Data Mining

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

Types of Data

- Outline
  - Data set
  - Attributes and Objects
  - Types of Data
  - Data Quality
  - Similarity and Distance
  - Data Preprocessing

What is Data?

- A data set
  - a collection of data objects.
    - Object is AKA record, point, case, sample, entity, or instance
- Data objects are described by a number of attributes that capture the characteristics of an object
  - Examples
    - eye color of a person, temperature, the mass of a physical object, the time at which an event occurred, etc.

Example

- A sample data set containing student information

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Freshman	3.51	...
1082246	Sophomore	3.62	...

  - Each row corresponds to a student
  - Each column is an attribute that describes some aspect of a student,
    - such as GPA or ID.

What is an Attribute?

- An attribute
  - a property or characteristic of an object that can vary, either from one object to another or from one time to another.
  - For example,
    - eye color varies from person to person,
      - a symbolic attribute with a small number of possible values {brown, black, blue, green, hazel, etc.}
    - the temperature of an object varies over time.
      - a numerical attribute with a potentially unlimited number of values
  - Attribute is AKA variable, field, characteristic, dimension, or feature

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute for a particular object
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
  - But properties of attribute can be different than the properties of the values used to represent the attribute
- To assign numbers or symbols to objects in a well-defined way, we need a measurement scale.

Measurement Scale

- A measurement scale is a rule (function) that associates a numerical or symbolic value with an attribute of an object.
- The process of measurement is the application of a measurement scale to associate a value with a particular attribute of a specific object.
  - For example,
    - we step on a bathroom scale to determine our weight,
    - we classify someone as male or female,
    - we count the number of chairs in a room to see if there will be enough to seat all the people coming to a meeting
  - In all these cases, the physical value of an attribute of an object is mapped to a numerical or symbolic value.

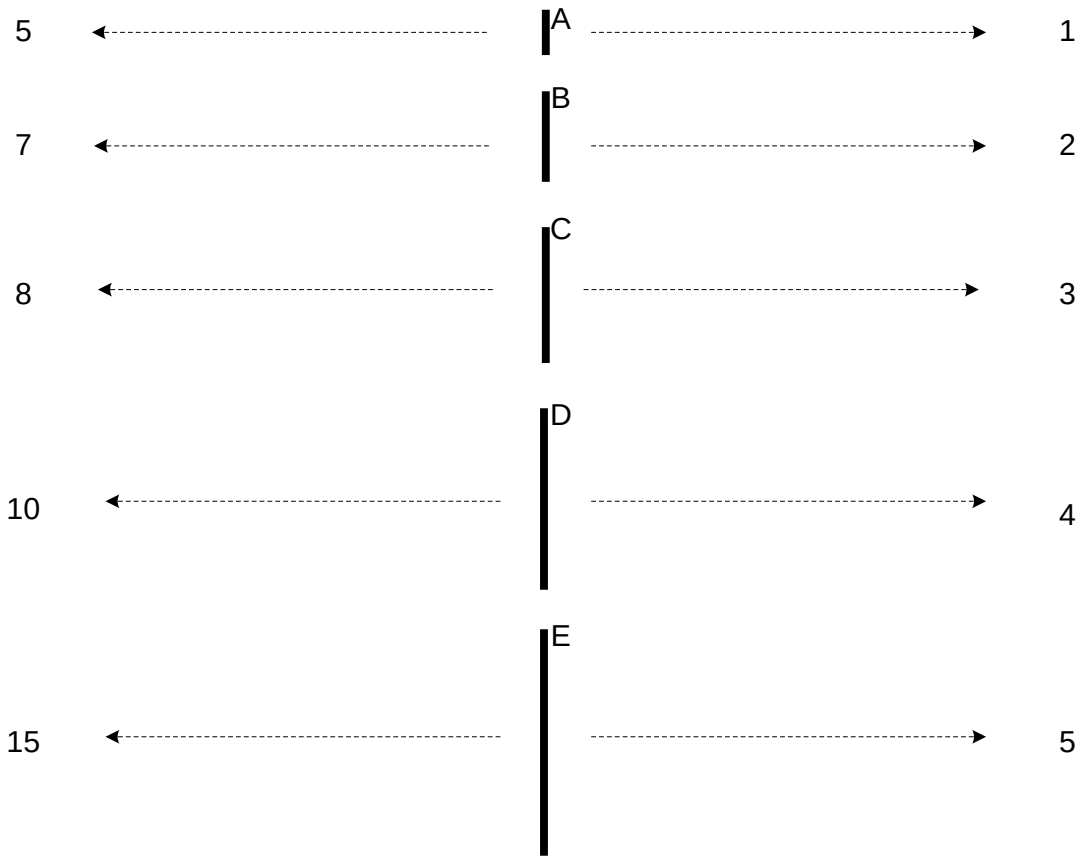
The Type of an Attribute

- It is common to refer to the type of an attribute as the type of a measurement scale.
- The values used to represent an attribute can have properties that are not properties of the attribute itself, and vice versa.
  - For example, two attributes that might be associated with an employee are ID and age (in years).
    - Both of these attributes can be represented as integers.
    - However, while it is reasonable to talk about the average age of an employee, it makes no sense to talk about the average employee ID.

The Type of an Attribute

– The measurement of the length of line segments on two different scales of measurement.

This scale preserves only the ordering property of length.



This scale preserves the ordering and additvity properties of length.

Types of Attributes

- There are four types of attributes
  - Categorical-Nominal (Qualitative)
    - The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. (=, ≠)
    - Operations: mode, entropy, contingency correlation, χ^2 test
      - Examples: ID numbers, eye color, zip codes
  - Categorical- Ordinal (Qualitative)
    - The values of an ordinal attribute provide enough information to order objects. (<, >)
    - Operations: median, percentiles, rank, correlation, run tests, sign tests
      - Examples: rankings (e.g., taste of food on a scale from 1-10), grades, height {tall, medium, short}

Types of Attributes

- Numeric-Interval (Quantitative)
  - For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, −)
  - Operations: mean, standard deviation, Pearson's correlation, *t* and *F* tests
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- Numeric-Ratio (Quantitative)
  - For ratio variables, both differences and ratios are meaningful. (×, /)
  - Operations: geometric mean, harmonic mean, percent variation
  - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness: $=, \neq$
  - Order : $<, \leq, >, \geq$
  - Differences are meaningful : $+, -$
  - Ratios are meaningful : $\times, /$
  - Nominal attribute : distinctness
  - Ordinal attribute : distinctness & order
  - Interval attribute : distinctness, order & meaningful differences
  - Ratio attribute : all 4 properties/operations

Difference Between Ratio and Interval

- Is it physically meaningful to say that a temperature of 10° is twice that of 5° on
  - the Celsius scale? ×
  - the Fahrenheit scale? ×
  - the Kelvin scale? √
- Consider measuring the height above average
  - If Ali's height is 3 cm above average and Veli's height is 6 cm above average, then would we say that Veli is twice as tall as Ali?
  - Is this situation analogous to that of temperature?

Categorization of Attributes

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are 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. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

Categorization of Attributes

- The types of attributes can also be described in terms of transformations that do not change the meaning of an attribute
  - For example, the meaning of a length attribute is unchanged if it is measured in meters instead of feet.
- The statistical operations that make sense for a particular type of attribute are those that will yield the same results when the attribute is transformed by using a transformation that preserves the attribute's meaning

Transformations that define attribute levels

	Attribute Type	Transformation	Comments
Categorical Qualitative	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}.
Numeric Quantitative	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.

Describing Attributes by the Number of Values

- Another way of distinguishing between attributes is by the number of values they can take.
- Discrete Attribute
  - Has only a finite or countably infinite set of values
    - Examples: zip codes (categorical), ID nos (categorical), counts (numeric)
  - 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.

Asymmetric Attributes

- For asymmetric attributes, only presence (a non-zero attribute value) is regarded as important
  - Words present in documents
  - Items present in customer transactions
- If we met a friend in the grocery store, would we ever say the following?
  "*I see our purchases are very similar since we didn't buy most of the same things*."
- Binary attributes where only non-zero values are important are called asymmetric binary attributes
  - which is particularly important for association analysis

Critiques of the attribute categorization

- Incomplete
  - Asymmetric binary
  - Cyclical
  - Multivariate
  - Partially ordered
  - Partial membership
  - Relationships between the data

- Real data is approximate and noisy
  - This can complicate recognition of the proper attribute type
  - Treating one attribute type as another may be approximately correct

Key Messages for Attribute Types

- The types of operations you choose should be meaningful for the type of data you have
  - Distinctness, order, meaningful intervals, and meaningful ratios are only four (among many possible) properties of data
  - The data type you see – often numbers or strings – may not capture all the properties or may suggest properties that are not present
  - Analysis may depend on these other properties of the data
    - Many statistical analyses depend only on the distribution
  - In the end, what is meaningful can be specific to domain

Types of Data Sets

- There are many types of data sets, and as the field of data mining develops and matures, a greater variety of data sets become available for analysis.
- Types of data sets:
  - record data,
  - graph-based data,
  - ordered data
- These categories do not cover all possibilities and other groupings are certainly possible.

Important Characteristics of Data

- Dimensionality (number of attributes)
  - The number of attributes that the objects in the data set possess
  - High dimensional data brings a number of challenges (curse of dimensionality)
    - an important motivation in preprocessing the data is dimensionality reduction
- Distribution (Sparsity)
  - the frequency of occurrence of various values or sets of values for the attributes comprising data objects
    - Equivalently, the distribution of a data set can be considered as a description of the concentration of objects in various regions of the data space

Important Characteristics of Data

- Resolution
  - Data can be obtained at different levels of resolution, and often the properties of the data are different at different resolutions
  - Patterns depend on the scale
    - if the resolution is too fine, a pattern may not be visible or may be buried in noise
    - if the resolution is too coarse, the pattern can disappear
- Record Data (Size)
  - Much data mining work assumes that the data set is a collection of records (data objects), each of which consists of a fixed set of data fields (attributes).
  - stored either in flat files or in relational databases
  - Type of analysis may depend on size of data

Different variations of record data

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
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

(a) Record data.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Soda, Diapers, Milk

(b) Transaction data.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

	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

(d) Document-term matrix.

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

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

Transaction or Market Basket Data

- Transaction data is a special type of record data, where each record (transaction) involves a set of items
- Consider a grocery store.
  - The set of products purchased by a customer during one shopping trip constitutes a transaction, while the individual products that were purchased are the items.
  - This type of data is called market basket data because the items in each record are the products in a person's market basket.

Transaction Data

- Transaction data is a collection of sets of items, but it can be viewed as a set of records whose fields are asymmetric attributes
  - Can represent transaction data as record data

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

The Data Matrix

- All the data objects in a collection of data have the same fixed set of numeric attributes.
  - data objects are points (vectors) in a multidimensional space, where each dimension represents a distinct attribute describing the object.
  - A set of such data objects can be interpreted as an *m* by *n* matrix, where there are *m* rows, one for each object, and *n* columns, one for each attribute.
  - This matrix is called a data matrix or a pattern matrix.

The Data Matrix

- A data matrix is a variation of record data, but because it consists of numeric attributes, standard matrix operation can be applied to transform and manipulate the data.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

The Sparse Data Matrix

- special case of a data matrix where the attributes are of the same type and are asymmetric;
  - i.e., only non-zero values are important.
  - Transaction data is an example of a sparse data matrix that has only 0–1 entries.
  - Another common example is document data.
    - If the order of the terms (words) in a document is ignored—the "bag of words" approach—then a document can be represented as a term vector, where each term is a component (attribute) of the vector and the value of each component is the number of times the corresponding term occurs in the document

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.

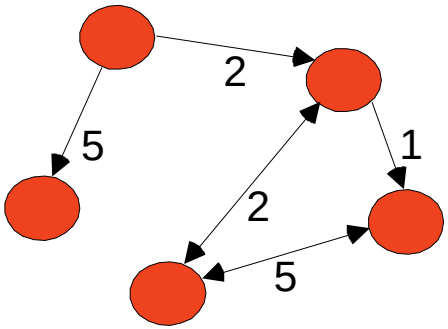
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

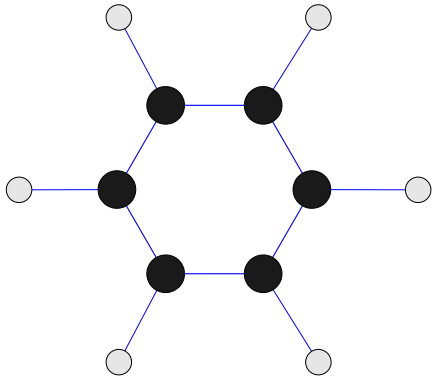
Graph-Based Data

- A graph can sometimes be a convenient and powerful representation for data.
- We consider two specific cases
  - The graph captures relationships among data objects
    - The relationships among objects frequently convey important information.
      - In such cases, the data is often represented as a graph.
  - The data objects themselves are represented as graphs.
    - If objects have structure, that is, the objects contain subobjects that have relationships, then such objects are frequently represented as graphs.

Graph-Based Data

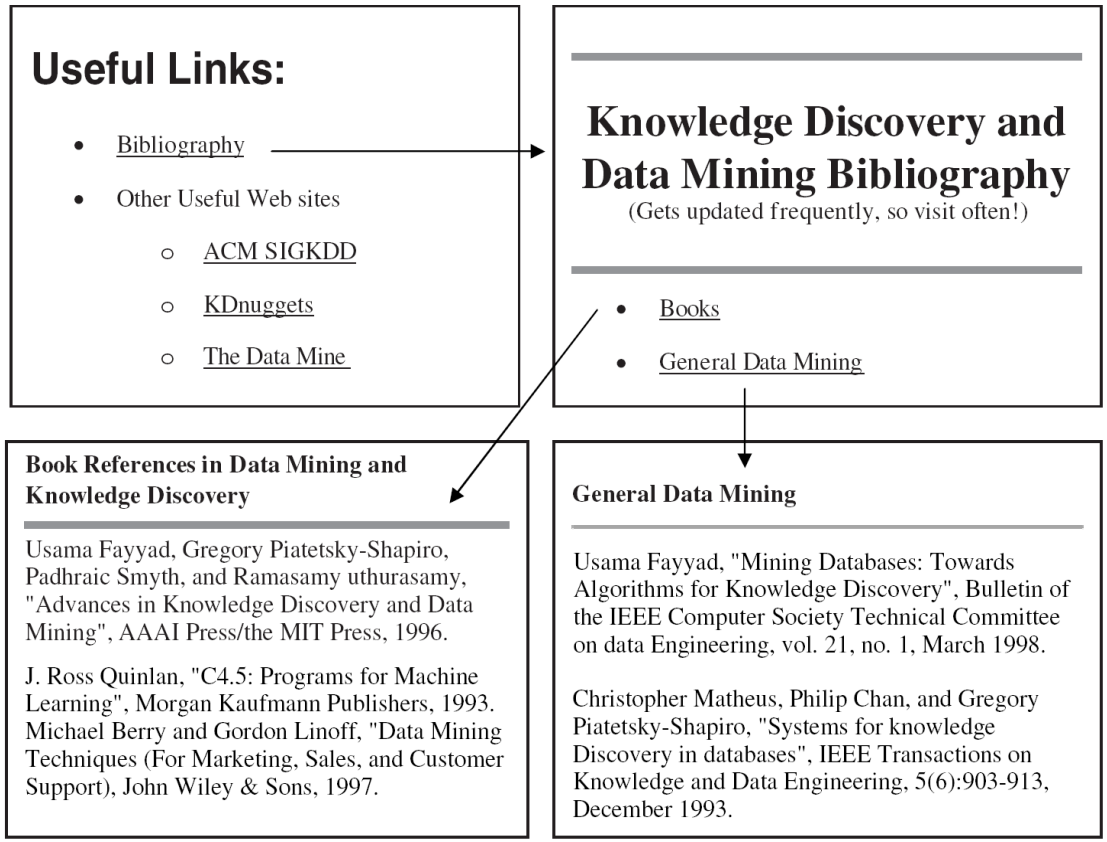
- Examples: Generic graph, a molecule, and webpages





Benzene Molecule: C6H6

(ball-and-stick diagram)



Linked web pages

Ordered Data

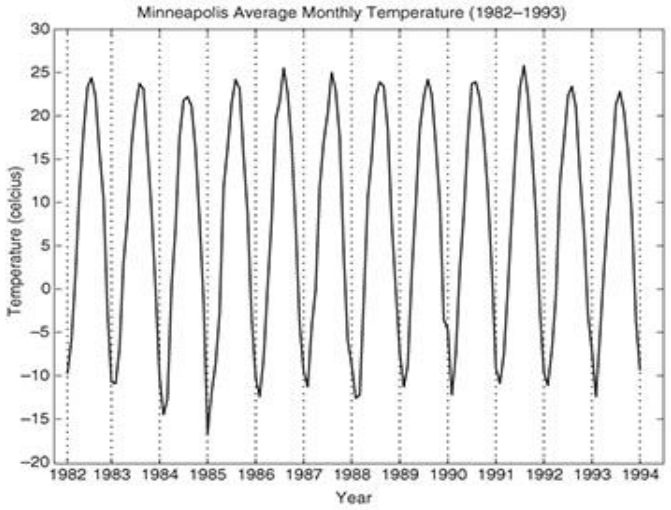
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

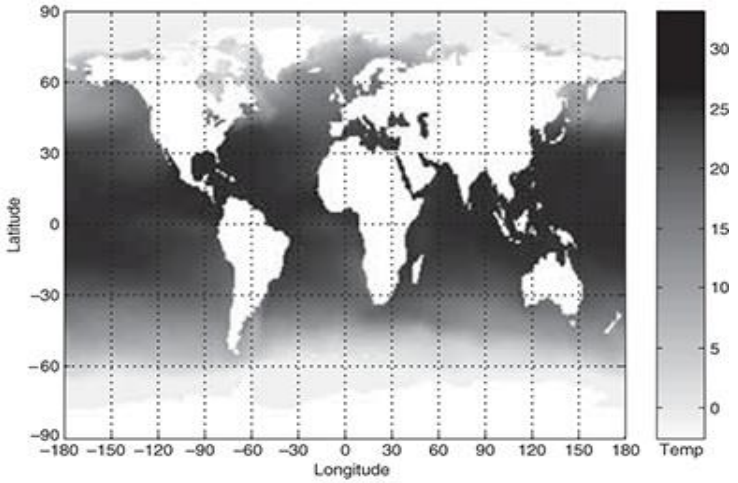
(a) Sequential transaction data.

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

(b) Genomic sequence data.



(c) Temperature time series.



(d) Spatial temperature data.

Ordered Data

- Attributes have relationships that involve order in time or space Sequences of transactions
- Sequential transaction data can be thought of as an extension of transaction data, where each transaction has a time associated with it.
  - Consider a retail transaction data set that also stores the time at which the transaction took place.
    - This time information makes it possible to find patterns such as “candy sales peak before Halloween.”
  - A time can also be associated with each attribute.

Sequential Transaction Data

- five different times:
  - *t*1, *t*2, *t*3, *t*4, and *t*5

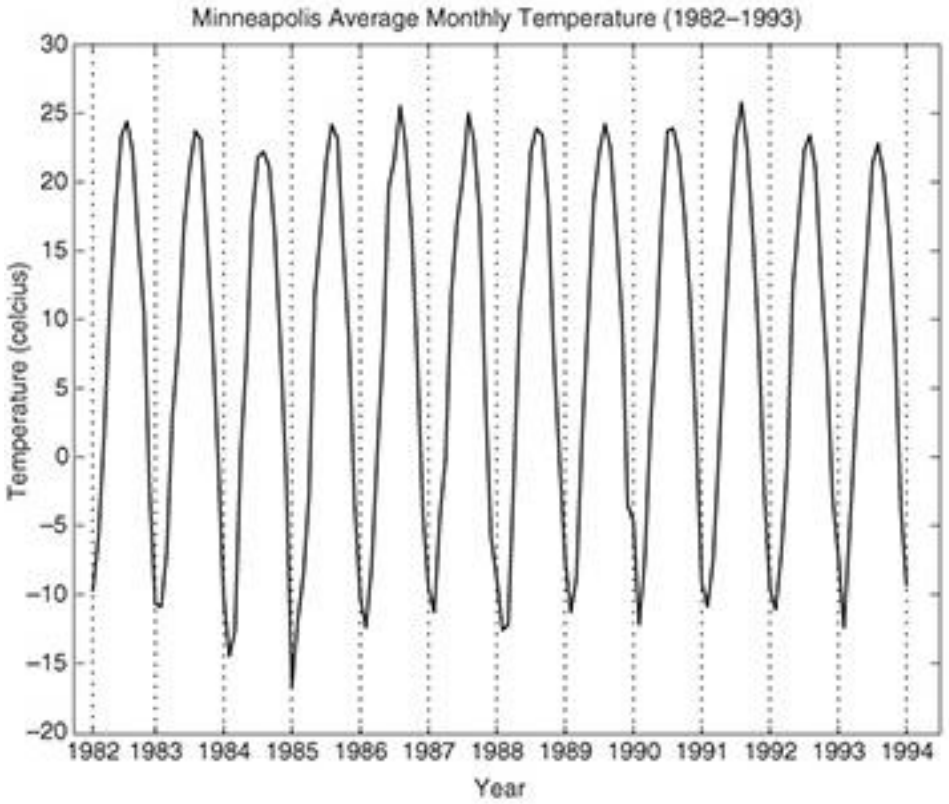
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

- three different customers:
  - C1, C2, and C3
- five different items:
  - A, B, C, D, and E

Time Series Data

- a special type of ordered data where each record is a time series ,
  - i.e., a series of measurements taken over time



- When working with temporal data, such as time series, it is important to consider temporal autocorrelation;
  - i.e., if two measurements are close in time, then the values of those measurements are often very similar.

Sequence Data

- Sequence data consists of a data set that is a sequence of individual entities, such as a sequence of words or letters.

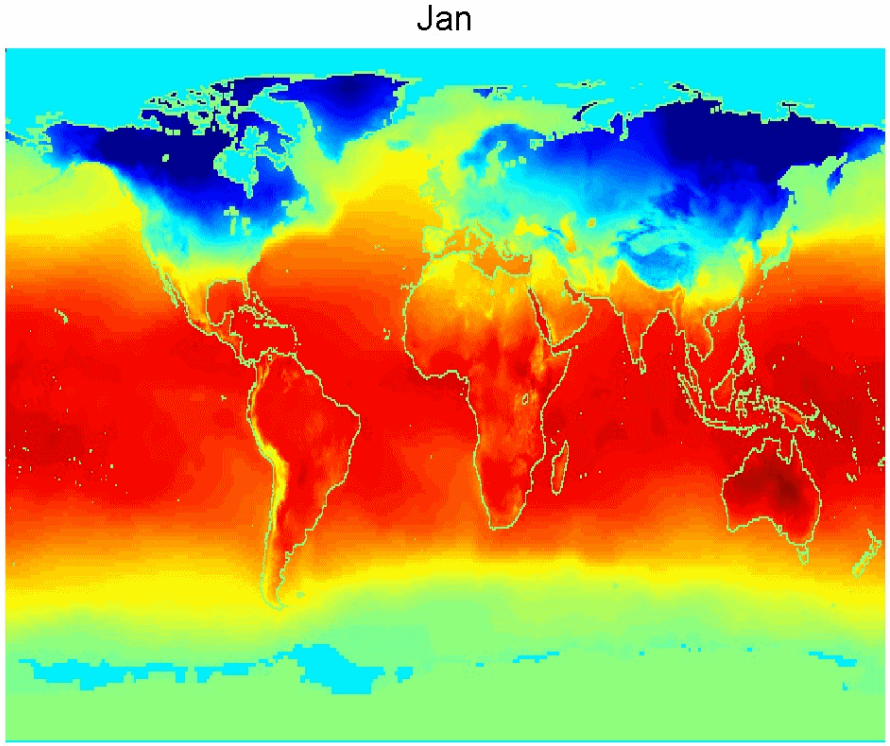
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

  - It is quite similar to sequential data, except that there are no time stamps;
  - instead, there are positions in an ordered sequence.

- The genetic information of plants and animals can be represented in the form of sequences of nucleotides.

Spatial and Spatio-Temporal Data

- Some objects have spatial attributes, such as positions or areas, in addition to other types of attributes.



  - An example of spatial data is weather data (precipitation, temperature, pressure) that is collected for a variety of geographical locations.

- Average Monthly Temperature of land and ocean

Spatial and Spatio-Temporal Data

- An important aspect of spatial data is spatial autocorrelation; i.e., objects that are physically close tend to be similar in other ways as well.
  - Thus, two points on the Earth that are close to each other usually have similar values for temperature and rainfall.
    - Note that spatial autocorrelation is analogous to temporal autocorrelation.
  - Important examples of spatial and spatio-temporal data are the science and engineering data sets that are the result of measurements or model output taken at regularly or irregularly distributed points on a two- or three dimensional grid or mesh.

Handling Non-Record Data

- Most data mining algorithms are designed for record data or its variations.
- Record-oriented techniques can be applied to non-record data by extracting features from data objects and using these features to create a record corresponding to each object.
  - Consider the chemical structure data that was described earlier.
    - Given a set of common substructures, each compound can be represented as a record with binary attributes that indicate whether a compound contains a specific substructure.
    - Such a representation is actually a transaction data set, where the transactions are the compounds, and the items are the substructures.

Data Quality

- Data mining algorithms are often applied to data that was collected for another purpose, or for future, but unspecified applications.
  - For that reason, data mining cannot usually take advantage of the significant benefits of "addressing quality issues at the source."
- Data mining focuses on
  - the detection and correction of data quality problems
  - the use of algorithms that can tolerate poor data quality.
- The first step, detection and correction, is often called data cleaning.

Data Quality

- Data is not perfect.
  - There may be problems due to
    - human error,
    - limitations of measuring devices,
    - flaws in the data collection process.
  - Values or even entire data objects can be missing.
  - There can be spurious or duplicate objects;
    - i.e., multiple data objects that all correspond to a single "real" object
      - For example, there might be two different records for a person who has recently lived at two different addresses.

Data Quality

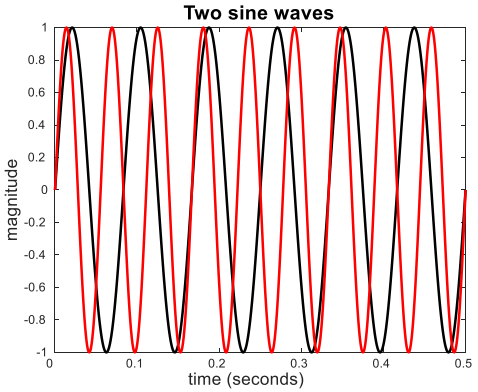
- Poor data quality negatively affects many data processing efforts
  - Data mining example:
    - a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems:
  - Noise and outliers, Wrong data, Fake data, Missing values, Duplicate data

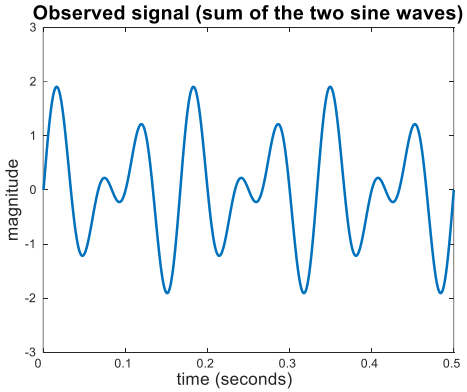
Measurement and Data Collection Errors

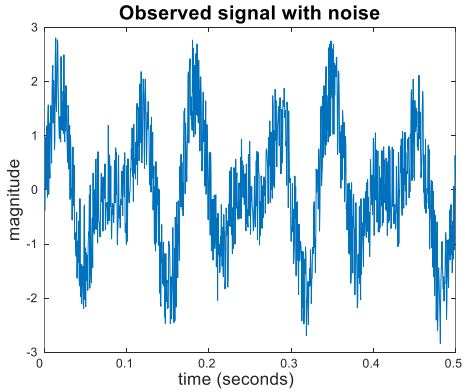
- The measurement error refers to any problem resulting from the measurement process.
  - A common problem is that the value recorded differs from the true value to some extent.
  - For continuous attributes, the numerical difference of the measured and true value is called the error.
- The data collection error refers to errors such as
  - omitting data objects or attribute values,
  - inappropriately including a data object.
- Both measurement errors and data collection errors can be either systematic or random.

Noise and Artifacts

- Noise is the random component of a measurement error.
  - It typically involves the distortion of a value or the addition of spurious objects
    - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen
- The figures below show two sine waves of the same magnitude and different frequencies, the waves combined, and the two sine waves with random noise

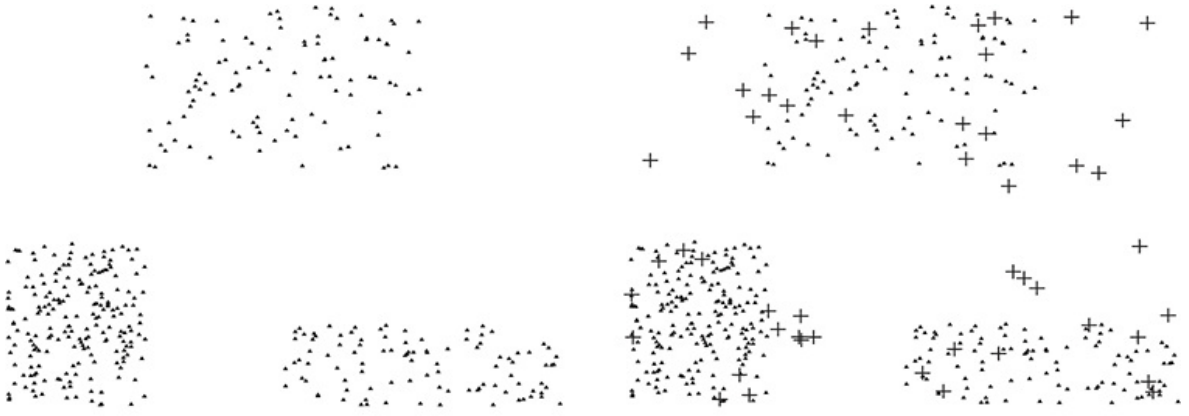






Noise and Artifacts

- The term noise is often used in connection with data that has a spatial or temporal component.



(a) Three groups of points. (b) With noise points (+) added.

Noise in a spatial context

- Data errors can be the result of a more deterministic phenomenon, such as a streak in the same place on a set of photographs.
  - Such deterministic distortions of the data are often referred to as artifacts.

Precision, Bias, and Accuracy

- In statistics and experimental science, the quality of the measurement process and the resulting data are measured by precision and bias
  - Precision:
    - The closeness of repeated measurements (of the same quantity) to one another.
      - often measured by the standard deviation of a set of values
  - Bias:
    - A systematic variation of measurements from the quantity being measured.
      - measured by taking the difference between the mean of the set of values and the known value of the quantity being measured

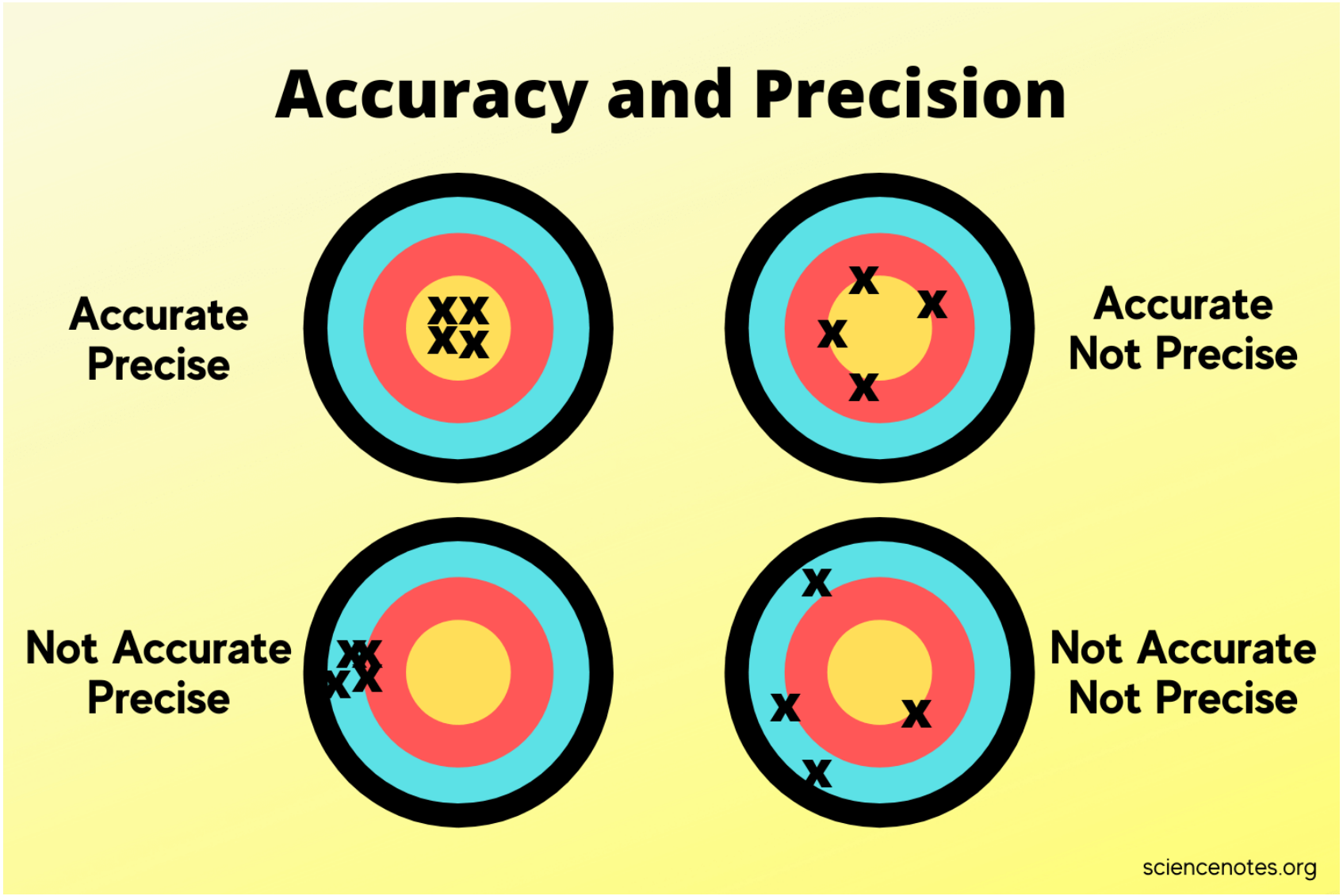
Precision, Bias, and Accuracy

- Example:
  - Suppose that we have a standard laboratory weight with a mass of 1g and want to assess the precision and bias of our new laboratory scale.
  - We weigh the mass five times, and obtain the following five values:
    - {1.015, 0.990, 1.013, 1.001, 0.986}.
  - The mean of these values is 1.001, and
    - hence, the bias is 0.001.
  - The precision, as measured by the standard deviation, is 0.013.

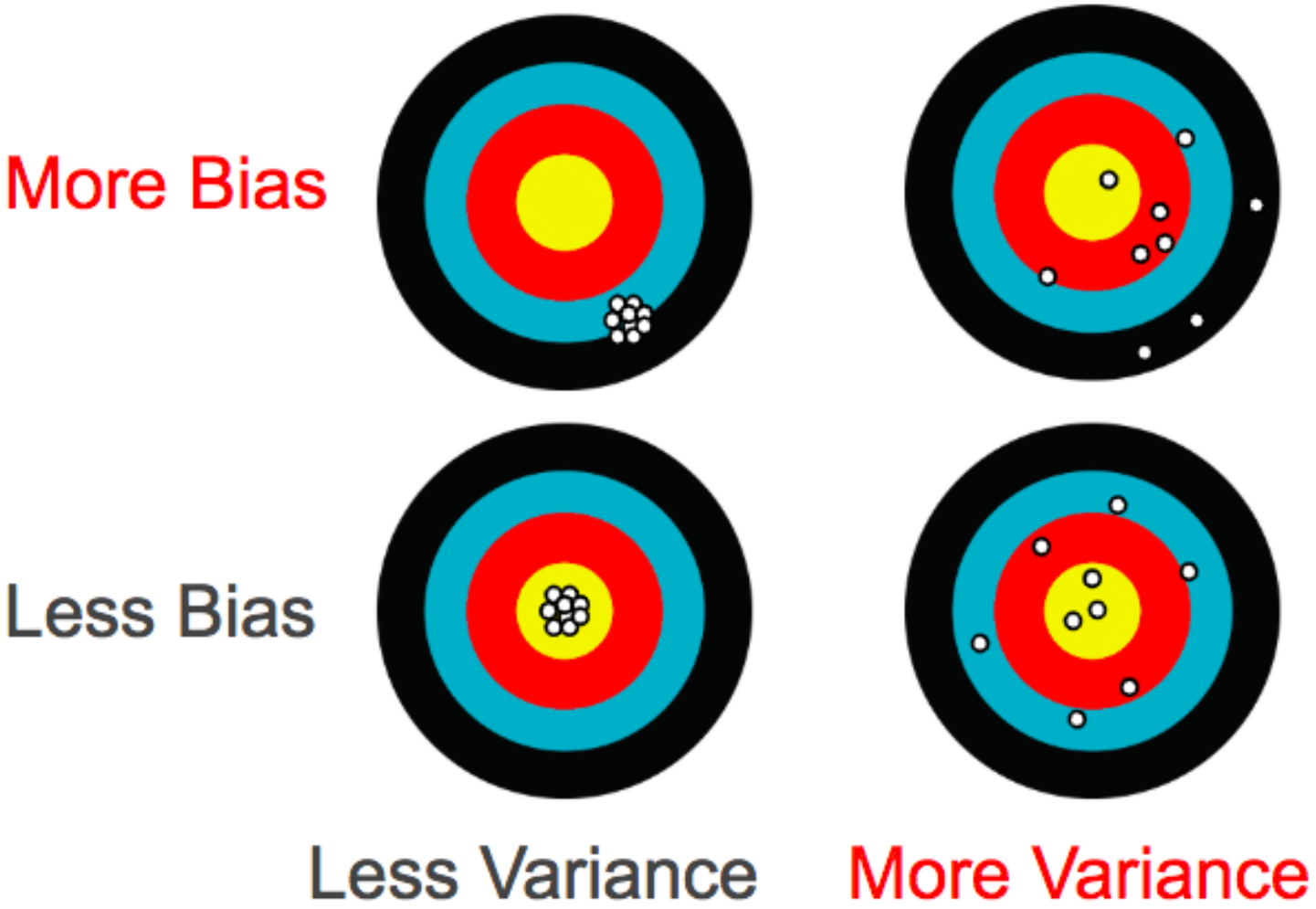
Precision, Bias, and Accuracy

- It is common to use the more general term, accuracy , to refer to the degree of measurement error in data.
  - Accuracy
    - The closeness of measurements to the true value of the quantity being measured.
  - Accuracy depends on precision and bias, but there is no specific formula for accuracy in terms of these two quantities.
  - One important aspect of accuracy is the use of significant digits.
    - The goal is to use only as many digits to represent the result of a measurement or calculation as are justified by the precision of the data.

Precision, Bias, and Accuracy



Precision, Bias, and Accuracy

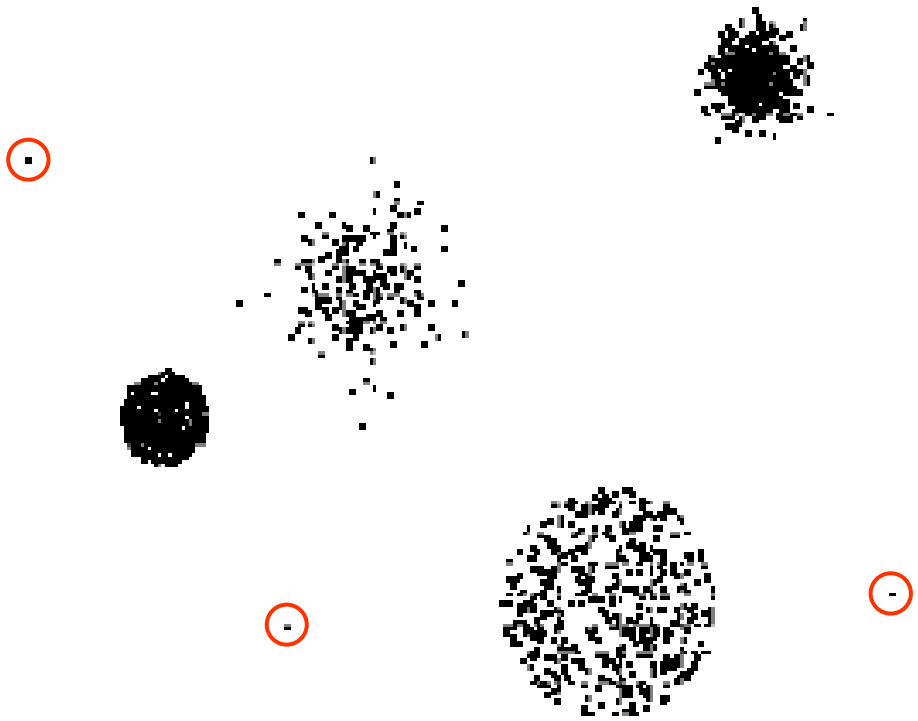


Outliers

- data objects that, in some sense, have characteristics that are different from most of the other data objects in the data set
- values of an attribute that are unusual with respect to the typical values for that attribute
- can be referred to as anomalous objects or values
  - Unlike noise, outliers can be legitimate data objects or values that we are interested in detecting
    - For instance, in fraud and network intrusion detection, the goal is to find unusual objects or events from among a large number of normal ones

Outliers

- Case 1: Outliers are noise that interferes with data analysis



- Case 2: Outliers are the goal of our analysis
  - Credit card fraud
  - Intrusion detection

Missing Values

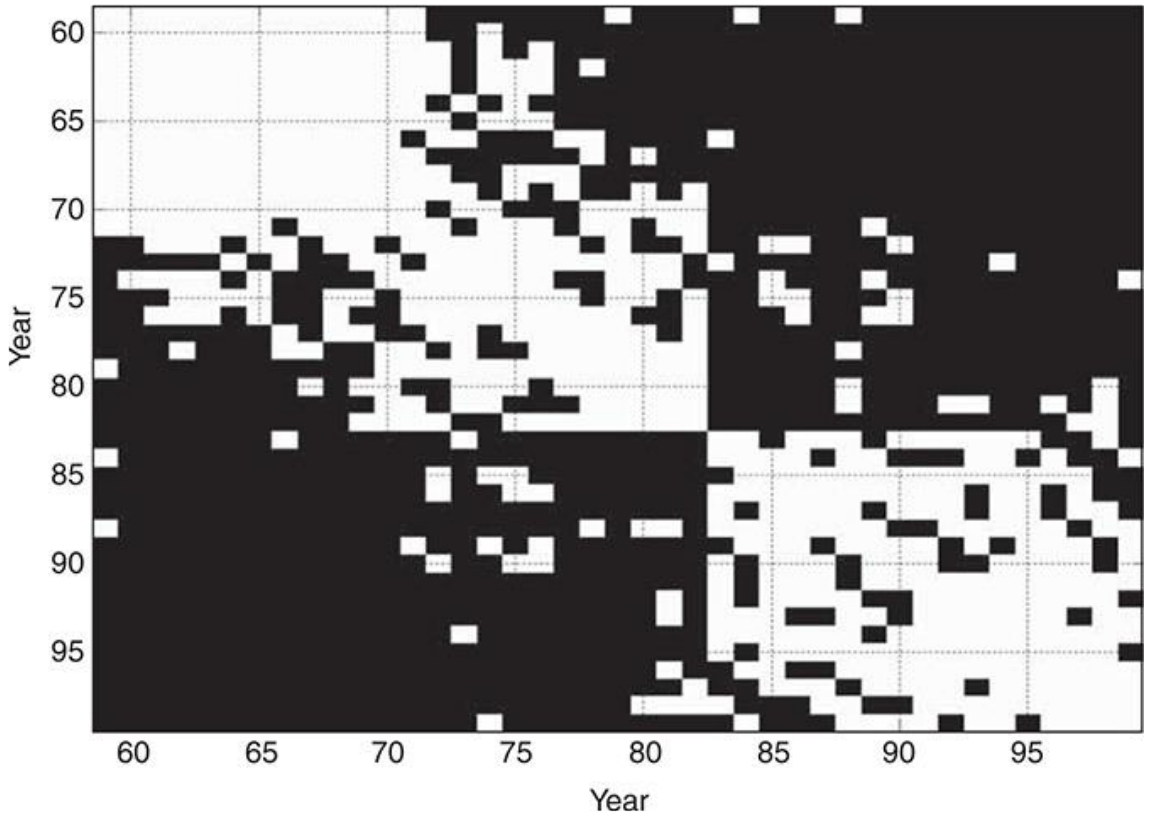
- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)
- Handling missing values
  - Eliminate data objects or variables
  - Estimate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

Inconsistent Values

- Data can contain inconsistent values
  - Consider an address field, where both a zip code and city are listed, but the specified zip code area is not contained in that city. It is possible that the individual entering this information transposed two digits
- Some types of inconsistences are easy to detect.
  - For instance, a person's height should not be negative.
- Once an inconsistency has been detected, it is sometimes possible to correct the data.
  - The correction of an inconsistency requires additional or redundant information.

Inconsistent Values

- Example (Inconsistent Sea Surface Temperature)



  - SST data was originally collected using ocean-based measurements from ships or buoys, but more recently, satellites have been used to gather the data.
  - To create a long-term data set, both sources of data must be used.
  - However, because the data comes from different sources, the two parts of the data are subtly different.

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
    - if there are two objects that actually represent a single object, then one or more values of corresponding attributes are usually different, and these inconsistent values must be resolved.
    - care needs to be taken to avoid accidentally combining data objects that are similar, but not duplicates, such as two distinct people with identical names.
  - The term deduplication is often used to refer to the process of dealing with these issues.

Duplicate Data

- Examples:
  - Same person with multiple email addresses
  - In some cases, two or more objects are identical with respect to the attributes measured by the database, but they still represent different objects.
    - Here, the duplicates are legitimate, but can still cause problems for some algorithms if the possibility of identical objects is not specifically accounted for in their design.
- Data cleaning
  - Process of dealing with duplicate data issues
- When should duplicate data not be removed?

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
- When should duplicate data not be removed?

Issues Related to Applications

- Data quality issues can also be considered from an application viewpoint as expressed by the statement “data is of high quality if it is suitable for its intended use.”
  - relevance to the specific purpose
  - completeness
  - accuracy
  - timeliness
  - format
  - cost

Data Preprocessing

- Data preprocessing is a broad area and fall into two categories:
  - selecting data objects and attributes for the analysis
  - creating/changing the attributes.
- The goal is to improve the data mining analysis with respect to time, cost, and quality through:
  - Aggregation
  - Sampling
  - Discretization and Binarization
  - Attribute Transformation
  - Dimensionality Reduction
  - Feature subset selection
  - Feature creation

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.
    - Days aggregated into weeks, months, or years
- More stable data
  - aggregated data tends to have less variability

Data set containing information about customer purchases

Transaction ID	Item	Store Location	Date	Price	...
⋮	⋮	⋮	⋮	⋮	
101123	Watch	Chicago	09/06/04	$25.99	...
101123	Battery	Chicago	09/06/04	$5.99	...
101124	Shoes	Minneapolis	09/06/04	$75.00	...
⋮	⋮	⋮	⋮	⋮	

Example: Precipitation in Australia

- This example is based on precipitation in Australia from the period 1982 to 1993.

  The next slide shows

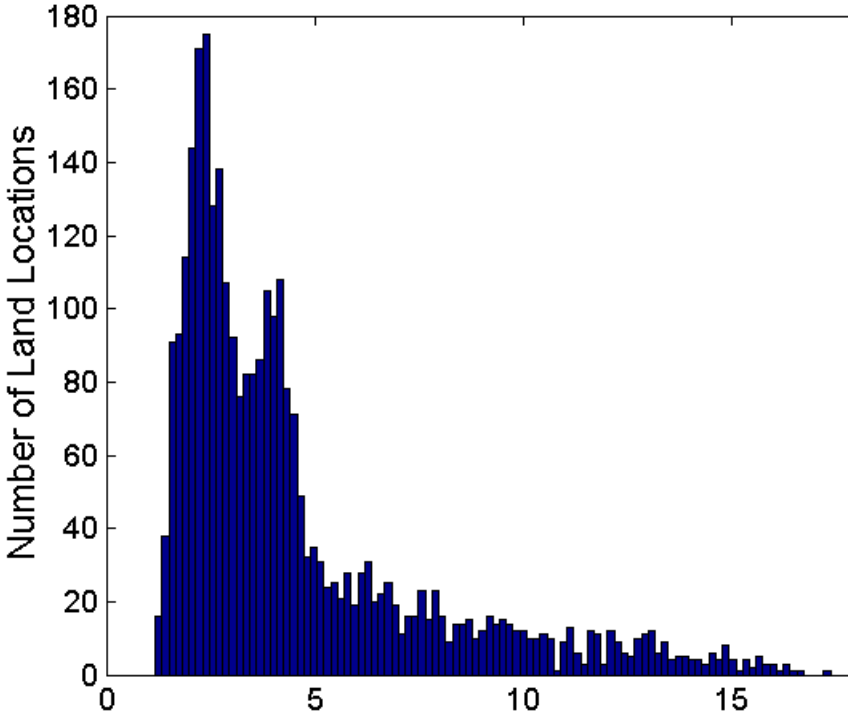
  – A histogram for the standard deviation of average monthly precipitation for 3,030 0.5° by 0.5° grid cells in Australia, and

  – A histogram for the standard deviation of the average yearly precipitation for the same locations.

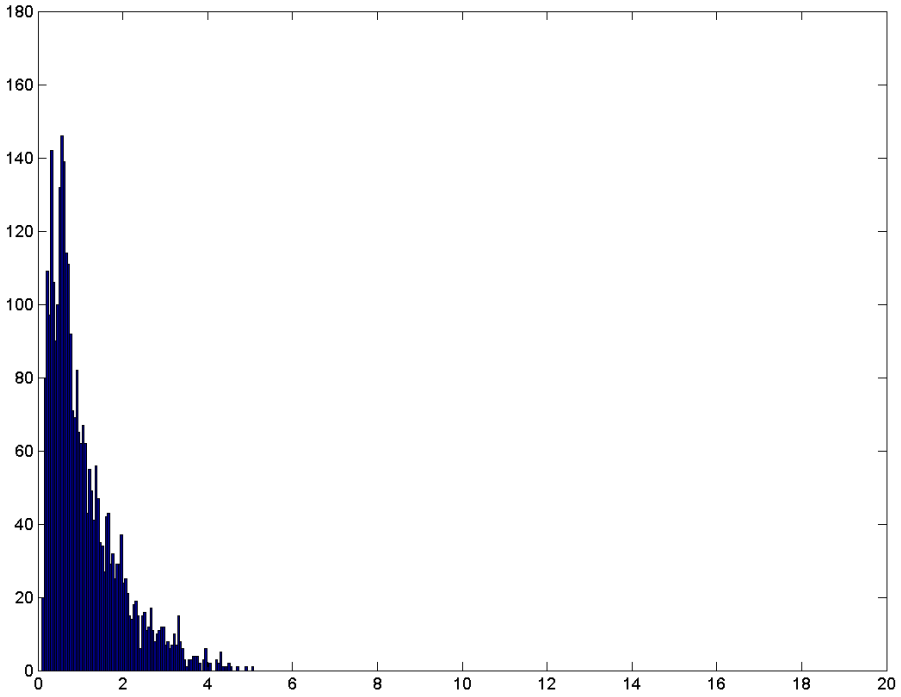
- The average yearly precipitation has less variability than the average monthly precipitation.
- All precipitation measurements (and their standard deviations) are in centimeters.

Example: Precipitation in Australia

- Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Sampling

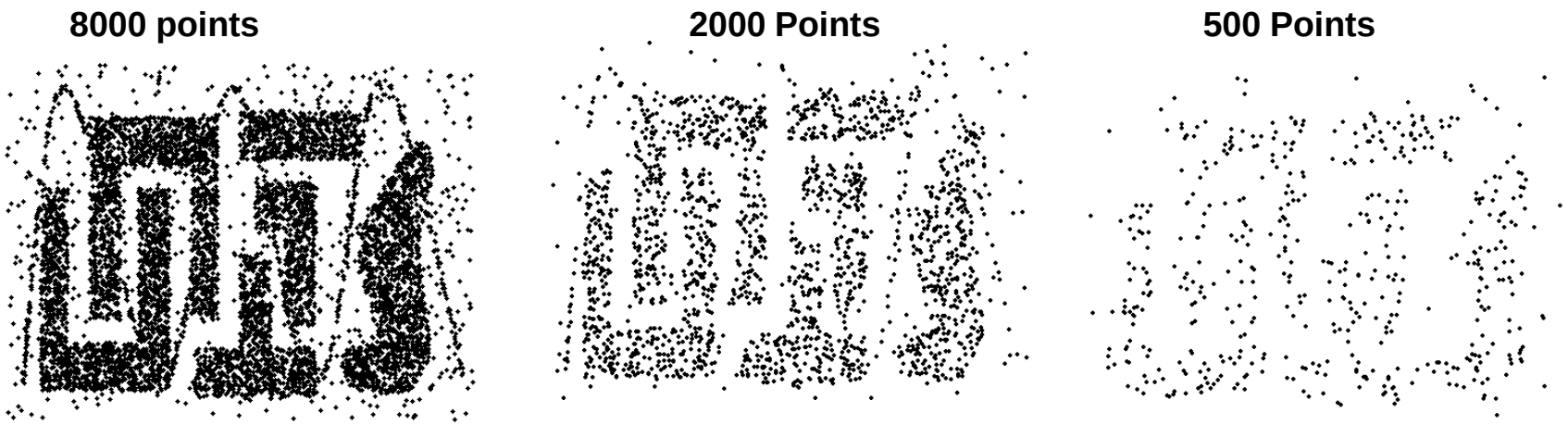
- Sampling is the main technique employed for data reduction.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

- Statisticians often sample because obtaining the entire set of data of interest is too expensive or time consuming.

- Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.

Sampling

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



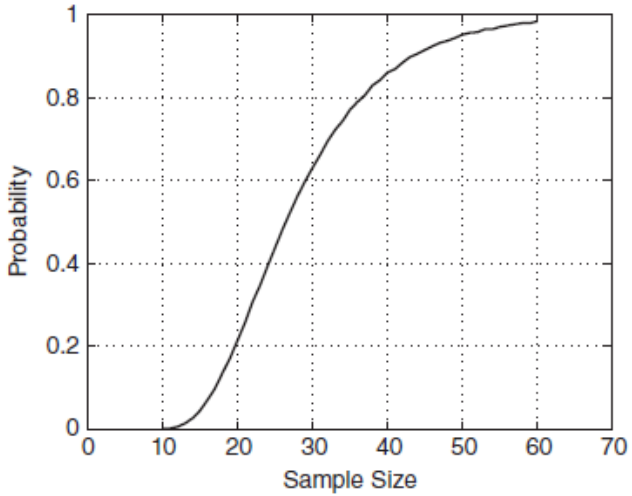
Types of Sampling

- 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

Types of Sampling

- Progressive Sampling
  - The proper sample size can be difficult to determine, so adaptive or progressive sampling schemes are sometimes used.
    - These approaches start with a small sample, and then increase the sample size until a sample of sufficient size has been obtained.
      - Finding representative points from 10 groups.

Ten groups of points

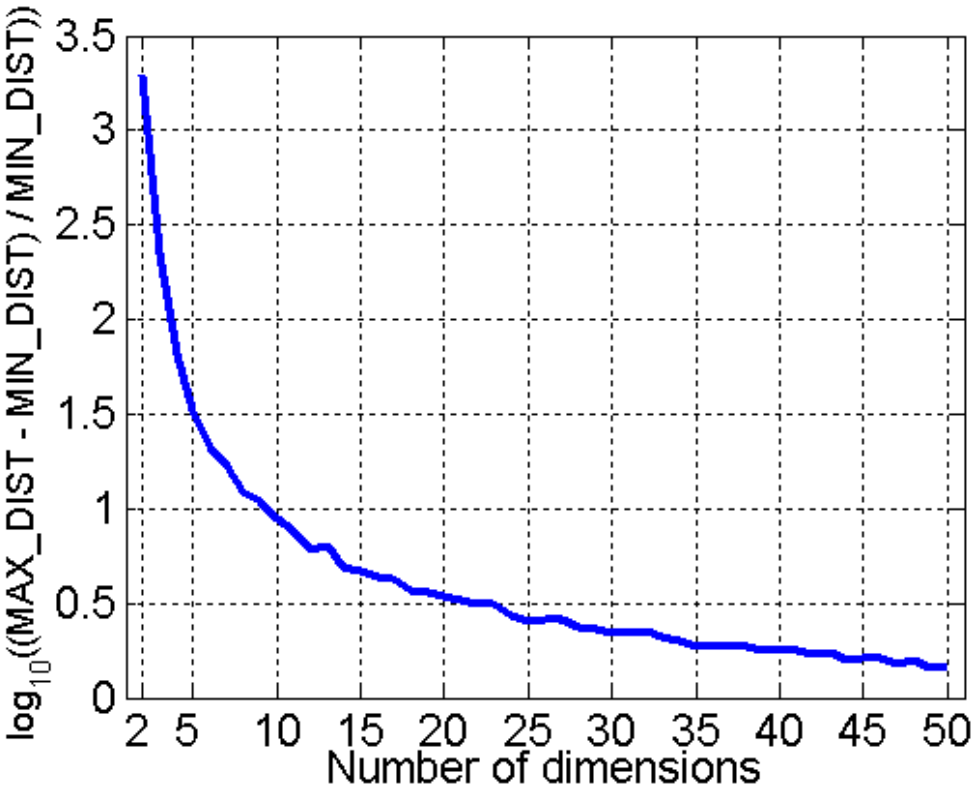


Probability a sample contains points from each of 10 groups

Dimensionality Reduction

- Data sets can have a large number of features
- There are a variety of benefits to dimensionality reduction.
- Many data mining algorithms work better if the dimensionality—the number of attributes in the data—is lower.
  - This is partly because dimensionality reduction can eliminate irrelevant features and reduce noise and partly because of the curse of dimensionality

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



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

Dimensionality Reduction

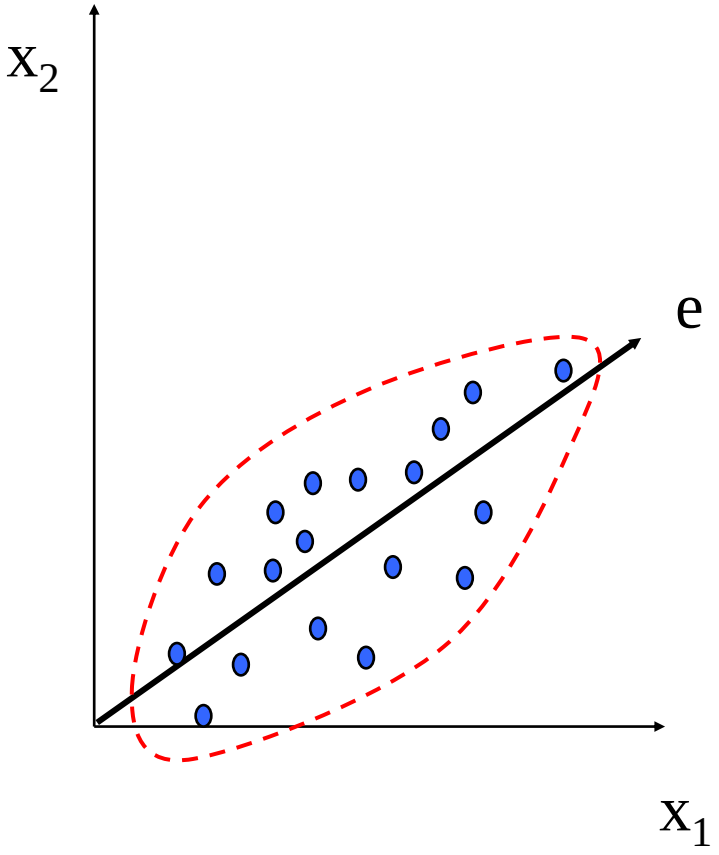
- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise
- Techniques
  - Principal Components Analysis (PCA)
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

Principal Components Analysis (PCA)

- a linear algebra technique for continuous attributes that finds new attributes (principal components) that
  - are linear combinations of the original attributes,
  - are orthogonal (perpendicular) to each other,
  - capture the maximum amount of variation in the data.
    - For example, the first two principal components capture as much of the variation in the data as is possible with two orthogonal attributes that are linear combinations of the original attributes.
- Singular Value Decomposition (SVD) is a linear algebra technique that is related to PCA and is also commonly used for dimensionality reduction.

Principal Components Analysis (PCA)

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



Principal Components Analysis (PCA)

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
- Many techniques developed, especially for classification

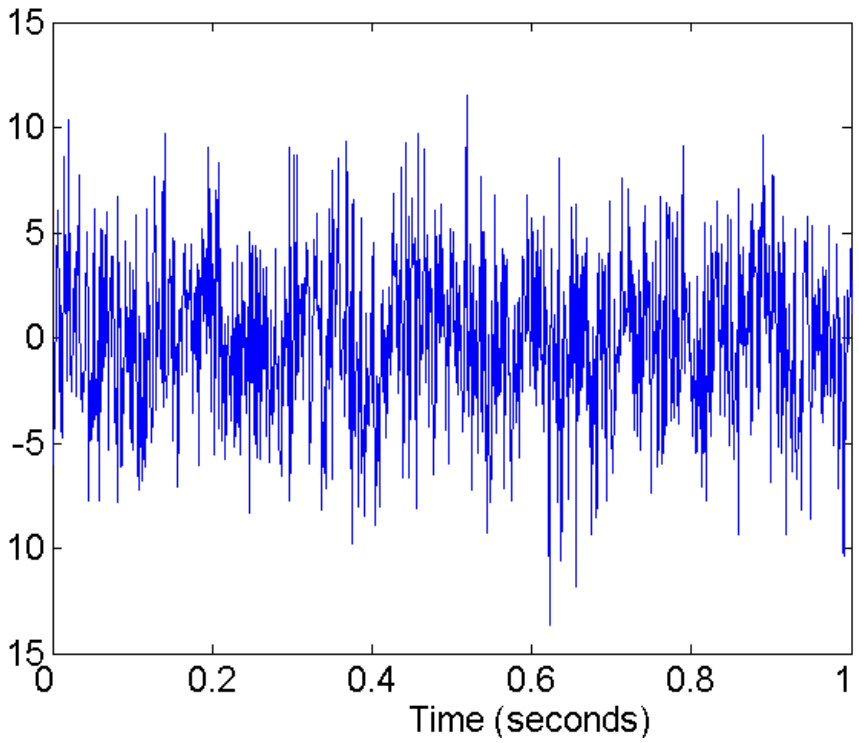
Feature Creation

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

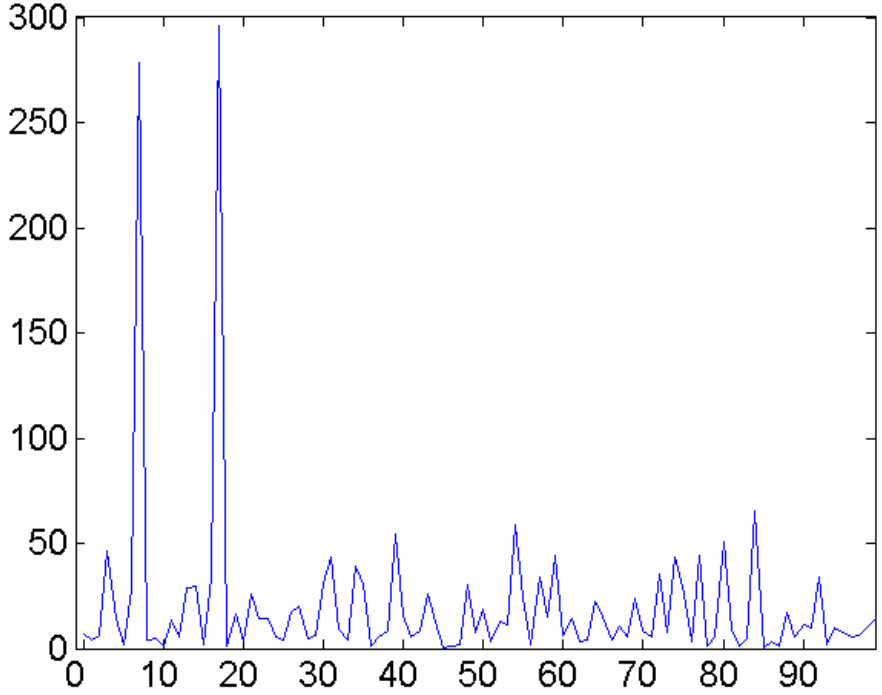
- Three general methodologies:
  - Feature extraction
    - Example: extracting edges from images
  - Feature construction
    - Example: dividing mass by volume to get density
  - Mapping data to new space
    - Example: Fourier and wavelet analysis

Mapping Data to a New Space

- Fourier and wavelet transform



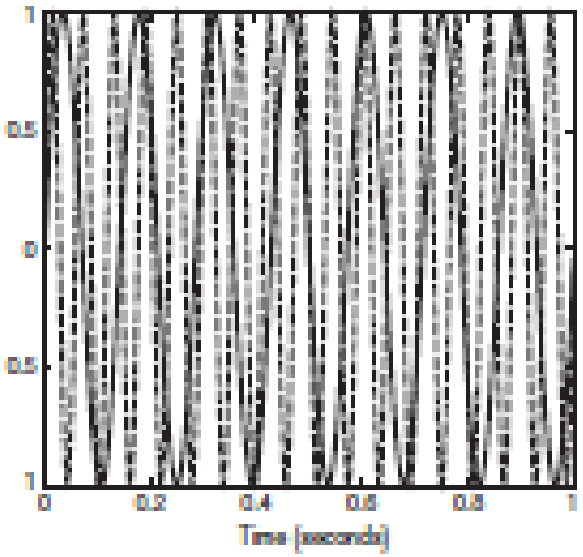
Two Sine Waves + Noise



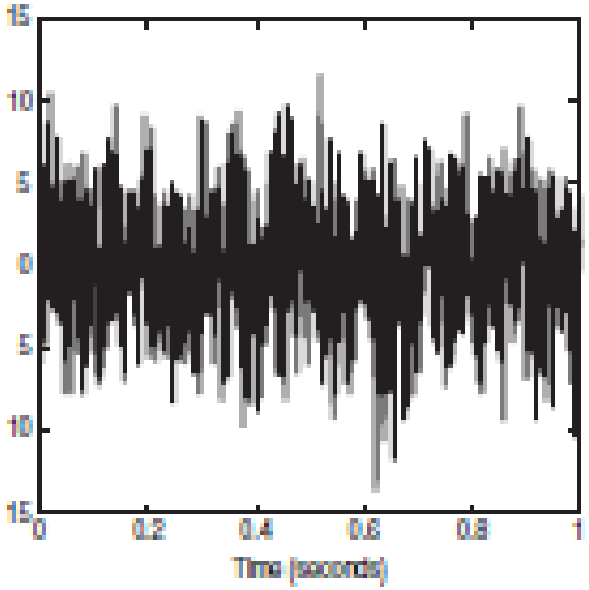
Frequency

Mapping Data to a New Space

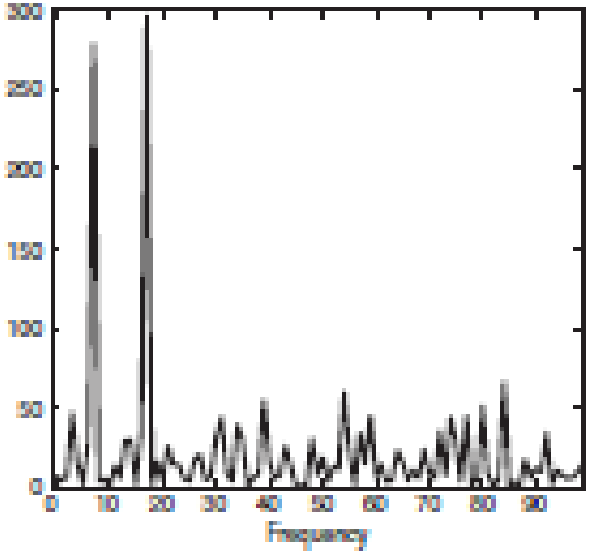
- Fourier and wavelet transform



(a) Two time series.



(b) Noisy time series.



(c) Power spectrum.

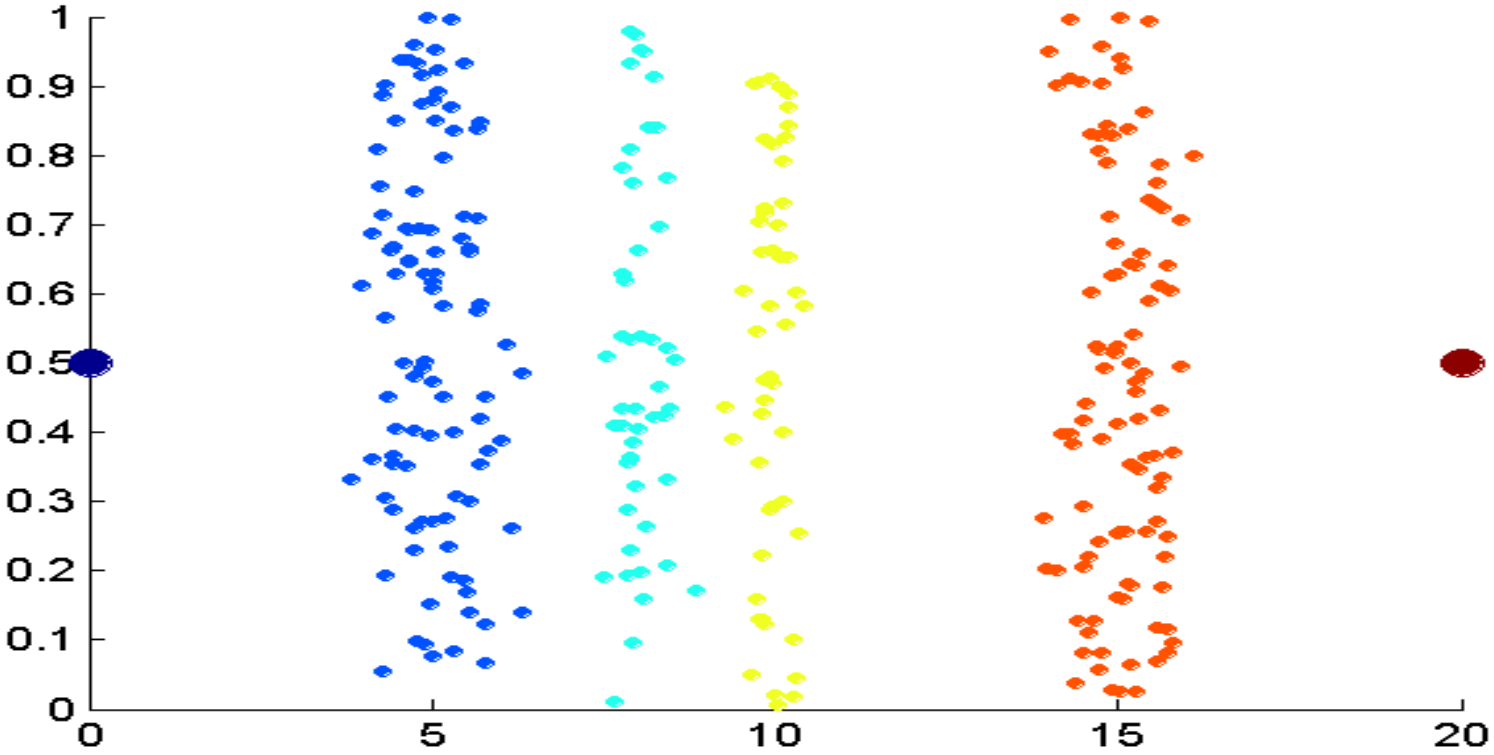
Discretization

- Discretization is the process of converting a continuous attribute into an ordinal attribute

- A potentially infinite number of values are mapped into a small number of categories

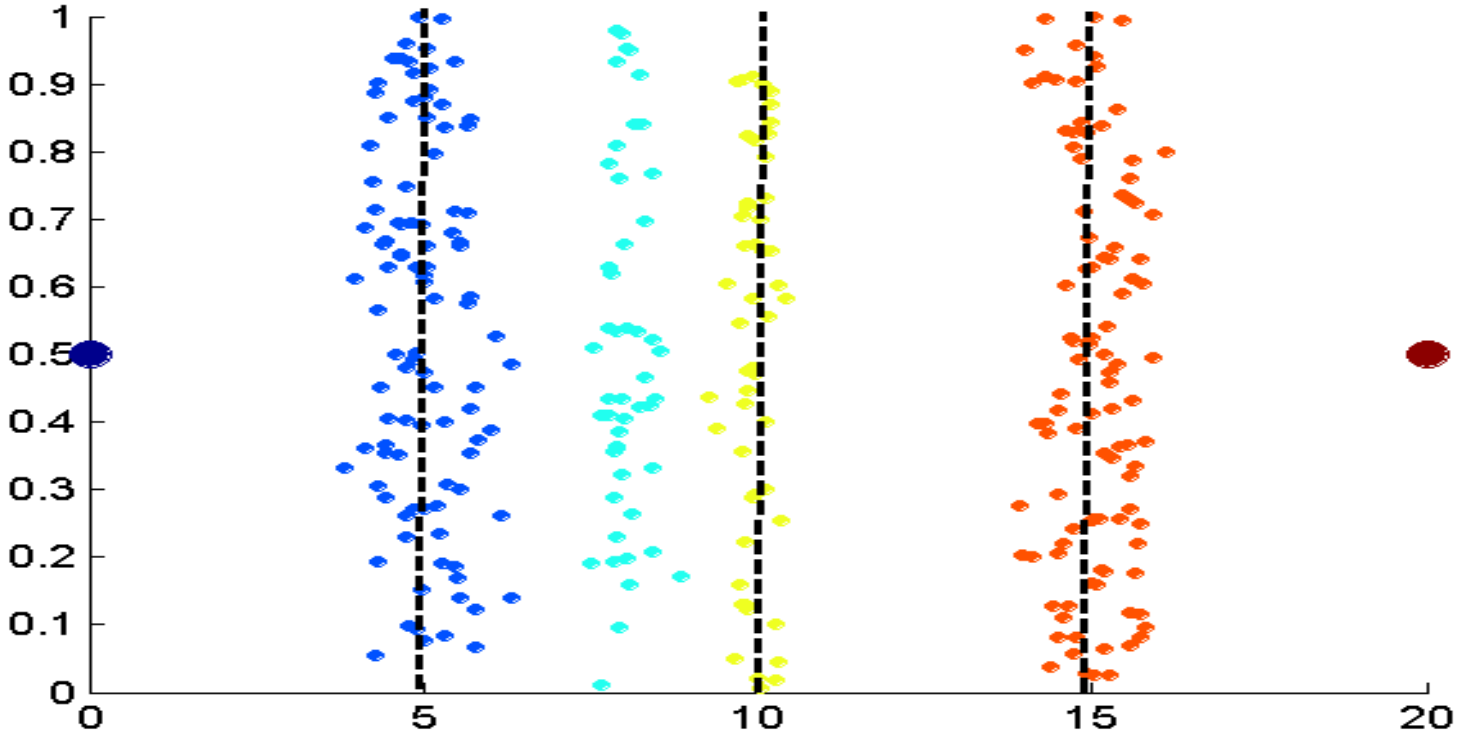
- Discretization is used in both unsupervised and supervised settings

Unsupervised Discretization



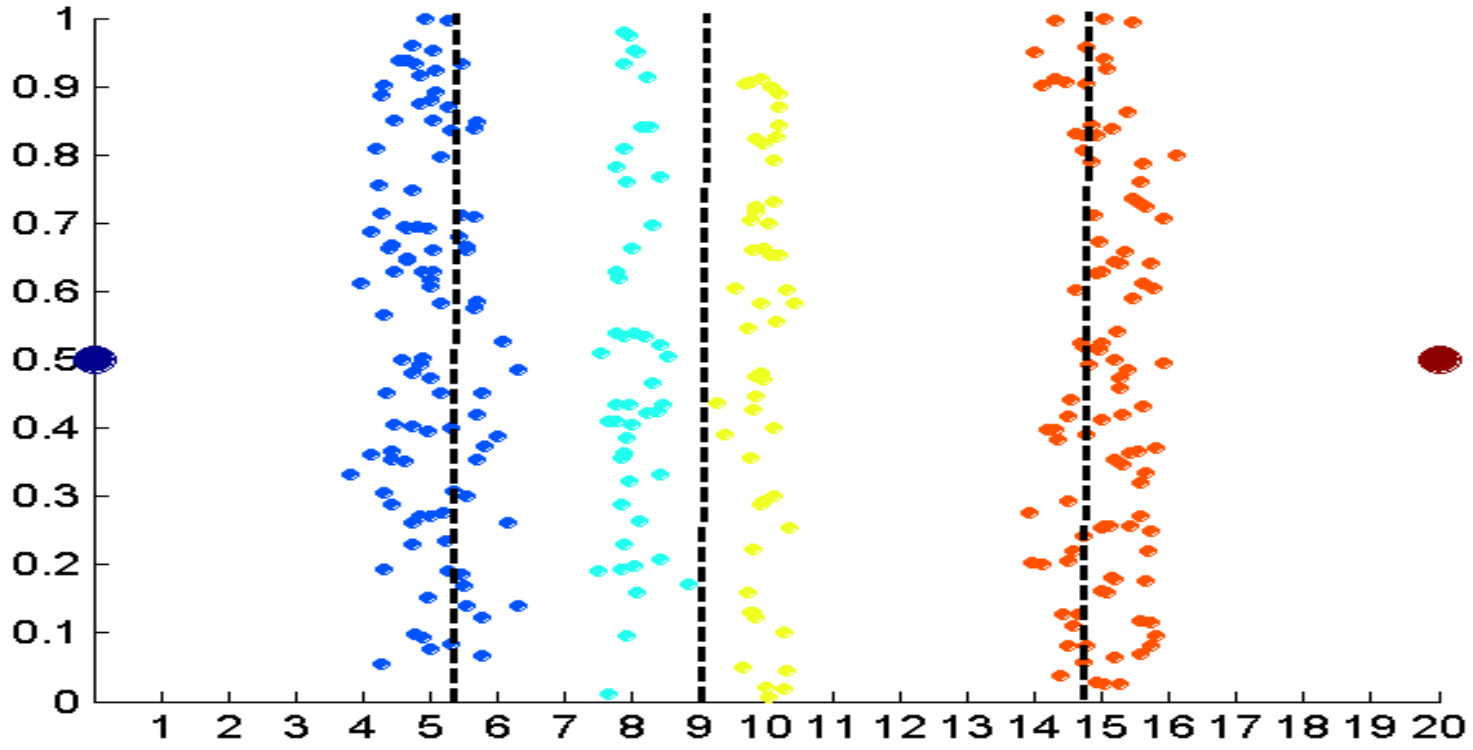
- Data consists of four groups of points and two outliers.
- Data is one-dimensional, but a random y component is added to reduce overlap.

Unsupervised Discretization



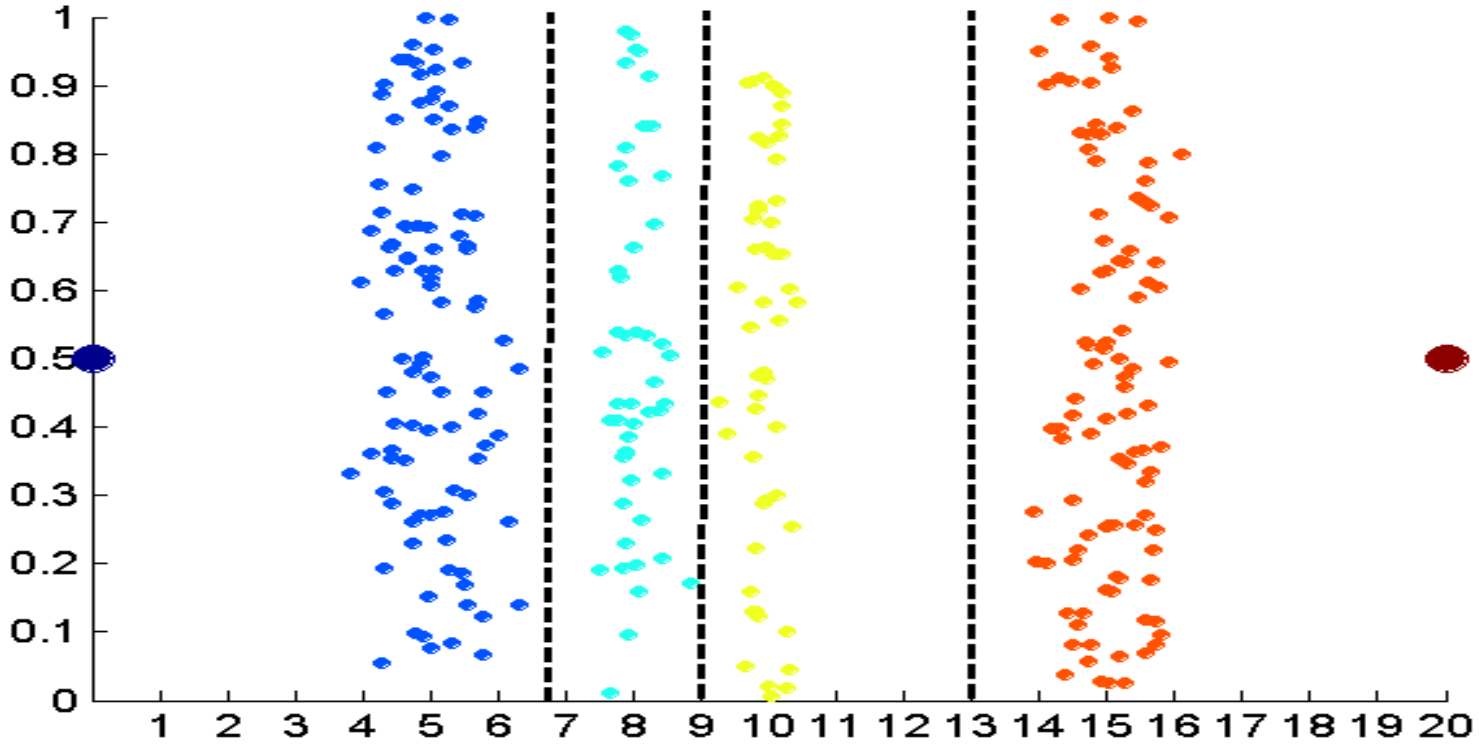
- Equal interval width approach used to obtain 4 values.

Unsupervised Discretization



- Equal frequency approach used to obtain 4 values.

Unsupervised Discretization

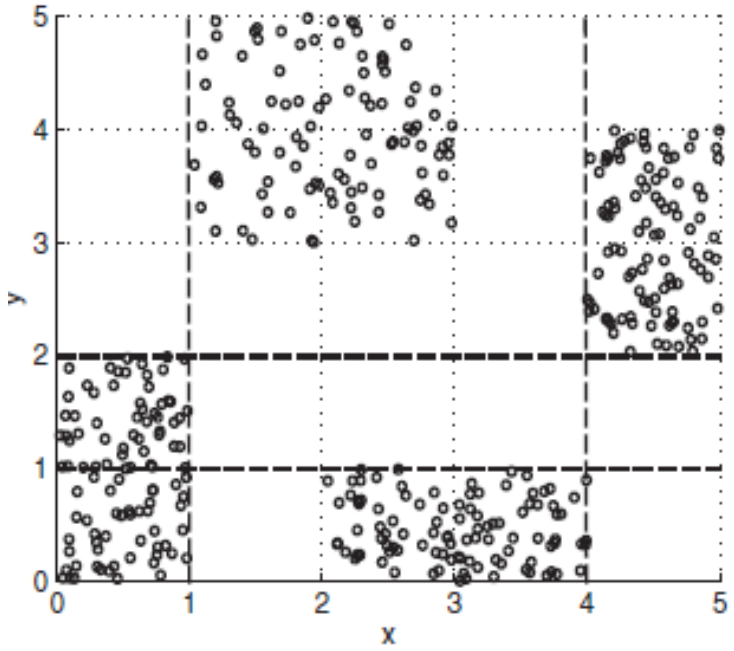


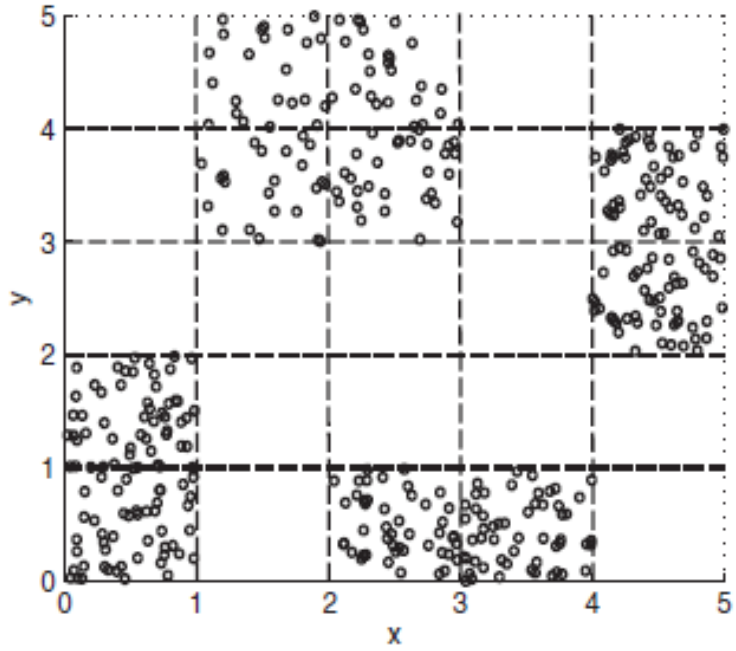
- K-means approach used to obtain 4 values.

Discretization in Supervised Settings

- Many classification algorithms work best if both the independent and dependent variables have only a few values
- We give an illustration of the usefulness of discretization using the following example.

  Discretizing *x* and *y* attributes for four groups (classes) of points:





Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables

Conversion of a categorical attribute to three binary attributes

Categorical Value	Integer Value	x_1	x_2	x_3
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

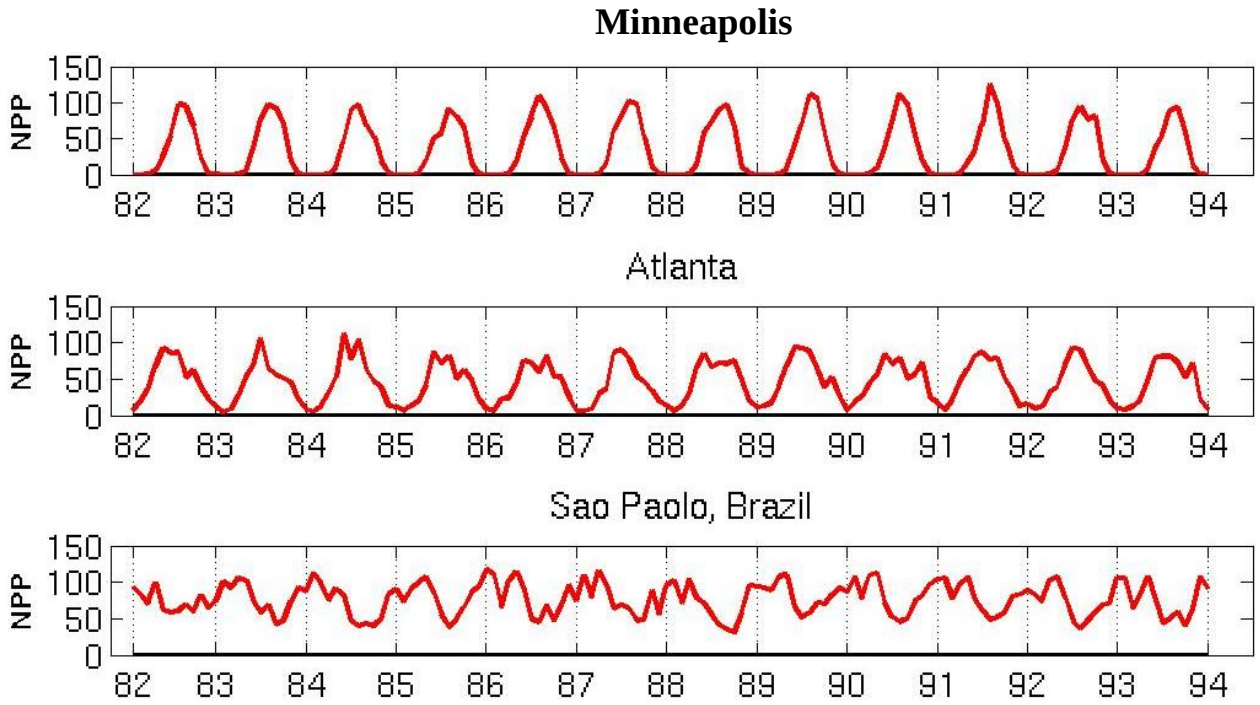
Conversion of a categorical attribute to five asymmetric binary attributes

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
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

Attribute Transformation

- An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, log(x), e^x, |x|
  - Normalization
    - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
    - Take out unwanted, common signal, e.g., seasonality
  - In statistics, standardization refers to subtracting off the means and dividing by the standard deviation

Example: Sample Time Series of Plant Growth

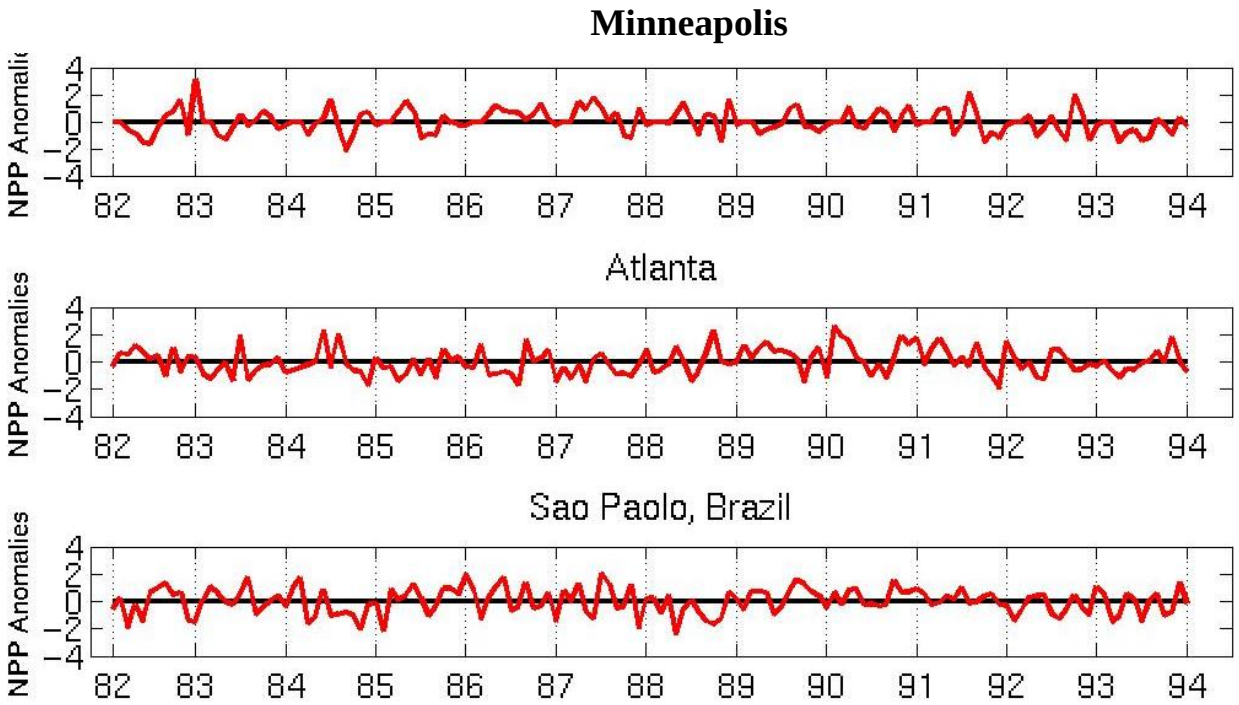


Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.

Correlations between time series

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

Seasonality Accounts for Much Correlation



Normalized using monthly Z Score:

Subtract off monthly mean and divide by monthly standard deviation

Correlations between time series

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