

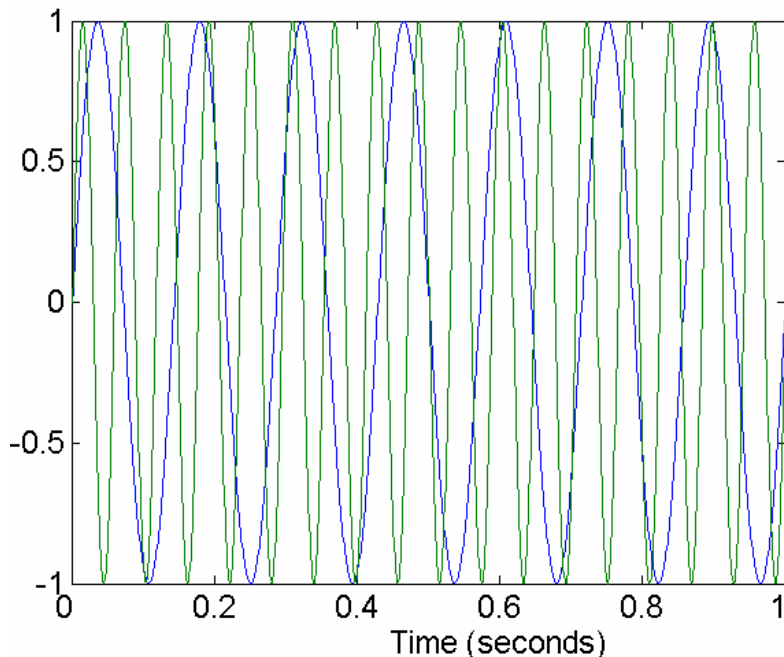
Data Quality

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

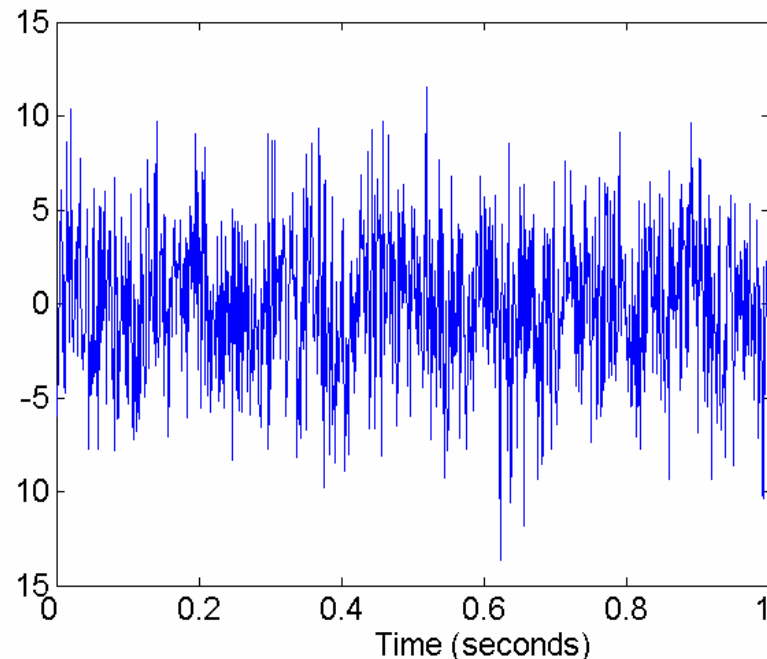
- Examples of data quality problems:
 - Noise and outliers
 - missing values
 - duplicate data

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



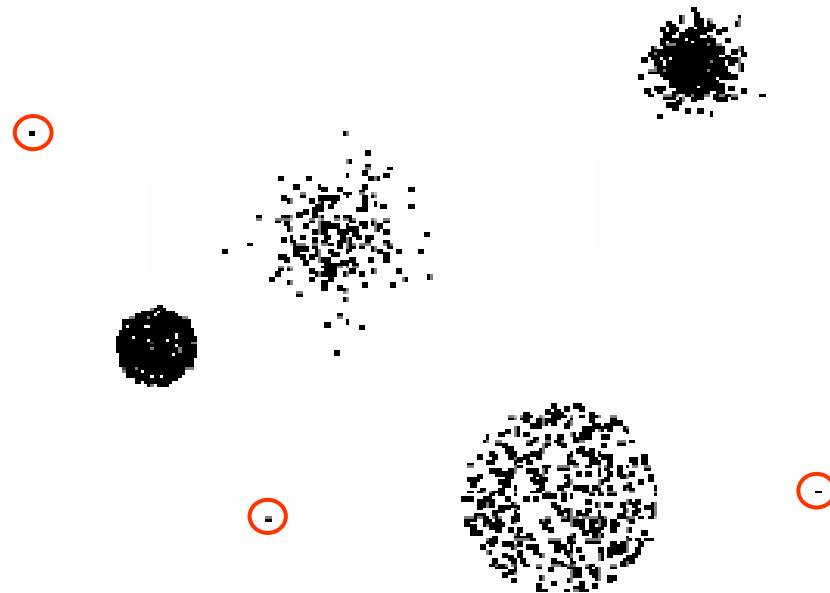
Two Sine Waves



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



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 heterogonous sources
- Examples:
 - Same person with multiple email addresses
- Data cleaning
 - Process of dealing with duplicate data issues

Data Preprocessing

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

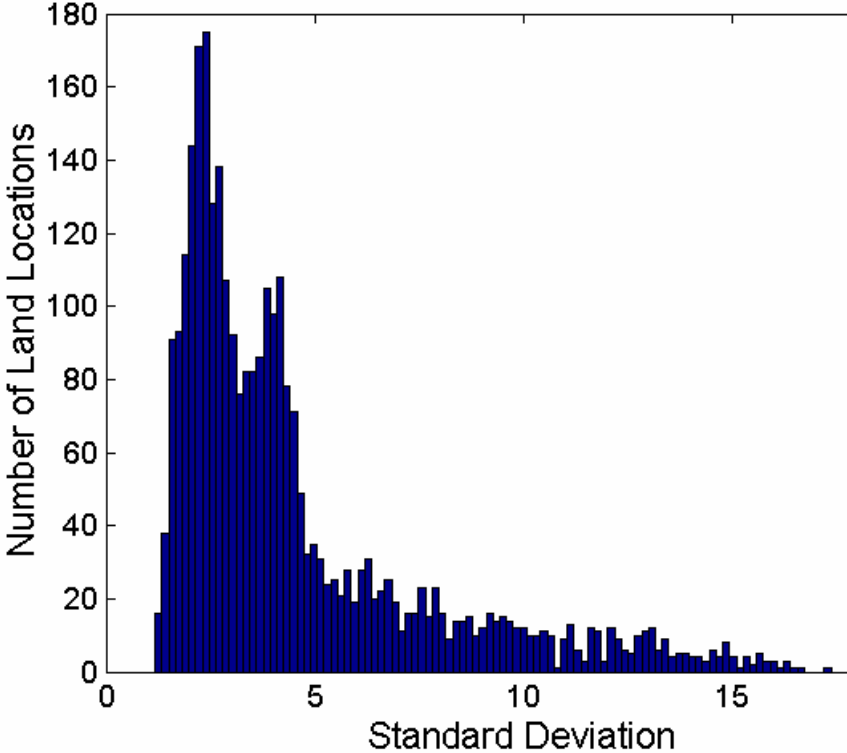
Aggregation

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

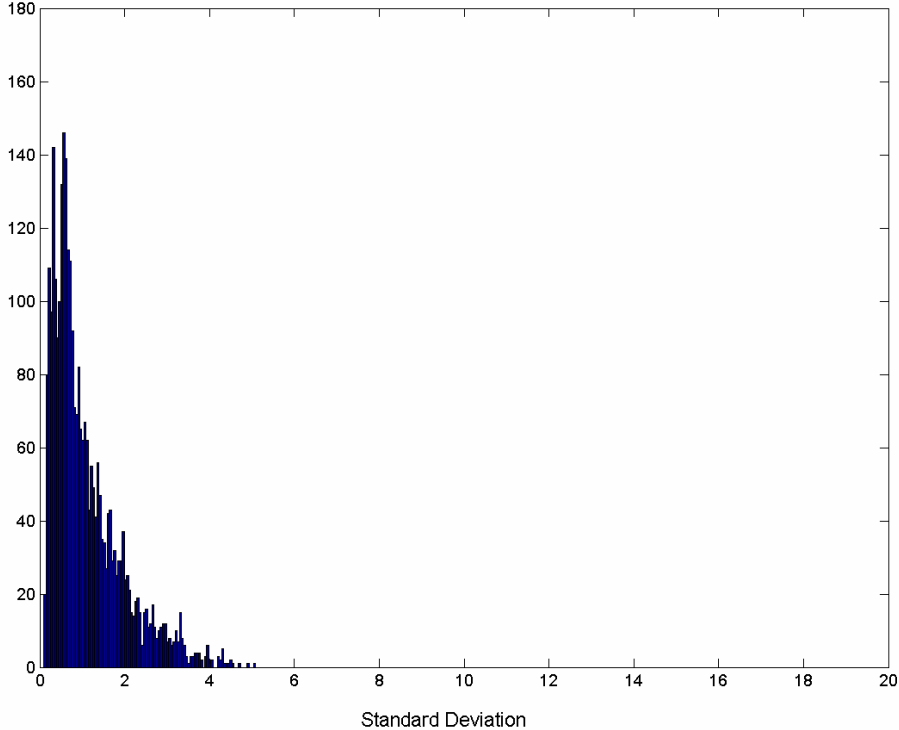
- Purpose
 - Data reduction
 - ◆ reduce the number of attributes or objects
 - Change of scale
 - ◆ cities aggregated into regions, states, countries, etc
 - More “stable” data
 - ◆ aggregated data tends to have less variability

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 it is too expensive or time consuming to process all the data

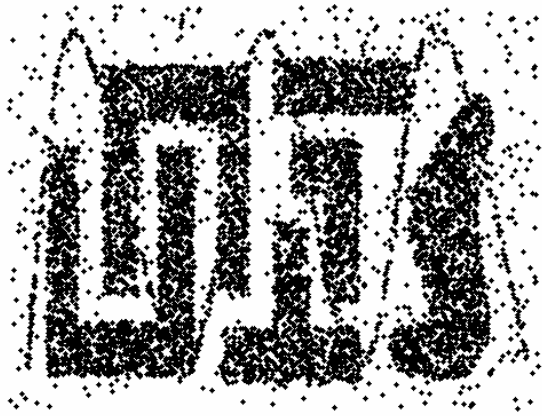
Sampling ...

- 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

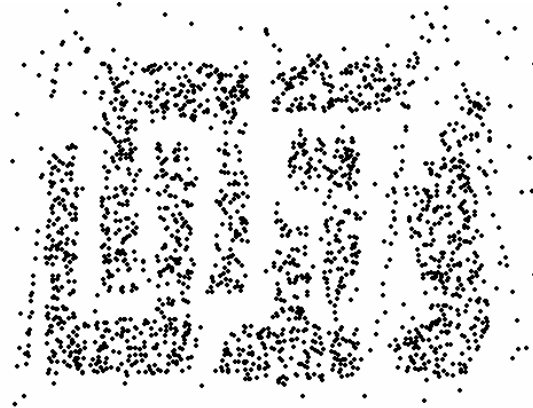
Types of Sampling

- 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

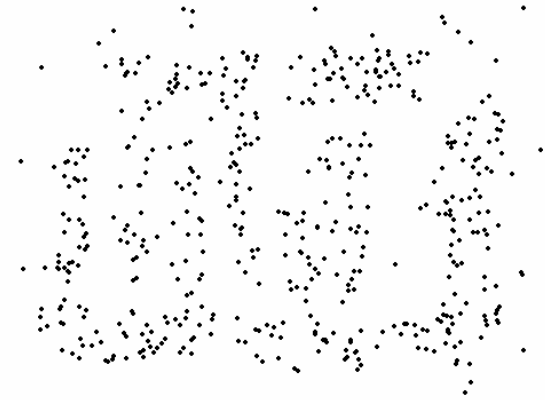
Sample Size



8000 points



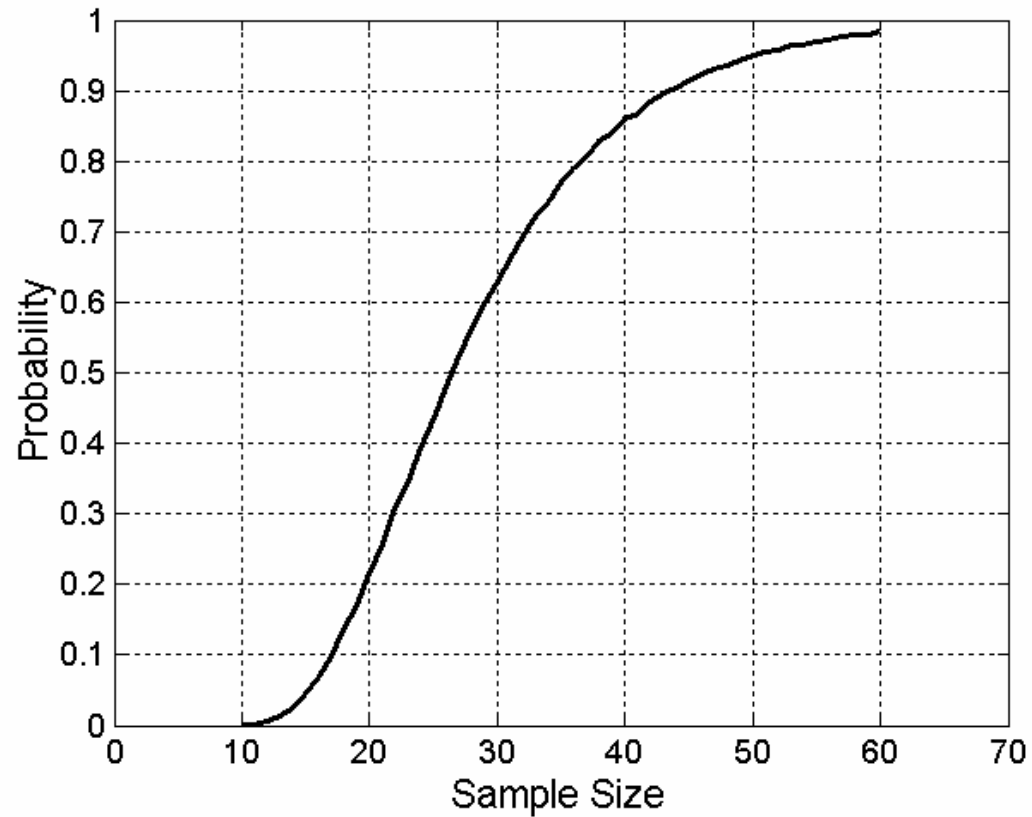
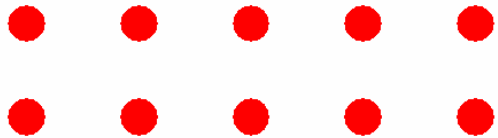
2000 Points



500 Points

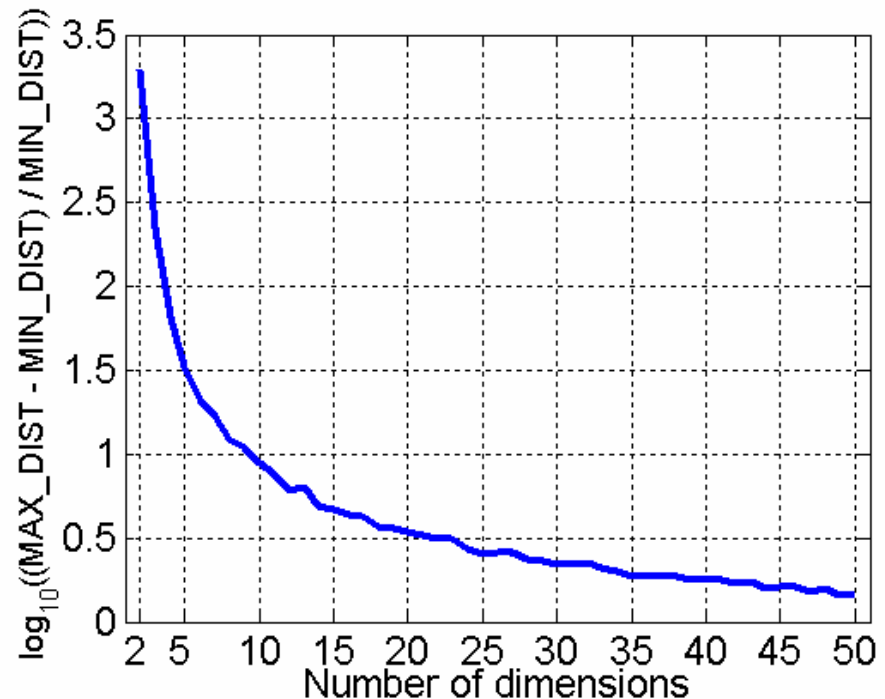
Sample Size

- **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

Feature Subset Selection

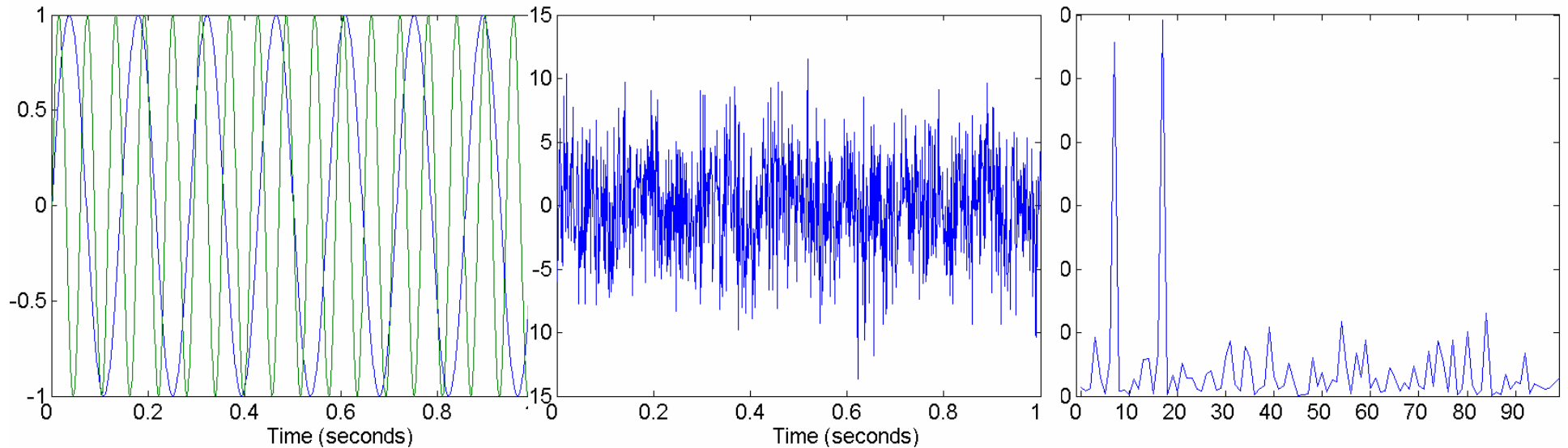
- Another way to reduce dimensionality of data
- Redundant features
 - duplicate much or all of the information contained in one or more other attributes
 - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
 - contain no information that is useful for the data mining task at hand
 - Example: students' ID is often irrelevant to the task of predicting students' GPA

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

Mapping Data to a New Space

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



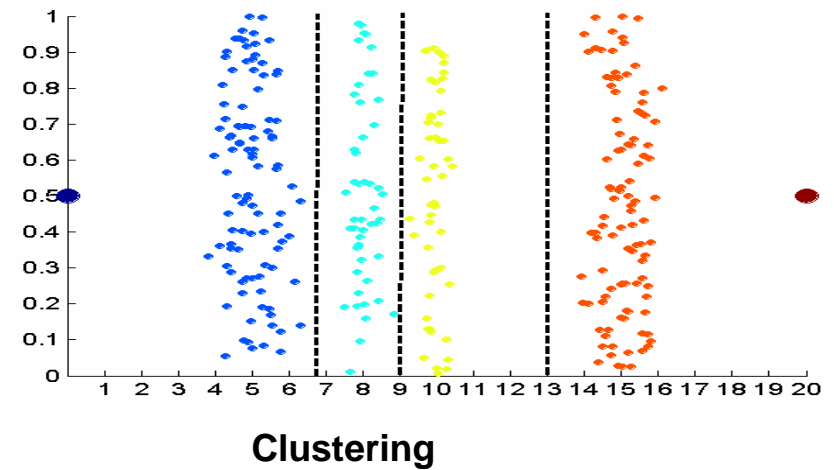
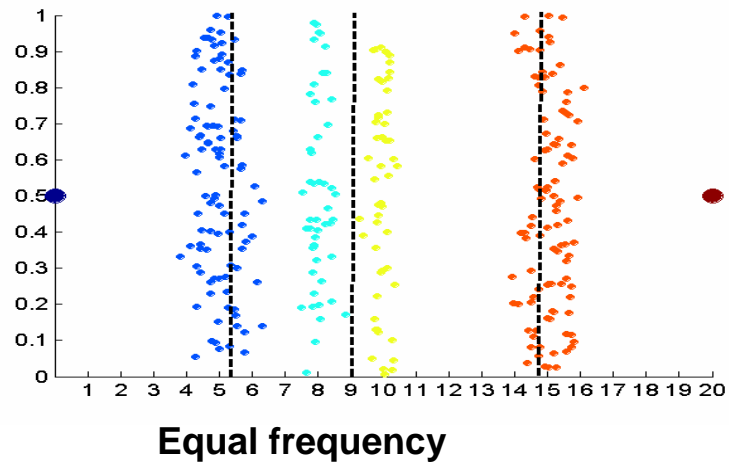
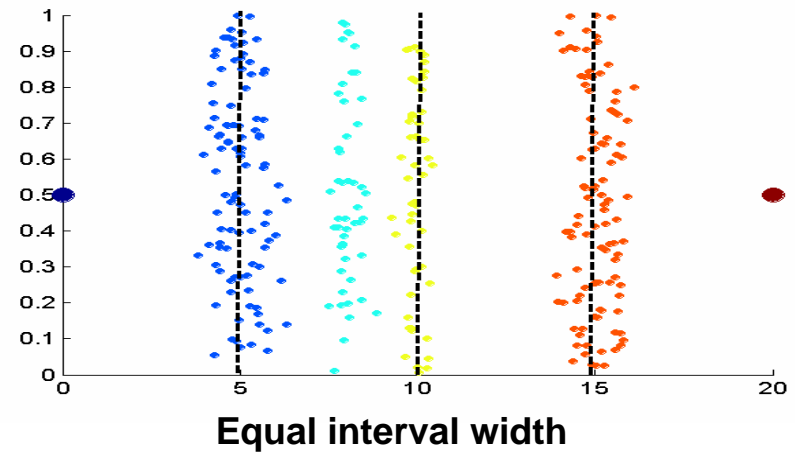
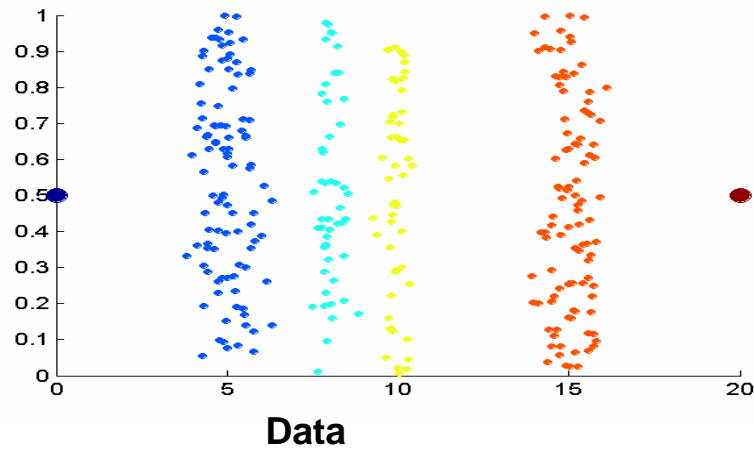
Two Sine Waves

Two Sine Waves + Noise

Frequency

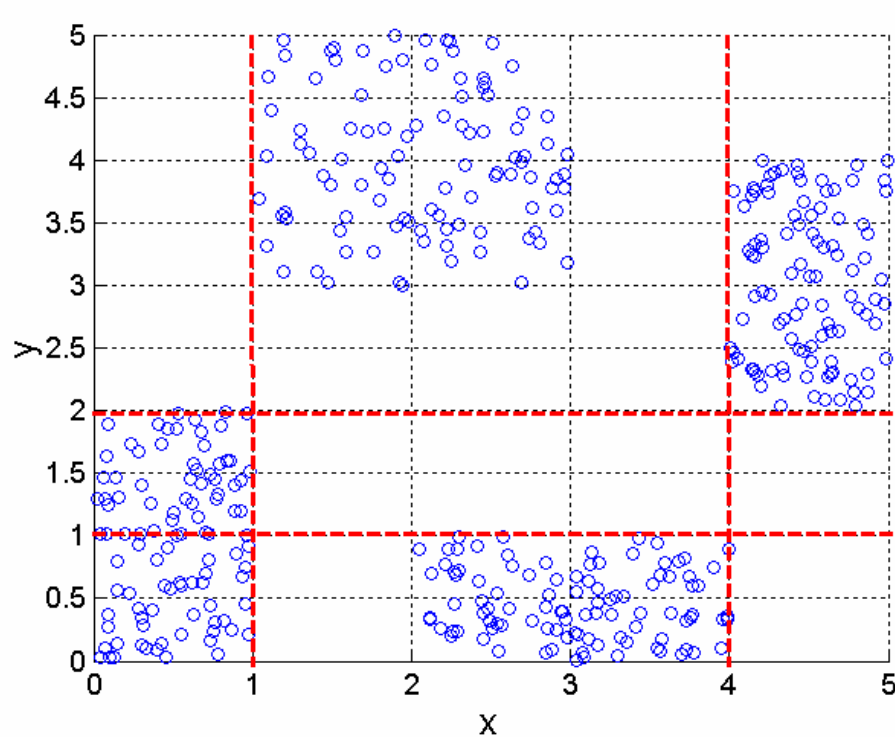
Discretization

- Some techniques don't use class labels.

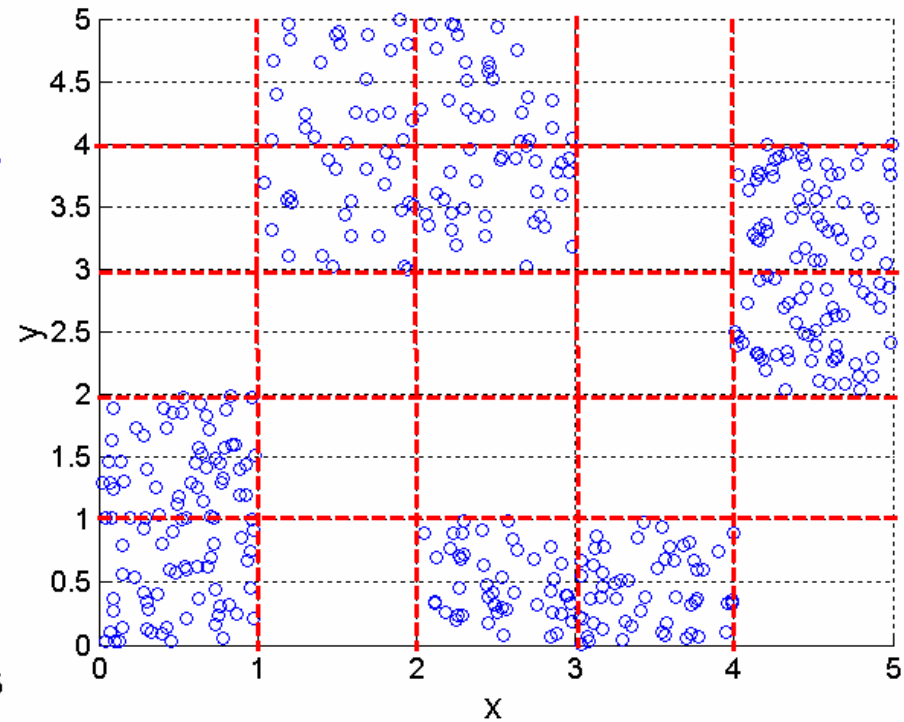


Discretization

- Some techniques use class labels.
- Entropy based approach



3 categories for both x and y



5 categories for both x and y

Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
 - Simple functions: x^k , $\log(x)$, e^x , $|x|$
 - Standardization and Normalization

