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:

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>

Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion
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)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: 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

Mean (algebraic measure) (sample vs. population):
\[ x = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \mu = \frac{\sum x}{N} \]

Note: \( n \) is sample size and \( N \) is population size.

Weighted arithmetic mean:
\[ \bar{x} = \frac{\sum w_i x_i}{\sum w_i} \]

Trimmed mean: chopping extreme values

Median:
- Middle value if odd number of values, or average of the middle two values otherwise

Estimated by interpolation (for grouped data):
\[ \text{median} = L_i + \left( \frac{n/2 - \left( \sum \text{freq} \right)}{\text{freq}_{\text{median}}} \right) \text{width} \]

Mode
- Value that occurs most frequently in the data
- Unimodal, bimodal, trimodal
- Empirical formula: \( \text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median}) \)
Symmetric vs. Skewed Data

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

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = Q_3 - Q_1
  - Five number summary: min, Q_1, median, Q_3, max
  - Boxplot: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (sample: s, population: σ)
  - Variance: (algebraic, scalable computation)
  - Standard deviation s (or σ) is the square root of variance s^2 (or σ^2)

Boxplot Analysis

- Five-number summary of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually

Visualization of Data Dispersion: 3-D Boxplots
Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From $\mu-\sigma$ to $\mu+\sigma$: contains about 68% of the measurements ($\mu$: mean, $\sigma$: standard deviation)
  - From $\mu-2\sigma$ to $\mu+2\sigma$: contains about 95% of it
  - From $\mu-3\sigma$ to $\mu+3\sigma$: contains about 99.7% of it

Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis represents frequencies
- **Quantile plot**: each value $x_i$ is paired with $f_i$ indicating that approximately 100 $f_i$% of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Histogram Analysis

- **Histogram**: Graph display of tabulated frequencies, shown as bars
  - It shows what proportion of cases fall into each of several categories
  - Diffs from a bar chart in that it is the area of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
  - The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent

Histograms Often Tell More than Boxplots

- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
  - But they have rather different data distributions
**Quantile Plot**
- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
  - For a data $x_i$, data sorted in increasing order, $f_i$ indicates that approximately $100f_i\%$ of the data are below or equal to the value $x_i$.

**Quantile-Quantile (Q-Q) Plot**
- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

**Scatter plot**
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

**Positively and Negatively Correlated Data**
- The left half fragment is positively correlated
- The right half is negative correlated
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.

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 (k^2/2 - k) scatterplots].

Data courtesy of UCCSA, University of Illinois at Urbana-Champaign.
Landschapes

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

Parallel Coordinates

- n equidistant axes which are parallel to one of the screen axes and correspond to the attributes
- The axes are scaled to the [minimum, maximum]: range of the corresponding attribute
- Every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to the value for the attribute

Parallel Coordinates of a Data Set

Icon-Based Visualization Techniques

- Visualization of the data values as features of icons
- Typical visualization methods
  - Chernoff Faces
  - Stick Figures
- General techniques
  - Shape coding: Use shape to represent certain information encoding
  - Color icons: Use color icons to encode more information
  - Tile bars: Use small icons to represent the relevant feature vectors in document retrieval
**Chernoff Faces**

- A way to display variables on a two-dimensional surface, e.g., let $x$ be eyebrow slant, $y$ be eye size, $z$ be nose length, etc.
- The figure shows faces produced using 10 characteristics—head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening): Each assigned one of 10 possible values, generated using Mathematica (S. Dickson)

**Hierarchical Visualization Techniques**

- Visualization of the data using a hierarchical partitioning into subspaces
- **Methods**
  - Dimensional Stacking
  - Worlds-within-Worlds
  - Tree-Map
  - Cone Trees
  - InfoCube

**Dimensional Stacking**

- Partitioning of the $n$-dimensional attribute space in 2-D subspaces, which are ‘stacked’ into each other
- Partitioning of the attribute value ranges into classes. The important attributes should be used on the outer levels.
- Adequate for data with ordinal attributes of low cardinality
- But, difficult to display more than nine dimensions
- Important to map dimensions appropriately
Visualization of oil mining data with longitude and latitude mapped to the outer x-, y-axes and ore grade and depth mapped to the inner x-, y-axes

**Dimensional Stacking**

- Assign the function and two most important parameters to innermost world
- Fix all other parameters at constant values - draw other (1 or 2 or 3 dimensional worlds choosing these as the axes)
- Software that uses this paradigm
  - N–vision: Dynamic interaction through data glove and stereo displays, including rotation, scaling (inner) and translation (inner/outer)
  - Auto Visual: Static interaction by means of queries

**Worlds-within-Worlds**

**Tree-Map**

- Screen-filling method which uses a hierarchical partitioning of the screen into regions depending on the attribute values
- The x- and y-dimension of the screen are partitioned alternately according to the attribute values (classes)

**Tree-Map of a File System (Schneiderman)**

- Screen-filling method which uses a hierarchical partitioning of the screen into regions depending on the attribute values
- The x- and y-dimension of the screen are partitioned alternately according to the attribute values (classes)
InfoCube

- A 3-D visualization technique where hierarchical information is displayed as nested semi-transparent cubes
- The outermost cubes correspond to the top level data, while the subnodes or the lower level data are represented as smaller cubes inside the outermost cubes, and so on

Three-D Cone Trees

- 3D cone tree visualization technique works well for up to a thousand nodes or so
- First build a 2D circle tree that arranges its nodes in concentric circles centered on the root node
- Cannot avoid overlaps when projected to 2D
- Graph from Nadeau Software Consulting website: Visualize a social network data set that models the way an infection spreads from one person to the next

Visualizing Complex Data and Relations

- Visualizing non-numerical data: text and social networks
- Tag cloud: visualizing user-generated tags
  - The importance of tag is represented by font size/color
  - Besides text data, there are also methods to visualize relationships, such as visualizing social networks

Newsmap: Google News Stories in 2005

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
**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

**Data Matrix and Dissimilarity Matrix**

- **Data matrix**
  - n data points with p dimensions
  - Two modes

- **Dissimilarity matrix**
  - n data points, but registers only the distance
  - A triangular matrix
  - Single mode

**Proximity Measure for Nominal Attributes**

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- **Method 1**: Simple matching
  - \( m = \# \) of matches, \( p = \) total \# of variables
  - \( d(i, j) = \frac{p - m}{p} \)
- **Method 2**: Use a large number of binary attributes
  - creating a new binary attribute for each of the \( M \) nominal states

**Proximity Measure for Binary Attributes**

- A contingency table for binary data

- Distance measure for symmetric binary variables:
  - \[ d(i, j) = \frac{r + s}{q + r + s + t} \]
- Distance measure for asymmetric binary variables:
  - \[ d(i, j) = \frac{r + s}{q + r + s + t} \]
- Jaccard coefficient (similarity measure for asymmetric binary variables):
  - \( \text{sim}_{\text{Jaccard}}(i, j) = \frac{q}{q + r + s} \)
- Note: Jaccard coefficient is the same as "coherence":
  - \[ \text{coherence}(i, j) = \frac{\text{sup}(i, j)}{\text{sup}(i) + \text{sup}(j) - \text{sup}(i, j)} = \frac{q}{(q + r) + (q + s) - q} \]
Dissimilarity between Binary Variables

- **Example**

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Fever</th>
<th>Cough</th>
<th>Test-1</th>
<th>Test-2</th>
<th>Test-3</th>
<th>Test-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack</td>
<td>M</td>
<td>Y</td>
<td>N</td>
<td>P</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Mary</td>
<td>F</td>
<td>Y</td>
<td>N</td>
<td>P</td>
<td>N</td>
<td>P</td>
<td>N</td>
</tr>
<tr>
<td>Jim</td>
<td>M</td>
<td>Y</td>
<td>P</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N 0

\[
d(jack, mary) = \frac{0+1}{2+0+1} = 0.33
\]

\[
d(jack, jim) = \frac{1+1}{1+1+1} = 0.67
\]

\[
d(jim, mary) = \frac{1+2}{1+1+2} = 0.75
\]

Standardizing Numeric Data

- **Z-score:**
  \[
z = \frac{X - \mu}{\sigma}
\]
  - \(X\): raw score to be standardized, \(\mu\): mean of the population, \(\sigma\): standard deviation
  - the distance between the raw score and the population mean in units of the standard deviation
  - negative when the raw score is below the mean, “+” when above

An alternative way: Calculate the mean absolute deviation

\[
s_f = \frac{1}{n}(|x_{f1} - m_f| + |x_{f2} - m_f| + ... + |x_{fn} - m_f|)
\]

where

\[
m_f = \frac{1}{n}(x_{f1} + x_{f2} + ... + x_{fn})
\]

- standardized measure (z-score):
  \[
z_f = \frac{x_f - m_f}{s_f}
\]
- Using mean absolute deviation is more robust than using standard deviation

Example: Data Matrix and Dissimilarity Matrix

<table>
<thead>
<tr>
<th>Point</th>
<th>Attribute1</th>
<th>Attribute2</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>x2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>x3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>x4</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

**Dissimilarity Matrix** (with Euclidean Distance)

\[
d(i, j) = \sqrt{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + ... + |x_{ip} - x_{jp}|^h}
\]

where \(i = (x_{i1}, x_{i2}, ..., x_{ip})\) and \(j = (x_{j1}, x_{j2}, ..., x_{jp})\) are two \(p\)-dimensional data objects, and \(h\) is the order (the distance so defined is also called \(L-h\) norm)

- Properties
  - \(d(i, j) > 0\) if \(i \neq j\), and \(d(i, i) = 0\) (Positive definiteness)
  - \(d(i, j) = d(j, i)\) (Symmetry)
  - \(d(i, j) \leq d(i, k) + d(k, j)\) (Triangle Inequality)

- A distance that satisfies these properties is a metric
Special Cases of Minkowski Distance

- $h = 1$: Manhattan (city block, $L_1$ norm) distance
  
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors
  
  \[ d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}| \]

- $h = 2$: (L_2 norm) Euclidean distance
  
  \[ d(i, j) = \sqrt{\sum_{f=1}^{p} |x_{if} - x_{jf}|^2} = \sqrt{|x_{i} - x_{j}|} \]

- $h \to \infty$. “supremum” ($L_{\infty}$ norm, $L_p$ norm) distance.
  - This is the maximum difference between any component (attribute) of the vectors
  
  \[ d(i, j) = \lim_{h \to \infty} \left( \sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^\frac{1}{h} = \max_{f} |x_{if} - x_{jf}| \]

Example: Minkowski Distance

Dissimilarity Matrices

<table>
<thead>
<tr>
<th>point</th>
<th>attribute 1</th>
<th>attribute 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>x2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>x3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>x4</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

Manhattan ($L_1$)

| x1    | 3           | 0           |
| x2    | 3.61        | 1           |
| x3    | 2.24        | 5.1         |
| x4    | 4.24        | 1.05        |

Euclidean ($L_2$)

| x1    | 0           | 3           |
| x2    | 3.61        | 0           |
| x3    | 2.24        | 5.1         |
| x4    | 4.24        | 1.05        |

Supremum

| x1    | 0           | 3           |
| x2    | 3.61        | 0           |
| x3    | 2.24        | 5.1         |
| x4    | 4.24        | 1.05        |

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace $x_{if}$ by their rank $r_{if} \in \{1, \ldots, M_f \}$
  - map the range of each variable onto $[0, 1]$ by replacing $\hat{f}$th object in the $\hat{f}$th variable by
  
  \[ z_{if} = \frac{r_{if} - 1}{M_f - 1} \]
  - compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects
  
  \[ d(i, j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}} \]
  - $f$ is binary or nominal:
    \[ d_{ij}^{(f)} = 0 \text{ if } x_{if} = x_{jf} \text{, or } d_{ij}^{(f)} = 1 \text{ otherwise} \]
  - $f$ is numeric: use the normalized distance
  - $f$ is ordinal
    - Compute ranks $r_{if}$ and
    - Treat $z_{if}$ as interval-scaled

\[ z_{if} = \frac{r_{if} - 1}{M_f - 1} \]
Cosine Similarity

- A document can be represented by thousands of attributes, each recording the frequency of a particular word (such as keywords) or phrase in the document.

<table>
<thead>
<tr>
<th>Document</th>
<th>team</th>
<th>coach</th>
<th>hockey</th>
<th>baseball</th>
<th>soccer</th>
<th>penalty</th>
<th>score</th>
<th>win</th>
<th>loss</th>
<th>season</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document1</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Document2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Document3</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Document4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

- Other vector objects: gene features in micro-arrays, ...
- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If $d_1$ and $d_2$ are two vectors (e.g., term-frequency vectors), then
  $$\cos(d_1, d_2) = \frac{(d_1 \cdot d_2)}{||d_1|| ||d_2||},$$
  where $\cdot$ indicates vector dot product, $||d||$: the length of vector $d$

Example: Cosine Similarity

- $\cos(d_1, d_2) = \frac{(d_1 \cdot d_2)}{||d_1|| ||d_2||}$
- Ex: Find the similarity between documents 1 and 2.

$d_1 = (5, 0, 3, 0, 2, 0, 2, 0, 0)$
$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$

$d_1 \cdot d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+2*1+0*0+0*0 = 25$
$||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$
$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$

$\cos(d_1, d_2) = 0.94$

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.
References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain, "Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- C. Yu, et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009