

Data

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



<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

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

Attribute Types

- ▶ There are different types of attributes
 - ▶ **Nominal**
 - ▶ Examples: ID numbers, eye color, zip codes
 - ▶ **Ordinal**
 - ▶ Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
 - ▶ **Interval**
 - ▶ Examples: calendar dates, 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: $=$ \neq
- ▶ 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. ($=$, \neq)	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. ($+$, $-$)	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.

Types of Datasets

▶ **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
- ▶ Resolution
 - ▶ Patterns depend on the scale

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 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.

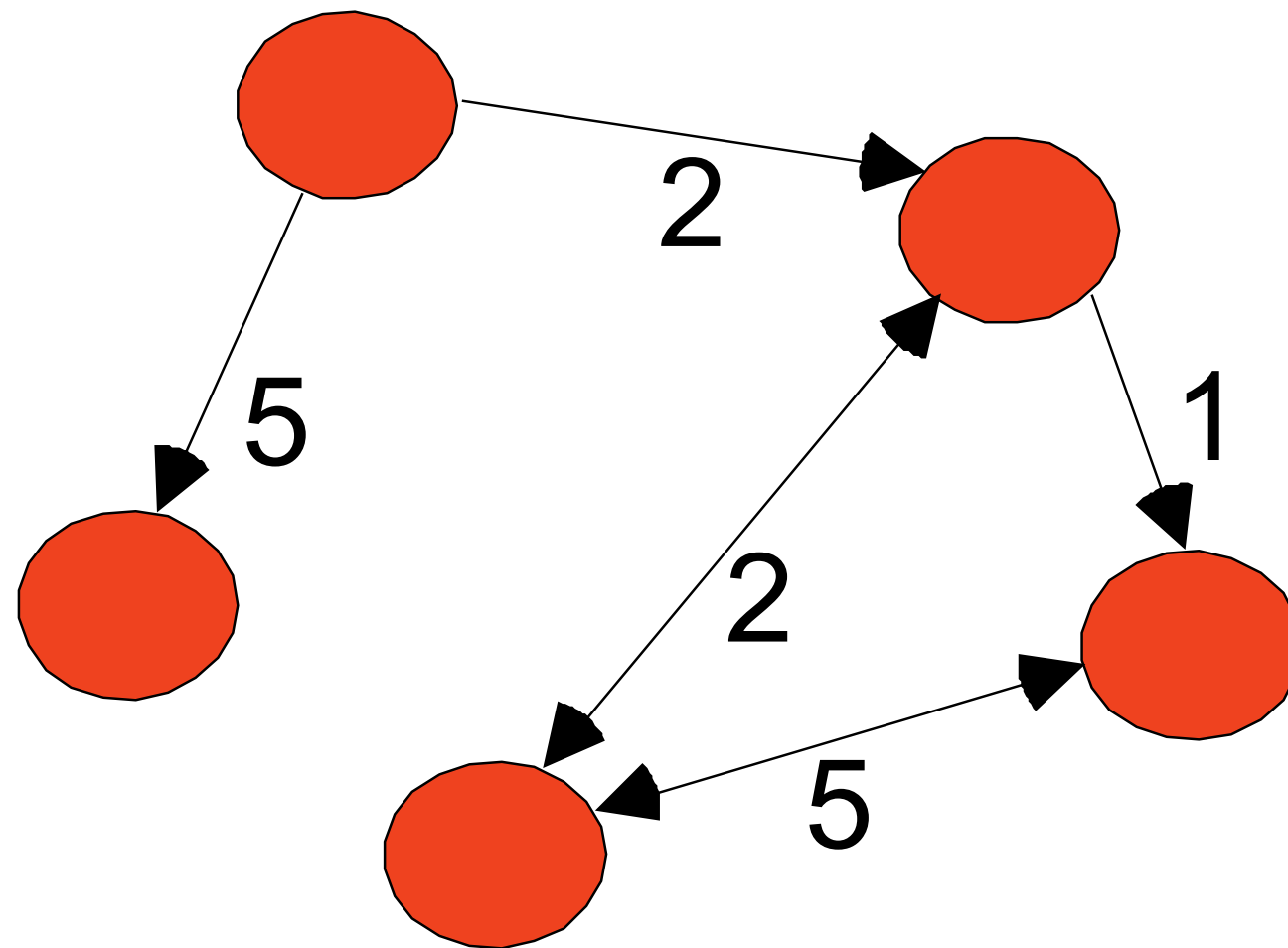
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

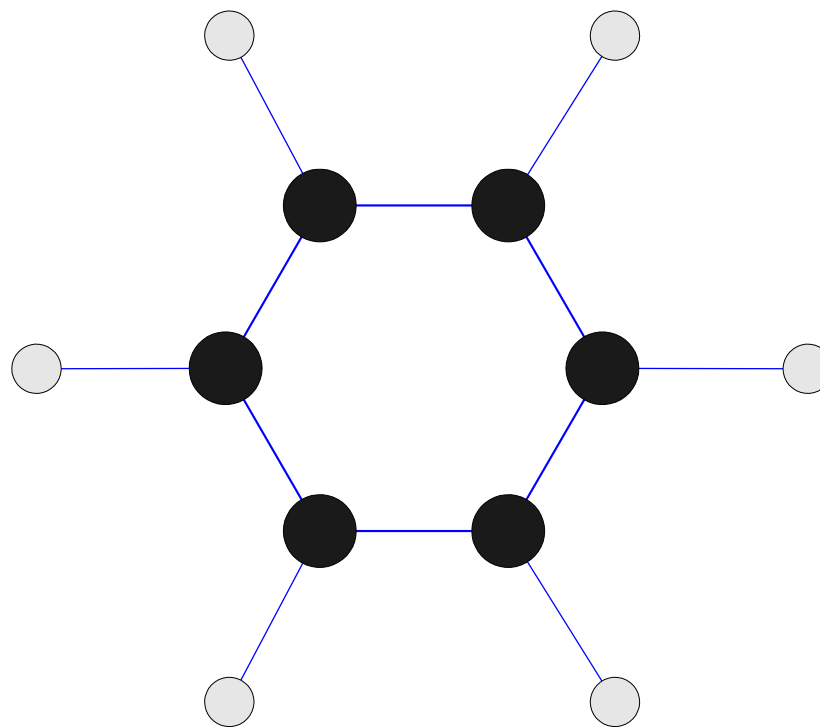
- ▶ A special type of record data, where
- ▶ each record (transaction) involves a set of items.
- ▶ For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

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

Graph Data

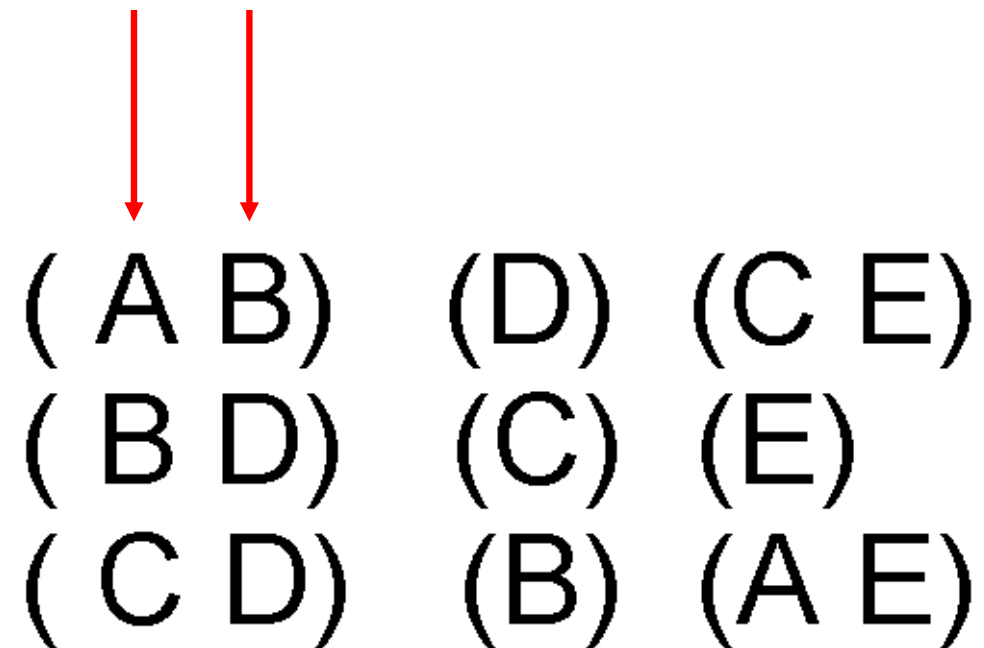


Chemical Structures



Sequential data

Items/Events



(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

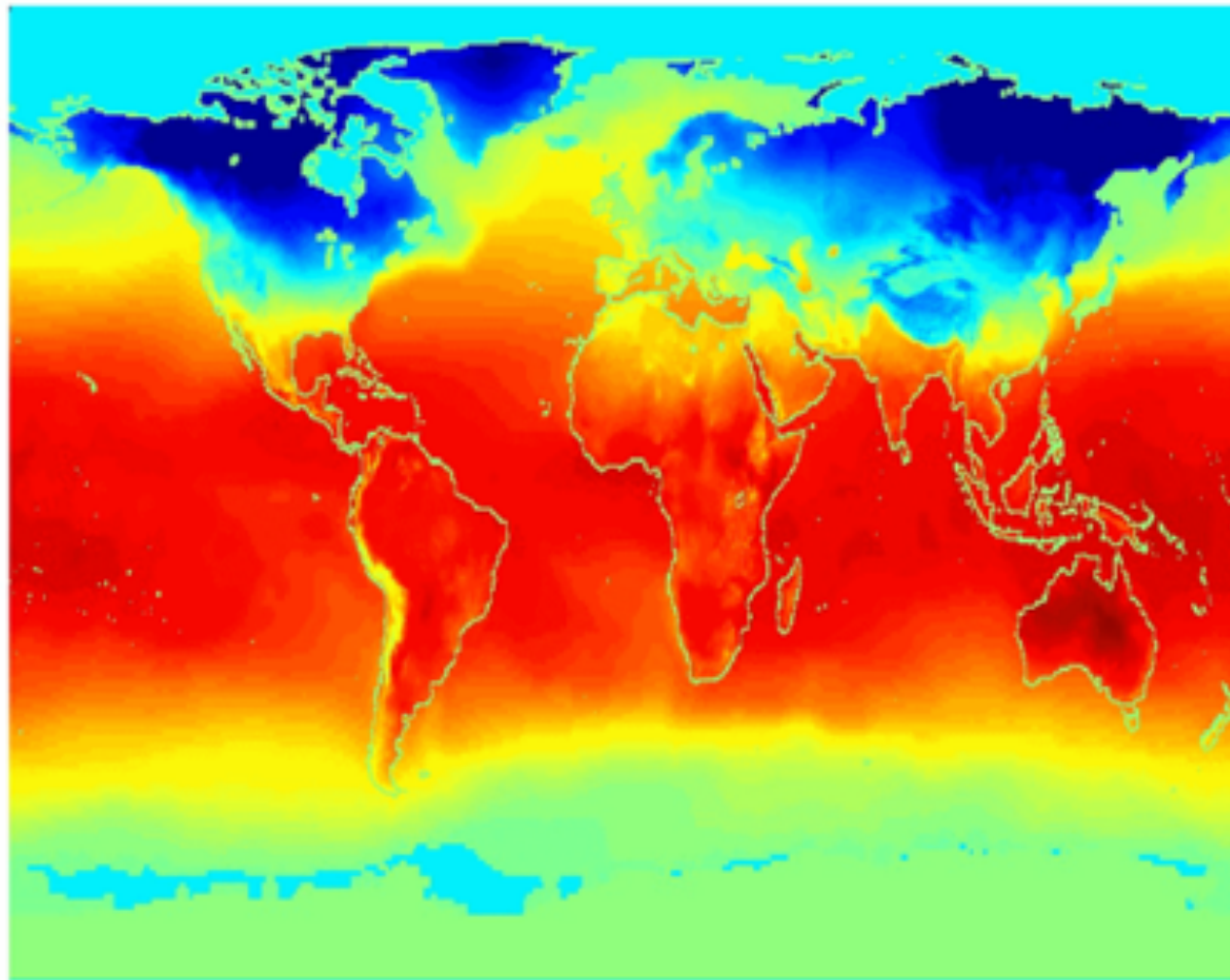
An element of
the sequence

Gene Sequence

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCCGCCTGGCGGGCG
GGGGGAGGCGGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Spatio-temporal Data

Jan



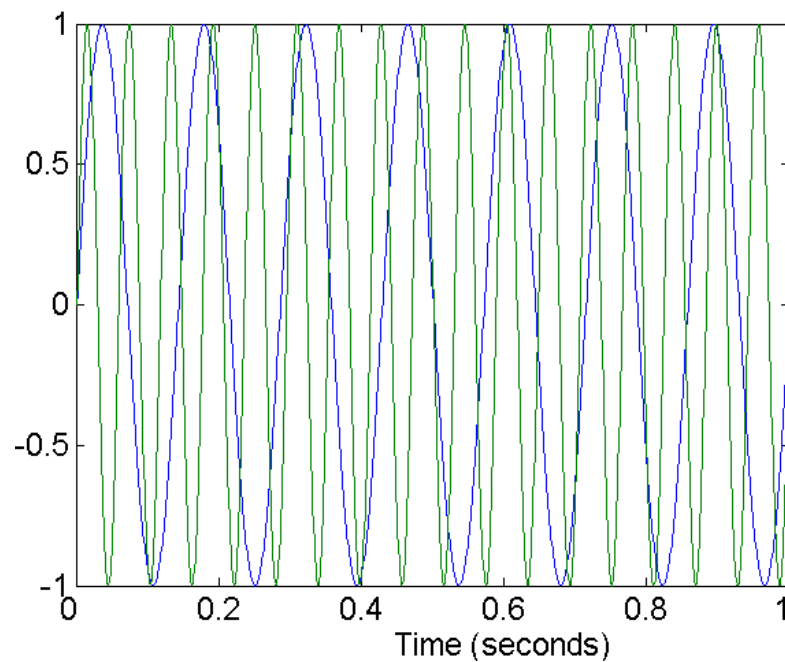
**Average Monthly
Temperature of
land and ocean**

Data Quality

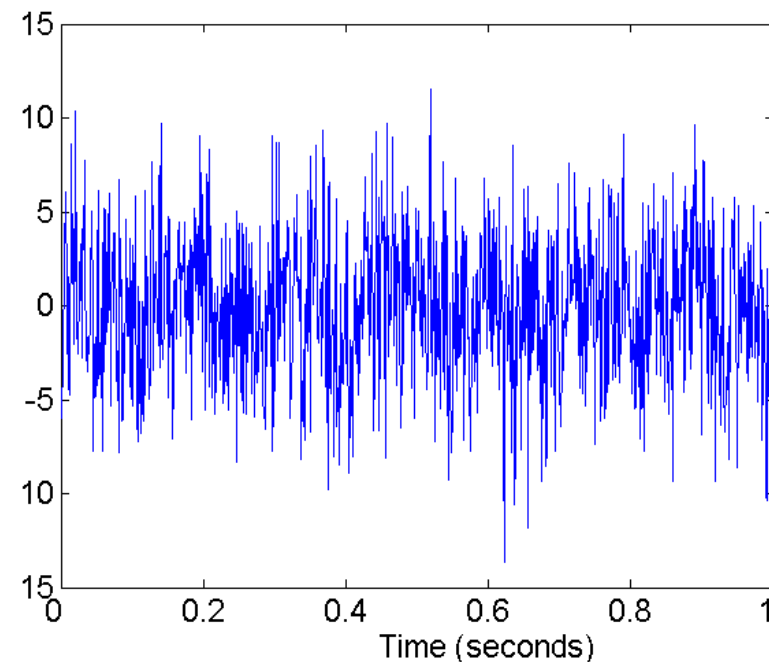
- ▶ What kinds of data quality problems?
- ▶ How can we detect problems with the data?
- ▶ What can we do about these problems?
- ▶ Examples of data quality problems:
 - ▶ Noise and outliers
 - ▶ missing values
 - ▶ duplicate data

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



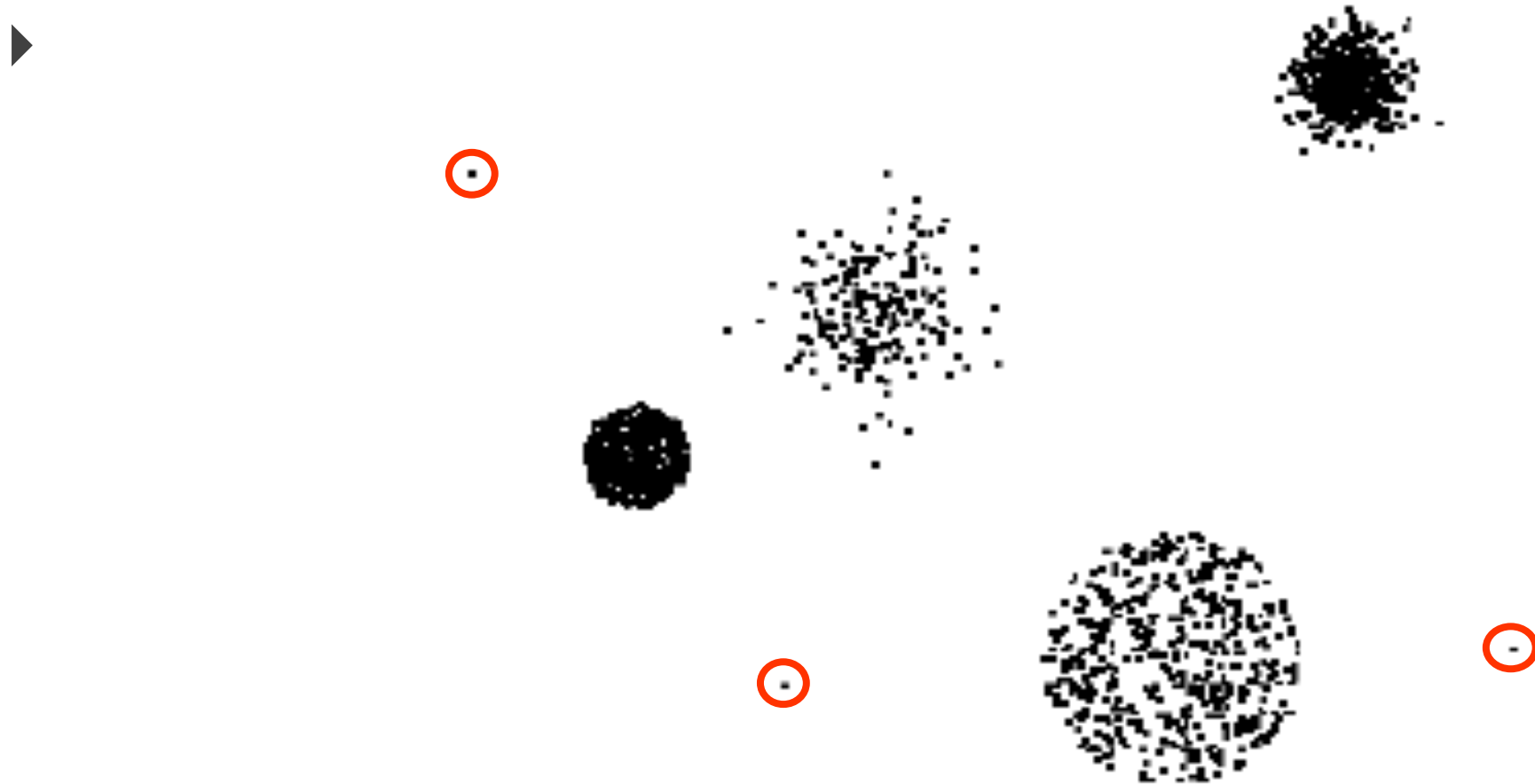
Two Sine Waves



Two Sine Waves + Noise

Outliers

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



-
- ▶ Question: What is the difference between noise and outliers?

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

Distance/Similarity Functions

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 $	$s = -d, s = \frac{1}{1+d} \text{ or } s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

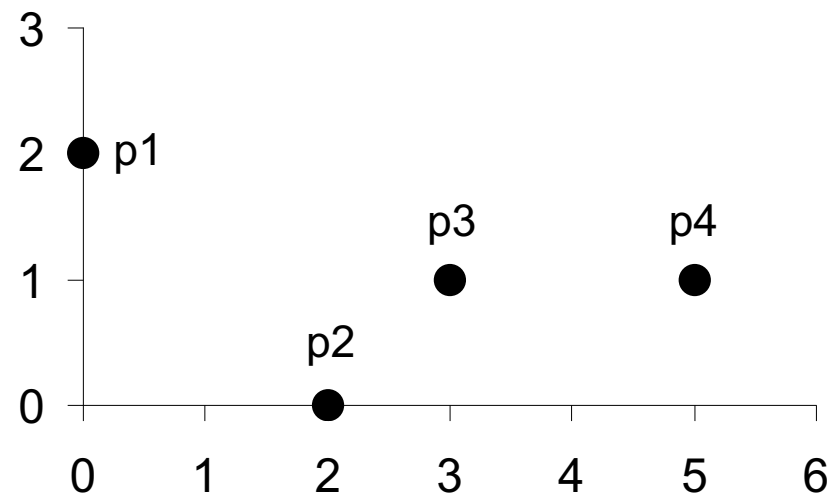
Euclidean Distance

- ▶ Euclidean Distance
- ▶ 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 .

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

- 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

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

- ▶ Minkowski Distance is a generalization of Euclidean Distance
- ▶ Where r is a parameter, 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 .

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

Properties of a Distance Function

- ▶ Distances, such as the Euclidean distance, have some well known properties.
 - ▶ $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$.
(Positive definiteness)
 - ▶ $d(p, q) = d(q, p)$ for all p and q . (Symmetry)
 - ▶ $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q , and r .
(Triangle Inequality)
- ▶ **A distance that satisfies these properties is a metric**

Properties of a Similarity Function

- ▶ 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 .

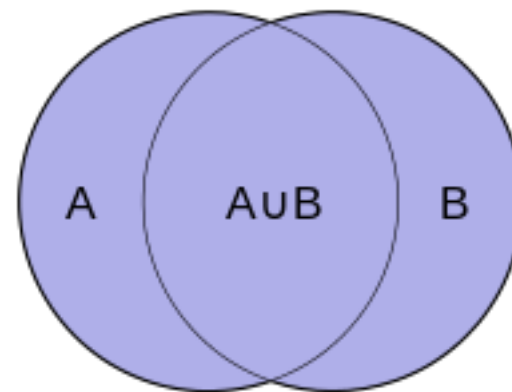
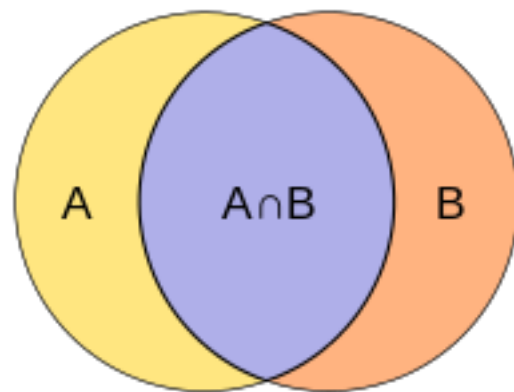
Simple Matching Coefficient

- ▶ 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
 - ▶ $SMC = \text{number of matches} / \text{number of attributes}$
 - ▶ $= (M11 + M00) / (M01 + M10 + M11 + M00)$
 - ▶ $J = \text{number of 11 matches} / \text{number of not-both-zero attributes values}$
 - ▶ $= (M11) / (M01 + M10 + M11)$

Jaccard Similarity/Coefficient

- ▶ Used for categorical attributes
- ▶ Can also be applied on sets

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



Vectors

- ▶ Is an $n \times 1$ matrix
 - ▶ Usually referred to as a lower case letter
 - ▶ n rows
 - ▶ 1 column
- ▶ e.g.

$$y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}$$

Dot product

- Given Two vectors a and b

$$\vec{a} = (a_1, a_2, a_3, \dots) \qquad \vec{b} = (b_1, b_2, b_3, \dots)$$

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^n a_i b_i = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$$

Cosine Similarity

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

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

- ▶ Example:

$$d_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

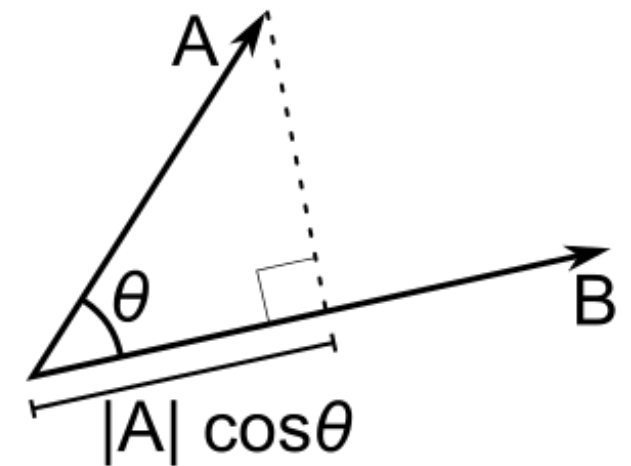
$$d_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

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

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

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

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



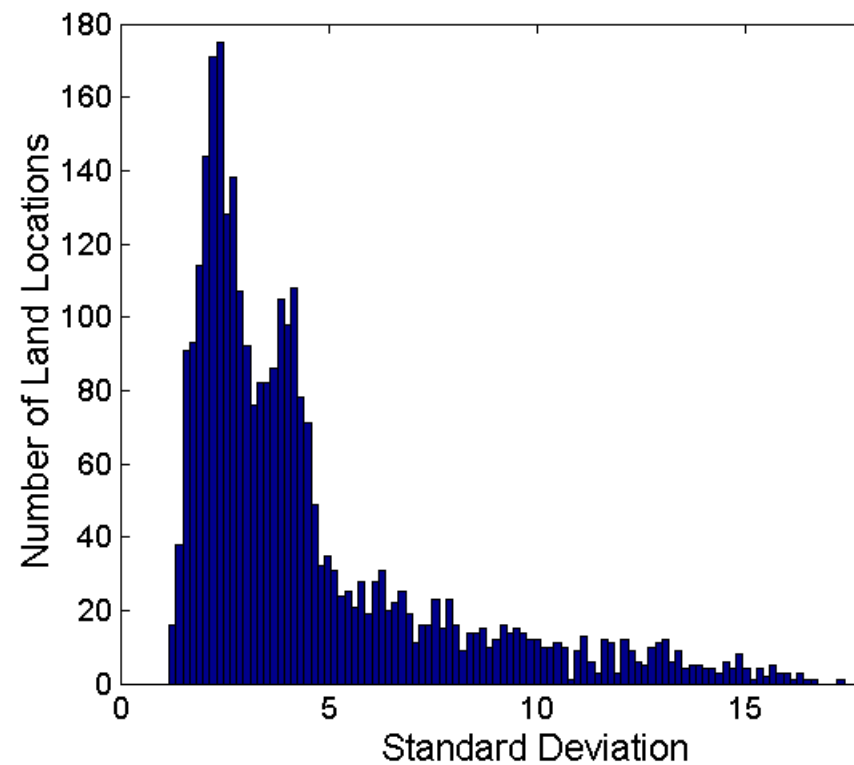
Data Preprocessing

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

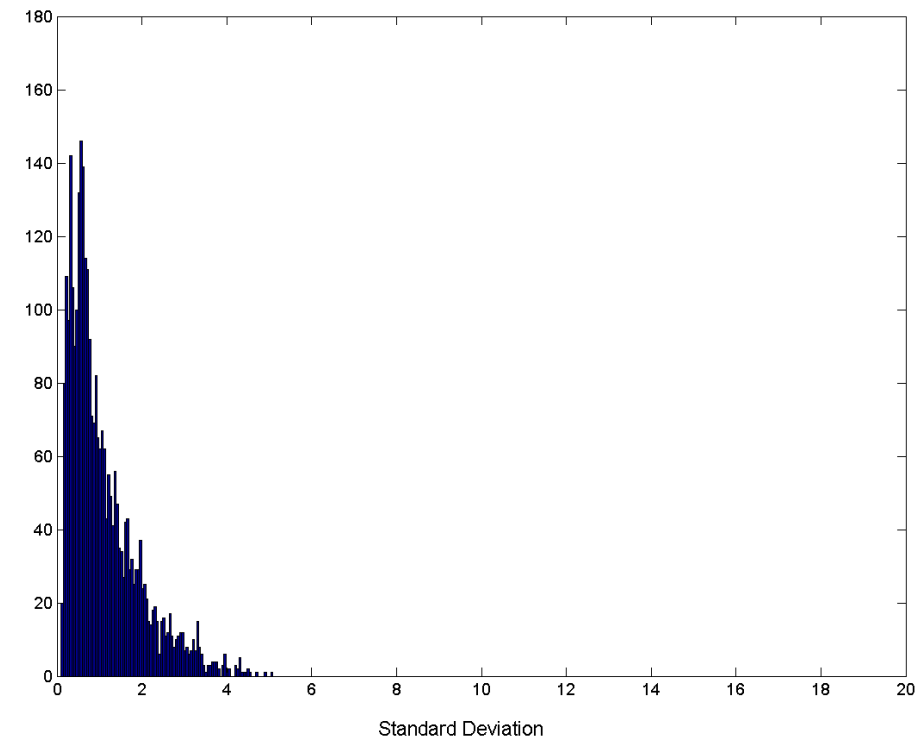
Aggregation

- ▶ 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 Example



**Standard Deviation of Average
Monthly Precipitation**



**Standard Deviation of Average
Yearly Precipitation**

Data 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**

How to Sample?

- ▶ 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**

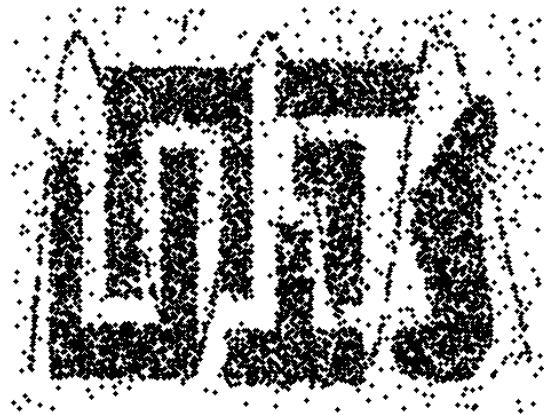
- ▶ Split the data into several partitions; then draw random samples from each partition

-
- ▶ Question: What are the pros and cons of sampling with/without replacement?

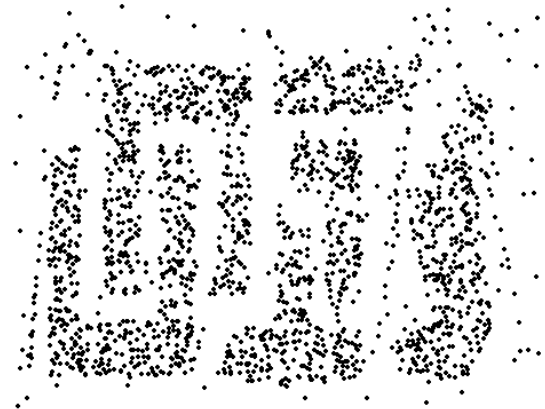
The Issue of Sample Size?

- ▶ Question: Can you suggest a good strategy to choose sample size?

Sample Size



8000 points

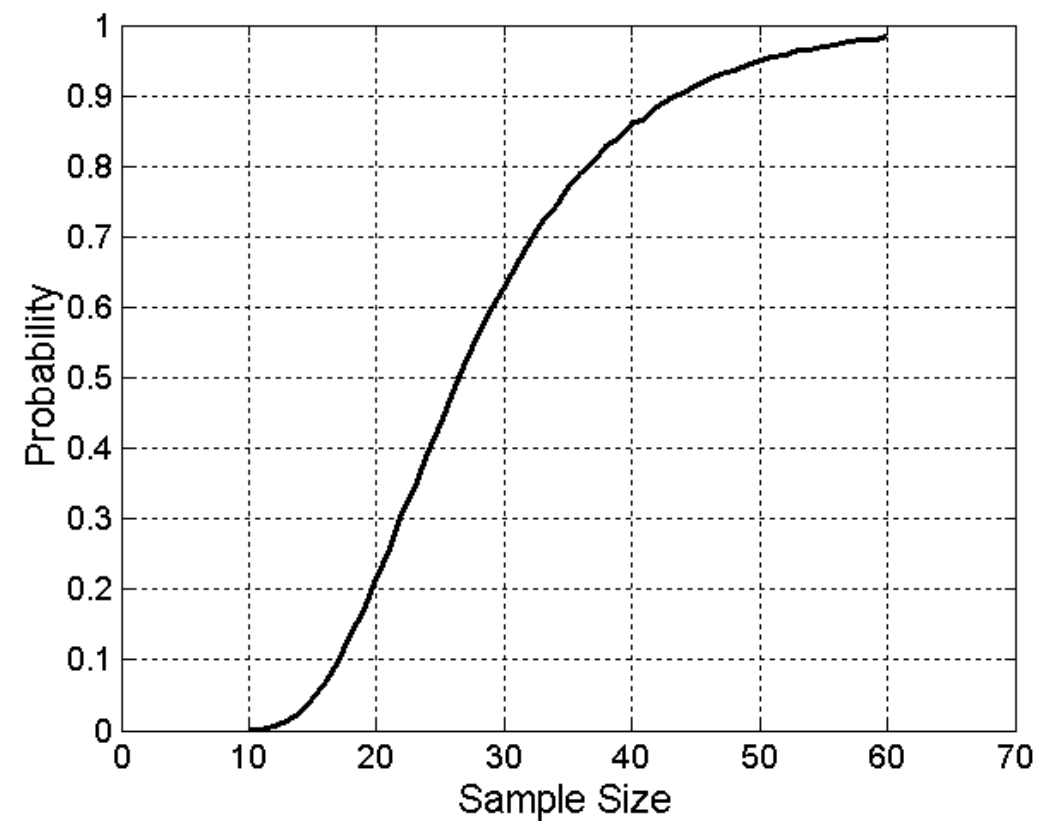
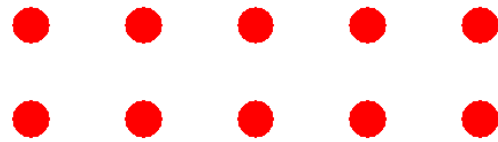


2000 Points



500 Points

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



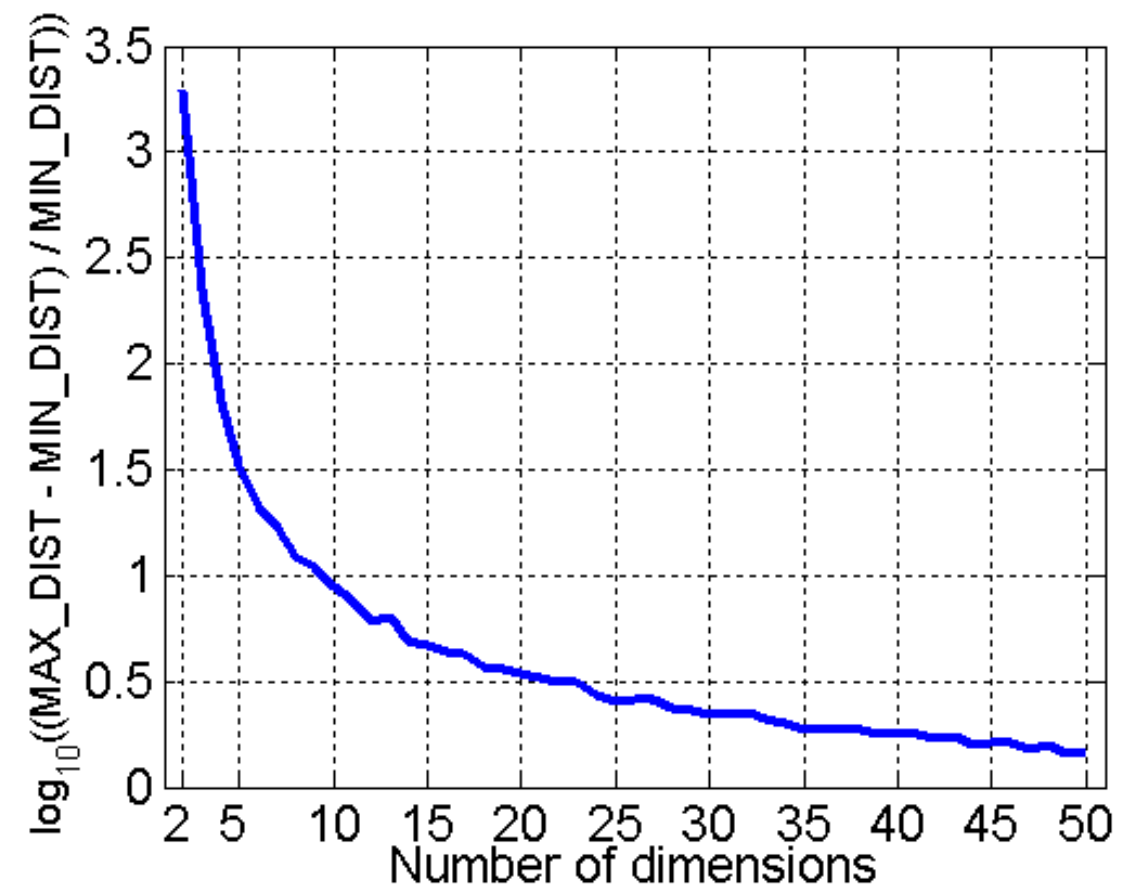
Reservoir Sampling

- ▶ How to sample when we don't know how big the data is? Or if the data is prohibitively huge.
- ▶ Solution keep a reservoir of r samples
 - ▶ Keep the first r items in memory.
 - ▶ When the i^{th} item arrives ($i > r$)
 - ▶ With probability r/i keep the new item (discard an old one, selecting which to replace at random, each with chance $1/r$)
 - ▶ With probability $1 - r/i$, keep old items (discard the new one)

Dimensionality Reduction

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