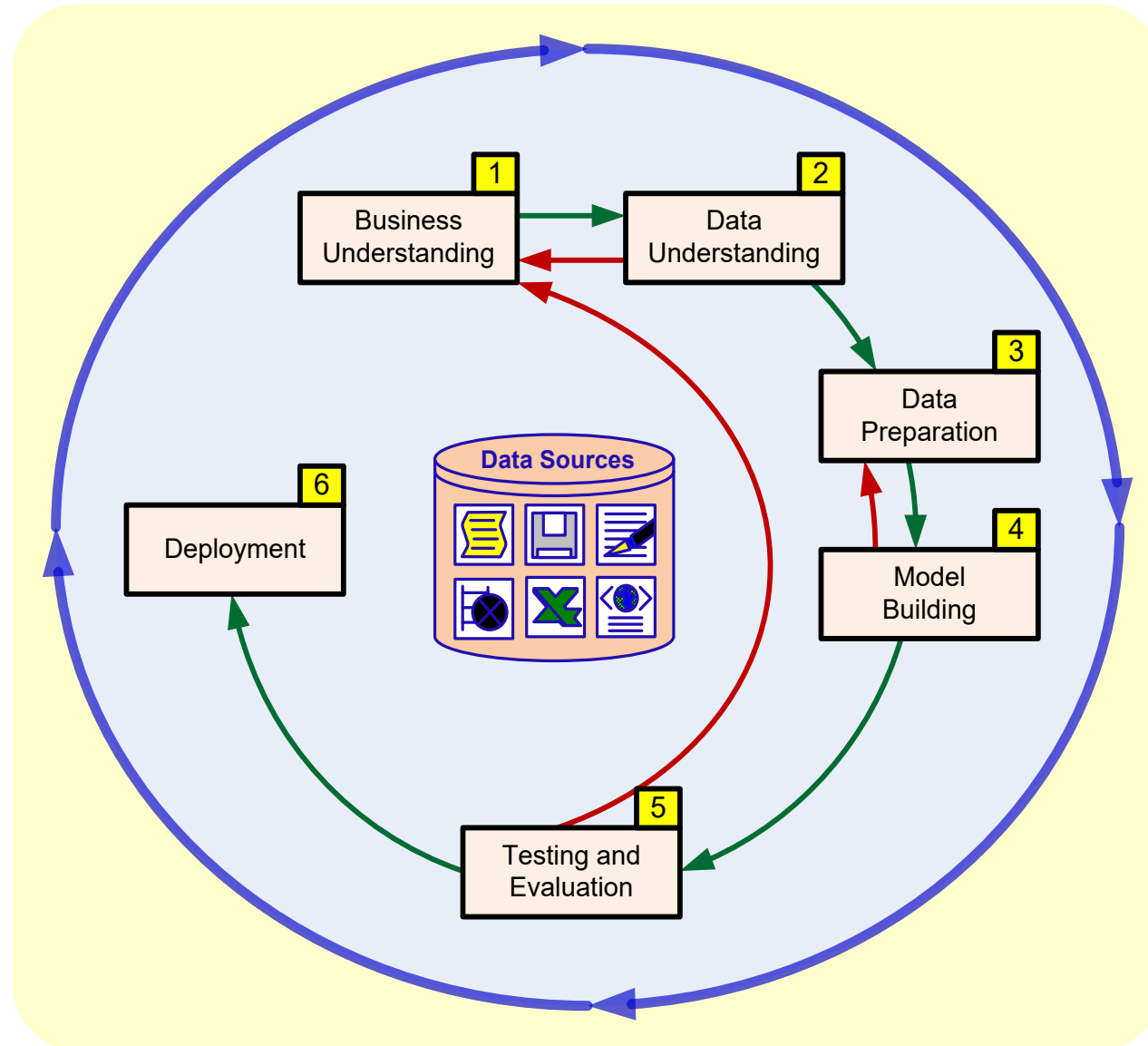


Chapter 2: Basic Techniques for Data Exploration and Preprocessing

Data Mining Process Model



Agenda

- Basic Techniques for Data Exploration
- Basic Techniques for Data Preprocessing

2.1: Basic Techniques for Data Exploration

What is data exploration?

- Key motivations of data exploration include:
 - Helping to select the right tool for preprocessing or analysis
 - Making use of humans' abilities to recognize patterns
 - ✓ People can recognize patterns not captured by data analysis tools
- Major techniques:
 - Summary statistics
 - Visualization

Summary Statistics

- Summary statistics are numbers that summarize properties of the data
 - Summarized properties include frequency, location and spread
 - ✓ Examples: location - mean
 spread - standard deviation
 - Most summary statistics can be calculated in a single pass through the data

Measures of Central Tendency: Mean and Median

- The mean is the most common measure of the location of a set of points with continuous attributes.

$$\text{mean}(x) = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$$

- However, the mean is very sensitive to outliers.
- The median is also commonly used.

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r + 1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Measures of Central Tendency: Frequency and Mode

- The notions of frequency and mode are typically used with categorical data
- The frequency of an attribute value is the percentage of time the value occurs in the data set
- The **mode** of an attribute is the most frequent attribute value(**s**)

Measures of Spread: Range and Variance

- Range is the difference between the max and min
- The variance or standard deviation is the most common measure of the spread of a set of points.

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

- However, this is also sensitive to outliers, so that other measures are often used.

$$\text{AAD}(x) = \frac{1}{m} \sum_{i=1}^m |x_i - \bar{x}| \quad \text{Average Absolute Deviation}$$

$$\text{MAD}(x) = \text{median} \left(\{|x_1 - \bar{x}|, \dots, |x_m - \bar{x}|\} \right) \quad \text{Median Absolute Deviation}$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Iris Sample Data Set

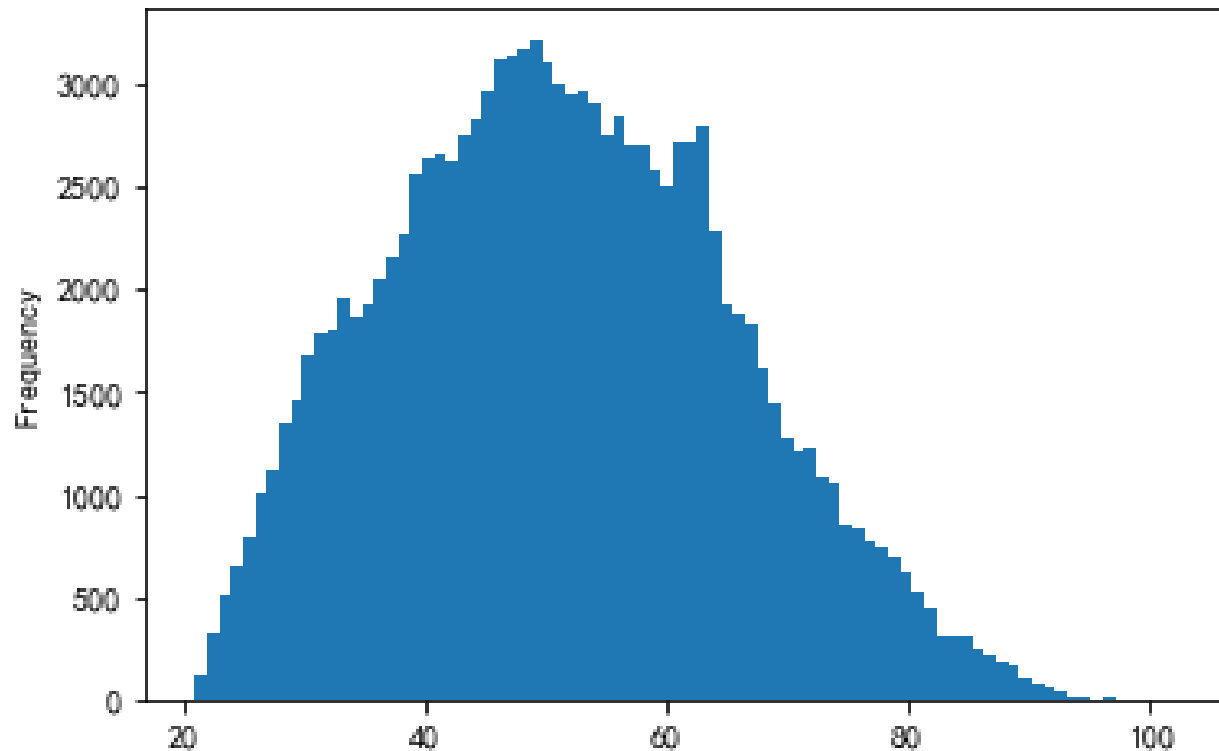
- Many data exploration techniques are illustrated with the Iris Plant data set.
 - Three flower types (classes):
 - ✓ Setosa (**Se**)
 - ✓ Virginica (**Vi**)
 - ✓ Versicolour (**Ve**)
 - Four attributes
 - ✓ Sepal width and length
 - ✓ Petal width and length



from the UCI Machine Learning Repository
<http://www.ics.uci.edu/~mlearn/MLRepository.html>

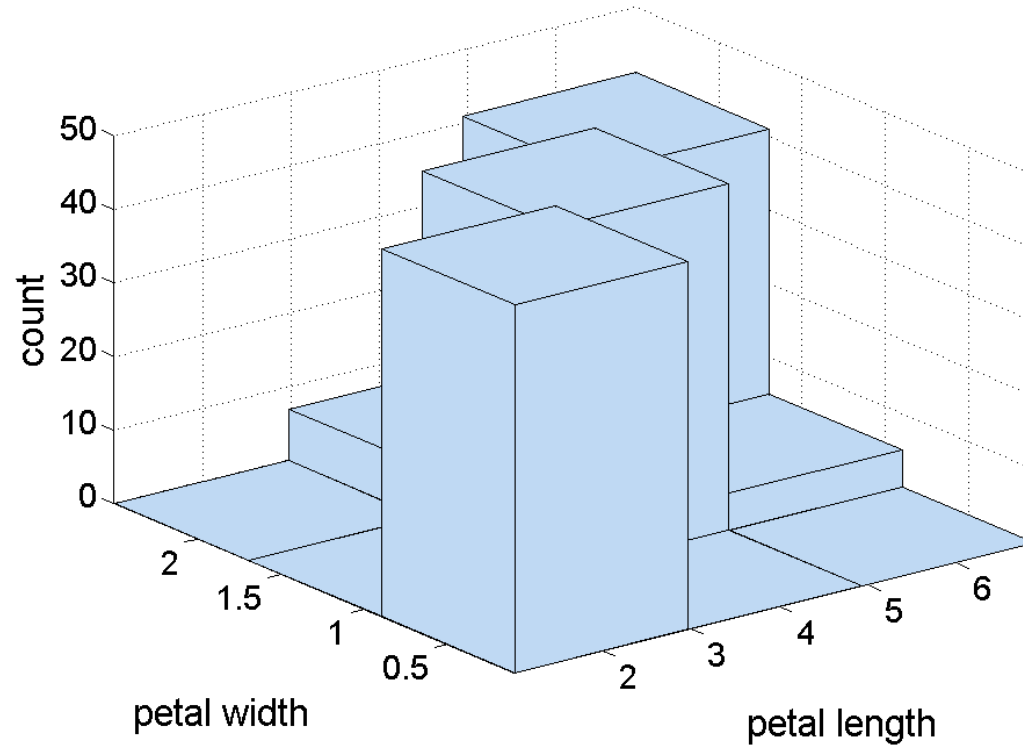
Visualization Techniques: Histograms

- Usually shows the distribution of values of a single variable.
- Divide the values into bins and show a bar plot of the number of objects in each bin.



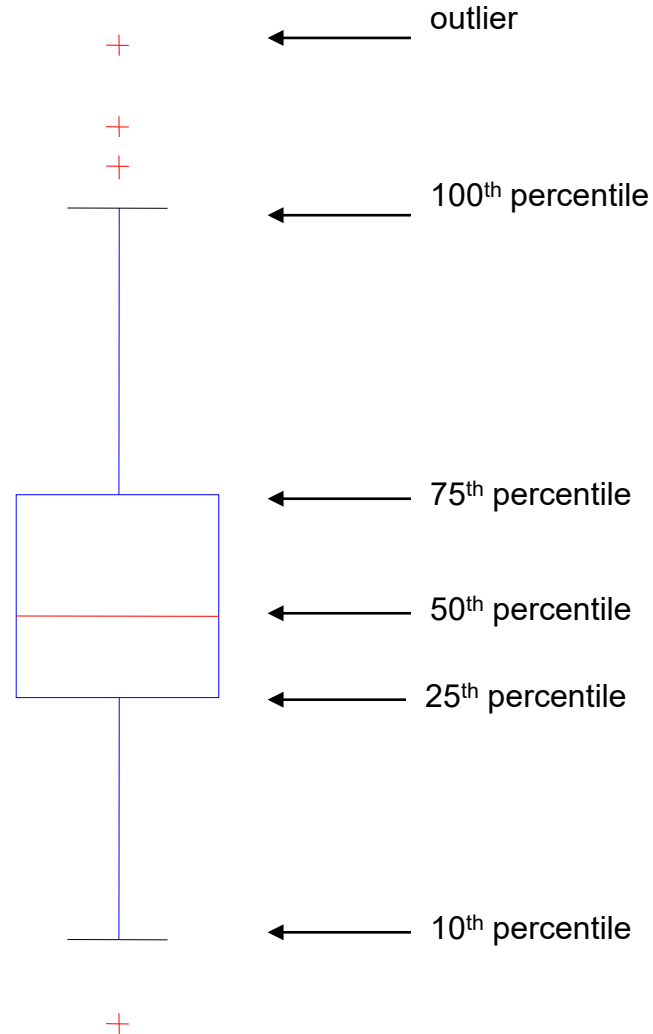
Two-Dimensional Histograms

- Show the joint distribution of the values of two attributes



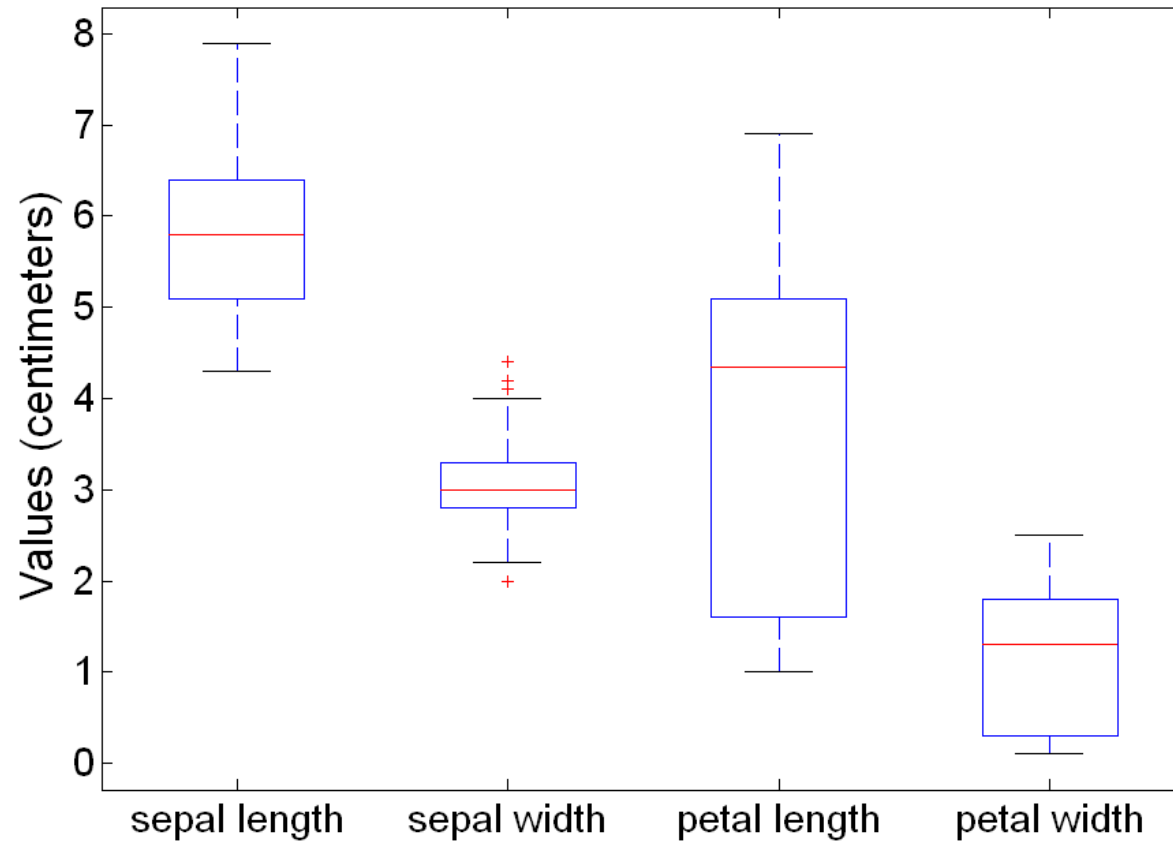
Visualization Techniques: Box Plots

- Another way of displaying the distribution of data

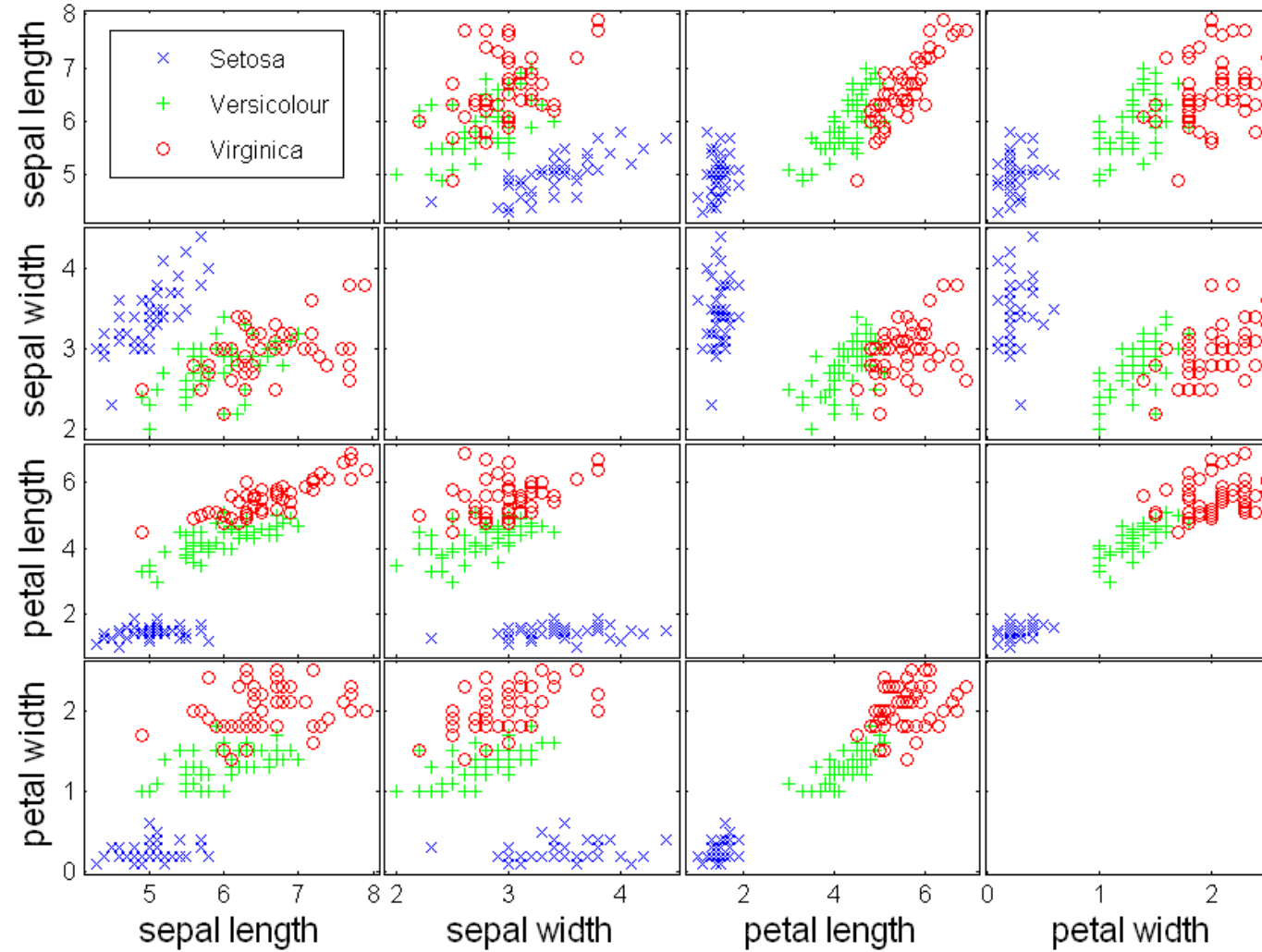


Example of Box Plots

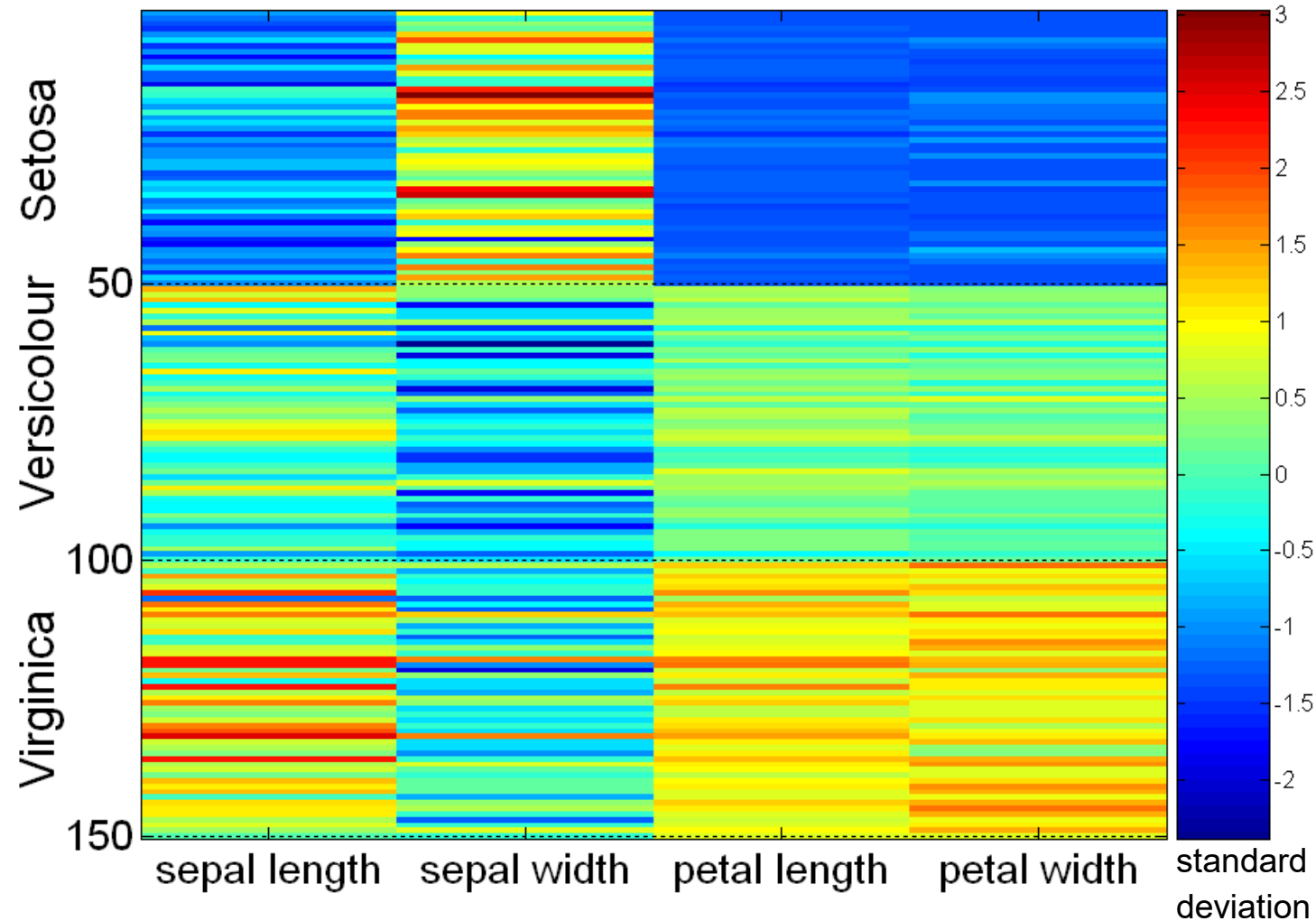
- Box plots can be used to compare attributes



Scatter Plot Array of Iris Attributes

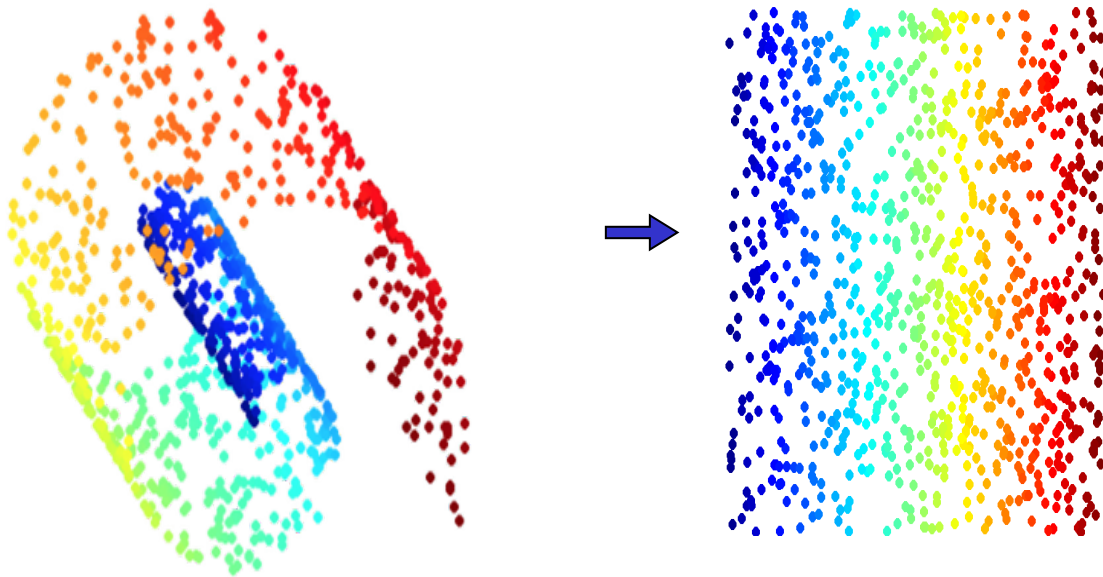


Visualization of the Iris Data Matrix



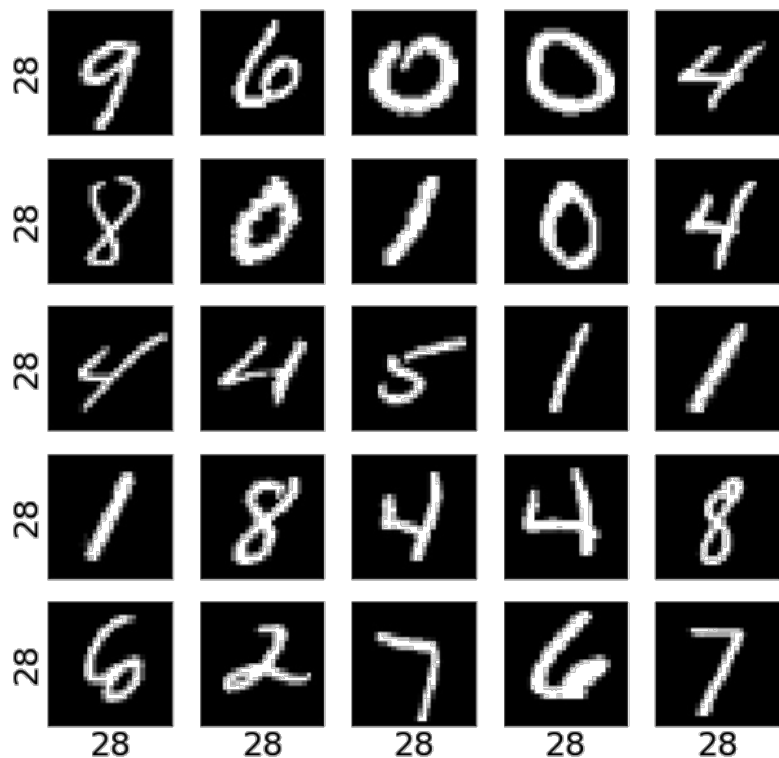
How to Visualize High-dimensional Data

- **Basic idea: low-dimensional embedding**
 - **Should keep similar data distribution in new low-dimensional subspace**

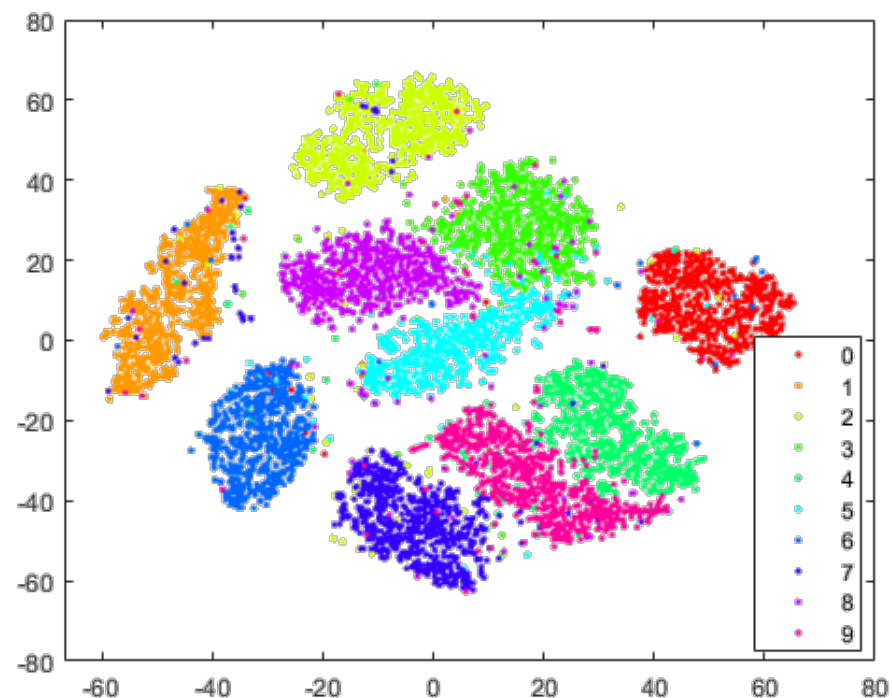
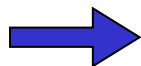


$$x_i \rightarrow y_i$$

$$x_i \in R^n, \quad y_i \in R^k, \quad k \ll n$$



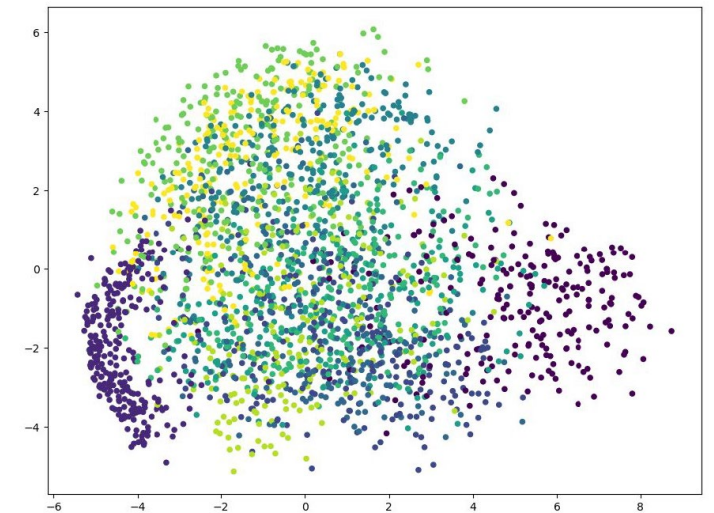
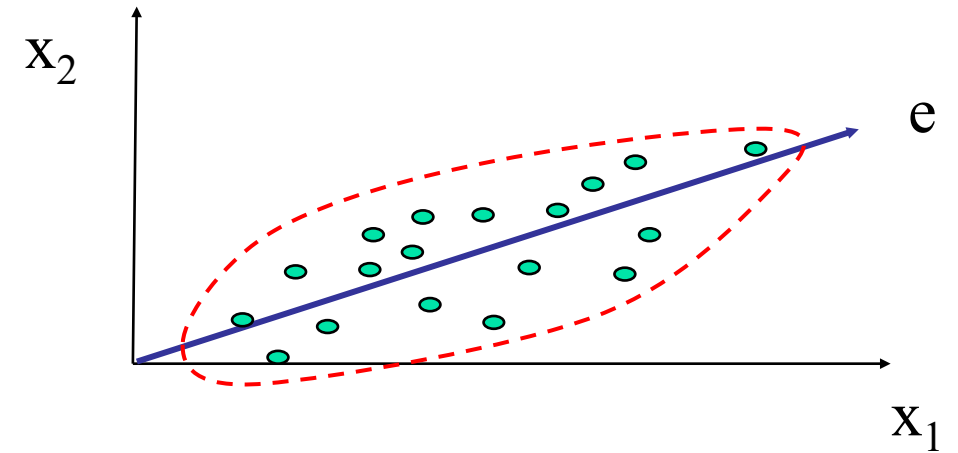
high dimensions
(28 x28)



two dimensions

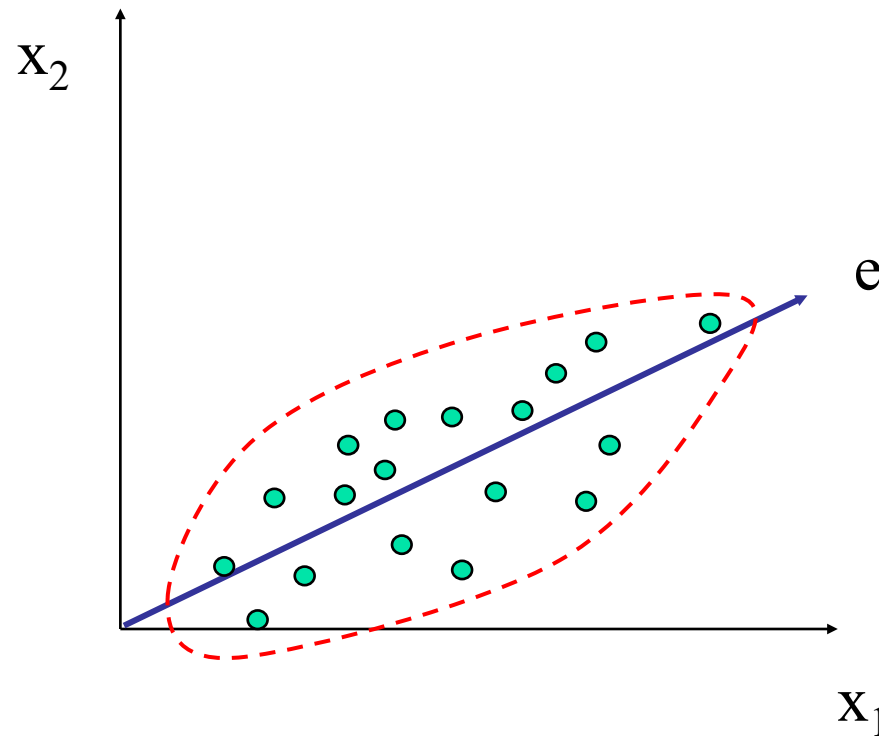
Classical Methods for Low-dimensional Embedding

- **PCA** (Principal Component Analysis)
 - Try to keep global data distribution
- **Stochastic Neighbor Embedding (SNE)**
 - Try to keep local data distribution
 - **t-SNE**
- **Other methods**
 - Autoencoder, IsoMap, LLE, MDS



Principal Component Analysis (PCA)

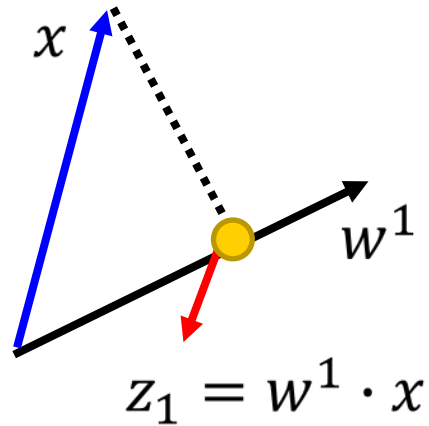
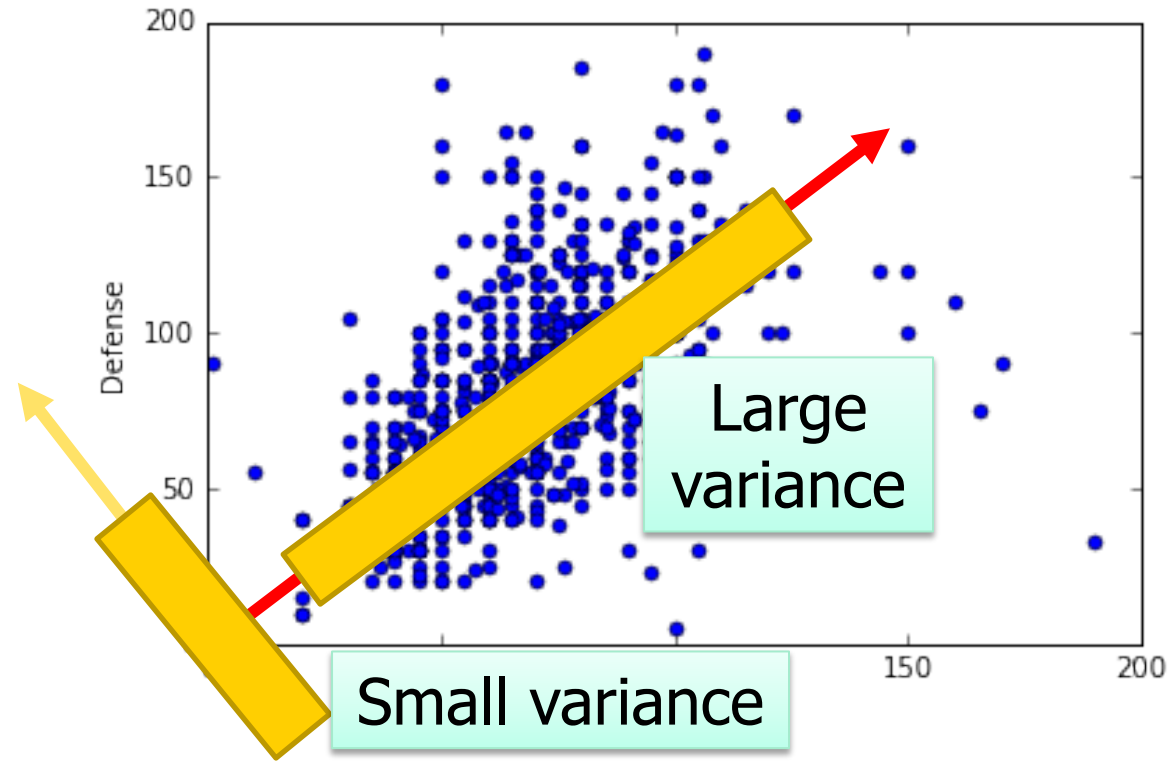
- Find a projection that captures the largest amount of variation in data.
- The original data are projected onto a much smaller space.
- Find the eigenvectors of the covariance matrix, and these eigenvectors define the new space.



$$z = Wx$$

Reduce to 1-D:

$$z_1 = w^1 \cdot x$$



Project all the data points x onto w^1 , and obtain a set of z_1

We want the variance of z_1 as large as possible

$$Var(z_1) = \frac{1}{N} \sum_{z_1} (z_1 - \bar{z}_1)^2 \quad \|w^1\|_2 = 1$$

$$z = Wx$$

Reduce to 1-D:

$$z_1 = w^1 \cdot x$$

$$z_2 = w^2 \cdot x$$

$$W = \begin{bmatrix} (w^1)^T \\ (w^2)^T \\ \vdots \end{bmatrix}$$

Orthogonal
matrix

Project all the data points x onto w^1 ,
and obtain a set of z_1

We want the variance of z_1 as large as
possible

$$Var(z_1) = \frac{1}{N} \sum_{z_1} (z_1 - \bar{z}_1)^2 \quad \|w^1\|_2 = 1$$

We want the variance of z_2 as large as
possible

$$Var(z_2) = \frac{1}{N} \sum_{z_2} (z_2 - \bar{z}_2)^2 \quad \|w^2\|_2 = 1$$

$$w^1 \cdot w^2 = 0$$

$$z_1 = w^1 \cdot x$$

$$\bar{z}_1 = \frac{1}{N} \sum z_1 = \frac{1}{N} \sum w^1 \cdot x = w^1 \cdot \frac{1}{N} \sum x = w^1 \cdot \bar{x}$$

$$Var(z_1) = \frac{1}{N} \sum_{z_1} (z_1 - \bar{z}_1)^2$$

$$= \frac{1}{N} \sum_x (w^1 \cdot x - w^1 \cdot \bar{x})^2$$

$$= \frac{1}{N} \sum (w^1 \cdot (x - \bar{x}))^2$$

$$= \frac{1}{N} \sum (w^1)^T (x - \bar{x})(x - \bar{x})^T w^1$$

$$= (w^1)^T \left[\frac{1}{N} \sum (x - \bar{x})(x - \bar{x})^T \right] w^1$$

$$= (w^1)^T Cov(x) w^1$$

$$S = Cov(x)$$

$$(a \cdot b)^2 = (a^T b)^2 = a^T b a^T b$$

$$= a^T b (a^T b)^T = a^T b b^T a$$

Find w^1 maximizing

$$(w^1)^T S w^1$$

$$\|w^1\|_2 = (w^1)^T w^1 = 1$$

Find w^1 maximizing $(w^1)^T S w^1$ $(w^1)^T w^1 = 1$

$S = \text{Cov}(x)$ Symmetric Positive-semidefinite
(non-negative eigenvalues)

Using Lagrange multiplier

$$g(w^1) = (w^1)^T S w^1 - \alpha((w^1)^T w^1 - 1)$$

$$\left. \begin{array}{l} \partial g(w^1) / \partial w_1^1 = 0 \\ \partial g(w^1) / \partial w_2^1 = 0 \\ \vdots \end{array} \right\} \begin{array}{l} S w^1 - \alpha w^1 = 0 \\ S w^1 = \alpha w^1 \quad w^1 : \text{eigenvector} \\ (w^1)^T S w^1 = \alpha (w^1)^T w^1 \\ = \alpha \quad \text{Choose the maximum one} \end{array}$$

w^1 is the eigenvector of the covariance matrix S
Corresponding to the largest eigenvalue λ_1

Find w^2 maximizing $(w^2)^T S w^2$ $(w^2)^T w^2 = 1$ $(w^2)^T w^1 = 0$

$$g(w^2) = (w^2)^T S w^2 - \alpha((w^2)^T w^2 - 1) - \beta((w^2)^T w^1 - 0)$$

$$\left. \begin{array}{l} \partial g(w^2)/\partial w_1^2 = 0 \\ \partial g(w^2)/\partial w_2^2 = 0 \\ \vdots \end{array} \right\} \begin{array}{l} S w^2 - \alpha w^2 - \beta w^1 = 0 \\ \underline{0} - \alpha \underline{0} - \beta \underline{1} = 0 \\ = ((w^1)^T S w^2)^T = (w^2)^T S^T w^1 \\ = (w^2)^T S w^1 = \lambda_1 (w^2)^T w^1 = 0 \end{array}$$

$$S w^1 = \lambda_1 w^1$$

$$\beta = 0: \quad S w^2 - \alpha w^2 = 0 \quad S w^2 = \alpha w^2$$

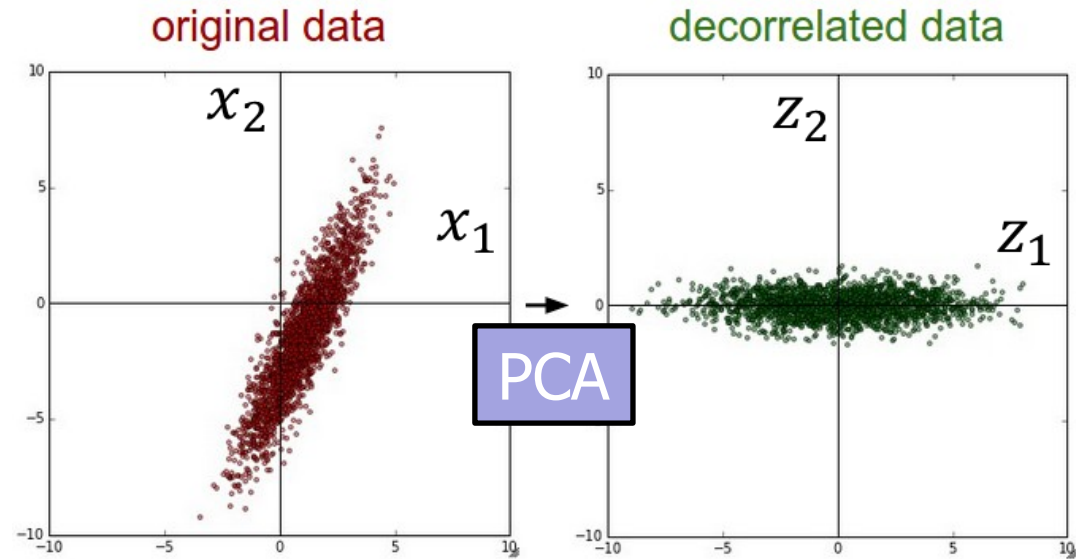
w^2 is the eigenvector of the covariance matrix S
Corresponding to the 2nd largest eigenvalue λ_2

Decorrelation Property

$$z = Wx$$

$$\text{Cov}(z) = D$$

Diagonal matrix



$$\text{Cov}(z) = \frac{1}{N} \sum (z - \bar{z})(z - \bar{z})^T = W S W^T \quad S = \text{Cov}(x)$$

$$= W S [w^1 \quad \dots \quad w^K] = W [S w^1 \quad \dots \quad S w^K]$$

$$= W [\lambda_1 w^1 \quad \dots \quad \lambda_K w^K] = [\lambda_1 W w^1 \quad \dots \quad \lambda_K W w^K]$$

$$= [\lambda_1 e_1 \quad \dots \quad \lambda_K e_K] = D$$



Learning Steps of PCA (1)

- **Data preprocessing**

Input: training set X - $[x^{(1)}, x^{(2)}, \dots, x^{(m)}]$

- Normalize input data: such as Z-score normalization

- Preprocessing(feature scaling/mean normalization)

$$\mu_j = \frac{1}{m} \sum_{i=1}^m x_j^{(i)}$$

- Replace each $x_j^{(i)}$ with $x_j^{(i)} - \mu_j$



Learning Steps of PCA (2)

- **Learning Principal Components**

Input: the transformed training set X' - $[x^{(1)}, x^{(2)}, \dots, x^{(m)}]$

- Compute "covariance matrix" Σ

$$\Sigma = \frac{1}{m} X' (X')^T$$

- Compute "eigenvectors" of matrix Σ

- Select k "eigenvectors" as Principal Components : $W = \begin{bmatrix} (w^1)^T \\ \vdots \\ (w^k)^T \end{bmatrix}$

● Compute "eigenvectors" with SVD

Input: the transformed training set $X' = [x^{(1)}, x^{(2)}, \dots, x^{(m)}]$


➤ Compute "eigenvectors" of matrix X' :

$$[U, S, V] = SVD(X');$$

$$U = \begin{bmatrix} | & | & & | \\ u^{(1)} & u^{(2)} & \dots & u^{(n)} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times n}$$

$\underbrace{\hspace{10em}}_k$

➤ Let: $W = (U_{reduce})^T$


$$U_{reduce} = \begin{bmatrix} | & | & & | \\ u^{(1)} & u^{(2)} & \dots & u^{(k)} \\ | & | & & | \end{bmatrix}$$



Projection Phase

- Simply project a test record $x^{(i)}$ of n-dimensional to k-dimensional by:

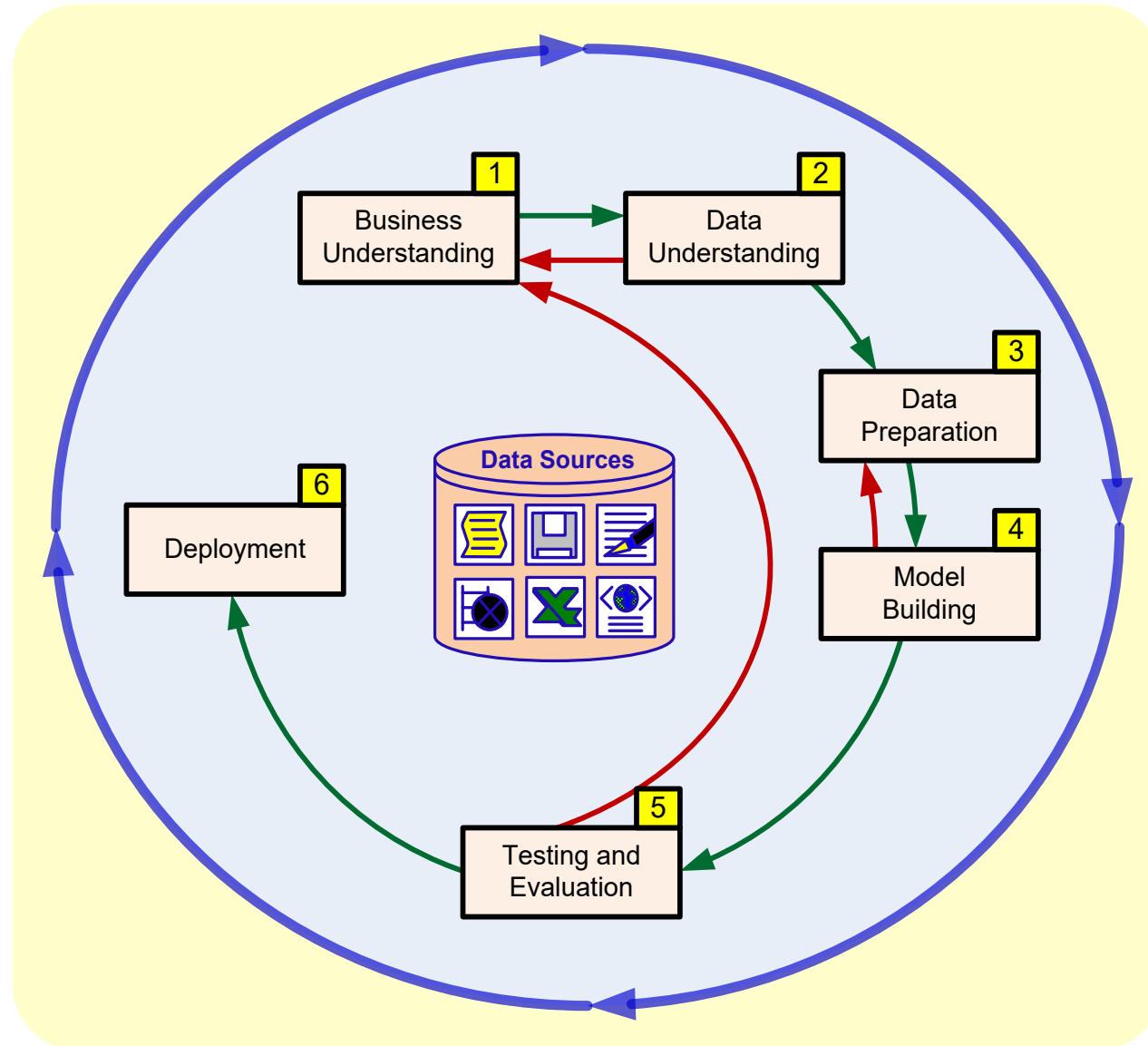
$$\mathbf{z}^{(i)} = \mathbf{W}x^{(i)}$$

$$\mathbf{x}^{(i)} \in \mathbb{R}^n \rightarrow \mathbf{z}^{(i)} \in \mathbb{R}^k$$

How to determine k ?

2.2: Basic Techniques for Data Preprocessing

Data Mining Process Model



Why Data Preprocessing?

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

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 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 warehouse needs consistent integration of quality data

Major Tasks in Data Preprocessing

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

Data Cleaning

- 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 (assuming the tasks in classification—not effective when the percentage 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 due to
 - faulty data collection instruments
 - data entry problems
 - technology limitation, inconsistency in naming convention
 - data transmission problems
- 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

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 equal-frequency (equi-depth) bins:
 - Bin 1: 4, 8, 9, 15
 - Bin 2: 21, 21, 24, 25
 - Bin 3: 26, 28, 29, 34
- * Smoothing by bin means:
 - Bin 1: 9, 9, 9, 9
 - Bin 2: 23, 23, 23, 23
 - Bin 3: 29, 29, 29, 29
- * Smoothing by bin boundaries:
 - Bin 1: 4, 4, 4, 15
 - Bin 2: 21, 21, 25, 25
 - Bin 3: 26, 26, 26, 34

Data Integration

- Data integration:
 - Combines data from multiple sources into a coherent store
- Schema integration: e.g., $A.cust-id \equiv B.cust-\#$
 - 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 (Categorical Data)

- χ^2 (chi-square) test

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

- The larger the χ^2 value, the more likely the variables are related
- The cells that contribute the most to the χ^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

- χ^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 (Numerical Data)

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

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

where n is the number of tuples, \bar{A} and \bar{B} are the respective means of A and B , σ_A and σ_B are the respective standard deviation of A and B , and $\sum(AB)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A 's values increase as B 's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{A,B} < 0$: negatively correlated

Data Transformation: Normalization

- Min-max normalization: to $[\text{new_min}_A, \text{new_max}_A]$

$$v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- 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 $\mu = 54,000$, $\sigma = 16,000$. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

- Normalization by decimal scaling

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

Data Reduction Strategies

- Why data reduction?
 - A database/data 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
 - 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
- Data reduction strategies
 - Data cube aggregation:
 - Dimensionality reduction
 - Data (record) reduction
 - Discretization and concept hierarchy generation
 - Other...

Data Cube Aggregation

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

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

Purpose:

- Data reduction: reduce the number of attributes or objects
- Change of scale: cities aggregated into regions, states, countries, etc
- More “stable” data: aggregated data tends to have less variability

Attribute Subset Selection

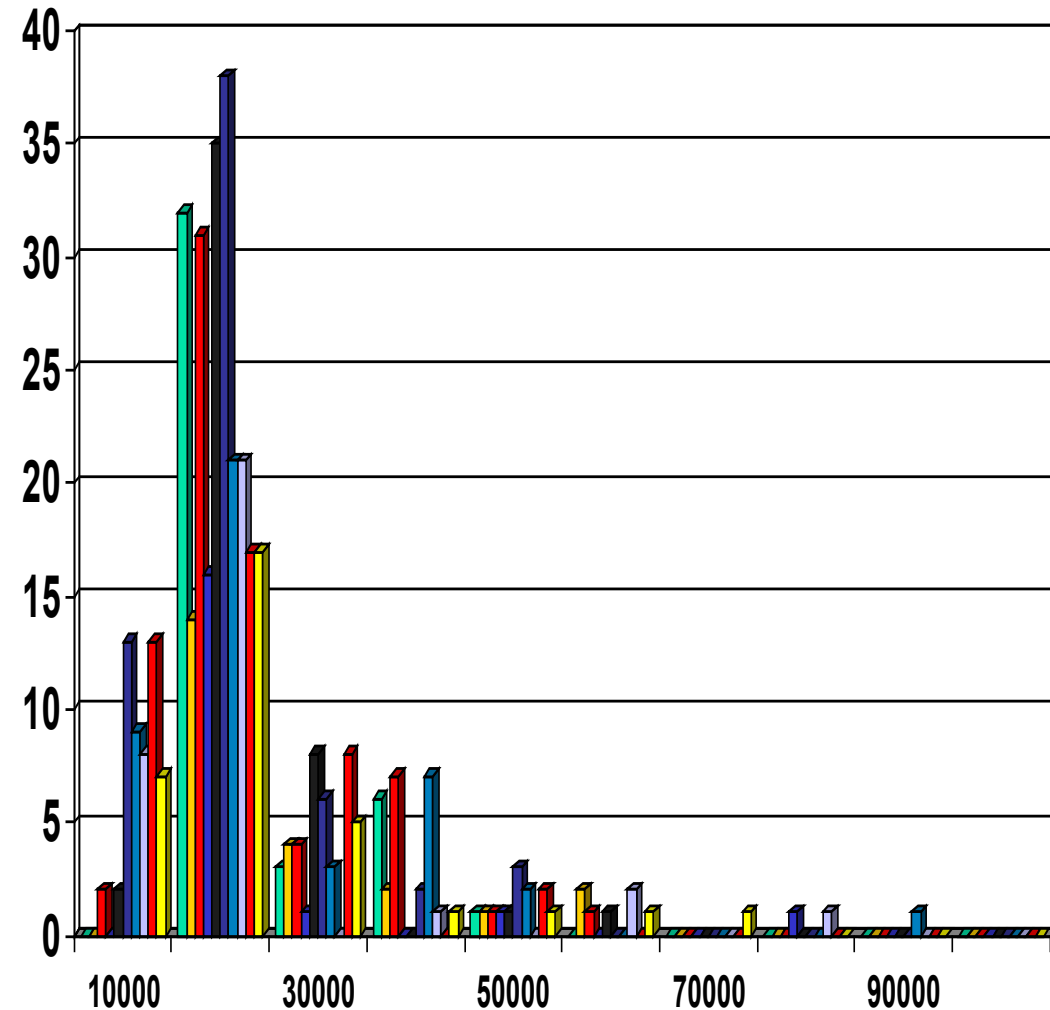
- Feature selection (i.e., attribute subset selection):
 - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
 - reduce # of patterns in the patterns, easier to understand
- Heuristic methods (due to exponential # of choices):
 - Step-wise forward selection
 - Step-wise backward elimination
 - Combining forward selection and backward elimination
 - Decision-tree induction

Numerosity Reduction

- Reduce data volume by choosing alternative, smaller forms of data representation
- Parametric methods
 - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
 - Example: Log-linear models—obtain value at a point in m -D space as the product on appropriate marginal subspaces
- Non-parametric methods
 - Do not assume models
 - Major families: histograms, clustering, sampling

Data Reduction Method: 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 $\beta-1$ largest differences



Data Reduction Method: 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

Data Reduction Method: 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
- Choose a **representative** subset of the data
 - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
 - Stratified sampling:
 - ✓ Approximate the percentage of each class (or subpopulation of interest) in the overall database
 - ✓ Used in conjunction with skewed data
- Note: Sampling may not reduce database I/Os (page at a time)

Sampling

- Sampling is the main technique employed for data selection
- It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming

Types of Sampling

- Sampling without replacement

- As each item is selected, it is removed from the population

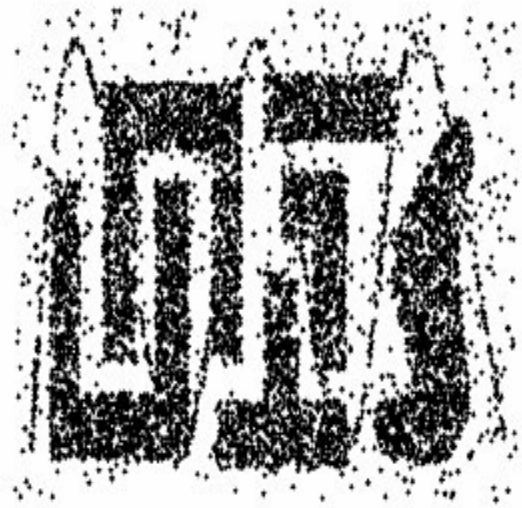
- Sampling with replacement

- Objects are not removed from the population as they are selected for the sample.
- In sampling with replacement, the same object can
- be picked up more than once

- Stratified sampling

- Split the data into several partitions
- Then draw random samples from each partition

Sample Size



8000 points

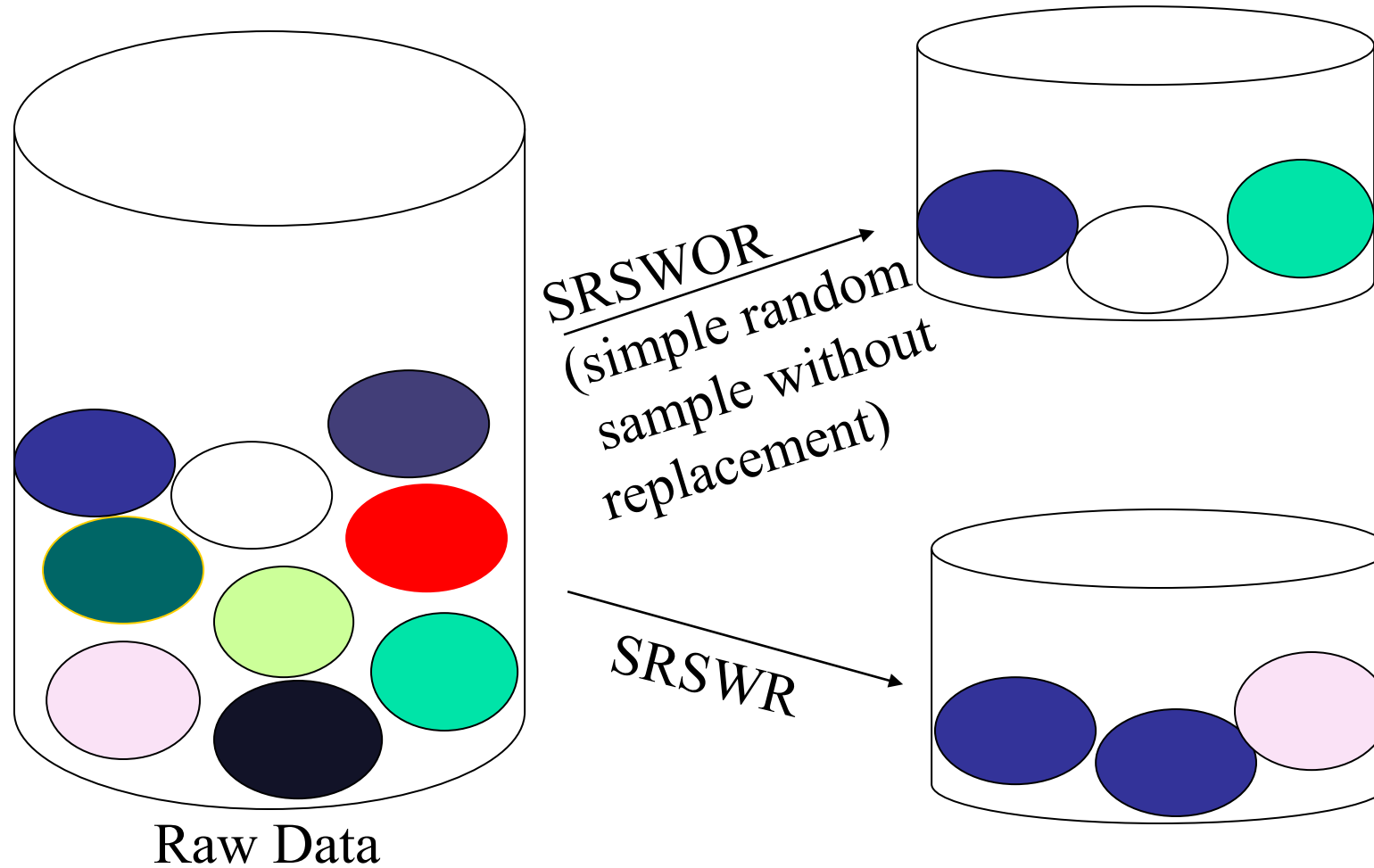


2000 Points



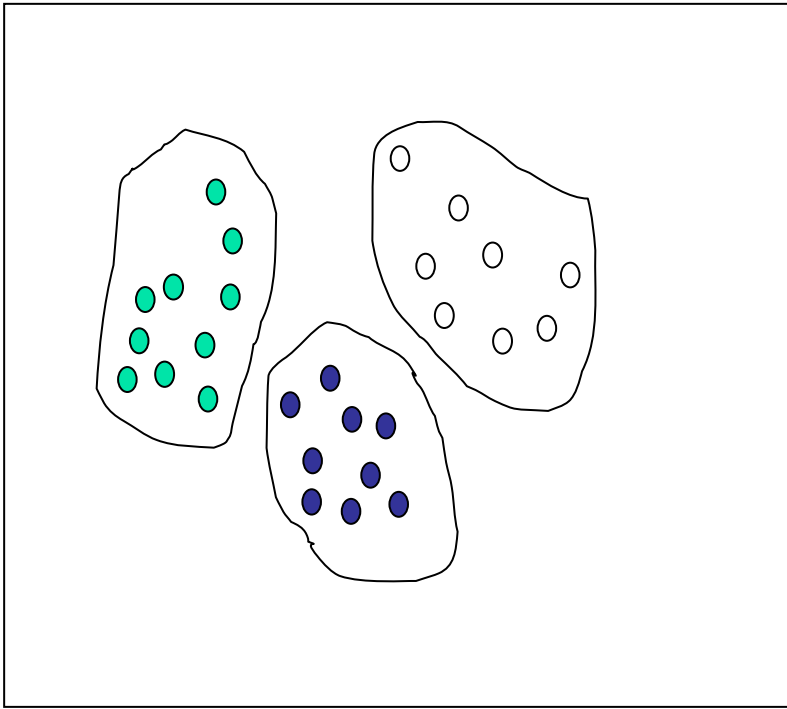
500 Points

Sampling: with or without Replacement

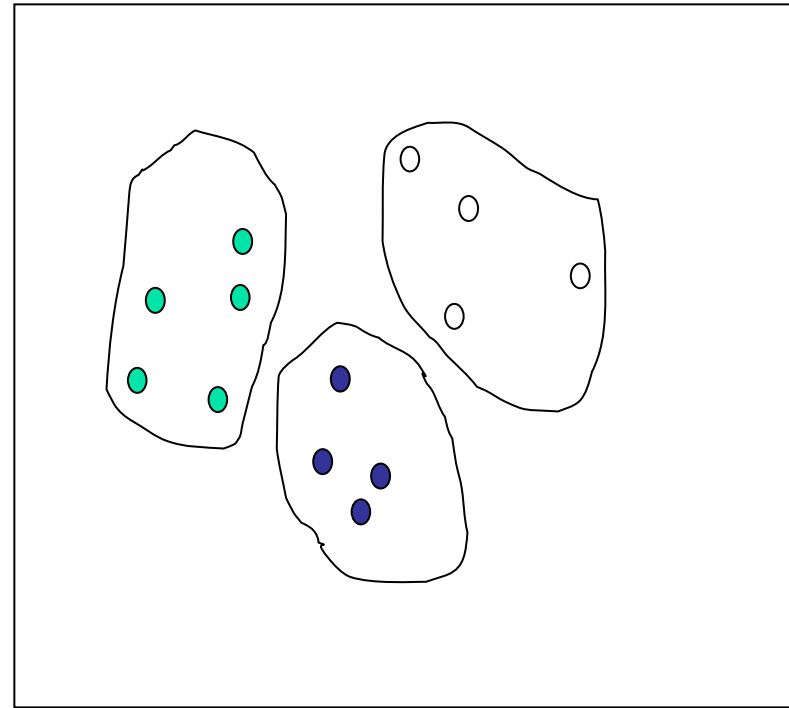


Sampling: Cluster or Stratified Sampling

Raw Data



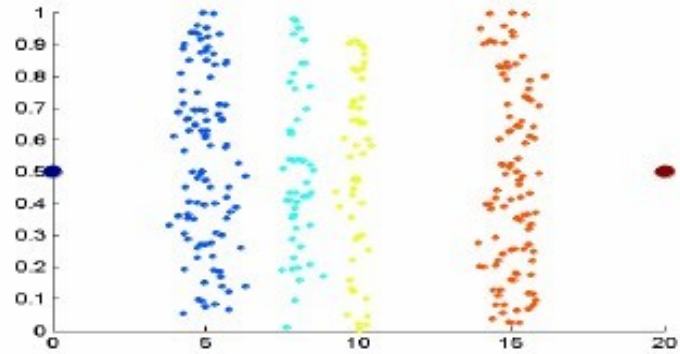
Cluster/Stratified Sample



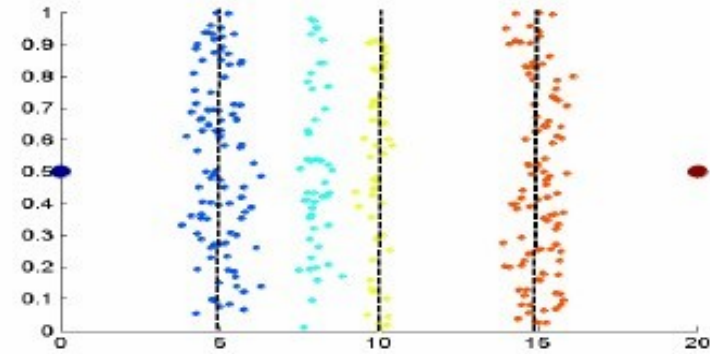
Discretization

- Discretization:
 - Divide the range of a continuous attribute into intervals
 - Some classification algorithms only accept categorical attributes.
 - Reduce data size by discretization
 - Prepare for further analysis

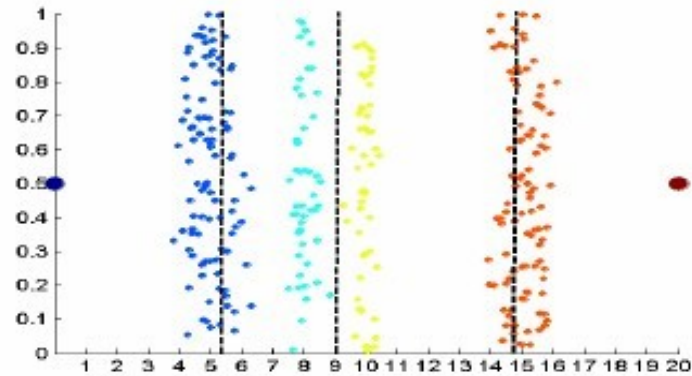
Discretization Without Using Class Labels



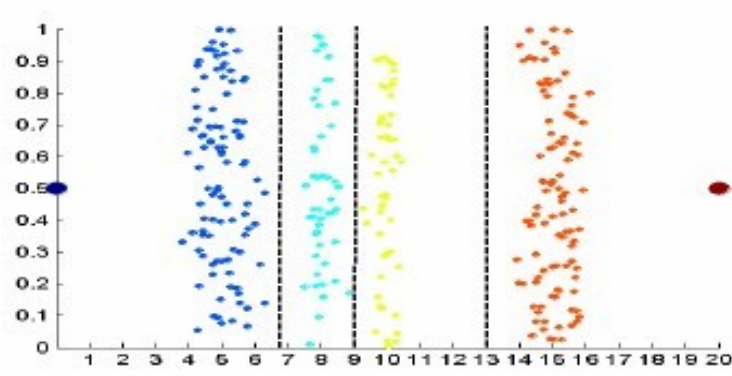
Data



Equal interval width



Equal frequency



K-means

Entropy-Based Discretization

- Given a set of samples S , if S is partitioned into two intervals S_1 and S_2 using boundary T , the information gain after partitioning is

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

- Entropy is calculated based on class distribution of the samples in the set. Given m classes, the entropy of S_1 is

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

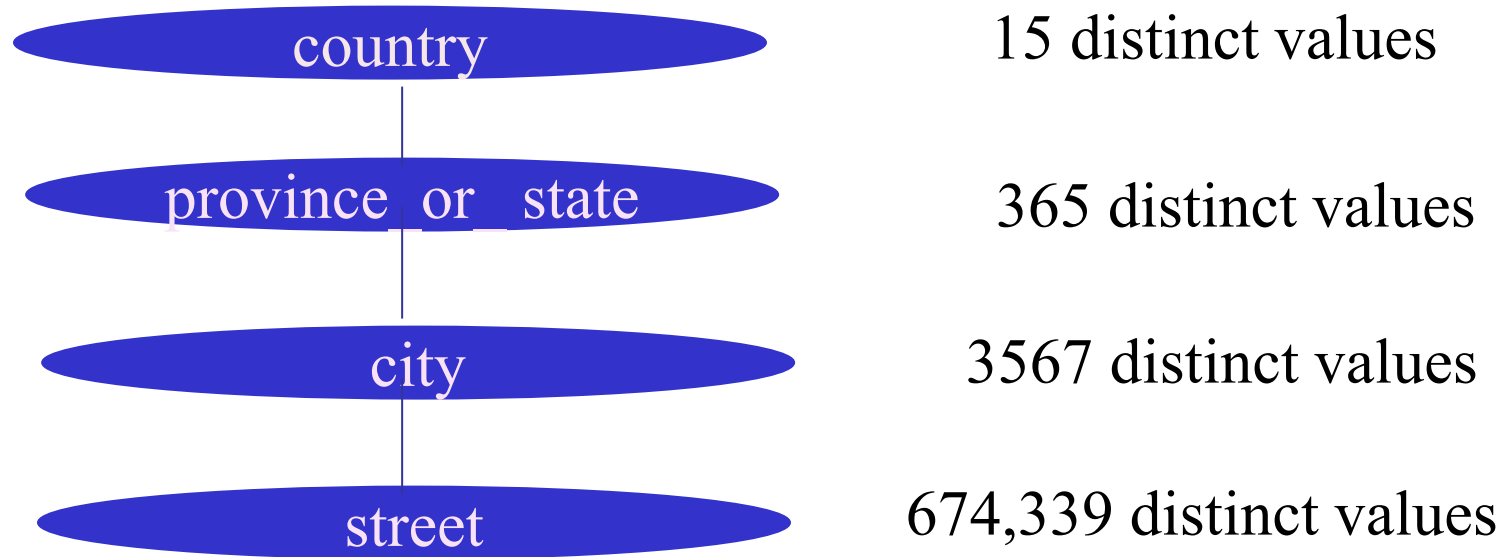
where p_i is the probability of class i in S_1

- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization
- The process is recursively applied to partitions obtained until some stopping criterion is met
- Such a boundary may reduce data size and improve classification accuracy

Interval Merge by χ^2 Analysis

- Merging-based (bottom-up) vs. splitting-based methods
- Merge: Find the best neighboring intervals and merge them to form larger intervals recursively
- ChiMerge [Kerber AAAI 1992, See also Liu et al. DMKD 2002]
 - Initially, each distinct value of a numerical attr. A is considered to be one interval
 - χ^2 tests are performed for every pair of adjacent intervals
 - Adjacent intervals with the least χ^2 values are merged together, since low χ^2 values for a pair indicate similar class distributions
 - This merge process proceeds recursively until a predefined stopping criterion is met (such as significance level, max-interval, max inconsistency, etc.)

Automatic Concept Hierarchy Generation



Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Descriptive data summarization is needed for quality data preprocessing
- Data preprocessing includes
 - Data cleaning and data integration
 - Data reduction and feature selection
 - Discretization
- Data preprocessing still an active area of research: especially, 'Representation Learning'

Acknowledgements

- Some text, figures and formulations are from WWW. Especially, some slides of PCA section are from Dr. Hongyi Lee (NTU). Thanks for their sharing. If you have copyright claim please contact with us at yym@hit.edu.cn.
- This lecture is distributed for nonprofit purpose.