Data Preparation for Data Mining

Lesson 2

Lesson 2 Overview

- Data Preparation as a Process
  - Inputs, Outputs, Models and Decisions
  - Modelling Tools and Data Preparation
  - Stages of Data Preparation
  - Overview of Basic Data Preparation Techniques

The Right Context

- Data exploration
- Identifying the problem to solve
- Defining the solution

Why Data Preprocessing?

- Data in the real world is far from clean
  - 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

Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility

The Process of Data Preparation

- Purpose: make the data better accessible for the mining tool
- No magical general purpose techniques, preparation is half art, half science
- Knowing the limitations and correct use is more important than thoroughly understanding the actual data preparation techniques

Inputs, Outputs and the In-Between

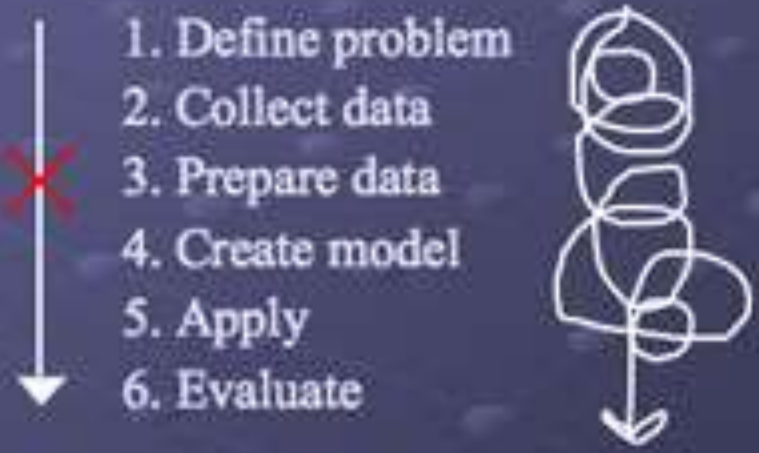
- Inputs:
  - raw data
  - decisions (problem, solutions, modeling tools, confidence limits)
- Outputs:
  - two data sets
  - PIE (Prepared Information Environment) modules

The In-Between

- Inputs and outputs first, then the middle
- How modeling tools affect what is done
- The stages of data preparation
- What needs to be decided at each stage

Modeling in data mining

Modeling is iterative:



1. Define problem
2. Collect data
3. Prepare data
4. Create model
5. Apply
6. Evaluate

Ten golden rules

1. Select clear problem with tangible benefit
2. Specify required solution
3. Define how solution is implemented
4. Understand the domain
6. Stipulate assumptions

5. Let the problem drive the modeling
7. Refine the model iteratively
8. Make the model as simple as possible (but no simpler)
9. Find areas of instability
10. Find areas of uncertainty

Data Mining Process (simplified)

1. Data Preparation

2. Data Survey

3. Data Modeling

Data Mining Process (simplified)

1. Data Preparation

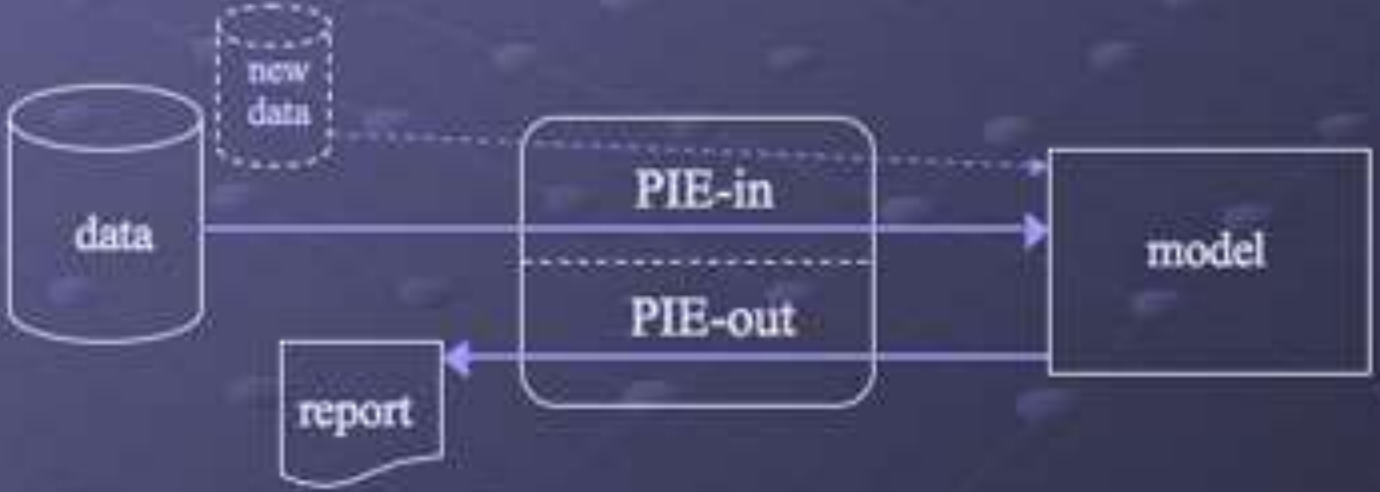
2. Data Survey

3. Data Modeling

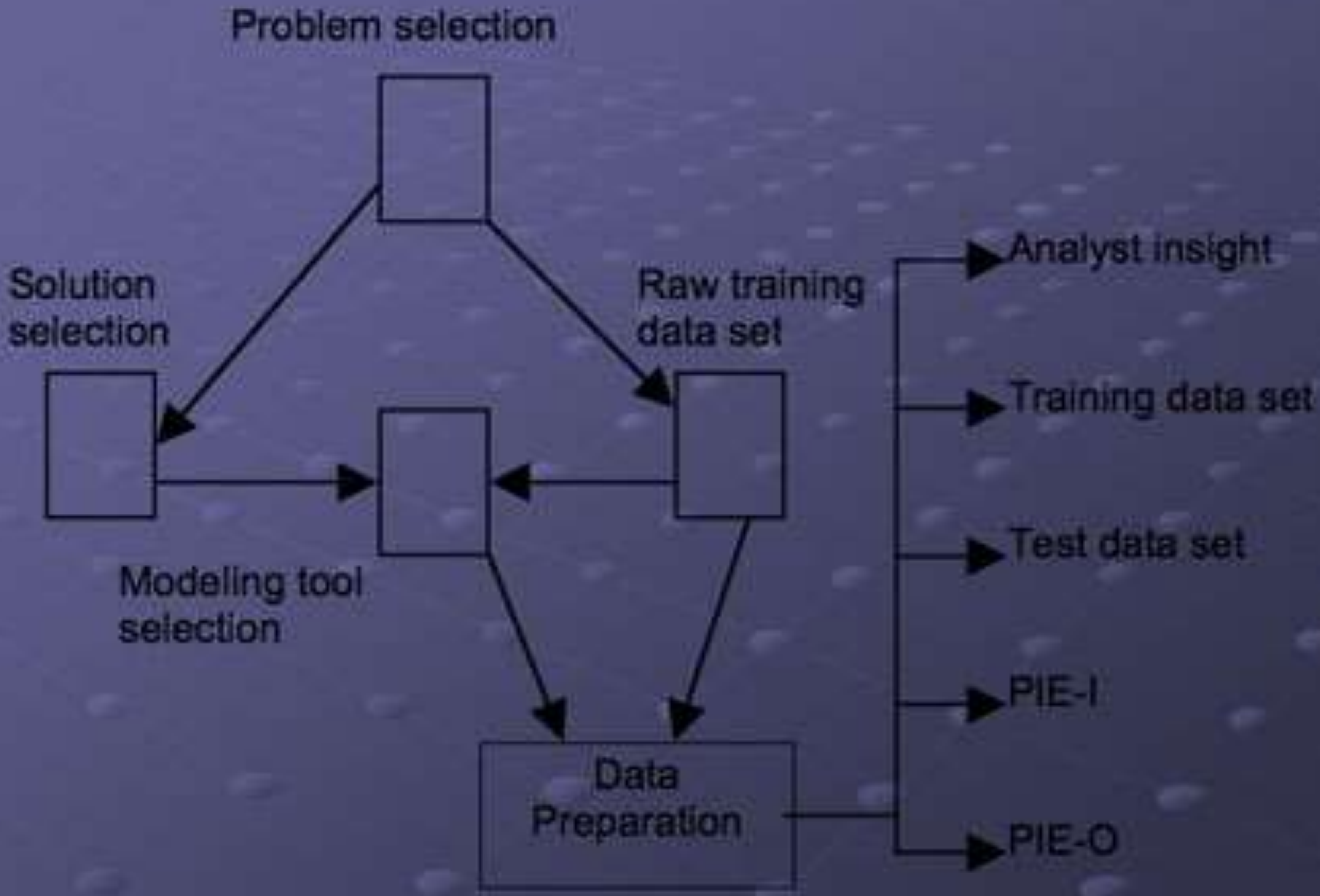
PIE

Prepared Information Environment

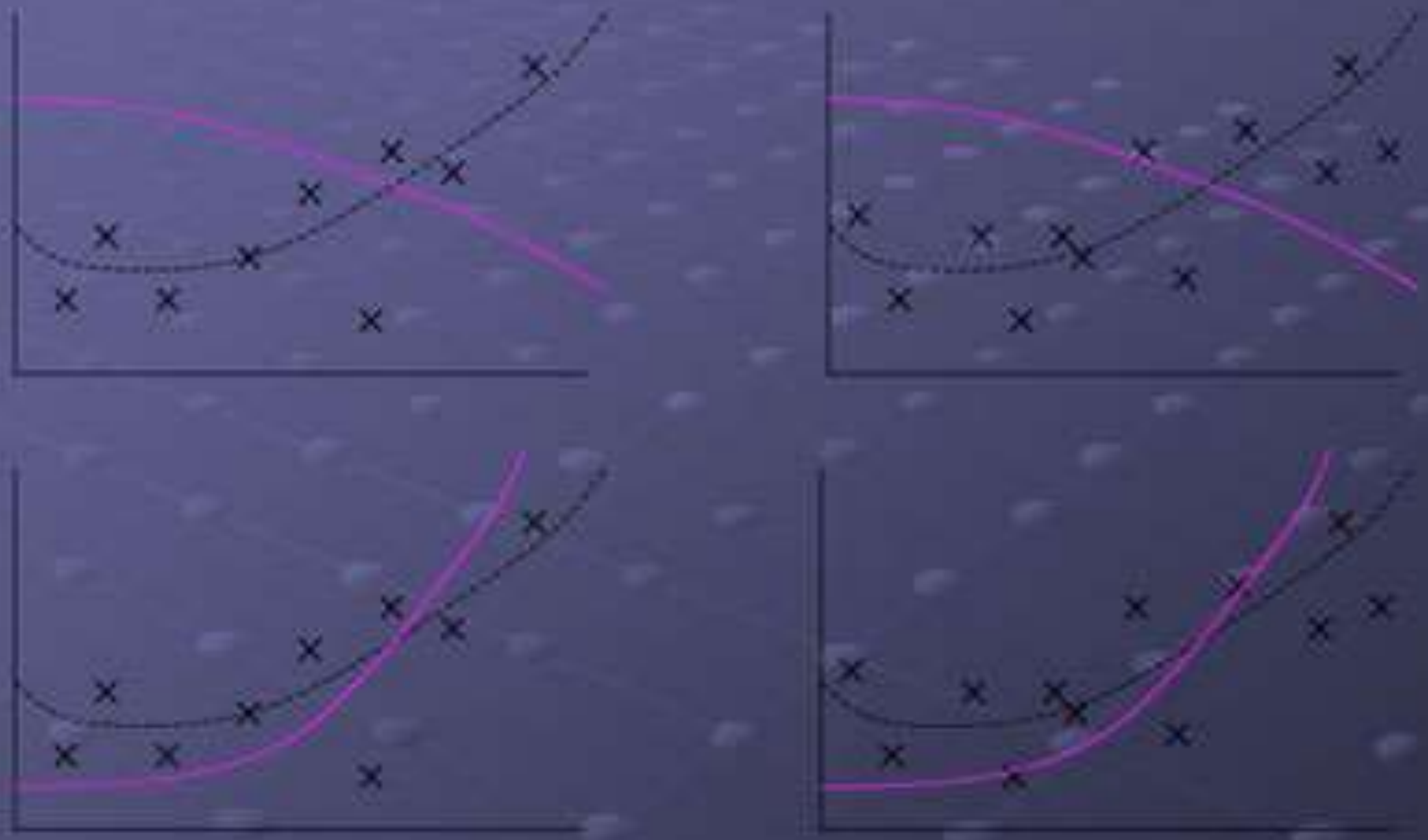
1. prepare the training/testing data
2. transform prepared values to original
3. apply the <u>same</u> preparation to new data

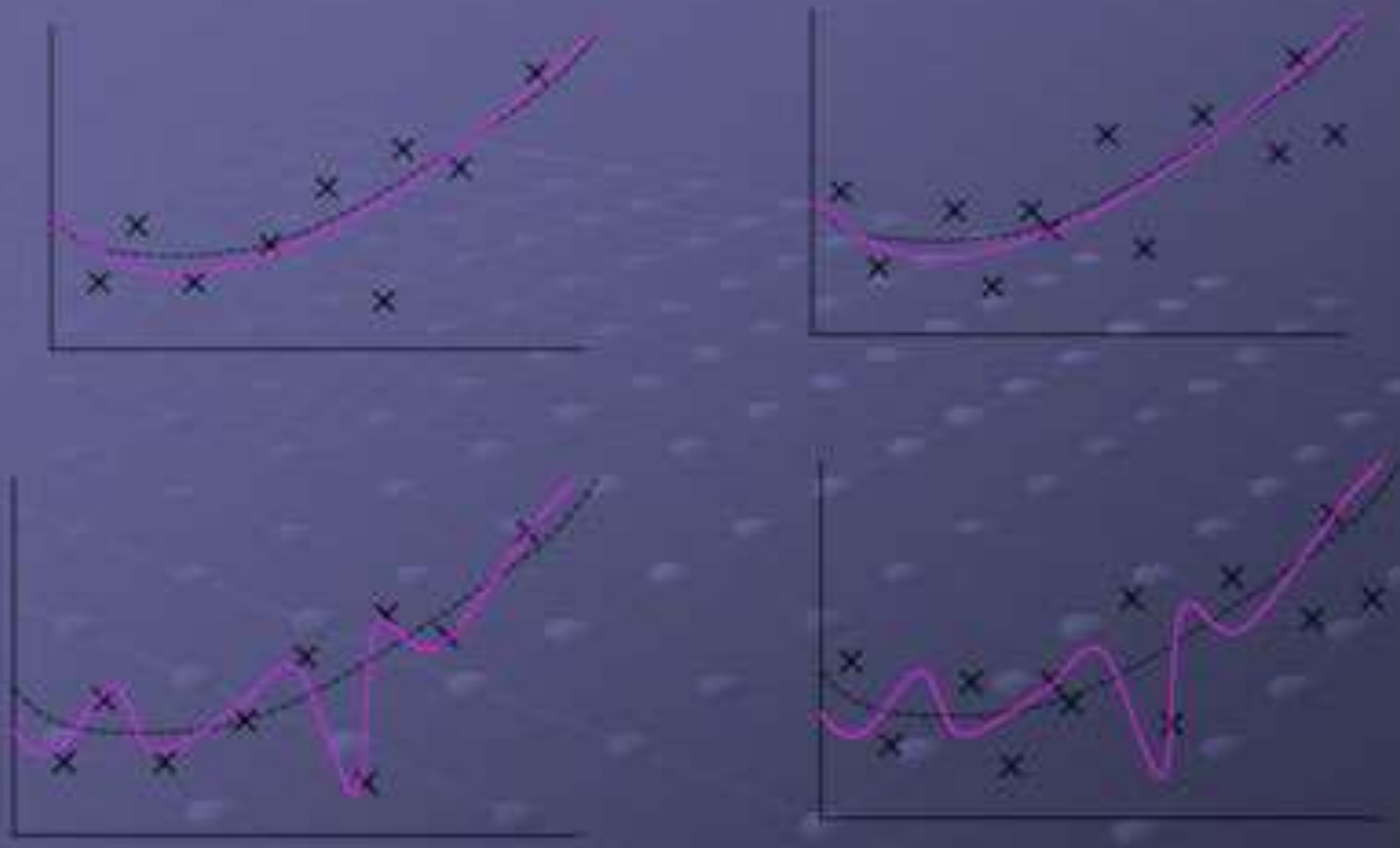


Data Preparation Process



Training and Test Data Sets





Prepared Information Environment Modules

- Input module transforms raw execution data:
  - categorical values into numerical
  - filling in / ignoring missing values
- Output module undoes the effect of PIE-I
- Used between the model and the real world

Data Mining Process (simplified)

1. Data Preparation

2. Data Survey

3. Data Modeling

Why survey?

Get a broad idea of the data:

- what is covered
- what is not covered, or is covered poorly

Dangerous areas:

- bias in data
- sparse data (in a dynamic area)

Is the data adequate?

Data Mining Process (simplified)

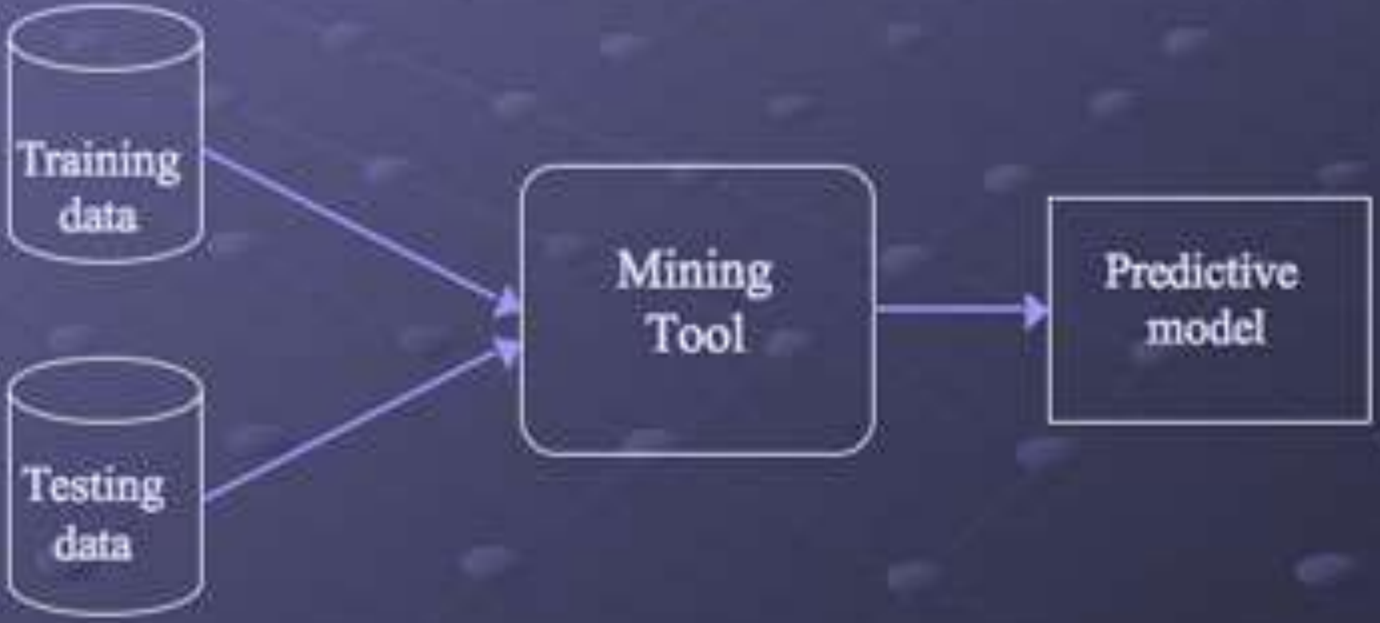
1. Data Preparation

2. Data Survey

3. Data Modeling

Mining for the Model

- High level view of the predictive modeling
- The two data sets are fed into the mining tool
- Output is the predictive model

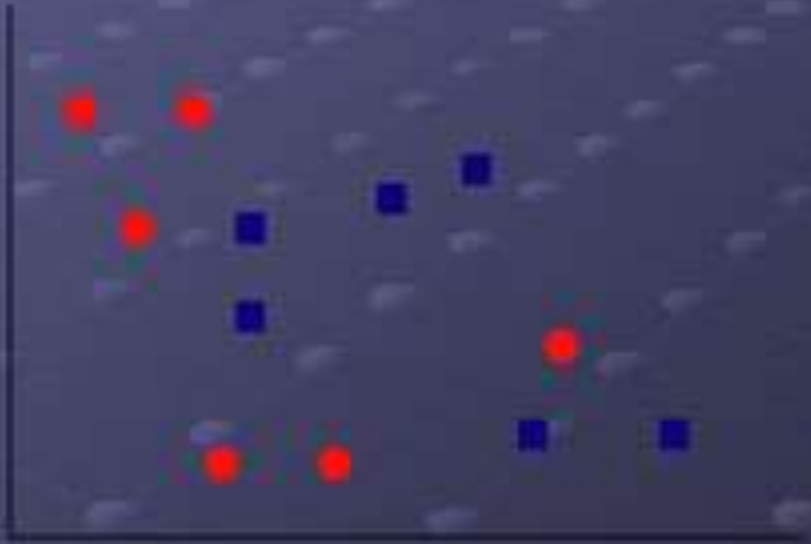


Modeling Tools and Data Preparation

- Right tool for the right job
- Early general-purpose mining tools were algorithm centric
- Modern tools concentrate on business problems
- Knowledge about tool essential!

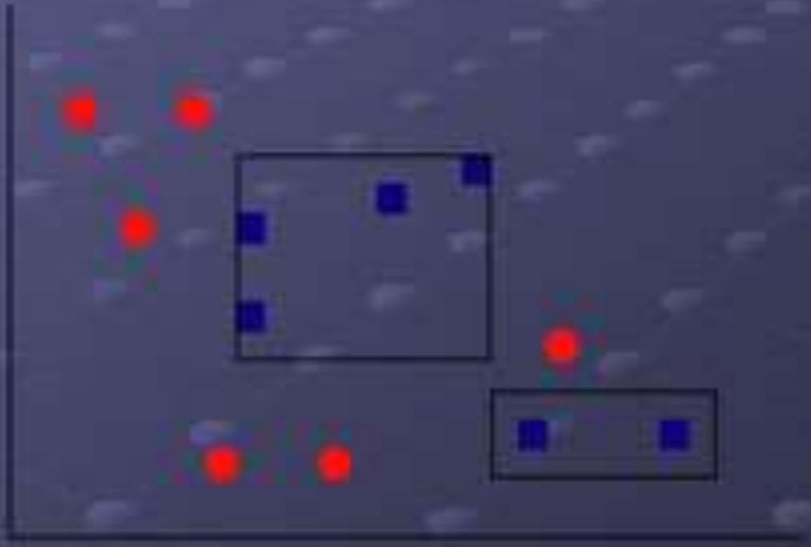
How Modeling Tools Influence Preparation: Data Separation

- Straight lines parallel to axes
- Straight lines not parallel to axes
- Curves
- Closed area
- Ideal arrangement



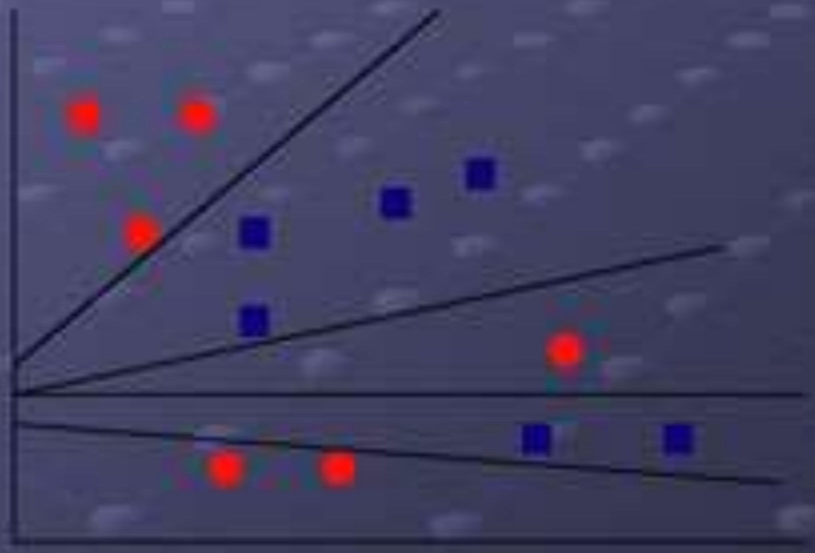
Data Separation

- Straight lines parallel to axes
- Straight lines not parallel to axes
- Curves
- Closed area
- Ideal arrangement



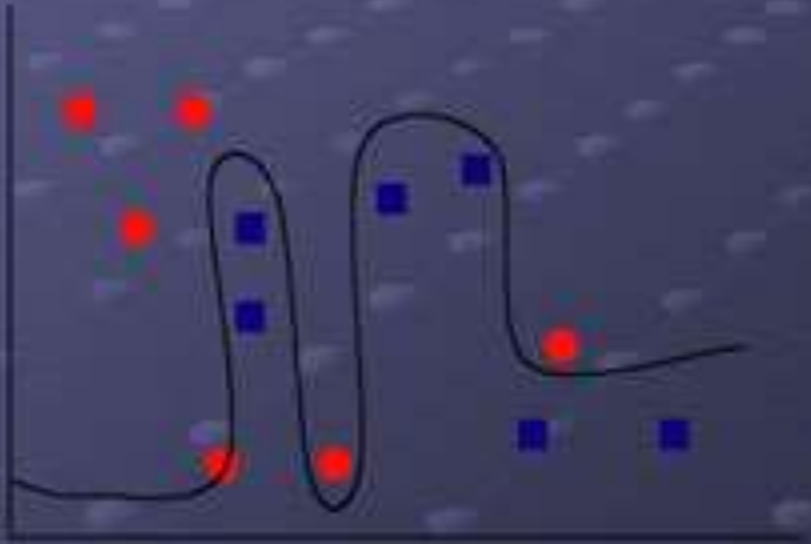
Data Separation

- Straight lines parallel to axes
- Straight lines not parallel to axes
- Curves
- Closed area
- Ideal arrangement



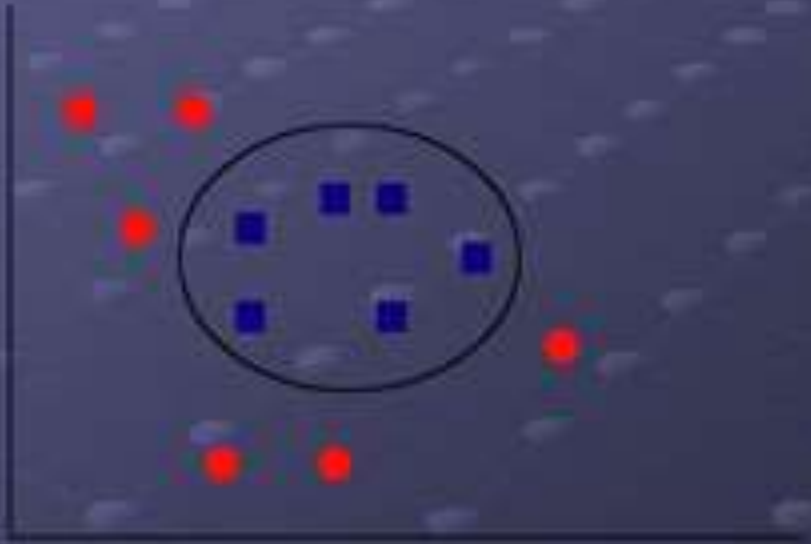
Data Separation

- Straight lines parallel to axes
- Straight lines not parallel to axes
- Curves
- Closed area
- Ideal arrangement



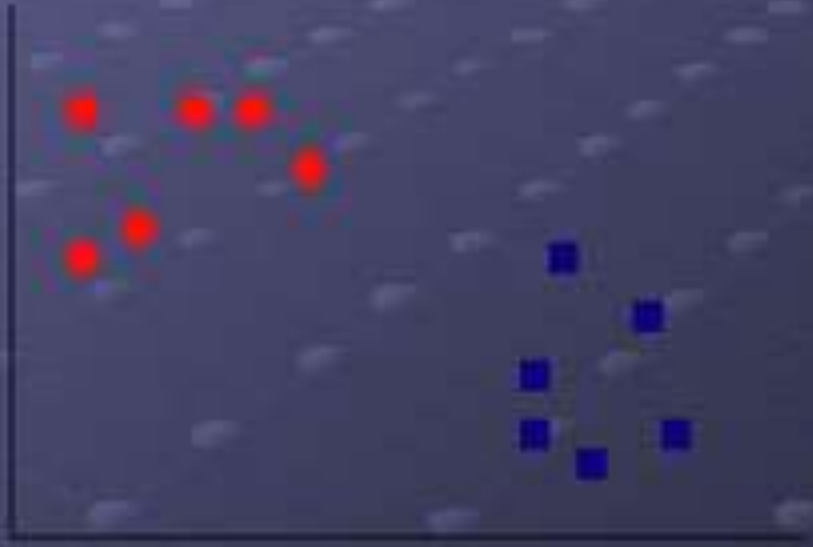
Data Separation

- Straight lines parallel to axes
- Straight lines not parallel to axes
- Curves
- Closed area
- Ideal arrangement



Data Separation

- Straight lines parallel to axes
- Straight lines not parallel to axes
- Curves
- Closed area
- Ideal arrangement

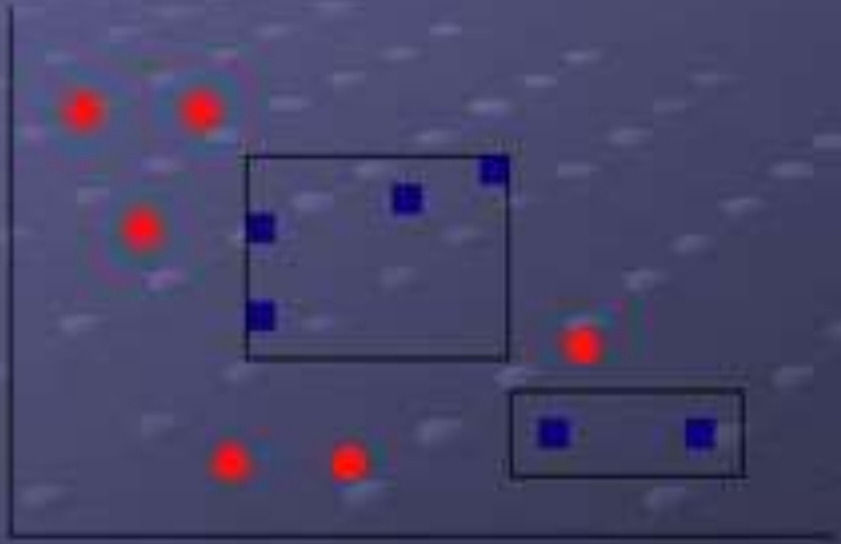


Algorithms for Data Separation

- Decision Trees
- Decision Lists
- Neural Networks
- Evolution Programs

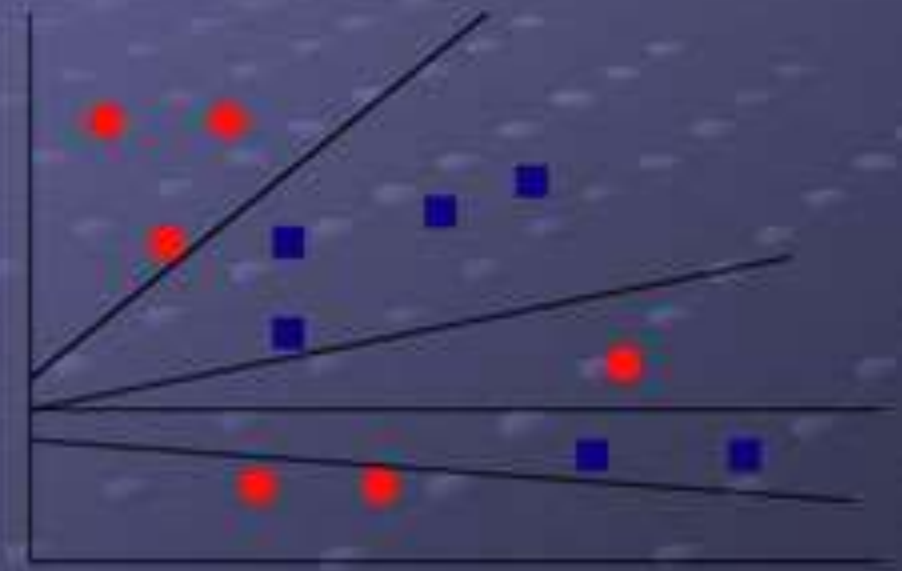
Algorithms for Data Separation

- Decision Trees
- Decision Lists
- Neural Networks
- Evolution Programs



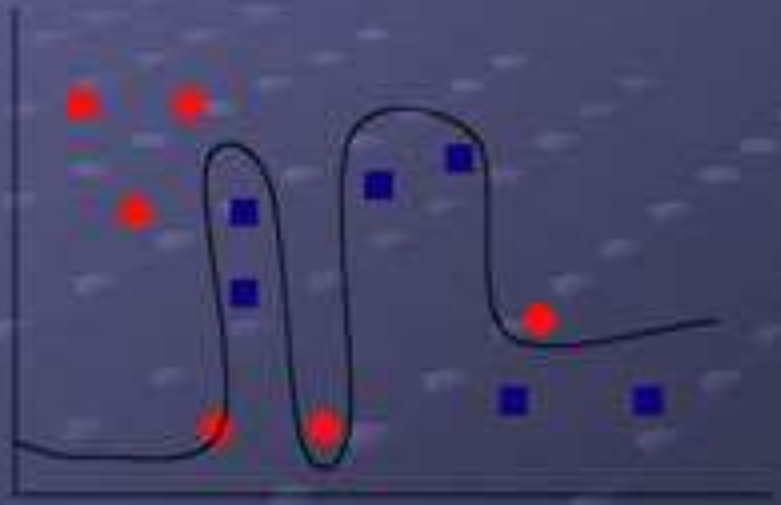
Algorithms for Data Separation

- Decision Trees
- Decision Lists
- Neural Networks
- Evolution Programs



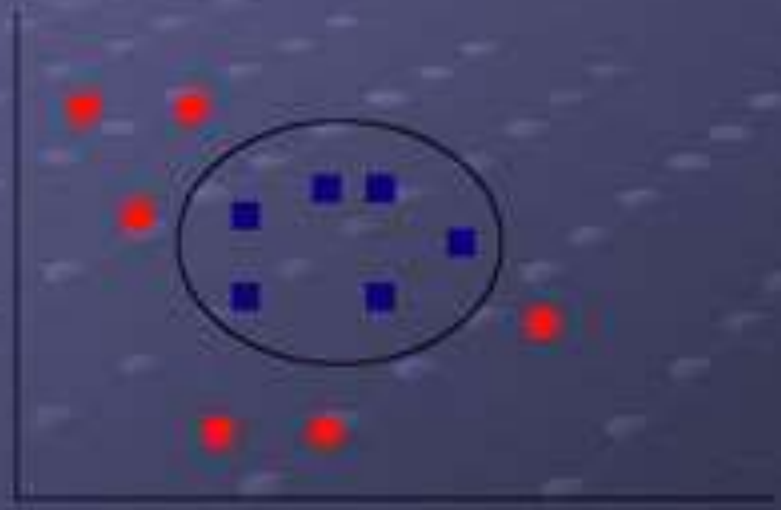
Algorithms for Data Separation

- Decision Trees
- Decision Lists
- Neural Networks
- Evolution Programs



Algorithms for Data Separation

- Decision Trees
- Decision Lists
- Neural Networks
- Evolution Programs



Modeling Data with the Tools

- Nominal and numeric tools - different approaches to different problems
- Binning vs. continuos algorithms
- It may be worthwhile trying different techniques for preparation
- Missing and empty values

Stages of Data Preparation

- Accessing the data
  - not trivial in many cases!
  - Very case dependent

Stages of Data Preparation

- Accessing the data
- Auditing the data
  - examining the quality, quantity and source of data
  - make sure the minimum requirements for solution are filled, forget unsupported hopes

Stages of Data Preparation

- Accessing the data
- Auditing the data
- Enhancing and enriching the data
  - add more data if needed
  - apply domain knowledge to ease the work of the tool

Stages of Data Preparation

- Accessing the data
- Auditing the data
- Enhancing and enriching the data
- Looking for sampling bias
  - data sets must accurately represent the population
  - failure may lead to useless models

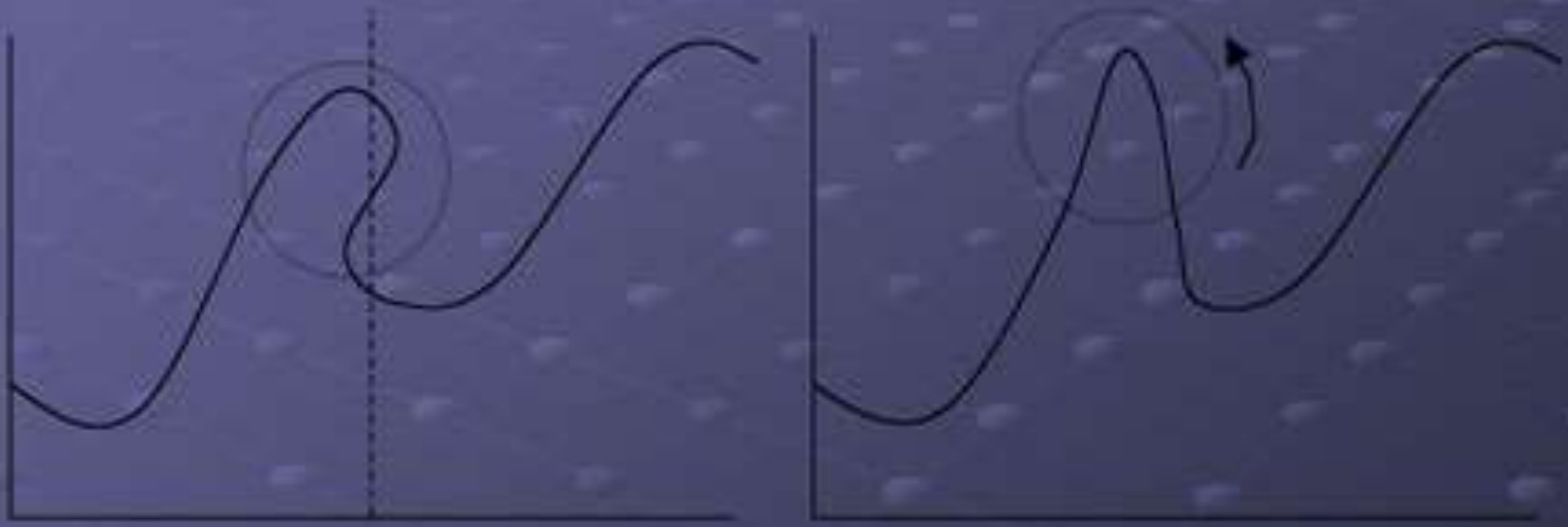
Stages of Data Preparation

- Accessing the data
- Auditing the data
- Enhancing and enriching the data
- Looking for sampling bias
- Determining data structure
  - superstructure: selected scaffolding
  - macrostructure: e.g. granularity
  - microstructure: relationships between variables

Stages of Data Preparation

- Building the PIE, data issues:
  - representative samples
  - categorical values
  - normalization
  - missing and empty values
  - reducing width and depth
  - well- and ill-formed manifolds

Correcting Problems with Ill-Formed Manifolds



Stages of Data Preparation

- Accessing the data
- Auditing the data
- Enhancing and enriching the data
- Looking for sampling bias
- Determining data structure
- Building the PIE
- Surveying the Data
- Modeling the Data

Summary

- Some data preparation is needed for all mining tools
- The purpose of preparation is to transform data sets so that their information content is best exposed to the mining tool
- Error prediction rate should be lower (or the same) after the preparation as before it
- The miner gains very good insight on the problem during the preparation process

Overview of Basic Data Preparation Techniques

- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Overview of Basic Data Preparation Techniques

- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Cleaning

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

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 Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing 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 (equal-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
- Regression
  - smooth by fitting the data into regression functions

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning:
  - It 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:
  - It 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

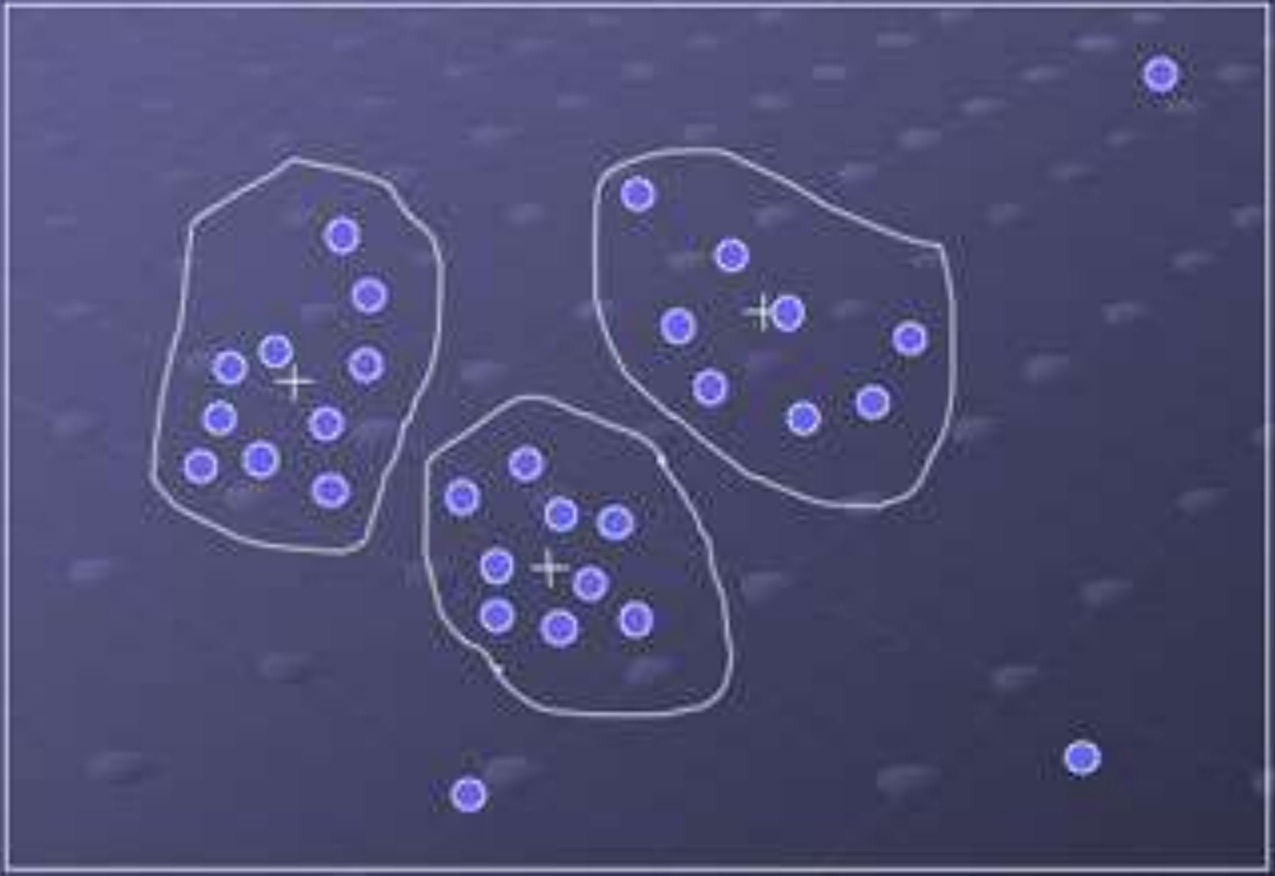
- Sorted data for price (in dollars):

  4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

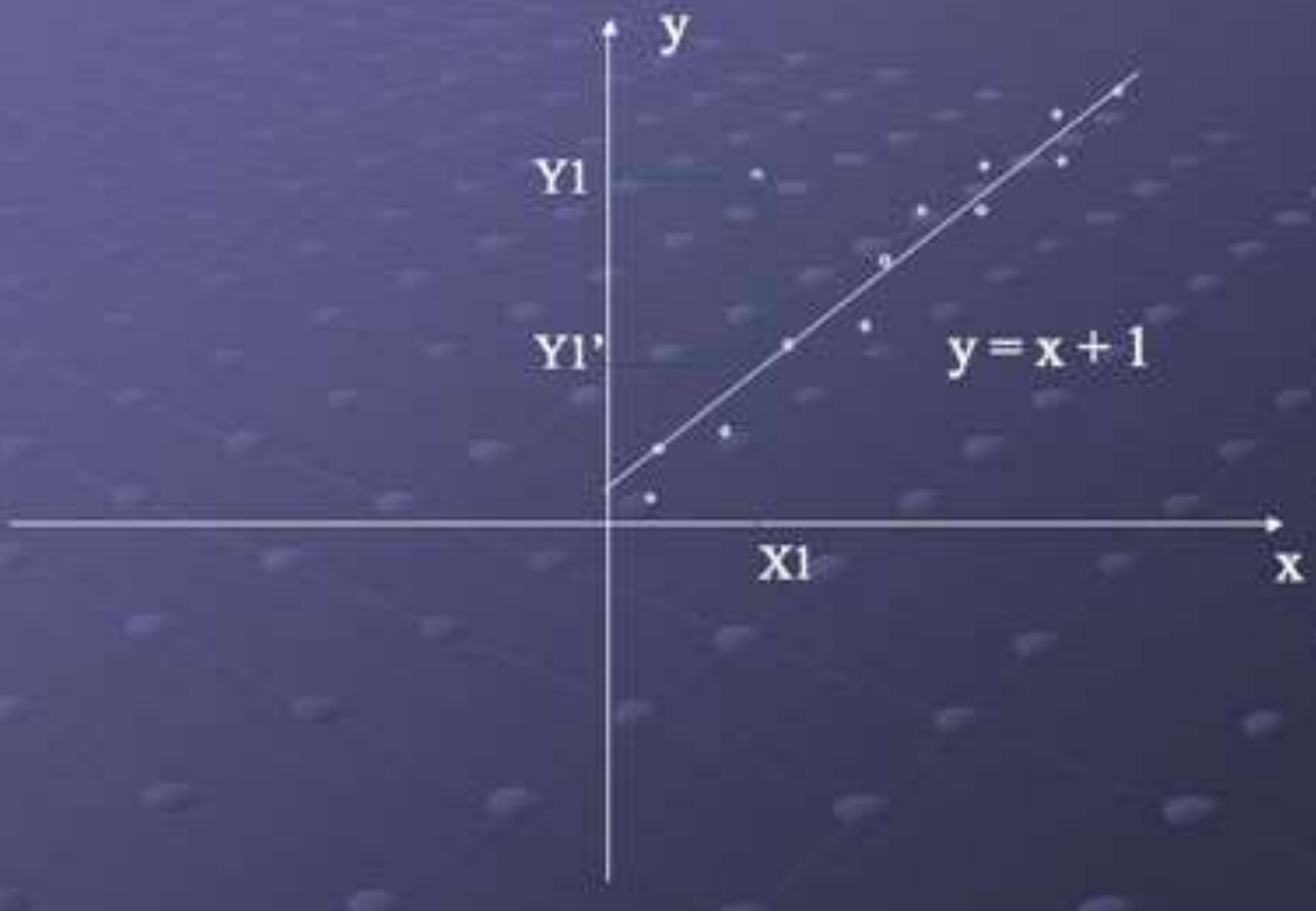
- Partition into (equal-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

*

Cluster Analysis



Regression



Overview of Basic Data Preparation Techniques

- 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 ≡ 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 Redundant Data 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 correlation 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

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

- 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

Overview of Basic Data Preparation Techniques

- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Reduction Strategies

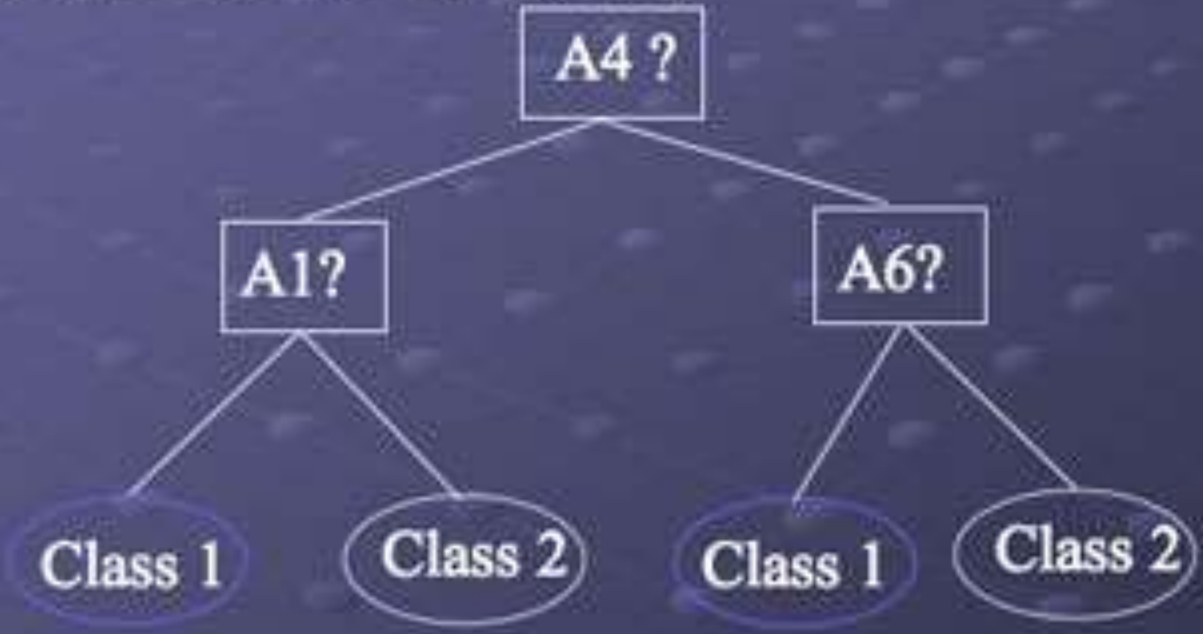
- Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
  - Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Data reduction strategies
  - Dimensionality reduction
  - Numerosity reduction
  - Discretization and concept hierarchy generation

Dimensionality Reduction

- Feature selection (i.e., attribute subset selection):
  - Select a minimum possible set of 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}



A4 ?

A1?

A6?

Class 1

Class 2

Class 1

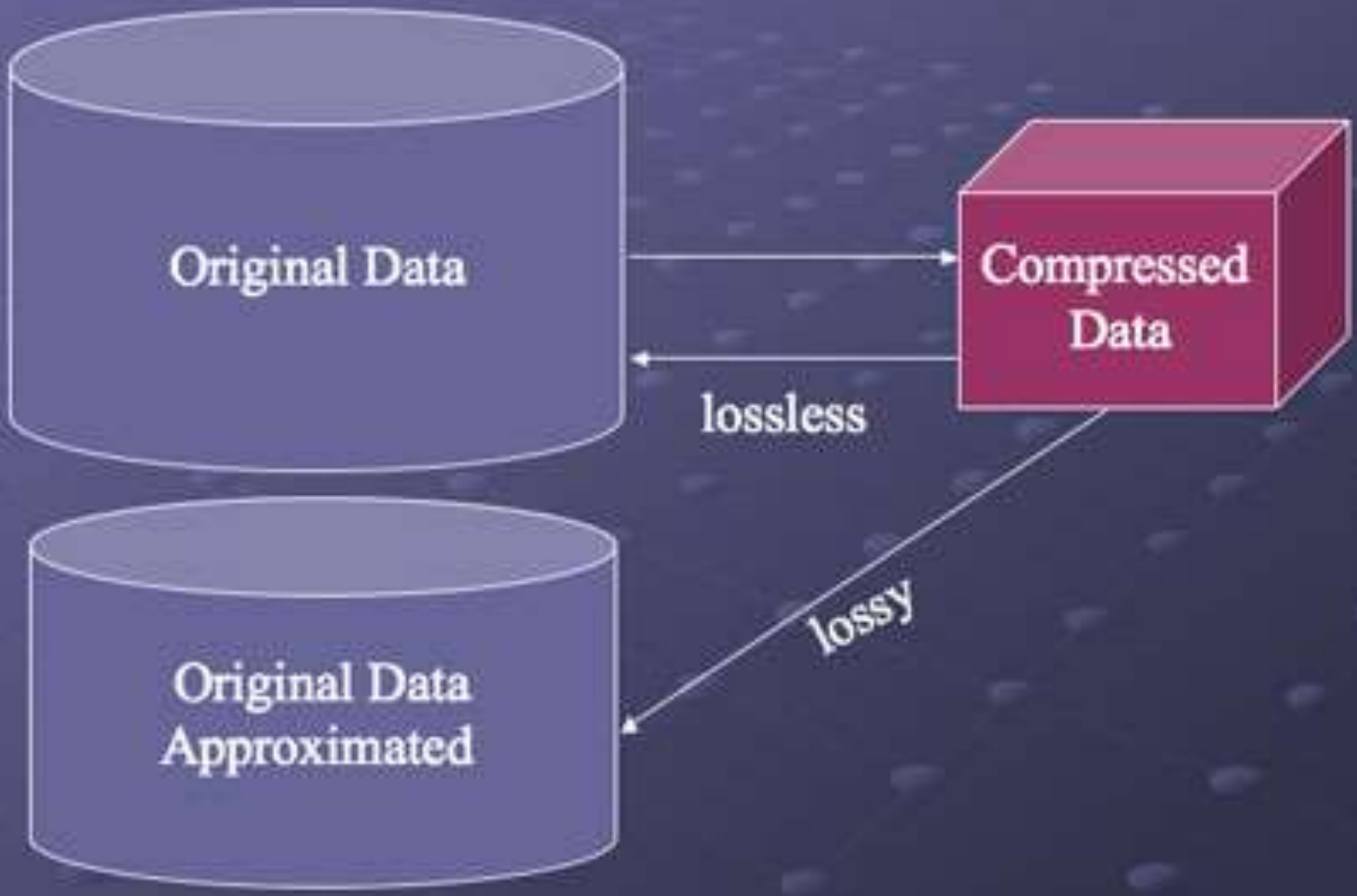
Class 2

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

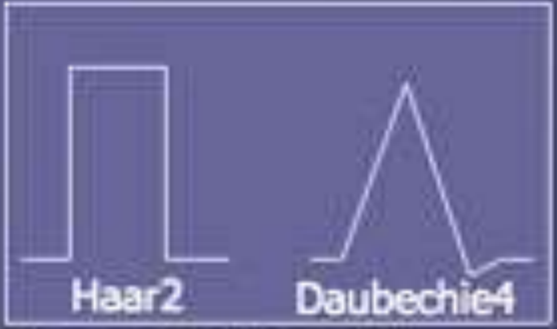
Data Compression

- 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



Wavelet Transforms

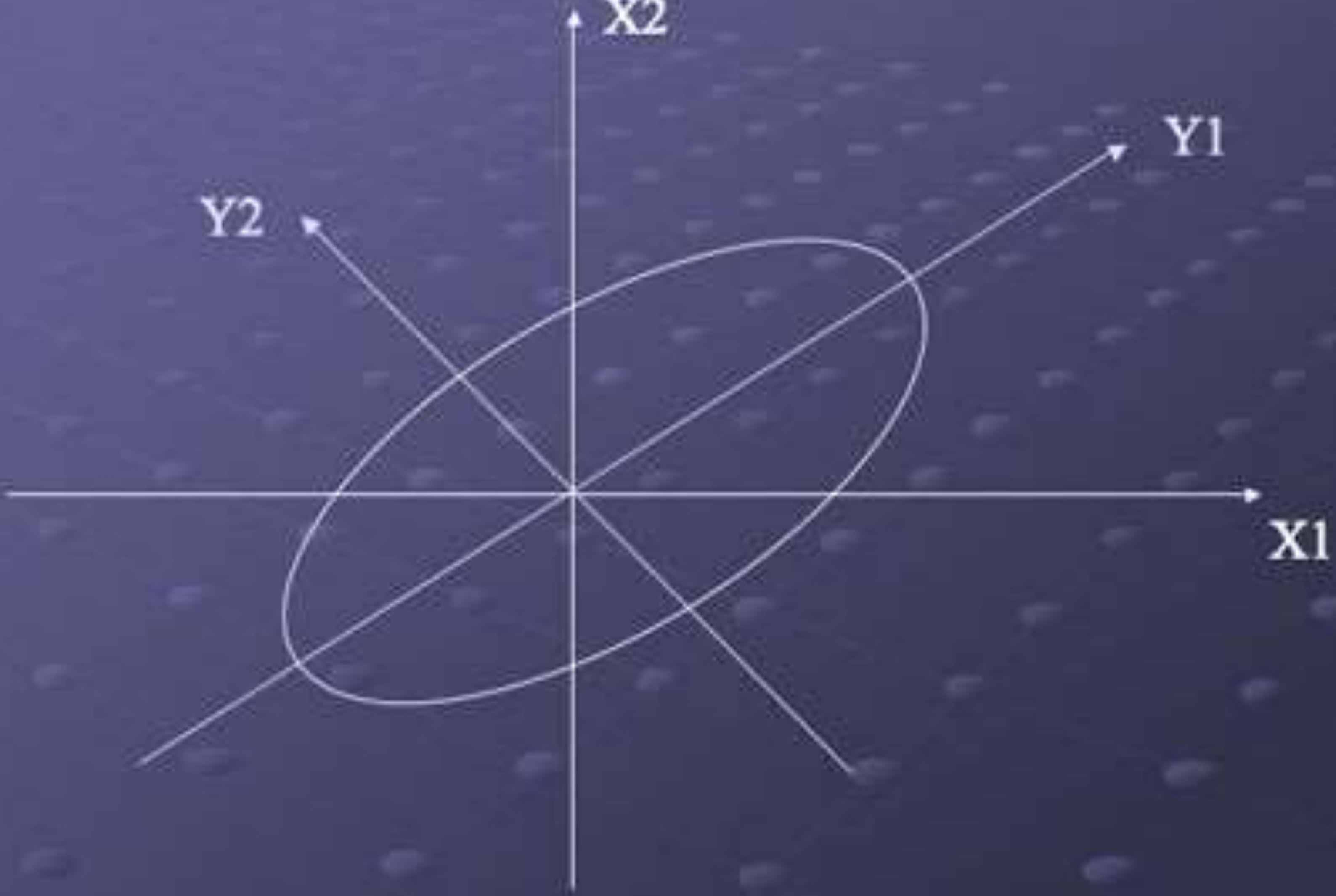


- Discrete wavelet transform (DWT): linear signal processing
- Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients
- Similar to discrete Fourier transform (DFT), but better lossy compression, localized in space

Principal Component Analysis

- Given N data vectors from k-dimensions, find c <= 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



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

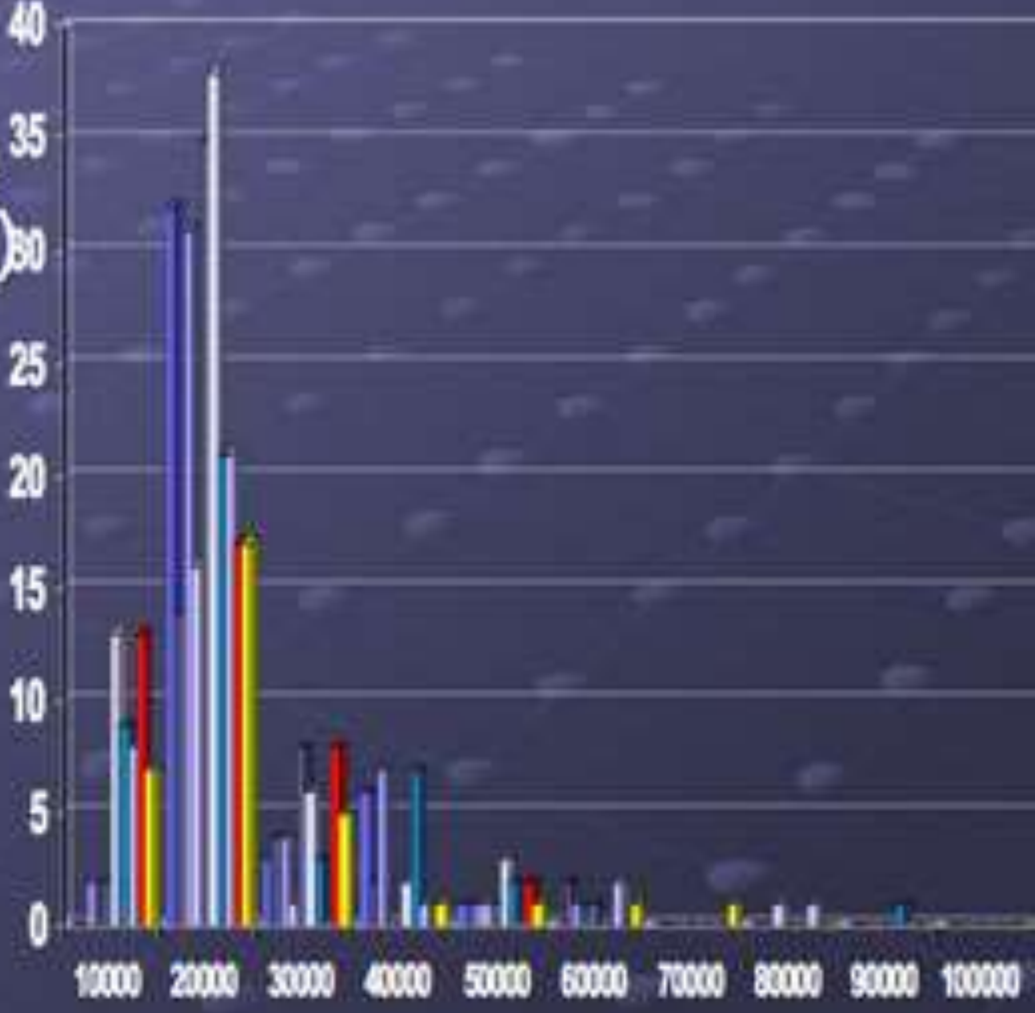
- Linear regression: Data modeled to fit a straight line
  - Often uses the least-square method to fit the line
- Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model: approximates discrete multidimensional probability 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 $Y_1, Y_2, \ldots, X_1, X2, \ldots.$
- Multiple regression: $Y = b0 + b1\ X1 + b2\ X2.$
  - 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.



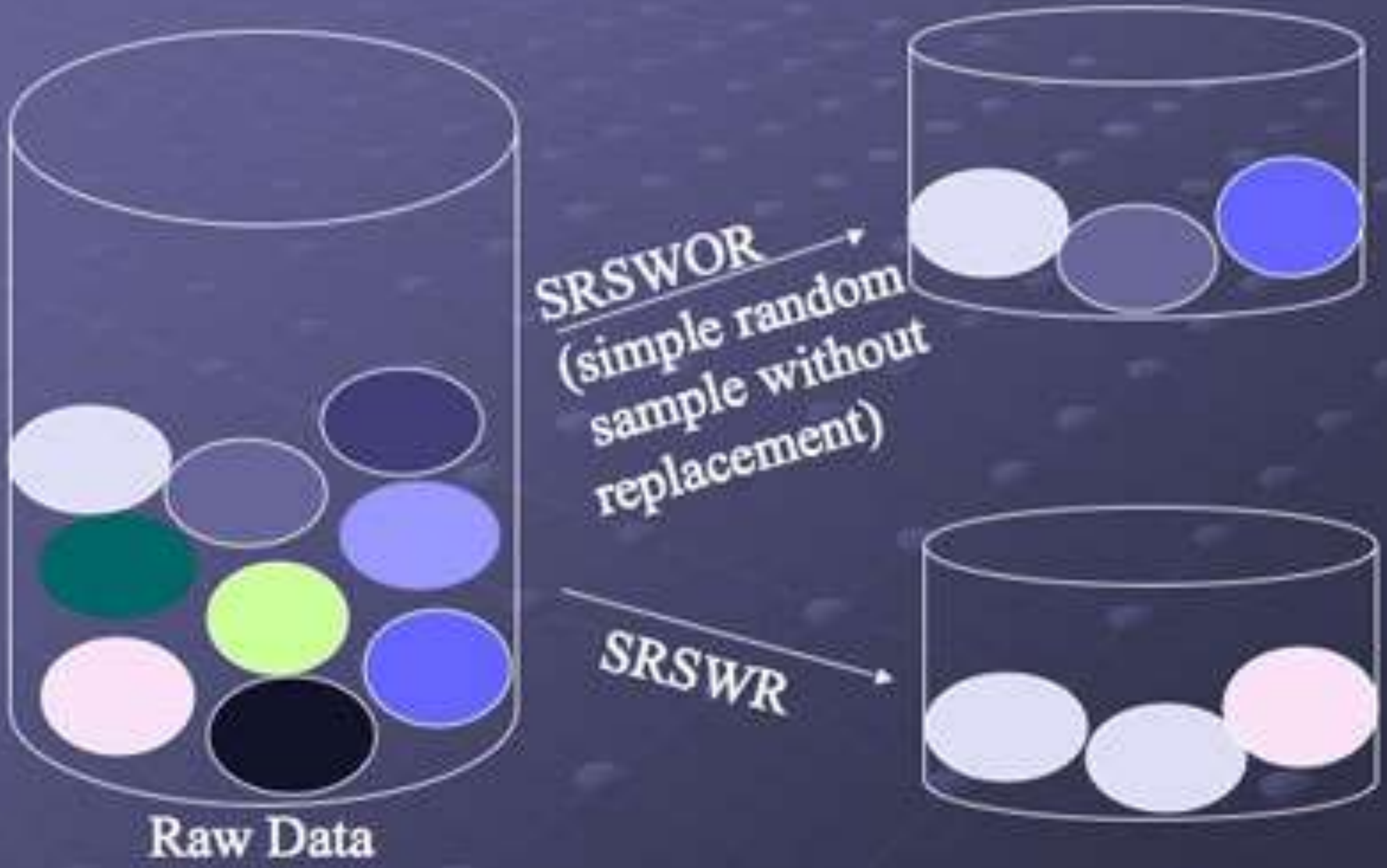
Clustering

- 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, as you have seen

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

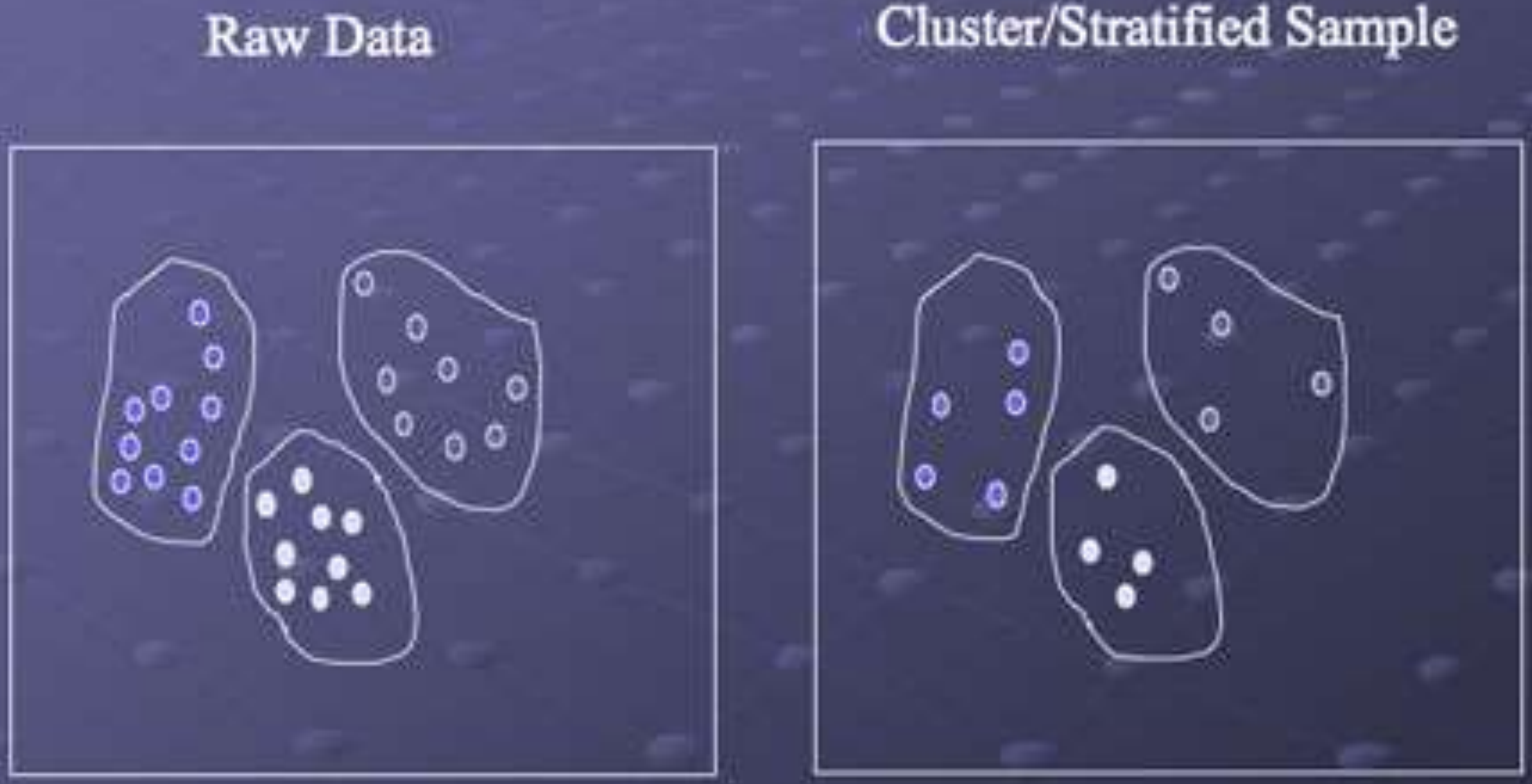


Raw Data

SRSWOR (simple random sample without replacement)

SRSWR

Sampling



Raw Data

Cluster/Stratified Sample

Hierarchical Reduction

- Use multi-resolution structure with different degrees of reduction
- Hierarchical clustering is often performed but tends to define partitions of data sets rather than "clusters"
- Parametric methods are usually not amenable to hierarchical representation
- Hierarchical aggregation
  - An index tree hierarchically divides a data set into partitions by value range of some attributes
  - Each partition can be considered as a bucket
  - Thus an index tree with aggregates stored at each node is a hierarchical histogram

Overview of Basic Data Preparation Techniques

- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- 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 hierarchy

- 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
- Histogram analysis
- Clustering analysis
- Entropy-based discretization
- Segmentation by natural partitioning

Entropy-Based Discretization

- Given a set of samples S, if S is partitioned into two 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.,

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

Segmentation by natural partitioning

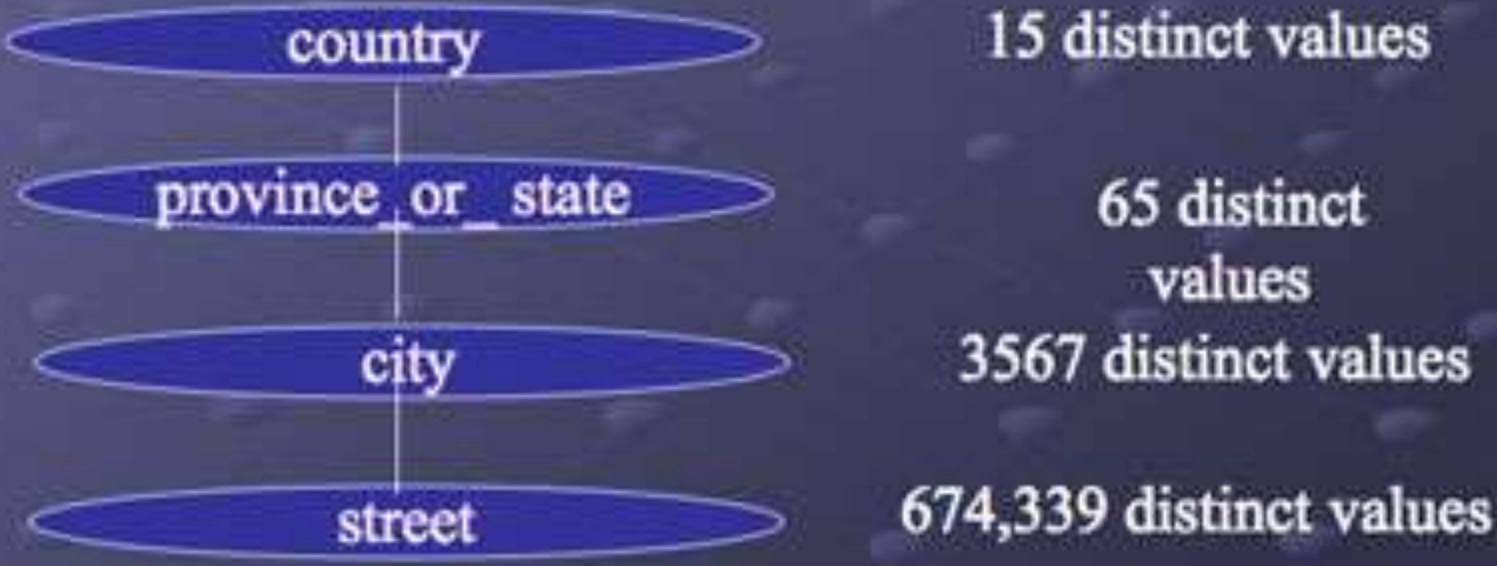
- “3-4-5” rule can be used to segment numeric data into relatively uniform, “natural” intervals.
- If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equal-width intervals
- If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals
- If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals

Concept hierarchy generation for categorical data

- Specification of a partial ordering of attributes explicitly at the schema level by users or experts
- Specification of a portion of a hierarchy by explicit data grouping
- Specification of a set of attributes, but not of their partial ordering
- Specification of only a partial set of attributes

Specification of a set of attributes

Concept hierarchy can be automatically generated based on the number of distinct values per attribute in the given attribute set. The attribute with the most distinct values is placed at the lowest level of the hierarchy.



country

15 distinct values

province_or_ state

65 distinct values

city

3567 distinct values

street

674,339 distinct values

Overview of Basic Data Preparation Techniques

- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Summary

- Data preparation is a big issue for 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

Assignment I