UNIVERSITÀ DI PISA

Dottorato di Ricerca in Ingegneria dell'Informazione

Data Mining and Soft Computing

Francisco Herrera

Research Group on Soft Computing and Information Intelligent Systems (SCI²S)

Dept. of Computer Science and A.I.

University of Granada, Spain

Email: herrera@decsai.ugr.es

http://sci2s.ugr.es

http://decsai.ugr.es/~herrera

DECSAI

Universidad de Granada

Summary

1. **Introduction to Data Mining and Knowledge Discovery**
2. **Data Preparation**
3. **Introduction to Prediction, Classification, Clustering and Association**
4. **Data Mining - From the Top 10 Algorithms to the New Challenges**
5. **Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation**
6. **Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning**
7. **Genetic Fuzzy Systems: State of the Art and New Trends**
8. **Some Advanced Topics I: Classification with Imbalanced Data Sets**
9. **Some Advanced Topics II: Subgroup Discovery**
10. **Some advanced Topics III: Data Complexity**
11. **Final talk: How must I Do my Experimental Study? Design of Experiments in Data Mining/Computational Intelligence. Using Non-parametric Tests. Some Cases of Study.**

Data Preparation

Outline

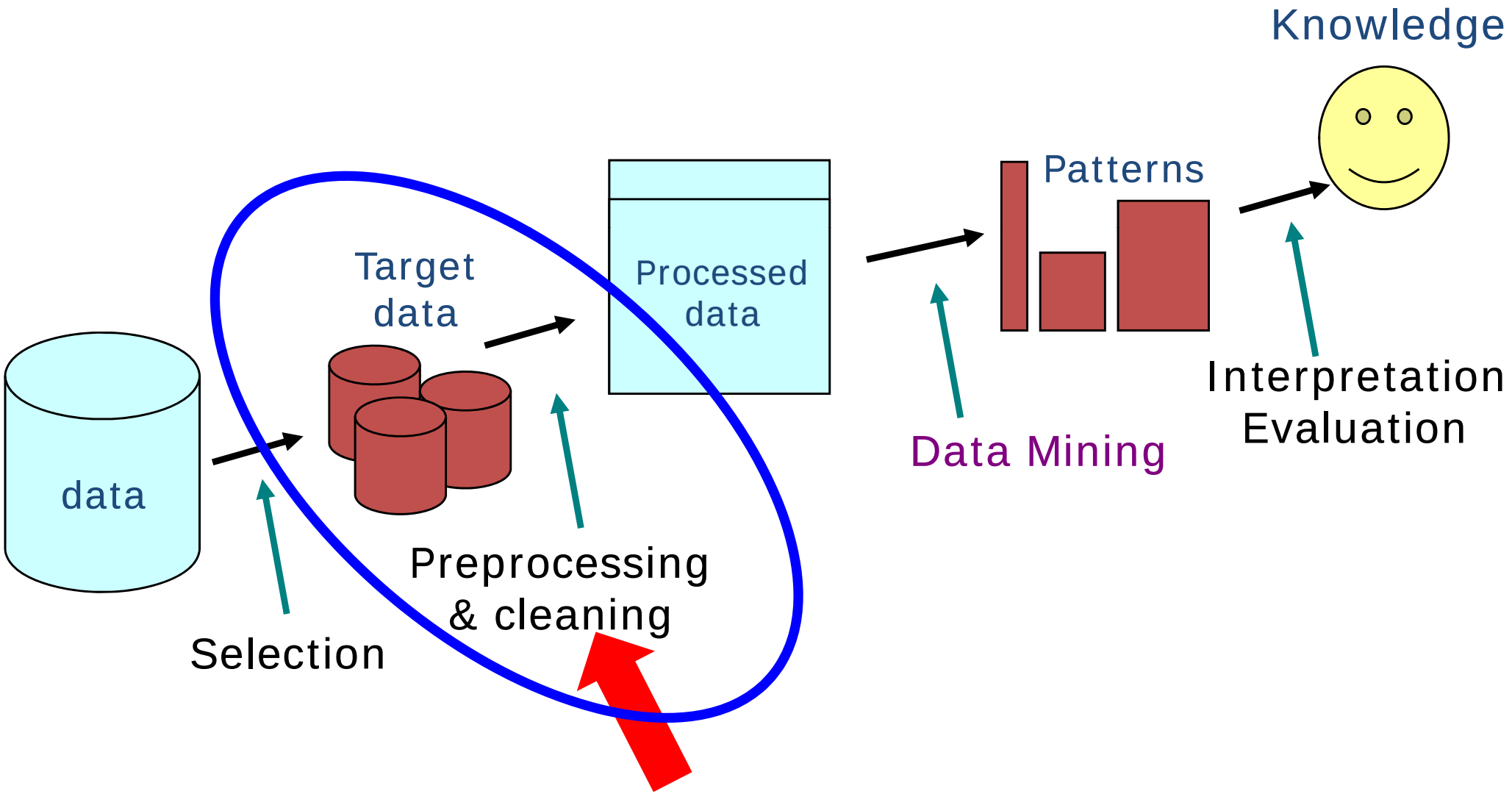
- ✓ Introduction
- ✓ Preprocessing
- ✓ Data Reduction
  Discretization, Feature Selection, Instance Selection
- ✓ Ex.: Instance Selection and Decision Trees
- ✓ Concluding Remarks

Data Preparation

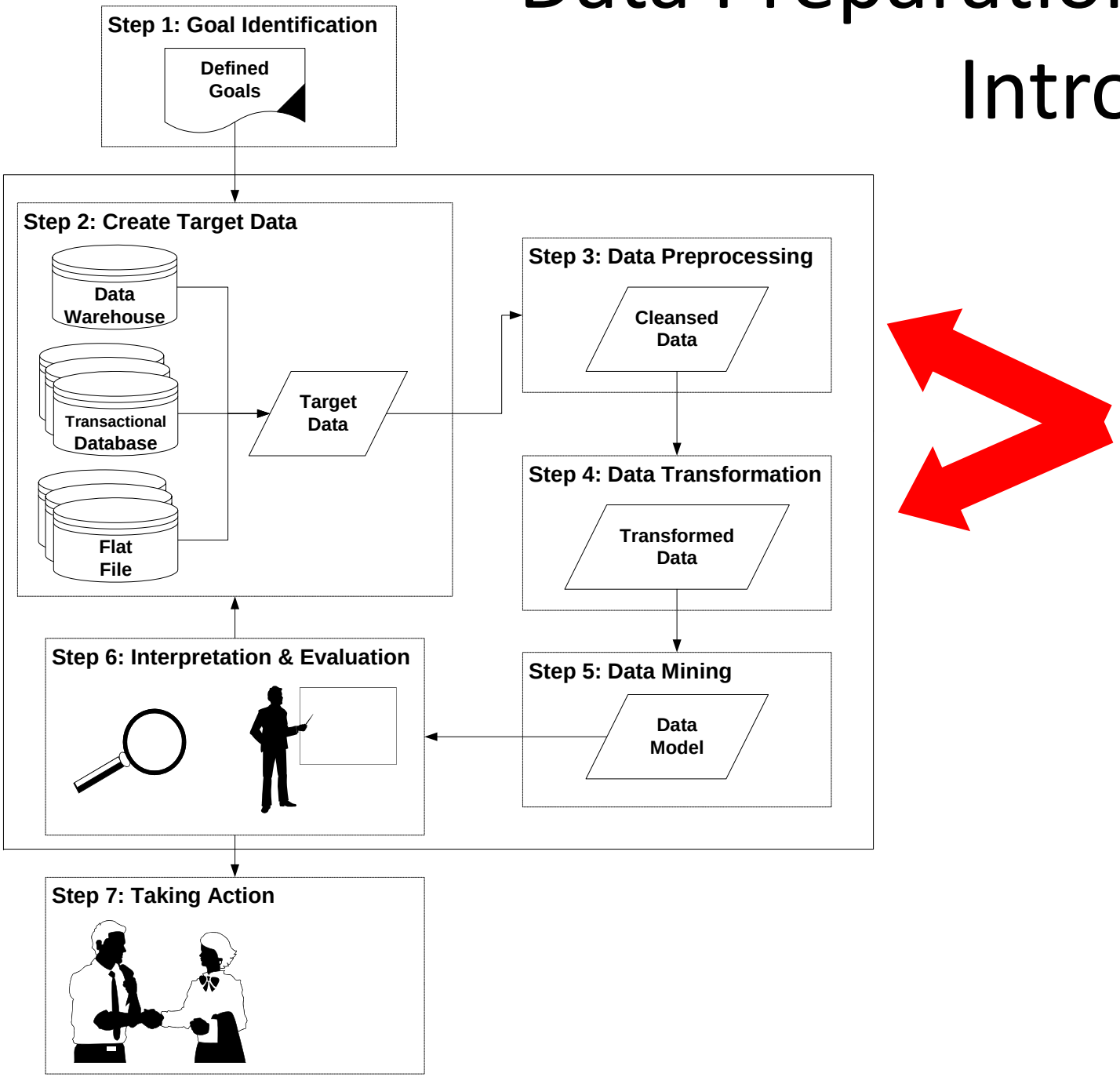
Outline

- ✓ Introduction
- ✓ Preprocessing
- ✓ Data Reduction
  Discretization, Feature Selection, Instance Selection
- ✓ Ex.: Instance Selection and Decision Trees
- ✓ Concluding Remarks

Data Preparation in KDD Introduction



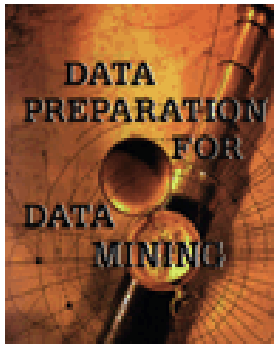
Data Preparation in KDD Introduction



Data Preparation in KDD
Introduction

D. Pyle, 1999, pp. 90:

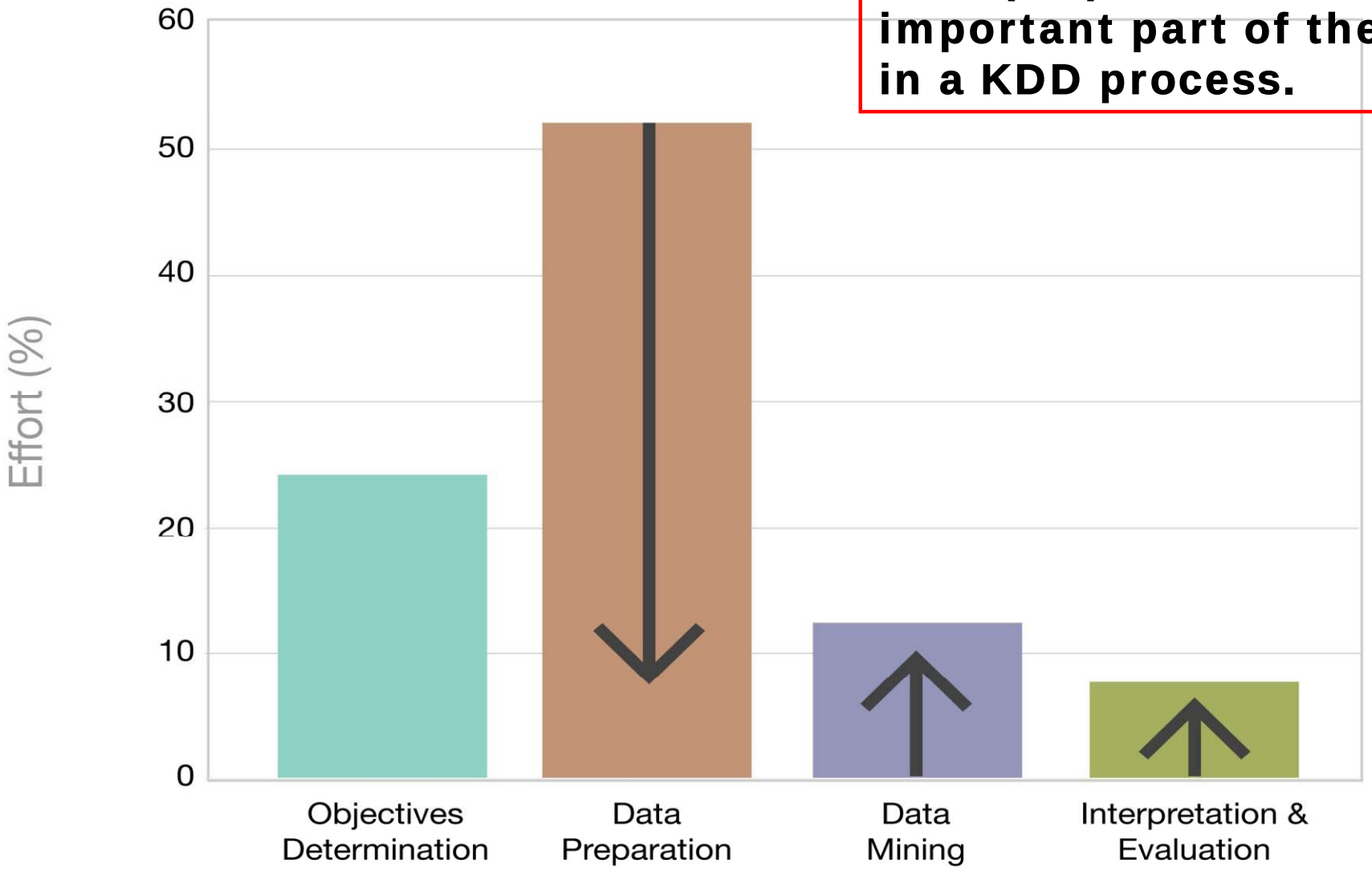
"The fundamental purpose of data preparation is to manipulate and transforrm raw data so that the information content enfolded in the data set can be exposed, or made more easily accesible."



Dorian Pyle
Data Preparation for Data Mining
Morgan Kaufmann Publishers, 1999

INTRODUCTION

Data preparation uses an important part of the time in a KDD process.



Data Preparation

Outline

- ✓ Introduction
- ✓ Preprocessing
- ✓ Data Reduction
  Discretization, Feature Selection, Instance Selection
- ✓ Ex.: Instance Selection and Decision Trees
- ✓ Concluding Remarks

Preprocessing

Why Data Preprocessing?

- **Data in the real world is dirty**
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- **No quality data, no quality mining results!**
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

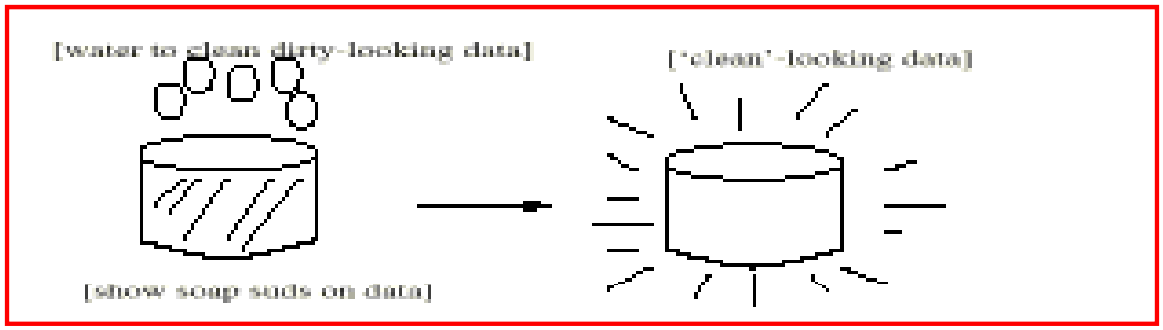
Preprocessing

Major Tasks in Data Preprocessing

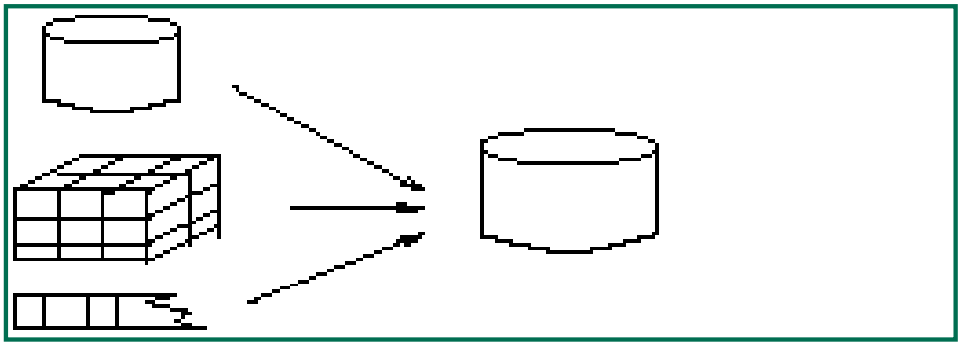
- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data transformation**
  - Normalization and aggregation
- **Data reduction**
  - Obtains reduced representation in volume but produces the same or similar analytical results
  - **Data discretization**
    - Part of data reduction but with particular importance, especially for numerical data

Preprocessing

Data Cleaning



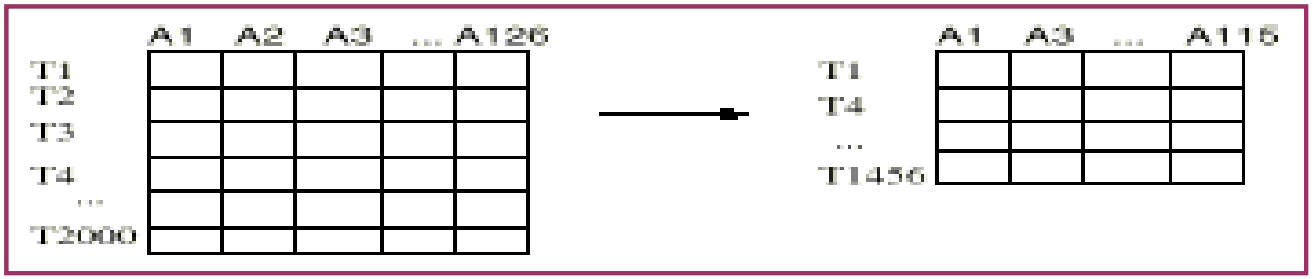
Data integration



Data transformation

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

Data reduction



Preprocessing

Data Cleaning

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data

Reference:
W. Kim, B. Choi, E-K. Hong, S-K. Kim. A Taxonomy of Dirty Data. Data Mining and Knowledge Discovery 7, 81-99, 2003.

R.K. Pearson. Mining Imperfect Data. Dealing with Contamination and Incomplete Records. SIAM, 2005.

Preprocessing

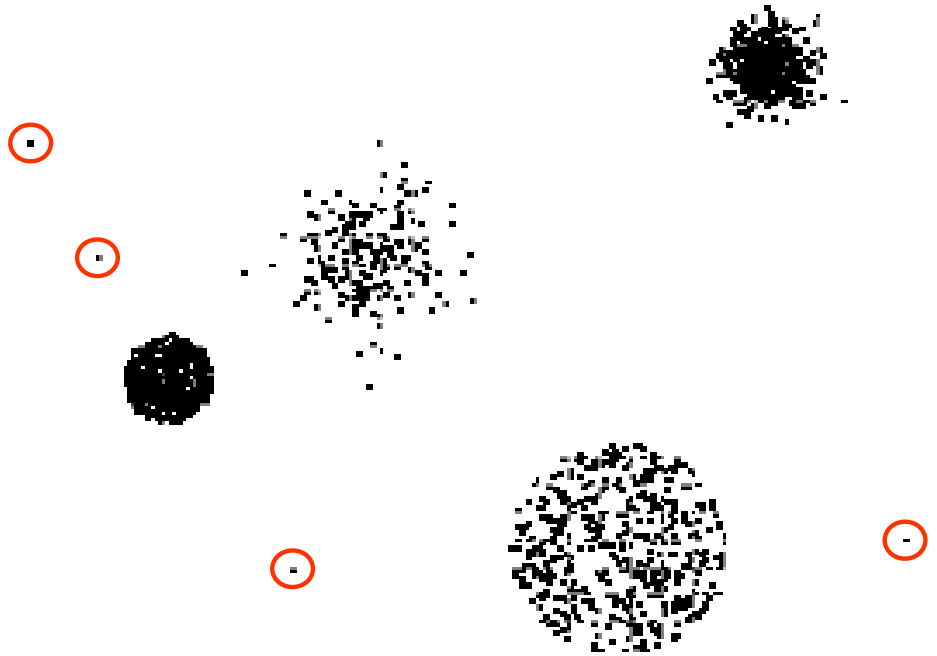
Data Cleaning: Missing values

- **Reasons for missing values**
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)

- **Handling missing values**
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

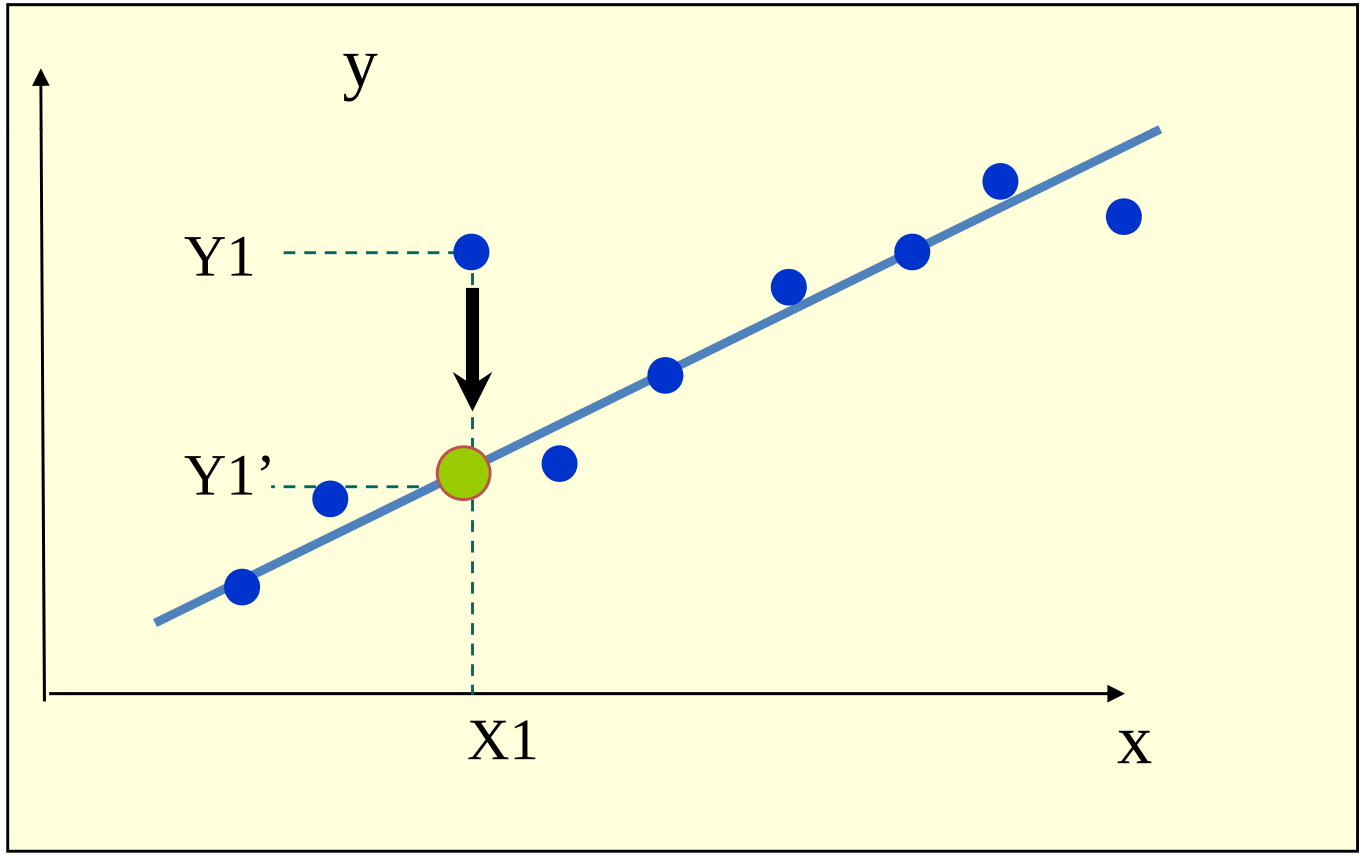
Preprocessing

Data Cleaning: Noise data. Outliers



Preprocessing

Data Cleaning: Noise data. Smooth by fitting the data into regression functions



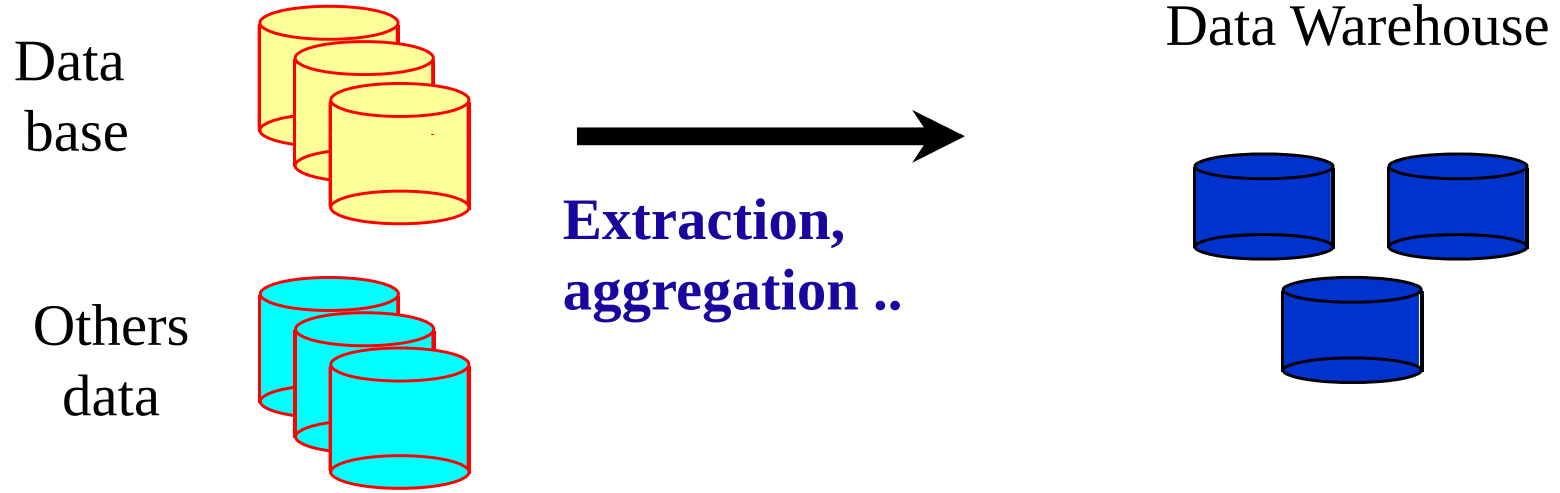
Preprocessing

Data Cleaning: Inconsistent data

Age="42"
Birth day="03/07/1997"

Preprocessing

Data Integration



Reference:
E.Schallehn, K. Sattler, G. Saake. Efficient Similarity-based Operations for Data Integration.
Data and Knowledge Engineering 48:3, 351-387, 2004.

Preprocessing

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

Reference:
T. Y. Lin. Attribute Transformation for Data Mining I: Theoretical Explorations.
International Journal of Intelligent Systems 17, 213-222, 2002.

Preprocessing

Data Transformation

- **Min-max normalizzation**

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

- **Z-score normalization**

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

- **normalization by decimal scaling**

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

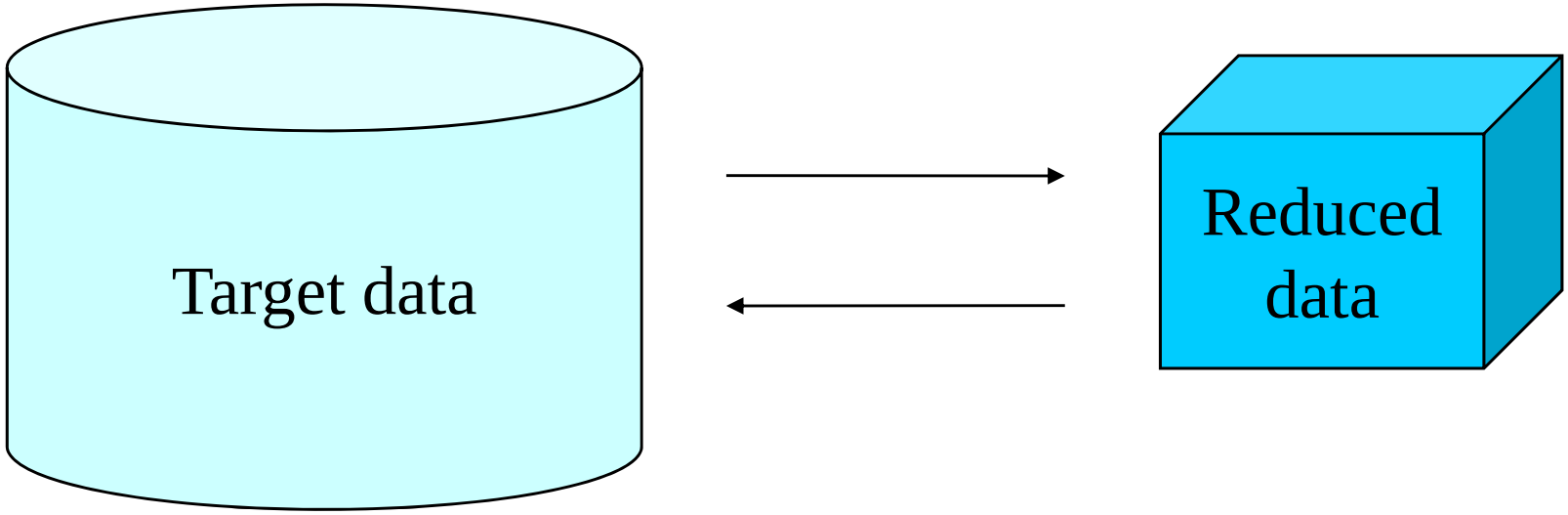
Where *j* is the smallest integer such that Max(| v'|)<1

Preprocessing

Data Reduction

Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set

Data reduction obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results



Preprocessing

Data Reduction strategies

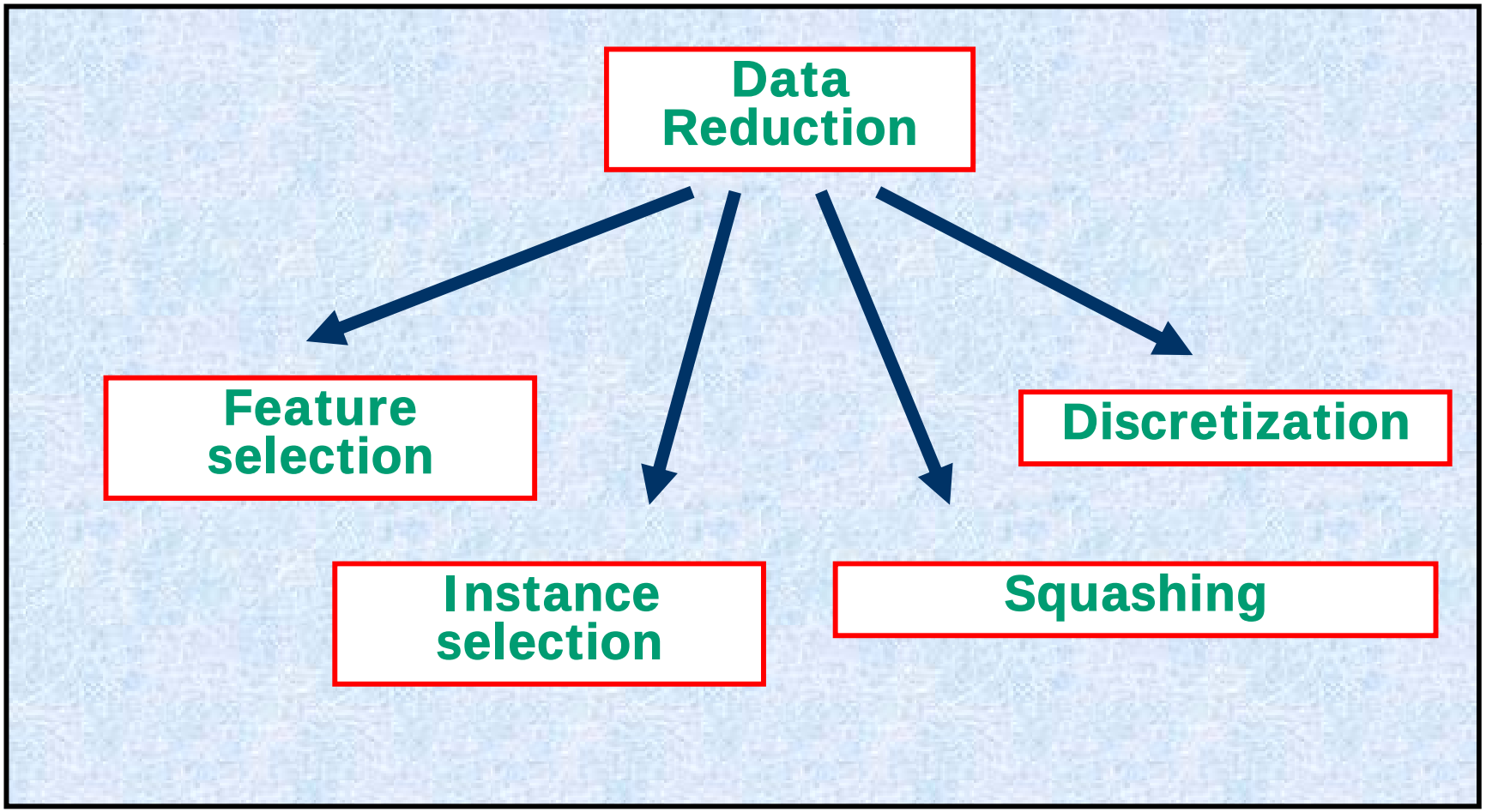
- Data cube aggregation
- Dimensionality reduction
- Numerosity reduction
- Discretization and concept hierarchy generation

Data Preparation

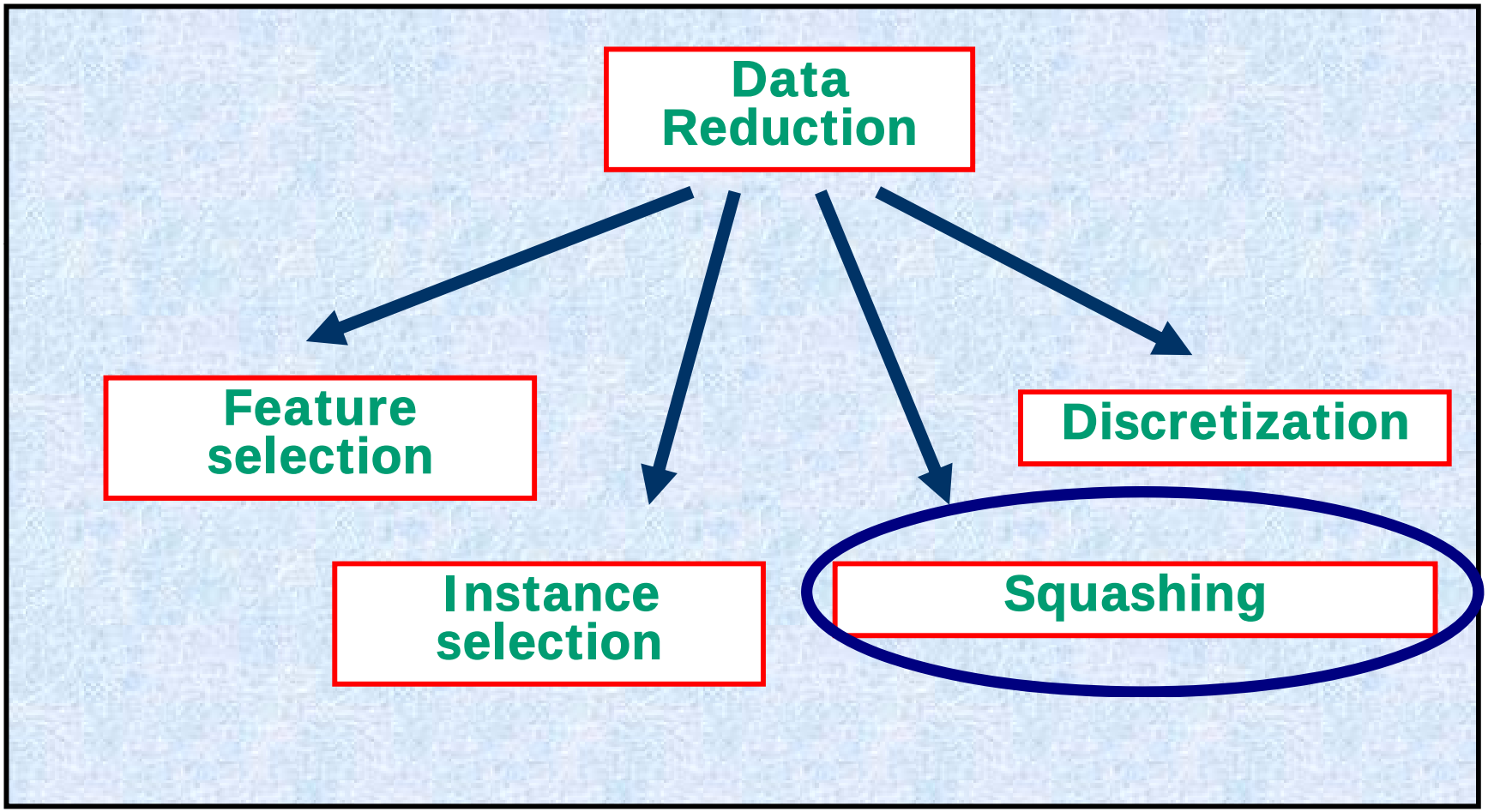
Outline

- ✓ Introduction
- ✓ Preprocessing
- ✓ Data Reduction
  Discretization, Feature Selection, Instance Selection
- ✓ Ex.: Instance Selection and Decision Trees
- ✓ Concluding Remarks

Data Reduction

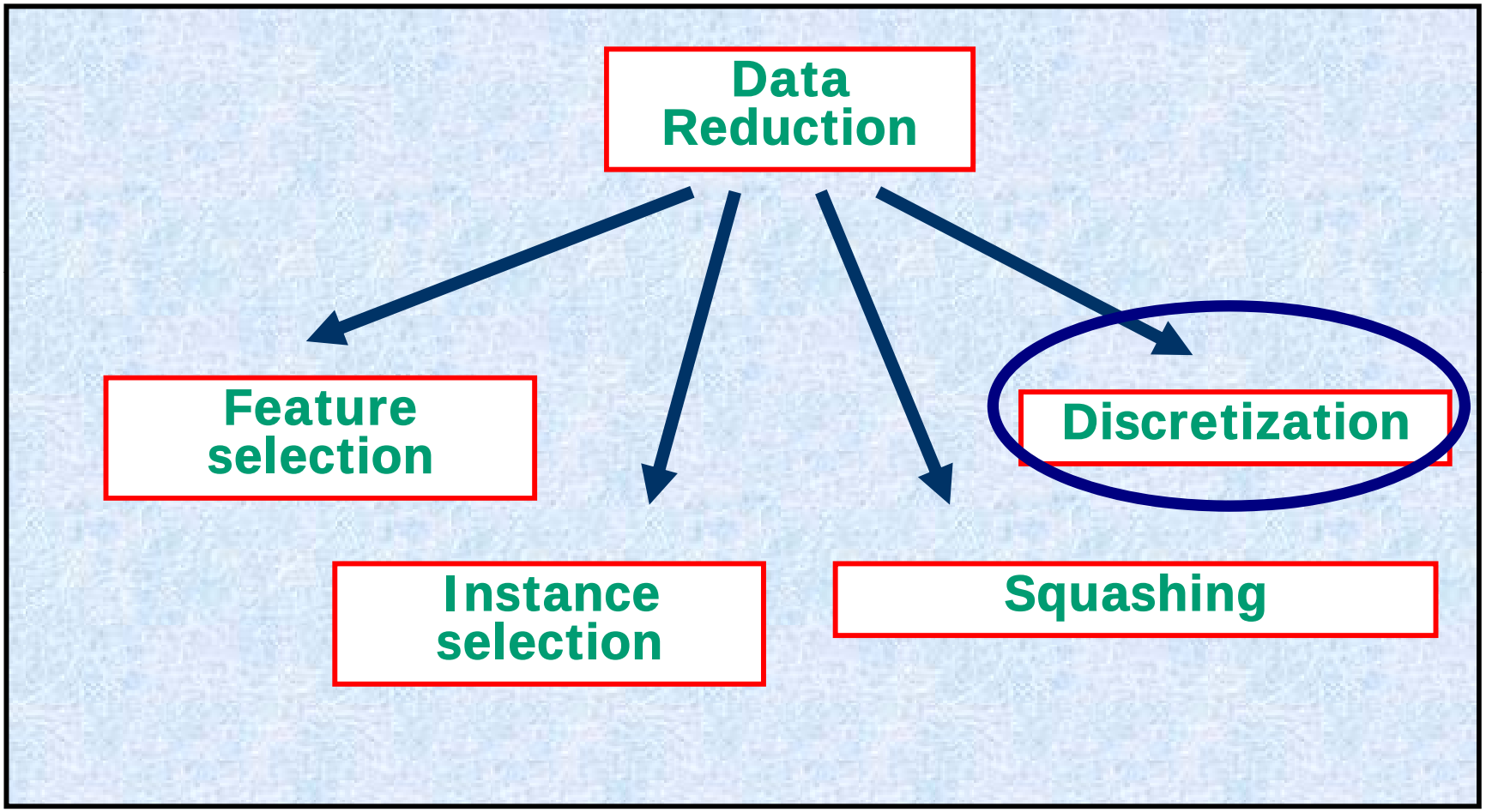


Data Reduction



Reference:
A. Owen, Data Squashing by Empirical Likelihood.
Data Mining and Knowledge Discovery 7, 101-113, 2003.

Data Reduction



Reference:
H. Liu, F. Hussain, C.L. Tan, M. Dash. Discretization: An Enabling Technique. Data mining and Knowledge Discovery 6, 393-423, 2002.

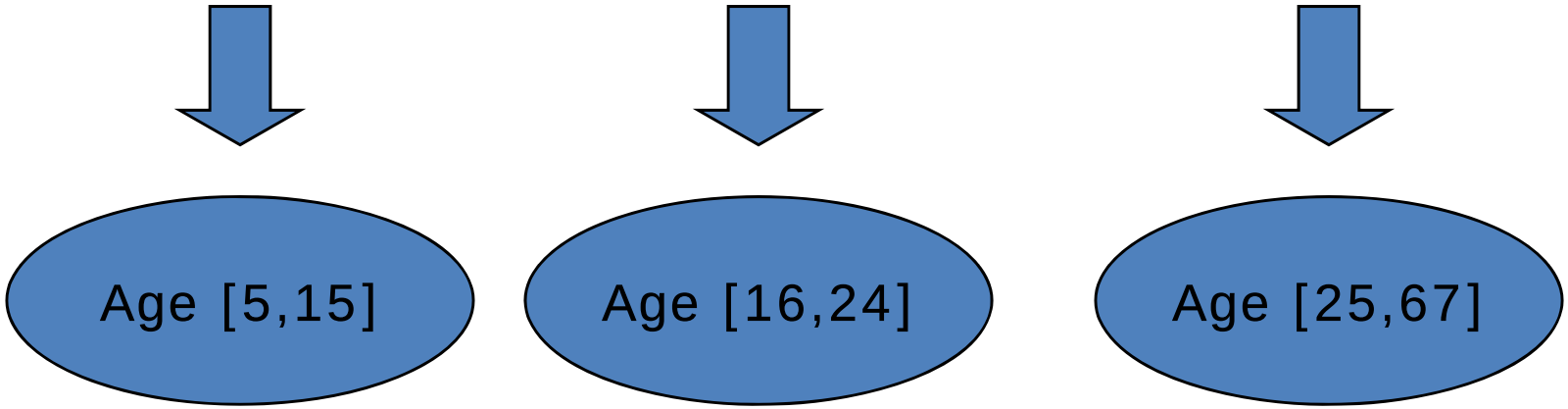
Discretization

- Three types of attributes:
  - Nominal — values from an unordered set
  - Ordinal — values from an ordered set
  - Continuous — real numbers
- Discretization:
  - ✉divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes, e.g. most versions of Naïve Bayes, CHAID
  - Reduce data size by discretization
  - Prepare for further analysis

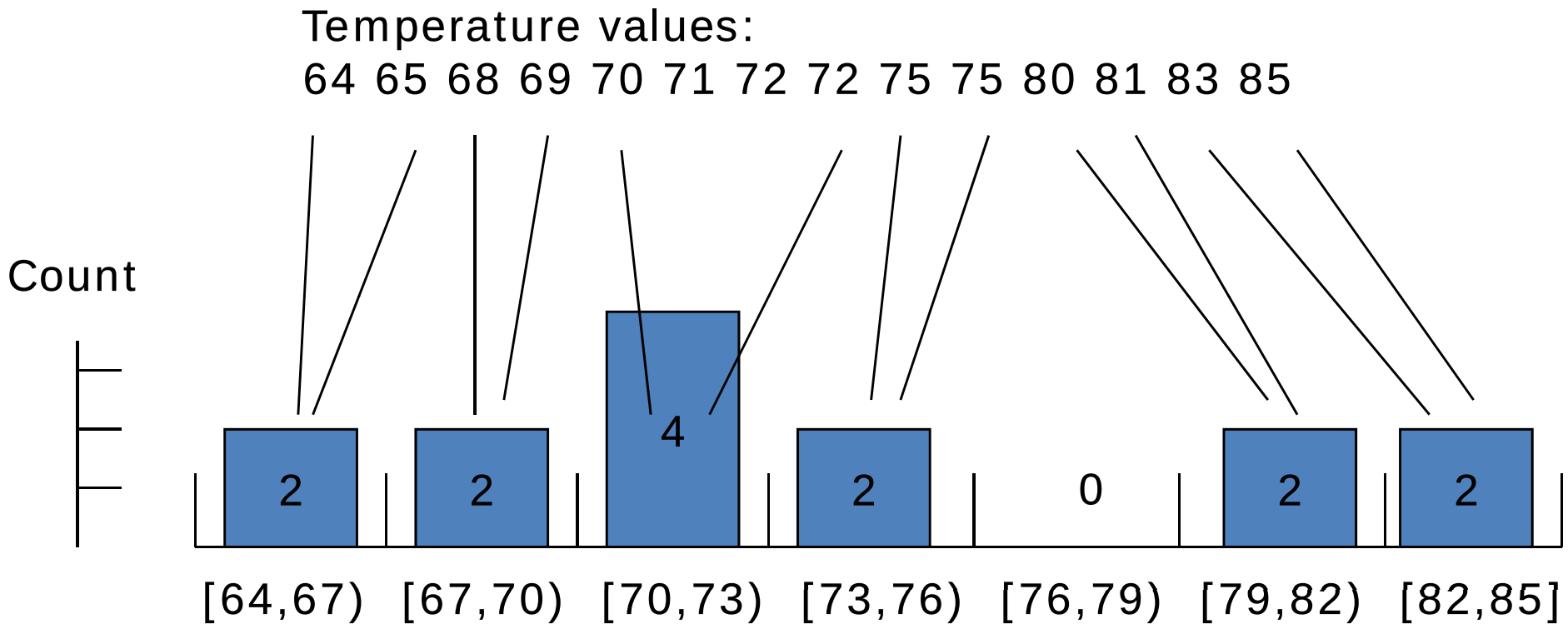
Discretization

- Divide the range of a continuous (numeric) attribute into intervals
- Store only the interval labels
- Important for association rules and classification

age	5	6	6	9	…	15	16	16	17	20	…	24	25	41	50	65	…	67
own a car	0	0	0	0	…	0	1	0	1	1	…	0	1	1	1	1	…	1

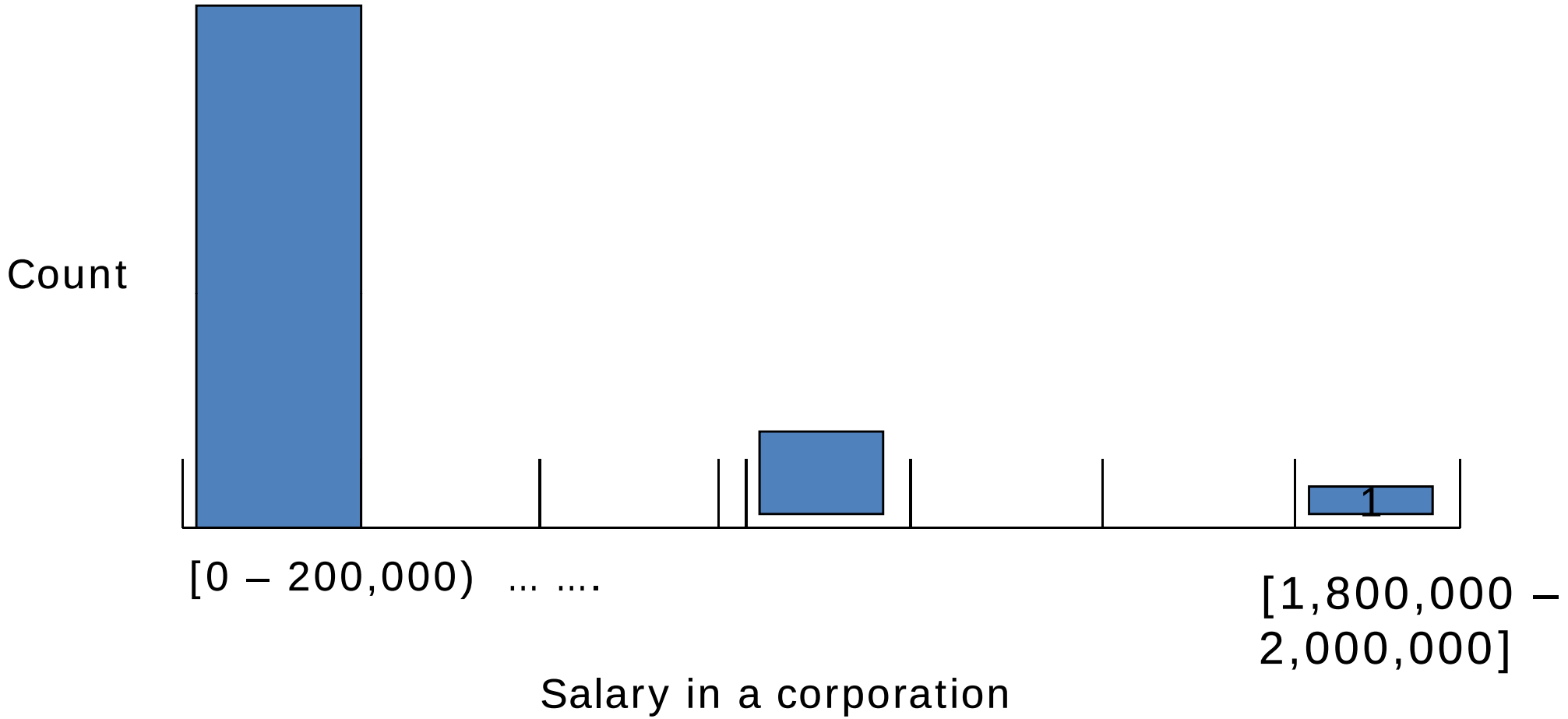


Discretization: Equal-width



Equal Width, bins Low <= value < High

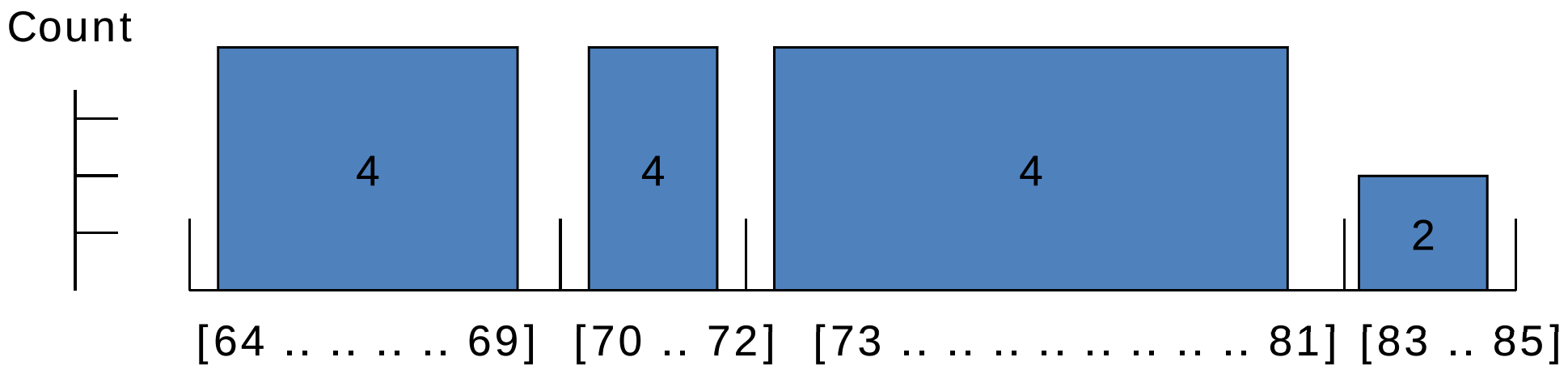
Discretization:
Equal-width may produce clumping



What can we do to get a more even distribution?

Discretization: Equal-height

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



Equal Height = 4, except for the last bin

Discretization: Equal-height advantages

- Generally preferred because avoids clumping
- In practice, “almost-equal” height binning is used which avoids clumping and gives more intuitive breakpoints
- Additional considerations:
  - don’t split frequent values across bins
  - create separate bins for special values (e.g. 0)
  - readable breakpoints (e.g. round breakpoints)

Discretization considerations

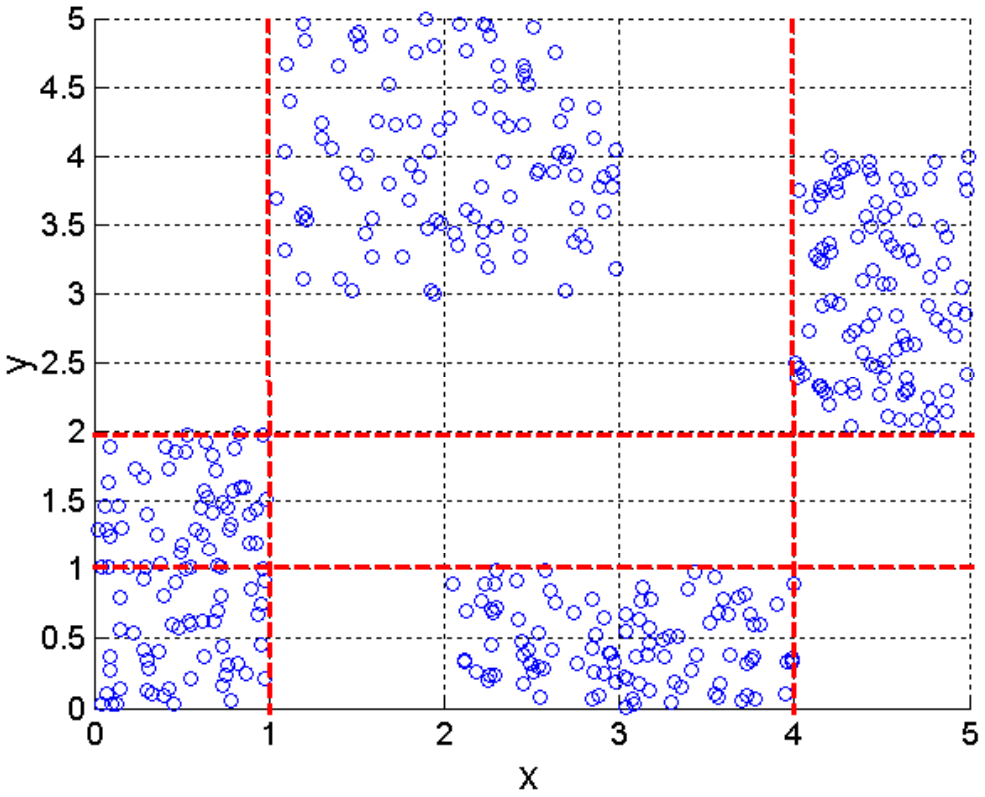
- Equal Width is simplest, good for many classes
  - can fail miserably for unequal distributions
- Equal Height gives better results

How else can we discretize?

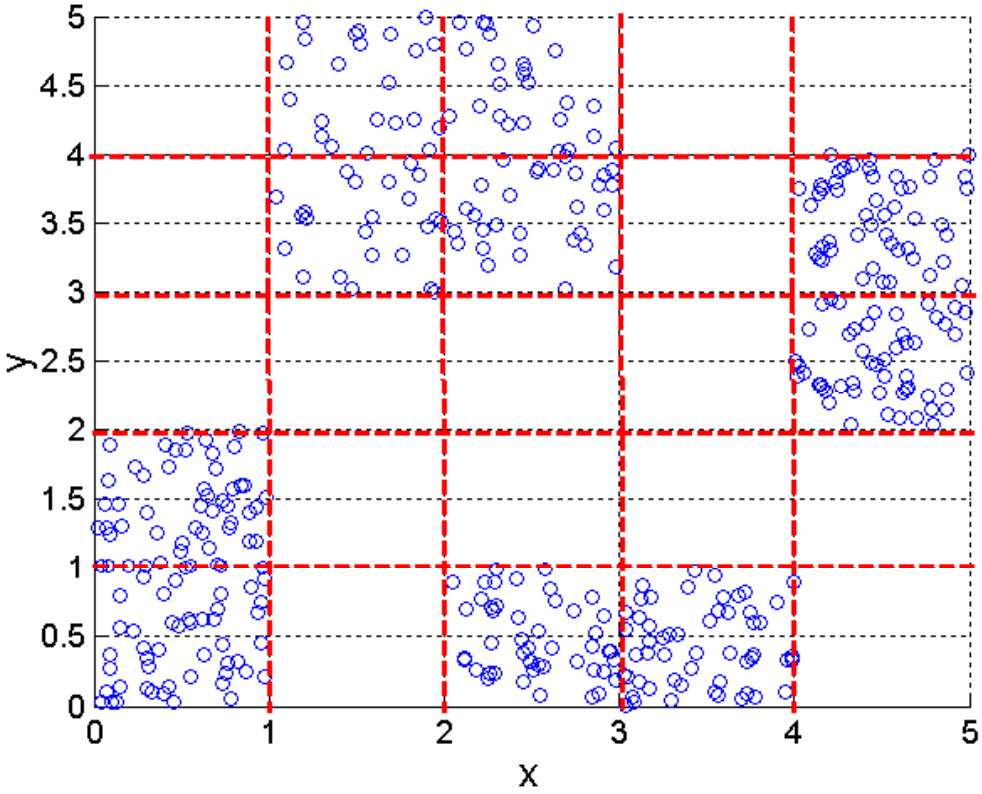
- Class-dependent can be better for classification
  - Note: decision trees build discretization on the fly
  - Naïve Bayes requires initial discretization
- Many other methods exist …

Discretization Using Class Labels

- **Entropy based approach**

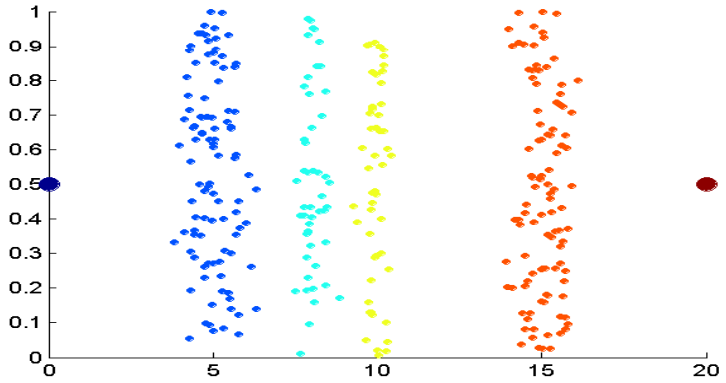


3 categories for both x and y

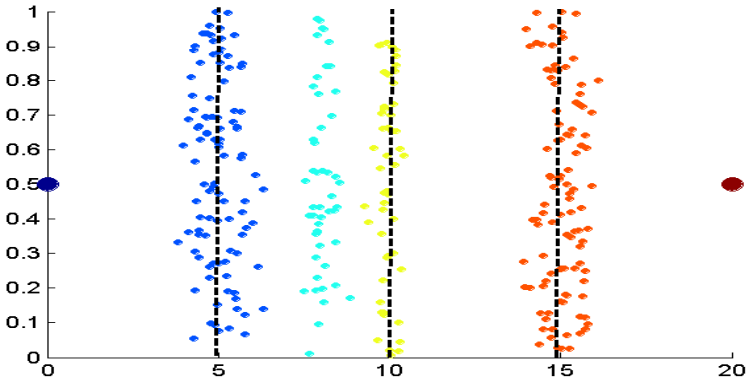


5 categories for both x and y

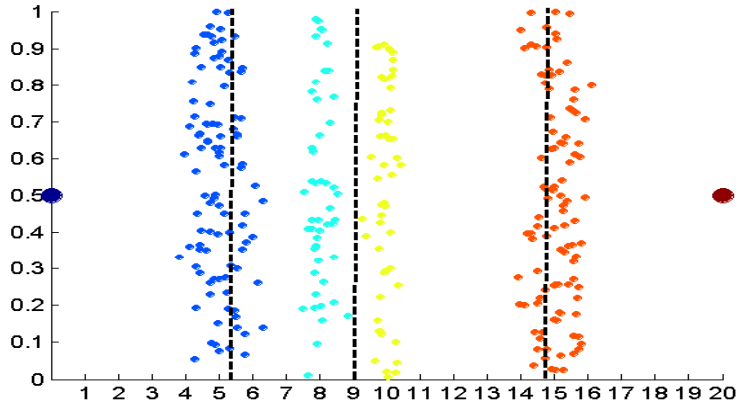
Discretization Without Using Class Labels



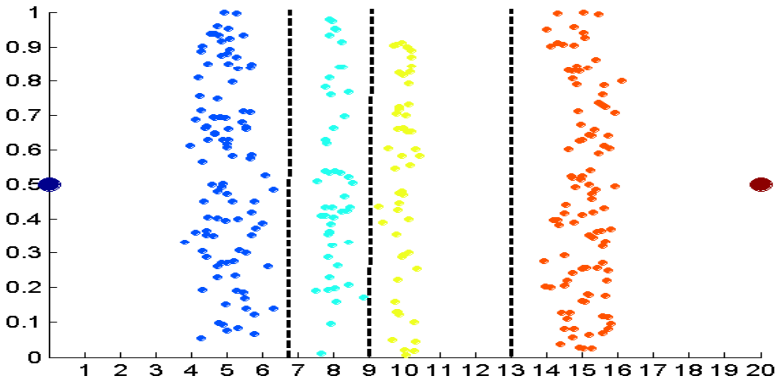
Data



Equal interval

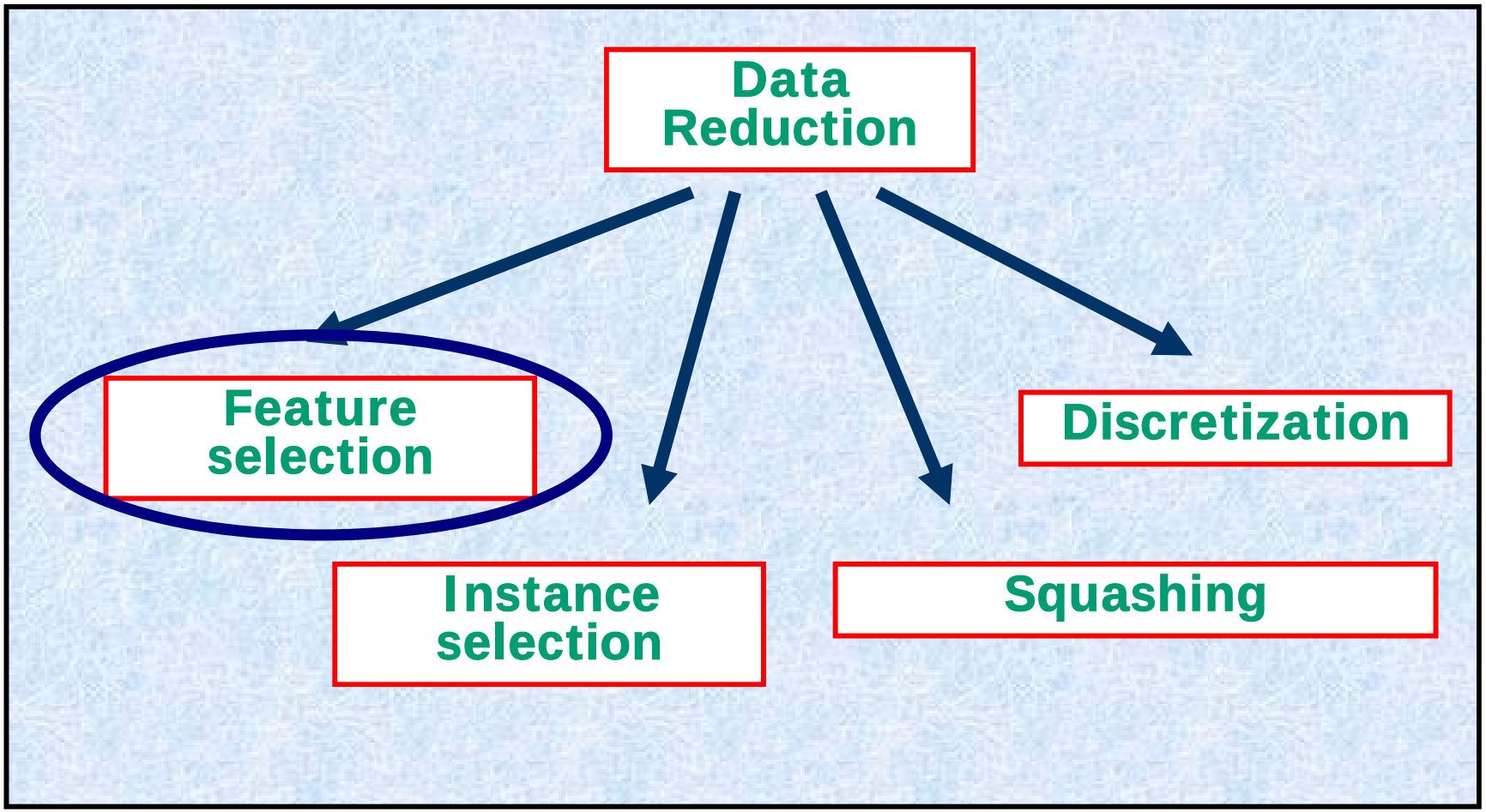


Equal frequency



K-means

Data Reduction



Reference:

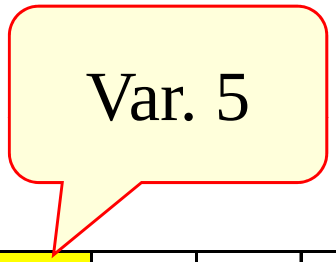
H. Liu, H. Motoda. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic, 1998.

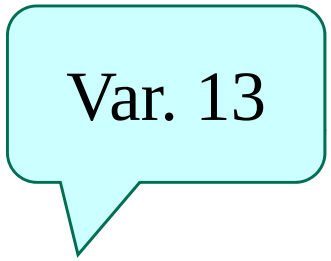
H. Liu, H. Motoda (Eds.) Feature Extraction, Construction, and Selection: A Data Mining Perspective, Kluwer Ac., 1998.

Feature Selection

- Another way to reduce dimensionality of data
- **Redundant features**
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- **Irrelevant features**
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

Feature Selection





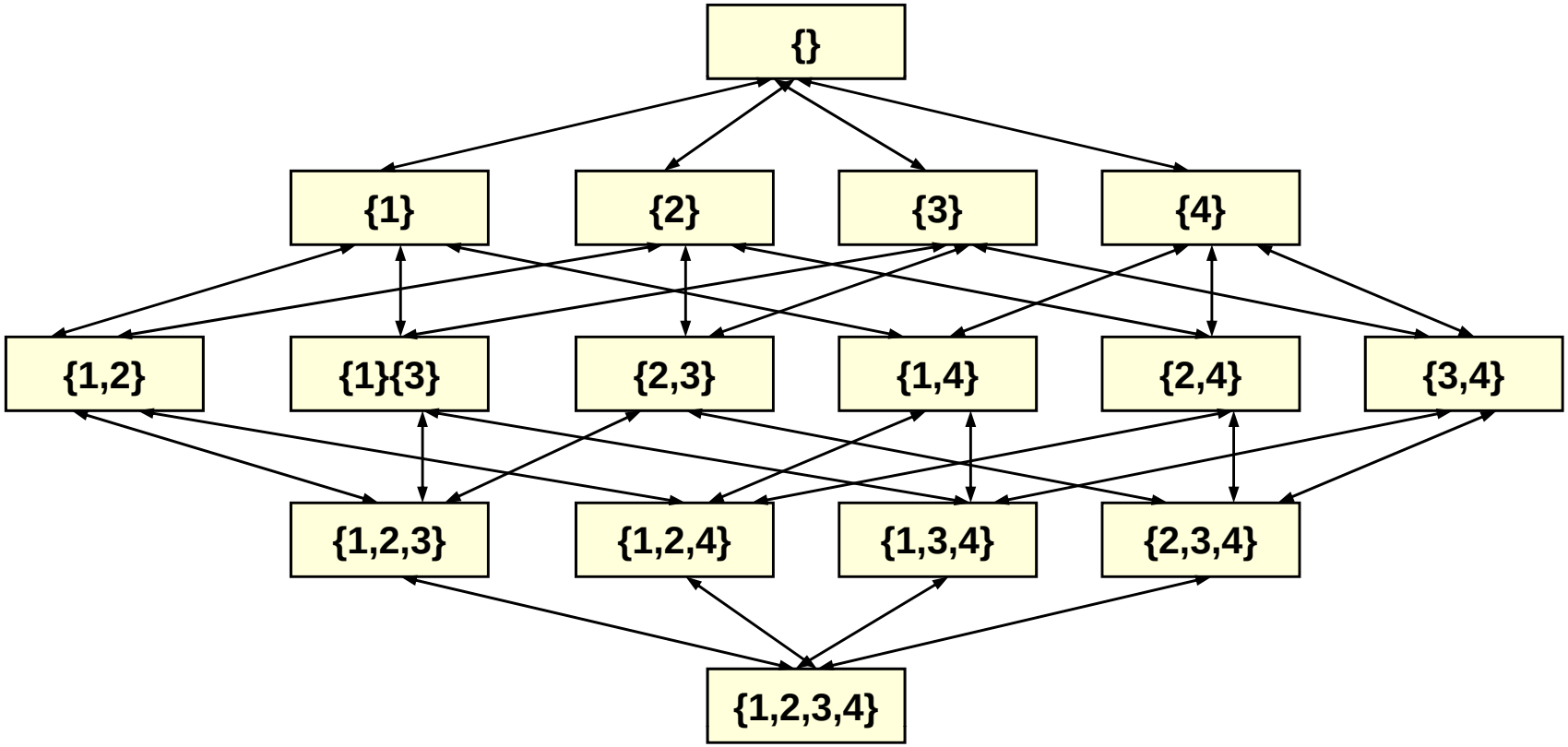
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
B	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
C	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
D	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
E	0	1	0	0	0	1	1	0	1	1	0	0	0	0	1	0
F	1	1	1	0	1	1	0	0	1	0	1	0	0	1	0	0

Feature Selection

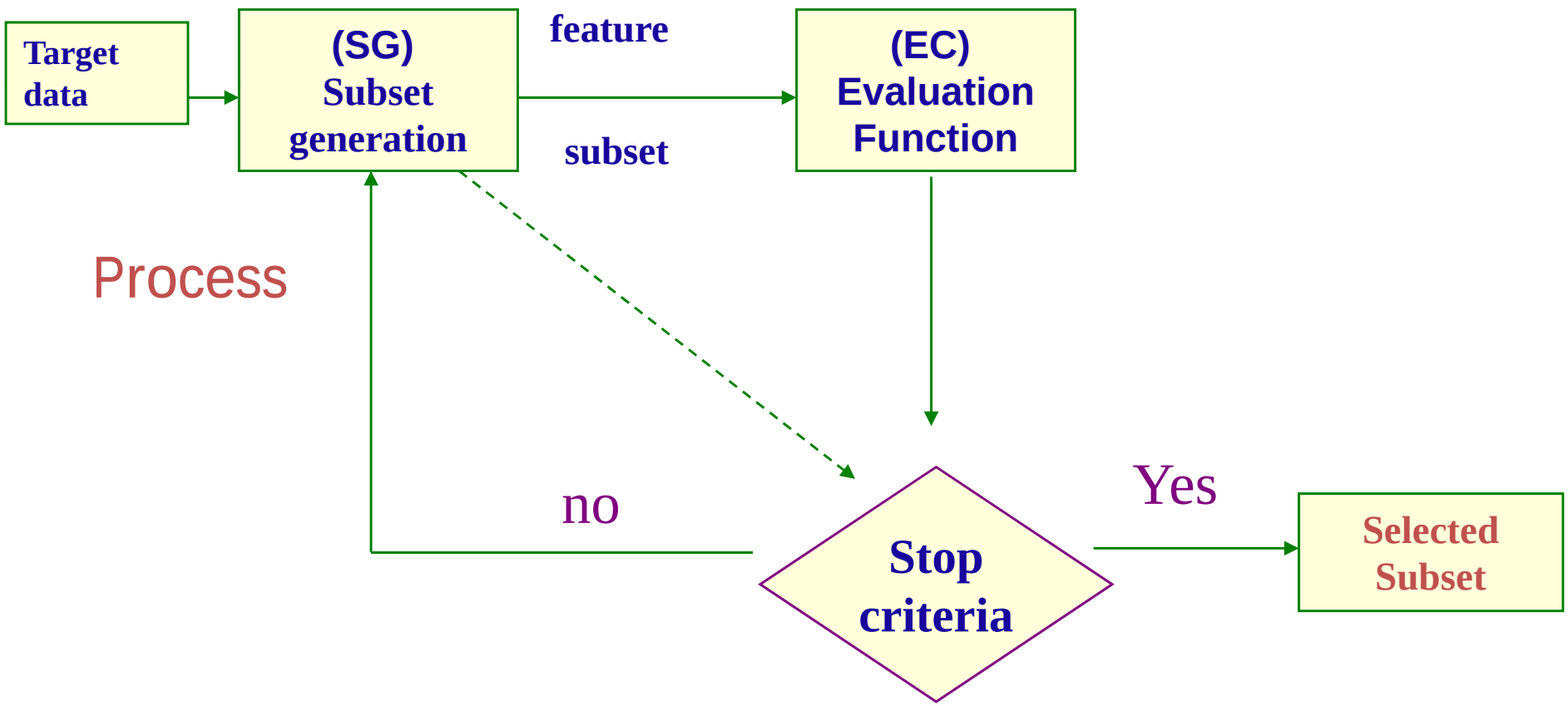
- **Techniques:**
  - Brute-force approch:
    - Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches:
    - Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches:
    - Features are selected before data mining algorithm is run
  - Wrapper approaches:
    - Use the data mining algorithm as a black box to find best subset of attributes

Feature Selection

It can be considered as a search problem



Feature Selection

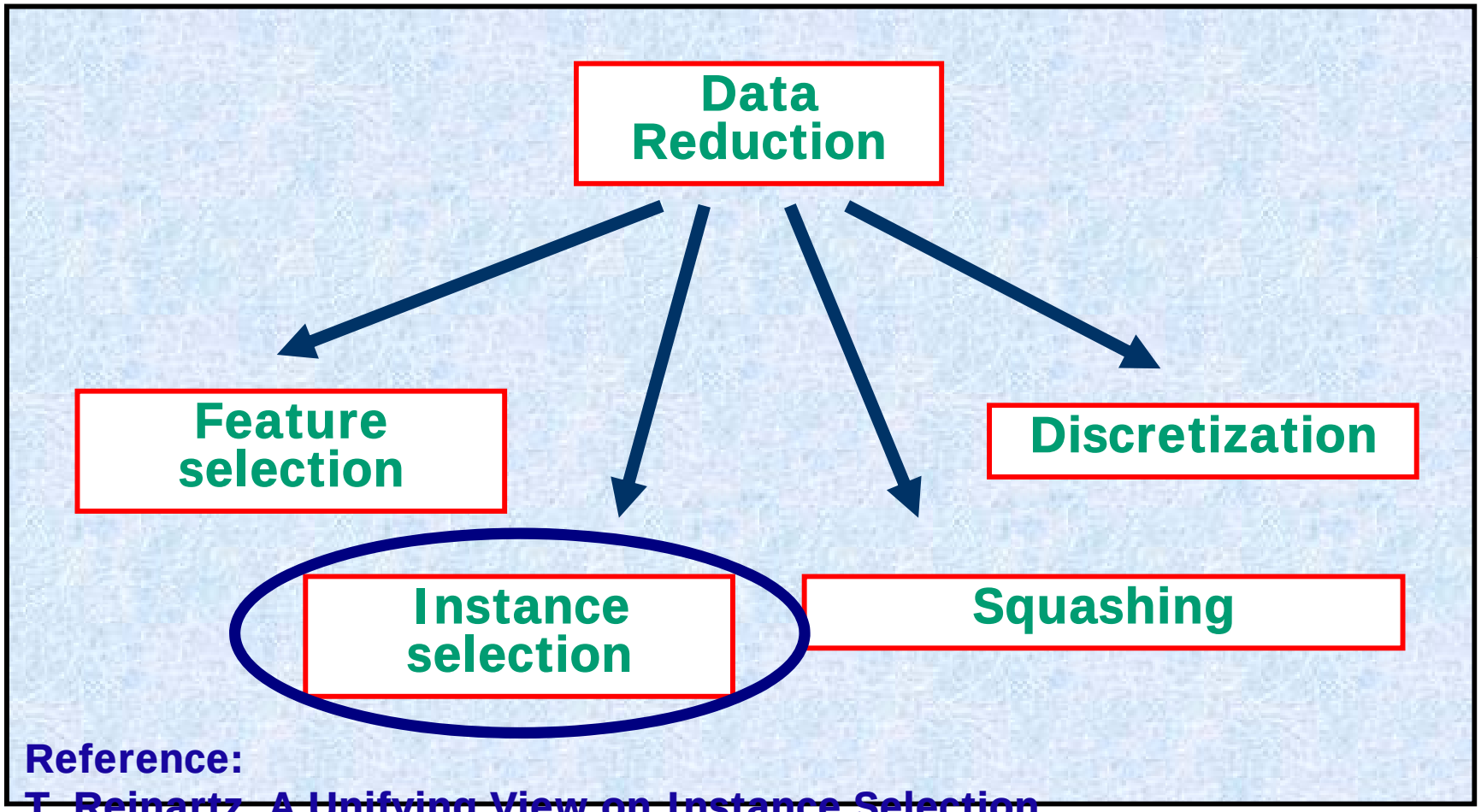


Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

- Three general methodologies:
  - Feature Extraction
    - domain-specific
  - Mapping Data to New Space
  - Feature Construction
    - combining features

Data Reduction



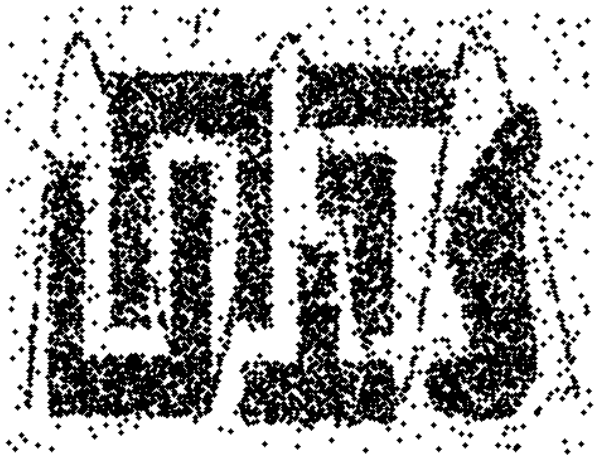
Reference:
T. Reinartz. A Unifying View on Instance Selection.
Data Mining and Knowledge Discovery 6, 191-210, 2002.

Instance Selection

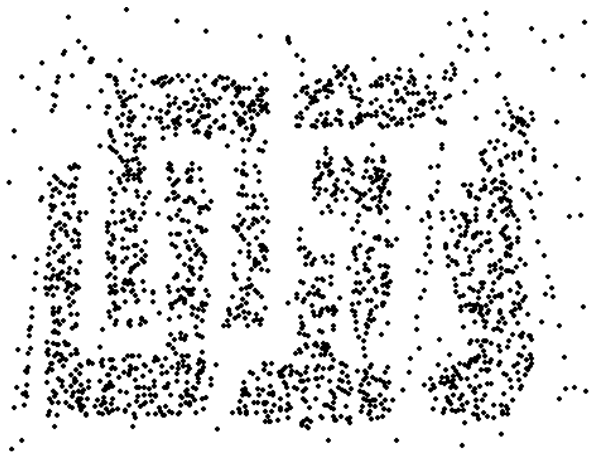
- IS obtains relevants patterns for getting the maximun model behaviour. The result is:
  - Reduced data set → fast algorithms
  - More precision → better algorithm accuracy
  - Simple results → high interpretability
- IS and Transformation (data extraction)

Instance Selection

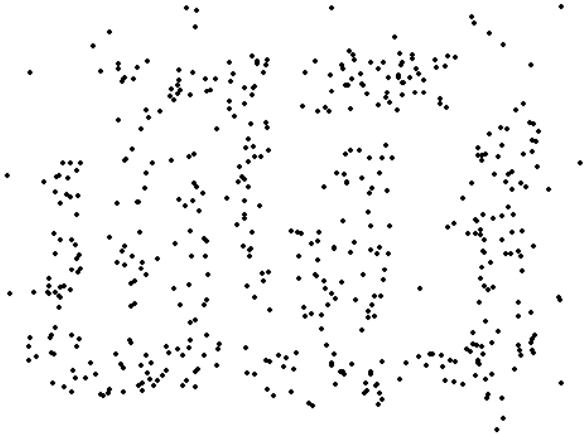
Example: Different sizes



8000 points

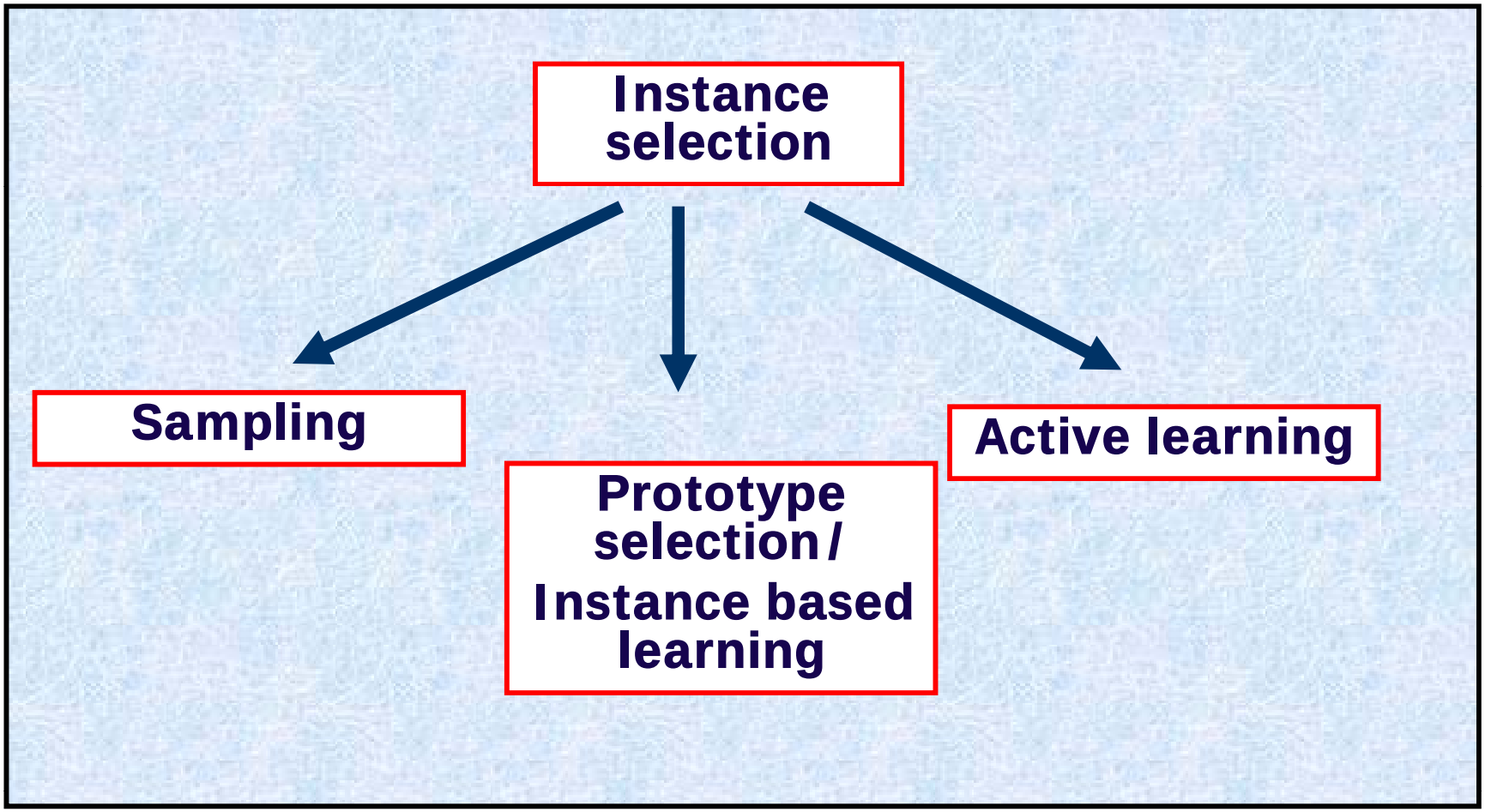


2000 points

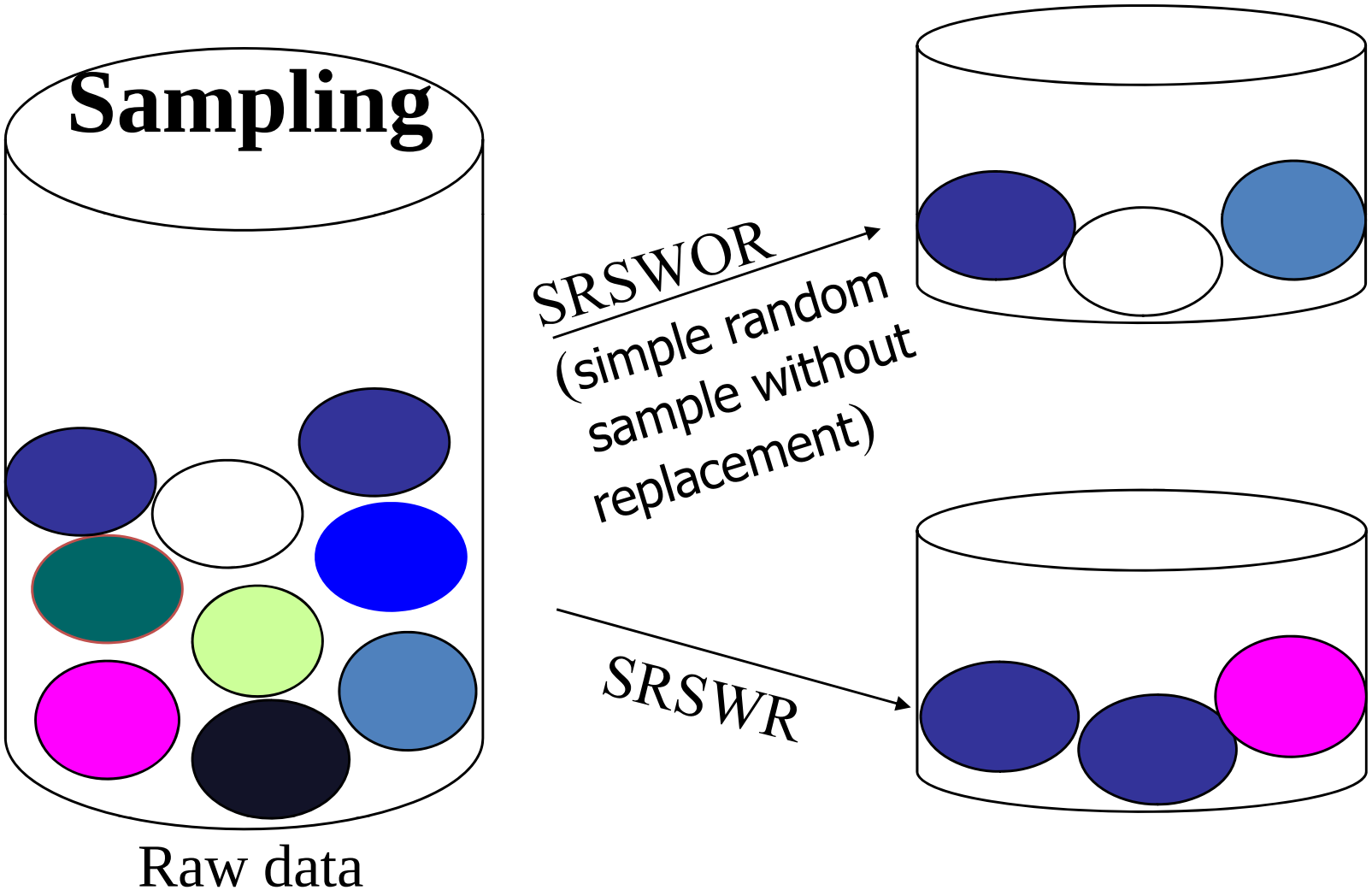


500 points

Instance Selection



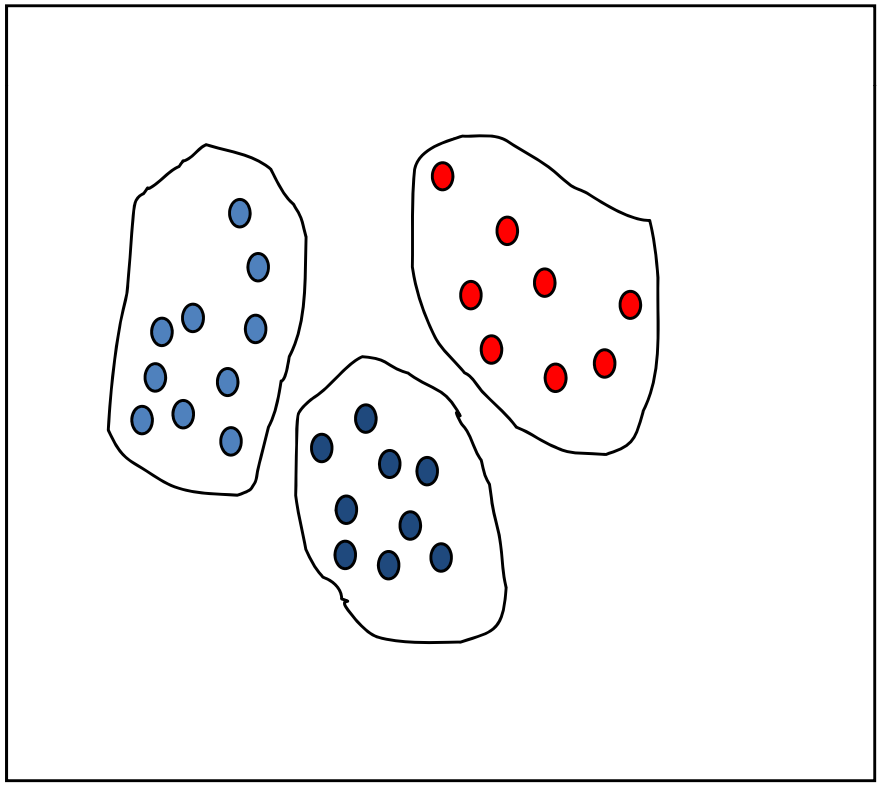
Instance Selection



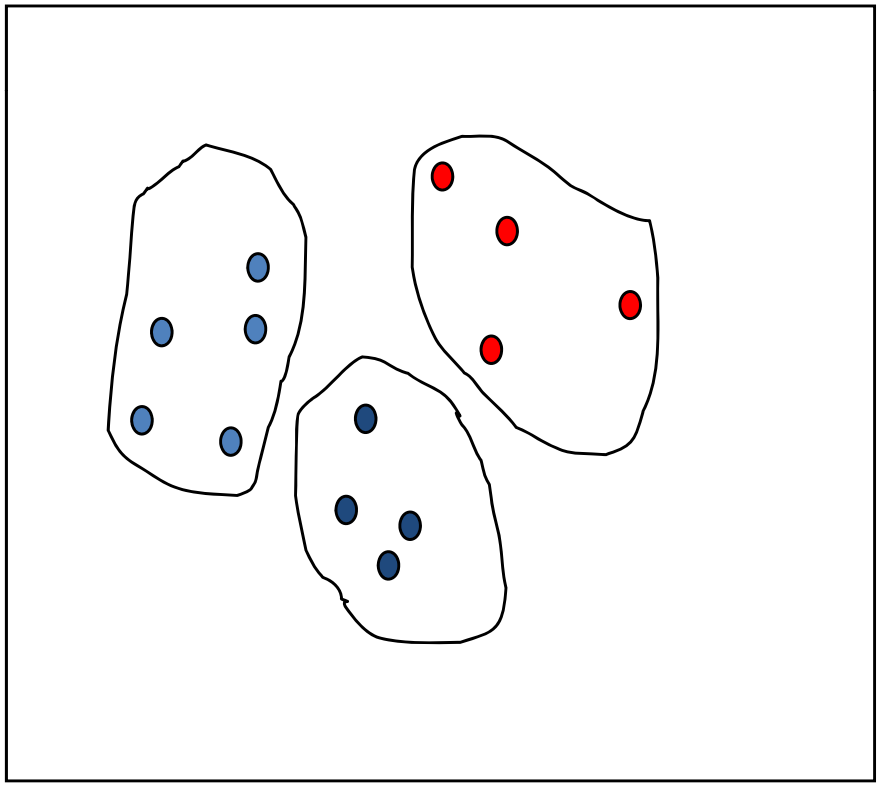
Instance Selection

Sampling

Raw data

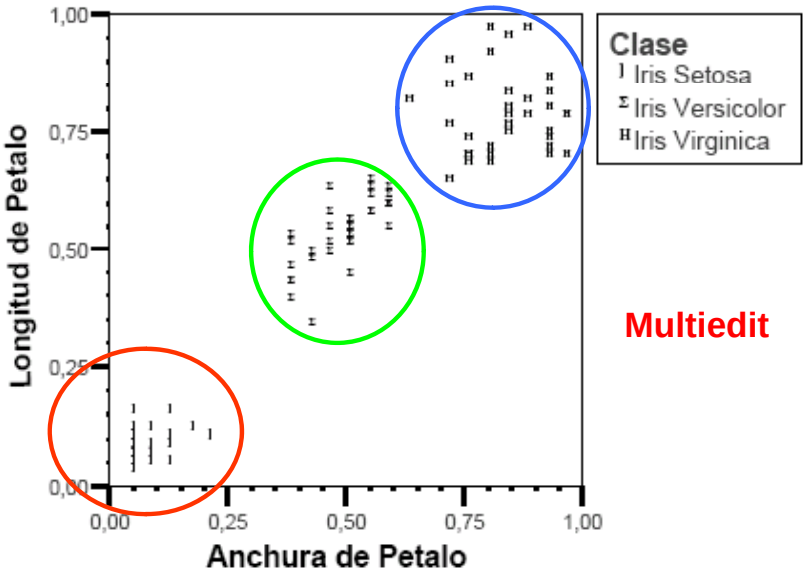


Reduced data

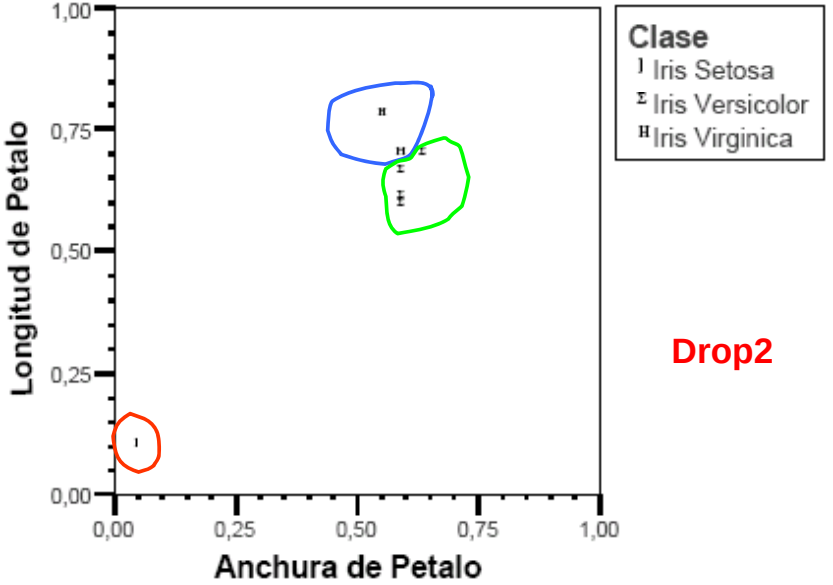


Instance Slelection

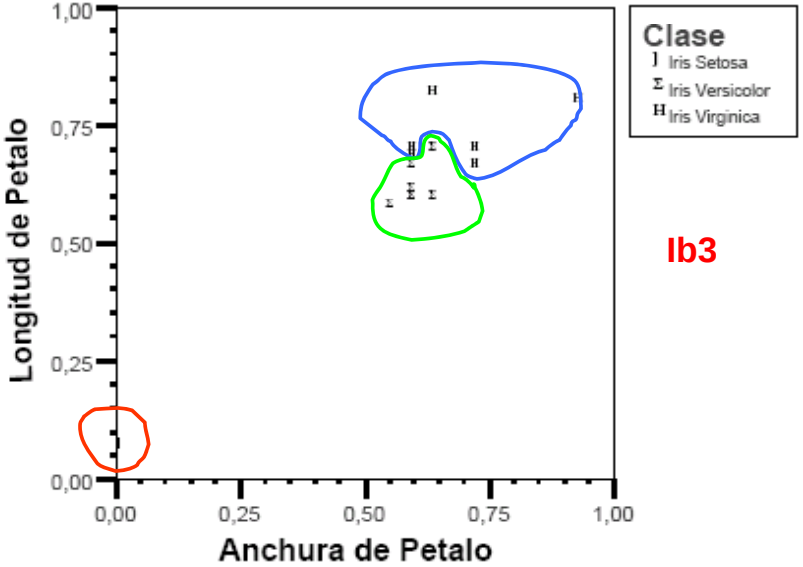
Example: Prototype selection methods



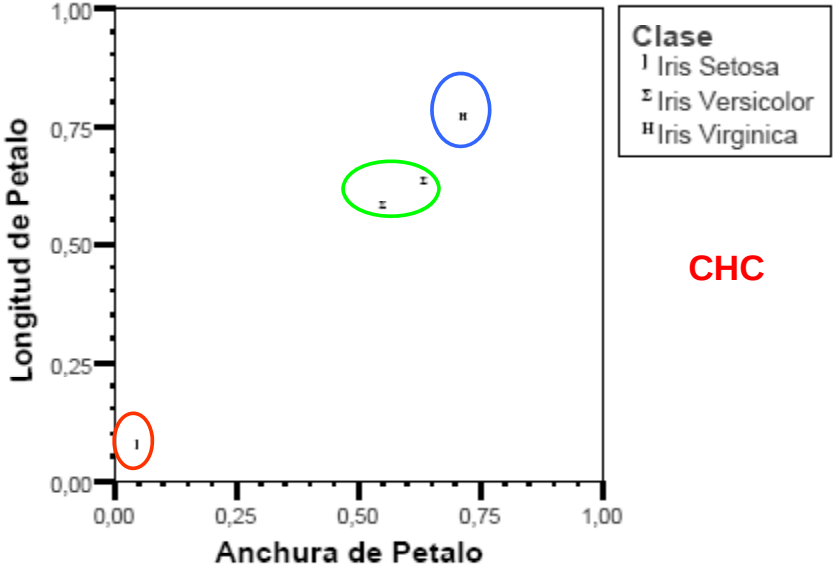
Multiedit



Drop2



Ib3



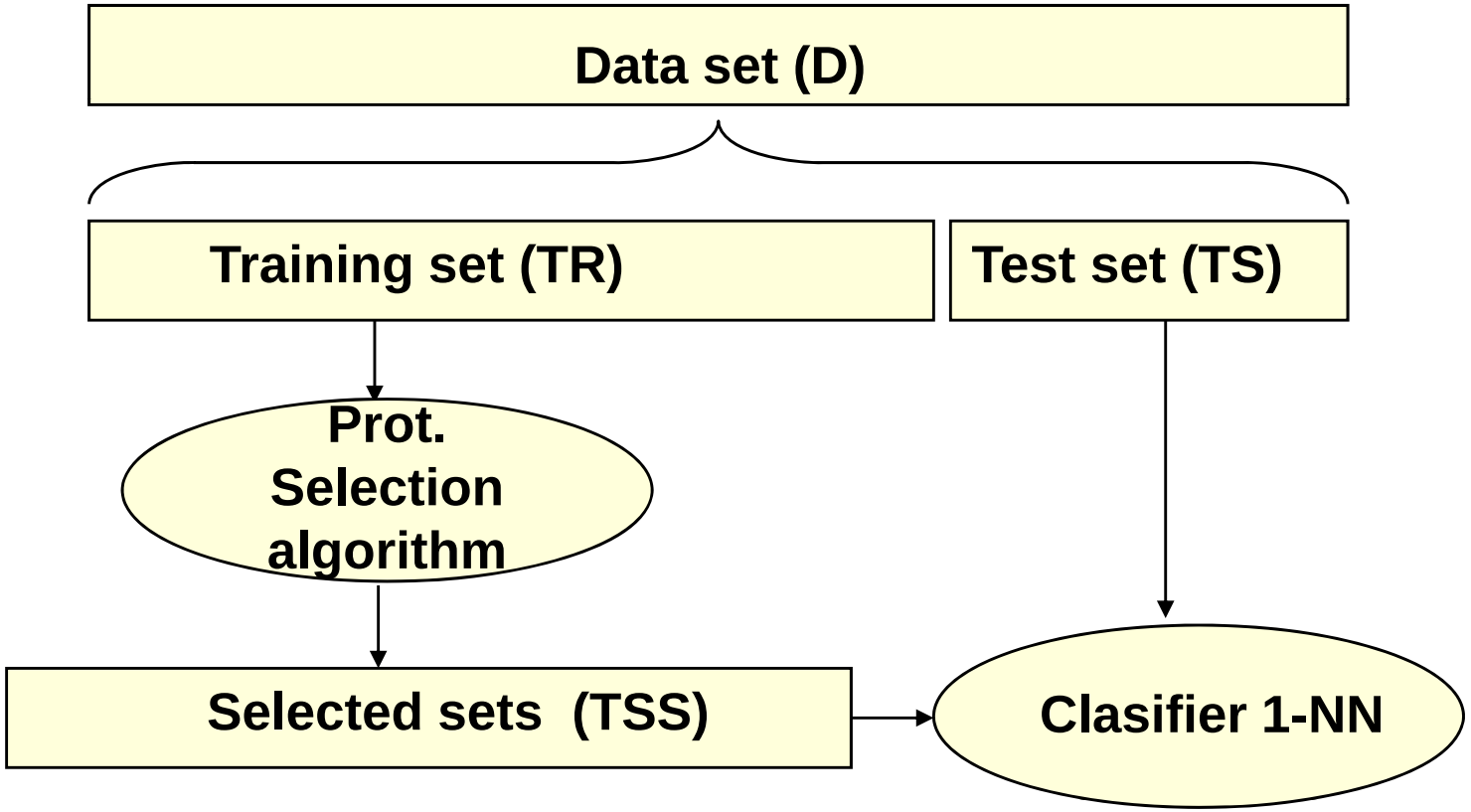
CHC

Reference: ***J.R. Cano, F. Herrera, M. Lozano.*** **Using Evolutionary Algorithms as Instance Selection for Data Reduction in KDD: An Experimental Study. IEEE Trans. on Evolutionary Computation 7:6 (2003) 561-575.**

Instance Selection

MODEL: Prototype selection

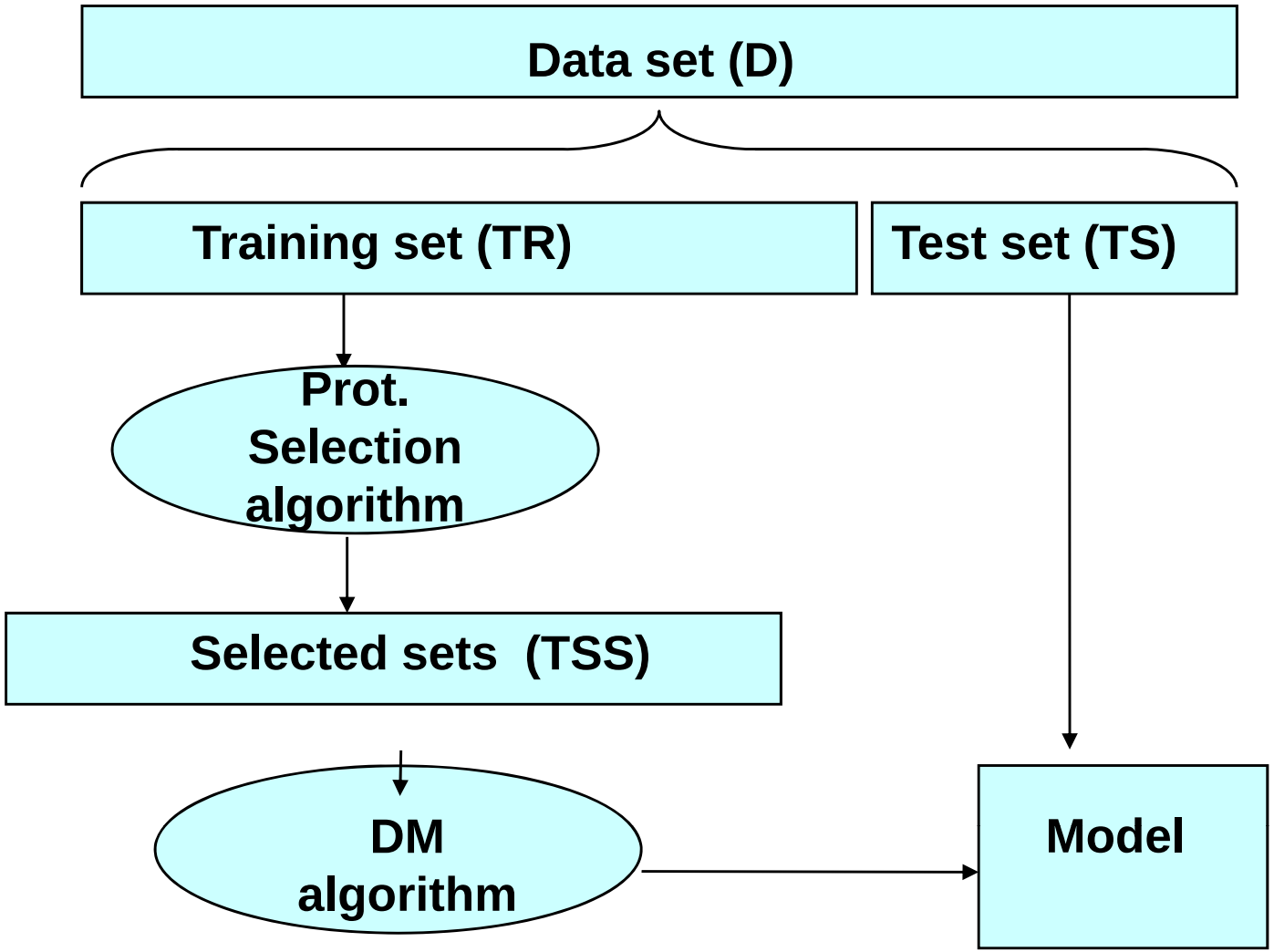
Using 1-NN



Instance Selection

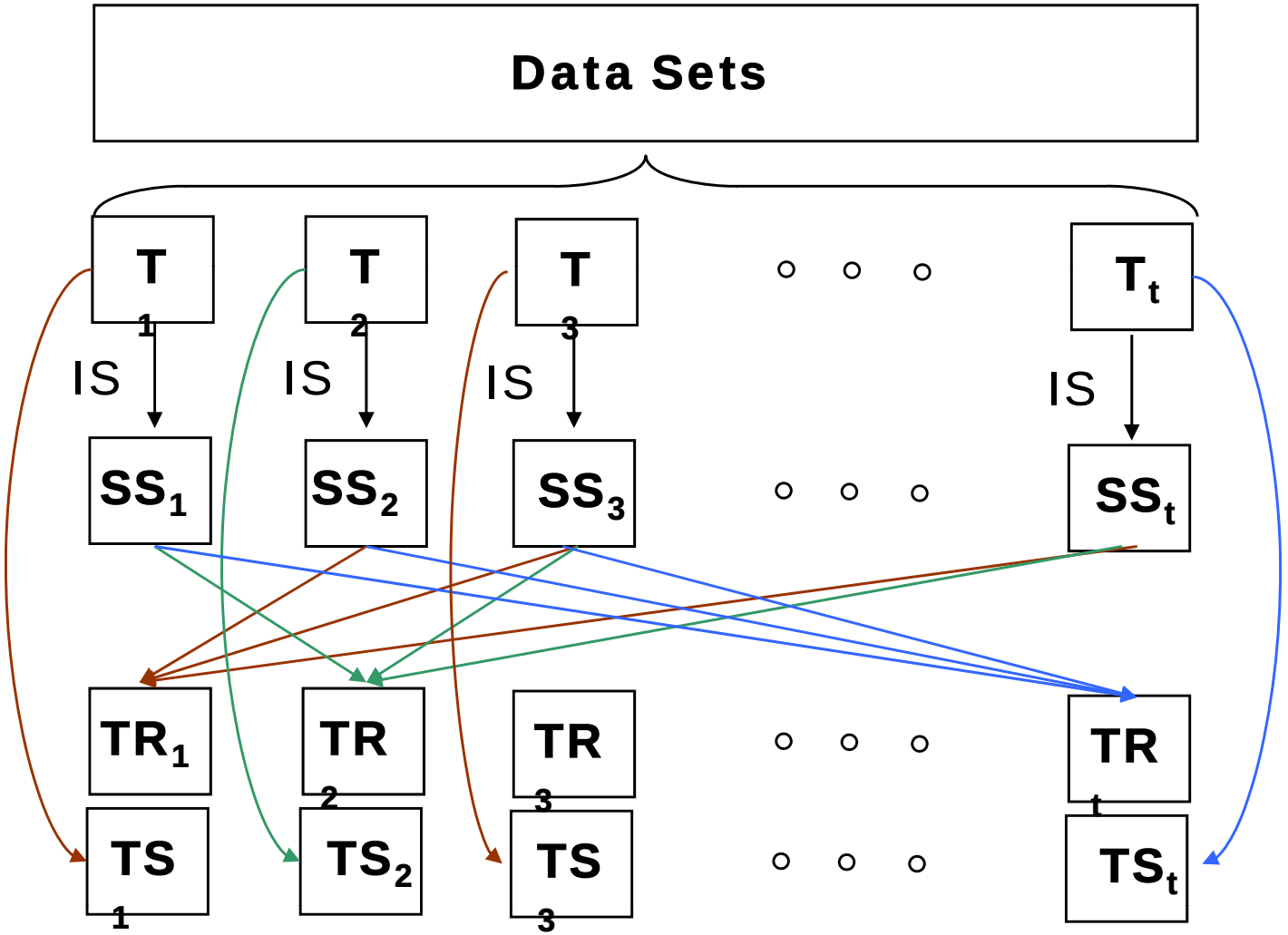
MODEL. Training set selection

Using DM algorithm



Instance Selection

Large data bases. **Stratification strategy.**



Referencia: *J.R. Cano, F. Herrera, M. Lozano.* **Stratification for Scaling Up Evolutionary Prototype Selection.** Pattern Recognition Letters 26:7 (2005) 953-963.

Instance Selection

Large data bases. Stratification strategy.

Example – Kdd Cup'99

Problem	Items	Features	Classes
Kdd Cup'99	494022	41	23

Instance Selection

Example – Kdd Cup'99

	Tpo	% Red	% Ac. Trn	% Ac Test
1-NN cl	18568		99.91	99.91
Cnn st 100	8	81.61	99.30	99.27
Cnn st 200	3	65.57	99.90	99.15
Cnn st 300	1	63.38	99.89	98.23
Ib2 st 100	0	82.01	97.90	98.19
Ib2 st 200	3	65.66	99.93	98.71
Ib2 st 300	2	60.31	99.89	99.03
Ib3 st 100	2	78.82	93.83	98.82
Ib3 st 200	0	98.27	98.37	98.93
Ib3 st 300	0	97.97	97.92	99.27
CHC st 100	1960	99.68	99.21	99.43
CHC st 200	418	99.48	99.92	99.23
CHC st 300	208	99.28	99.93	99.19

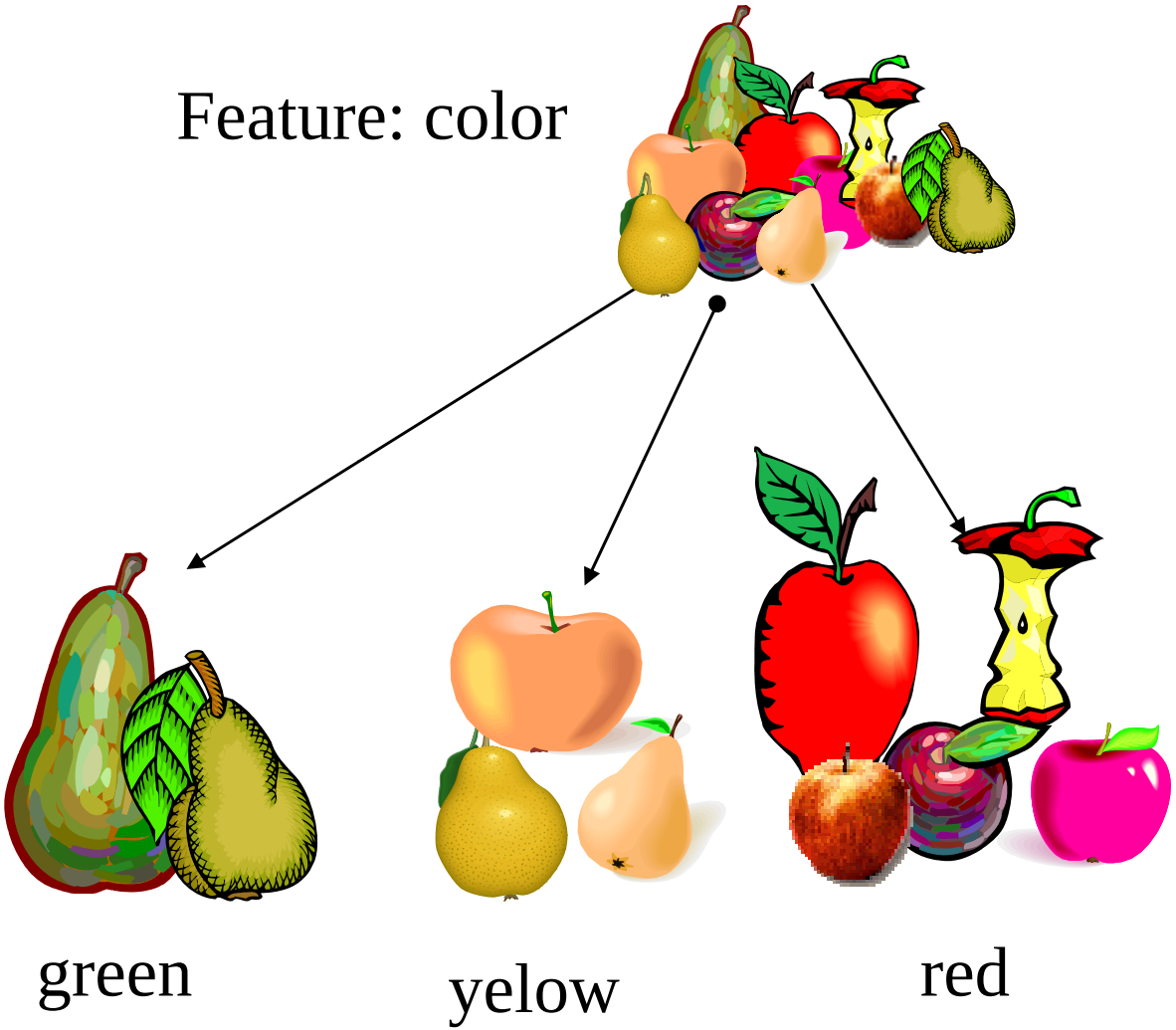
J.R. Cano, F. Herrera, M. Lozano, **Stratification for Scaling Up Evolutionary Prototype Selection**. *Pattern Recognition Letters, 26, (2005), 953-963.*

Data Preparation

Outline

- ✓ Introduction
- ✓ Preprocessing
- ✓ Data Reduction
  Discretization, Feature Selection, Instance Selection
- ✓ Ex.: Instance Selection and Decision Trees
- ✓ Concluding Remarks

Ex.: Instance Selection and Decision Trees

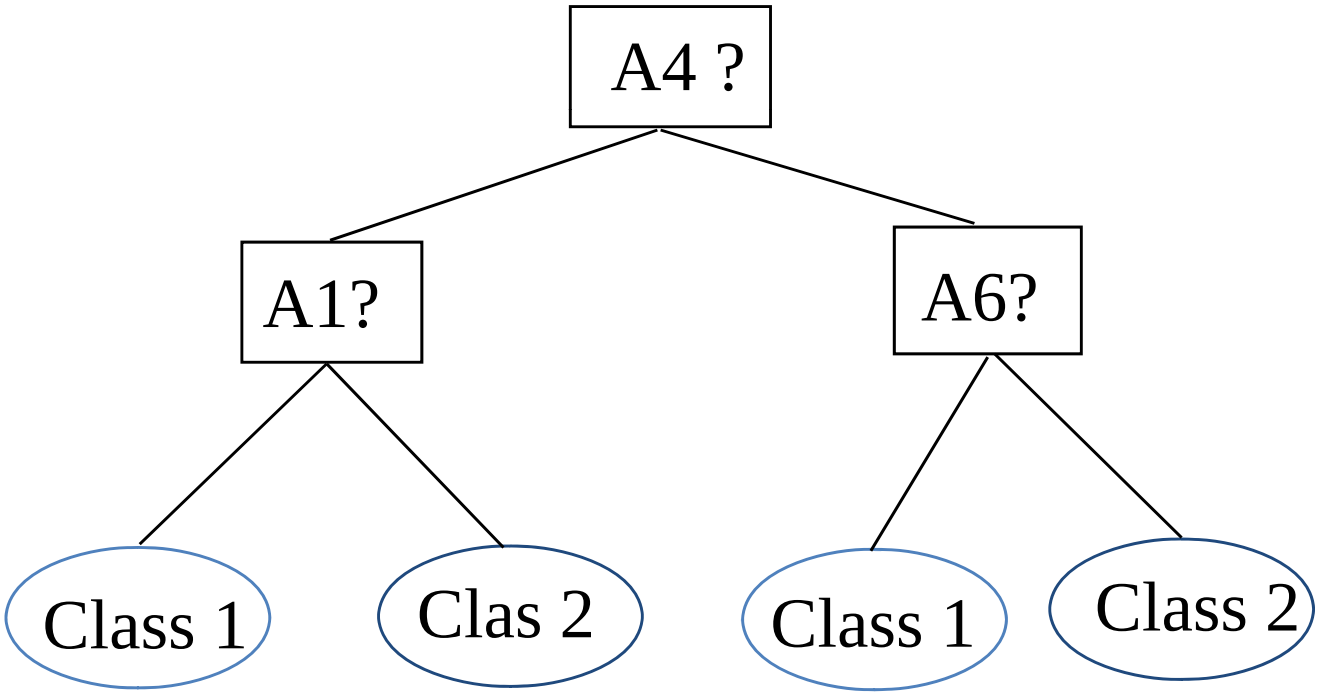


J.R. Cano, F. Herrera, M. Lozano, **Evolutionary Stratified Training Set Selection for Extracting Classification Rules with Trade-off Precision-Interpretability**. *Data and Knowledge Engineering 60 (2007) 90-108.*

Instance Selection and Decision Trees

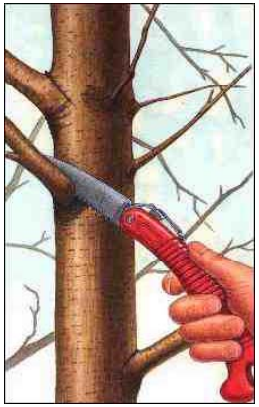
Example: Decision Tree

Attribute set:
{A1, A2, A3, A4, A5, A6}



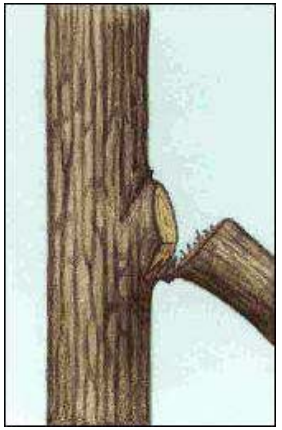
-----> Reduced attribute set: {A1, A4, A6}

Instance Selection and Decision Trees



Comprehensibility: small Decision trees

Pruning: We can cut / eliminate nodes



Instance selection strategies allow us to build decision tree for large data base reducing the tree size.

Instance Selection can increase the decision tree interpretability.

Instance Selection and Decision Trees

Training set selection. Example – Kdd Cup'99

Problem	Items	Features	Classes
Kdd Cup'99	494022	41	23

Instance Selection and Decision Trees

Kdd Cup'99. Strata: 100

	Rules Number	% Reduction	*C4.5*	
			%Ac Trn	%Ac Test
C4.5	252		99.97%	99.94%
Cnn Strat	83	81.61%	98.48%	96.43%
Drop1 Strat	3	99.97%	38.63%	34.97%
Drop2 Strat	82	76.66%	81.40%	76.58%
Drop3 Strat	49	56.74%	77.02%	75.38%
Ib2 Strat	48	82.01%	95.81%	95.05%
Ib3 Strat	74	78.92%	99.13%	96.77%
Icf Strat	68	23.62%	99.98%	99.53%
CHC Strat	**9**	**99.68%**	**98.97%**	**97.53%**

Reduced number of rules (reduced number of variables)

Instance Selection and Decision Trees

Adult Data Set

	Instance Number N	Varia-bles	Rule numbers		Variables per rule		Rules confidence N(Cond,Clas) / N	
Adult 2 classes	30132	14	C4.5	IS-CHC/ C4.5	C4.5	IS-CHC/ C4.5	C4.5	IS-CHC/ C4.5
			359	**5**	**14**	**3**	**0.003**	**0.167**

Instance selection allows us to get more interpretable rule sets (low number of rules and variables).

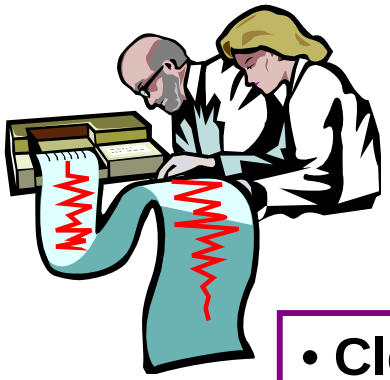
Outline

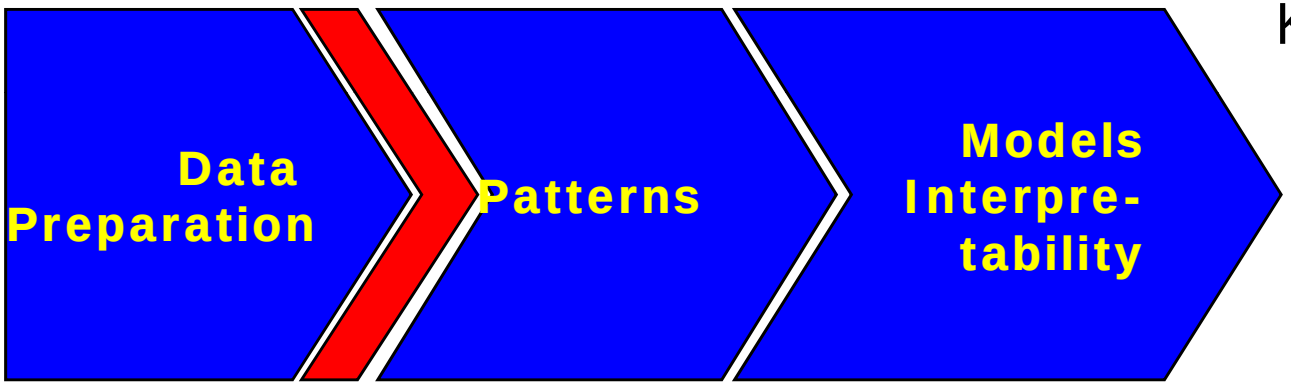
- ✓ Introduction
- ✓ Preprocessing
- ✓ Data Reduction
  Discretization, Feature Selection, Instance Selection
- ✓ Ex.: Instance Selection and Decision Trees
- ✓ Concluding Remarks

Concluding Remarks

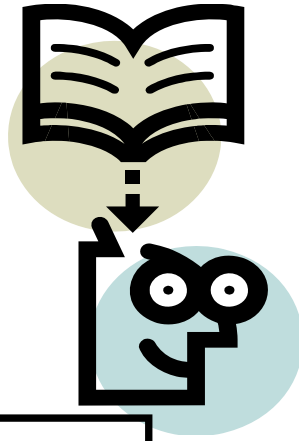
Data preprocessing is a necessity when we work with real applications.

Target data





Knowledge



- Cleaning
- Integration
- Transformation
- Reduction

- Association rules
- Classification
- Prediction
- Clustering

- Visualization
- Validation

Bibliography: http://sci2s.ugr.es/keel (List of references by Specific Areas)

Concluding Remarks

Advantage: Data preparation allows us to apply the data mining algorithms in a quicker/simpler way, getting high quality models: high precision and/or high interpretability.

All is not advantage: The data preparation is not a structured area with a specific methodology for managing a new problem. Every problem can need a different preprocessing process, using different tools.

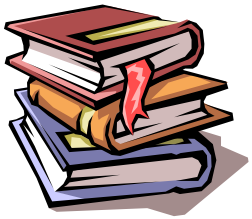
Concluding Remarks

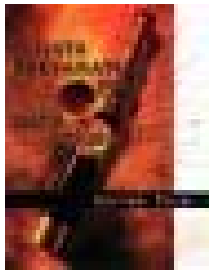
Summary

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

- Data preparation is a big issue for both warehousing and mining
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and data transformation
- A lot a methods have been developed but still an active area of research
- The cooperation between DM algorithms and data preparation methods is an interesting/active area.

Bibliography





Dorian Pyle
Data Preparation for Data Mining
Morgan Kaufmann, Mar 15, 1999

Mamdouh Refaat
Data Preparation for Data Mining Using SAS
Morgan Kaufmann, Sep. 29, 2006)

Tamraparni Dasu, Theodore Johnson
Exploratory Data Mining and Data Cleaning
Wiley, 2003

Bibliography of the Research Group SCI²S

J.R. Cano, S. García, F. Herrera, **Subgroup Discovery in Large Size Data Sets Preprocessed Using Stratified Instance Selection for Increasing the Presence of Minority Classes**. *Pattern Recognition Letters 29 (2008) 2156-2164, doi:10.1016/j.patrec.2008.08.*.

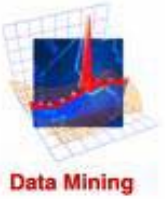
S. García, J.R. Cano, F. Herrera, **A Memetic Algorithm for Evolutionary Prototype Selection: A Scaling Up Approach**. *Pattern Recognition 41:8 (2008) 2693-2709, doi:10.1016/j.patcog.2008.02.006*.

J.R. Cano, F. Herrera, M. Lozano, **Evolutionary Stratified Training Set Selection for Extracting Classification Rules with Trade-off Precision-Interpretability**. *Data and Knowledge Engineering 60 (2007) 90-108, doi:10.1016/j.datak.2006.01.008*.

J.R. Cano, F. Herrera, M. Lozano, **Stratification for Scaling Up Evolutionary Prototype Selection**. *Pattern Recognition Letters, 26, (2005), 953-963, doi: 10.1016/j.patrec.2004.09.043* .

J.R. Cano, F. Herrera, M. Lozano, **Using Evolutionary Algorithms as Instance Selection for Data Reduction in KDD: An Experimental Study**. *IEEE Trans. on Evolutionary Computation 7:6 (2003) 561-575, doi: 10.1109/TEVC.2003.819265*.

Available at: http://sci2s.ugr.es/publications/byAll.php



Data Mining and Soft Computing

Summary

1. Introduction to Data Mining and Knowledge Discovery
2. Data Preparation
3. Introduction to Prediction, Classification, Clustering and Association
4. Data Mining - From the Top 10 Algorithms to the New Challenges
5. Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation
6. Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning
7. Genetic Fuzzy Systems: State of the Art and New Trends
8. Some Advanced Topics I: Classification with Imbalanced Data Sets
9. Some Advanced Topics II: Subgroup Discovery
10. Some advanced Topics III: Data Complexity
11. Final talk: How must I Do my Experimental Study? Design of Experiments in Data Mining/Computational Intelligence. Using Non-parametric Tests. Some Cases of Study.