Data Mining
Data Preprocessing

Lecturer: JERZY STEFANOWSKI
Institute of Computing Sciences
Poznan Univeristy of Technology
Poznan, Poland

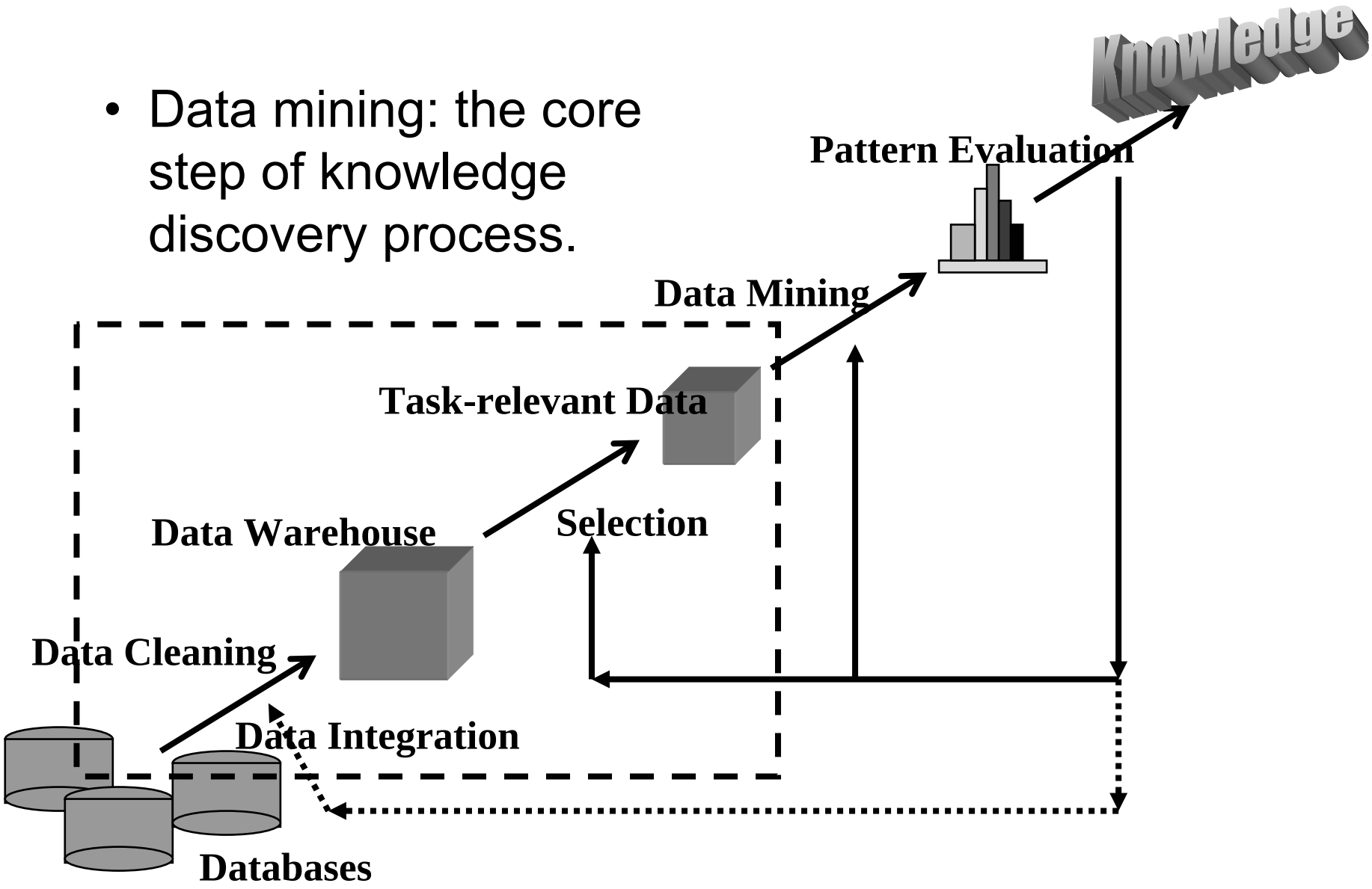
Lecture 3b→4
SE Master Course
Update for edition 2009/2010

Outline

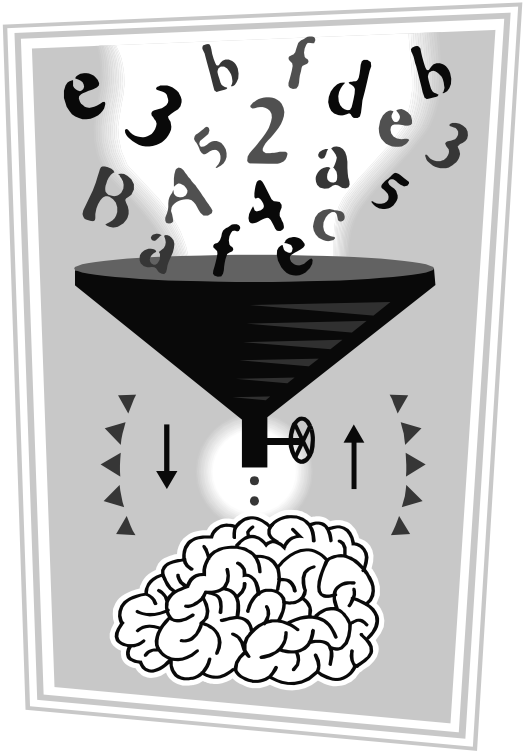
1. Motivations
2. Data integration and cleaning
   - Errors
   - Missing values
   - Noisy Data
   - Outliers
3. Transformations
4. Discretization of numeric values
5. Data reduction (Attribute selection)

Data Mining a step in A KDD Process

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



Data Preparation for Knowledge Discovery



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

Data Understanding: Relevance

- What data is available for the task?
- Is this data relevant?
- Is additional relevant data available?
- How much historical data is available?
- Who is the data expert ?

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 imbalanced, use stratified sampling or specific preprocessing (SMOTE, NCR, etc.)

Why Data Preprocessing?

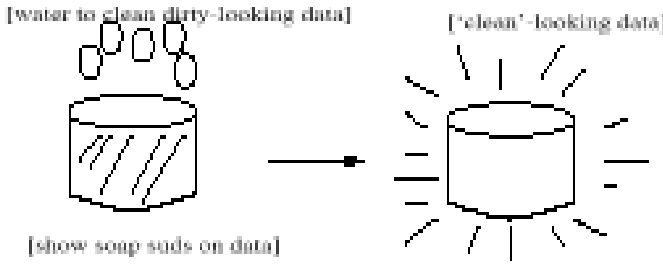
- Data in the real world is „dirty" …
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation="”
  - noisy: containing errors or outliers
    - e.g., Salary="-10"
  - inconsistent: containing discrepancies (disagreements) 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

Why Is Data Preprocessing Important?

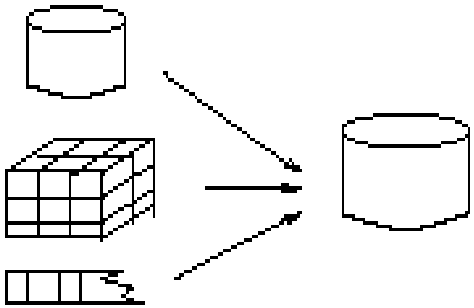
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading basic descriptive statistics.
  - Data warehouse needs consistent integration of quality data!
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse [J. Han].
- Economic benefits of data cleaning:
  - Data warehouse contains data that is analyzed for business decisions.
  - Knowledge discovered from data will be used in future.
  - Detecting data anomalies and rectifying them early has huge payoffs.

Basic forms of data preprocessing

Data Cleaning

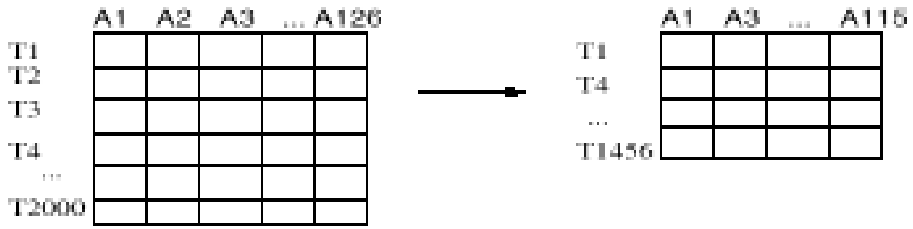


Data Integration



Data Transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



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

Data Integration

- Data can be in DBMS
  - ODBC, JDBC protocols
- Data in a flat file
  - Fixed-column format
  - Delimited format: tab, comma “,” , other
  - E.g. C4.5 and Weka “arff” use comma-delimited data
  - Attention: Convert field delimiters inside strings
- Verify the number of fields before and after

Data Integration and Cleaning: Metadata

- **Field types:**
  - **binary, nominal (categorical), ordinal, numeric, …**
  - **For nominal fields: tables translating codes to full descriptions**
- **Field role:**
  - input : inputs (condition attributes) for modeling
  - target : output
  - id/auxiliary : keep, but not use for discovering
  - ignore : don’t use for discovering
  - weight : instance weight
  - …
- **Field descriptions**

Data Integration from different sources

- Schema integration
  - integrate metadata from different sources
- Entity identification problem
  - to identify real world entities from multiple data sources
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different
  - possible reasons: different representation, different scale

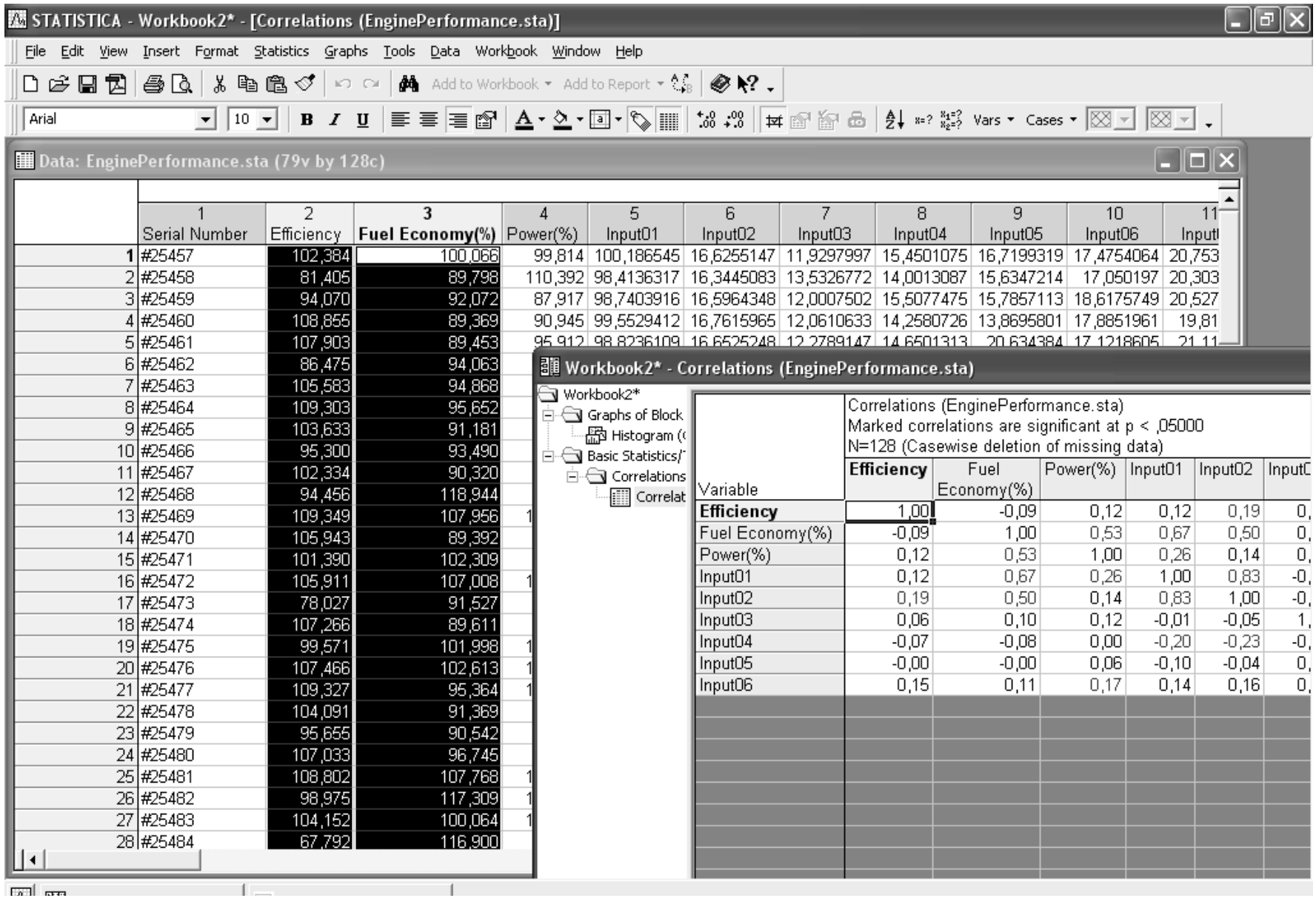
Data Cleaning: Unified Date Format

- We want to transform all dates to the same format internally
- Some systems accept dates in many formats
  - e.g. “Sep 24, 2003” , 9/24/03, 24.09.03, etc
  - dates are transformed internally to a standard value
- Frequently, just the year (YYYY) is sufficient
- For more details, we may need the month, the day, the hour, etc
- Representing date as YYYYMM or YYYYMMDD can be OK, but has problems
- ***Q: What are the problems with YYYYMMDD dates?***
  - A: Ignoring for now the Looming Y10K (year 10,000 crisis …)
  - YYYYMMDD does not preserve intervals:
  - 20040201 - 20040131 /= 20040131 – 20040130
  - This can introduce bias into models

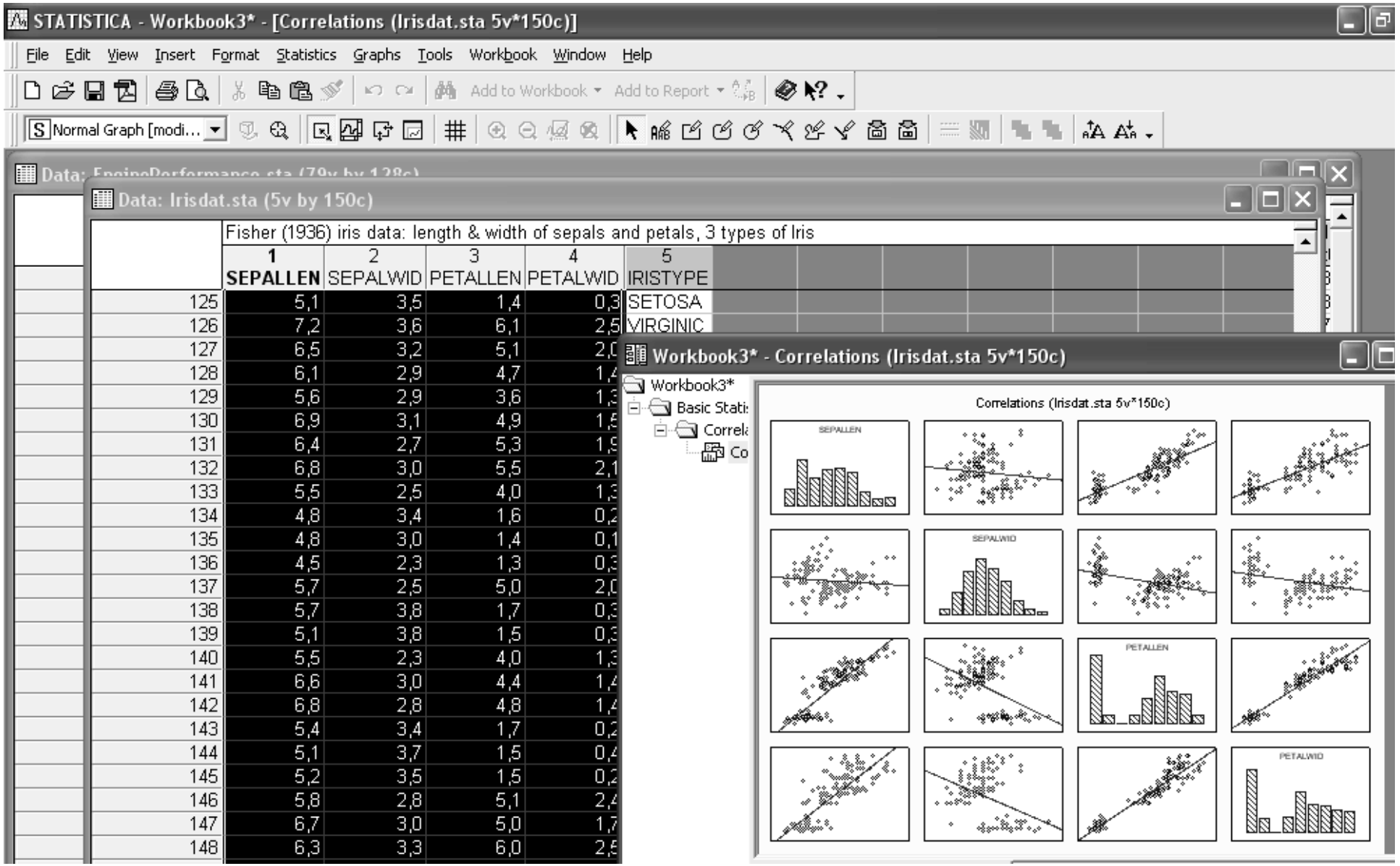
Redundant Data

- 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



KDnuggets : **Software** : **Data Transformation and Cleaning**

Data Transformation and Cleaning Software

- Ab Initio, provides high-performance software library and graphical environment for data transformation
- AMADEA, data Extraction, Transformation, and Real Time Reporting software
- BioComp iManageData(tm), Accesses, cleans, filters, converts and transforms data from files, Excel, Oracle, SQL Server, process control systems and more.
- ChoiceMaker 2.2 data quality and database record matching, merging, & deduplication software based on patented AI and machine learning techniques.
- COMGEN - Disk, tape and data conversion and data recovery experts, Commercial and General Systems.
- Data Manager, windows GUI application for data transformation and cleansing before data mining.
- DataFlux, provides Data Management solutions including Data profiling, Data quality, Data integration and Data augmentation
- Datatect, a powerful program for generating realistic test data to ASCII flat files or directly to RDBMS including Oracle, Sybase, SQL Server, and Informix.
- Dataskope, department-level tools to map, transform, alarm, output and view high volumes of binary or ASCII input data.
- DQ Now, profiling, cleansing, and dedup tools, providing a clear view of the data
- DQ Global, data cleansing, data management software, including de-duplication, merge/purge, address correction and suppression.
- GritBot, for identifying anomalies in data (compatible with See5 and Cubist).
- Hummingbird ETL, powerful data integration solution.
- IBM Datajoiner, allows you to view IBM, multi-vendor, relational, nonrelational, local, remote, and now geographic data as local and access and join tables without knowing the source location.
- MiningMart platform, for the preparation of relational data for Knowledge Discovery, free for research and non-commercial applications.
- NewView from SPSS
- proMISS, imputes missing values in databases.
- Relational Tools streamline application testing by allowing moving, editing and comparing referentially intact sets of complex relational data.
- Sagent, provides a suite of data transformation and loading tools
- Syncsort, fast high-volume sorting, filtering, reformatting, aggregating, and more
- The TrueData COMponent, functions to programmatically standardise your data, process it phonetically, and output a match key.
- WinPure, powerful data cleaning software, including duplication removal, email suggestions, statistics and more.

KDnuggets recomendations for data transformation and cleaning software

More at

- http://www.kdnuggets.com/software/index.html

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.

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**
- Some programs can infer missing values.

Hospital Check-in Database

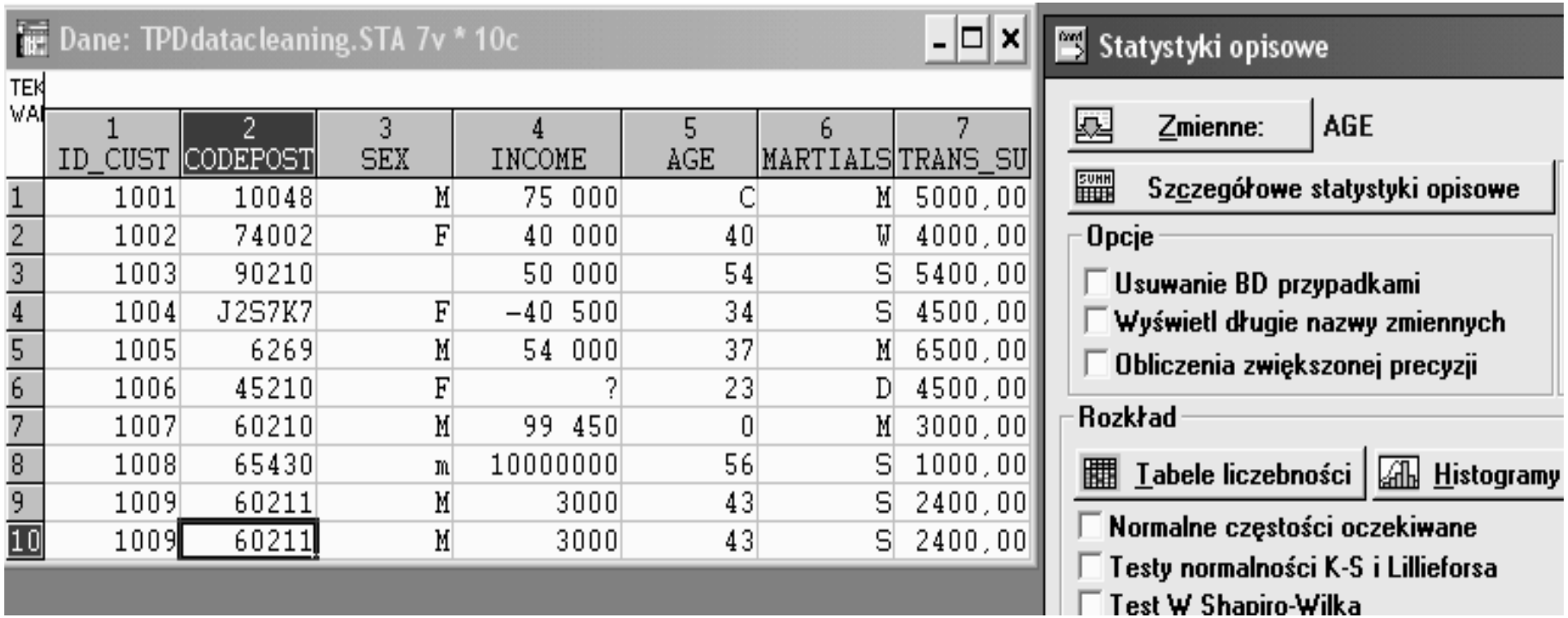
Name	Age	Sex	Pregnant	..
Mary	25	F	N	
Jane	27	F	?	
Joe	30	M	-	
Anna	2	F	?	

Handle Missing Values – Fill in (Substitute)

- 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?!
  - 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
  - the subset of possible values
  - result of global closest fit (distance base approaches)

Erroneous / Incorrect values

- What suspicious can you see in this table?



Dane: TPDdatacleaning.STA 7v * 10c

	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

Statystyki opisowe

Zmienne: AGE

Szczegółowe statystyki opisowe

Opcje

Usuwanie BD przypadkami

Wyświetl długie nazwy zmiennych

Obliczenia zwiększonej precyzji

Rozkład

Tabele liczebności

Histogramy

Normalne częstości oczekiwane

Testy normalności K-S i Lillieforsa

Test W Shapiro-Wilka

Inaccurate values

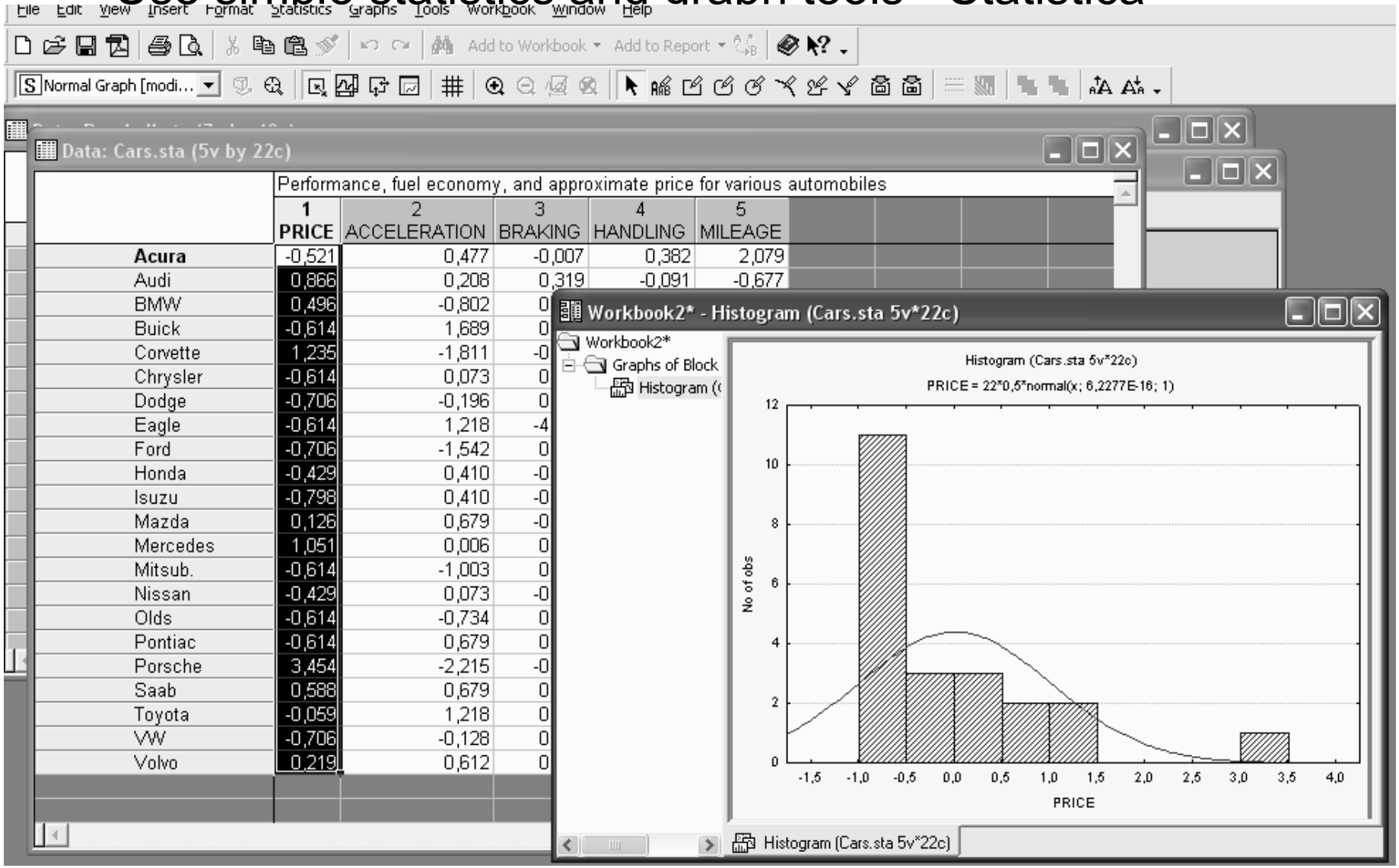
- Reason: data has not been collected for mining it
- Result: errors and omissions that don't affect original purpose of data (e.g. age of customer)
- Typographical errors in nominal attributes
  $\Rightarrow$ values need to be checked for consistency
- Typographical and measurement errors in numeric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

Noise and Incorrect Data

- Data could be noisy: Incorrect attribute values
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

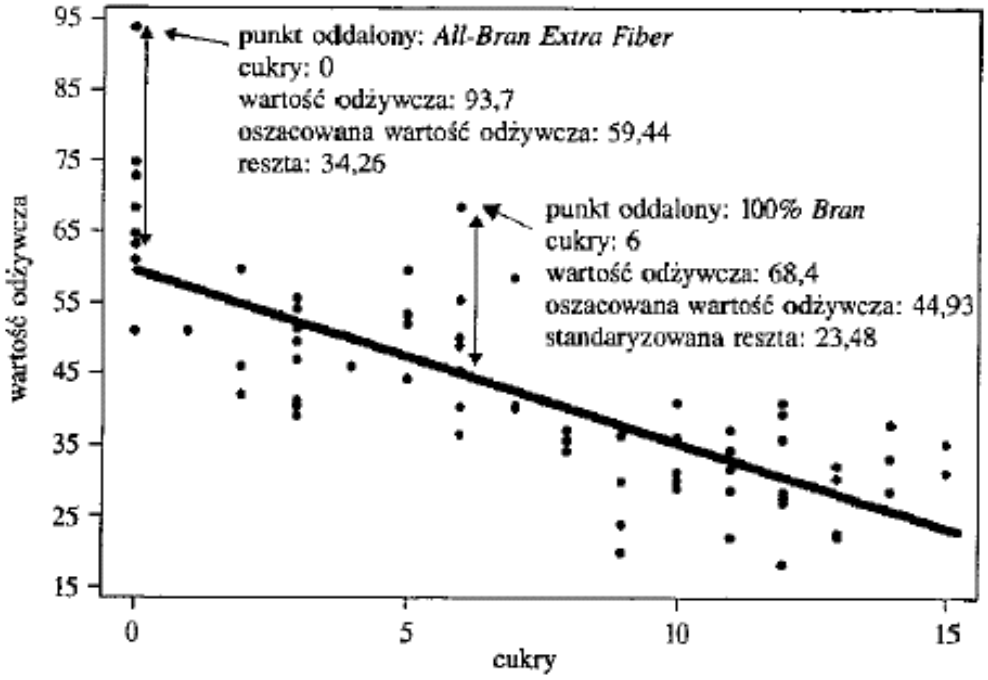
Outliers – graphical identification

- Use simple statistics and graph tools - Statistica



Regression - outliers

- An example of „corn flakes” [Larose 08] – 2 points could be outliers for a linear regression → standardized residuals

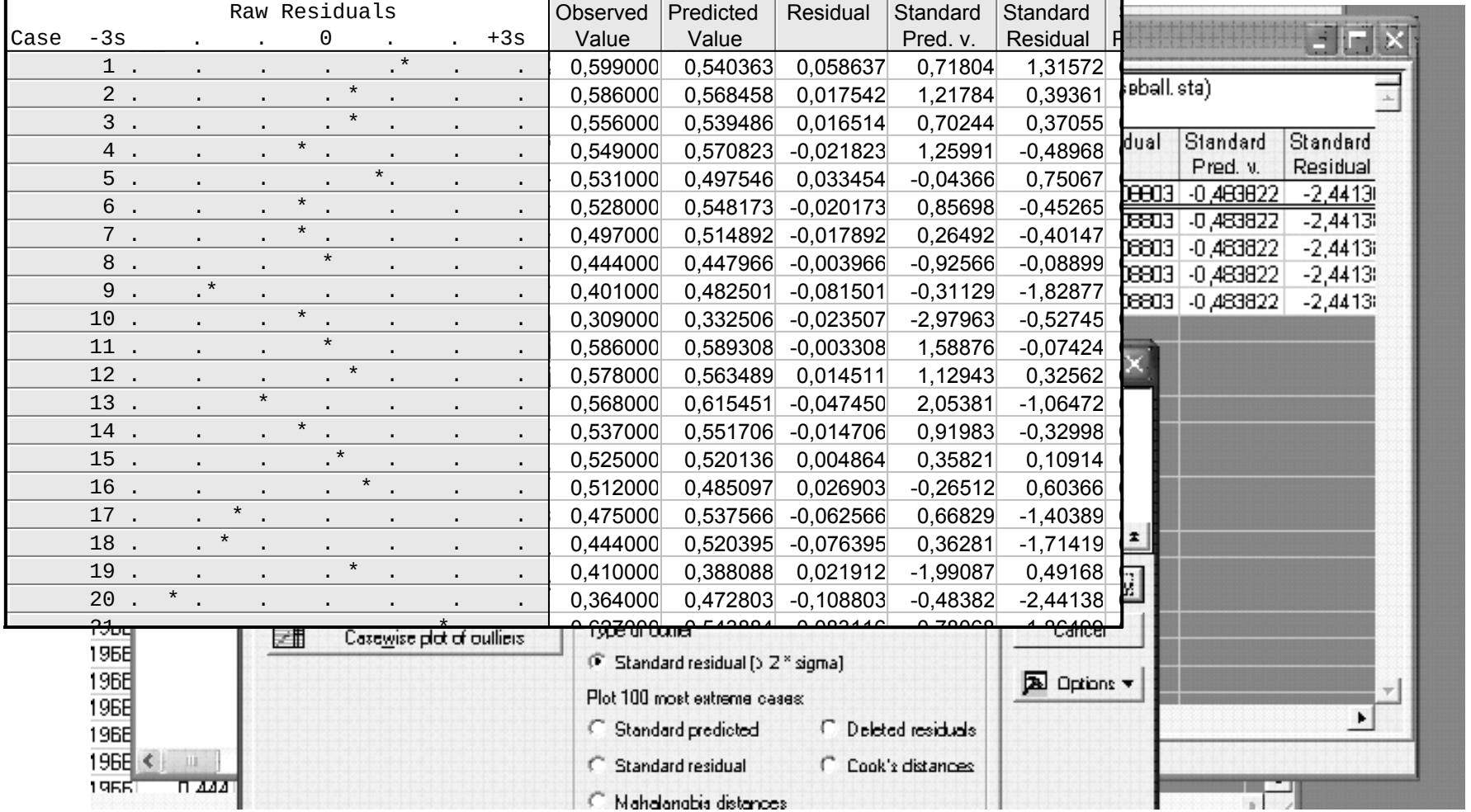


Rys. 2.3. Identyfikacja punktów oddalonych dla regresji zmiennej *wartość odżywcza* względem zmiennej *cukry*

Analysis of residulas

Raw Residual (Baseball.sta)
Dependent variable: WIN

Case	Raw Residuals -3s . . 0 . . +3s	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual
1	. . . . .* . .	0,599000	0,540363	0,058637	0,71804	1,31572
2	. . . . * . . .	0,586000	0,568458	0,017542	1,21784	0,39361
3	. . . . * . . .	0,556000	0,539486	0,016514	0,70244	0,37055
4	. . . *	0,549000	0,570823	-0,021823	1,25991	-0,48968
5	. . . . *. . .	0,531000	0,497546	0,033454	-0,04366	0,75067
6	. . . *	0,528000	0,548173	-0,020173	0,85698	-0,45265
7	. . . *	0,497000	0,514892	-0,017892	0,26492	-0,40147
8	. . . *	0,444000	0,447966	-0,003966	-0,92566	-0,08899
9	. .*	0,401000	0,482501	-0,081501	-0,31129	-1,82877
10	. . . *	0,309000	0,332506	-0,023507	-2,97963	-0,52745
11	. . . *	0,586000	0,589308	-0,003308	1,58876	-0,07424
12	. . . . * . . .	0,578000	0,563489	0,014511	1,12943	0,32562
13	. . *	0,568000	0,615451	-0,047450	2,05381	-1,06472
14	. . . *	0,537000	0,551706	-0,014706	0,91983	-0,32998
15	. . . .* . . .	0,525000	0,520136	0,004864	0,35821	0,10914
16	. . . . * . . .	0,512000	0,485097	0,026903	-0,26512	0,60366
17	. . *	0,475000	0,537566	-0,062566	0,66829	-1,40389
18	. . *	0,444000	0,520395	-0,076395	0,36281	-1,71419
19	. . . . * . . .	0,410000	0,388088	0,021912	-1,99087	0,49168
20	. *	0,364000	0,472803	-0,108803	-0,48382	-2,44138
21	. . . . . *	0,627000	0,543884	0,083116	0,78068	1,86409



Getting to know the data

- Simple visualization tools are very useful
  - Nominal attributes: histograms (Distribution consistent with background knowledge?)
  - Numeric attributes: graphs (Any obvious outliers?)
- 2-D and 3-D plots show dependencies
- Need to consult domain experts
- Too much data to inspect? Take a sample!

Nice examples of using simple graph tools

- Prof. M.Lasek „Data mining” – banking data [in Polish]
- Larose D. Odkrywanie wiedzy z danych, PWN. (Original version in English)
- Stanisz. Przystępny kurs statystyki (3 tomy), Statsoft [in Polish]
- U. Fayyad, G.Gristen, A.Wierse, Information Visulization in Data Mining and Knowledge Discovery, Morgan Kaufmann Publisher.

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- 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

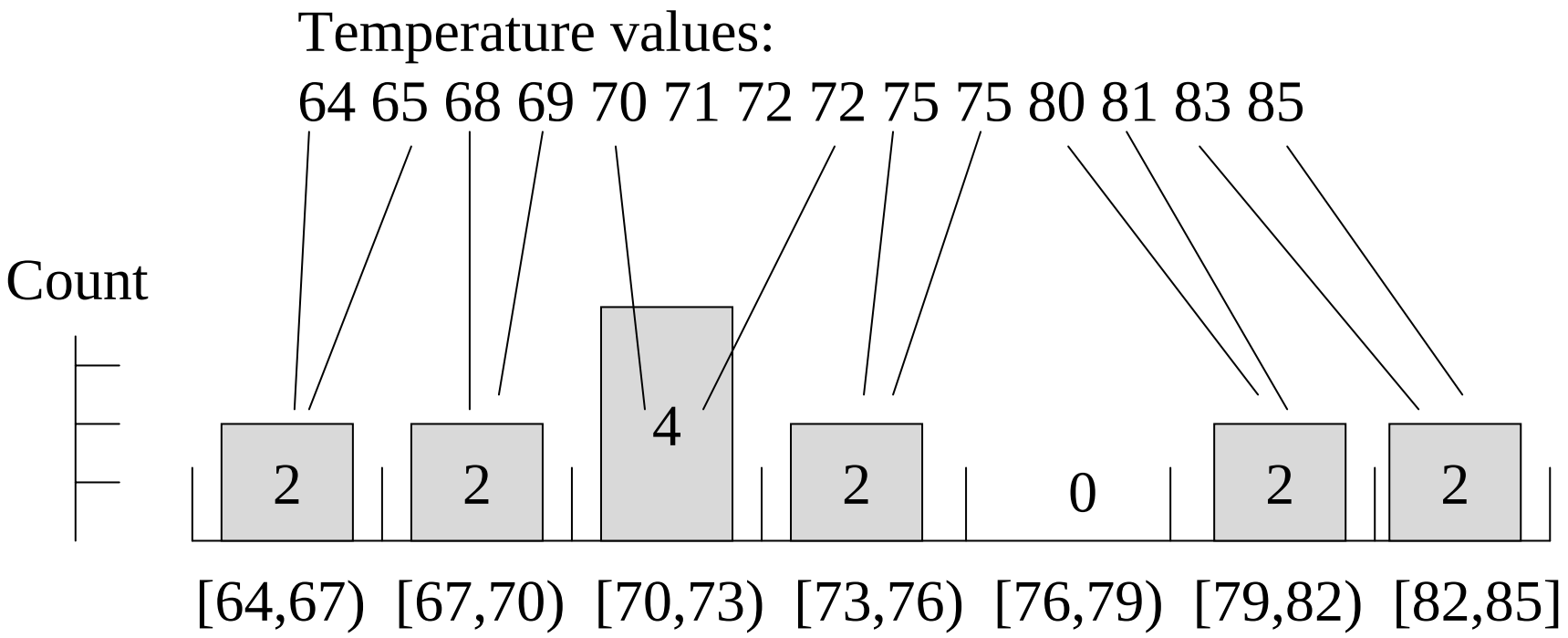
$$v' = \frac{v}{10^j}$$

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

Discretization

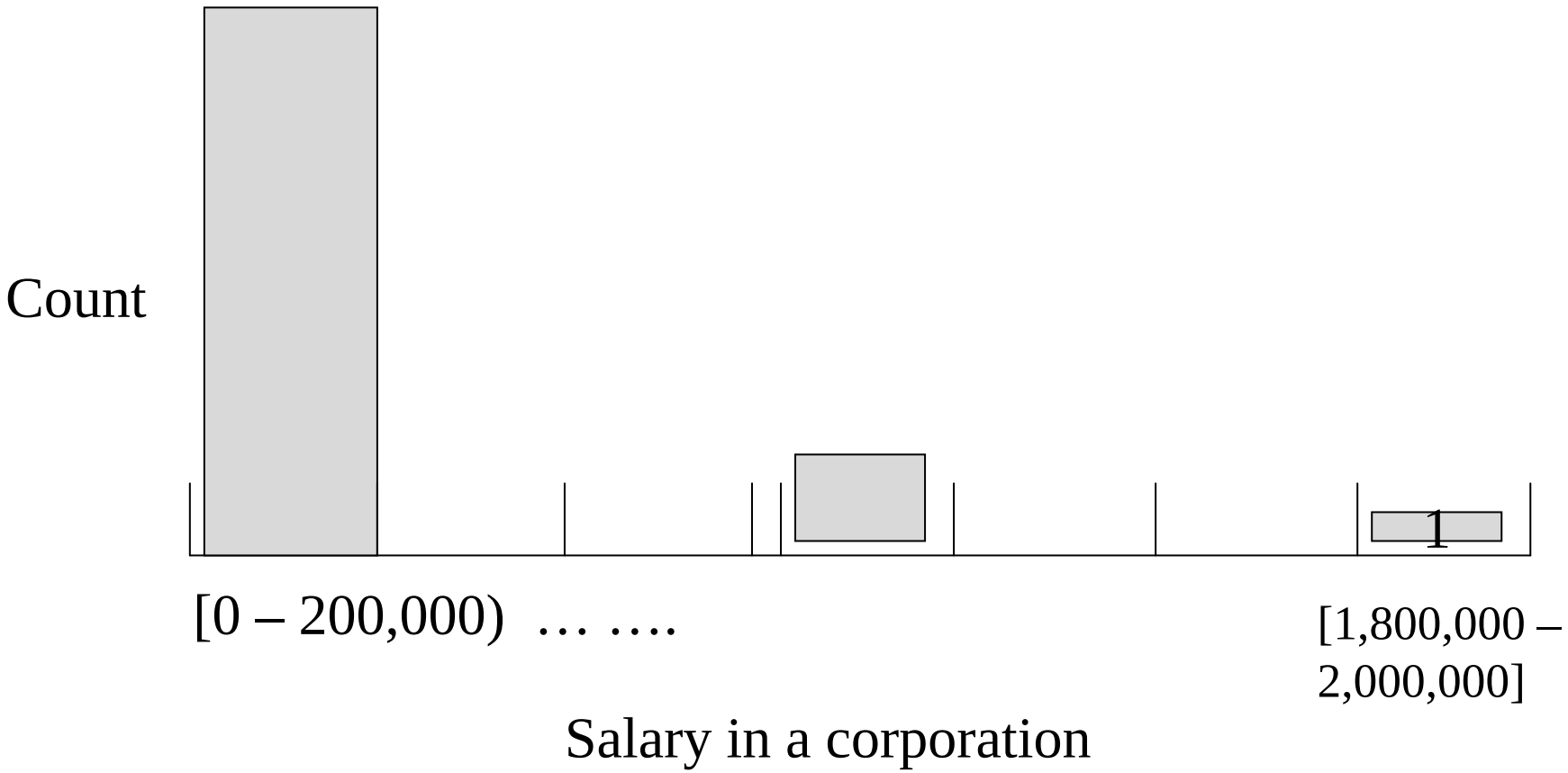
- Some methods require discrete values, e.g. most versions of Naïve Bayes, CHAID
- Discretization $\rightarrow$ transformation of numerical values into codes / values of ordered subintervals defined over the domain of an attribute.
- Discretization is very useful for generating a summary of data
- Many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

Discretization: Equal-Width (Length)



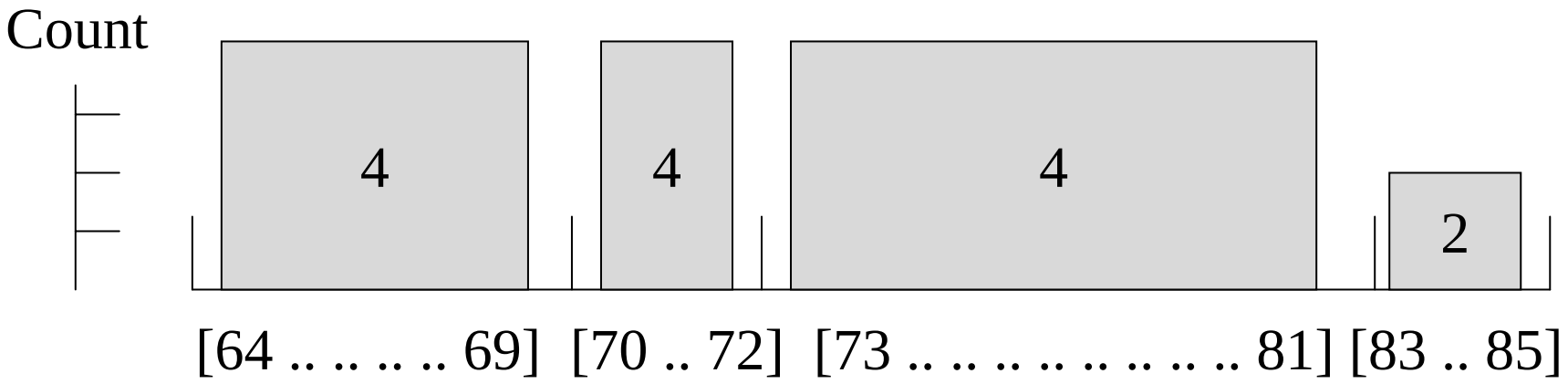
Equal Width, bins Low <= value < High

Discretization: Equal-Width may produce clumping



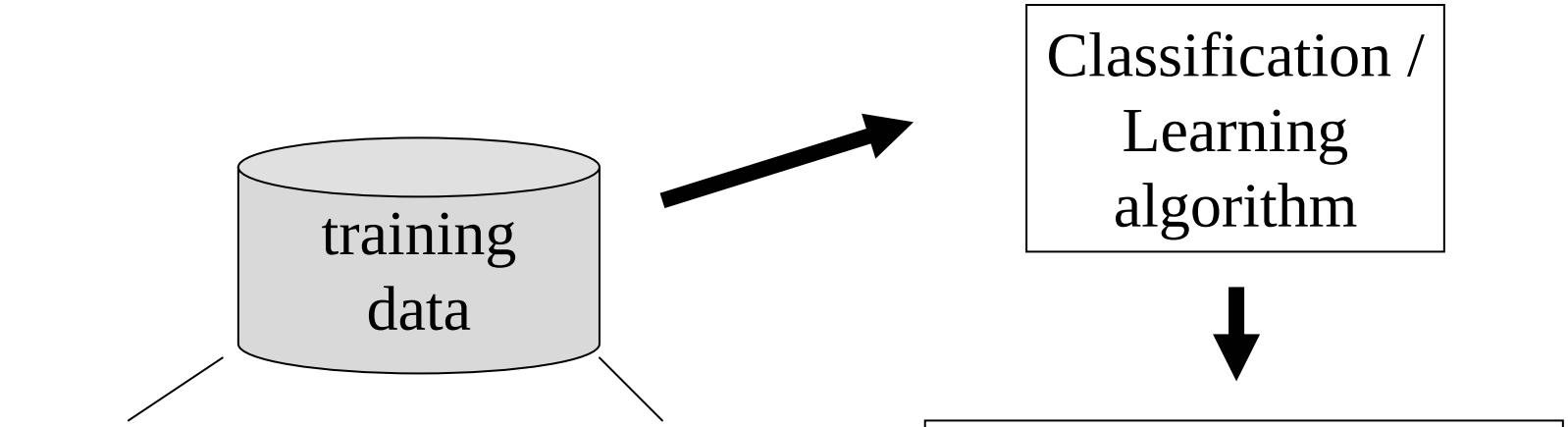
Discretization: Equal-Frequency

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



Equal Height = 4, except for the last bin

Input data for classification problems

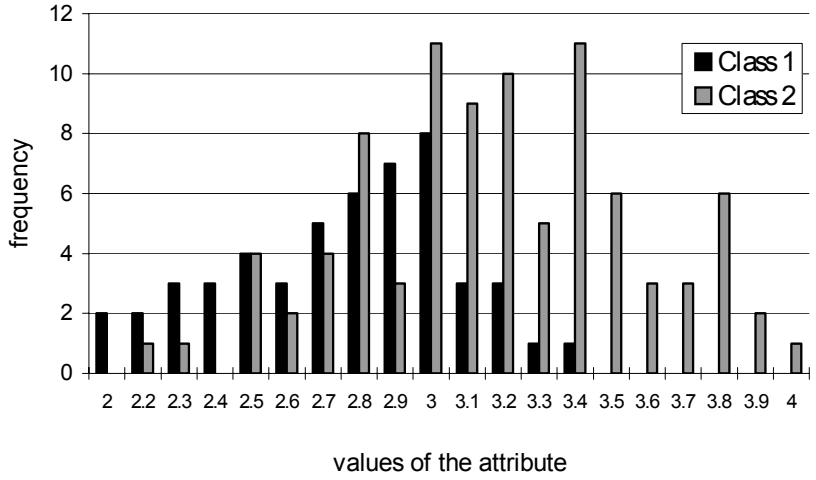


Age	Car Type	...	Risk
20	Combi	...	High
18	Sports	...	High
40	Sports	...	High
50	Family	...	Low
35	Minivan	...	Low
30	Combi	...	High
32	Family	...	Low

Classifier e.g. rules
If (Car = Minivan or Family) then (Risk=Low)
If (Age $\in$[16,30]) then (Risk=High)

Supervised (class) discretization

- Use information about attribute value distribution + class assignment.



- Minimal entropy based approaches; Chi-Merge, others
  - WEKA uses a version of class entropy

Entropy-Based Discretization

- Given a set of samples S, if S is partitioned into two sub-intervals S1 and S2 using boundary T, the entropy after partitioning is

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

- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization.
- The process is recursively applied to partitions obtained until some stopping criterion is met, e.g.,

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

- Experiments show that it may reduce data size and improve classification accuracy

An example of discretization [Grzymala 97]

Attributes			Decision
Caliber	Length	Weight	Recoil
5.56	45	55	light
6.5	55	120	light
6.5	55	142	medium
7	57	100	medium
7.5	55	150	medium
7.62	39	123	light
7.62	63	150	heavy
7.62	63	168	heavy
8	57	198	heavy

If (Weight=55) then (Decision=light)
If (Weight=120) then (Decision=light)
If (Weight=100) then (Decision=medium)
If (Weight=142) then (Decision=medium)

…

If (Length=63) then (Decision=heavy)

Using discretization before rule induction

Attributes			Decision
Caliber	Length	Weight	Recoil
5.56..7.62	39..57	55..142	light
5.56..7.62	39..57	55..142	light
5.56..7.62	39..57	142..198	medium
5.56..7.62	57..63	55..142	medium
5.56..7.62	57..63	142..198	medium
7.62..8	39..57	55..142	light
7.62..8	57..63	142..198	heavy
7.62..8	57..63	142..198	heavy
7.62..8	57..63	142..198	heavy

If (length,39..57) & (weight,55..142) then (recoil,light)

If (caliber,5.56..7.62) & (weight,142..198) then (recoil,medium)

If (weight,55..142) & (length,57..63) then (recoil,medium)

If (caliber,7.62..8) & (length,57..63) then (recoil,heavy)

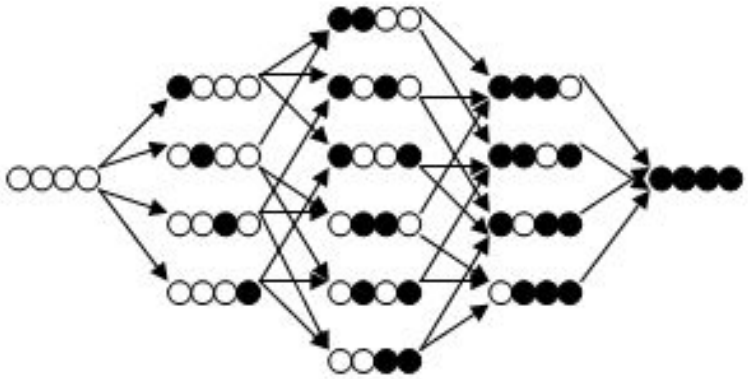
Data Preprocessing: Attribute Selection

First: Remove fields with no or little variability

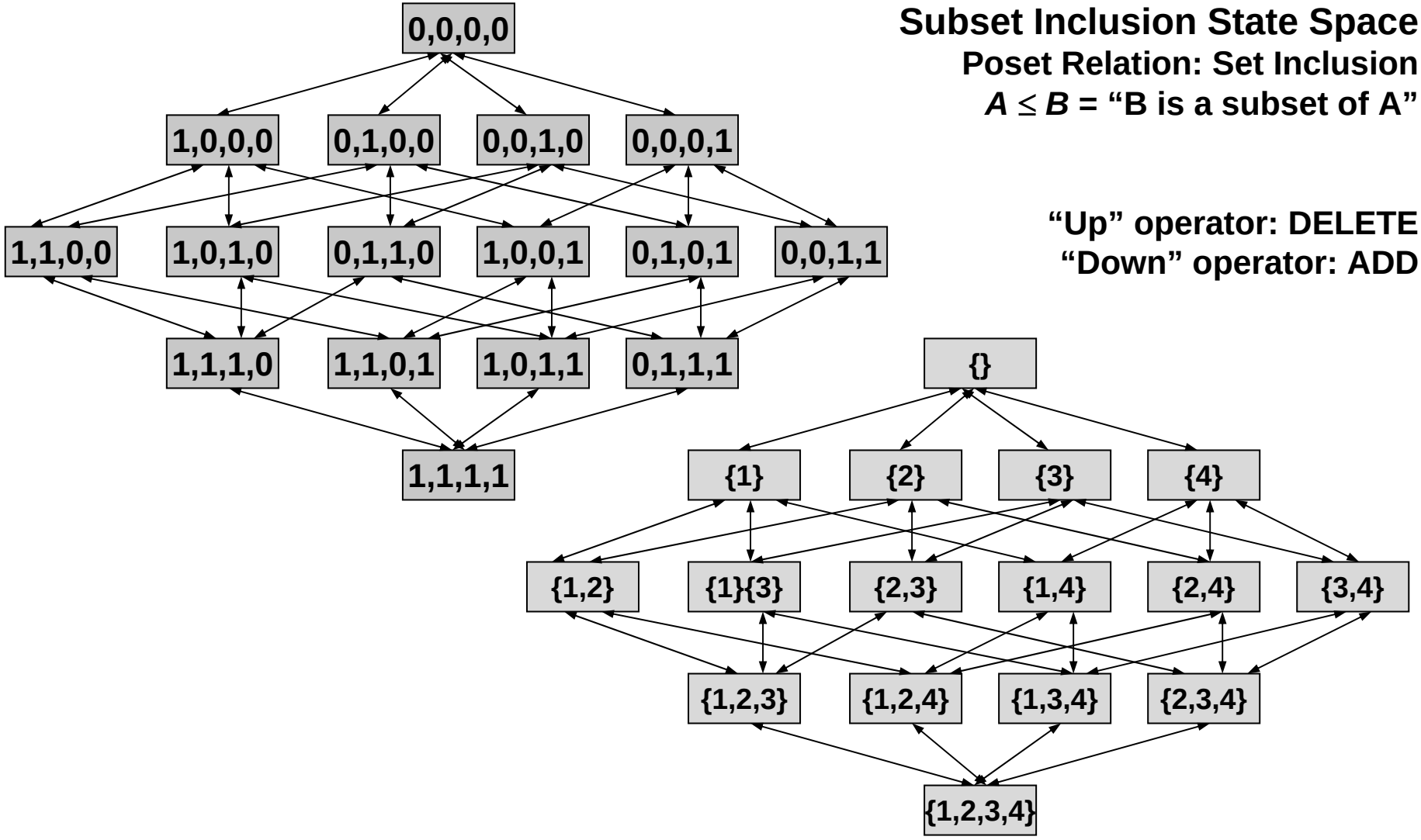
- Examine the number of distinct field values
  - *Rule of thumb: remove a field where almost all values are the same (e.g. null), except possibly in minp % or less of all records.*
  - *minp* could be 0.5% or more generally less than 5% of the number of targets of the smallest class
- More sophisticated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

A few remarks on selecting attributes

- Irrelevant attributes (features) in the input data may decrease the classification performance (supervised approaches)
- Attribute (feature) selection:
  - Find the smallest subset of attributes leading to a higher classification accuracy than all attributes
- Search problem in the space of attribute subsets
- Three components:
  - Search algorithm
  - Evaluation function
  - Classifier



Relevance Feature Determination



Subset Inclusion State Space

Poset Relation: Set Inclusion

A **≤** ***B*** **= "B is a subset of A"**

"Up" operator: DELETE

"Down" operator: ADD

Heuristic Feature Selection Methods

- There are *2^d* possible sub-features of *d* features
- Several heuristic feature selection methods:
  - Best single features under the feature independence assumption: choose by significance tests.
  - Best step-wise feature selection:
    - The best single-feature is picked first
    - Then next best feature condition to the first, ...
  - Step-wise feature elimination:
    - Repeatedly eliminate the worst feature
  - Best combined feature selection and elimination:
  - Optimal branch and bound:
    - Use feature elimination and backtracking

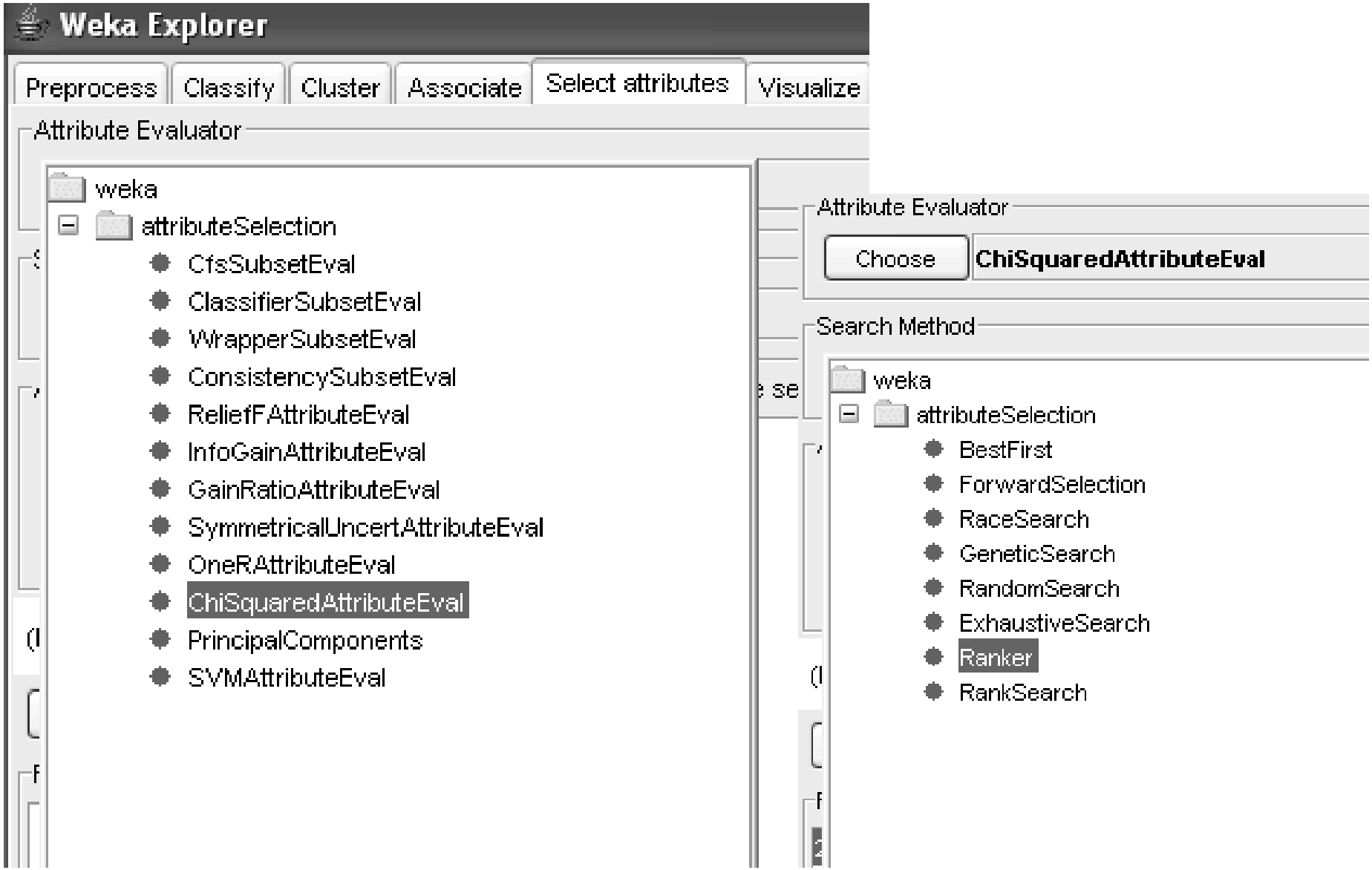
Different attribute selection methods

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

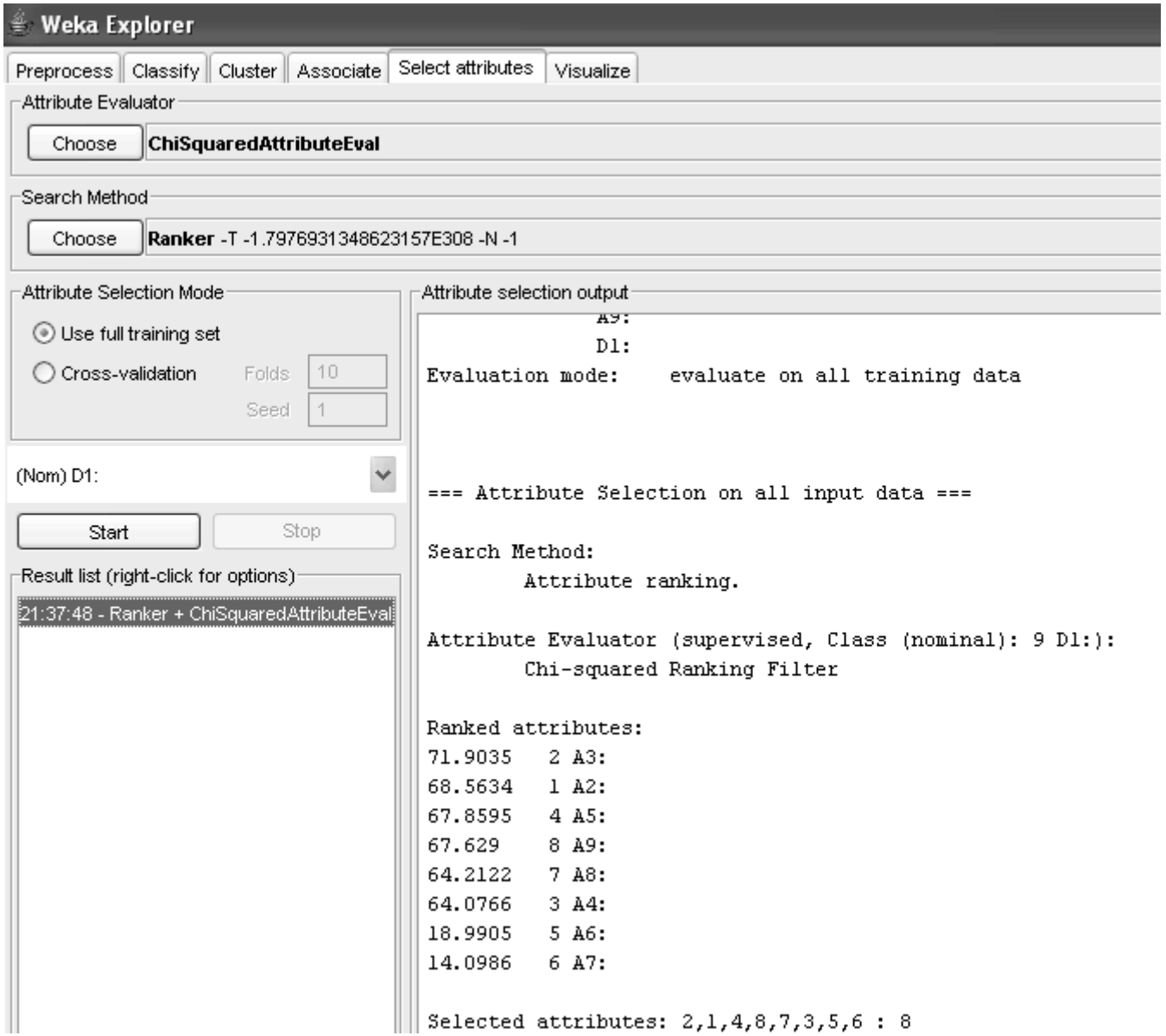
  and

  - Relief method
  - Wrapper model

WEKA – attribute selection tools

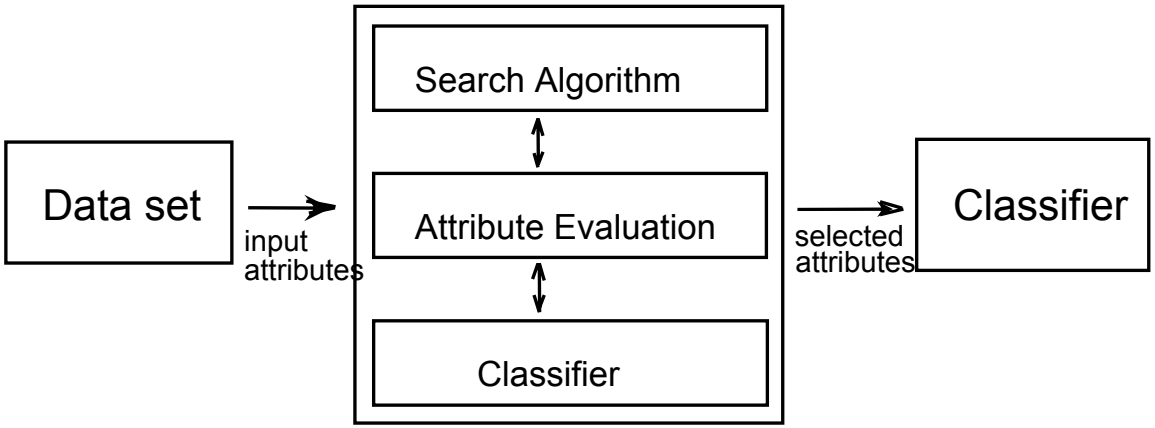


Ranking with …? WEKA



Wrapper approach

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



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

Conclusion

Good data preparation is key to producing valid and reliable models!

Any questions, remarks?