# Subject : Data Warehouse and Data Mining

MIT-WPU
॥ विश्वशान्तिर्ध्रुवं ध्रुवा ॥

Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

## (Computer Engineering and Technology- AI-DS)
## (SY B.Tech)

**Disclaimer:**

1. Information included in this slides came from multiple sources. We have tried our best to cite the sources. Please refer to the References to learn about the sources, when applicable.
2. The slides should be used only for academic purposes (e.g., in teaching a class), and should not be used for commercial purposes.

# UNIT II - DATA PREPROCESSING

- Data Preprocessing: An Overview, Methods: Data Cleaning, Data Integration, Data Reduction, Data Transformation, Data Discretization.
- Data Cleaning: Handling Missing Values, Noisy Data, Data Cleaning as a ProcessData Integration: Entity Identification Problem, Redundancy and Correlation Analysis, Tuple Duplication, Data Value Conflict Detection and Resolution, Data Reduction: Attribute Subset
- Selection, Histograms, Sampling, Data discretization – binning, histogram analysis

# What is Data Mining?

- Data mining is also called *knowledge discovery and data mining* (KDD)
- Data mining is
  - Extraction of useful patterns from data sources, e.g.- Databases, texts, web, image
- Patterns must be:
  - Valid, novel, potentially useful, understandable

# What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything “data mining”?
  - Simple search and query processing
  - (Deductive) expert systems

# Why is Data Mining important?

- Rapid computerization of businesses produce huge amount of data
- How to make best use of data?
- It is used to discover patterns and relationships in the data in order to help make better business decisions
- Data mining technology can generate new business opportunities by:
  - *Automated prediction of trends and behaviors*
  - *Automated discovery of previously unknown patterns*

# Example of Discovered Patterns

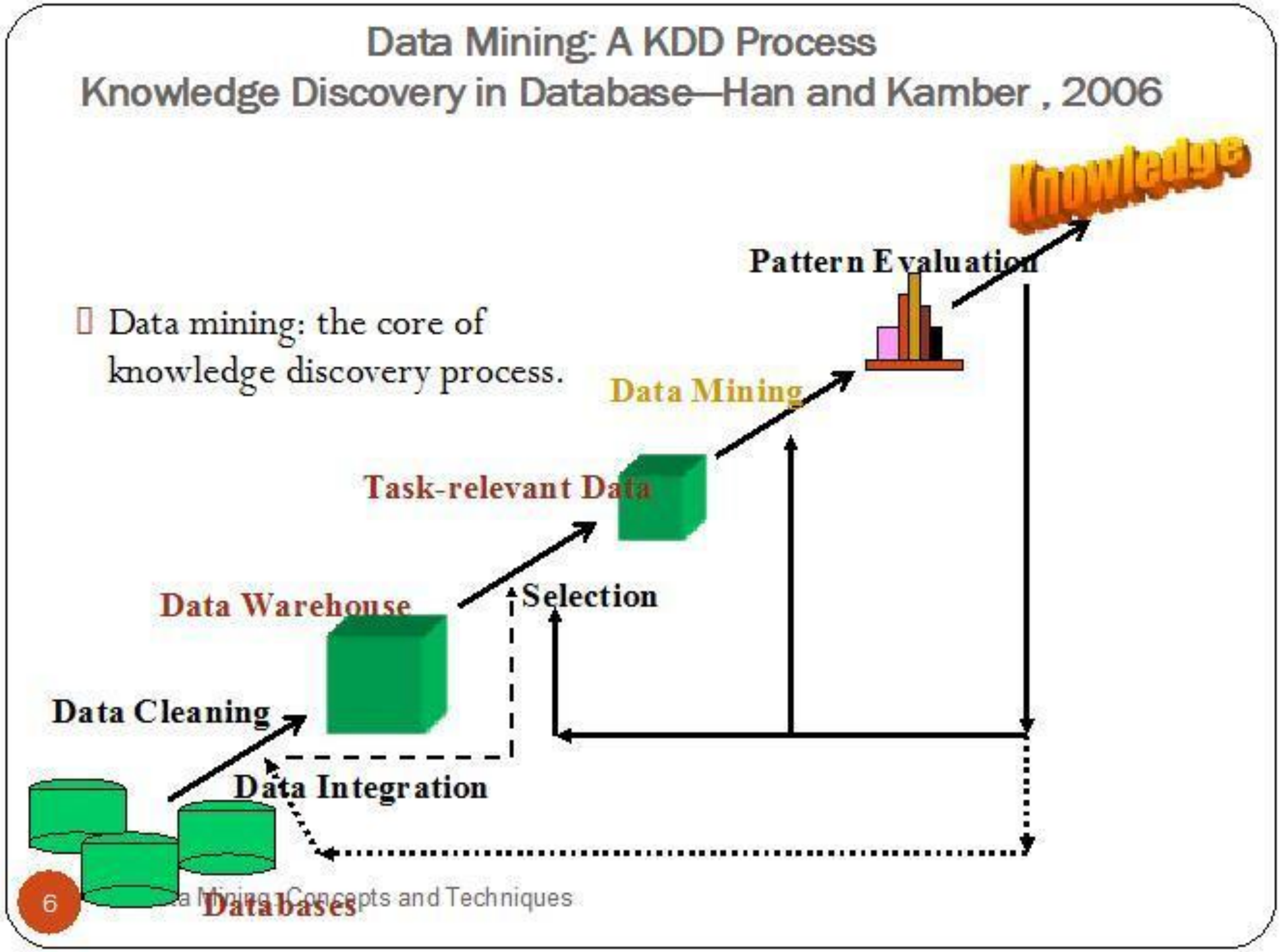
- Association rules:

  “80% of customers who buy *cheese* and *milk* also buy *bread*, and 5% of customers buy all of them together”

  Cheese, Milk→ Bread  [sup =5%, confid=80%]

# Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discover process



Data Mining: A KDD Process

Knowledge Discovery in Database—Han and Kamber , 2006

□ Data mining: the core of knowledge discovery process.

# Sequence of the steps

- **Data Cleaning:** To remove noise and inconsistent data.
- **Data Integration:** where multiple data sources may be combined.
- **Data Selection:** Where data relevant to the analysis task are retrieved from the database.
- **Data Transformation:** where data transformed and consolidated into forms appropriate  for mining by performing summary or aggregation operations.
- **Data Mining:** An essential process where intelligent methods are applied to extract data patterns.
- **Pattern evaluation:** To identify the truly interesting pattern representing knowledge based on interesting measures.
- **Knowledge Presentation:** where visualization and knowledge representation tech. are used to present mined knowledge to users.

# Data Warehouse and Data Mining

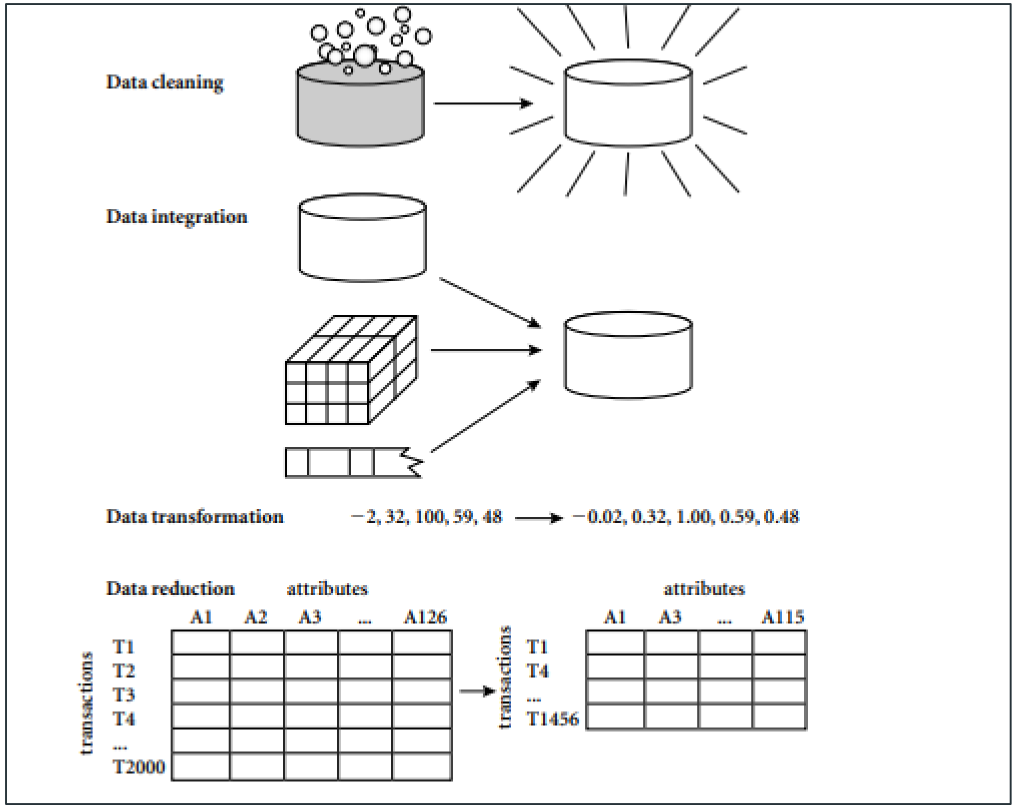
- Introduction to KDD
- **Data Preprocessing: An Overview**
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

# Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - **Quality** decisions must be based on quality data
  - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
- Data preparation, cleaning, and transformation comprises the majority of the work in a data mining application (90%).

# 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**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation



# Data Cleaning

- Importance
  - “Data cleaning is the first step”
- Data cleaning tasks
  - Fill in missing values
  - Smoothing noisy data
  - Identify or removing outliers
  - Resolve inconsistencies

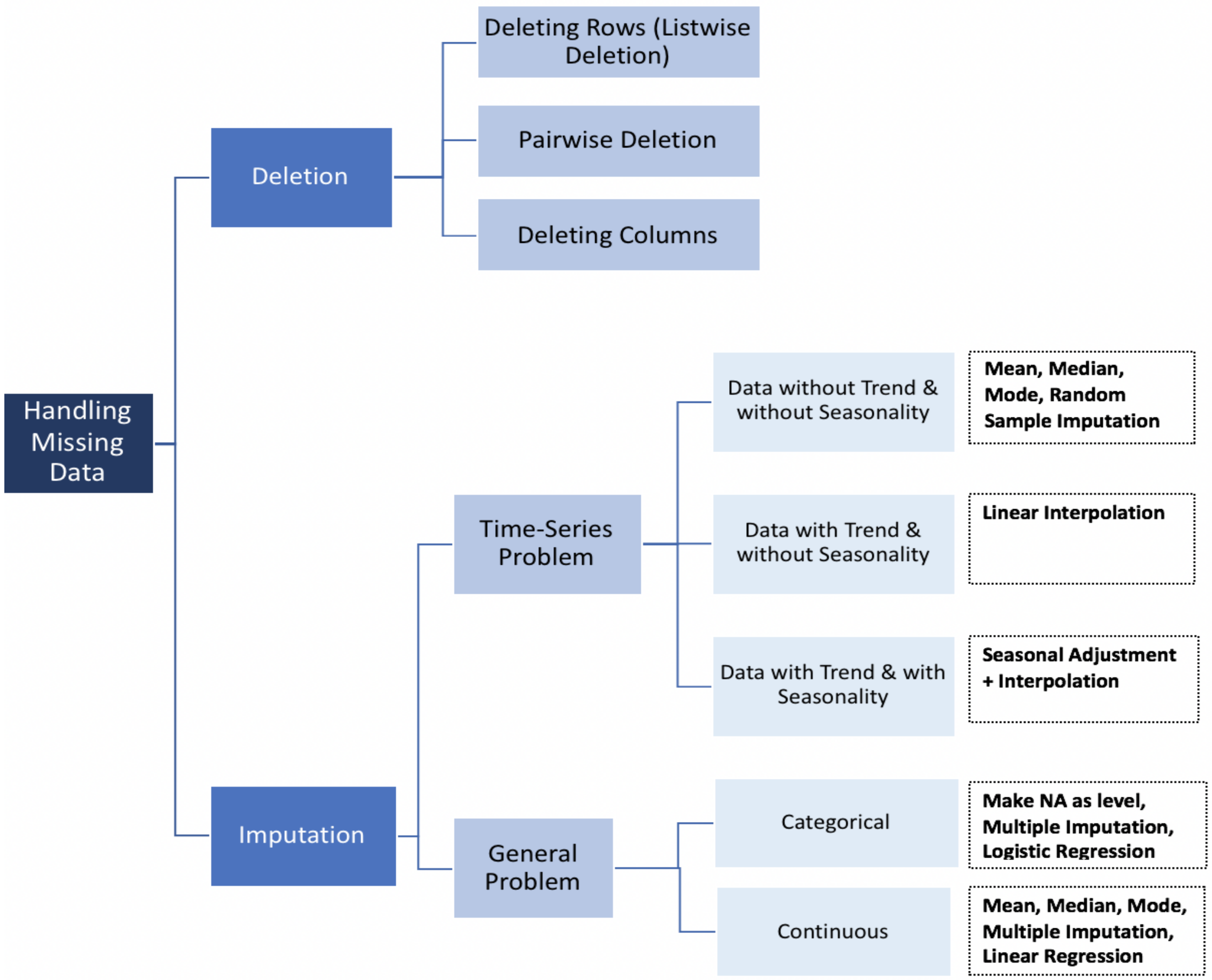
# Incomplete (**Missing**) Data

- Data is not always available
  - Many tuples may not have recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - Equipment malfunction
  - May not be available at the time of entry
  - Data not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry
  - Data inconsistent with other recorded data may have been deleted
  - Recording of data or its modifications may have been overlooked
- Missing data may need to be inferred

# How to Handle Missing Data?

- Ignore the tuple: not effective unless tuple contains several attributes with missing values
- Fill in the missing value manually: tedious + infeasible?
- Fill it automatically with
  - A global constant : eg., “Unknown”
  - The attribute mean or median
  - The attribute mean for all samples belonging to the same class as the tuple: smarter
  - The most probable value: inference-based/Impute

# How to Handle Missing Data?



Handling Missing Data

Deletion

Deleting Rows (Listwise Deletion)

Pairwise Deletion

Deleting Columns

Imputation

Time-Series Problem

Data without Trend & without Seasonality

**Mean, Median, Mode, Random Sample Imputation**

Data with Trend & without Seasonality

**Linear Interpolation**

Data with Trend & with Seasonality

**Seasonal Adjustment + Interpolation**

General Problem

Categorical

**Make NA as level, Multiple Imputation, Logistic Regression**

Continuous

**Mean, Median, Mode, Multiple Imputation, Linear Regression**

# 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

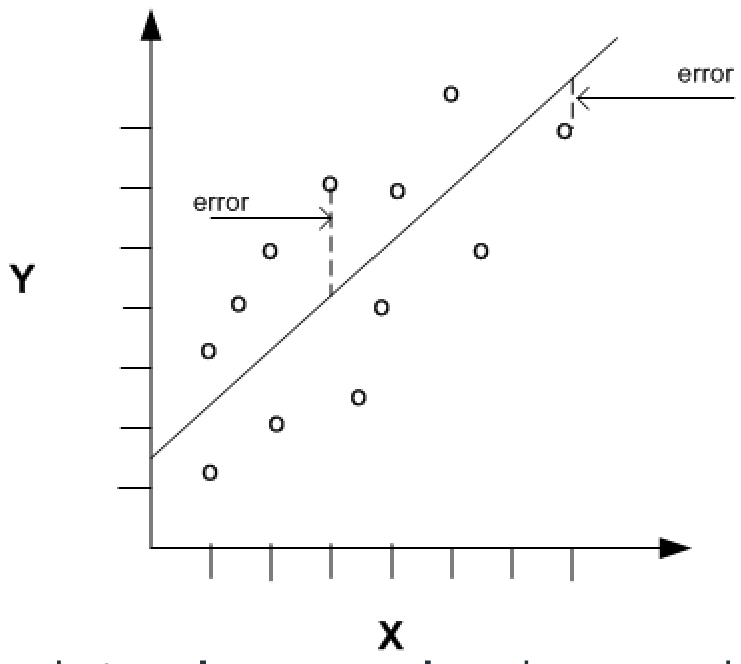
# How to Handle Noisy Data?

- 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 Methods for Data Smoothing

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

# Regression for Data Smoothing



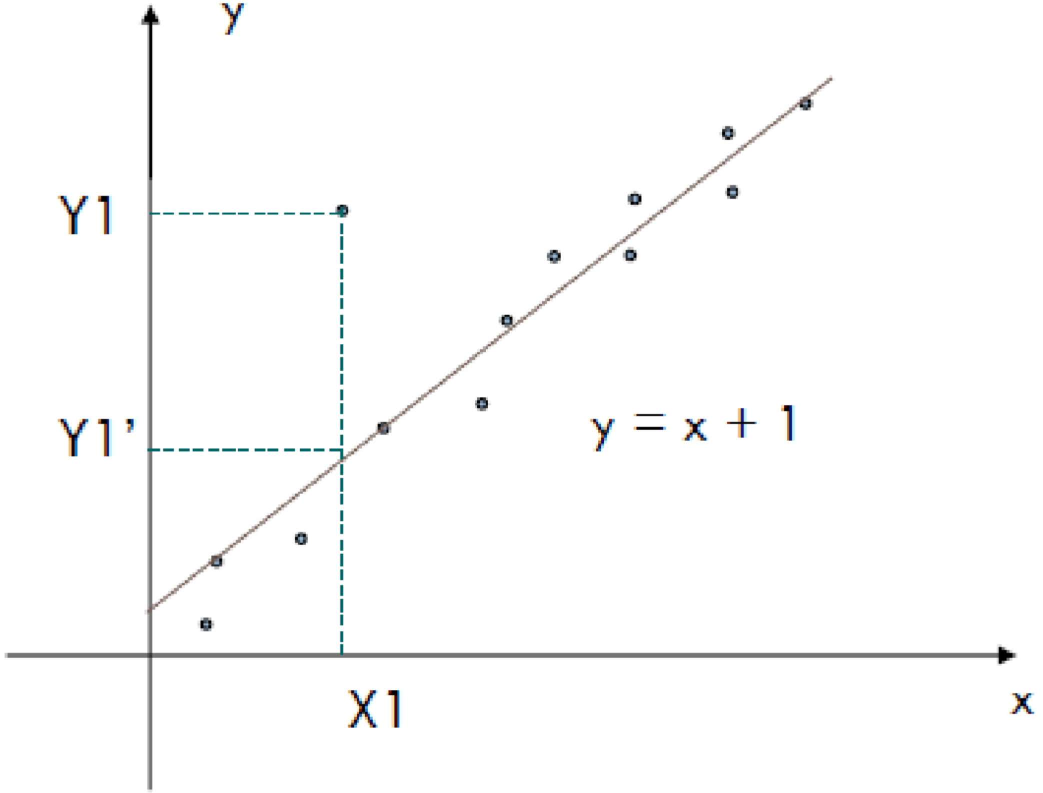
For **linear** and **stepwise regression**, the regression formula is:

*Dependent variable = intercept + Sum$_i$ (coefficient$_i$ * independent variable$_i$ ) + error*

- The regression functions are used to determine the relationship between the dependent variable (target field) and one or more independent variables. The dependent variable is the one whose values you want to predict, whereas the independent variables are the variables that you base your prediction on.
- A RegressionModel defines three types of regression models: linear, polynomial, and logistic regression.
- The **modelType** attribute indicates the type of regression used.
- Linear and stepwise-polynomial regression are designed for numeric dependent variables having a continuous spectrum of values. These models should contain exactly one regression table.
- Logistic regression is designed for categorical dependent variables.

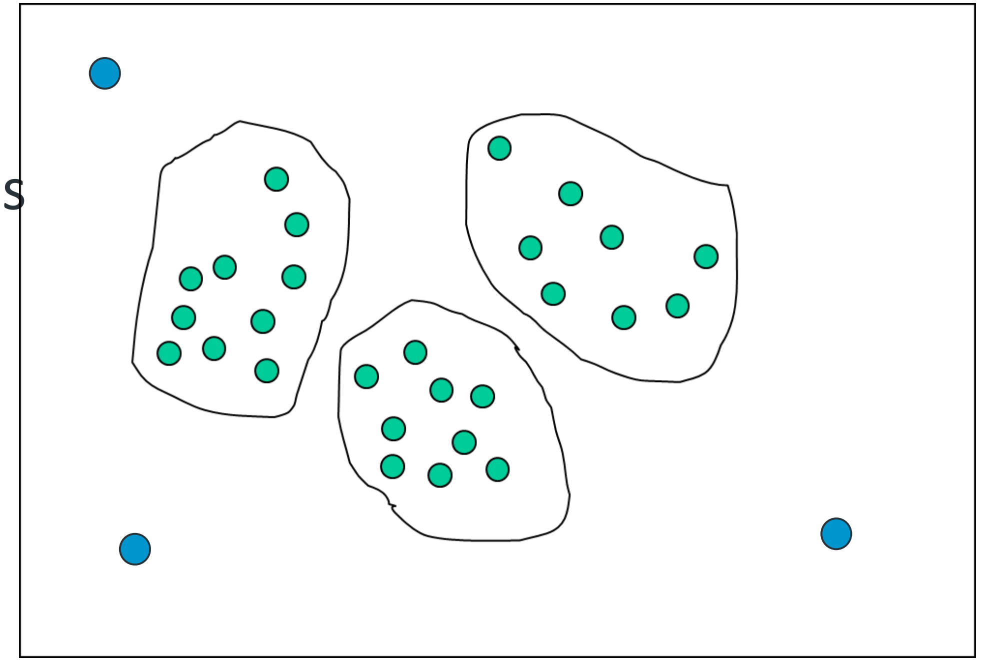
# Regression for Data Smoothing

- Replace noisy or missing values by predicted values
- Requires model of attribute dependencies(may be wrong!)
- Can be used for data smoothing or for handling missing data



# Regression for Data Smoothing: Outlier Removal

- Data points inconsistent with the majority of data
- Different outliers
  - Valid: CEO’s salary,
  - Noisy: One’s age = 200, widely deviated points
- Removal methods
  - Clustering
  - Curve-fitting
  - Hypothesis-testing with a given model



# Exercise

- Suppose a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215 Partition them into three bins solve it by each of the following methods:
  (a) equi-depth partitioning
  (b) Smoothing by bin boundaries

- Solution:
  (a) equi-depth partitioning -Bin-1: 5, 10, 11, 13, Bin-2: 15, 35, 50, 55, Bin-03: 72, 92, 204, 215
  (b) Smoothing by bin boundaries-Smoothing by bin boundaries: Bin-1: 5,13,13,13 , Bin-2: 15,15,55,55, Bin-3:72,72,215,215

# Data Warehouse and Data Mining

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

# What is Data Integration

- **Data integration**:
  - Combines data from multiple sources into a coherent store
  - Careful integration can help reduce & avoid redundancies and inconsistencies
  - This helps to improve accuracy & speed of subsequent data mining
  - Heterogeneity & structure of data pose great challenges
  - Issues that need to be addressed:
    1. How to match schema & objects from different sources? (Entity identification problem)
    2. Are any attributes correlated?
    3. Tuple duplication
    4. Detection & resolution of data value conflicts

# Data Integration Issues

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton, Use Metadata to handle this problem
- **Redundancy**:
  - Inconsistencies in attribute or dimensions naming can cause redundancy
- 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

**Redundancy & Correlation Analysis:**

- 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

# Detection of Data Redundancy-Correlation

- Redundancies can be detected using following methods:
  - χ2Test (Used for nominal Data or categorical or qualitative data)
  - Correlation coefficient and covariance (Used for numeric Data or quantitative data)
- The term "correlation" refers to a mutual relationship or association between quantities.
- In almost any business, it is useful to express one quantity in terms of its relationship with others.
  - For example, sales might increase when the marketing department spends more on TV advertisements,
  - Customer's average purchase amount on an e-commerce website might depend on a number of factors related to that customer.
- Often, correlation is the first step to understanding these relationships and subsequently building better business and statistical models.

# Why is correlation a useful metric?

- Correlation can help in predicting one quantity from another
- Correlation can (but often does not, as we will see in some examples below) indicate the presence of a causal relationship
- Correlation is used as a basic quantity and foundation for many other modeling techniques
- More formally, correlation is a statistical measure that describes the association between random variables.
- There are several methods for calculating the correlation coefficient, each measuring different types of strength of association

# 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)

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

- It shows that like_science_fiction and play_chess are correlated in the group

# Correlation Analysis (Numeric data )

- Evaluate correlation between 2 attributes, A & B, by computing **correlation coefficient(Pearson's product moment coefficient**)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_ib_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

Where *n* is the number of tuples, $a_i$ and $b_i$ are the respective values of *A* and *B* in tuple *i*,

$\bar{A}$ and $\bar{B}$ are the respective mean values of *A* and *B*,

$\sigma_A$ and $\sigma_B$ are the respective standard deviations of *A* and *B*

$\Sigma(a_ib_i)$ is the sum of the *AB* cross-product (i.e., for each tuple, the value for *A* is multiplied by the value for *B* in that tuple)

*The Correlations coefficient can range between +1 and -1*

# Correlation Analysis (Numeric data )

- Note that $-1 \leq r_{A,B} \leq +1$
- If resulting value is greater than 0, then *A* and *B* are *positively correlated*, meaning that the values of *A* increase as the values of *B* increase
- Higher the value, the stronger the correlation
- Higher value may indicate that *A* (or *B*) may be removed as a redundancy
- If the resulting value is equal to 0, then *A* and *B* are *independent* and there is no
- correlation between them
- If the resulting value is less than 0, then *A* and *B* are *negatively correlated*, where the values of one attribute increase as the values of the other attribute decrease

# Covariance Analysis(Numeric Data)

- Correlation and covariance are two similar measures for assessing how much two attributes change together
- Consider two numeric attributes *A* and *B*, and a set of *n* observations

$$\{(a_1, b_1), \ldots, (a_n, b_n)\}$$

- Mean values of *A* and *B*, respectively, are also known as the **expected values** on *A* and *B*, that is

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$

$$E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}.$$

- **covariance** between *A* and *B* is defined as $Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n}.$

# Covariance Analysis(Numeric Data)

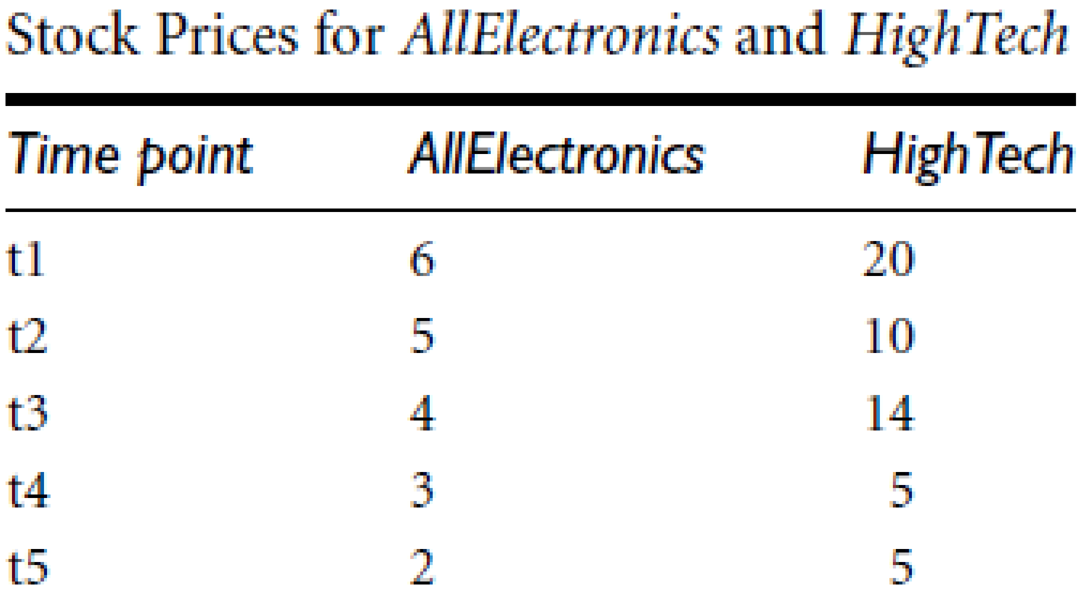
$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

- Also, for simplified calculations

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

- For two attributes *A* and *B* that tend to change together,
  - if *A* is larger than $\bar{A}$ , then *B* is likely to be larger than $\bar{B}$ . Hence, the covariance between *A* and *B* is ***positive***
  - if one of the attributes tends to be above its expected value when the other attribute is below its expected value, then the covariance of *A* and *B* is ***negative***
  - covariance value is 0 means attributes are independent

# Covariance Analysis(Numeric Data)

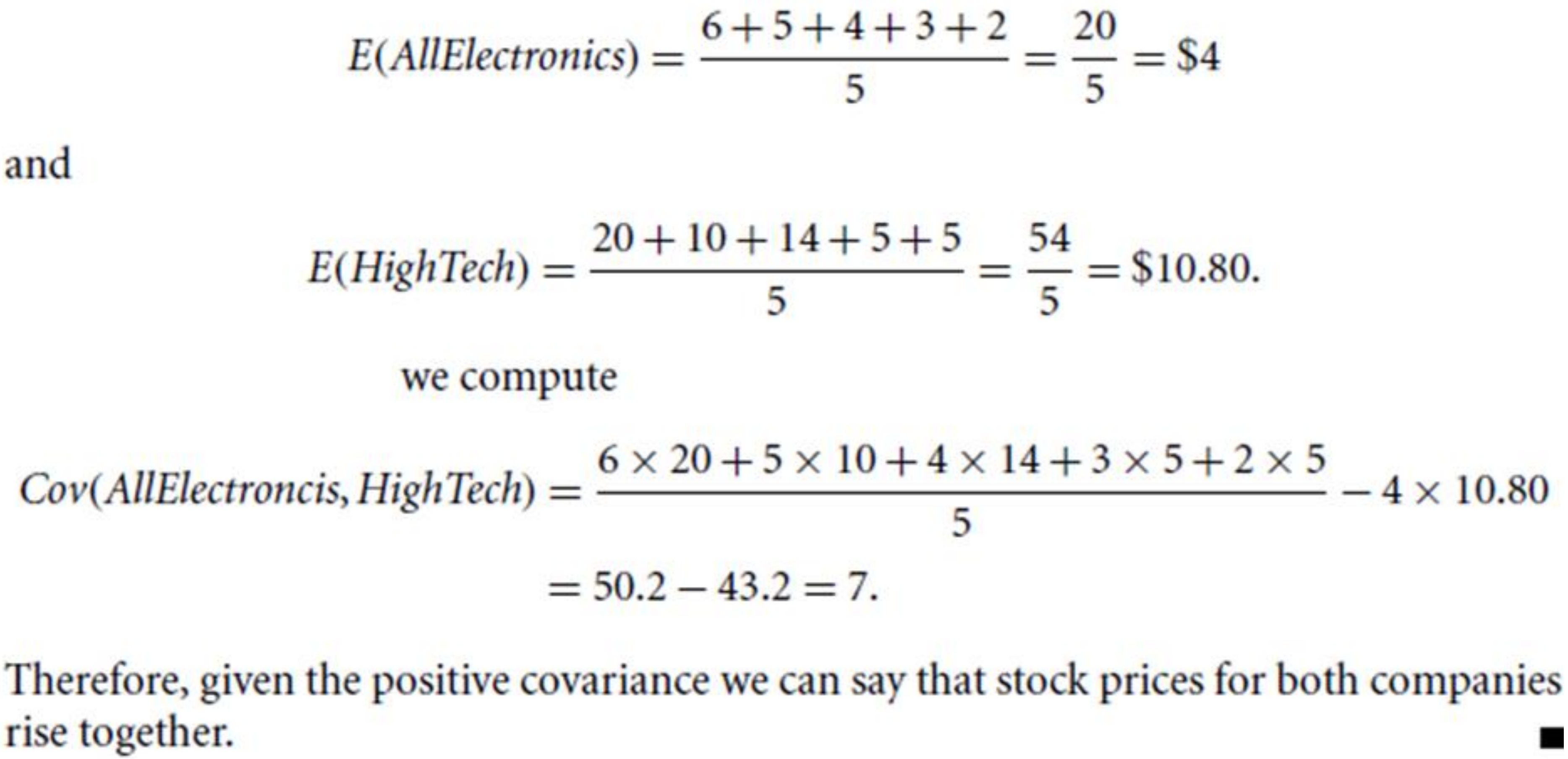


Stock Prices for *AllElectronics* and *HighTech*

| *Time point* | *AllElectronics* | *HighTech* |
|---|---|---|
| t1 | 6 | 20 |
| t2 | 5 | 10 |
| t3 | 4 | 14 |
| t4 | 3 | 5 |
| t5 | 2 | 5 |

- Table shows stock prices of two companies at five time points. If the stocks are affected by same industry trends, determine whether their prices rise or fall together?

# Covariance Analysis(Numeric Data)



$$E(AllElectronics) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \$4$$

and

$$E(HighTech) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.80.$$

we compute

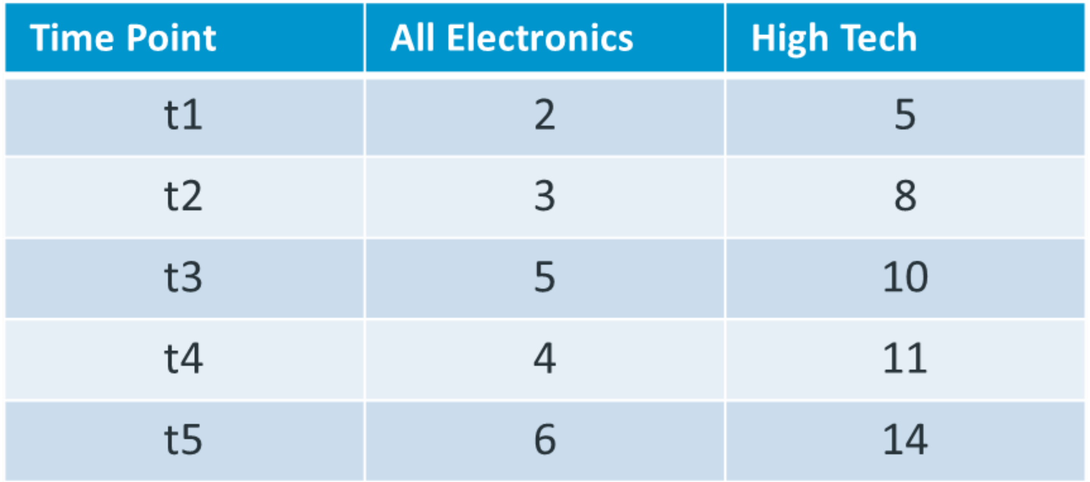
$$Cov(AllElectroncis, HighTech) = \frac{6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10.80$$

$$= 50.2 - 43.2 = 7.$$

Therefore, given the positive covariance we can say that stock prices for both companies rise together. ■

# Correlation: example

An example of stock prices observed at five time points for *AllElectronics* and *HighTech*, a high-tech company. If the stocks are affected by the same industry trends, will their prices rise or fall together?



| Time Point | All Electronics | High Tech |
|---|---|---|
| t1 | 2 | 5 |
| t2 | 3 | 8 |
| t3 | 5 | 10 |
| t4 | 4 | 11 |
| t5 | 6 | 14 |

- 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?

# Solution

  - **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 Value Conflict Detection and Resolution

- For the same real-world entity, attribute values from different sources may differ
  - Eg. Prices of rooms in different cities may involve different currencies
- Attributes may also differ on the abstraction level, where an attribute in one system is recorded at, say, a lower abstraction level than the “same” attribute in another.
  - Eg. total sales in one database may refer to one branch of All_Electronics, while an attribute of the same name in another database may refer to the total sales for All_Electronics stores in a given region.
- To resolve, data values have to be converted into consistent form

# Data Transformation

*In data transformation, the data are transformed or consolidated into forms appropriate for mining. Strategies for data transformation include the following*

- 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 ones
  - E.g. add attribute *area* based on the attributes *height* and *width*

# 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

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

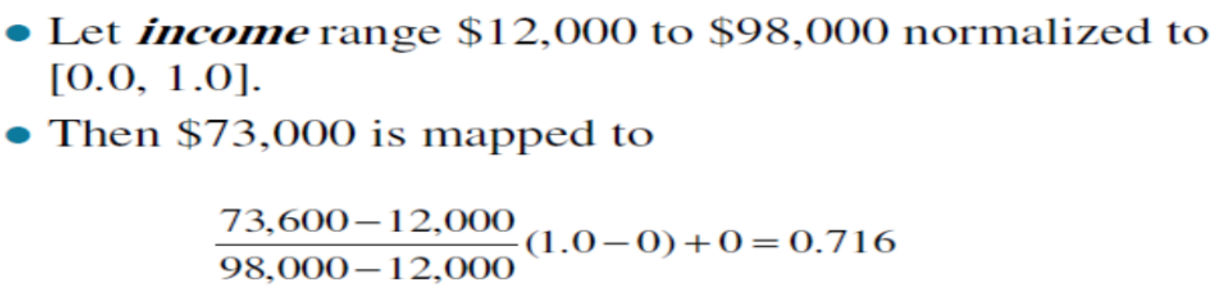
- Normalization by decimal scaling

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

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

# Data Transformation: Min Max-Normalization

- Min-max normalization: to [new_minA, new_maxA]
  - Performs a linear transformation on the original data.
  - Suppose that $min_a$ and $max_a$ are the minimum and maximum values of an attribute, $a$.
  - Min-max normalization maps a value, $v_i$ , of $a$ to the range $[new\_min_a, new\_max_a]$ by computing



- Let ***income*** range \$12,000 to \$98,000 normalized to [0.0, 1.0].
- Then \$73,000 is mapped to

$$\frac{73{,}600-12{,}000}{98{,}000-12{,}000}(1.0-0)+0=0.716$$

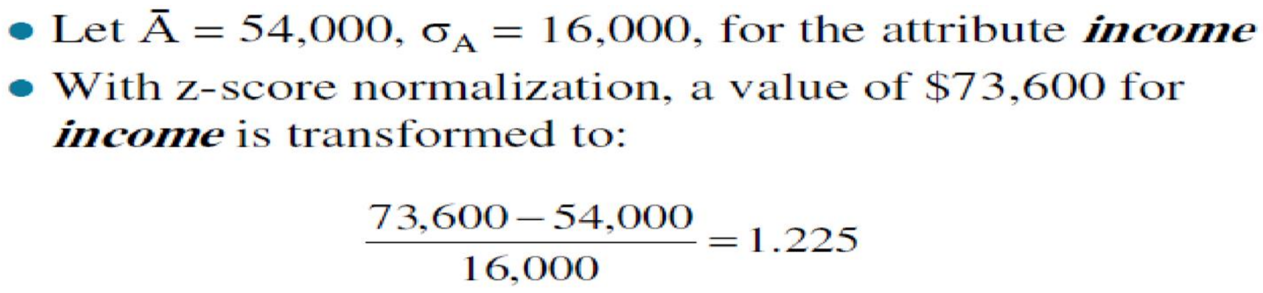
  - Eg. Suppose that the minimum and maximum values for the attribute income are \$12,000 and \$98,000, respectively. We would like to map income to the range [0.0, 1.0]. By min-max normalization, a value of \$73,600 for income is transformed to **0.716**

# Data Transformation: Z-score Normalization

- Z-score normalization (uses  mean &  σ: standard deviation):
- Values for an attribute, A, are normalized based on the mean (i.e., average) and standard deviation of A

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$$

- Eg. Suppose that the mean and 'on of the values for the attribute income are $54,000 and $16,000, respectively.



- Let $\bar{A} = 54{,}000$, $\sigma_A = 16{,}000$, for the attribute ***income***
- With z-score normalization, a value of $73,600 for ***income*** is transformed to:

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

# Data Transformation: Decimal scaling Normalization

- Suppose that the recorded values of A ranges from -986 to 917
- The maximum absolute value of A is 986
- To normalize by decimal scaling, we therefore divide each value by 1000
- So that -986 normalize to -0.986 and
- 917 normalize to -0.917

# Exercise

- Define Normalization.
- What is the value range of min-max. Use min-max normalization to normalize the following group of data: 8,10,15,20.

- Solution:

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

| Marks | Marks after Min-Max normalization |
|---|---|
| 8 | 0 |
| 10 | 0.16 |
| 15 | 0.58 |
| 20 | 1 |

# Data Warehouse and Data Mining

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

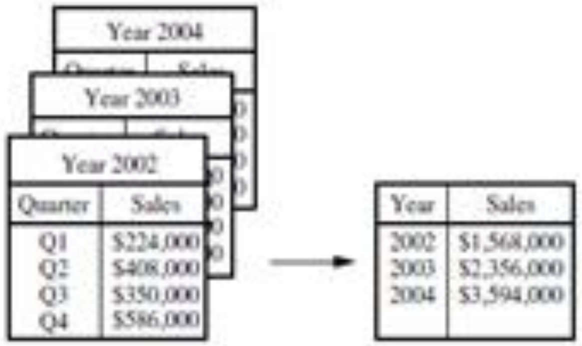
# Data Reduction Strategies

- Data is too big to work with
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- 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 — remove unimportant attributes
  - Aggregation and clustering
  - Sampling

# Data Reduction Strategies

- 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

# Data Cube Aggregation



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 |

**Figure 2.13** Sales data for a given branch of *AllElectronics* for the years 2002 to 2004. On the left, the sales are shown per quarter. On the right, the data are aggregated to provide the annual sales.

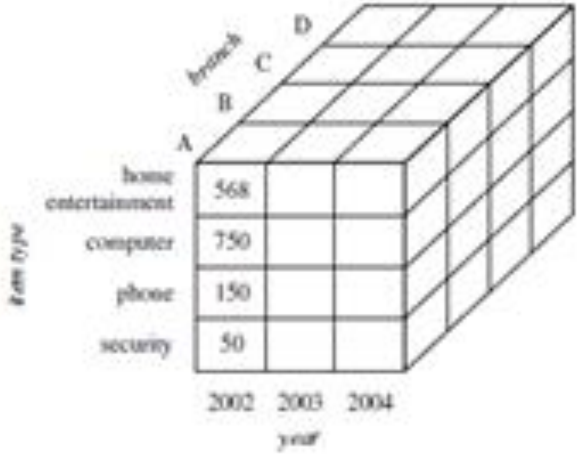


**Figure 2.14** A data cube for sales at *AllElectronics*.

- 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

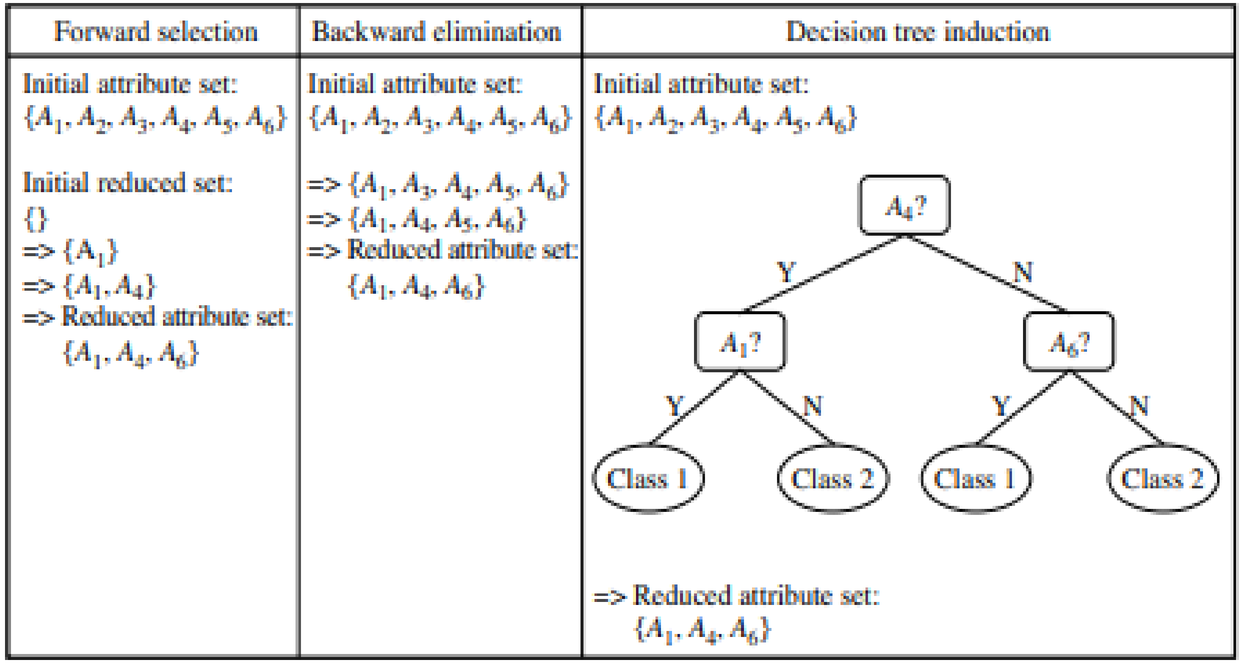
# Data Reduction Strategies

- Data reduction strategies:
  - Dimensionality reduction:
    - Wavelet transforms, Principal Components Analysis (PCA) : Methods to project original data onto smaller space
    - Attribute subset selection :Redundant, irrelevant attribute are detected and removed
  - Numerosity reduction: replaces original data volume by alternative smaller forms
    - Regression :Models are used to estimate data, so only data parameters are stored
    - Histograms, clustering, sampling :To stored reduced representations of data
  - Data compression : Lossless transformations are applied for compression and original data can be retained without any loss

# Dimensionality Reduction :Attribute Subset Selection

- Attribute subset selection reduces the data set size by removing irrelevant or redundant attributes (or dimensions).
- The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attribute
- 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

# Attribute Subset Selection Methods



| Forward selection | Backward elimination | Decision tree induction |
|---|---|---|
| Initial attribute set:<br>$\{A_1, A_2, A_3, A_4, A_5, A_6\}$<br><br>Initial reduced set:<br>{}<br>=> $\{A_1\}$<br>=> $\{A_1, A_4\}$<br>=> Reduced attribute set:<br>$\{A_1, A_4, A_6\}$ | Initial attribute set:<br>$\{A_1, A_2, A_3, A_4, A_5, A_6\}$<br><br>=> $\{A_1, A_3, A_4, A_5, A_6\}$<br>=> $\{A_1, A_4, A_5, A_6\}$<br>=> Reduced attribute set:<br>$\{A_1, A_4, A_6\}$ | Initial attribute set:<br>$\{A_1, A_2, A_3, A_4, A_5, A_6\}$<br><br>=> Reduced attribute set:<br>$\{A_1, A_4, A_6\}$ |

$A_4$?
Y
N
$A_1$?
$A_6$?
Y
N
Y
N
Class 1
Class 2
Class 1
Class 2

"How can we find a 'good' subset of the original attributes?

- **Heuristic Search Method** to explore reduced search space
- **Greedy Methods**: Find the best choice of attribute at given time
- **Local optimal choice** will lead to global optimal attribute selection
- **Best/Worst choice for attribute determined by statistical** significance/attribute evaluation methods

# Heuristic(Greedy) methods for attribute subset selection

1. **Stepwise Forward Selection:**
   - Starts with an empty set of attributes as the reduced set
   - Best of the relevant attributes is determined and added to the reduced set
   - In each iteration, best of remaining attributes is added to the set
2. **Stepwise Backward Elimination:**
   - Here all the attributes are considered in the initial set of attributes
   - In each iteration, worst attribute remaining in the set is removed
3. **Combination of Forward Selection and Backward Elimination:**
   - Stepwise forward selection and backward elimination are combined
   - At each step, the procedure selects the best attribute and removes the worst from among the remaining attributes

# Heuristic(Greedy) methods for attribute subset selection(cont)

## 4. Decision Tree Induction:

- This approach uses decision tree for attribute selection.
- It constructs a flow chart like structure having nodes denoting a test on an attribute.
- Each branch corresponds to the outcome of test and leaf nodes is a class prediction.
- The attribute that is not the part of tree is considered irrelevant and hence discarded

# Data Reduction 2: Numerosity Reduction

- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
- Non-parametric methods
  - Do not assume models
  - Major families: histograms, clustering, sampling

# Regression Models

- Linear regression: Data are modeled to fit a straight line: $Y = \alpha + \beta X$

- Multiple regression: allows a response variable y to be modeled as a linear function of multidimensional feature vector (predictor variables) $Y = b0 + b1\ X1 + b2\ X2.$

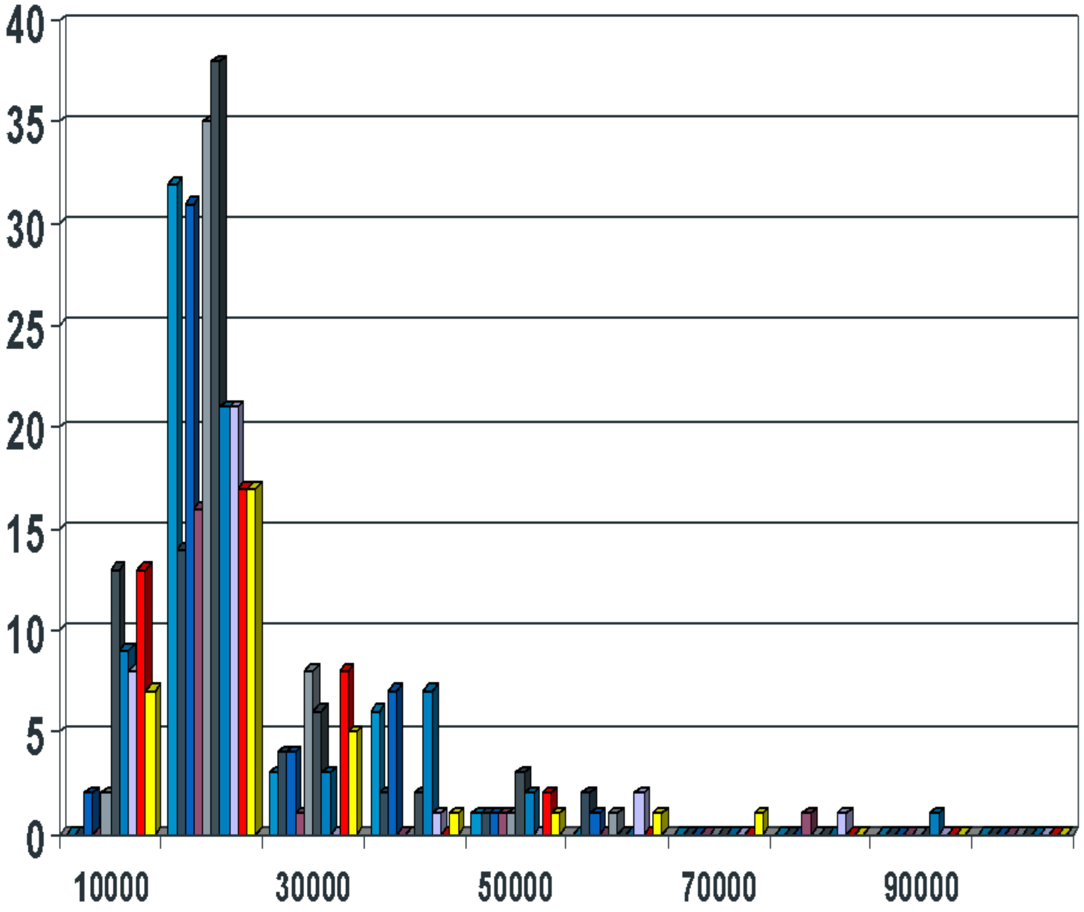
# Histograms

- Histograms (or frequency histograms) are at least a century old and are widely used.
- Plotting histograms is a graphical method for summarizing the distribution of a given attribute, X.
- Height of the bar indicates the frequency (i.e., count) of that X value
- Range of values for X is partitioned into disjoint consecutive subranges.
- Subranges, referred to as **buckets or bins**, are disjoint subsets of the data distribution for X.
- Range of a bucket is known as the width
- Typically, the buckets are of equal width.
- Eg. a price attribute with a value range of $1 to $200 can be partitioned into subranges 1 to 20, 21 to 40, 41 to 60, and so on.
- For each subrange, a bar is drawn with a height that represents the total count of items observed within the subrange

# Histogram

- Divide data into buckets and store average (sum) for each bucket
- Approximate data distributions
- Divide data into buckets and store average (sum) for each bucket
- A bucket represents an attribute-value/frequency pair

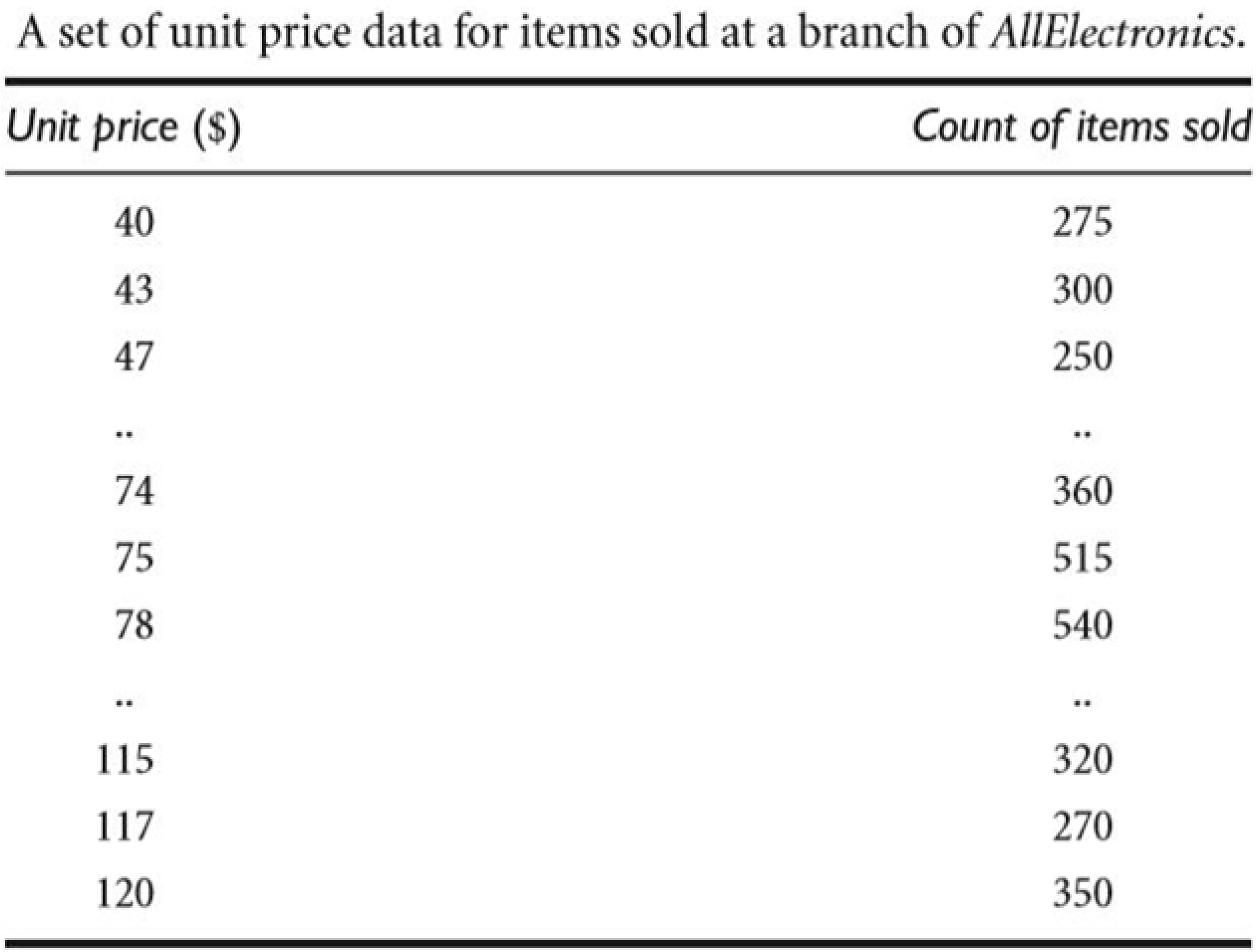
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)



# Histogram Analysis- Explanation & Example

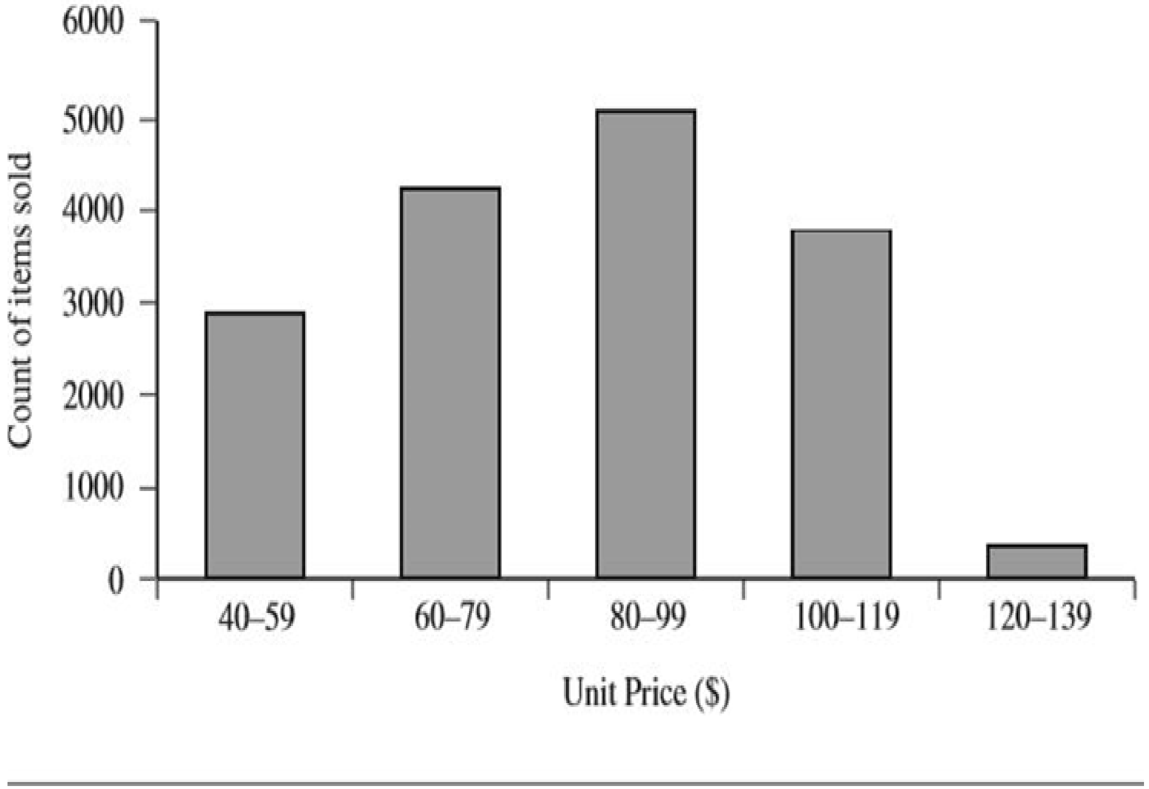
- Plotting histograms, or frequency histograms, is a graphical method for summarizing the distribution of a given attribute.
- A histogram for an attribute A partitions the data distribution of A into disjoint subsets, or buckets. Typically, the width of each bucket is uniform.
- Each bucket is represented by a rectangle whose height is equal to the count or relative frequency of the values at the bucket.
- The following data are a list of AllElectronics prices for commonly sold items (rounded to the nearest dollar).

# Histogram Analysis- Explanation & Example



A set of unit price data for items sold at a branch of *AllElectronics*.

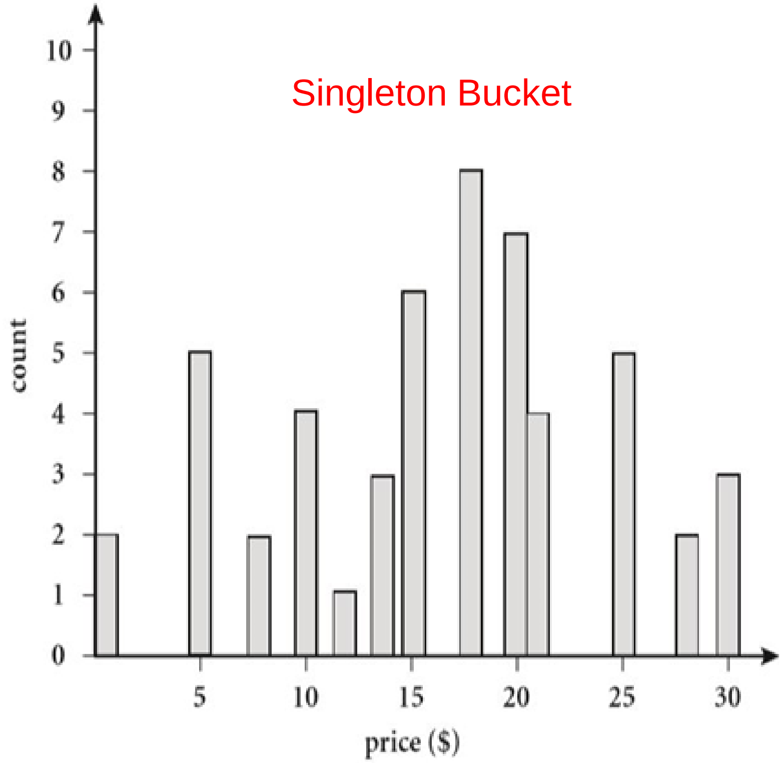
| *Unit price ($)* | *Count of items sold* |
|---|---|
| 40 | 275 |
| 43 | 300 |
| 47 | 250 |
| .. | .. |
| 74 | 360 |
| 75 | 515 |
| 78 | 540 |
| .. | .. |
| 115 | 320 |
| 117 | 270 |
| 120 | 350 |



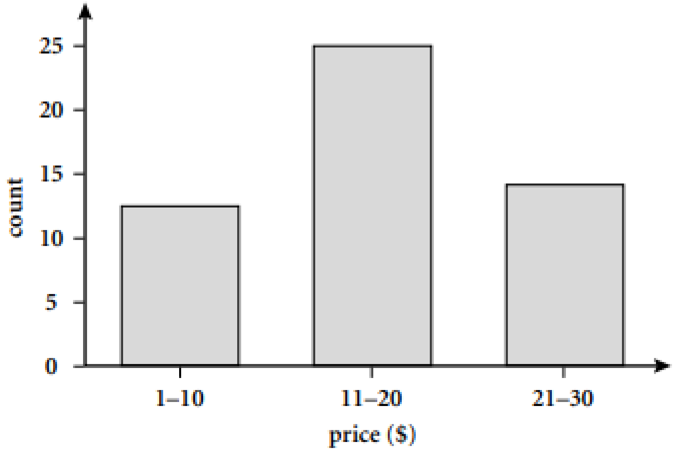
# Histogram Analysis- Explanation & Example

- The following data are a list of prices of commonly sold items at AllElectronics (rounded to the nearest dollar).
- The numbers have been sorted:
- 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

Singleton Bucket



Equi Width Bucket

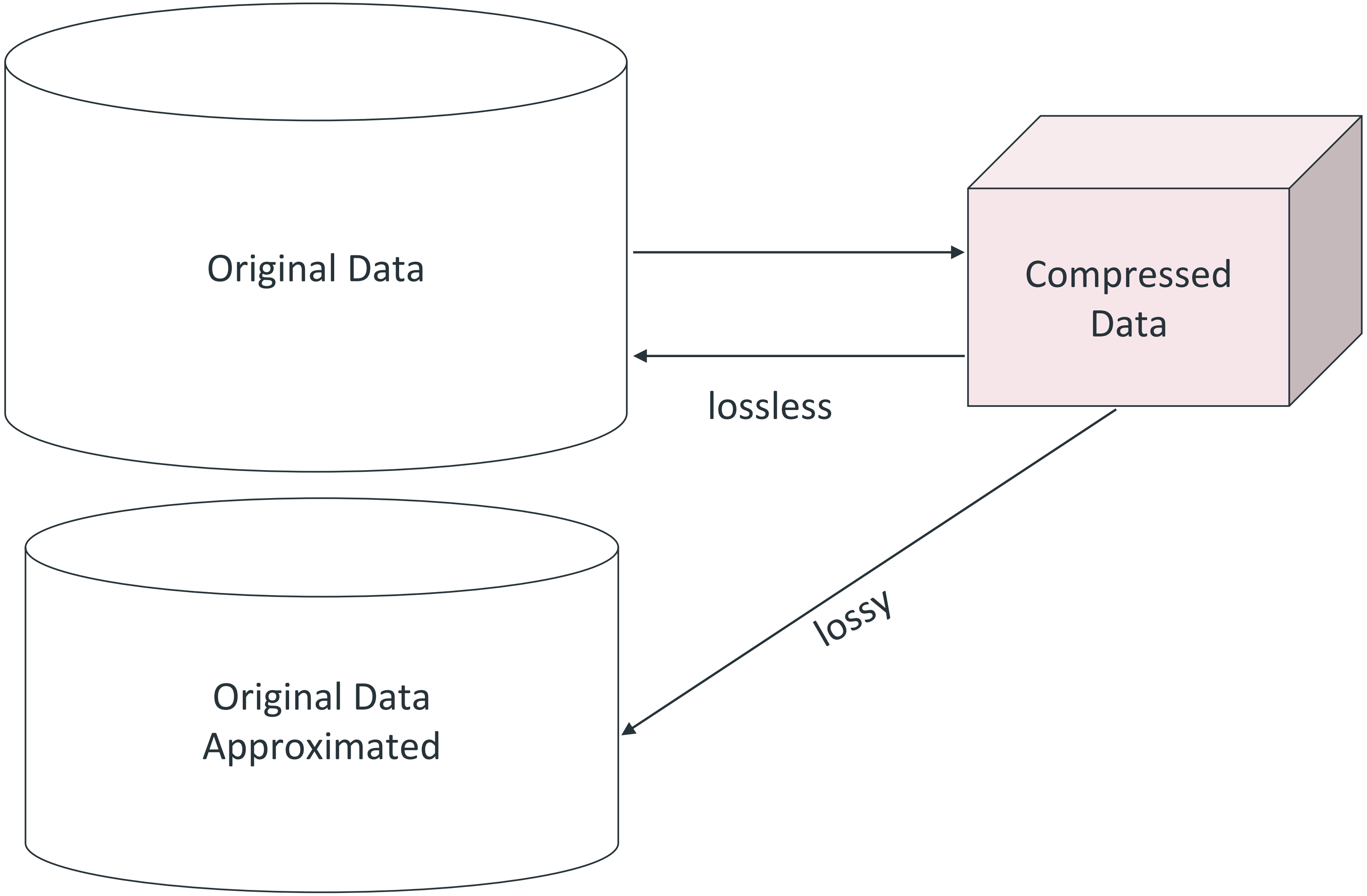


A histogram for *price* using singleton buckets—each bucket represents one price-value/ frequency pair.

# Data Compression

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible without expansion
- Audio/video, image compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole

# Data Compression



# Clustering

- Partition data set into clusters, and store cluster representation only
- Quality of clusters measured by their diameter (max distance between any two objects in the cluster) or centroid distance (avg. distance of each cluster object from its centroid)
- Can be very effective if data is clustered but not if data is “smeared”

# 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

# Sampling

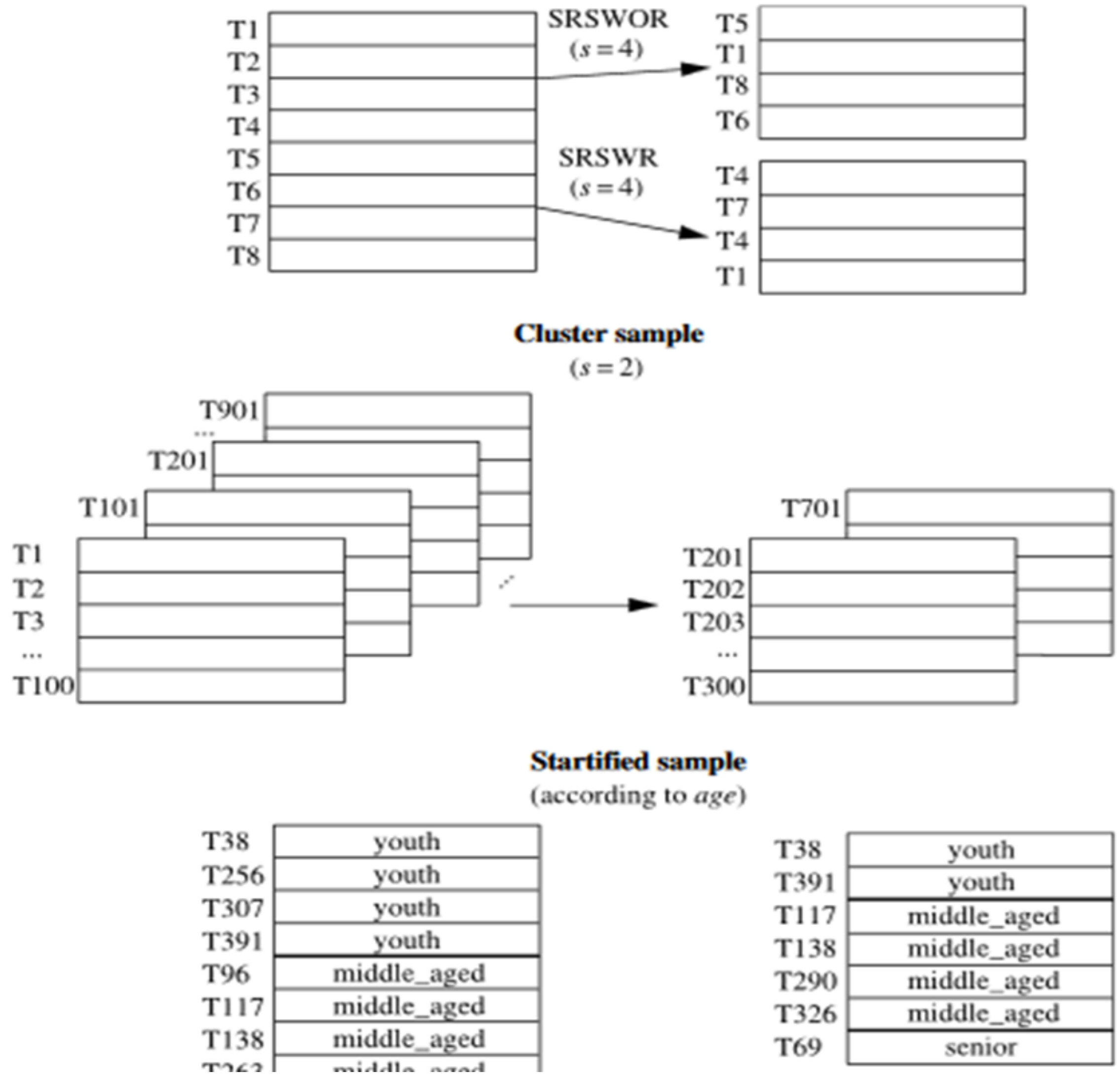
- Simple random sampling may have very poor performance in the presence of skew

- Stratified sampling:
  - Approximate the percentage of each class (or subpopulation of interest) in the overall database
  - Used in conjunction with skewed data

# 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

# Types of Sampling



# 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

# 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

# Reference

- Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, and Jian Pei, 3rd edition