Machine Learning

Data

Data topics

- Types of attributes
- Data quality issues
- Transformations
- Visualization
- Types of datasets
- Preprocessing
- Summary statistics

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
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Types of attributes

- 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 attributes

- 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 four properties

Attribute Type	Description	Examples	Statistical 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. (+, -)	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	Allowed Transformations	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 values {1, 2, 3} or {0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	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 (time series) 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

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* x *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

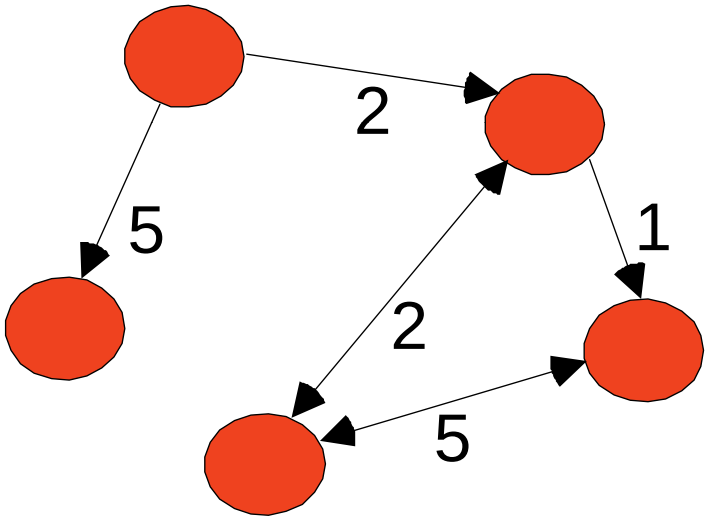
Transaction data

- 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.

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

Graph data

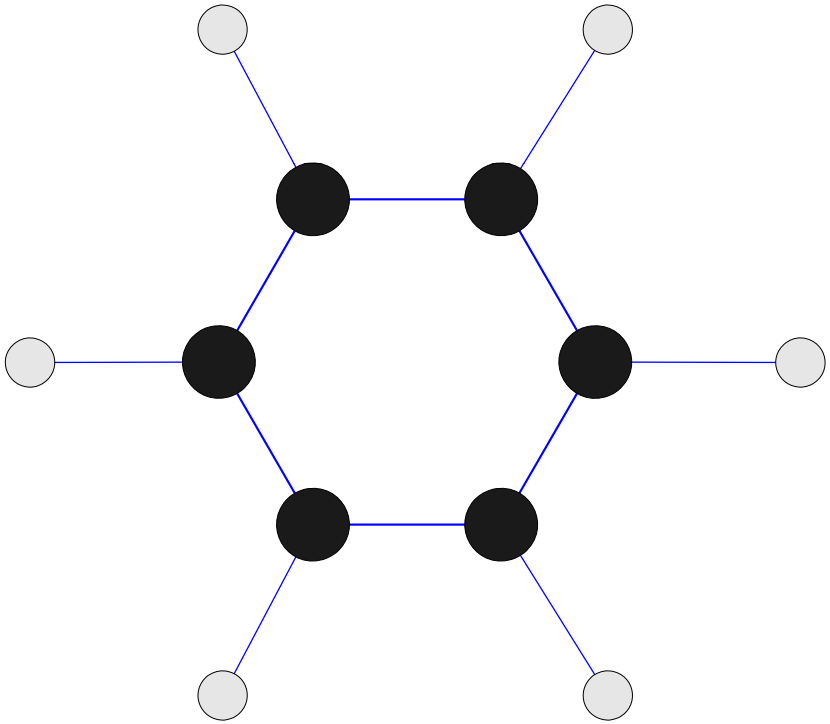
- Examples: Generic graph and HTML Links




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

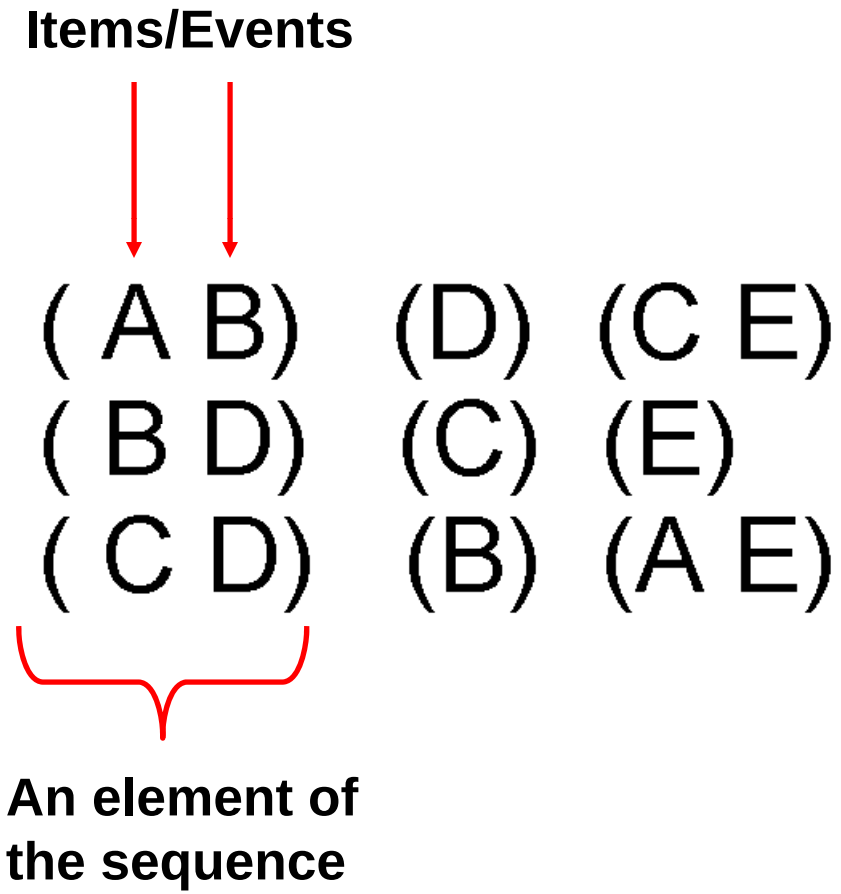
Chemical data

- Benzene molecule: C_6H_6



Ordered data

- Sequences of transactions



Ordered data

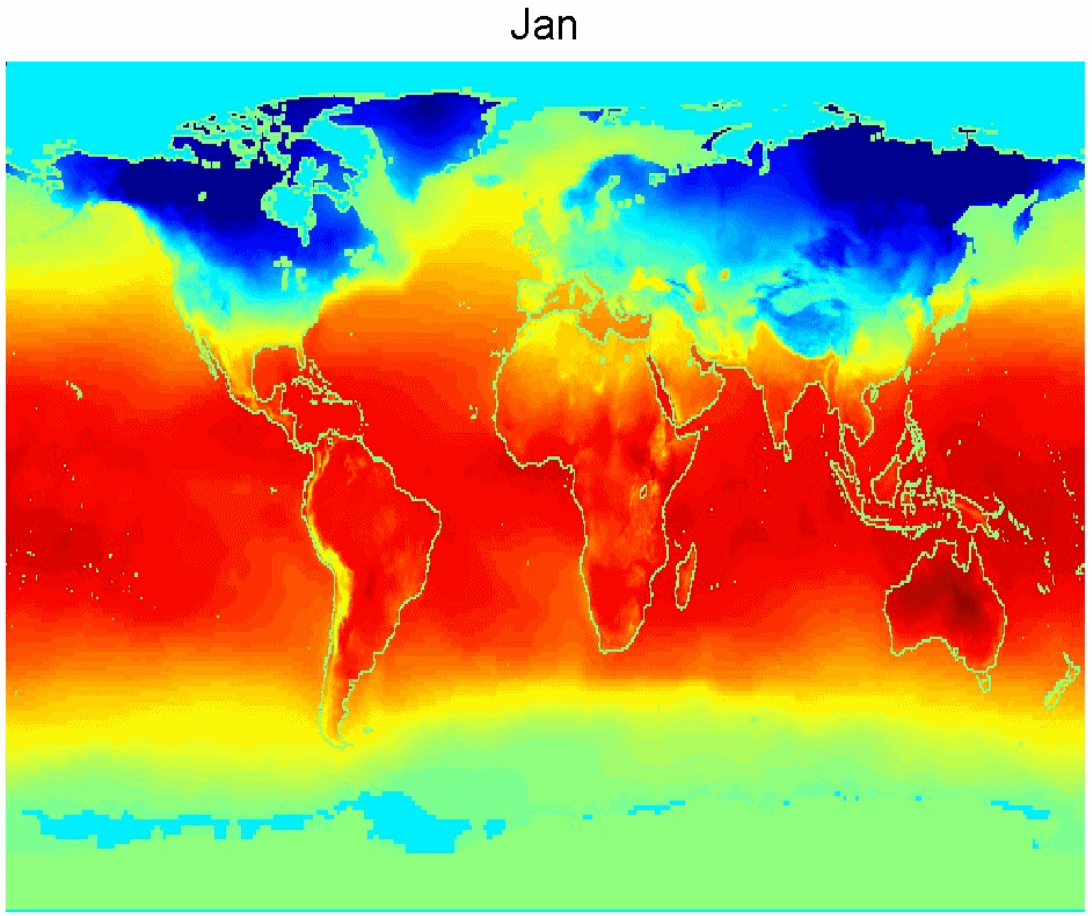
- Genomic sequence data

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

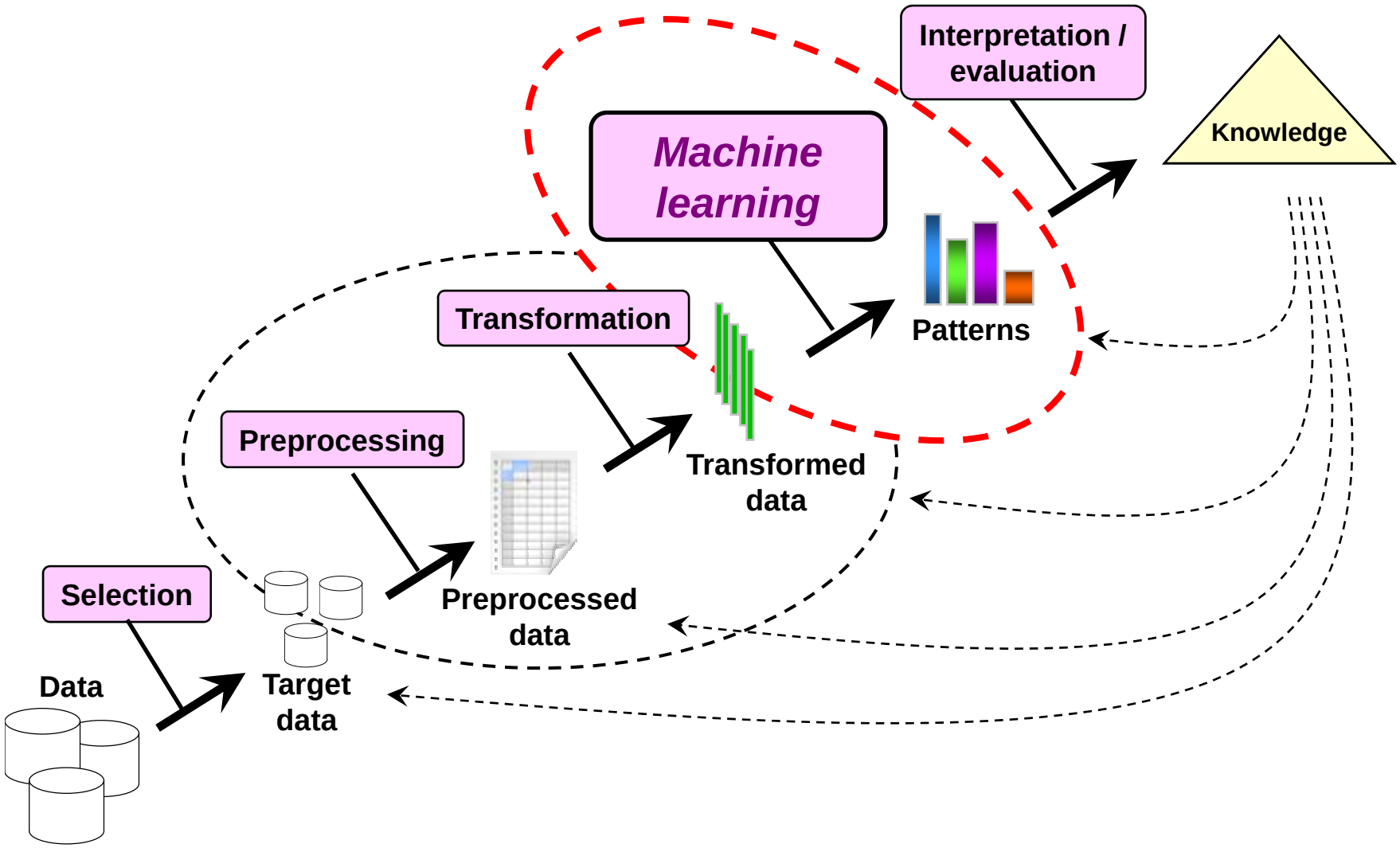
Ordered data

- Spatio-temporal data

Average monthly temperature of land and ocean



Stages of knowledge extraction



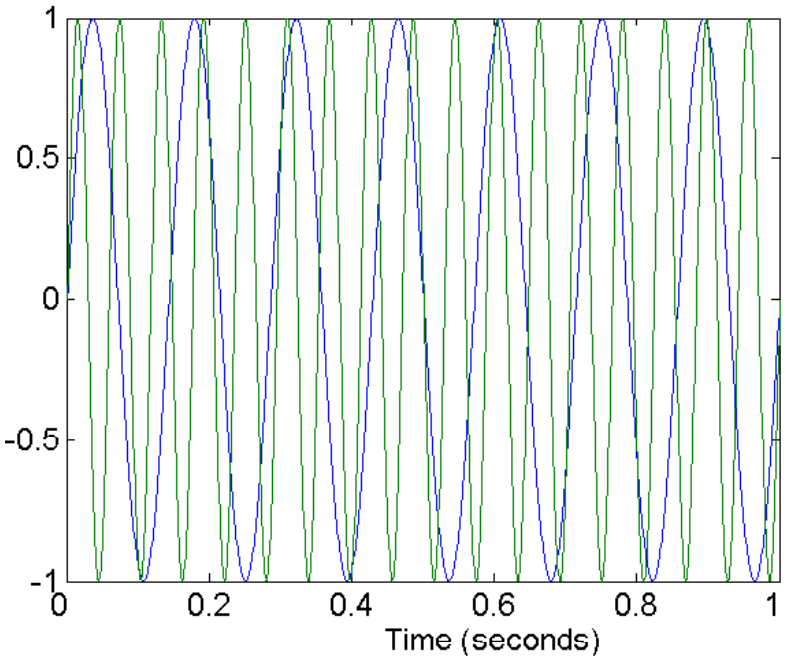
Data quality

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

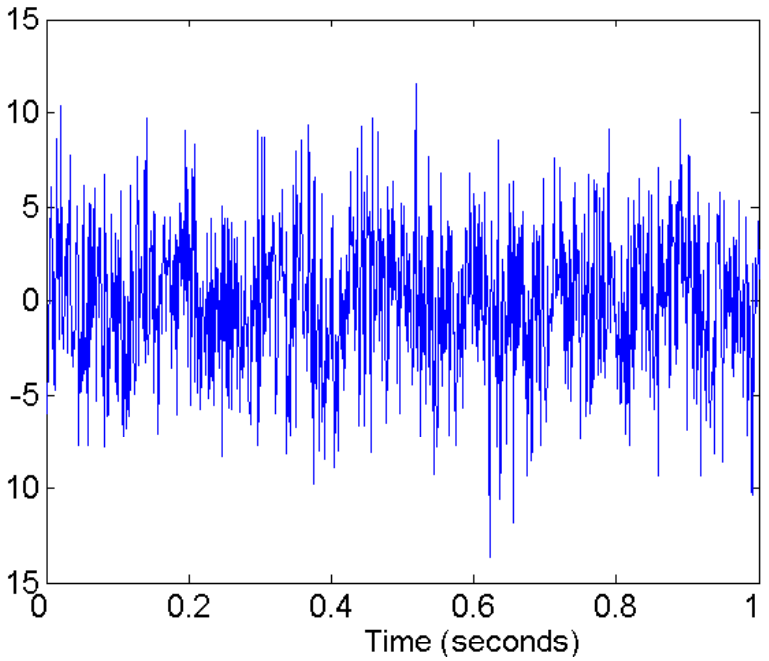
- Examples of data quality problems:
  - noise and outliers
  - missing values
  - duplicate data

Noise

- Noise refers to random modification of original values
- Examples:
  - distortion of a person's voice when talking on a poor phone
  - "snow" on television screen



Two sine waves



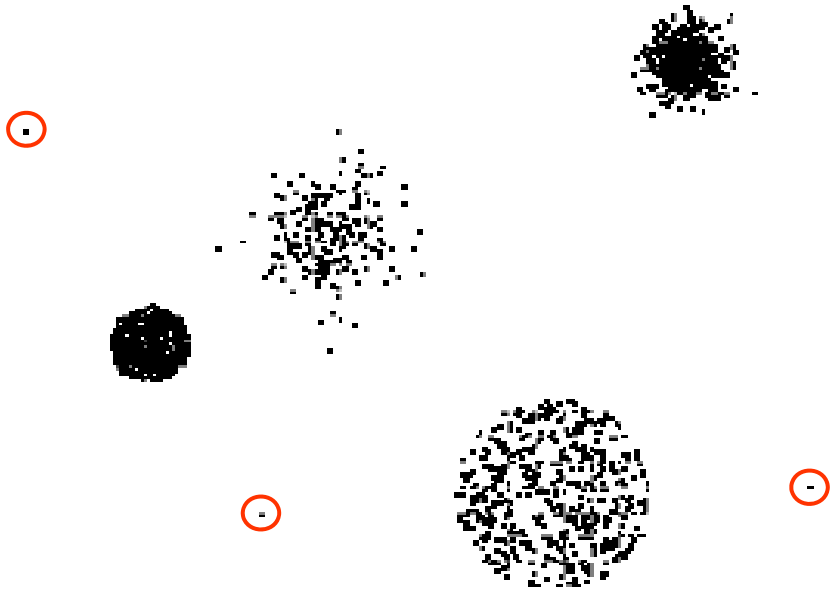
Two sine waves + noise

Noise

- Dealing with noise
  - Mostly you have to live with it
  - Certain kinds of smoothing or averaging can be helpful
  - In the right domain (e.g. signal processing), transformation to a different space can get rid of majority of noise

Outliers

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



Outliers

- Dealing with outliers
  - There are robust statistical methods for detecting outliers
  - In some situations, you want to get rid of outliers
    - but be judicious – they may carry useful, even important information
  - In other situations, the outliers are the objects of interest
    - anomaly 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
  - 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 heterogeous sources
- Example:
  - Same person with multiple email addresses
- Data cleaning
  - Includes process of dealing with duplicate data issues

Data preprocessing

- Aggregation
- Sampling
- Discretization and binarization
- Attribute transformation
- Feature creation
- Feature selection
  - Choose subset of existing features
- Dimensionality reduction
  - Create smaller number of new features through linear or nonlinear combination of existing features

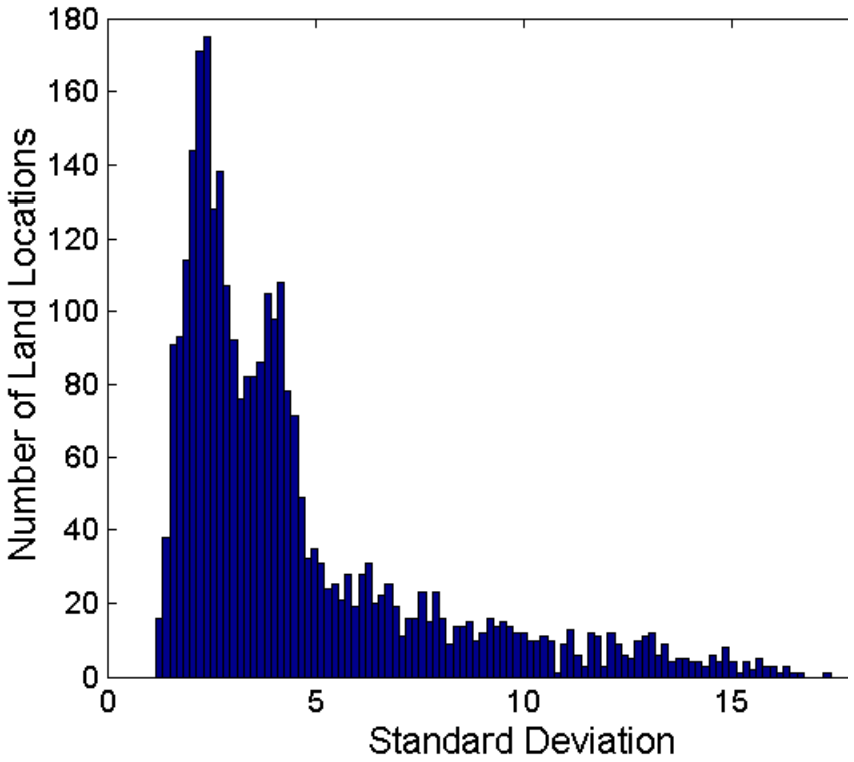
Aggregation

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

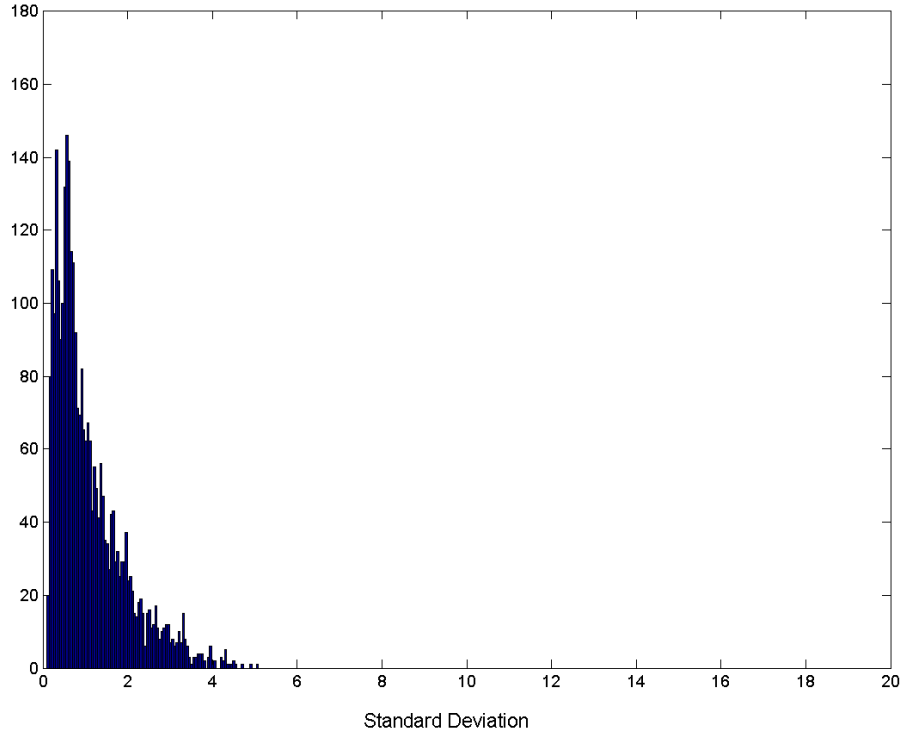
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

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.
  - Often used for both 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.

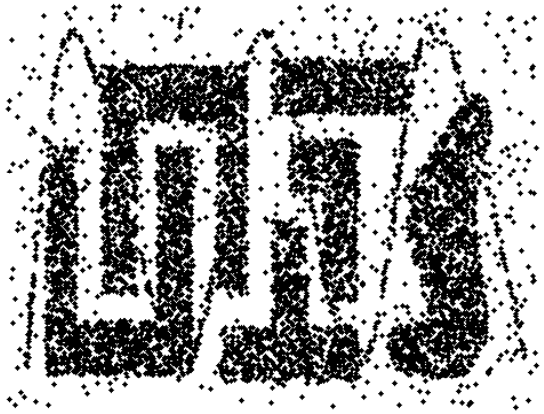
Sampling

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

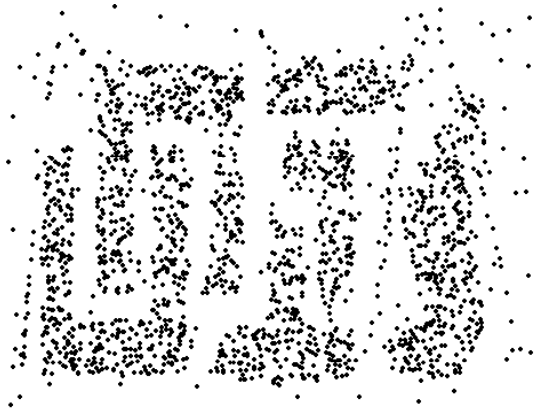
Types of sampling

- 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.
    - Same object can be selected more than once.
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition.
  - Example: During polling, you might want equal numbers of male and female respondents. You create separate pools of men and women, and sample separately from each.

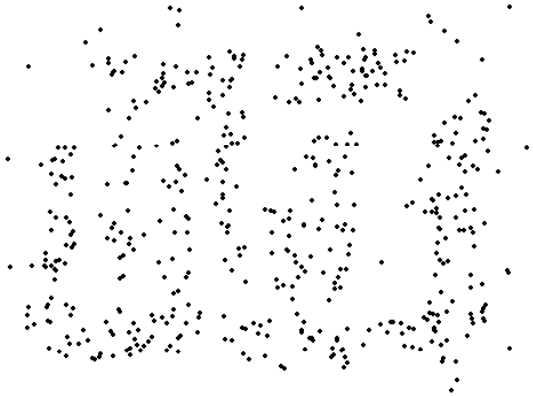
Sample size



8000 points



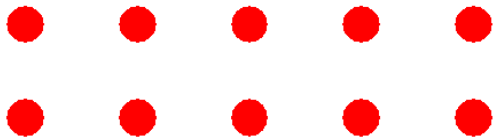
2000 Points

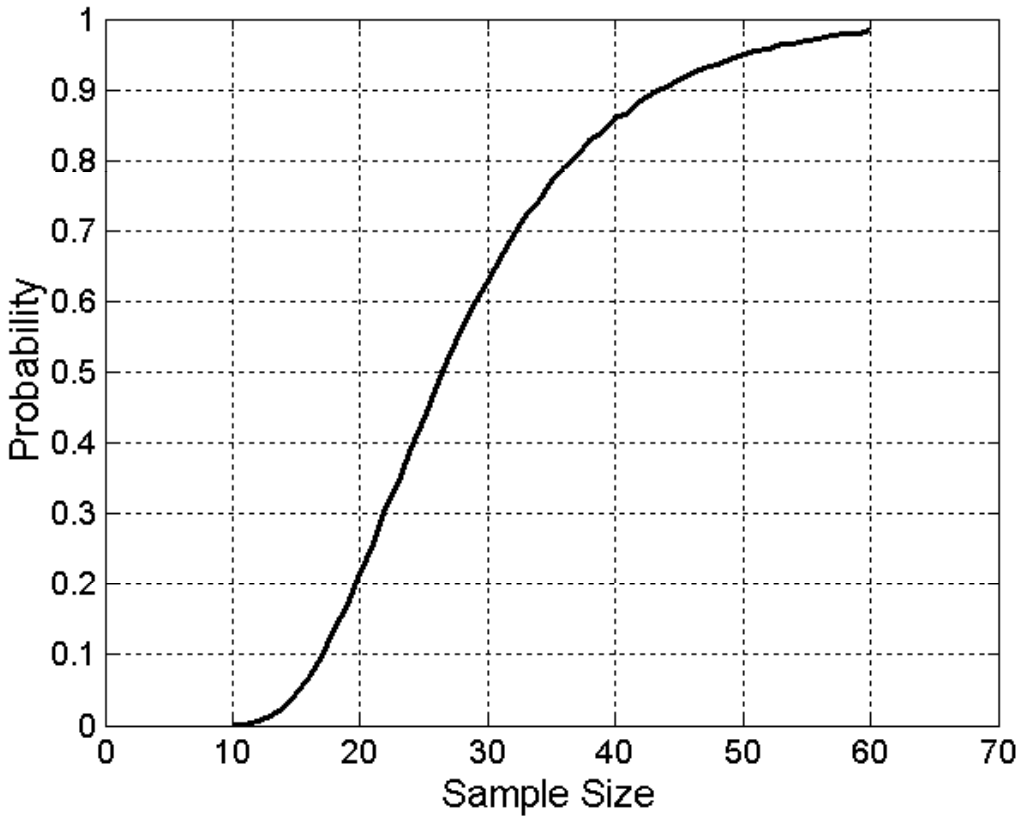


500 Points

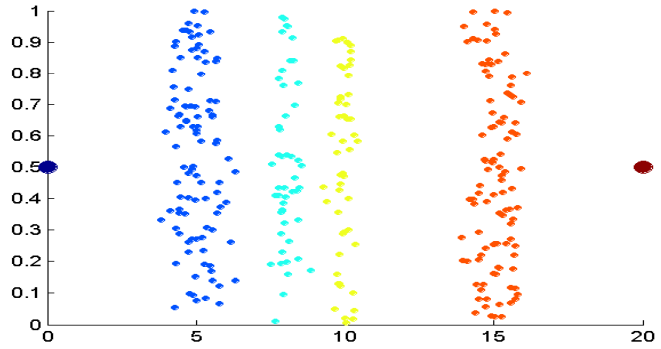
Sample size

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

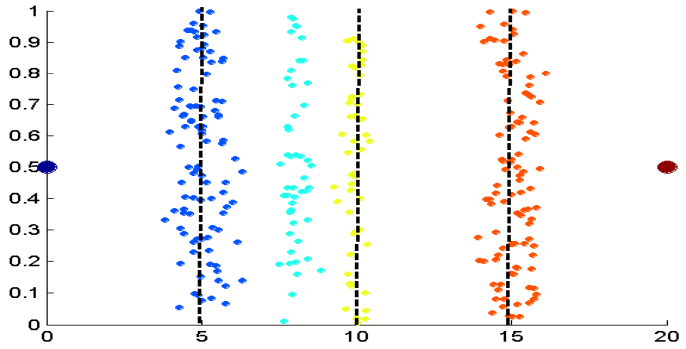




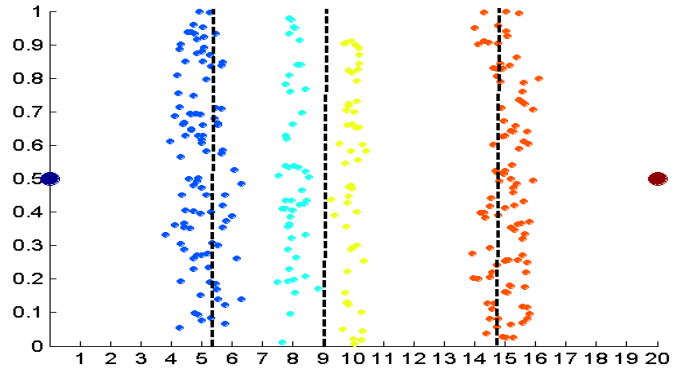
Discretization without using class labels



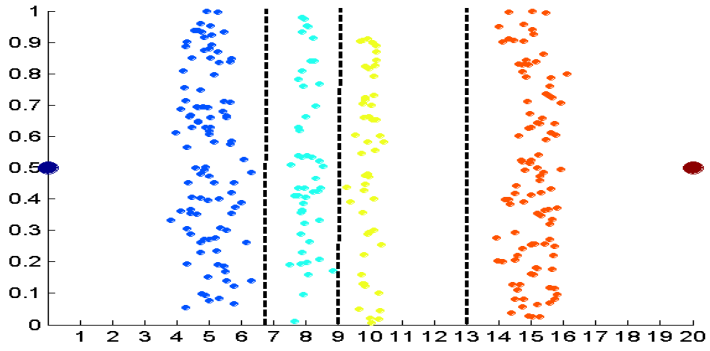
Data



Equal interval width



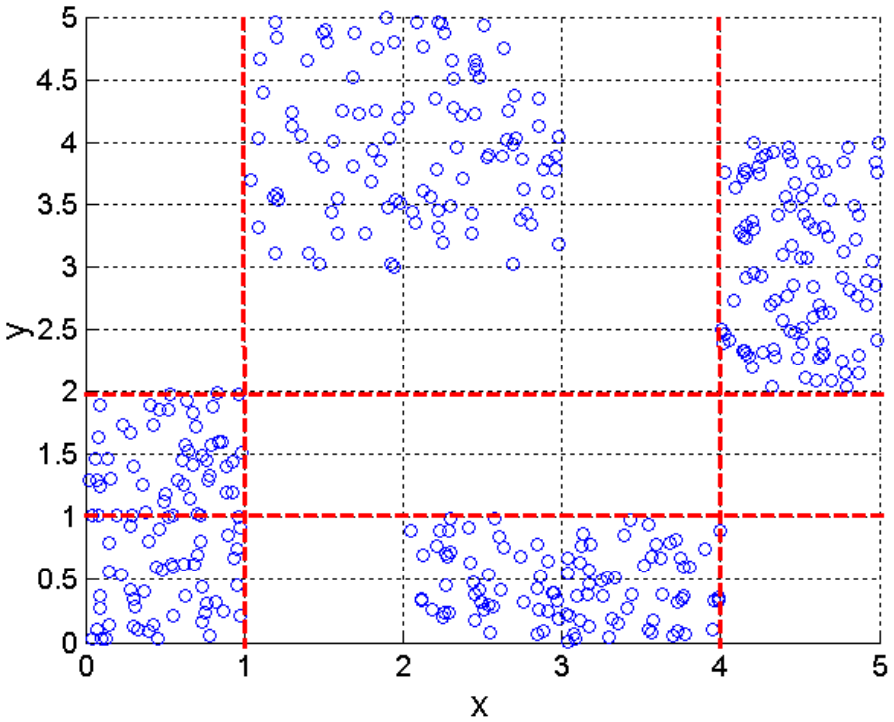
Equal frequency



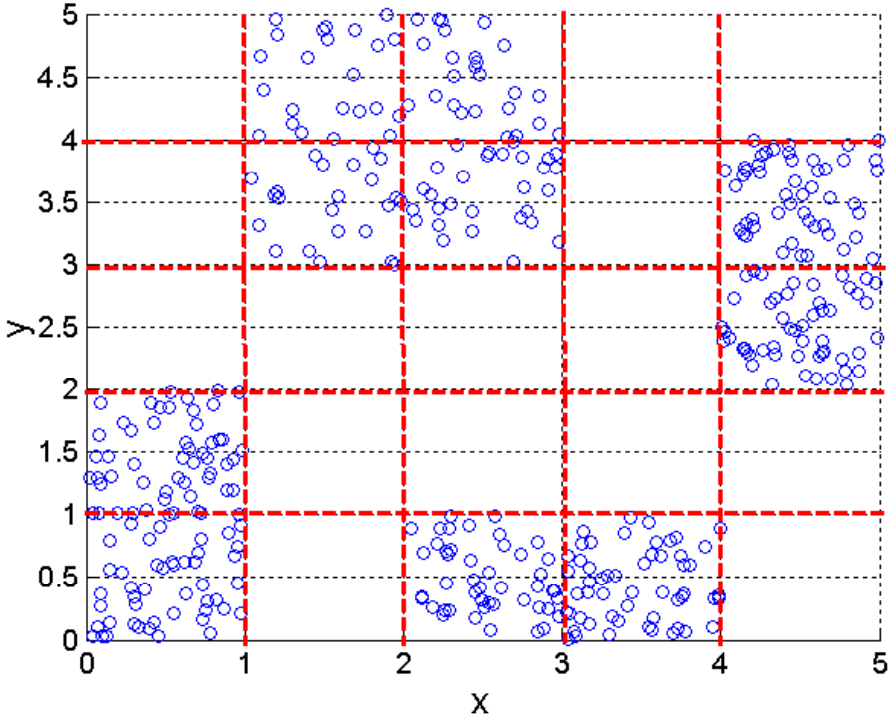
k-means

Discretization using class labels

- Entropy based approach



3 categories for both x and y



5 categories for both x and y

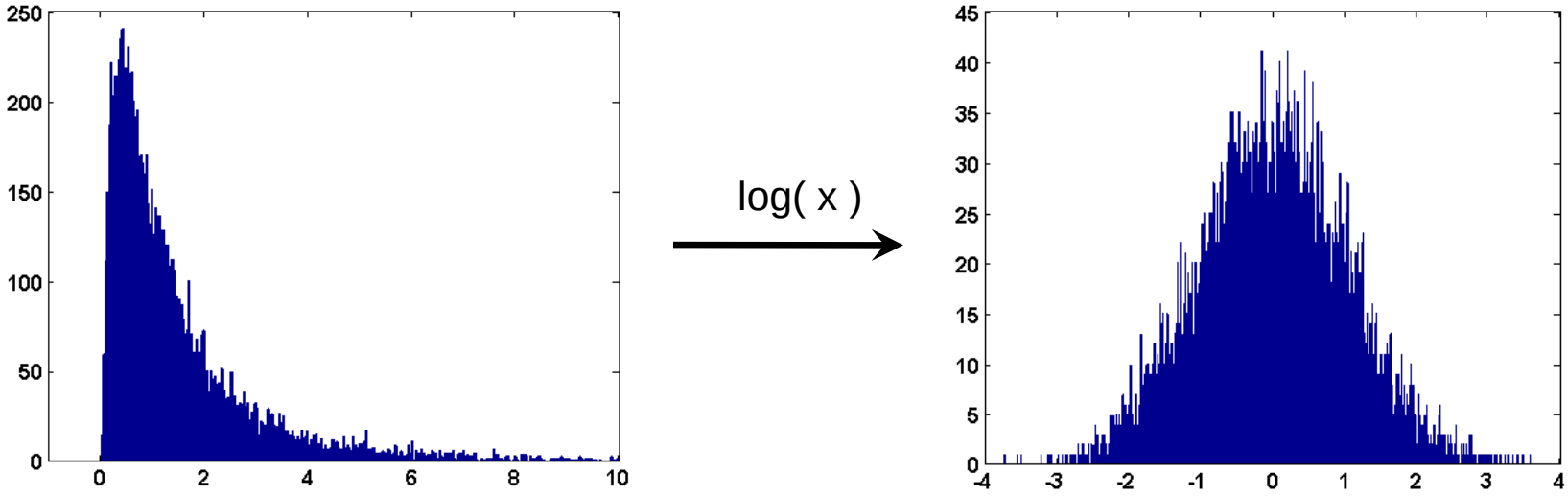
Attribute transformation

Definition:

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.

Attribute transformation

- Simple functions
  - Examples of transform functions:
    x^k $\log(x)$ e^x $|x|$
  - Often used to make the data more like some standard distribution, to better satisfy assumptions of a particular algorithm.
    - Example: discriminant analysis explicitly models each class distribution as a multivariate Gaussian



Attribute transformation

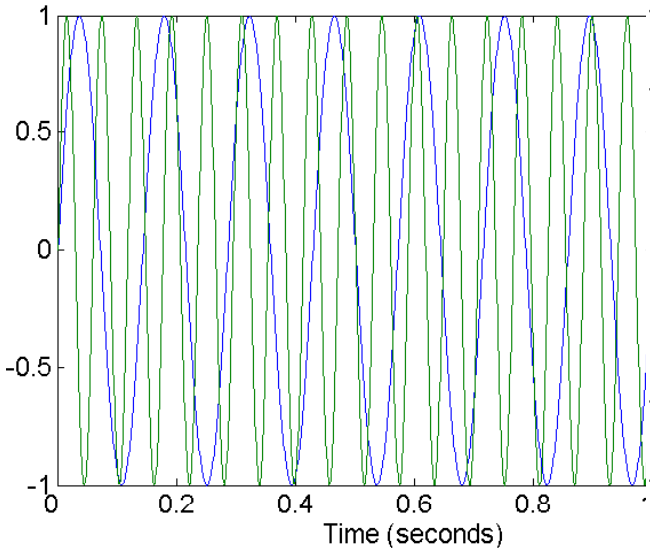
- Standardization or normalization
  - Usually involves making attribute:

    mean = 0

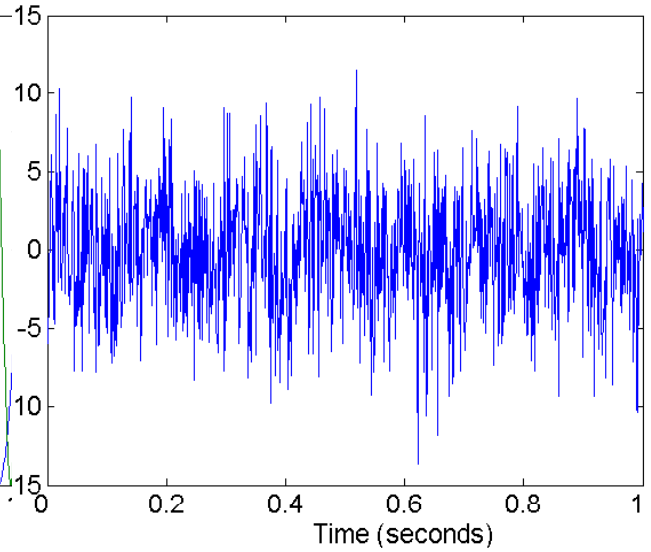
    standard deviation = 1
    - in MATLAB, use zscore() function
  - Important when working in Euclidean space and attributes have very different numeric scales.
  - Also necessary to satisfy assumptions of certain algorithms.
    - Example: principal component analysis (PCA) requires each attribute to be mean-centered (i.e. have mean subtracted from each value)

Transform data to a new space

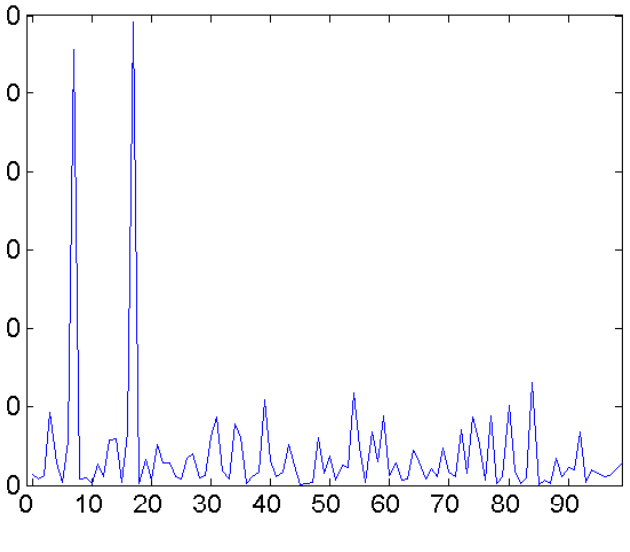
- Fourier transform
  - Eliminates noise present in time domain



Two sine waves



Two sine waves + noise



Frequency

Summary statistics and visualization

Let’s use a tool that’s good at those things …

PowerPoint isn’t it

Iris dataset

- Many exploratory data techniques are nicely illustrated with the iris dataset.
  - Dataset created by famous statistician Ronald Fisher
  - 150 samples of three species in genus *Iris* (50 each)
    - *Iris setosa*
    - *Iris versicolor*
    - *Iris virginica*
  - Four attributes
    - sepal width
    - sepal length
    - petal width
    - petal length
  - Species is class label

***Iris virginica*. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.**