

# Data preprocessing



Data Base and Data Mining Group of Politecnico di Torino

Elena Baralis, Tania Cerquitelli

*Politecnico di Torino*



# Outline

- Data types and properties
- Data preparation
- Data preparation for document data
- Similarity and dissimilarity
- Correlation

# Data types and properties



Data Base and Data Mining Group of Politecnico di Torino



# What is Data?

- Collection of **data objects** and their **attributes**
- An **attribute** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describes an **object**
  - Object is also known as record, point, case, sample, entity, or instance

## 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



# Attribute Values

- ***Attribute values*** are numbers or symbols assigned to an attribute for a particular object
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different



# Attribute types

- There are different types of attributes
  - **Nominal**
    - Examples: ID numbers, eye color, zip codes
  - **Ordinal**
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - **Interval**
    - Examples: calendar dates
  - **Ratio**
    - Examples: temperature in Kelvin, length, time, counts



# Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness:  $= \neq$
  - Order:  $< >$
  - Addition:  $+ -$
  - Multiplication:  $* /$
- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & addition
- Ratio attribute: all 4 properties



# Discrete and Continuous Attributes

## ■ Discrete Attribute

- Has only a finite or countably 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.





# More Complicated Examples

- ID numbers
  - Nominal, ordinal, or interval?
  
- Number of cylinders in an automobile engine
  - Nominal, ordinal, or ratio?



# Key Messages for Attribute Types

- The types of operations you choose should be “meaningful” for the type of data you have
  - Distinctness, order, meaningful intervals, and meaningful ratios are only four properties of data
  - The data type you see – often numbers or strings – may not capture all the properties or may suggest properties that are not there
  - Analysis may depend on these other properties of the data
    - Many statistical analyses depend only on the distribution
  - Many times what is meaningful is measured by statistical significance
  - But in the end, what is meaningful is measured by the domain



# Data set types

- Record
  - Tables
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data



# Tabular Data

- A collection of records
  - Each record is characterized by a fixed set of attributes

<i>Tid</i>	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



# Document data

- It includes **textual data** that can be **semi-structured** or **unstructured**
  - Plain text can be organized in sentences, paragraphs, sections, documents
- Text acquired in different contexts may have a structure and/or a semantics
  - **Web pages** are enriched with tags
  - **Documents in digital libraries** are enriched with metadata
  - **E-learning documents** can be annotated or partly highlighted



# Document Data

- Each document becomes a 'term' vector
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0



# Transaction Data

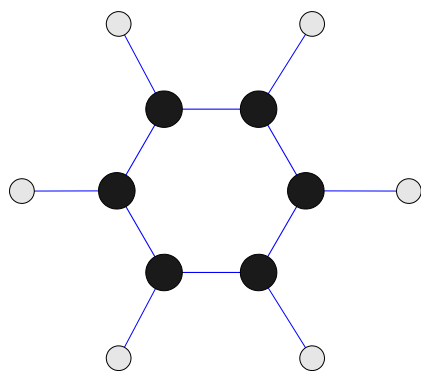
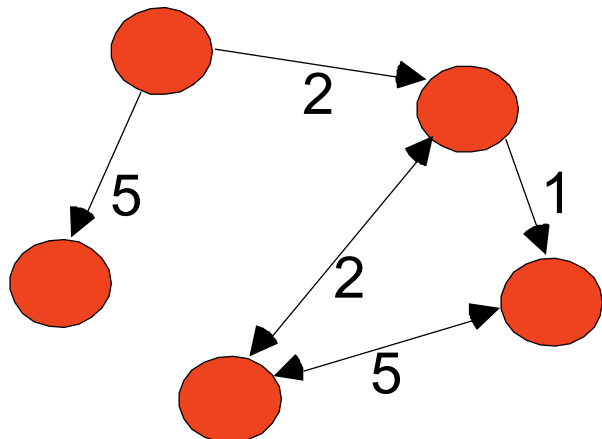
- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk



# Graph Data

- Examples: Generic graph, a molecule, and webpages



Benzene Molecule: C<sub>6</sub>H<sub>6</sub>

## Useful Links:

- [Bibliography](#)
- Other Useful Web sites
  - [ACM SIGKDD](#)
  - [KDnuggets](#)
  - [The Data Mine](#)

## Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- [Books](#)
- [General Data Mining](#)

## Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.  
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support)", John Wiley & Sons, 1997.

## General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

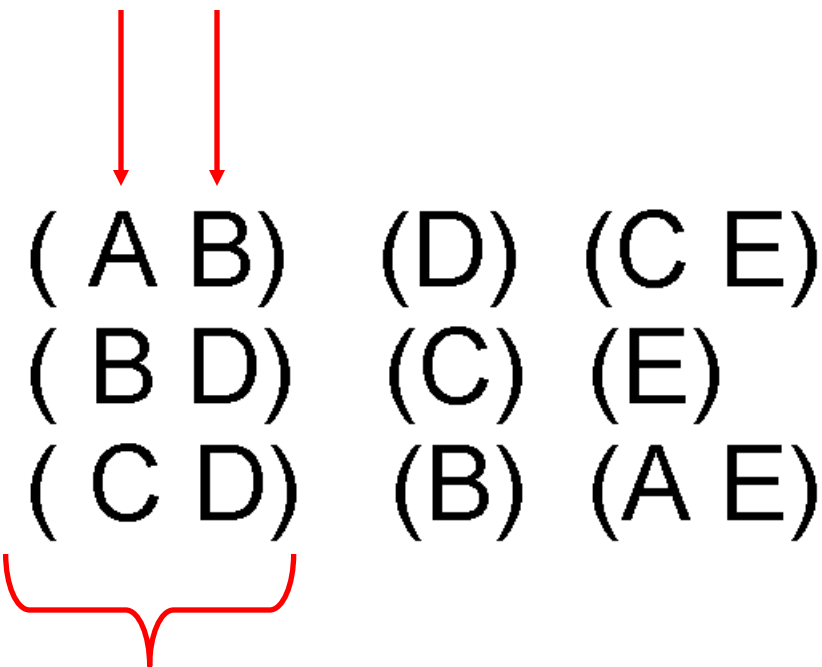




# Ordered Data

- Sequences of transactions

Items/Events



An element of  
the sequence



# Ordered Data

- Genomic sequence data

GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG

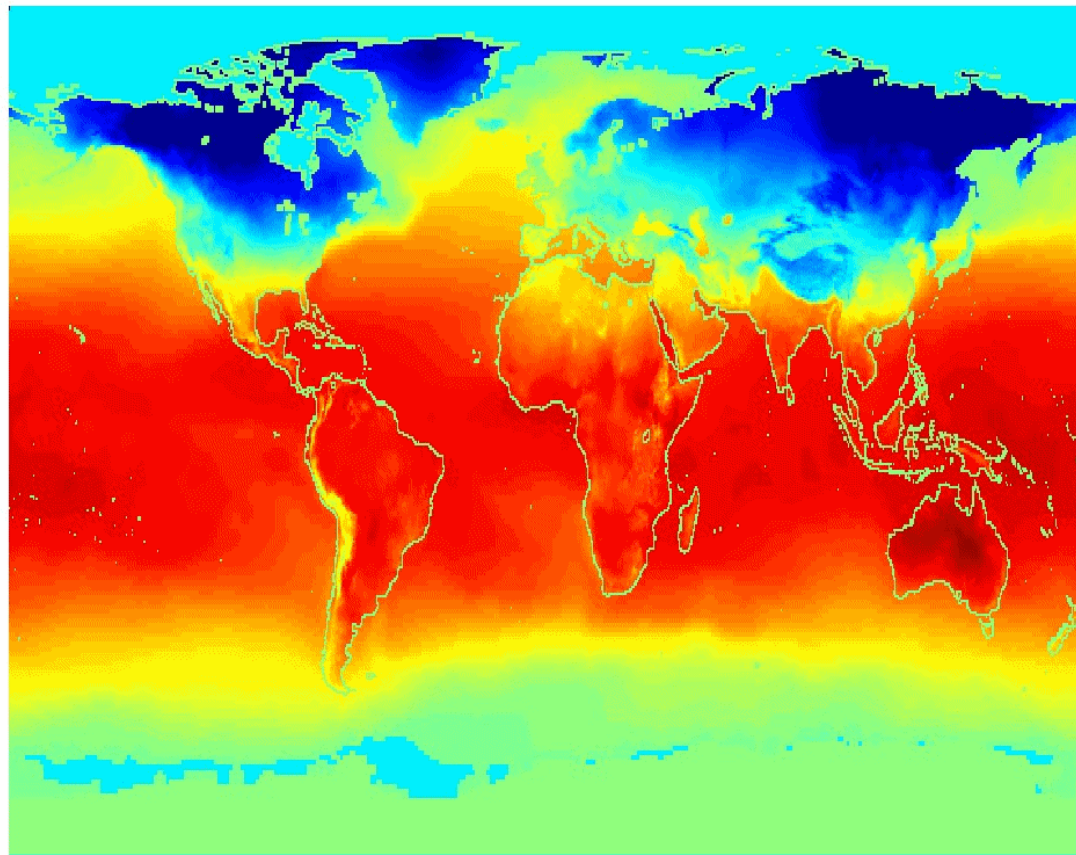


# Ordered Data

- Spatio-Temporal Data

Average Monthly  
Temperature of  
land and ocean

Jan





# Data Quality

- Poor data quality negatively affects many data processing efforts

“The most important point is that poor data quality is an unfolding disaster. Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate.”

Thomas C. Redman, DM Review, August 2004

- Data mining example: a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default



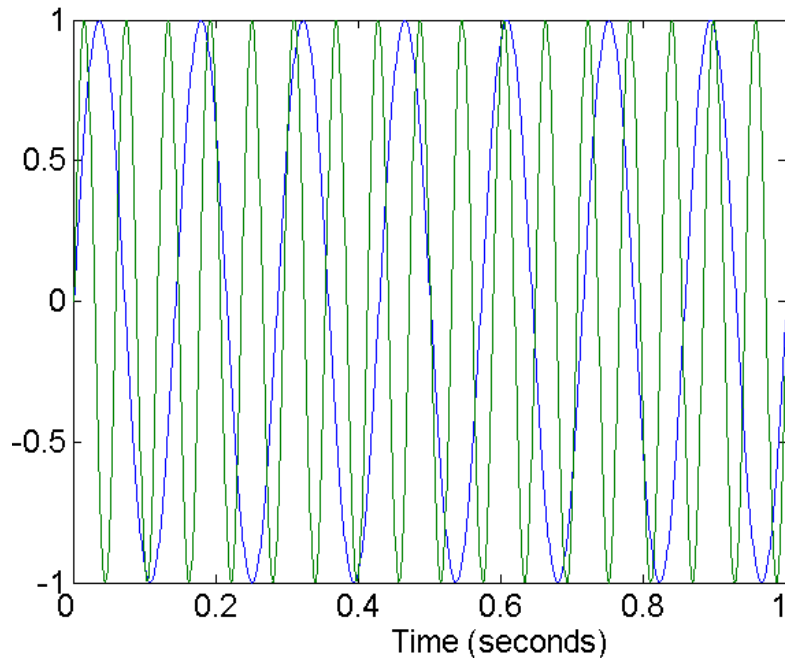
# Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems
  - Noise and outliers
  - Missing values
  - Duplicate data
  - Wrong data

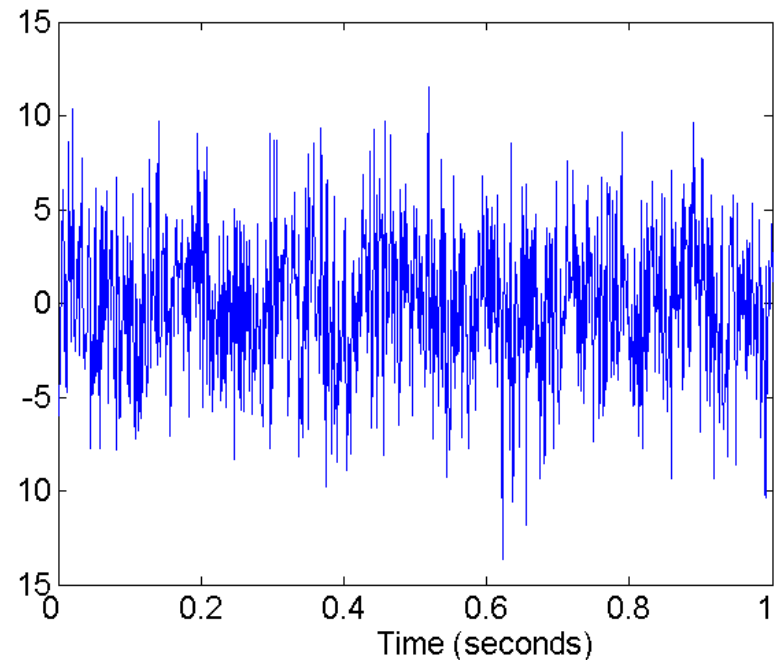


# Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



Two Sine Waves

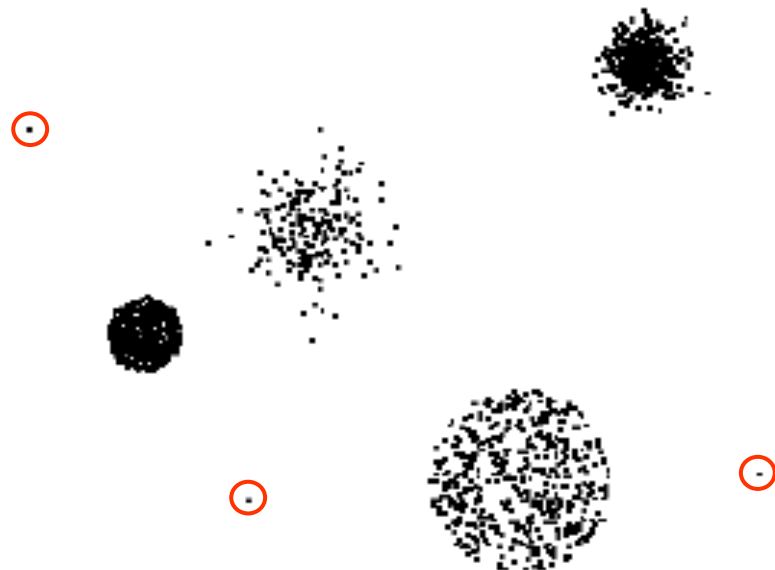


Two Sine Waves + Noise



# Outliers

- **Outliers** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection





# Missing Values

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- Handling missing values
  - Eliminate data objects or variables
  - Estimate missing values
    - Example: time series of temperature
  - Ignore the missing value during analysis





# Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples
  - Different words/abbreviations for the same concept (e.g., Street, St.)
- Data cleaning
  - Process of dealing with duplicate data issues
- When should duplicate data not be removed?

# Data preparation



Data Base and Data Mining Group of Politecnico di Torino



# Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation



# 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



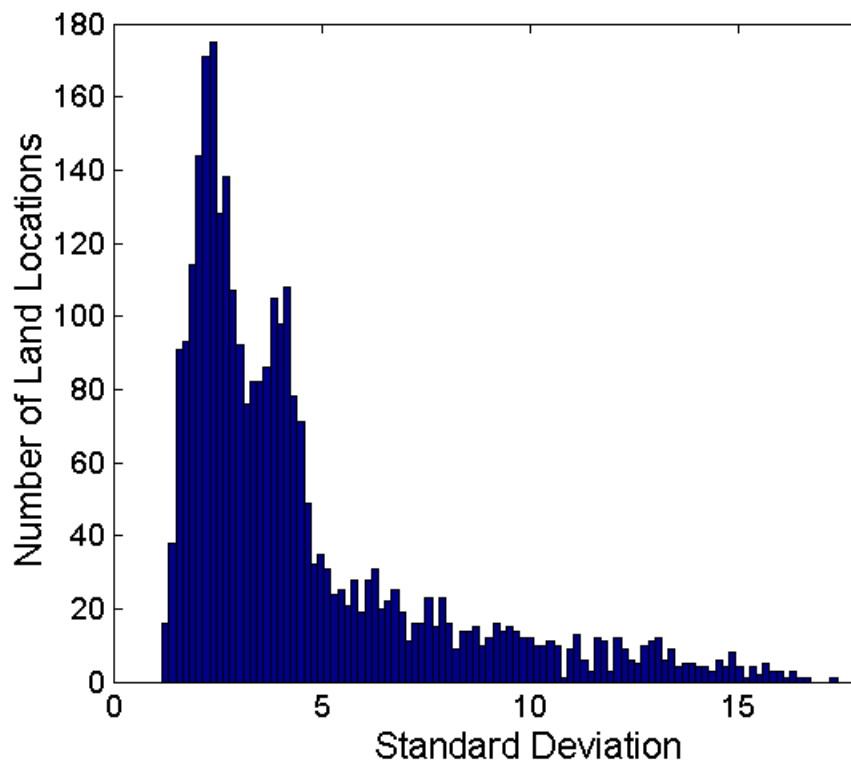
# Example: Precipitation in Australia

- The next slide shows precipitation in Australia from the period 1982 to 1993
  - A histogram for the standard deviation of average monthly precipitation for 3,030  $0.5^\circ$  by  $0.5^\circ$  grid cells in Australia
  - A histogram for the standard deviation of the average yearly precipitation for the same locations.
- The average yearly precipitation has less variability than the average monthly precipitation.
- All precipitation measurements (and their standard deviations) are in centimeters.

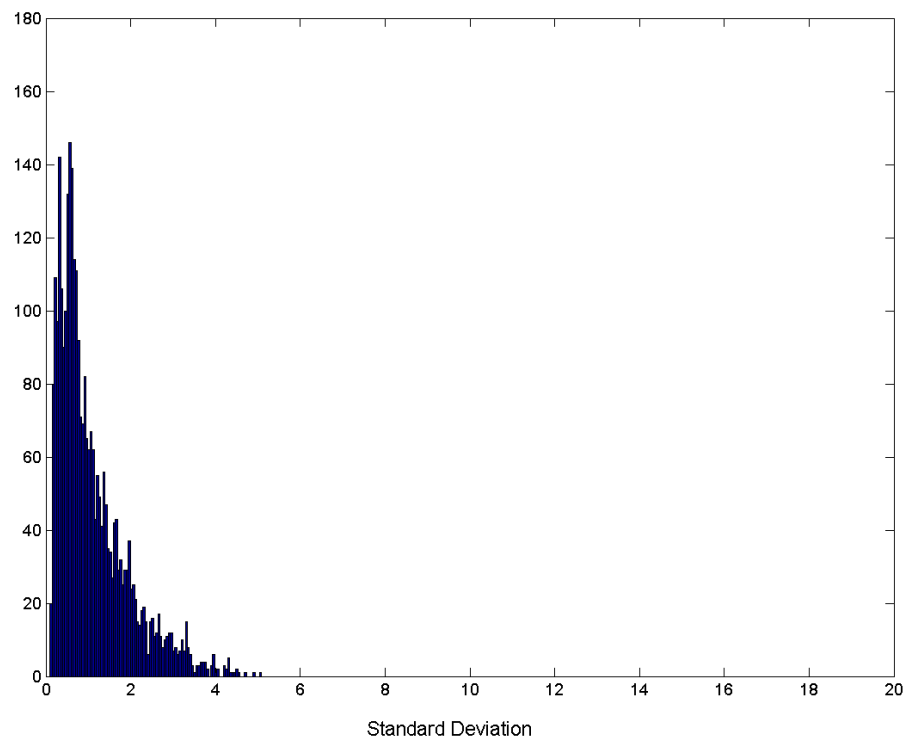


# Aggregation

## Variation of Precipitation in Australia



Standard Deviation of Average  
Monthly Precipitation



Standard Deviation of Average  
Yearly Precipitation



# Data reduction

- Generates a reduced representation of the dataset
- This representation is smaller in volume, but it can provide similar analytical results
  - sampling
    - reduces the cardinality of the set
  - feature selection
    - reduces the number of attributes
  - discretization
    - reduces the cardinality of the attribute domain



# 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.



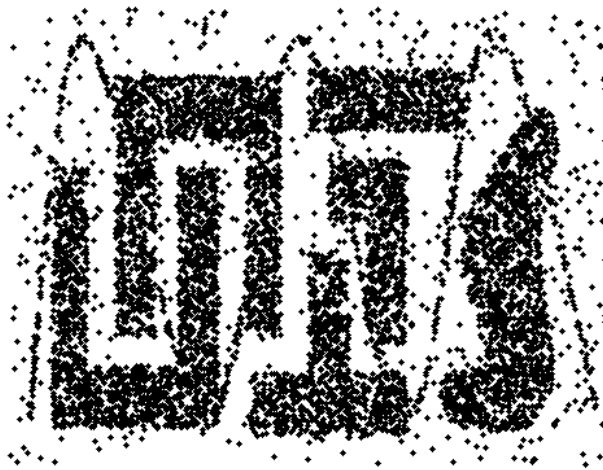


# Sampling ...

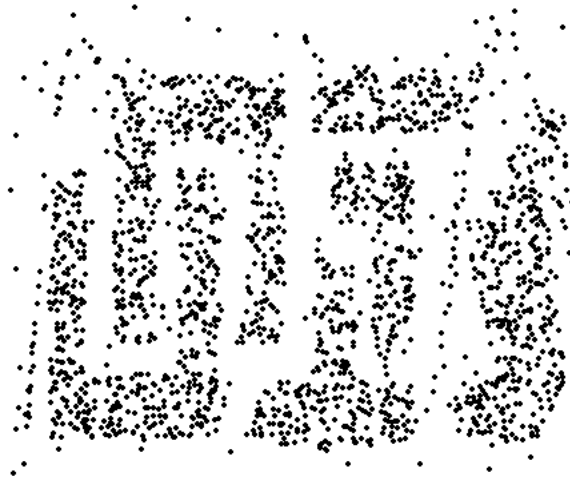
- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data set, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data



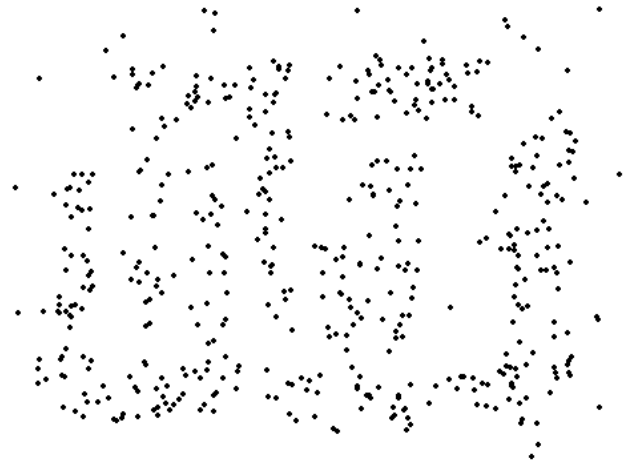
# Sample Size: examples



8000 points



2000 Points



500 Points



# Types of Sampling

## ■ Simple Random Sampling

- There is an equal probability of selecting any particular item
- 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

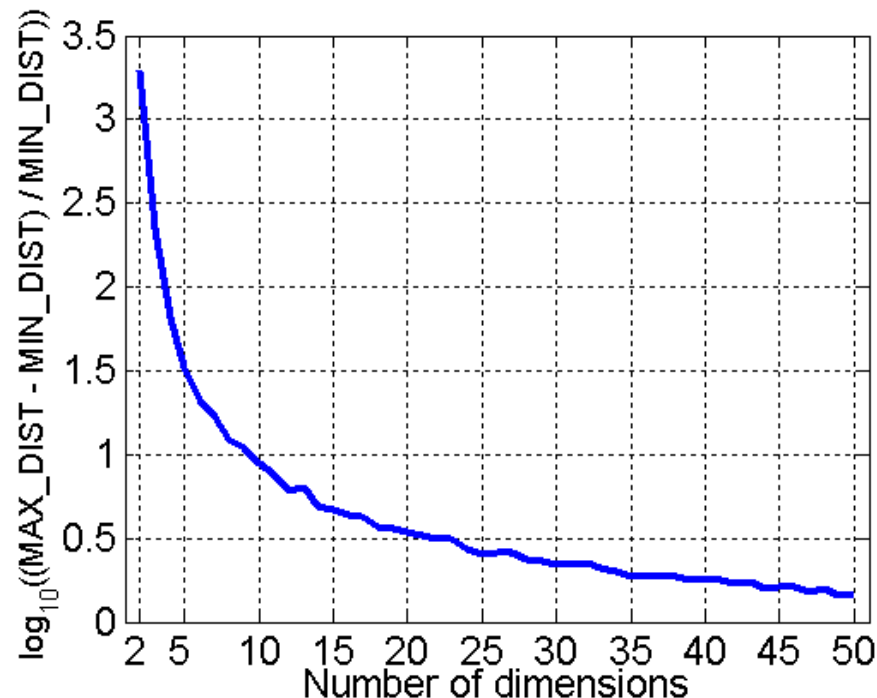
## ■ Stratified sampling

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



# Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points



# Dimensionality Reduction

## ■ Purpose

- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise

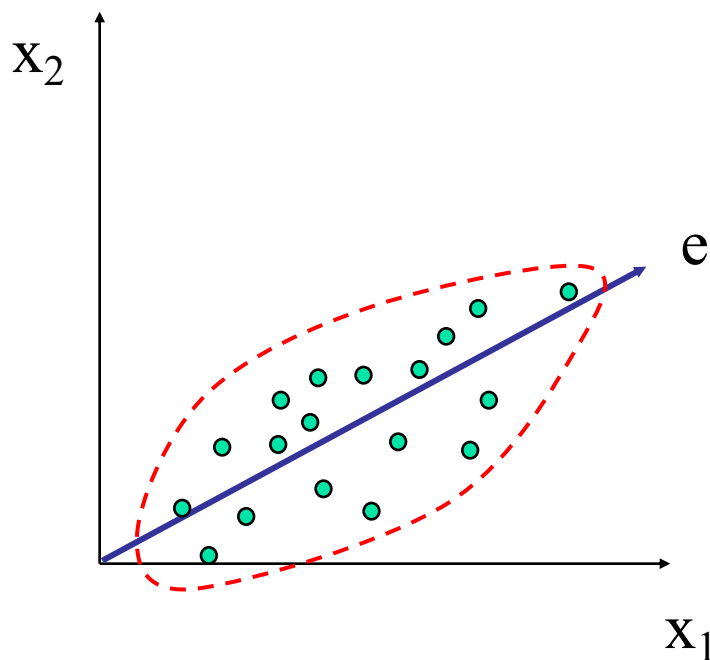
## ■ Techniques

- Principal Component Analysis (PCA)
- Singular Value Decomposition
- Others: supervised and non-linear techniques



# Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data





# Dimensionality Reduction: PCA

256





# Feature Subset Selection

- Another way to reduce dimensionality of data
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is irrelevant to the task of predicting students' GPA





# Feature Subset Selection

## ■ Techniques

### ■ Brute-force approach

- Try all possible feature subsets as input to data mining algorithm

### ■ Embedded approaches

- Feature selection occurs naturally as part of the data mining algorithm

### ■ Filter approaches

- Features are selected before data mining algorithm is run

### ■ Wrapper approaches

- Use the data mining algorithm as a black-box to find best subset of attributes



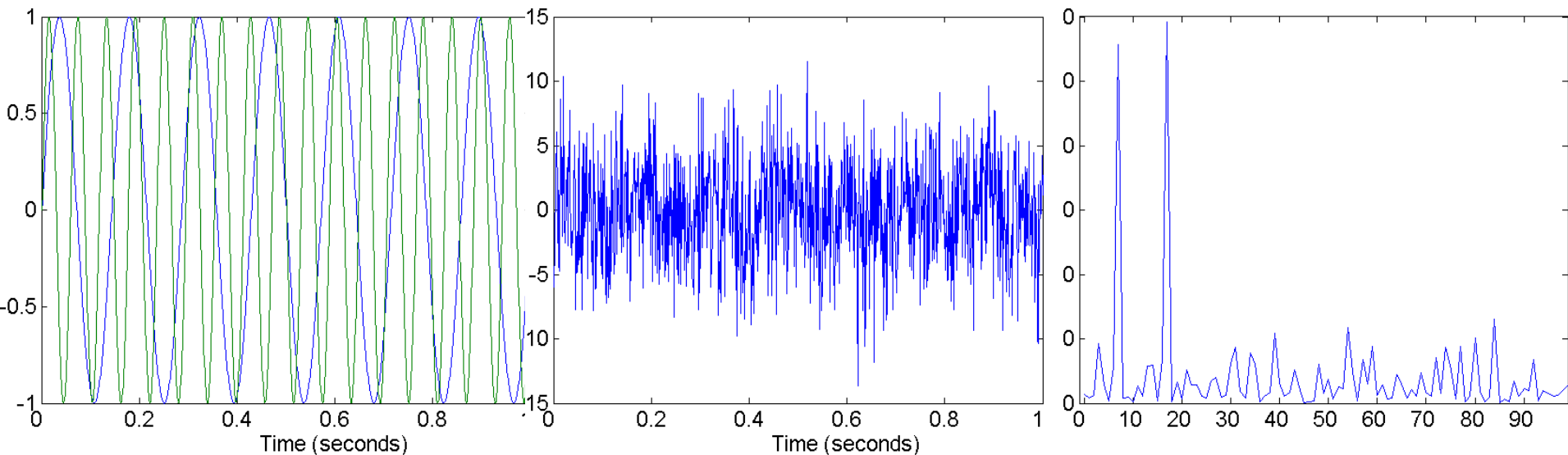
# Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- Three general methodologies
  - Feature Extraction
    - domain-specific
  - Mapping Data to New Space
  - Feature Construction
    - combining features



# Mapping Data to a New Space

- Fourier transform
- Wavelet transform



Two Sine Waves

Two Sine Waves + Noise

Frequency



# Discretization

- **Discretization** is the process of converting a continuous attribute into an ordinal attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is commonly used in classification
    - Many classification algorithms work best if both the independent and dependent variables have only a few values



# Iris Sample Data Set

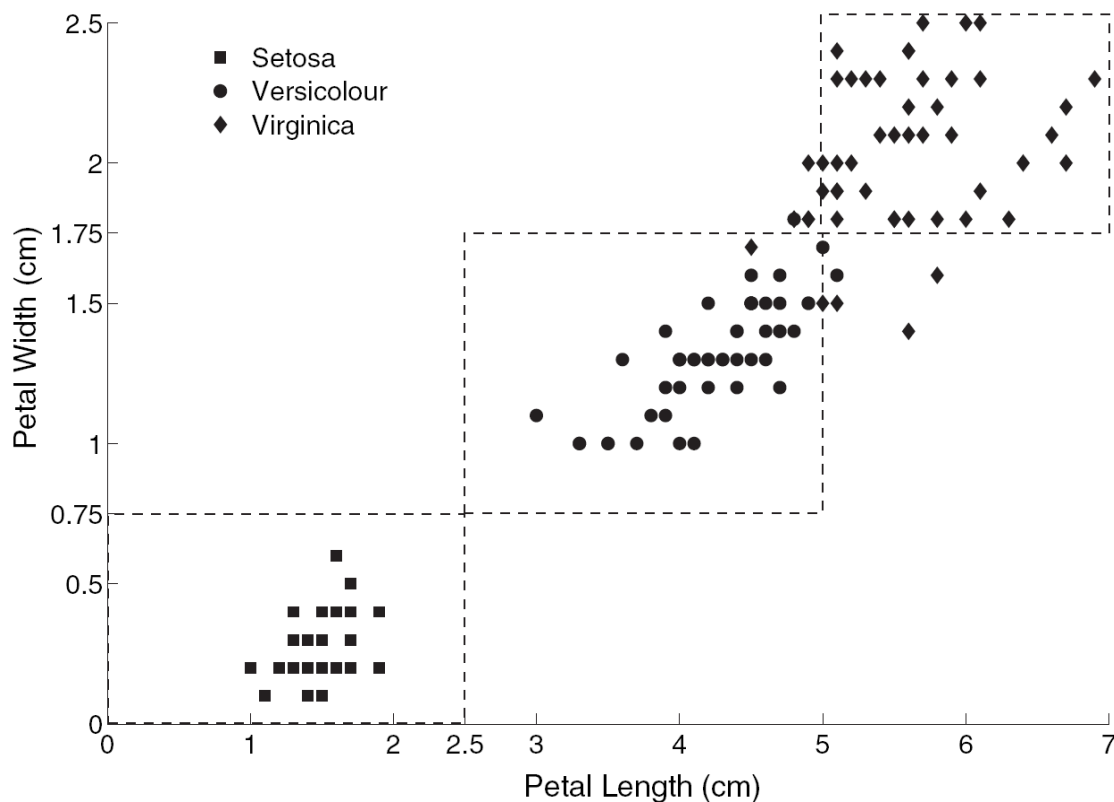
- Iris Plant data set
  - Can be obtained from the UCI Machine Learning Repository  
<http://www.ics.uci.edu/~mlearn/MLRepository.html>
  - From the statistician Douglas Fisher
  - Three flower types (classes)
    - Setosa
    - Versicolour
    - Virginica
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length



Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.



# Discretization: Iris Example



Petal width low or petal length low implies Setosa.

Petal width medium or petal length medium implies Versicolour.

Petal width high or petal length high implies Virginica.

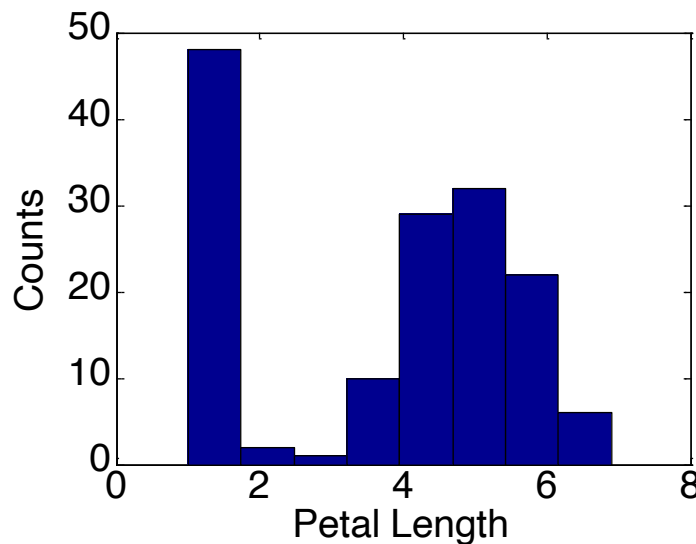


# Discretization: Iris Example ...

- How can we tell what the best discretization is?

- **Unsupervised discretization:** find breaks in the data values

- Example:  
Petal Length



- **Supervised discretization:** Use class labels to find breaks



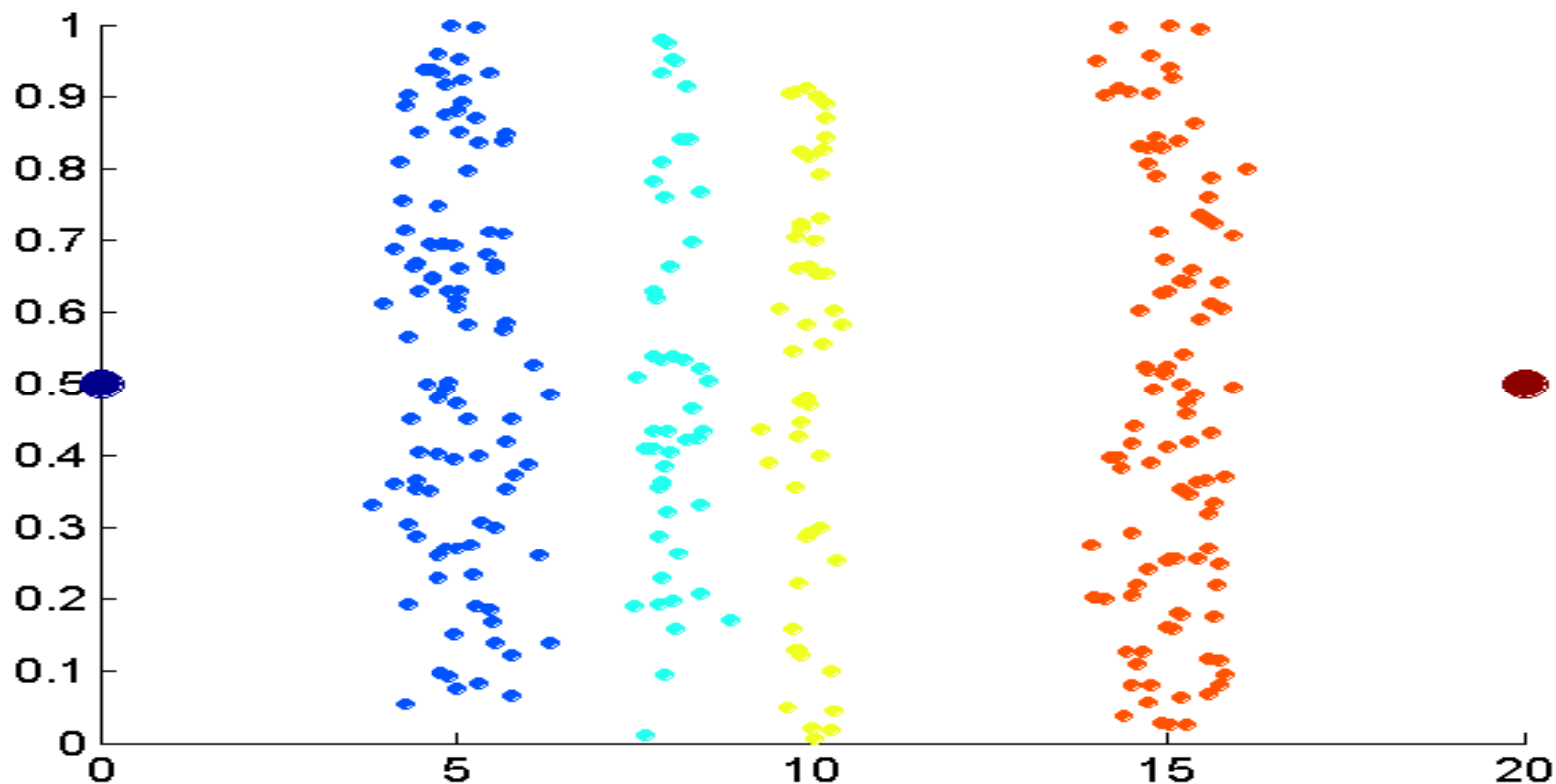
# Discretization

- Examples of **unsupervised discretization** techniques
  - N intervals with the same width  $W = (v_{\max} - v_{\min})/N$ 
    - Easy to implement
    - It can be badly affected by outliers and sparse data
    - Incremental approach
  - N intervals with (approximately) the same cardinality
    - It better fits sparse data and outliers
    - Non incremental approach
  - clustering
    - It fits well sparse data and outliers
  - analysis of data distribution
    - e.g., 4 intervals, one for each quartile





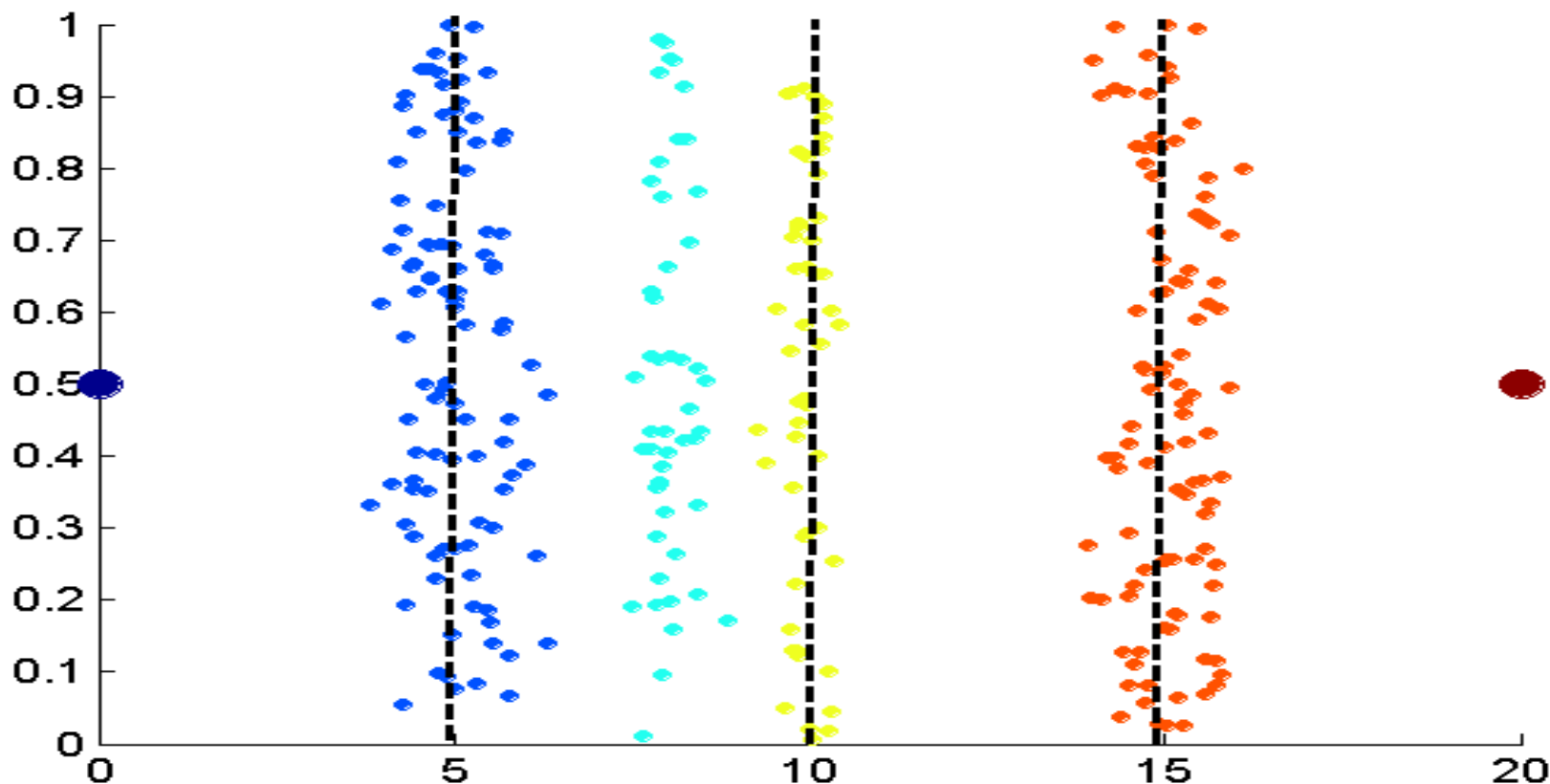
# Example: unsupervised discretization technique



**Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.**



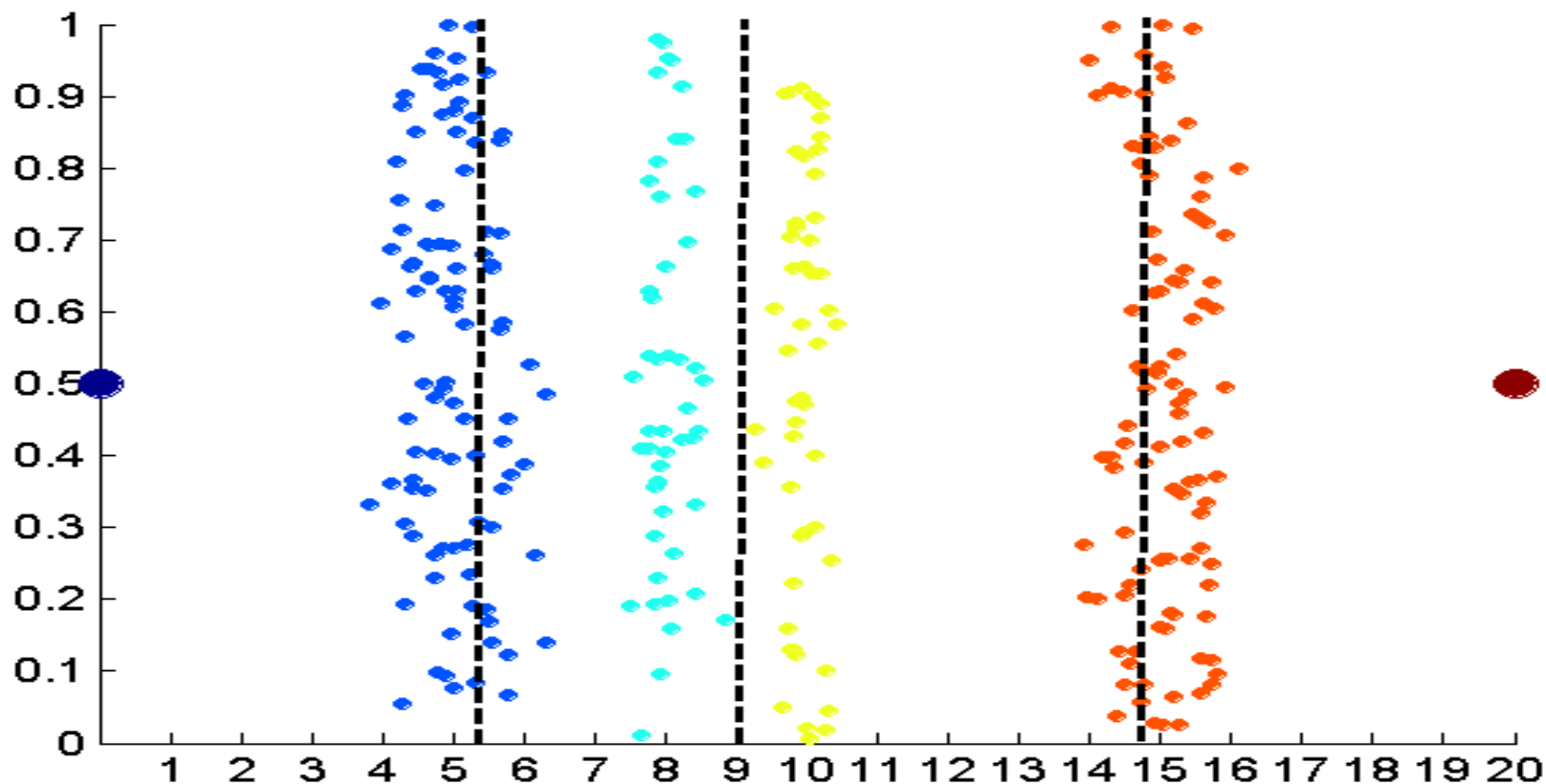
# Example: unsupervised discretization technique



**Equal interval width** approach used to obtain 4 values.



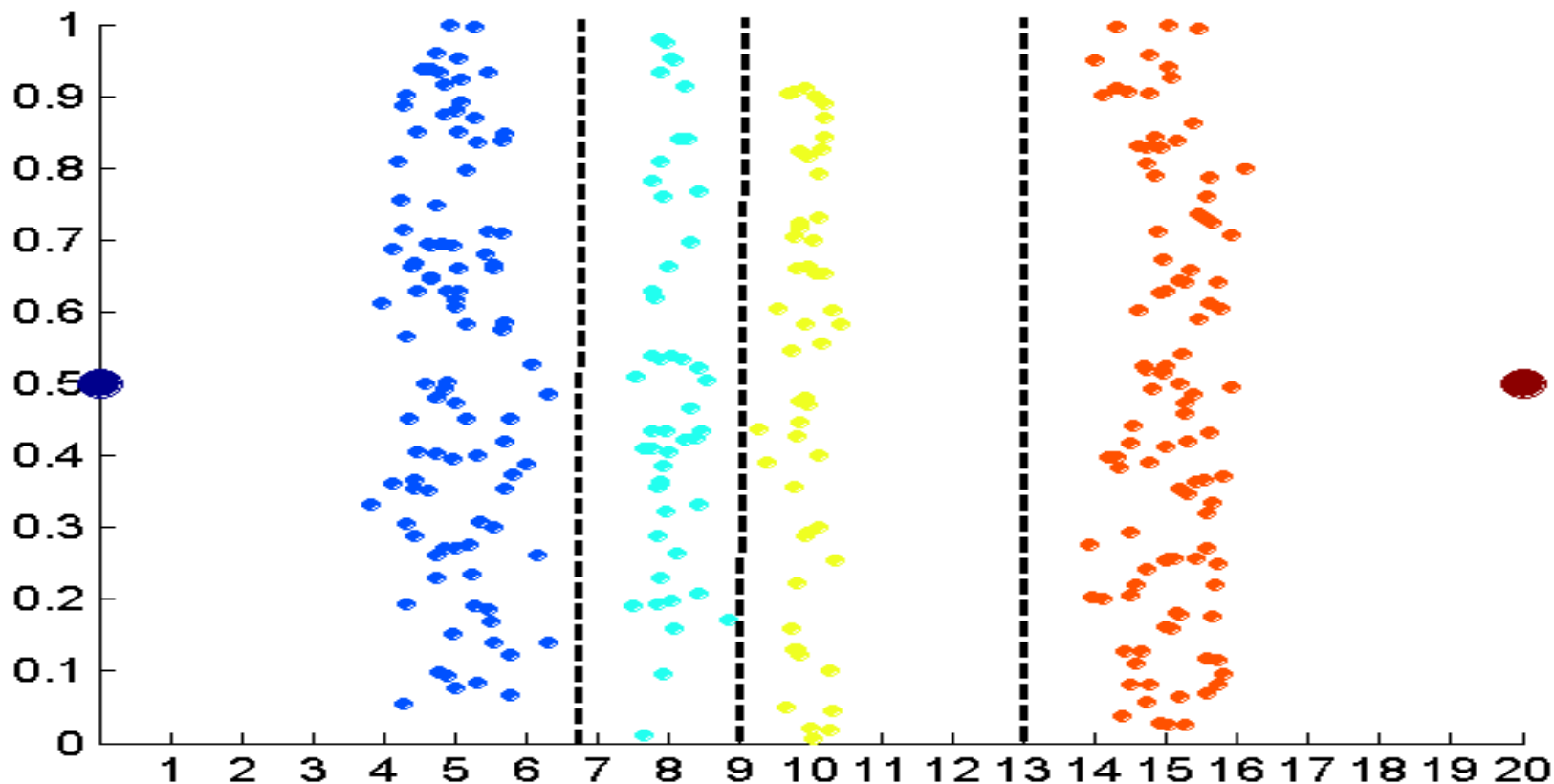
# Example: unsupervised discretization technique



**Equal frequency** approach used to obtain 4 values.



# Example: unsupervised discretization technique



**K-means** approach to obtain 4 values.



# Binarization

- Binarization maps an attribute into one or more binary variables
- **Continuous** attribute: first map the attribute to a categorical one
  - Example: height measured as {low, medium, high}
- **Categorical** attribute
  - Mapping to a set of binary attributes
  - Example: Low, medium, high as 1 0 0, 0 1 0, 0 0 1
  - **One-hot encoding**
    - Only 1 bit takes value 1
    - It represents the specific value taken by the attribute



# Attribute Transformation

- An **attribute transform** is 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
  - Simple functions:  $x^k$ ,  $\log(x)$ ,  $e^x$ ,  $|x|$
- **Normalization**
  - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
  - Take out unwanted, common signal, e.g., seasonality
- In statistics, **standardization** refers to subtracting off the means and dividing by the standard deviation



# Normalization

- It is a type of data transformation
  - The values of an attribute are scaled so as to fall within a small specified range, typically  $[-1,+1]$  or  $[0,+1]$
- Techniques
  - min-max normalization

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

- z-score normalization
- decimal scaling

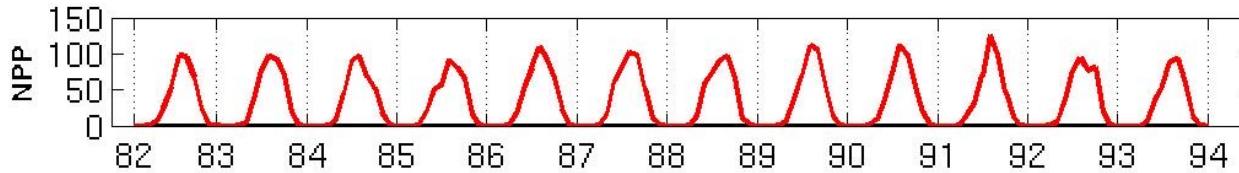
$$v' = \frac{v - \text{mean}_A}{\text{stand\_dev}_A}$$

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

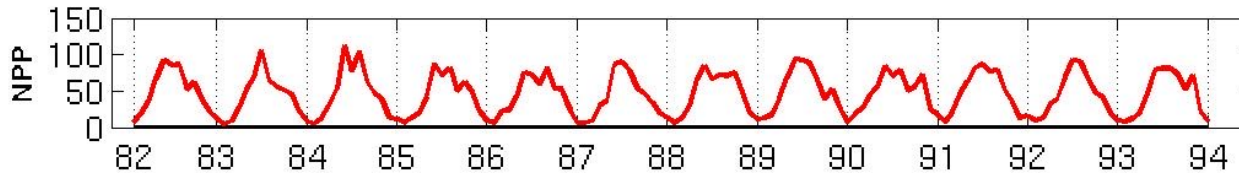


# Example: Sample Time Series of Plant Growth

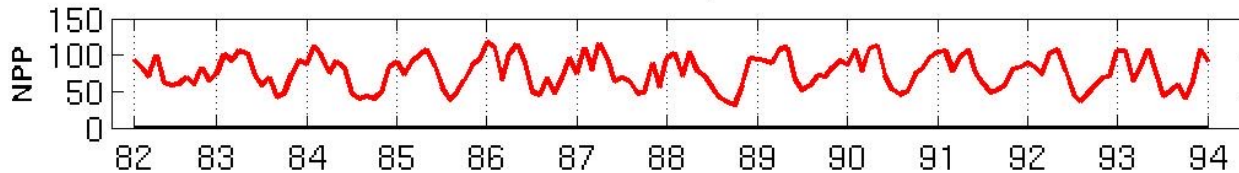
Minneapolis



Atlanta



Sao Paulo, Brazil



**Net Primary Production (NPP)** is a measure of plant growth used by ecosystem scientists.

Correlations between time series

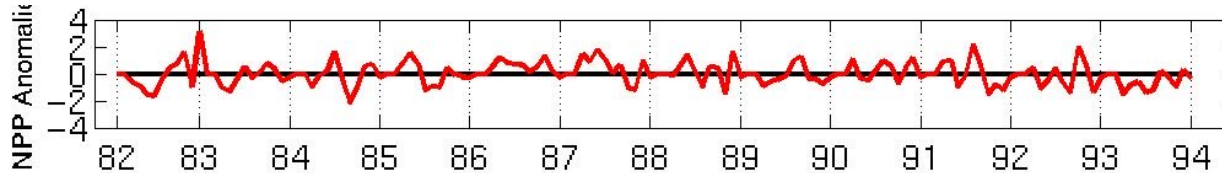
	Minneapolis	Atlanta	Sao Paulo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paulo	-0.7581	-0.5739	1.0000



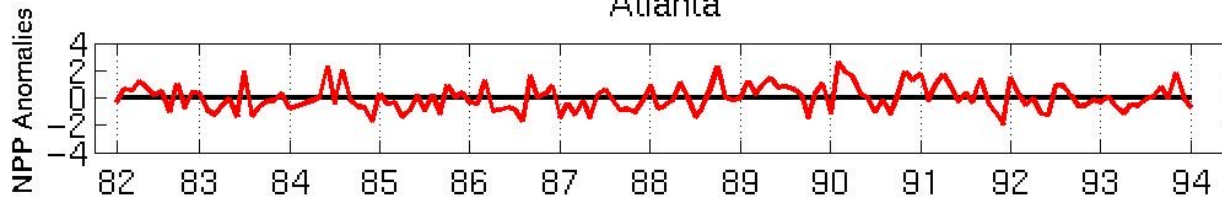


# Example: Sample Time Series of Plant Growth

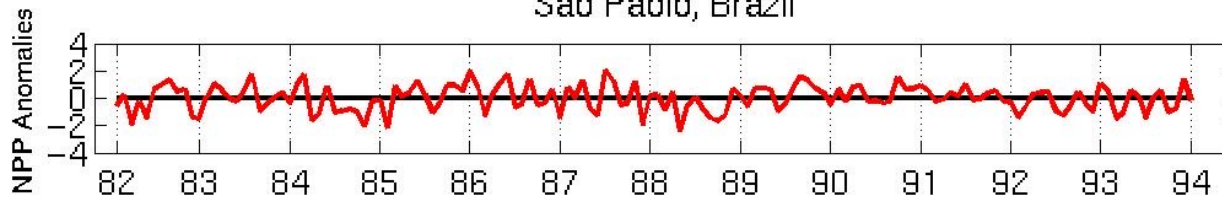
Minneapolis



Atlanta



Sao Paulo, Brazil



Normalized using  
monthly Z Score:  
Subtract off  
monthly mean and  
divide by monthly  
standard deviation

Correlations between time series

	Minneapolis	Atlanta	Sao Paulo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paulo	0.0906	-0.0154	1.0000

# Similarity and dissimilarity



Data Base and Data Mining Group of Politecnico di Torino



# Similarity and Dissimilarity

## ■ Similarity

- Numerical measure of how alike two data objects are
- Is higher when objects are more alike
- Often falls in the range  $[0,1]$

## ■ Dissimilarity

- Numerical measure of how different are two data objects
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

## ■ Proximity refers to a similarity or dissimilarity



# Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects,  $x$  and  $y$ , with respect to a single, simple attribute.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d =  x - y  / (n - 1)$ (values mapped to integers 0 to $n-1$ , where $n$ is the number of values)	$s = 1 - d$
Interval or Ratio	$d =  x - y $	$s = -d, s = \frac{1}{1+d}, s = e^{-d},$ $s = 1 - \frac{d - \min\_d}{\max\_d - \min\_d}$



# Euclidean Distance

- Euclidean Distance

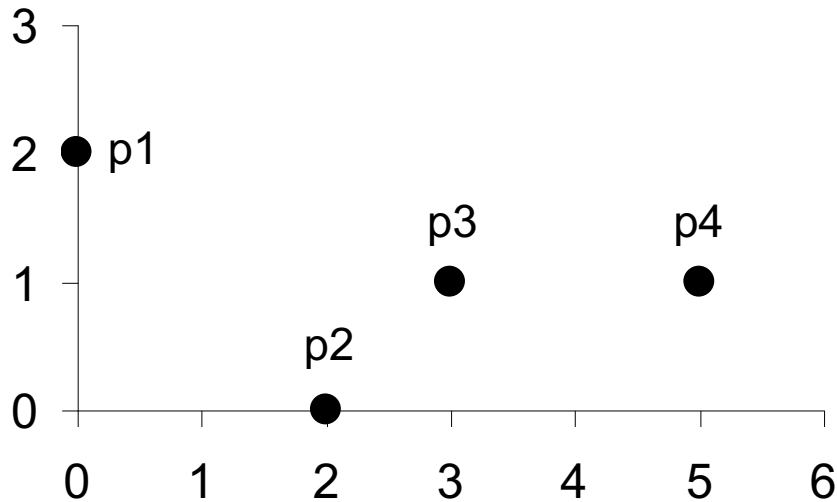
$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

where  $n$  is the number of dimensions (attributes) and  $x_k$  and  $y_k$  are, respectively, the  $k^{th}$  attributes (components) of data objects  $\mathbf{x}$  and  $\mathbf{y}$ .

- Standardization is necessary, if scales differ.



# Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

**Distance Matrix**



# Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left( \sum_{k=1}^n |x_k - y_k|^r \right)^{1/r}$$

Where  $r$  is a parameter,  $n$  is the number of dimensions (attributes) and  $x_k$  and  $y_k$  are, respectively, the  $k^{\text{th}}$  attributes (components) or data objects  $x$  and  $y$ .



# Minkowski Distance: Examples

- $r = 1$ . City block (Manhattan, taxicab,  $L_1$  norm) distance
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$ . Euclidean distance
- $r \rightarrow \infty$ . “supremum” ( $L_{\max}$  norm,  $L_{\infty}$  norm) distance
  - This is the maximum difference between any component of the vectors
- Do not confuse  $r$  with  $n$ , i.e., all these distances are defined for all numbers of dimensions.





# Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

$L_{\infty}$	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix



# Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well-known properties.
  1.  $d(\mathbf{x}, \mathbf{y}) \geq 0$  for all  $\mathbf{x}$  and  $\mathbf{y}$  and  $d(\mathbf{x}, \mathbf{y}) = 0$  only if  $\mathbf{x} = \mathbf{y}$ . (Positive definiteness)
  2.  $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$  for all  $\mathbf{x}$  and  $\mathbf{y}$ . (Symmetry)
  3.  $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$  for all points  $\mathbf{x}$ ,  $\mathbf{y}$ , and  $\mathbf{z}$ . (Triangle Inequality)

where  $d(\mathbf{x}, \mathbf{y})$  is the distance (dissimilarity) between points (data objects)  $\mathbf{x}$  and  $\mathbf{y}$ .

- A distance that satisfies these properties is a **metric**



# Common Properties of a Similarity

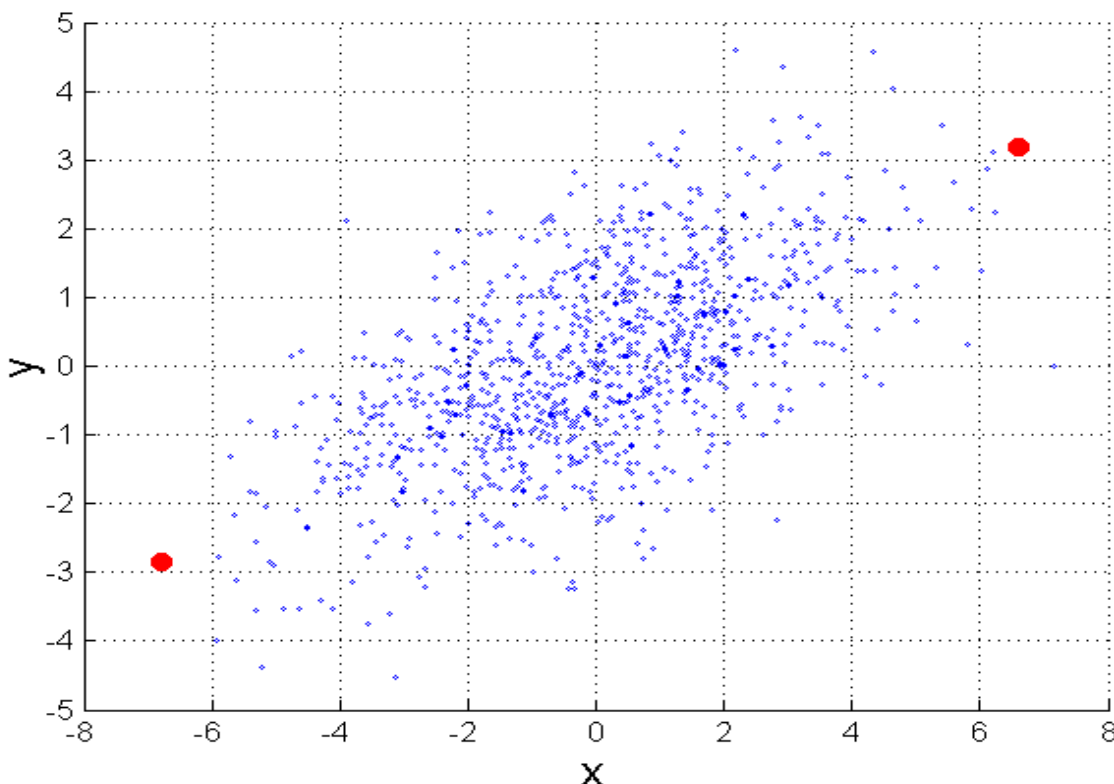
- Similarities also have some well known properties
  1.  $s(\mathbf{x}, \mathbf{y}) = 1$  (or maximum similarity) only if  $\mathbf{x} = \mathbf{y}$
  2.  $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$  for all  $\mathbf{x}$  and  $\mathbf{y}$  (symmetry)

where  $s(\mathbf{x}, \mathbf{y})$  is the similarity between points (data objects)  $\mathbf{x}$  and  $\mathbf{y}$



# Mahalanobis Distance

$$\text{mahalanobis}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y})$$

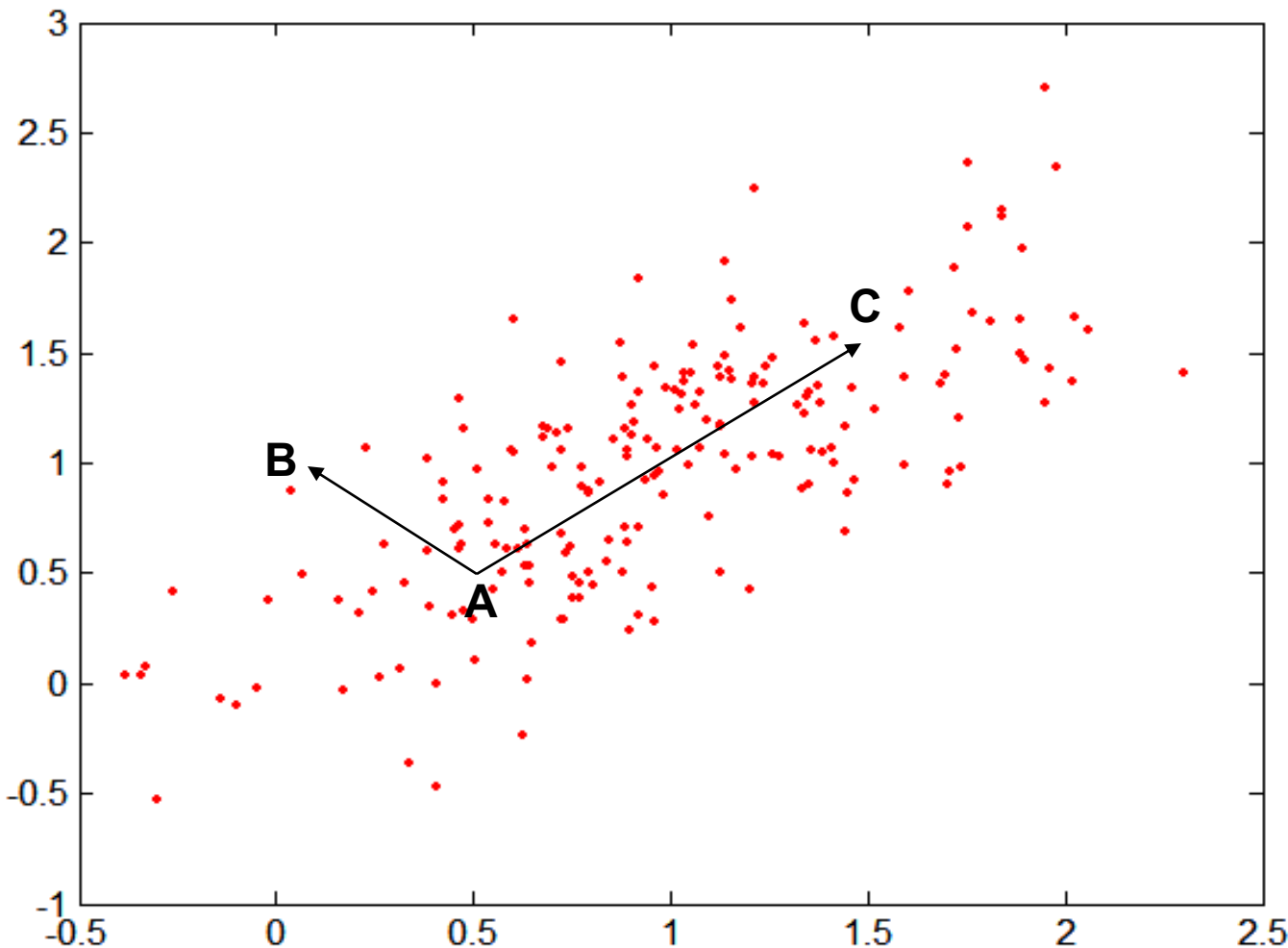


$\Sigma$  is the covariance matrix

**For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.**



# Mahalanobis Distance



**Covariance  
Matrix:**

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

**A: (0.5, 0.5)**

**B: (0, 1)**

**C: (1.5, 1.5)**

**Mahal(A,B) = 5**

**Mahal(A,C) = 4**



# Similarity Between Binary Vectors

- Common situation is that objects  $p$  and  $q$  have only binary attributes
- Compute similarities using the following quantities
  - $M_{01}$  = the number of attributes where  $p$  was 0 and  $q$  was 1
  - $M_{10}$  = the number of attributes where  $p$  was 1 and  $q$  was 0
  - $M_{00}$  = the number of attributes where  $p$  was 0 and  $q$  was 0
  - $M_{11}$  = the number of attributes where  $p$  was 1 and  $q$  was 1
- Simple Matching and Jaccard Coefficients
  - SMC = number of matches / number of attributes  
=  $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$
  - J = number of 11 matches / number of not-both-zero attributes values  
=  $(M_{11}) / (M_{01} + M_{10} + M_{11})$



# SMC versus Jaccard: Example

$$p = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$$

$$q = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$$

$M_{01} = 2$  (the number of attributes where p was 0 and q was 1)

$M_{10} = 1$  (the number of attributes where p was 1 and q was 0)

$M_{00} = 7$  (the number of attributes where p was 0 and q was 0)

$M_{11} = 0$  (the number of attributes where p was 1 and q was 1)

$$SMC = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$$

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$$



# Cosine Similarity

- If  $d_1$  and  $d_2$  are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| ||d_2|| ,$$

where  $\bullet$  indicates vector dot product and  $||d||$  is the norm of vector  $d$

- Example:

$$d_1 = \mathbf{3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0}$$

$$d_2 = \mathbf{1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2}$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$||d_1|| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$||d_2|| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150$$





# General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.
- 1: For the  $k^{\text{th}}$  attribute, compute a similarity,  $s_k(\mathbf{x}, \mathbf{y})$ , in the range  $[0, 1]$ .
- 2: Define an indicator variable,  $\delta_k$ , for the  $k^{\text{th}}$  attribute as follows:
  - $\delta_k = 0$  if the  $k^{\text{th}}$  attribute is an asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing value for the  $k^{\text{th}}$  attribute
  - $\delta_k = 1$  otherwise
- 3. Compute 
$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \delta_k}$$



# Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use non-negative weights  $\omega_k$

- $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \omega_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \omega_k \delta_k}$

- Can also define a weighted form of distance

$$d(\mathbf{x}, \mathbf{y}) = \left( \sum_{k=1}^n w_k |x_k - y_k|^r \right)^{1/r}$$

# Correlation



Data Base and Data Mining Group of Politecnico di Torino



# Data correlation

- Measure of the linear relationship between two data objects
  - having binary or continuous variables
- Useful during the data exploration phase
  - To be better aware of data properties
- Analysis of feature correlation
  - Correlated features should be removed
    - simplifying the next analytics steps
    - improving the performance of the data-driven algorithms



# Pearson's correlation

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x s_y},$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})$$

$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}$$

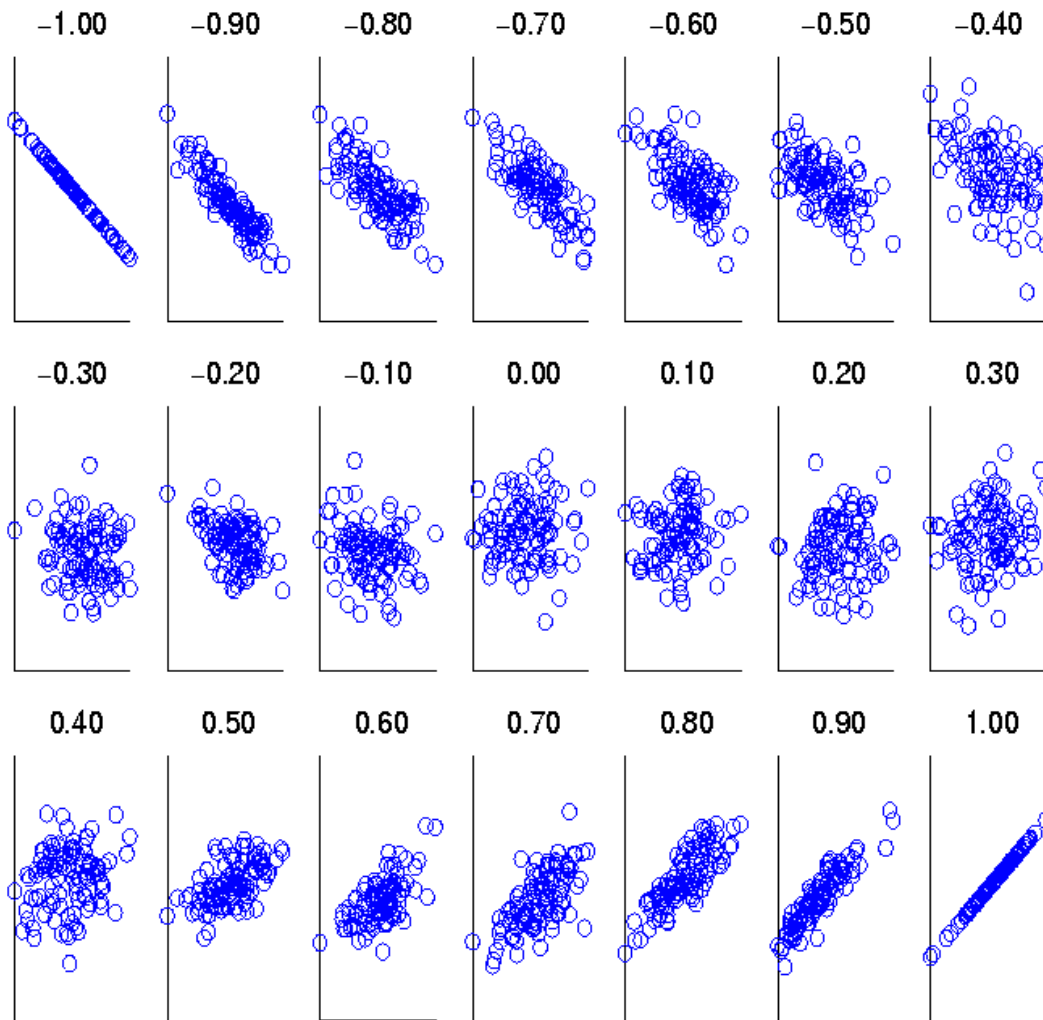
$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k \text{ is the mean of } \mathbf{x}$$

$$\bar{y} = \frac{1}{n} \sum_{k=1}^n y_k \text{ is the mean of } \mathbf{y}$$



# Visually Evaluating Correlation



Scatter plots showing the similarity from  $-1$  to  $1$ .

Perfect linear correlation when value is  $1$  or  $-1$



# Drawback of Correlation

- $\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$

- $\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$

$$y_i = x_i^2$$

- $\text{mean}(\mathbf{x}) = 0, \text{mean}(\mathbf{y}) = 4$

- $\text{std}(\mathbf{x}) = 2.16, \text{std}(\mathbf{y}) = 3.74$

$$\begin{aligned} \text{corr} &= (-3)(5) + (-2)(0) + (-1)(-3) + (0)(-4) + (1)(-3) + (2)(0) + 3(5) / (6 * 2.16 * 3.74) \\ &= 0 \end{aligned}$$