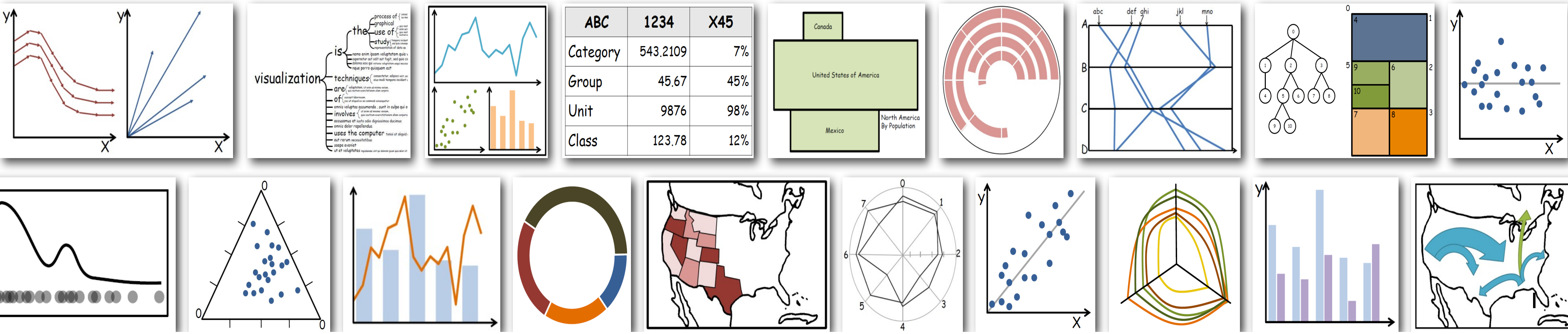


Lecture : Visualization model & pipeline

+ Data Preparation

DATA ANALYSIS & VISUALIZATION

FALL 2021

Dr. Muhammad Faisal Cheema

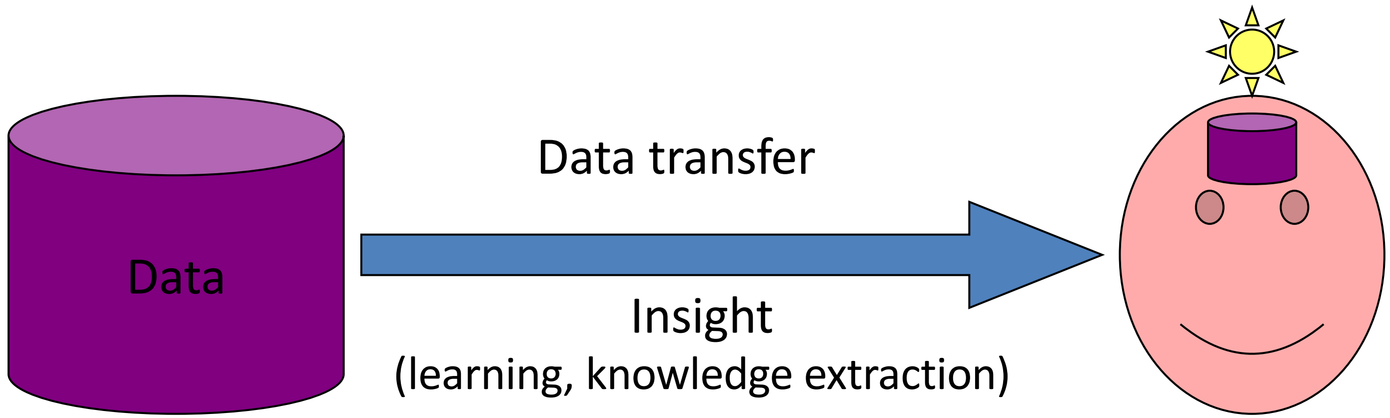
FAST-NU

Basic Visualization Model

The purpose of computing is about insight, not numbers

- R. W. Hamming

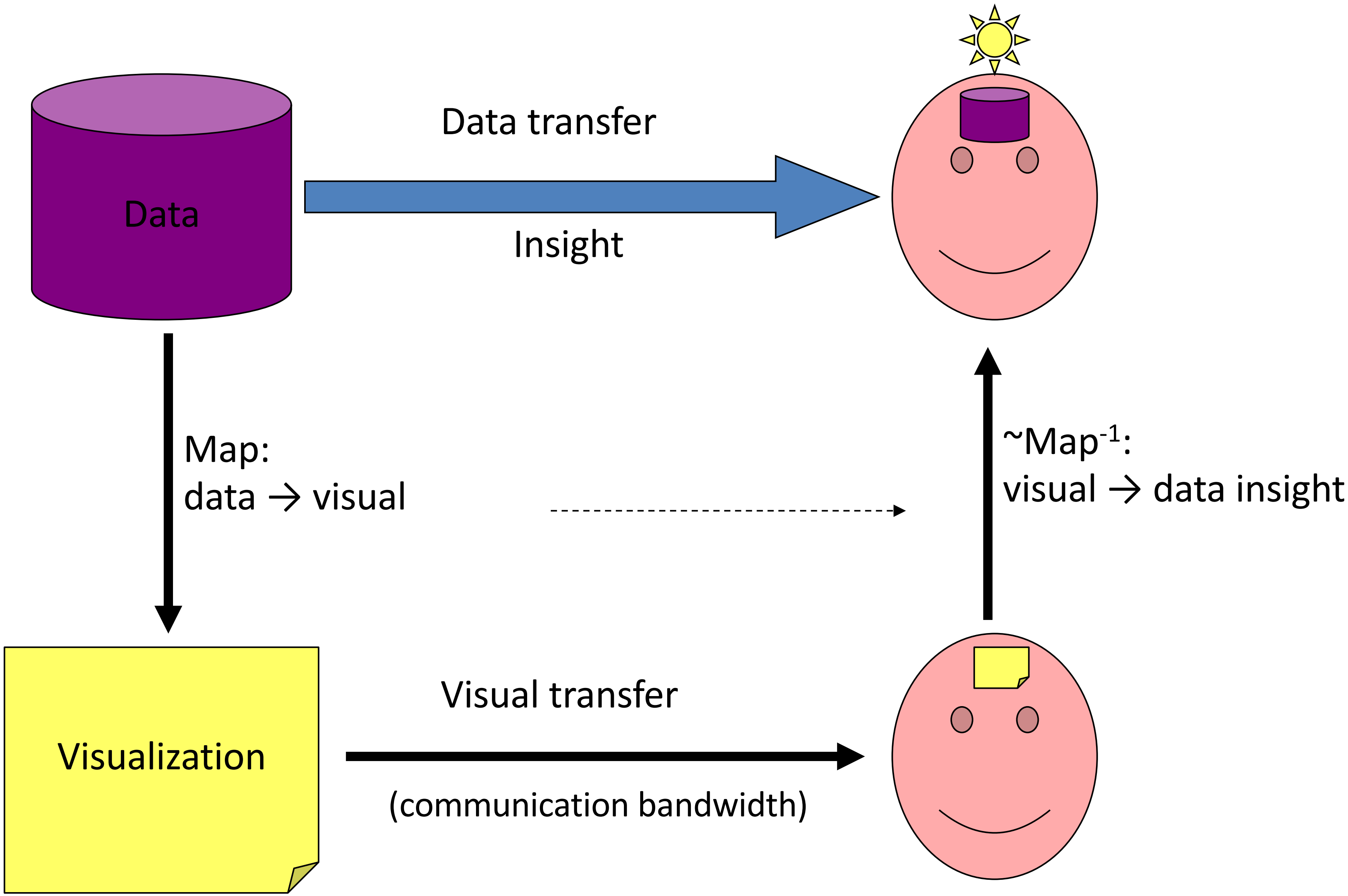
Goal


Data
Data transfer
Insight
(learning, knowledge extraction)

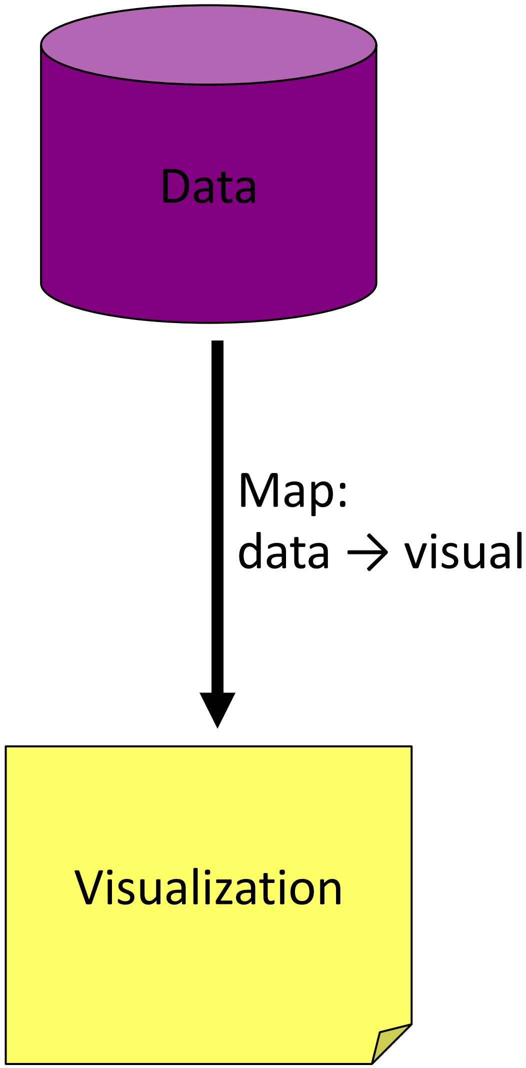
The purpose of visualization is about insight, not pictures

- Card, Mackinlay, Schneiderman

Method


Data transfer
Data
Insight
Map:
data → visual
~Map^{-1}:
visual → data insight
Visualization
Visual transfer
(communication bandwidth)

Visual Mappings



Visual Mappings must be:

- Computable (math)

  visual = f(data)

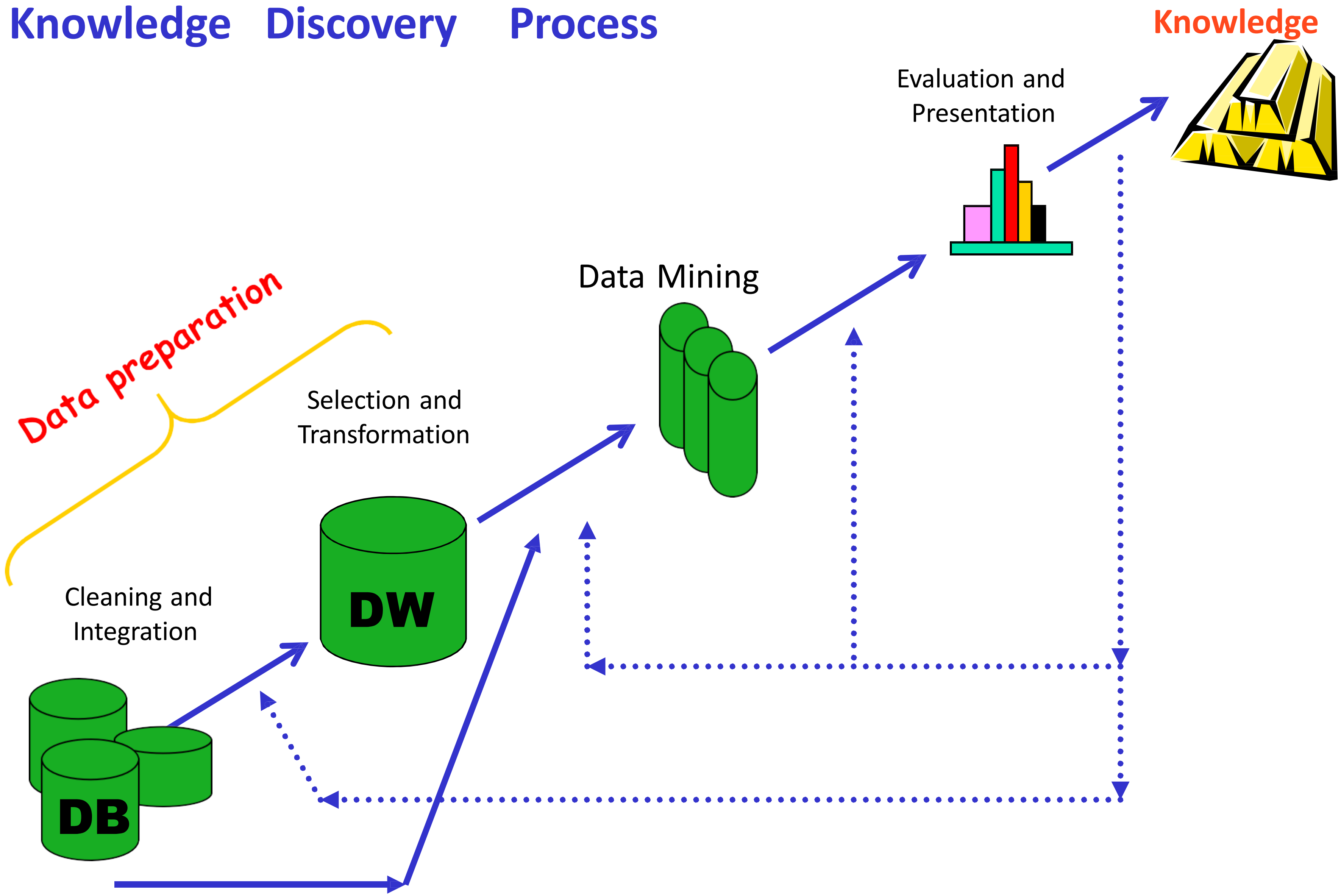
- Comprehensible (invertible)

  data = f^{-1}(visual)

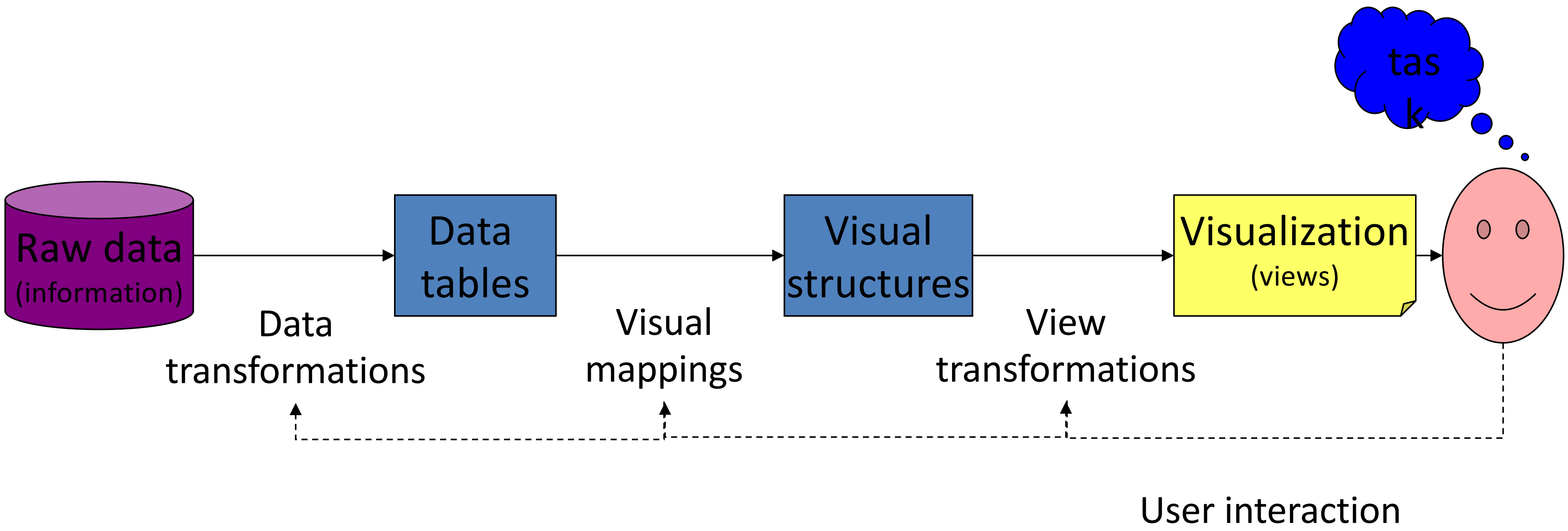
- *Creative!*

The Visualization Pipeline

Data Preparation as a step in the Knowledge Discovery Process



The Visualization Pipeline (InfoVis)



Data Preprocessing and transformation

==

Data preparation

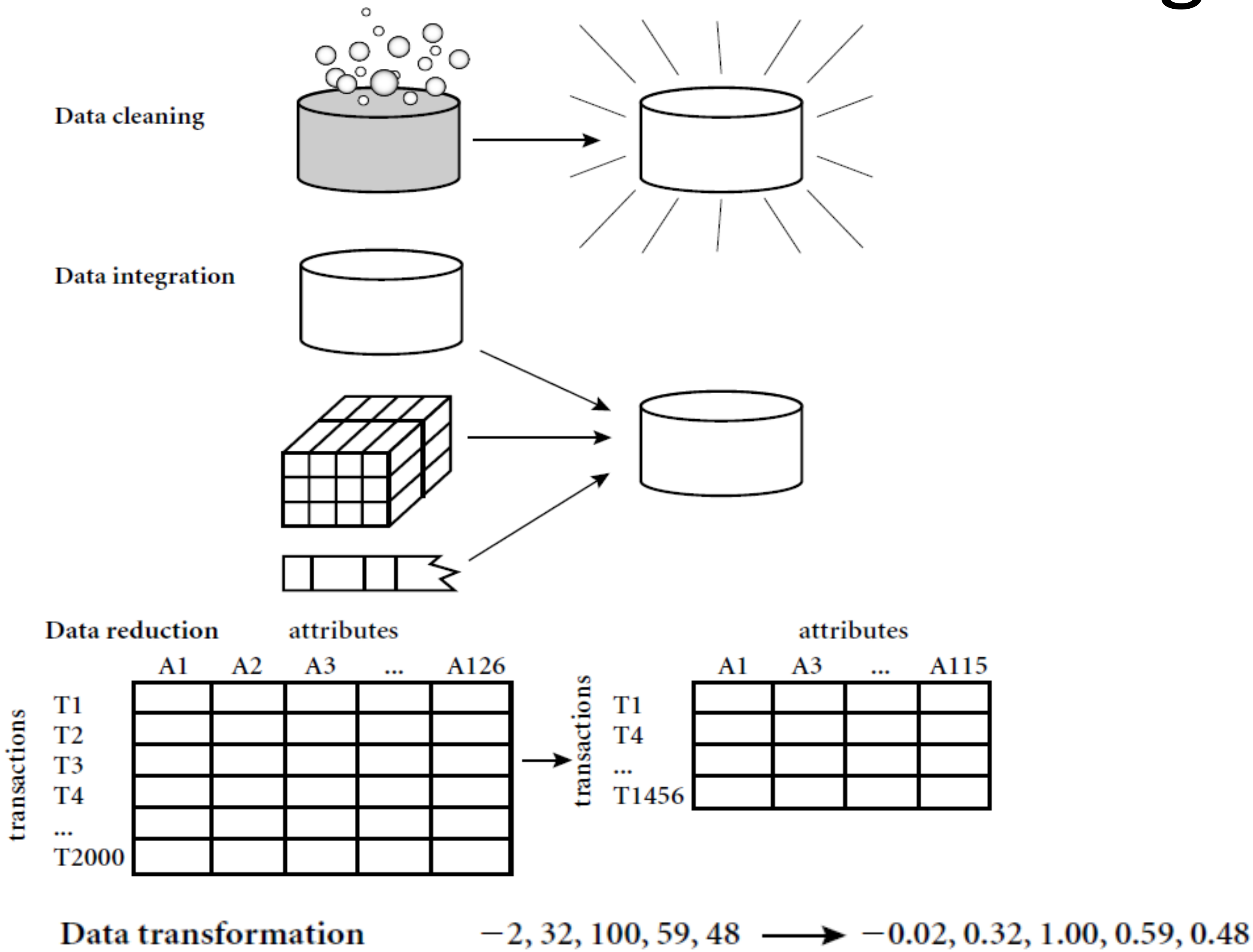
Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

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

Forms of Data Preprocessing



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

15

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - 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*=“03/07/2010”
    - Was rating “1, 2, 3”, now rating “A, B, C”
    - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone’s birthday?

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

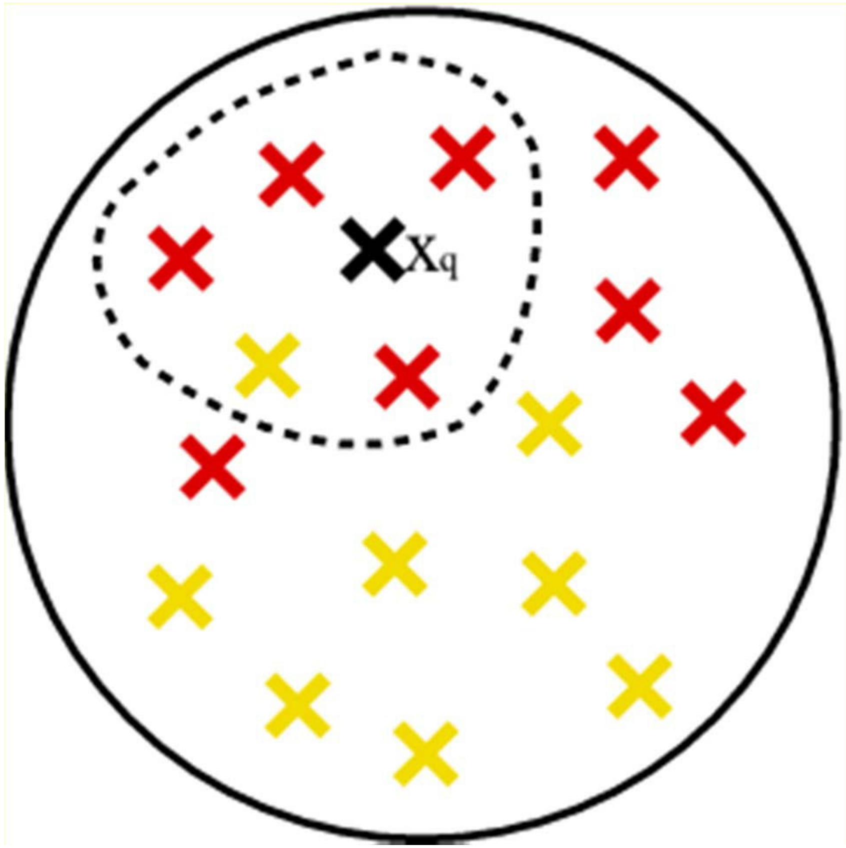
How to Handle Missing Data?

- 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”, or “-∞” a new class?!
  - the attribute mean (symmetric) or median (skewed)
  - the attribute mean for all samples belonging to the same class, e.g. average income in same *credit_risk*
  - the most probable value: inference-based such as Bayesian formula or decision tree

How to Handle Missing Data?

- Fill in it automatically with
  - the most probable value:
    - Inference-based such as Bayesian formula or decision tree
    - Identify relationships among variables
      - Linear regression, Multiple linear regression, Nonlinear regression
    - Nearest-Neighbour estimator
      - Finding the k neighbours nearest to the point and fill in the most frequent value or the average value
      - Finding neighbours in a large dataset may be slow

Nearest-Neighbour



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.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

How to Handle Noisy Data?

- Regression
  - smooth by fitting the data into regression functions
  - *Linear regression* involves finding “best” line to fit two attributes, so that one attribute can be used to predict the other.
  - *Multiple Linear regression* – more than two attributes involved and data fit to a multidimensional surface
- Clustering
  - detect and remove outliers
  - Outliers – values outside of the set of clusters
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

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

24

Data Integration

- **Tuple Duplication**
  - The use of denormalized tables (improve performance by avoiding joins) creates data redundancy
  - Inconsistencies often arise between various duplicates, due to inaccurate data entry
- **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
  - Hotel chain – price difference in currencies and services and taxes
  - Attributes may differ on level of abstraction, e.g. total_sales – at branch level or region level

Data Integration- Entity identification problem

- **Data integration**:
  - Combines data from multiple sources into a coherent store
  - Integrate metadata from different sources
- Entity identification problem:
  - *Schema integration* and *object matching*: e.g., A.cust-id ≡ B.cust-#
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
  - Metadata – name, meaning, data type, range, null rules
  - Metadata can help avoid errors in schema integration
  - Metadata may help transform the data
  - When matching attributes from two databases, *structure* of data should be checked

26

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

27

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

28

Data Reduction Strategies

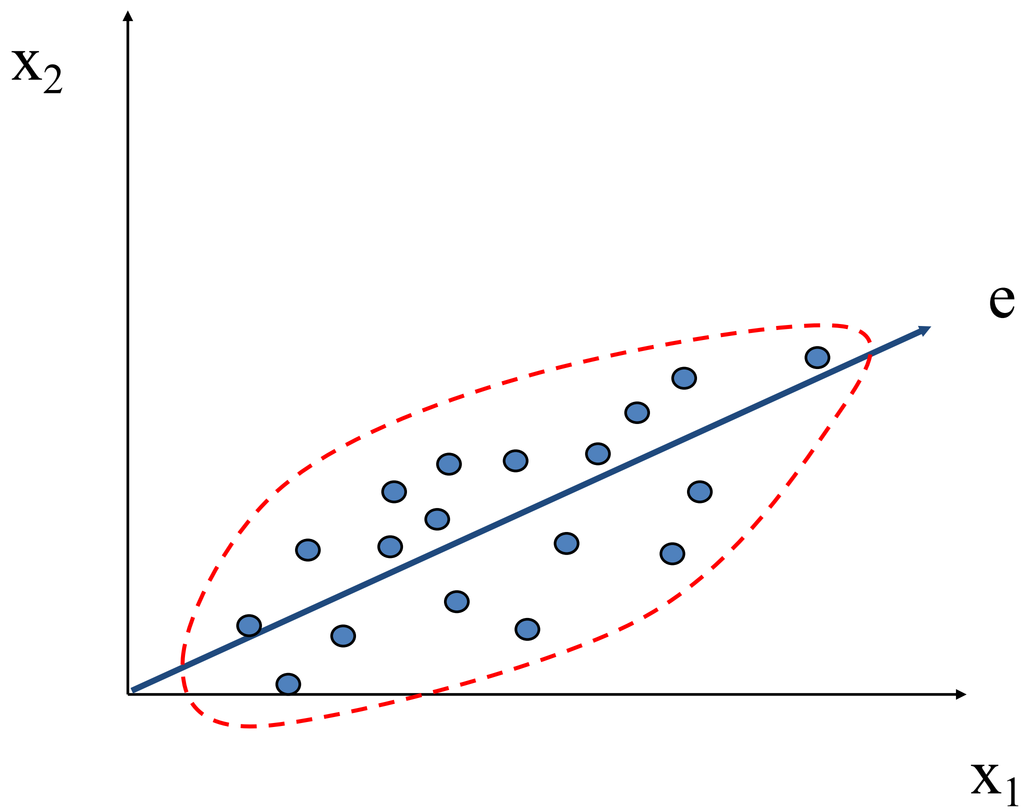
- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run 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)
    - Parametric - Regression and Log-Linear Models
    - Non-parametric - Histograms, clustering, sampling
    - Data cube aggregation
  - Data compression
    - Lossless - Reconstruction without any loss of information
    - Lossy – reconstruct only an approximation of the original data

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)

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

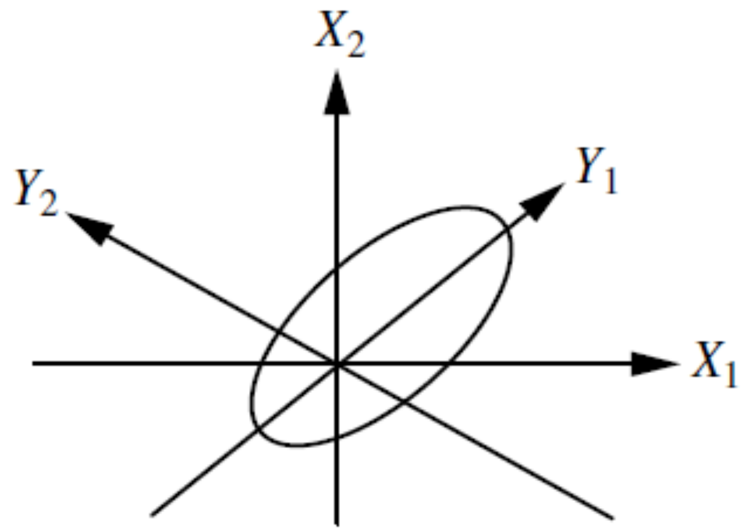


Principal Component Analysis (Steps)

- Given *N* data vectors from *n*-dimensions, find $k \leq n$ orthogonal vectors (*principal components*) that can be best used to represent data
  - Normalize input data: Each attribute falls within the same range
  - Compute *k* orthonormal (unit) vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the *k* principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength. The principal components serve as new set of axes for the data, giving important information on variance
  - Since the components are sorted, the size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)

Principal Component Analysis

- Works for numeric data only
- PCA can be applied to ordered and unordered attributes and can handle sparse and skewed data
- Multidimensional handled by reducing to two-dimensional
- PCA handles sparse data better than wavelet transforms



Y_1 and Y_2 are first two principal components

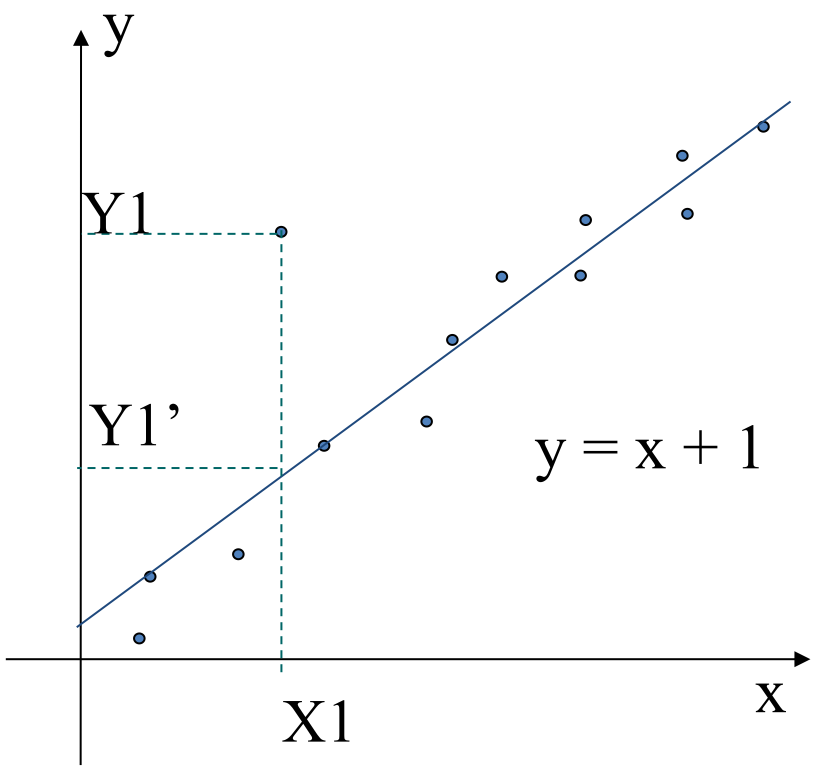
Data Reduction 2: Numerosity Reduction

- 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)
  - Ex.: Log-linear models—obtain value at a point in n-dimensional 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 Linear 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
  - Consider each tuple as a point in an n-dimensional space

Regression Analysis

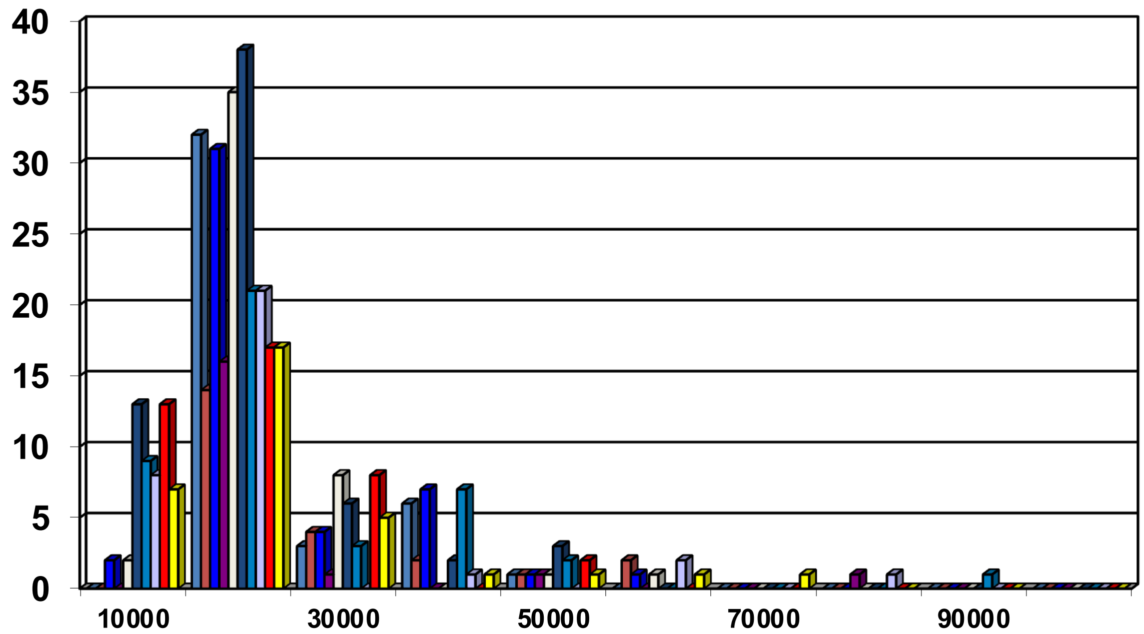


- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a ***dependent variable*** (also called ***response variable*** or *measurement*) and of one or more *independent variables* (aka. ***explanatory variables*** or ***predictors***)
- 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

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): frequency of each bucket is constant



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

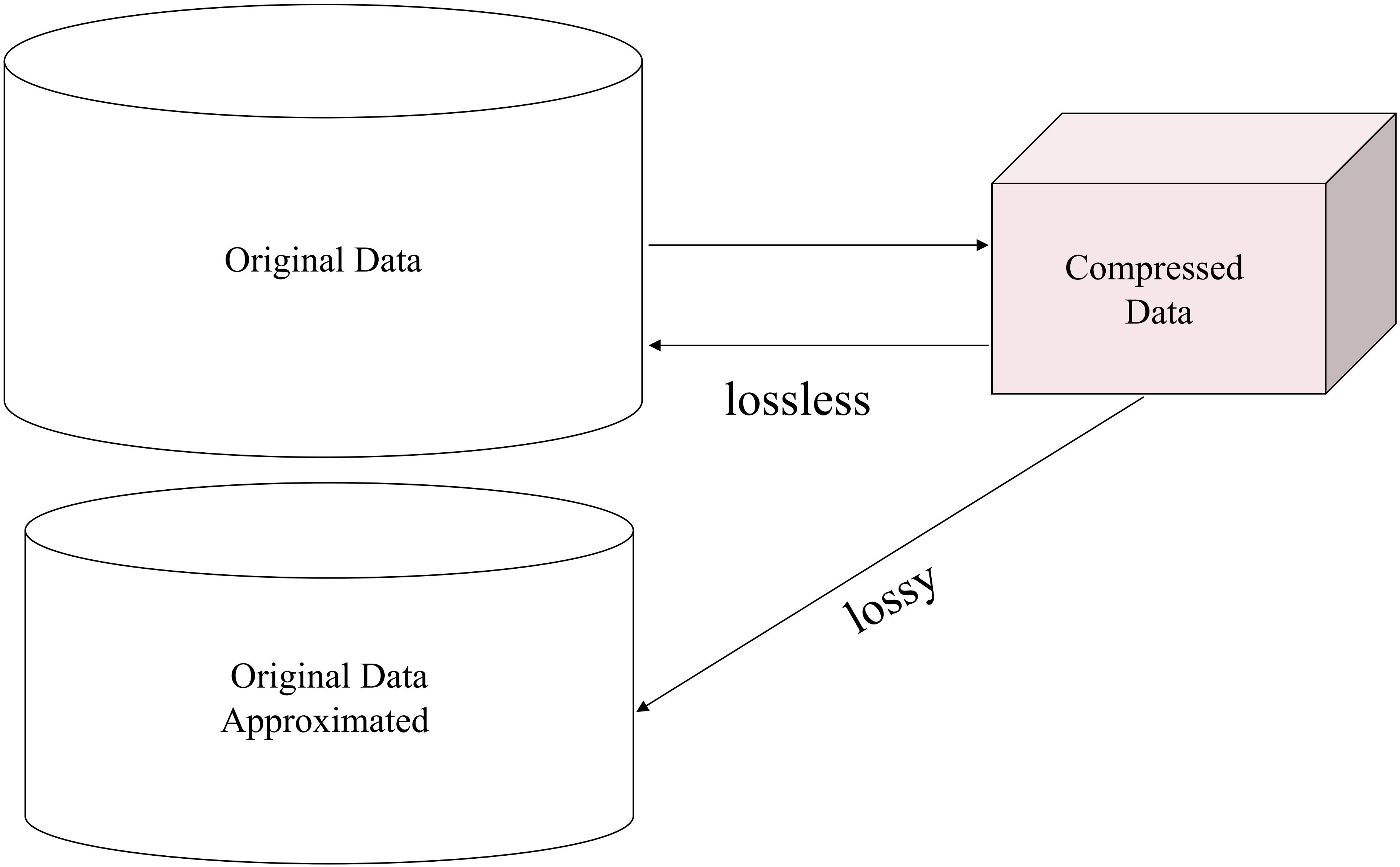
Sampling

- 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 (page at a time)

Data Reduction 3: Data Compression

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless, but only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
  - Typically short and vary slowly with time
- Dimensionality and numerosity reduction may also be considered as forms of data compression

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
  - Statistics: Descriptive and Distribution
  - Smoothing: Remove noise from data – binning, regression, clustering
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, used in 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

Descriptive Statistics: Univariate

- Range, Min/Max
  - Difference between minimum and maximum values in a data set
  - Larger range usually (but not always) indicates a large spread or deviation in the values of the data set.
- Average
  - Sum of all values divided by the number of values in the data set.
  - One measure of central **location** in the data set.
- Median
  - The middle value in a sorted data set. Half the values are greater and half are less than the median.
- Mode
  - The most frequent occurring value.
  - Another measure of central location in the data set.

Distribution Statistics

- Variance
  - One measure of dispersion (deviation from the mean) of a data set. The larger the variance, the greater is the average deviation of each datum from the average value
- Standard Deviation
  - the average deviation from the mean of a data set.
- Histograms and Normal Distribution

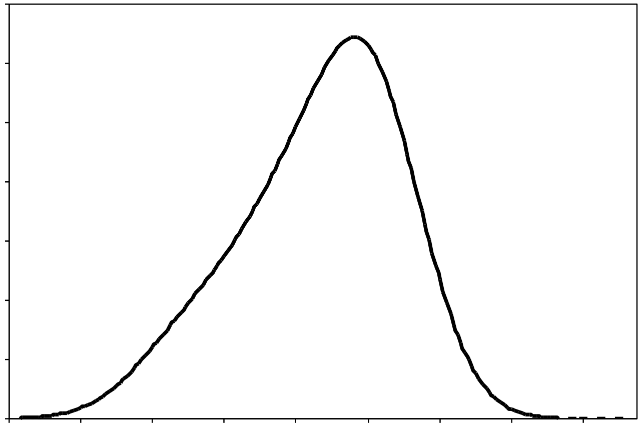
- Variance and SD are critical in analyzing your data distribution and determining how “meaningful” is the chosen average

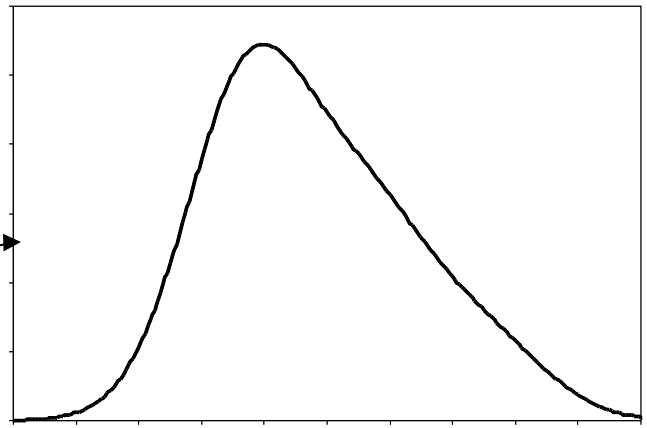
Distribution Statistics: Normal and Skewed Distributions

- When data are skewed, the mean and SD can be misleading
- Skewness

  sk= 3(mean-median)/SD

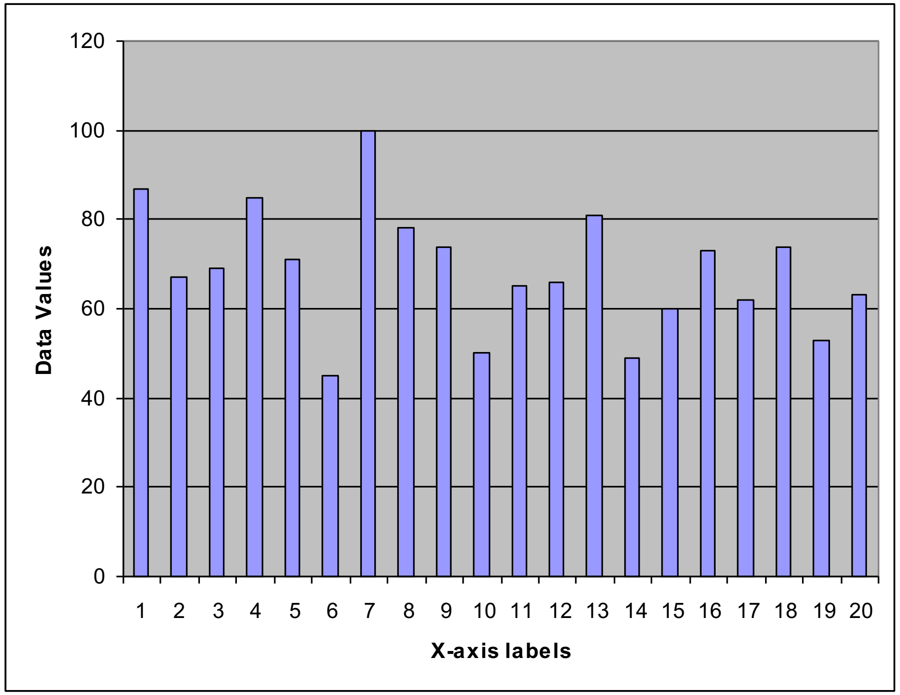
  If sk>|1| then distribution is non-symetrical
- Negatively skewed
  - Mean<Median
  - Sk is negative
- Positively Skewed
  - Mean>Median
  - Sk is positive





Distribution Statistics: Problems in reading distribution

- We can't really tell much about this data set
- Even Min and Max are hard to see

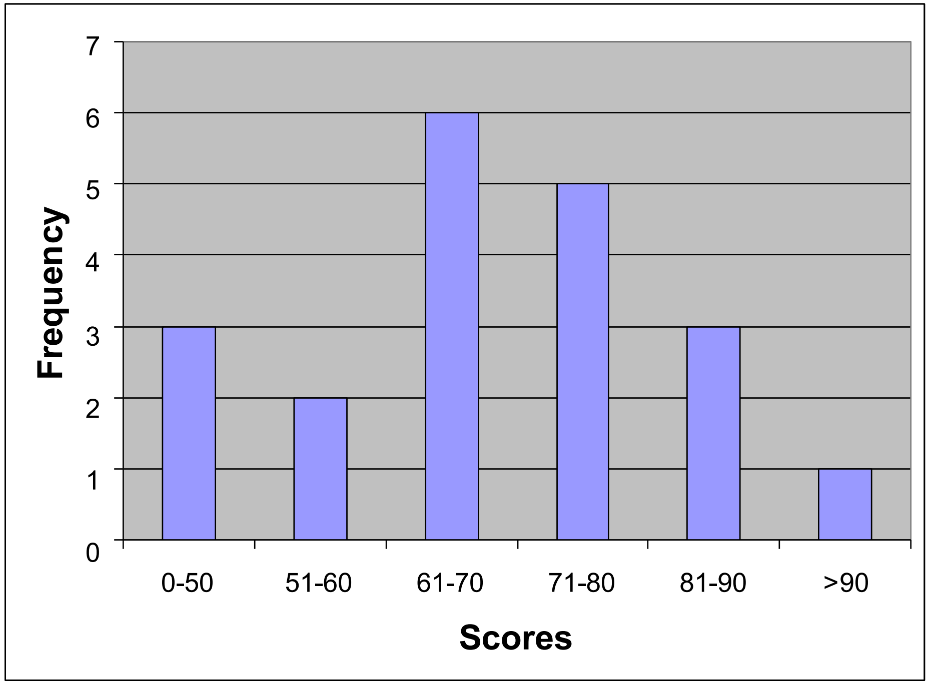


The data can be presented such that more statistical info can be estimated from the chart (average, standard deviation).

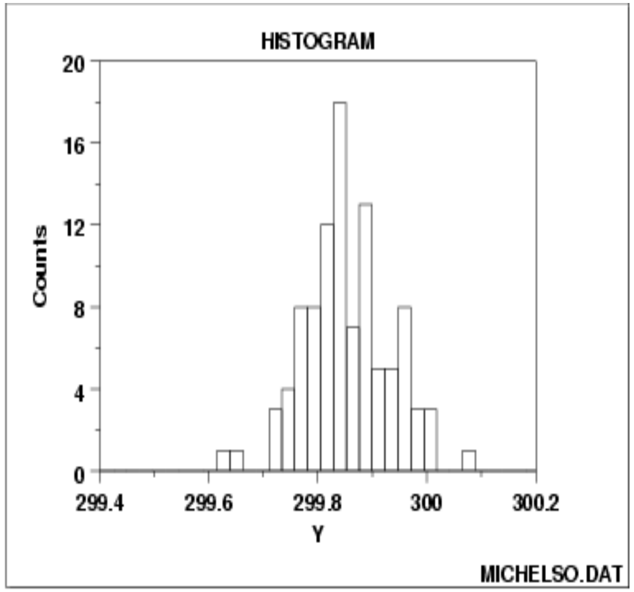
Distribution Statistics: Plotting the distribution

- Determine a frequency table (bins)
- A **histogram** is a column chart of the frequencies

Category Labels	Frequency
0-50	3
51-60	2
61-70	6
71-80	5
81-90	3
>90	1



Distribution Statistics: Histogram



- The histogram graphically shows the following:
  1. center (i.e., the location) of the data;
  2. spread (i.e., the scale) of the data;
  3. skewness of the data;
  4. presence of outliers; and
  5. presence of multiple modes in the data
- For small data sets, histograms can be misleading. Small changes in the data or to the bucket boundaries can result in very different histograms.
- For large data sets, histograms can be quite effective at illustrating general properties of the distribution.
- Histograms effectively only work with 1 variable at a time
  - Difficult to extend to 2 dimensions, not possible for >2
  - So histograms tell us nothing about the relationships among variables

Normalization

- The measurement unit can affect the data analysis
- Smaller unit leads to larger range and thus give more weight to an attribute
- Normalize data between [-1,1] or [0,1] to avoid dependence on choice of measurement unit
- **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

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

  – Min-max normalization preserves the relationships among the original data values

Normalization

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

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

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

  – Useful when the actual minimum and maximum values are unknown, or when there are outliers that dominate the min-max normalization

  – A variation replaces standard deviation by mean absolute deviation

$$s_A = \frac{1}{n}(|v_1 - \bar{A}| + |v_2 - \bar{A}| + \cdots + |v_n - \bar{A}|). \quad v_i' = \frac{v_i - \bar{A}}{s_A}.$$

- **Normalization by decimal scaling**

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

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

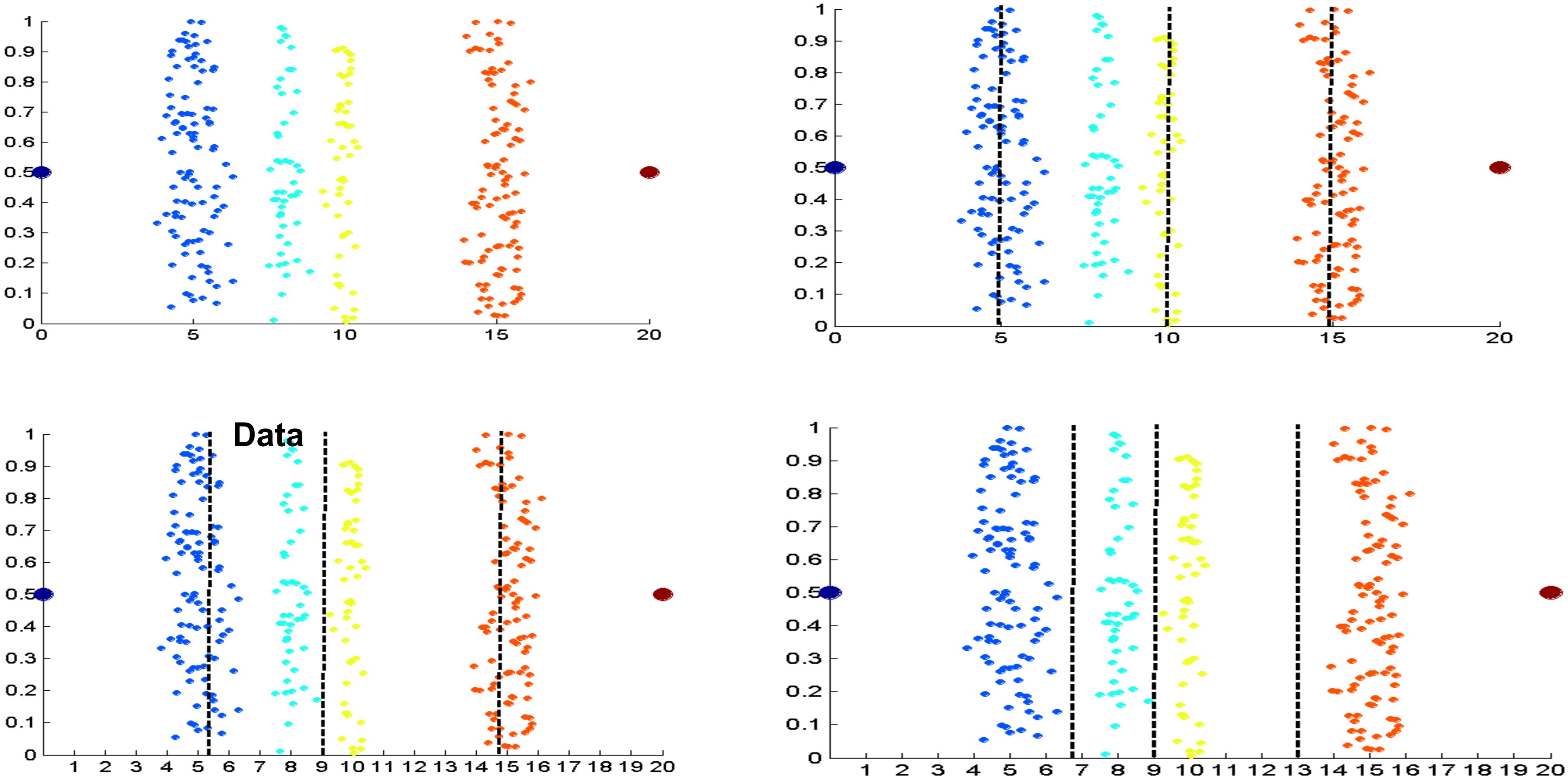
Discretization

- 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 (does not use class information)
  - Histogram analysis
    - Top-down split, unsupervised
  - 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)

Discretization Without Using Class Labels (Binning vs. Clustering)



Equal frequency (binning)

K-means clustering leads to better results

Discretization by Histogram Analysis

- Histogram analysis is an unsupervised discretization technique as it does not use class information
- Equal-width – values are partitioned into equal sized partitions or ranges
- Equal frequency – values are partitioned so each partition contains the same number of data tuples
- Histogram analysis algorithm can be applied recursively to each partition to automatically generate multilevel concept hierarchy
- Histogram can be partitioned based on cluster analysis of the data distribution

Discretization by Classification & Correlation Analysis

- Classification (e.g., decision tree analysis)
  - Supervised: Given class labels, e.g., cancerous vs. benign
  - Using *entropy* to determine split point (discretization point)
  - Top-down, recursive split
- Correlation analysis (e.g., Chi-merge: χ^2-based discretization)
  - Supervised: use class information
  - Bottom-up merge: find the best neighboring intervals (those having similar distributions of classes, i.e., low χ^2 values) to merge
  - Merge performed recursively, until a predefined stopping condition

Concept Hierarchy Generation

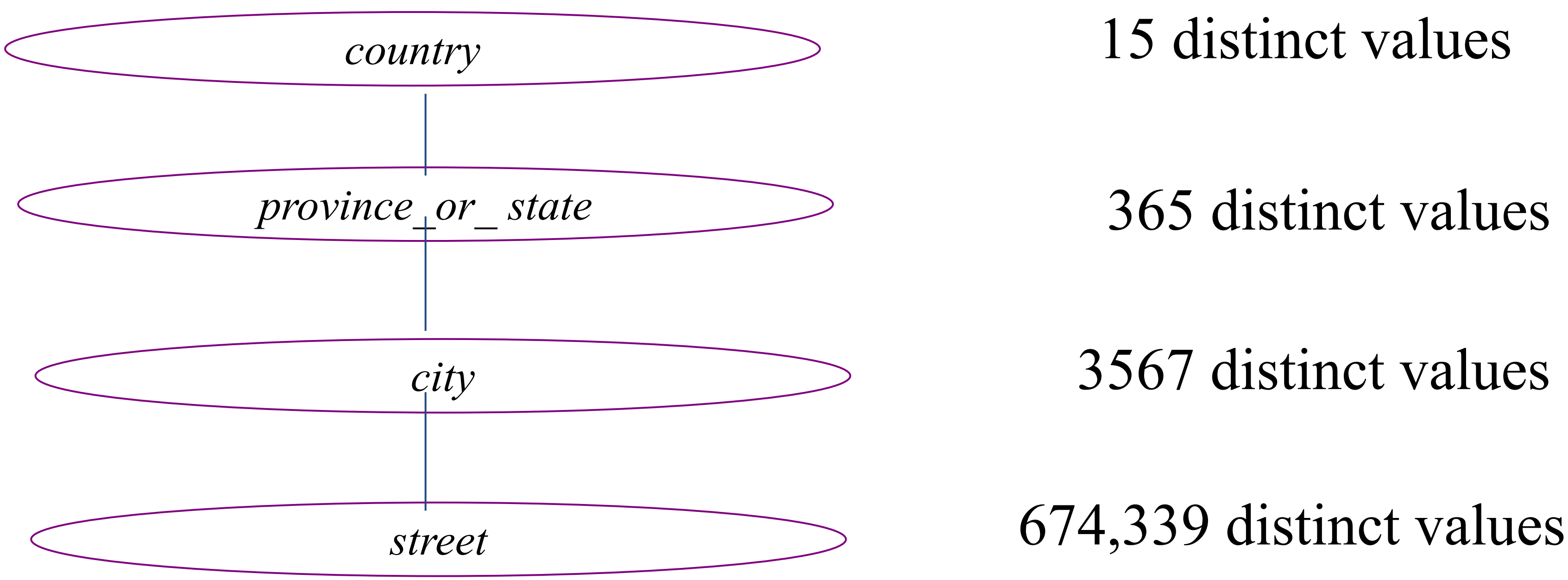
- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate <u>drilling and rolling</u> in data warehouses to view data in multiple granularity
- Concept hierarchy formation: Recursively 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} $\subset$ Illinois
- Specification of only a partial set of attributes
  – E.g., only *street* < *city*, not others
- Specification of a set of attributes, but not of their partial ordering
  – Concept hierarchy based on 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



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