# DATA & DATA PREPROCESSING

# Outline

- General data characteristics
- Why preprocess the data?
- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Summary

# Outline

- ✓ **General data characteristics**
- Why preprocess the data?
- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Summary

# What is Data?

- Information about objects with attributes
- Attribute = characteristics or descriptions
- Data = Data Set = Record = Entity
- Data can be **qualitative** or **quantitative**
- Data can be **structured** or **unstructured**

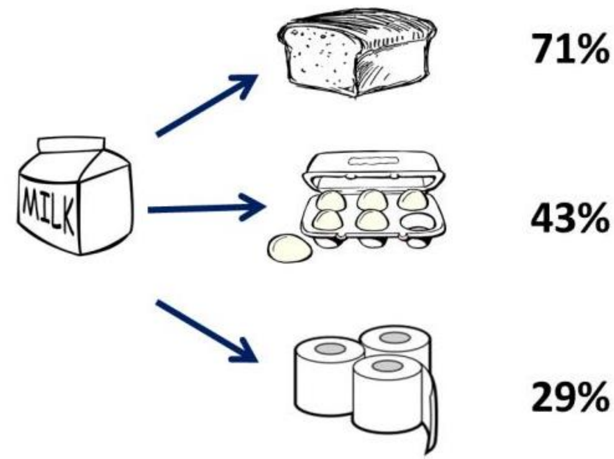
**Attributes**

**Objects**

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

# Transaction Data

- Also known as "**Market Basket Data**"
- Each transaction involves **a set of items**




**Of transactions that included milk:**
- 71% included bread
- 43% included eggs
- 29% included toilet paper

# Source of Data

- **Structured**
  - **Database** – **information** stored in company repository
  - E.g., HR, Finance, Product Inventory, Sales Record

- **Unstructured**
  - **Documents** – **process, ISO documents , emails etc.** stored in company repository
  - Web Pages / Sites
  - Online Social Media

# What is an Attribute?

- **Description** about the data
- An attribute will have a **value** (attribute value) assigned to it
- Attribute value has its own **properties:**
  - **Data Type**: Integer, Real, Character
  - **Limit**: Upper & lower values. Some no limit.
  - **Measurement scale**: Meter or Feet
  - **Numerical or Symbolic** (black, while, brown)
- It may **vary from one object to another** – person's eye colour
- It may **vary from time to another for the same object** – person's weight
- Attribute can have **discrete** or **continuous** values

To analyze objects. We need to assign **numbers or symbols** to an attribute

# Discrete vs. Continuous Attributes

- **Discrete** Attribute
  - Has only a **finite** or **countable infinite set of values**
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as **integer variables**.
  - Note: binary attributes are a special case of discrete attributes

- **Continuous** Attribute
  - Has **real numbers** as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a **finite number of digits**.
  - Continuous attributes are typically represented as **floating-point variables**.

# Types of Attribute Values

- Nominal
  - E.g., profession, ID numbers, eye color, zip codes
- Ordinal
  - E.g., rankings (e.g., army, professions), grades, height in {tall, medium, short}
- Binary
  - E.g., medical test (positive vs. negative)
- Interval
  - E.g., calendar dates, body temperatures
- Ratio
  - E.g., temperature in Kelvin, length, time, counts

# Nominal

- **Categorical / Qualitative**
- Nominal values provide only **enough information** to **distinguish** one object from another
- Examples are eye color, NRIC number, Postal codes, marital status
- The attribute value has **distinctiveness**. It means that you can check whether an attribute is **equal or not equal to a value**.
- Operators are: =, <>
- Any transformation done must have a **one-to-one mapping** and result with a permutation of values

# Binary

- **Categorical/Qualitative**
- Nominal attribute with only **2 states** (**0 and 1**)
- 2 types of binary attribute
  - **Symmetric** binary: both outcomes **equally important**
    - E.g., gender
  - **Asymmetric** binary: outcomes **not equally important**.
    - E.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

# Ordinal

- **Categorical/Qualitative**
- Ordinal values provide only **enough information** to **distinguish** one object from another and **order** the objects
- E.g., length = {short, medium, long}, exam grades
- The attribute value has **distinctiveness** and **order** . It means that you can check whether an attribute is **equal to, not equal to, greater than or smaller than a value.**
- Operators are: **=, <>, > , <**
- Any transformation done must **ensure the order is preserved**
- Transformation formula : **new_value = f(old_value)**. Must be a **monotonic** function.
  - {short, medium, long} → {1,2,3}

# Interval

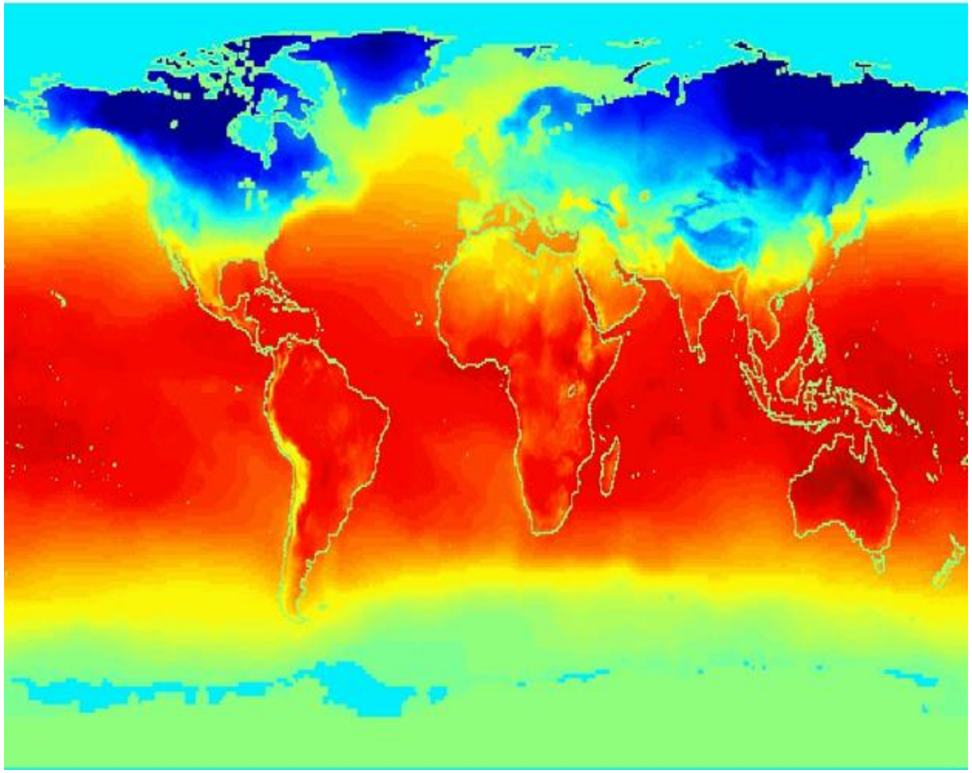
- **Numeric/Quantitative**
- Measured on a **scale of equal-sized units**
- **Differences between attribute values are meaningful**
- E.g., calendar dates, temperature in Fahrenheit /Celsius
- The attribute value can be **added or subtracted** from a value
- Operators are: **=, <>, > , < , +, -**
- Transformation formula : **new_value = a * old_value + b**
  - F = (9/5) C + 32; F = Fahrenheit, C = Celsius

# Ratio

- **Numeric/Quantitative**
- Both **differences and ratio between attribute values are meaningful**
- Values as being **an order of magnitude larger than the unit of measurement**
- E.g., age, temperature in Kelvin, mass, length
- The attribute value can be **added, subtracted or multiplied** from a value
- Operators are: **=, <>, > , < , +, -, * , /**
- Transformation formula : **new_value = a * old_value**
  - 1 hour = 3600 seconds
  - 2 hours is always twice of 1 hour
  - 10 K° is twice as high as 5 K°

# Spatial Data

- Data that have a **spatial component**, it means that data are connected to a place in the Earth.
- Data set that is based on **geographical locations**
- Diagram shows "Average Monthly Temperature of Land and Ocean"



# Outline

- General data characteristics
- ✓ **Why preprocess the data?**
- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Summary

# Why preprocess data?

Dirty data

incomplete

- lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- e.g., occupation=“ ” (missing data)

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

noise

- containing noise, errors, or outliers
- e.g., Salary=“−10” (an error)

# Why Is Data 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.
  - Human/hardware/software problems
- 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

# Why Preprocessing is Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data

- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse.

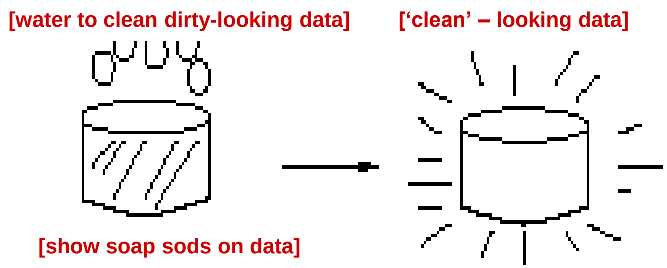
  — Bill Inmon

# Major Tasks in Data Preprocessing

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

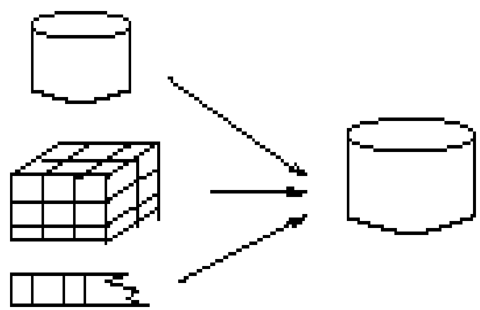
# Forms of Data Preprocessing

**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**

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

Normalization and aggregation

**Data Reduction & / Data discretization**
***(with particular importance, esp for numerical values)***

| | A1 | A2 | A3 | ... | A126 |
|---|---|---|---|---|---|
| T1 | | | | | |
| T2 | | | | | |
| T3 | | | | | |
| T4 | | | | | |
| ... | | | | | |
| T2000 | | | | | |

→

| | A1 | A3 | ... | A115 |
|---|---|---|---|---|
| T1 | | | | |
| T4 | | | | |
| ... | | | | |
| T1456 | | | | |

Obtains reduced representation in volume but produces the same or similar analytical results.

# 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

# Outline

- General data characteristics
- Why preprocess the data?
- ✓ **Data cleaning**
- Data integration
- Data transformation
- Data reduction
- Summary

# Data Cleaning

- To clean data from:
  - Incomplete /missing data
  - Noisy data
  - Inconsistent /outliers
- Importance
  - "Data cleaning is one of the three biggest problems in data warehousing"—Ralph Kimball
  - "Data cleaning is the number one problem in data warehousing"—DCI survey
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

# Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

# How to Handle Missing Data?

- Ignore the tuple, usually done when class label is missing (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 requires 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)

# Simple Discretization Methods - Binning

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

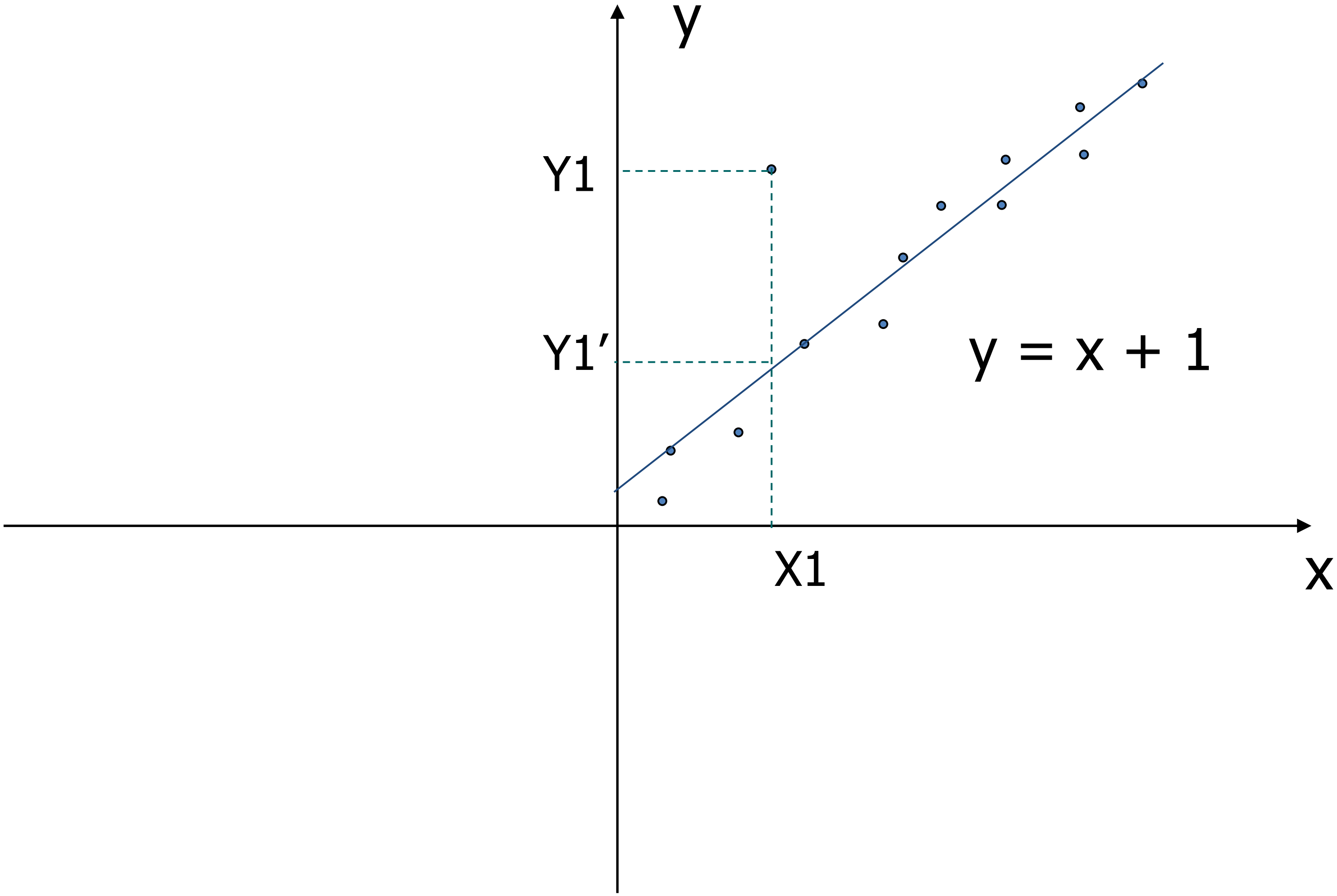
# Binning Methods for Data Smoothing

- Sorted data for price (in dollars):

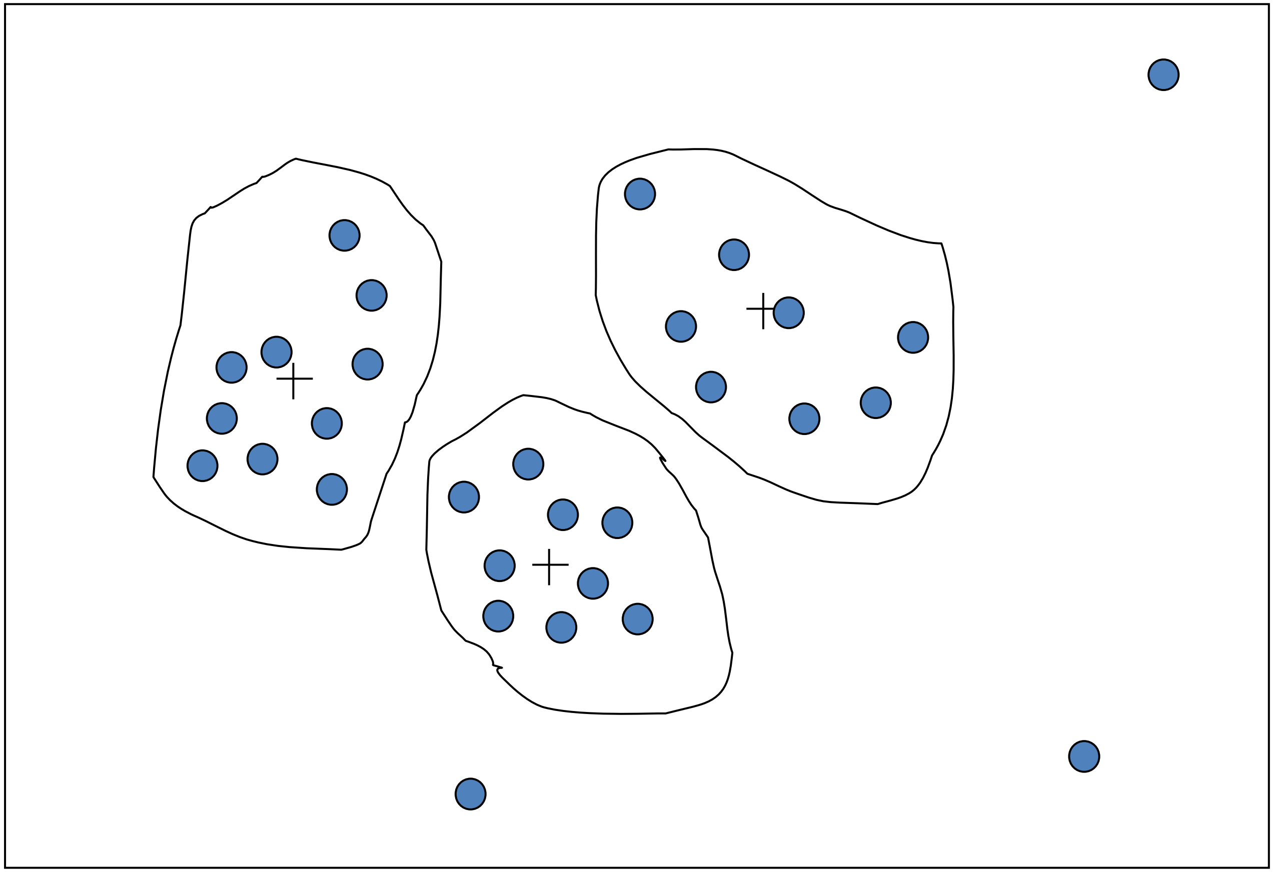
  4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

- Partition into (equi-depth) 4 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



# Cluster Analysis



# Outline

- General data characteristics
- Why preprocess the data?
- Data cleaning
- ✓ **Data integration**
- Data transformation
- Data reduction
- Summary

# Data Integration

- 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
- 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

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

# Correlation Analysis (Numerical 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, $\sigma_p$ and $\sigma_q$ are the respective standard deviation of p and q, and $\Sigma(pq)$ is the sum of the pq cross-product.
- If $r_{p,q} > 0$, p and q are positively correlated (p's values increase as q's). The higher, the stronger correlation.
- $r_{p,q} = 0$: independent
- $r_{pq} < 0$: negatively correlated

# 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

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

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

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

- Visually Evaluating Correlation
- Scatter plots showing the similarity from –1 to 1.

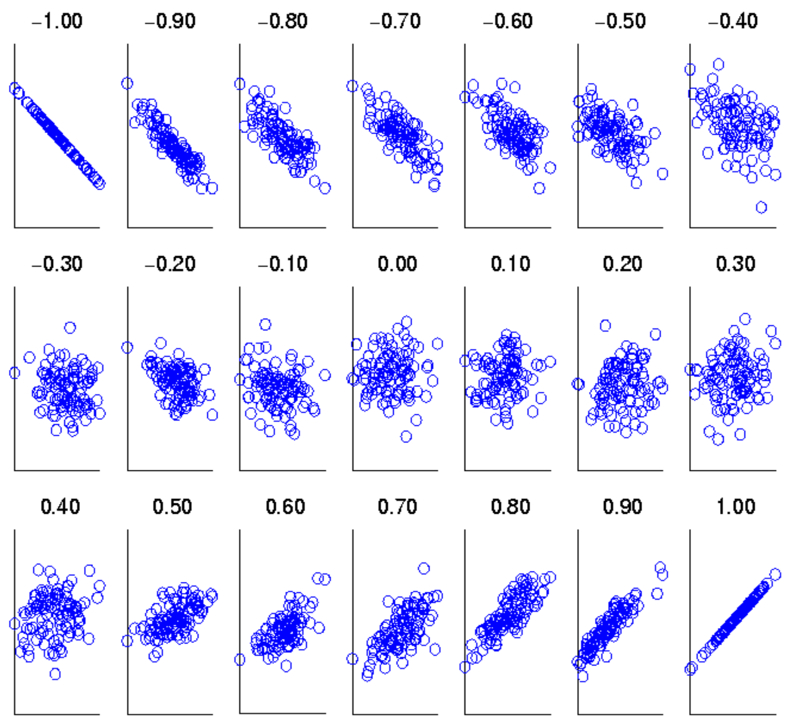



**Figure 5.11.** Scatter plots illustrating correlations from -1 to 1.

# Outline

- General data characteristics
- Why preprocess the data?
- Data cleaning
- Data integration
- ✓ **Data transformation**
- Data reduction
- 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
  - Aggregation: Summarization, data cube construction
  - Generalization: Concept hierarchy climbing
  - Normalization: Scaled to fall within a small, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Attribute/feature construction
    - New attributes constructed from the given ones

# Data Transformation: Normalization

- Min-max normalization: to [new_minA, new_maxA]

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

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 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 by decimal scaling

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

# Z-Score (Example)

| v | v’ | | | v | v’ | | |
|---|---|---|---|---|---|---|---|
| 0.18 | -0.84 | Avg | 0.68 | 20 | -.26 | Avg | 34.3 |
| 0.60 | -0.14 | sdev | 0.59 | 40 | .11 | sdev | 55.9 |
| 0.52 | -0.27 | | | 5 | .55 | | |
| 0.25 | -0.72 | | | 70 | 4 | | |
| 0.80 | 0.20 | | | 32 | -.05 | | |
| 0.55 | -0.22 | | | 8 | -.48 | | |
| 0.92 | 0.40 | | | 5 | -.53 | | |
| 0.21 | -0.79 | | | 15 | -.35 | | |
| 0.64 | -0.07 | | | 250 | 3.87 | | |
| 0.20 | -0.80 | | | 32 | -.05 | | |
| 0.63 | -0.09 | | | 18 | -.30 | | |
| 0.70 | 0.04 | | | 10 | -.44 | | |
| 0.67 | -0.02 | | | -14 | -.87 | | |
| 0.58 | -0.17 | | | 22 | -.23 | | |
| 0.98 | 0.50 | | | 45 | .20 | | |
| 0.81 | 0.22 | | | 60 | .47 | | |
| 0.10 | -0.97 | | | -5 | -.71 | | |
| 0.82 | 0.24 | | | 7 | -.49 | | |
| 0.50 | -0.30 | | | 2 | -.58 | | |
| 3.00 | 3.87 | | | 4 | -.55 | | |

# Outline

- General data characteristics
- Why preprocess the data?
- Data cleaning
- Data integration
- Data transformation
- ✓ **Data reduction**
- Summary

# Data Reduction Strategies

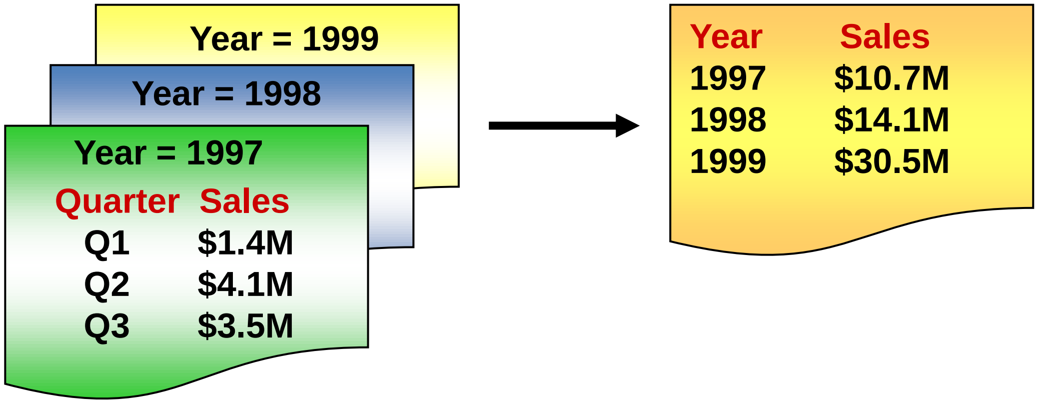
- Warehouse may store terabytes of data
  - Complex data analysis/mining may take a very long time to run on the complete data set

- Data reduction
  - Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

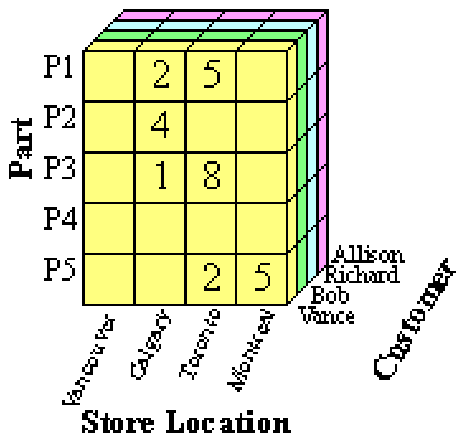
- Data reduction strategies
  - Data cube aggregation
  - Dimensionality reduction ➔ remove unimportant attributes
  - Data compression
  - Numerosity reduction
  - Discretization and concept hierarchy generation

# Data Cube Aggregation

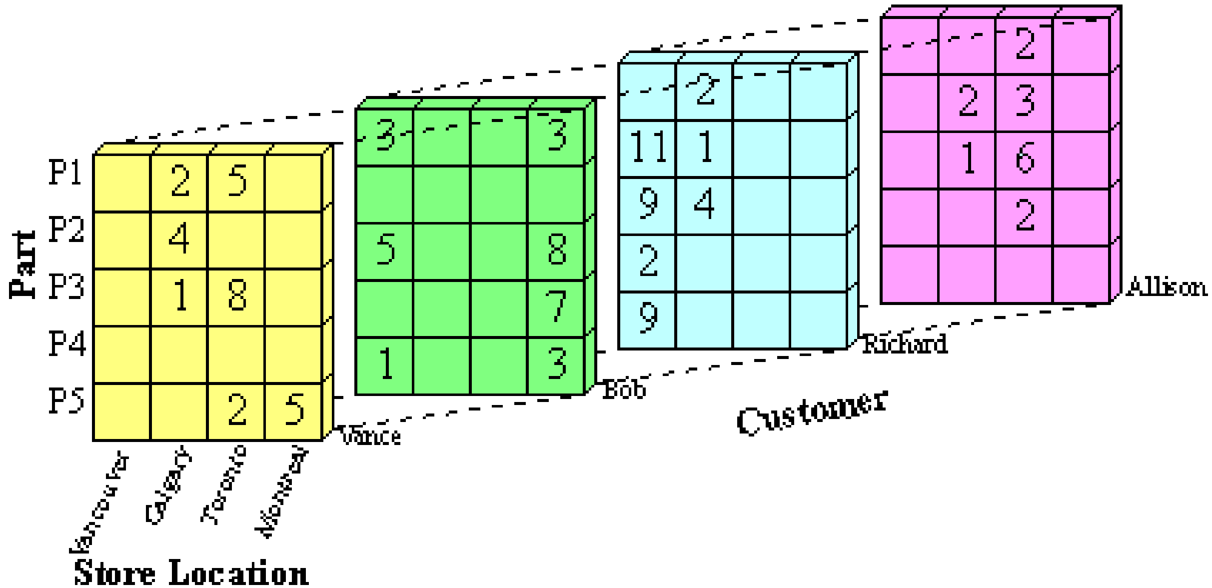
- The lowest level of a data cube
  - the aggregated data for an individual entity of interest
  - e.g., a customer in a phone calling data warehouse.
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation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



# Data reduction: Data Cube Aggregation




**Front View of Sample Data Cube**




**Entire View of Sample Data Cube**

# Attribute Reduction

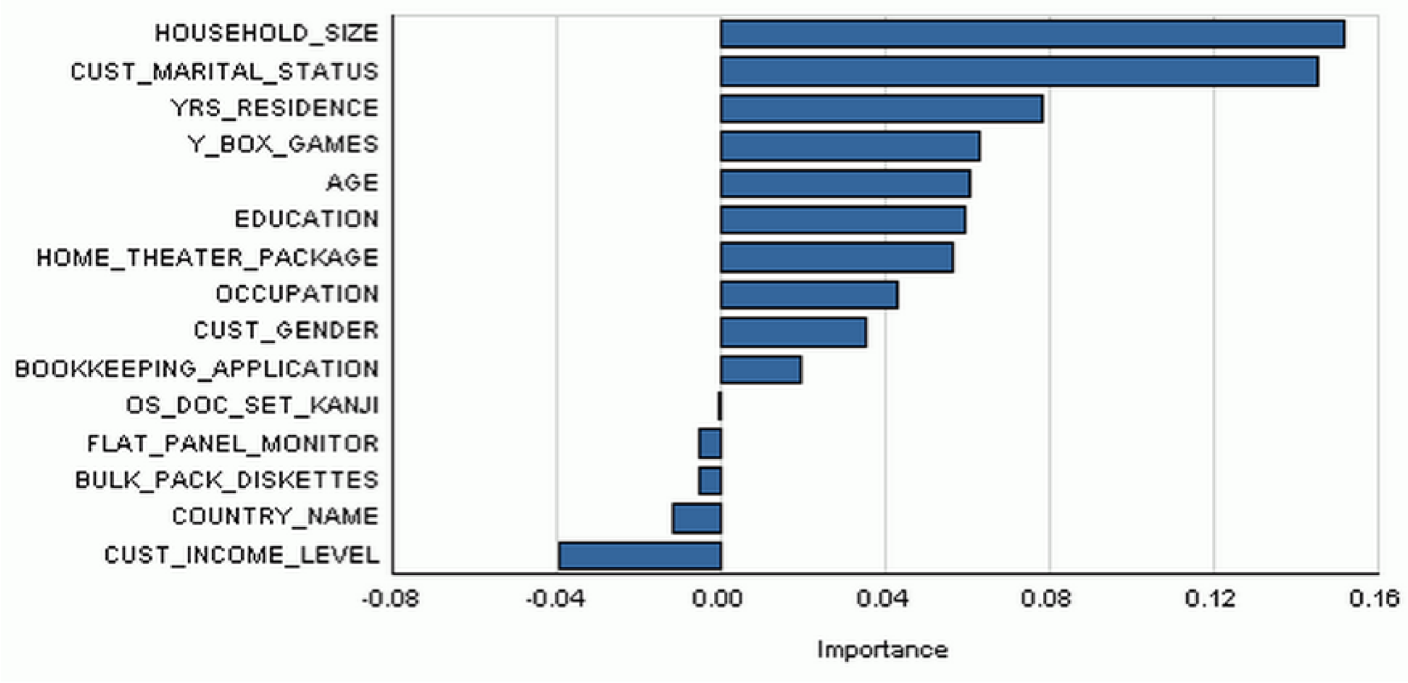
- When there are a **large number of variables** in the database.
- It is very likely that **subsets of variables are highly correlated** with each other.
- The accuracy and reliability of a classification or prediction model will **suffer** if we include **highly correlated variables** or **variables that are unrelated** to the outcome of interest because of over fitting. In model deployment, it can increase costs due to collection and processing of these variables.
- The dimensionality of a model is the **number of independent or input variables used by the model**. One of the key steps in data mining is therefore finding ways to **reduce dimensionality without sacrificing accuracy**.

# Attribute Reduction - Dimensionality Reduction

- Feature Selection
- Feature Extraction

*** Feature is another word for "attribute" or "variable".*

Which attribute or attributes can be considered **important** and should be **selected** ?



# Attribute Reduction – Feature Selection

- Also known as variable selection or attribute selection
- Selection of **optimal subset of attributes for better performance and accuracy**
- One common method in feature subset selection is called **StepWise regression**.
- StepWise regression has 2 strategies:-
  - Forward Selection (FS)
  - Backward Elimination (BE)

# Attribute Reduction – Feature Selection (Stepwise)

- **Forward selection**:
  - involves **starting with no attributes** in the model, testing the **addition of each attribute** using a chosen model comparison criterion, adding the attribute (if any) that improves the model the most, and **repeating** this process **until none improves** the model.

- **Backward elimination**:
  - involves **starting with all candidate attribute** , testing the **deletion of each attribute** using a chosen model comparison criterion, deleting the attribute (if any) that improves the model the most by being deleted, and **repeating** this process **until no further improvement** is possible.

# Dimensionality Reduction: Example of Heuristic Method

- **Forward selection**

Initial attribute set:

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

Initial reduced set:

{ }

➔ {A1}

➔ {A1, A4}

➔ Reduced attribute set:

{A1, A4, A6}

- **Backward selection**

Initial attribute set:

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

➔ {A1, A3, A4, A5, A6}

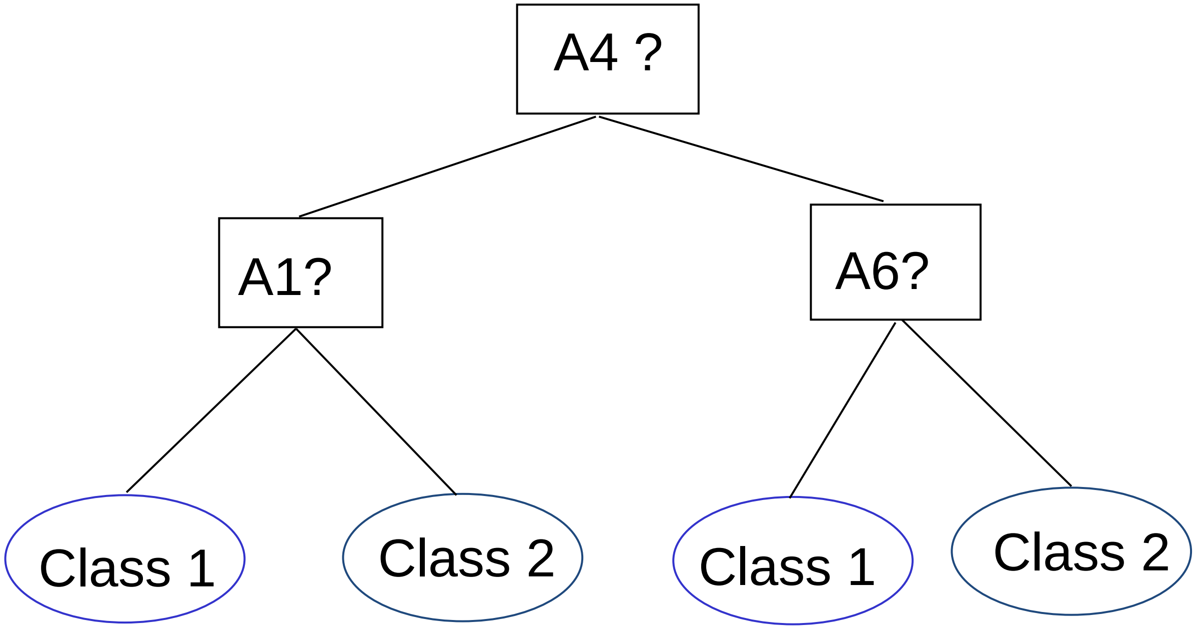
➔ {A1, A4, A5, A6}

➔ Reduced attribute set:

{A1, A4, A6}

# Dimensionality Reduction: Example of Decision Tree Induction

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




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

# Numerosity reduction

"Can we reduce the data volume by choosing alternative, 'smaller' forms of data representation?"

- Parametric methods
  - A model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data (outliers may also be stored)
  - e.g: Log-linear models: estimate discrete multidimensional probability distributions
- Non-parametric methods
  - Do not assume models; storing reduced representations of data
  - Major methods: histograms, clustering, & sampling

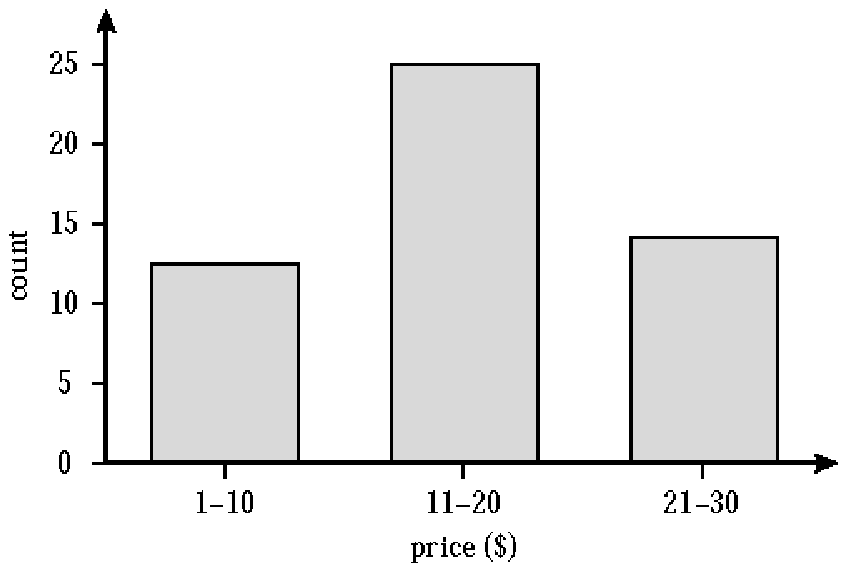
# Numerosity Reduction: Histograms

- 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

Example:

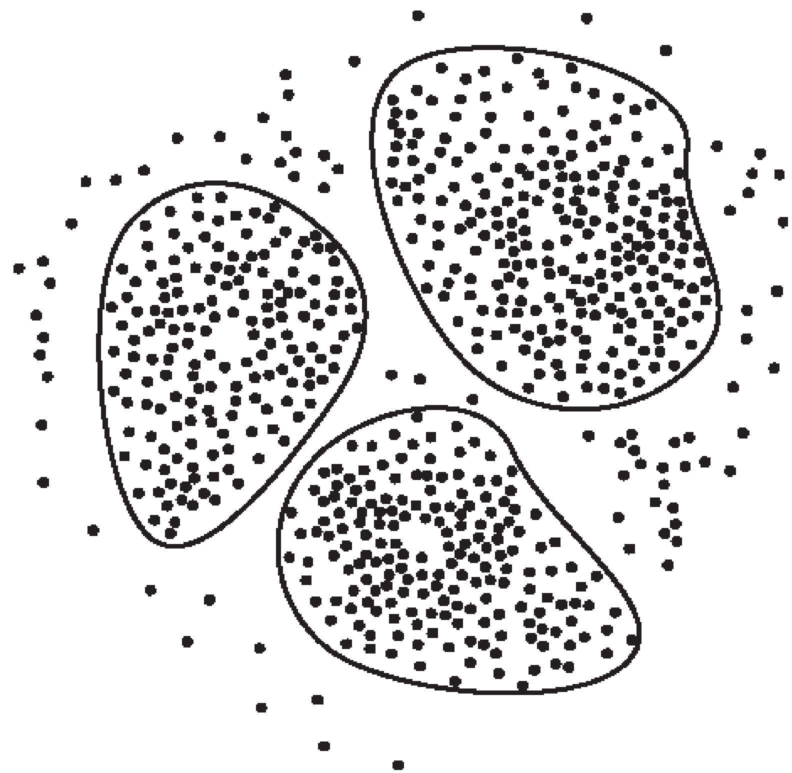
*list of prices of sold items:*

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.

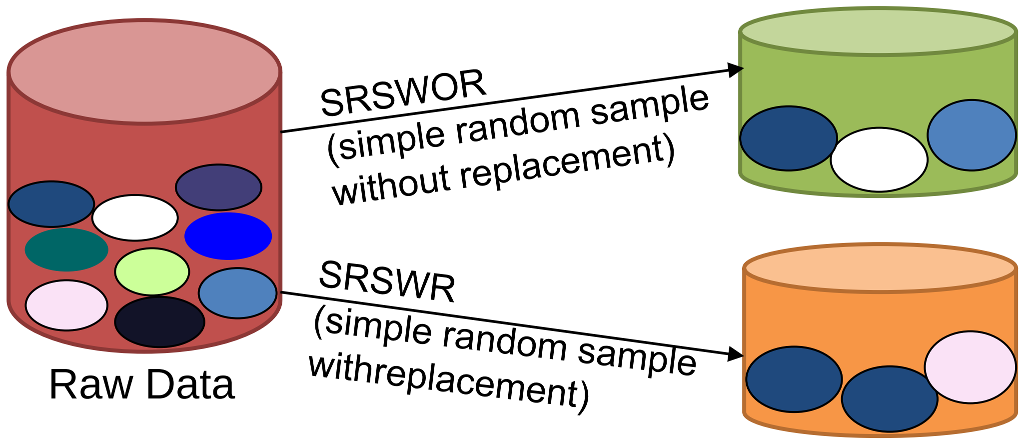


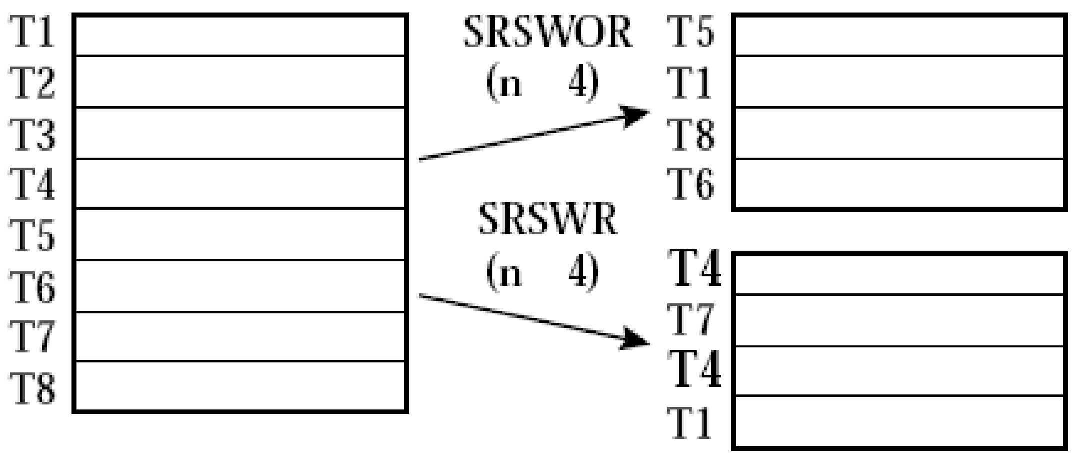
# Numerosity Reduction: Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) 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 (Chapter 8)

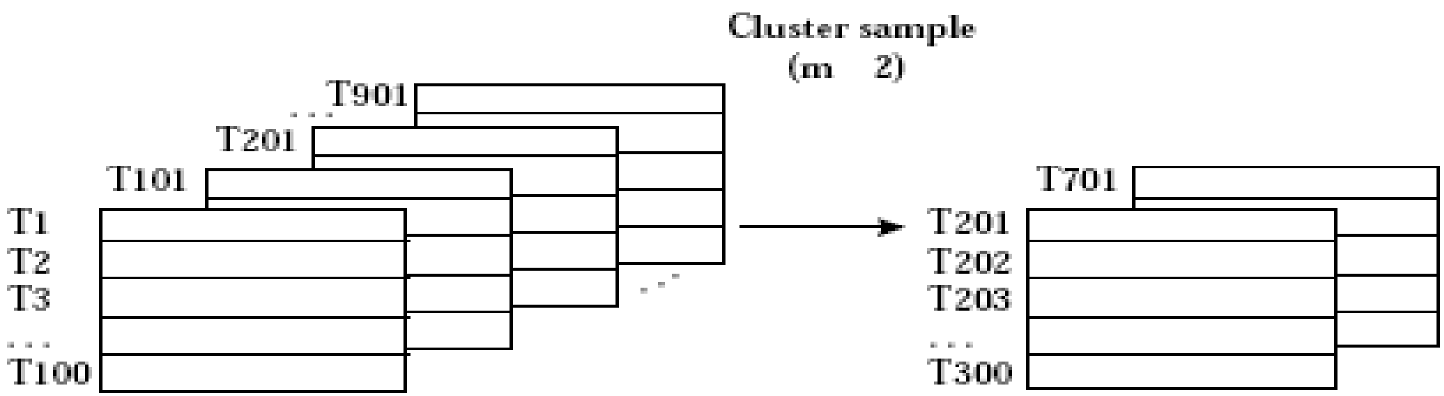


# Numerosity Reduction: Sampling (Sampling: With or without Replacement)



# Numerosity Reduction: Sampling (Cluster or Stratified Sampling)



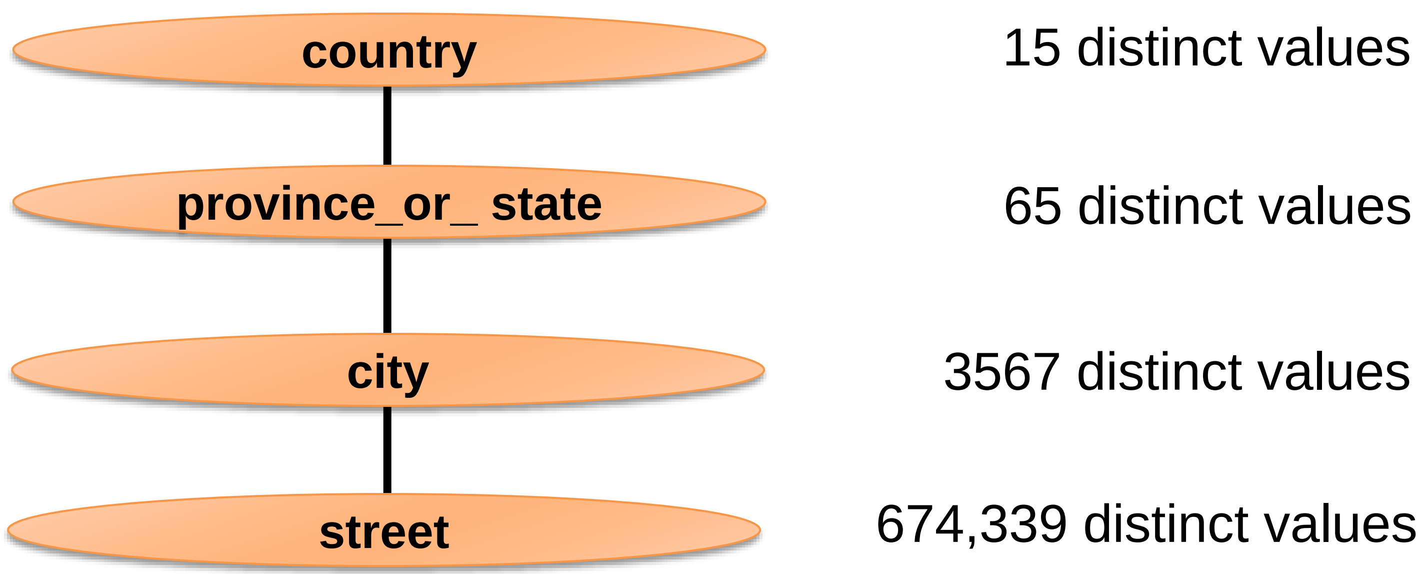

Stratified sample
(according to *age*)

| | |
|---|---|
| T38 | young |
| T256 | young |
| T307 | young |
| T391 | young |
| T96 | middle-aged |
| T117 | middle-aged |
| T138 | middle-aged |
| T263 | middle-aged |
| T290 | middle-aged |
| T308 | middle-aged |
| T326 | middle-aged |
| T387 | middle-aged |
| T69 | senior |
| T284 | senior |

| | |
|---|---|
| T38 | young |
| T391 | young |
| T117 | middle-aged |
| T138 | middle-aged |
| T290 | middle-aged |
| T326 | middle-aged |
| T69 | senior |

# 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



# Outline

- General data characteristics
- Why preprocess the data?
- Data cleaning
- Data integration
- Data transformation
- Data reduction
- ✓ **Summary**

# Summary of Data Preprocessing Methods

Data preprocessing

| | | |
|---|---|---|
| Data cleaning | Missing / incomplete data | 1. Ignore the tuple<br>2. Fill in it manually<br>3. Fill in it automatically |
| | Noisy data / inconsistence / outliers | 1. Binning<br>2. Regression<br>3. Clustering<br>4. Combined computer with human inspection |
| Data integration | 1. Correlation analysis | |
| Data transformation | 1. Smoothing<br>2. Aggregation<br>3. Generalization<br>4. Normalization<br>5. Attribute / feature construction | |
| Data reduction | 1. Data cube aggregation<br>2. Attribute Reduction - Dimensionality reduction<br>3. Numerosity reduction<br>5. Discretization and concept hierarchy generation | |

# Summary

- Data preparation/preprocessing: A big issue for data mining
- Data description, data exploration, and measure data similarity set the base for quality data preprocessing
- Data preparation includes
  - Data cleaning
  - Data integration and data transformation
  - Data reduction (dimensionality and numerosity reduction)
- A lot a methods have been developed but data preprocessing still an active area of research