



DATA ANALYTICS

Unit 1: Data Integration, Cleaning and Reduction

Mamatha.H.R

Department of Computer Science and
Engineering

DATA ANALYTICS

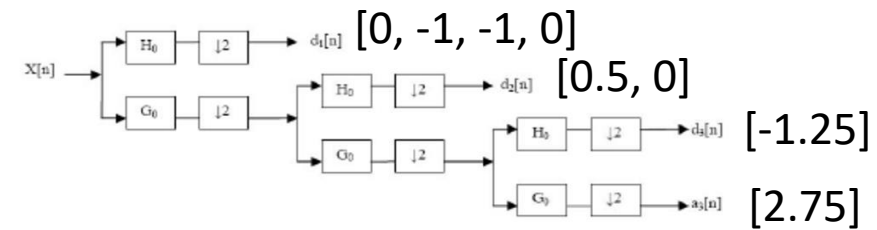
Unit 1:Data Reduction (contd.)

Mamatha H R, Gowri Srinivasa

Department of Computer Science and Engineering

- Wavelets: A math tool for space-efficient hierarchical decomposition of functions
- $S = [2, 2, 0, 2, 3, 5, 4, 4]$ can be transformed to
 $S_{\wedge} = [2^3/4, -1^{1/4}, 1/2, 0, 0, -1, -1, 0]$
- Compression: many small detail coefficients can be replaced by 0's, and only the significant coefficients are retained

Resolution	Averages	Detail Coefficients
8	$[2, 2, 0, 2, 3, 5, 4, 4]$	
4	$[2, 1, 4, 4]$	$[0, -1, -1, 0]$
2	$[1\frac{1}{2}, 4]$	$[\frac{1}{2}, 0]$
1	$[2\frac{3}{4}]$	$[-1\frac{1}{4}]$



$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$H_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

$$H_8 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Coefficient “Supports”

2.75 +

-1.25 + -

0.5 + -

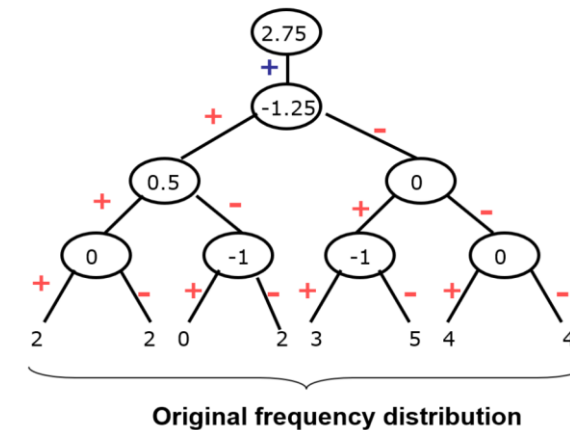
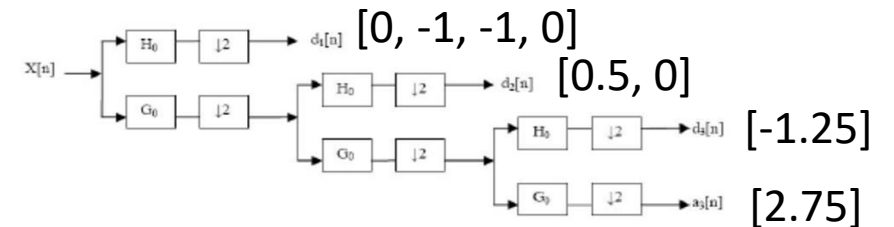
0 + -

0 + -

-1 + -

-1 + -

0 + - -



Hierarchical decomposition structure (a.k.a. “error tree”)

DATA ANALYTICS

Zeroing out detailed coefficients



JPEG



JPEG 2000



JPEG



JPEG 2000

Why Wavelet Transform?

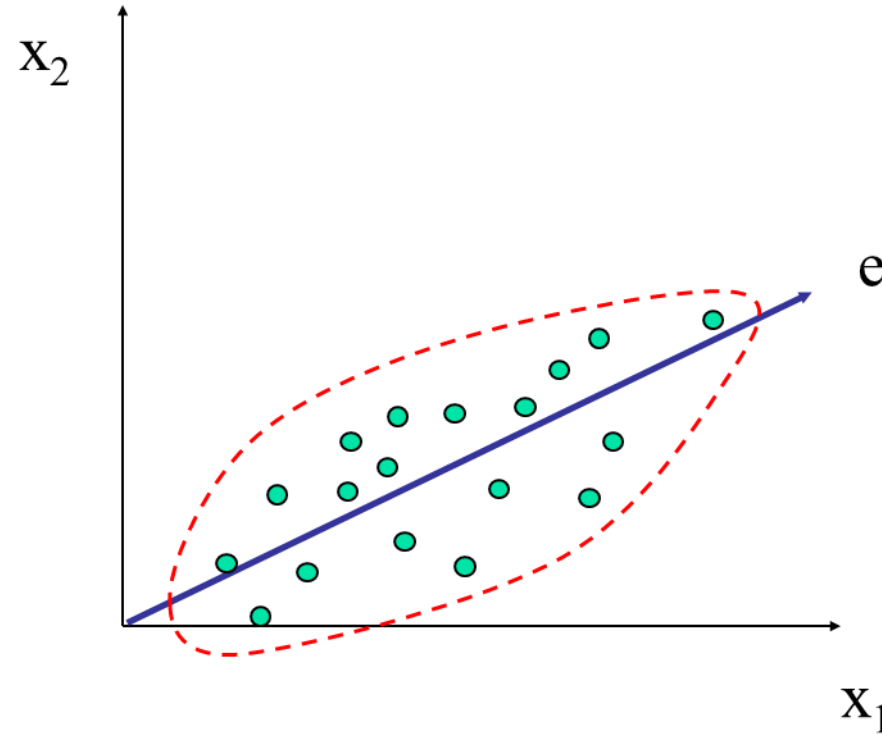
- Use hat-shape filters
 - Emphasize region where points cluster
 - Suppress weaker information in their boundaries
- Effective removal of outliers
 - Insensitive to noise, insensitive to input order
- Multi-resolution
 - Detect arbitrary shaped clusters at different scales
- Efficient
 - Complexity $O(N)$
- Only applicable to low dimensional data

Principal component analysis

- Simplify data
- Understand relationship between variables
- Get an insight to patterns

Principal Component Analysis (PCA)

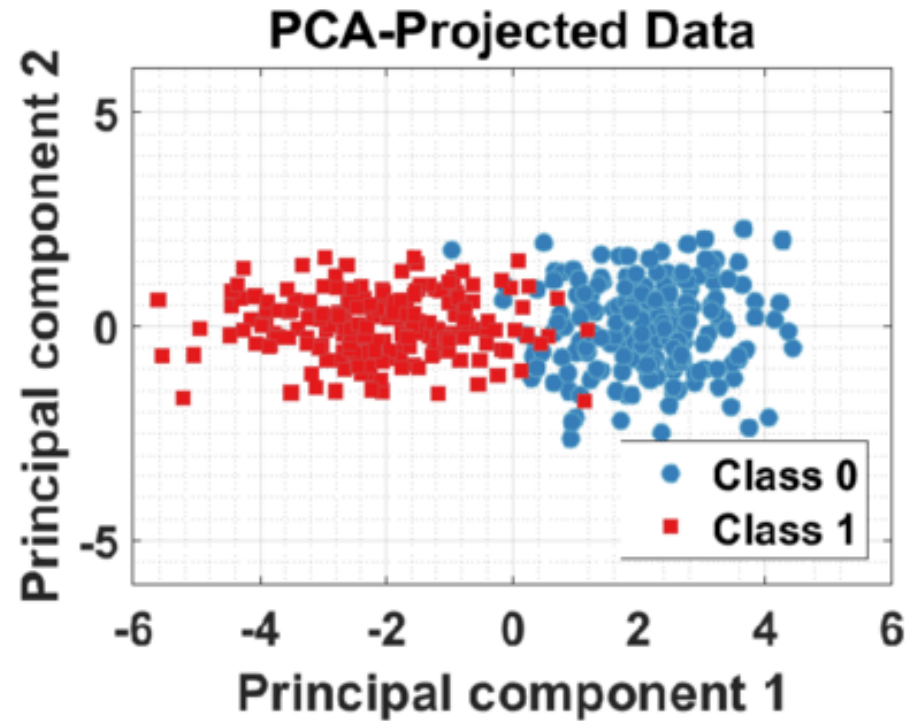
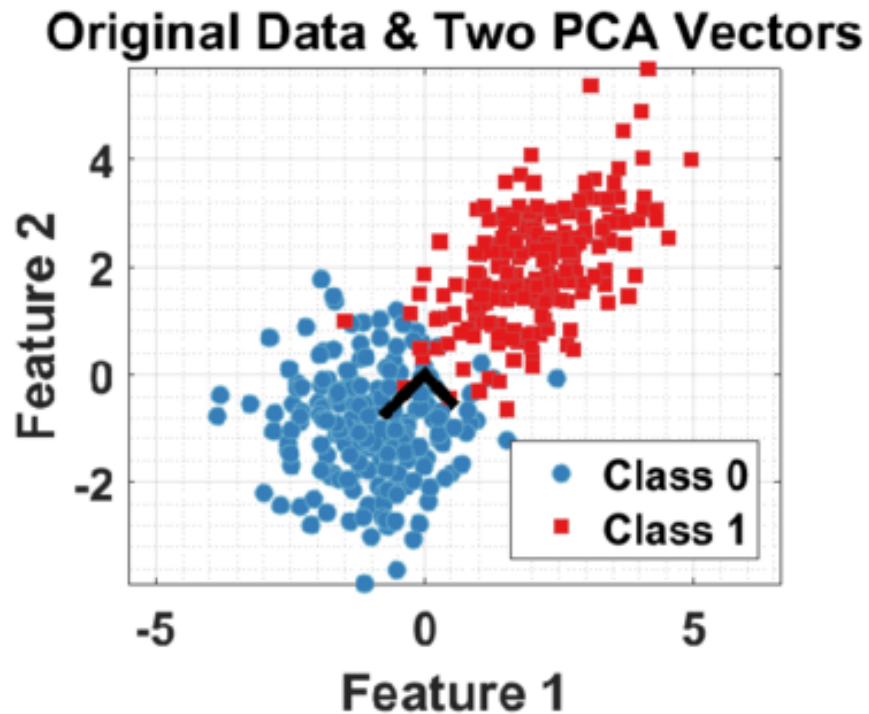
- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



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
 - 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)
 - Works for numeric data only

PCA: Data in the Eigen Space – A different representation



https://www.researchgate.net/publication/320410861_Physically_Motivated_Feature_Development_for_Machine_Learning_Applications/figures?lo=1

Principal component analysis

- Get data
- Subtract mean
(or bring it to zero mean, unit standard deviation form)
- Compute the covariance matrix
- Find Eigen values and Eigen vectors
- Select principal Eigen vectors (PCA)
 - Use proportion of variance retained by an eigen vector
(using eigen values)
- Project data onto selected Eigen vectors
- Plot data

PCA example

		x	y			x	y
Data =		2.5	2.4	DataAdjust =		.69	.49
		0.5	0.7			-1.31	-1.21
		2.2	2.9			.39	.99
		1.9	2.2			.09	.29
		3.1	3.0			1.29	1.09
		2.3	2.7			.49	.79
		2	1.6			.19	-.31
		1	1.1			-.81	-.81
		1.5	1.6			-.31	-.31
		1.1	0.9			-.71	-1.01

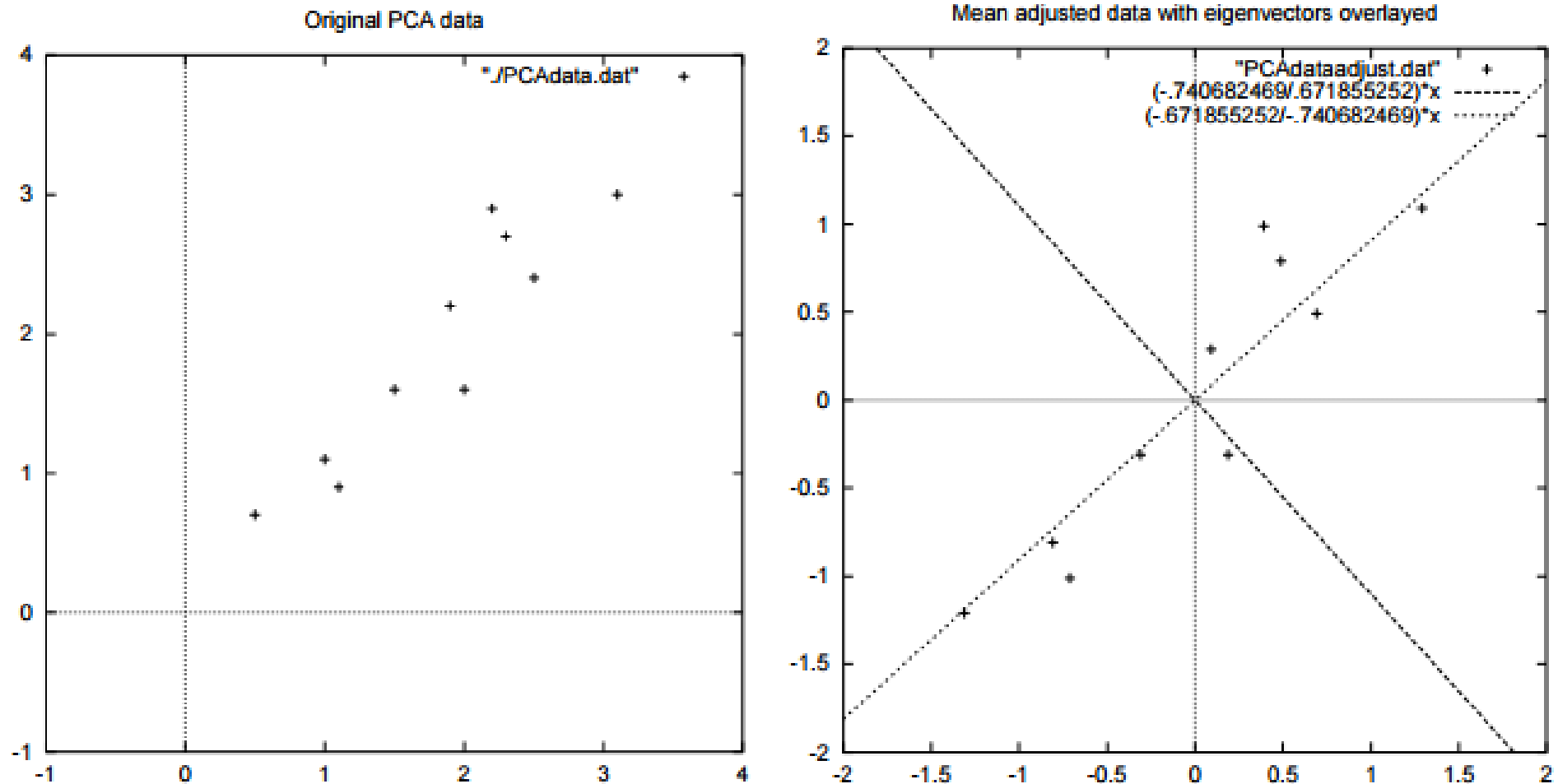
Covariance, Eigen analysis

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

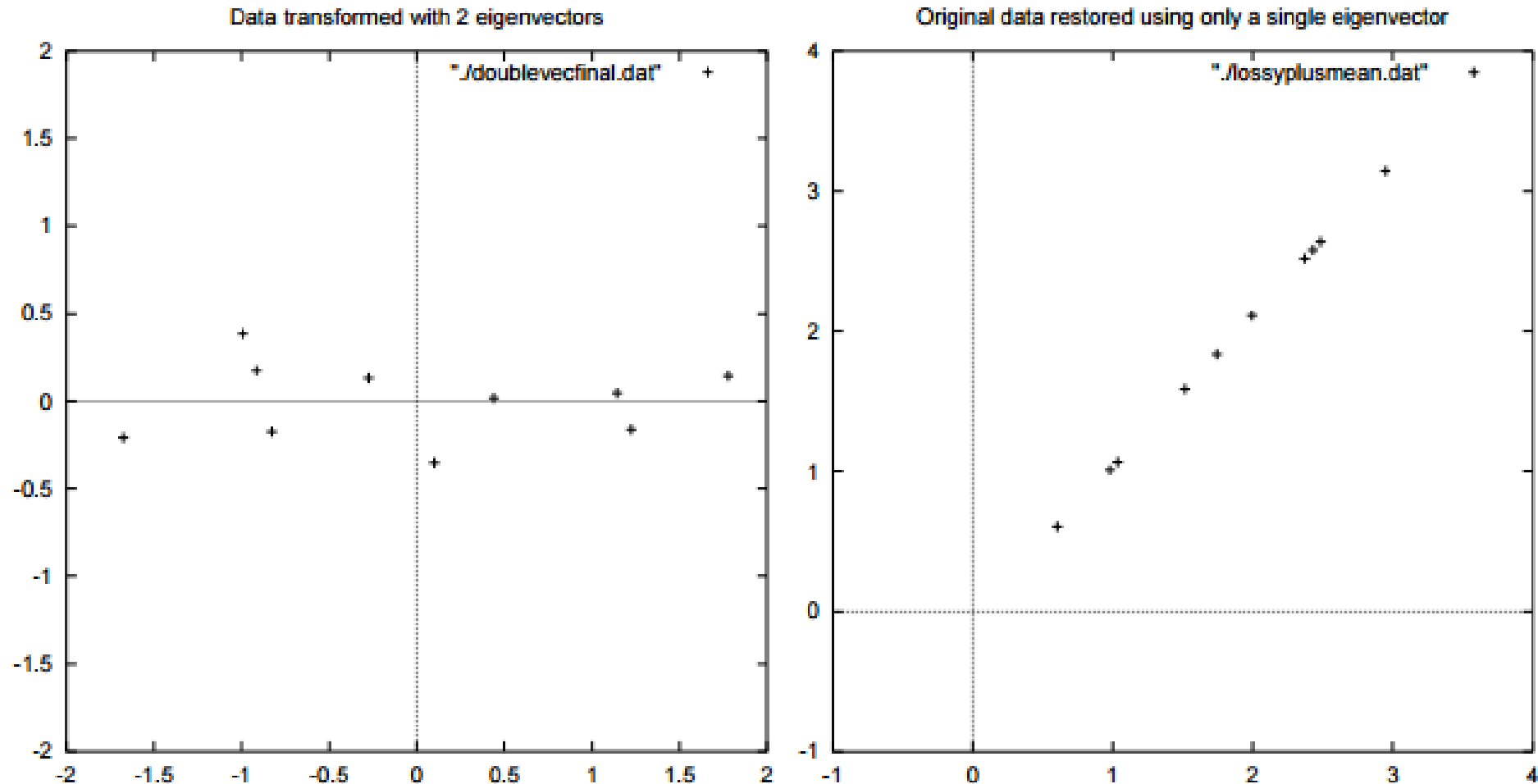
$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

Choosing an appropriate 'axis'



A new representation



PCA using R – 1

(factoMineR, factoextra)

name	100m	Long.jump	//	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58		63.19	291.7	1	8217	Decastar
CLAY	10.76	7.4		60.15	301.5	2	8122	Decastar
Macey	10.89	7.47		58.46	265.42	4	8414	OlympicG
Warners	10.62	7.74		55.39	278.05	5	8343	OlympicG
\\								
Zsivoczky	10.91	7.14		63.45	269.54	6	8287	OlympicG
Hernu	10.97	7.19		57.76	264.35	7	8237	OlympicG
Pogorelov	10.95	7.31		53.45	287.63	11	8084	OlympicG
Schoenbeck	10.9	7.3		60.89	278.82	12	8077	OlympicG
Barras	11.14	6.99		64.55	267.09	13	8067	OlympicG
KARPOV	11.02	7.3		50.31	300.2	3	8099	Decastar
WARNERS	11.11	7.6		51.77	278.1	6	8030	Decastar
Nool	10.8	7.53		61.33	276.33	8	8235	OlympicG
Drews	10.87	7.38		51.53	274.21	19	7926	OlympicG

	Active individuals
	Active variables
	Supplementary quantitative variables
	Supplementary qualitative variable
	Supplementary individuals

<http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/>

PCA using R - 2

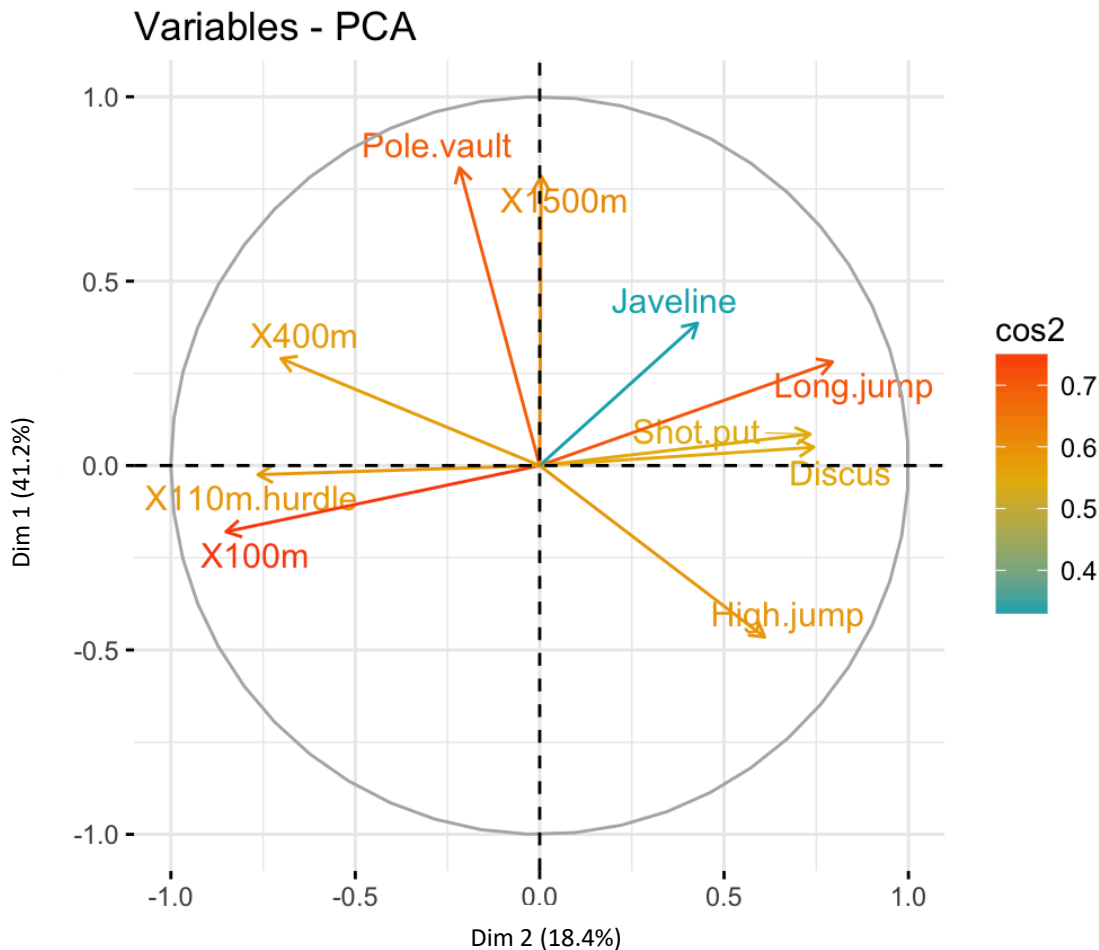
- Selecting the principal components

```
library("factoextra") eig.val <- get_eigenvalue(res.pca)
eig.val
```

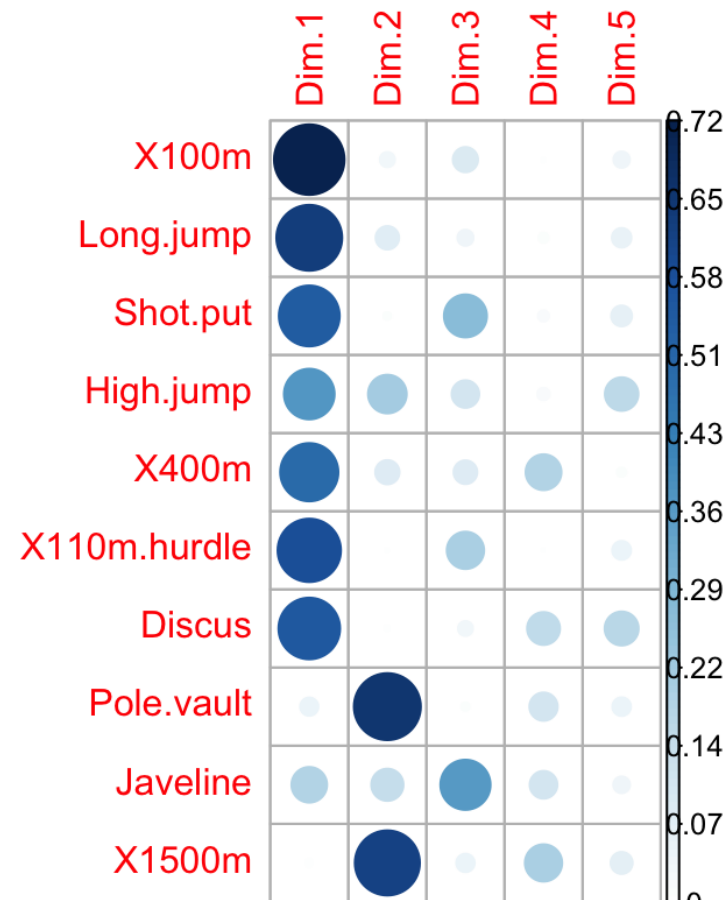
##	eigenvalue	variance.percent	cumulative.variance.percent
## Dim.1	4.124	41.24	41.2
## Dim.2	1.839	18.39	59.6
## Dim.3	1.239	12.39	72.0
## Dim.4	0.819	8.19	80.2
## Dim.5	0.702	7.02	87.2
## Dim.6	0.423	4.23	91.5
## Dim.7	0.303	3.03	94.5
## Dim.8	0.274	2.74	97.2
## Dim.9	0.155	1.55	98.8
## Dim.10	0.122	1.22	100.0

PCA using R - 3

- Correlation circle

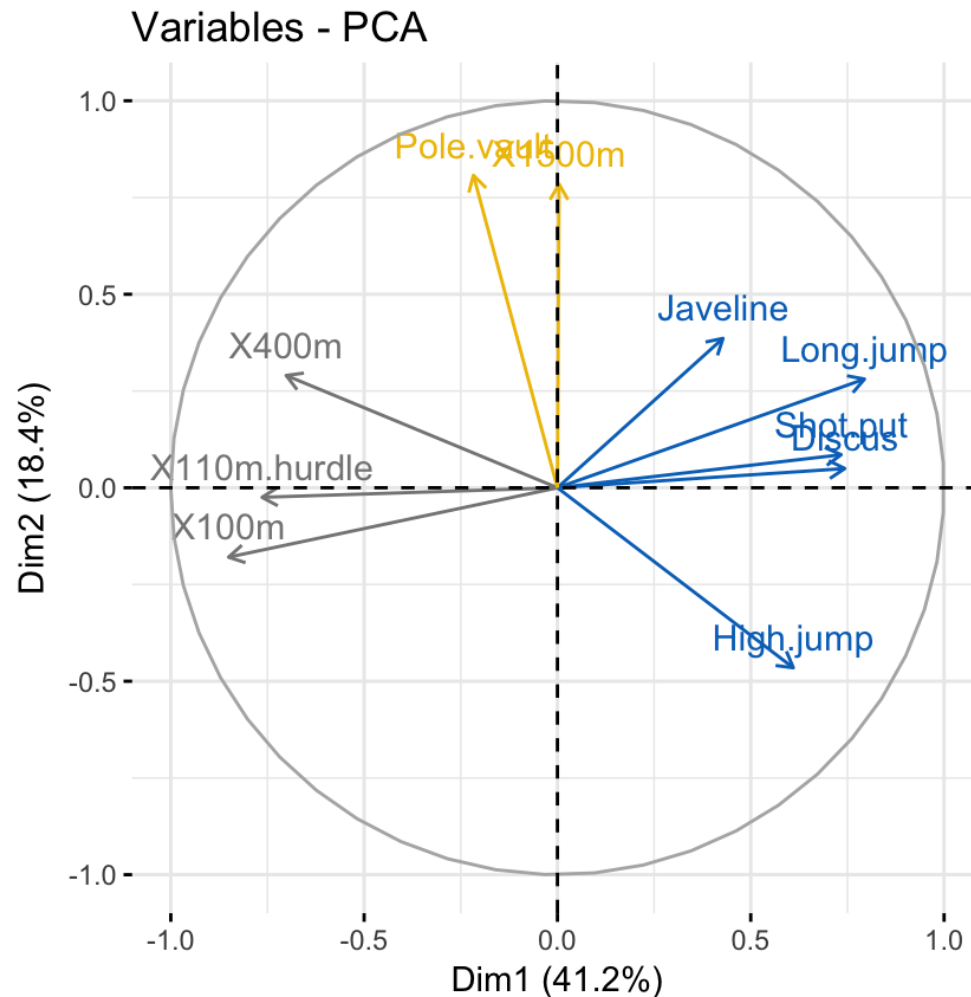


var.cos2 =
var.coord * var.coord

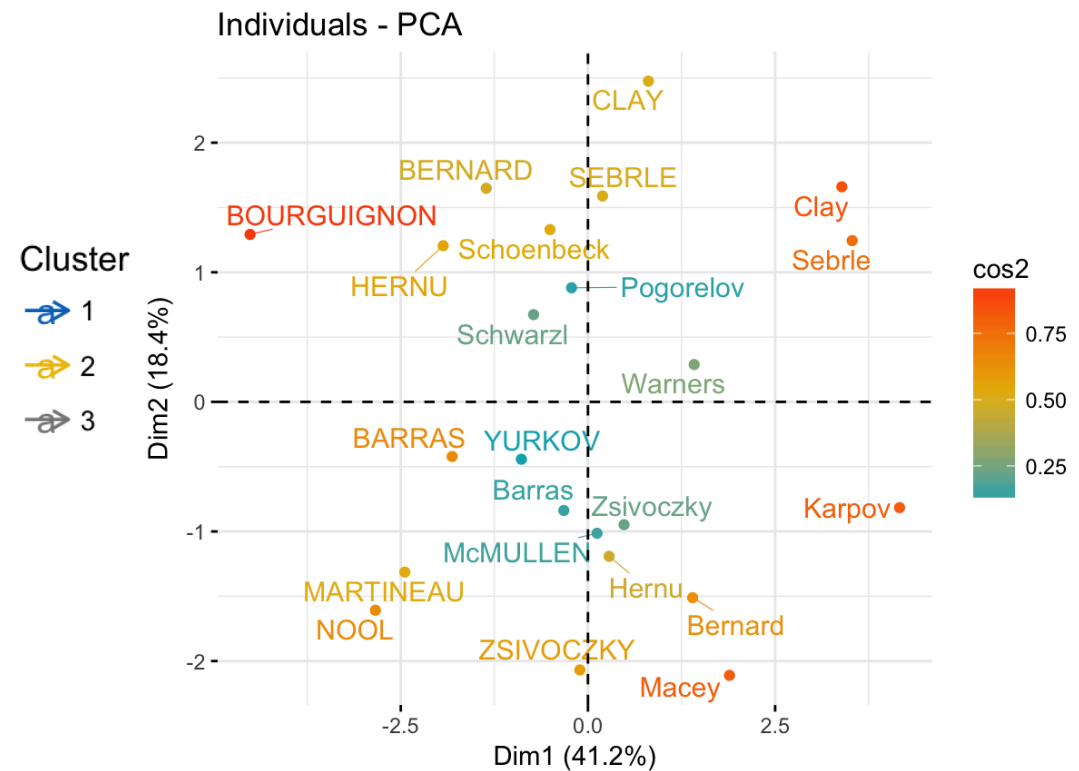


PCA using R - 4

- Which events are similar?

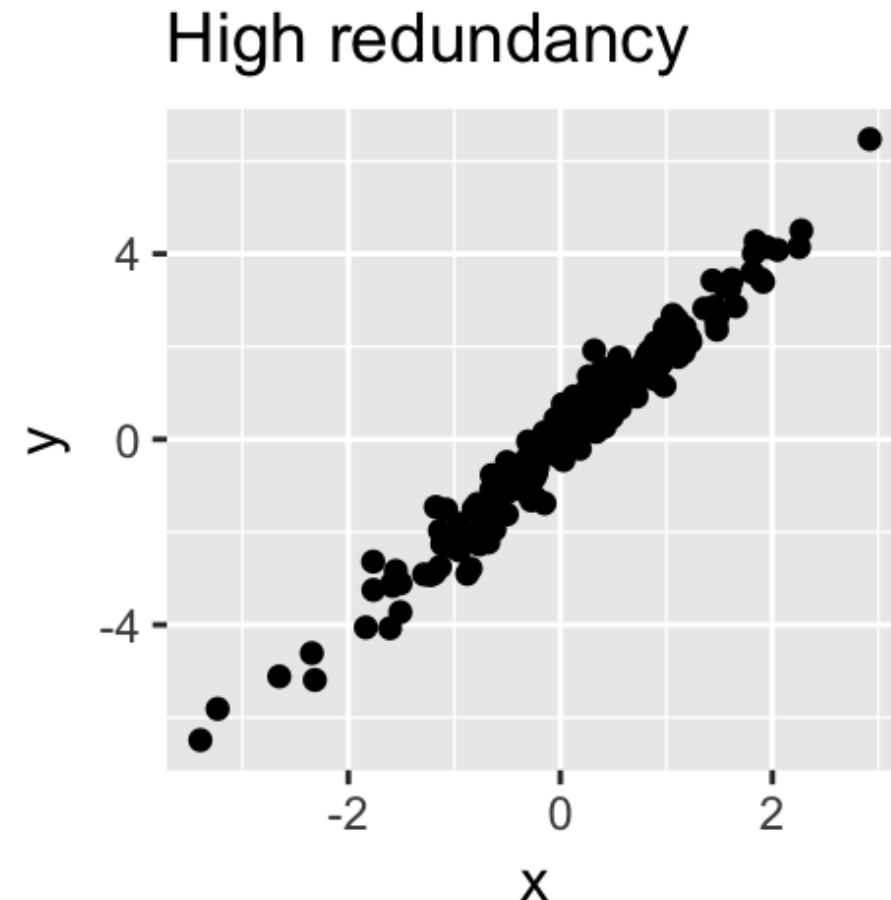
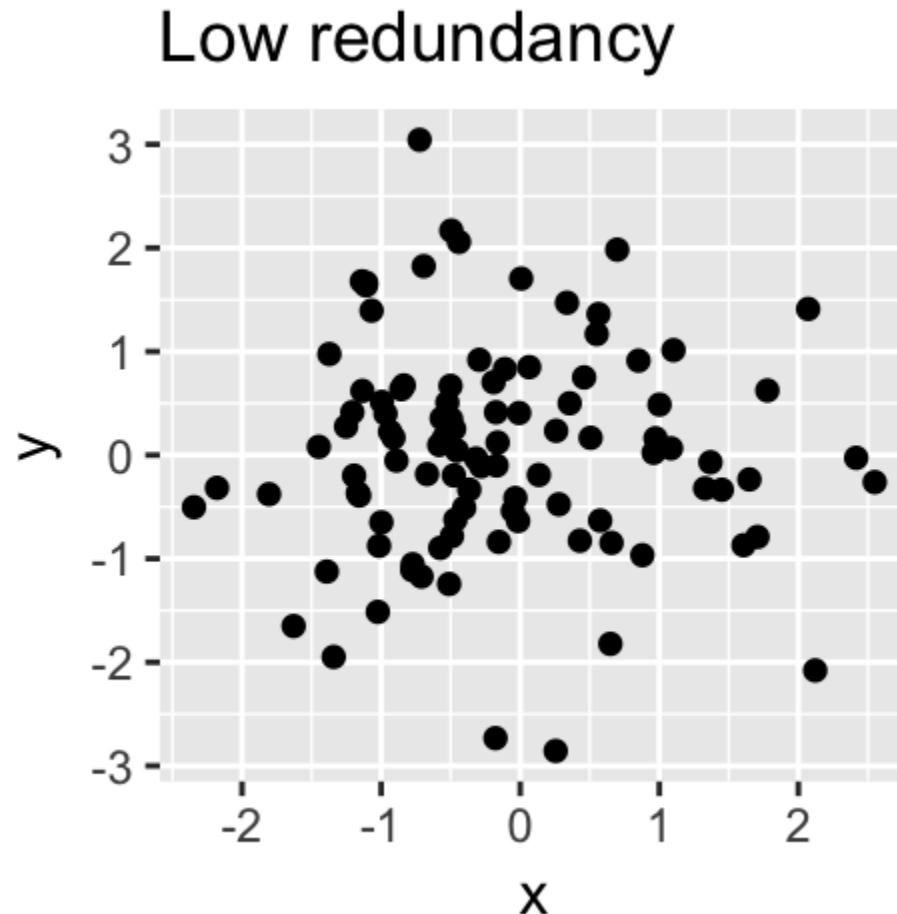


- Which athletes are similar?



Can PCA be used on all data?

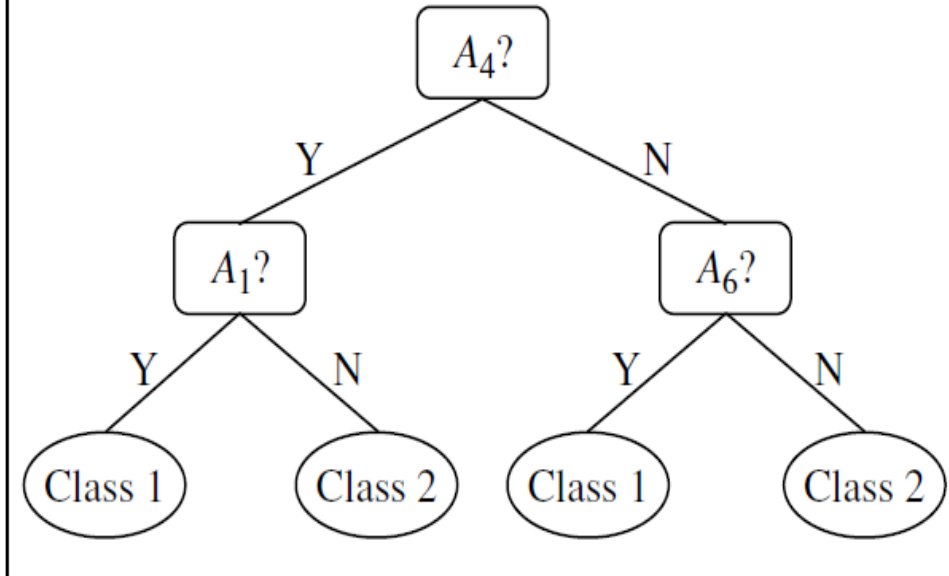
- PCA helps to un-correlate data



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

Heuristic Search in Attribute Selection

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

Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p>  <pre> graph TD A4["A4?"] -- Y --> A1["A1?"] A4 -- N --> A6["A6?"] A1 -- Y --> C1_1((Class 1)) A1 -- N --> C2_1((Class 2)) A6 -- Y --> C1_2((Class 1)) A6 -- N --> C2_2((Class 2)) </pre> <p>\Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>

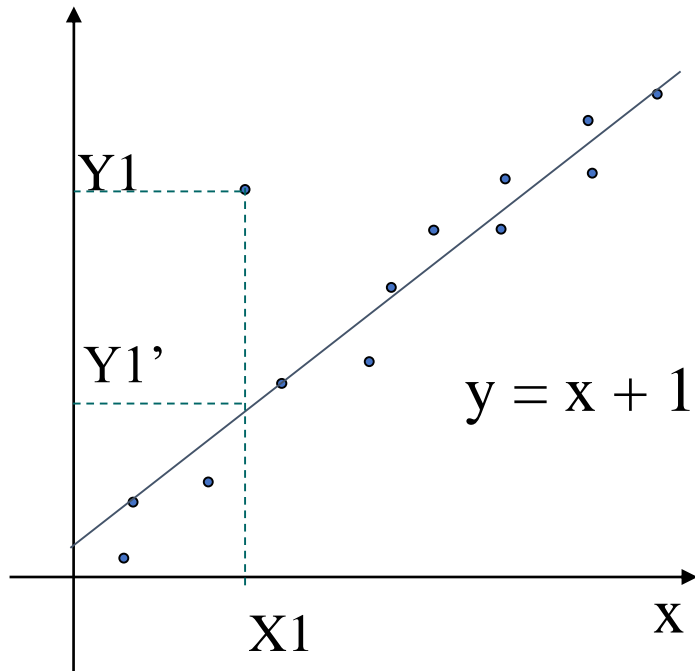
- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
 - Attribute extraction
 - Domain-specific
 - Mapping data to new space (see: data reduction)
 - E.g., Fourier transformation, wavelet transformation,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

- ☐ Mention and explain the different data reduction strategies.
- ☐ Explain how Wavelet transform and Principal Component Analysis are used in the process of data 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 m -D space as the product on appropriate marginal subspaces
- Non-parametric methods
 - Do not assume models
 - Major families: histograms, clustering, sampling, ...

- **Linear regression**
 - Data modeled to fit a straight line
 - Often uses the least-square method to fit the line
- **Multiple regression**
 - Allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- **Log-linear model**
 - Approximates discrete multidimensional probability distributions

- Regression analysis: 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

- **Linear regression:** $Y = w X + b$
- Two regression coefficients, w and b , specify the line and are to be estimated by using the data at hand
- Using the least squares criterion to the known values of $Y_1, Y_2, \dots, X_1, X_2, \dots$
- **Multiple regression:** $Y = b_0 + b_1 X_1 + b_2 X_2$
- Many nonlinear functions can be transformed into the above

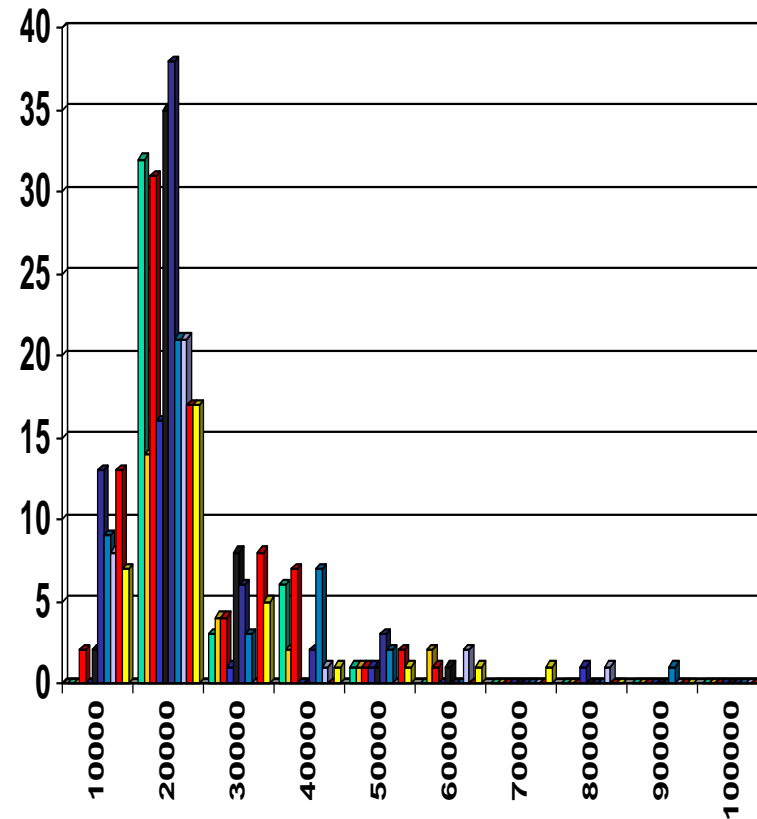
Log-linear models:

- Approximate discrete multidimensional probability distributions
- Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations
- Useful for dimensionality reduction and data smoothing

DATA ANALYTICS

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)



- 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

The process of identifying a subset from a population of elements (aka observations or cases) is called **sampling process** or **simply sampling**

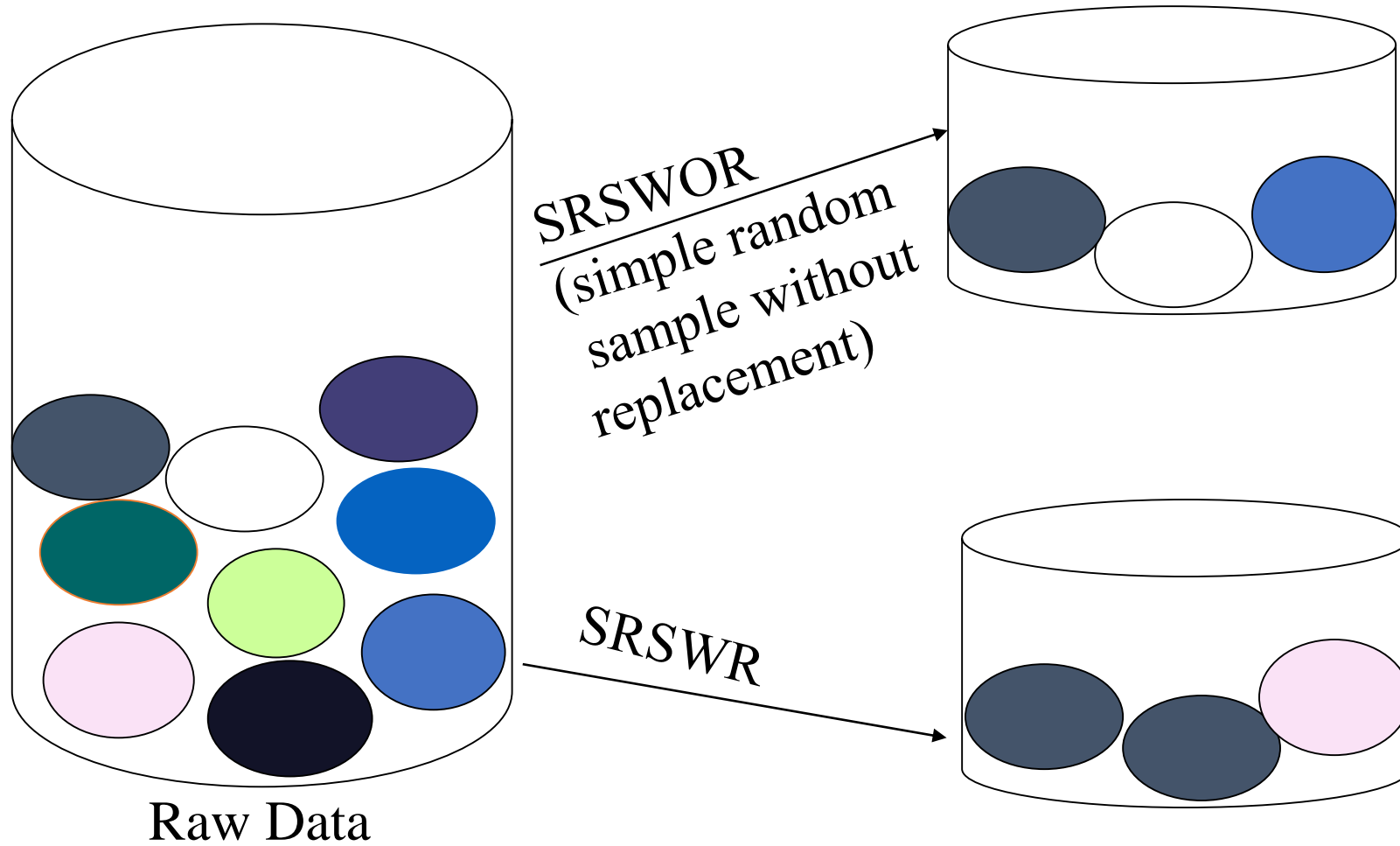
Steps used in any Sampling process:

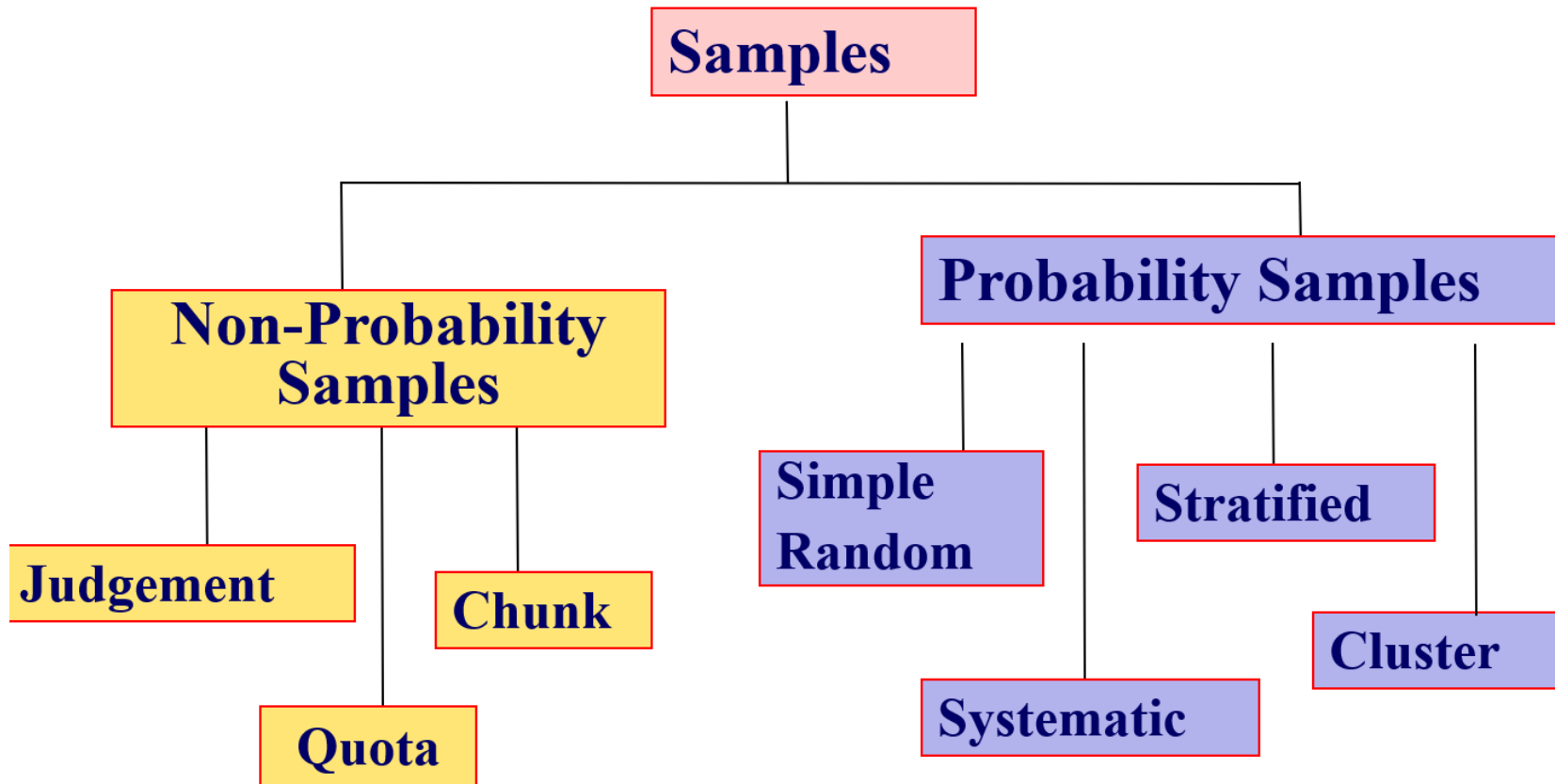
- Identification of target population that is important for a given problem under study
- Decide the sampling frame.
- Determine the sample size
- Sampling method

- Sampling: obtaining a small sample s to represent the whole data set N
- Allow an analytics 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 ANALYTICS

Sampling: With or without Replacement





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

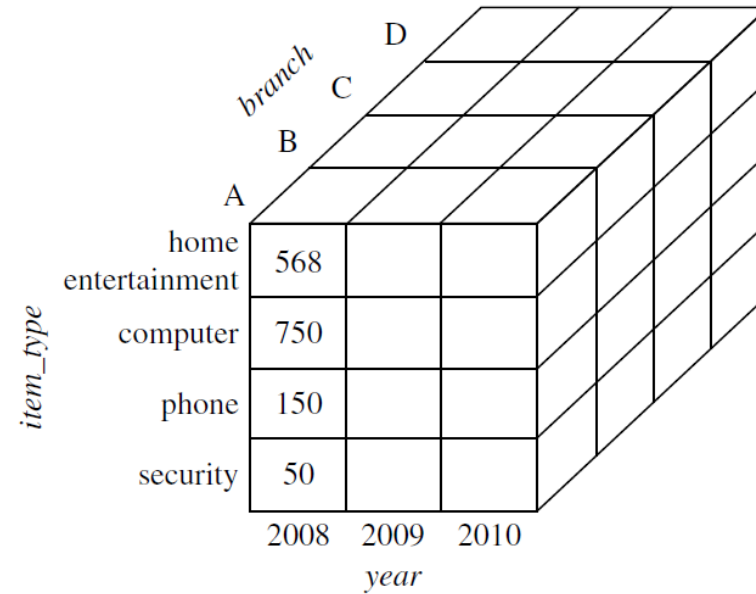
DATA ANALYTICS

Data Cube Aggregation

Year 2010	
Quarter	Sales
Year 2009	
Quarter	Sales
Year 2008	
Quarter	Sales
Q1	\$224,000
Q2	\$408,000
Q3	\$350,000
Q4	\$586,000

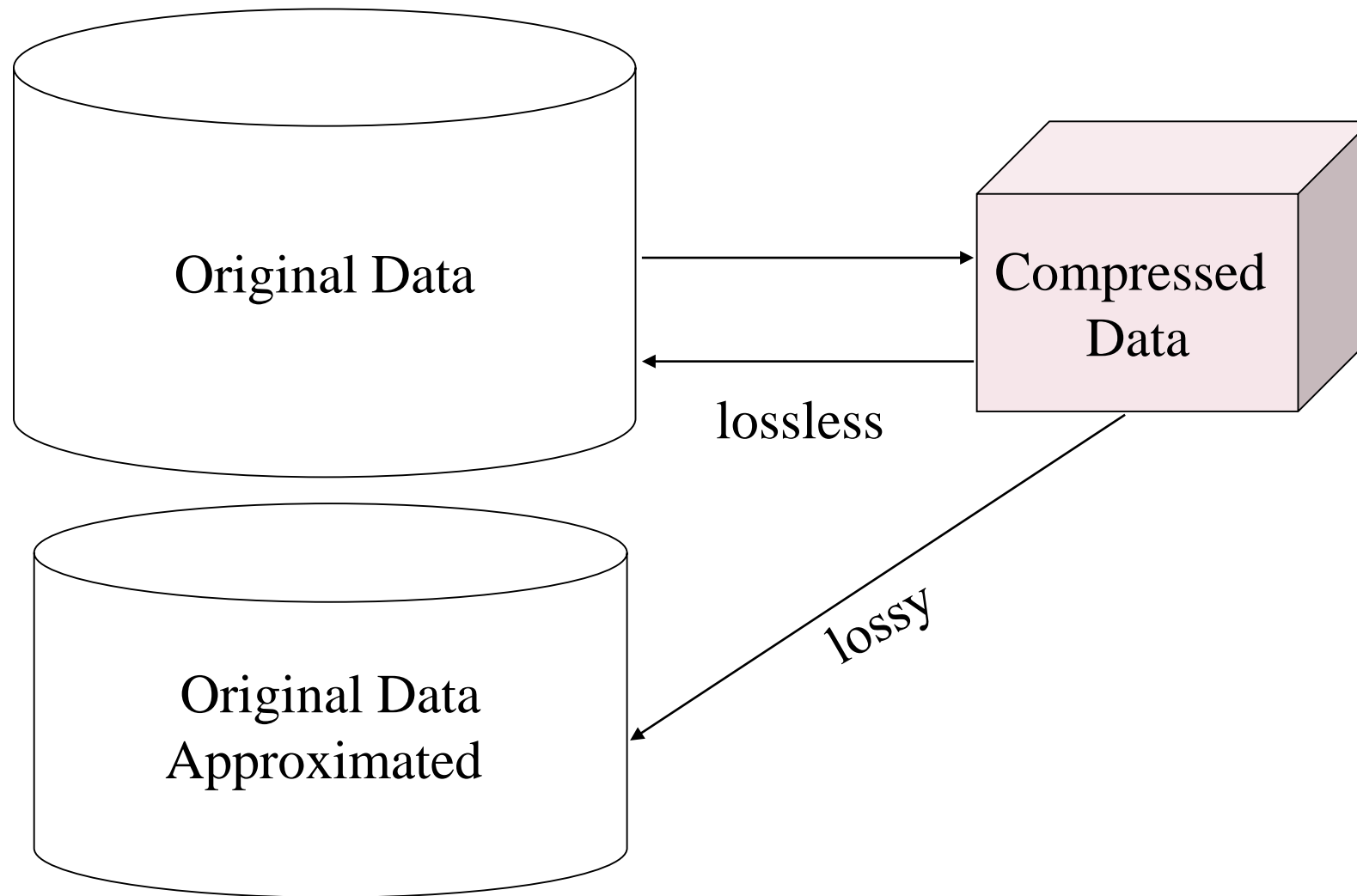


Year	Sales
2008	\$1,568,000
2009	\$2,356,000
2010	\$3,594,000



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



- ☐ Mention and explain the different parametric and non parametric methods used in data reduction.
- ☐ Compare and contrast the probability and non probability sampling methods.

Text Book:

- [Data Mining: Concepts and Techniques](#) by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.

DATA ANALYTICS

Unit 1: Data Transformation and Data Discretization

Mamatha H R

Department of Computer Science and Engineering

A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values

■ Methods

- Smoothing: Remove noise from data
 - Simple average or weighted average or Gaussian
- Attribute/feature construction
 - New attributes constructed from the given ones
- Aggregation: Summarization, data cube construction
- Normalization: Scaled to fall within a smaller, specified range
 - min-max normalization
 - z-score normalization
 - normalization by decimal scaling
- Discretization: Concept hierarchy climbing

- **Min-max normalization:** to $[new_min_A, new_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

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

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

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

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

- **Normalization by decimal scaling**

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

- 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

- Typical methods: All the methods can be applied recursively
 - Binning
 - Top-down split, unsupervised
 - 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)

- **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

- ❑ 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

- ❑ Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

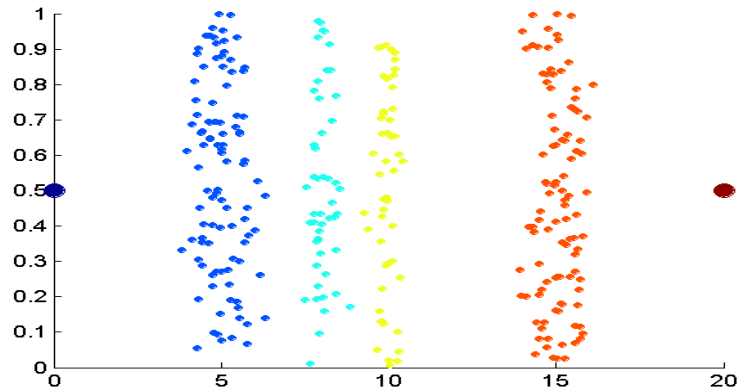
- ❑ Smoothing by **bin means**:
 - Bin 1: 9, 9, 9, 9
 - Bin 2: 23, 23, 23, 23
 - Bin 3: 29, 29, 29, 29

- ❑ Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

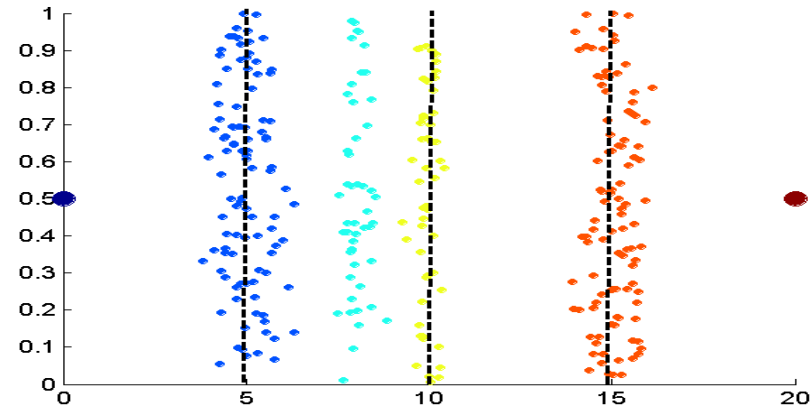
* Smoothing by **bin boundaries**:

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

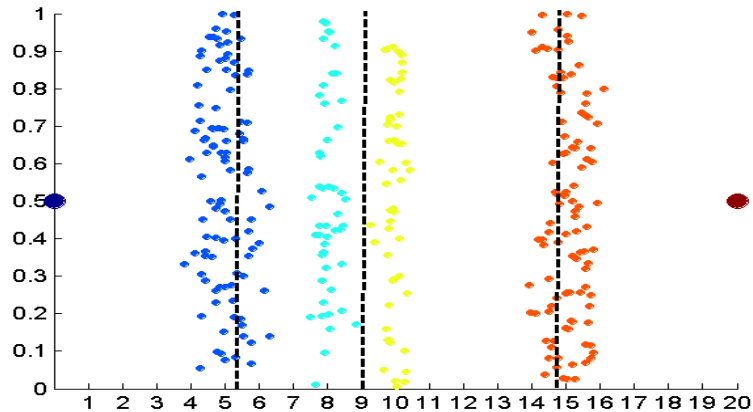
Discretization Without Using Class Labels (Binning vs. Clustering)



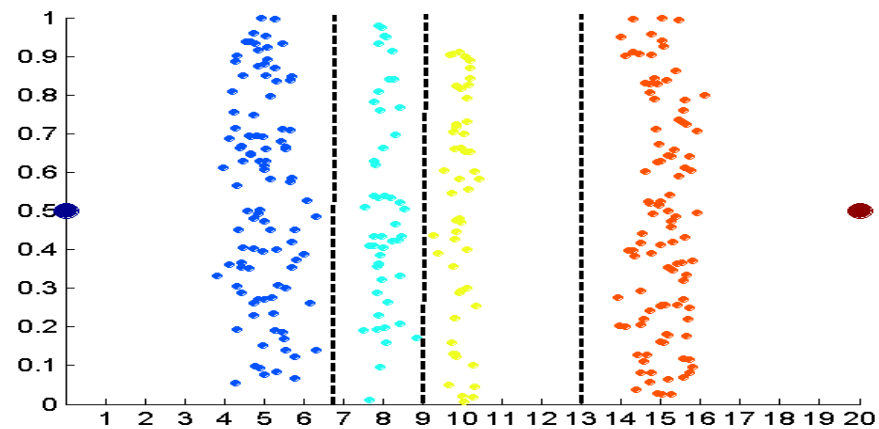
Original Data



Equal Width(binning)



Equal frequency (binning)

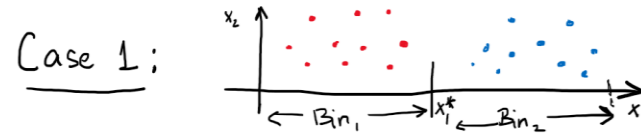


K-means clustering leads to
better results

- 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

$$\text{Entropy } E = \sum_{i=1}^n P(C_i) \log_2(1/P(C_i)), \quad n \text{ is the \# categories}$$

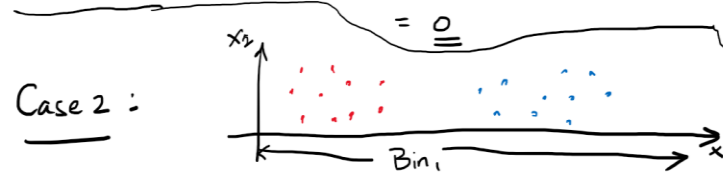
$$= - \sum_{i=1}^n P(C_i) \log_2(P(C_i))$$



$$\begin{aligned} \text{Entropy}(\text{Bin}_1) &= - (P(\text{red}) \log_2(P(\text{red})) + P(\text{blue}) \log_2(P(\text{blue}))) \\ &= - (1 \log_2(1) + 0 \log_2(0)) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{Entropy}(\text{Bin}_2) &= - (P(\text{red}) \log_2(P(\text{red})) + P(\text{blue}) \log_2(P(\text{blue}))) \\ &= - (0 + 1 \log_2(1)) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{Overall entropy} &= \frac{10}{20} E(\text{Bin}_1) + \frac{10}{20} E(\text{Bin}_2) \\ &= 0 \end{aligned}$$



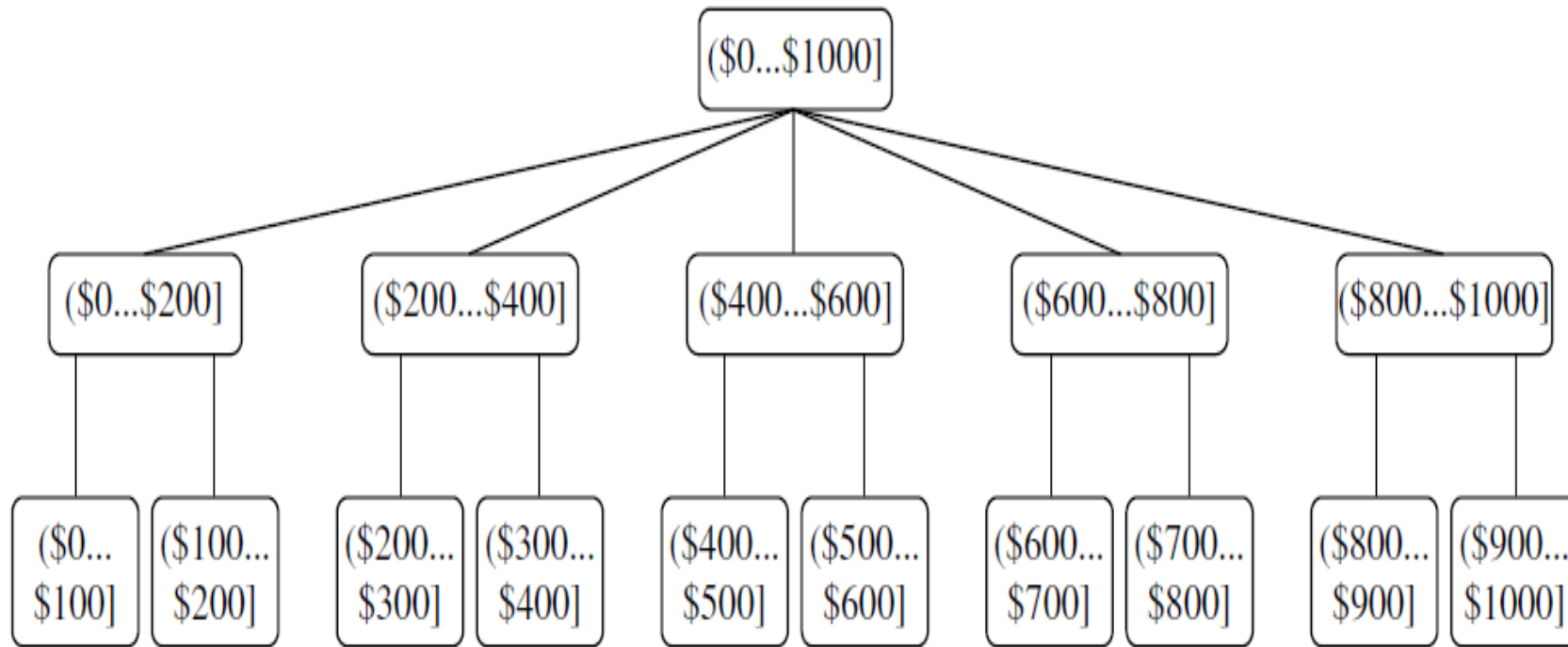
$$\begin{aligned} \text{Entropy}(\text{Bin}_1) &= - (P(\text{red}) \log_2(P(\text{red})) + P(\text{blue}) \log_2(P(\text{blue}))) \\ &= - (1/2 \log_2(1/2) + 1/2 \log_2(1/2)) \\ &= - (-1/2 - 1/2) \\ &= 1 \end{aligned}$$

Since $\text{Entropy}(\text{Case}_1) < \text{Entropy}(\text{Case}_2)$, $\{(0, x_1^*), (x_1^*, 1)\}$ is a better binning strategy than $(0, 1)$ as one Bin.

- 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** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity

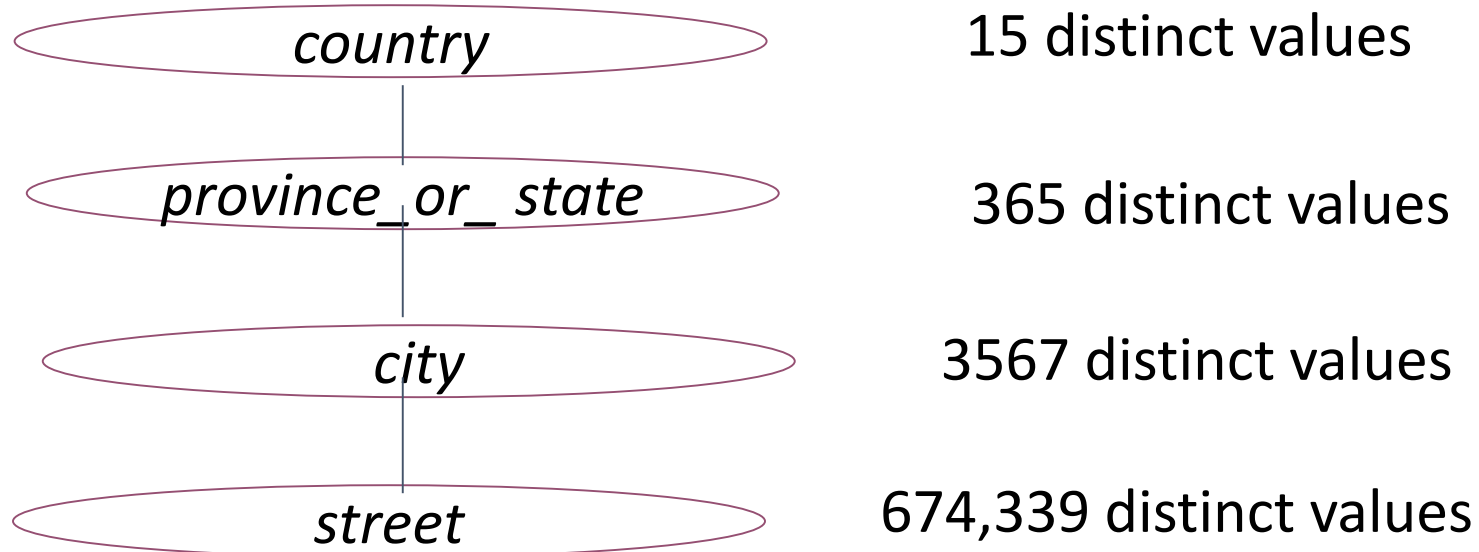
- Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth*, *adult*, or *senior*)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed for both numeric and nominal data. For numeric data, use discretization methods shown.



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

Automatic Concept Hierarchy Generation

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



- ☐ Mention and explain the different data normalization techniques.
- ☐ How classification and correlation analysis is used in data discretization.

Text Book:

- [Data Mining: Concepts and Techniques](#) by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.



THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834