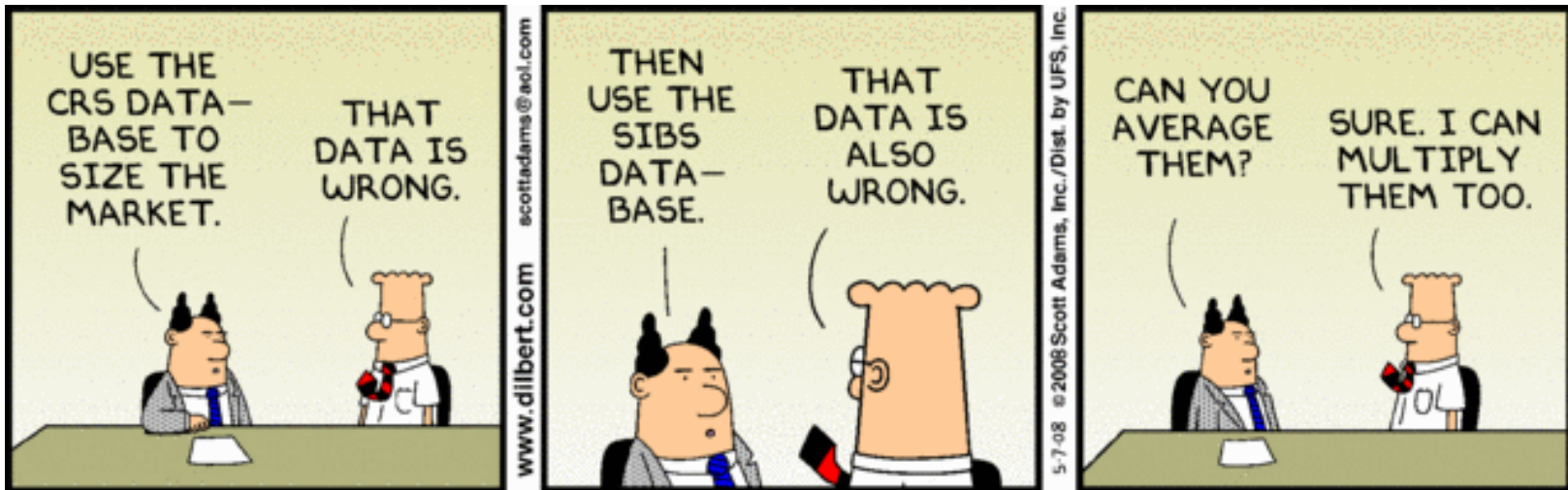


# Data Mining: Data



## Lecture Notes for Chapter 2

Slides by Tan, Steinbach, Kumar adapted by Michael Hahsler



*Look for accompanying R  
code on the course web site.*



## Topics

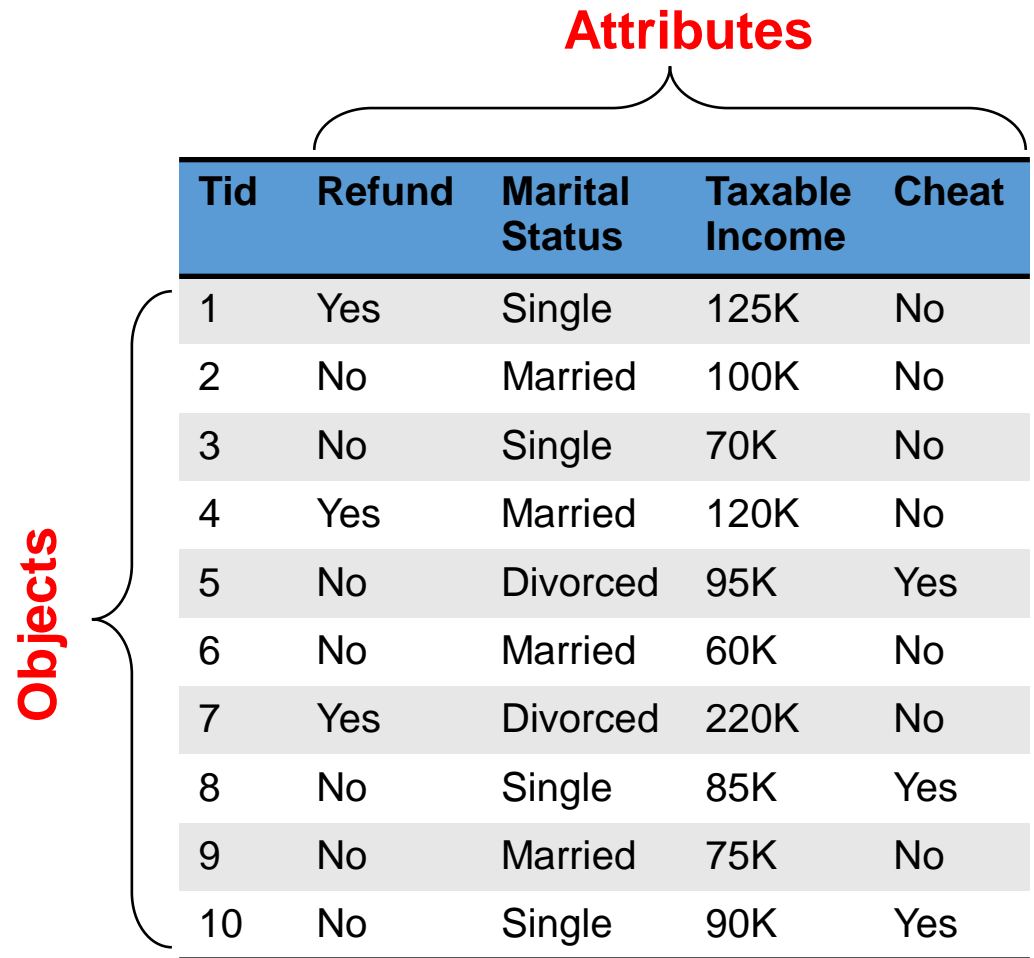
---

- **Attributes/Features**
- Types of Data Sets
- Data Quality
- Data Preprocessing
- Similarity and Dissimilarity
- Density



# What is Data?

- Collection of data objects and their attributes
- An attribute (in Data Mining and Machine learning often "feature") is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance



The diagram illustrates the relationship between data objects and their attributes. A table contains 10 rows of data. A bracket labeled 'Attributes' in red spans the columns: Tid, Refund, Marital Status, Taxable Income, and Cheat. A bracket labeled 'Objects' in red spans the rows, indicating each row represents a single data object.

Attributes				
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
- 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
      - ID has no limit but age has a maximum and minimum value

# Types of Attributes - Scales

- 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, temperatures in Celsius or Fahrenheit.

- Ratio

- Examples: temperature in Kelvin, length, time, counts

Categorical,  
Qualitative

Quantitative

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. ( $=$ , $\neq$ )	zip codes, employee ID numbers, eye color, sex: $\{male, female\}$	mode, entropy, contingency correlation, $\chi^2$ test
Ordinal	The values of an ordinal attribute provide enough information to order objects. ( $<$ , $>$ )	hardness of minerals, $\{good, better, best\}$ , grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. ( $+$ , $-$ )	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, $t$ and $F$ tests
Ratio	For ratio variables, both differences and ratios are meaningful. ( $*$ , $/$ )	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., $new\_value = f(old\_value)$ where $f$ is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by {0.5, 1, 10}.
Interval	$new\_value = a * old\_value + b$ where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	$new\_value = a * old\_value$	Length can be measured in meters or feet.

# 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.



# Examples

- What is the scale of measurement of:
  - Number of cars per minute (count data)
  - Age data grouped in:  
0-4 years, 5-9, 10-14, ...
  - Age data grouped in:  
<20 years, 21-30, 31-40, 41+



## Topics

---

- Attributes/Features
- **Types of Data Sets**
- Data Quality
- Data Preprocessing
- Similarity and Dissimilarity
- Density



# 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 (e.g., from a relational database)

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

# 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 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

		n attributes			
m objects	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	
	5.6	2.7	4.2	1.3	
	6.5	3.0	5.8	2.2	
	6.8	2.8	4.8	1.4	
	5.7	3.8	1.7	0.3	
	5.5	2.5	4.0	1.3	
	4.8	3.0	1.4	0.1	
	5.2	4.1	1.5	0.1	

# 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.

	Terms									
	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

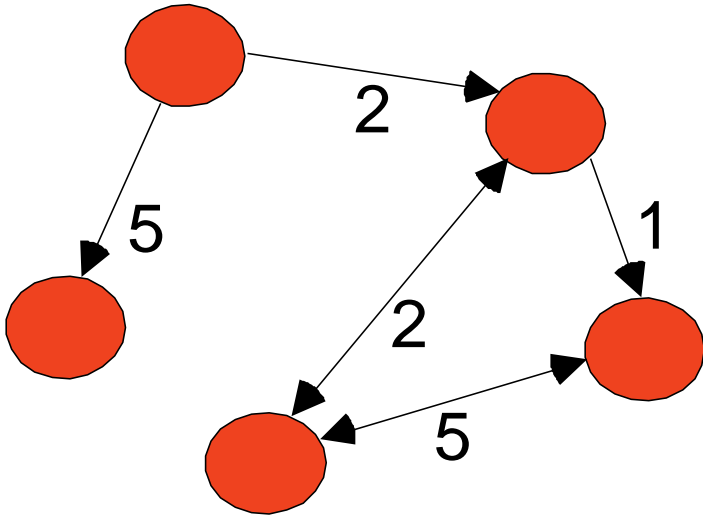
# Transaction Data

- 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.

TID	Items
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 and HTML Links



`<a href="papers/papers.html#bbbb">`  
Data Mining `</a>`

`<li>`

`<a href="papers/papers.html#aaaa">`  
Graph Partitioning `</a>`

`<li>`

`<a href="papers/papers.html#aaaa">`  
Parallel Solution of Sparse Linear System of Equations `</a>`

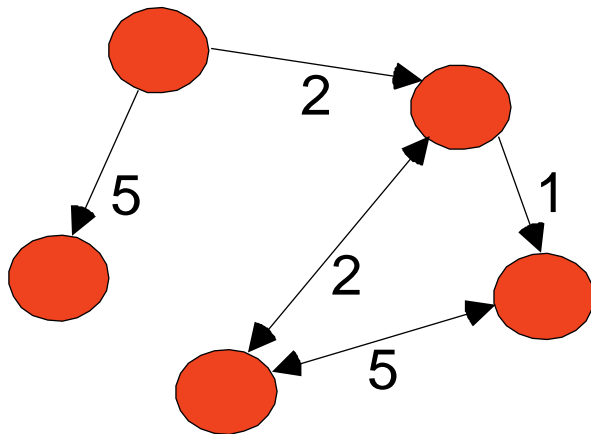
`<li>`

`<a href="papers/papers.html#ffff">`  
N-Body Computation and Dense Linear System Solvers



# 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

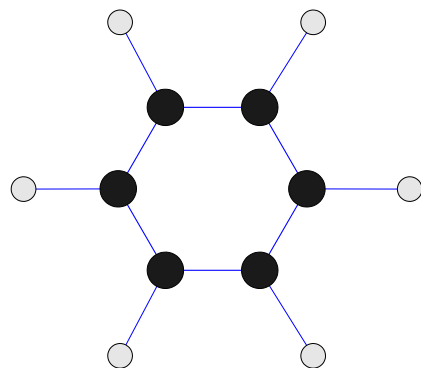
Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

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.

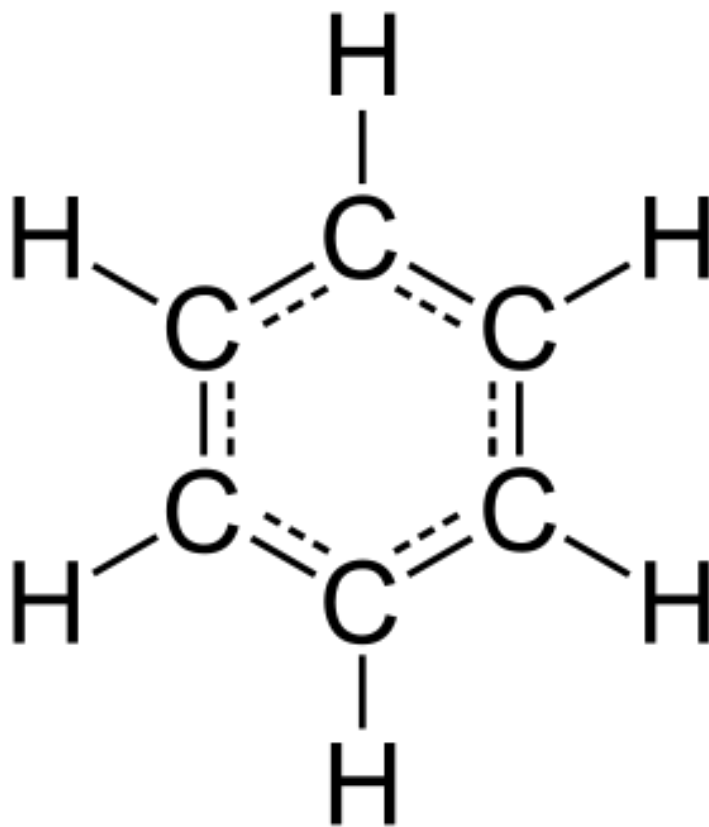
Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.



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

# Chemical Data

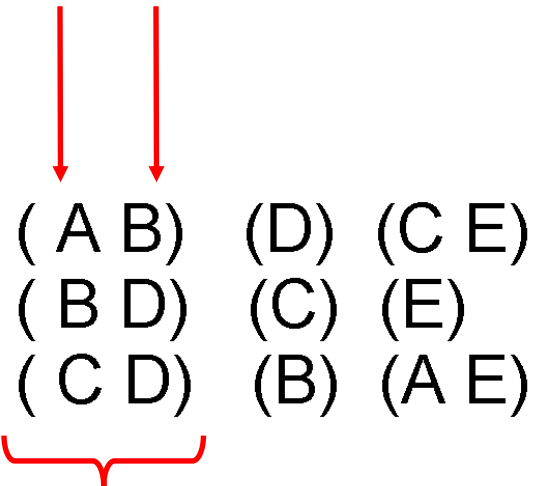
- Benzene Molecule: C<sub>6</sub>H<sub>6</sub>



# Ordered Data

- Sequences of transactions

**Items/Events**



**An element of  
the sequence**

# Ordered Data

- Genomic sequence data

GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG

Subsequences

# Ordered Data: Time Series Data

## S&P 500 Index

*April 1, 2016 – March 31, 2017*

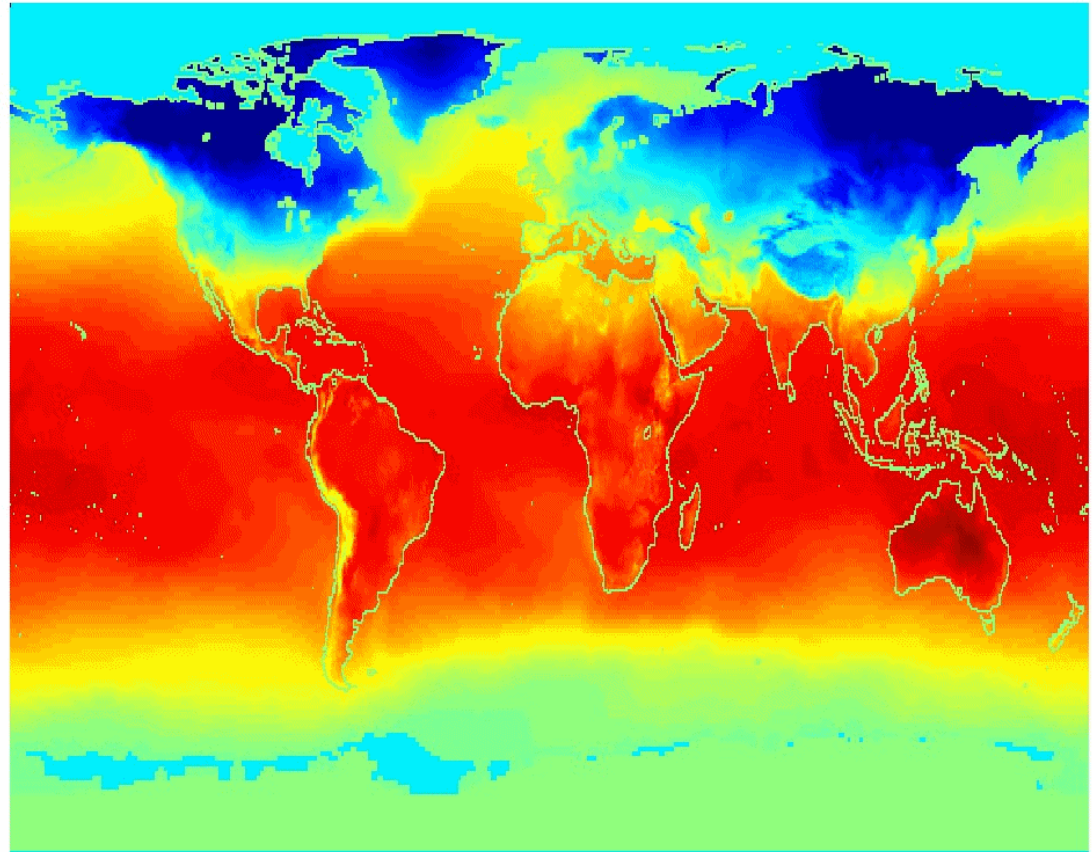


Source: FactSet

# Ordered Data: Spatio-Temporal

Jan, Feb, Mar, ...

**Average Monthly  
Temperature of  
land and ocean**





## Topics

---

- Attributes/Features
- Types of Data Sets
- **Data Quality**
- Data Preprocessing
- Similarity and Dissimilarity
- Density



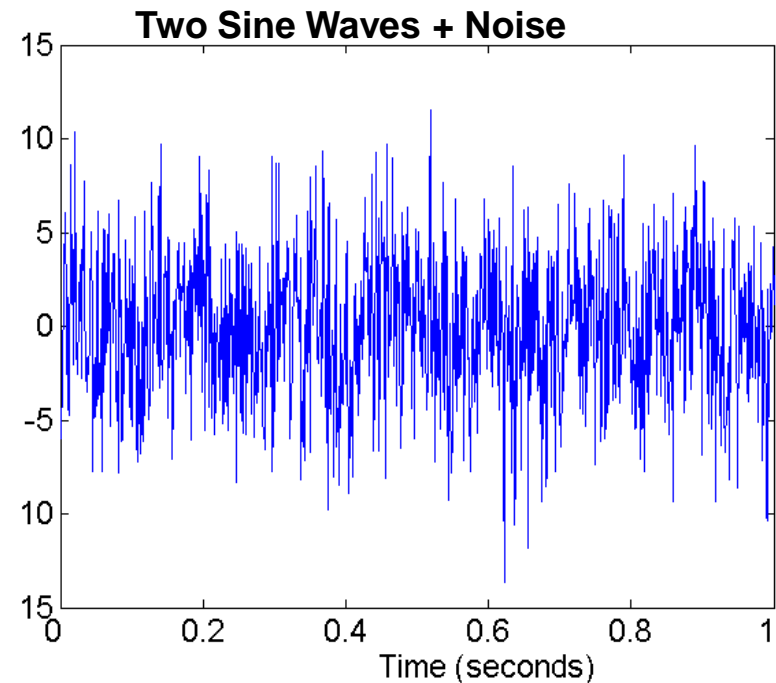
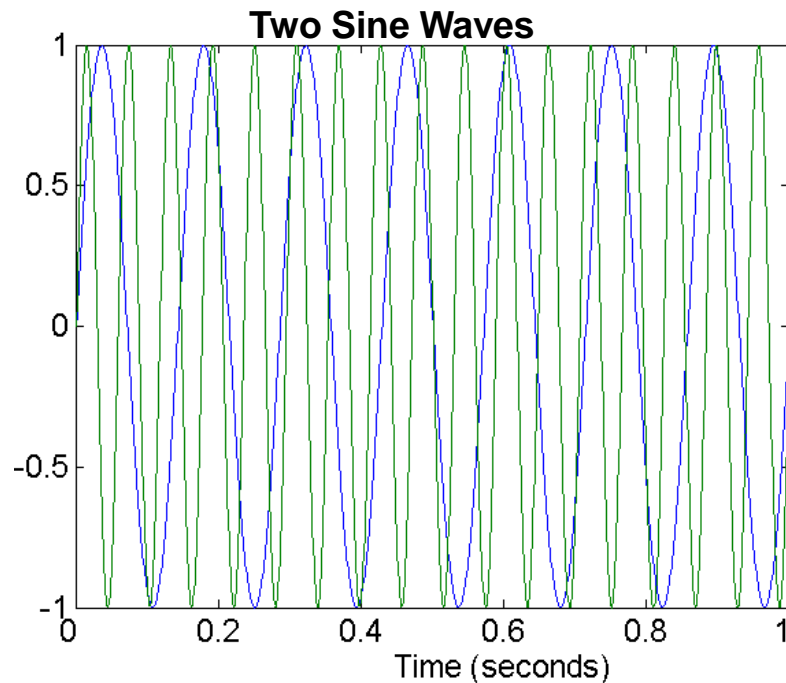
# Data Quality

- 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



# Noise

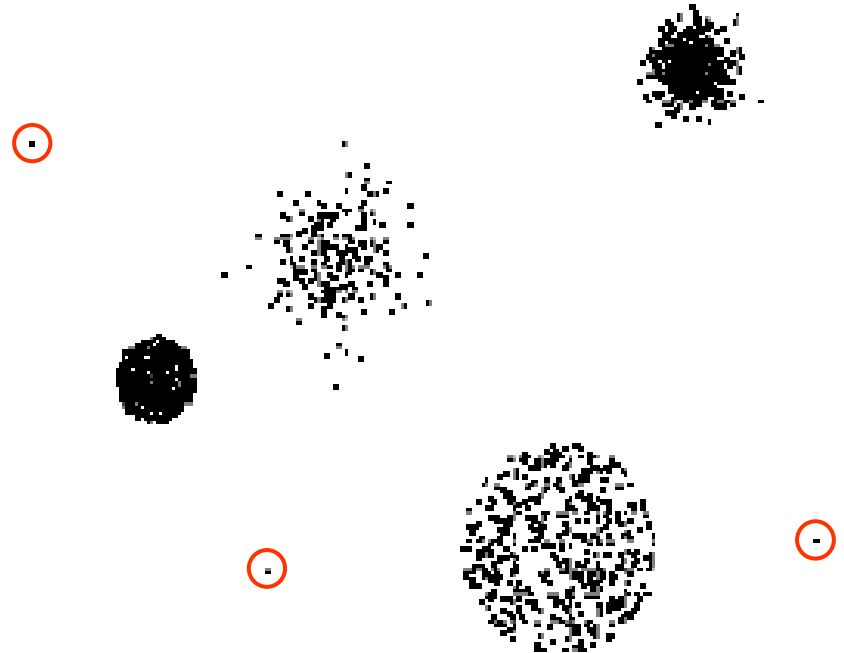
- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone, “snow” on television screen, measurement errors.



- Find less noisy data
- De-noise (signal processing)

# 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
- Causes?



# 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

- Estimate missing values = Imputation  
(e.g., replace with mean or weighted mean where all possible values are weighted by their probabilities)
- Eliminate data objects with missing value
- Eliminate feature with missing values
- Ignore the missing value during analysis

# Duplicate Data

- Data set may include data objects that are duplicates, or "close duplicates" of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues
  - ETL tools typically support deduplication





## Topics

---

- Attributes/Features
- Types of Data Sets
- Data Quality
- **Data Preprocessing**
- Similarity and Dissimilarity
- Density



# Data Preprocessing

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

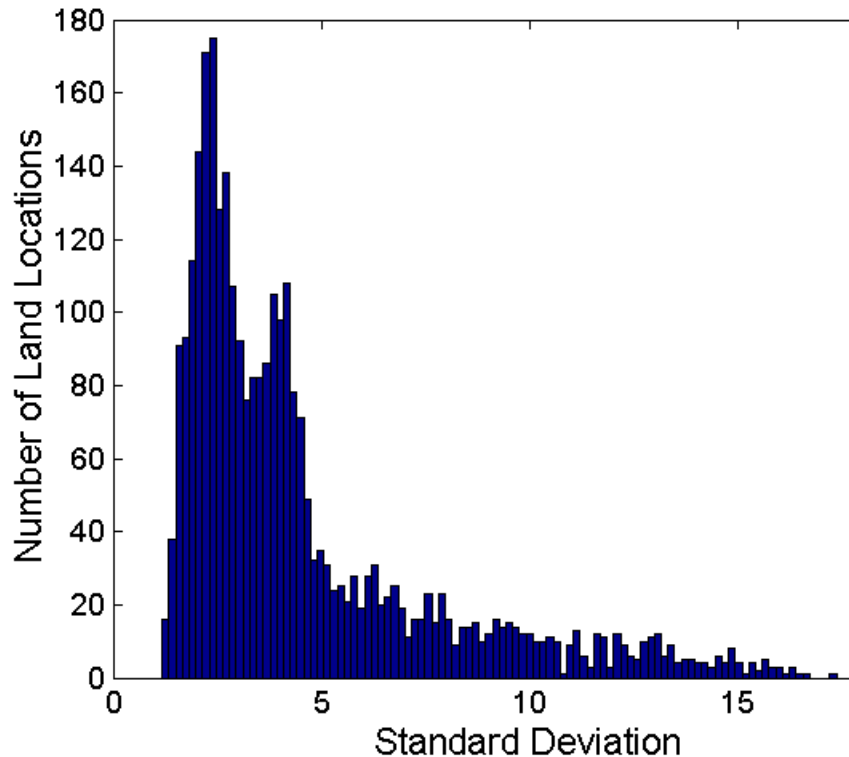
# Aggregation

- 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 (e.g., reduce seasonality by aggregation to yearly data)

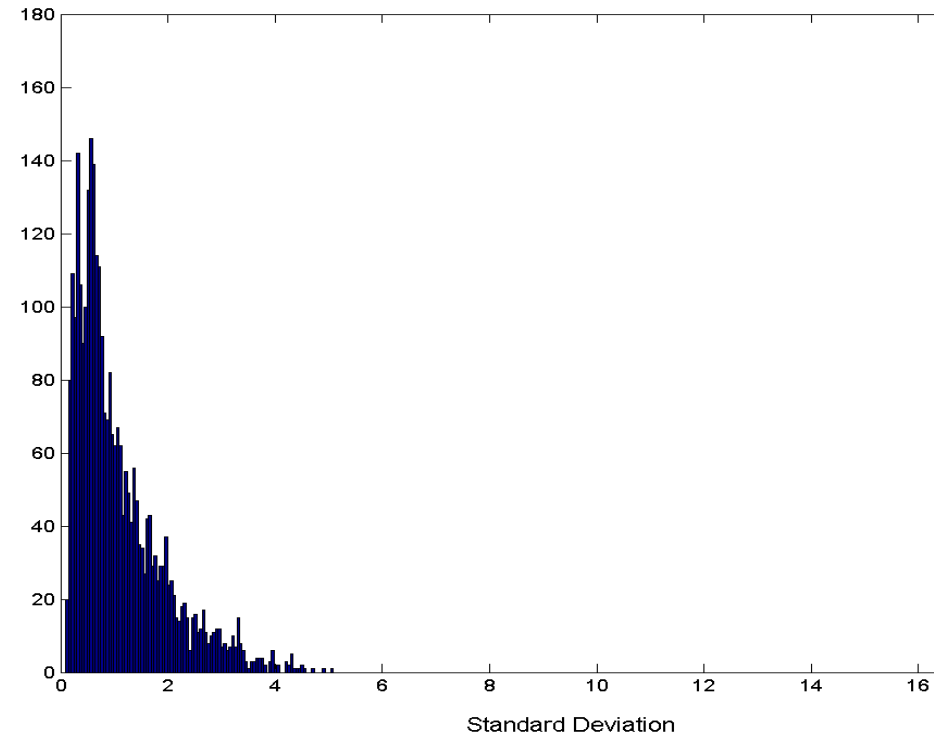


# Aggregation

## Variation of Precipitation in Australia



**Histogram of Standard Deviation of Average Monthly Precipitation**



**Histogram of Standard Deviation of Average Yearly Precipitation**

**Figure 2.8.**

Histograms of standard deviation for monthly and yearly precipitation in Australia for the period 1982–1993.



# 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 (e.g., does not fit into memory or is too slow).

# Sampling ...

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

# Types of Sampling

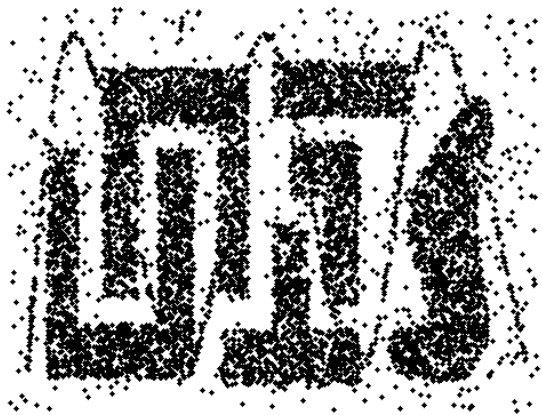
## Replacement?

- 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. Note: the same object can be picked up more than once

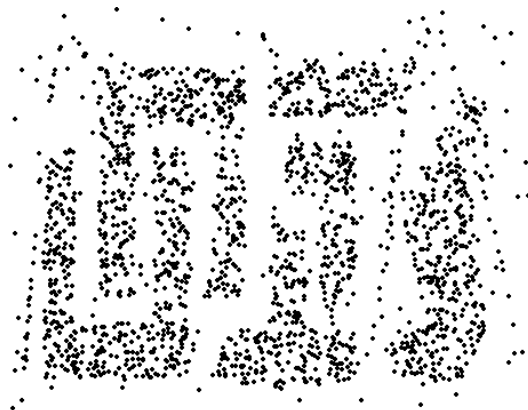
## Selection?

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

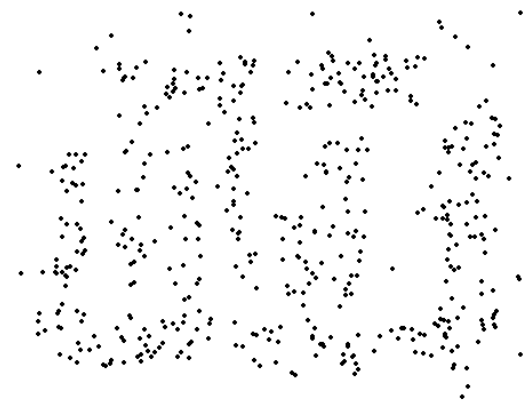
# Sample Size



**8000 points**



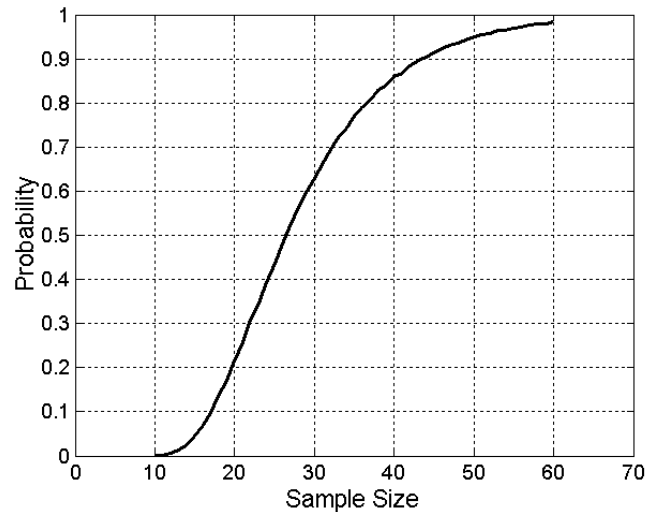
**2000 Points**



**500 Points**

# Sample Size

- What sample size is necessary to get at least one object from each of 10 groups.

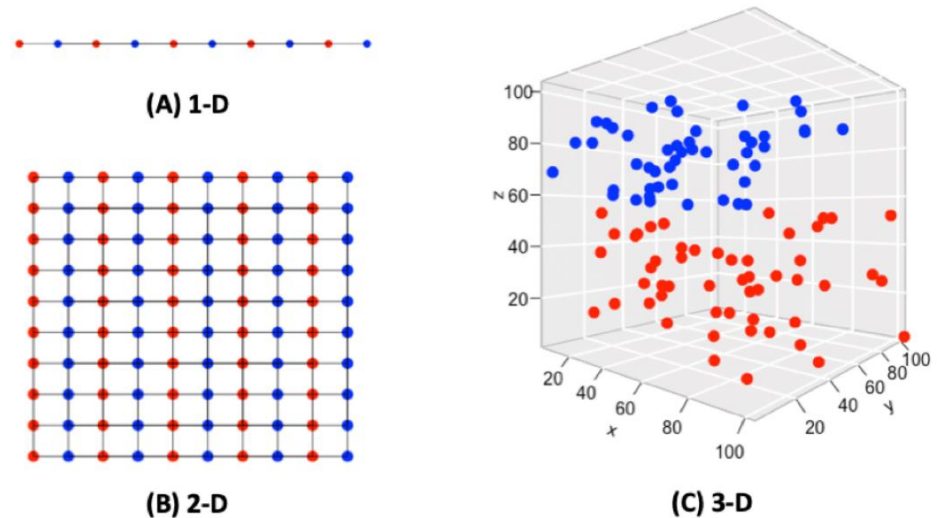


- Sample size determination:
  - Statistics: confidence interval for parameter estimate or desired statistical power of test.
  - Machine learning: often more is better, cross-validated accuracy.



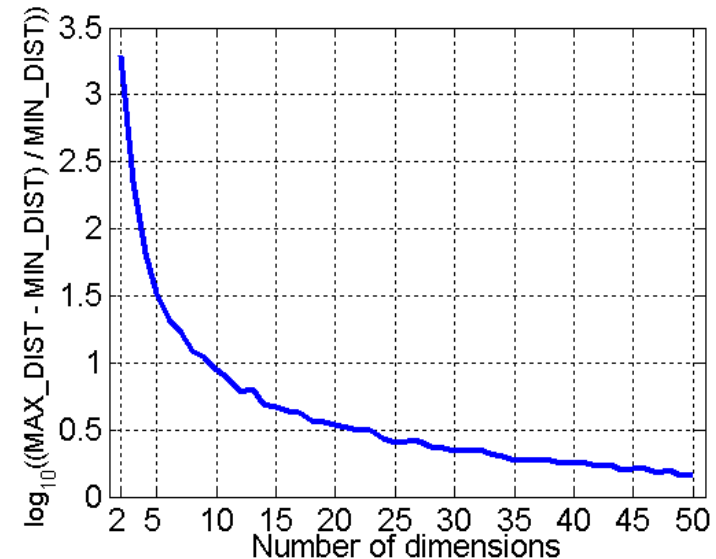
# Curse of Dimensionality

- When dimensionality increases, the size of the data space grows exponentially.
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful
  - Density  $\rightarrow 0$
  - All points tend to have the same Euclidean distance to each other.



Points and space

**Experiment:** Randomly generate 500 points. Compute difference between max and min distance between any pair of points



# Dimensionality Reduction

- Purpose:

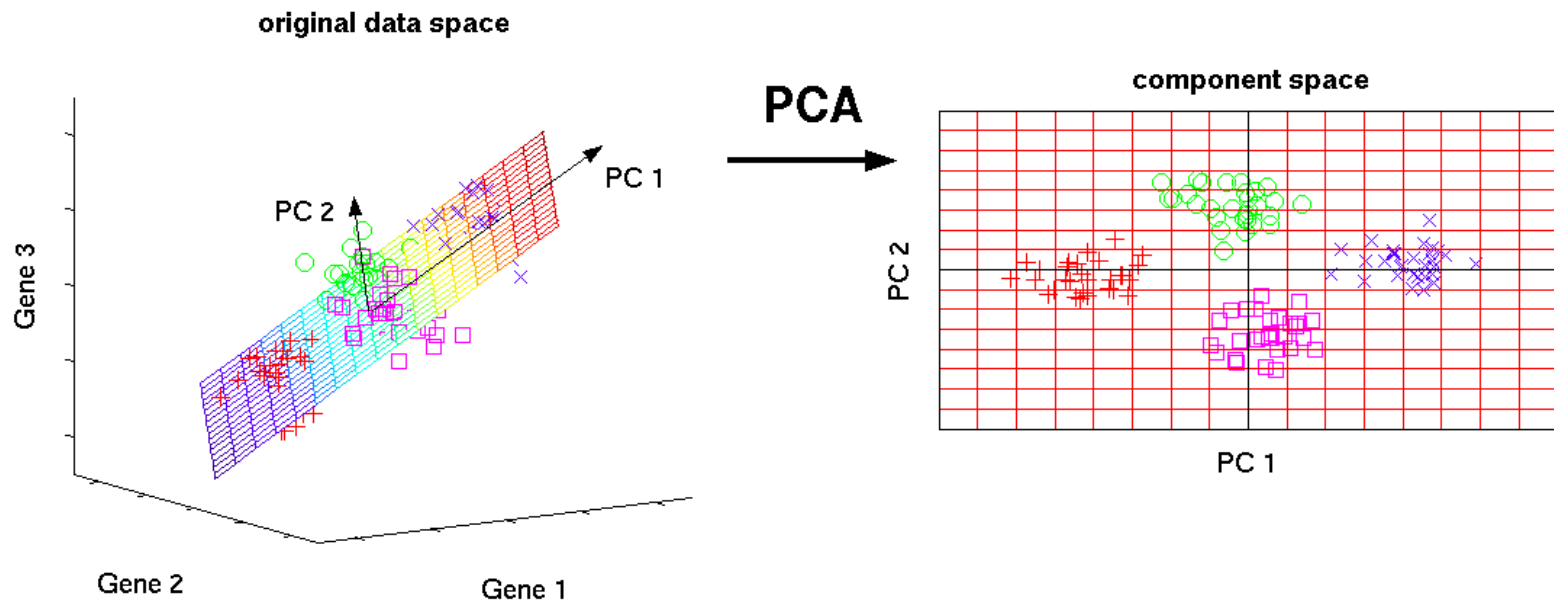
- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise

- Techniques

- Principle Component Analysis
- Singular Value Decomposition
- Others: supervised and non-linear techniques

# Dimensionality Reduction: Principal Components Analysis (PCA)

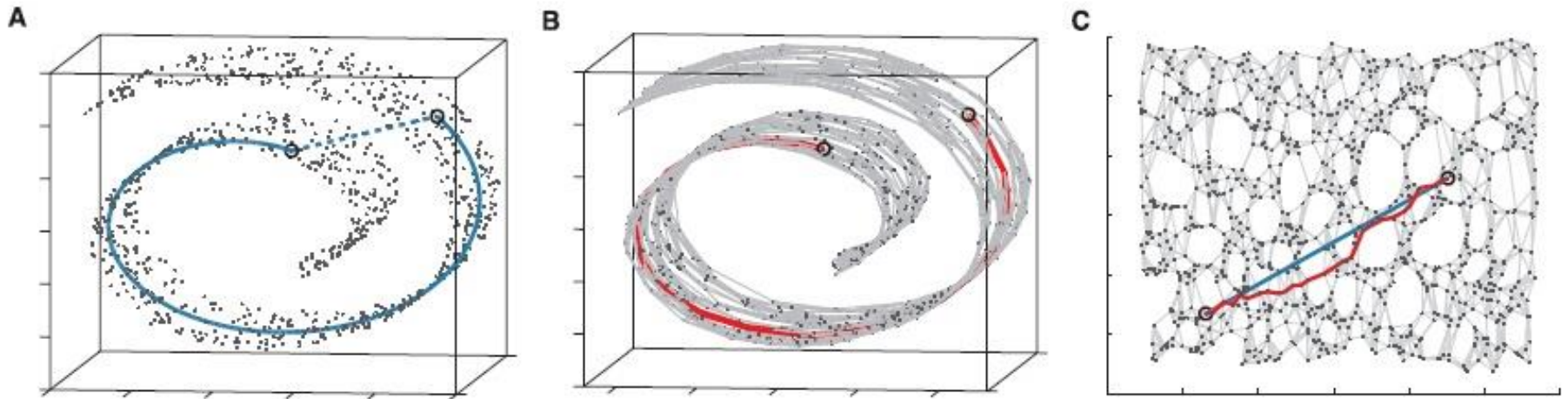
- **Goal:** Map points to a lower dimensional space while preserving distance information.



- **Method:** Find a projection (new axes) that captures the largest amount of variation in data. This can be done using eigenvectors of the covariance matrix or SVD (singular value decomposition).



# Dimensionality Reduction: ISOMAP

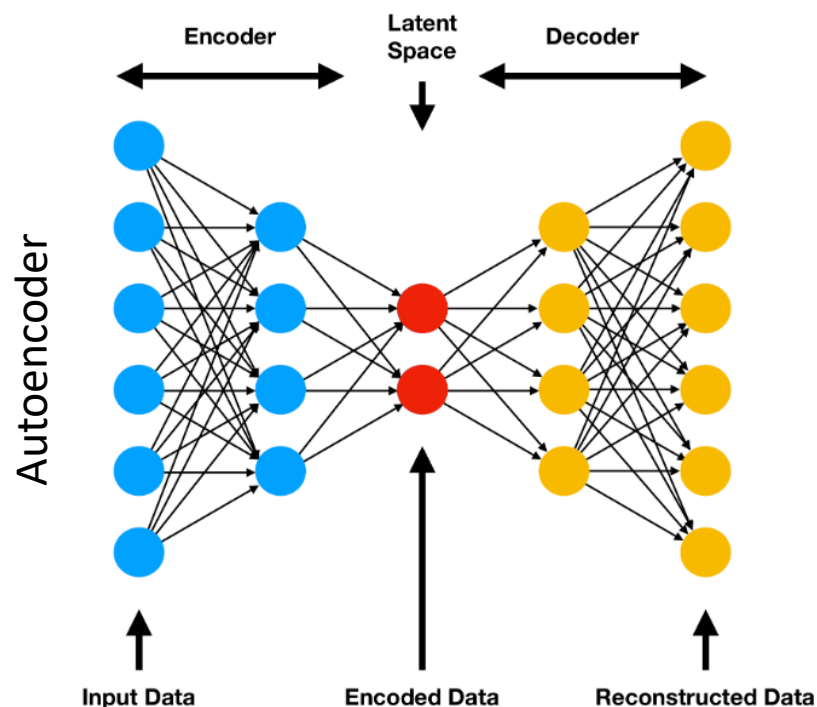
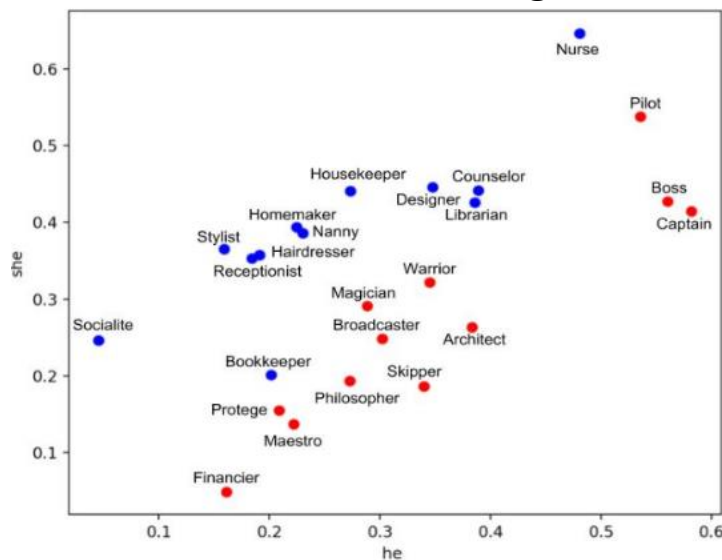


- **Goal:** Unroll the “swiss roll!” (i.e., preserve distances on the roll)
- **Method:** Use a non-metric space, i.e., distances are not measured by Euclidean distance, but along the surface of the roll (geodesic distances).
  1. Construct a neighbourhood graph (k-nearest neighbors or within a radius).
  2. For each pair of points in the graph, compute the shortest path distances = geodesic distances.
  3. Create a lower dimensional embedding using the geodesic distances (multi-dimensional scaling; MDS)

# Low-dimensional Embedding

- General notion of representing objects described in one space (i.e., set of features) in a different space using a map  $f : X \rightarrow Y$
- PCA is an example where  $Y$  is the space spanned by the principal components and objects close in the original space  $X$  are embedded in space  $Y$ .
- Low-dimensional embeddings can be produced with various other methods:
  - T-SNE: T-distributed Stochastic Neighbor Embedding; non-linear for visualization of high-dimensional datasets.
  - Autoencoders (deep learning): non-linear
  - Word embedding: Word2vec, GloVe, BERT

Word Embedding



# 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 (are correlated)
  - 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 often irrelevant to the task of predicting students' GPA

# Feature Subset Selection

- Embedded approaches:
  - Feature selection occurs naturally as part of the data mining algorithm (e.g., regression, decision trees).
- Filter approaches:
  - Features are selected before data mining algorithm is run
  - (e.g., highly correlated features)
- Brute-force approach:
  - Try all possible feature subsets as input to data mining algorithm and choose the best.
- Wrapper approaches:
  - Use the data mining algorithm as a black box to find best subset of attributes (often using greedy search)

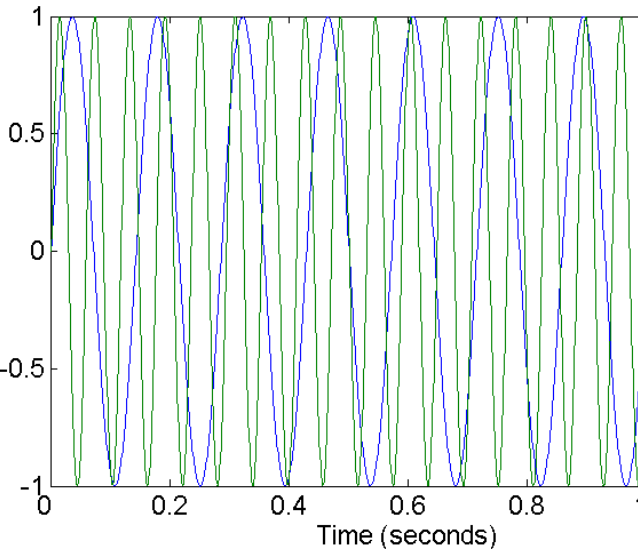
# Feature Creation

- 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 (e.g., face recognition in image mining)
  - Feature Construction / Feature Engineering
    - combining features (interactions: multiply features)
  - Mapping Data to New Space

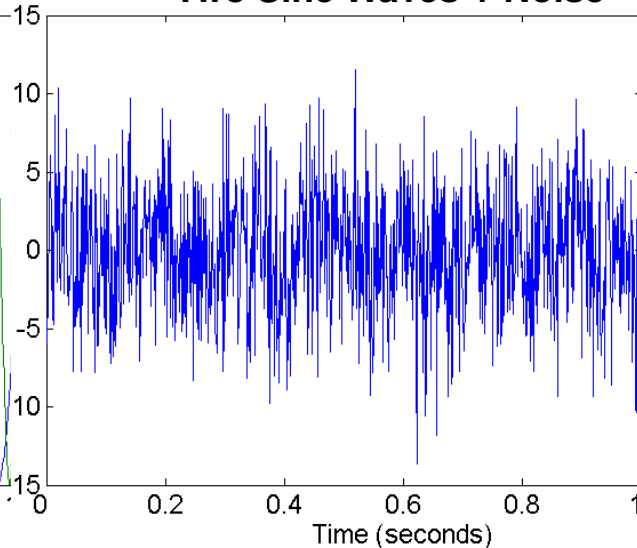
# Mapping Data to a New Space

- Fourier transform
- Wavelet transform

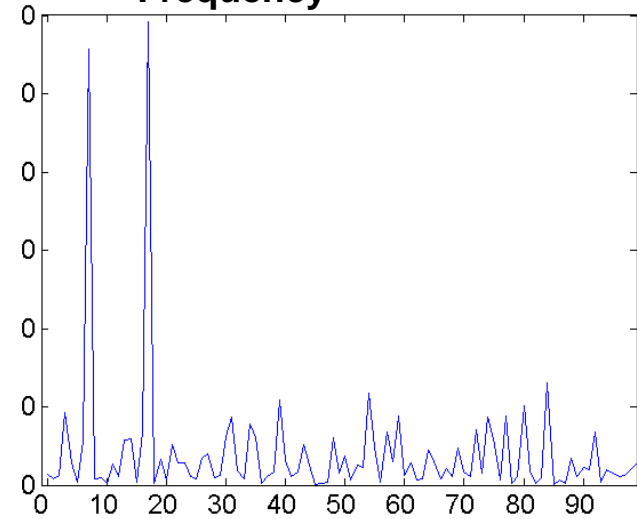
**Two Sine Waves**



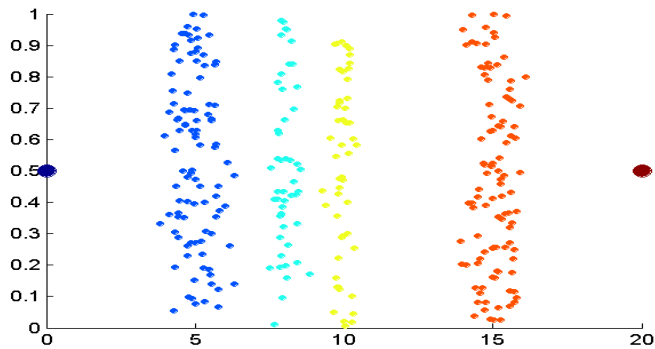
**Two Sine Waves + Noise**



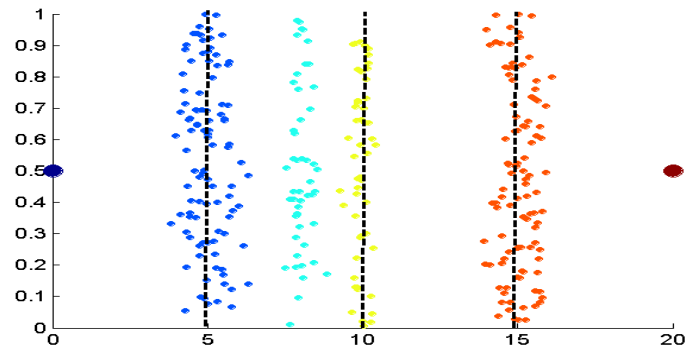
**Frequency**



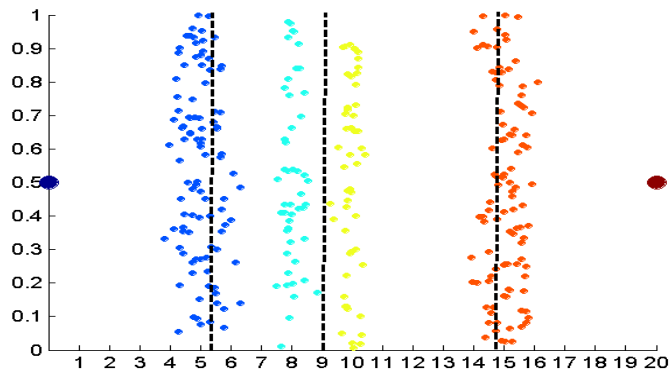
# Unsupervised Discretization



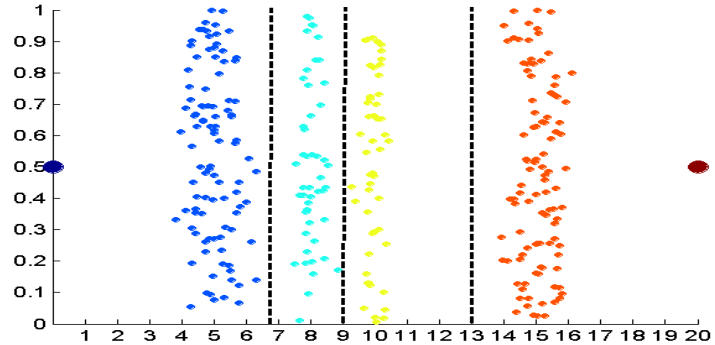
Data



Equal interval width



Equal frequency



K-means



# Attribute Transformation

- 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:  $x^k$ ,  $\log(x)$ ,  $e^x$ ,  $|x|$
  - Standardization and Normalization
    - The z-score normalizes data roughly to an interval of  $[-3,3]$ .

$$x' = \frac{x - \bar{x}}{s_x}$$

$\bar{x}$  ... column (attribute) mean

$s_x$  ... column (attribute) standard deviation







## Topics

---

- Attributes/Features
- Types of Data Sets
- Data Quality
- Data Preprocessing
- **Similarity and Dissimilarity**
- Density



# 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

$p$  and  $q$  are the attribute values for two data objects.

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

$$s = f(d)$$

$f$  can be any strictly decreasing function.

# Euclidean Distance

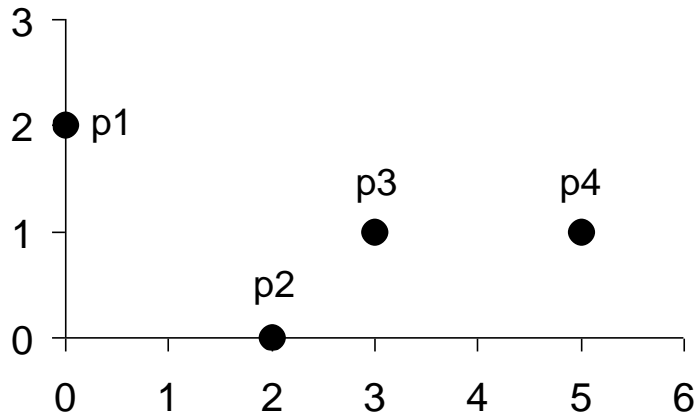
point	x	y
<b>p</b>	0	2
<b>q</b>	2	0

- Euclidean Distance (for quantitative attribute vectors)

$$d_E = \sqrt{\sum_{k=1}^n (p_k - q_k)^2} = \|\mathbf{p} - \mathbf{q}\|_2$$

- Where  $\mathbf{p}$  and  $\mathbf{q}$  are two objects represented by vectors.  $n$  is the number of dimensions (attributes) of the vectors and  $p_k$  and  $q_k$  are, respectively, the  $k$ th attributes (components) or data objects  $p$  and  $q$ .
  - $\|\cdot\|_2$  is the  $L^2$  vector norm (i.e., length of a vector in Euclidean space).
- **Note:** If ranges differ between components of  $\mathbf{p}$  then standardization (z-scores) is necessary to avoid one variable to dominate the distance.

# Euclidean Distance



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

	p1	p2	p3	p4
p1	0.00	2.83	3.16	5.10
p2	2.83	0.00	1.41	3.16
p3	3.16	1.41	0.00	2.00
p4	5.10	3.16	2.00	0.00

**Distance Matrix**

# Minkowski Distance

point	x	y
<b>p</b>	0	2
<b>q</b>	2	0

- Minkowski Distance is a generalization of Euclidean Distance

$$d_M = \left( \sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}} = \|\mathbf{p} - \mathbf{q}\|_r$$

- Where  $\mathbf{p}$  and  $\mathbf{q}$  are two objects represented by vectors.  $n$  is the number of dimensions (attributes) of the vectors and  $p_k$  and  $q_k$  are, respectively, the  $k$ th attributes (components) of data objects  $\mathbf{p}$  and  $\mathbf{q}$ .
- **Note:** If ranges differ then standardization (z-scores) is necessary to avoid one variable to dominate the distance.

# 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 ( $L^2$  norm)
- $r = \infty$ . “supremum” (maximum 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 Distances

## Distance Matrix

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

$L^1$	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

$L^2$	p1	p2	p3	p4
p1	0.00	2.83	3.16	5.10
p2	2.83	0.00	1.41	3.16
p3	3.16	1.41	0.00	2.00
p4	5.10	3.16	2.00	0.00

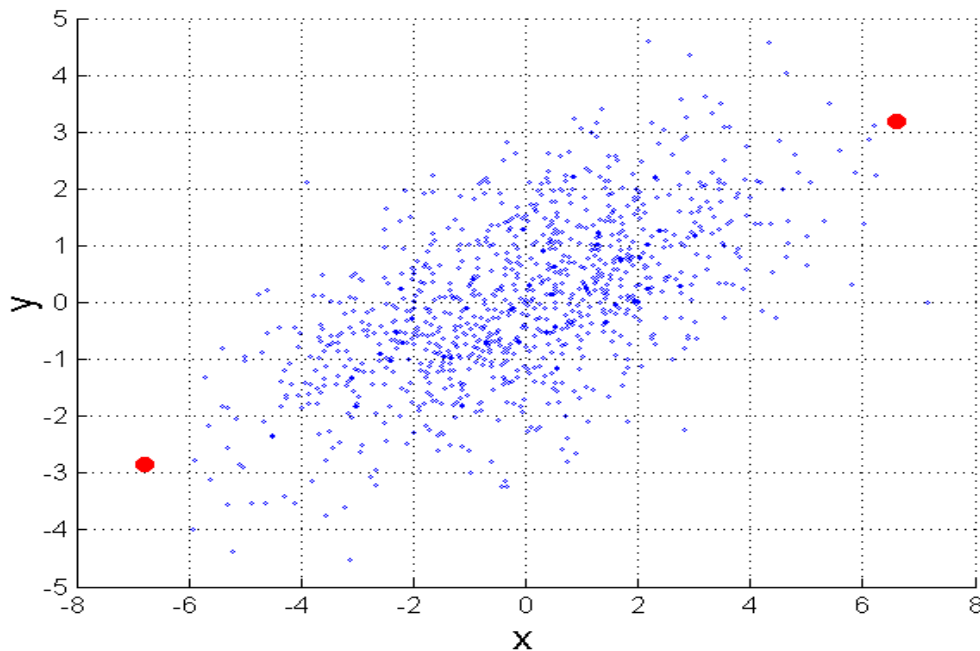
$L^\infty$	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0





# Mahalanobis Distance

$$d_{mahalanobis}(\mathbf{p}, \mathbf{q}) = \sqrt{(\mathbf{p} - \mathbf{q})^T \mathbf{S}^{-1} (\mathbf{p} - \mathbf{q})}$$

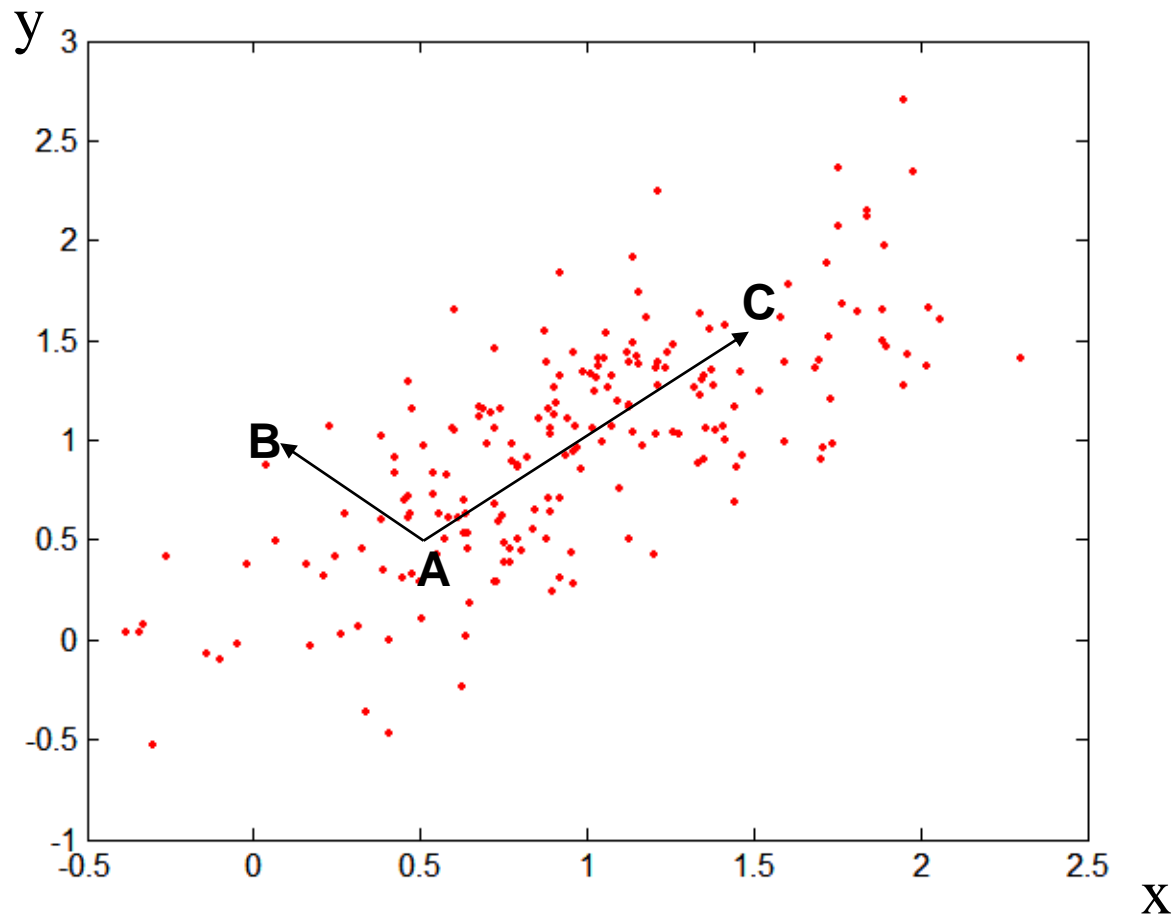


$\mathbf{S}^{-1}$  is the inverse of the covariance matrix of the input data

Measures how many standard deviations two points are away from each other → scale invariant measure

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

# Mahalanobis Distance



Covariance Matrix:

$$S = \begin{bmatrix} .3 & .2 \\ .2 & .3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

$$d_{mahal}(A, B) = 5$$

$$d_{mahal}(A, C) = 4$$

Data varies in direction A-C more than in A-B!

# Cosine Similarity

For two vector A and B, the cosine similarity is defined as

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Example:

A = 3 2 0 5 0 0 0 2 0 0

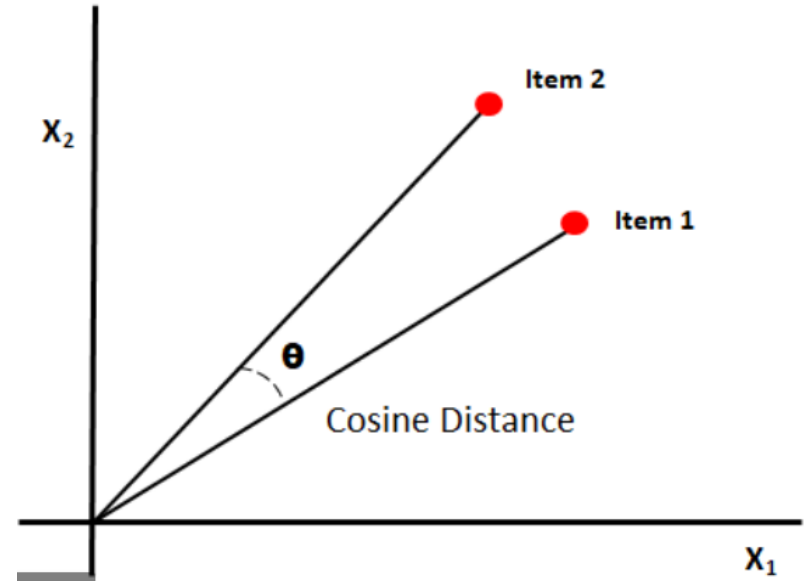
B = 1 0 0 0 0 0 0 1 0 2

$$\mathbf{A} \cdot \mathbf{B} = 3 * 1 + 2 * 0 + 0 * 0 + 5 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 2 * 1 + 0 * 0 + 0 * 2 = 5$$

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

$$\|\mathbf{B}\| = (1 * 1 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 1 * 1 + 0 * 0 + 2 * 2)^{0.5} = 2.245$$

$$s_{\cosine} = .3150$$



Cosine similarity is often used for word count vectors to compare documents.

# Similarity Between Binary Vectors

- Common situation is that objects, p and q, have only binary attributes

- Compute similarities using the following quantities

M01 = the number of attributes where p was 0 and q was 1

M10 = the number of attributes where p was 1 and q was 0

M00 = the number of attributes where p was 0 and q was 0

M11 = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

$s_{SMC}$  = number of matches / number of attributes

$$= (M11 + M00) / (M01 + M10 + M11 + M00)$$

$s_J$  = number of 11 matches / number of not-both-zero attribute values

$$= (M11) / (M01 + M10 + M11)$$

Note: Jaccard ignores 0s!

# 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

M01 = 2 (the number of attributes where p was 0 and q was 1)

M10 = 1 (the number of attributes where p was 1 and q was 0)

M00 = 7 (the number of attributes where p was 0 and q was 0)

M11 = 0 (the number of attributes where p was 1 and q was 1)

$$s_{SMC} = (M11 + M00) / (M01 + M10 + M11 + M00) = (0+7) / (2+1+0+7) = 0.7$$

$$s_J = (M11) / (M01 + M10 + M11) = 0 / (2 + 1 + 0) = 0$$

# Extended Jaccard Coefficient (Tanimoto)

- Variation of Jaccard for continuous or count attributes:

$$T(p, q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q}$$

where  $\bullet$  is the dot product between two vectors and  $\|\cdot\|^2$  is the Euclidean norm (length of the vector).

- Reduces to Jaccard for binary attributes

# Dis(similarities) With Mixed Types

- Sometimes attributes are of many different types (nominal, ordinal, ratio, etc.), but an overall similarity is needed.
  - Gower's (dis)similarity:
    - Ignores missing values
    - Final (dis)similarity is a weighted sum of variable-wise (dis)similarities
1. For the  $k^{th}$  attribute, compute a similarity,  $s_k$ , in the range  $[0, 1]$ .
  2. Define an indicator variable,  $\delta_k$ , for the  $k_{th}$  attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have} \\ & \text{a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^n \delta_k s_k}{\sum_{k=1}^n \delta_k}$$



# Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well-known properties.

1.  $d(p, q) \geq 0$  for all  $p$  and  $q$  and  $d(p, q) = 0$  only if  $p = q$ . (Positive definiteness)
2.  $d(p, q) = d(q, p)$  for all  $p$  and  $q$ . (Symmetry)
3.  $d(p, r) \leq d(p, q) + d(q, r)$  for all points  $p$ ,  $q$ , and  $r$ . (Triangle Inequality)

where  $d(p, q)$  is the distance (dissimilarity) between points (data objects),  $p$  and  $q$ .

- A distance that satisfies these properties is a **metric** and forms a **metric space**.



# Common Properties of a Similarity

- Similarities, also have some well-known properties.

- $s(p, q) = 1$  (or maximum similarity) only if  $p = q$ .

- $s(p, q) = s(q, p)$  for all  $p$  and  $q$ . (Symmetry)

where  $s(p, q)$  is the similarity between points (data objects),  $p$  and  $q$ .

# Exercise

	x	y
A	2	1
B	4	3
C	1	1

- Calculate the Euclidean and the Manhattan distances between A and C and A and B
- Calculate the Cosine similarity between A and C and A and B

# Correlation

- Correlation measures the (linear) relationship between two variables.
- To compute Pearson correlation (Pearson's Product Moment Correlation), we standardize data objects,  $p$  and  $q$ , and then take their dot product

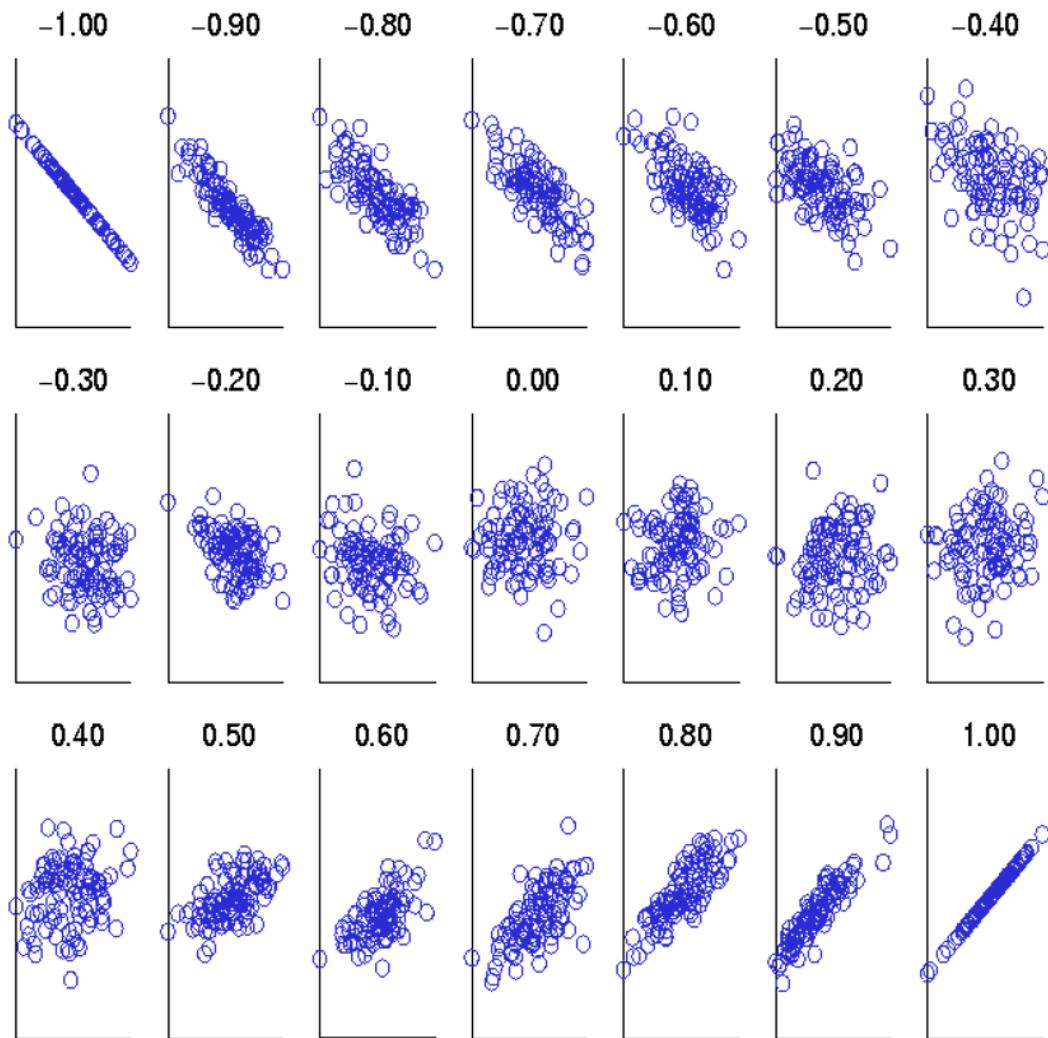
$$\rho = \frac{\text{cov}(X, Y)}{\text{sd}(X) \text{sd}(Y)}$$

- Estimation:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- Correlation is often used as a measure of similarity.

# Visually Evaluating Correlation



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

# Rank Correlation

- Measure the degree of similarity between two ratings (e.g., ordinal data).
- Is more robust against outliers and does not assume normality of data or linear relationship like Pearson Correlation.
- Measures (all are between -1 and 1)
  - Spearman's Rho: Pearson correlation between ranked variables.
  - Kendall's Tau

$$\tau = \frac{N_s - N_d}{\frac{1}{2}n(n-1)}$$

$N_s$ ... concordant pair  
 $N_d$ ... discordant pair

- Goodman and Kruskal's Gamma

$$\gamma = \frac{N_s - N_d}{N_s + N_d}$$





## Topics

---

- Attributes/Features
- Types of Data Sets
- Data Quality
- Data Preprocessing
- Similarity and Dissimilarity
- **Density**



# Density

- Density-based clustering require a notion of density
- Examples:
  - Probability density (function) = describes the likelihood of a random variable taking a given value
  - Euclidean density = number of points per unit volume
  - ~~— Graph-based density = number of edges compared to a complete graph~~
  - ~~— Density of a matrix = proportion of non-zero entries.~~

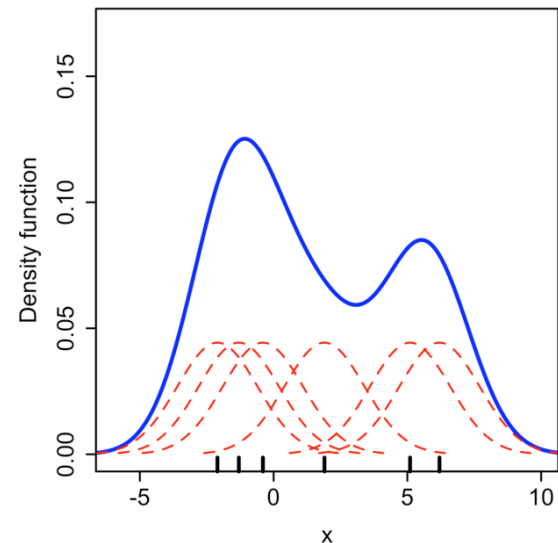
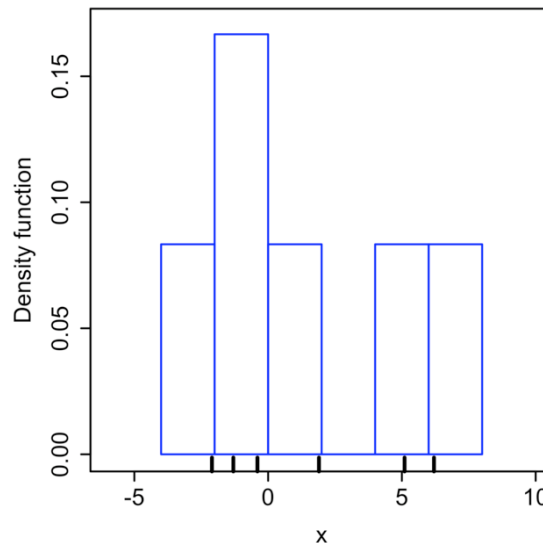
# Kernel Density Estimation (KDE)

- KDE is a non-parametric way to estimate the probability density function of a random variable.

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right),$$

- K is the kernel (a non-negative function that integrates to one) and  $h > 0$  is a smoothing parameter called the bandwidth. Often a Gaussian kernel is used.

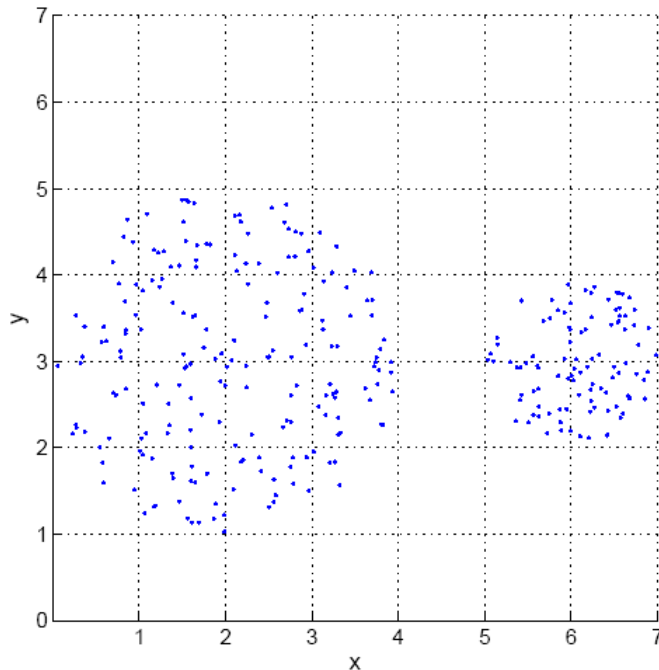
- Example:





# Euclidean Density – Cell-based

- Simplest approach is to divide region into a number of rectangular cells of equal volume and define density as # of points the cell contains.



**Figure 7.13.** Cell-based density.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

**Table 7.6.** Point counts for each grid cell.

# Euclidean Density – Center-based

- Euclidean density is the number of points within a specified radius of the point

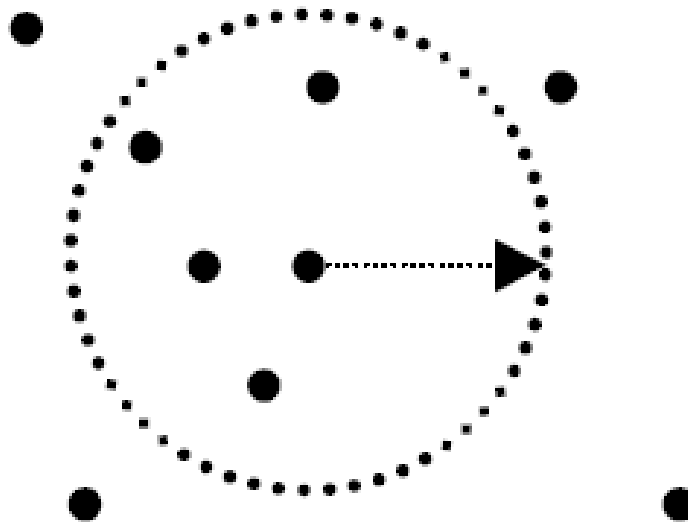


Figure 7.14. Illustration of center-based density.

## You should know now about...

---

- Attributes/Features
- Types of Data Sets
- Data Quality
- Data Preprocessing
- Similarity and Dissimilarity
- Density



# Examples of Classification Task

Task	Attribute set, $x$	Class label, $y$
Categorizing email messages	Features extracted from email message header and content	spam or non-spam
Identifying tumor cells	Features extracted from x-rays or MRI scans	malignant or benign cells
Cataloging galaxies	Features extracted from telescope images	Elliptical, spiral, or irregular-shaped galaxies