Unit III

Data Preprocessing

Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

	team	coach	pla y	ball	score	game	wi n	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

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

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples* , *examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

Attributes

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

- **Nominal:** categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - <u>Symmetric binary</u>: both outcomes equally important
    - e.g., gender
  - <u>Asymmetric binary</u>: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., 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

Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Basic Statistical Descriptions of Data

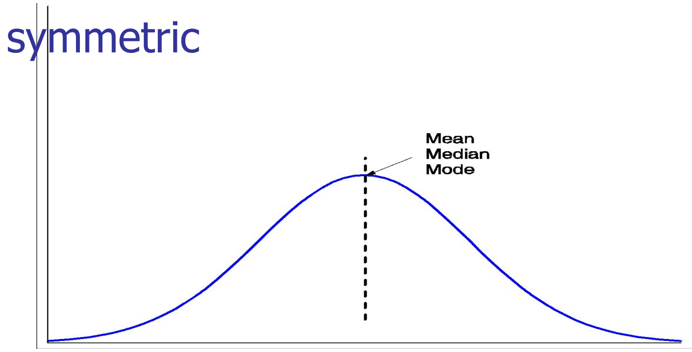
- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

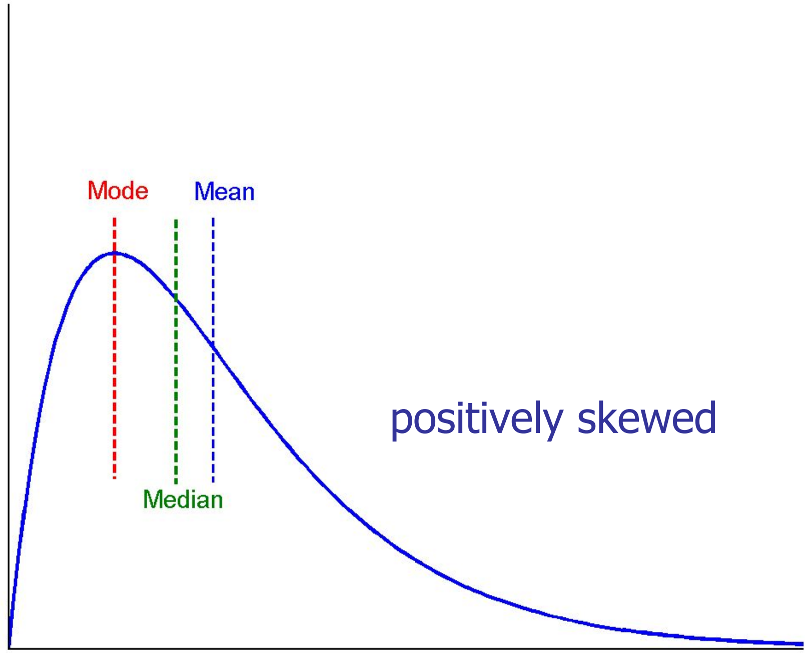
Measuring the Central Tendency

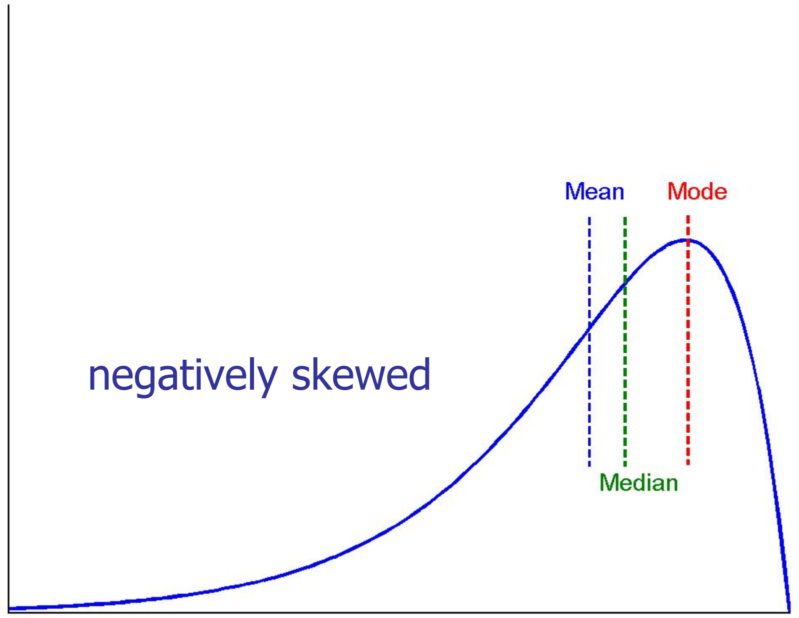
- <u>Mean (algebraic measure) (sample vs. population):</u>
  - Weighted arithmetic mean:
- <u>Median</u>:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):
- <u>Mode</u>
  - Value that occurs most frequently in the data

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







Measuring the Dispersion of Data

Statistical dispersion means the extent to which a numerical data is likely to vary about an average value

Measure of Dispersion

1. Range
2. Variance
3. Standard Deviation
4. Skewness
5. IQR

Measuring the Dispersion of Data

RANGE

Range: Range is the measure of the difference between the largest and smallest value of the data variability. The range is the simplest form of Measures of Dispersion.

Example: 1,2,3,4,5,6,7

Range = Highest value – Lowest value

= (7 – 1) = 6

Measuring the Dispersion of Data

MEAN

Example: 1,2,3,4,5,6,7,8

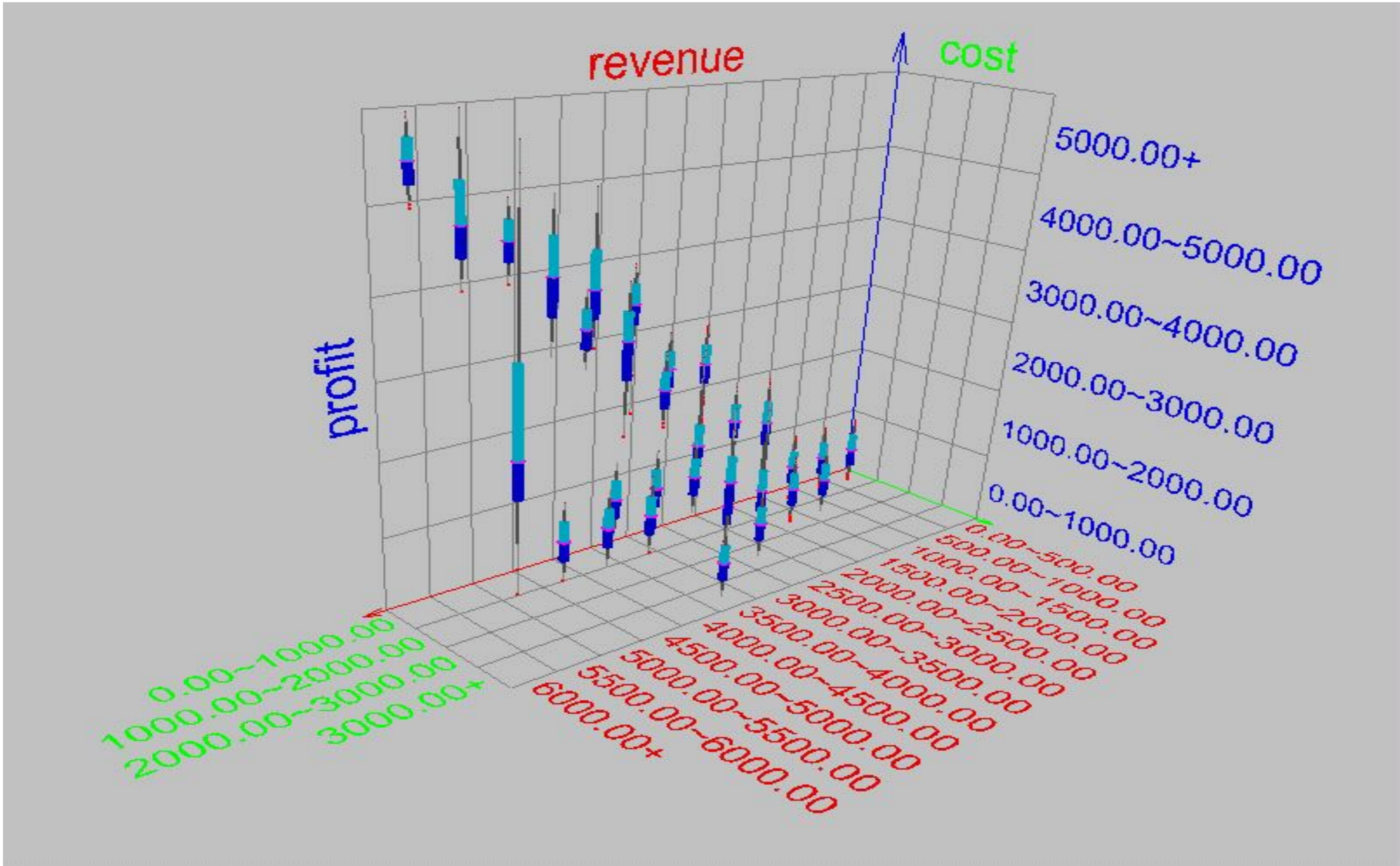
Mean = (sum of all the terms / total number of terms)

= (1 + 2 + 3 + 4 + 5 + 6 + 7 + 8) / 8

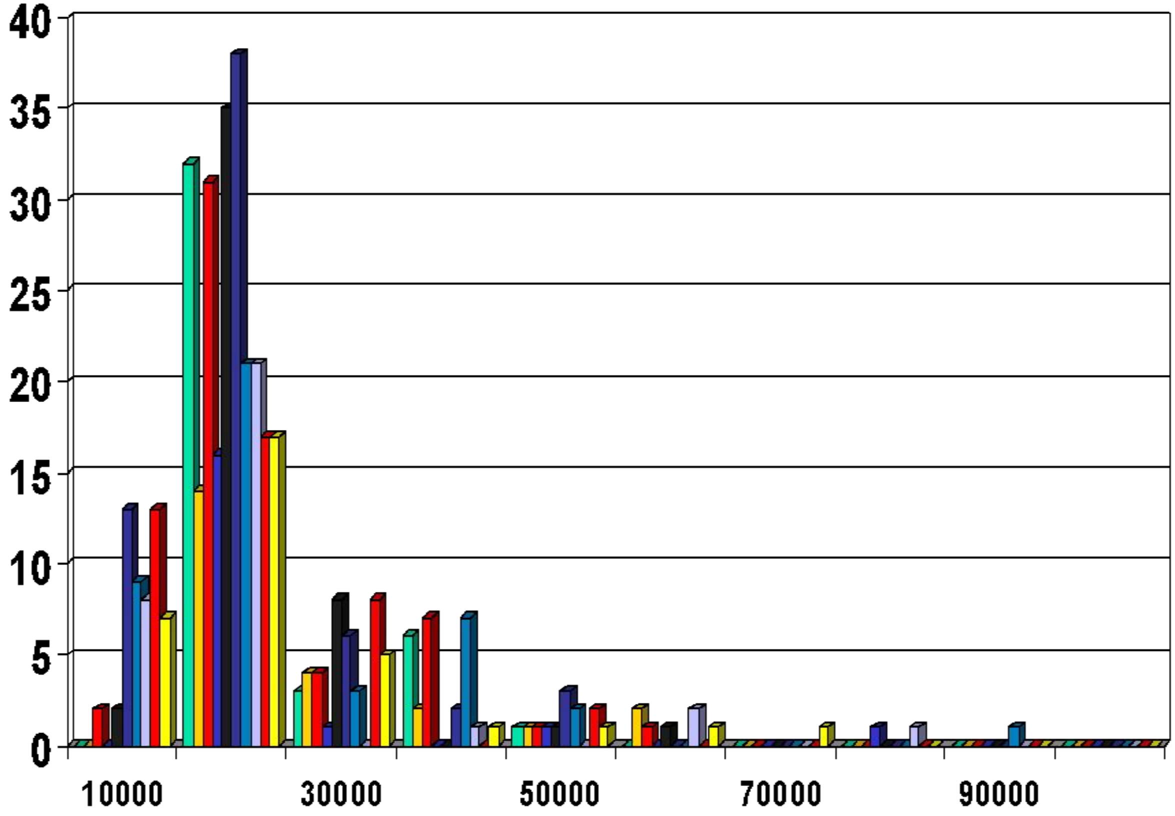
= 36 / 8

= 4.5

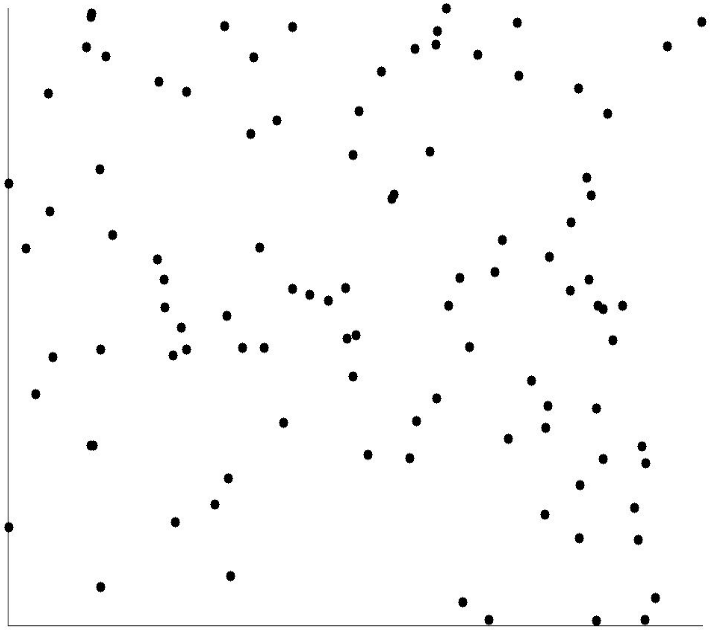
Visualization of Data Dispersion: 3-D Boxplots

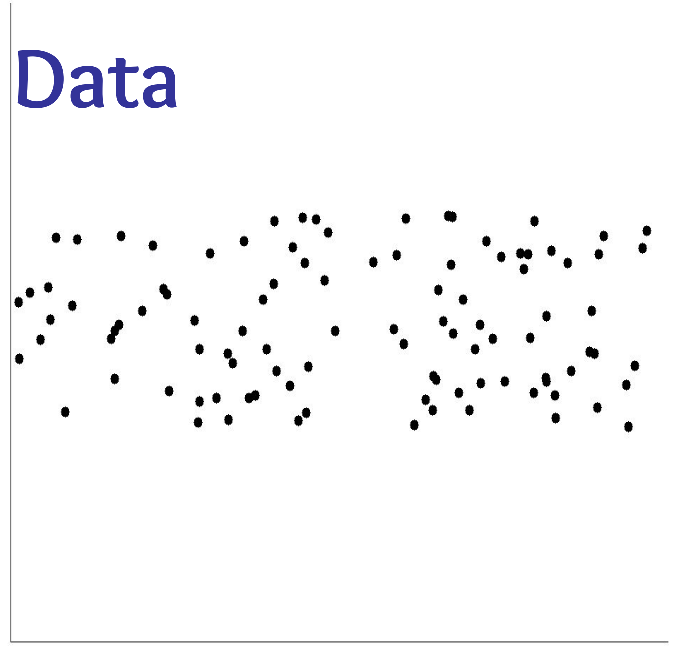


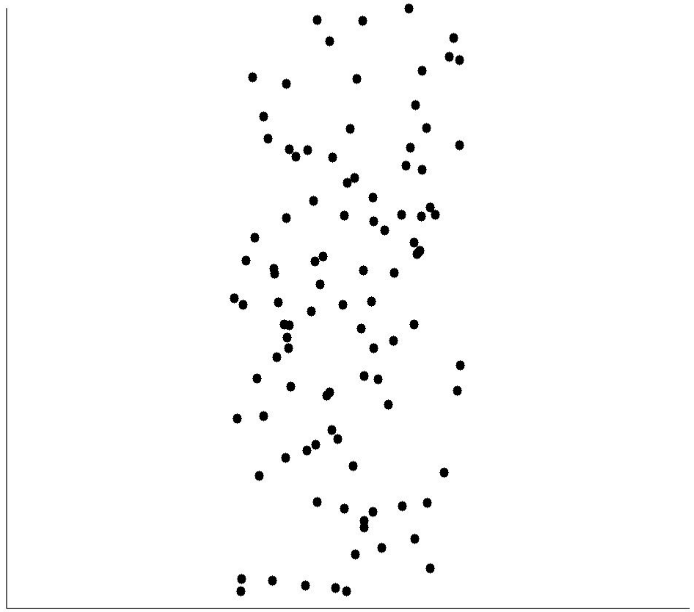
Histogram Analysis



Uncorrelated Data







Chapter 2: Getting to Know Your Data

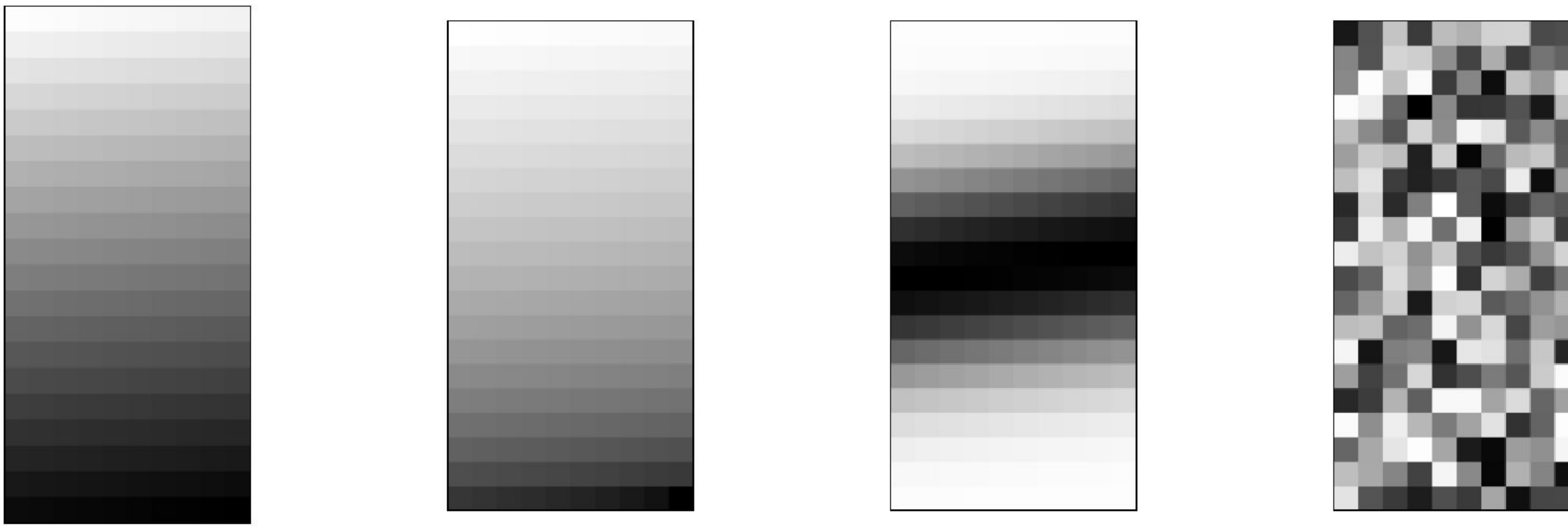
- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Data Visualization

- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters for further quantitative analysis
  - Provide a visual proof of computer representations derived
- Categorization of visualization methods:
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques
  - Icon-based visualization techniques
  - Hierarchical visualization techniques
  - Visualizing complex data and relations

Pixel-Oriented Visualization Techniques

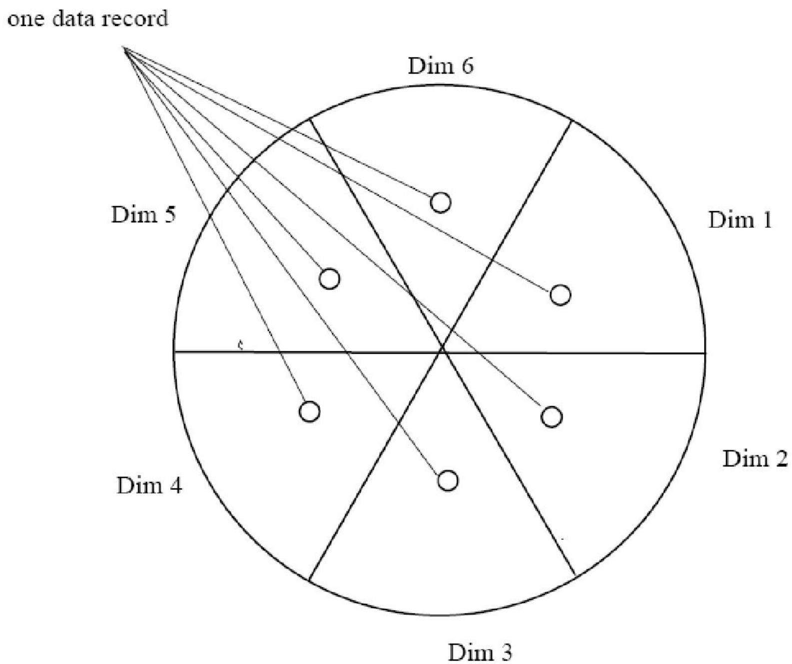
- For a data set of m dimensions, create m windows on the screen, one for each dimension
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values



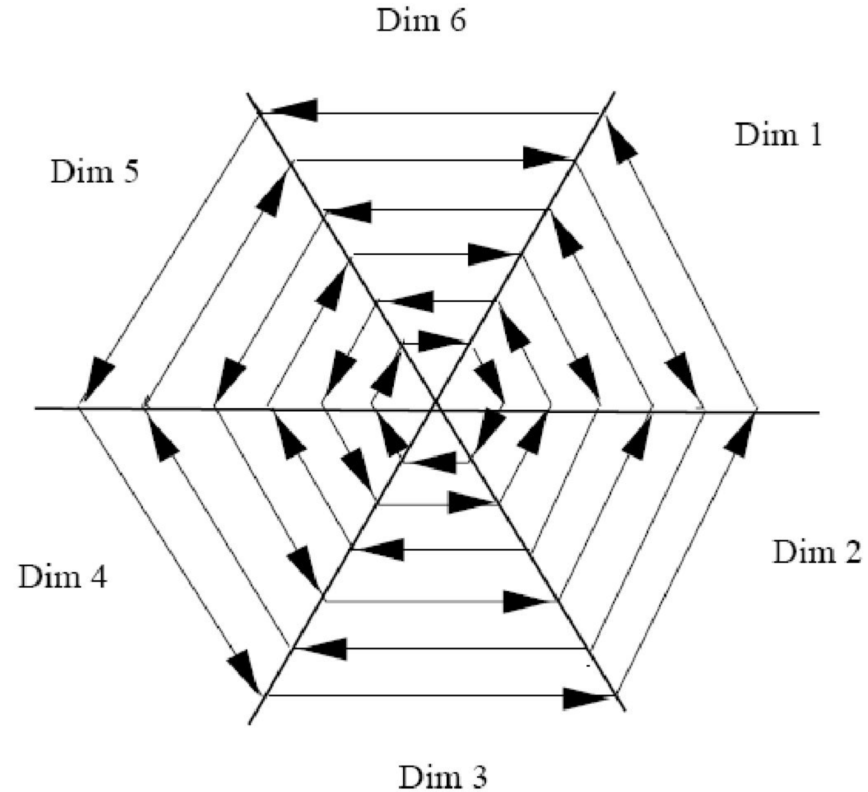
(a) Income (b) Credit Limit (c) transaction volume (d) age

Laying Out Pixels in Circle Segments

- To save space and show the connections among multiple dimensions, space filling is often done in a circle segment



(a) Representing a data record in circle segment



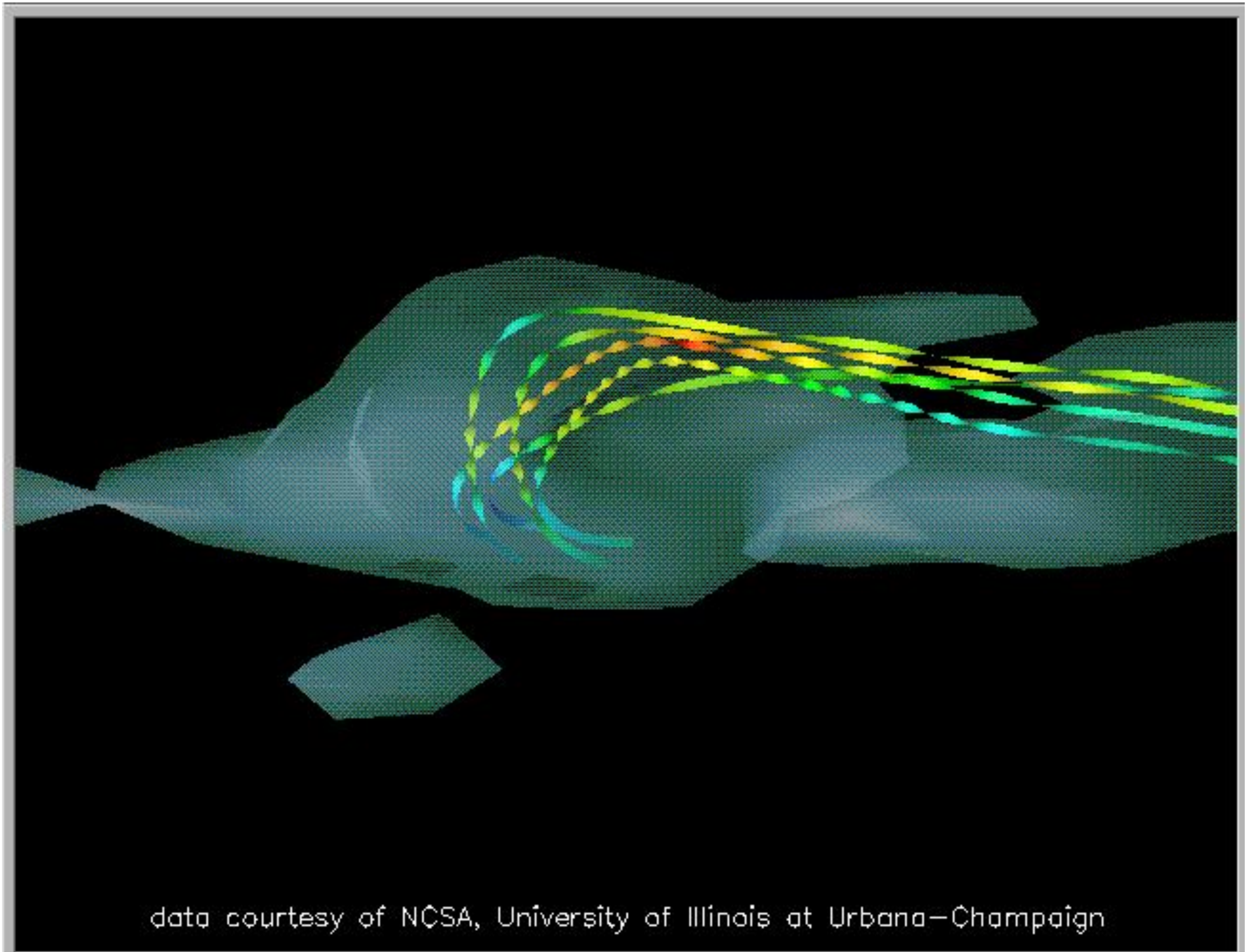
(b) Laying out pixels in circle segment

Geometric Projection Visualization Techniques

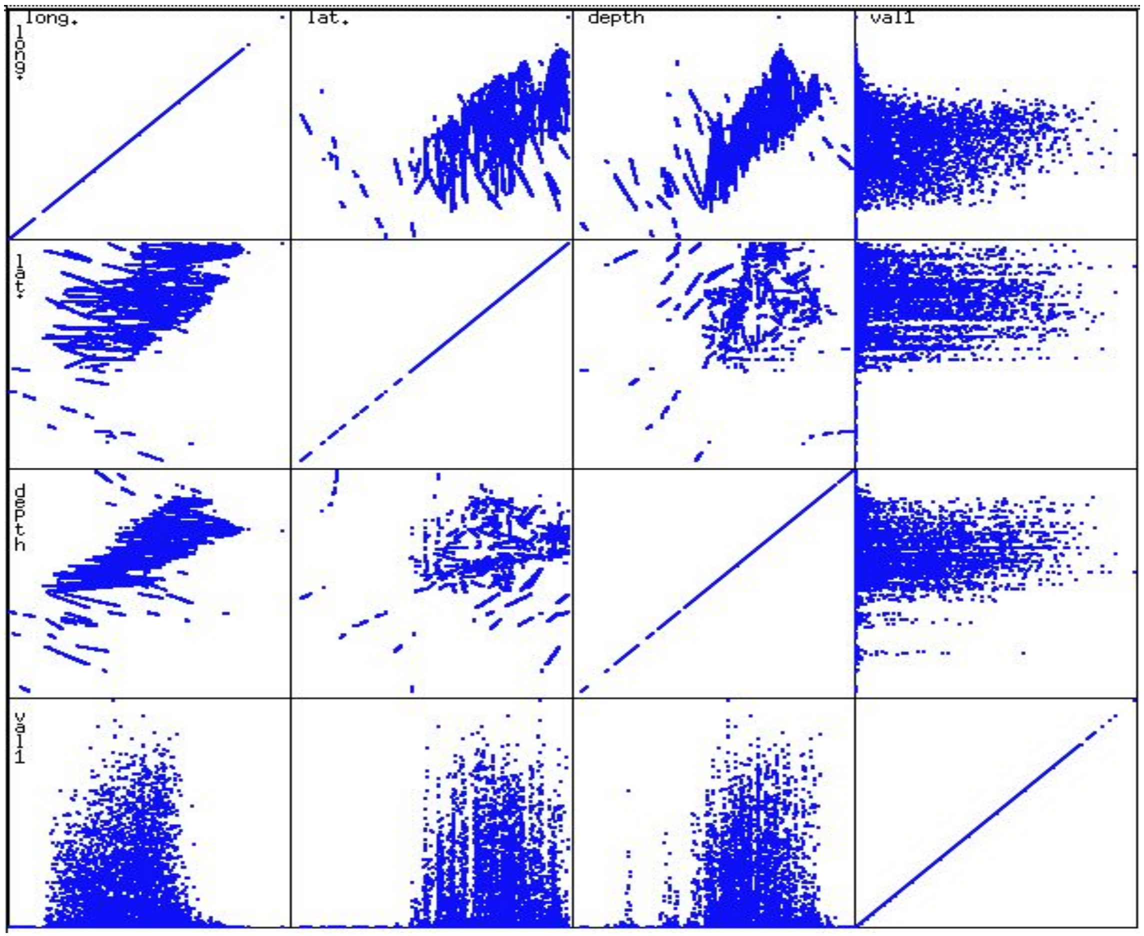
- Visualization of geometric transformations and projections of the data
- Methods
  - Direct visualization
  - Scatterplot and scatterplot matrices
  - Landscapes
  - Projection pursuit technique: Help users find meaningful projections of multidimensional data
  - Prosection views
  - Hyperslice
  - Parallel coordinates

Direct Data Visualization

Ribbons with Twists Based on Vorticity

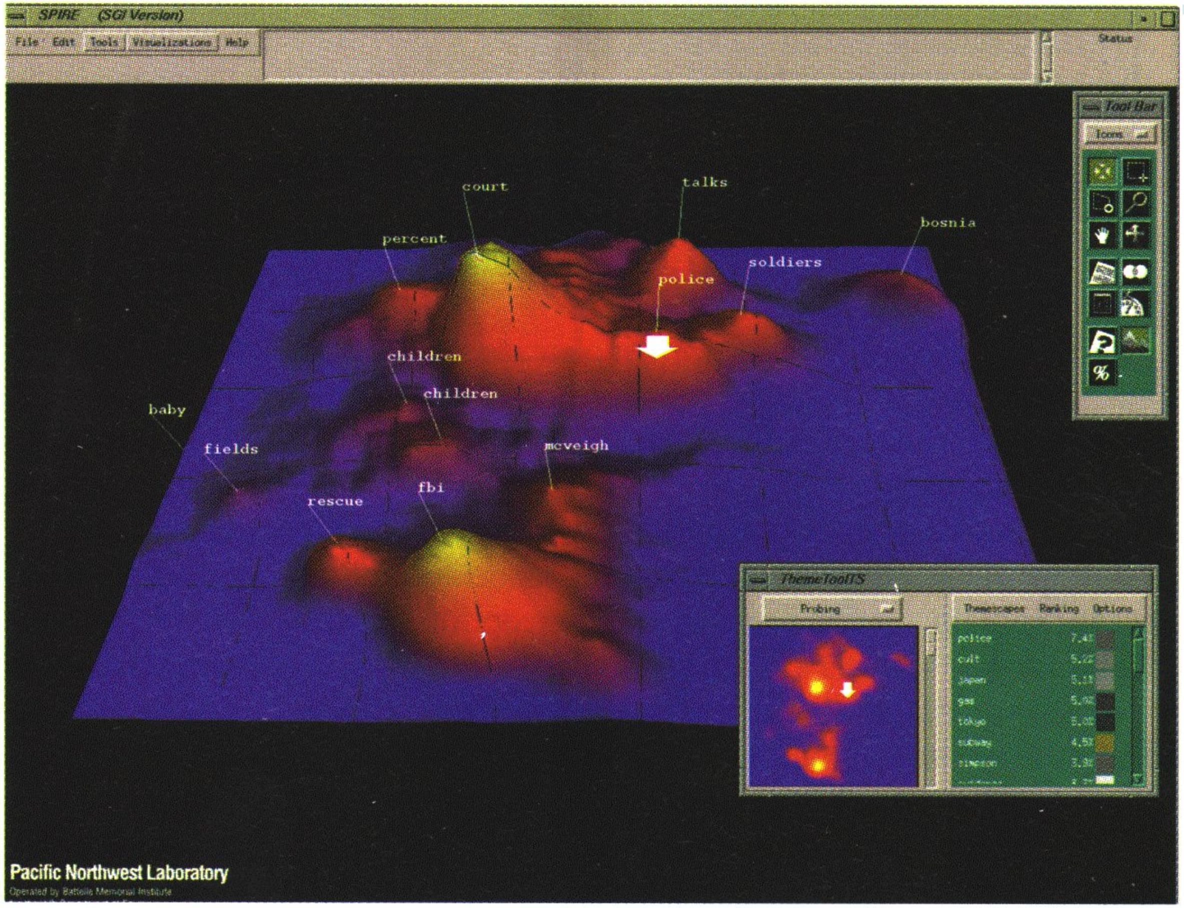


Scatterplot Matrices



Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of (k2/2-k) scatterplots]

Landscapes



news articles visualized as a landscape

- Visualization of the data as perspective landscape
- The data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

Similarity and Dissimilarity

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

Similarity and Dissimilarity

The similarity measure is a way of measuring how data samples are related or closed to each other.

On the other hand, the dissimilarity measure is to tell how much the data objects are distinct.

Similarity and Dissimilarity

Similarity/Dissimilarity for Simple Attributes

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

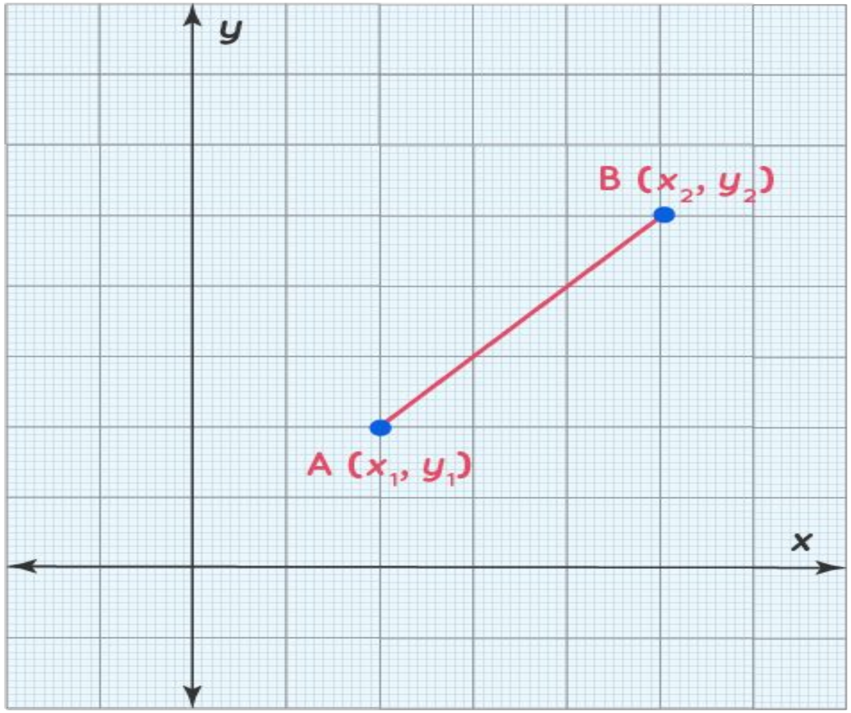
Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is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

Similarity and Dissimilarity

- Nominal attributes only tell us about the distinctness of objects. Hence, in this case similarity is defined as 1 if attribute values match, and 0 otherwise and oppositely defined would be dissimilarity.
- For objects with a single **ordinal** attribute, the situation is more complicated because information about order needs to be taken into account. Consider an attribute that measures the quality of a product, on the scale {poor, fair, OK, good, wonderful}. We have 3 products P1, P2, & P3 with quality as wonderful, good, & OK respectively. In order to compare **ordinal** quantities, they are mapped to successive integers. In this case, if the scale is mapped to {0, 1, 2, 3, 4} respectively. Then, dissimilarity(P1, P2) = 4−3 = 1.
- For **interval or ratio** attributes, the natural measure of dissimilarity between two objects is the absolute difference of their values. For example, we might compare our current weight and our weight a year ago by saying "I am ten pounds heavier."

Similarity and Dissimilarity

Euclidean Distance



$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Similarity and Dissimilarity

Euclidean Distance

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

p, q	= two points in Euclidean n-space
q_i, p_i	= Euclidean vectors, starting from the origin of the space (initial point)
n	= n-space

Similarity and Dissimilarity

Euclidean Distance - Example 1

Table 1

	1 Var1	2 Var2
Person 1	20	80
Person 2	30	44
Person 3	90	40

Using equation 1 ...

$$d = \sqrt{\sum_{i=1}^{v} (p_{1i} - p_{2i})^2}$$

For the distance between person **1** and **2**, the calculation is:

$$d = \sqrt{(20-30)^2 + (80-44)^2} = 37.36$$

For the distance between person **1** and **3**, the calculation is:

$$d = \sqrt{(20-90)^2 + (80-40)^2} = 80.62$$

For the distance between person **2** and **3**, the calculation is:

$$d = \sqrt{(30-90)^2 + (44-40)^2} = 60.13$$

Similarity and Dissimilarity

Euclidean Distance - Example - 2

As an example, the (Euclidean) distance between points (2, -1) and (-2, 2) is found to be

$$
\begin{aligned}
\text{dist}((2, -1), (-2, 2)) &= \sqrt{(2 - (-2))^2 + ((-1) - 2)^2} \\
&= \sqrt{(2 + 2)^2 + (-1 - 2)^2} \\
&= \sqrt{(4)^2 + (-3)^2} \\
&= \sqrt{16 + 9} \\
&= \sqrt{25} \\
&= 5.
\end{aligned}
$$

Similarity and Dissimilarity

Other Distance Measures

- Manhattan distance
- Pearson correlation distance
- Cosine Similarity Measure
- Minkowski distance

Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Data visualization: map data onto graphical primitives
  - Measure data similarity
- Above steps are the beginning of data preprocessing.
- Many methods have been developed but still an active area of research.