Data Preprocessing

Introduction to Data Mining

CS412, 2021 Spring

Gaoang Wang

Data Preprocessing

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

Example: Precipitation in Australia

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

The next slide shows

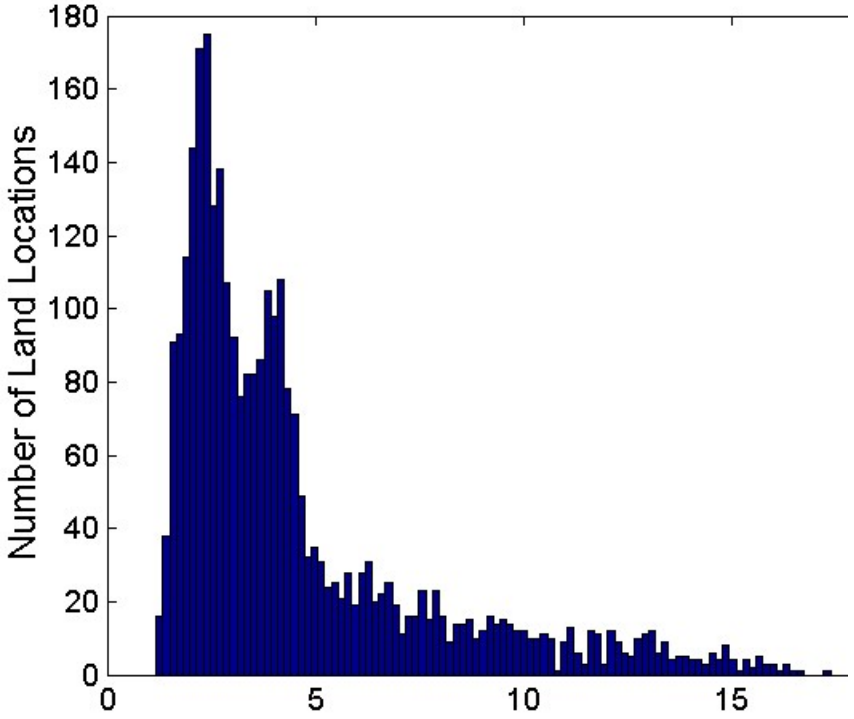
- A histogram for the standard deviation of average monthly precipitation for 3,030 0.5◦ by 0.5◦ grid cells in Australia, and
- A histogram for the standard deviation of the average yearly precipitation for the same locations.

The average yearly precipitation has less variability than the average monthly precipitation.

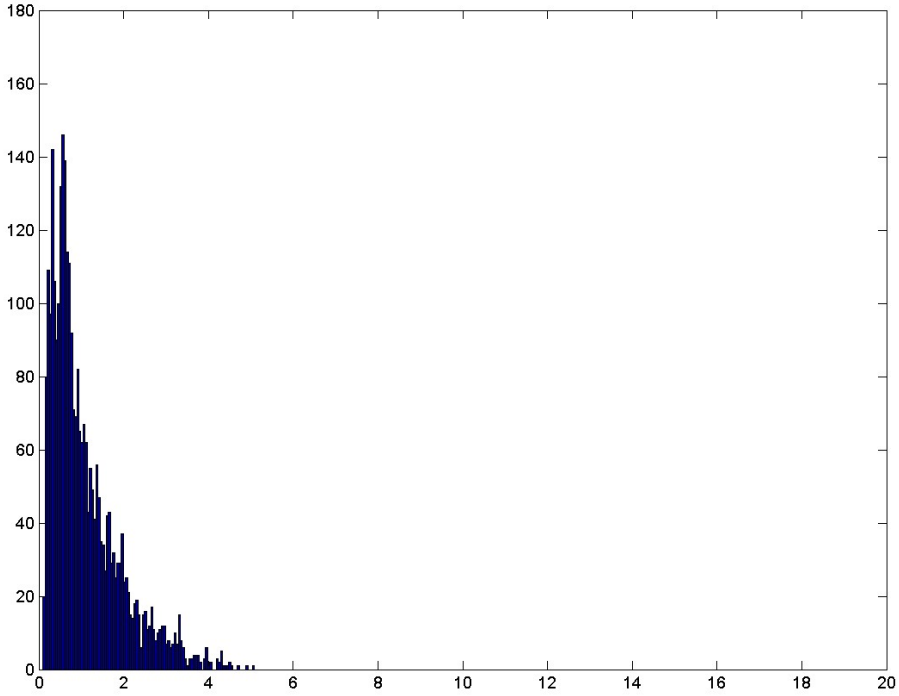
All precipitation measurements (and their standard deviations) are in centimeters.

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.

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

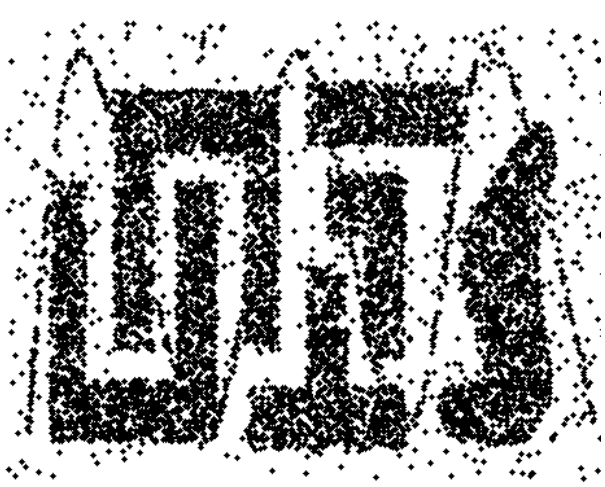
Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.

Sampling …

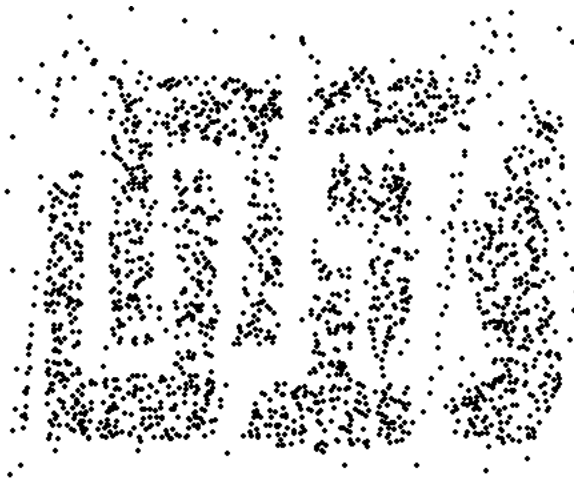
The key principle for effective sampling is the following:

- 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

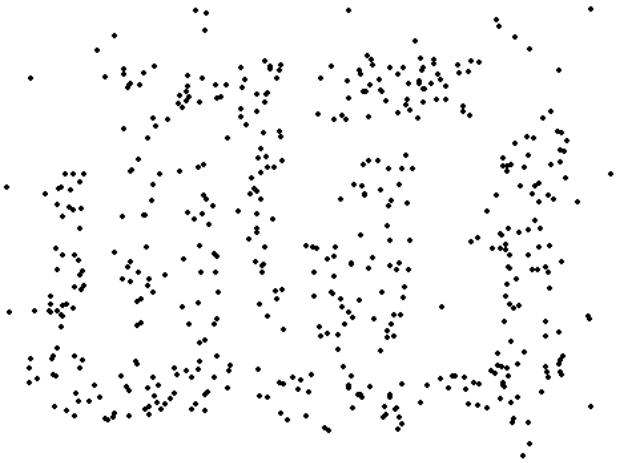
Sample Size



8000 points



2000 Points



500 Points

Types of Sampling

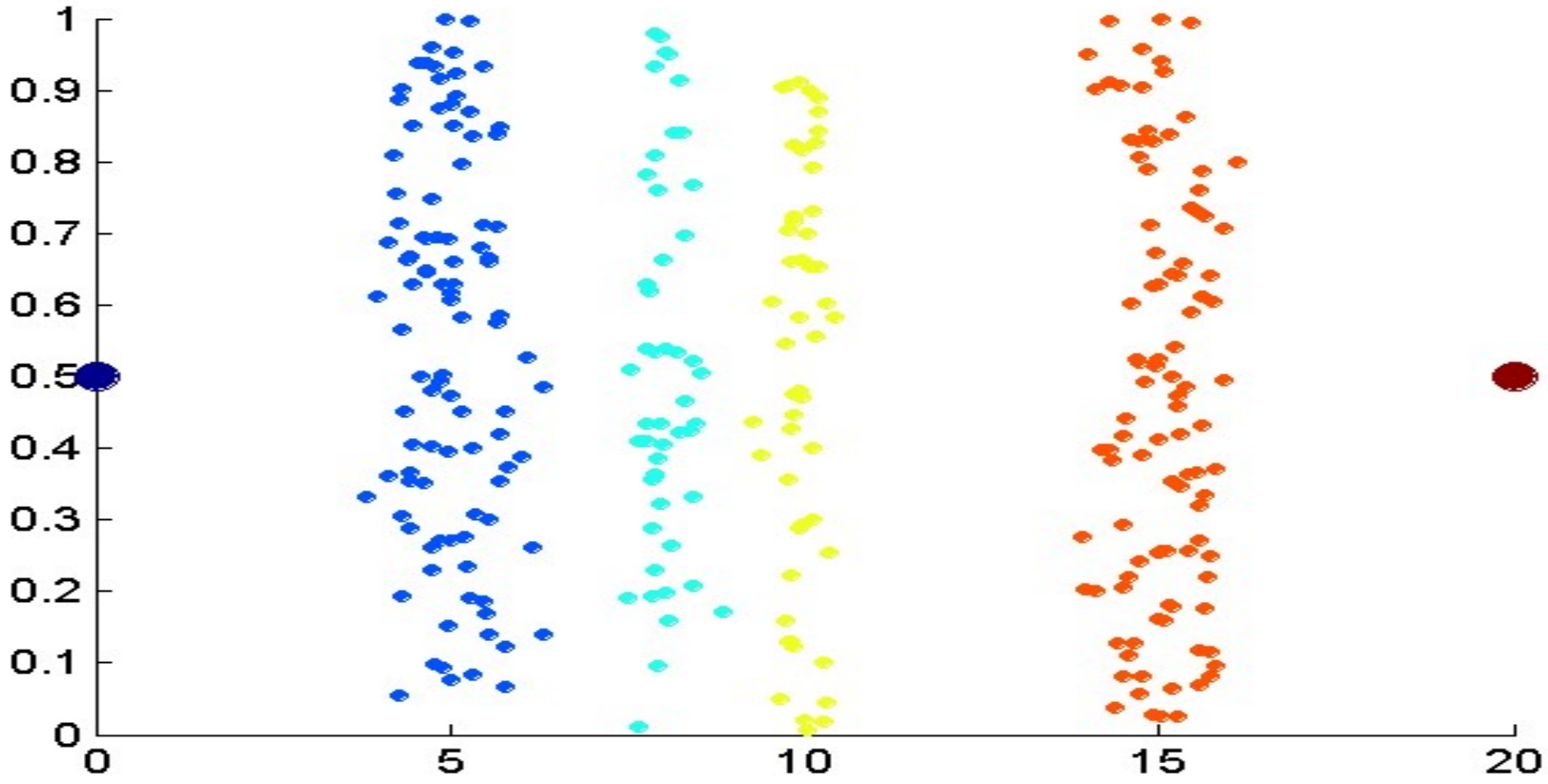
- Simple Random Sampling
  - There is an equal probability of selecting any particular item
  - Sampling without replacement
    - As each item is selected, it is removed from the population
  - Sampling with replacement
    - Objects are not removed from the population as they are selected for the sample.
    - 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

Discretization

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

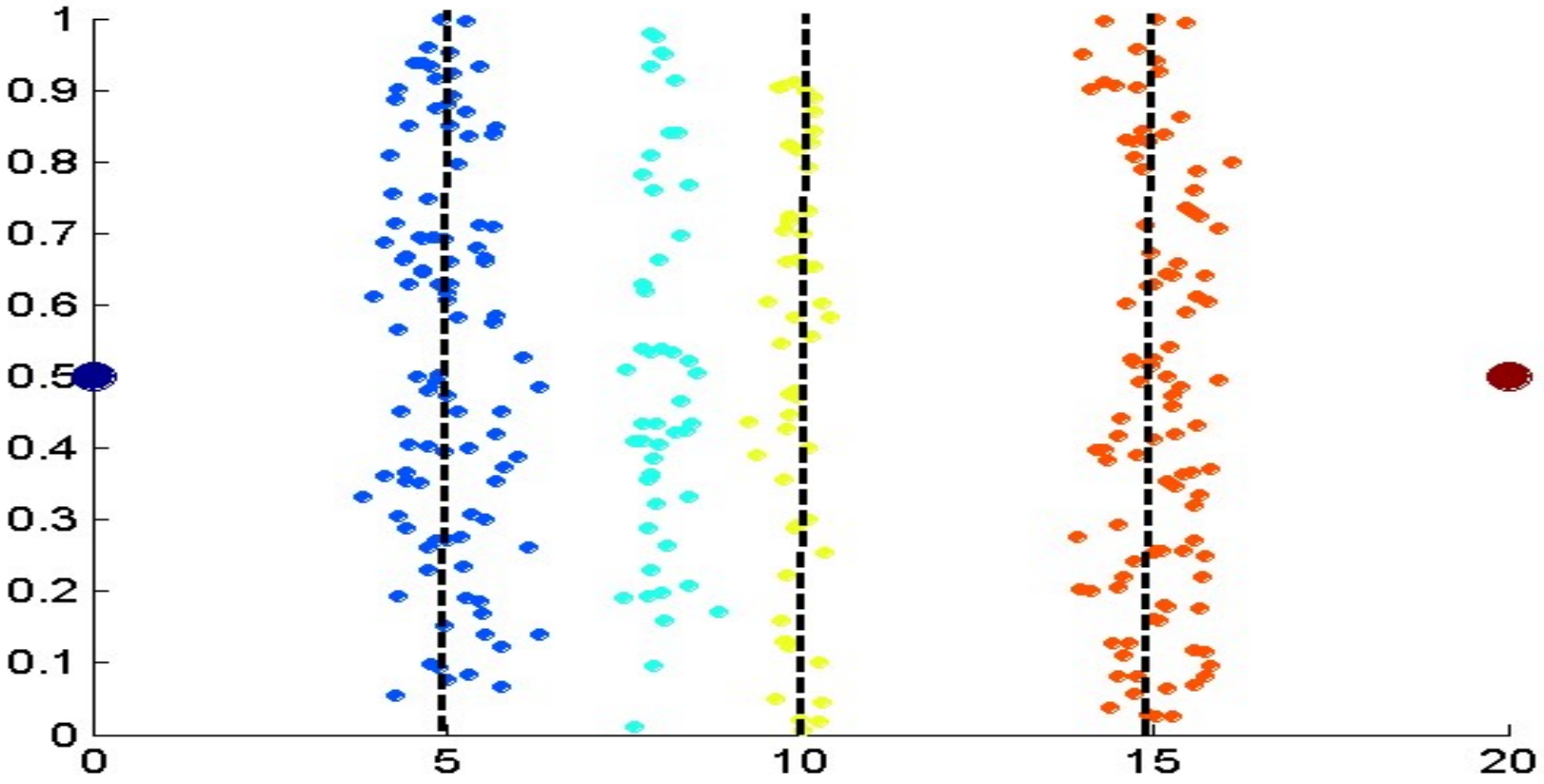
- A potentially infinite number of values are mapped into a small number of categories
- Discretization is used in both unsupervised and supervised settings

Unsupervised Discretization



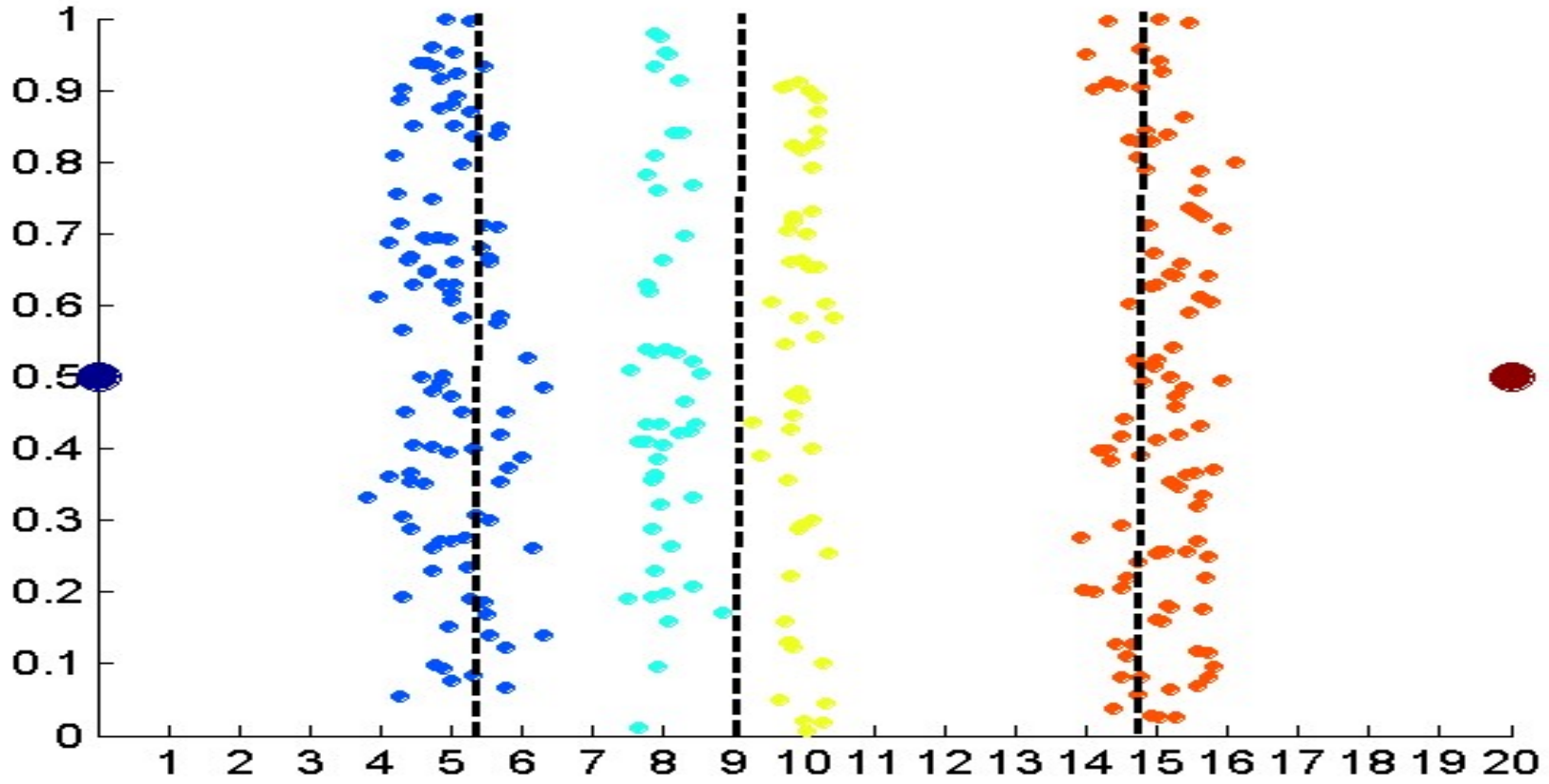
Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

Unsupervised Discretization



Equal interval width approach used to obtain 4 values.

Unsupervised Discretization



Equal frequency approach used to obtain 4 values.

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 Iris data set

Iris Sample Data Set

Iris Plant data set.

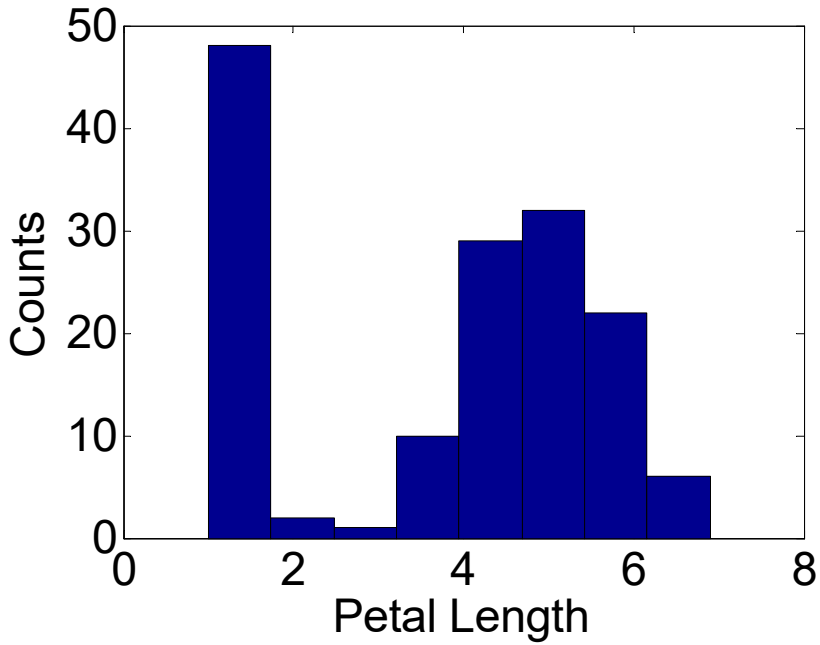
- Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
- From the statistician Douglas Fisher
- Three flower types (classes):
  - Setosa
  - Versicolour
  - Virginica
- Four (non-class) attributes
  - Sepal（萼片）width and length
  - Petal （花瓣）width and length

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.

Discretization: Iris Example …

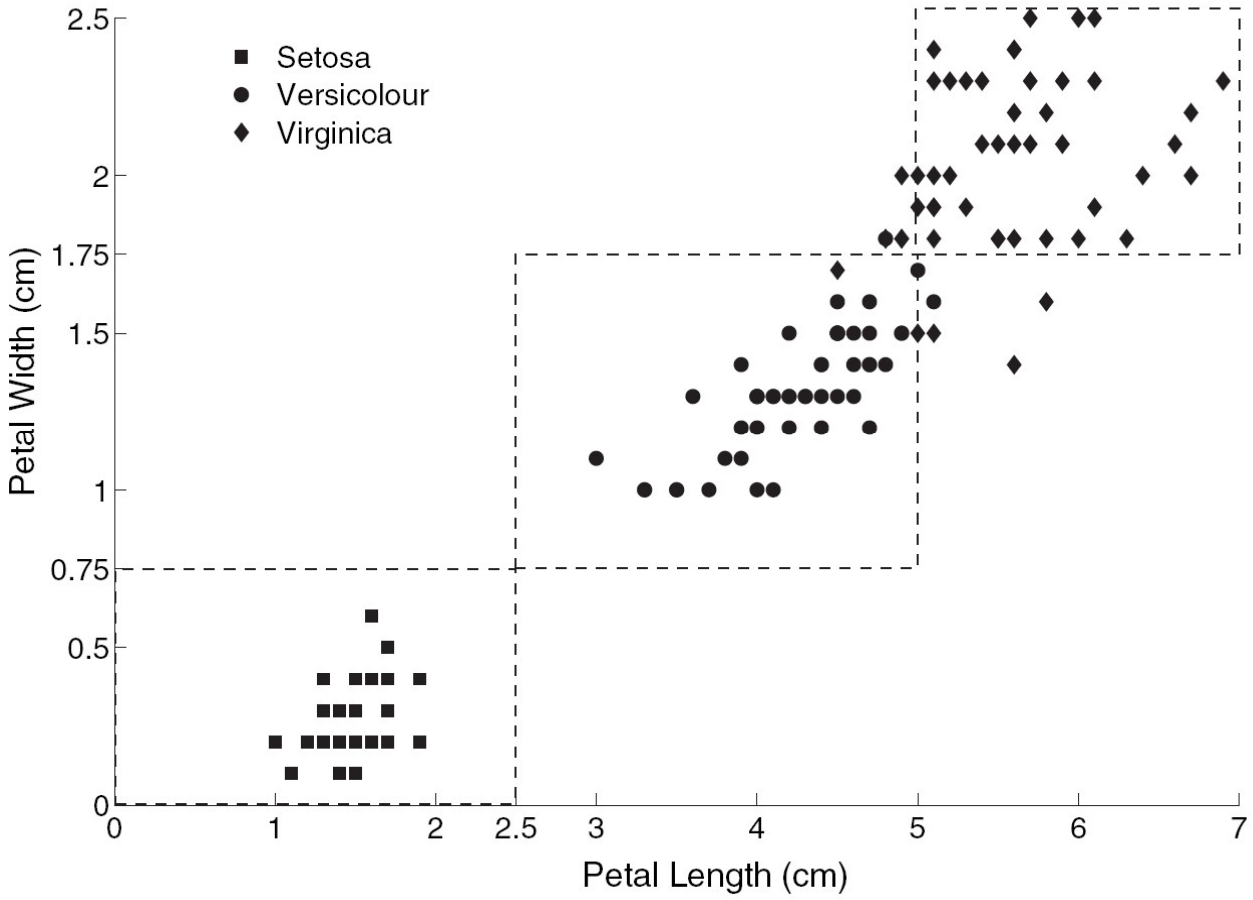
How can we tell what the best discretization is?

- Unsupervised discretization: find breaks in the data values
  - Example: Petal Length



- Supervised discretization: Use class labels to find breaks

Discretization: Iris Example



Petal width low or petal length low implies Setosa.
Petal width medium or petal length medium implies Versicolour.
Petal width high or petal length high implies Virginica.

Binarization

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

Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes

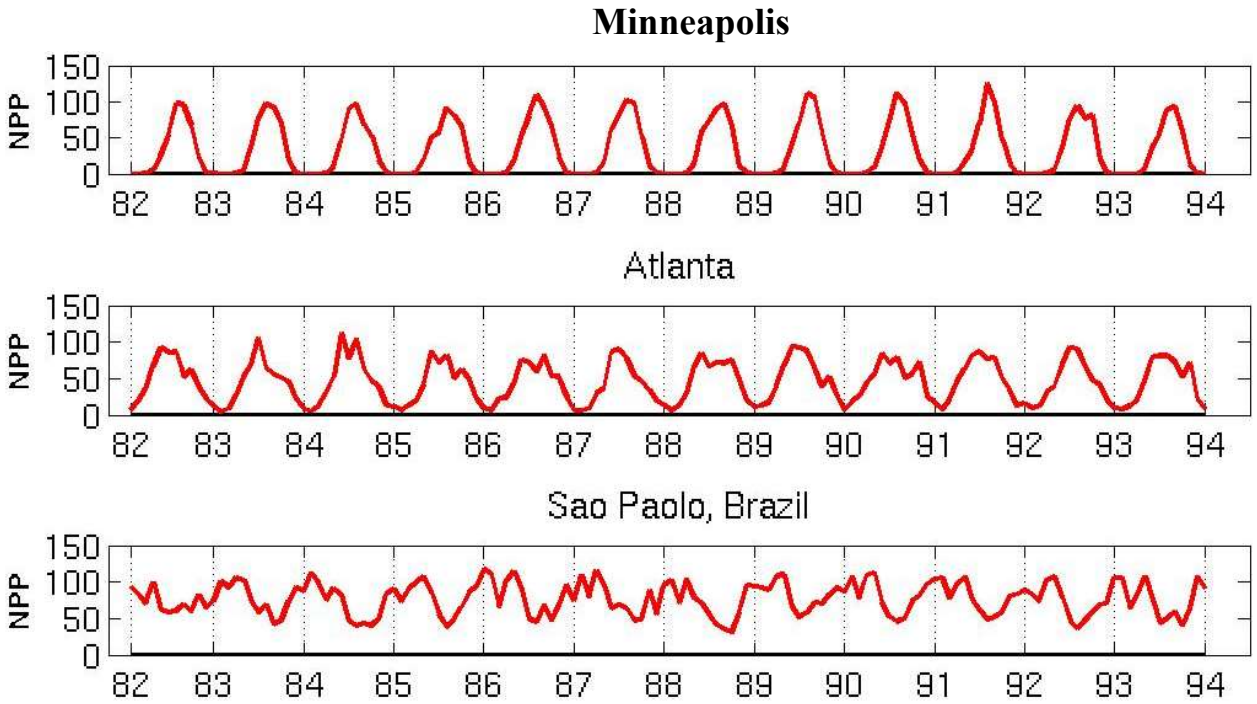
- Examples: eye color and height measured as {low, medium, high}

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
- 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.

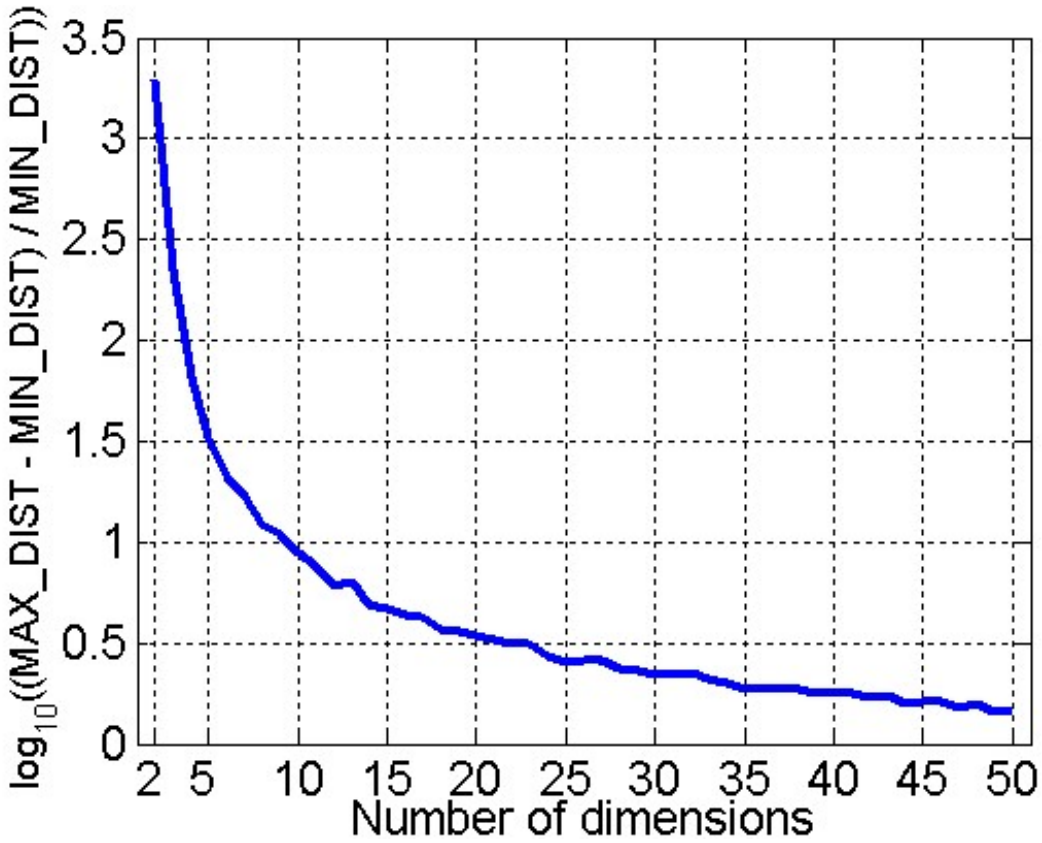
Correlations between time series

	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

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:

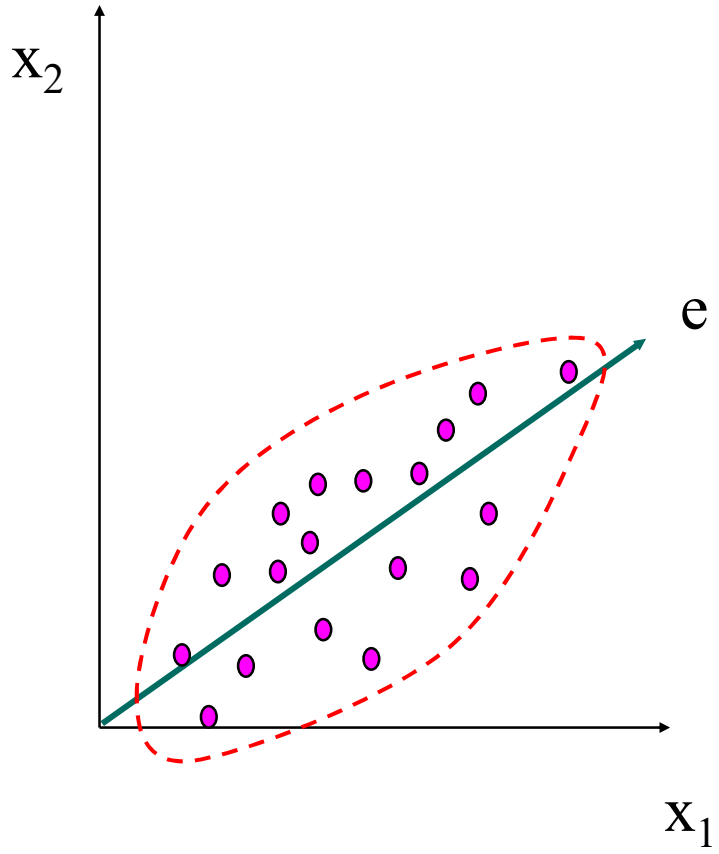
- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise

Techniques

- Principal Components Analysis (PCA)
- Singular Value Decomposition
- Others: supervised and non-linear techniques

Dimensionality Reduction: PCA

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



Dimensionality Reduction: 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

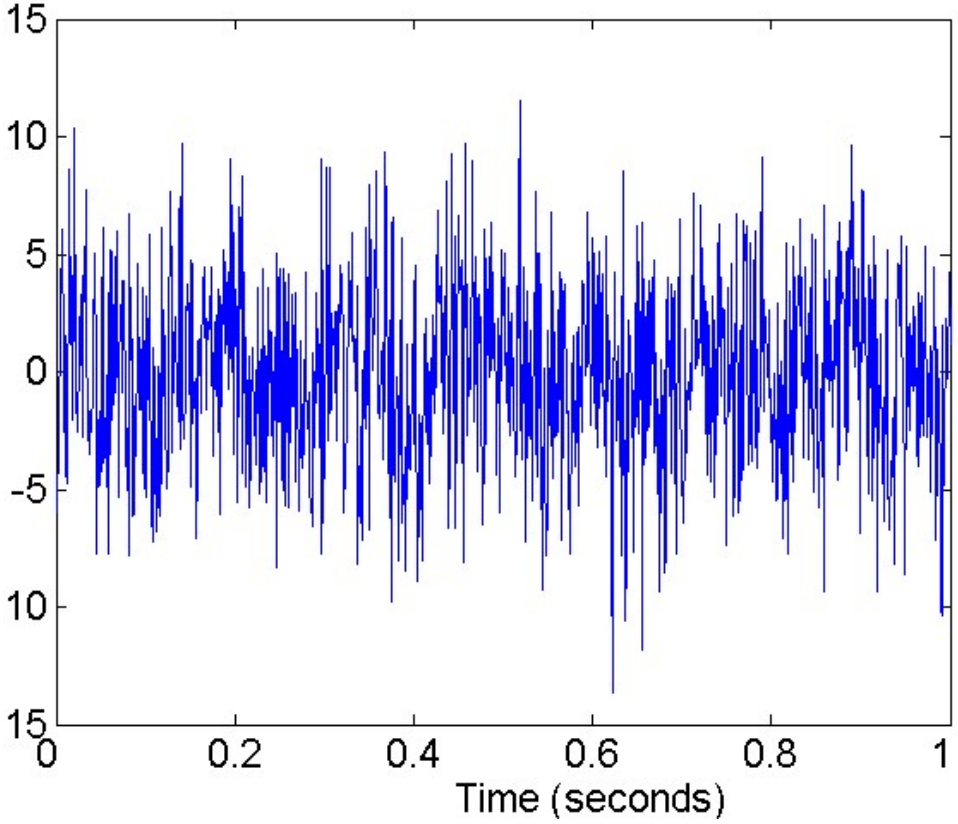
Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

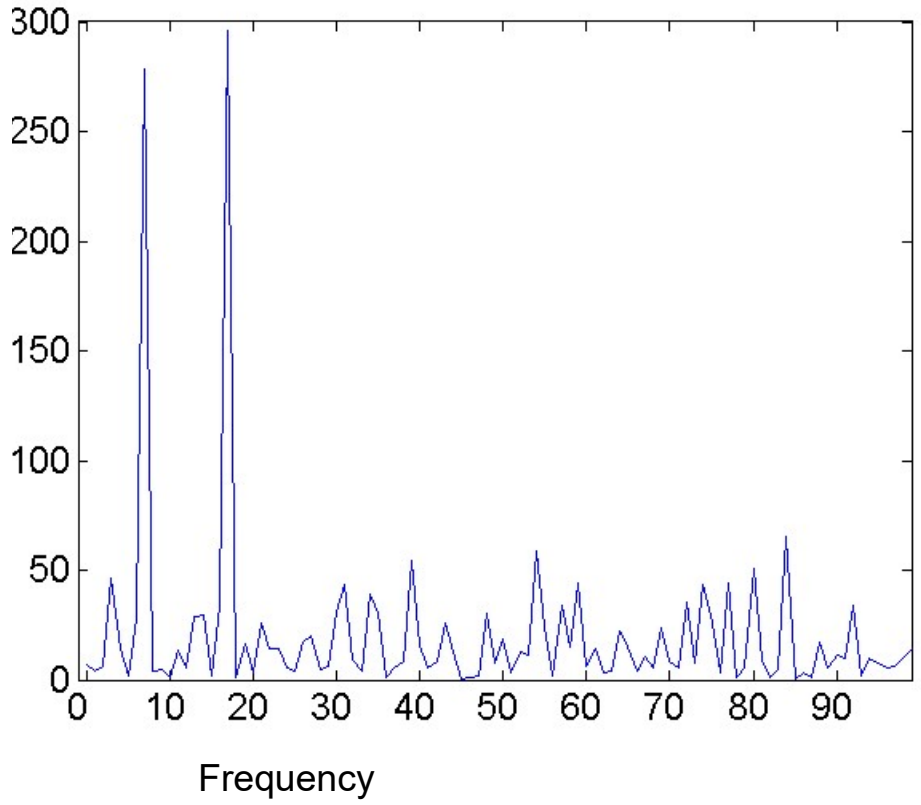
Three general methodologies:

- 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