

## Sample Data Sets

- Machine learning: compiled and hosted by the University of California at Irvine (UCI) <http://archive.ics.uci.edu/ml/>
- Data mining: compiled by the University of Edinburgh (School of Informatics) <http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html>
- <https://www.microsoft.com/en-us/research/project/urban-computing>
- What dataset to use in homework #1?
  - If you have to ask, use the Iris data set from UCI
- What is "interpreting data"?
  - *Make some sense out of the data and/or results, bonus up to 20%*
- What to submit to Blackboard
  - Report: descriptions, figures, tables, examples (separate file)
  - Code
  - Data if applicable (link to a well-known public dataset is sufficient)

CSC 240/440 by Prof. Jiebo Luo

3

## Code of Data Mining Algorithms

- Source codes for Frequent Pattern Mining, Clustering, *Time Series* and Web Mining algorithms implemented by Chinese Univ. of Hong Kong: <http://apprv.cse.cuhk.edu.hk/~kdd/program.html>
- FIMI workshops: Datasets and source codes for frequent itemset mining implementations: <http://fimi.cs.helsinki.fi/>
- Frequent itemset mining algorithm implementations by Bart Goethals: <http://www.adrem.ua.ac.be/~goethals/software/>
- Repository of implementations of UIUC data mining research package: IlliMine: <http://illimine.cs.uiuc.edu/>
- Weka: Weka 3 - Data Mining with Open Source Machine Learning Software in Java: <http://www.cs.waikato.ac.nz/ml/weka/>
- *Graph mining* algorithm implementations: gSpan and CloseGraph [java implementation](#)

CSC 240/440 by Prof. Jiebo Luo

7

## It's a Wild World Out There

- Known and Unknown  
*"there are known knowns;  
 there are things that we know that we know.  
 We also know there are known unknowns; that  
 is to say we know there are some things we do  
 not know.  
 But there are also unknown unknowns, the ones  
 we don't know we don't know."*

- Get Used to It



9

## Data Mining: Concepts and Techniques

### — Chapter 2 —

Jiawei Han, Micheline Kamber, and Jian Pei

10

## 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

11

## 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:

| ID | Items                     |
|----|---------------------------|
| 1  | Bread, Coke, Milk         |
| 2  | Beer, Bread               |
| 3  | Beer, Coke, Diaper, Milk  |
| 4  | Beer, Bread, Diaper, Milk |
| 5  | Coke, Diaper, Milk        |

12

## Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale (remember dots?)
- Distribution
  - Centrality and dispersion

13

## 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.

14

14

## 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

15

## 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*



17

## 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



16

16

## 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

18

18

## 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

19

## 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

20

## Measuring the Central Tendency

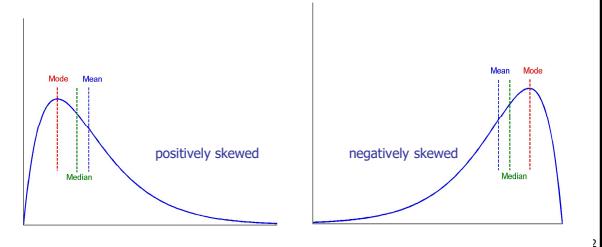
- Mean (algebraic measure) (sample vs. population):  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$      $\mu = \frac{\sum x}{N}$   
Note:  $n$  is sample size and  $N$  is population size.
- Weighted arithmetic mean:
- Trimmed mean: chopping extreme values  $\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise

| age    | frequency |
|--------|-----------|
| 1–5    | 200       |
| 6–15   | 450       |
| 16–20  | 300       |
| 21–50  | 1500      |
| 51–80  | 700       |
| 81–110 | 44        |
- Mode  $median = L_1 + (\frac{n/2 - (\sum freq)}{freq_{median}})width$ 
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula (for unimodal data that is moderately skewed):  $mean - mode = 3 \times (mean - median)$

21

## Symmetric vs. Skewed

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



22

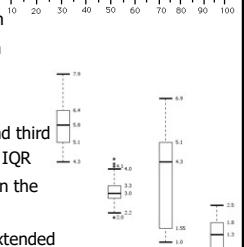
## Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
    - **Quartiles:**  $Q_1$  (25<sup>th</sup> percentile),  $Q_3$  (75<sup>th</sup> 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  $1.5 \times IQR$
  - Variance and standard deviation (sample STD:  $s$ , population STD:  $\sigma$ )
    - **Variance:** (algebraic, scalable computation)
- $$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} [\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2] \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$
- **Standard deviation**  $s$  (or  $\sigma$ ) is the square root of variance  $s^2$  (or  $\sigma^2$ )

23

## Boxplot Analysis

- **Five-number summary** of a distribution
  - Minimum,  $Q_1$ , **Median**,  $Q_3$ , 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 (e.g.,  $1.5 \times IQR$ , plotted individually)



24

## Project Proposal

### For CSC 440

- The 2-page proposal should summarize the papers and outline your project plan (e.g., problem statement, data acquisition, algorithm choices/ideas, experiment design, and performance evaluation). It would naturally be added to your final project report as the introduction after being expanded to include additional information if needed.
- The proposal may contain a 1-2 figures.

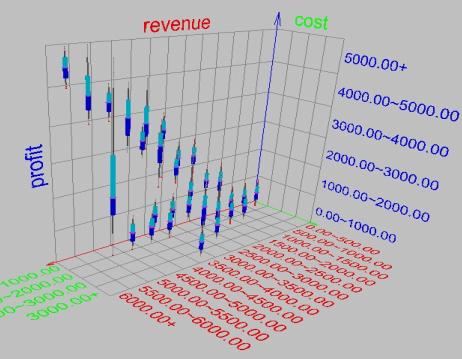
25

### For CSC 240

- Introduction: Background, problem statement, and problem importance
- Methods: Major algorithms (hint: not 1, not 2, 5+) used to solve the problem
- Results: Major results (from literature) and their implications
- Critical Evaluation: Comments on the methods used and the major results. Advantages and disadvantages of the methods used. Discussion on alternative methods that may be considered for future work.

25

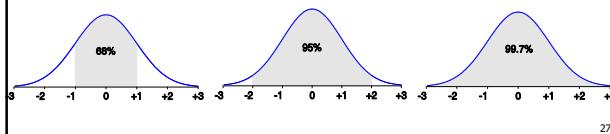
## Visualization of Data Dispersion: 3-D Boxplots



26

## 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



27

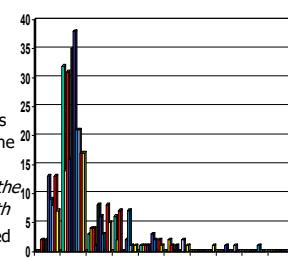
## Graphic Displays of Basic Statistical Descriptions

- Boxplot:** graphic display of five-number summary
- Histogram:** x-axis are values, y-axis repres. 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

28

## Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs 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



29

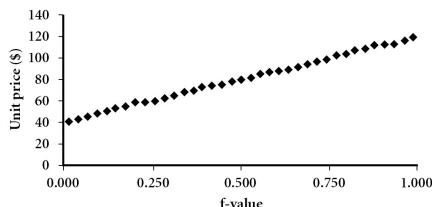
## Histograms Often Tell More than Boxplots

- 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

30

## Quantile Plot

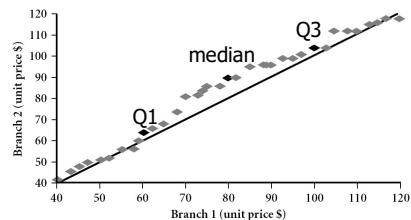
- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data  $x_i$  data sorted in increasing order,  $f_i$  indicates that approximately  $100 f\%$  of the data are below or equal to the value  $x_i$



31

## Quantile-Quantile (Q-Q) Plot

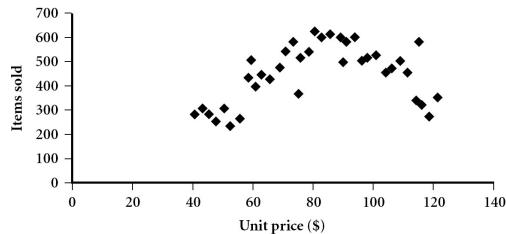
- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be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



32

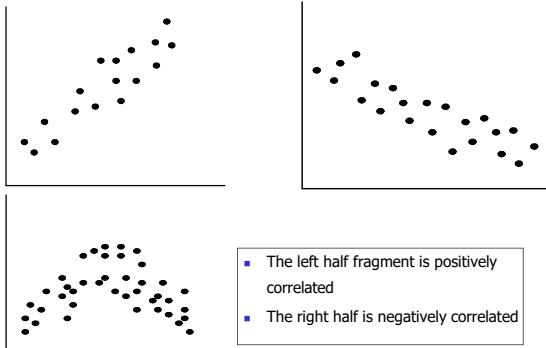
## Scatter plot

- 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



33

## Positively and Negatively Correlated Data



34

## Uncorrelated Data



35

## 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

36

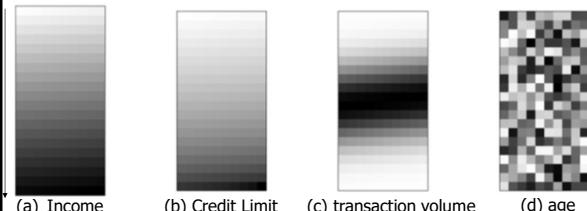
## Data Visualization

- 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

37

## Pixel-Oriented Visualization Techniques

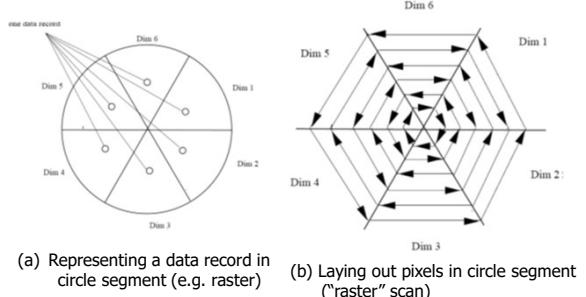
- For a data set of  $m$  dimensions, create  $m$  windows on the screen, one for each dimension
- The  $m$  dimension values of a record are mapped (raster scan order) to  $m$  pixels at the corresponding (2D) positions in the windows
- The colors of the pixels reflect the corresponding values



38

## 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



39

## Example



40

## Example

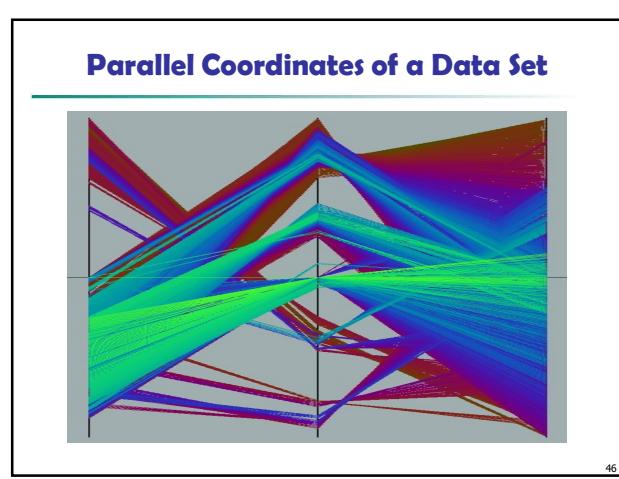
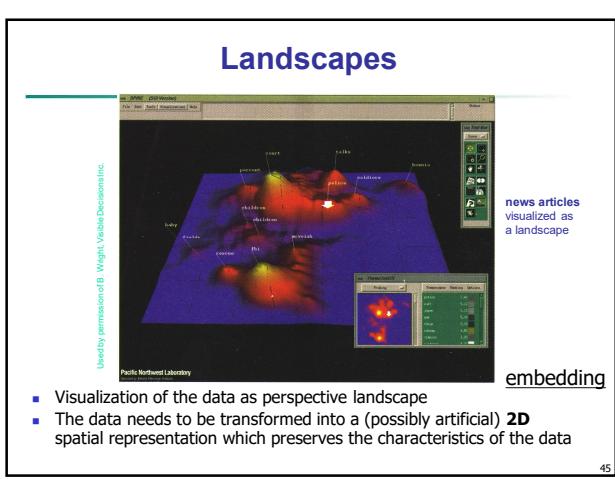
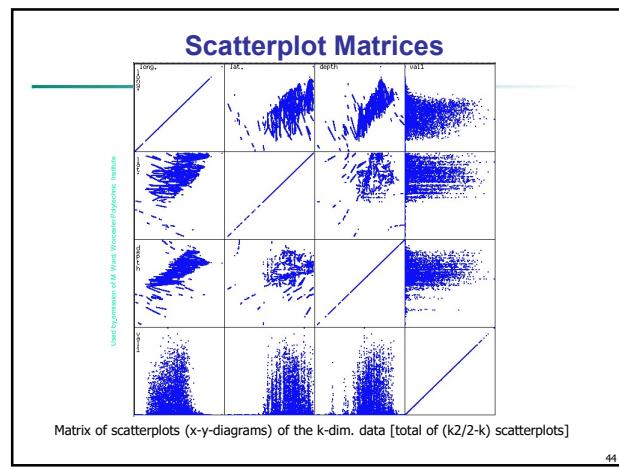
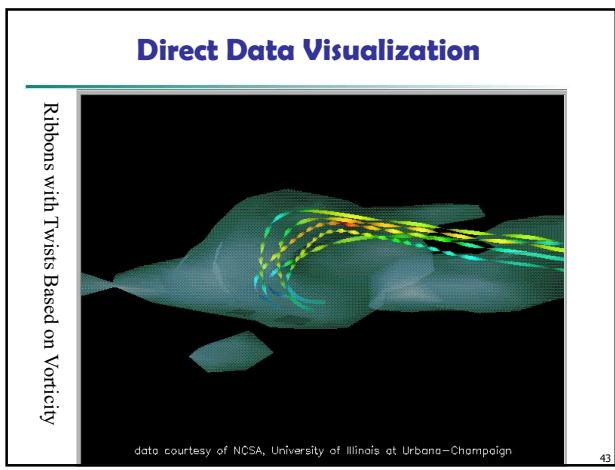


41

## Geometric Projection Visualization Techniques

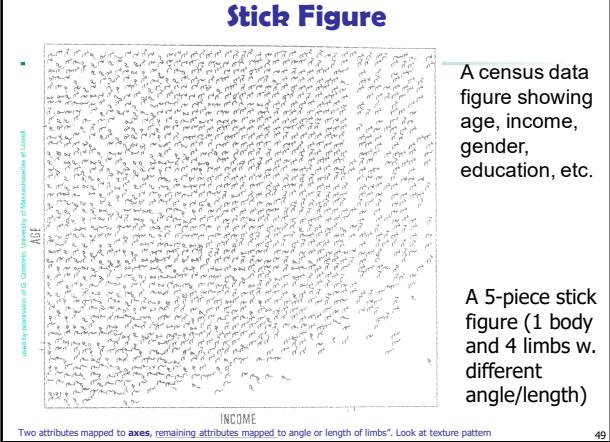
- Visualization of geometric transformations and projections of the data
- Methods
  - Direct visualization
  - Scatterplot and scatterplot matrices
  - Landscapes
  - Projection pursuit technique: Help users find meaningful projections of multidimensional data
  - Prosection views
  - Hyperslice
  -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

42



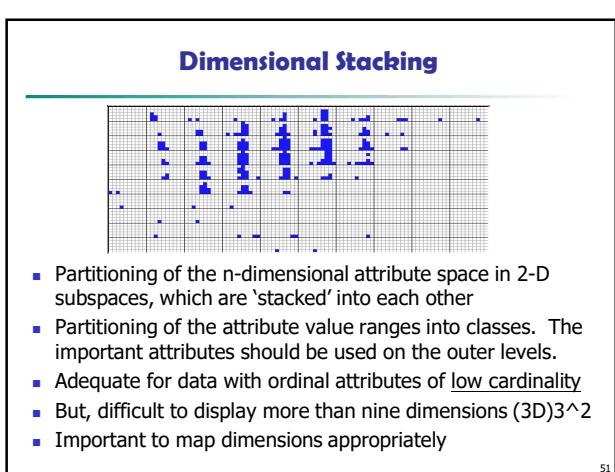
- ### Icon-Based Visualization Techniques
- Visualization of the data values as features of icons
  - Typical visualization methods
    - Chernoff Faces
    - Stick Figures
  - General techniques
    - Shape coding: Use shape to represent certain information encoding
    - Color icons: Use color icons to encode more information
    - Tile bars: Use small icons to represent the relevant feature vectors in document retrieval
- 47

- ### Chernoff Faces
- A way to display 2+ 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)
  - REFERENCE: Gonick, L. and Smith, W. *The Cartoon Guide to Statistics*. New York: Harper Perennial, p. 212, 1993
  - Weisstein, Eric W. "Chernoff Face." From *MathWorld--A Wolfram Web Resource*. [mathworld.wolfram.com/ChernoffFace.html](http://mathworld.wolfram.com/ChernoffFace.html)
- Imagine a big crowd, visualize clusters as "tribes"
- 48

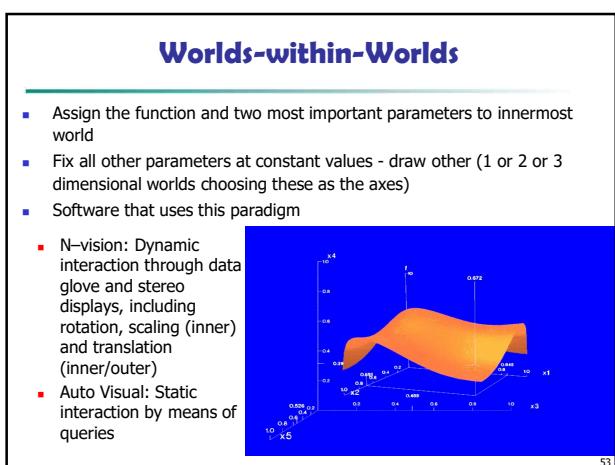
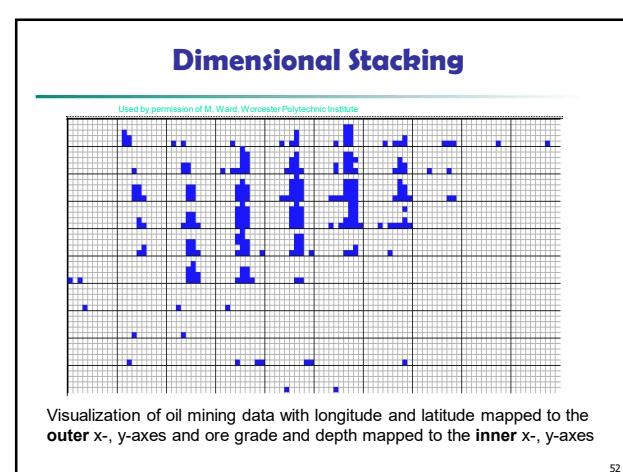


49

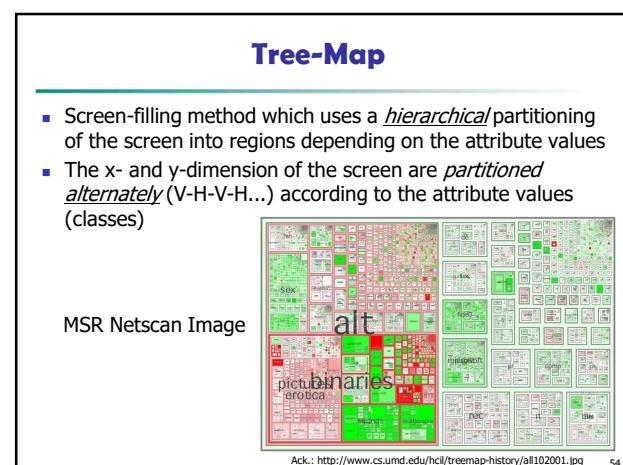
- ## Hierarchical Visualization Techniques
- Visualization of the data using a hierarchical partitioning into subspaces
  - Methods
    - Dimensional Stacking
    - Worlds-within-Worlds
    - Tree-Map
    - Cone Trees
    - InfoCube
- 50



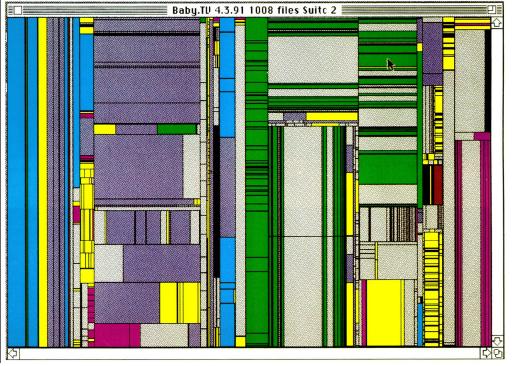
51



53



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



55

## 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



56

## 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
- G. Robertson, J. Mackinlay, S. Card. "Cone Trees: Animated 3D Visualizations of Hierarchical Information", *ACM SIGCHI'91*
- 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



Ack.: <http://nadeausoftware.com/circles.html> 57

57

## 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



Data Cloud vs Tag Cloud

Newsmap: Google News Stories in 2005

58

## 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

59

## 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
  - E.g.,  $\text{sim}(i, j) = 1 - d(i, j)$  **Sim = 1 – Diff? Not always!!!**

60

## Data Matrix and Dissimilarity Matrix

### Data matrix

- n data points with p dimensions ( $n \times p$ )
- Two modes (attributes)

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

### Dissimilarity matrix

- n data points, but registers only the distance ( $n \times n$ )
- A triangular matrix (why?)
- Single mode (attribute)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Confused? Wait until confusion matrix?

## 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

62

61

## Proximity Measure for Binary Attributes

- A contingency table for binary data

|            |         | Object $j$ |     | sum     |
|------------|---------|------------|-----|---------|
| Object $i$ | 1       | 0          | sum |         |
|            | 0       | $q$        | $r$ | $q + r$ |
| sum        | $q + s$ | $r + t$    | $p$ | $s + t$ |

- Distance measure for symmetric binary variables:

$$d(i, j) = \frac{r + s}{q + r + s + t} \text{ easy}$$

- Distance measure for *asymmetric* binary variables:

$$d(i, j) = \frac{r + s}{q + r + s} \text{ tricky (t ignored)}$$

- 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":

$$\text{coherence}(i, j) = \frac{\text{sup}(i, j)}{\text{sup}(i) + \text{sup}(j) - \text{sup}(i, j)} = \frac{q}{(q + r) + (q + s) - q}$$

63

## 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*

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

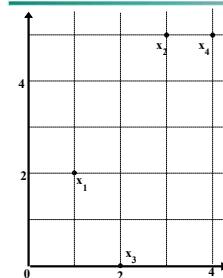
$$\text{where } m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf})$$

$$\text{standardized measure (z-score): } z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using mean absolute deviation is *more robust* than using standard deviation (why?)

64

## Example: Data Matrix and Dissimilarity Matrix



Data Matrix

| point | attribute1 | attribute2 |
|-------|------------|------------|
| x1    | 1          | 2          |
| x2    | 3          | 5          |
| x3    | 2          | 0          |
| x4    | 4          | 5          |

Dissimilarity Matrix  
(with Euclidean Distance)

|    | x1   | x2  | x3  | x4   |
|----|------|-----|-----|------|
| x1 | 0    |     |     |      |
| x2 | 3.61 | 0   |     |      |
| x3 |      | 5.1 | 5.1 | 0    |
| x4 | 4.24 |     | 1   | 5.39 |
|    |      |     |     | 0    |

65

66

## Distance on Numeric Data: Minkowski Distance

- Minkowski distance:** A popular distance measure

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

where  $i = (x_{i1}, x_{i2}, \dots, x_{ip})$  and  $j = (x_{j1}, x_{j2}, \dots, x_{jp})$  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**

67

## 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

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

- $h = 2$ : ( $L_2$  norm) **Euclidean** distance

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2}$$

- $h \rightarrow \infty$ : **"supremum"** ( $L_{\max}$  norm,  $L_{\infty}$  norm) distance.

- This is the **maximum difference** between any component (attribute) of the vectors

$$d(i, j) = \lim_{h \rightarrow \infty} \left( \sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

68

68

## Example: Minkowski Distance

| point | attribute 1 | attribute 2 | Dissimilarity Matrices              |      |     |      |
|-------|-------------|-------------|-------------------------------------|------|-----|------|
| x1    | 1           | 2           | <b>Manhattan (<math>L_1</math>)</b> |      |     |      |
| x2    | 3           | 5           | L1                                  | x1   | x2  | x3   |
| x3    | 2           | 0           | x1                                  | 0    |     |      |
| x4    | 4           | 5           | x2                                  | 5    | 0   |      |
|       |             |             | x3                                  | 3    | 6   | 0    |
|       |             |             | x4                                  | 6    | 1   | 7    |
|       |             |             |                                     |      |     | 0    |
|       |             |             | <b>Euclidean (<math>L_2</math>)</b> |      |     |      |
|       |             |             | L2                                  | x1   | x2  | x3   |
|       |             |             | x1                                  | 0    |     |      |
|       |             |             | x2                                  | 3.61 | 0   |      |
|       |             |             | x3                                  | 2.24 | 5.1 | 0    |
|       |             |             | x4                                  | 4.24 | 1   | 5.39 |
|       |             |             |                                     |      |     | 0    |
|       |             |             | <b>Supremum</b>                     |      |     |      |
|       |             |             | L $\infty$                          | x1   | x2  | x3   |
|       |             |             | x1                                  | 0    |     |      |
|       |             |             | x2                                  | 3    | 0   |      |
|       |             |             | x3                                  | 2    | 5   | 0    |
|       |             |             | x4                                  | 3    | 1   | 5    |
|       |             |             |                                     |      |     | 0    |

69

## 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  $r_{if} \in \{1, \dots, M_f\}$
  - map the range of each variable onto  $[0, 1]$  by replacing  $f$ -th object in the  $f$ -th variable by
- $$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$
- compute the dissimilarity using methods for interval-scaled (numeric) variables

70

## 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 into a single dissimilarity matrix

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

- $f$  is binary or nominal:
  - $d_{ij}^{(f)} = 0$  if  $x_{if} = x_{jf}$ , or  $d_{ij}^{(f)} = 1$  otherwise
- $f$  is numeric: use the normalized distance
- $f$  is ordinal
  - Compute ranks  $r_{if}$  and  $r_{jf}$
  - Treat  $r_{if}$  as interval-scaled

71

## 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.

| Document  | team | coach | hockey | baseball | soccer | penalty | score | win | loss | season |
|-----------|------|-------|--------|----------|--------|---------|-------|-----|------|--------|
| Document1 | 5    | 0     | 3      | 0        | 2      | 0       | 0     | 2   | 0    | 0      |
| Document2 | 3    | 0     | 2      | 0        | 1      | 1       | 0     | 1   | 0    | 1      |
| Document3 | 0    | 7     | 0      | 2        | 1      | 0       | 0     | 3   | 0    | 0      |
| Document4 | 0    | 1     | 0      | 0        | 1      | 2       | 2     | 0   | 3    | 0      |

- 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

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

where  $\bullet$  indicates vector dot product,  $||d||$ : the length of vector  $d$

72

71

11

## Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| ||d_2||$ ,  
where  $\bullet$  indicates vector dot product,  $||d|$ : the length of vector  $d$
  - Ex: Find the **similarity** between documents 1 and 2.
- $d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$   
 $d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$
- $$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+2*1+0*0+0*1 = 25$$
- $$||d_1|| = (5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$$
- $$||d_2|| = (3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2)^{0.5} = (17)^{0.5} = 4.12$$
- $$\cos(d_1, d_2) = 0.94$$

73

## 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

74

74

## 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

75

## References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- 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
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu , et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009

76

76

## Data about Changing Times



77

## Pictures about Times



78

78

## Pictures about Times

- Now and Then



79

## Big Data News

- "Dropbox Acquires KBVT's Computer Vision Geniuses To Mine Its Photos"

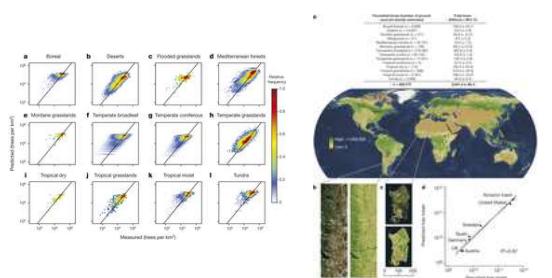
- Dropbox or any other company could be sitting on a gold mine if they can figure out how to squeeze more strategic value out of photos or offer added benefits to users who store their treasured moments with them.



80

## How many trees on earth?

Mapping tree density at a global scale  
Nature 520, 201–205, (10 September 2015)

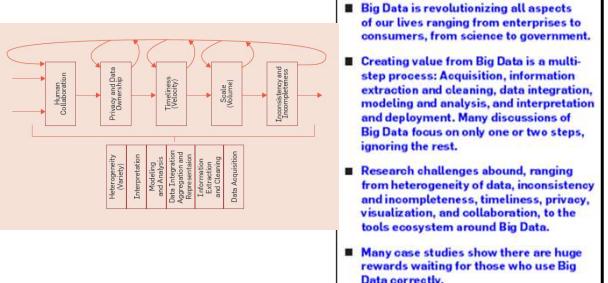


81

## Position Paper

- Big Data and Its Technical Challenges

- Communications of the ACM, Vol. 57 No. 7, Pages 86-94



- Big Data** is revolutionizing all aspects of our lives ranging from enterprises to consumers, from science to government.
- Creating value from **Big Data** is a multi-step process: Acquisition, information extraction and cleaning, data integration, modeling and analysis, and interpretation and deployment. Many discussions of **Big Data** focus on only one or two steps, ignoring the rest.
- Research challenges abound, ranging from heterogeneity of data, inconsistency and incompleteness, timeliness, privacy, visualization, and collaboration, to the tools ecosystem around **Big Data**.
- Many case studies show there are huge rewards waiting for those who use **Big Data** correctly.

82