

# Outline

---

- Attributes and Objects
- Types of Data
- Data Quality
- Similarity and Distance

# 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 describe an **object**
  - Object is also known as record, point, case, sample, entity, or instance

**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

**Objects**

# 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 can be different than the properties of the values used to represent the attribute

# Attribute Types

---

- **Nominal:** categories, states, or “names of things”
  - *Hair\_color* = {*auburn*, *black*, *blond*, *brown*, *grey*, *red*, *white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - ◆ e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - ◆ e.g., medical test (positive vs. negative)
    - ◆ Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small*, *medium*, *large*}, grades, army rankings

# Numeric Attribute Types

---

## □ Interval

- ◆ Measured on a scale of **equal-sized units**
- ◆ Values have order
  - E.g., *temperature in  $C^\circ$  or  $F^\circ$ , calendar dates*
- ◆ No true zero-point

## □ Ratio

- ◆ Inherent **zero-point**
- ◆ We can speak of values as being an order of magnitude larger than the unit of measurement (10  $K^\circ$  is twice as high as 5  $K^\circ$ ).
  - e.g. *length, counts, monetary quantities*

---

Differences between measurements, true zero exists

**Ratio Data**

Quantitative Data

Differences between measurements but no true zero

**Interval Data**

Ordered Categories (rankings, order, or scaling)

**Ordinal Data**

Qualitative Data

Categories (no ordering or direction)

**Nominal Data**

<https://www.graphpad.com/support/faq/what-is-the-difference-between-ordinal-interval-and-ratio-variables-why-should-i-care/>

Features	Interval scale	Ratio scale
<b>Variable property</b>	All variables measured in an interval scale can be added, subtracted, and multiplied. You cannot calculate a ratio between them.	Ratio scale has all the characteristics of an interval scale, in addition, to be able to calculate ratios. That is, you can leverage numbers on the scale against 0.
<b>Absolute Point Zero</b>	Zero-point in an interval scale is arbitrary. For example, the temperature can be below 0 degrees Celsius and into negative temperatures.	The ratio scale has an absolute zero or character of origin. Height and weight cannot be zero or below zero.
<b>Calculation</b>	Statistically, in an interval scale, the arithmetic mean is calculated.	Statistically, in a ratio scale, the geometric or harmonic mean is calculated.
<b>Measurement</b>	Interval scale can measure size and magnitude as multiple factors of a defined unit.	Ratio scale can measure size and magnitude as a factor of one defined unit in terms of another.
<b>Example</b>	A classic example of an interval scale is the temperature in Celsius. The difference in temperature between 50 degrees and 60 degrees is 10 degrees; this is the same difference between 70 degrees and 80 degrees.	Classic examples of a ratio scale are any variable that possesses an absolute zero characteristic, like age, weight, height, or sales figures.

# 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.



# Basic Statistical Descriptions of Data

---

## □ Motivation

- To better understand the data: central tendency, variation and spread

## □ Data dispersion characteristics

- median, max, min, quantiles, outliers, variance, etc.

## □ Numerical dimensions correspond to sorted intervals

- Data dispersion: analyzed with multiple granularities of precision
- Boxplot or quantile analysis on sorted intervals

## □ Dispersion analysis on computed measures

- Folding measures into numerical dimensions
- Boxplot or quantile analysis on the transformed cube

# Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

Note:  $n$  is sample size and  $N$  is population size.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \mu = \frac{\sum x}{N}$$

- Weighted arithmetic mean:
- Trimmed mean: chopping extreme values

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

- Median:

- Middle value if odd number of values, or average of the middle two values otherwise
- Estimated by interpolation (for *grouped data*):

$$median = L_1 + \left( \frac{n/2 - (\sum freq)l}{freq_{median}} \right) width$$

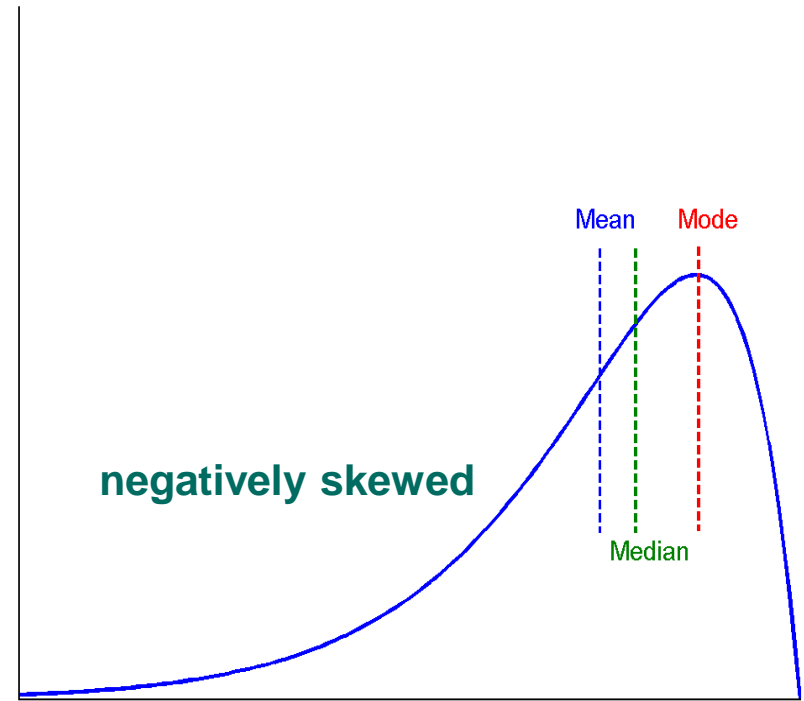
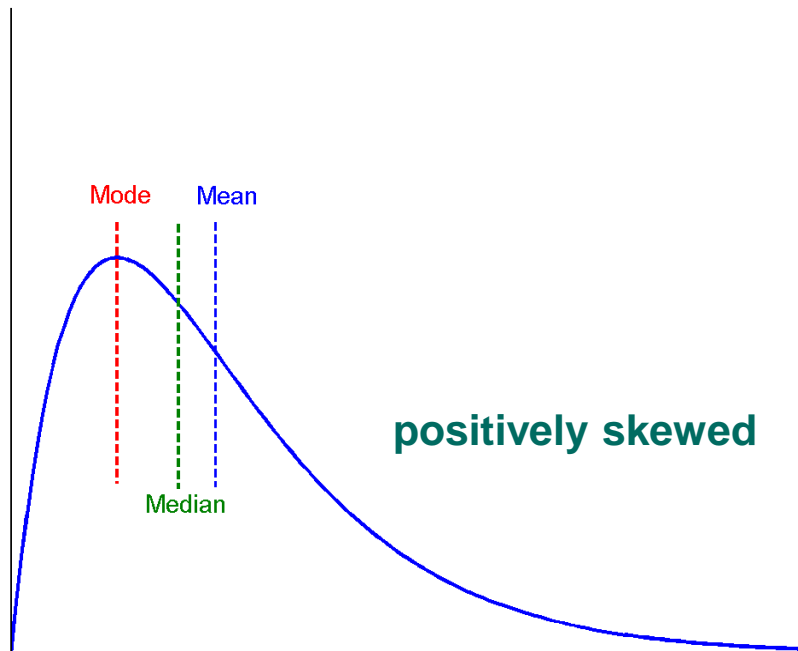
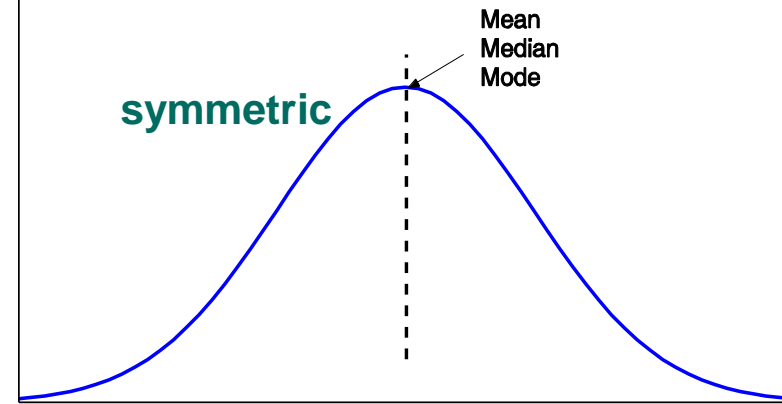
<i>age</i>	<i>frequency</i>
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

- Mode

- Value that occurs most frequently in the data

# Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



# Measuring the Dispersion of Data

## □ Quartiles, outliers and boxplots

- **Quartiles:**  $Q_1$  (25<sup>th</sup> percentile of data below this point),  $Q_3$  (75<sup>th</sup> percentile)
- **Inter-quartile range:**  $IQR = Q_3 - Q_1$
- **Five number summary:** min,  $Q_1$ , median,  $Q_3$ , max
- **Boxplot:** ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
- **Outlier:** usually, a value higher/lower than  $1.5 \times IQR$

## □ Variance and standard deviation (*sample: $s$ , population: $\sigma$* )

- **Variance:** (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right] \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

- **Standard deviation**  $s$  (*or*  $\sigma$ ) is the square root of variance  $s^2$  (*or*  $\sigma^2$ )

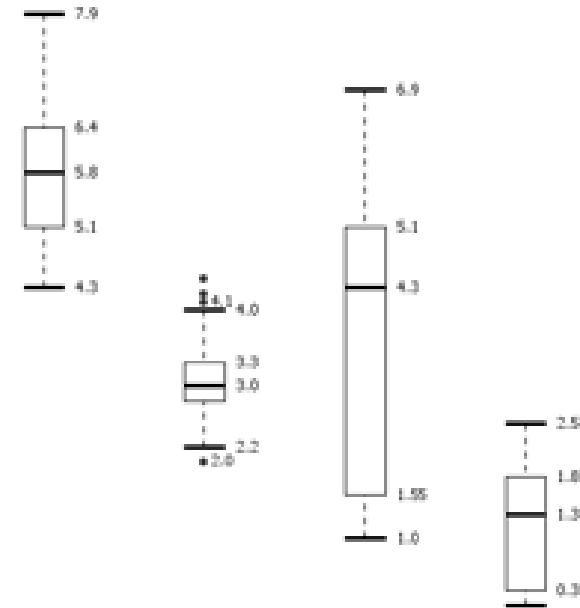
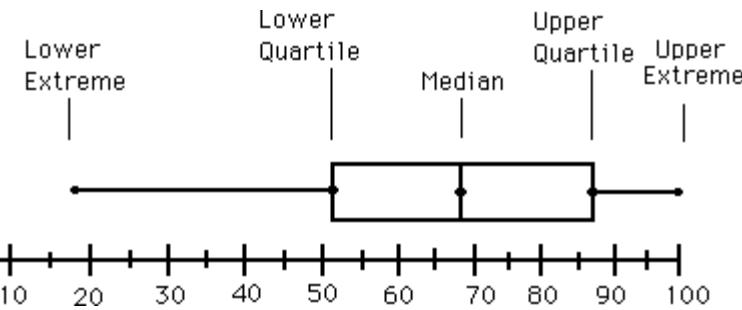
# Boxplot Analysis

## □ Five-number summary of a distribution

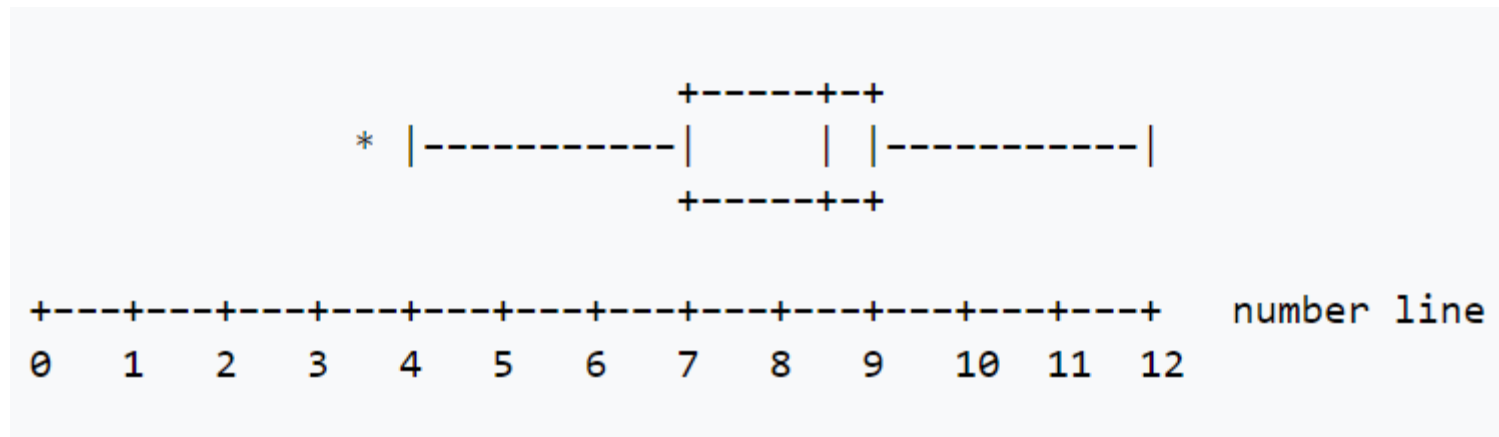
- Minimum, Q1, Median, Q3, Maximum

## □ Boxplot

- Data is represented with a box
- The **ends of the box are at the first and third quartiles**, i.e., the height of the box is IQR
- The median is marked by a line within the box
- Outliers: points beyond a specified outlier threshold, plotted individually.



# Example



For the data set in this box plot:

- lower (first) quartile  $Q_1 = 7$
- median (second quartile)  $Q_2 = 8.5$
- upper (third) quartile  $Q_3 = 9$
- interquartile range,  $IQR = Q_3 - Q_1 = 2$
- lower  $1.5 \times IQR$  whisker =  $Q_1 - 1.5 \times IQR = 7 - 3 = 4$ . (If there is no data point at 4, then the lowest point greater than 4.)
- upper  $1.5 \times IQR$  whisker =  $Q_3 + 1.5 \times IQR = 9 + 3 = 12$ . (If there is no data point at 12, then the highest point less than 12.)

# Example

i	x[i]	Median	Quartile
1	7	$Q_2 = 87$ (median of whole table)	$Q_1 = 31$ (median of upper half, from row 1 to 6)
2	7		
3	31		
4	31		
5	47		
6	75		
7	87		
8	115		$Q_3 = 119$ (median of lower half, from row 8 to 13)
9	116		
10	119		
11	119		
12	155		
13	177		

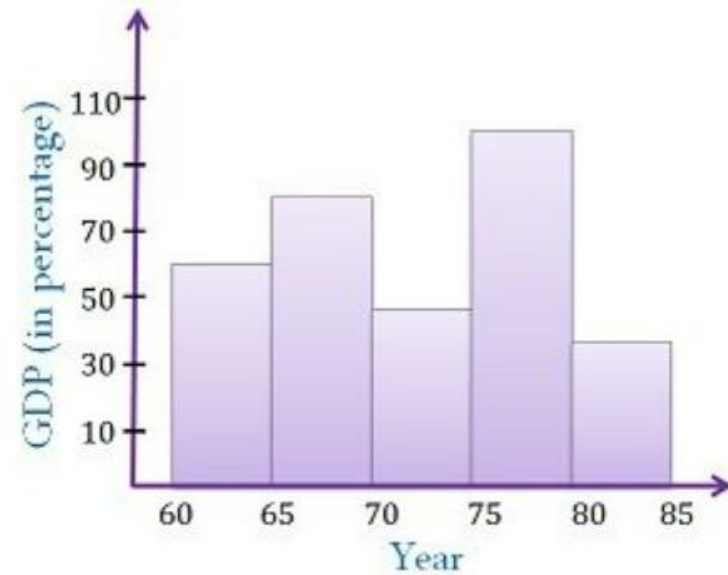
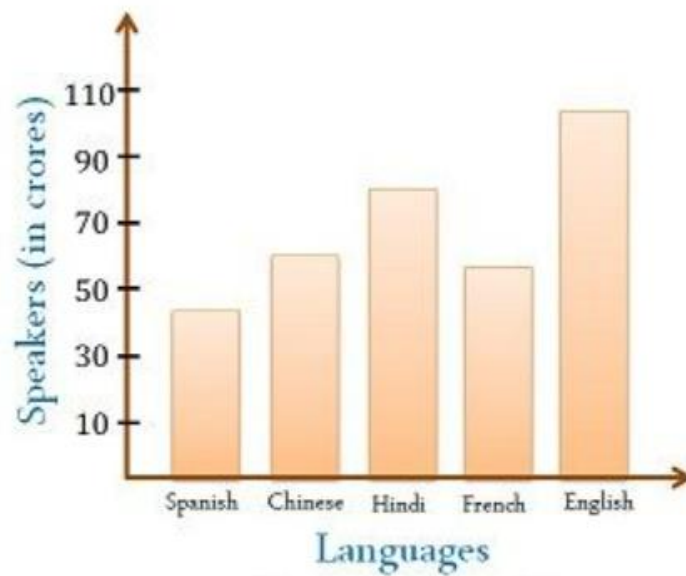
For the data in this table the interquartile range is  $IQR = Q_3 - Q_1 = 119 - 31 = 88$ .

# Graphic Displays of Basic Statistical Descriptions

---

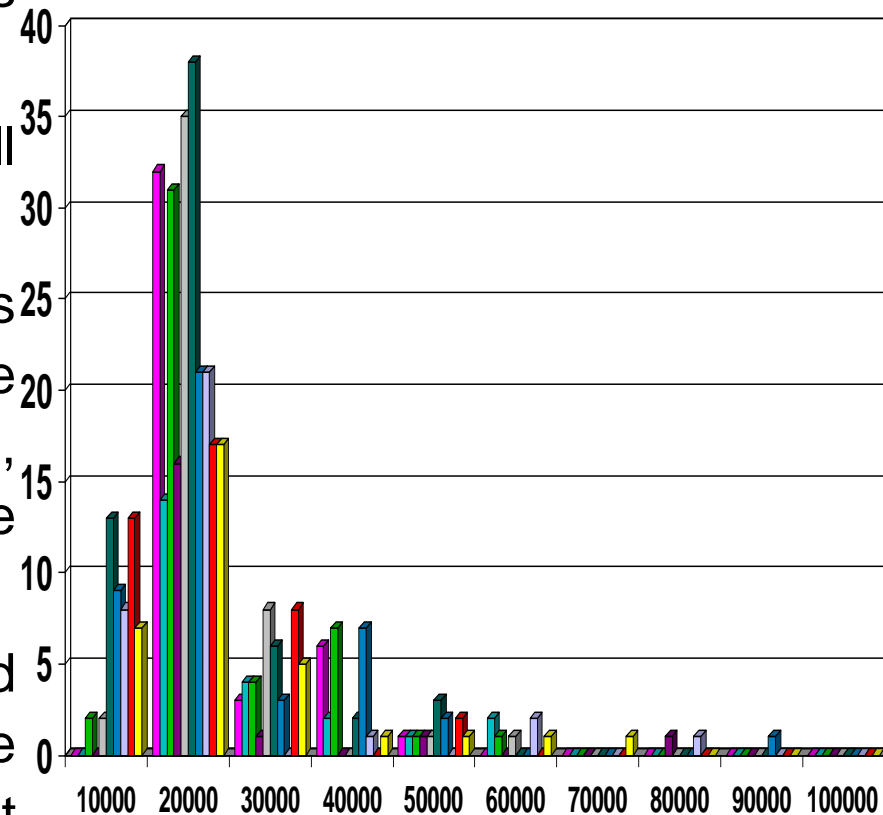
- **Boxplot:** graphic display of five-number summary
- **Histogram:** x-axis are values, y-axis repres. frequencies
- **Quantile plot:** each value  $x_i$  is paired with  $f_i$  indicating that approximately  $100 f_i\%$  of data are  $\leq x_i$
- **Quantile-quantile (q-q) plot:** graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot:** each pair of values is a pair of coordinates and plotted as points in the plane





# Histogram Analysis

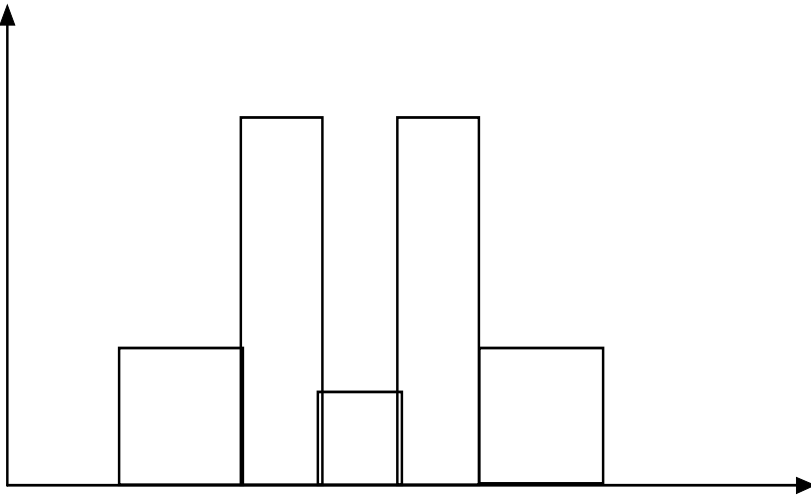
- **Histogram:** Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



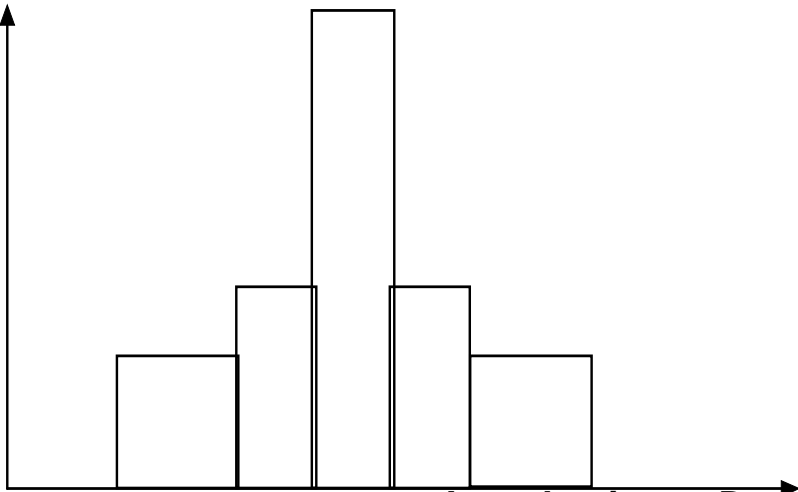
# Histogram vs. Bar Graph

Histogram	Bar Graph
The histogram is a term that refers to a graphical representation that shows data by way of bars to display the frequency of numerical data.	The bar graph is a graphical representation of data that uses bars to compare different categories of data.
Distribution of non-discrete variables.	Comparison of discrete variables.
Bars touch each other, so there are no spaces between bars.	Bars never touch each other, so there are spaces between bars.
In this type of graph, elements are grouped so that they are considered as ranges.	In this type of graph, elements are taken as individual entities.
Histogram width may vary.	The bar chart is mostly of equal width.
To display the frequency of occurrences.	To compare different categories of data.
In Histogram, the data points are grouped and rendered based on its bin value.	In the Bar graph, each data point is rendered as a separate bar.

# Histograms Often Tell More than Boxplots

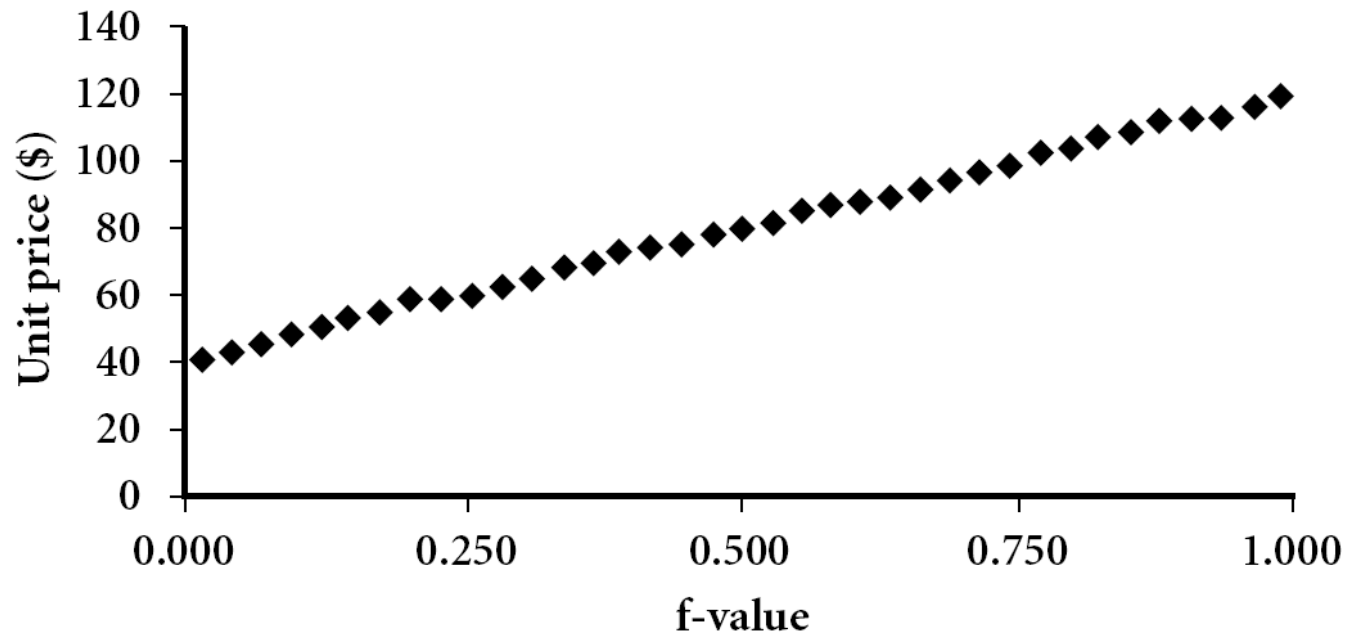


- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions



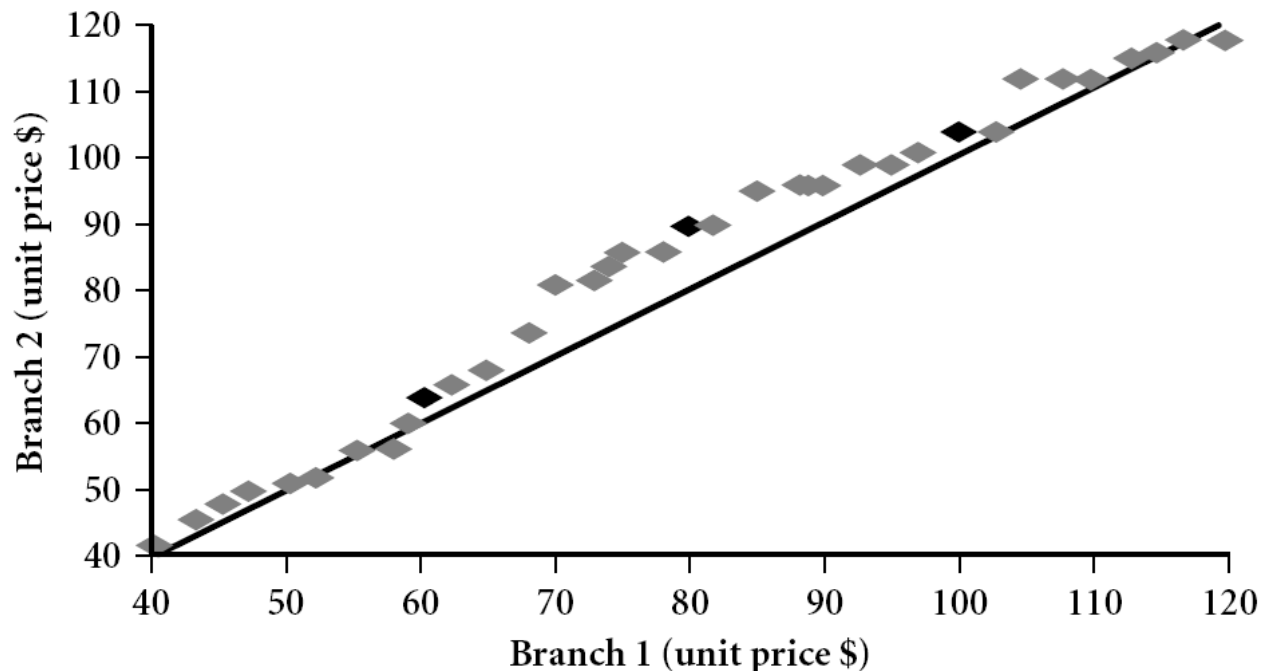
# Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data  $x_i$  data sorted in increasing order,  $f_i$  indicates that approximately 100  $f_i$ % of the data are below or equal to the value  $x_i$



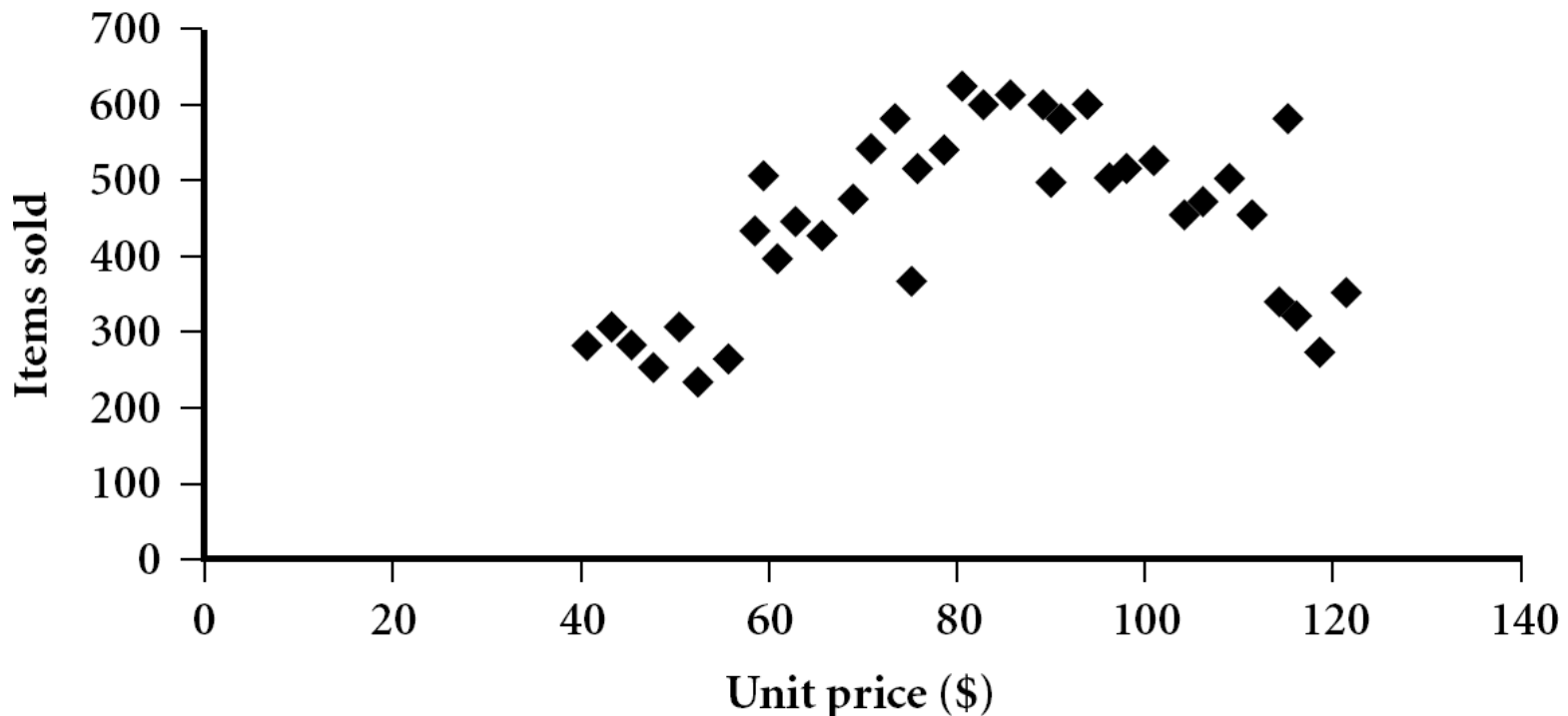
# Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- **View: Is there is a shift in going from one distribution to another?**
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



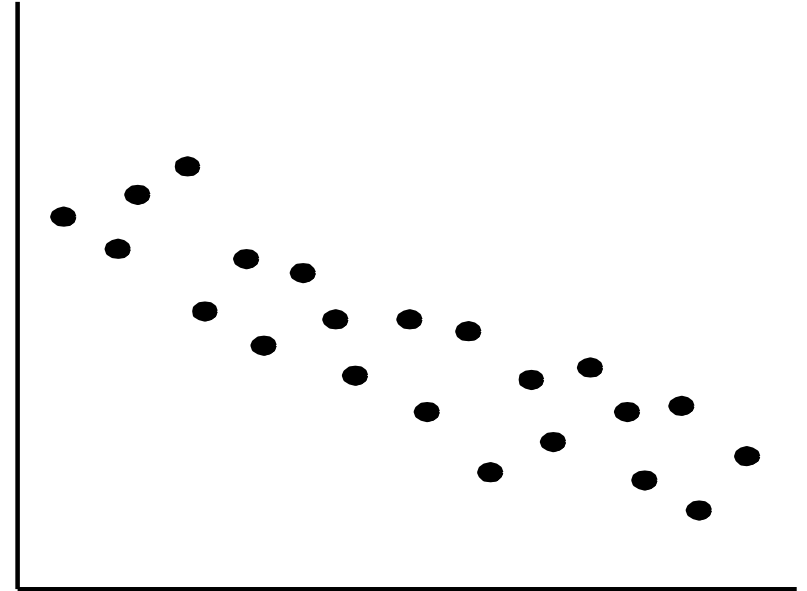
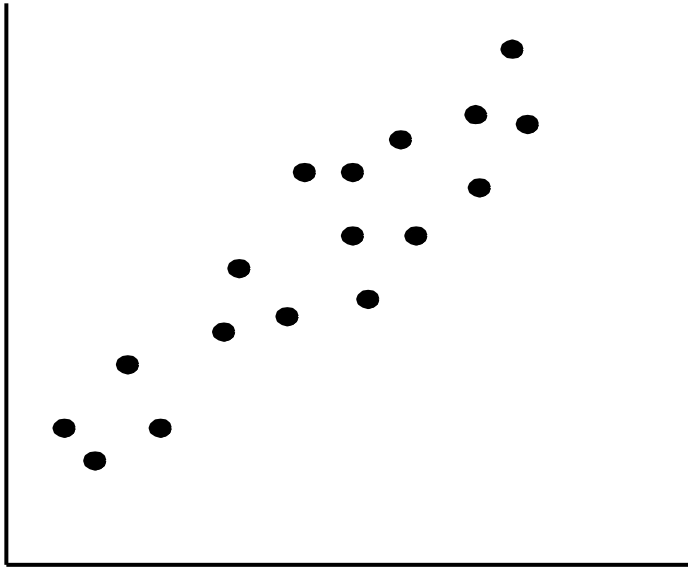
# Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



# Positively and Negatively Correlated Data

---



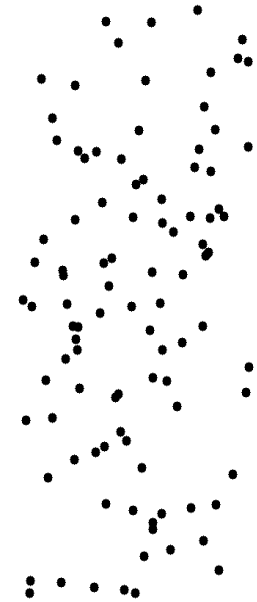
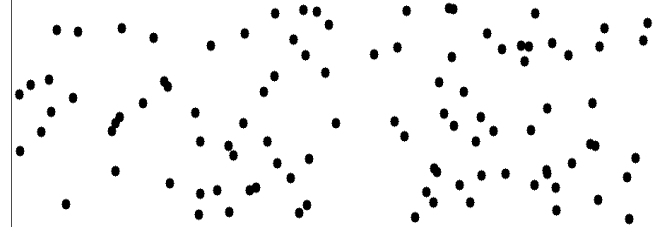
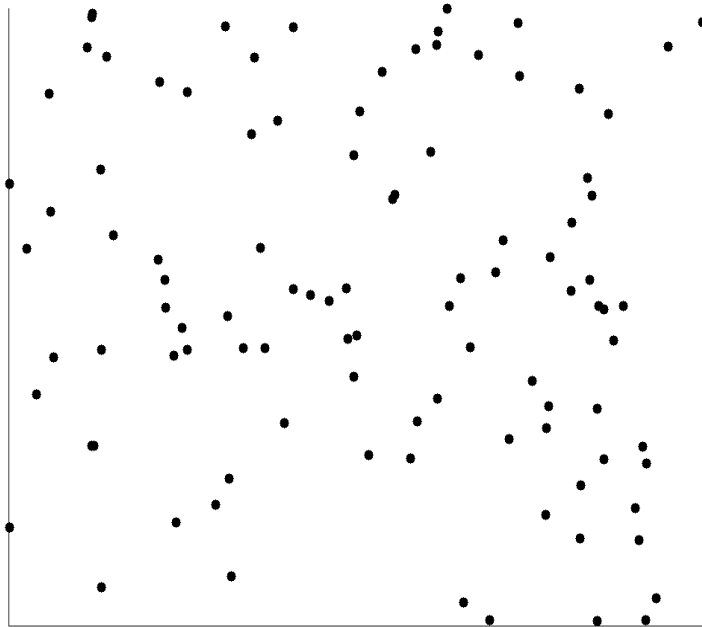
- The left half fragment is positively correlated
- The right half is negative correlated



# Uncorrelated Data

---

---



# Important Characteristics of Data

---

- Dimensionality (number of attributes)
  - ◆ High dimensional data brings a number of challenges
- Sparsity
  - ◆ Only presence counts
- Resolution
  - ◆ Patterns depend on the scale
- Size
  - ◆ Type of analysis may depend on size of data

# Types of data sets

---

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

# Record Data

- Data that consists of a collection of records, each of which consists of 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

# Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such a data set can be represented by an  $m$  by  $n$  matrix, where there are  $m$  rows, one for each object, and  $n$  columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

# 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

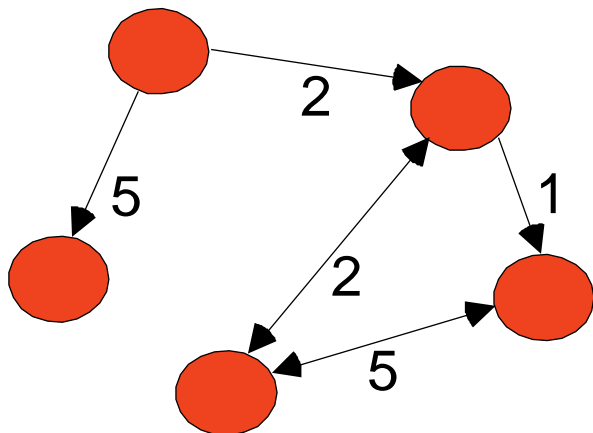
# Transaction Data

- A special type of data, where
  - Each 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.
  - Can represent transaction data as record data

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

# Graph Data

- Examples: Generic graph, a molecule, and webpages



## 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

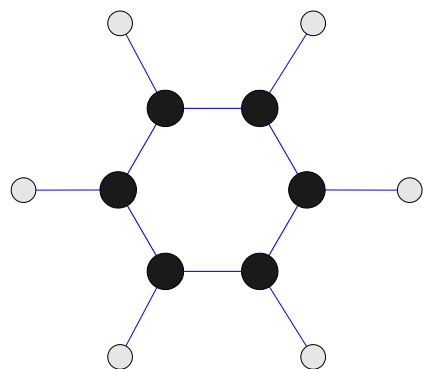
Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

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.



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

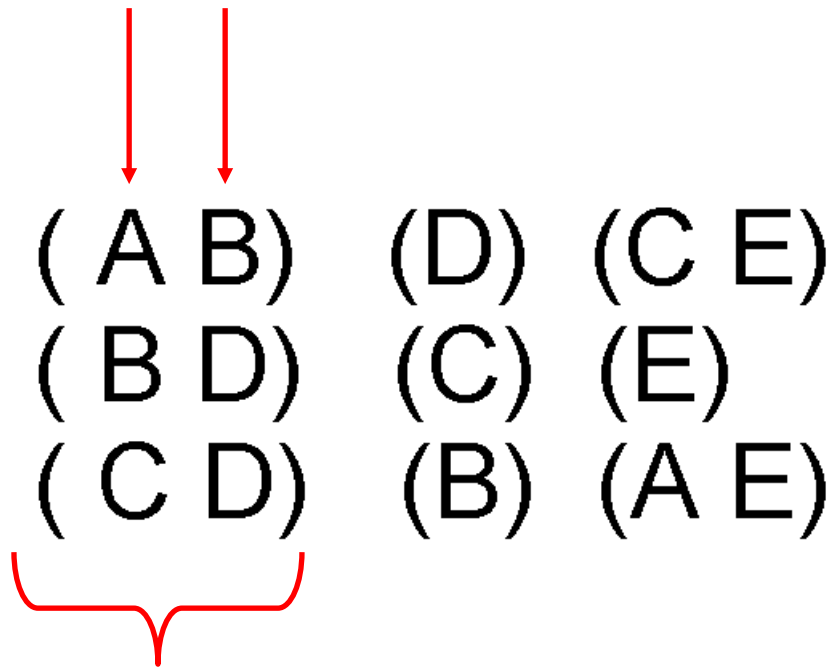


# Ordered Data

---

## □ Sequences of transactions

**Items/Events**



**An element of  
the sequence**

# Ordered Data

---

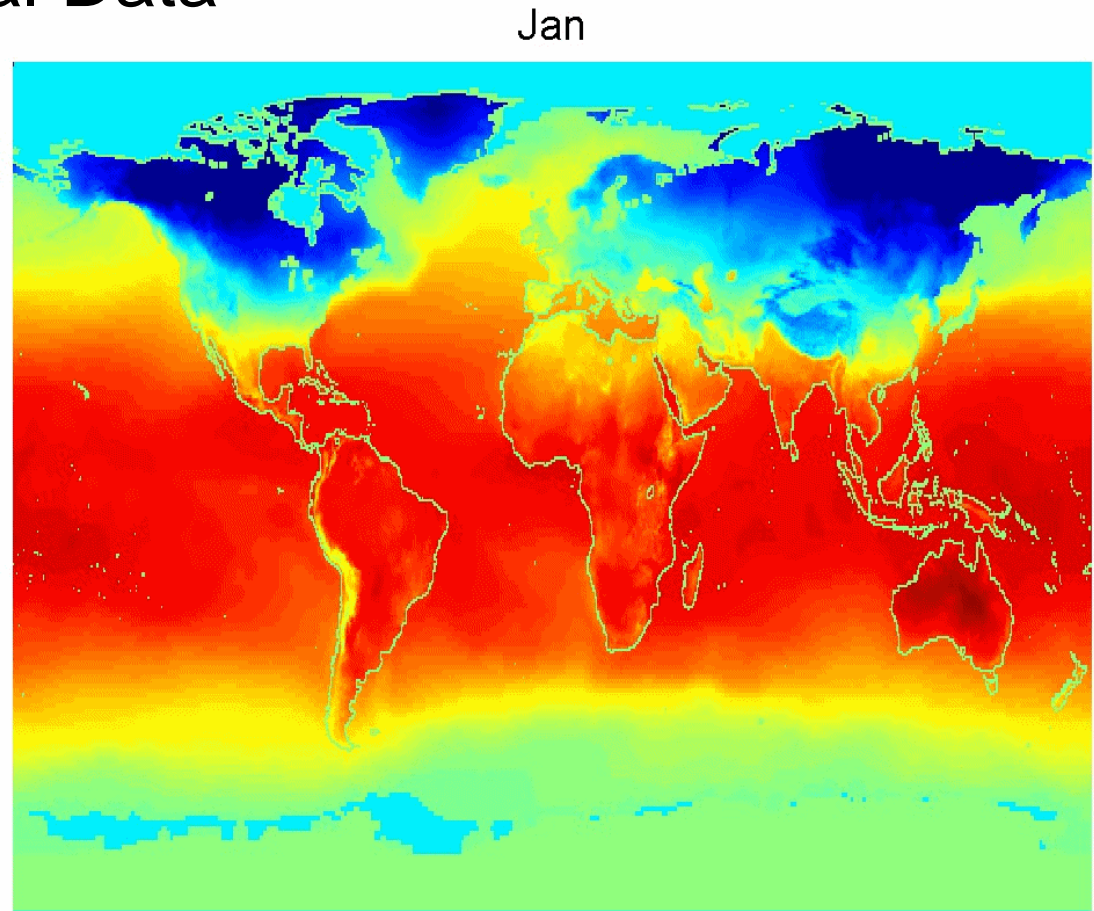
## □ Genomic sequence data

**GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG**

# Ordered Data

## □ Spatio-Temporal Data

**Average Monthly  
Temperature of  
land and ocean**



# Similarity and Dissimilarity Measures

---

## □ Similarity measure

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

## □ Dissimilarity measure

- Numerical measure of how different two data objects are
- 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

---

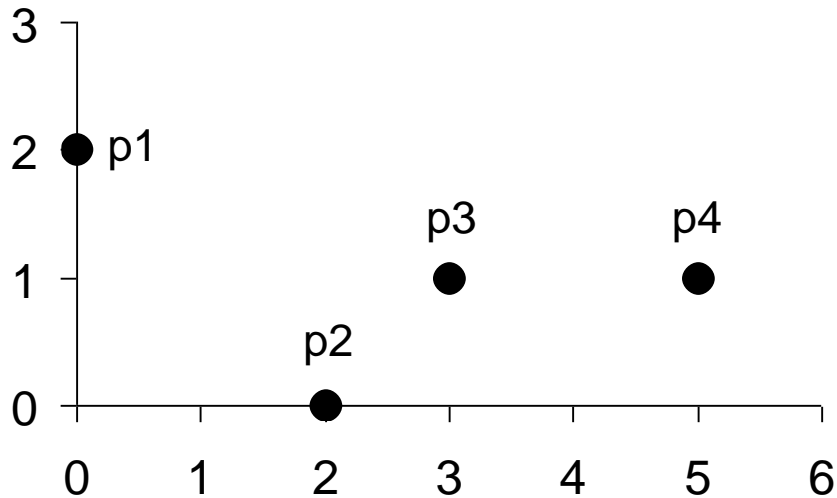
## □ Euclidean Distance

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

where  $n$  is the number of dimensions (attributes) and  $x_k$  and  $y_k$  are, respectively, the  $k^{th}$  attributes (components) or 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 for binary vectors 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

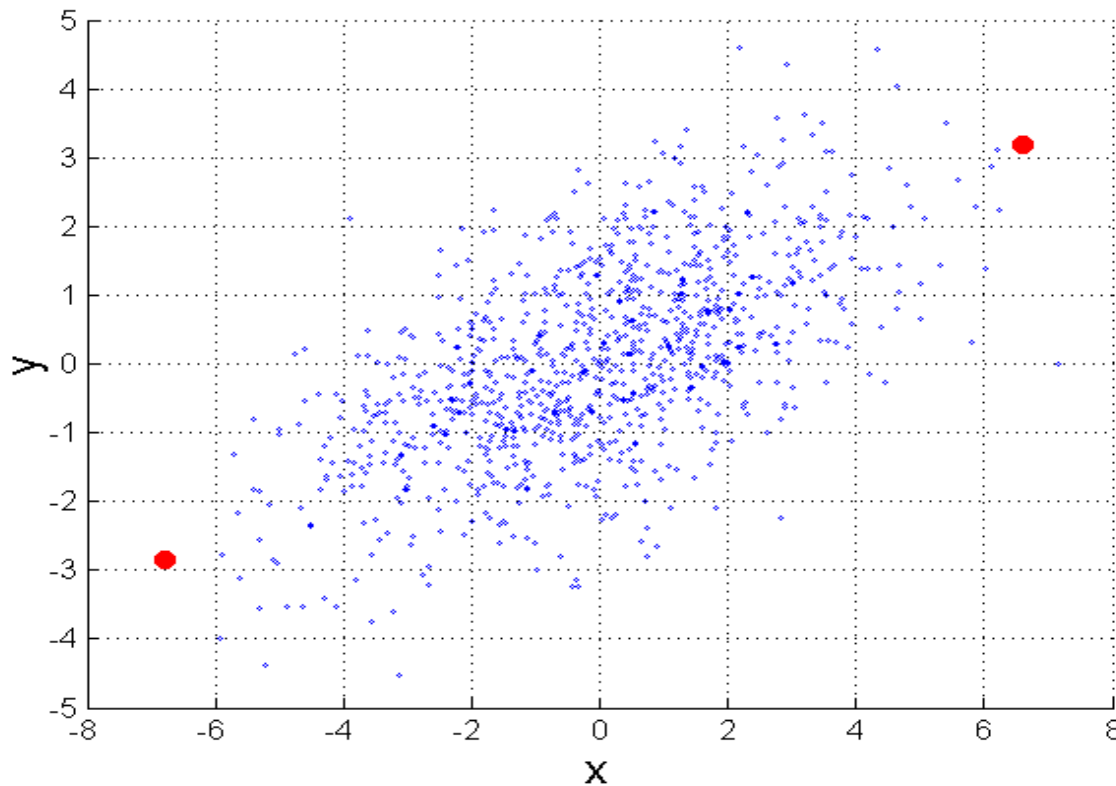
$L_{\infty}$	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

## Distance Matrix

# Mahalanobis Distance

$$\text{mahalanobis}(\mathbf{x}, \mathbf{y}) = ((\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y}))^{-0.5}$$

$\Sigma$  is the covariance matrix



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

# 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  $d(\mathbf{x}, \mathbf{y}) = 0$  if and only if  $\mathbf{x} = \mathbf{y}$ .
  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  $\mathbf{x} = \mathbf{y}$ .  
(does not always hold, e.g., cosine)
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  $\mathbf{y}$ .

# Similarity Between Binary Vectors

- Common situation is that objects,  $\mathbf{x}$  and  $\mathbf{y}$ , have only binary attributes
- Compute similarities using the following quantities
  - $f_{01}$  = the number of attributes where  $\mathbf{x}$  was 0 and  $\mathbf{y}$  was 1
  - $f_{10}$  = the number of attributes where  $\mathbf{x}$  was 1 and  $\mathbf{y}$  was 0
  - $f_{00}$  = the number of attributes where  $\mathbf{x}$  was 0 and  $\mathbf{y}$  was 0
  - $f_{11}$  = the number of attributes where  $\mathbf{x}$  was 1 and  $\mathbf{y}$  was 1
- **Simple Matching and Jaccard Coefficients**

counts both presences and absences equally and it is normally used for symmetric binary attributes

$$\begin{aligned}\mathbf{SMC} &= \text{number of matches} / \text{number of attributes} \\ &= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})\end{aligned}$$

# Similarity Between Binary Vectors

- Common situation is that objects,  $\mathbf{x}$  and  $\mathbf{y}$ , have only binary attributes
- Compute similarities using the following quantities
  - $f_{01}$  = the number of attributes where  $\mathbf{x}$  was 0 and  $\mathbf{y}$  was 1
  - $f_{10}$  = the number of attributes where  $\mathbf{x}$  was 1 and  $\mathbf{y}$  was 0
  - $f_{00}$  = the number of attributes where  $\mathbf{x}$  was 0 and  $\mathbf{y}$  was 0
  - $f_{11}$  = the number of attributes where  $\mathbf{x}$  was 1 and  $\mathbf{y}$  was 1
- **Jaccard Coefficients**

counts only presences and it is frequently for asymmetric binary attributes.

$J$  = number of 11 matches / number of non-zero attributes

$$= (f_{11}) / (f_{01} + f_{10} + f_{11})$$

# SMC versus Jaccard: Example

---

$$\mathbf{x} = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$$

$$\mathbf{y} = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$$

$f_{01} = 2$  (the number of attributes where  $\mathbf{x}$  was 0 and  $\mathbf{y}$  was 1)

$f_{10} = 1$  (the number of attributes where  $\mathbf{x}$  was 1 and  $\mathbf{y}$  was 0)

$f_{00} = 7$  (the number of attributes where  $\mathbf{x}$  was 0 and  $\mathbf{y}$  was 0)

$f_{11} = 0$  (the number of attributes where  $\mathbf{x}$  was 1 and  $\mathbf{y}$  was 1)

$$\begin{aligned}\text{SMC} &= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00}) \\ &= (0+7) / (2+1+0+7) = 0.7\end{aligned}$$

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



# Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the frequency of a particular word or phrase in the document.

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- A similarity measure for documents needs to ignore 0–0 matches like the Jaccard measure, but also must be able to handle non-binary vectors.
- Cosine similarity** is one of the most common measure of document similarity.

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

where  $\mathbf{x}$  and  $\mathbf{y}$  are two document vectors

- where  $\bullet$  indicates vector dot product  $\mathbf{x} \cdot \mathbf{y} = \sum_{k=1}^n x_k y_k$  and  $\|\mathbf{x}\|$  is the length of vector  $\mathbf{x}$ .  $\|\mathbf{x}\| = \sqrt{\sum_{k=1}^n x_k^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}}$

# Cosine Similarity

□ If  $\mathbf{d}_1$  and  $\mathbf{d}_2$  are two document vectors, then

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\| ,$$

where  $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$  indicates inner product or vector dot product of vectors,  $\mathbf{d}_1$  and  $\mathbf{d}_2$ , and  $\|\mathbf{d}\|$  is the length of vector  $\mathbf{d}$ .

□ Example:

$$\mathbf{d}_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

$$\mathbf{d}_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

$$\langle \mathbf{d}_1, \mathbf{d}_2 \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$\|\mathbf{d}_1\| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|\mathbf{d}_2\| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} = 2.449$$

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.3150$$

# Correlation measures the linear relationship between objects

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x s_y}, \quad (2.11)$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \quad (2.12)$$

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

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

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

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

$$\text{corr}(x, y) = \frac{s_{xy}}{s_x \cdot s_y} \Rightarrow \frac{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2} \cdot \sqrt{\frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2}}$$

$$\Rightarrow \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^n (x_k - \bar{x})^2} \cdot \sqrt{\sum_{k=1}^n (y_k - \bar{y})^2}}$$

$x$	$y$
6	12
8	10
10	20

$$\Rightarrow \frac{\sum (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum (x_k - \bar{x})^2 \sum (y_k - \bar{y})^2}}$$

$x$	$y$
6	12
8	10
10	20

$$\bar{x} = \frac{6+8+10}{3} = 8$$

$$\bar{y} = \frac{12+10+20}{3} = 14$$

$$\sum (x_i - \bar{x})^2 = (6-8)^2 + (8-8)^2 + (10-8)^2 = 4+0+4 = 8$$

$$\sum (y_i - \bar{y})^2 = (12-14)^2 + (10-14)^2 + (20-14)^2 = 4+16+36 = 56$$

$$\sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) = (-2)(-2) + (0)(-4) + (2)(6) = 16$$

$$\text{Cor}(x, y) \Rightarrow \frac{16}{\sqrt{8 \cdot 56}} \Rightarrow 0.756$$

# Correlation

---

- Correlation is always in the range -1 to 1.
- A correlation of 1 (-1) means that x and y have a perfect positive (negative) linear relationship.

- A perfect negative linear relationship (correlation: -1)

$$\begin{array}{lll} x = (-3, 6, 0, 3, -6) & s_{xy} = -7.5 & s_x = 4.74341649 \quad s_y = 1.58113883 \\ y = (1, -2, 0, -1, 2) & \text{corr}(x,y) = -1 & \end{array}$$

- A perfect positive linear relationship (correlation: +1)

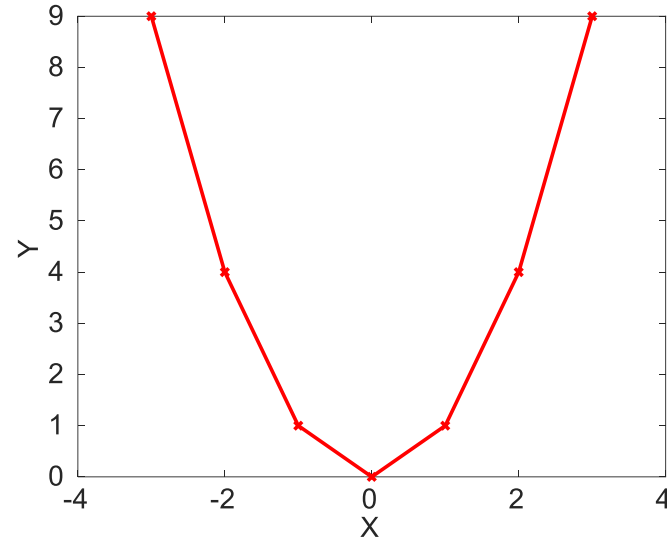
$$\begin{array}{lll} x = (3, 6, 0, 3, 6) & s_{xy} = 2.1 & s_x = 2.50998008 \quad s_y = 0.836660027 \\ y = (1, 2, 0, 1, 2) & \text{corr}(x,y) = +1 & \end{array}$$

# Drawback of Correlation (Non-linear Data)

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

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

$$y_i = x_i^2$$



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

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

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

# Correlation vs cosine vs Euclidean distance

---

- Choice of the right proximity measure depends on the domain
- What is the correct choice of proximity measure for the following situations?
  - Comparing documents using the frequencies of words
    - ◆ Documents are considered similar if the word frequencies are similar
  - Comparing the temperature in Celsius of two locations
    - ◆ Two locations are considered similar if the temperatures are similar in magnitude
  - Comparing two time series of temperature measured in Celsius
    - ◆ Two time series are considered similar if their “shape” is similar, i.e., they vary in the same way over time, achieving minimums and maximums at similar times, etc.