“DATA AND WEB MINING”

DATA

Salvatore Orlando

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, or feature
- A collection of attributes describe 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

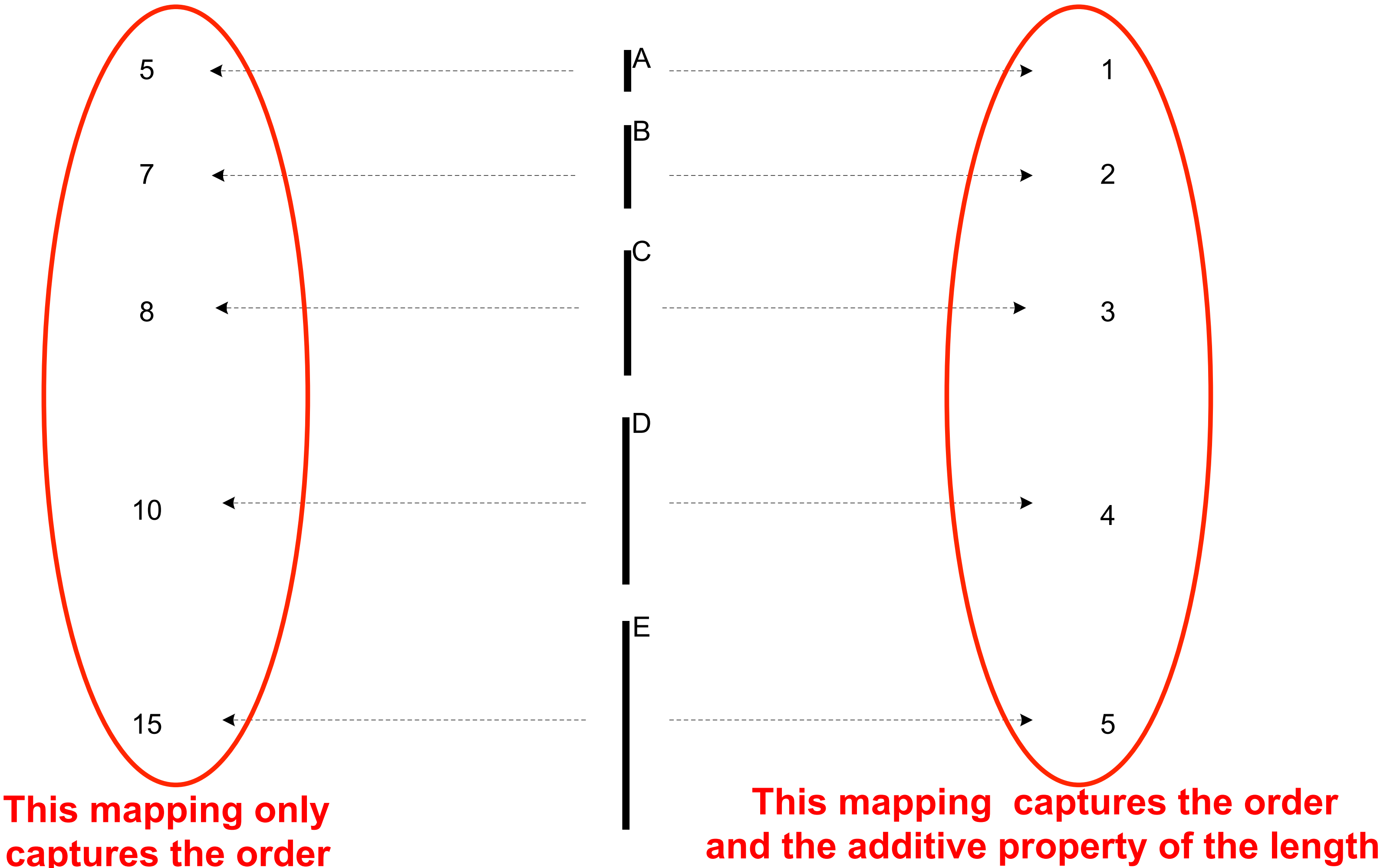
Attribute Values

- **Attribute values are numbers or symbols assigned to an attribute**

- **Distinction between attributes and attribute values**
  - **Same attribute can be mapped to different attribute values**
    - **Depends on the** measurement scale
    - **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

Measurement of Length

- **The way you measure an attribute is somewhat may not match the attributes properties.**



Types of Attributes

- 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

Properties of Attribute Values

- **The type of an attribute depends on which of the following properties it possesses:**
  - **Distinctness: = ≠**
  - **Order: < >**
  - **Addition: + -**
  - **Multiplication: * /**

  - **Nominal attribute: distinctness**
  - **Ordinal attribute: distinctness & order**
  - **Interval attribute: distinctness, order & addition**
  - **Ratio attribute: all 4 properties**

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. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^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. The ratio of two measures is not meaningful (+, -)	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. (*, /) I can say measure 500 is two times measure 250, since 500/250=2	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.

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

Important Characteristics of Structured Data

- Dimensionality
  - Curse of Dimensionality

- Sparsity
  - Only presence counts
  - Many absent attributes (0)

- Resolution
  - Pattern identification depends on the scale

Record Data

- **Data that consists of a collection of records, each of which consists of a fixed set of 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

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**

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	pla y	ball	score	game	wi n	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

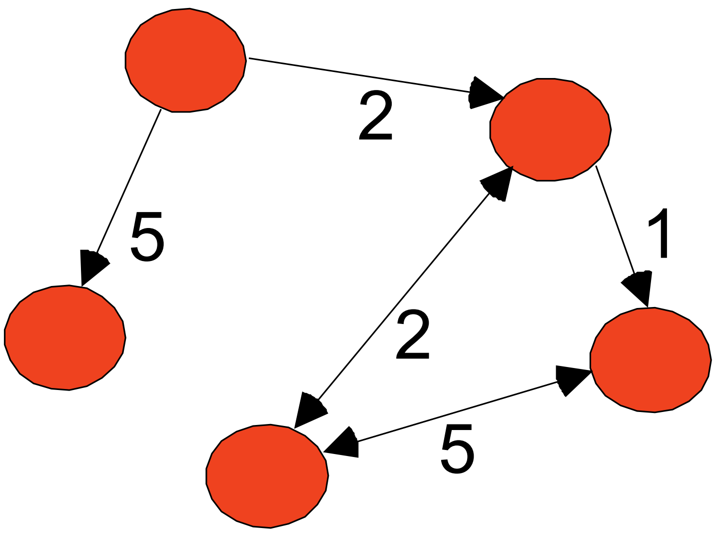
Transaction Data

- 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

- **The graph captures relationships among data objects**
  - **Example: Generic graph and HTML Links**



Data Mining

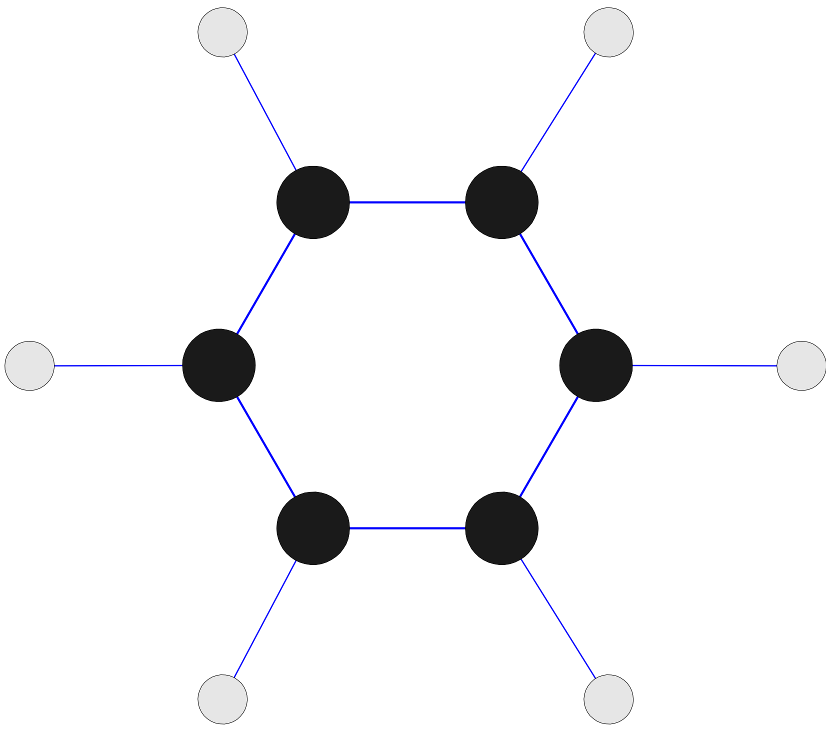
Graph Partitioning

Parallel Solution of Sparse Linear System of Equations

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

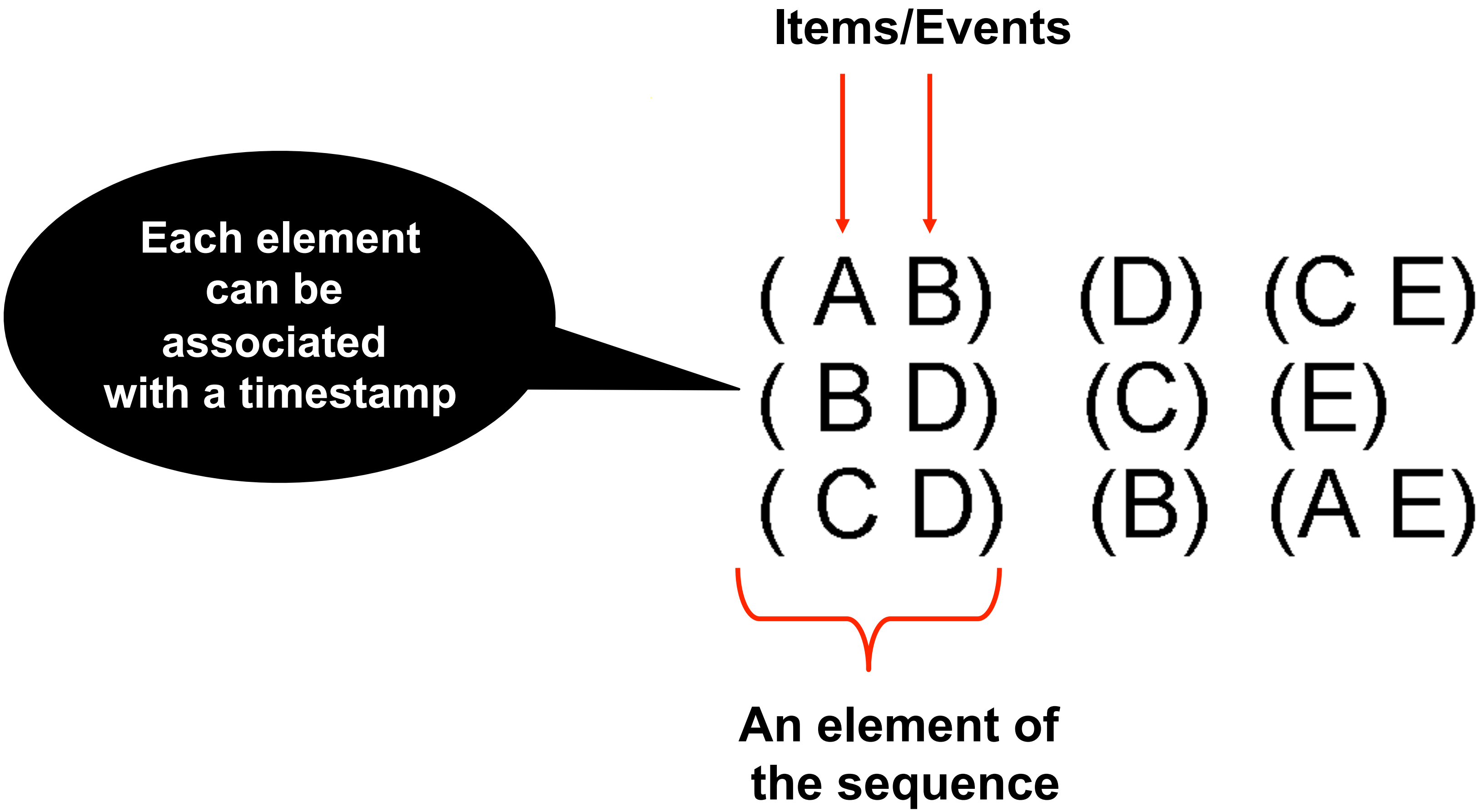
Graph Data

- The data objects themselves are represented as graphs
  - Example: Benzene Molecule: C_6H_6



Ordered Data

- **Sequential data**
  - Example: Sequences of transactions (temporal data)



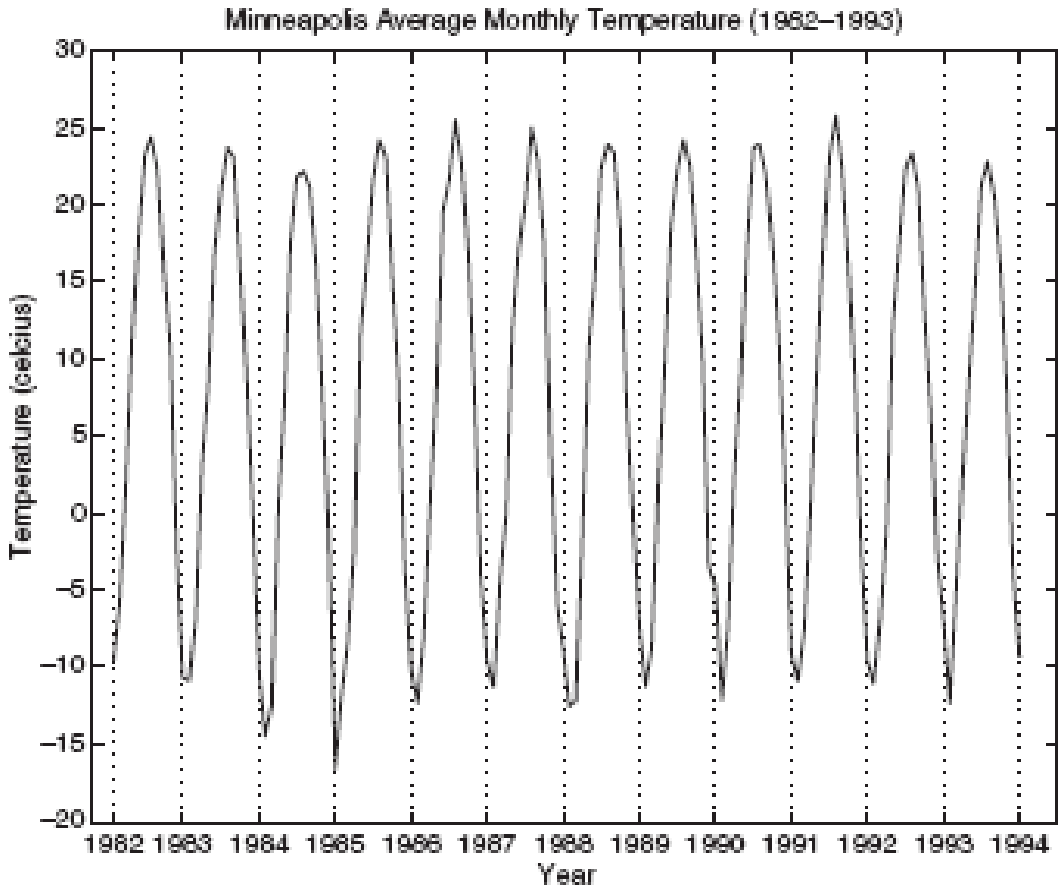
Ordered Data

- **Sequence data**
  - Example: Genomic sequence expressed in term of the four nucleoutides from which all DNA is constructed: A, T, G, C

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

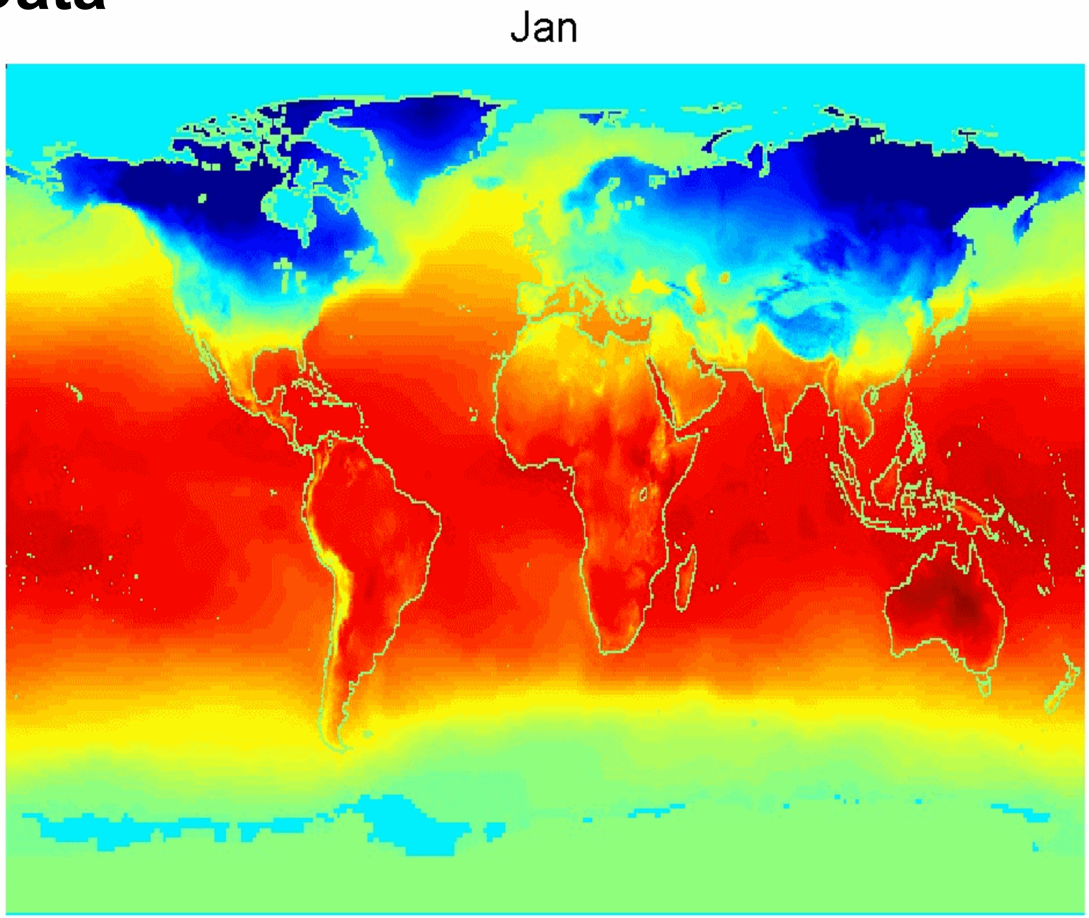
Ordered data

- **Time series**



Ordered Data

- **Spatio-Temporal Data**



Average Monthly Temperature of land and ocean

Data Quality

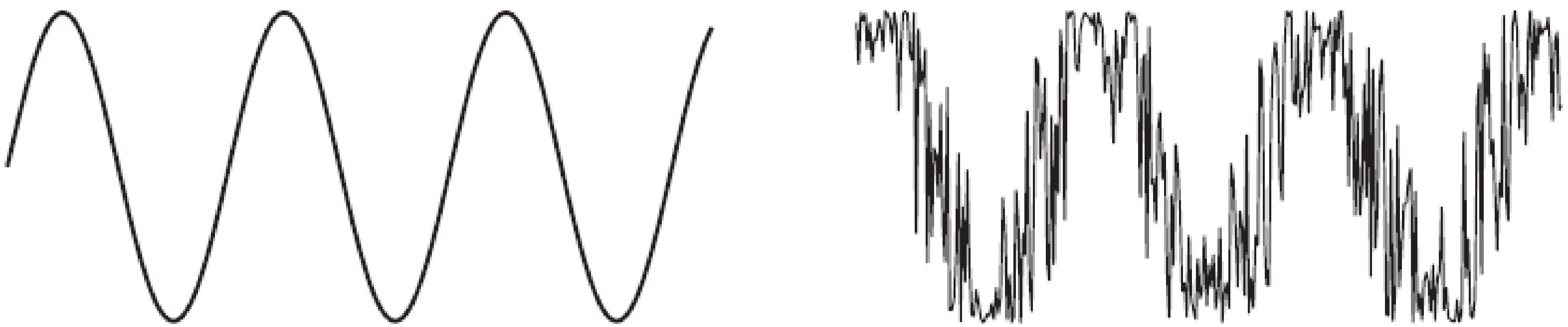
- **What kinds of data quality problems?**
- **How can we detect problems with the data?**
- **What can we do about these problems?**

- **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 and "snow" on television screen
- Due to measurement and data collection errors

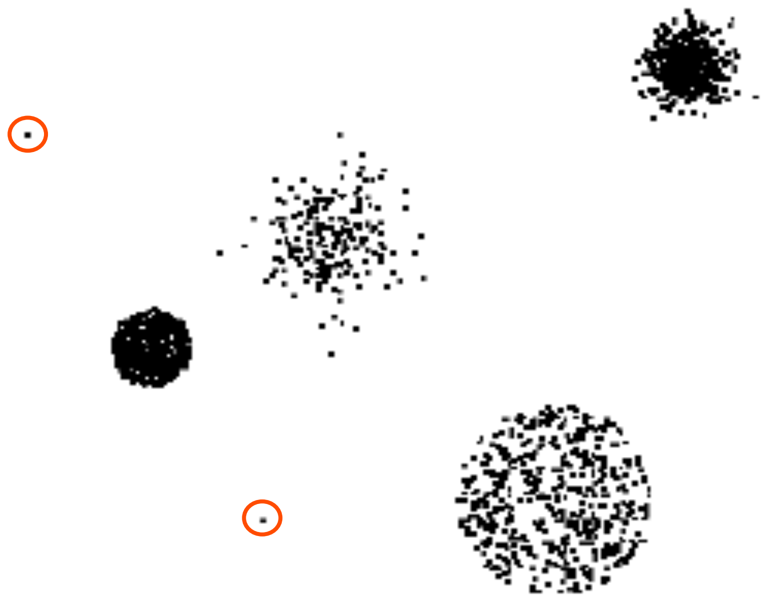
- Noise is the random component of a measurement error
  – Example: a time series disrupted by random noise



(a) Time series. (b) Time series with noise.

Outliers

- **Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set**



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
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

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:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues
- However, the occurrence of duplicate data may not imply an error
  - Consider two distinct customers that buy exactly the same combination of items in a market, thus generating two identical transactions

Data Preprocessing

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

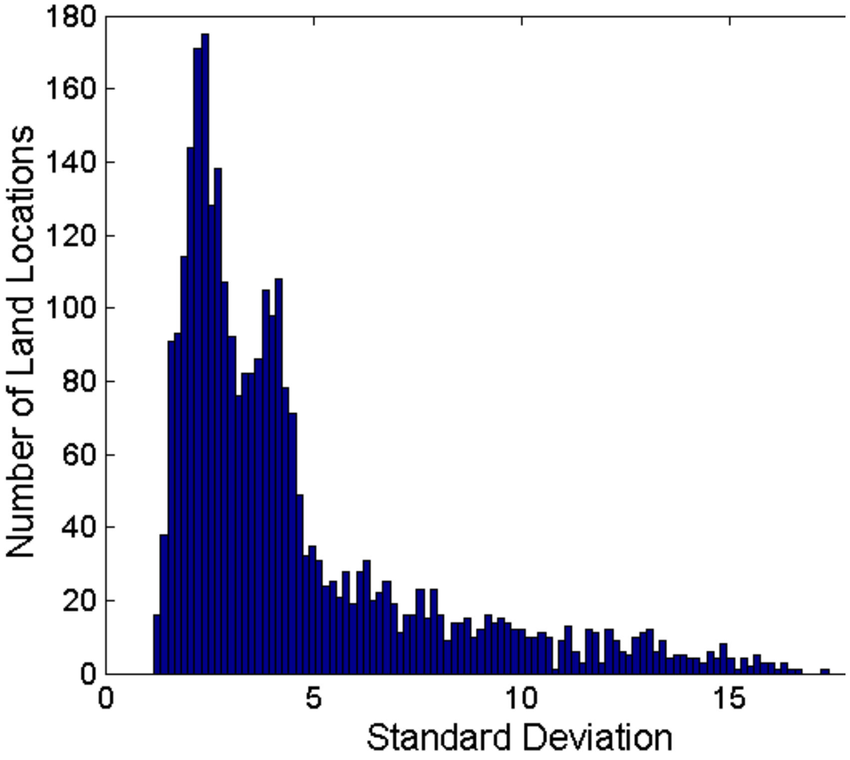
Aggregation

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

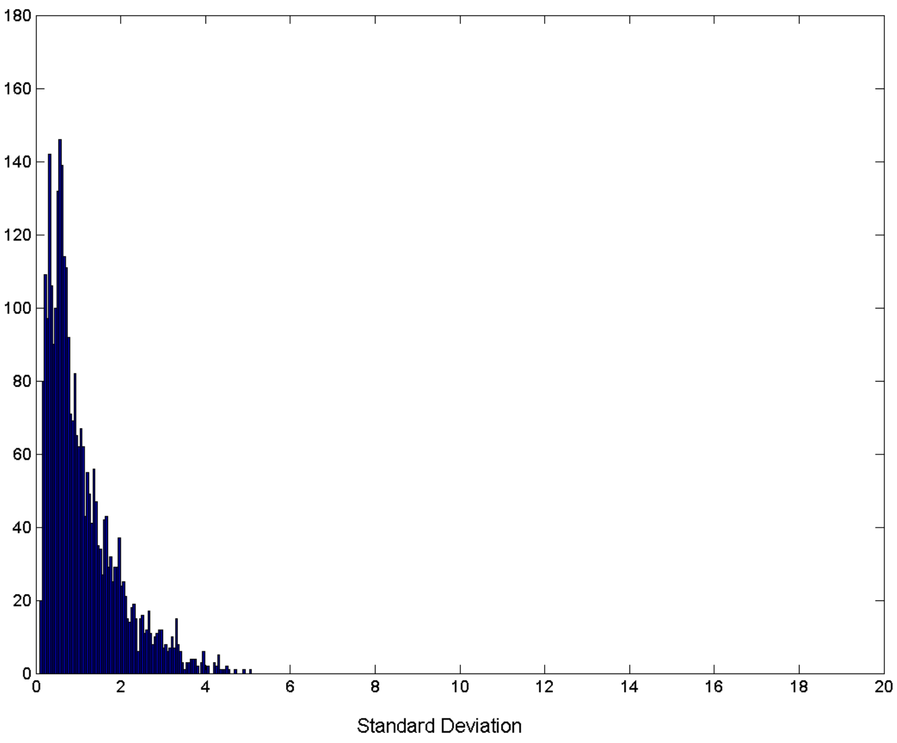
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

Aggregation

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

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.**

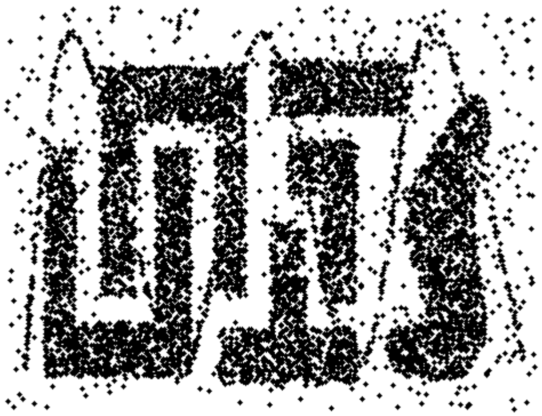
Sampling …

- **The key principle for effective sampling is the following:**
  - **using a sample will work almost as well as using the entire data 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**

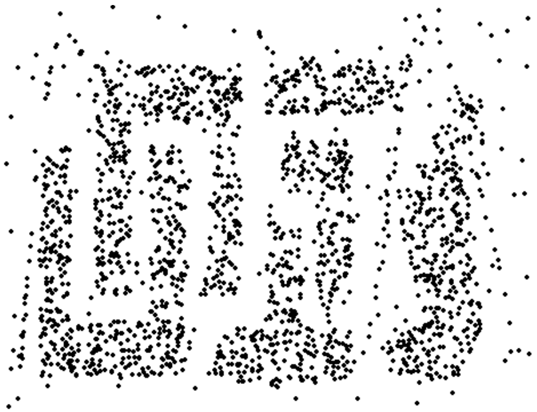
Types of Sampling

- **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

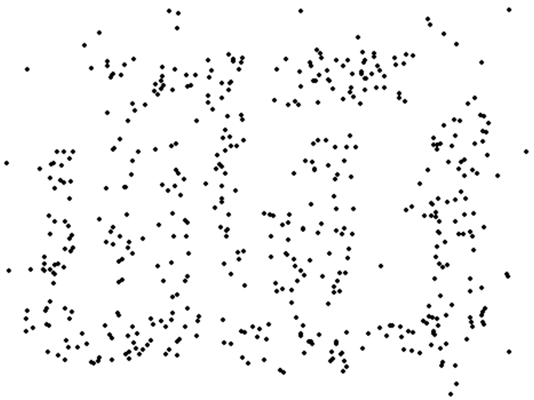
Sample Size



8000 points



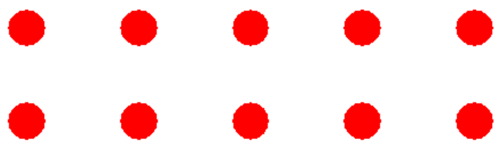
2000 Points

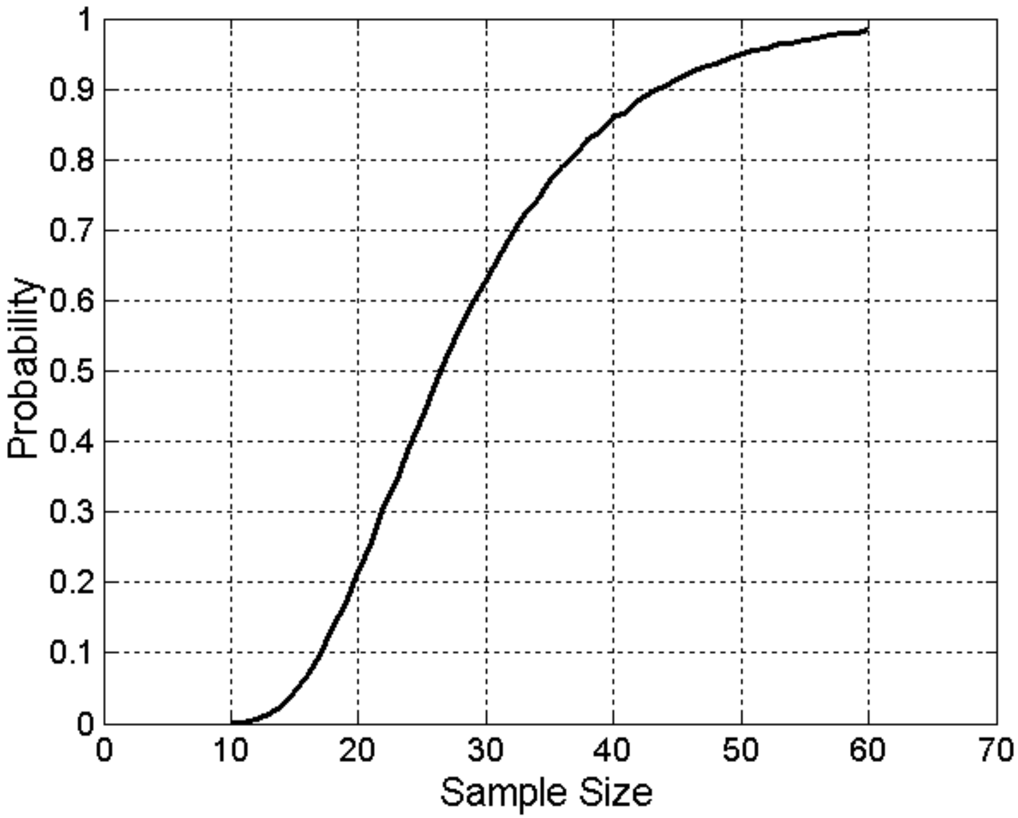


500 Points

Sample Size

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

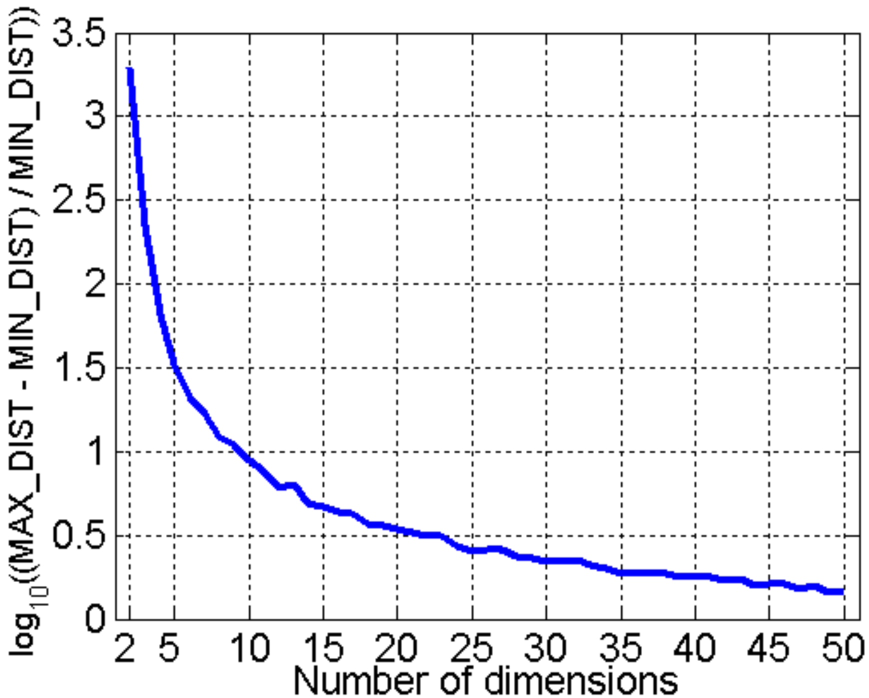




Curse of Dimensionality

- **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**

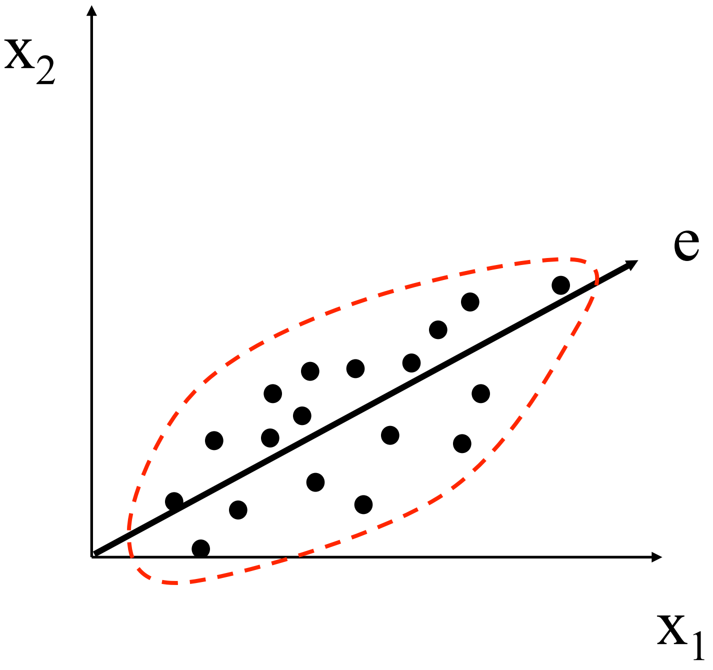
Dimensionality Reduction

- **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**
  - Principle Component Analysis
  - 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**



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 often irrelevant to the task of predicting students' GPA (grade point average)

Feature Subset Selection

- **Techniques:**
  - **Brute-force approch:**
    - **Try all possible feature subsets as input to data mining algorithm: impractical**
  - **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, but typically without enumerating all possible subsets**
- **Feature Weighting is a related approach**

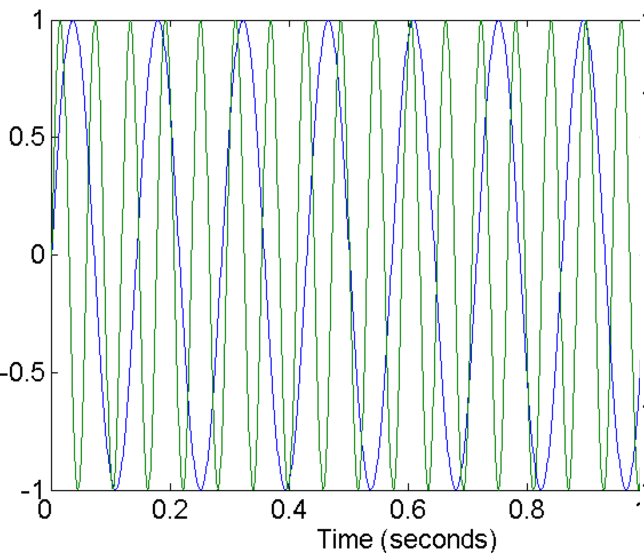
Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

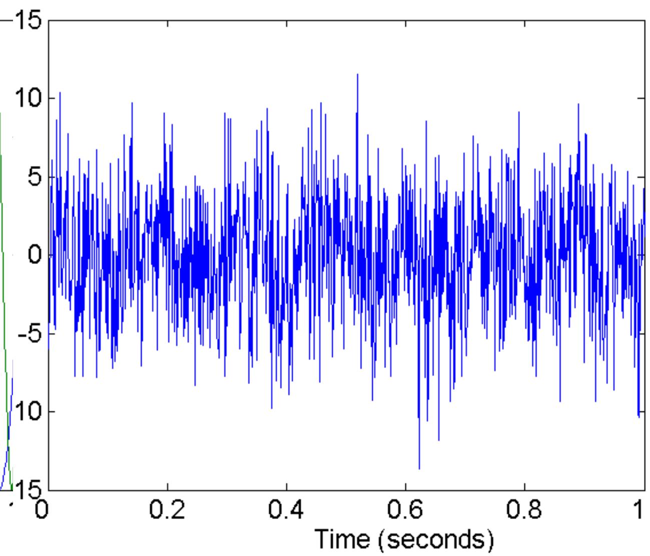
- Three general methodologies:
  - Feature Extraction
    - domain-specific
  - Mapping Data to New Space
  - Feature Construction
    - combining features (usually performed by domain expert)

Mapping Data to a New Space

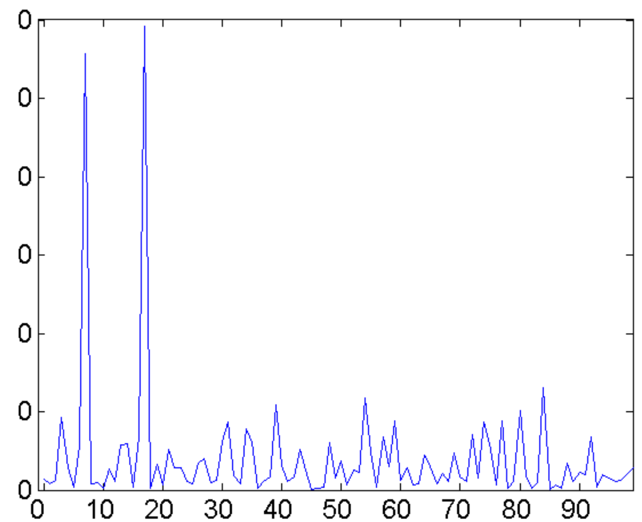
- **Fourier transform**
- **Wavelet transform**



Two Sine Waves



Two Sine Waves + Noise



Frequency (Fourier)

Discretization and Binarization

- **Discretization**
  - Transformation of continuous attributes into categorical ones
- **Binarization**
  - Transformation of continuous/categorical attributes into into one or binary attributes
- **As for feature selection, the best discretization/binarization is the one that produce the best result for the DM algorithm used to analyze the data**

Binarization

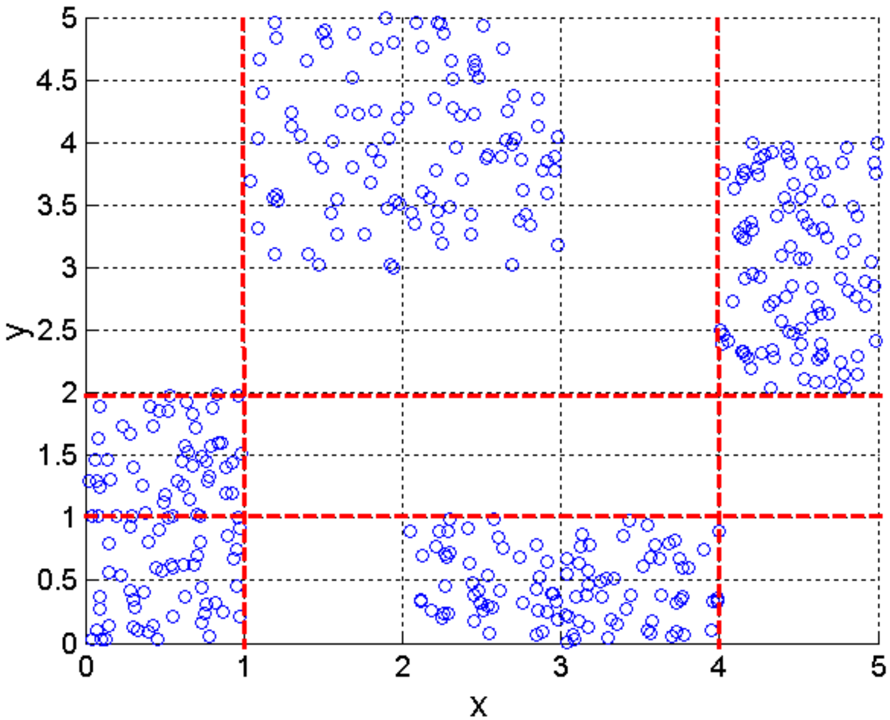
Categorical / Ordinal Value	Integer Value	x_0	x_1	x_2
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

- **For some algorithms (like association rules) the binary attributes are asymmetric: only 1 (presence) is important**
- **So another binarization must be used:**

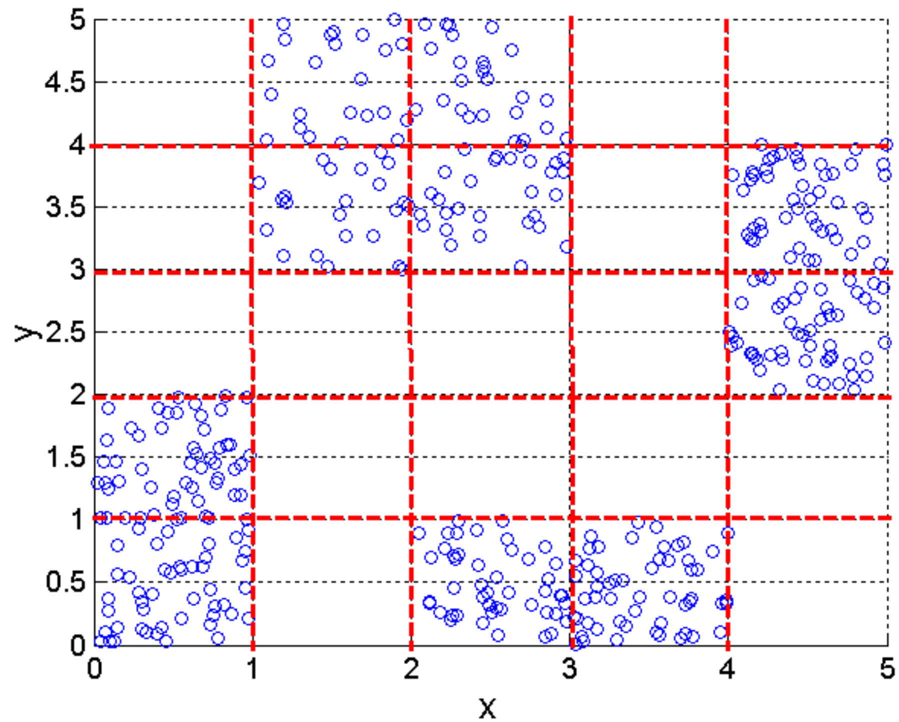
Categorical / Ordinal Value	Integer Value	x_0	x_1	x_2	x_3	x_4
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Discretization Using Class Labels

- Entropy based approach
  - 4 classes of objects, for each interval compute the entropy (intuitively the purity of an interval)
  - Incrementally split each dimension, choosing the best split and recomputing the entropy
  - With five intervals it is better than 3 intervals. With 6 intervals the entropy becomes worse



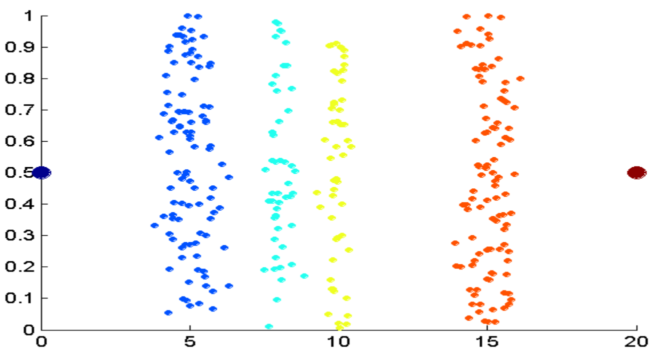
3 intervals for both x and y



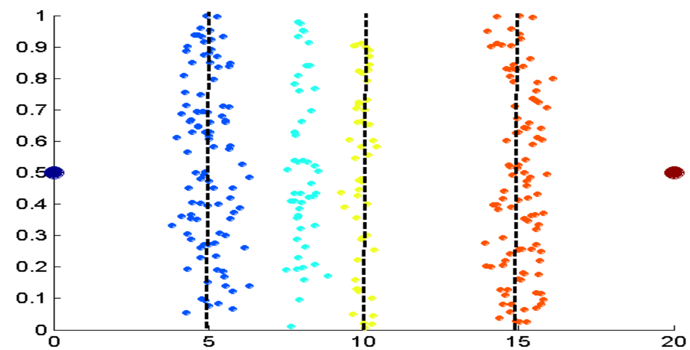
5 intervals for both x and y

Discretization Without Using Class Labels

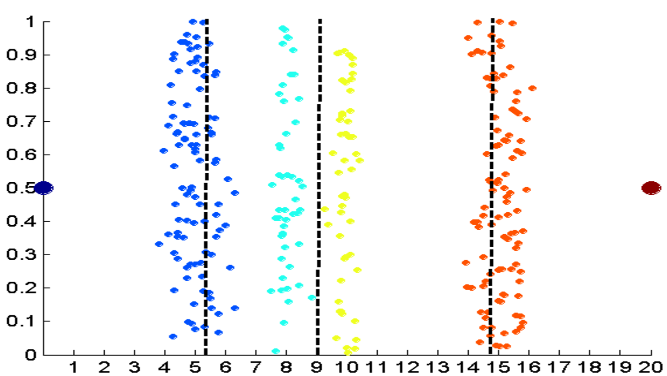
- **If we measure the performance of a discretization technique by the extent to which different objects in different groups are assigned to the same intervals, k-means (clustering) is the best, then equi-depth, and finally equi-width**



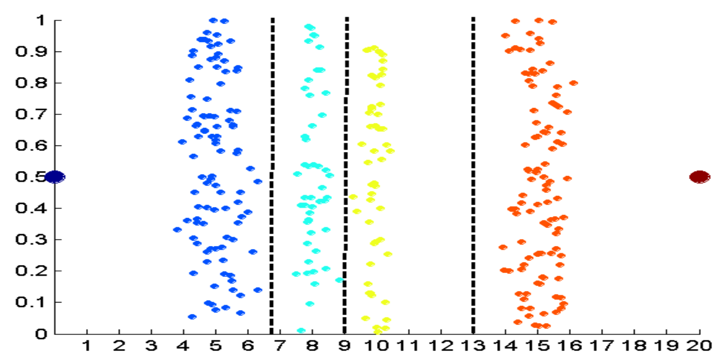
Data



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, $|x|$
  - In statistics variable transformations often force a Gaussian distribution
  - Standardization and Normalization
  - to avoid dependendencies from the single units of measure, when multiple attributes are combined

Standardization of numerical attributes

- For **interval-scaled** attributes, values are measured on a linear scale
  - The difference are always significant and independent of the absolute values of the attributes
- Compute the absolute average deviation, where m_f is the mean:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|) \qquad m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf}).$$

- **Standardized measure** (*z-score*) to replace the original one

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

$z_{if}=0$ if $x_{if}=m_f$

$z_{if}=\pm 1$ if $|x_{if}-m_f| = |s_f|$

- Use of absolute average deviation is more robust
  - $|x_{if} - m_f|$ is not squared, thus reducing issues due to outliers
- The median rather than mean could be used to reduce the outlier issues

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
  - The distance is a measure of dissimilarity

- **Proximity** refers to a similarity or dissimilarity
  - Measures of proximity are used in clustering, classification (kNN), or outlier detection

Similarity/Dissimilarity for Simple Attributes

p and *q* are the single 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	$d = \frac{\vert p-q \vert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\vert p-q \vert}{n-1}$
Interval or Ratio	$d = \vert p - q \vert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

Euclidean Distance between two *n*-vectors

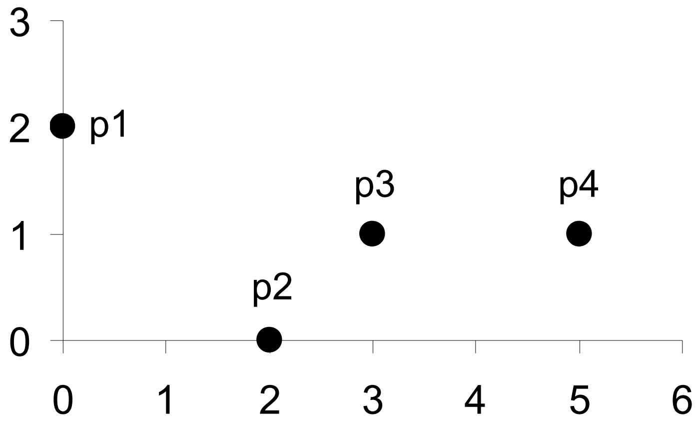
- **Euclidean Distance to compute the distance between two vectors with numerical attributes**

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where *n* is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^th attributes (components) or data objects *p* and *q*.

- **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 between two *n*-vectors

- **Minkowski Distance is a generalization of Euclidean Distance**

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where *r* is a parameter, *n* is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects *p* and *q*.

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_{∞} 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_∞	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. $d(p, q) \geq 0$ 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 defined as metric

Are all the distances metric?

- There exist non-metric distances
- Example: distance between two time attributes:
  - The same day or consecutive days
  - $d(1PM, 2PM) = 1$ *hours*
  - $d(2PM, 1PM) = 23$ *hours*

  } Symmetry does not hold
- Example: ***set difference***
  - $d(A,B) = |A-B|$
  - If $A=\{1,2,3,4\}$ and $B=\{2,3,4\}$
    - $d(A, B) = |A-B| = |\{1\}| = 1$
    - $d(B, A) = |B-A| = |emptyset| = 0$

    } Symmetry does not hold
- A common way to make a measure symmetric is the following:
  - $d'(x, y) = d'(y, x) = (d(x, y) + d(y, x)) / 2$

Common Properties of a Similarity

- Similarities, also have some well known properties.

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

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

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

Similarity Between Binary Vectors

- A 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)

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

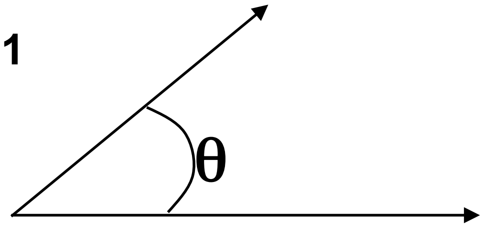
$$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$$

Cosine Similarity between *n*-vectors

- 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 • is the vector dot product, and $||d||$ is the length of vector d.
- COSINE of the angle θ between the two vectors, independent of the lenght of the vectors
  - 90° = similarity equal to 0 0° = similarity equal to 1



- Example:

d_1 = 3 2 0 5 0 0 0 2 0 0
d_2 = 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

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

Pearson's Correlation between *n*-vectors

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

Pearson's Correlation between *n*-vectors

- **A correlation of 1 (-1) means that *x* and *y* have a positive (negative) correlazione**
  - **Ex. of negative correlation = -1 (mean = 0):**
    x = (-3, 6, 0, 3, -6)
    y = (1, -2, 0, -1, 2)

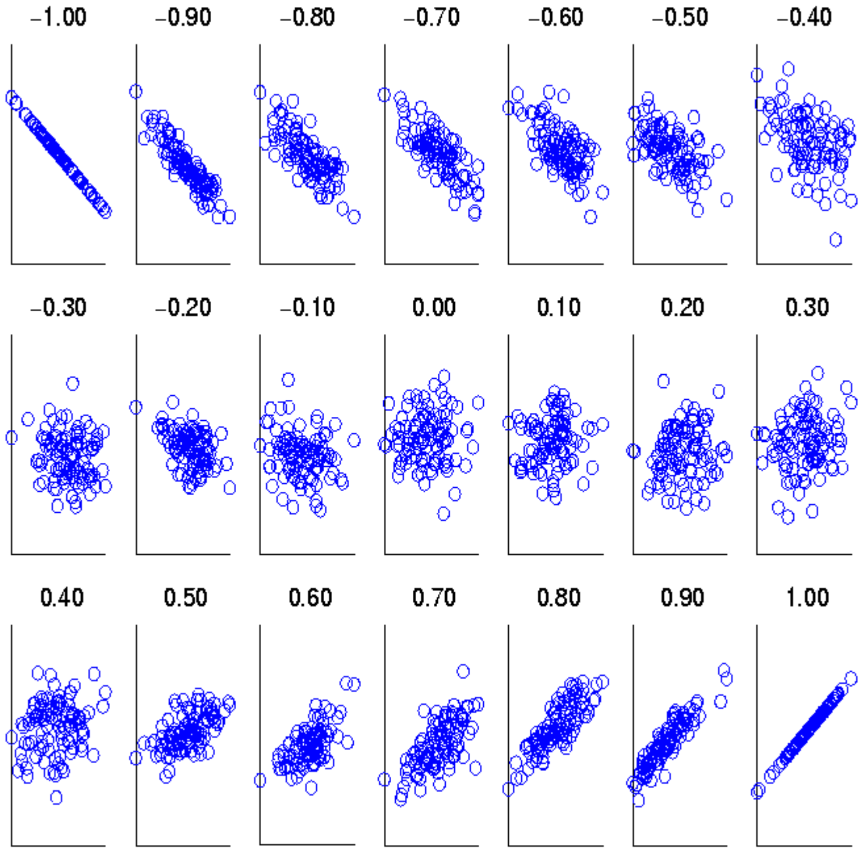
    $x_k = -3\ y_k$
  - **Ex. of positive correlation = 1:**
    x = (3, 6, 0, 3, 6)
    y = (1, 2, 0, 1, 2)

    $x_k = 3\ y_k$
  - **Example of correlation = 0:**
    x = (-3, -2, -1, 0, 1, 2, 3)
    y = (9, 4, 1, 0, 1, 4, 9)

    $y_k = x_k^2$

- **The correlation is used to compare the ranks of two customers for a set of *n* items:**
  - **Two customers are similar even if their max/min rank is different (see the example with correlation = 1)**

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1

- ***21* pairs of random vectors, with 30 attributes**
- **Each plot corresponds to a pair of vector attribute**
  - **Each point corresponds to the standardized components x_i and y_i of the vectors**

General Approach for Combining Similarities

- **Sometimes attributes are of many different types, but an overall similarity is needed.**

1. For the k^{th} attribute, compute a similarity, s_k, in the range $[0, 1]$.
2. Define an indicator variable, δ_k, for the k_{th} attribute as follows:

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

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

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

- **Note that the denominator is greater or equal to the numerator, and corresponds to the number of <u>attribute that exist in both</u> the vectors, and that are <u>not binary asymmetric</u>**

General Approach for Combining Similarities

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

- **If the attribute p_k and q_k are binary or nominal:**

  $s_k = 1$ **if** $p_k = q_k$ **,** $s_k = 0$ **otherwise**

- **If the attribute p_k and q_k are numeric interval-scaled:**

  $s_k = 1 / (1 + |p_k - q_k|)$

- **If the attribute p_k and q_k are ordinal or ratio-scaled, substitute with the a rank in $[0, n\text{-}1]$ and then compute the similarity:**

  $s_k = 1 - |p_k - q_k| / (n\text{-}1)$

Using Weights to Combine Similarities

- **May not want to treat all attributes the same.**
  - **Use weights w_k which are between 0 and 1 and sum to 1.**

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

$$distance(p, q) = \left(\sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}$$

Density

- Density-based clustering require a notion of density
- Examples:
  - Euclidean density
    - Euclidean density = number of points per unit volume
  - Probability density
  - Graph-based density

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**

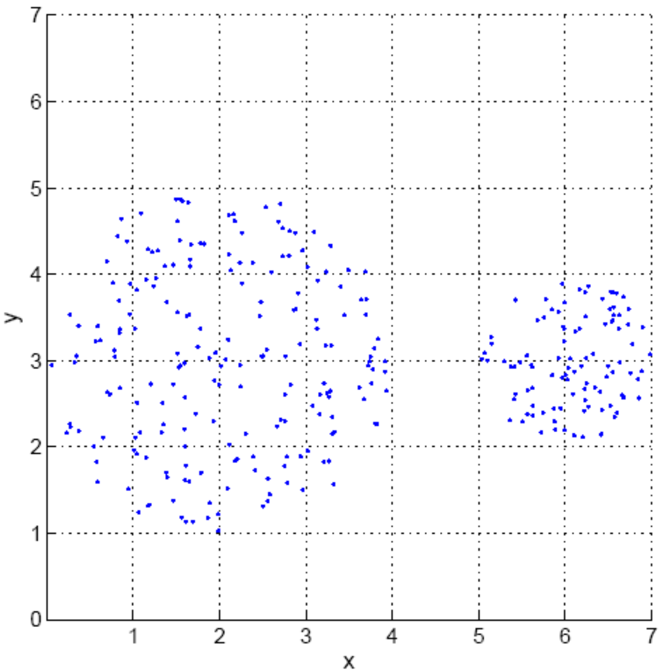


Figure 7.13. Cell-based density.

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**

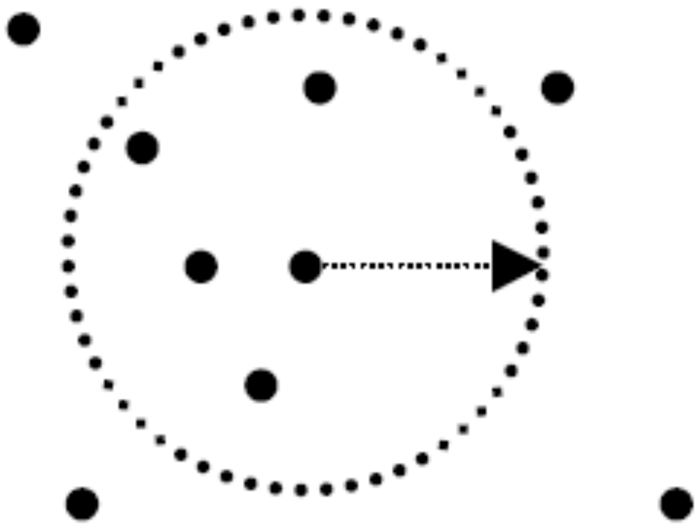


Figure 7.14. Illustration of center-based density.