Chapter 2.
Getting to Know Your Data

Meng Jiang

CS412 Summer 2017:

Introduction to Data Mining

Chapter 2. Getting to Know Your Data

- **Data Objects and Attribute Types**
- Basic Statistical Descriptions (Assign. 1)
- Data Visualization
- Measuring Data Similarity and Dissimilarity

Types of Data Sets: (1) Record Data

- Relational records in relational tables: highly structured
- Data matrix: numerical matrix
- Transaction data
- Document data: Term-frequency matrix of text documents

	PLAYER	TEAM	AGE	GP	W	L	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	DD2	TD3	+/-
1	Russell Westbrook	OKC	28	81	46	35	34.6	31.6	10.2	24.0	42.5	2.5	7.2	34.3	8.8	10.4	84.5	1.7	9.0	10.7	10.4	5.4	1.6	0.4	2.3	62	42	3.1
2	James Harden	HOU	27	81	54	27	36.4	29.1	8.3	18.9	44.0	3.2	9.3	34.7	9.2	10.9	84.7	1.2	7.0	8.1	11.2	5.7	1.5	0.5	2.7	64	22	5.2
3	Isaiah Thomas	BOS	28	76	51	25	33.8	28.9	9.0	19.4	46.3	3.2	8.5	37.9	7.8	8.5	90.9	0.6	2.1	2.7	5.9	2.8	0.9	0.2	2.2	5	0	3.6
4	Anthony Davis	NOP	24	75	31	44	36.1	28.0	10.3	20.3	50.5	0.5	1.8	29.9	6.9	8.6	80.2	2.3	9.5	11.8	2.1	2.4	1.3	2.2	2.2	49	0	0.7
5	DeMar DeRozan	TOR	27	74	47	27	35.4	27.3	9.7	20.9	46.7	0.4	1.7	26.6	7.4	8.7	84.2	0.9	4.3	5.2	3.9	2.4	1.1	0.2	1.8	5	0	2.0
6	Damian Lillard	POR	26	75	38	37	35.9	27.0	8.8	19.8	44.4	2.9	7.7	37.0	6.5	7.3	89.5	0.6	4.3	4.9	5.9	2.6	0.9	0.3	2.0	11	0	1.1
7	DeMarcus Cousins	NOP	26	72	30	42	34.2	27.0	9.0	19.9	45.2	1.8	5.0	36.1	7.2	9.3	77.2	2.1	8.9	11.0	4.6	3.7	1.4	1.3	3.9	46	2	-0.3
8	LeBron James	CLE	32	74	51	23	37.8	26.4	9.9	18.2	54.8	1.7	4.6	36.3	4.8	7.2	67.4	1.3	7.3	8.6	8.7	4.1	1.2	0.6	1.8	42	13	6.5
9	Kawhi Leonard	SAS	25	74	54	20	33.4	25.5	8.6	17.7	48.5	2.0	5.2	38.0	6.3	7.2	88.0	1.1	4.7	5.8	3.5	2.1	1.8	0.7	1.6	9	0	5.9
10	Stephen Curry	GSW	29	79	65	14	33.4	25.3	8.5	18.3	46.8	4.1	10.0	41.1	4.1	4.6	89.8	0.8	3.7	4.5	6.6	3.0	1.8	0.2	2.3	9	0	12.8

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
- Types:
  - Nominal (e.g., red, blue)
  - Binary (e.g., {true, false})
  - Ordinal (e.g., {freshman, sophomore, junior, senior})
  - Numeric: quantitative
    - Interval-scaled: 100°C is interval scales
    - Ratio-scaled: 100°K is ratio scaled since it is twice as high as 50 °K

Attribute Types

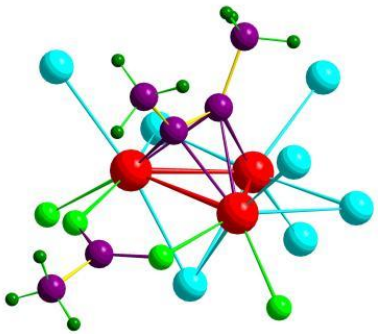
- **Nominal:** categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - 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

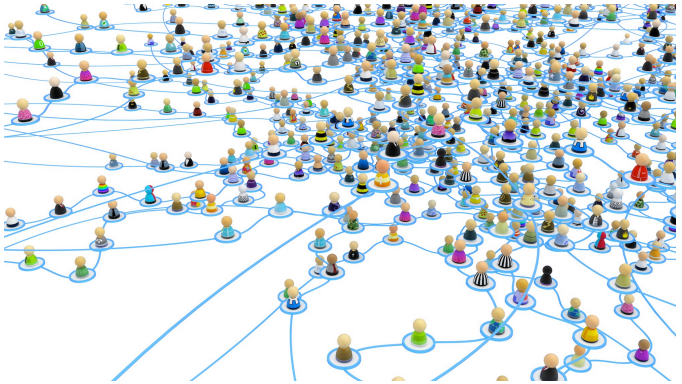
Discrete vs Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countable 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

Types of Data Sets: (2) Graphs and Networks

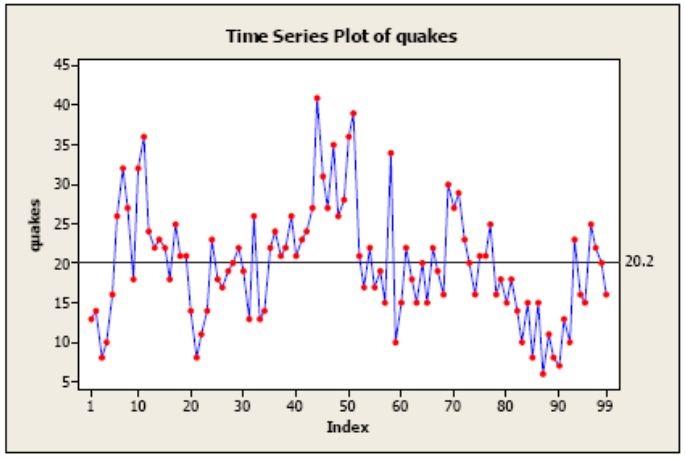
- Transportation networks
- World Wide Web
- Molecular structures
- Social or information networks





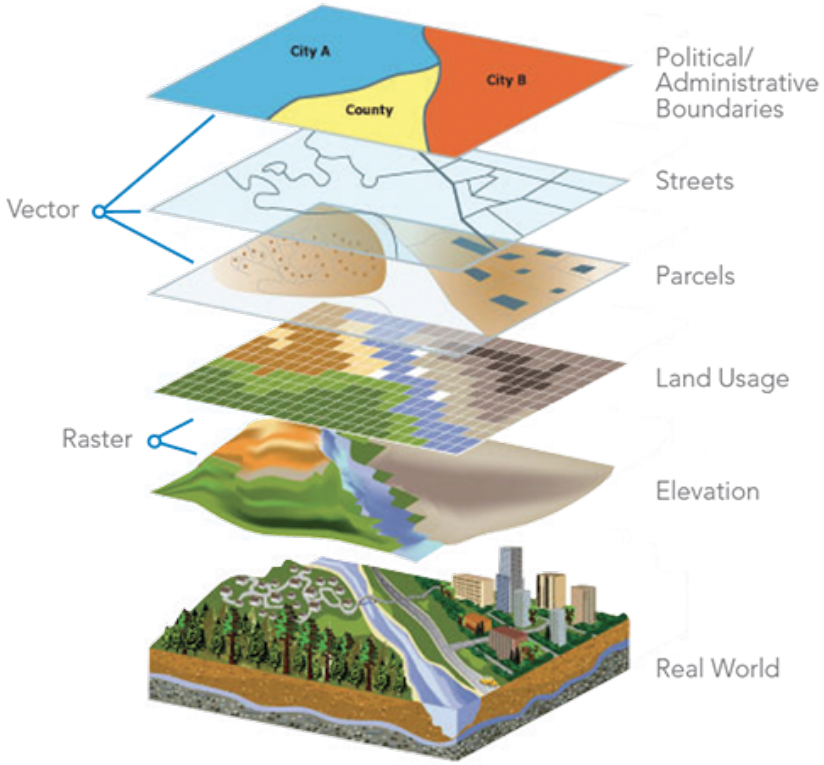
Types of Data Sets: (3) Ordered Data

- Video data: sequence of images
- Temporal data: time-series
- Sequential Data: transaction sequences
- Genetic sequence data



Other Types of Data Sets

- Spatial data
- Image and multimedia data



Characterisitcs of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion

Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- **Basic Statistical Descriptions (Assign. 1)**
- Data Visualization
- Measuring Data Similarity and Dissimilarity

Basic Statistical Descriptions of Data

- Motivation: to better understand the data
- Data dispersion characteristics
  - Median, max, min, quantiles, outliers, variance, …
- 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: (1) Mean

- Mean (algebraic measure) (sample vs. population):
  - Note: n is sample size and N is population size.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

- Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Trimmed mean: Chopping extreme values (e.g., Olympics gymnastics score computation)

Measuring the Central Tendency: (2) Median

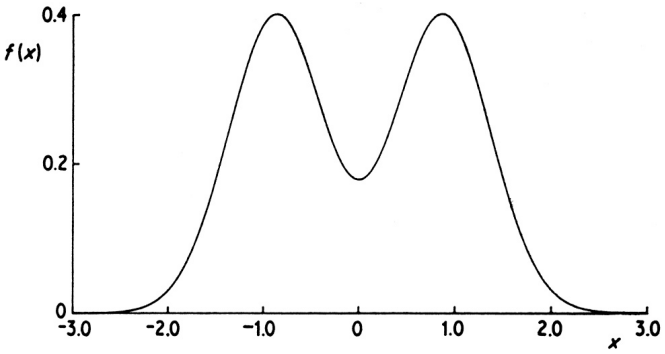
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
- Estimated by interpolation (for grouped data):

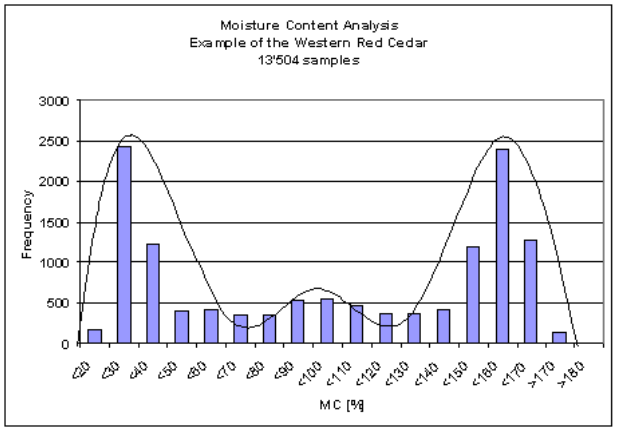
age	$frequency$
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

$$median = L_1 + (\frac{n/2 - (\sum freq)_l}{freq_{median}})width$$

Measuring the Central Tendency: (3) Mode

- Mode: Value that occurs most frequently in the data
- Multi-modal
  - Bimodal
  - Trimodal

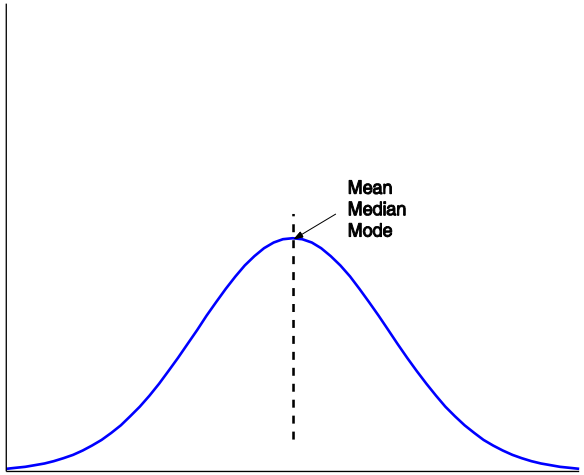


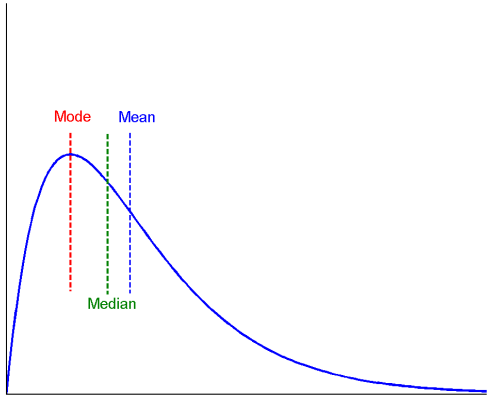


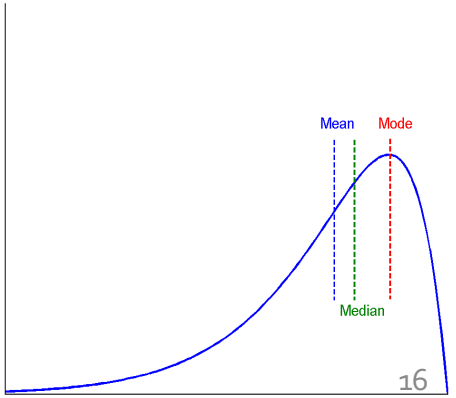
Symmetric data

Positively skewed data

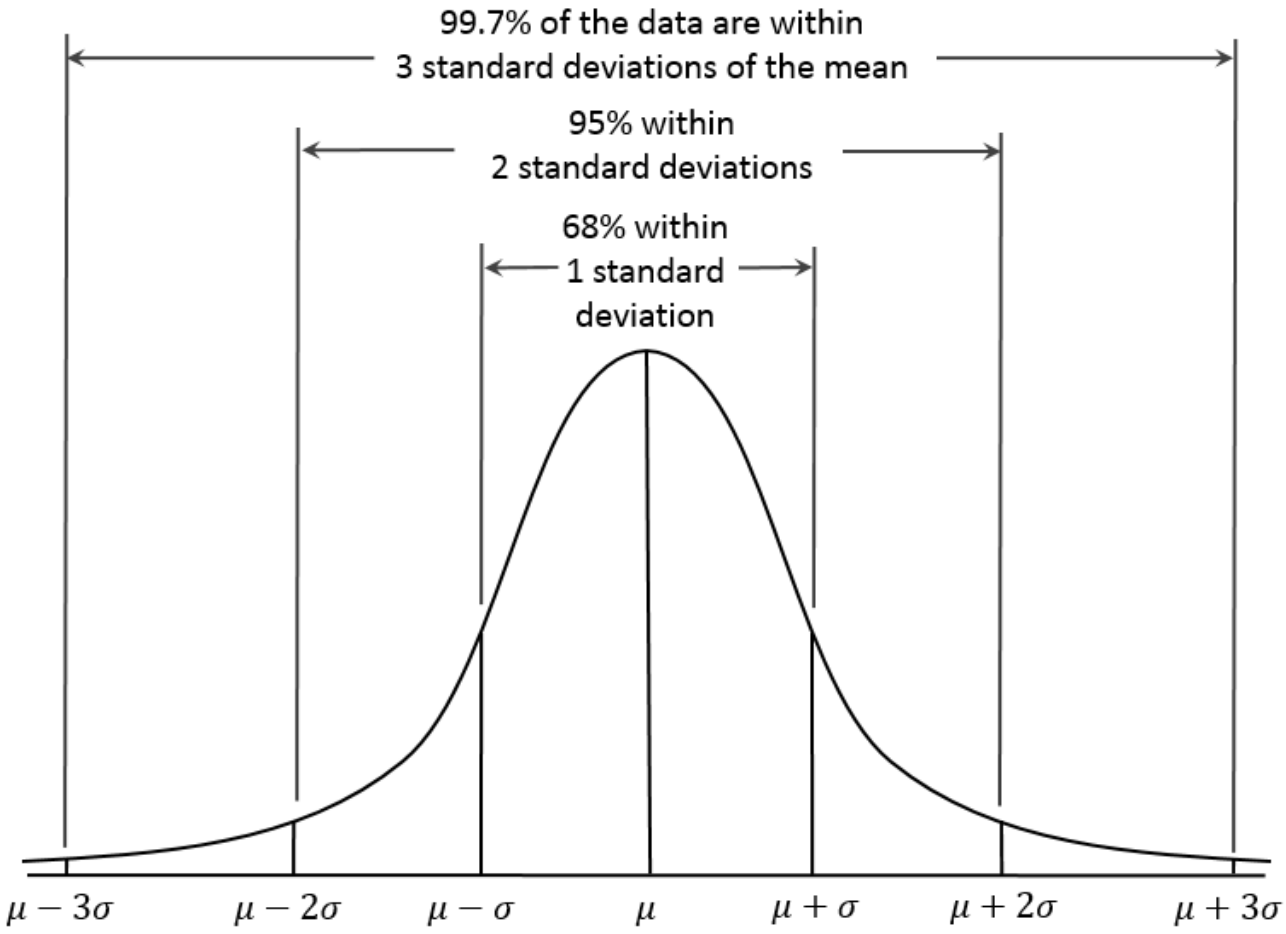
Negatively skewed data







Properties of Normal Distribution Curve



Measures Data Distribution: Variance and Standard Deviation

- Variance and standard deviation (sample: s, population: σ)
  - Variance: (algebraic, scalable computation)
    - Q: Can you compute it incrementally and efficiently?

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

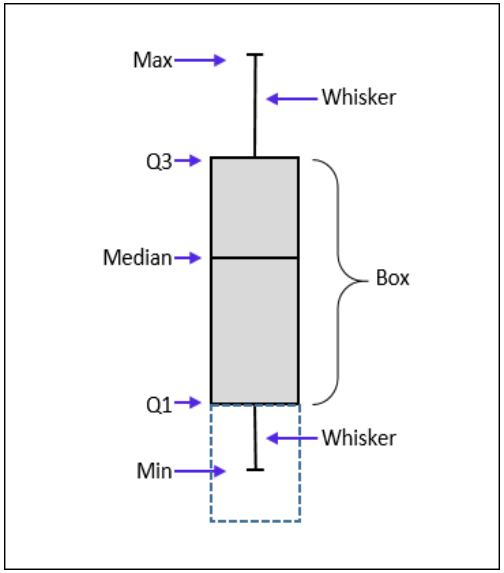
- Standard deviation s (or σ) is the square root of variance s2 (or σ2)

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 univariant 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

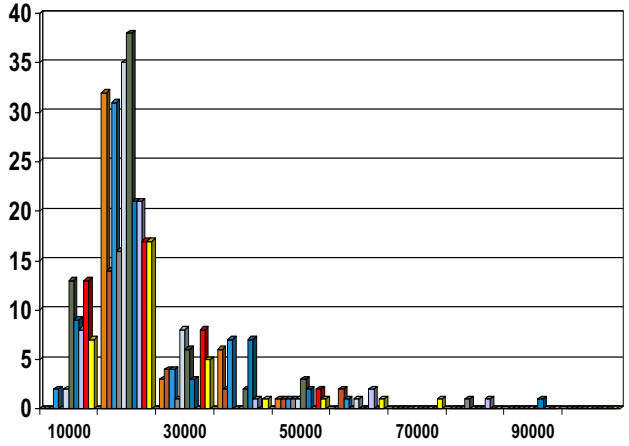
Measuring the Dispersion of Data: (1) Quartiles & Boxplots

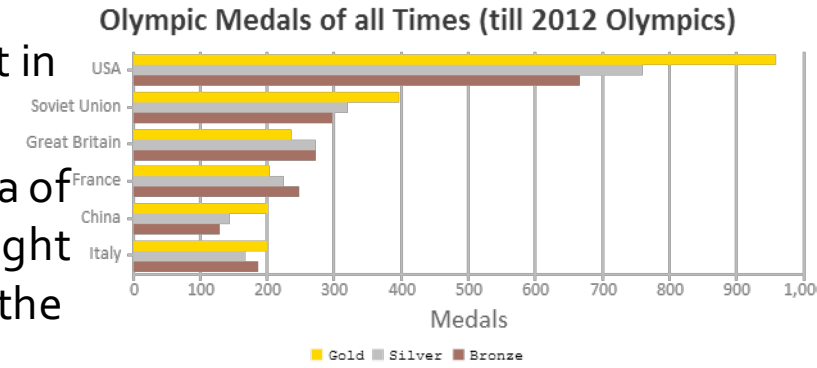
- **Quartiles**: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
- **Inter-quartile range**: $IQR = Q_3 - Q_1$
- **Five number summary**: min, Q_1, median, Q_3, max
- **Boxplot**: Data is represented with a box
  - Q_1, Q_3, IQR: The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - Median (Q_2) 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
    - Outlier: usually, a value higher/lower than 1.5 x IQR



(2) Histogram Analysis

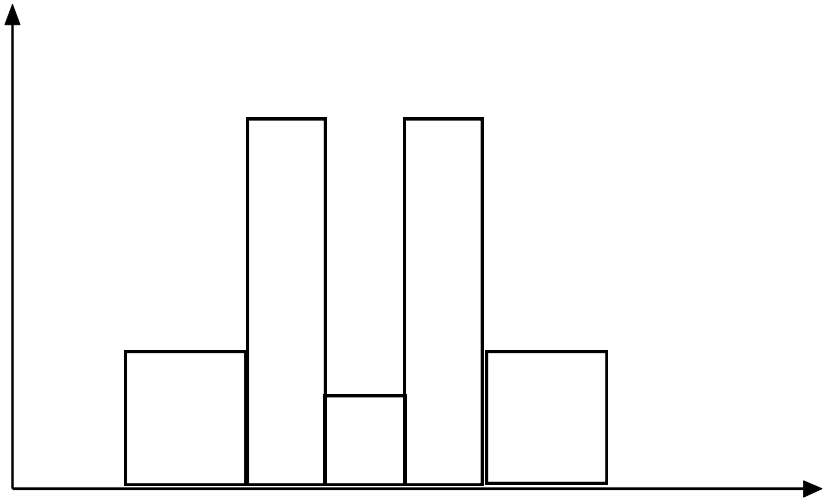
- Histogram: Graph display of **tabulated frequencies**, shown as bars
- Between histograms and bar charts
  - Histograms are used to show distributions of variables while bar charts are used to compare variables
  - Histograms plot **binned quantitative data** while bar charts plot **categorical data**
  - Bars can be reordered in bar charts but not in histograms
  - 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

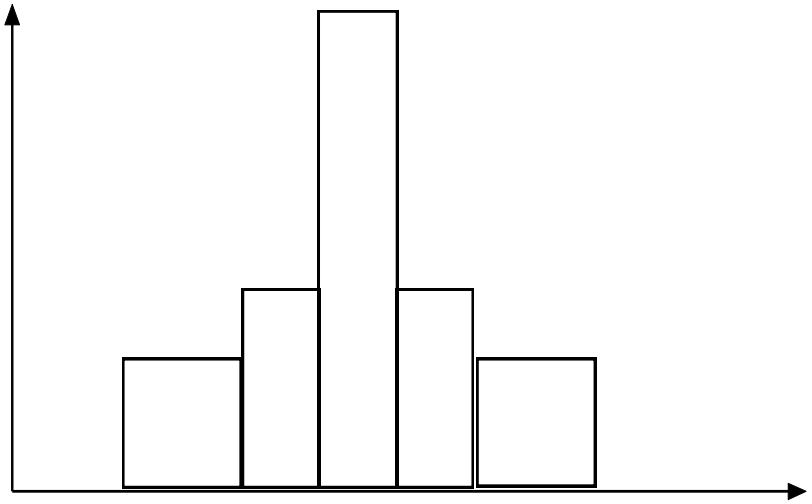




Histograms Often Tell More than Boxplots

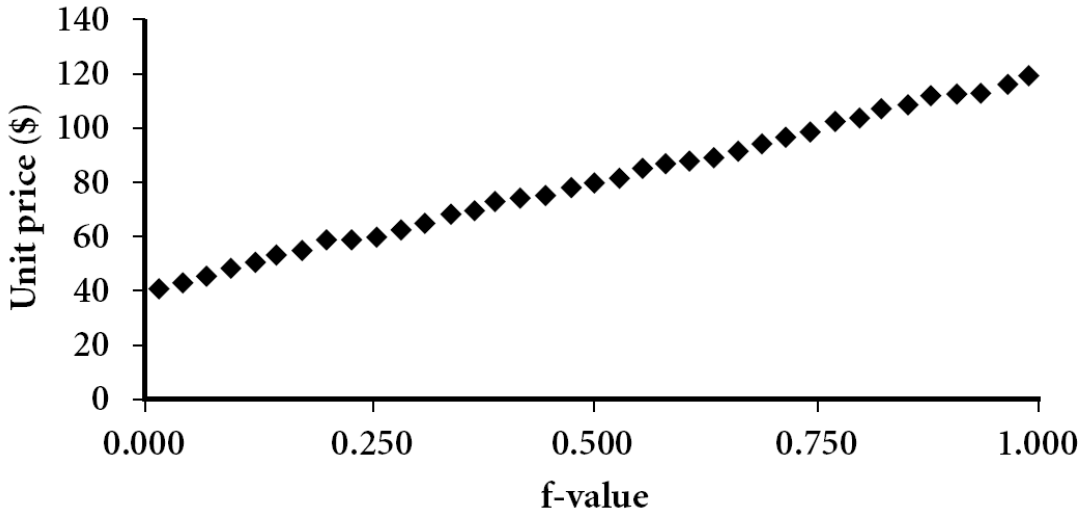
- The two histograms may have the same boxplot
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions





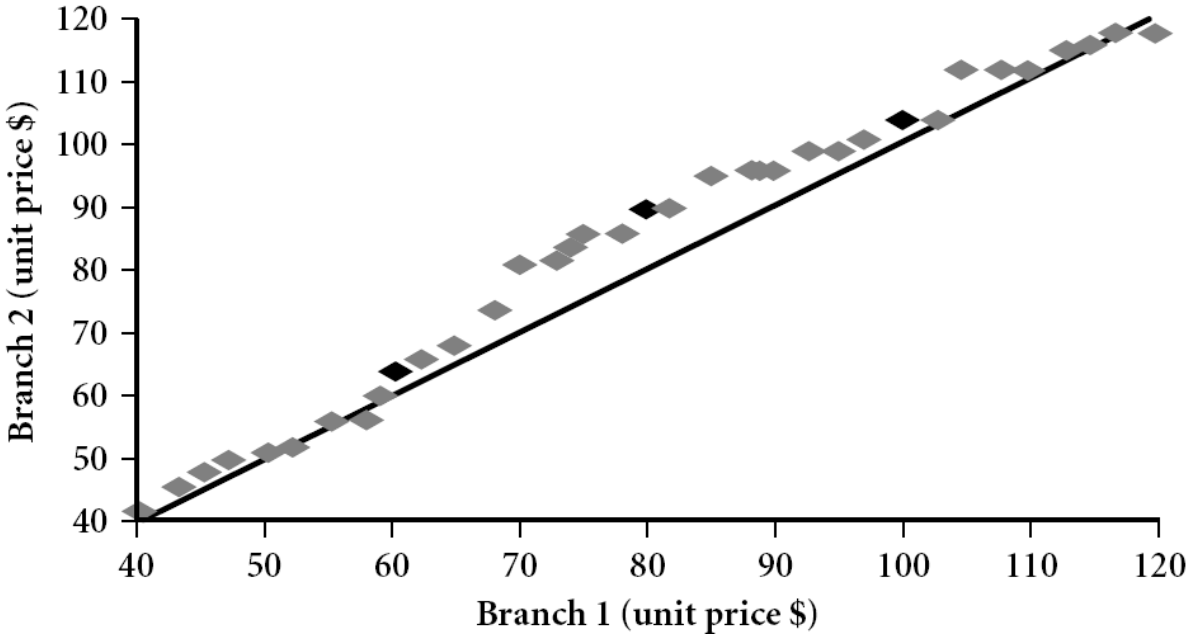
(3) Quantile Plot

- 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_i% of the data are below or equal to the value x_i



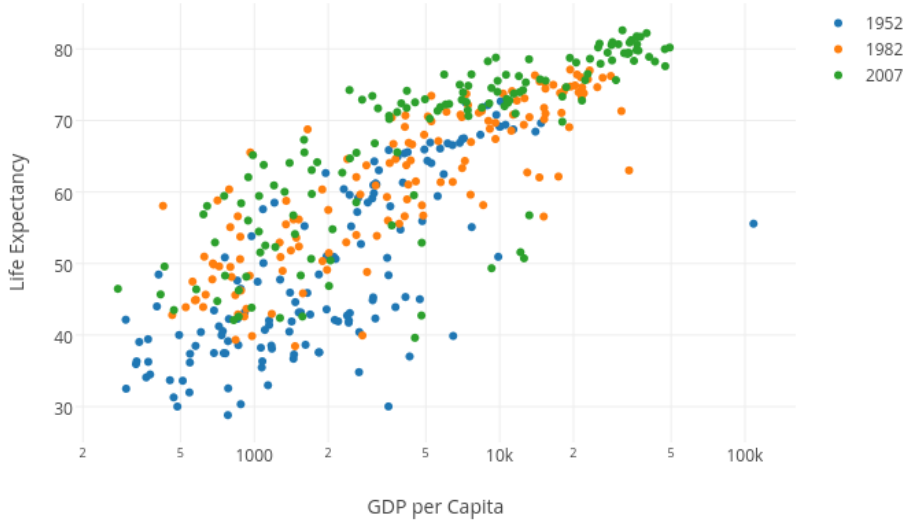
(4) Quantile-Quantile (Q-Q) Plot

- 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

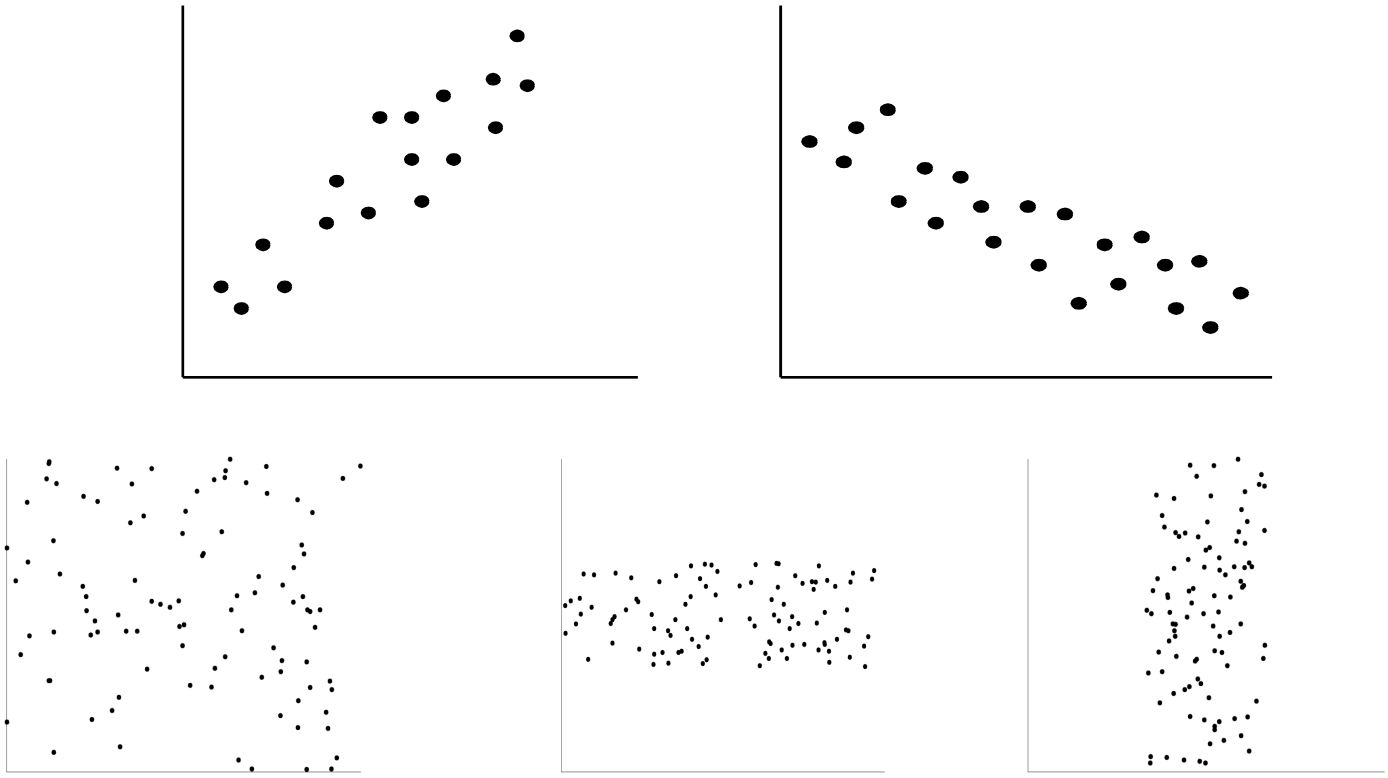


(5) Scatter plot

- 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, Negatively Correlated, and Uncorrelated Data



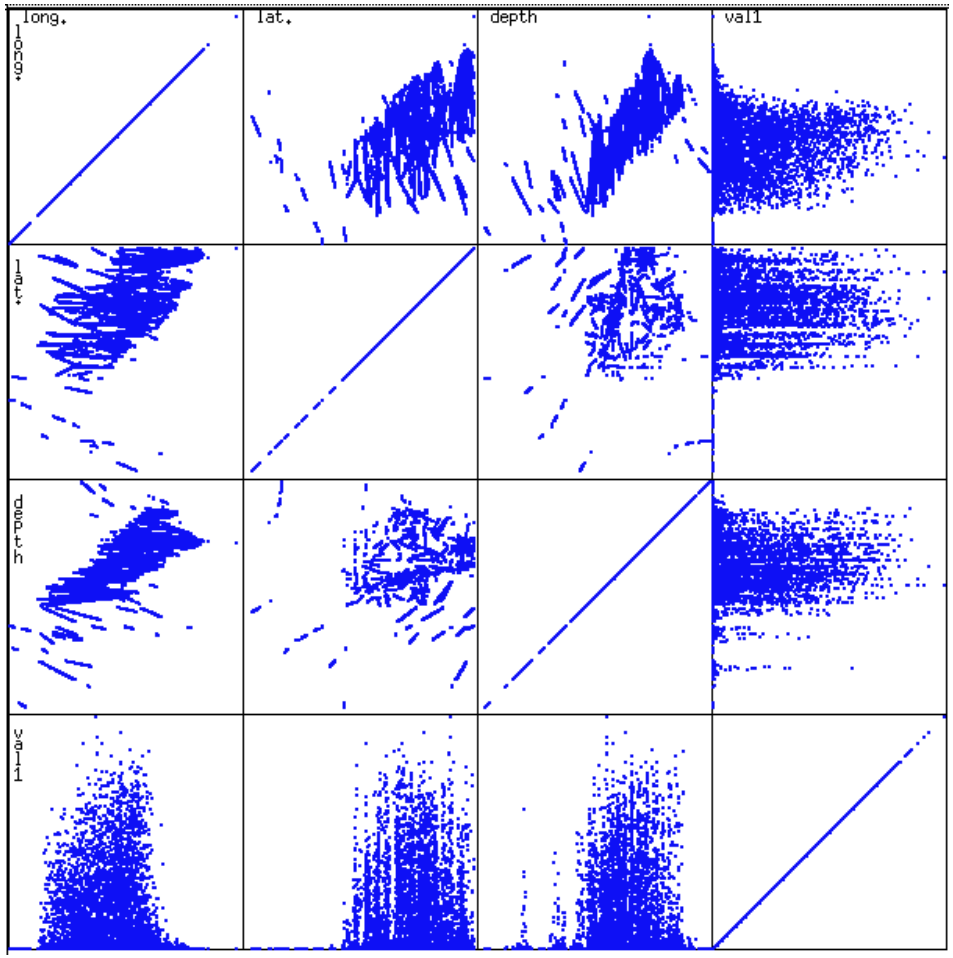
Announcements: Assign. 1 and 4th Credit Project

- CS412: The First Homework
  - Assignment #1 is ready and is distributed today!
  - Please check lecture page linking to the assignment #1
- Project for the 4th Credit
  - Reading KDD'15-'16 proceedings (430 papers)
  - Manually structuring paper **full text** into a "data object" (paper ID) – "attribute" (**phrases of research problem, methodology, technique, datasets…**) table
  - Clustering analysis based on similarity/dissimilarity
  - Multi-faceted clustering: vs. author, affiliation, etc.
  - A PDF report and a TabSV (.tsv) table: individual proj.

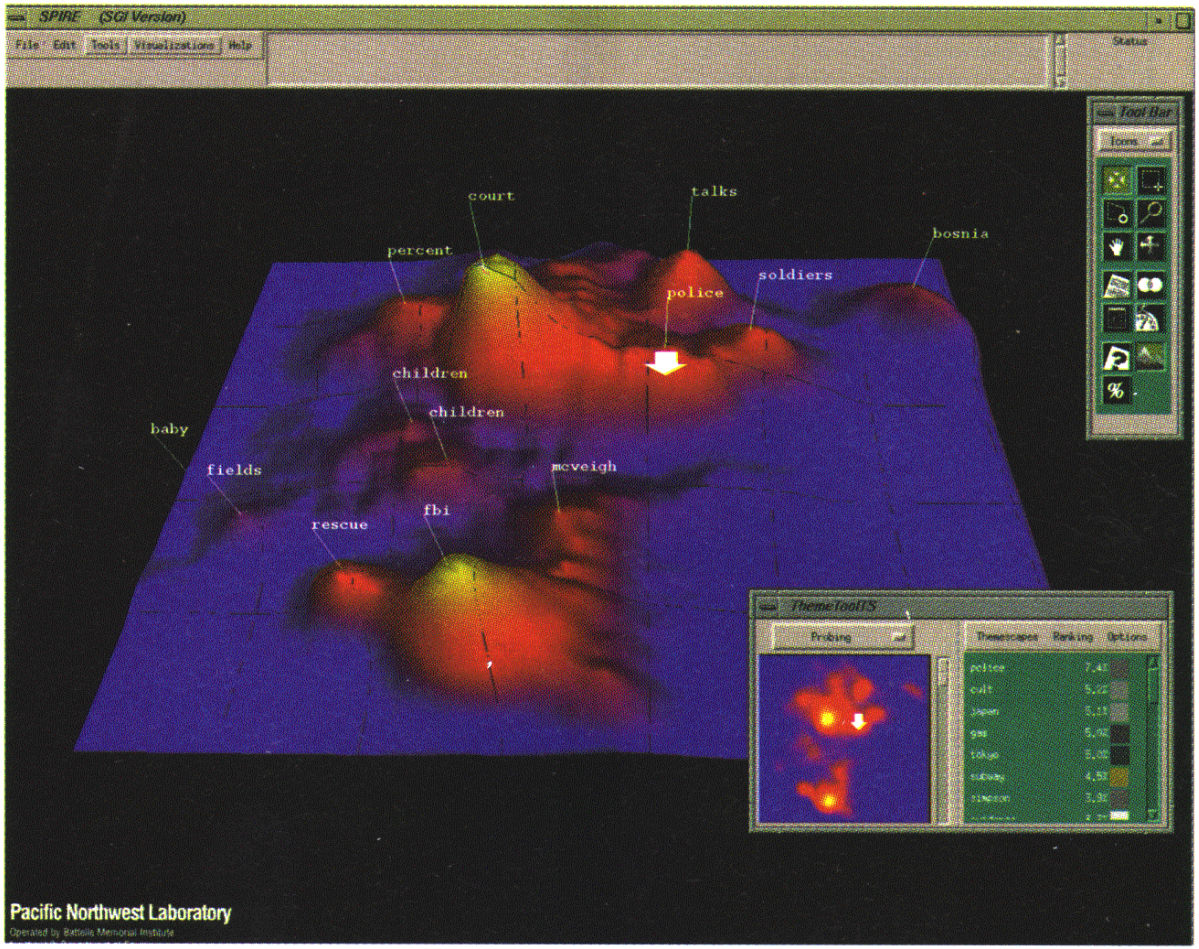
Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions (Assign. 1)
- **Data Visualization**
- Measuring Data Similarity and Dissimilarity

Scatterplot Matrices



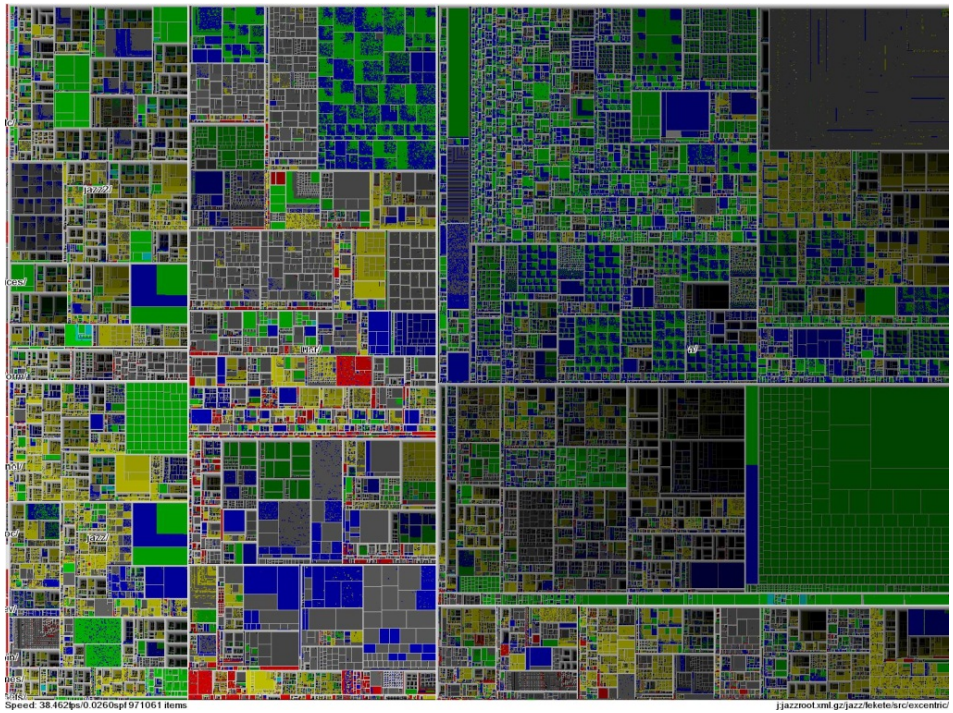
Landscapes



news articles visualized as a landscape

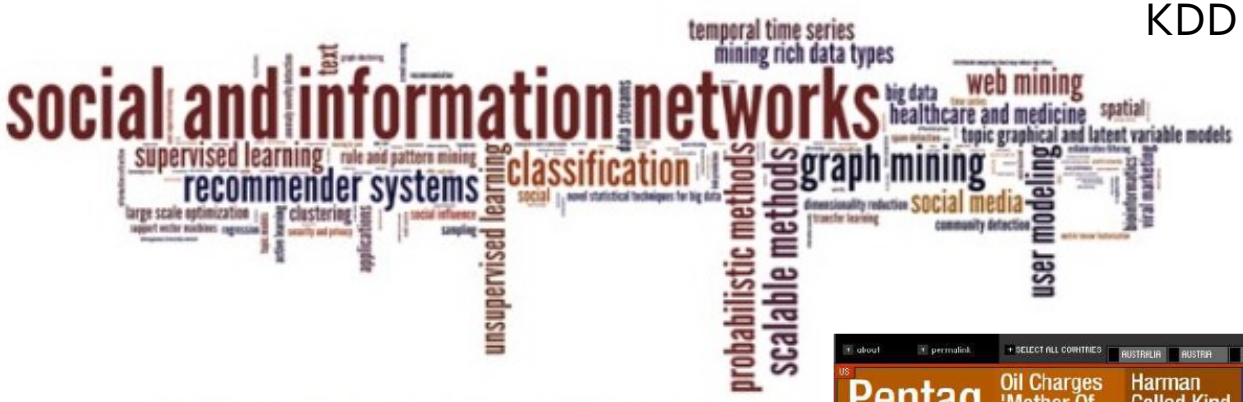
Tree-Map

- 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)

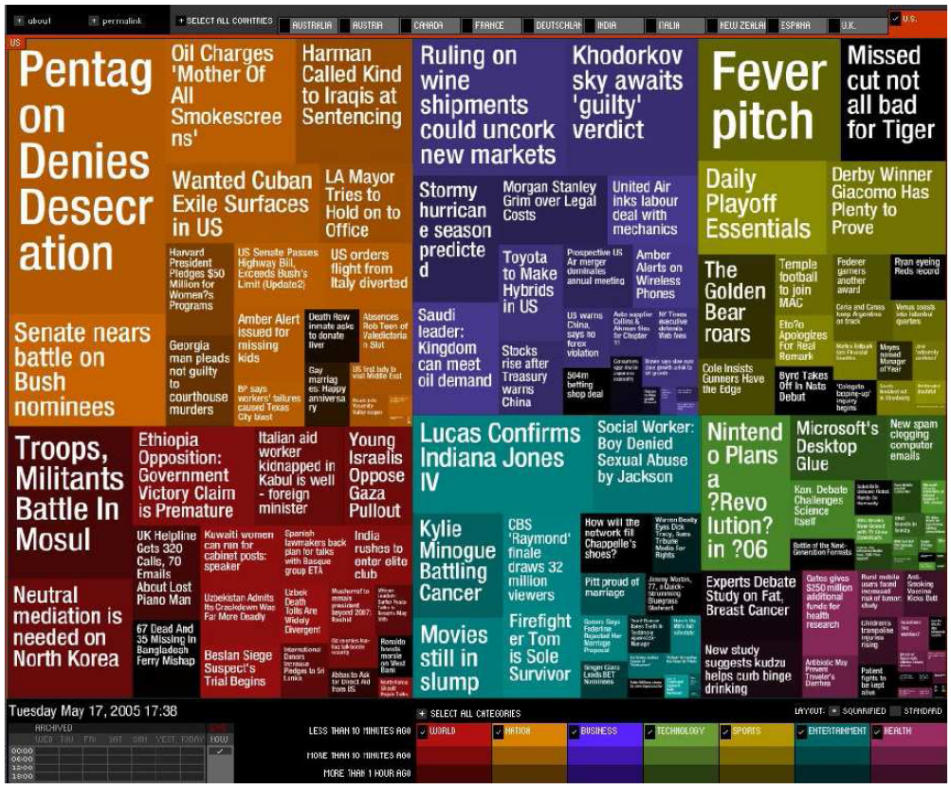


Tag Cloud

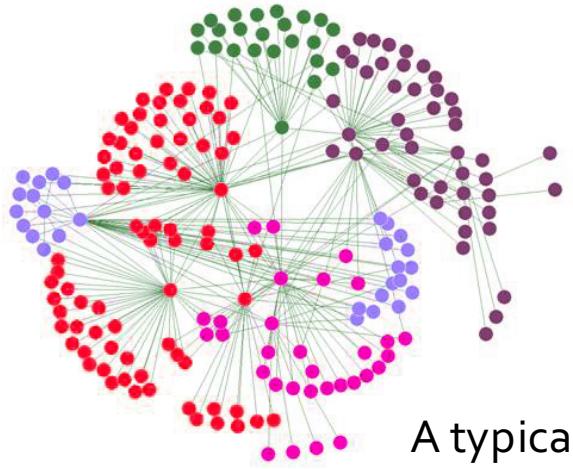
KDD 2013 Research Paper Title



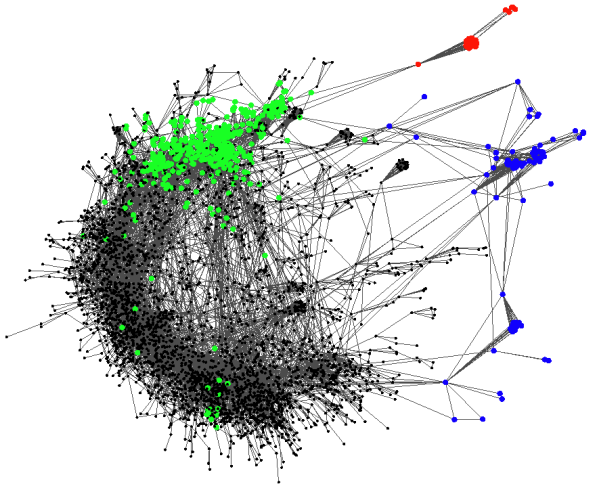
Newsmap: Google News Stories in 2005



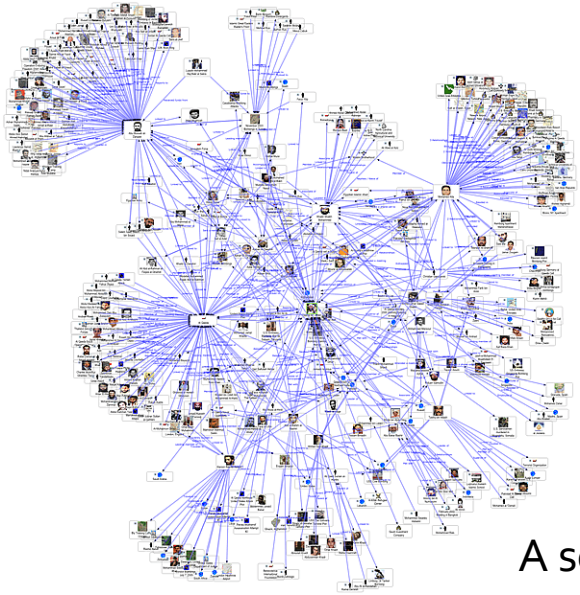
Networks



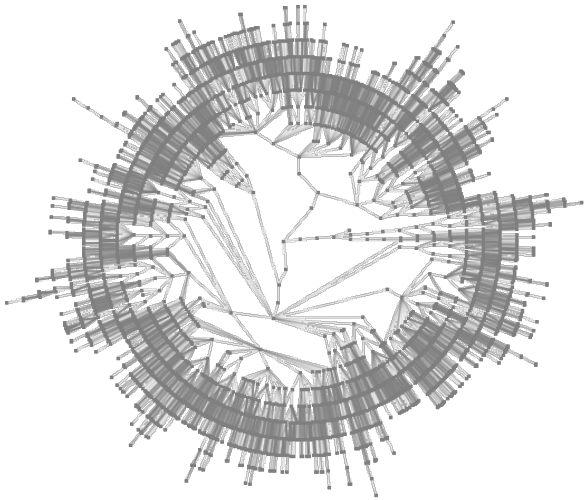
A typical network structure



organizing information networks



A social network



Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions (Assign. 1)
- Data Visualization
- **Measuring Data Similarity and Dissimilarity**

Similarity, Dissimilarity, and Proximity

- **Similarity measure** or **similarity function**
  - A real-valued function that quantifies the similarity between two objects
  - Measure how two data objects are alike: The higher value, the more alike
  - Often falls in the range [0,1]: 0: no similarity; 1: completely similar
- **Dissimilarity** (or **distance**) measure
  - Numerical measure of how different two data objects are
  - In some sense, the inverse of similarity: The lower, the more alike
  - Minimum dissimilarity is often 0 (i.e., completely similar)
  - Range [0, 1] or [0, ∞) , depending on the definition
- **Proximity** usually refers to either similarity or dissimilarity

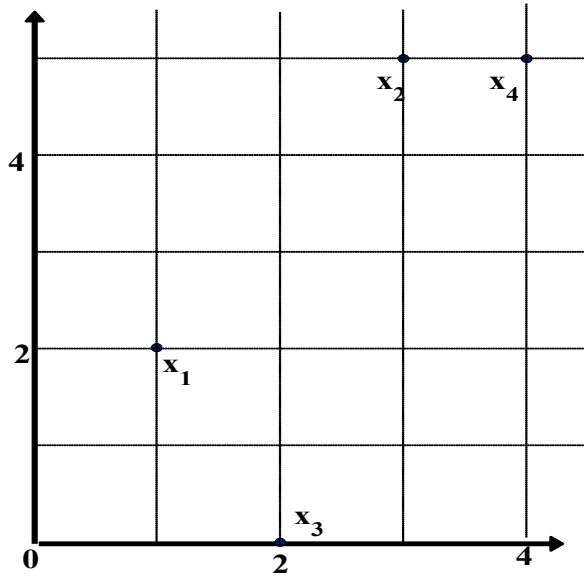
Data Matrix and Dissimilarity Matrix

- Data matrix
  - A data matrix of n data points with l dimensions
- Dissimilarity (distance) matrix
  - n data points, but registers only the distance d(i, j)
  - Usually symmetric, thus a triangular matrix
  - Distance functions are usually different for real, boolean, categorical, ordinal, ratio, and vector variables
  - Weights can be associated with different variables based on applications and data semantics

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1l} \\ x_{21} & x_{22} & \cdots & x_{2l} \\ \text{M} & \text{M} & \text{O} & \text{M} \\ x_{n1} & x_{n2} & \cdots & x_{nl} \end{pmatrix}$$

$$\begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ \text{M} & \text{M} & \text{O} & \\ d(n,1) & d(n,2) & \cdots & 0 \end{pmatrix}$$

Example: Euclidean Distance



Data Matrix

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

Dissimilarity Matrix (by Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Minkowski Distance

- Minkowski distance: A popular distance measure

$$d(i, j) = \sqrt[p]{| x_{i1} - x_{j1} |^p + | x_{i2} - x_{j2} |^p + \text{L} + | x_{il} - x_{jl} |^p}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{il})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jl})$ are two l-dimensional data objects, and p is the order (the distance so defined is called L-p norm)

- Properties
  - d(i, j) > o if i ≠ j, and d(i, i) = o (Positivity)
  - d(i, j) = d(j, i) (Symmetry)
  - d(i, j) ≤ d(i, k) + d(k, j) (Triangle Inequality)
- A distance that satisfies these properties is a metric
- Note: There are nonmetric dissimilarities, e.g., set differences

Special Cases of Minkowski Distance

- $p = 1$: (L_1 norm) Manhattan (or city block) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i, j) = | x_{i1} - x_{j1} | + | x_{i2} - x_{j2} | + \mathrm{L} + | x_{il} - x_{jl} |$$

- $p = 2$: (L_2 norm) Euclidean distance

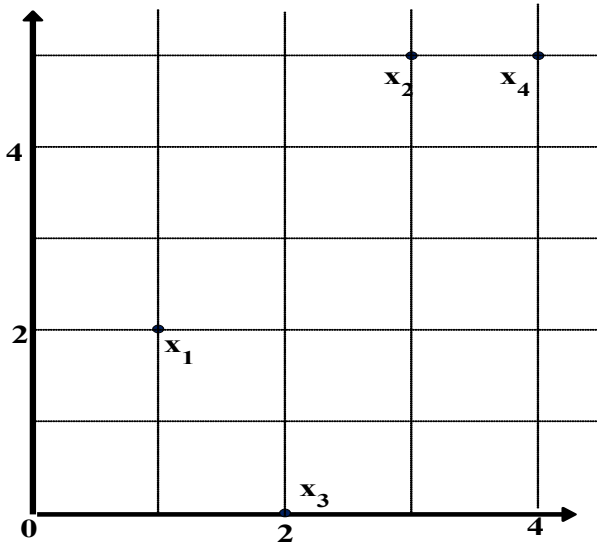
$$d(i, j) = \sqrt{| x_{i1} - x_{j1} |^2 + | x_{i2} - x_{j2} |^2 + \mathrm{L} + | x_{il} - x_{jl} |^2}$$

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

$$d(i, j) = \lim_{p \to \infty} \sqrt[p]{| x_{i1} - x_{j1} |^p + | x_{i2} - x_{j2} |^p + \mathrm{L} + | x_{il} - x_{jl} |^p} = \max_{f=1}^{l} | x_{if} - x_{if} |$$

Example: Minkowski Distance at Special Cases

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum (L_∞)

L∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Proximity Measure for Binary Attributes

- A contingency table for binary data

Object *i* \ Object *j*	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- Distance measure for symmetric binary variables:

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

- Distance measure for asymmetric binary variables:

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

- Jaccard coefficient (similarity measure for asymmetric binary variables):

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

- Note: Jaccard coefficient is the same as "coherence":

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

Proximity Measure for Categorical Attributes

- Categorical data, also called nominal attributes
  - Example: Color (red, yellow, blue, green), profession, etc.
- 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

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank (e.g., freshman, sophomore, junior, senior)
- Can be treated like interval-scaled
  - Replace an ordinal variable value by its rank:
  - Map the range of each variable onto [0, 1] by replacing i-th object in the f-th variable by
    - Example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1
      - Then distance: d(freshman, senior) = 1, d(junior, senior) = 1/3
  - Compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Types

- A dataset may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, and ordinal
- One may use a weighted formula to combine their effects:

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

  - If f is numeric: Use the normalized distance
  - If f is binary or nominal: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; or $d_{ij}^{(f)} = 1$ otherwise
  - If f is ordinal
    - Compute ranks z_{if} (where $z_{if} = \frac{r_{if} - 1}{M_f - 1}$)
    - Treat z_{if} as interval-scaled

Cosine Similarity of Two Vectors

- A **document** can be represented by a bag of terms or a long vector, with each attribute recording the *frequency* of a particular term (such as word, keyword, 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

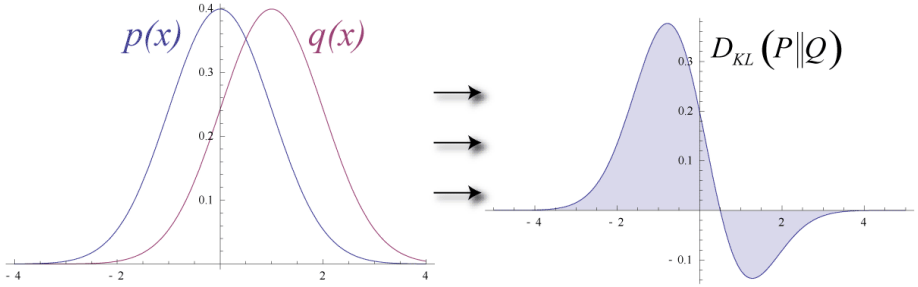
- Other vector objects: Gene features in micro-arrays
- Applications: Information retrieval, biologic taxonomy, gene feature mapping, etc.
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

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

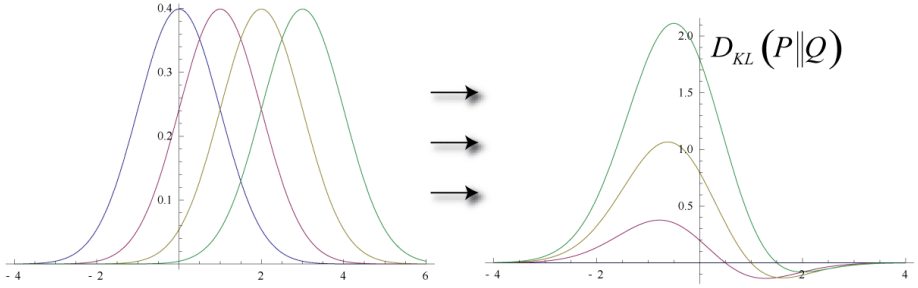
KL Divergence: Comparing Two Probability Distributions

- *The Kullback-Leibler (KL) divergence:* Measure the difference between two probability distributions over the same variable *x*
  - From information theory, closely related to *relative entropy*, *information divergence*, and *information for discrimination*
- $D_{KL}(p(x) \,||\, q(x))$: divergence of $q(x)$ from $p(x)$, measuring the information lost when $q(x)$ is used to approximate $p(x)$



Original Gaussian PDF's

KL Area to be Integrated



Discrete form

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Continuous form

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

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
- 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

Announcement

- CS412: Assignment #1 was distributed yesterday!
  - The due date is June 15. No late homework will be accepted!!
- Project for the 4th Credit was announced!
  - Reading KDD'15-'16 proceedings (430 papers)
  - Manually structuring paper full text into a "data object" (paper ID) – "attribute" (phrases of research problem, methodology, technique, datasets…) table
  - Clustering analysis based on **similarity/dissimilarity**
  - **Multi-faceted** clustering: vs. author, affiliation, etc.
  - A PDF report and a TabSV (.tsv) table: individual proj.