

Data Mining: Data

Lecture Notes for Chapter 2

Introduction to Data Mining , 2nd Edition
by
Tan, Steinbach, Kumar

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

Table 2.4. 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

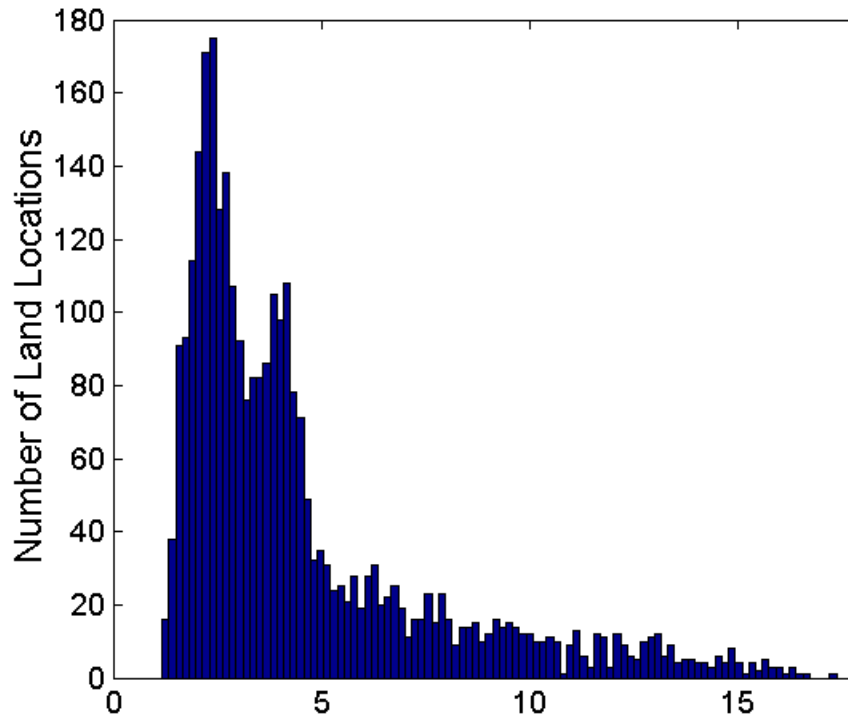
- This example is based on precipitation in Australia from the period 1982 to 1993.

The next slide shows

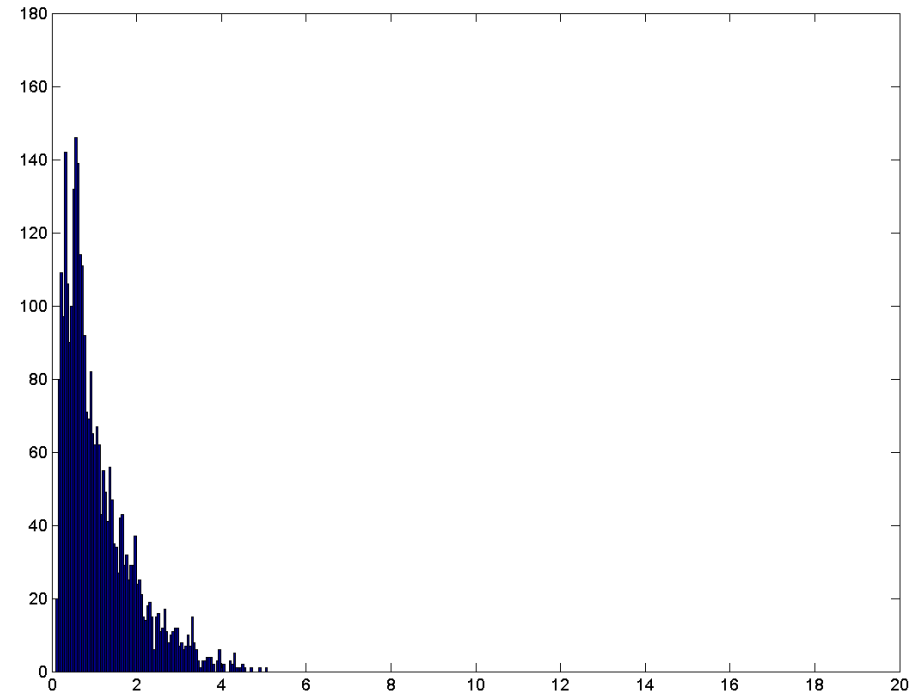
- A histogram for the standard deviation of average monthly precipitation for 3,030 0.5° by 0.5° grid cells in Australia, and
 - A histogram for the standard deviation of the average yearly precipitation for the same locations.
- The average yearly precipitation has less variability than the average 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**

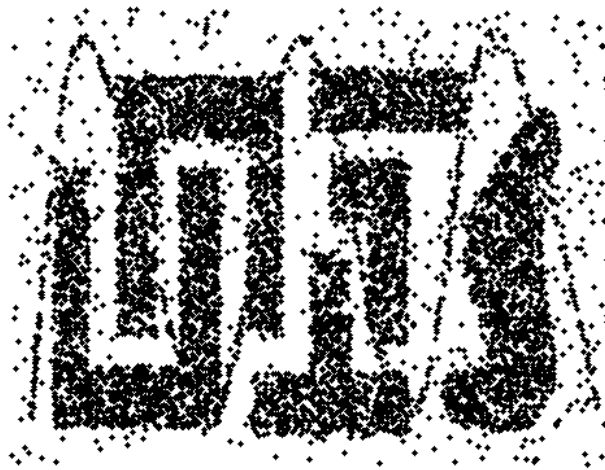
Sampling

- 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.

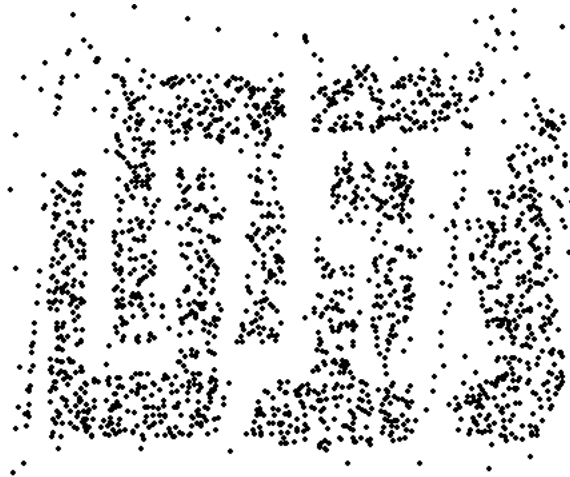
Sampling ...

- 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

Sample Size



8000 points



2000 Points



500 Points

Types of Sampling

□ Simple Random Sampling

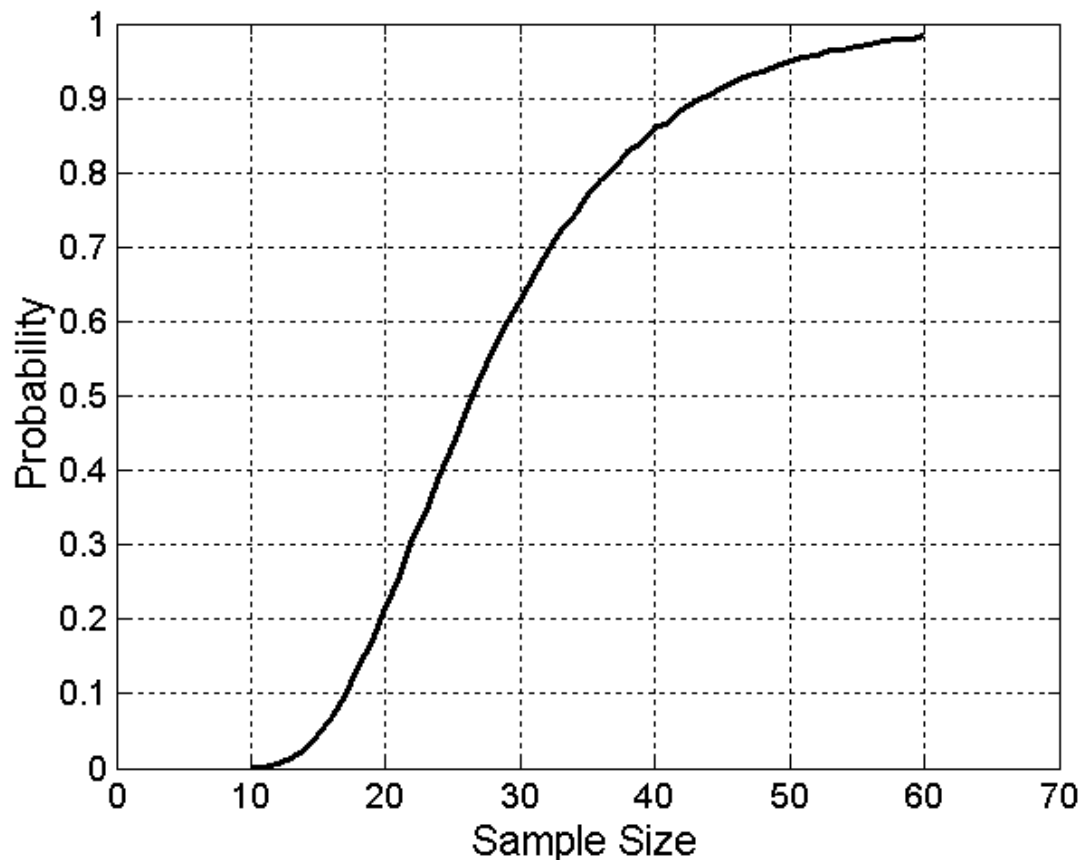
- There is an equal probability of selecting any particular item
- Sampling without replacement
 - ◆ As each item is selected, it is removed from the population
- Sampling with replacement
 - ◆ Objects are not removed from the population as they are selected for the sample.
 - ◆ In sampling with replacement,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

Sample Size

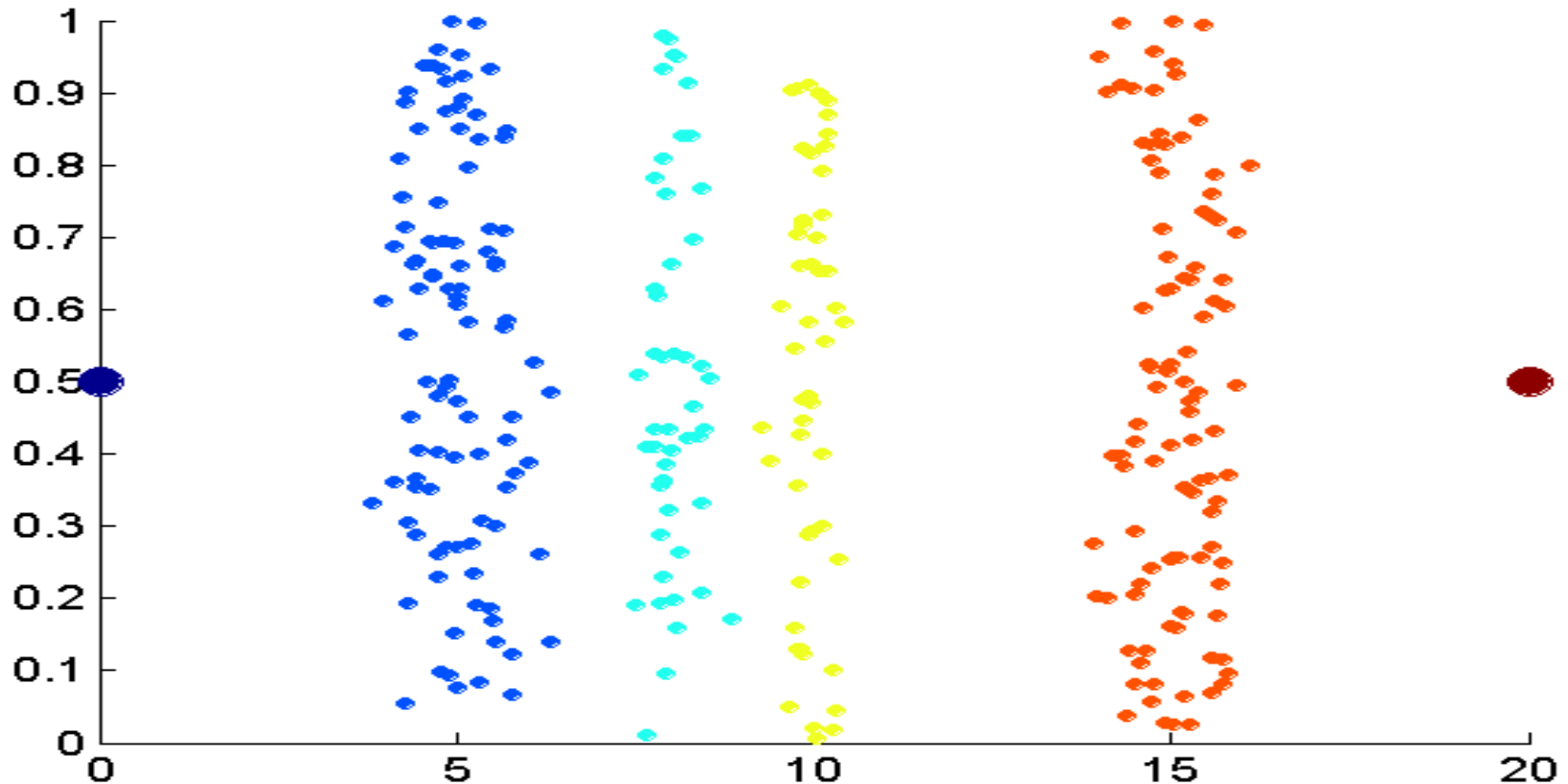
- What sample size is necessary to get at least one object from each of 10 equal-sized groups.



Discretization

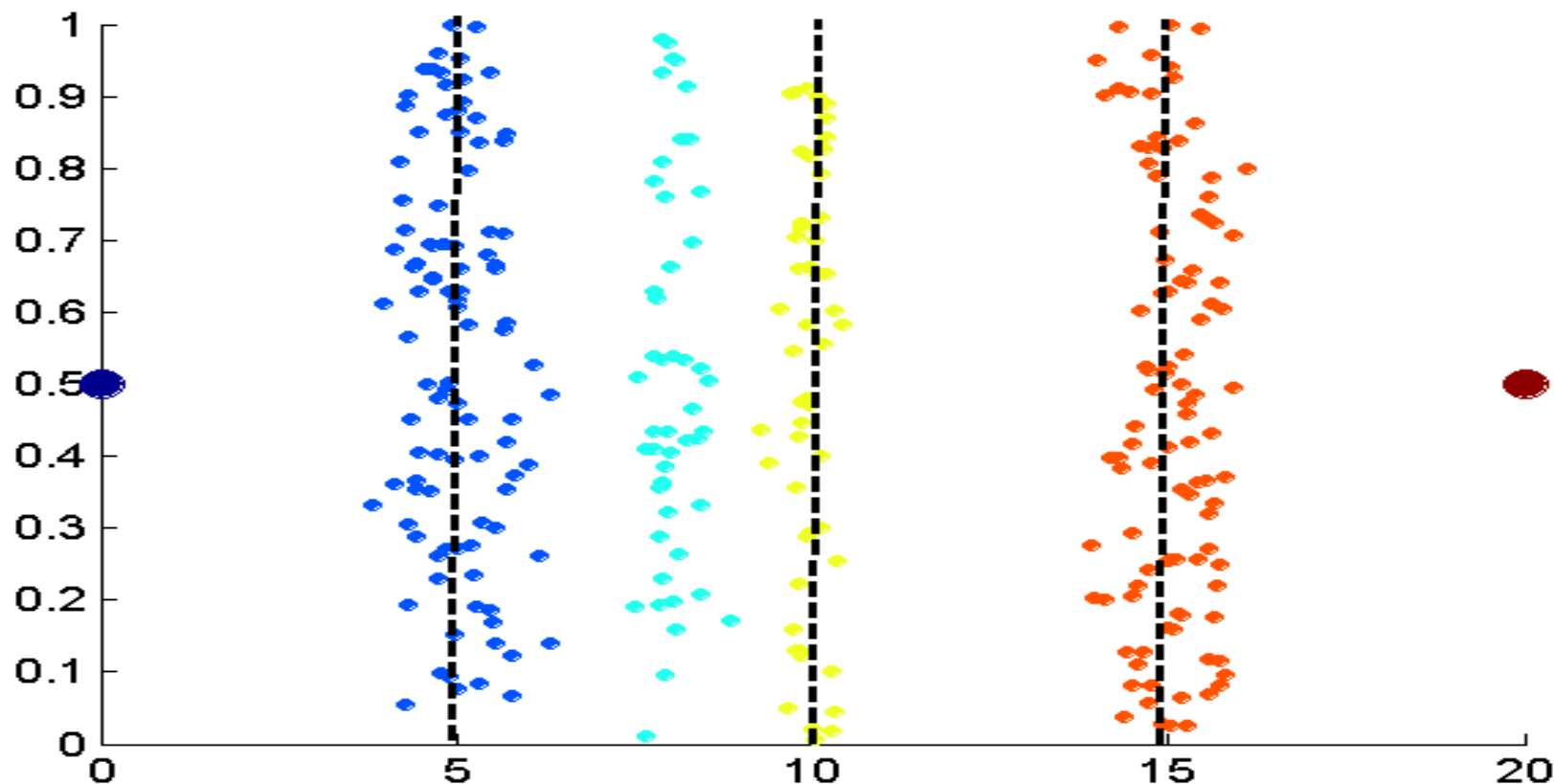
- **Discretization** is the process of converting a continuous attribute into an ordinal attribute
 - A potentially infinite number of values are mapped into a small number of categories
 - Discretization is used in both unsupervised and supervised settings

Unsupervised Discretization



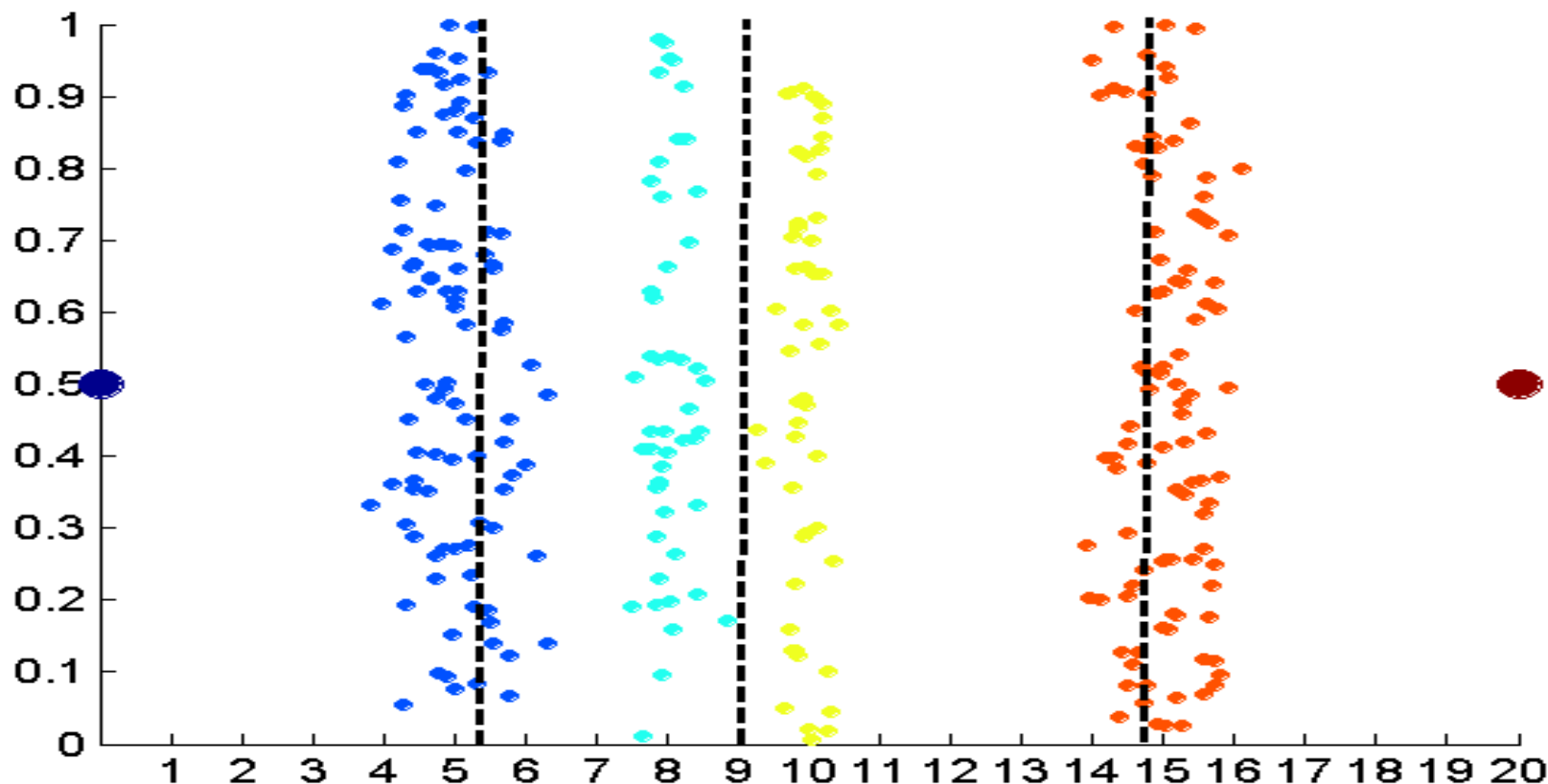
Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

Unsupervised Discretization



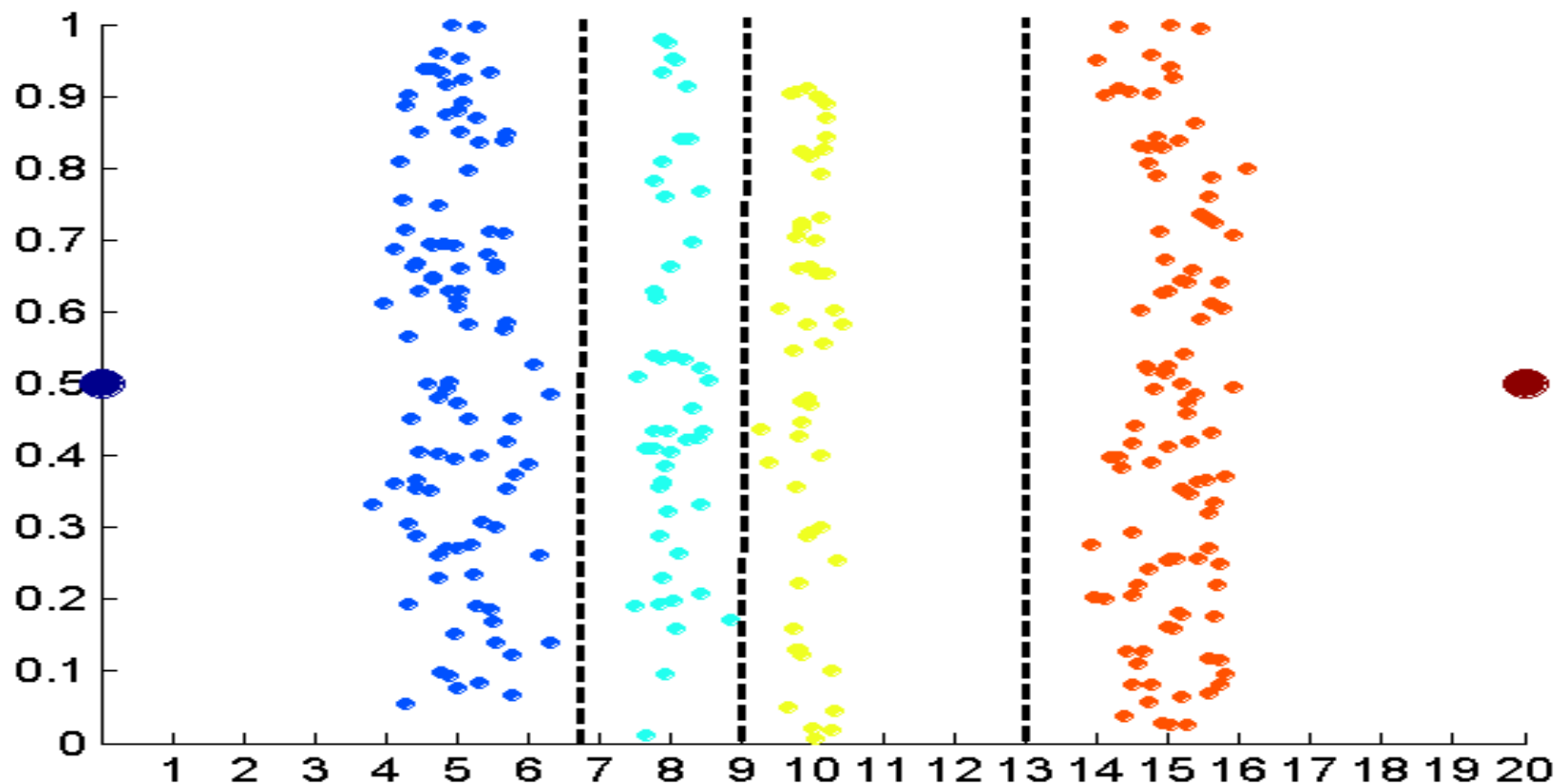
Equal interval width approach used to obtain 4 values.

Unsupervised Discretization



Equal frequency approach used to obtain 4 values.

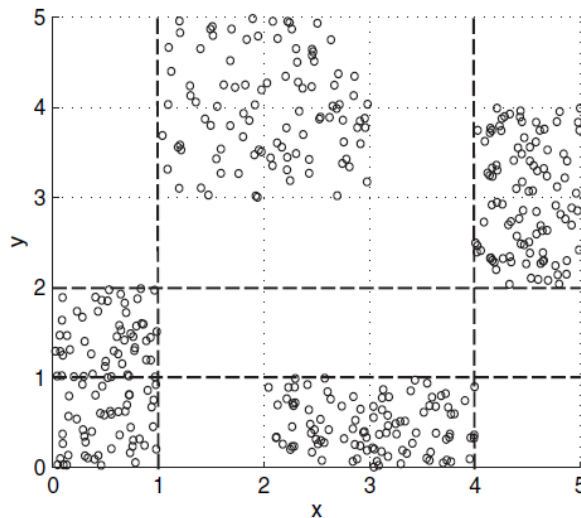
Unsupervised Discretization



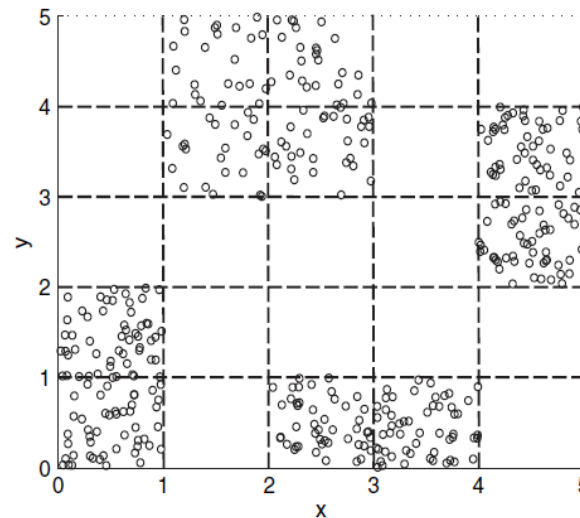
K-means approach to obtain 4 values.

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.



(a) Three intervals



(b) Five intervals

Figure 2.14. 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

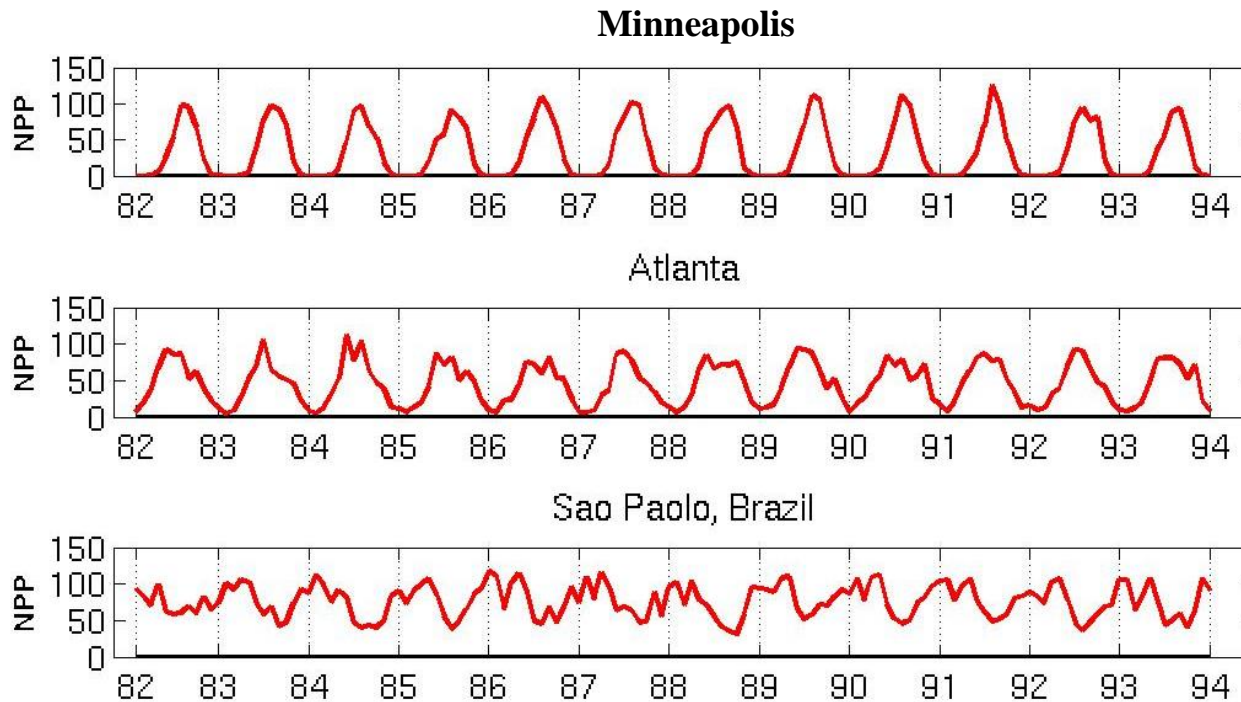
Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
<i>awful</i>	0	1	0	0	0	0
<i>poor</i>	1	0	1	0	0	0
<i>OK</i>	2	0	0	1	0	0
<i>good</i>	3	0	0	0	1	0
<i>great</i>	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

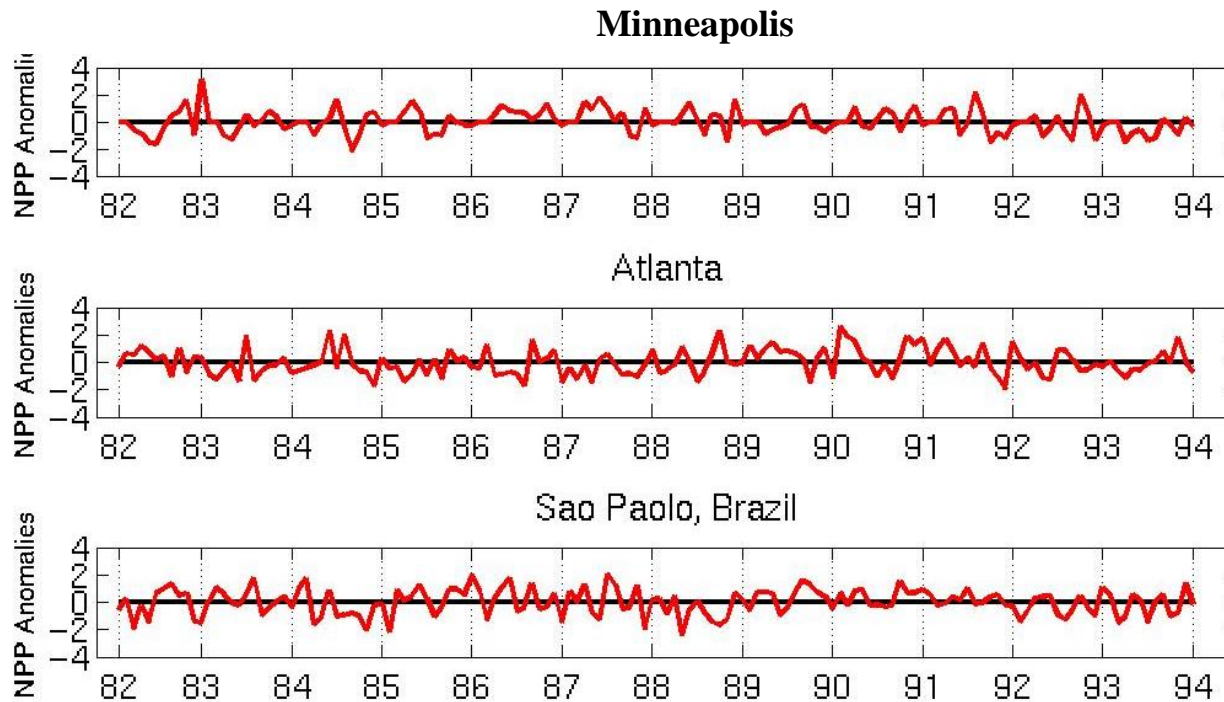


Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.

Correlations between time series

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

Seasonality Accounts for Much Correlation



Normalized using
monthly Z Score:

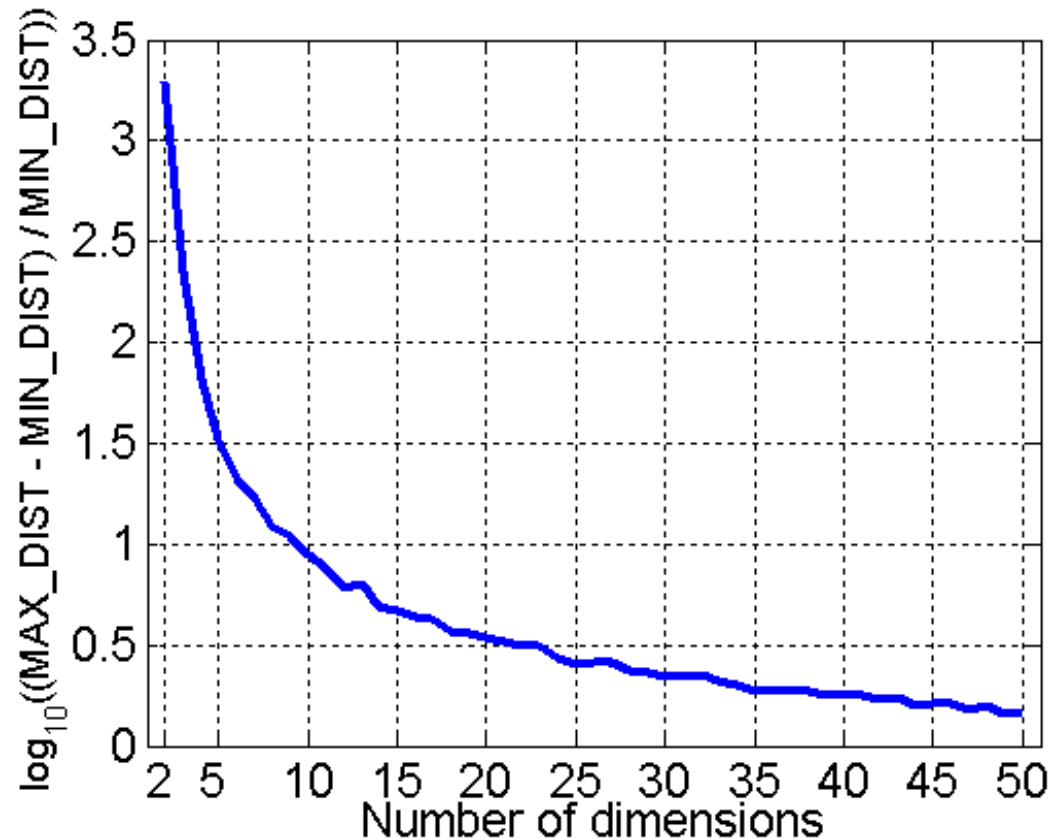
Subtract off monthly
mean and divide by
monthly standard
deviation

Correlations between time series

	Minneapolis	Atlanta	Sao Paulo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paulo	0.0906	-0.0154	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:

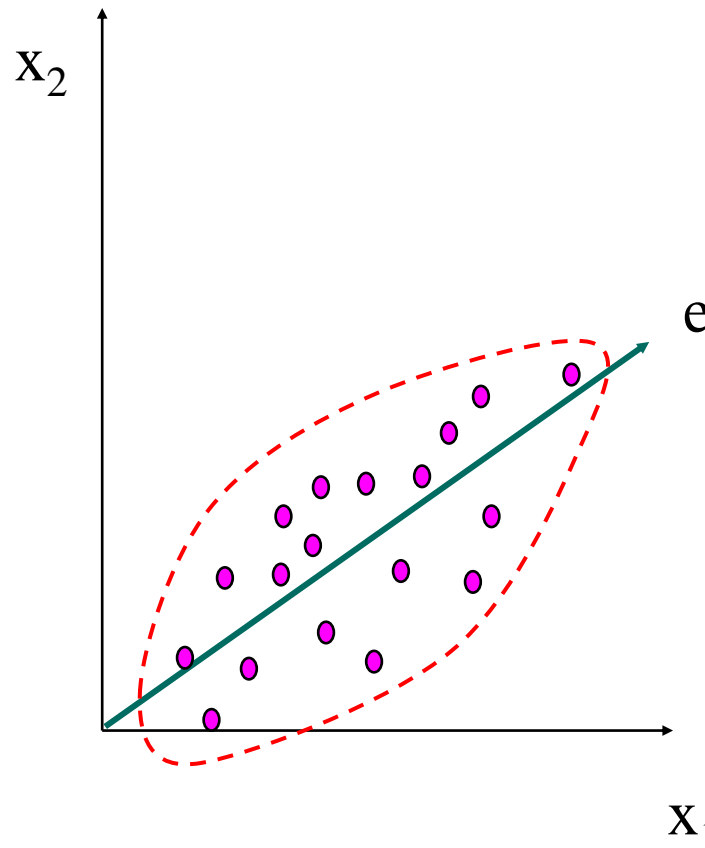
- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise

□ 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

256



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

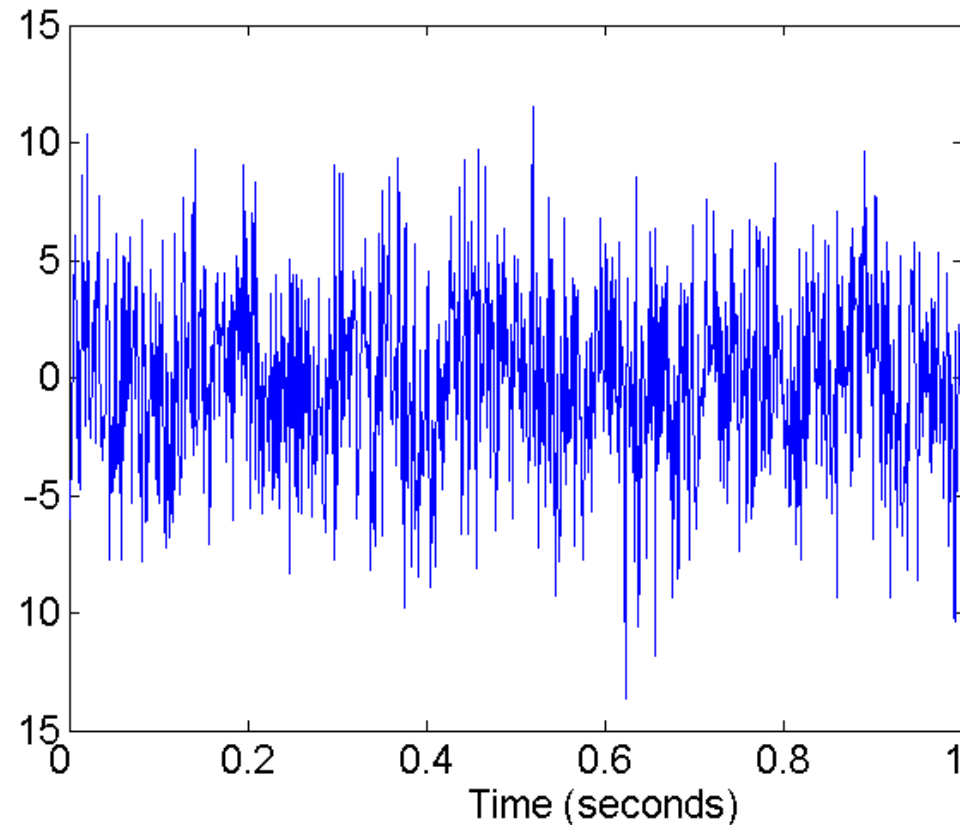
Feature Creation

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

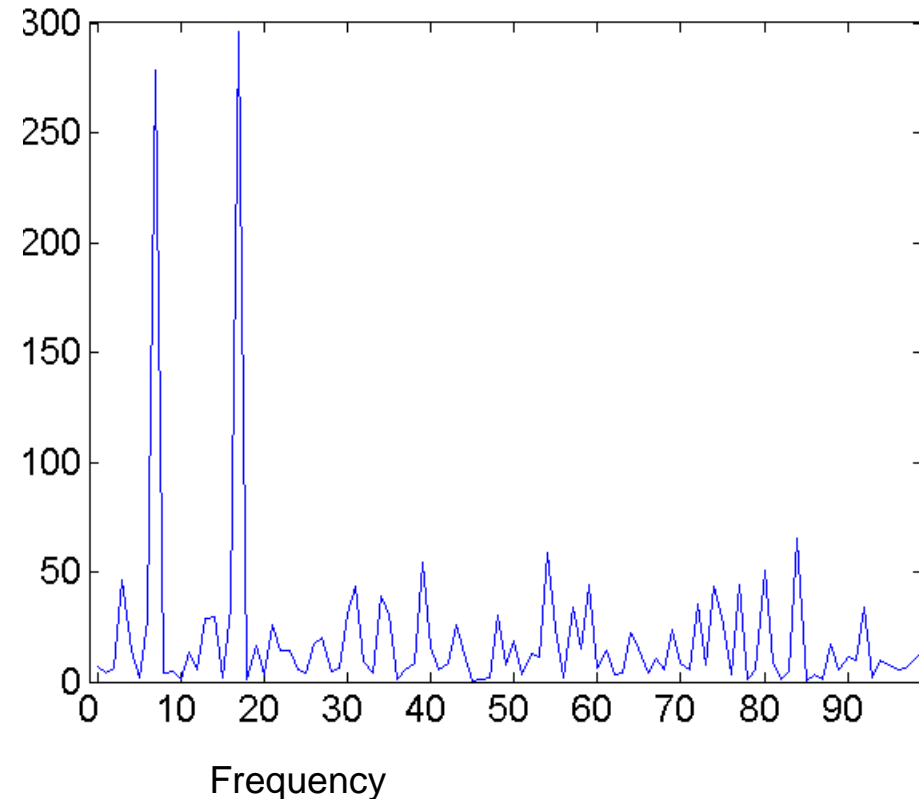
- Three general methodologies:
 - Feature extraction
 - ◆ Example: extracting edges from images
 - Feature construction
 - ◆ Example: dividing mass by volume to get density
 - Mapping data to new space
 - ◆ Example: Fourier and wavelet analysis

Mapping Data to a New Space

□ Fourier and wavelet transform



Two Sine Waves + Noise



Frequency