BBS654
Data Mining

Pinar Duygulu

Slides are adapted from

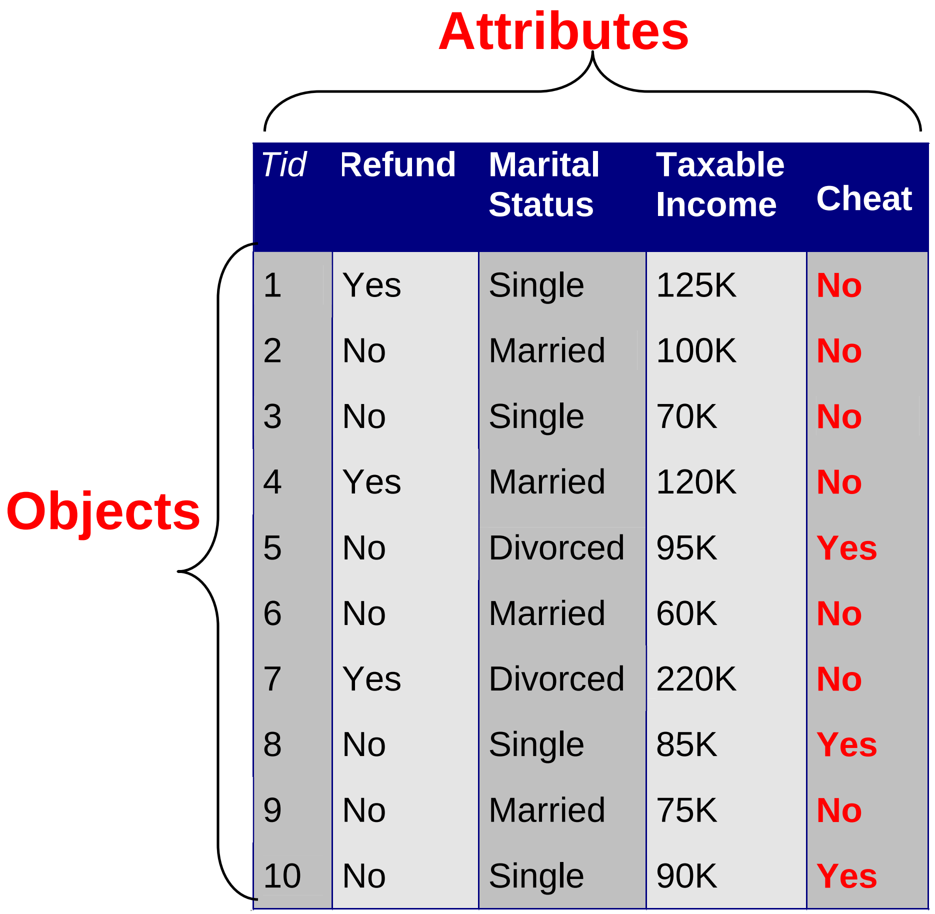
Nazli Ikizler, Sanjay Ranka

Topics

- What is data?
  - Definitions, terminology
  - Types of data and datasets
- Data preprocessing
  - Data Cleaning
  - Data integration
  - Data transformation
  - Data reduction
- Data similarity

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



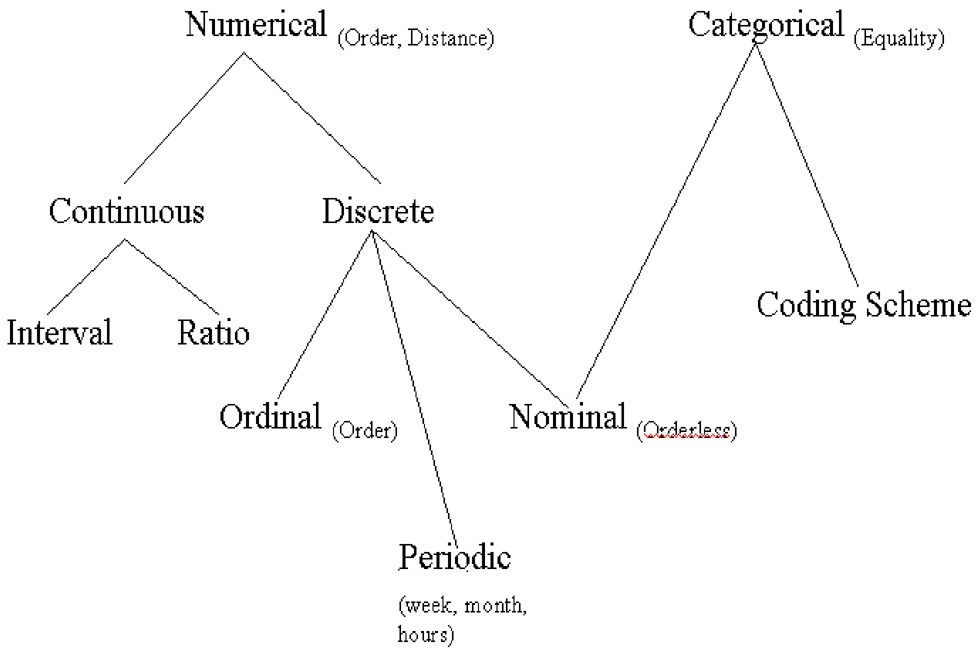
Attributes

Objects

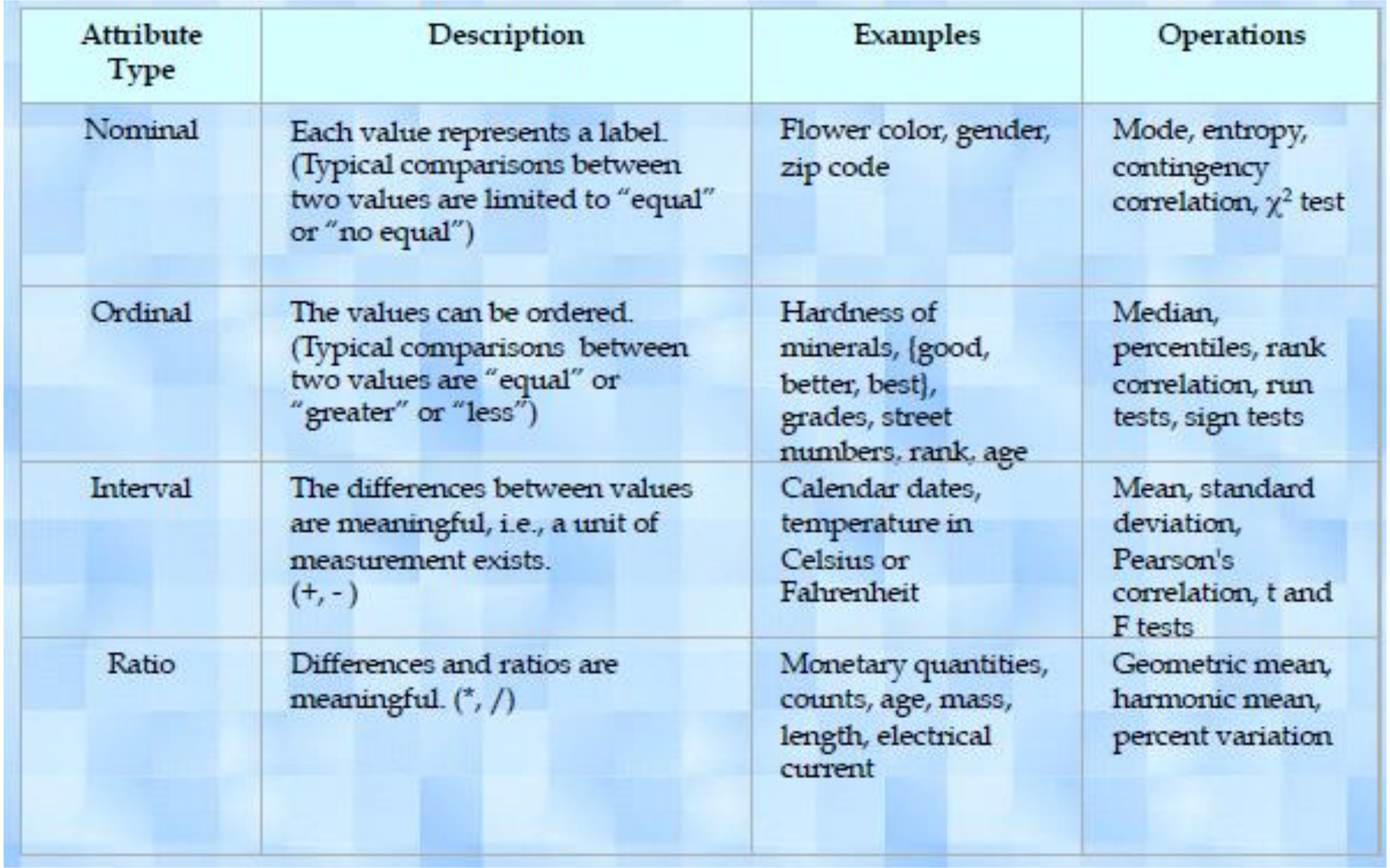
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

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}
- Binary
  - E.g., medical test (positive vs. negative)
- Interval
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- Ratio
  - Examples: temperature in Kelvin, length, time, counts

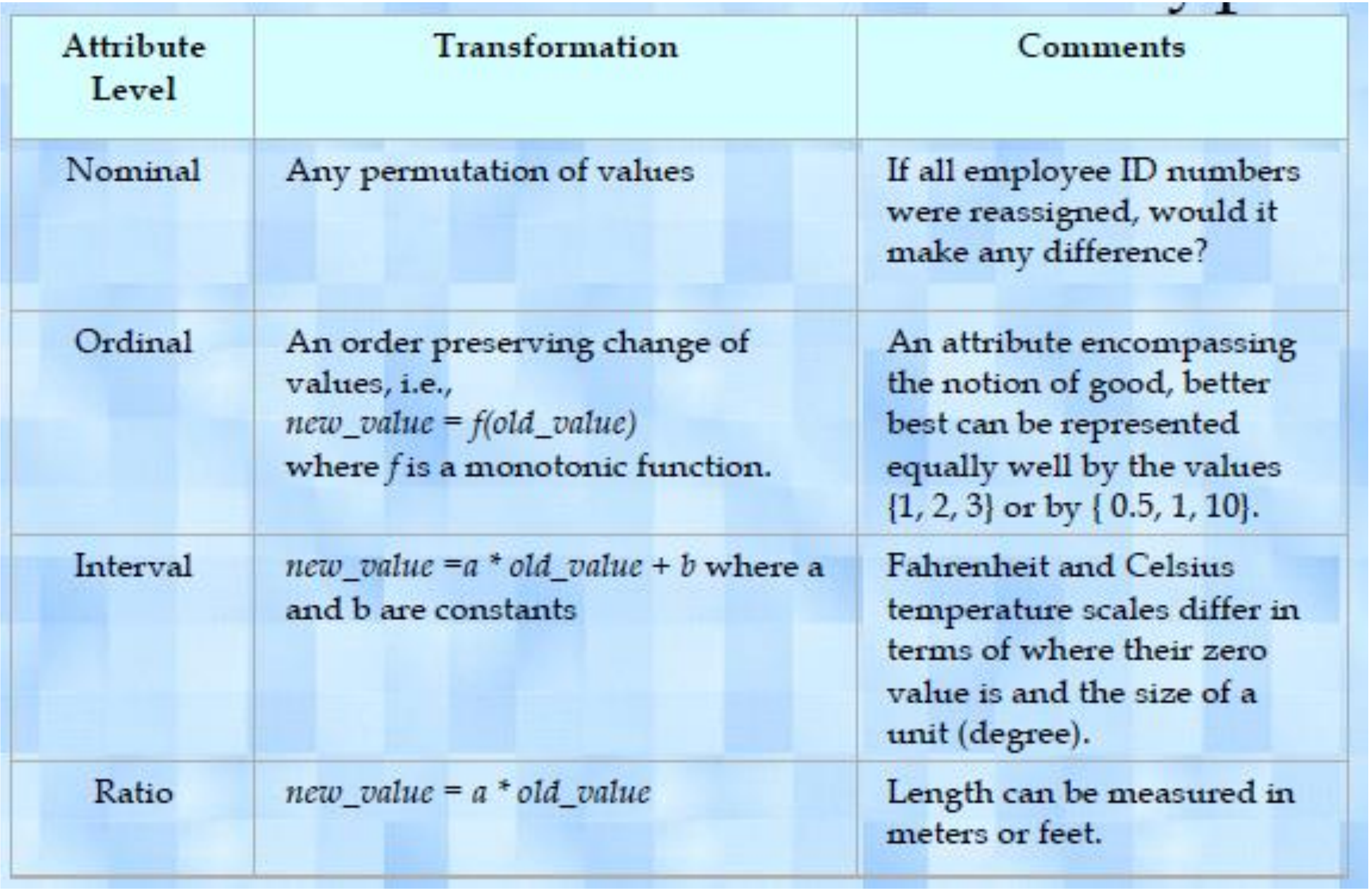


Types of attributes



Attribute Type	Description	Examples	Operations
Nominal	Each value represents a label. (Typical comparisons between two values are limited to "equal" or "no equal")	Flower color, gender, zip code	Mode, entropy, contingency correlation, χ^2 test
Ordinal	The values can be ordered. (Typical comparisons between two values are "equal" or "greater" or "less")	Hardness of minerals, {good, better, best}, grades, street numbers, rank, age	Median, percentiles, rank correlation, run tests, sign tests
Interval	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	Differences and ratios are meaningful. (*, /)	Monetary quantities, counts, age, mass, length, electrical current	Geometric mean, harmonic mean, percent variation

Transformations for different types



Attribute Level	Transformation	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 the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	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 attributes

– A discrete attribute has only a finite or countably infinite set of values.

Eg. Zip codes, counts, or the set of words in a document

– Discrete attributes are often represented as integer variables

Binary attributes are a special case of discrete attributes and assume only two values

– E.g. Yes/no, true/false, male/female

– Binary attributes are often represented as Boolean variables, or as integer variables that take on the values 0 or 1

Continuous attributes

– A continuous attribute has real number values.

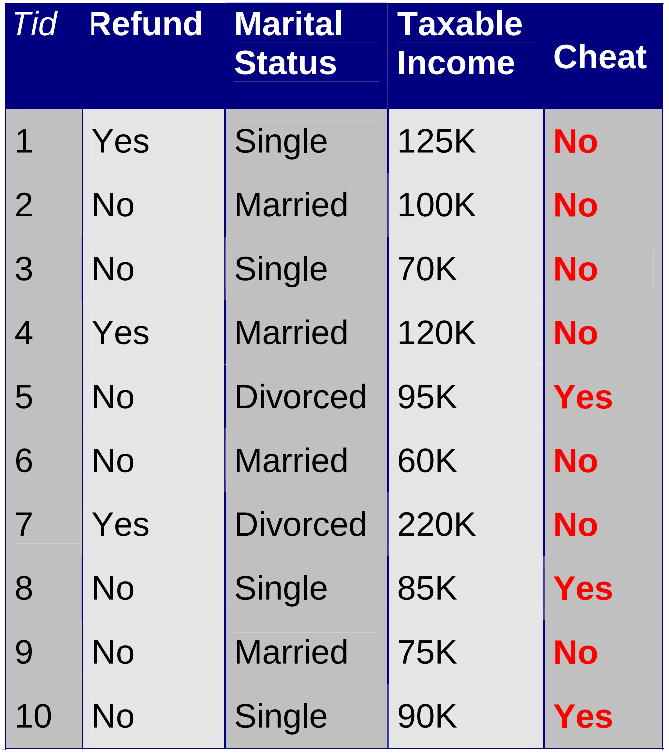
E.g. Temperature, height or weight

(Practically real values can only be measured and represented to a finite number of digits)

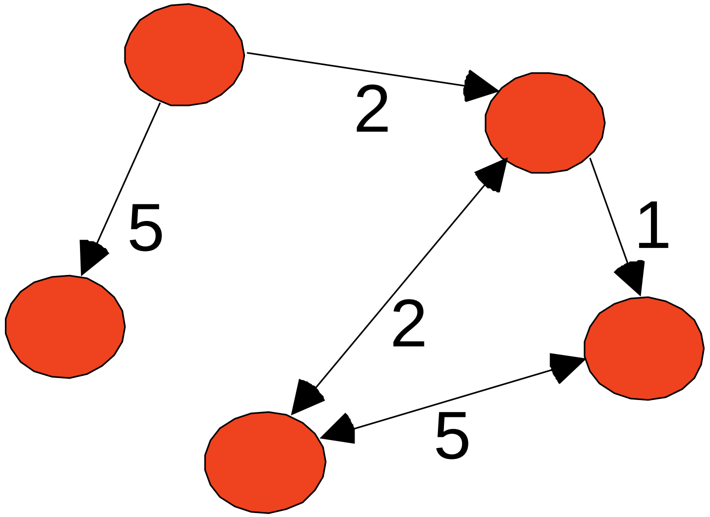
– Continuous attributes are typically represented as floating point variables

Types of data

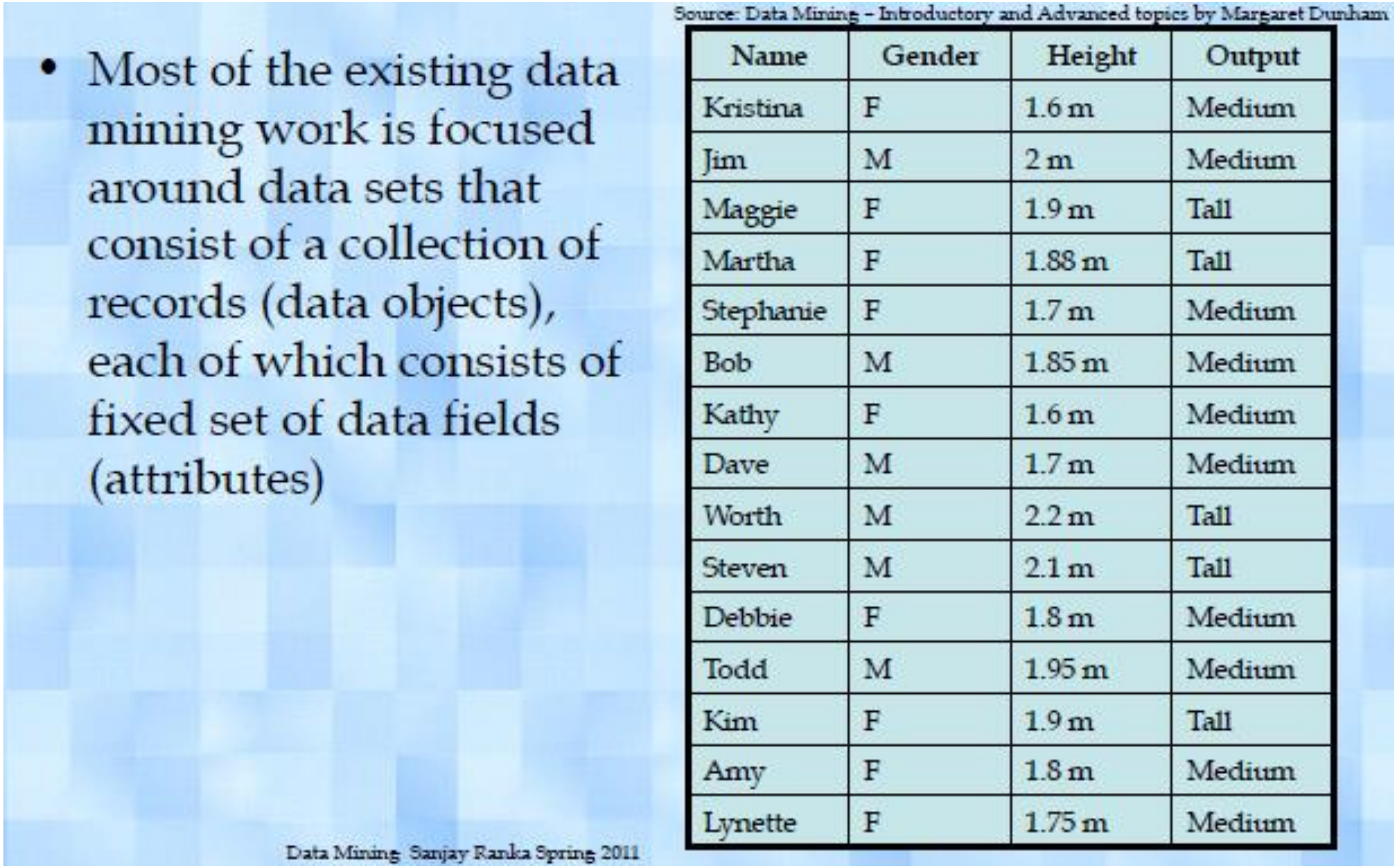
- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data



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**



Record data



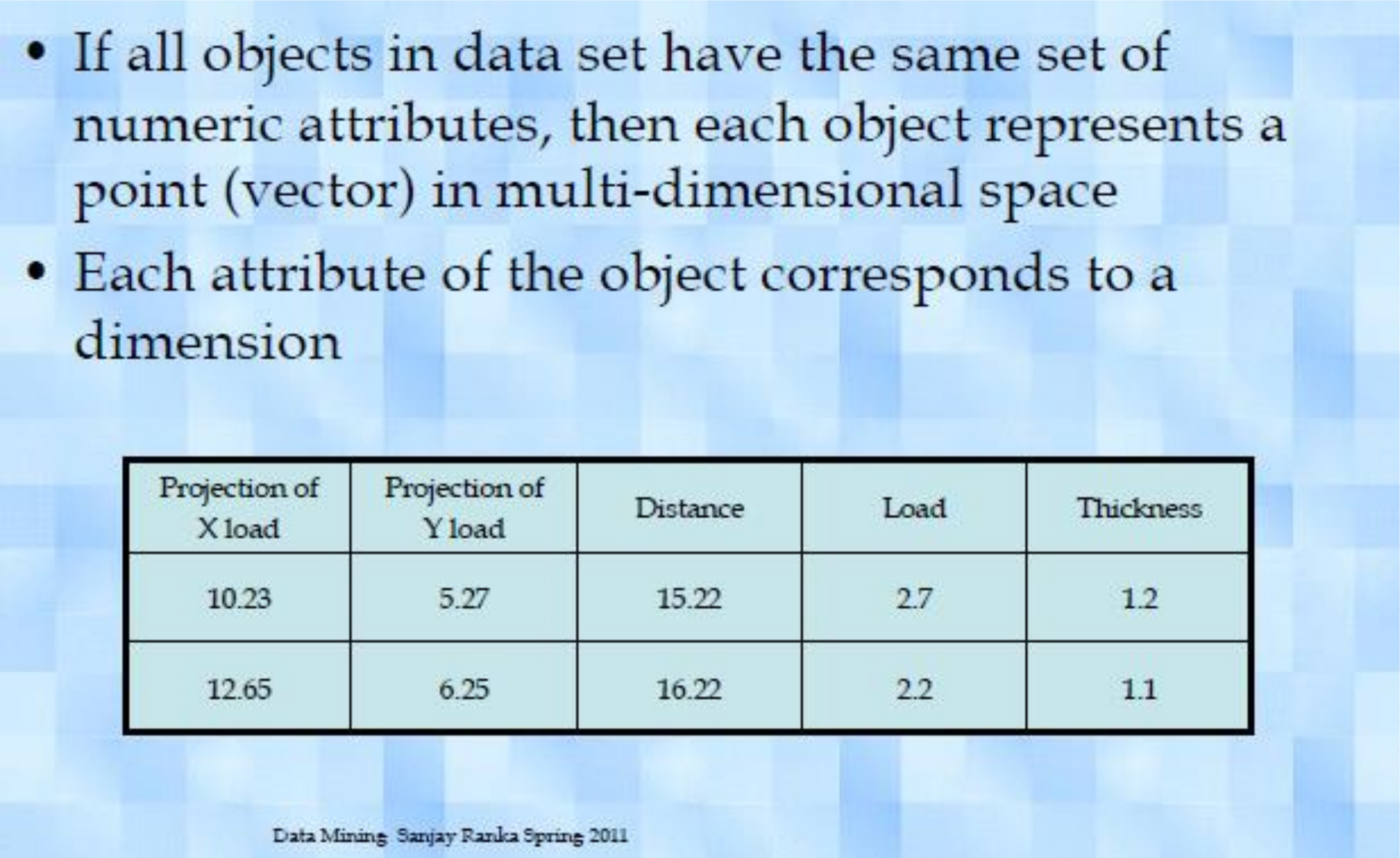
- Most of the existing data mining work is focused around data sets that consist of a collection of records (data objects), each of which consists of fixed set of data fields (attributes)

Source: Data Mining – Introductory and Advanced topics by Margaret Dunham

Name	Gender	Height	Output
Kristina	F	1.6 m	Medium
Jim	M	2 m	Medium
Maggie	F	1.9 m	Tall
Martha	F	1.88 m	Tall
Stephanie	F	1.7 m	Medium
Bob	M	1.85 m	Medium
Kathy	F	1.6 m	Medium
Dave	M	1.7 m	Medium
Worth	M	2.2 m	Tall
Steven	M	2.1 m	Tall
Debbie	F	1.8 m	Medium
Todd	M	1.95 m	Medium
Kim	F	1.9 m	Tall
Amy	F	1.8 m	Medium
Lynette	F	1.75 m	Medium

Data Mining Sanjay Ranka Spring 2011

Data matrix

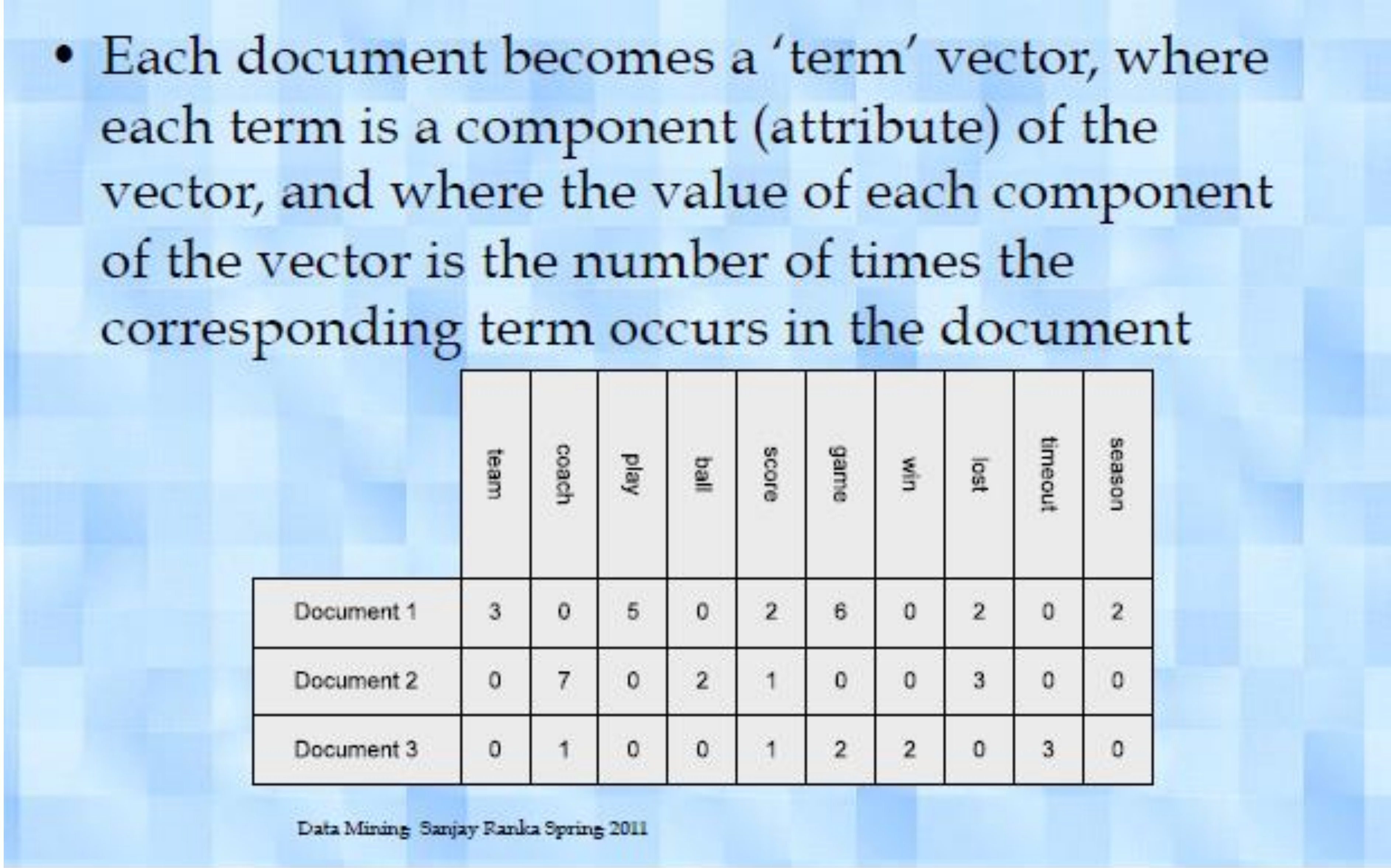


- If all objects in data set have the same set of numeric attributes, then each object represents a point (vector) in multi-dimensional space
- Each attribute of the object corresponds to a dimension

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

Data Mining Sanjay Ranka Spring 2011

Document data

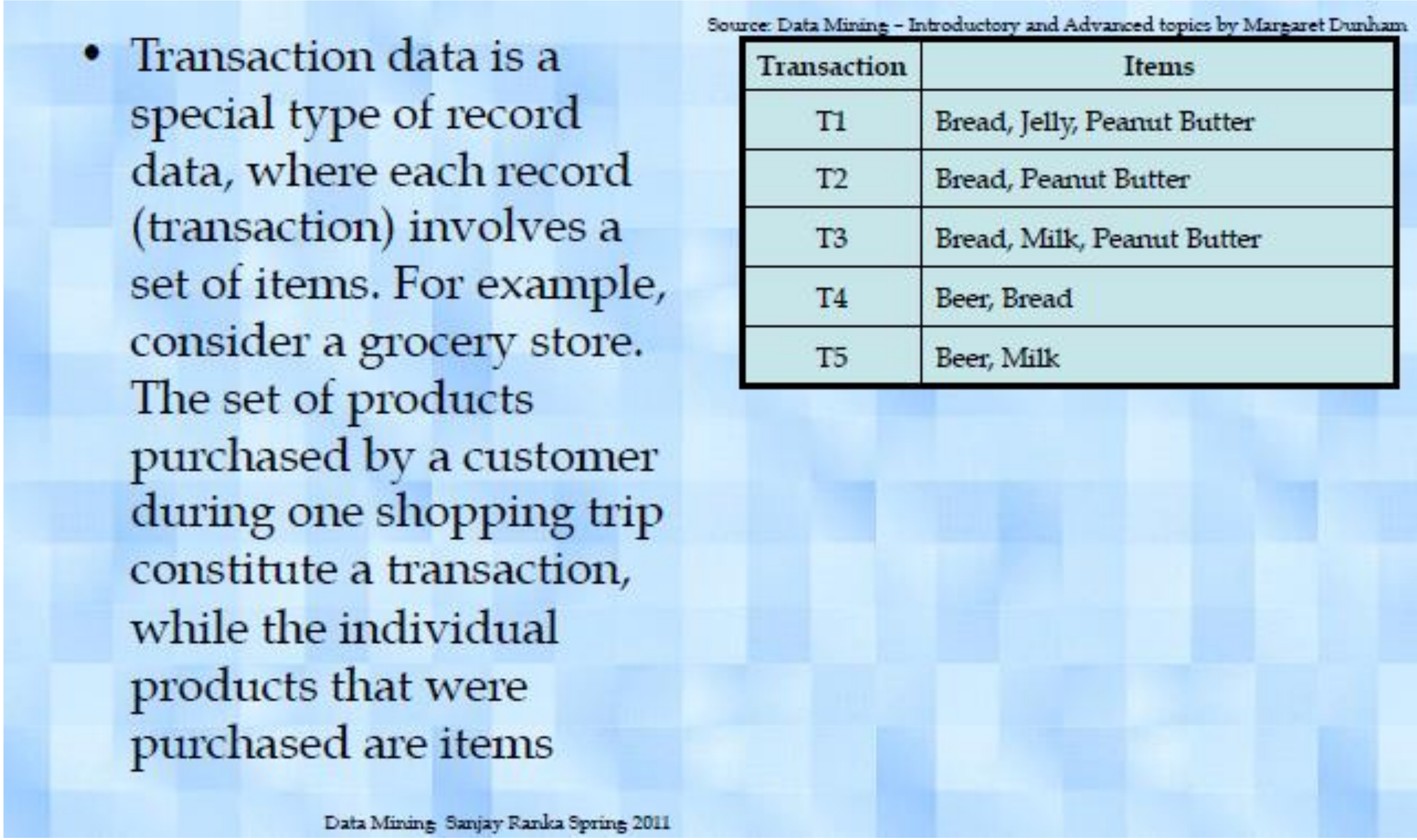


- Each document becomes a 'term' vector, where each term is a component (attribute) of the vector, and where the value of each component of the vector is the number of times the corresponding term occurs in the document

	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

Data Mining Sanjay Ranka Spring 2011

Transaction Data



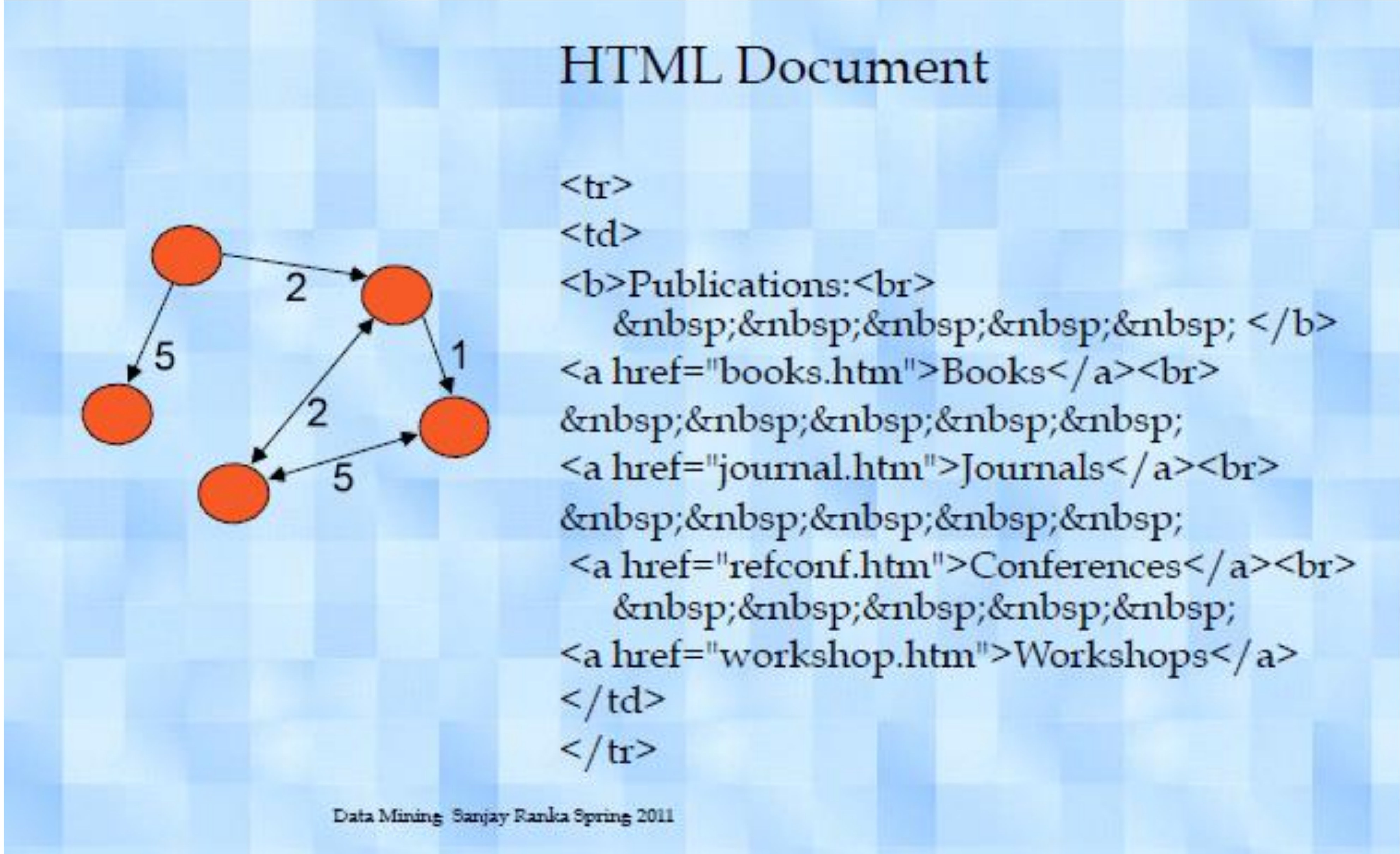
Source: Data Mining – Introductory and Advanced topics by Margaret Dunham

- Transaction data is 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 items

Transaction	Items
T1	Bread, Jelly, Peanut Butter
T2	Bread, Peanut Butter
T3	Bread, Milk, Peanut Butter
T4	Beer, Bread
T5	Beer, Milk

Data Mining Sanjay Ranka Spring 2011

Graph data



HTML Document

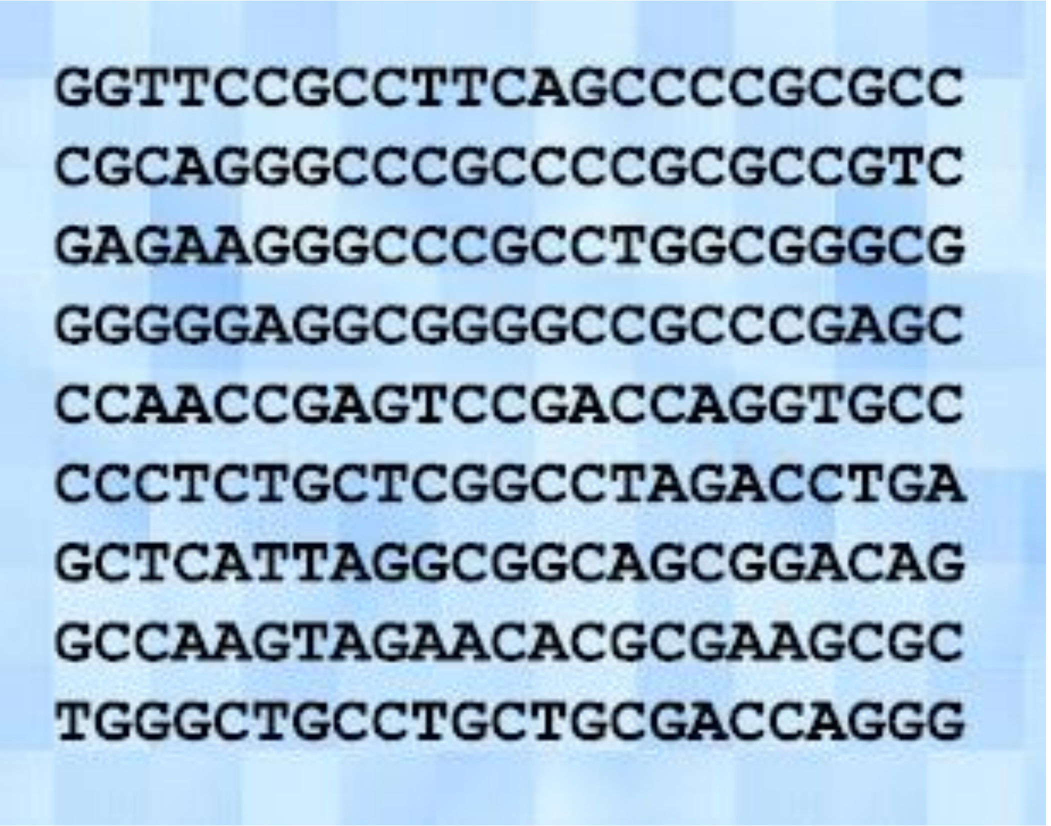
```
<tr>
<td>
<b>Publications:<br>
    &nbsp;&nbsp;&nbsp;&nbsp;&nbsp; </b>
<a href="books.htm">Books</a><br>
&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;
<a href="journal.htm">Journals</a><br>
&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;
 <a href="refconf.htm">Conferences</a><br>
    &nbsp;&nbsp;&nbsp;&nbsp;&nbsp;
<a href="workshop.htm">Workshops</a>
</td>
</tr>
```

Data Mining Sanjay Ranka Spring 2011

Ordered Data: Transaction Data

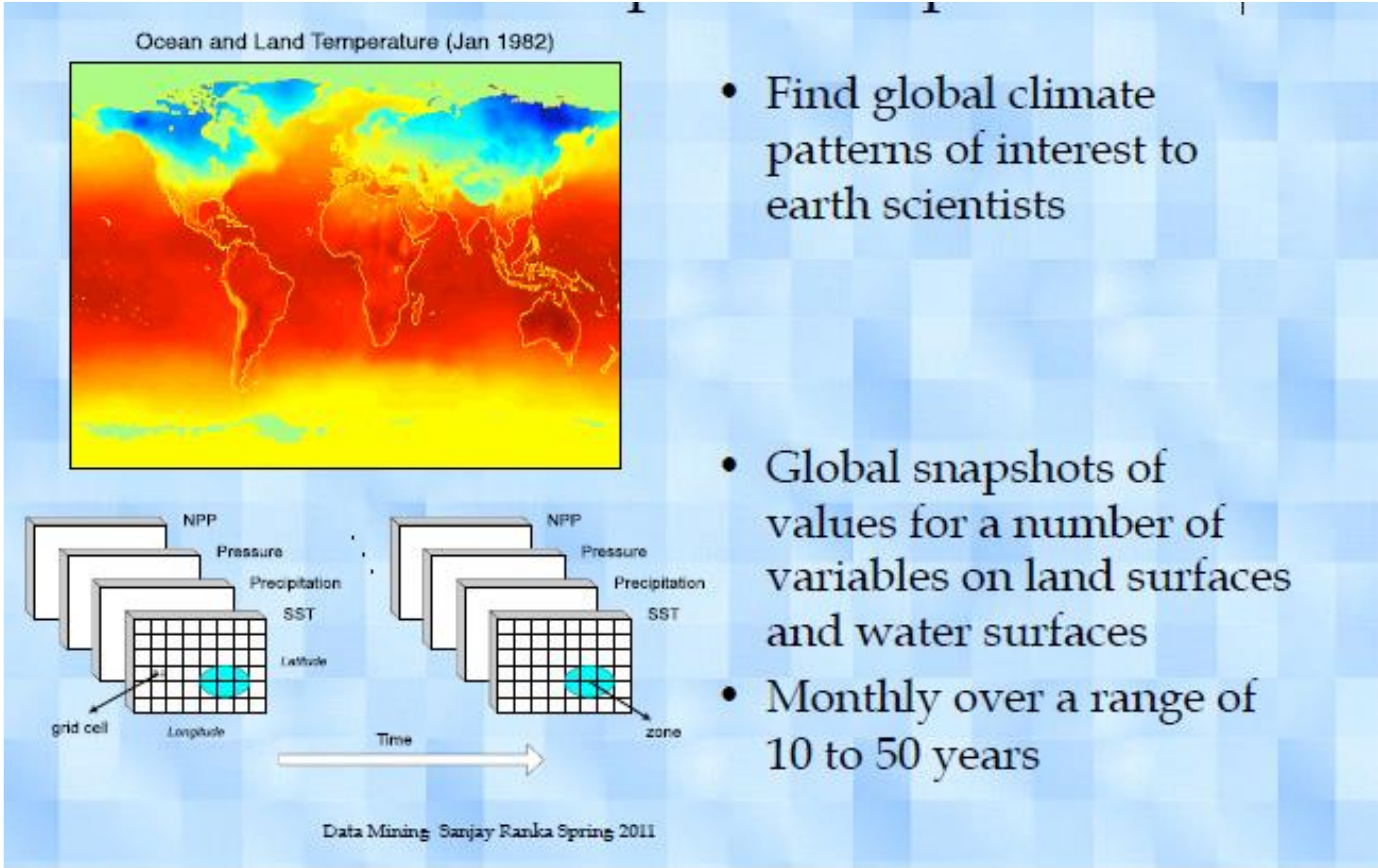
- (AB) (D) (CE)
- (BD) (C) (E)
- (CD) (B) (AE)

Ordered Data: Genomic Sequence Data



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

Ordered Data: Spatio-temporal Data



- Find global climate patterns of interest to earth scientists

- Global snapshots of values for a number of variables on land surfaces and water surfaces
- Monthly over a range of 10 to 50 years

Data Mining Sanjay Ranka Spring 2011

Data Quality

The following are some well known issues

– Noise and outliers

– Missing values

– Duplicate data

– Inconsistent values

Noise and Outliers

Noise –Modification of original value

– random

– non-random (artifact of measurement)

Noise can be

– temporal

– spatial

Signal processing can reduce (generally not eliminate) noise

Outliers

Small number of points

with characteristics different from rest of the data

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Data Cleaning:
Data in the Real World Is Dirty

- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=“ ” (missing data)
- noisy: containing noise, errors, or outliers
  - e.g., Salary=“−10” (an error)
- inconsistent: containing discrepancies in codes or names, e.g.,
  - Age=“42” Birthday=“03/07/2010”
  - Was rating “1,2,3”, now rating “A, B, C”
  - discrepancy between duplicate records
- Intentional (e.g., *disguised missing* data)
  - Jan. 1 as everyone’s birthday?

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Missing Values:

- A simple and effective strategy is to eliminate those records which have missing values. A related strategy is to eliminate attribute that have missing values
- • Drawback: you may end up removing a large number of objects

Missing Values: Estimating Them

- Price of the IBM stock changes in a reasonably smooth fashion. The missing values can be estimated by Interpolation

- For a data set that has many similar data points, a nearest neighbor approach can be used to estimate the missing value. If the attribute is continuous, then the average attribute value of the nearest neighbors can be used. While if the attribute is categorical, then the most commonly occurring attribute value can be taken

Missing Values: Using the Missing Value As Another Value

- Many data mining approaches can be modified to operate by ignoring missing values
- E.g. Clustering - Similarity between pairs of data objects needs to be calculated. If one or both objects of a pair have missing values for some attributes, then the similarity can be calculated by using only the other attributes.

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Simple Discretization Methods: Binning

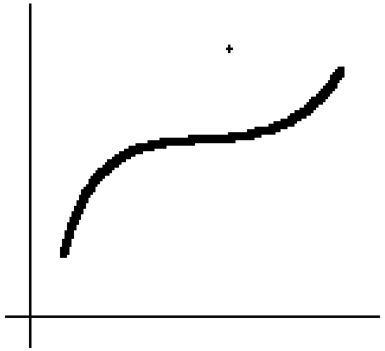
- Equal-width (distance) partitioning:
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B –A)/N.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well.
- Equal-depth (frequency) partitioning:
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky.

Binning Methods for Data Smoothing

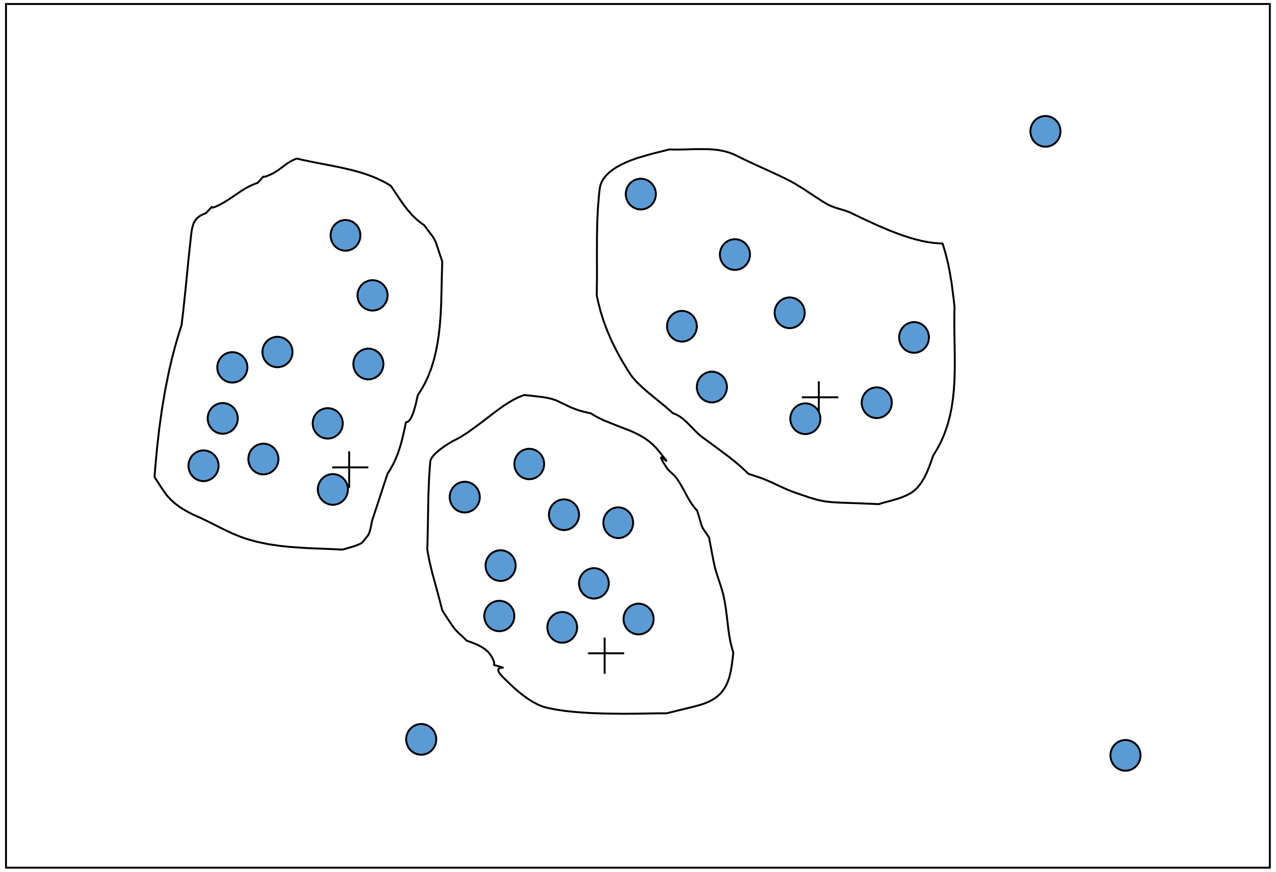
- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Outlier Removal

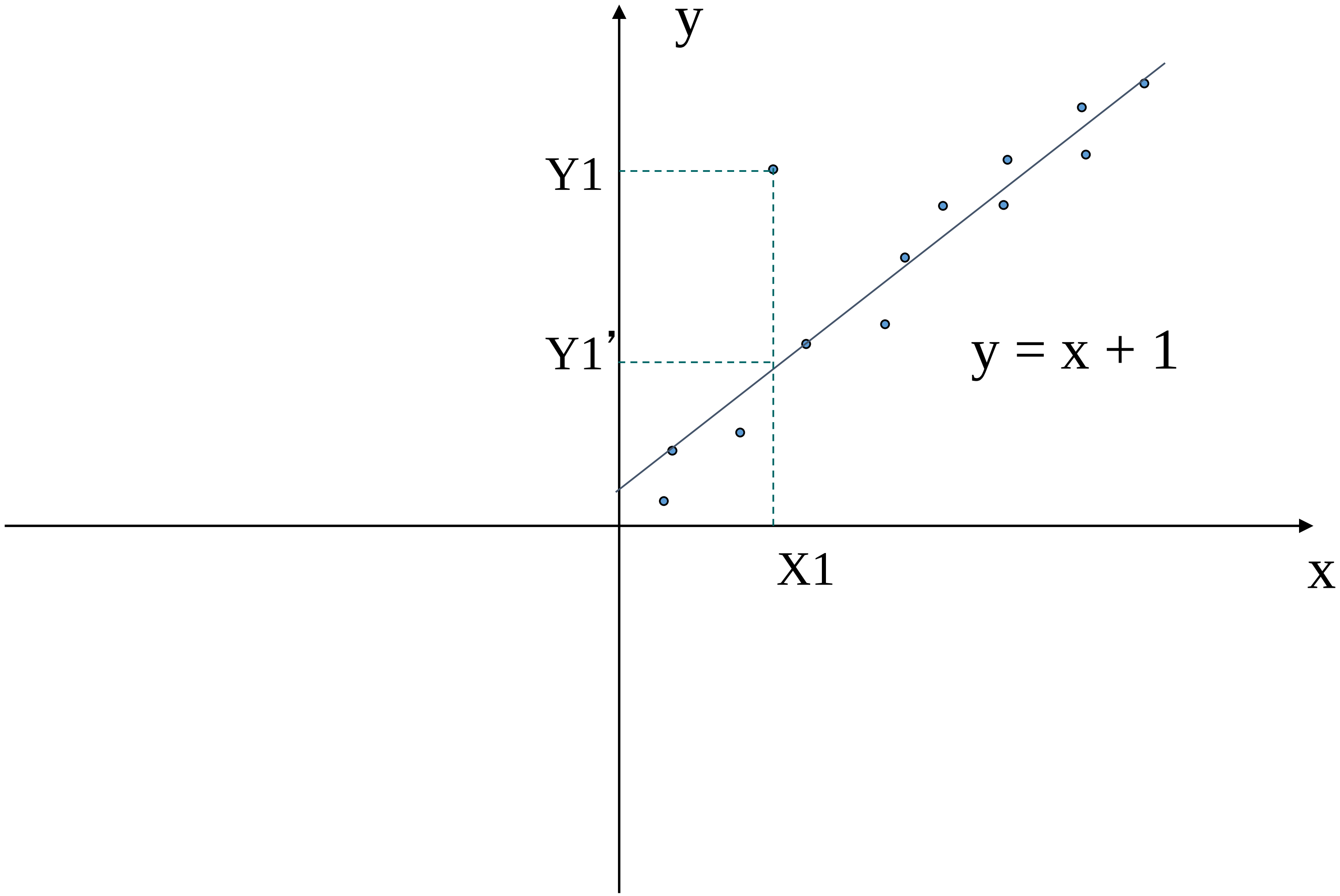
- Data points inconsistent with the majority of data
- Different outliers
  - Valid: CEO’s salary,
  - Noisy: One’s age = 200, widely deviated points
- Removal methods
  - Clustering
  - Curve-fitting
  - Hypothesis-testing with a given model



Cluster Analysis



Regression



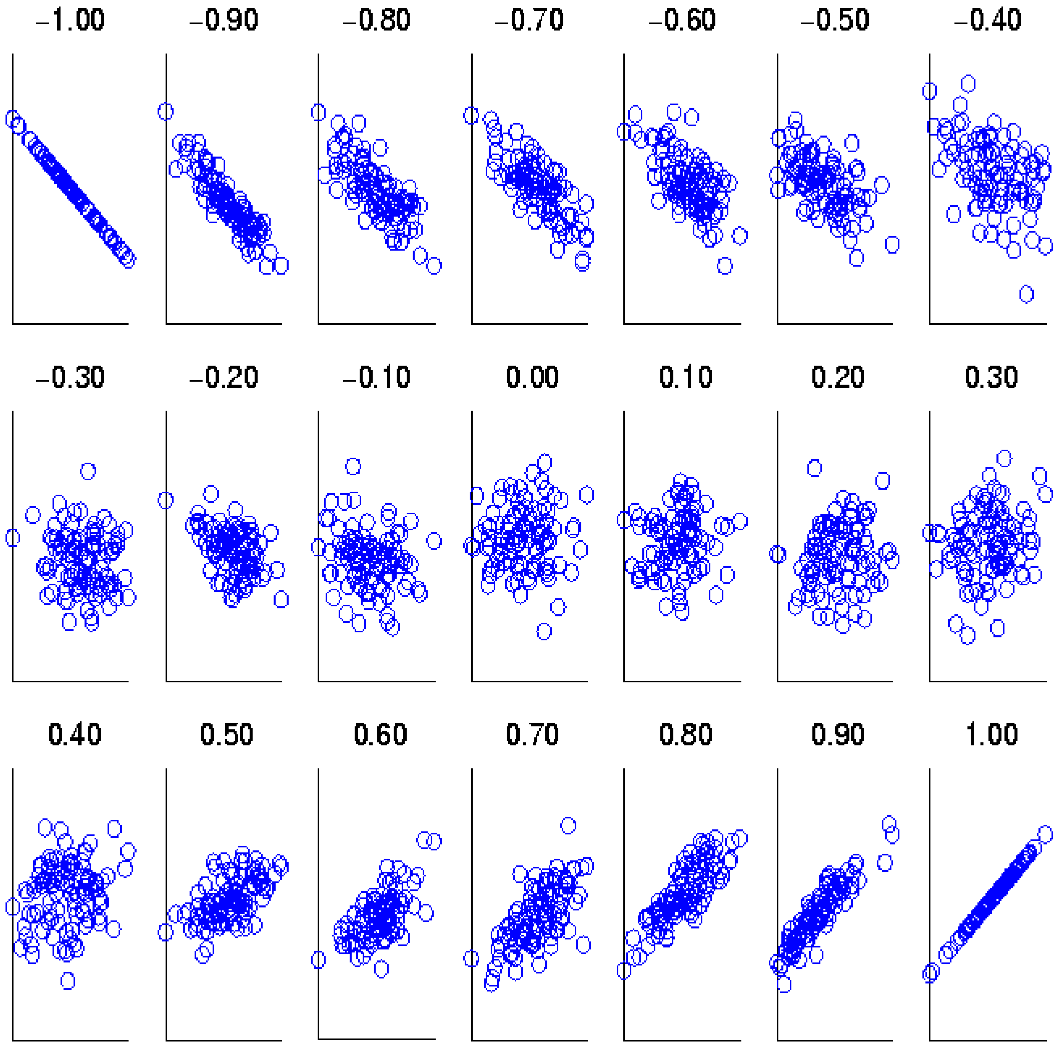
Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Data Transformation

- Methods
  - Aggregation:
    - Summarization
  - Sampling
    - Attribute selection
  - Normalization:
    - Scaled to fall within a small, specified range
  - Attribute/feature construction
    - New attributes constructed from the given ones

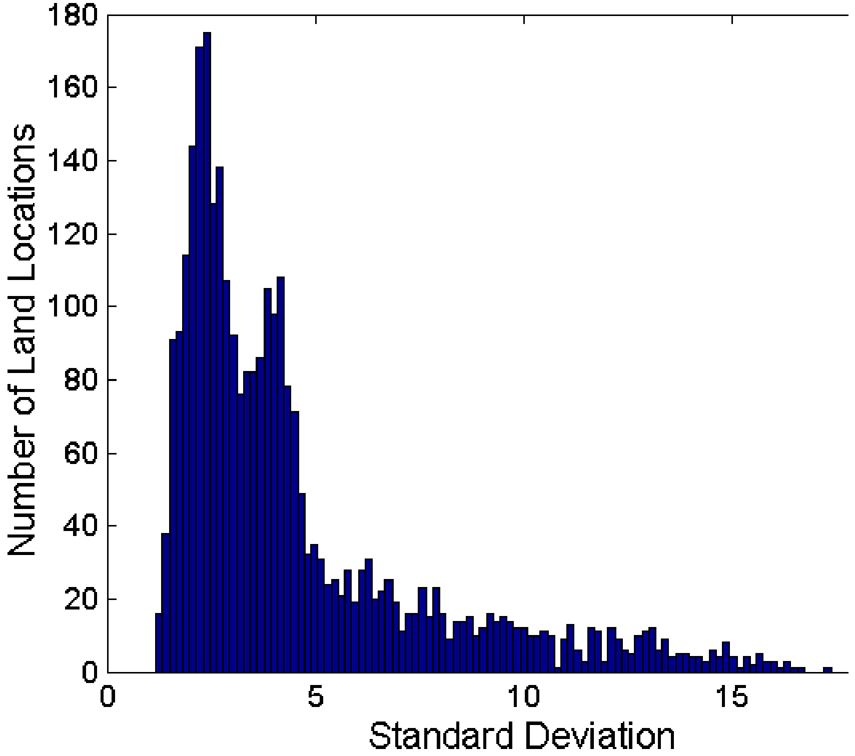
Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)
- For example, merging daily sales figures to obtain monthly sales figures
- 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

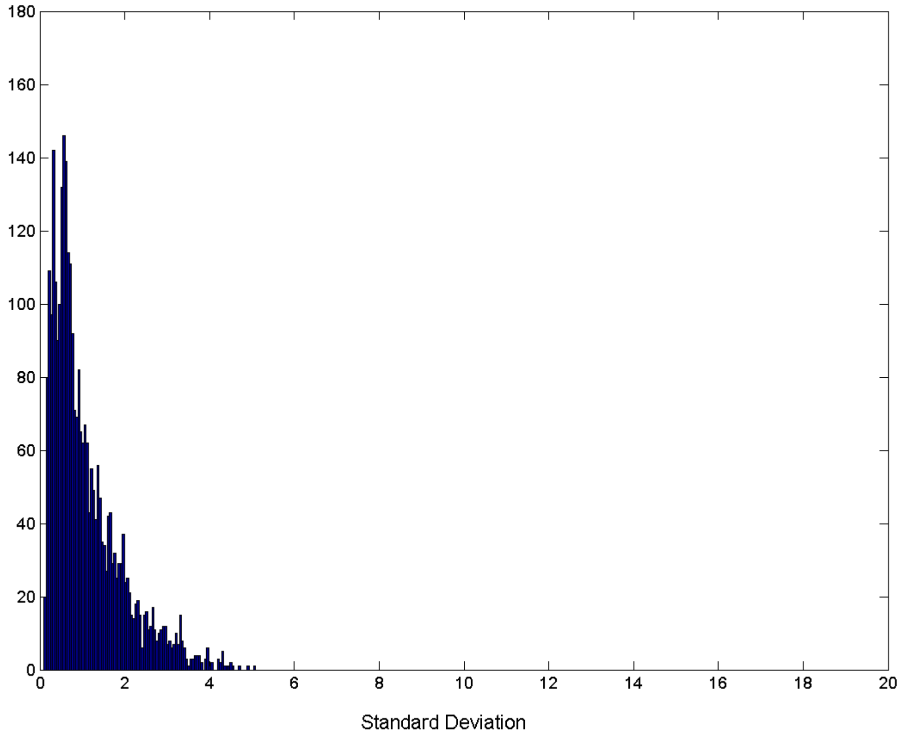
Aggregation

Behavior of group of objects in more stable than that of individual objects
The aggregate quantities have less *variability* than the individual objects being aggregated

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 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 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

Sampling

Stratified Sampling

– When subpopulations vary considerably, it is advantageous to sample each subpopulation (stratum) independently

– *Stratification* is the process of grouping members of the population into relatively homogeneous subgroups before sampling

– The strata should be mutually exclusive : every element in the population must be assigned to only one stratum. The strata should also be collectively exhaustive : no population element can be excluded

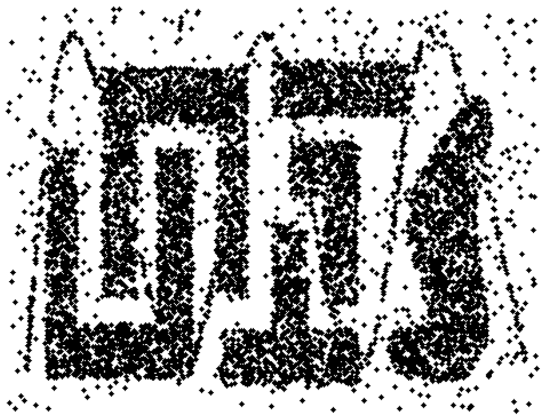
– Then random sampling is applied within each stratum. This often improves the representative-ness

of the sample by reducing sampling error

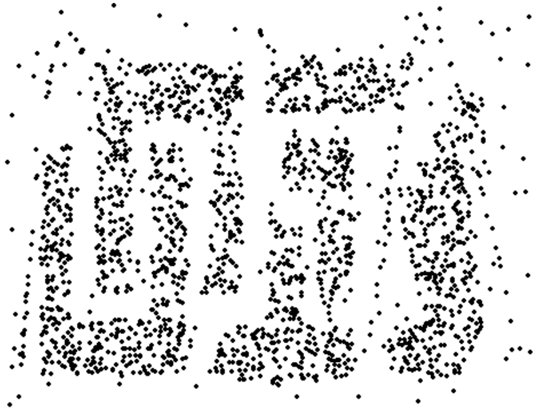
Sample size

Even if proper sampling technique is known, it is important to choose proper sample size

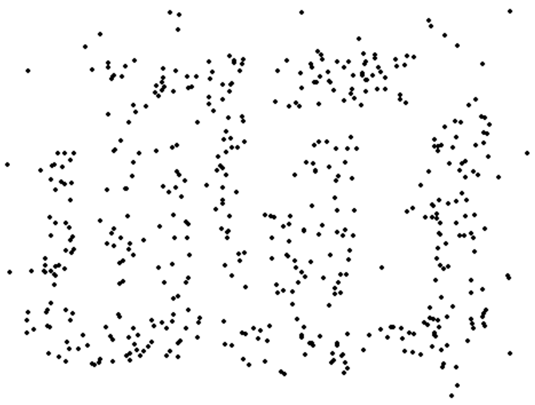
- Larger sample sizes increase the probability that a sample will be *representative*, but also eliminate much of the advantage of sampling
- With smaller sample size, patterns may be missed or erroneous patters detected



8000 points



2000 Points



500 Points

Feature Creation

Sometimes, a small number of *new* attributes can capture the important information in a data set much more efficiently than the original attributes

- Also, the number of new attributes can be often smaller than the number of original attributes. Hence, we get benefits of dimensionality reduction
- Three general methodologies:

– Feature Extraction

– Mapping the Data to a New Space

– Feature Construction

Feature extraction

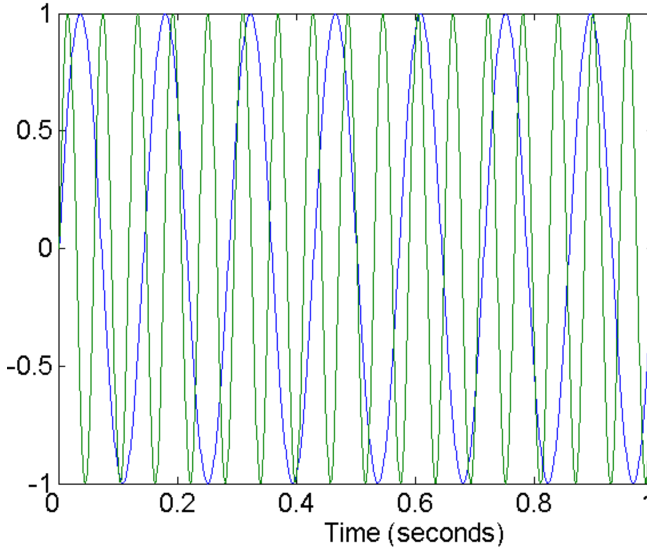
One approach to dimensionality reduction is feature extraction, which is creation of a new, smaller set of features from the original set of features

- For example, consider a set of photographs, where each photograph is to be classified whether its human face or not
- The raw data is set of pixels, and as such is not suitable for many classification algorithms
- However, if data is processed to provide high-level features like presence or absence of certain types of edges or areas correlated with presence of human faces, then a broader set of classification techniques can be applied to the problem

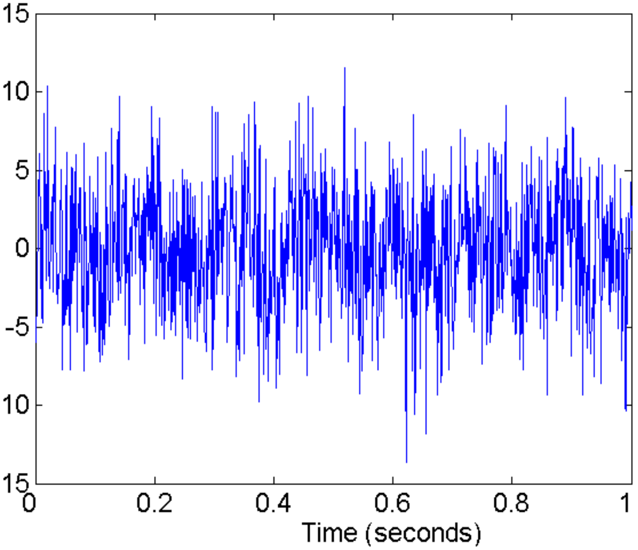
Mapping Data to a New Space

Sometimes, a totally different view of the data can reveal important and interesting features

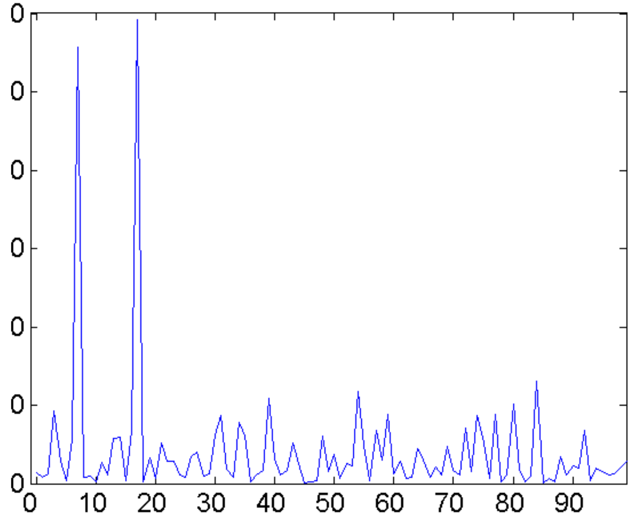
- Example: Applying Fourier transformation to data to detect time series patterns



Two Sine Waves



Two Sine Waves + Noise



Frequency

Feature Construction

Sometimes features have the necessary information, but not in the form necessary for the data mining algorithm. In this case, one or more new features constructed out of the original features may be useful

- Example, there are two attributes that record volume and mass of a set of objects
- Suppose there exists a classification model based on material of which the objects are constructed
- Then a density feature constructed from the original two features would help classification

Discretization and Binarization

Discretization is the process of converting a continuous attribute to a discrete attribute

- A common example is rounding off real numbers to integers
- Some data mining algorithms require that the data be in the form of categorical or binary attributes. Thus, it is often necessary to convert

continuous attributes in to categorical attributes and / or binary attributes

- Its pretty straightforward to convert categorical attributes in to discrete or binary attributes

Discretization of Continuous Attributes

Transformation of continuous attributes to a categorical attributes involves

– Deciding how many categories to have

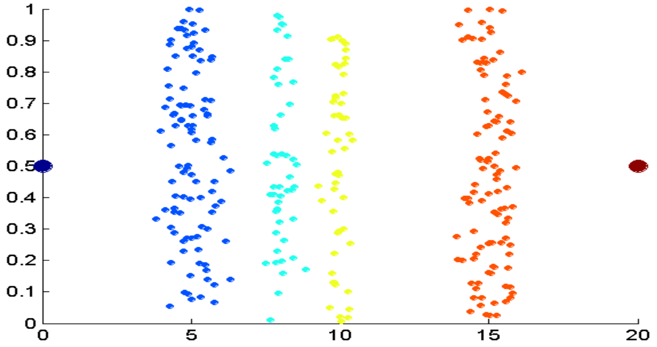
– How to map the values of the continuous attribute to categorical attribute

• A basic distinction between discretization methods for classification is whether class information is used (supervised) or not (unsupervised)

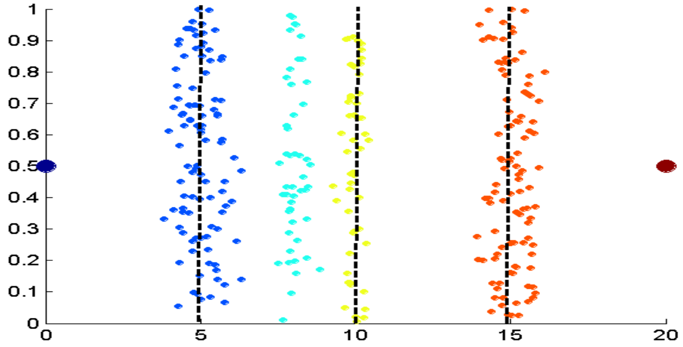
Data Discretization Methods

- Discretization: Divide the range of a continuous attribute into intervals
- Typical methods: All the methods can be applied recursively
  - Binning
    - Top-down split, unsupervised
  - Histogram analysis
    - Top-down split, unsupervised
  - Clustering analysis (unsupervised, top-down split or bottom-up merge)
  - Decision-tree analysis (supervised, top-down split)
  - Correlation (e.g., χ^2) analysis (unsupervised, bottom-up merge)

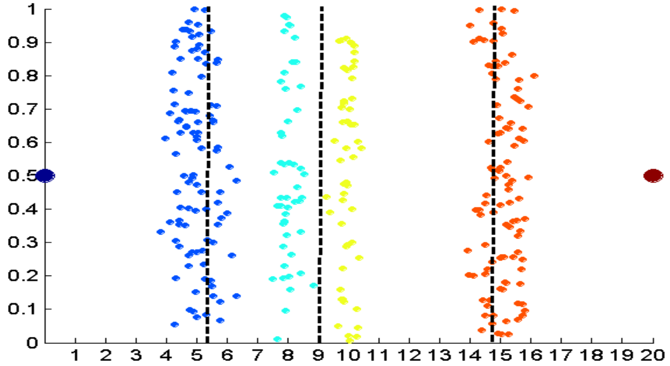
Discretization Without Using Class Labels



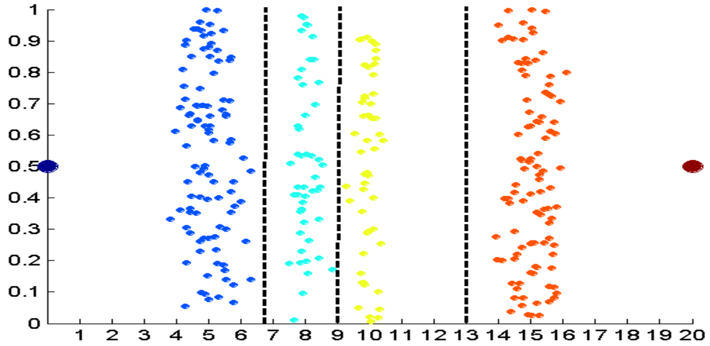
Data



Equal interval



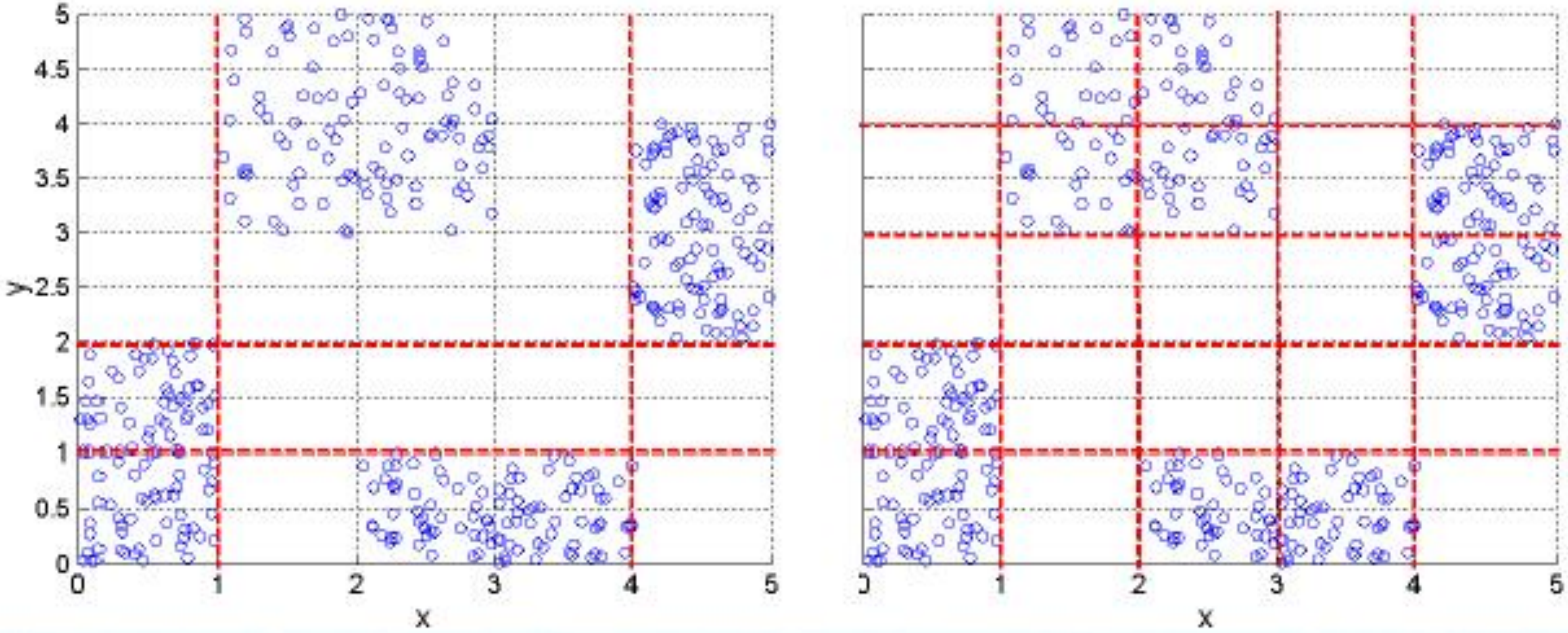
Equal frequency



K-means

Different Discretization Techniques

- Entropy based approach



3 categories for both x and y

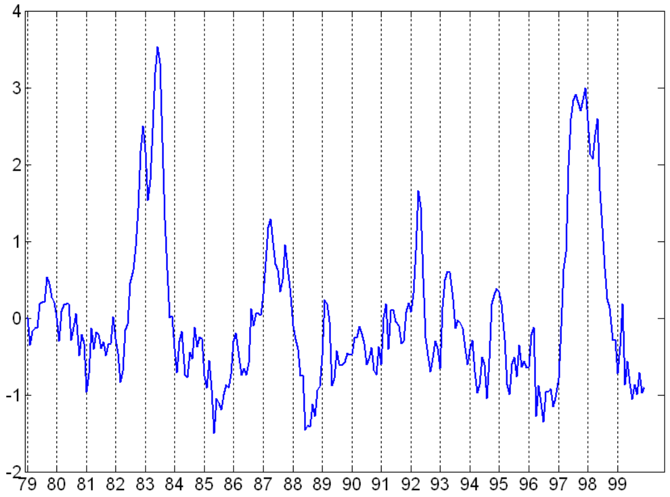
5 categories for both x and y

Discretization by Classification & Correlation Analysis

- Classification (e.g., decision tree analysis)
  - Supervised: Given class labels, e.g., cancerous vs. benign
  - Using *entropy* to determine split point (discretization point)
  - Top-down, recursive split
  - Details to be covered in Chapter "Classification"
- Correlation analysis (e.g., Chi-merge: χ^2-based discretization)
  - Supervised: use class information
  - Bottom-up merge: find the best neighboring intervals (those having similar distributions of classes, i.e., low χ^2 values) to merge
  - Merge performed recursively, until a predefined stopping condition

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



Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max(|v’ |) < 1

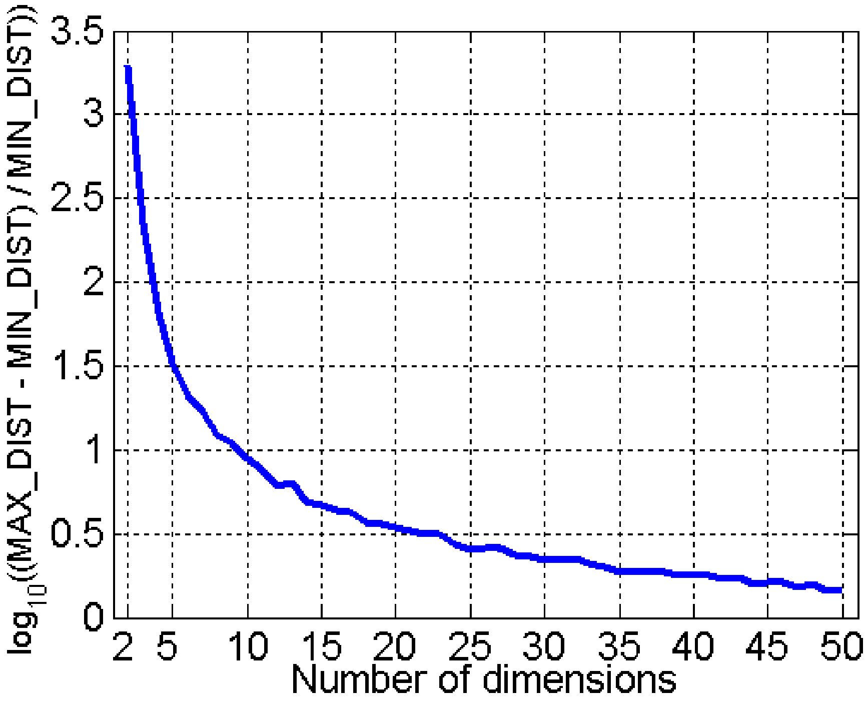
Topics

- What is data?
  - Definitions, terminology
  - Types of data and datasets
- Data preprocessing
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction
- Data similarity

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

Determining dimensions (or combinations of dimensions) that are important for modeling

- Why dimensionality reduction?

– Many data mining algorithms work better if the dimensionality of data (i.e. number of attributes) is lower

– Allows the data to be more easily visualized

– If dimensionality reduction eliminates irrelevant features or reduces noise, then quality of results may improve

– Can lead to a more understandable model

Dimensionality Reduction

Redundant features duplicate much or all of the information contained in one or more attributes

– The purchase price of product and the sales tax paid contain the same information

• *Irrelevant features* contain no information that is useful for data mining task at hand

– Student ID numbers would be irrelevant to the task of predicting their GPA

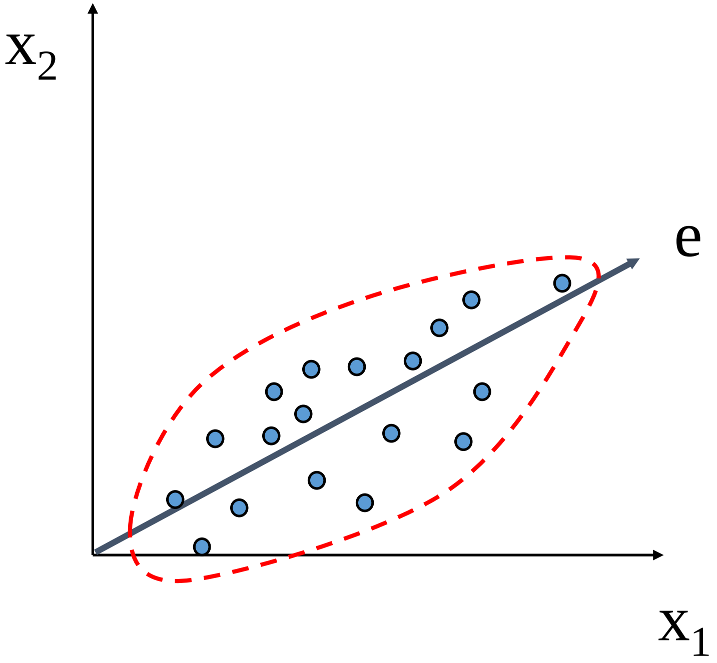
Dimensionality Reduction

- 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

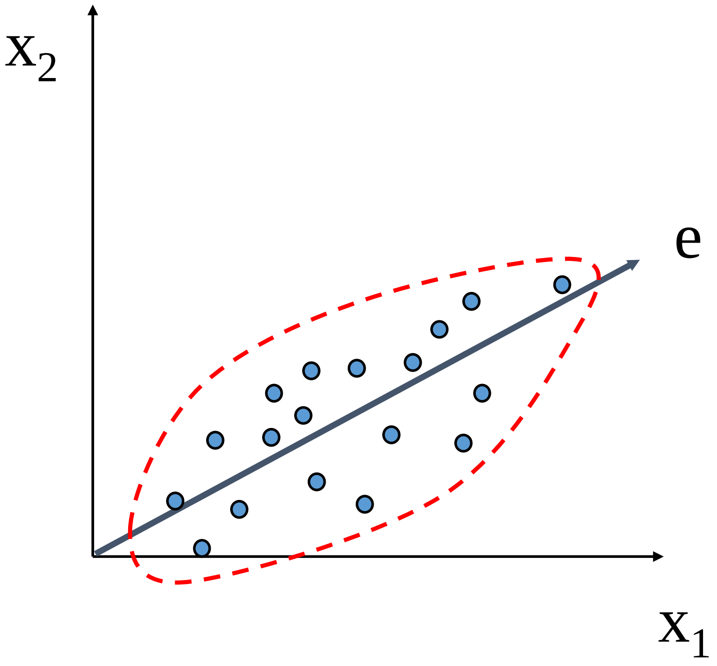
Dimensionality Reduction: PCA

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



Dimensionality Reduction: PCA

- Find the eigenvectors of the covariance matrix
- The eigenvectors define the new space



Principal Component Analysis (Steps)

- Given *N* data vectors from *n*-dimensions, find $k \leq n$ orthogonal vectors (*principal components*) that can be best used to represent data
  - Normalize input data: Each attribute falls within the same range
  - Compute *k* orthonormal (unit) vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the *k* principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)
- Works for numeric data only

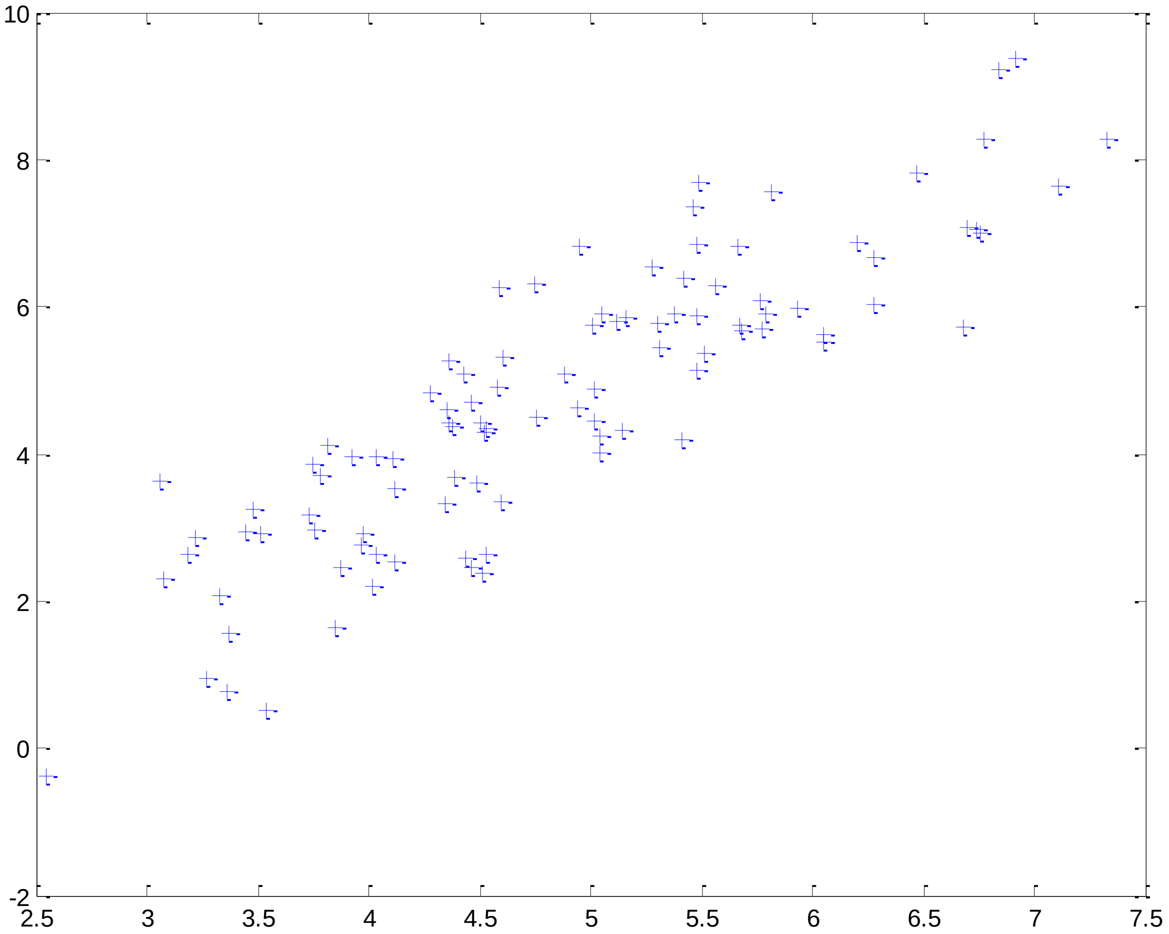
PCA Algorithm

- PCA algorithm:
  - 1. X ← Create N x d data matrix, with one row vector x_n per data point
  - 2. X subtract mean *x* from each row vector x_n in X
  - 3. Σ ← covariance matrix of X
  - Find eigenvectors and eigenvalues of Σ
  - PC’s ← the M eigenvectors with largest eigenvalues

PCA Algorithm in Matlab

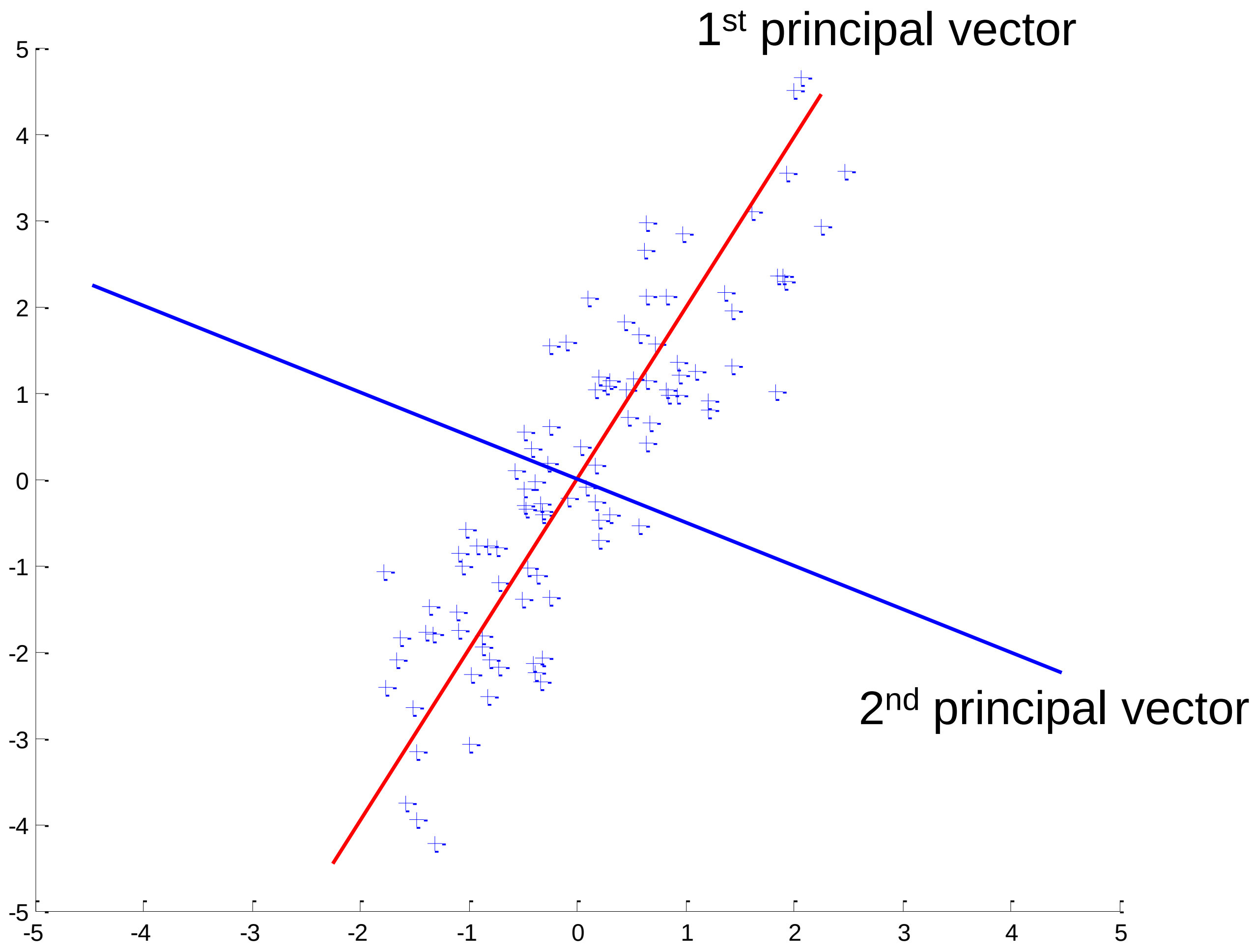
```
% generate data
Data = mvnrnd([5, 5],[1 1.5; 1.5 3], 100);
figure(1); plot(Data(:,1), Data(:,2), '+');
%center the data
for i = 1:size(Data,1)
  Data(i, :) = Data(i, :) - mean(Data);
end

DataCov = cov(Data); %covariance matrix
[PC, variances, explained] = pcacov(DataCov); %eigen

% plot principal components
figure(2); clf; hold on;
plot(Data(:,1), Data(:,2), '+b');
plot(PC(1,1)*[-5 5], PC(2,1)*[-5 5], '-r’)
plot(PC(1,2)*[-5 5], PC(2,2)*[-5 5], '-b’); hold off

% project down to 1 dimension
PcaPos = Data * PC(:, 1);
```

2d Data



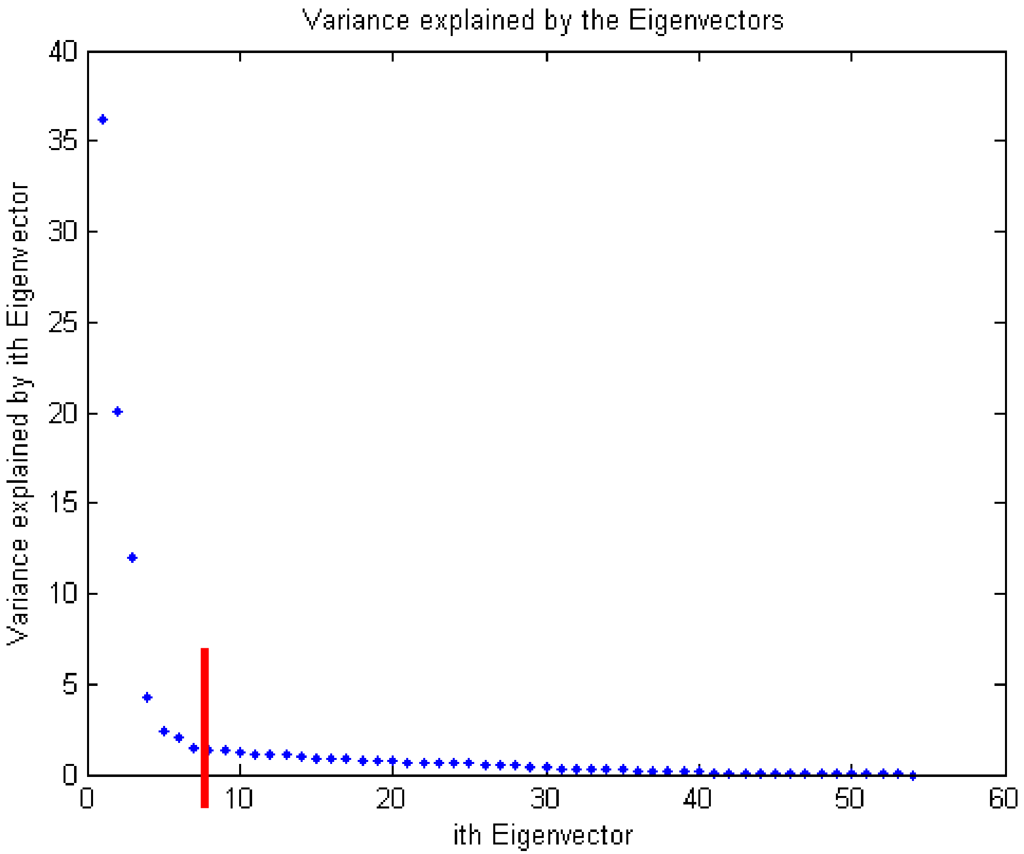
Principal Components

- Gives best axis to project
- Principal vectors are orthogonal



How many components?

- Check the distribution of eigen-values
- Take enough many eigen-vectors to cover 80-90% of the variance



Problems and limitations

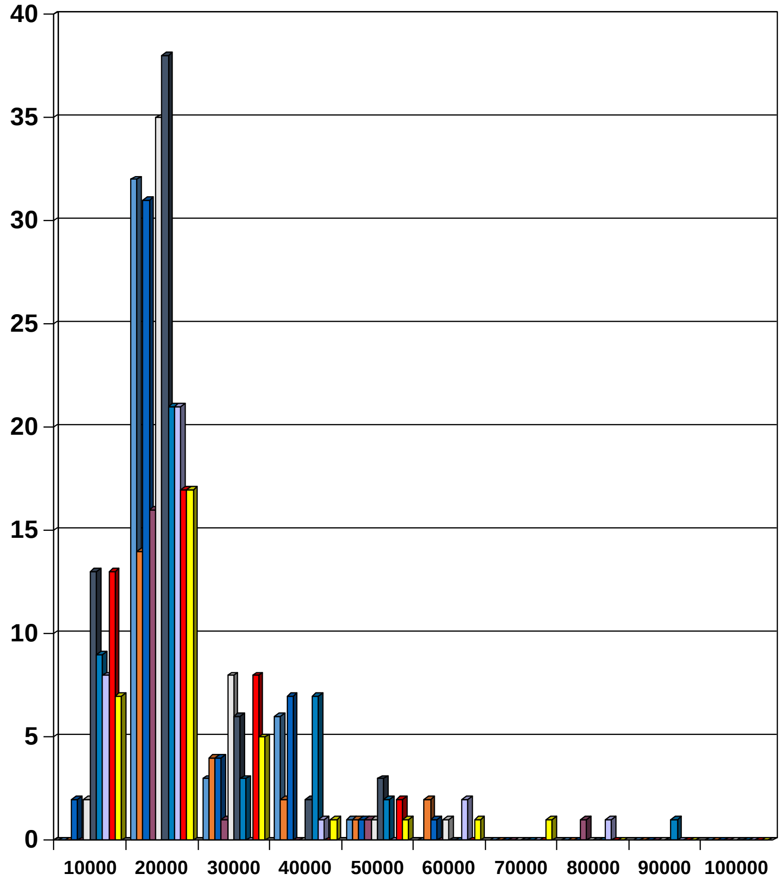
- What if very large dimensional data?
  - e.g., Images ($d \geq 10^4$)
- Problem:
  - Covariance matrix Σ is size (d^2)
  - $d=10_4$ → $|\Sigma| = 10^8$
- Singular Value Decomposition (SVD)!
  - efficient algorithms available (Matlab)
  - some implementations find just top N eigenvectors

Numerosity Reduction

- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Log-linear models: obtain value at a point in m-D space as the product on appropriate marginal subspaces
- Non-parametric methods
  - Do not assume models
  - Major families: histograms, clustering, sampling

Histograms

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Can be constructed optimally in one dimension using dynamic programming
- Related to quantization problems.



Clustering

- Partition data set into clusters, and one can store cluster representation only
- Can be very effective if data is clustered
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms, further will be discussed

Feature Subset Selection

- Another way to reduce dimensionality of data

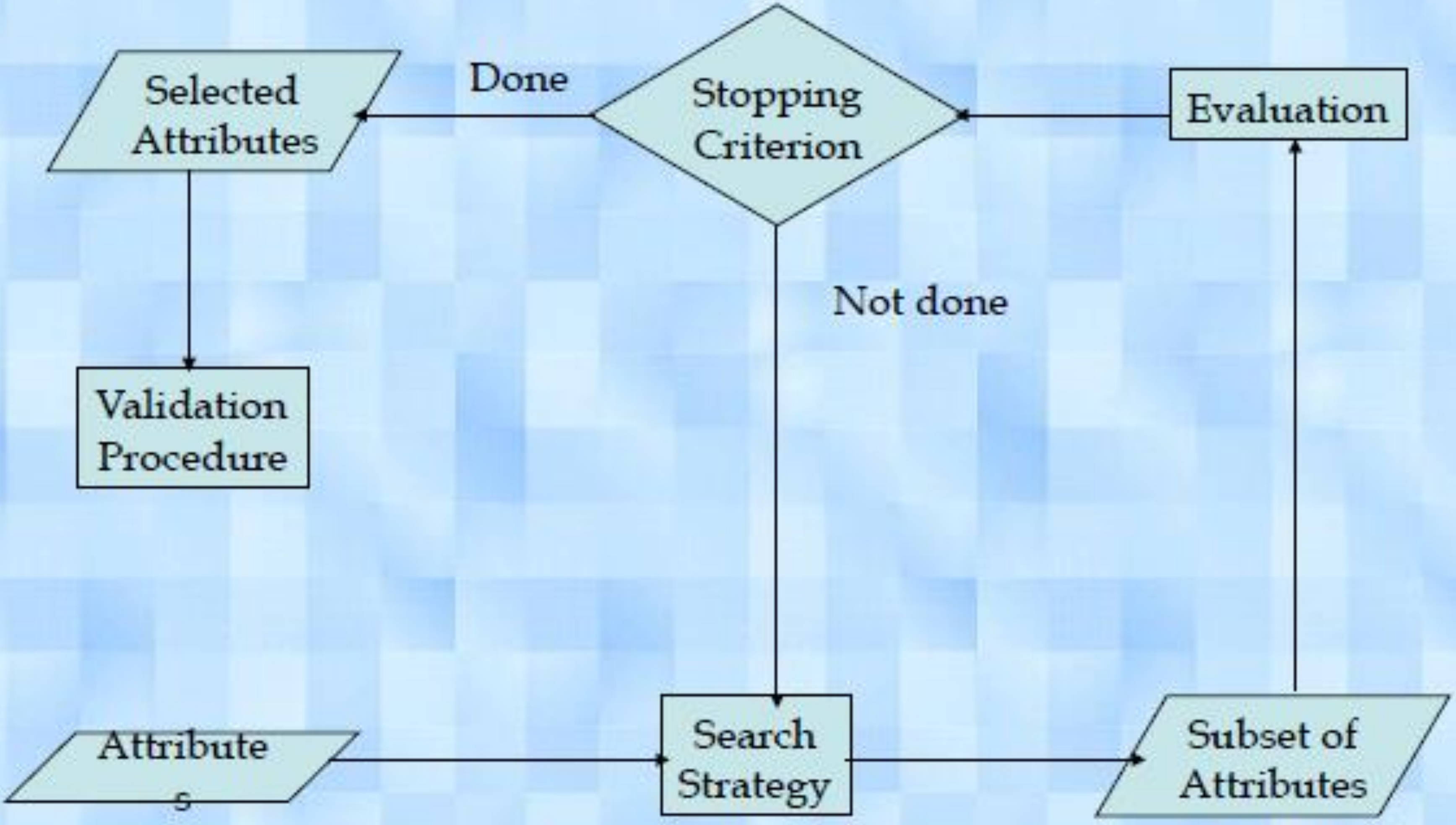
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes

- Irrelevant features
  - contain no information that is useful for the data mining task at hand

Feature Subset Selection

- Techniques:
  - Brute-force approch:
    - Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches:
    - Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches:
    - Features are selected before data mining algorithm is run
  - Wrapper approaches:
    - Use the data mining algorithm as a black box to find best subset of attributes

Architecture for Feature Subset Selection



Flowchart of a feature subset selection process

Topics

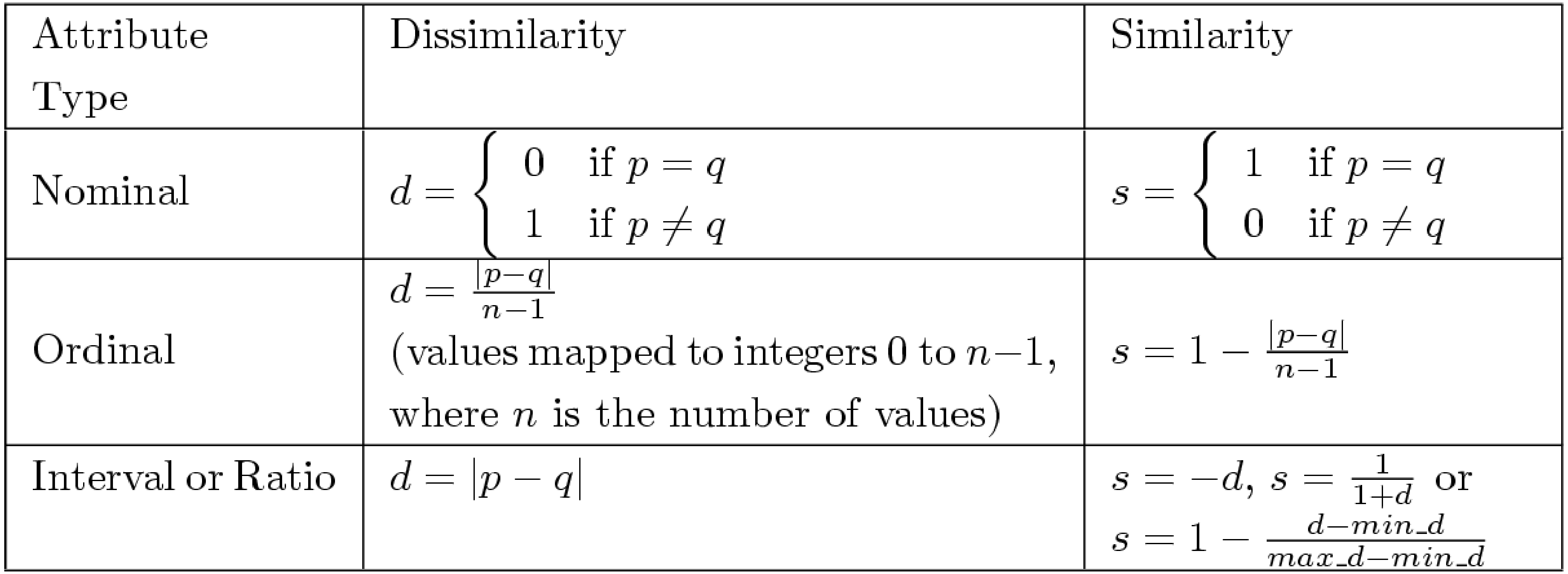
- What is data?
    - Definitions, terminology
    - Types of data and datasets
- Data preprocessing
    - Data Cleaning
    - Data integration
    - Data transformation
    - Data reduction
- Data similarity

Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

p and *q* are the attribute values for two data objects.



Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d, \; s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

Euclidean Distance

- Euclidean Distance (L_2 norm)

Where *n* is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects *p* and *q*.

- Standardization is necessary, if scales differ.

$$\boldsymbol{dist} = \sqrt{\sum_{\boldsymbol{k}=1}^{\boldsymbol{n}} (\boldsymbol{p_k} - \boldsymbol{q_k})^2}$$

Minkowski Distance

- Minkowski Distance (L_p norm) is a generalization of Euclidean Distance

Where *r* is a parameter, *n* is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects *p* and *q*.

$$dist = \left(\sum_{k=1}^{n} |p_k - q_k|^p\right)^{\frac{1}{p}}$$

Minkowski Distance: Examples

- p = 1. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors

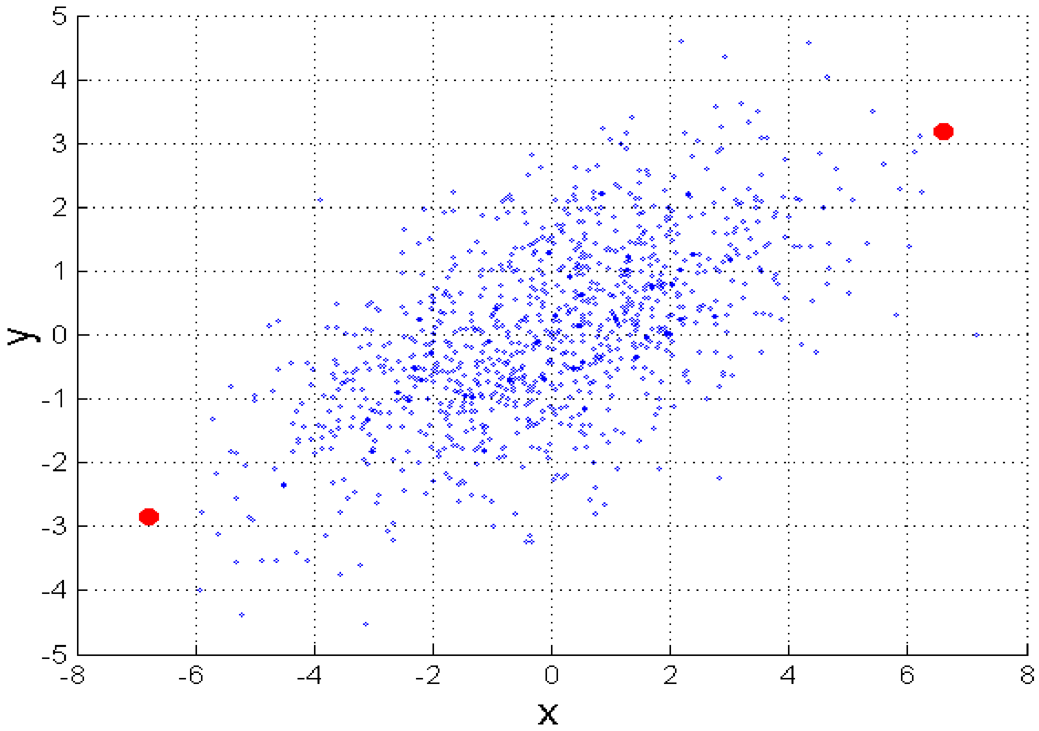
- p = 2. Euclidean distance

- $p \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors

- Do not confuse p with n, i.e., all these distances are defined for all numbers of dimensions.

Mahalanobis Distance

$$\boldsymbol{mahalanobis(p,q) = (p-q)\Sigma^{-1}(p-q)^T}$$

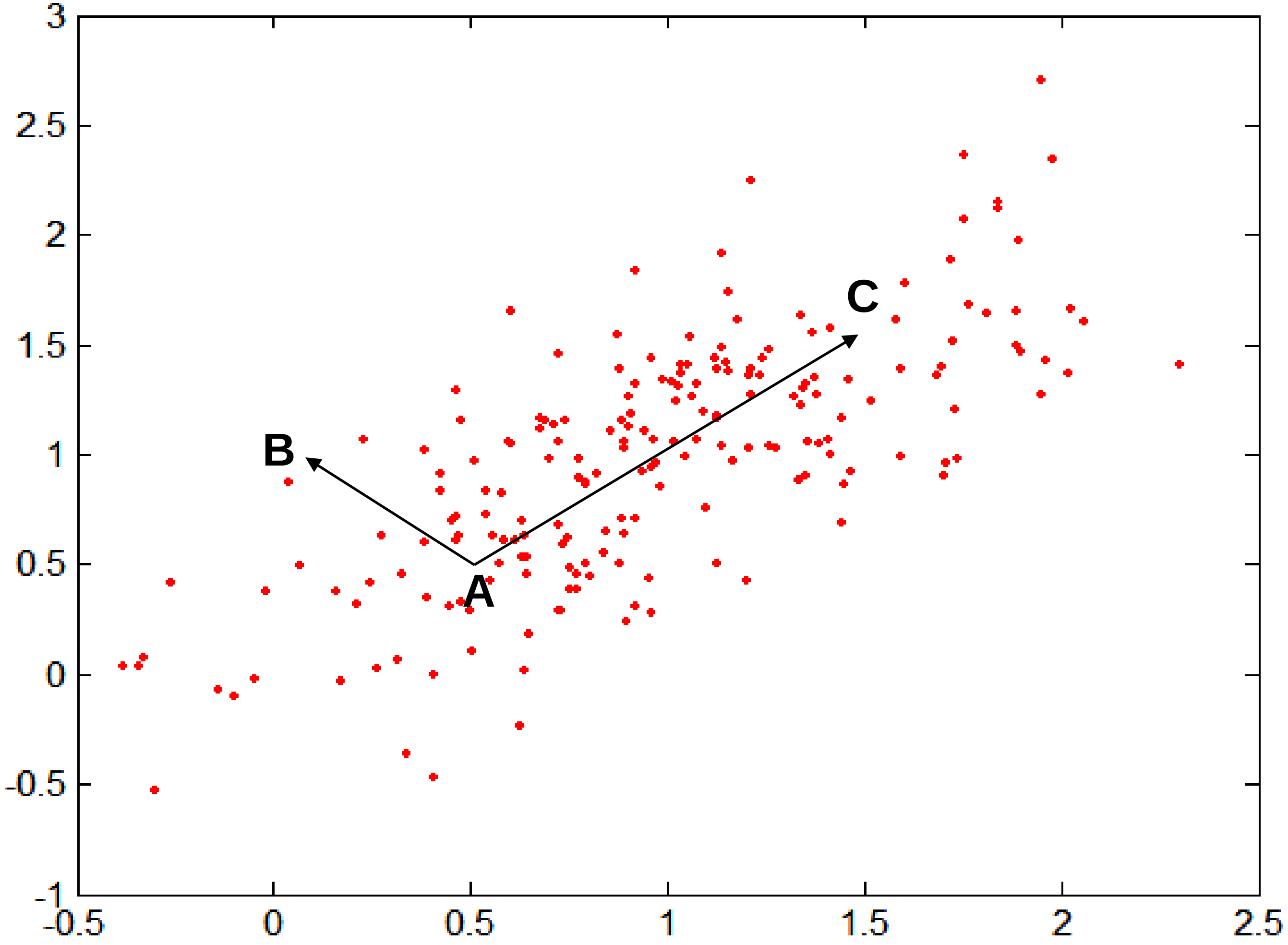


Σ is the covariance matrix of the input data *X*

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\overline{X}_j)(X_{ik}-\overline{X}_k)$$

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes

- Compute similarities using the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1
  M_{10} = the number of attributes where p was 1 and q was 0
  M_{00} = the number of attributes where p was 0 and q was 0
  M_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes
  = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = number of 11 matches / number of not-both-zero attributes values
  = $(M_{11}) / (M_{01} + M_{10} + M_{11})$

SMC versus Jaccard: Example

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

M_{01} = 2 (the number of attributes where p was 0 and q was 1)

M_{10} = 1 (the number of attributes where p was 1 and q was 0)

M_{00} = 7 (the number of attributes where p was 0 and q was 0)

M_{11} = 0 (the number of attributes where p was 1 and q was 1)

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$

J = $(M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

where $\bullet$ indicates vector dot product and $|| d ||$ is the length of vector d.

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$||d_1||$ = $(3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{\mathbf{0.5}}$ = $(42)^{\mathbf{0.5}}$ = 6.481

$||d_2||$ = $(1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{\mathbf{0.5}}$ = $(6)^{\mathbf{0.5}}$ = 2.245

$\cos(d_1, d_2)$ = 0.3150

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use weights w_k which are between 0 and 1 and sum to 1.

$$similarity(p,q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

$$distance(p,q) = \left(\sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}.$$

Summary

- Data preparation is a big issue for data mining
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- Many methods have been proposed but still an active area of research
- Method to choose depends on the nature of the data