Hes·so
Haute Ecole Spécialisée
de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences
Western Switzerland

"Data Preprocessing"

MAS-ICT

Professor:
Dr. Laura E. RAILEANU

heig-vd
Haute Ecole d'Ingénierie et de Gestion
du Canton de Vaud

Reference

- Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”, 3rd edition, Morgan Kaufmann, 2011.

Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Quality: Why Preprocess the Data?

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

Accuracy, completeness, consistency (1|2)

- E.g.,
  - You are a manager at *AllElectronics* and have been charged with analyzing the company's data with respect to the sales at your branch.
  - You identify and select the attributes to be included in your analysis (*item*, *price*, and *units sold*).
  - You notice that several of the attributes for various tuples have no recorded value.
  - For your analysis, you would like to include information as to whether each item purchased was advertised as on sale, yet you discover that this information has not been recorded.
  - Users of your database system have reported errors, unusual values, and inconsistencies in the data recorded for some transactions.

Accuracy, completeness, consistency (2|2)

- The data you wish to analyze by data mining techniques are:
  - *incomplete* (lacking attribute values or certain attributes of interest, or containing only aggregate data)
  - *inaccurate* or *noisy* (containing errors, or *outlier* values that deviate from the expected)
  - *inconsistent* (e.g., containing discrepancies in the department codes used to categorize items)

Timeliness

- E.g.,
  - Suppose that you are overseeing the distribution of monthly sales bonuses to the top sales representatives at *AllElectronics.*
  - Several sales representatives, however, fail to submit their sales records on time at the end of the month. There are also a number of corrections and adjustments that flow in after the month's end.
  - For a period of time following each month, the data stored in the database is incomplete.
  - However, once all of the data is received, it is correct.
  - The fact that the month-end data is not updated in a timely fashion has a negative impact on the data quality.

Believability

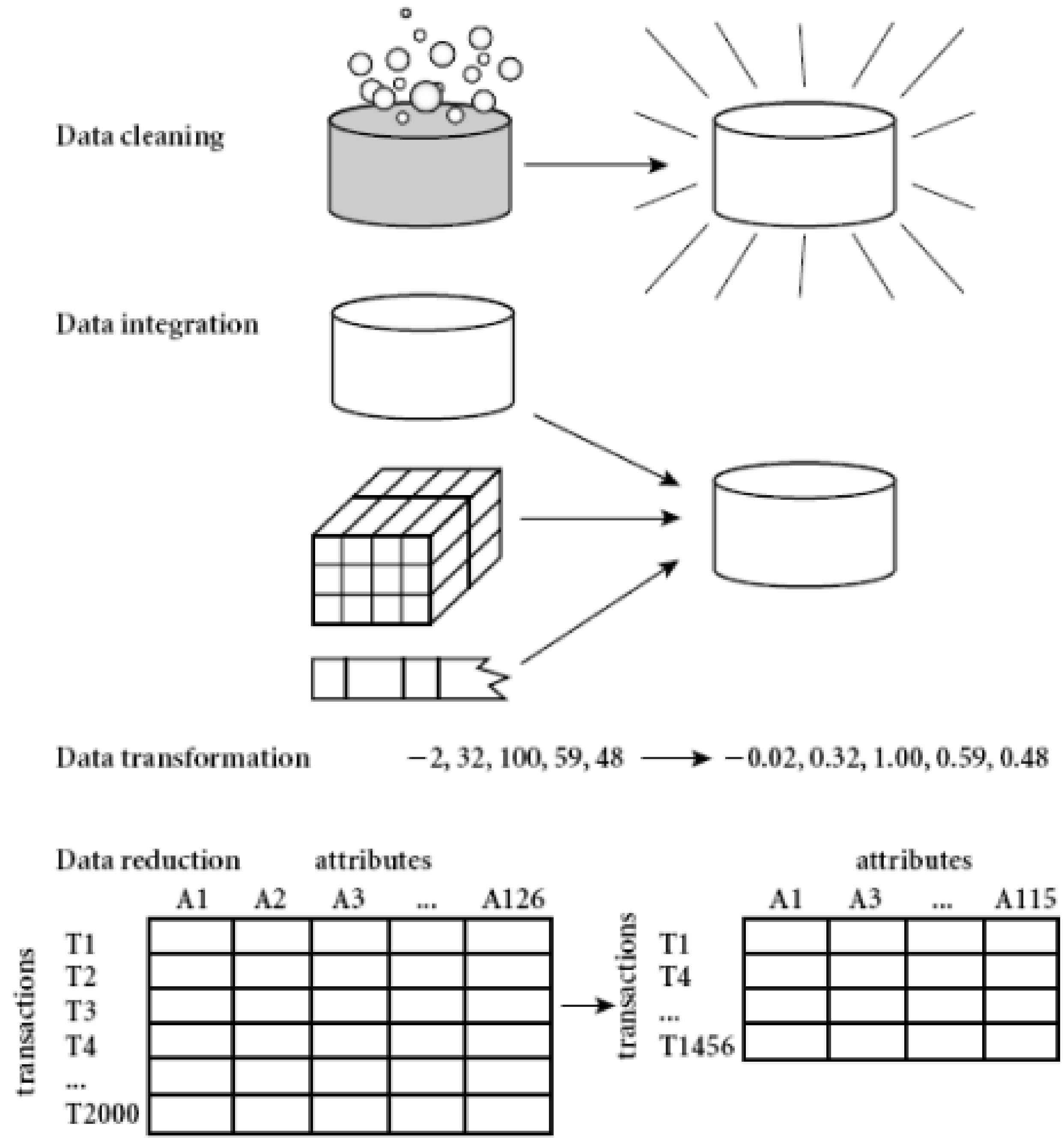
- E.g.,
  - A few years back, a programming error miscalculated the sales commissions for its sales representatives so that these employees received 20% less than was due.
  -
  - The software bug was quickly fixed and the data corrected.
  - Even though the database is now accurate, complete, consistent, and timely, it is still not trusted because of the memory users have of the past error.
  - Sales managers prefer to compute the expected commissions by hand based on their employees hand-submitted reports rather than believe the data stored in the database.

Interpretability

- E.g.,
  - Consider a sales database, where it is common to create "adjustment" orders to handle complaints and returns.
  - This procedure assigns new order numbers to the adjustment and replacement orders.
  - The accounting department knows how to interpret the resulting data.
  - A business analyst may have a hard time understanding the data, thinking that each order number represents a distinct order.
  -
  - Thus, to the business analyst, the data is of low quality due to poor interpretability.

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 reduction (a reduced representation of data set, smaller in volume, producing the same (almost) analytical results)**
  - Dimensionality reduction (by applying data encoding schemes)
  - Numerosity reduction (replace data by alternative, smaller representations using parametric or non parametric models
  - Data compression
- **Data transformation and data discretization**
  - Normalization (scale data to ranges)
  - Concept hierarchy generation


Data cleaning
Data integration
Data transformation
−2, 32, 100, 59, 48
−0.02, 0.32, 1.00, 0.59, 0.48
Data reduction
attributes
A1 A2 A3 ... A126
transactions
T1
T2
T3
T4
...
T2000
attributes
A1 A3 ... A115
transactions
T1
T4
...
T1456

Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Cleaning

- Data in the real world is dirty:
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=" " (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary*="−10" (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*="42" *Birthday*="03/07/1997"
    - was rating "1,2,3", now rating "A, B, C"
    - discrepancy between duplicate records

Incomplete (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 (when doing classification); not effective when the % 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 be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

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

Binning

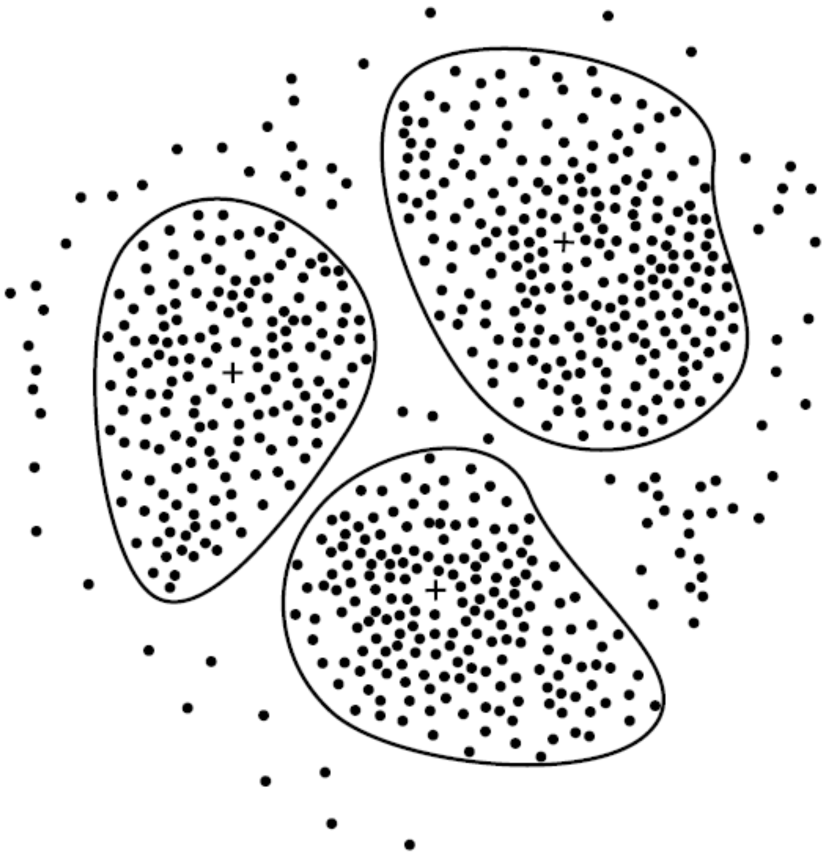
- Smooth a sorted data value by consulting its "neighborhood" ; the sorted values are distributed into a number of "buckets" *(bins)*
- Data are first sorted and then partitioned into:
  - *equal-frequency* bins of same size (each bin contains the same number of values)
  - *equal-width* bins (the interval range of values in each bin is constant)
- Smoothing is done by *bin means*, *bin medians* or *bin boundaries*
- Smooth methods perform *local* smoothing and the larger the width of bin is, the greater is the effect of the smoothing.

Regression

- *Linear regression* involves finding the “best” line to fit two attributes (or variables), so that one attribute can be used to predict the other.

- *Multiple linear regression* is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

Clustering



Data Cleaning as a Process (1|5)

- Data discrepancy detection
  - Use **metadata**: the data type and domain of each attribute, the acceptable values for each attribute
  - Basic **statistical data descriptions** to grasp data trends and identify anomalies (values that are more than two standard deviations away from the mean for a given attribute may be flagged as potential outliers)
  - Write your **own scripts**

Data Cleaning as a Process (2|5)

- Data discrepancy detection
  - Inconsistent use of codes
    - e.g., “2010/12/25” and “25/12/2010” for *date*
  - Field overloading, when developers squeeze new attribute definitions into unused (bit) portions of already defined attributes
    - e.g., using an unused bit of an attribute whose value range uses only, say, 31 out of 32 bits

Data Cleaning as a Process (3|5)

- Data discrepancy detection
  - Data should be examined regarding:
    - unique rules: each value of the given attribute must be different from all other values for that attribute
    - consecutive rules: there can be no missing values between the lowest and highest values for the attribute and that all values must also be unique (e.g., as in check numbers)
    - null rules: specify the use of blanks, question marks, special characters, or other strings that may indicate the null condition (e.g., where a value for a given attribute is not available), and how such values should be handled

Data Cleaning as a Process (4|5)

- Use commercial tools to detect discrepancy
  - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
  - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)

Data Cleaning as a Process (5|5)

- Use commercial tools for data transformation step
    - Data migration tools: allow simple transformations to be specified (e.g., replace the string "gender" by "sex)
    - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
    - They support only a restricted set of transforms: write custom scripts
- Integration of the two processes (discrepancy detection and transformation)
    - Iterative and interactive (e.g., Potter's Wheels):
        - *http://control.cs.berkeley.edu/abc*

Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Integration

- Combines data from multiple sources into a coherent store
- Schema integration
  - Integrate metadata from different sources
  - e.g., A.cust-id ≡ B.cust-#
- Entity identification problem
  - Identify real world entities from multiple data sources
  - e.g., Bill Clinton = William Clinton
- 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
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Nominal Data)

- **X^2 (chi-square) test**

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- The X^2 statistic tests the hyp that A and B are independent. The test is based on a significance level , with (r-1)x(c-1)= (2-1)x(2-1)=1 degrees of freedom. For 1 degree of freedom, the X^2 value needed to reject the hypothesis at 0.001 significance level is 10.828. Since 507.93>10.828, we can reject the hypothesis that like_science_fiction and play_chess are independent and conclude that they are strongly correlated for the given group of people.

- 90=(300*450)/1500;360=(450*1200)/1500
- 210=(300*1050)/1500; 840=1200*1050/1500

Correlation Analysis (Numeric Data)

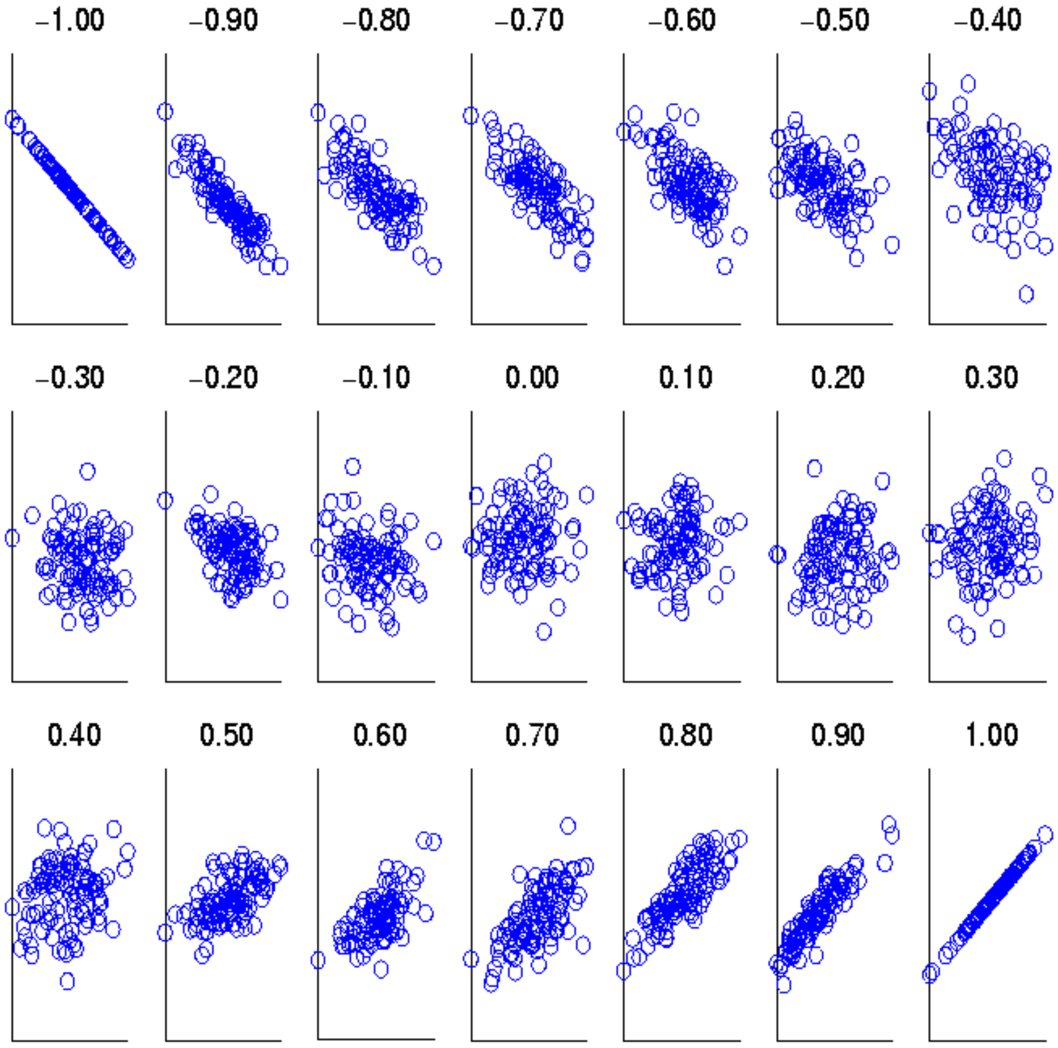
- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{p,q} = \frac{\sum (p-\bar{p})(q-\bar{q})}{(n-1)\sigma_p \sigma_q} = \frac{\sum (pq) - n\bar{p}\bar{q}}{(n-1)\sigma_p \sigma_q}$$

  where n is the number of tuples, $\bar{p}$ and $\bar{q}$ are the respective means of p and q, σ_p and σ_q are the respective standard deviation of p and q, and $\Sigma(pq)$ is the sum of the pq cross-product.

- $r_{p,q} > 0$, p and q are positively correlated (p's values increase as q's). $r_{p,q} = 0$: independent; $r_{pq} < 0$: negatively correlated
- $-1 \le r_{p,q} \le 1$

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Correlation (viewed as linear relationship)

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

Covariance (Numeric Data)

- Covariance is similar to correlation

$$Cov(p,q) = E((p-\bar{p})(q-\bar{q})) = \frac{\sum_{i=1}^{n}(p_i-\bar{p})(q_i-\bar{q})}{n}$$

$$r_{p,q} = \frac{Cov(p,q)}{\sigma_p \sigma_q}$$

where n is the number of tuples, $\bar{p}$ and $\bar{q}$ are the respective mean or **expected values** of p and q, σ_p and σ_q are the respective standard deviation of p and q.

- **Positive covariance**: If $Cov_{p,q} > 0$, then p and q both tend to be larger than their expected values.
- **Negative covariance**: If $Cov_{p,q} < 0$ then if p is larger than its expected value, q is likely to be smaller than its expected value.
- **Independence**: $Cov_{p,q} = 0$ but the converse is not true:
  - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Co-Variance: An Example

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
  - E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 − 4 × 9.6 = 4
- Thus, A and B rise together since Cov(A, B) > 0.

Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Reduction Strategies

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
  - **Dimensionality reduction**, e.g., remove unimportant attributes
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature creation
  - **Numerosity reduction** (some simply call it: Data Reduction)
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - **Data compression**

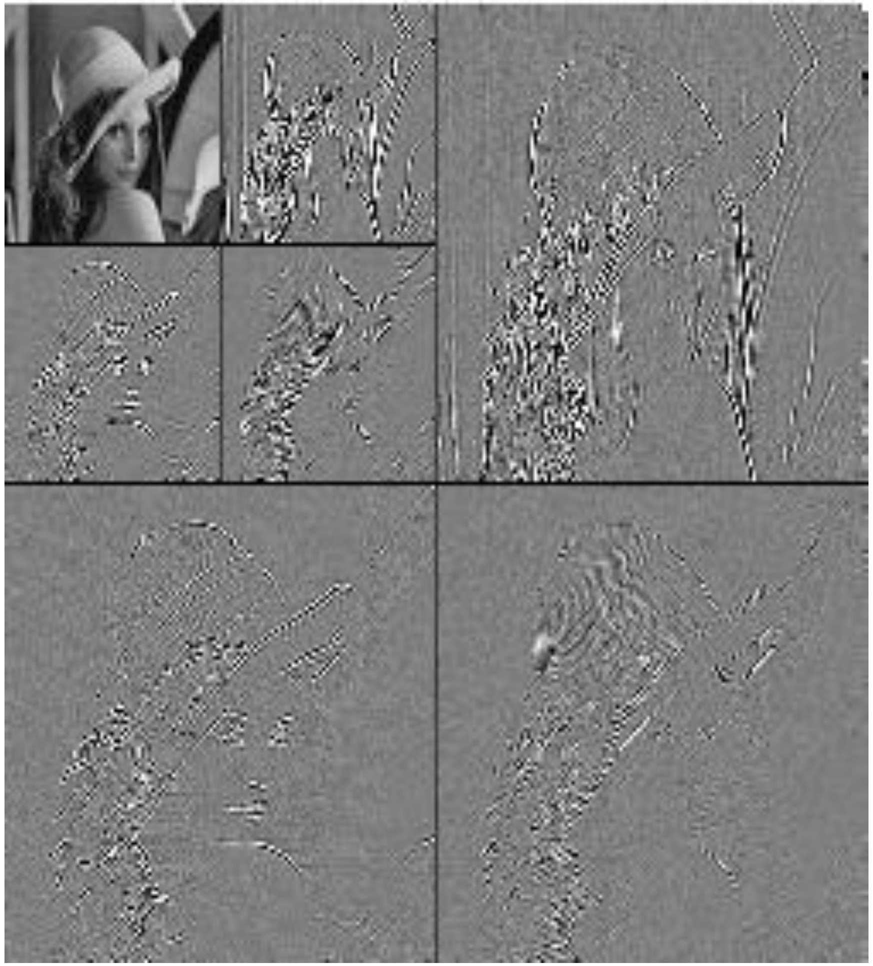
Data Reduction 1: Dimensionality Reduction

- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization
- **Dimensionality reduction techniques**
  - Wavelet transforms
  - Principal Component Analysis
  - Supervised and nonlinear techniques (e.g., feature selection)

38

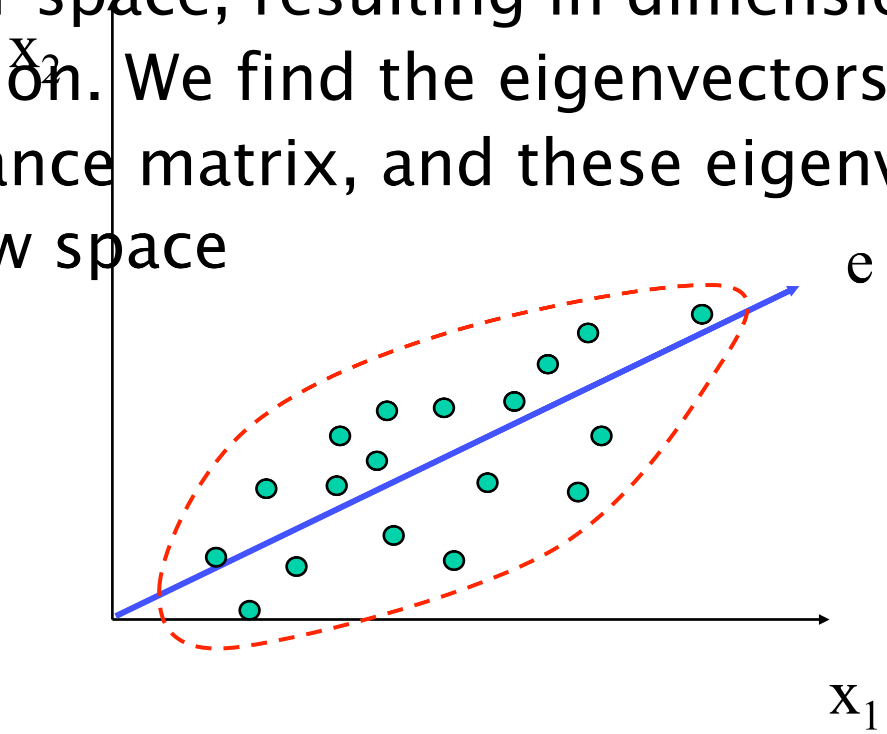
What Is Wavelet Transform?

- Decomposes a signal into different frequency subbands
  - Applicable to n-dimensional signals
- Data are transformed to preserve relative distance between objects at different levels of resolution
- Allow natural clusters to become more distinguishable
- Used for image compression



Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



Attribute Subset Selection

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

Heuristic Search in Attribute Selection

- There are 2^d possible attribute combinations of d attributes
- Typical heuristic attribute selection methods:
  - Best single attribute under the attribute independence assumption: choose by significance tests
  - Best step-wise feature selection:
    - The best single-attribute is picked first
    - Then next best attribute condition to the first, ...
  - Step-wise attribute elimination:
    - Repeatedly eliminate the worst attribute
  - Best combined attribute selection and elimination

Attribute Creation (Feature Generation)

- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
  - Attribute extraction
    - domain-specific
  - Mapping data to new space (see: data reduction)
    - E.g., Fourier transformation, wavelet transformation
  - Attribute construction
    - Combining features
    - Data discretization

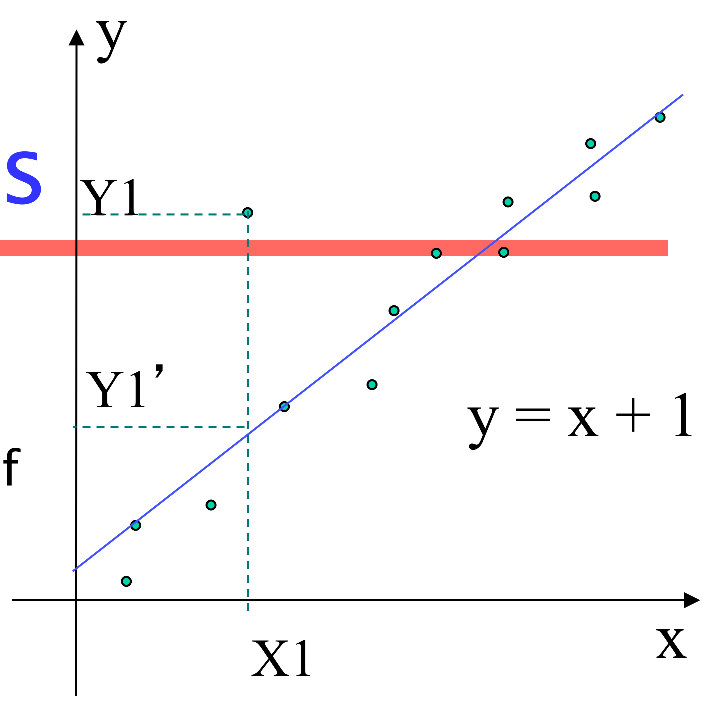
Data Reduction 2: Numerosity Reduction

- Reduce data volume by choosing alternative, *smaller forms* of data representation
- **Parametric methods** (e.g., regression)
  - 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
  - Major families: histograms, clustering, sampling, …

Parametric Data Reduction: 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

Regression Analysis



- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a *dependent variable* (also called *response variable* or *measurement*) and of one or more *independent variables* (*explanatory variables* or *predictors*)
- The parameters are estimated so as to give a "best fit" of the data
- Most commonly the best fit is evaluated by using the *least squares method*, but other criteria have also been used

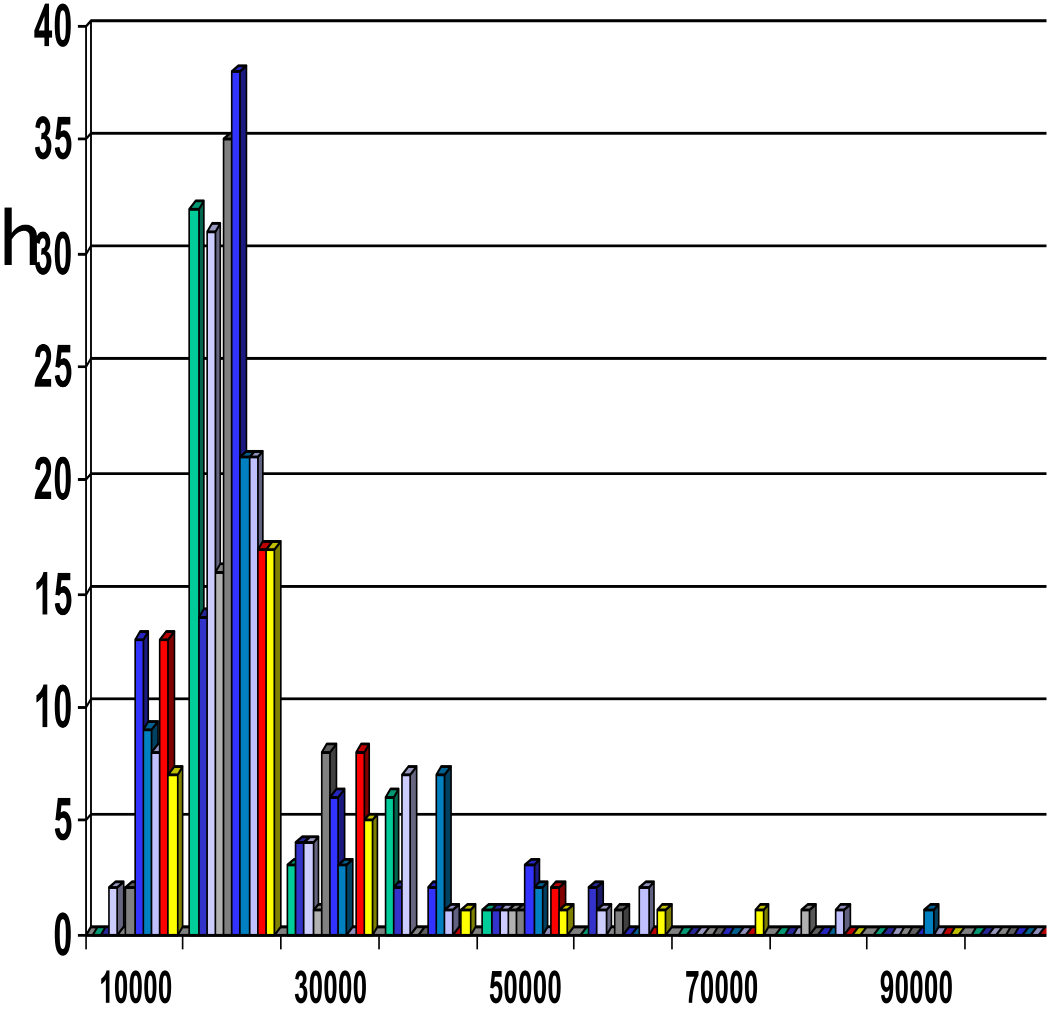
- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Regress Analysis and Log-Linear Models

- Linear regression: $Y = w\,X + b$
  - Two regression coefficients, *w* 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 Y_1, Y_2, …, X_1, X_2, ….
- 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

Histogram Analysis

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)



Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms

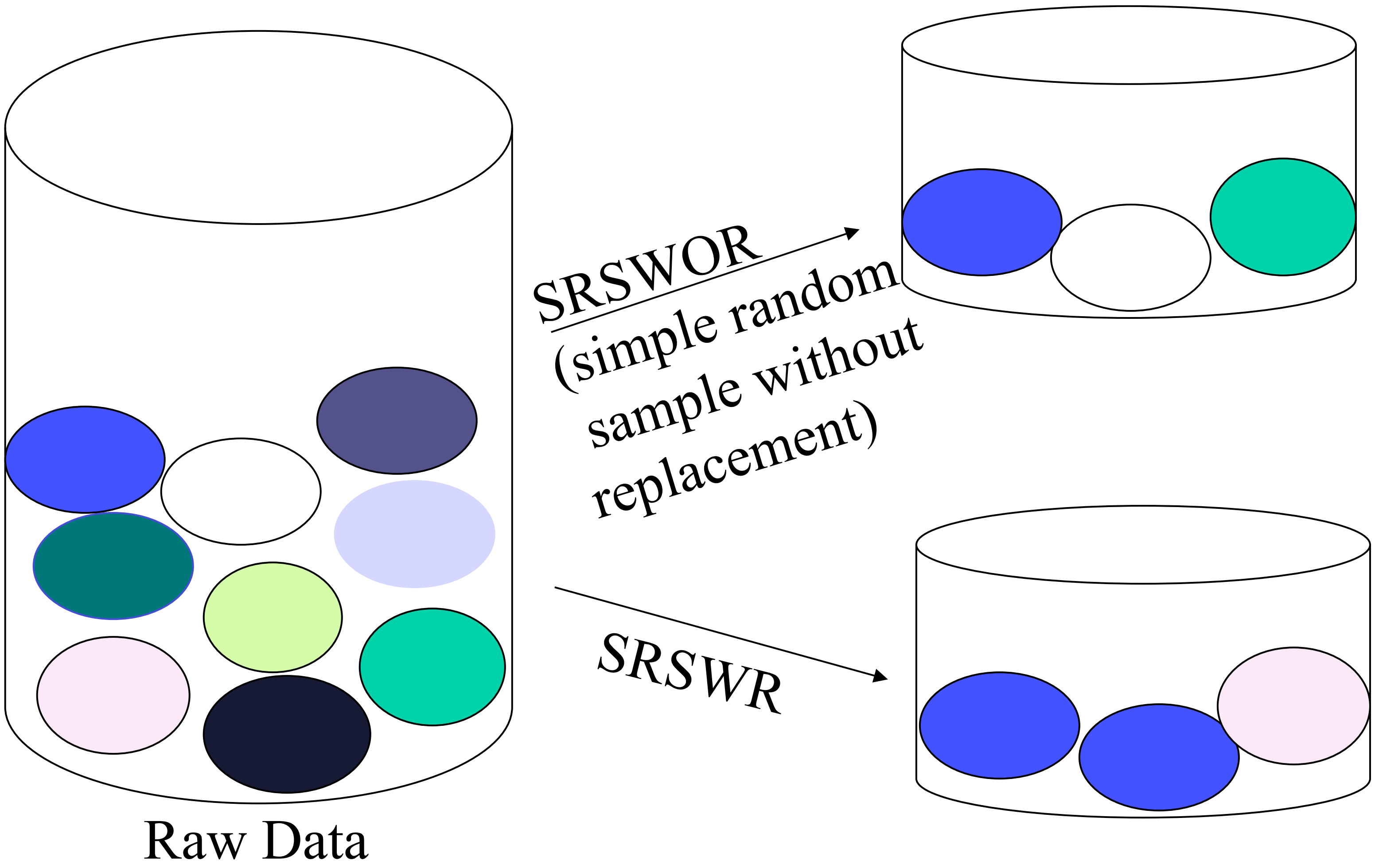
Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling:
- Note: Sampling may not reduce database I/Os

Types of Sampling

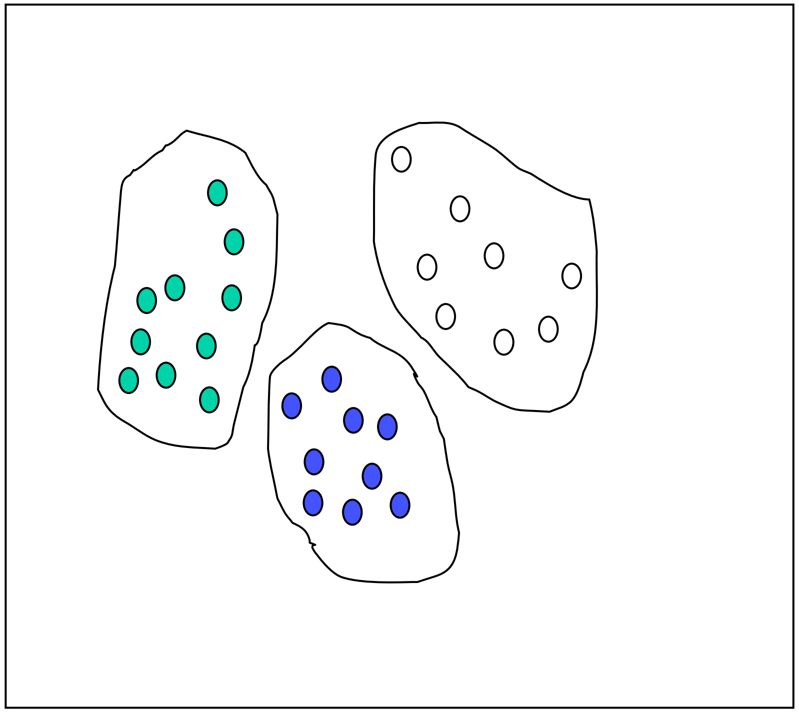
- **Simple random sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement**
  - A selected object is not removed from the population
- **Stratified sampling:**
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

Sampling: With or without Replacement

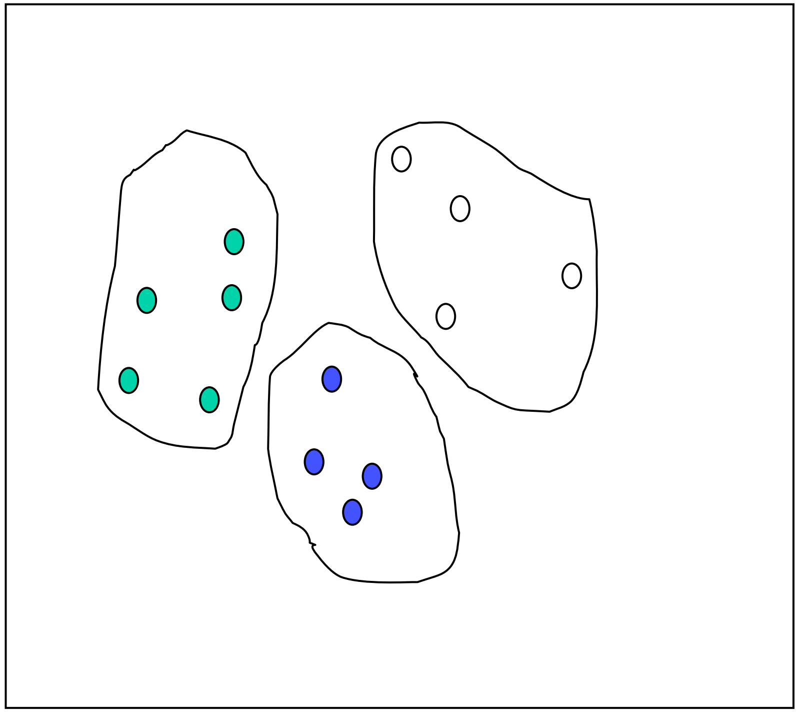


Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



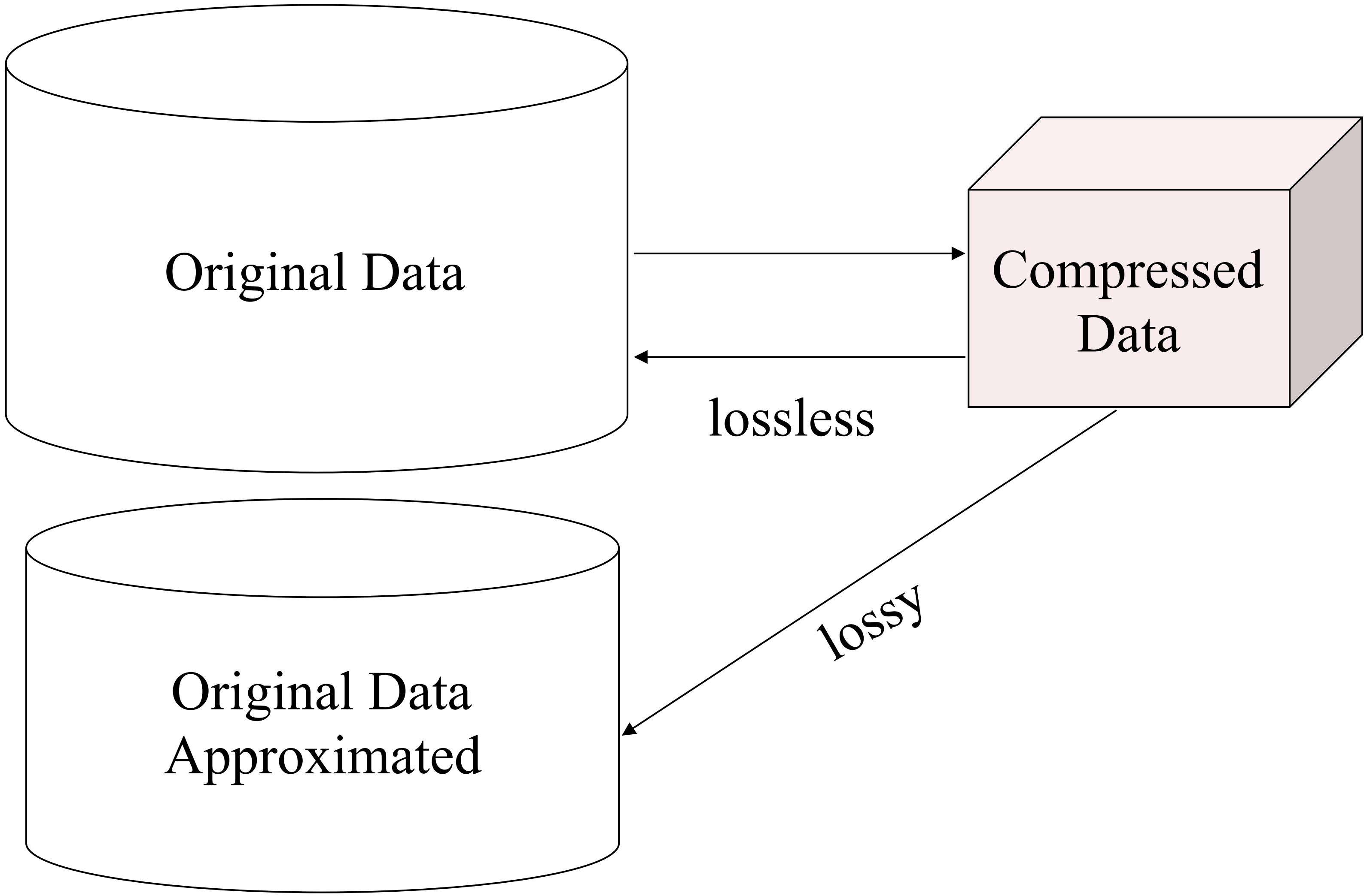
Data Cube Aggregation

- The lowest level of a data cube (base cuboid)
  - The aggregated data for an individual entity of interest
  - E.g., a customer in a phone calling data warehouse
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

Data Reduction 3: Data Compression

- String compression
  - There are extensive theories and well-tuned algorithm
  - 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
- Dimensionality and numerosity reduction may also be considered as forms of data compressior

Data Compression



Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: Concept hierarchy climbing

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

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

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,600 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

Normalization

- **Normalization by decimal scaling**

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

- Suppose that the recorded values of A range from −986 to 917. The maximum absolute value of A is 986. To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., j = 3) so that −986 normalizes to −0.986 and 917 normalizes to 0.917

Discretization

- Three types of attributes
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers
- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

Data Discretization Methods

- Typical methods: All the methods can be applied recursively
  - Binning
    - Top-down split, unsupervised
  - Histogram analysis
    - Top-down split, unsupervised
  - Other Methods
    - Clustering analysis (unsupervised, top-down split or bottom-up merge)
    - Decision-tree analysis (supervised, top-down split)
    - Correlation (e.g., χ^2) analysis (unsupervised, bottom-up merge)

Simple Discretization: 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 equal-frequency (**equi-depth**) bins:

  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

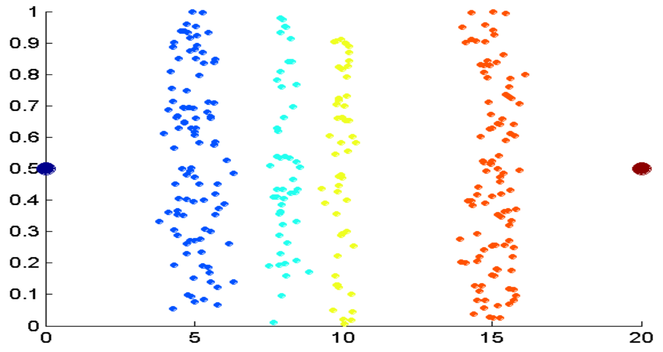
* Smoothing by **bin means**:

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

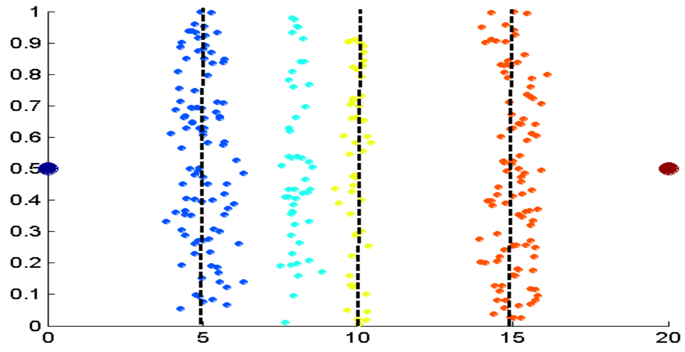
* Smoothing by **bin boundaries**:

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

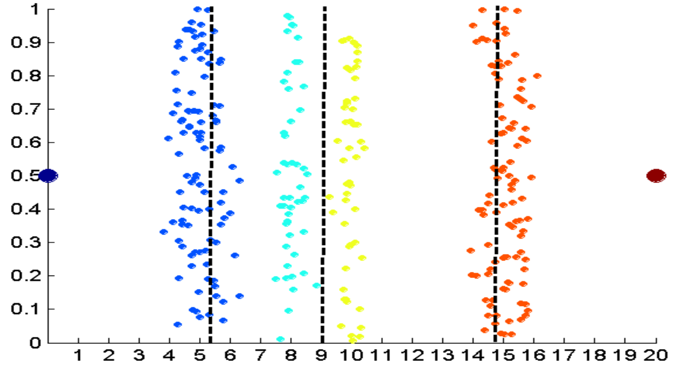
Discretization Without Using Class Labels (Binning vs. Clustering)



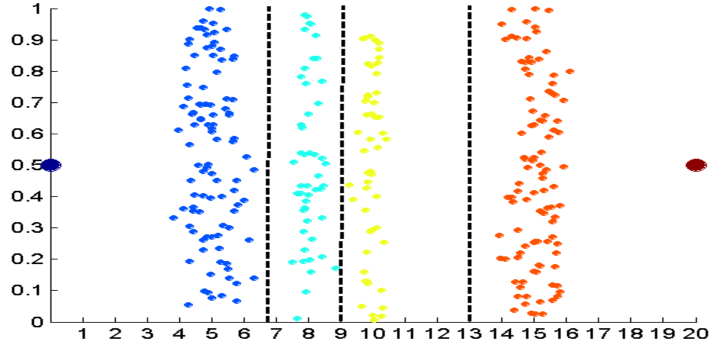
Data



Equal interval width (binning)



Equal frequency (binning)



K-means clustering leads to better results

Concept Hierarchy Generation

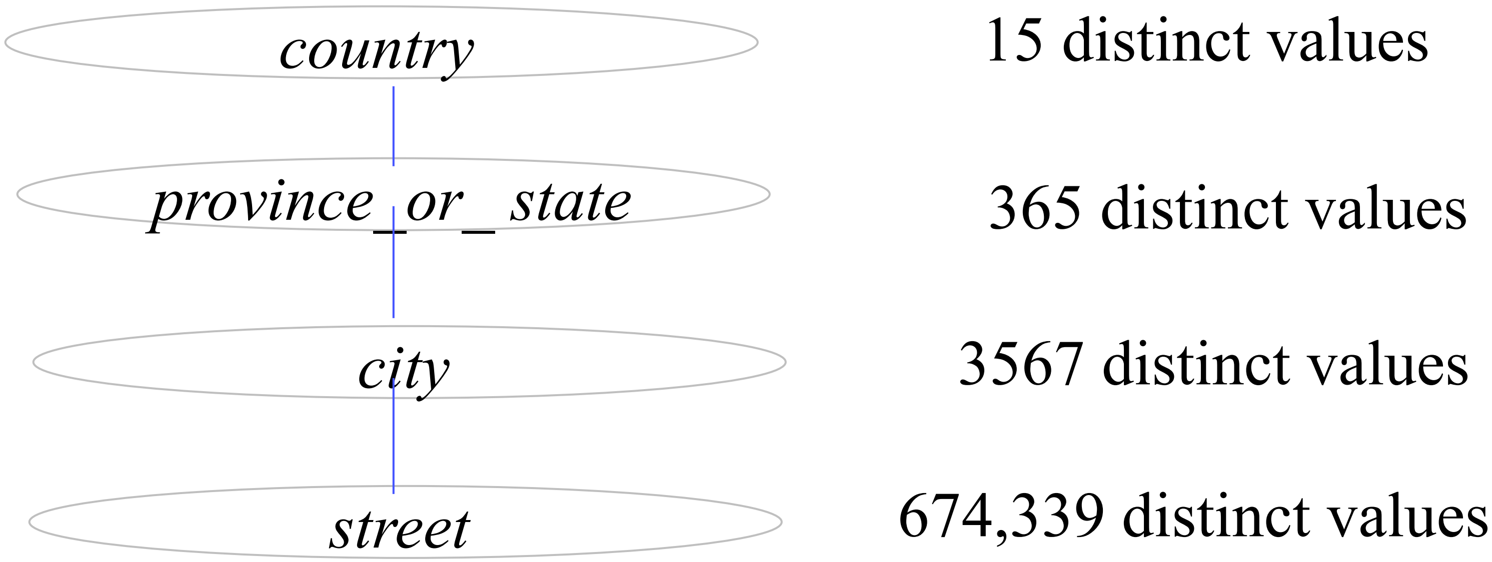
- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity
- Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth, adult*, or *senior*)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed for both numeric and nominal data. For numeric data, use discretization methods shown.

Concept Hierarchy Generation for Nominal Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - *street* < *city* < *state* < *country*
- Specification of a hierarchy for a set of values by explicit data grouping
  - {Urbana, Champaign, Chicago} < Illinois
- Specification of only a partial set of attributes
  - E.g., only *street* < *city*, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {*street, city, state, country*}

Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year



Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Summary

- **Data quality**: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
  - Entity identification problem
  - Remove redundancies
  - Detect inconsistencies
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation