Data Warehousing & Mining
Unit-II
Data Preprocessing

Dr. K. Raghava Rao
Professor of CSE
Dept. of MCA
KL University.

Data Preprocessing

Definition: Data preprocessed in order to help improve the quality of the data and , consequently improve efficiency and ease of mining process and the results.

There are no. of data preprocessing techniques, they are:

>-data cleaning-

>-data integration-

>-data transformation& Discretization

>-data reduction-

Data Preprocessing:
Why Preprocess the Data?

>-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

Data Preprocessing:
Why Preprocess the Data?

Why Data is Dirty?

>-Incomplete data may come from “Not applicable” data value when collected Different considerations between the time when the data was collected and when it is analyzed.

>-Noisy data (incorrect values) may come from Faulty data collection instruments Human or computer error at data entry Errors in data transmission.

>-Inconsistent data may come from Different data sources Functional dependency violation (e.g., modify some linked data).

>-Duplicate records also need data cleaning.

>-Human/hardware/software problems.

Data Preprocessing:
Why Preprocess the Data?

>-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

Data Preprocessing:
Why Preprocess the Data?

Multi-Dimensional Measure of Data Quality

>-A well-accepted multidimensional view:

- -Accuracy
- -Completeness
- -Consistency
- -Timeliness
- -Believability
- -Value added
- -Interpretability
- -Accessibility

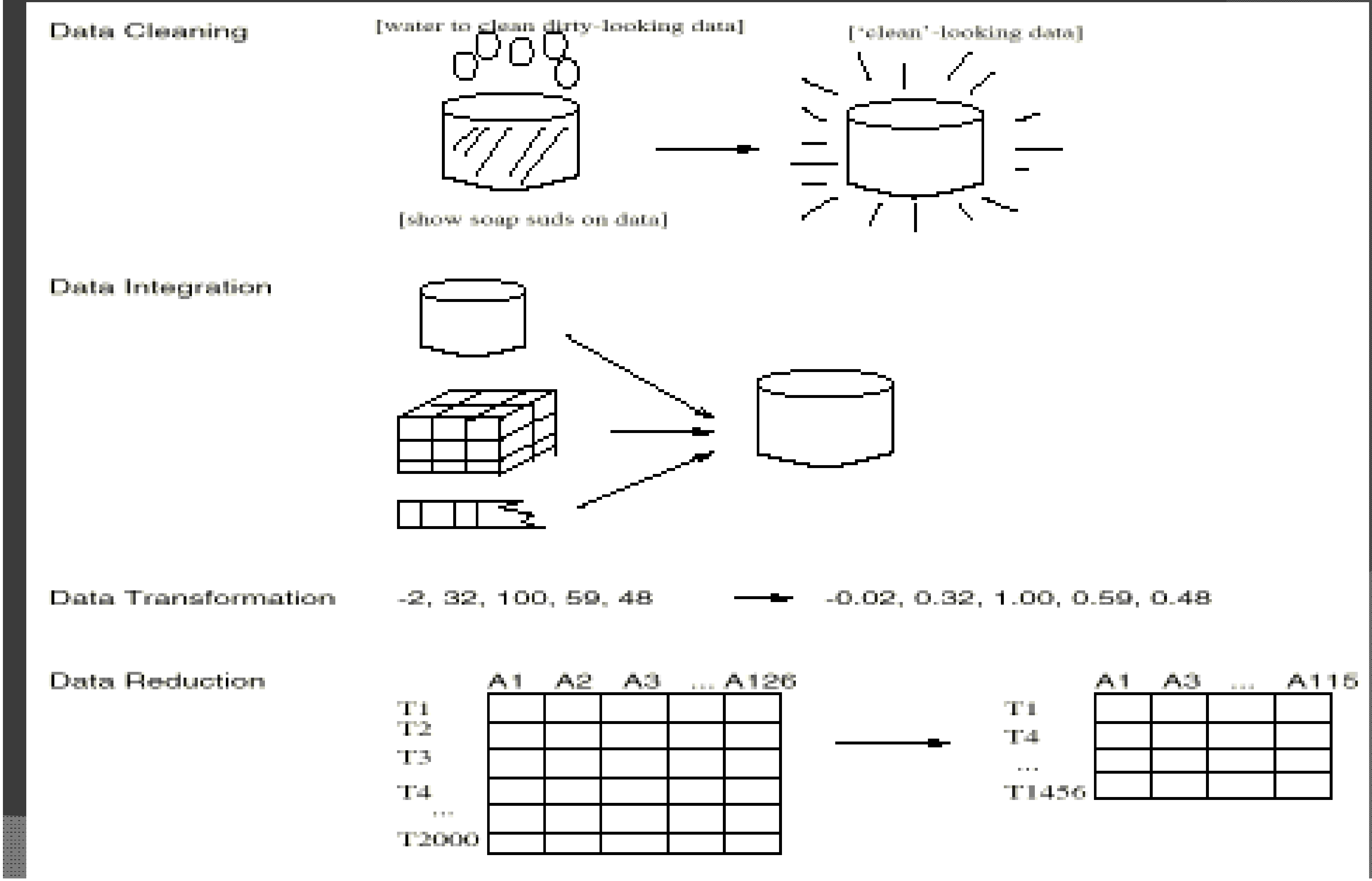
>-Broad categories:

-Intrinsic, contextual, representational, and accessibility

Data Preprocessing:

Why Preprocess the Data?

Fig. Forms of data preprocessing



Data Preprocessing:
Why Preprocess the Data?

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

Data Preprocessing:
Descriptive Data Summarization

Definition: DDS techniques can be used to identify the typical properties of your data and highlight which data values should be treated as noise or outliers.

>-Motivation

-To better understand the data: central tendency, variation and spread

>-Data dispersion characteristics

-median, max, min, quintiles, outliers, variance, etc.

>-Numerical dimensions correspond to sorted intervals

-Data dispersion: analyzed with multiple granularities of precision

-Boxplot or quintile analysis on sorted intervals

>-Dispersion analysis on computed measures

-Folding measures into numerical dimensions

-Boxplot or quintile analysis on the transformed cube

Data Preprocessing: Descriptive Data Summarization
Measuring the Central Tendency

>-Mean (algebraic measure) (sample vs. population): $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ $\mu = \frac{\sum x}{N}$

- Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Trimmed mean: chopping extreme values

>-Median: A holistic measure

- Middle value if odd number of values, or average of the middle two values otherwise
- Estimated by interpolation (for *grouped data*):

$$median = L_1 + \left(\frac{n/2 - (\sum f)l}{f_{median}}\right)c$$

>-Mode

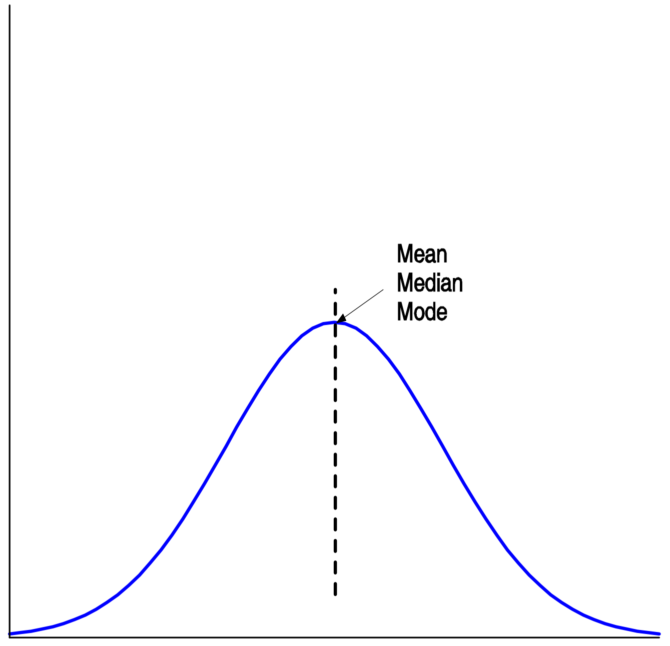
- Value that occurs most frequently in the data
- Unimodal, bimodal, trimodal
- Empirical formula:

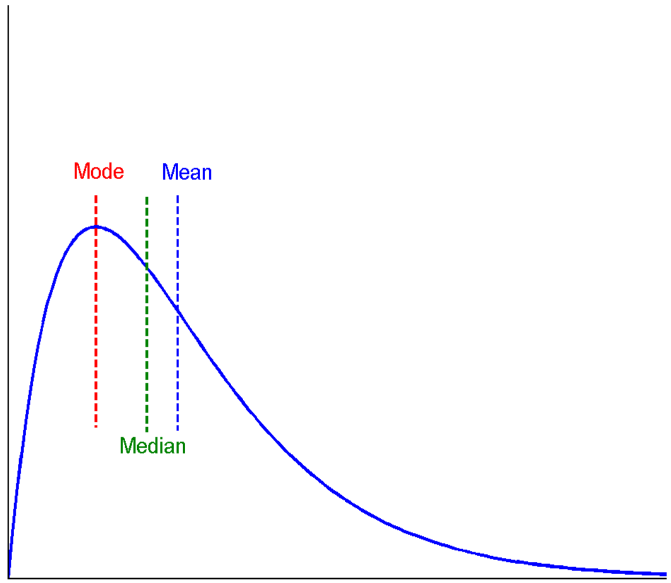
$$mean - mode = 3 \times (mean - median)$$

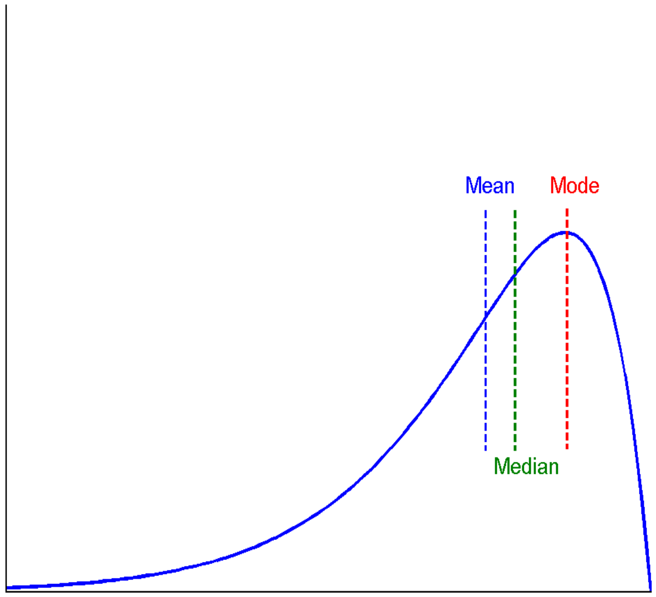
Data Preprocessing: Descriptive Data Summarization Measuring the Central Tendency

Symmetric vs. Skewed Data

>-Median, mean and mode of symmetric, positively and negatively skewed data







Data Preprocessing: Descriptive Data Summarization
Measuring the Dispersion of Data

>-Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)

Inter-quartile range: IQR = $Q_3 – Q_1$

Five number summary: min, Q_1, M, Q_3, max

Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually

Outlier: usually, a value higher/lower than 1.5 x IQR

>-Variance and standard deviation (*sample: s, population: σ)*

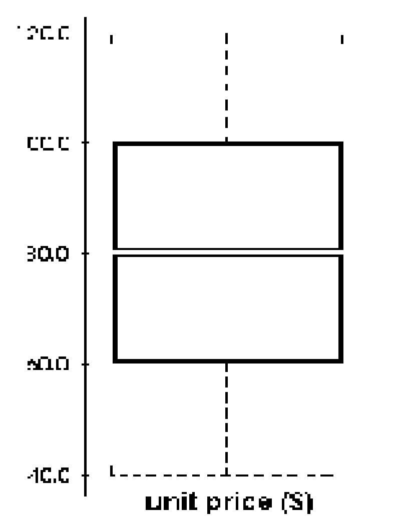
Variance: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^2 - \mu^2$$

>-Standard deviation *s (or σ)* is the square root of variance *s^2 (or σ^2)*

Data Preprocessing: Descriptive Data Summarization
Measuring the Dispersion of Data



Boxplot Analysis

>-Five-number summary of a distribution:

Minimum, Q1, M, Q3, Maximum

>-Boxplot

Data is represented with a box

The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ

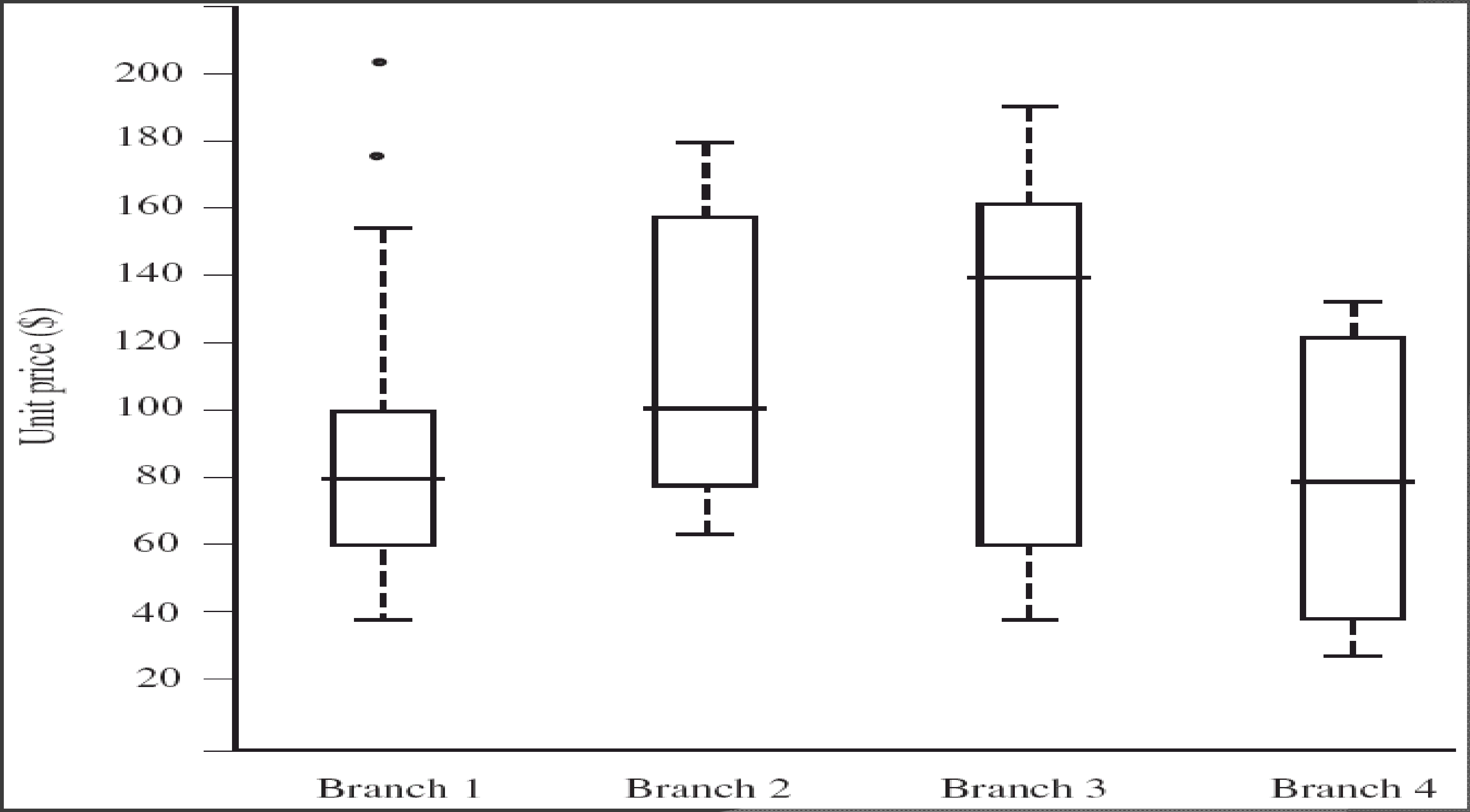
The median is marked by a line within the box

Whiskers: two lines outside the box extend to Minimum and Maximum

Data Preprocessing: Descriptive Data Summarization

Measuring the Dispersion of Data

Visualization of Data Dispersion: Boxplot Analysis



Data Preprocessing: Descriptive Data Summarization

Measuring the Dispersion of Data

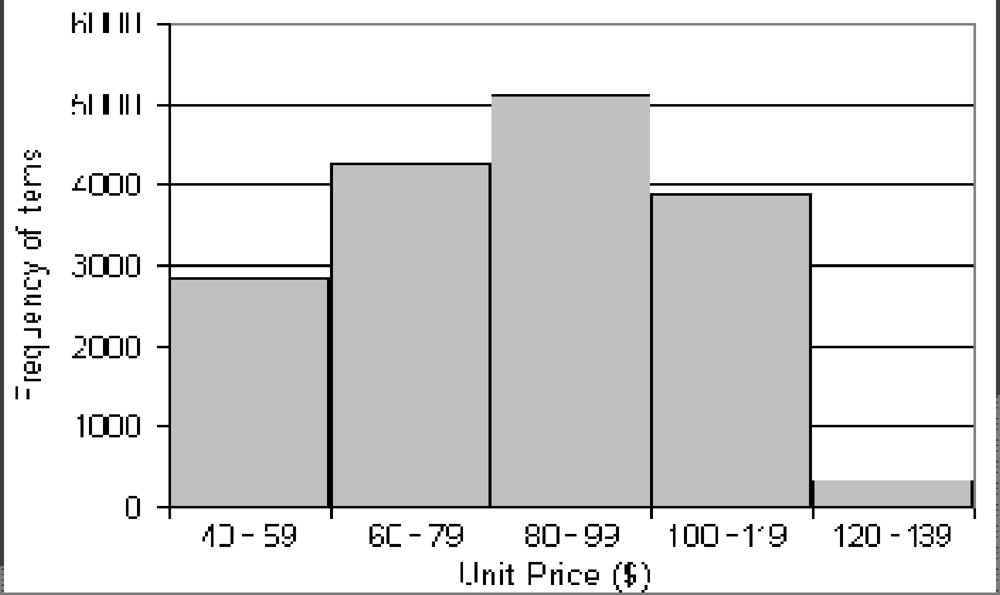
Histogram Analysis

>-Graph displays of basic statistical class descriptions

-Frequency histograms

--A univariate graphical method

-- Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data



Data Preprocessing: Descriptive Data Summarization

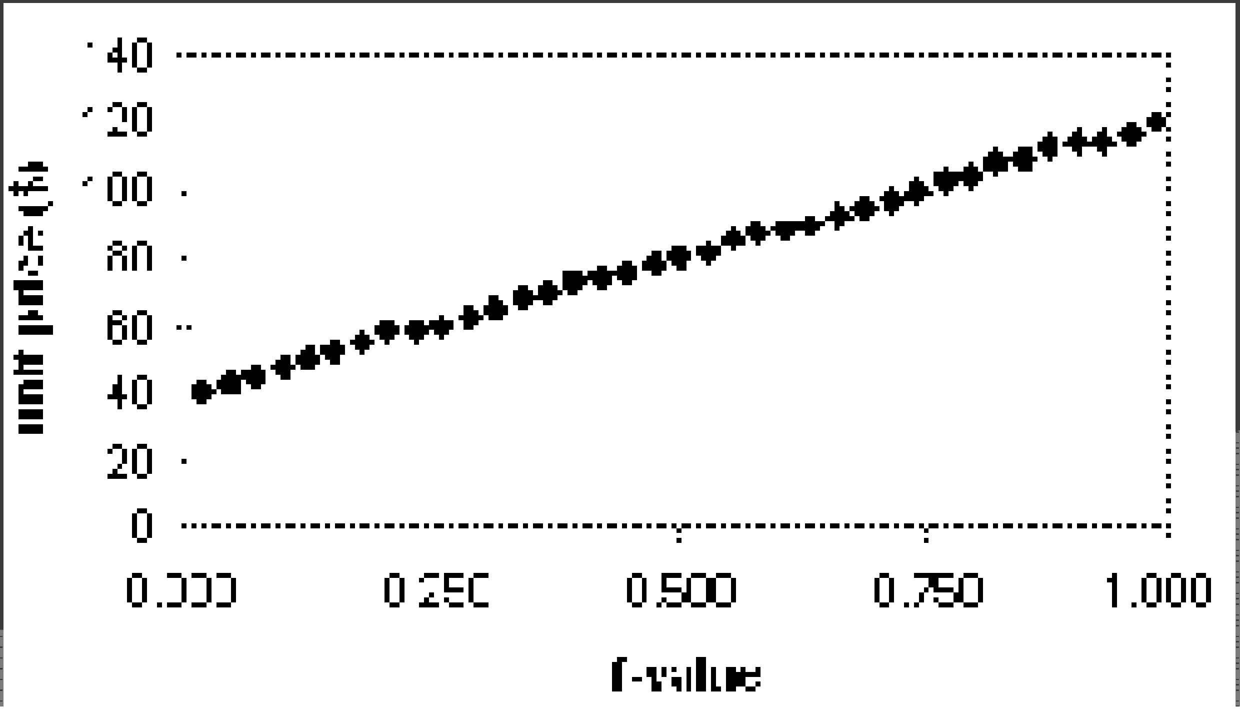
Measuring the Dispersion of Data

Quantile Plot

>-Displays all of the data
(allowing the user to assess both the overall behavior and unusual occurrences)

>-Plots quantile information

For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



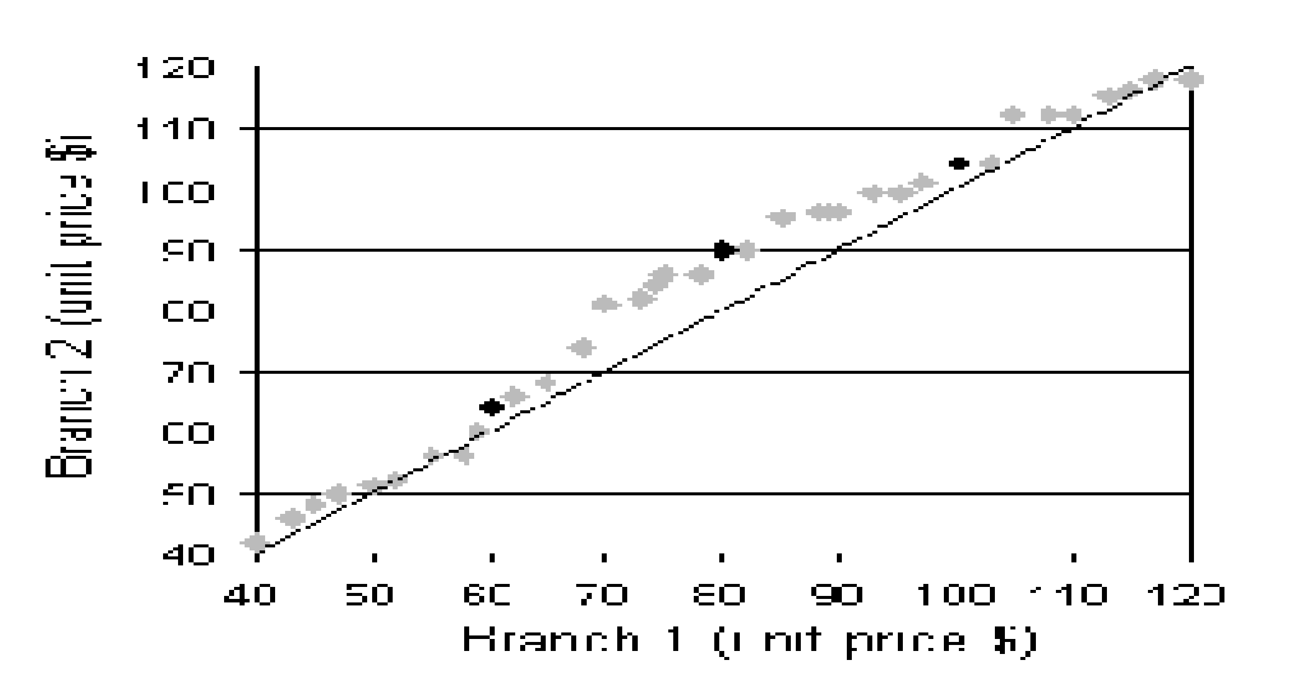
Data Preprocessing: Descriptive Data Summarization

Measuring the Dispersion of Data

Quantile-Quantile (Q-Q) Plot

>-Graphs the quantiles of one univariate distribution against the corresponding quantiles of another

>-Allows the user to view whether there is a shift in going from one distribution to another



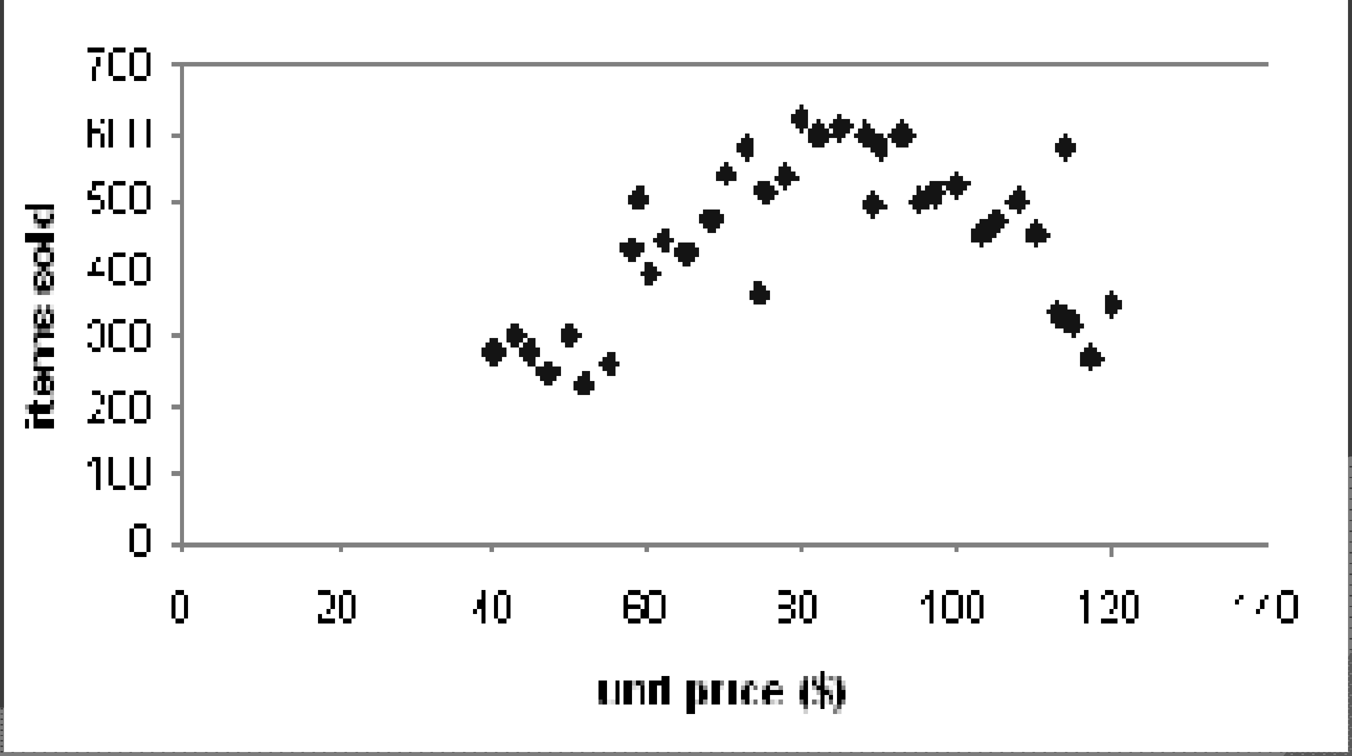
Data Preprocessing: Descriptive Data Summarization

Measuring the Dispersion of Data

Scatter plot

>-Provides a first look at bivariate data to see clusters of points, outliers, etc.

>-Each pair of values is treated as a pair of coordinates and plotted as points in the plane.

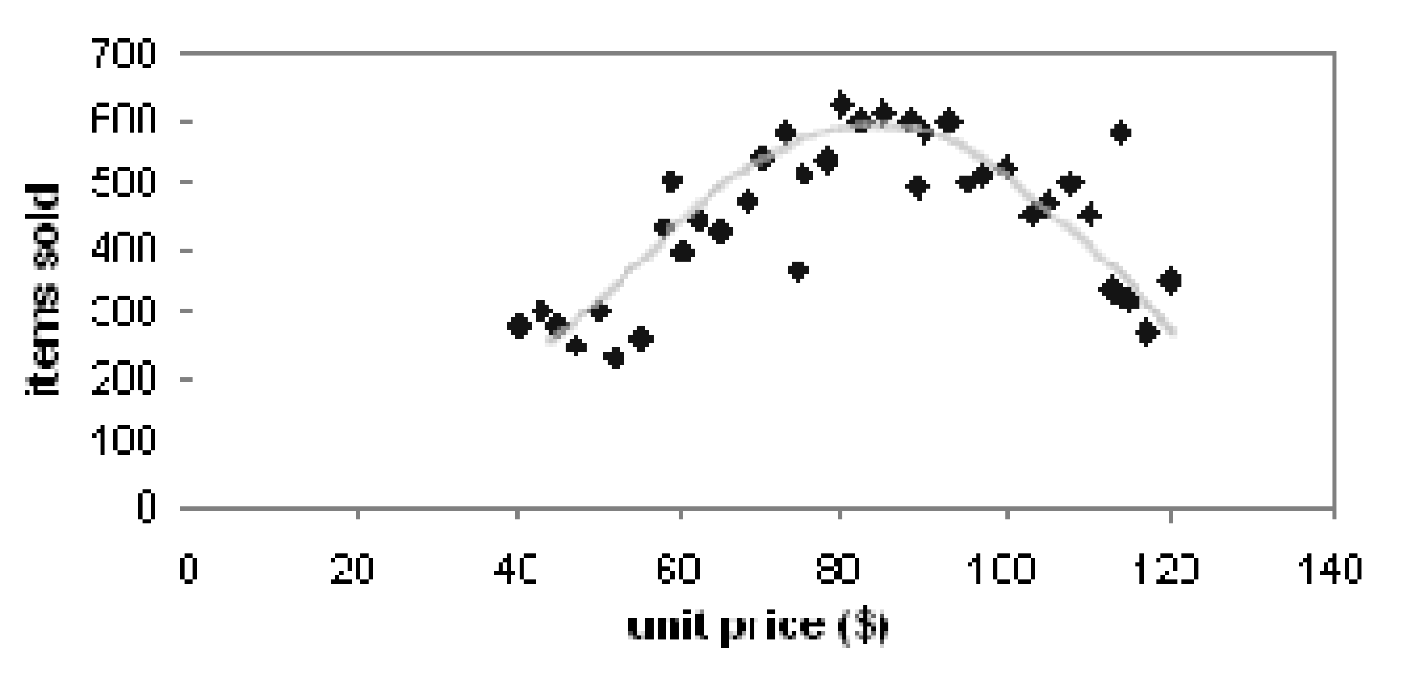


Data Preprocessing: Descriptive Data Summarization
Measuring the Dispersion of Data

Loess Curve

>-Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence

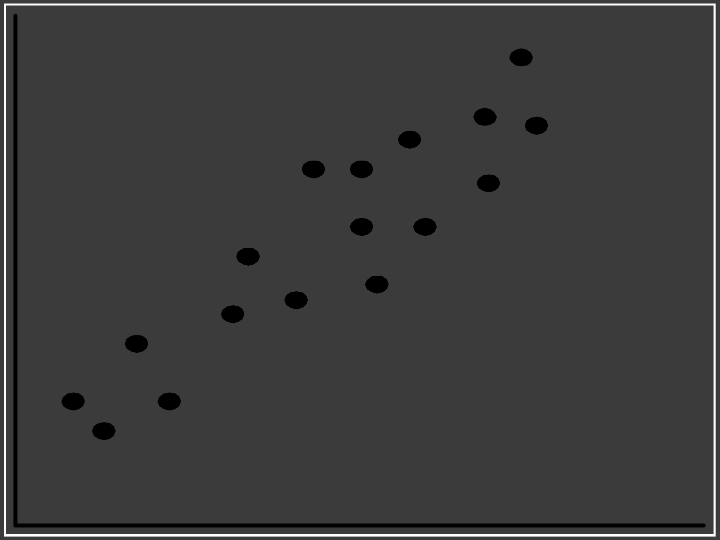
>-Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression

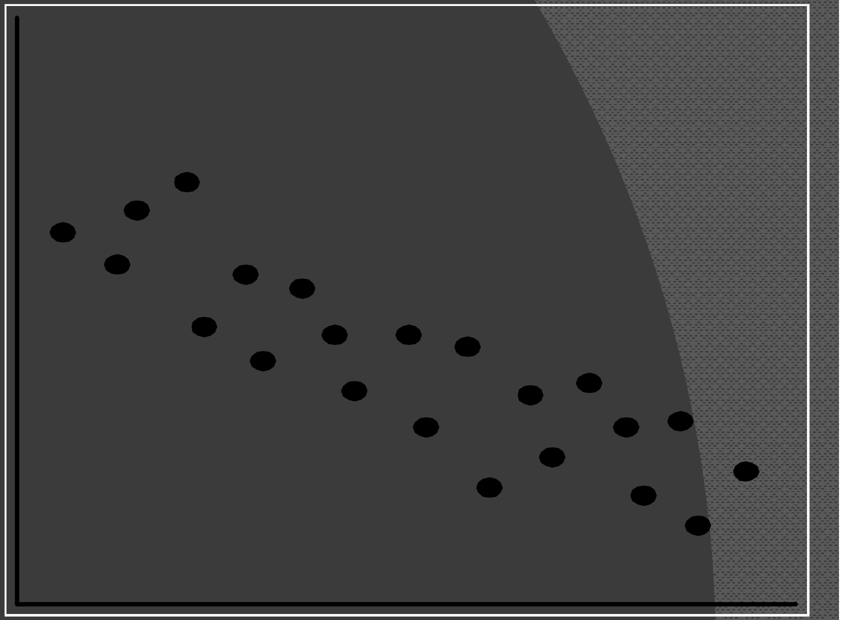


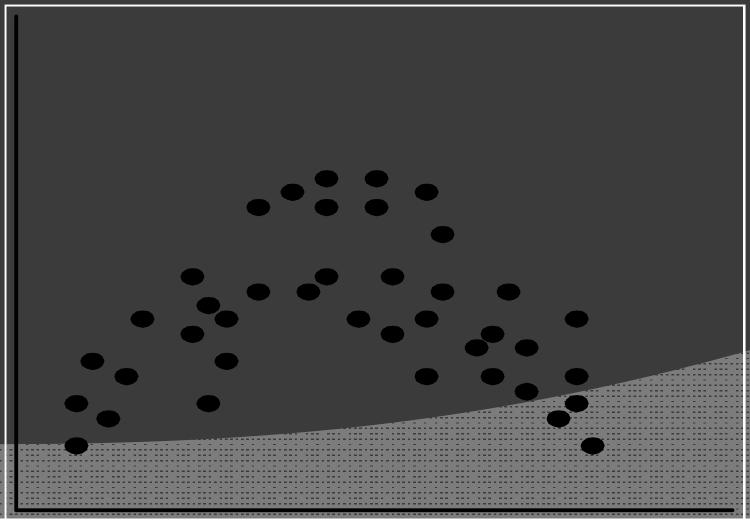
Data Preprocessing: Descriptive Data Summarization

Measuring the Dispersion of Data

Positively and Negatively Correlated Data



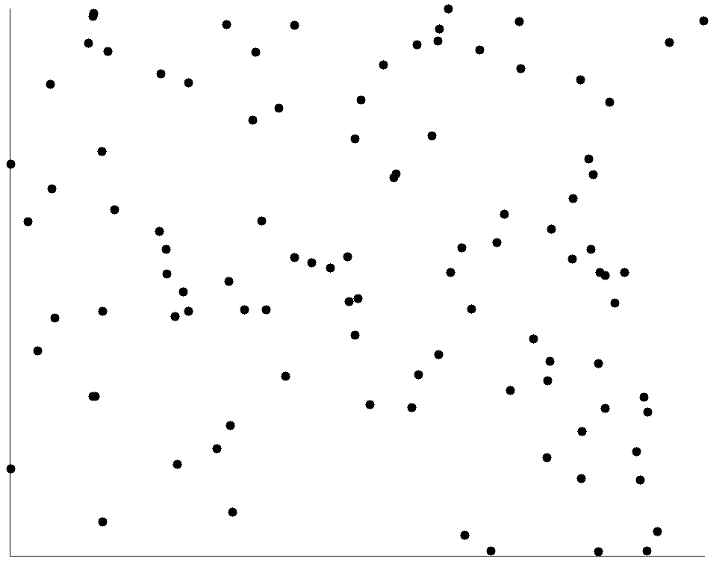


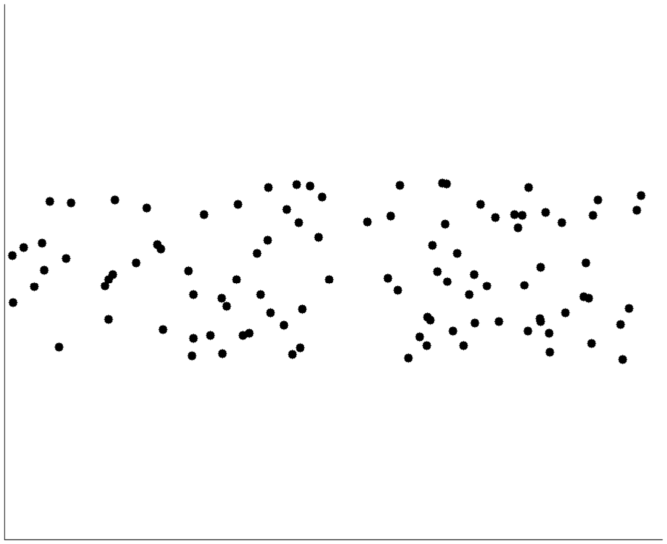


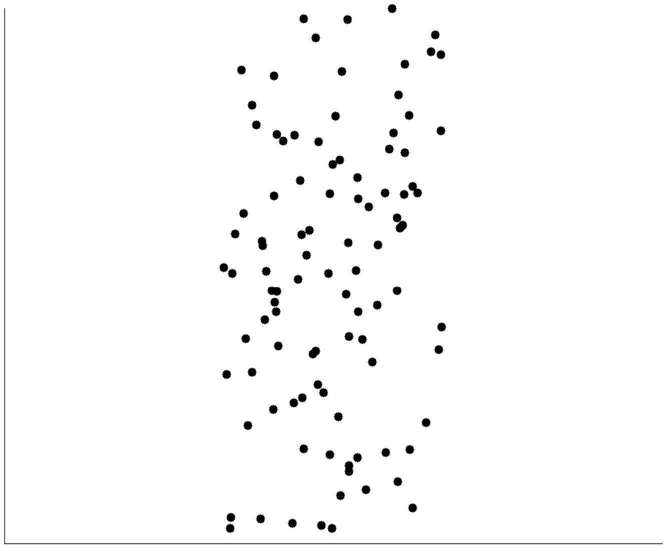
Data Preprocessing: Descriptive Data Summarization

Measuring the Dispersion of Data

Not Correlated Data







Data Preprocessing: Descriptive Data Summarization

Measuring the Dispersion of Data

Graphic Displays of Basic Statistical Descriptions

>-Histogram: (shown before)

>-Boxplot: (covered before)

>-Quantile plot: each value x_i is paired with f_i indicating that approximately 100 f_i% of data are $\leq x_i$.

>-Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another.

>-Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane.

>-Loess (local regression) curve: add a smooth curve to a scatter plot to provide better perception of the pattern of dependence.

Data Preprocessing: Data cleaning

Definition: Data cleaning or cleansing are routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

Data Preprocessing: Data cleaning

Missing Values

>-Missing values handled using following Methods:

1)Ignore the tuple : usually done when class label is missing

(assuming the task is classification—not effective in certain cases)

2)Fill in the missing values manually: tedious + infeasible?

3) Use a global constant to fill in the missing value: e.g., "unknown", a new class?!

4) Use the attribute mean to fill in the missing value: use this value to replace missing value.

5) Use the attribute mean for all samples belonging to same class as given

Tuple: smarterUse the attribute mean to fill in the missing value.

6)Use the most probable value to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Data Preprocessing: Data cleaning
Noisy Data

>-Q: What is noise?
Ans:It is a Random error in a measured variable.

>-Incorrect attribute values may be 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

Data Preprocessing: Data cleaning
Noisy 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.

used also for discretization (discussed later)

>-Clustering:

--detect and remove outliers.

>-Semi-automated method:

--combined computer and human inspection

--detect suspicious values and check manually.

>-Regression:

--smooth by fitting the data into regression functions.

Data Preprocessing: Data cleaning
Noisy Data

Binning method for data smoothig:

>-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:

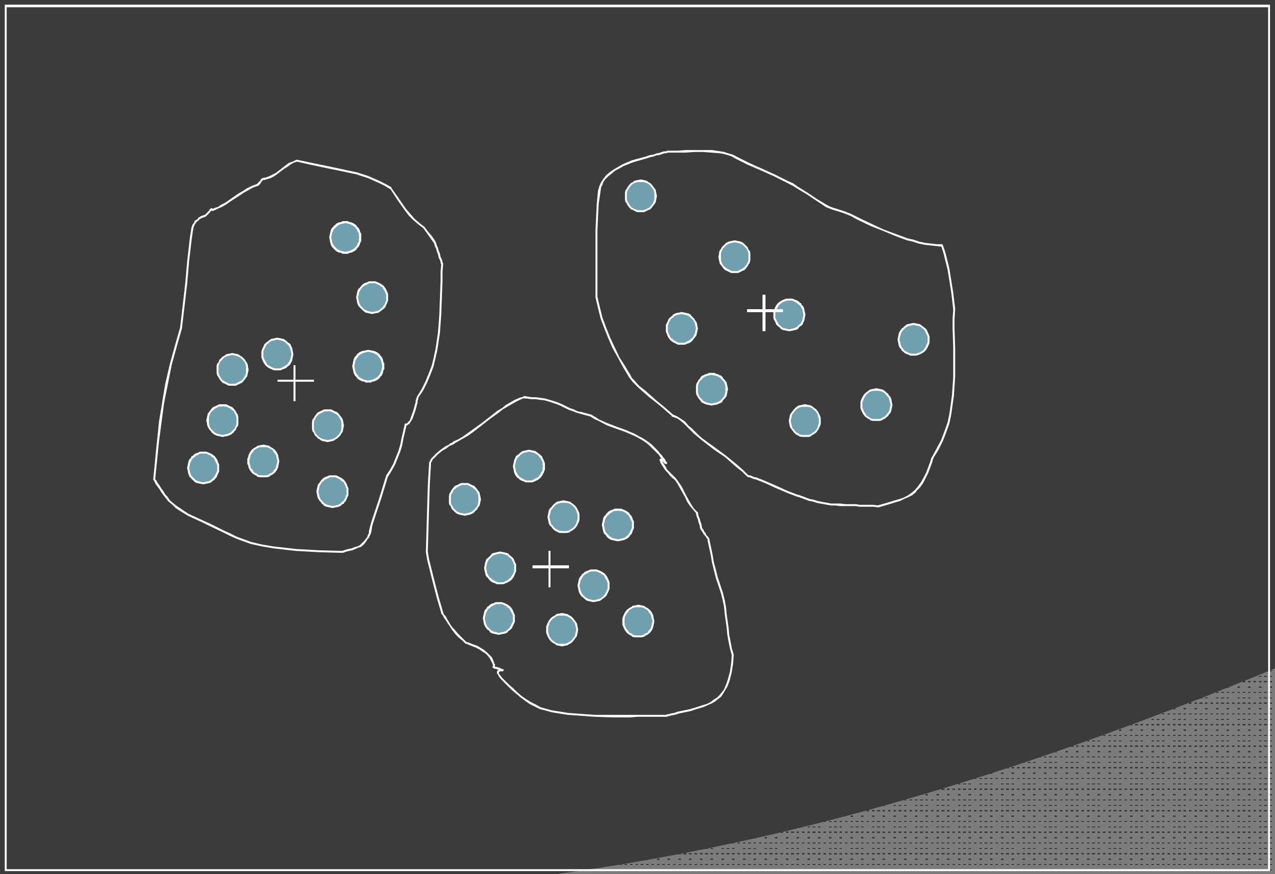
- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

Smoothing by bin boundaries:

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

Data Preprocessing: Data cleaning
Noisy Data

Cluster Analysis



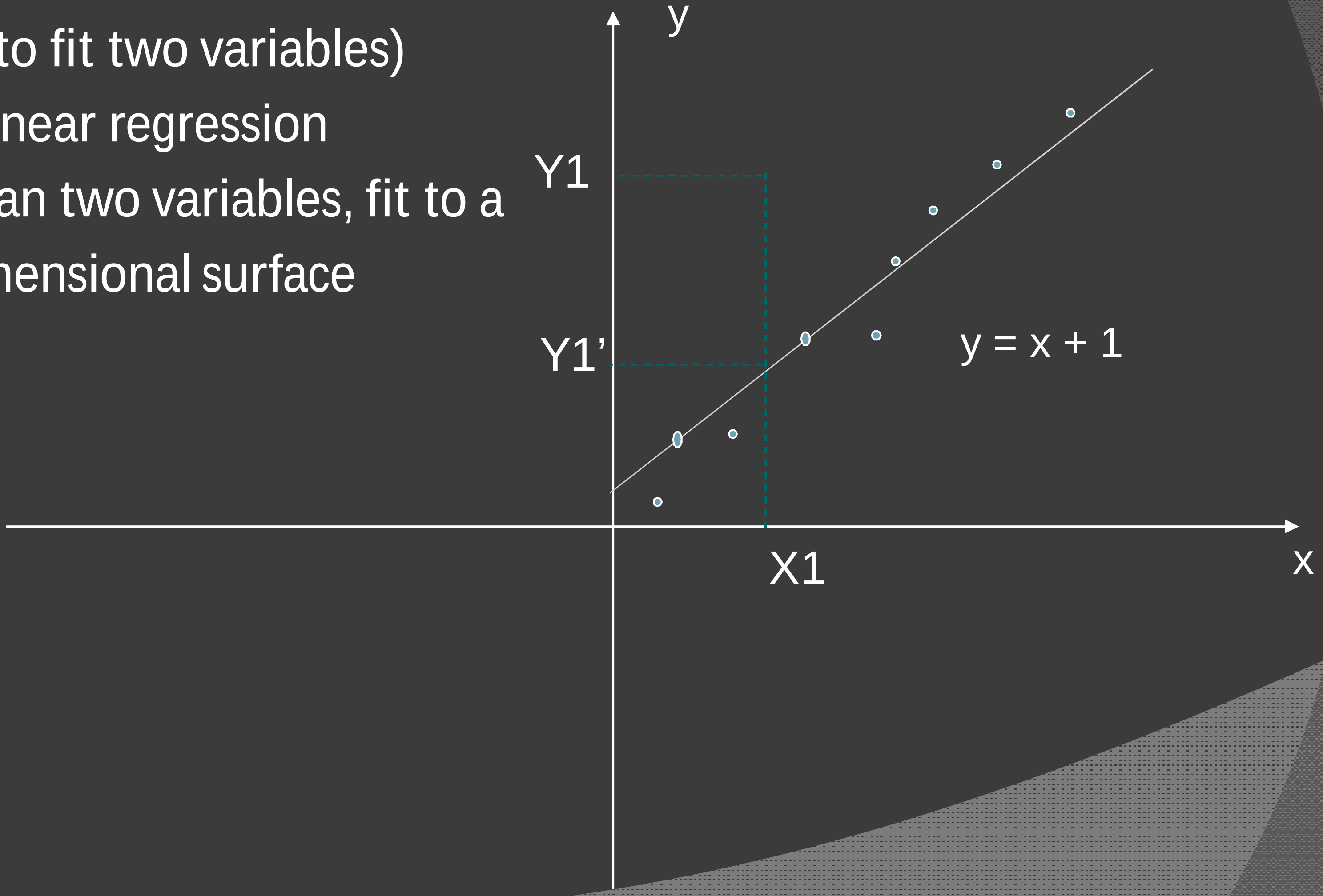
Data Preprocessing: Data cleaning
Noisy Data

Linear regression

(best line to fit two variables)

Multiple linear regression

(more than two variables, fit to a multidimensional surface



Data Preprocessing: Data cleaning
Data Cleaning as a Process

>-As a first step in data cleaning as a process is discrepancy detection. Discrepancy caused by the following:

- -poorly designed data entry
- -human error in data entry
- -deliberate errors and data decay.
- -inconsistent data representation and inconsistent of use codes

>- This can be prevented in one way with the knowledge using metadata.

>- field overloading : when developers squeeze new attribute definitions into unused (bit) portions of already defined attributes.

>- By applying unique rules : consecutive rules , null rule .
There are tools to detect data discrepancy
-data scrubbing tools: using domain knowledge to detect errors and make corrections in data.

-data auditing tools : find discrepancies by analyzing the data to discover rules and relationships and detecting data that violate such conditions.

Data Preprocessing: Data Integration & Transformation
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-#

--Meta data can be to help avoid errors in schema integration and transform the data as well.

>-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, different currency.

Data Preprocessing: Data Integration & Transformation
Data Integration

>-Redundant data occur often when integrating multiple DBs

--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

$$r_{A,B} = \frac{\Sigma(A - \overline{A})(B - \overline{B})}{(n-1)\sigma_A \sigma_B}$$

-1<= $r_{A,B}$ <=+1, if $r_{A,B}$ is greater than 0, then A,B positively correlated,it means strong close between two ,so A or B may be removed as redundancy. If $r_{A,B}$ Equal to 0 or <0 ,A and B negatively correlated. Correlation does not imply casualtiy.

>-Careful integration can help reduce/avoid redundancies and inconsistencies and improve mining speed and quality.

Data Preprocessing: Data Integration & Transformation
Data Integration

>- For categorical data, a correlation relationship between two attributes A and B can be discovered by chi-square test

$$\chi 2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(oi, j - ei, j)}{ei, j}$$

Where Oi,j observed frequency of the joint event ,Ai,Bj is the expected frequency of which can be computed as

$$Ci,j = \frac{count(A=ai) \text{ x } count(B= bj)}{N}$$

Data Preprocessing: Data Integration & Transformation

Data Transformation

>-Smoothing: remove noise from data (binning, clustering, regression)

>-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 on

Data Preprocessing: Data Integration & Transformation

Data Transformation

>-Particularly useful for classification

(NNs, distance measurements,nn classification, etc)

>-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(| |)<1

Data Preprocessing: Data Reduction

>-Problem:

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

>-Solution?

--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

–Data cube aggregation

–Dimensionality reduction

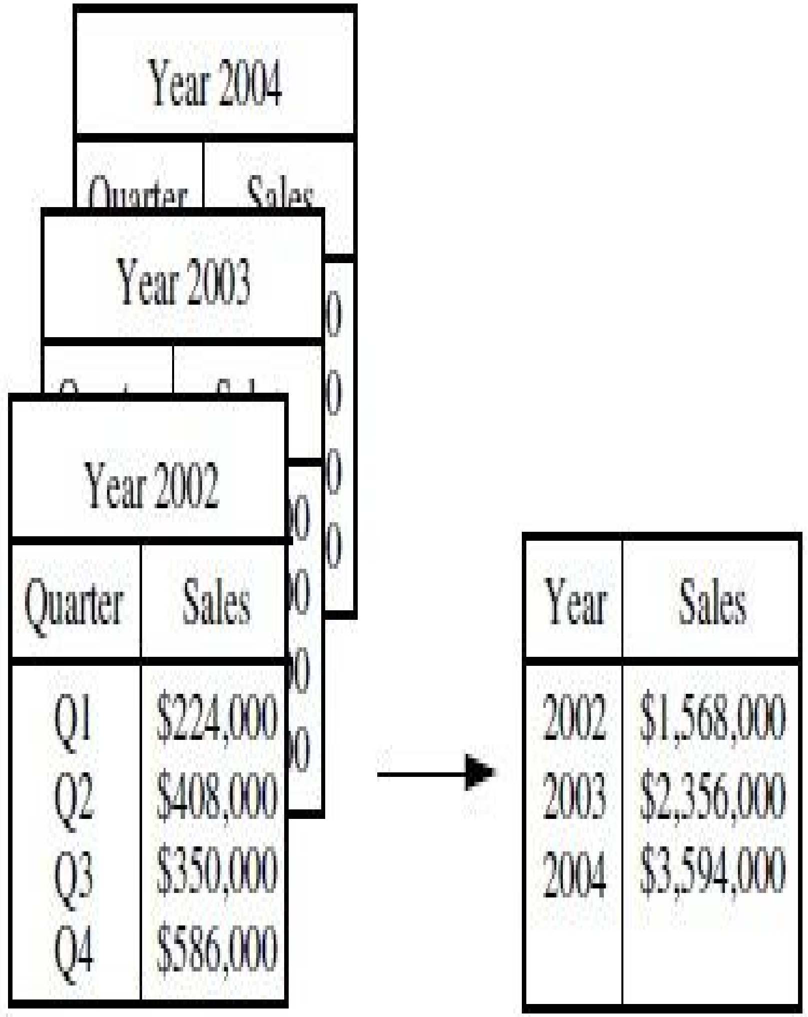
–Data compression

–Numerosity reduction

–Discretization and concept hierarchy generation

Data Preprocessing: Data Reduction
Data Cube Aggregation

Fig.-Sales data for given branch

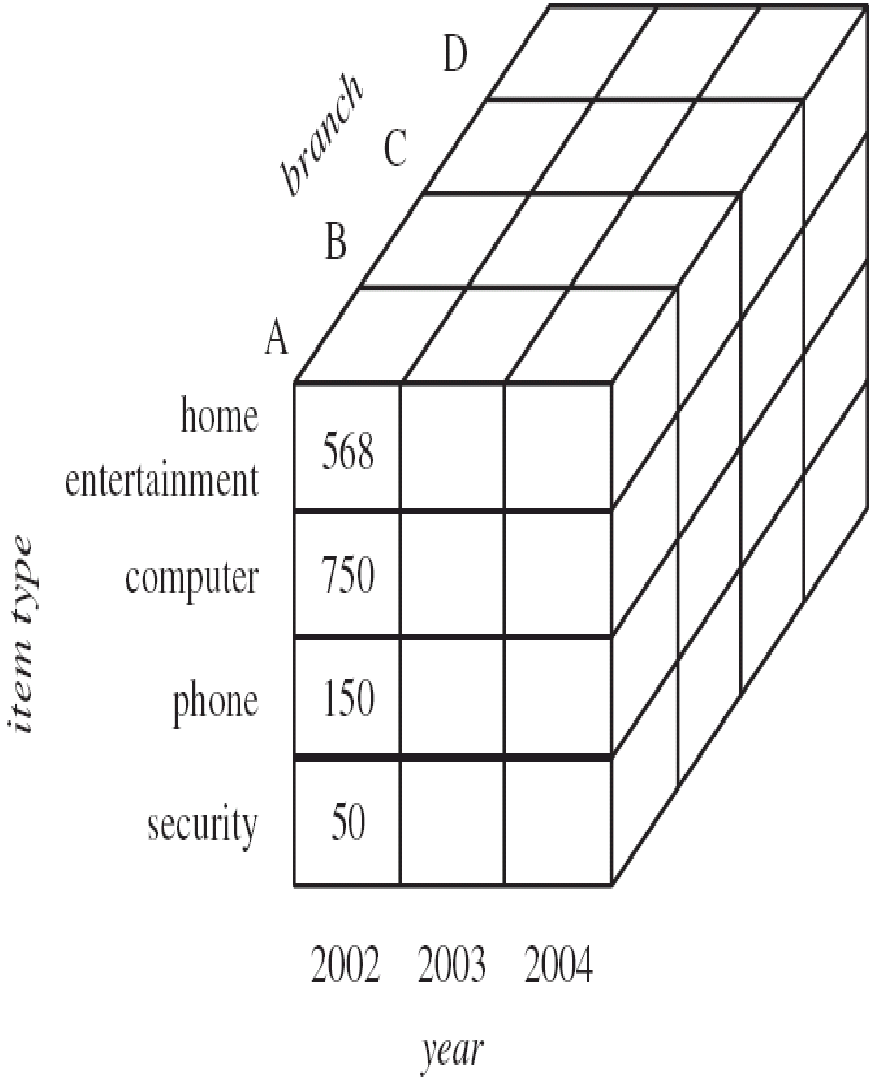

Year 2004
Year 2003

Year 2002

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2002	$1,568,000
2003	$2,356,000
2004	$3,594,000

Fig.-Data cube for sales at Allelectronics



Data Preprocessing: Data Reduction
Data Cube Aggregation

>-Multiple levels of aggregation in data cubes

-Further reduce the size of data to deal with

>-Reference appropriate levels

-Use the smallest representation capable to solve the task

>-Queries regarding aggregated information should be answered using data cube, when possible

>-By using the Data cube, all the quarter sales can be aggregated to yearly sales. Hence the Huge data of quarterly is Reduced to yearly..

>-Multiple levels of aggregation in data cube Further reduce the size of data to deal with Reference appropriate levels.

>-Use the smallest representation capable to solve the task.

>-Queries regarding aggregated information should be answered using data cube, when possible.

Data Preprocessing: Data Reduction
Attribute Subset Selection

>-Problem: 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

--Nice side-effect: reduces # of attributes in the discovered patterns (which are now easier to understand)

>-Solution: Heuristic methods (due to exponential # of choices) usually greedy:

--step-wise forward selection

--step-wise backward elimination

--combining forward selection and backward elimination

--decision-tree induction

Data Preprocessing: Data Reduction

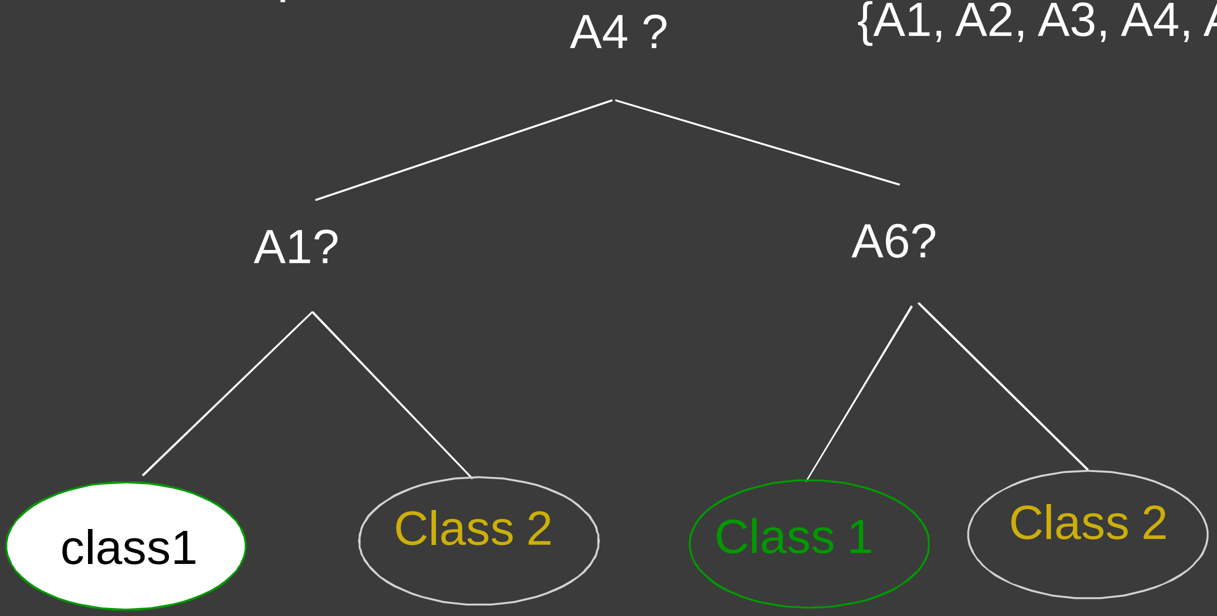
Attribute Subset Selection

nonleaf nodes: tests

branches: outcomes of tests

leaf nodes: class prediction

Initial attribute set:
{A1, A2, A3, A4, A5, A6}



A4 ?

A1?

A6?

class1

Class 2

Class 1

Class 2

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

Data Preprocessing: Data Reduction

Dimensionality Reduction

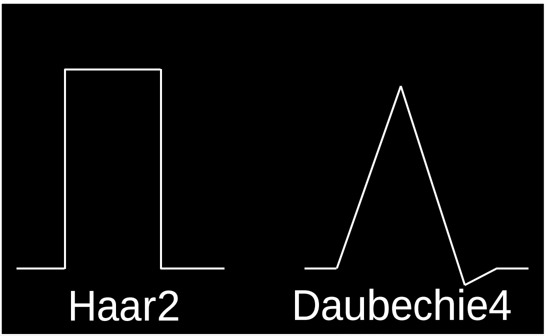
Definition: Data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data.

>-There are two data dimensionality reduction techniques : 1) lossless 2) lossy

>- There are two popular methods of lossy dimensionality reduction:

1) Wavelet transforms 2)Principal Component Analysis

Data Preprocessing: Data Reduction
Dimensionality Reduction



Wavelet Transforms

>-Discrete wavelet transform (DWT) linear signal processing technique.

>-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 (conserves local details)

>-Method (hierarchical pyramid algorithm):

-Length, L, must be an integer power of 2 (padding with 0s, when necessary)

-Each transform has 2 functions:

--smoothing (e.g., sum, weighted avg.), weighted difference

-Applies to pairs of data, resulting in two sets of data of length L/2

-Applies the two functions recursively, until reaches the desired length

Data Preprocessing: Data Reduction

Dimensionality Reduction

Wavelet Transforms

>-A wavelet transformed data usefulness lies in the fact the wavelet transformed data can be truncated.

>-A compressed approximation of the data can be retained by storing only a small faction of the strongest wavelet coefficients.

>-For example , all wavelet coefficients larger than some user-specified threshold can be retained. All other coefficients set to 0

>-Given a set of coefficients , an approximation of the original data can be constructed by applying the inverse of DWT used.

Data Preprocessing: Data Reduction

Dimensionality Reduction

Haar wavelets

>-The first DWT was invented by the Hungarian mathematician Alfréd Haar. For an input represented by a list of 2^n numbers, the Haar wavelet transform may be considered to simply pair up input values, storing the difference and passing the sum. This process is repeated recursively, pairing up the sums to provide the next scale: finally resulting in $2^n - 1$ differences and one final sum.

>- Daubechies wavelets

The most commonly used set of discrete wavelet transforms was formulated by the Belgian mathematician Ingrid Daubechies in 1988. This formulation is based on the use of recurrence relations to generate progressively finer discrete samplings of an implicit mother wavelet function; each resolution is twice that of the previous scale. In her seminal paper, Daubechies derives a family of wavelets, the first of which is the Haar wavelet. Interest in this field has exploded since then, and many variations of Daubechies' original wavelets were developed.

Data Preprocessing: Data Reduction
Dimensionality Reduction

Principal Components Analysis (PCA)Karhunen-Loeve (K-L) method

>-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 (projected) 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 ordered and unordered attributes

>-Used when the number of dimensions is large

Data Preprocessing: Data Reduction

Dimensionality Reduction

The basic procedure of PCA is as follows:

1) The input data are normalized, so that each attribute falls within same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.

2) PCA computes k orthonormal vectors that provides a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to the others, these vectors are referred to as the principal components. The input data are a linear combination of the principal components.

3) The principal components are sorted in order of decreasing “significance” or strength. The principal component essentially serve as a new set of axes for the data ,providing important information bout variance.

4) Because the components are sorted according to decreasing order of “singificance” the size of the data can be reduced by eliminating the weaker components ,that is , those with low variance. Using strongest principal component s,it should be possible to reconstruct a good approximation of the original data.

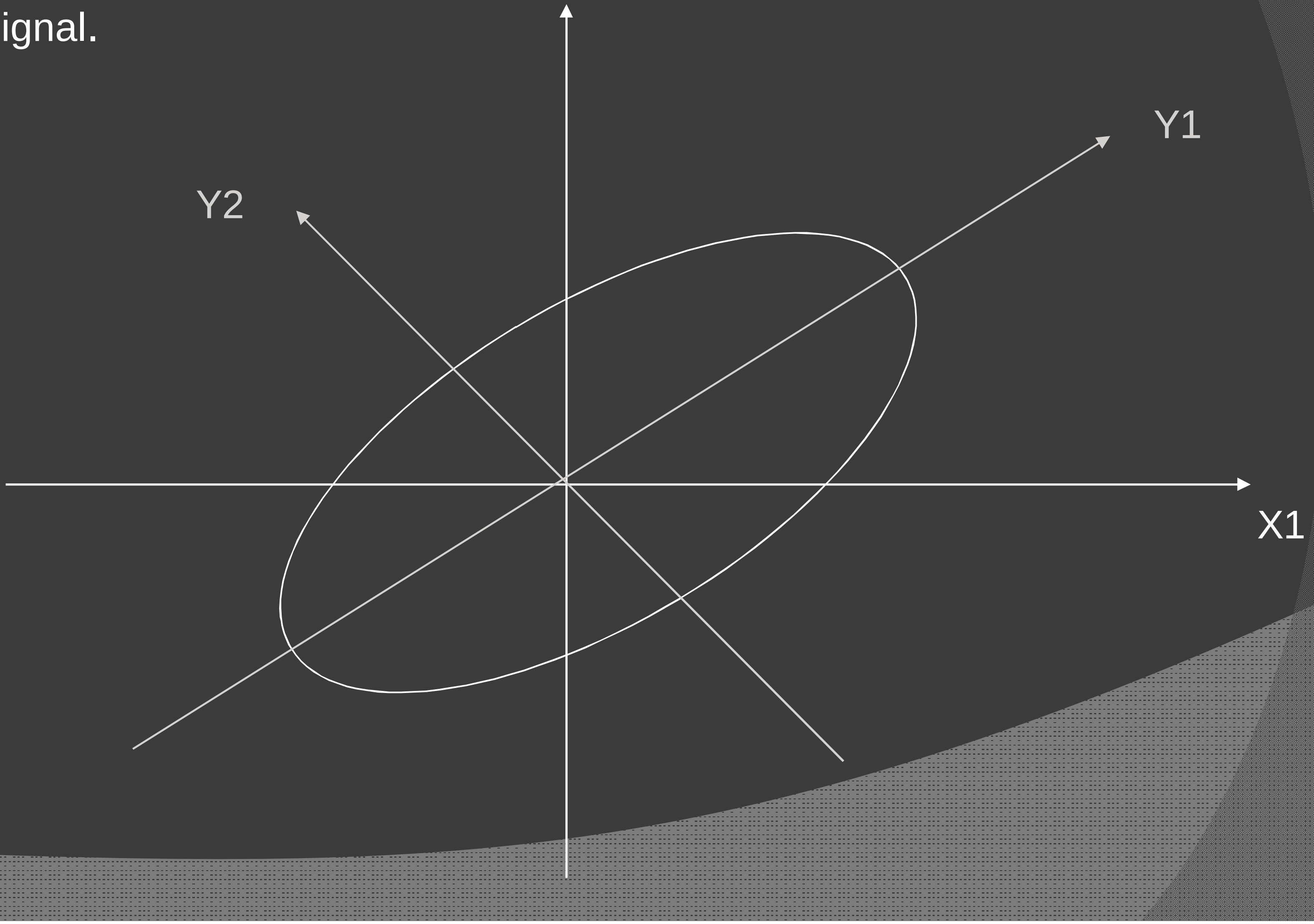
Data Preprocessing: Data Reduction

Dimensionality Reduction

Principal Components Analysis (PCA)Karhunen-Loeve (K-L) method

>-The principal components (new set of axes) give important information about variance.

>-Using the strongest components one can reconstruct a good approximation of the original signal.



Data Preprocessing: Data Reduction

Numerosity Reduction

Definition: Reduce data volume by choosing alternative, smaller forms of data representation.

>-Parametric methods

--Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)

--Example: 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

Example Techniques of Numerosity Reduction are:

1)histograms

2)Clustering

3) sampling

Data Preprocessing: Data Reduction

Numerosity Reduction

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.

Data Preprocessing: Data Reduction

Numerosity Reduction

Regression and Log-Linear Models

>-Linear regression: $Y = a + b X$

Two parameters , a and b specify the line and are to be estimated by using the data at hand.

>-Using the least squares criterion to the known values of *Y1, Y2, …, X1, X2, ….*

Multiple régression: $Y = b0 + b1 X1 + b2 X2.$

>-Many nonlinear functions can be transformed into the above.

>-Log-linear models:

--The multi-way table of joint probabilities is

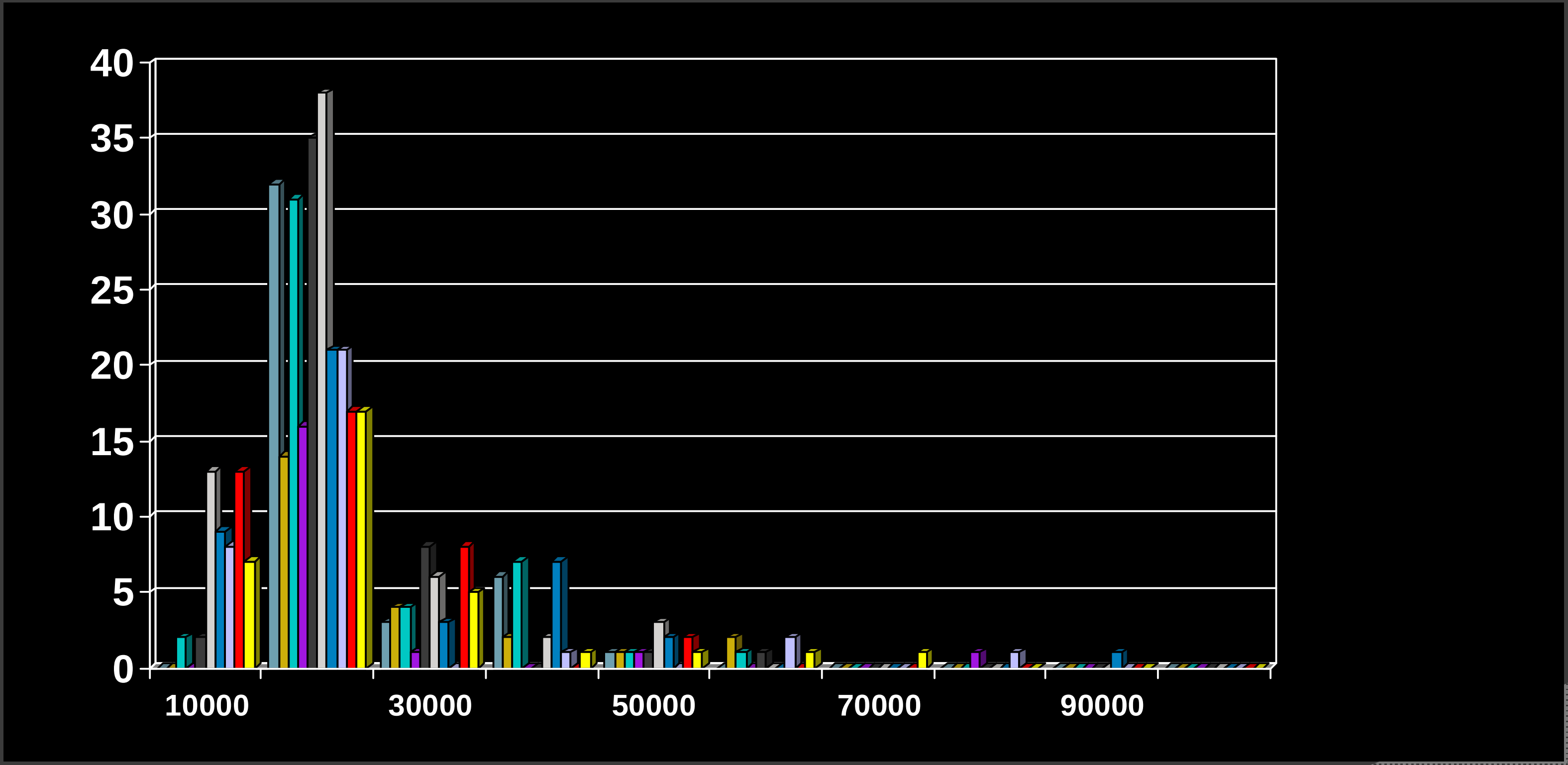
--approximated by a product of lower-order tables.

>-Probability: *p(a, b, c, d) = aab baccad dbcd*

Data Preprocessing: Data Reduction

Numerosity Reduction

Histograms: Histrogram for an attribute ,A,partitions the data distribution of A into disjoint subsets, or buckets.



>-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.

Data Preprocessing: Data Reduction

Numerosity Reduction

Histograms

>-Divide data into buckets and store average (sum) for each bucket

>-Partitioning rules:

-Equal-width: equal bucket range

-Equal-frequency (or equal-depth)

-V-optimal: with the least *histogram variance* (weighted sum of the original values that each bucket represents)

-MaxDiff: set bucket boundary between each pair for pairs have the β–1 largest differences

Data Preprocessing: Data Reduction

Numerosity Reduction

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 Coming up unit 3

Data Preprocessing: Data Reduction

Numerosity Reduction

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 (orsubpopulation of interest) in the overall database

>-Used in conjunction with skewed data.

>-Sampling may not reduce database I/Os (page at a time

Data Preprocessing: Data Reduction

Numerosity Reduction

Sampling

>-Sampling: obtaining a small sample *s* to represent the whole data set *N*

-Simple random sample without replacement

-Simple random sample with replacement

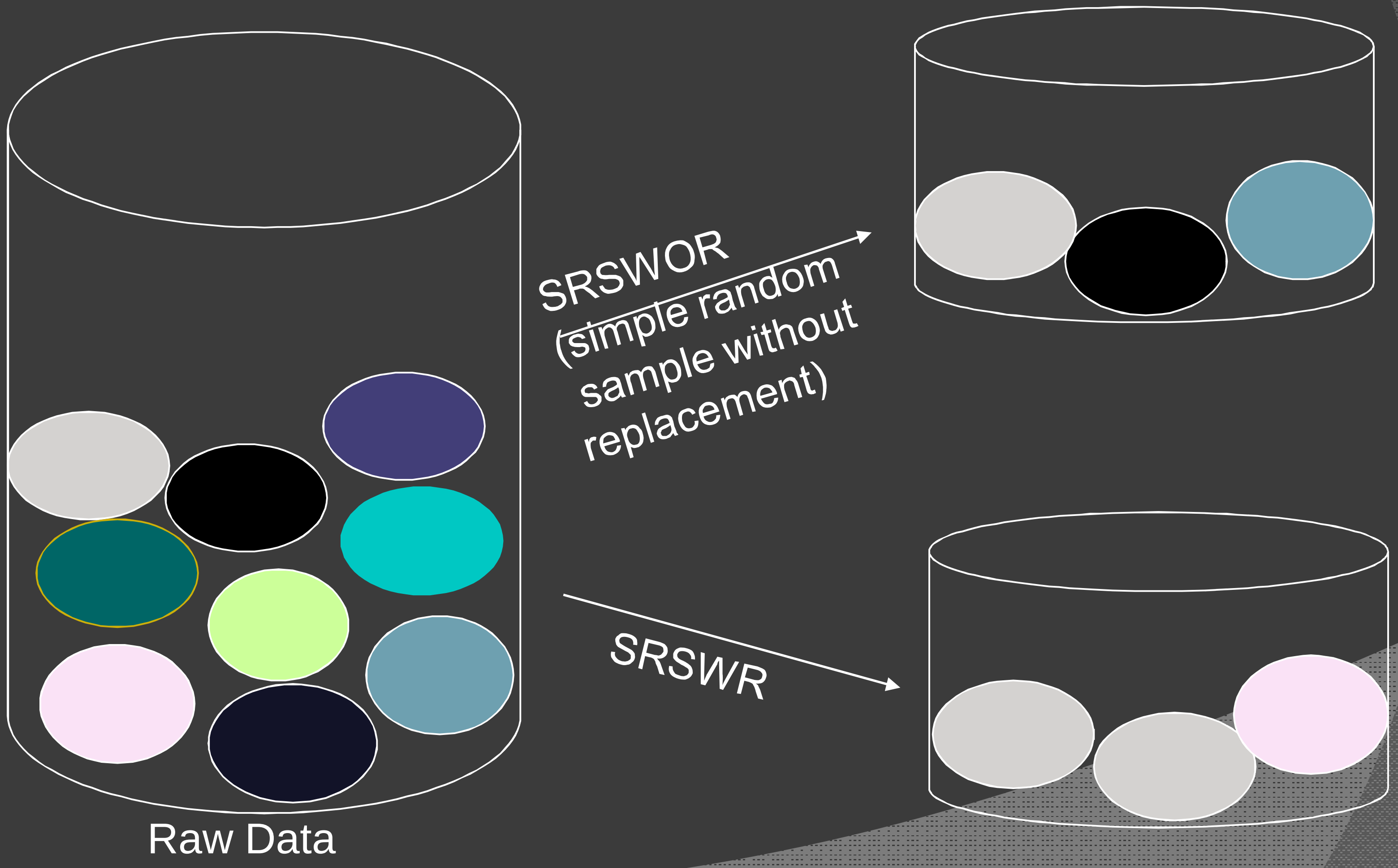
-Cluster sample: if the tuples in D are grouped into M mutually disjoint clusters, then an Simple Random Sample can be obtained, where s < M

-Stratified sample

Data Preprocessing: Data Reduction

Numerosity Reduction

Sampling

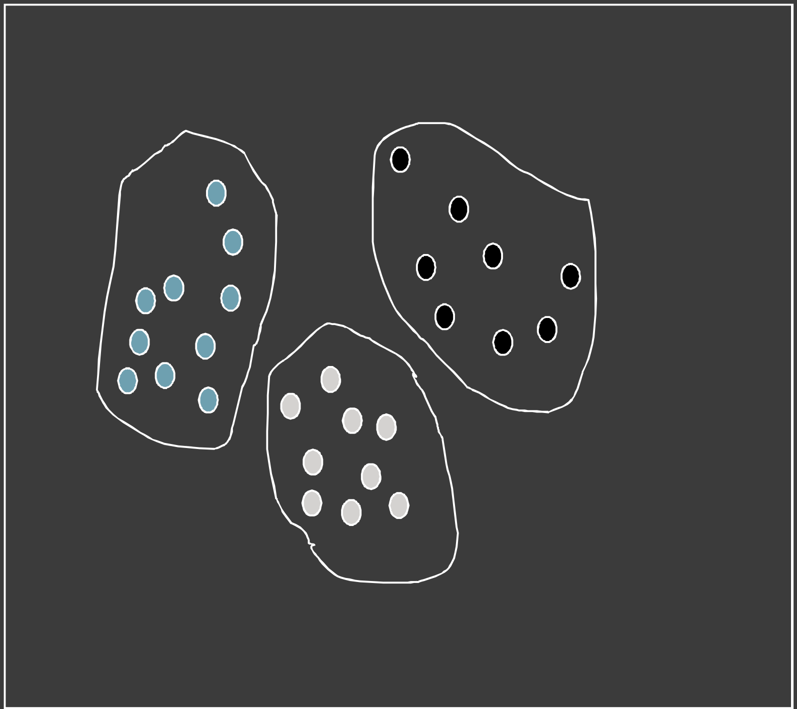


Data Preprocessing: Data Integration & Transformation

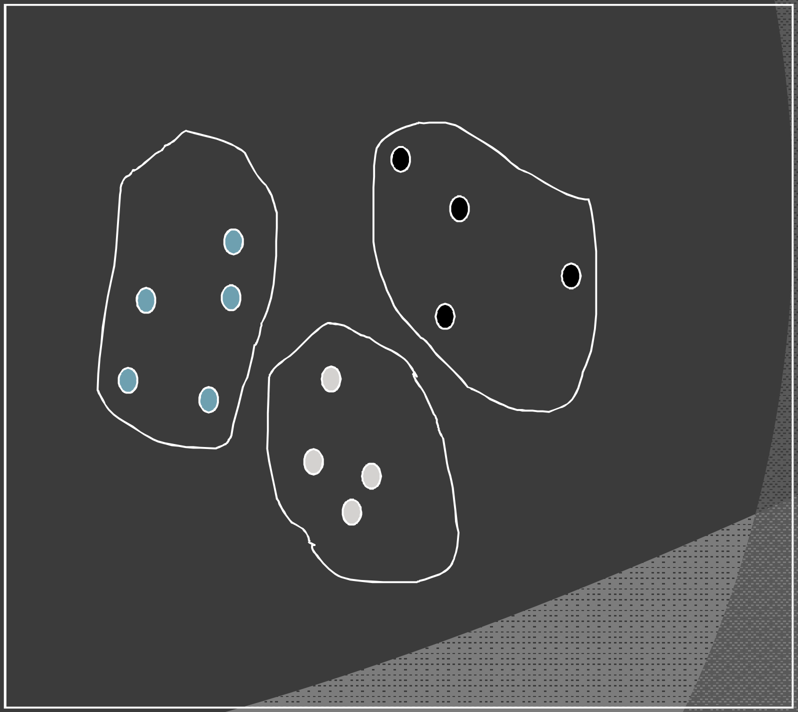
Data Reduction

Sampling

Raw Data



Cluster/Stratified Sample



Data Preprocessing: Discretization and concept hierarchy generation

>-Three types of attributes:

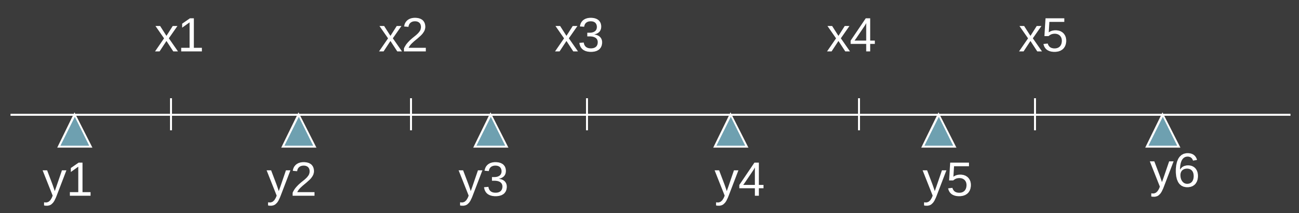
-Nominal — values from an unordered set

-Ordinal — values from an ordered set

-Continuous — real numbers

>-Discretization/Quantization:

-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.

Data Preprocessing: Discretization and concept hierarchy generation

>-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).

Data Preprocessing: Discretization and concept hierarchy generation-Discretization and concept hierarchy generation for numeric data

>-Hierarchical and recursive decomposition using:

-Binning (data smoothing)

-Histogram analysis (numerosity reduction)

-Clustering analysis (numerosity reduction)

>-Entropy-based discretization

>-Segmentation by natural partitioning

Data Preprocessing: Discretization and concept hierarchy generation

Discretization and concept hierarchy generation for numeric data

Entropy-Based Discretization

>-Given a set of samples S, if S is partitioned into two intervals S1 and S2 using threshold T on the value of attribute A, the information gain resulting from the partitioning is:

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

>-where the entropy function E for a given set is calculated based on the class distribution of the samples in the set. Given m classes the entropy of S1 is:

$$E(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

>where pi is the probability of class i in S1.

-The threshold that maximizes the information gain over all possible thresholds is selected as a binary discretization.

-The process is recursively applied to partitions obtained until some stopping criterion is met, e.g.,

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

Data Preprocessing: Discretization and concept hierarchy generation -for numeric data

Interval Merging by X^2 Analysis

>-ChiMerge employs a bottom-up approach by finding best neighboring intervals and then merging these to form larger intervals, recursively.

>-This method is supervised in that it uses class information.

>-if two adjacent intervals have a very similar distribution of classes , then the intervals can be merged else they are separate.

>-ChiMerge proceeds as follows:

1)each distinct value of a numerical attribute A is considered to be one interval.

2)X2 tests are performed for every pair of adjacent intervals.

3) adjacent intervals with the least X^2 values are merged together, because low X^2 values for a pair indicate similar class distributions.

4)This merging proceed recursively until a predefined criterion is met.

Data Preprocessing: Discretization and concept hierarchy generation for numeric data

Interval Merging by X^2 Analysis

>-The stopping criterion is typically determined by three conditions:

1) Merging stops when X^2 values of all pairs of adjacent intervals exceed some threshold, which is determined by a specified significance level.

2) The no. of intervals can not be over prescribed max-interval, such as 10 to 15.

3) chiMerge is that the relative class frequencies should be fairly consistent with in interval.

Data Preprocessing: Discretization and concept hierarchy generation for numeric data

Discretization by intuitive partitioning

>-3-4-5 rule can be used to segment numeric data into relatively uniform, "natural" intervals.

>-It partitions a given range into 3,4, or 5 equiwidth intervals recursively level-by-level based on the value range of the most significant digit.

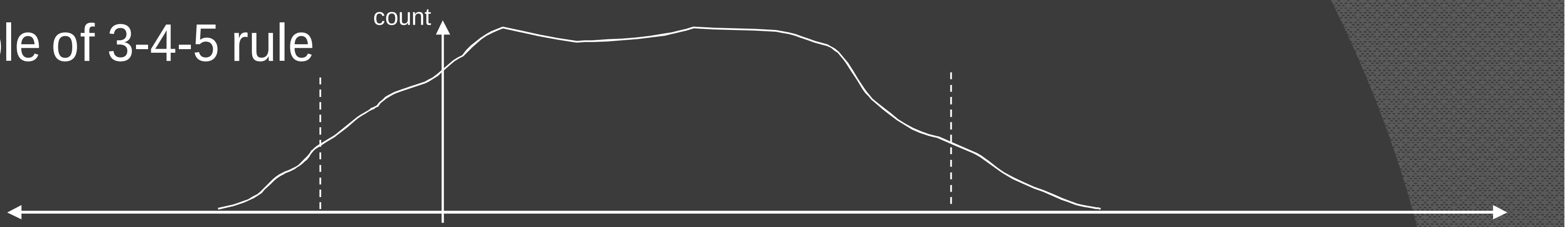
>-If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equi-width intervals

>-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

Data Preprocessing: Discretization and concept hierarchy generation for numeric data

Example of 3-4-5 rule

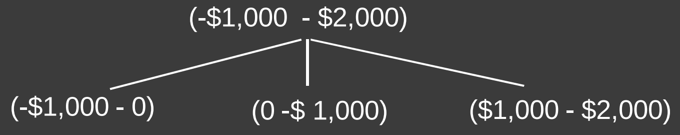


Step 1: -$351 -$159 profit $1,838 $4,700

Min Low (i.e, 5%-tile) High(i.e, 95%-0 tile) Max

Step 2: msd=1,000 Low=-$1,000 High=$2,000

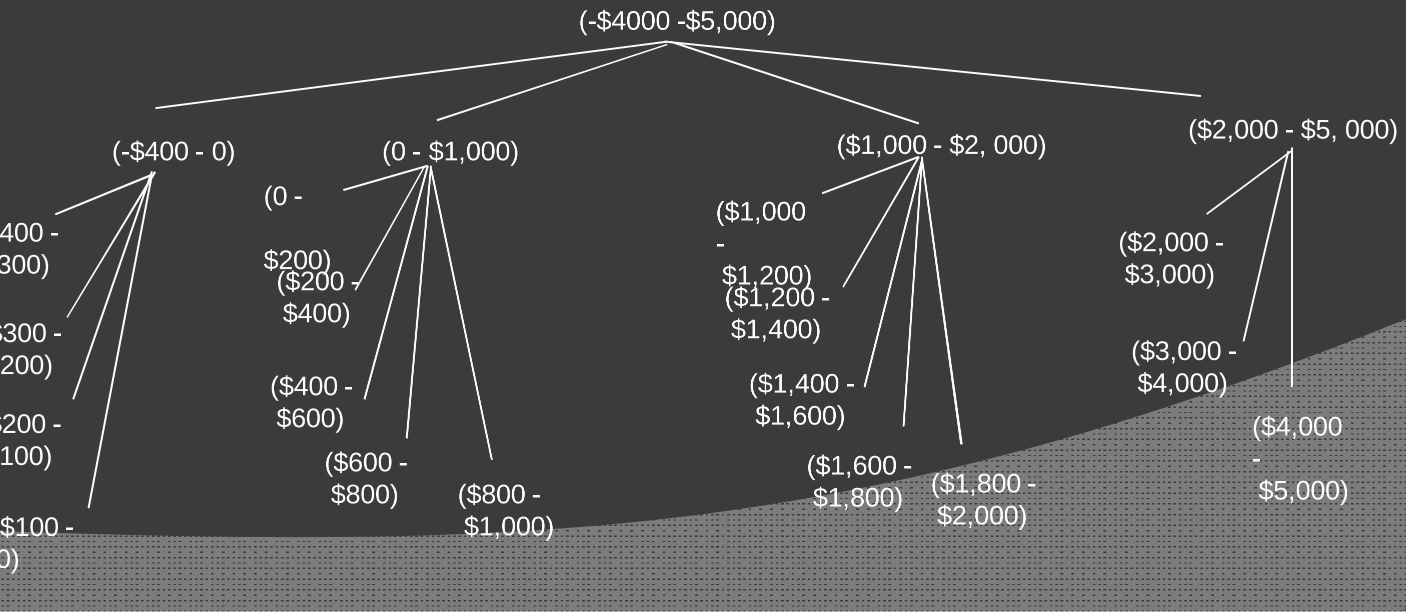
Step 3:



(-$1,000 - $2,000)

(-$1,000 - 0) (0 -$ 1,000) ($1,000 - $2,000)

Step 4:



(-$4000 -$5,000)

(-$400 - 0)

(-$400 - -$300)

(-$300 - -$200)

(-$200 - -$100)

(-$100 - 0)

(0 - $1,000)

(0 - $200)

($200 - $400)

($400 - $600)

($600 - $800)

($800 - $1,000)

($1,000 - $2, 000)

($1,000 - $1,200)

($1,200 - $1,400)

($1,400 - $1,600)

($1,600 - $1,800)

($1,800 - $2,000)

($2,000 - $5, 000)

($2,000 - $3,000)

($3,000 - $4,000)

($4,000 - $5,000)

Data Preprocessing: Discretization and concept hierarchy generation for categorical data

>-Categorical data: no ordering among values

>-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

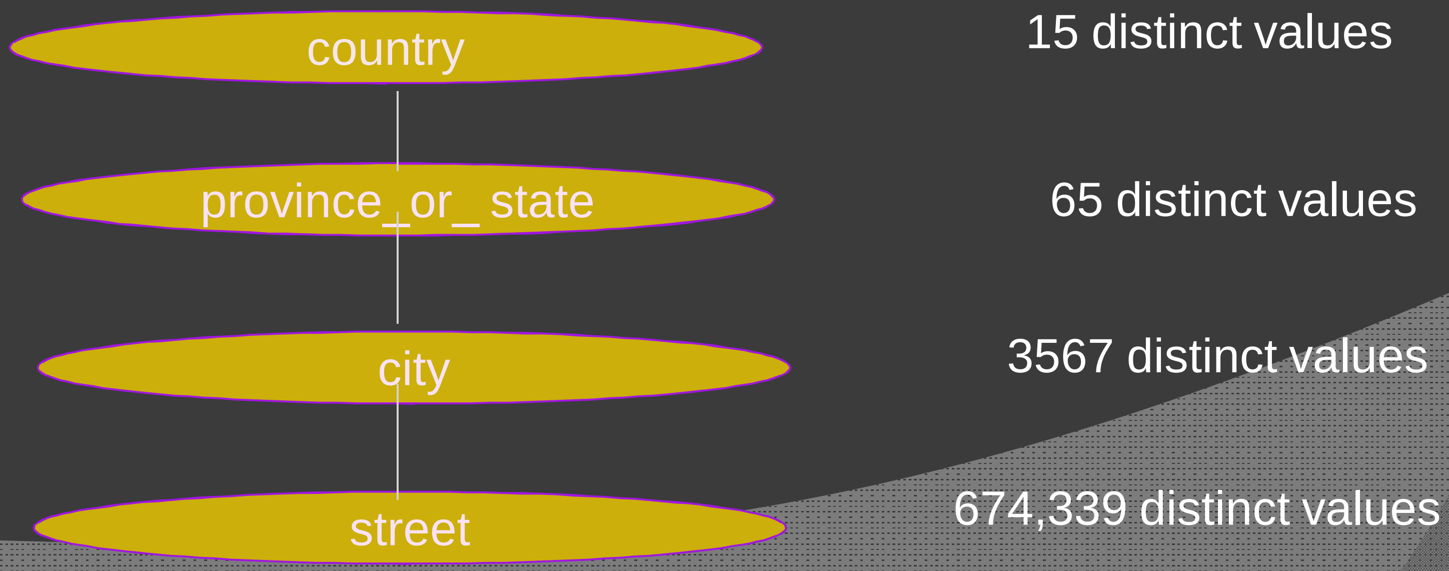
>-Specification of a set of attributes, but not of their partial ordering

>-Specification of only a partial set of attributes

Data Preprocessing: Discretization and concept hierarchy generation for categorical data

>-Concept hierarchy generation w/o data semantics - 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 (limitations?)



country

15 distinct values

province_or_ state

65 distinct values

city

3567 distinct values

street

674,339 distinct values

End of Part-1 of 2 unit