
COMP9318: Data Warehousing and Data Mining

— L3: Data Preprocessing and Data Cleaning —

Abridged from Prof. Jiawei Han's Slides

Chapter 3: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

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 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 Dirty?

- Incomplete data comes from
 - n/a data value when collected
 - different consideration between the time when the data was collected and when it is analyzed.
 - human/hardware/software problems
- Noisy data comes from the process of data
 - collection
 - entry
 - transmission
- Inconsistent data comes from
 - Different data sources
 - Functional dependency violation

Why Is Data Preprocessing Important?

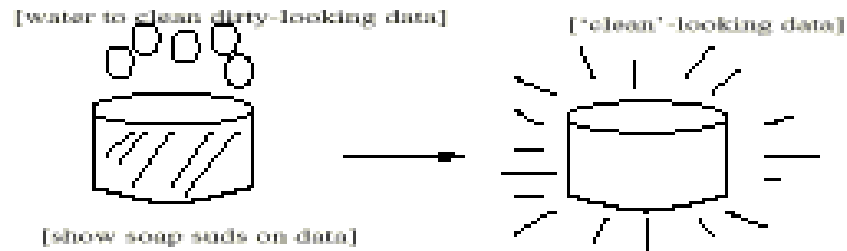
- No quality data, no quality mining results!
 - Quality decisions must be based on quality data
 - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
 - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the **majority** of the work of building a data warehouse. — Bill Inmon

Major Tasks in Data Preprocessing

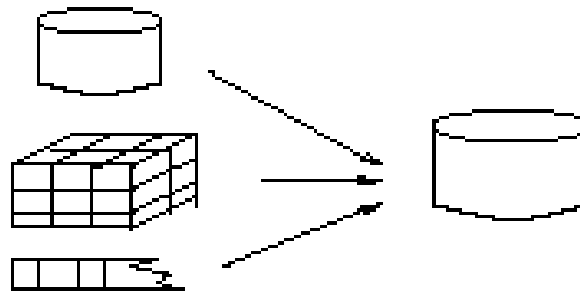
- Data cleaning
 - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
 - Integration of multiple databases, data cubes, or files
- Data transformation
 - Normalization and aggregation
- Data reduction
 - Obtains reduced representation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

Forms of data preprocessing

Data Cleaning



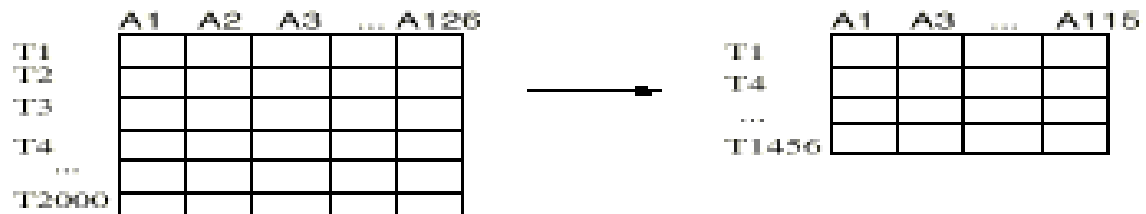
Data Integration



Data Transformation

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

Data Reduction



Chapter 3: Data Preprocessing

- Why preprocess the data?
- **Data cleaning**
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Cleaning

- **Importance**
 - “Data cleaning is one of the three biggest problems in data warehousing”—Ralph Kimball
 - “Data cleaning is the number one problem in data warehousing”—DCI survey
- Data cleaning tasks
 - Fill in missing values
 - Identify outliers and smooth out noisy data
 - Correct inconsistent data
 - Resolve redundancy caused by data integration

Missing Data

- Data is not always available
 - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- 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.

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
 - a global constant : e.g., “unknown”, a new class?!
 - the attribute mean
 - the attribute mean for all samples belonging to the same class: smarter
 - the most probable value: inference-based such as Bayesian formula or decision tree

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
 - 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

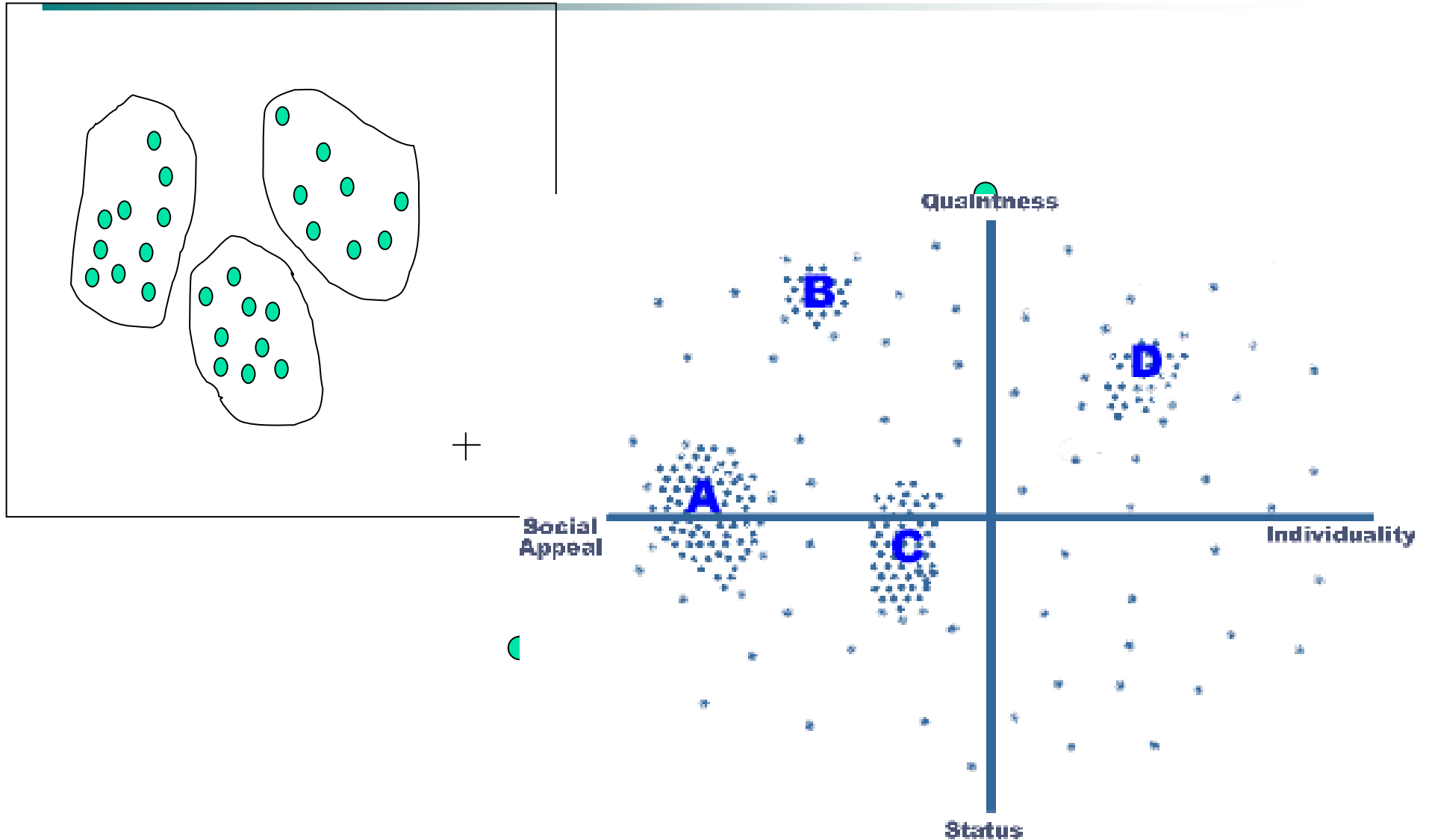
How to Handle Noisy Data?

- Binning method:
 - first sort data and partition into (equi-depth) bins
 - then one can **smooth by bin means, smooth by bin median, smooth by bin boundaries**, etc.
- Clustering
 - detect and remove outliers
- Combined computer and human inspection
 - detect suspicious values and check by human (e.g., deal with possible outliers)
- Regression
 - smooth by fitting the data into regression functions

Cluster Analysis

Application in Marketing:

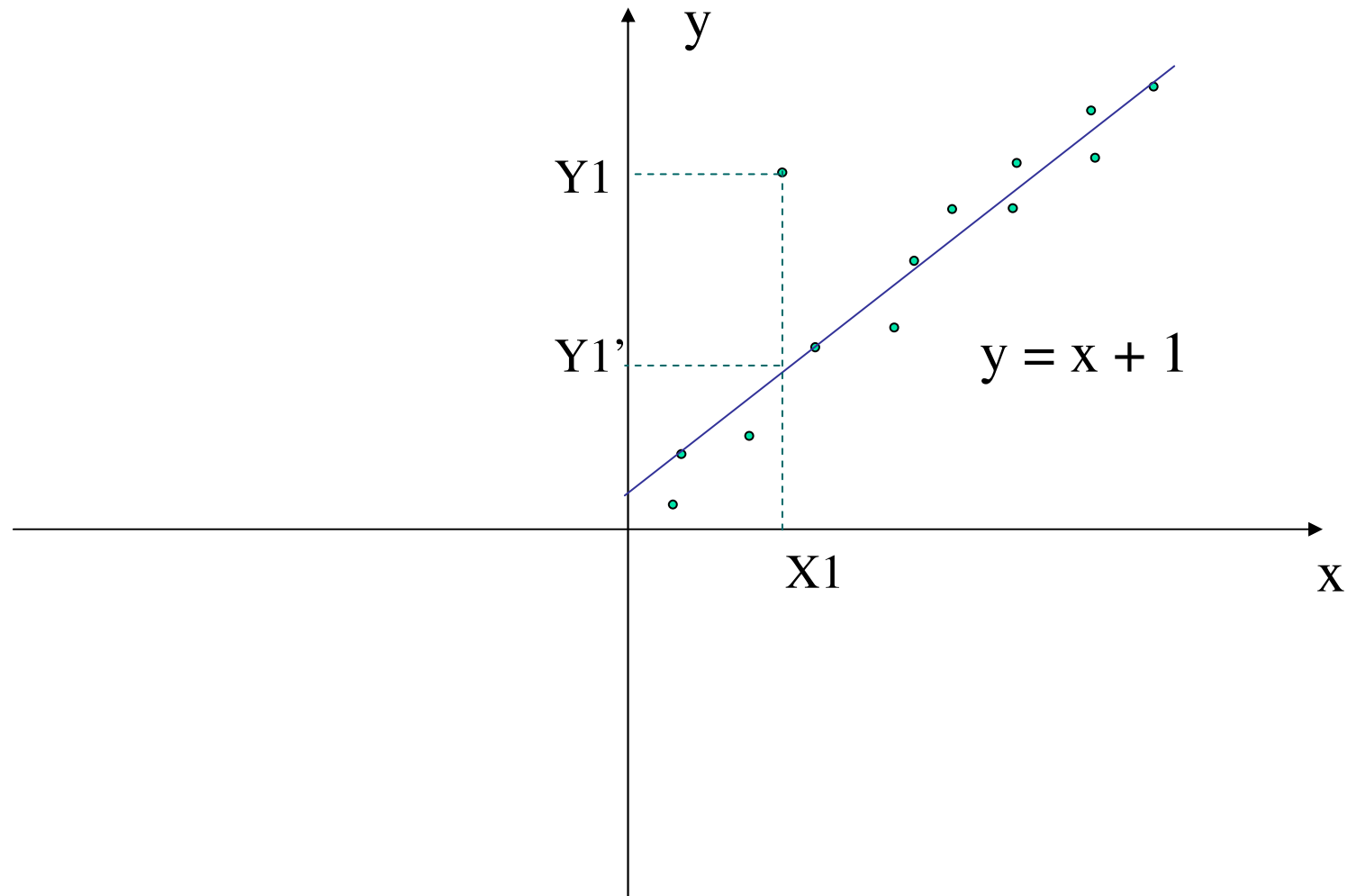
http://en.wikipedia.org/wiki/Cluster_analysis_%28in_marketing%29



Regression

Demo:

<http://www.wessa.net/esteq.wasp>



Simple Discretization Methods: Binning

- **Equal-width** (distance) partitioning:
 - Divides the range into N intervals of equal size:
uniform grid
 - if A and B are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A) / N$.
 - The most straightforward, but outliers may dominate presentation
 - Skewed data is not handled well.
- **Equal-depth** (frequency) partitioning:
 - Divides the range into N intervals, each containing approximately same number of samples
 - Good data scaling
 - Managing categorical attributes can be tricky.

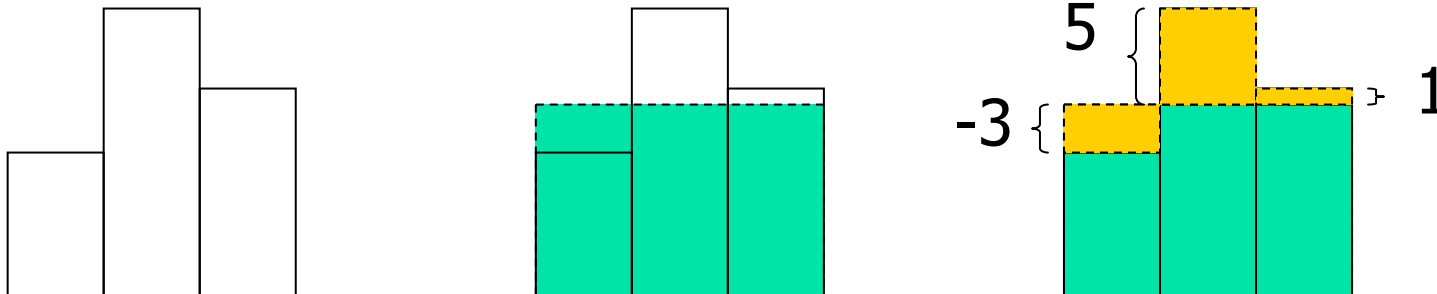
Binning Methods for Data Smoothing

- * Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- * Partition into (equi-depth) bins:
 - Bin 1: 4, 8, 9, 15
 - Bin 2: 21, 21, 24, 25
 - Bin 3: 26, 28, 29, 34
- * Smoothing by bin means:
 - Bin 1: 9, 9, 9, 9
 - Bin 2: 23, 23, 23, 23
 - Bin 3: 29, 29, 29, 29
- * Smoothing by bin boundaries:
 - Bin 1: 4, 4, 4, 15
 - Bin 2: 21, 21, 25, 25
 - Bin 3: 26, 26, 26, 34

Optimal Binning/Histogram

Not in the textbook

- Problem definition:
 - Given an array of n values, construct an optimal binning/histogram with B bins/buckets, such that some error measure is minimized.
- Error measure
 - SSE: Sum Square Error
 - $SSE = \sum_i (x_i - E(x_i))^2$
 - In the example, $SSE = ((-3)^2 + 5^2 + 1^2) = 35$



Dynamic Programming

$$\text{Since } \text{SSE}(x[1]..x[n]) = \text{SSE}(x[1]..x[i]) + \text{SSE}(x[i+1]..x[n])$$

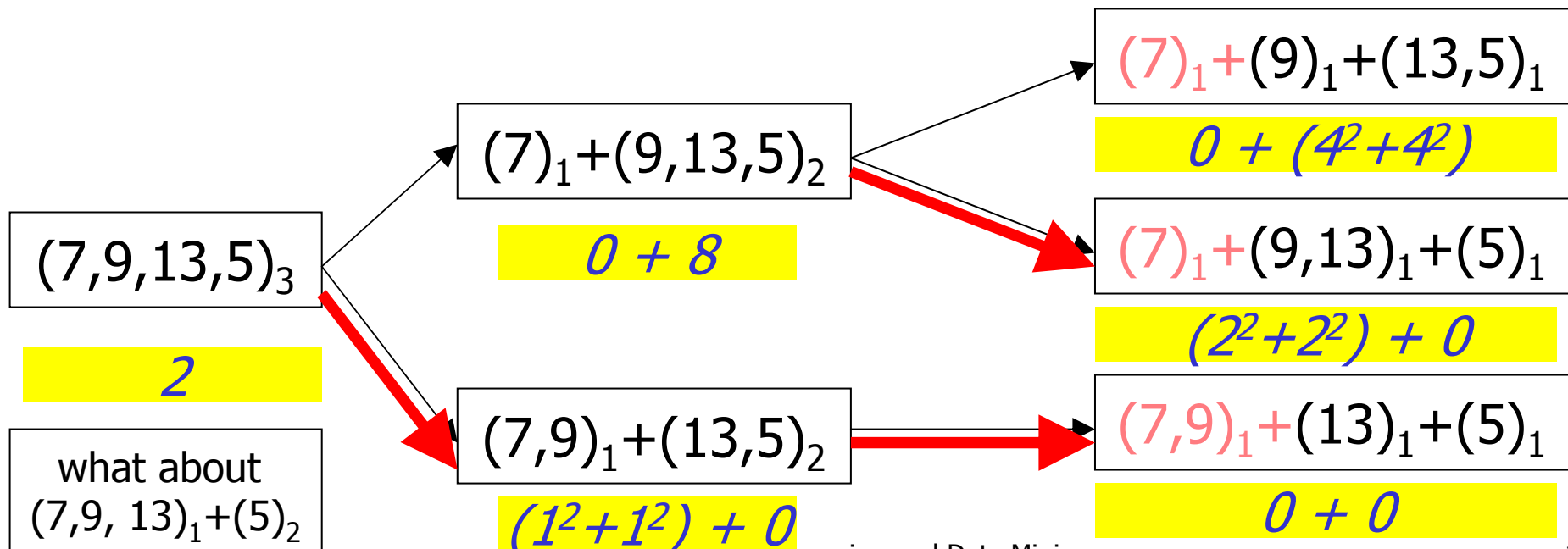
- Observation

- $\text{OPT}(x[1]..x[n], B) = \min_{1 <= i <= n} \{ \text{OPT}(x[1]..x[i], 1) + \text{OPT}(x[i+1]..x[n], B-1) \}$

- Example

- $n=4, B=3$

x[1]	x[2]	x[3]	x[4]
7	9	13	5



MaxDiff

- Complexity of the DP algorithm:
 - $O(n^2 * B)$ running time!
- Consider a heuristic method: MaxDiff
 - Idea: use the top-(B-1) max "gaps" in the data as the bin/bucket boundary

- Example:

$n=4, B=3$

x[1]	x[2]	x[3]	x[4]
7	9	13	5

gap(1-2)	gap(2-3)	gap(3-4)
2	4	8

$(7,9)_1$	$(13)_1$	$(5)_1$
-----------	----------	---------

Chapter 3: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Integration

- Data integration:
 - combines data from multiple sources into a coherent store
- Schema integration
 - integrate metadata from different sources
 - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id \equiv B.cust-#
- Detecting and resolving data value conflicts
 - for the same real world entity, attribute values from different sources are different
 - possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- 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 correlational analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

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

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- z-score normalization

$$v' = \frac{v - \text{mean}_A}{\text{stddev}_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \max(|v'|) < 1$$

Chapter 3: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- **Data reduction**
- Discretization and concept hierarchy generation
- Summary

Data Reduction Strategies

- A data 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
 - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- **Data reduction strategies**
 - **Dimensionality reduction**—remove unimportant attributes
 - **Data Compression**
 - **Numerosity reduction**—fit data into models
 - **Discretization and concept hierarchy generation**

Another method is combine features into new features. c.f., PCA, LSI

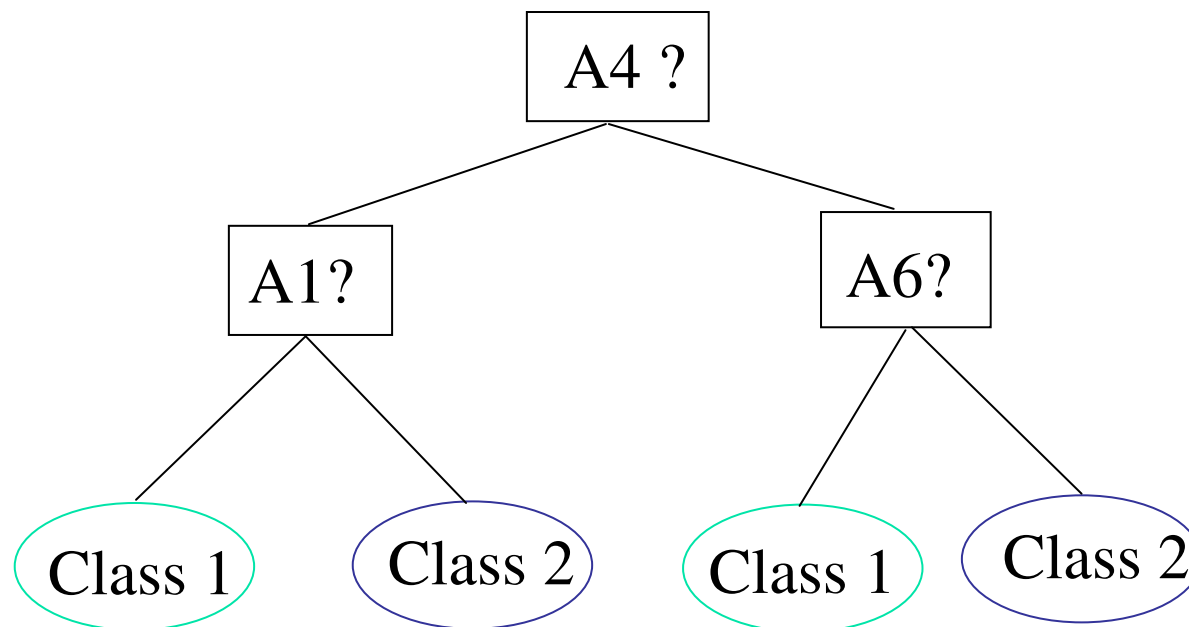
Dimensionality Reduction

- **Feature** selection (i.e., attribute subset selection):
 - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
 - reduce # of patterns in the patterns, easier to understand
- Heuristic methods (due to exponential # of choices):
 - step-wise forward selection
 - step-wise backward elimination
 - combining forward selection and backward elimination
 - decision-tree induction

Example of Decision Tree Induction

Initial attribute set:

{A1, A2, A3, A4, A5, A6}



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

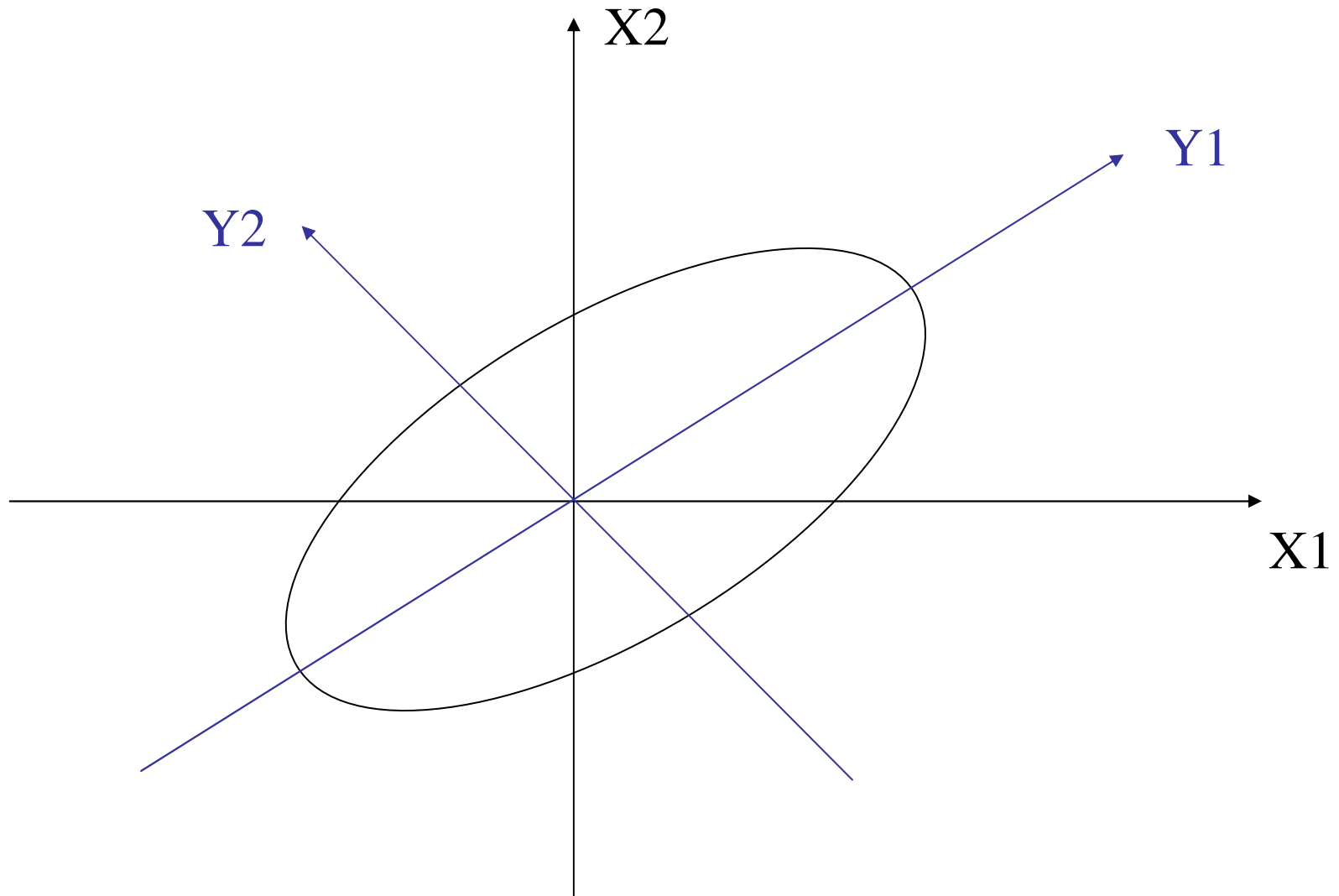
Heuristic Feature Selection Methods

- 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

Principal Component Analysis

- Given N data vectors from k -dimensions, find $c \leq k$ orthogonal vectors that can be best used to represent data
 - The original data set is reduced to one consisting of N data vectors on c principal components (reduced dimensions)
- Each data vector is a linear combination of the c principal component vectors
- Works for numeric data only
- Used when the number of dimensions is large

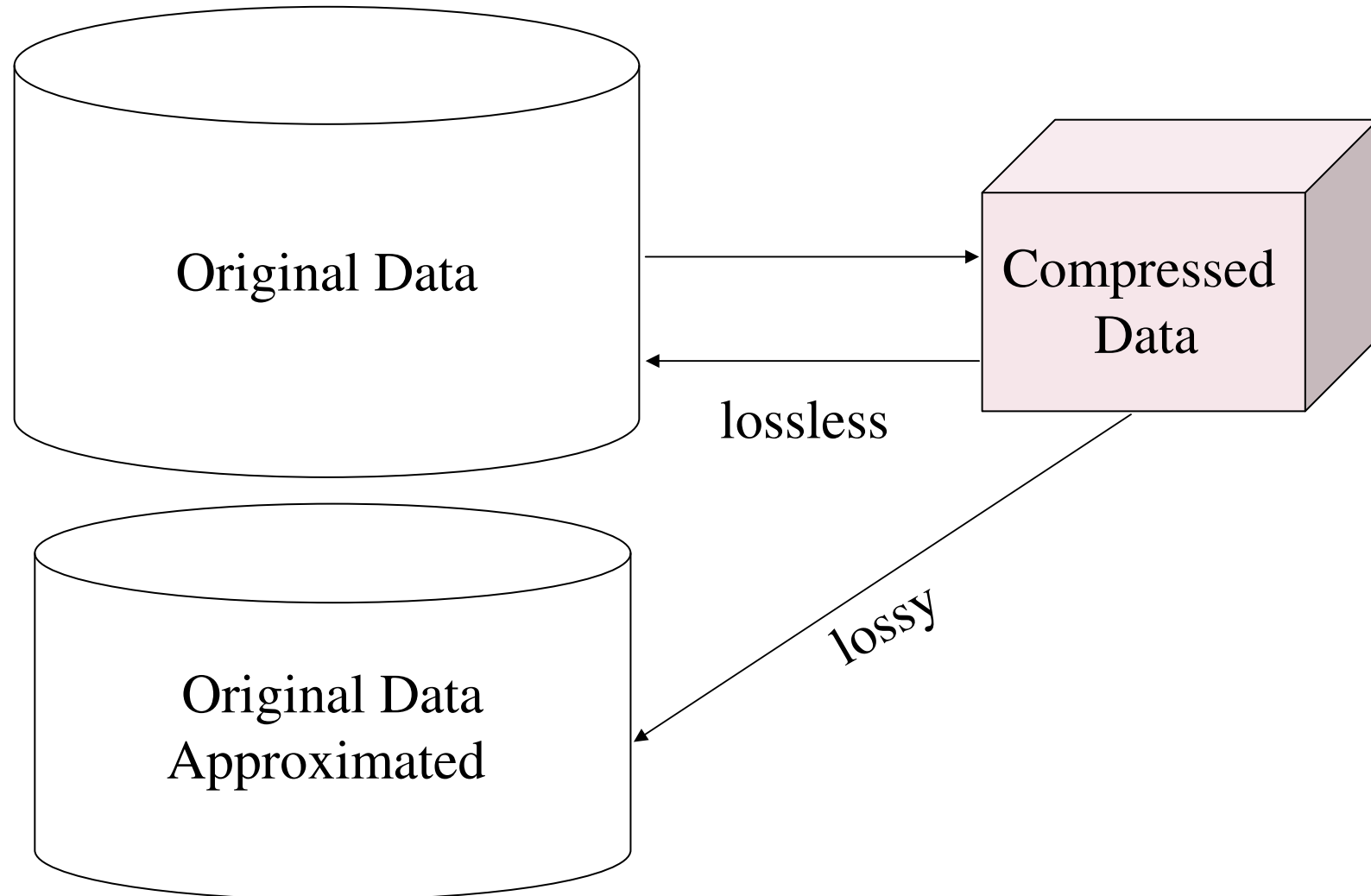
Principal Component Analysis



Data Compression

- String compression
 - There are extensive theories and well-tuned algorithms
 - Typically lossless
 - But only limited manipulation is possible without expansion
- Audio/video compression
 - Typically lossy compression, with progressive refinement
 - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
 - Typically short and vary slowly with time

Data Compression



Numerosity Reduction

- Parametric methods
 - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
 - Log-linear models: obtain value at a point in m-D space as the product on appropriate marginal subspaces
- Non-parametric methods
 - Do not assume models
 - Major families: histograms, clustering, sampling

Regression and Log-Linear Models

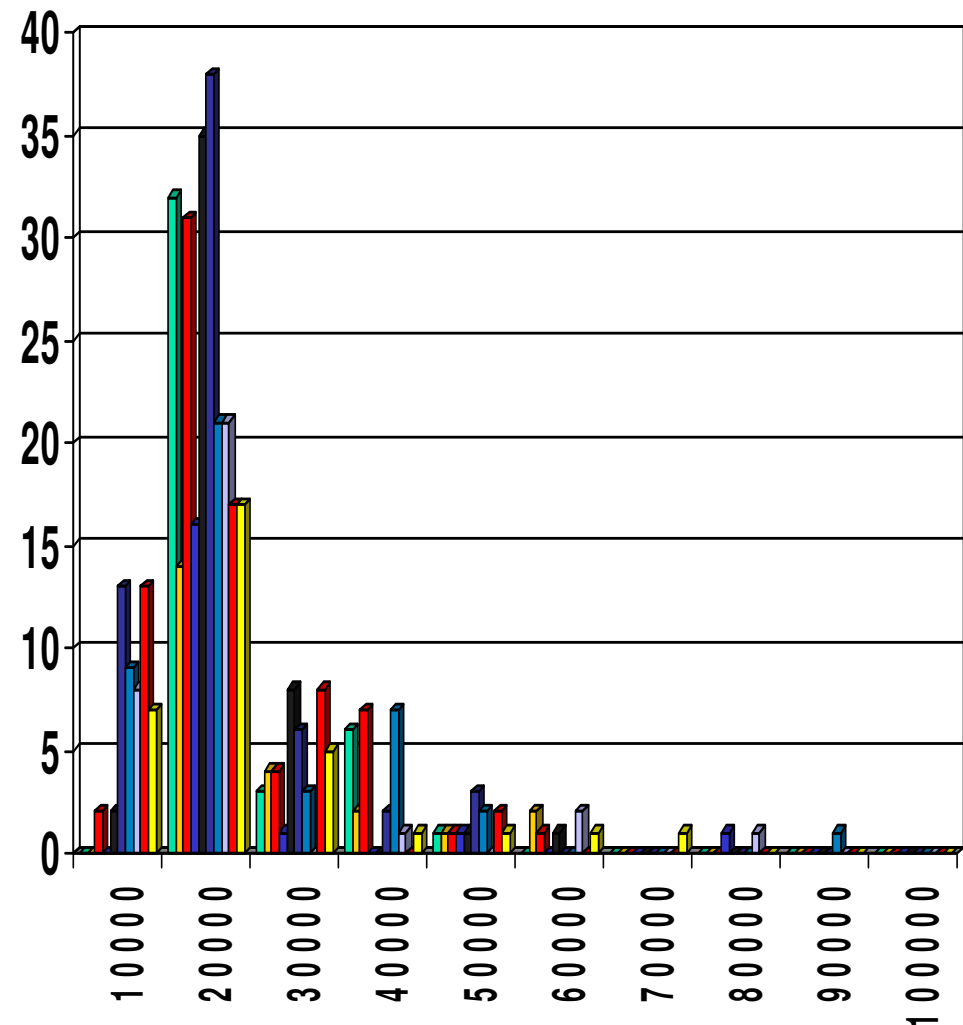
- Linear regression: Data are modeled to fit a straight line
 - Often uses the least-square method to fit the line
- Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model: approximates discrete multidimensional probability distributions

Regress Analysis and Log-Linear Models

- Linear regression: $Y = \alpha + \beta X$
 - Two parameters , α and β specify the line and are to be estimated by using the data at hand.
 - using the least squares criterion to the known values of $Y_1, Y_2, \dots, X_1, X_2, \dots$
- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$.
 - Many nonlinear functions can be transformed into the above.
- Log-linear models:
 - The multi-way table of joint probabilities is approximated by a product of lower-order tables.
 - Probability: $p(a, b, c, d) = \alpha_{ab} + \beta_{ac} + \chi_{ad} + \delta_{bcd}$

Histograms

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Can be constructed optimally in one dimension using dynamic programming
- Related to quantization problems.



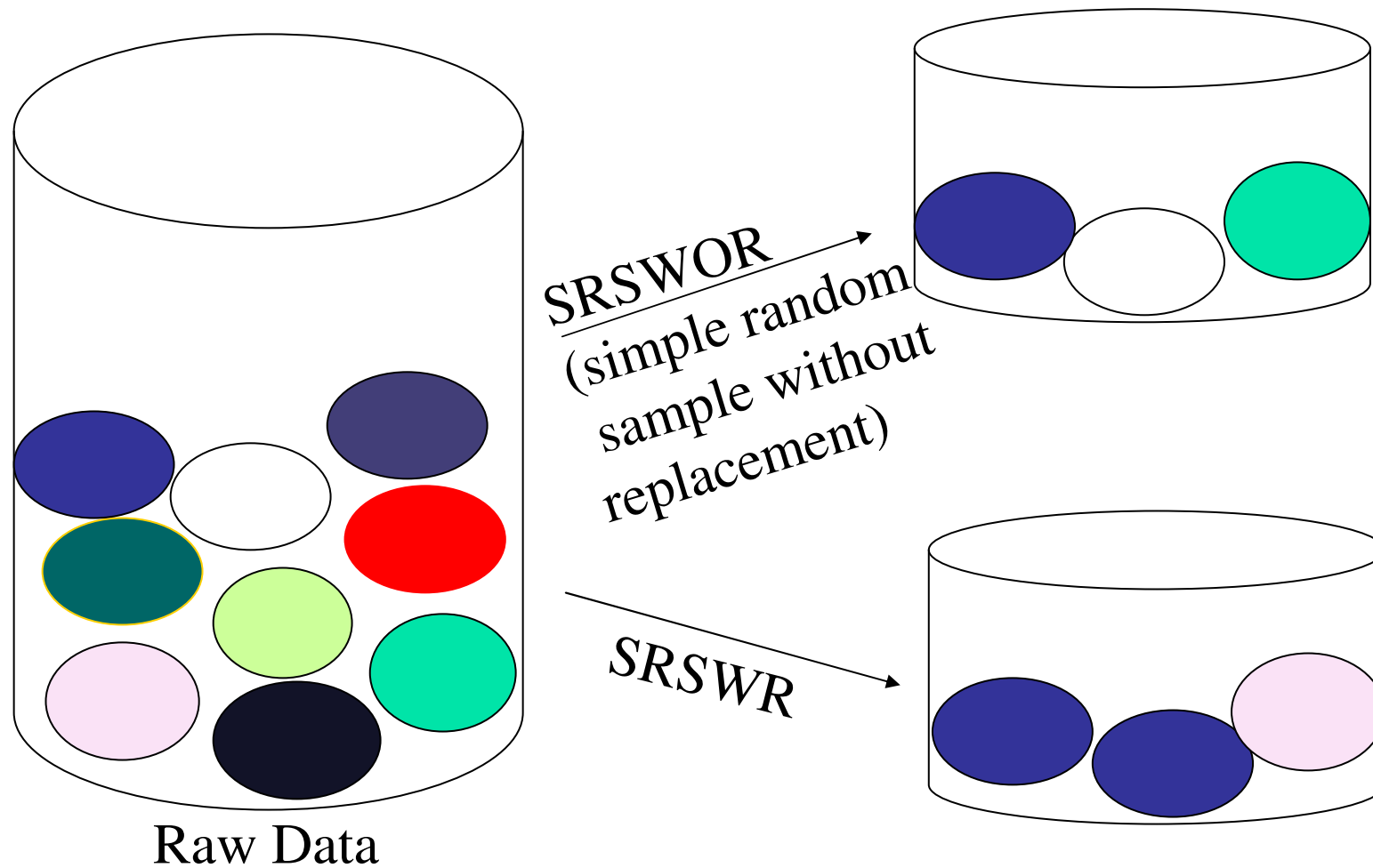
Clustering

- Partition data set into clusters, and one can store cluster representation only
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms, further detailed in Chapter 8

Sampling

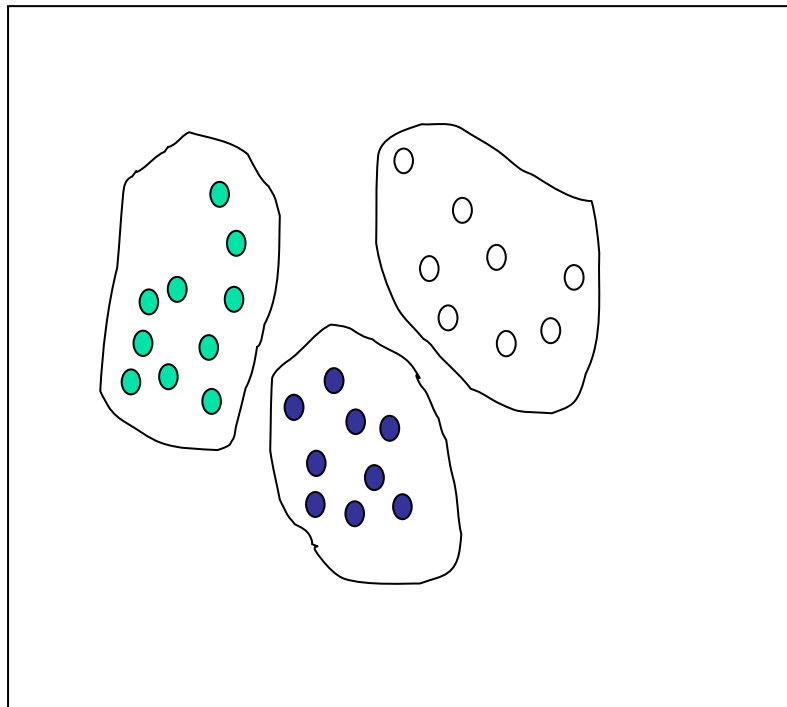
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a **representative** subset of the data
 - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
 - Stratified sampling:
 - Approximate the percentage of each class (or subpopulation of interest) in the overall database
 - Used in conjunction with skewed data
- Sampling may not reduce database I/Os (page at a time).

Sampling

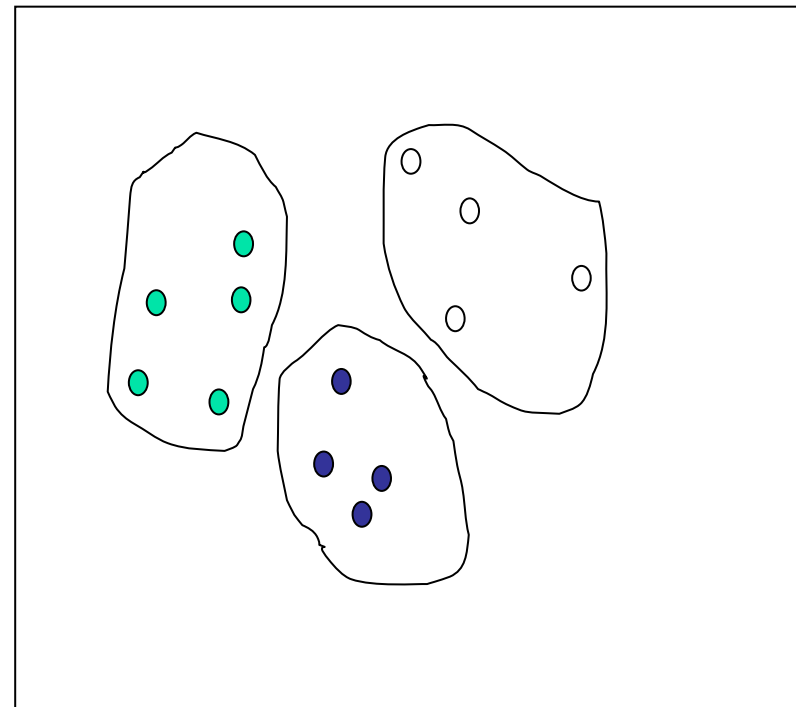


Sampling

Raw Data



Cluster/Stratified Sample



Chapter 3: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- **Discretization**
- Summary

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.
 - Reduce data size by discretization
 - Prepare for further analysis

Discretization and Concept hierachy

- Discretization

- reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values

- Concept hierarchies

- reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior)

Discretization and Concept Hierarchy Generation for Numeric Data

- Binning (see sections before)
- Histogram analysis (see sections before)
- Clustering analysis (see sections before)
- Entropy-based discretization
- Segmentation by natural partitioning

Entropy-Based Discretization

- Given a set of samples S , if S is partitioned into two intervals S_1 and S_2 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.,

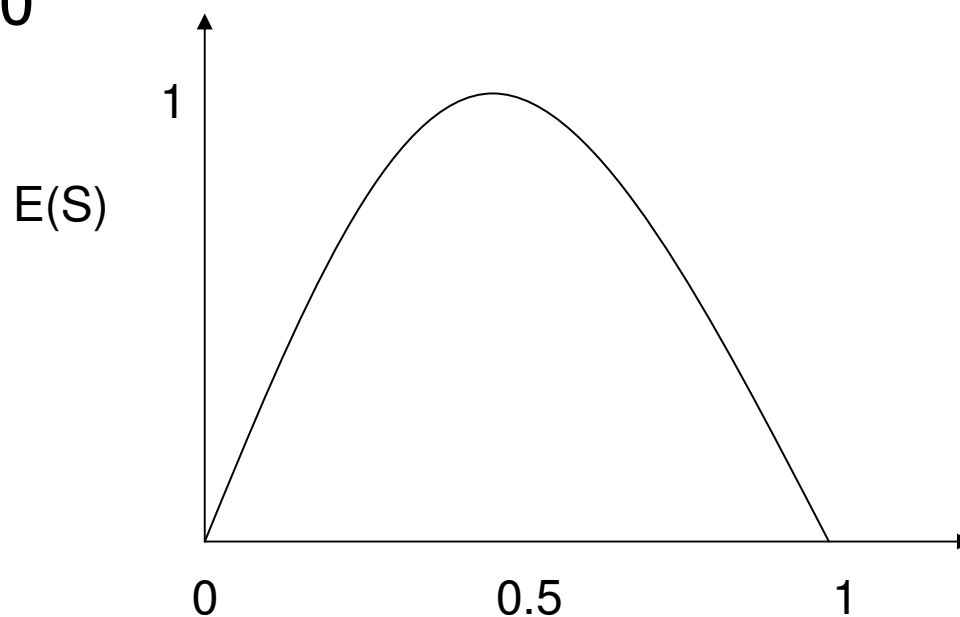
$$\text{before } Ent(S) - E(T, S) > \delta \text{ after}$$

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

Entropy

- Give a set S of instances with binary classes $\{+, -\}$. Say proportions of $+$ and $-$ are p_+ and p_- respectively.
- Then Entropy of S is defined as

$$\text{Ent}(S) = - (p_+ \log_2 p_+ + p_- \log_2 p_-) \quad \text{-- assuming } \log 0 = 0$$



From information theory, number of bits to encode the class label.

Can be generalized to Multi-class.

Chapter 3: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- **Summary**

Summary

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

References

- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*. Vol.23, No.4
- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. *Communications of ACM*, 42:73-78, 1999.
- H.V. Jagadish et al., Special Issue on Data Reduction Techniques. *Bulletin of the Technical Committee on Data Engineering*, 20(4), December 1997.
- A. Maydanchik, Challenges of Efficient Data Cleansing (DM Review - Data Quality resource portal)
- D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.
- D. Quass. A Framework for research in Data Cleaning. (Draft 1999)
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001.
- T. Redman. *Data Quality: Management and Technology*. Bantam Books, New York, 1992.
- Y. Wand and R. Wang. Anchoring data quality dimensions ontological foundations. *Communications of ACM*, 39:86-95, 1996.
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engineering*, 7:623-640, 1995.
- <http://www.cs.ucla.edu/classes/spring01/cs240b/notes/data-integration1.pdf>