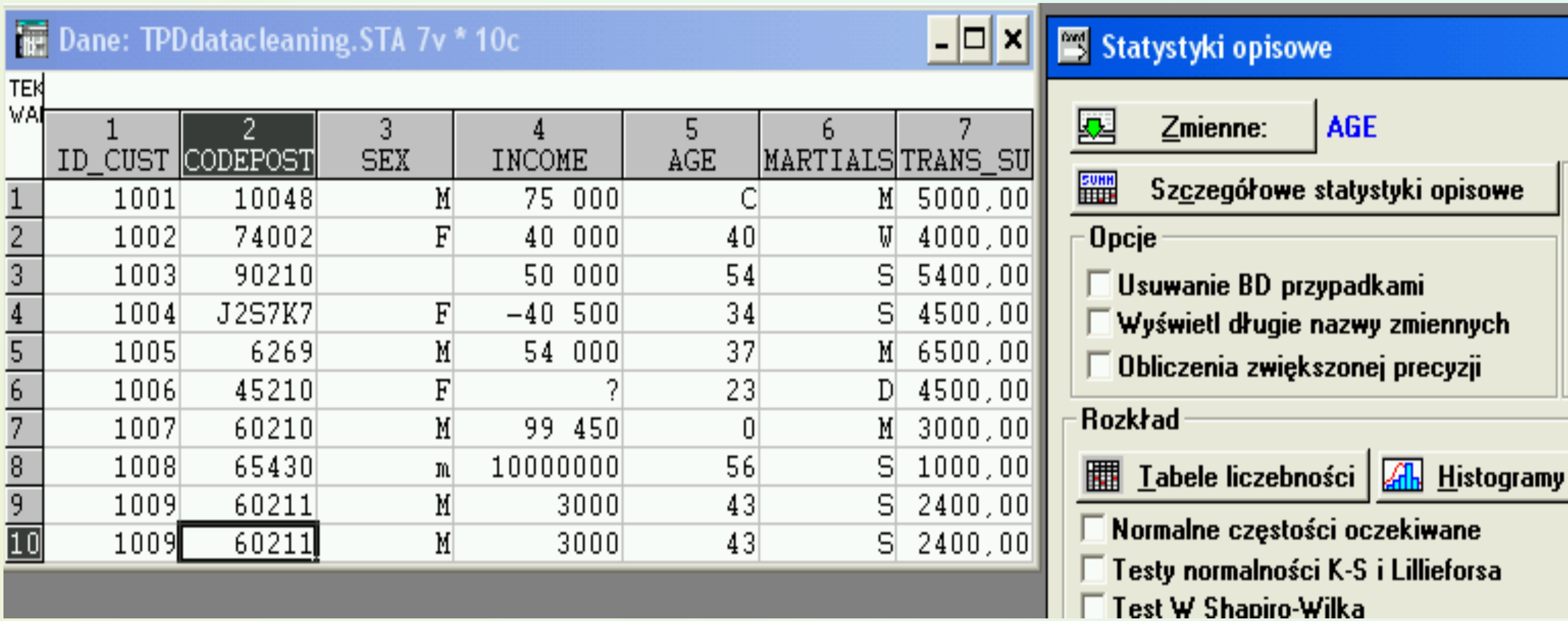
From J.Han's book

Basis problems in „Data Cleaning”

- Data „acquisition” / integration and metadata
- Unified formats and other transformations
- Erroneous values
- Missing values
- Data validation and statistics

Erroneous / Incorrect values

- What suspicious can you see in this table?



	1 ID_CUST	2 CODEPOST	3 SEX	4 INCOME	5 AGE	6 MARTIALS	7 TRANS_SU
1	1001	10048	M	75 000	C	M	5000,00
2	1002	74002	F	40 000	40	W	4000,00
3	1003	90210		50 000	54	S	5400,00
4	1004	J2S7K7	F	-40 500	34	S	4500,00
5	1005	6269	M	54 000	37	M	6500,00
6	1006	45210	F	?	23	D	4500,00
7	1007	60210	M	99 450	0	M	3000,00
8	1008	65430	m	10000000	56	S	1000,00
9	1009	60211	M	3000	43	S	2400,00
10	1009	60211	M	3000	43	S	2400,00

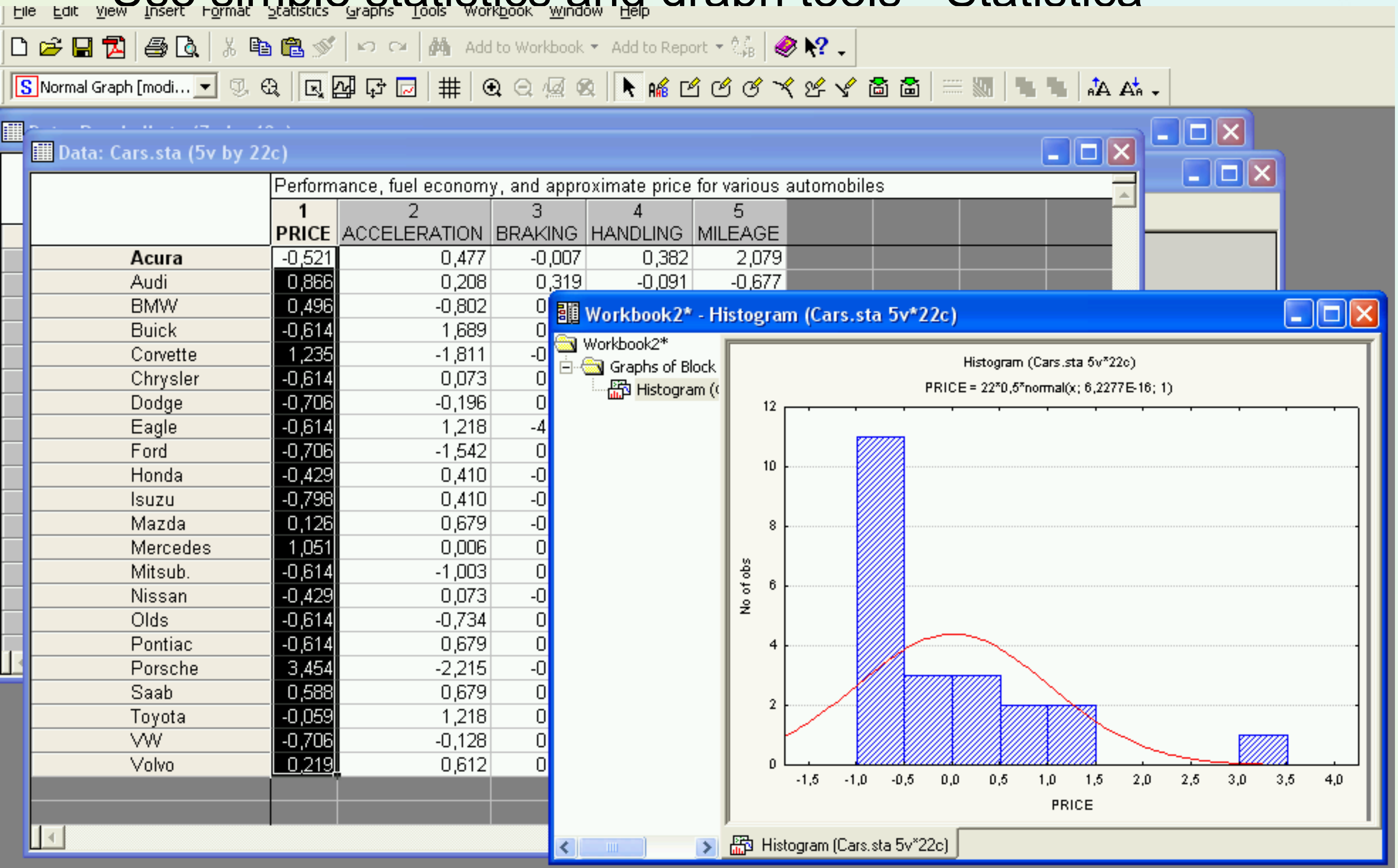
Incorrect values

- Reason: data has not been collected for mining it
- Result: errors and omissions that don't affect original purpose of data (e.g. age of customer)
- Typographical errors in nominal attributes $\Rightarrow$ values need to be checked for consistency
- Typographical and measurement errors in numeric attributes $\Rightarrow$ outliers need to be identified
- Errors may be deliberate (e.g. wrong zip codes)
- Other problems: duplicates, …

Tools?

Outliers – graphical identification

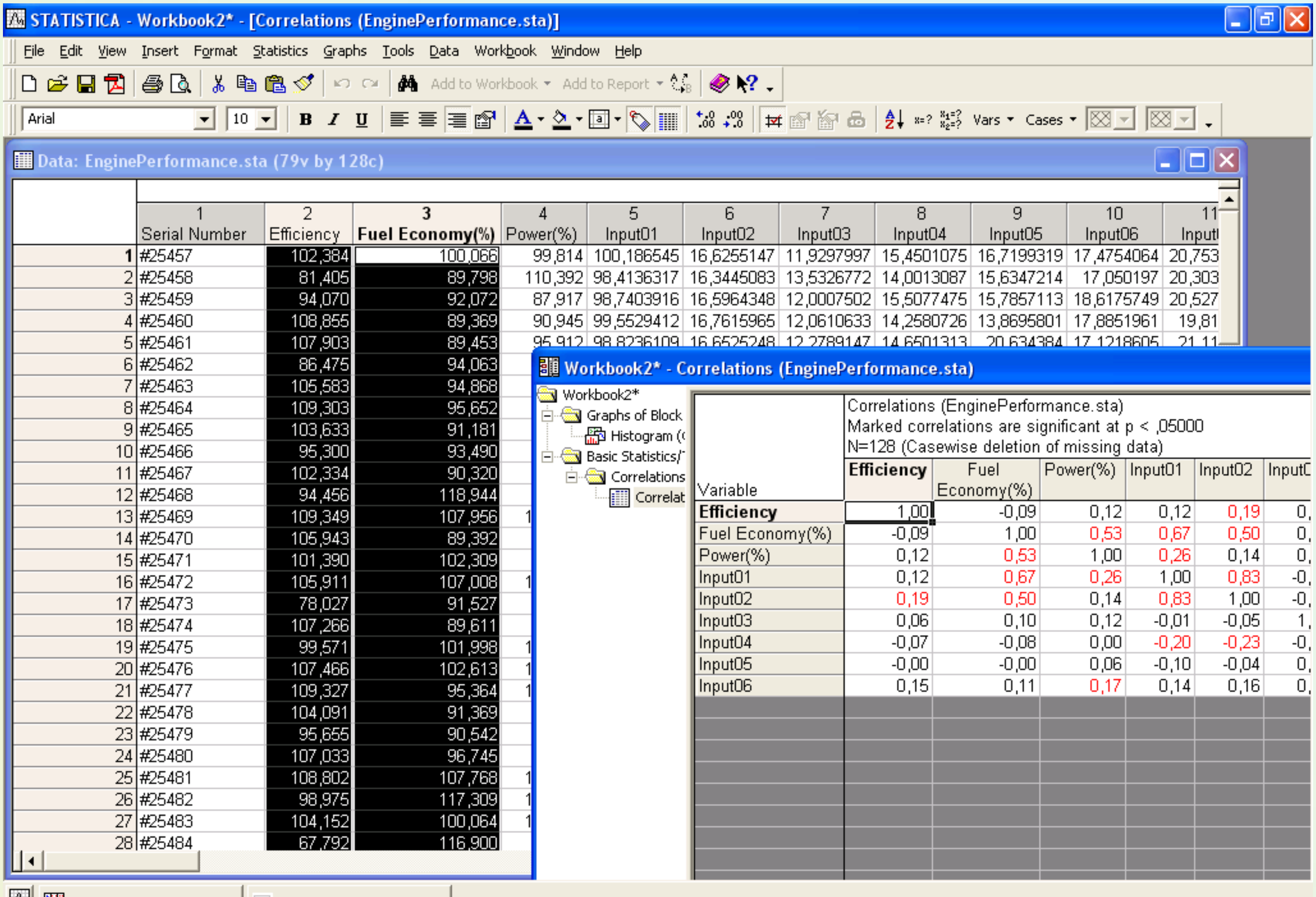
- Use simple statistics and graph tools - Statistica



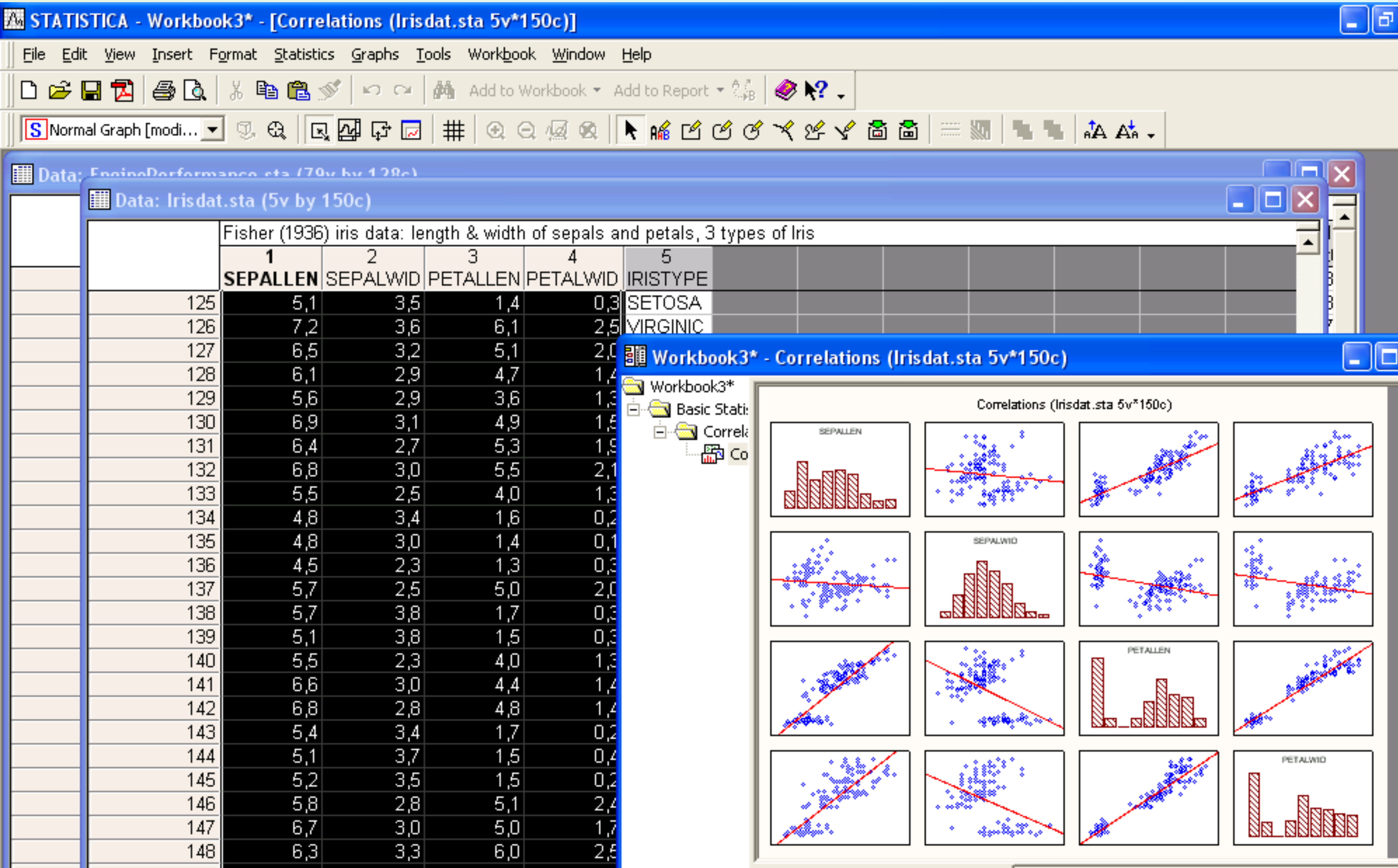
Redundant Data

- Redundant data occur often when integration of multiple databases
  - The same attribute may have different names in different databases
  - One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant data may be able to be detected by correlation analysis
- Large number of redundant data may slow-down or confuse knowledge discovery process.

Looking for correlated columns



Scatterplot matrix



Data Cleaning: Missing Values

- Missing data can appear in several forms:
  - <empty field> ? "0"   "." "999" "NA" …
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred – inputation!

Missing and other absent values of attributes

- Value may be missing because it is unrecorded or because it is inapplicable
- In medical data, value for **Pregnant?** attribute for **Jane** or **Anna** is missing, while for **Joe** should be considered **Not applicable**
- Don't care values

Hospital Check-in Database

Name	Age	Sex	Pregnant	..
Mary	25	F	N	
Jane	27	F	?	
Joe	30	M	-	
Anna	2	F	?	

Handle Missing Values

- Ignore / delete the instance: (not effective when the percentage of missing values per attribute varies considerably).
- Fill in the missing value manually: expert based + infeasible?
- Fill in a more advanced way :
  - a global constant : e.g., “unknown”, a new class? – don’t use it!
  - the attribute mean or the most common value.
  - the attribute mean for all examples belonging to the same class.
  - the most probable value: inference-based such as Bayesian formula or decision tree // prediction - regression model
  - result of global closest fit (distance base approaches)
  - Use a prediction technique

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a smaller, specified range
  - min-max normalization
  - z-score normalization
- Attribute/feature construction
  - New attributes constructed from the given ones

Data Transformation: Normalization

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

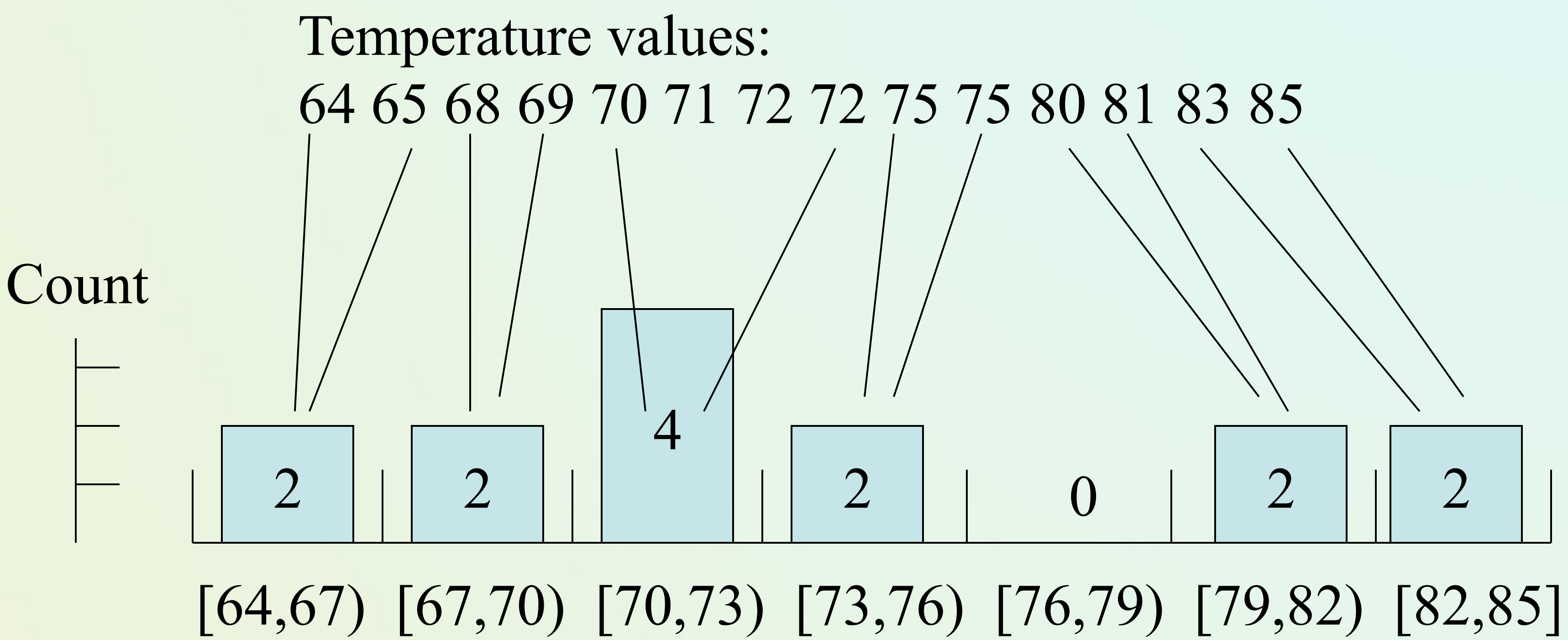
$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max($|v'|$)<1

Discretization

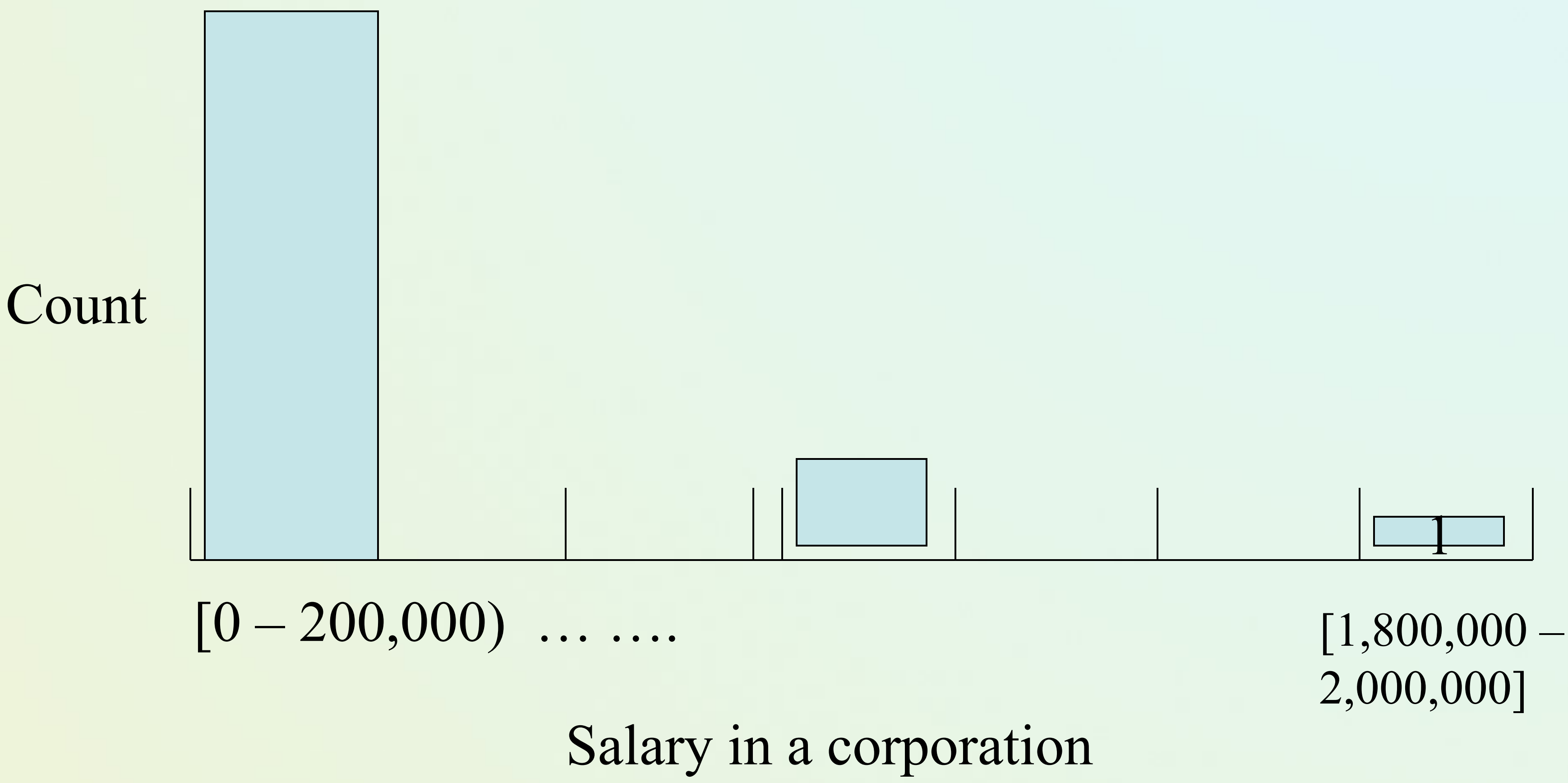
- Some methods require discrete values, e.g. most versions of Naïve Bayes, CHAID, Associations,
- Discretization → transformation of numerical values into codes / values of ordered subintervals defined over the domain of an attribute.
- Discretization is very useful for generating a summary of data
- Many approaches have been proposed:
  - Supervised vs. unsupervised,
  - Global vs. local (attribute point of view),
  - Dynamic vs. static choice of parameters

Discretization: Equal-Width (Length)



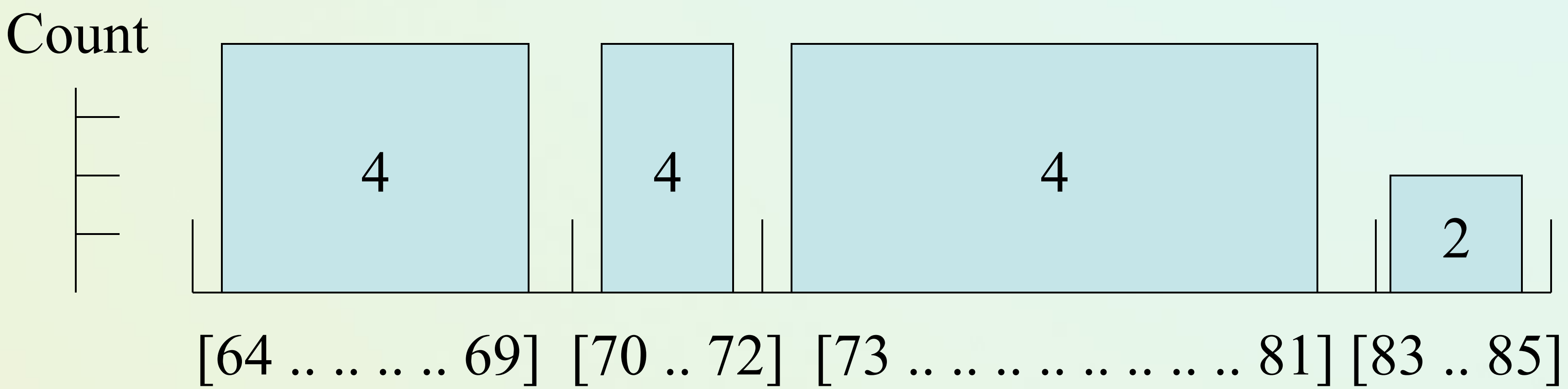
Equal Width, bins Low <= value < High

Discretization: Equal-Width may produce clumping



Discretization: Equal-Frequency

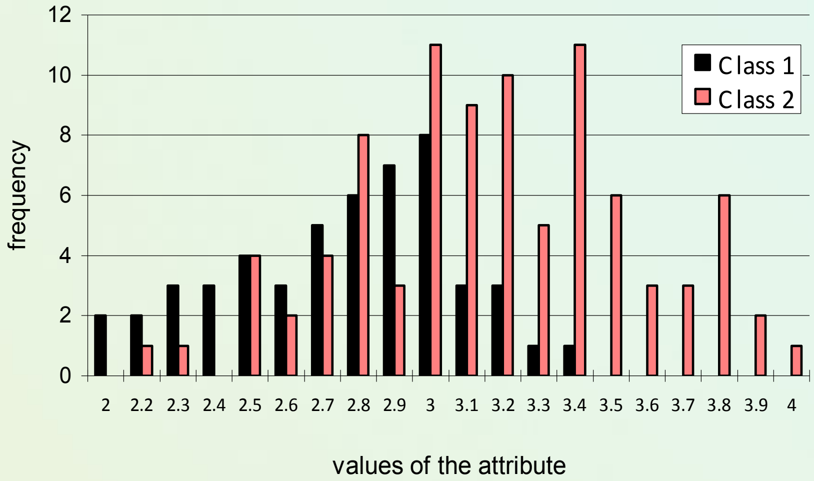
Temperature values:
64 65 68 69 70 71 72 72 75 75 80 81 83 85



Equal Height = 4, except for the last bin

Supervised (class) discretization

- Use information about attribute value distribution + class assignment.

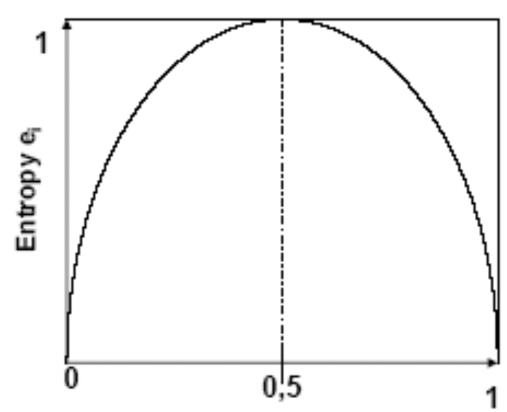


- Minimal entropy based approaches; Chi-Merge, others

Class entropy discretization

Evaluate purity of information about learning examples with **Entropy** (similar to decision tree split)

$$e_i = -\sum_{j=1}^{k} p_{ij} \log_2 p_{ij}$$



However, you also need a conditional entropy (with and attribute splitting)

Entropy-Based Discretization

- For learning examples S; If S discretized into two subintervals S1 i S2 using (cut point) T, conditional entropy is defined as:

$$E(S,T)=\frac{|S_1|}{|S|}Ent(S_1)+\frac{|S_2|}{|S|}Ent(S_2)$$

- Scan all possible cut points
- Choose the one mininizing the entropy.
- Continue until a stopping conditions such as

$$Ent(S)-E(T,S)>\delta$$

- MDL principle could be also exploited

A Toy example

- Starting entropy $$Ent(S) = -\frac{3}{6}\cdot\lg\frac{3}{6} - \frac{3}{6}\cdot\lg\frac{3}{6} = 1$$
- Attribute *IQ* and a cut point T=107 (inter. Left < T)

105	107	107	109	113	115
yes	no	no	no	yes	yes

$$Ent(S\,|\,T) = \frac{1}{6}(-1\cdot\lg 1) + \frac{5}{6}(-\frac{3}{5}\cdot\lg\frac{3}{5} - \frac{2}{5}\lg\frac{2}{5}) = 0.811$$

- Yet another cut point T=113 Ent(S|T) = 0.541 - the better choice.
- Fayyad and Irani theoretical advice – limit tested cut points

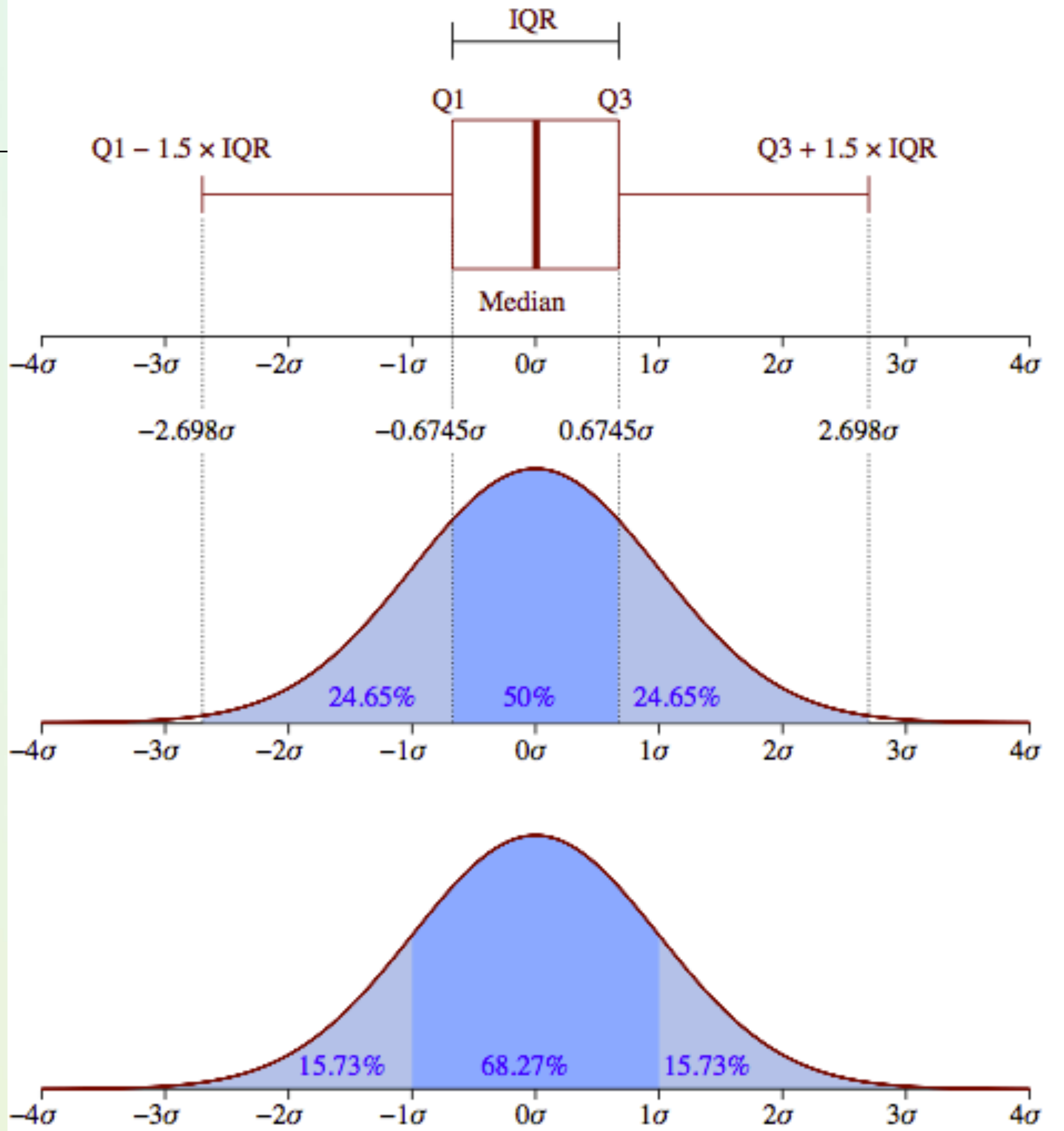
Outliers and Errors

- Outliers are values thought to be out of range.
- Approaches:
  - do nothing
  - enforce upper and lower bounds
  - let binning handle the problem

Examine Data Statistics

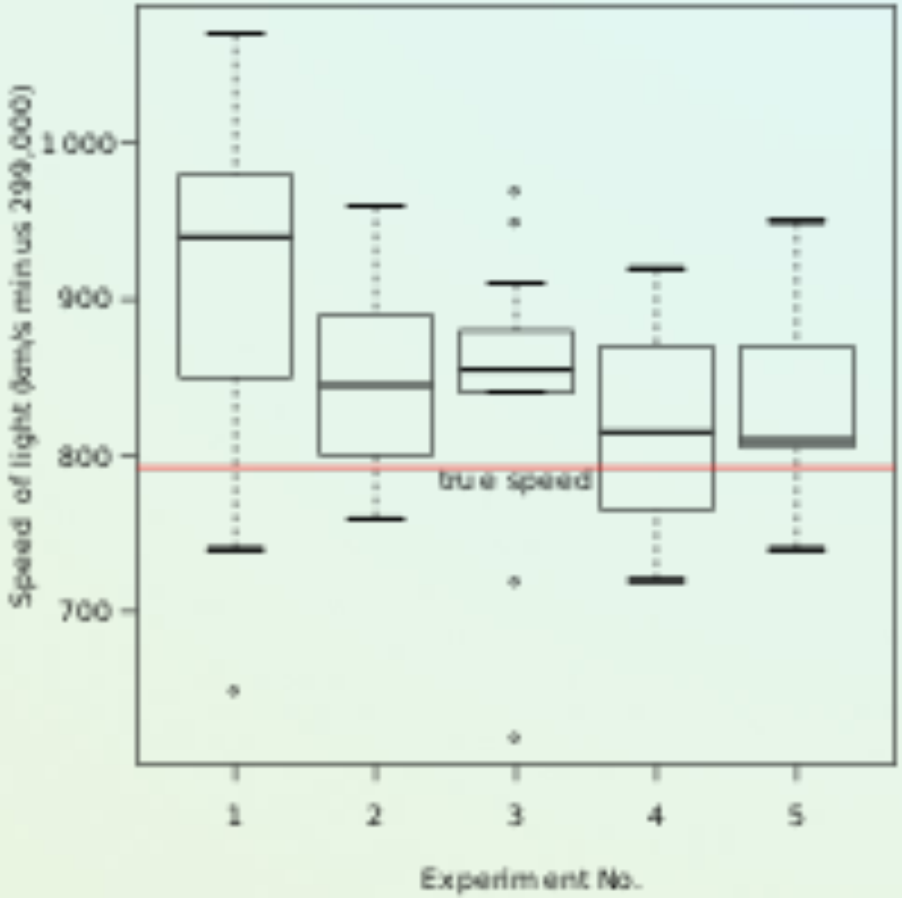
```
*****************      Field 9:  MILES_ACCUMULATED

Total entries = 865636 (23809 different values).  Contains non-numeric values.    Missing
data indicated by "" (and possibly others).

Numeric items = 165161,  high = 418187.000,  low = -95050.000
       mean = 4194.557,  std = 10505.109,  skew =    7.000

Most frequent entries:

       Value   Total
           :        700474 ( 80.9%)
          0:        32748 (  3.8%)
          1:        416 (  0.0%)
          2:        337 (  0.0%)
         10:        321 (  0.0%)
          8:        284 (  0.0%)
          5:        269 (  0.0%)
          6:        267 (  0.0%)
         12:        262 (  0.0%)
          7:        246 (  0.0%)
          4:        237 (  0.0%)
```

Outliers – quite far from typical distributions

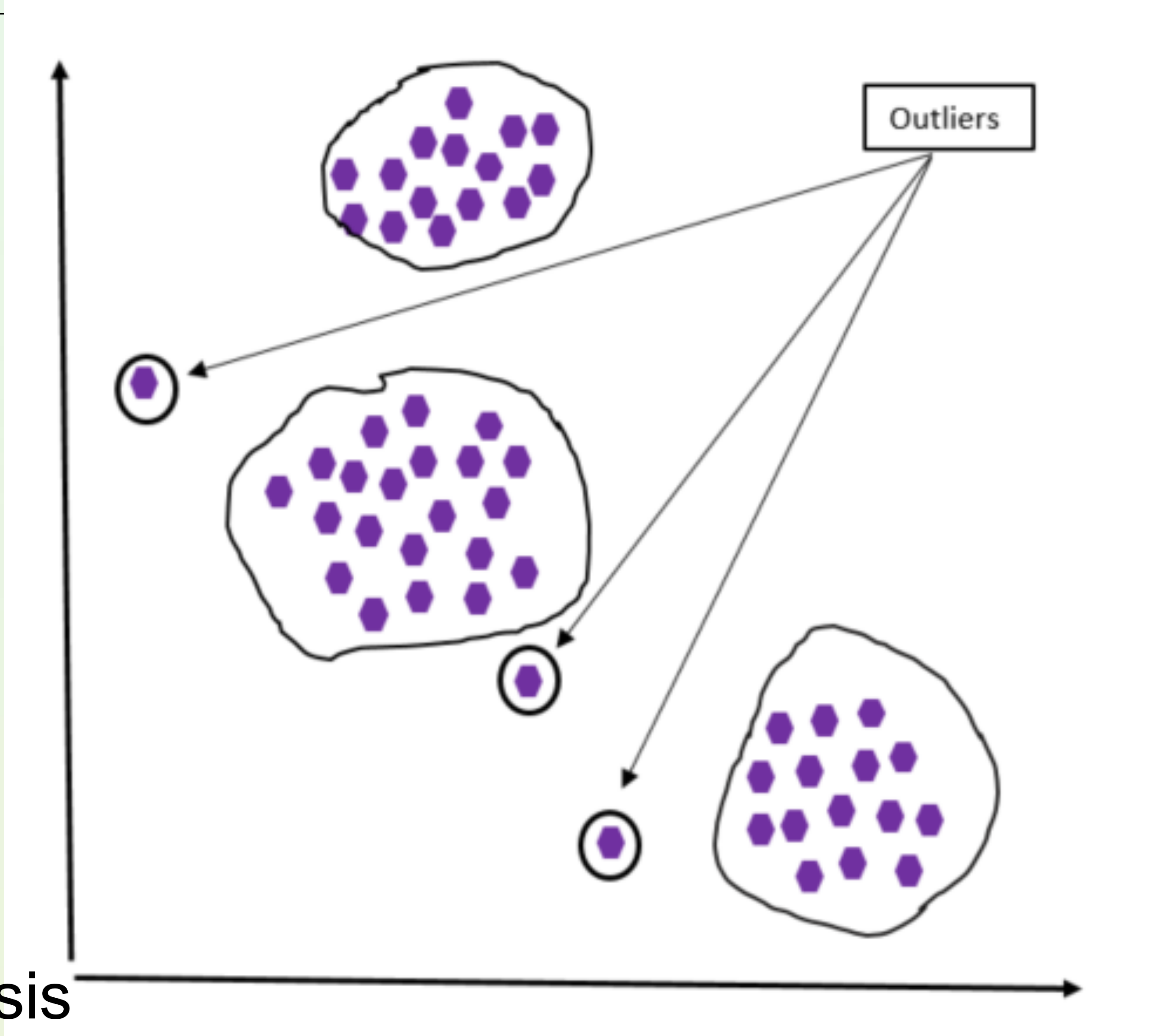


Box-plots

- Box plot of data from the Michelson–Morley experiment displaying four outliers in the middle column, as well as one outlier in the first column.

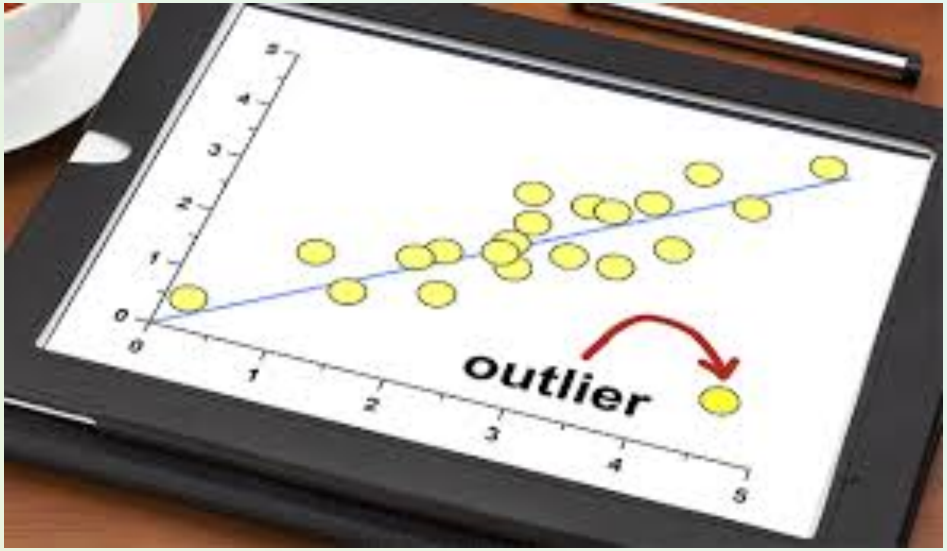


Multi-dimensional case



Cluster analysis

Another perspective - regression



- Numeric prediction – regression model
- Linear model $y = a_1x_1+a_2x_2+\ldots+a_mx_m$

Data Preprocessing: Attribute Selection

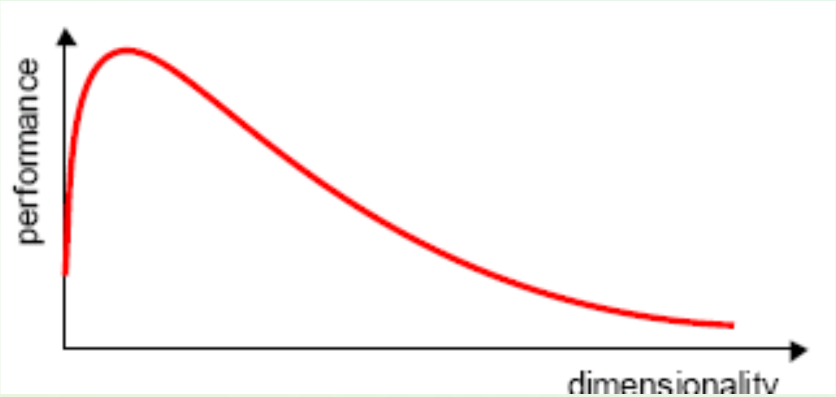
First: Remove fields with no or little variability

- Examine the number of distinct field values
  - *Rule of thumb: remove a field where almost all values are the same (e.g. null), except possibly in minp % or less of all records.*
  - *minp* could be 0.5% or more generally less than 5% of the number of targets of the smallest class
- More sophisticated (statistical or ML) techniques specific for data mining tasks
  - In WEKA see attribute selection

Too many attributes!

„Curse of dimensionality" [Bellman 1961]

- For a given sample size, there is a maximum number of features above which the performance of our classifier will degrade rather than improve!
- „the number of samples required per variable increases exponentially with the number of variables"



Toy classification example [D.Mladenic 2005]

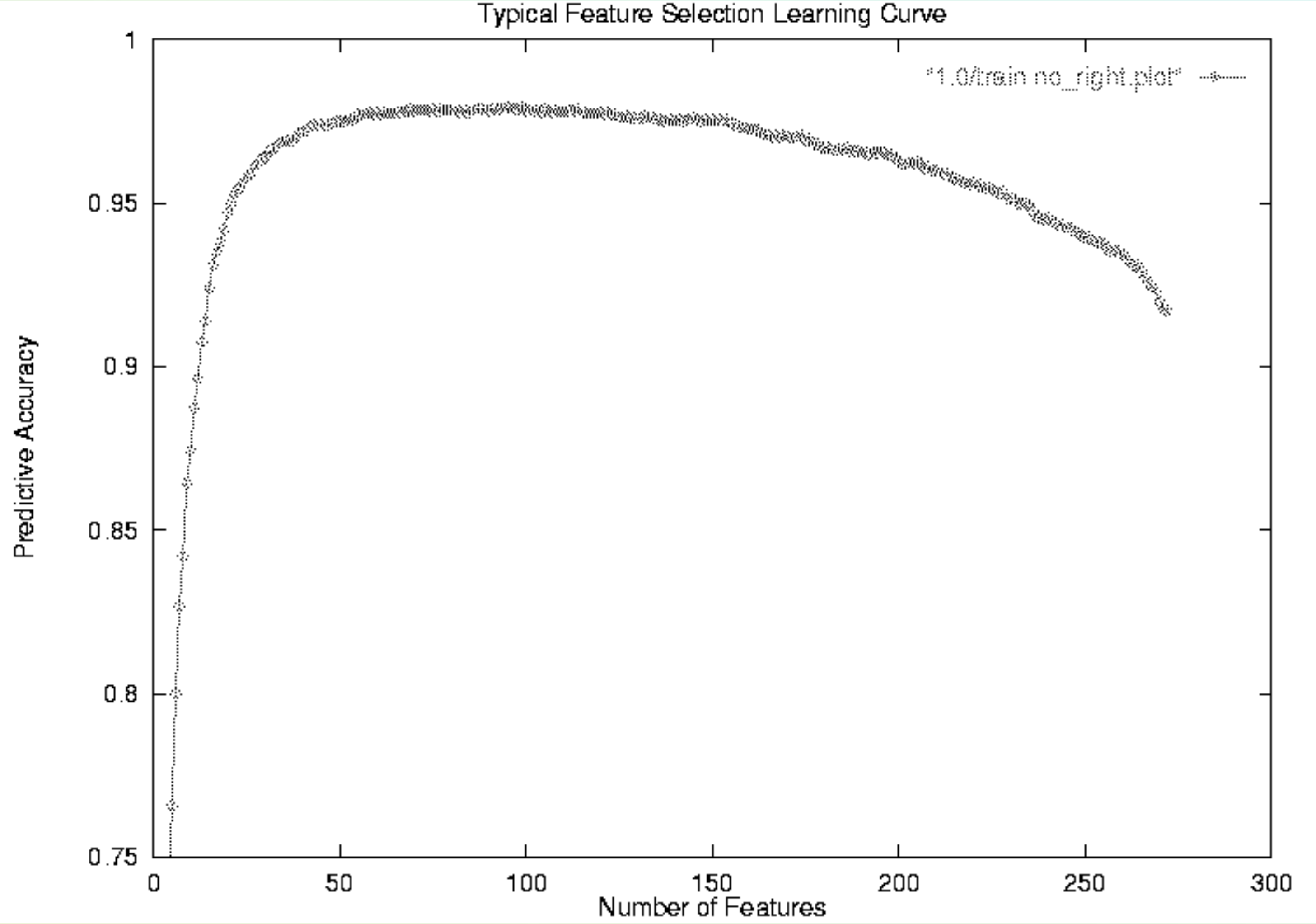
F_1	F_2	F_3	F_4	F_5	C
0	0	1	0	1	0
0	1	0	0	1	1
1	0	1	0	1	1
1	1	0	0	1	1
0	0	1	1	0	0
0	1	0	1	0	1
1	0	1	1	0	1
1	1	0	1	0	1

- Data set
  - Five Boolean features
  - $C = F_1 \vee F_2$
  - $F_3 = \neg F_2$, $F_5 = \neg F_4$
  - Optimal subset:

    $\{F_1, F_2\}$ or $\{F_1, F_3\}$
- optimization in space of all feature subsets (2^F possibilities)

(tutorial on genomics [Yu 2004])

Real working K-NN with many attributes



Different attribute selection methods

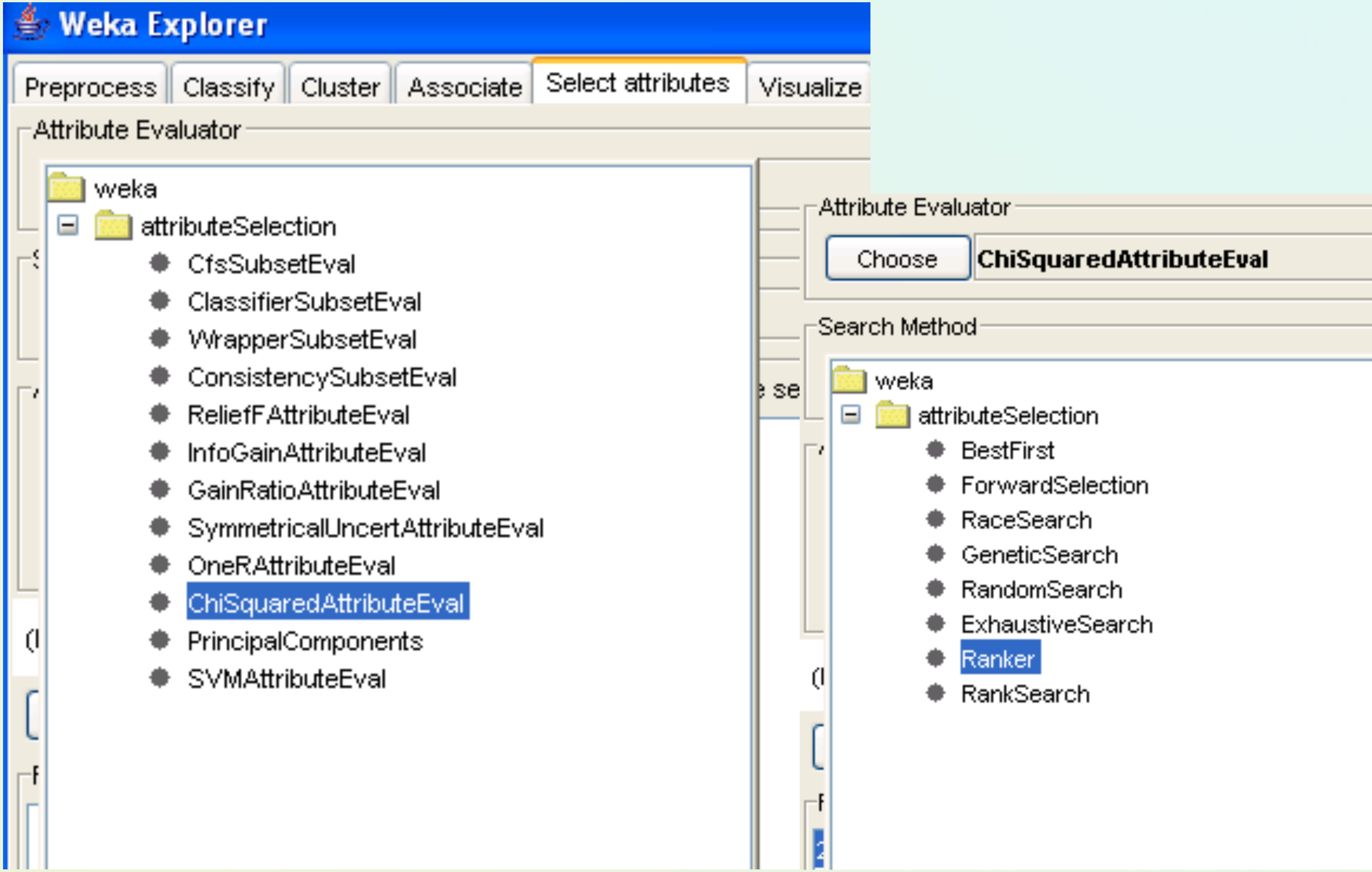
Filtering (single attributes)

- Correlation-based measure.
- Contextual-merit.
- Info-Gain.
  - Gain ratio
- Chi-squared statistic
- Liu Consistency measure

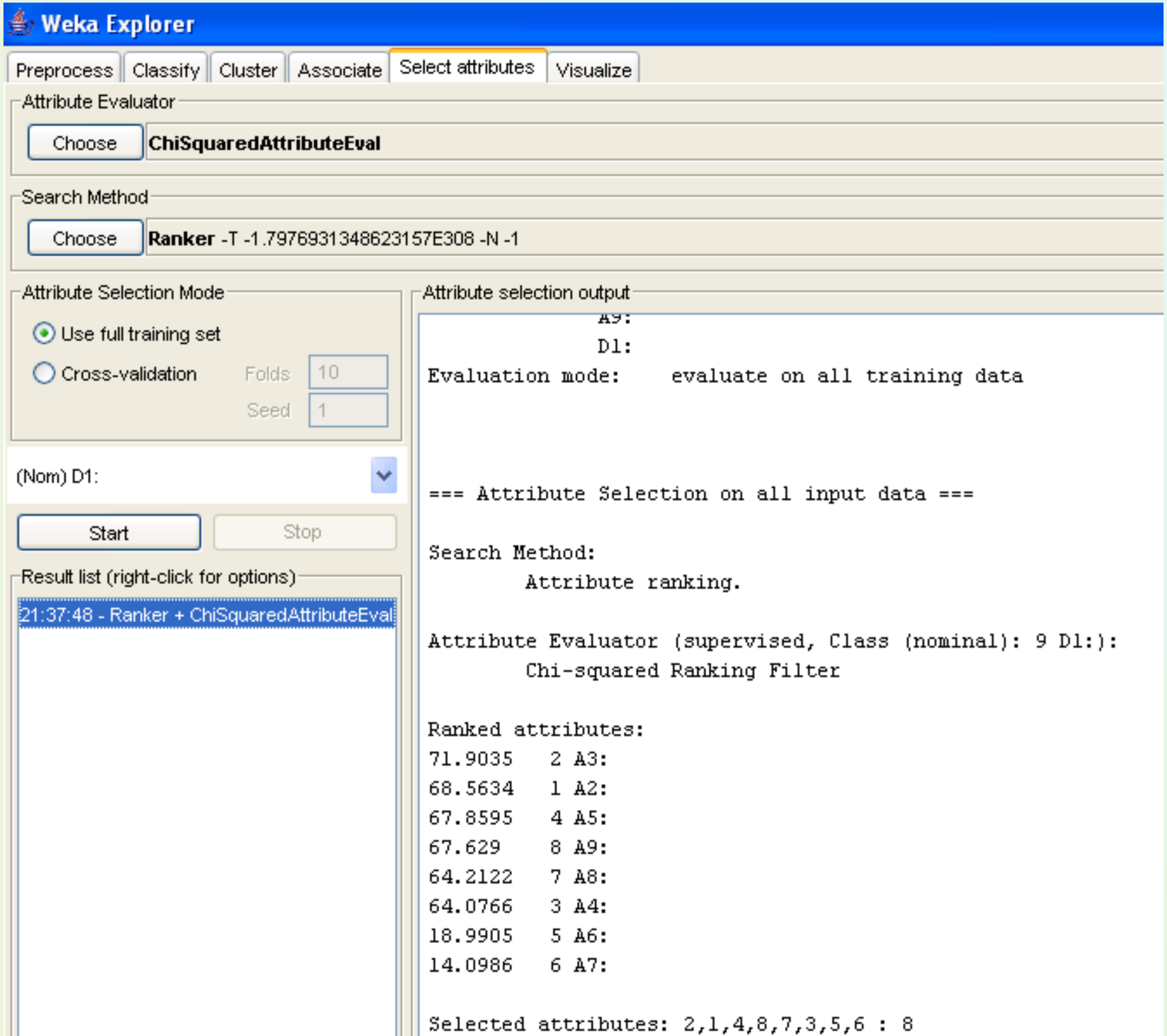
Subsets and more advanced search

- Relief method
- Wrapper model

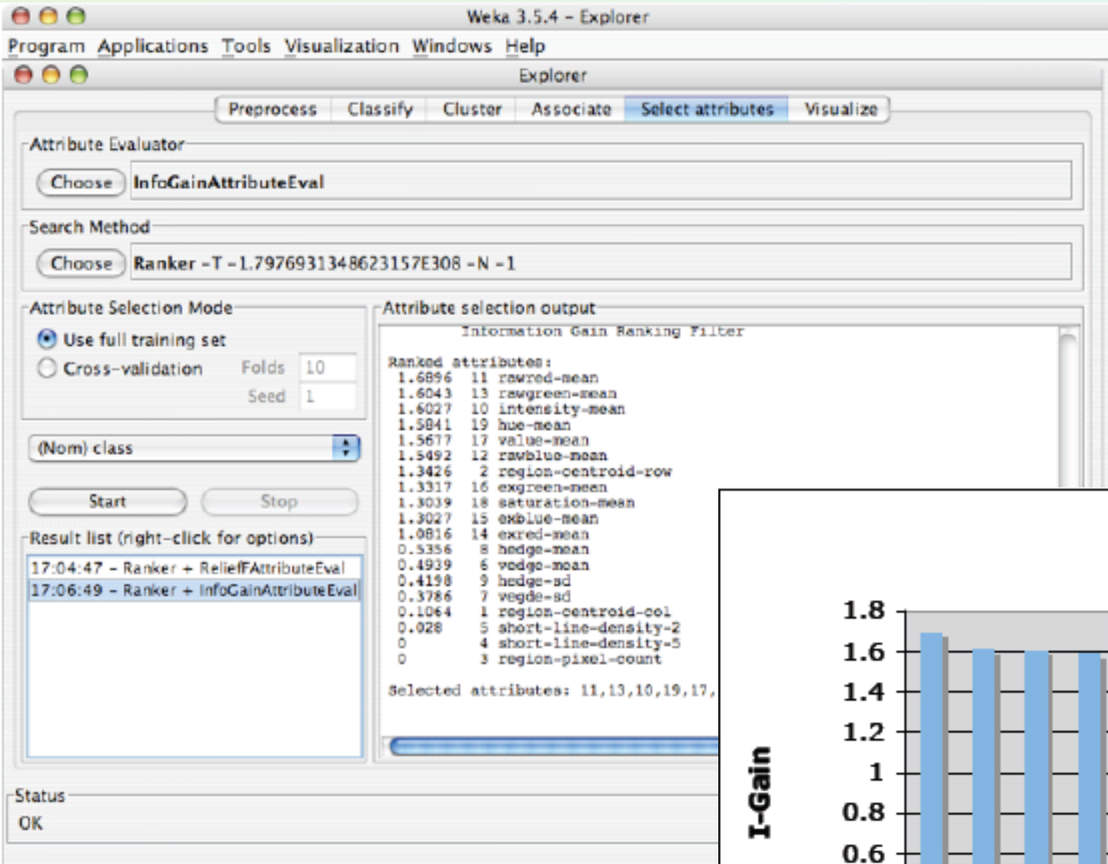
WEKA – attribute selection tools



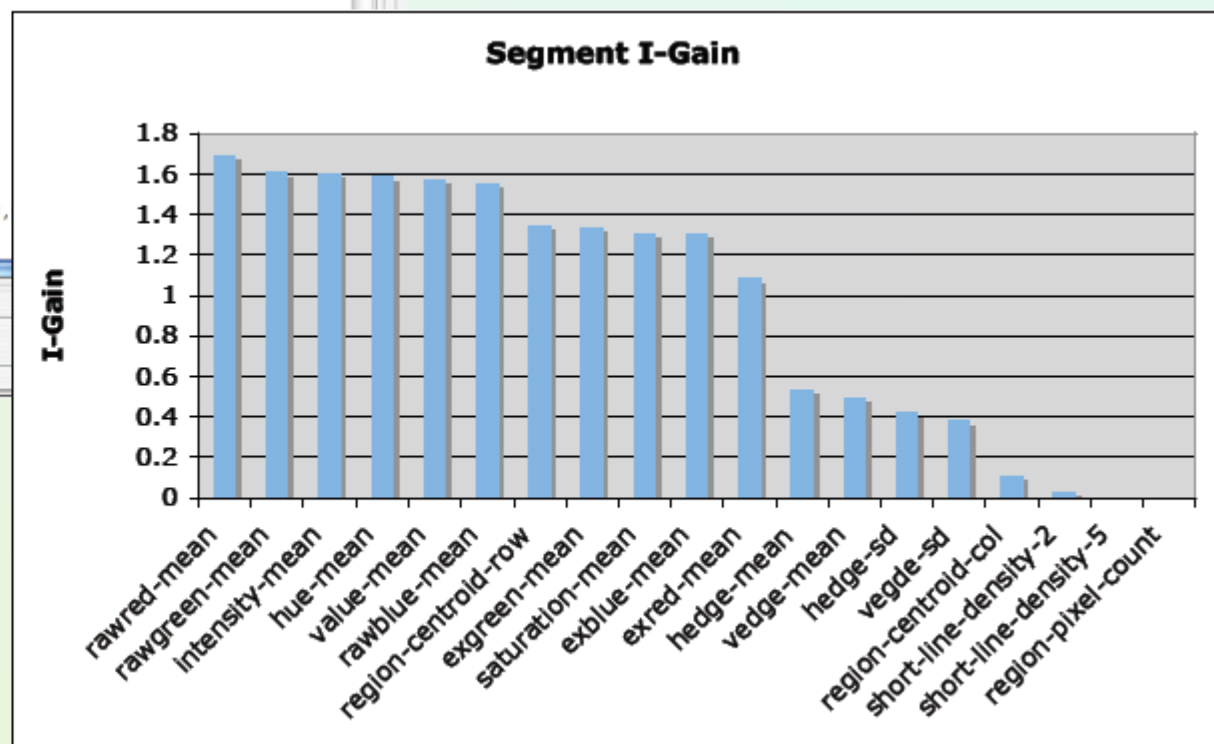
Ranking with …? WEKA



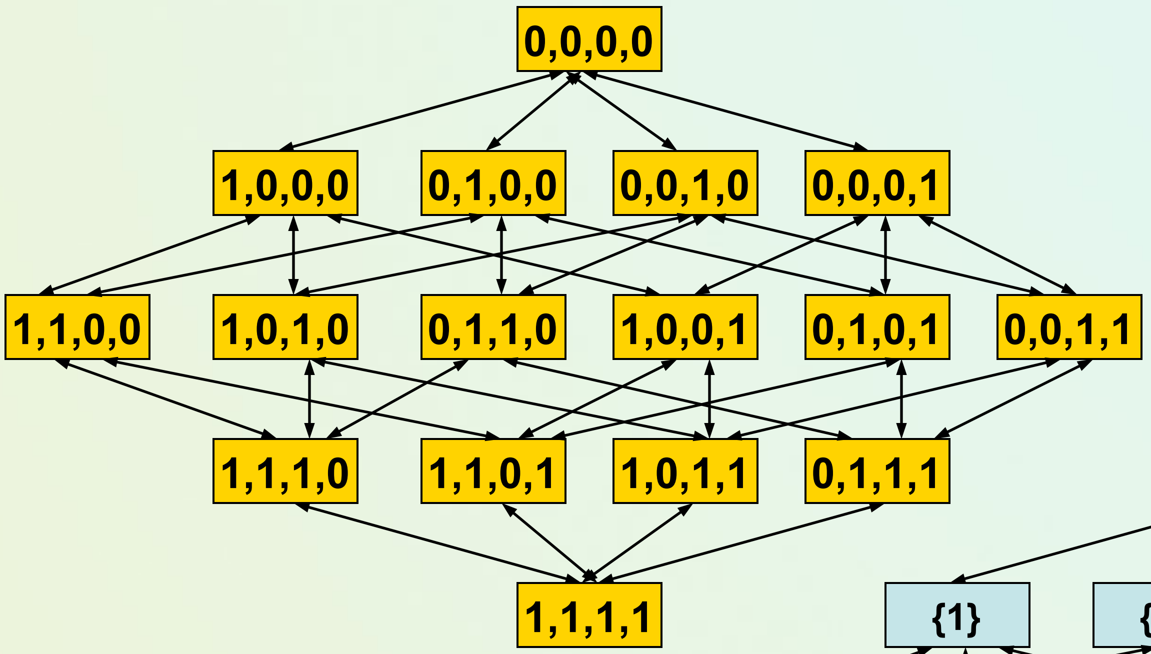
How could you exploit rankng



- Find thershold τ
- Median or smth else?



Search in Subset Space



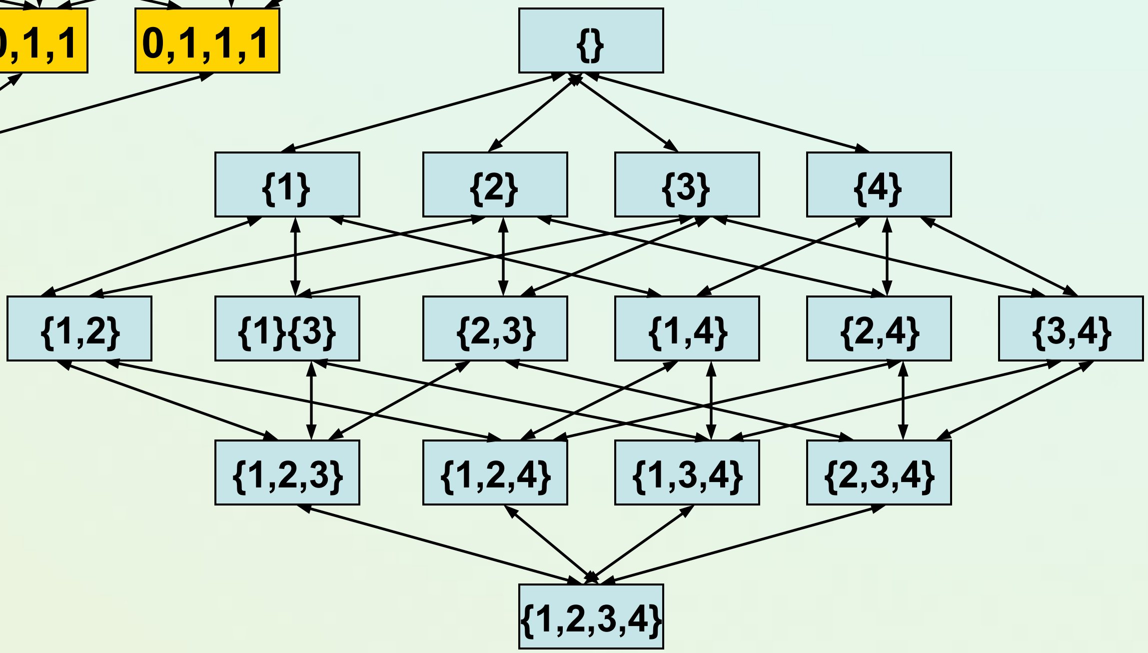
Subset Inclusion State Space

Poset Relation: Set Inclusion

***A* ≤ *B* = "B is a subset of A"**

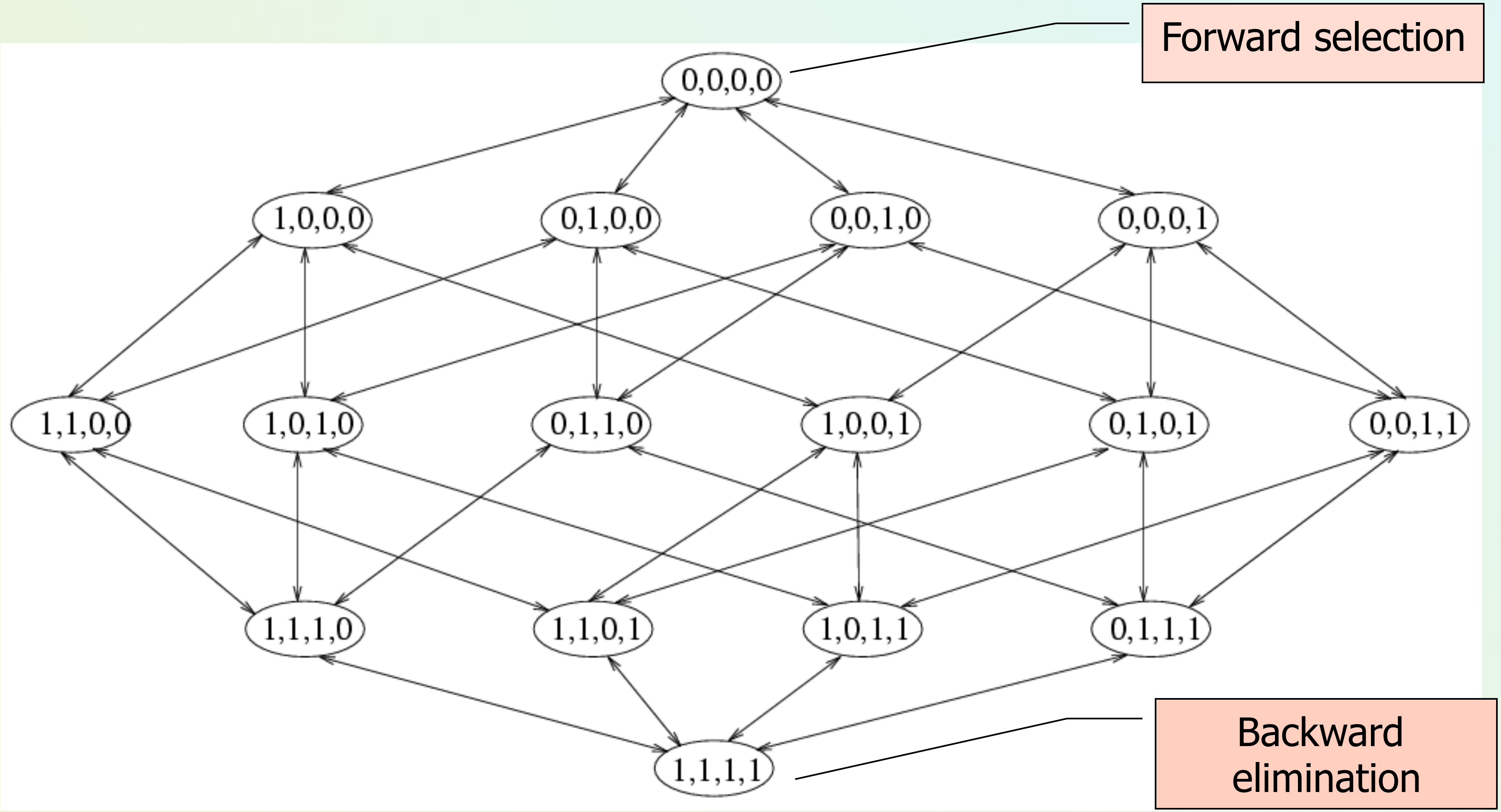
"Up" operator: DELETE

"Down" operator: ADD



How to move in the space

- An example of search space (*John & Kohavi 1997*)

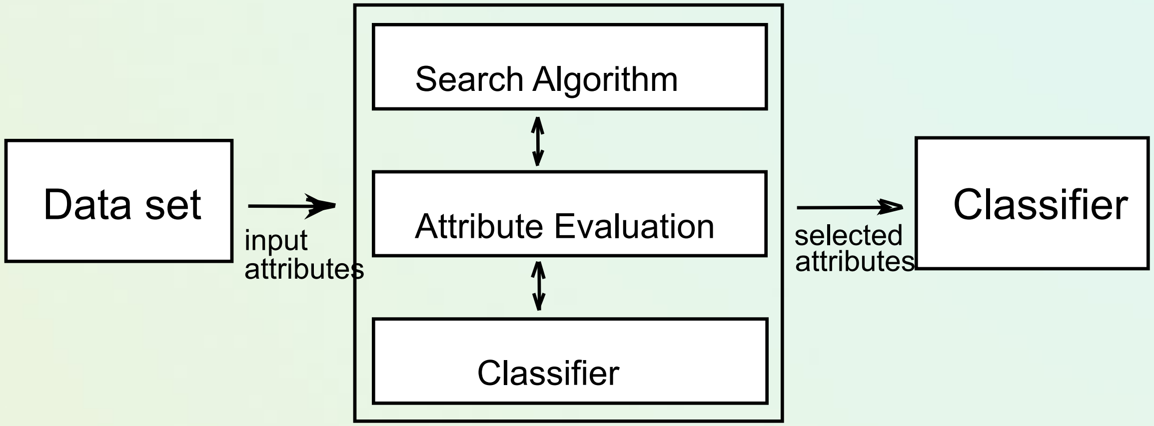


Heuristic Subset Search Techniques

- There are *2^d* possible sub-features of *d* features
- Several heuristic feature selection methods:
  - Best step-wise feature selection (forward):
    - The best single-feature is picked first
    - Then next best feature condition to the first, ...
  - Step-wise feature elimination (backward):
    - Repeatedly eliminate the worst feature
  - Best combined feature selection and elimination
  - Partly non-deterministic seach (genetic and other techqniques)

Wrapper approach

- Filter vs. Wrapper approach (Kohavi et al. 94, and …)



- The classifier is used by the evaluation function
- Search algorithms:
  - Forward selection
  - Backward elimination
  - …

Constructing new attribute

- Following A.Berge – find new attributes

- In general - two approaches for dimensionality reduction
  - Feature selection: choose a subset of the features

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} \longrightarrow \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_m} \end{bmatrix}$$

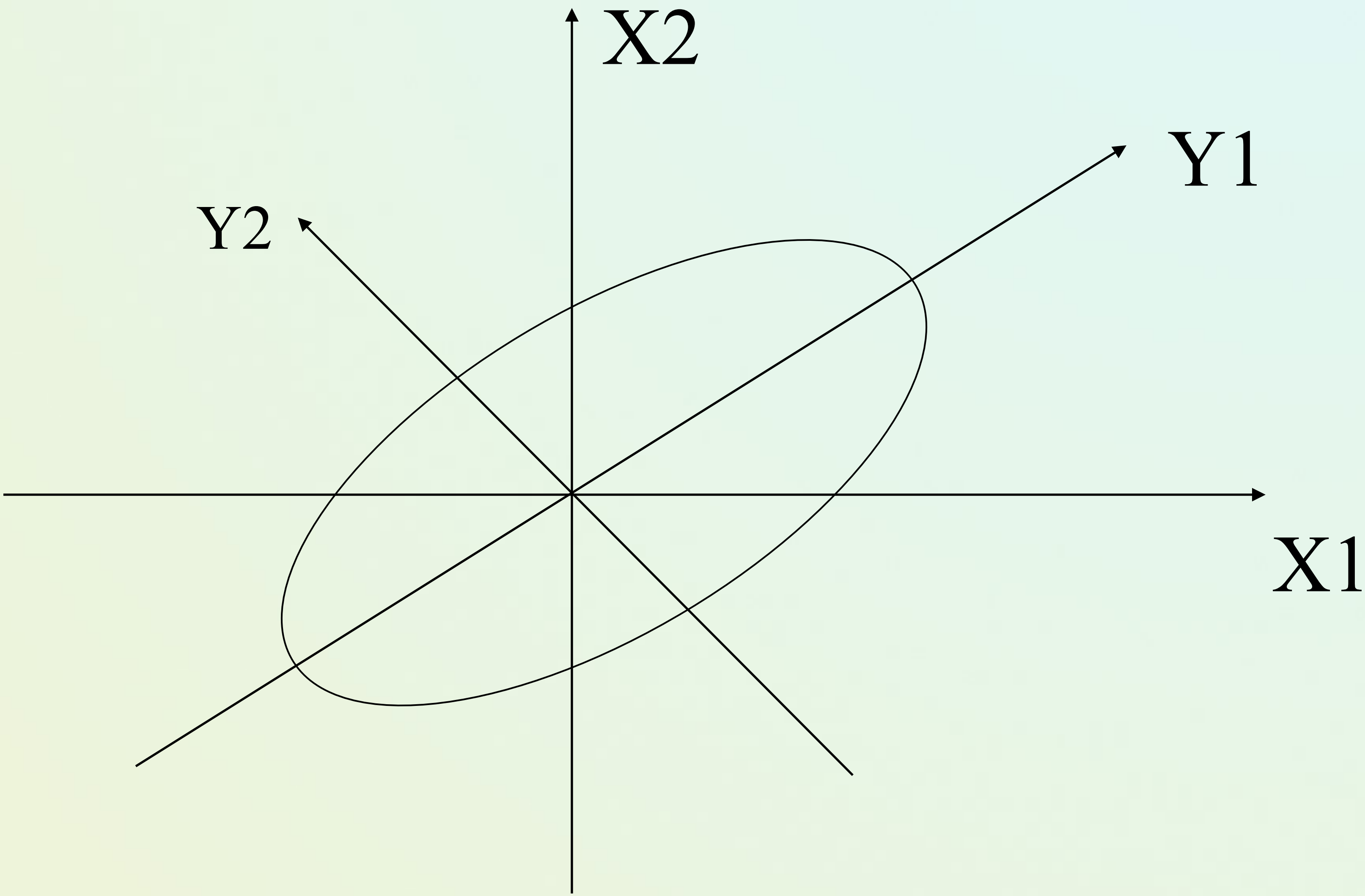
  - Feature extraction: create a subset of new features by combining existing features

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} \longrightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix}\right)$$

Principal Component Analysis (PCA)

- Given N data vectors from n-dimensions, find $k \le n$ orthogonal vectors (*principal components*) that can be best used to represent data
- Steps
  - Normalize input data: Each attribute falls within the same range
  - Compute k orthonormal (unit) vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the k principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance. (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data
- Works for numeric data only
- Used when the number of dimensions is large

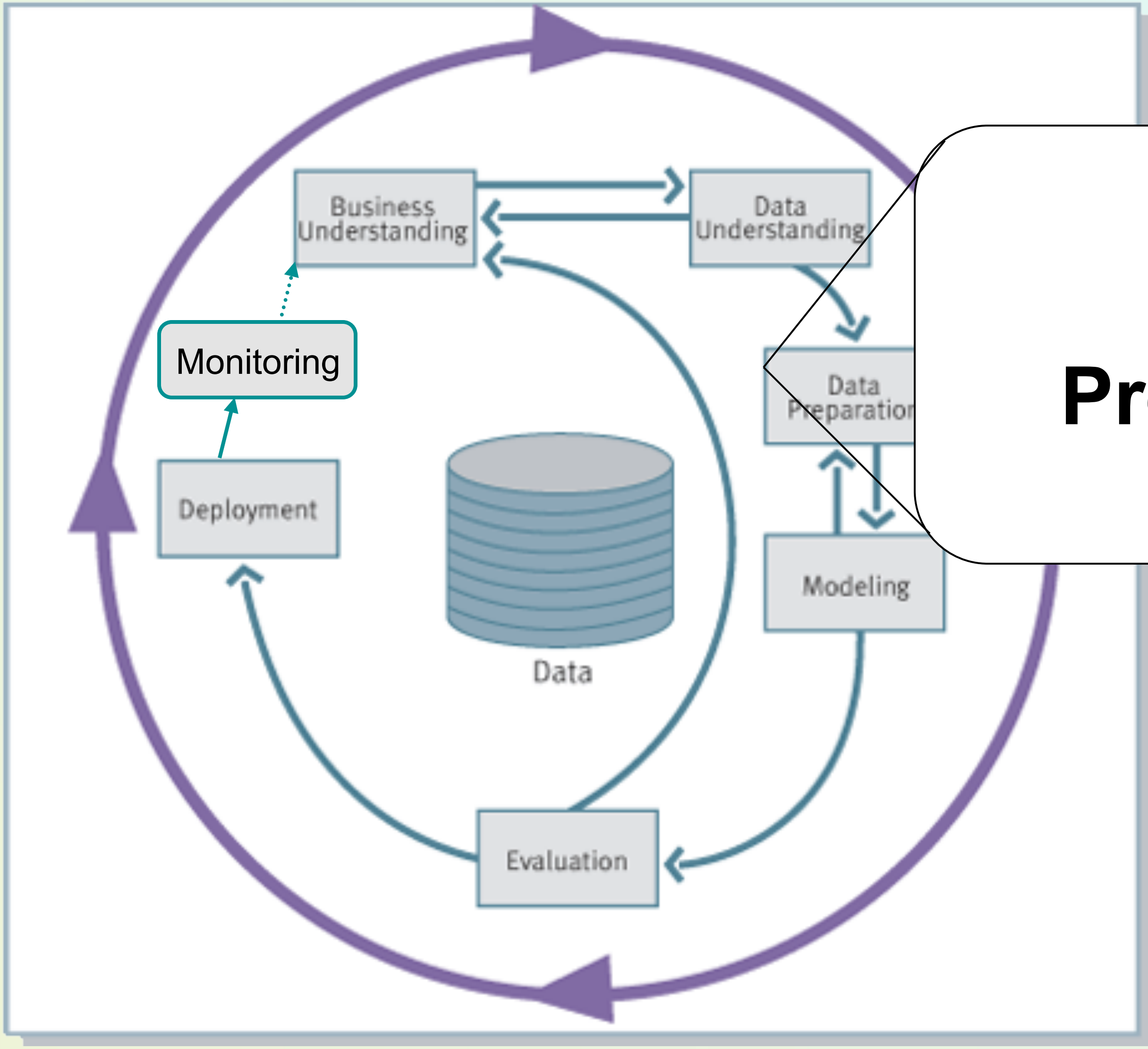
Principal Component Analysis



Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Descriptive data summarization is need for quality data preprocessing
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot a methods have been developed but data preprocessing still an active area of research

Knowledge Discovery Process, in practice



Data Preparation

Data Preparation estimated to take 70-80% of the time and effort

Any questions, remarks?

?