



Chapter 2. Getting to Know Your Data

Meng Jiang

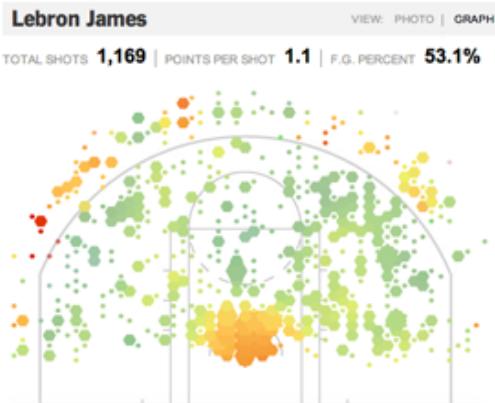
CS412 Summer 2017:

Introduction to Data Mining

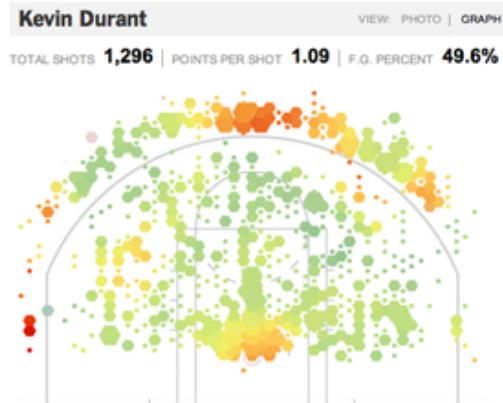
From Data to Knowledge

PLAYER	P	MIN	PTS	FGM	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	
LeBron James	F	38	34	15	27	55.6	1	6	16.7	3	4	75.0	0	5	5	6
Kevin Love	F	41	17	6	13	46.2	3	5	60.0	2	2	100	2	15	17	5
Tristan Thompson	C	37	7	3	4	75.0	0	0	0.0	1	2	50.0	1	6	7	0
JR Smith	G	26	5	2	4	50.0	1	2	50.0	0	0	0.0	1	0	1	3
Kyrie Irving	G	41	42	15	22	68.2	4	7	57.1	8	9	88.9	0	3	3	4
Richard Jefferson		6	0	0	0	0.0	0	0	0.0	0	0	0.0	0	0	0	0
Deron Williams		18	2	1	2	50.0	0	1	0.0	0	0	0.0	0	1	1	1
Kyle Korver		20	0	0	0	0.0	0	0	0.0	0	0	0.0	0	3	3	4
Iman Shumpert		10	5	2	2	100	1	1	100	0	0	0.0	0	0	0	0

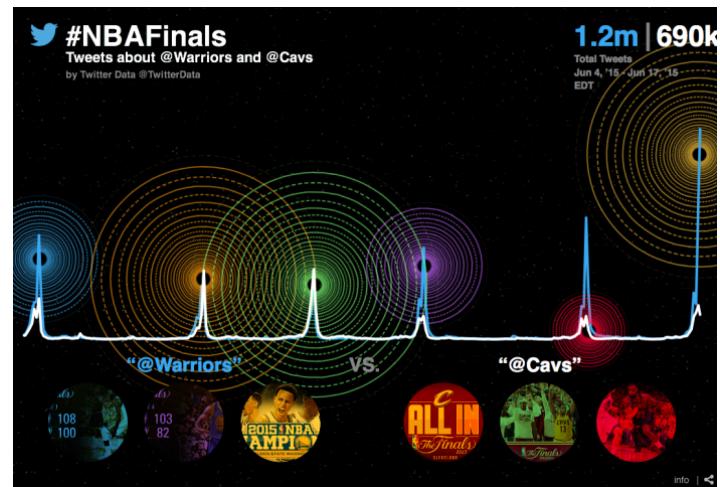
Eastern Conference								
TEAM	W	L	WIN%	GB	CONF	DIV	HOME	ROAD
1 🐱 - Boston	53	29	.646	-	36 - 16	11 - 5	30 - 11	23 - 18
2 🏹 - Cleveland	51	31	.622	2	35 - 17	8 - 8	31 - 10	20 - 21
3 🏀 * - Toronto	51	31	.622	2	34 - 18	14 - 2	28 - 13	23 - 18
4 🏀 se - Washington	49	33	.598	4	32 - 20	8 - 8	30 - 11	19 - 22
5 🐍 * - Atlanta	43	39	.524	10	30 - 22	6 - 10	23 - 18	20 - 21
6 🦌 * - Milwaukee	42	40	.512	11	27 - 25	10 - 6	23 - 18	19 - 22
7 🎖 * - Indiana	42	40	.512	11	26 - 26	8 - 8	29 - 12	13 - 28



His athleticism and ball-handling create a lot of high-percentage shots near the basket. He prefers the wing locations beyond the 3-point line. His midrange game is his weakest.



Despite his size, he is a very effective midrange shooter, taking nearly half his shots from that zone and another 25 percent from beyond the 3-point arc.



Your Data

	Number
Students who provide us data	?
Count of words in total	?
Average count of words per student	?
Count of non-stop-words in total	?
Average count of non-stop-words per student	?
Number of words (size of vocabulary)	?
Number of non-stop-words	?

Your Data

	Number
Students who provide us data	?
Count of words in total	1,175
Average count of words per student	?
Count of non-stop-words in total	?
Average count of non-stop-words per student	?
Number of words (size of vocabulary)	?
Number of non-stop-words	?

Your Data

	Number
Students who provide us data	29
Count of words in total	1,175
Average count of words per student	40.5
Count of non-stop-words in total	494
Average count of non-stop-words per student	17.0
Number of words (size of vocabulary)	426
Number of non-stop-words	298

Your Data

Word	Count
i	74
to	45
a	43
in	39
and	38
am	32
the	29
of	26
data	23
my	15
this	14

Word	#Students
i	27
a	25
to	22
am	22
in	20
and	19
the	18
data	15
of	13
this	11
my	11

Your Data

Word	Count
data	23
student	10
major	9
engineering	9
class	8
mining	7
students	7
science	6
grad	6
learning	6
interested	6

Word	#Students
data	15
student	9
major	8
engineering	8
class	7
mining	6
science	5
learning	5
graduate	5
interested	5
students	5

Knowledge (Patterns) You Want

- Students' personal interests (activities) and connect them with like-minded individuals * 2
- Correlation between classes students have taken and their grades
- Correlation between nationalities and interests in data mining
- Correlation between handwriting patterns and academic performance
- Classify students by their background knowledge of data mining
- Clustering students by their interests or backgrounds
- ...

Knowledge (Patterns) You Want

- Students' personal interests (activities) and connect them with like-minded individuals * 2
 - Correlation between classes students have taken and their grades
 - Correlation between nationalities and interests in data mining
 - Correlation between academic performance and interests in data mining
 - Classify students based on their academic performance
 - Clustering based on their interests in data mining
 - Clustering based on their academic performance
 - ...
- During the 5-minute break (11:40-11:45), if you'd like to contribute, please grab a paper and write down your self-introduction and give it to me at the end of class. Pay attention to privacy issue!

Knowledge/Patterns We Can Have: Phrase Structures

Phrase	Score
civil engineering	0.94
machine learning	0.93
big data	0.93
classify group	0.84
master student	0.82
cs knowledge	0.81
students' academic performance matches	0.81
human body	0.81
favoriate colors	0.80
skill levels	0.80
algorithm design	0.80

Phrase	Score
masters in computer science	0.70
academic history and transcripts	0.70
sophomore majored in statistics	0.70
professor of music history	0.70
addition to computer science	0.70
control modeling and management	0.69

Phrase	Score
4 th year phd candidate	0.68
soccer and practice guitar	0.67

Phrase	Score
a computer	0.08
the data	0.08
and major	0.08

Knowledge/Patterns We Can Have: Phrase Structures

Phrase	Score	Phrase	Score
civil engineering	0.94	masters in computer science	0.70
machine learning	0.93	academic history and transcripts	0.70
big data	0.93	sophomore majored in statistics	0.70
classify group	0.84	professor of music history	0.70
master student	0.82	addition to computer science	0.70
cs knowledge	0.81	control modeling and management	0.69
students' academic performance match		Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, and Jiawei Han, “Automated Phrase Mining from Massive Text Corpora”. arXiv: 1702:04457	Score
human body			0.68
favoriate colors			0.67
skill levels			Score
algorithm design			0.08
		and major	0.08

Chapter 2. Getting to Know Your Data

- **Data Objects and Attribute Types**
- Basic Statistical Descriptions
- 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	P	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST
LeBron James	F	38	34	15	27	55.6	1	6	16.7	3	4	75.0	0	5	5	6
Kevin Love	F	41	17	6	13	46.2	3	5	60.0	2	2	100	2	15	17	5
Tristan Thompson	C	37	7	3	4	75.0	0	0	0.0	1	2	50.0	1	6	7	0
JR Smith	G	26	5	2	4	50.0	1	2	50.0	0	0	0.0	1	0	1	3
Kyrie Irving	G	41	42	15	22	68.2	4	7	57.1	8	9	88.9	0	3	3	4
Richard Jefferson		6	0	0	0	0.0	0	0	0.0	0	0	0.0	0	0	0	0
Deron Williams		18	2	1	2	50.0	0	1	0.0	0	0	0.0	0	1	1	1
Kyle Korver		20	0	0	0	0.0	0	0	0.0	0	0	0.0	0	3	3	4
Iman Shumpert		10	5	2	2	100	1	1	100	0	0	0.0	0	0	0	0

PLAYER	TEAM	AGE	GP	W	L	MIN	PTS	FGM	FGA	FG%
1 Russell Westbrook	OKC	28	81	46	35	34.6	31.6	10.2	24.0	42.5
2 James Harden	HOU	27	81	54	27	36.4	29.1	8.3	18.9	44.0
3 Isaiah Thomas	BOS	28	76	51	25	33.8	28.9	9.0	19.4	46.3
4 Anthony Davis	NOP	24	75	31	44	36.1	28.0	10.3	20.3	50.5
5 DeMar DeRozan	TOR	27	74	47	27	35.4	27.3	9.7	20.9	46.7
6 Damian Lillard	POR	26	75	38	37	35.9	27.0	8.8	19.8	44.4
7 DeMarcus Cousins	NOP	26	72	30	42	34.2	27.0	9.0	19.9	45.2
8 LeBron James	CLE	32	74	51	23	37.8	26.4	9.9	18.2	54.8

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

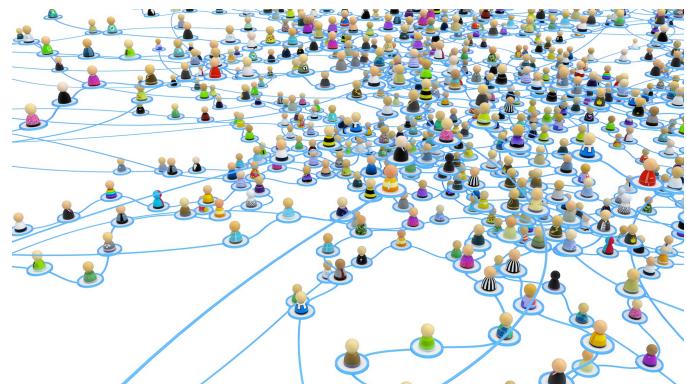
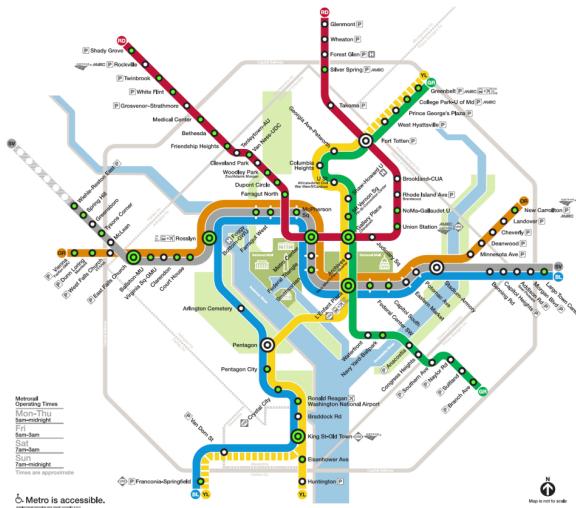
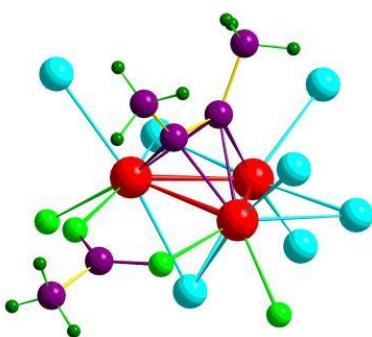
- **Nominal:** categories, states, or “names of things”
 - $\text{Hair_color} = \{\text{auburn}, \text{black}, \text{blond}, \text{brown}, \text{grey}, \text{red}, \text{white}\}$
 - marital status, occupation, ID numbers, zip codes
- **Binary**
 - Nominal attribute with only 2 states (0 and 1)
 - Symmetric 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
 - $\text{Size} = \{\text{small}, \text{medium}, \text{large}\}$, 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

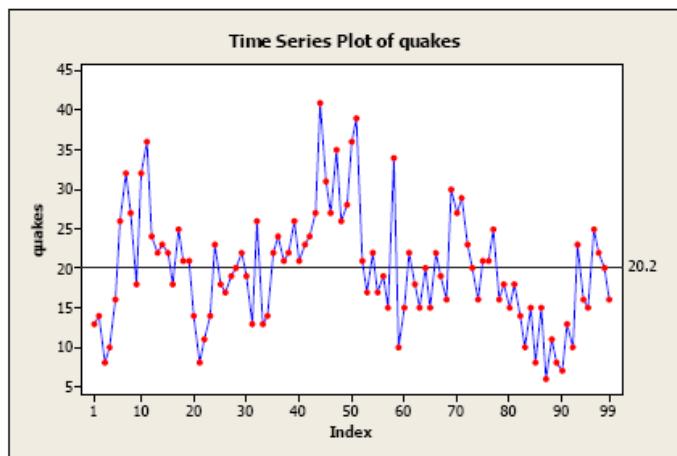
Types of Data Sets: (2) Graphs and Networks

- 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

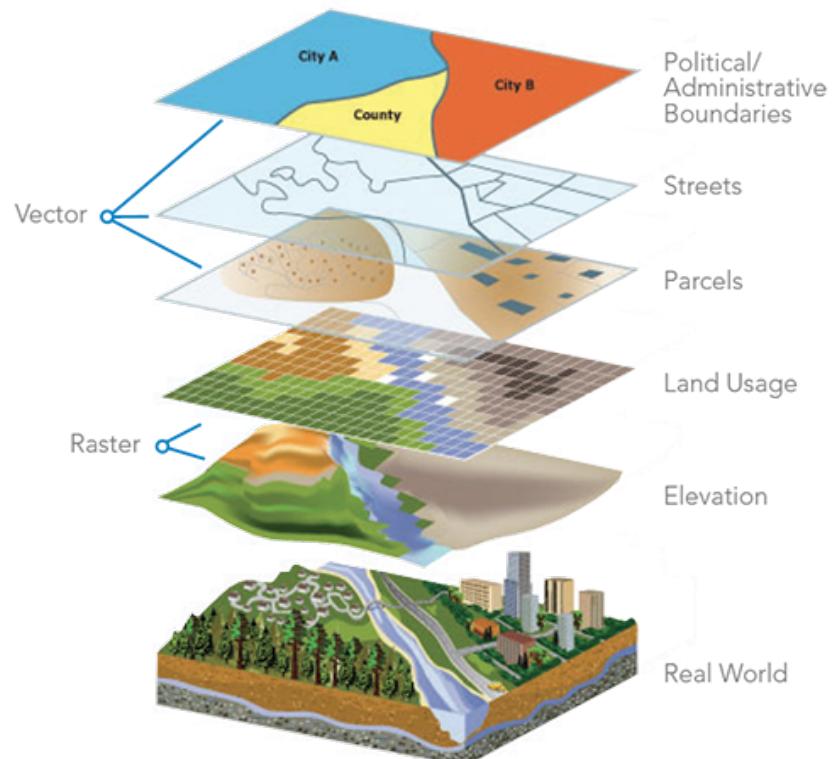


Start

Human	GTTTGAGG	- - ATGTTCAACAAATGCTCCTTCATTCCCTTCTATTACAGACCTGCCGCA
Chimpanzee	GTTTGAGG	- - ATGTTCAATAAATGCTGCTTCACTCCTTCTATTACAGACCTGCCGCA
Macaque	GTTTGAGG	- - ATGCTCAATAAATGCTCCTTCATTCCCTCATTACAAACTTGCCGCA
Human	GACAATTCTGCTAGCAGCCTTGTGCTATTATCTGTTTCTAAACCTTAGTAATTGAGTGT	
Chimpanzee	GACAATTCTGCTAGCAGCCTTGTGCTATTATCTGTTTCTAAACCTTAGTAATTGAGTGT	
Macaque	GACAATTCTGCTAGCAGCCTTGTGCTATTATCTGTTTCTAAACCTTAGTAATTGAGTGT	
Human	GATCTGGAGACTAA-CTCTGAAATAAAATAAGCTGATTATTTATTTCTCAAAACAA	
Chimpanzee	GATCTGGAGACTAAACTCTGAAATAAAATAAGCTGATTATTTATTTCTCAAAACAA	
Macaque	TATCTGGAGACTAAACTCTGAAATAAAATAAGCTGATTATTTATTTATTTCTCAAAACAA	
Human	CAGAAATACGATTTAGCAAATTACTCTTAAGATAATTATTTACATTTCTATATTCTCCTA	
Chimpanzee	CAGAAATACGATTTAGCAAATTACTCTTAAGATACTATTACATTTCTATATTCTCCTA	
Macaque	CAGAAATATGATTTAGCAAATTACCTCTTAAGATAATTATTTGCACATTCTATATTCTCCTA	
Human	CCCTGAGTTGATGTGTGAGCAATATGTCACTTTCATAAAAGCCAGGTATAACA-----TTATG	
Chimpanzee	CCCTGAGTTGATGTGTGAGCGTATGTCACTTTCATAAAAGCCAGGTATAACA-----TTATG	
Macaque	CCCTGAGTTGATGTGTGAGCAATATGTCACTTCCACAAAGCCAGGTATAATAACATTACG	
Human	GACAGGTAAAGTAAAAAACATATTATTATTCAGTTTGTCCAAAATTTAAATTTC	
Chimpanzee	GACAGGTAAAGTAAAAAACATATTATTATTCAGTTTGTCCAAAAGATTAAATTTC	
Macaque	GACAGGTAAAGTAAAAA-CATATTATTATTCAGGTTTTGTCCAAGAGTTAAATTTC	
Human	AACCTGTTGCGCGTGTGTTGGTAA-----TGTAAAACAAACTCAGTACA	
Chimpanzee	AACCTGTTGCGCGTGTGTTGGTAA-----TGTAAAACAAACTCAGTACA	
Macaque	AACCTGTTGCGCATGTTGGTAA-----CGTAAAACAAATTCAGTACG	

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**
- 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):

Finding the **median** from a grouped frequency table

Example: Parcels

Calculate an **estimate of the median** weight, shown to the nearest gram, in the following grouped frequency table:

Weight (g)	1-10	11-20	21-30	31-40	41-50
Frequency	10	13	28	15	9

Which interval does the median lie?

Cumulative frequency

Finding the median from a grouped frequency table

Weight (g)	1-10	11-20	21-30	31-40	41-50
Frequency	10	13	28	15	9
Cumulative Frequency	10	23	51	66	75

The median lies in the $\frac{1}{2} (75) = 37.5$ therefore:
 38^{th} position

Therefore the **median** lies in the **21 – 30** class.

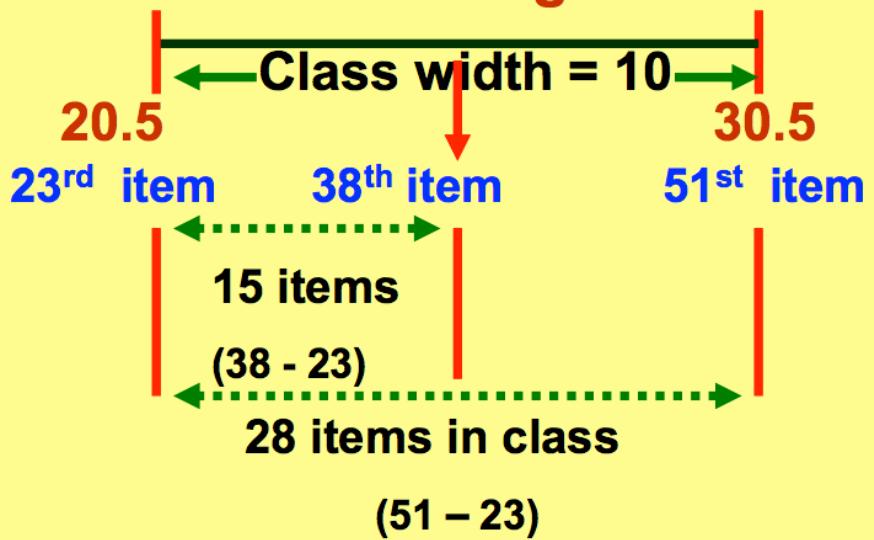
BUT

Remember it was rounded:

ACTUAL CLASS BOUNDARY: 20.5 – 30.5

Lower class boundary + number of items up to median x class width
 Number of items in the class

Useful diagram:



Assumption:

data are evenly spread over each class interval

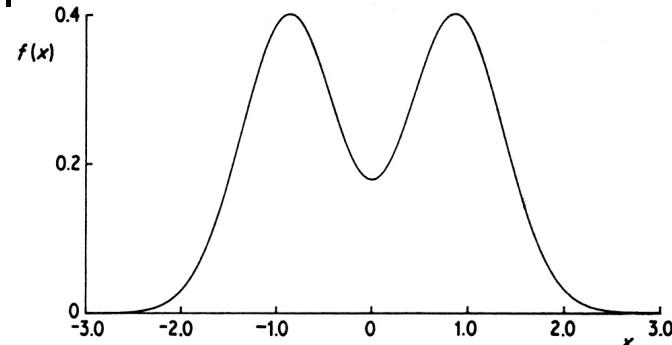
$$\begin{aligned}
 & 20.5 + \frac{15}{28} \times 10 \\
 & = 25.85714.... \\
 & = 25.9(1d.p.)
 \end{aligned}$$

$$\text{median} = L_1 + \left(\frac{n/2 - (\sum \text{freq})_l}{\text{freq}_{\text{median}}} \right) \text{width}$$

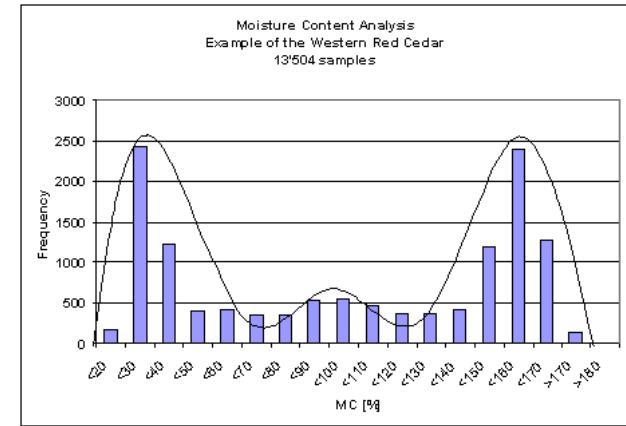
Measuring the Central Tendency:

(3) Mode

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

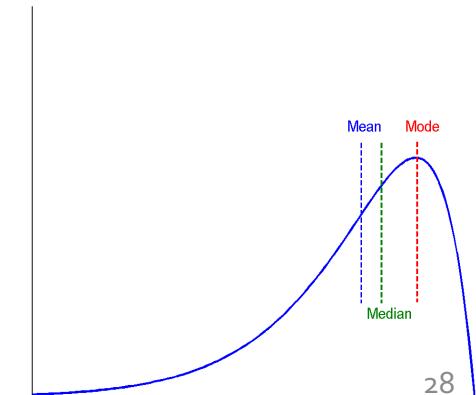
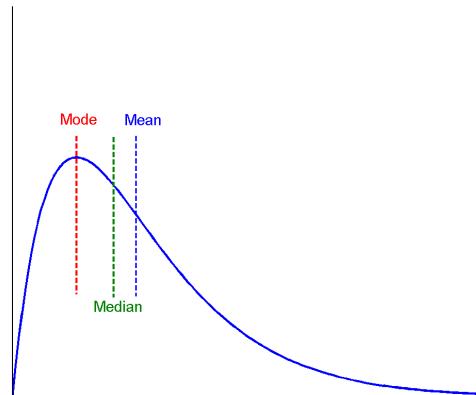
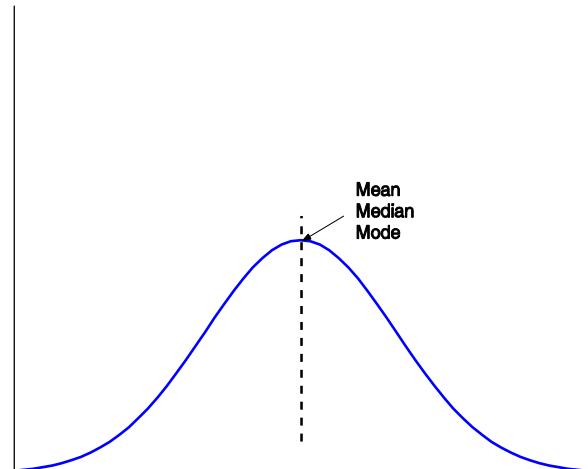


Symmetric data

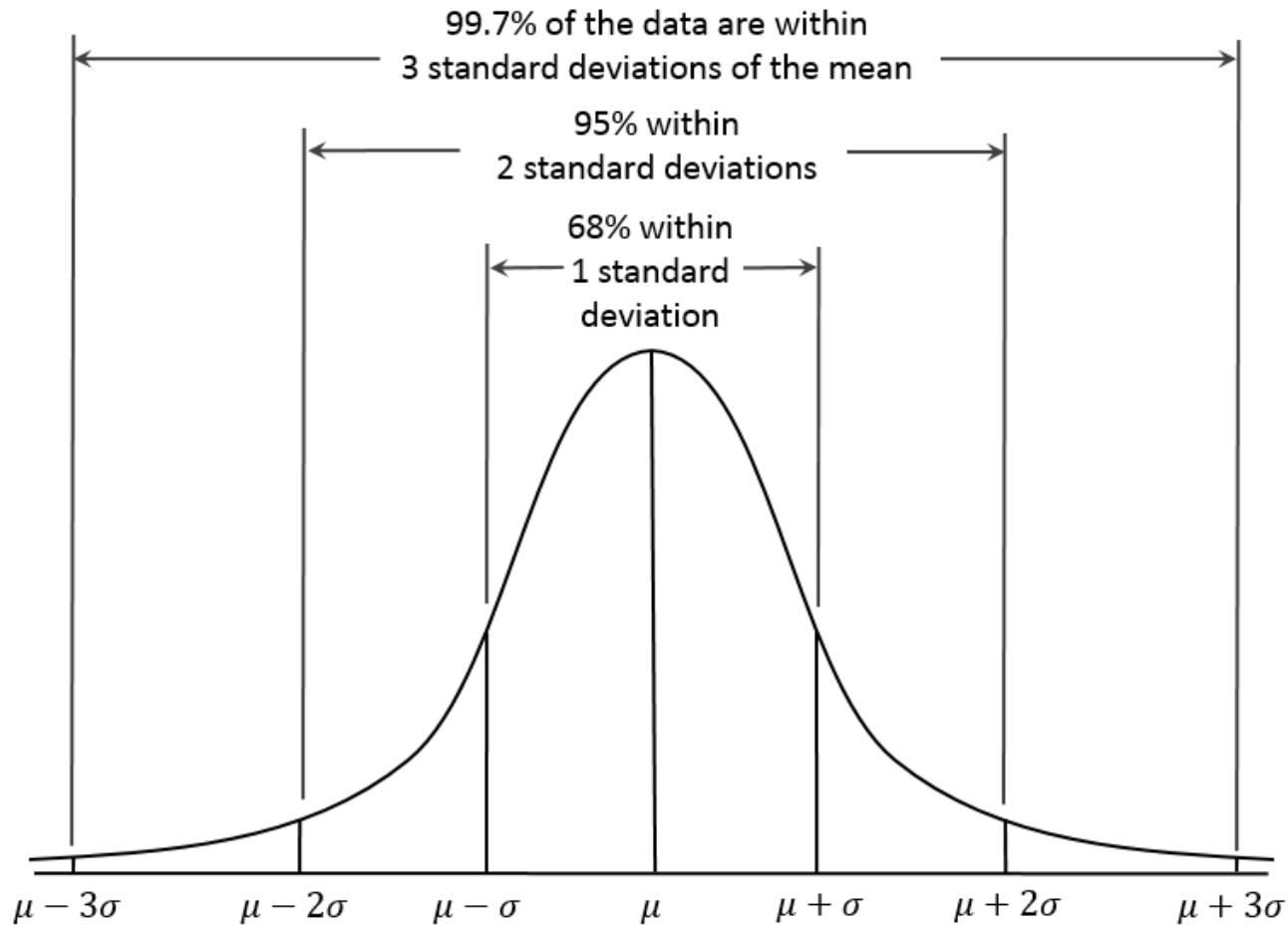


Positively skewed data

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?

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

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

- Standard deviation s (or σ) is the square root of variance s^2 (or σ^2)

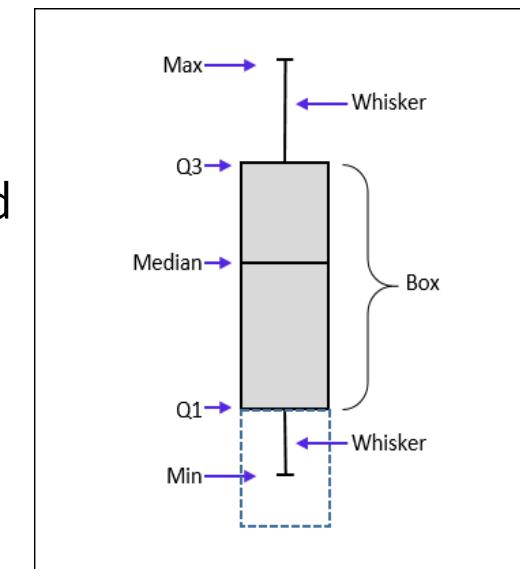
Graphic Displays of Basic Statistical Descriptions

- **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 $100f_i\%$ 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:

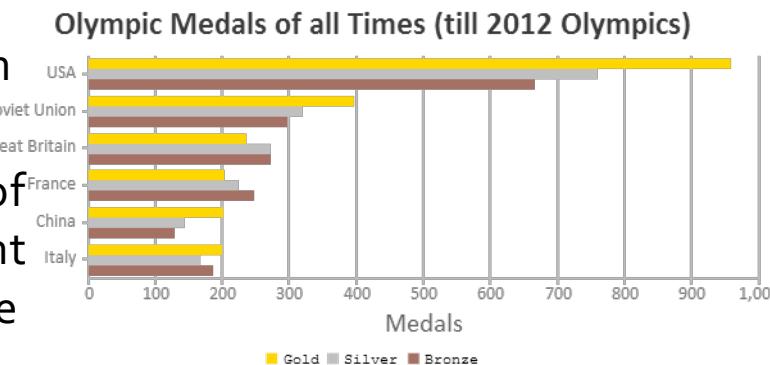
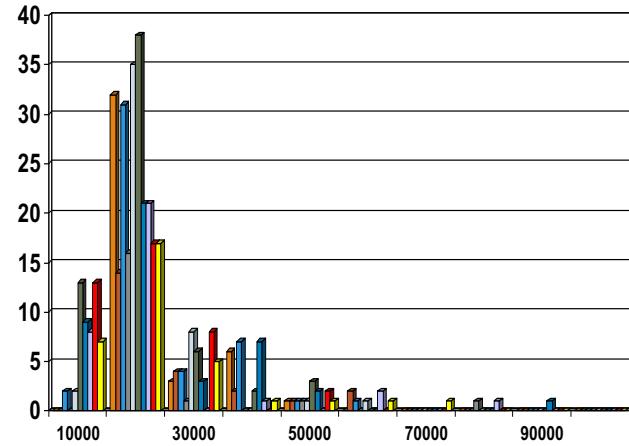
(1) Quartiles & Boxplots

- **Quartiles:** Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
- **Inter-quartile range:** $\text{IQR} = Q_3 - Q_1$
- **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 $1.5 \times \text{IQR}$



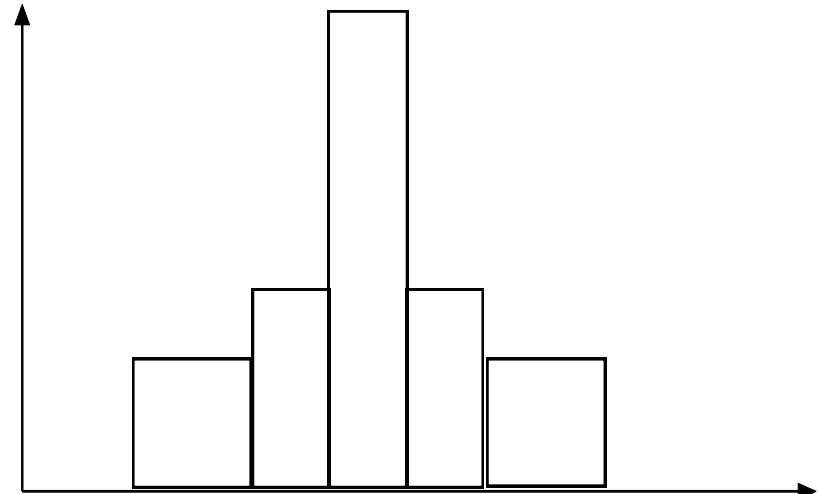
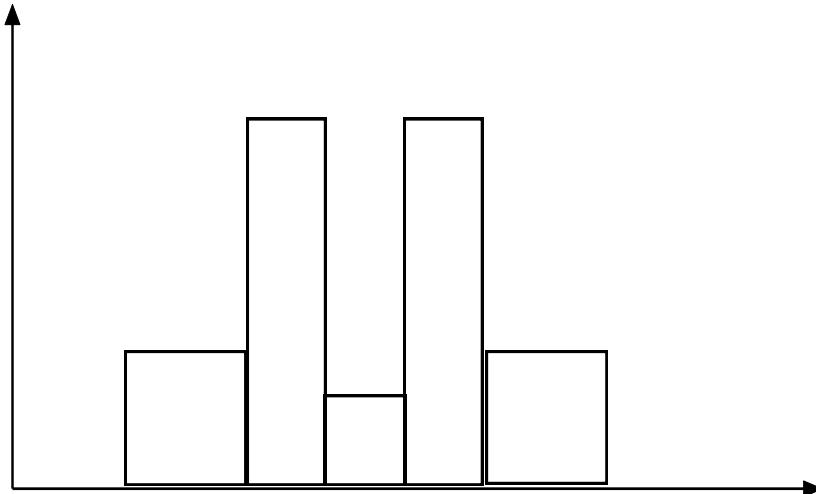
(2) Histogram Analysis

- 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



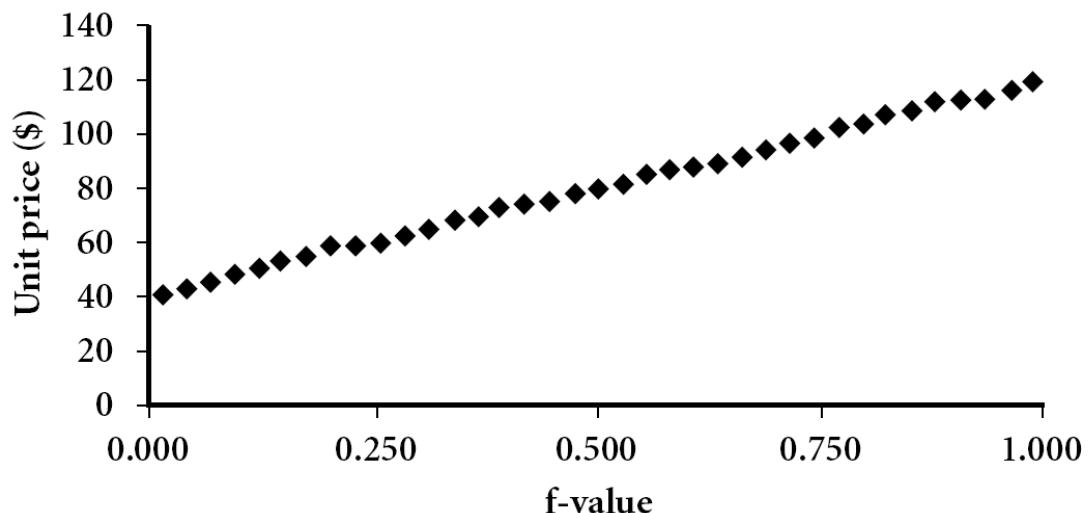
Histograms Often Tell More than Boxplots

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



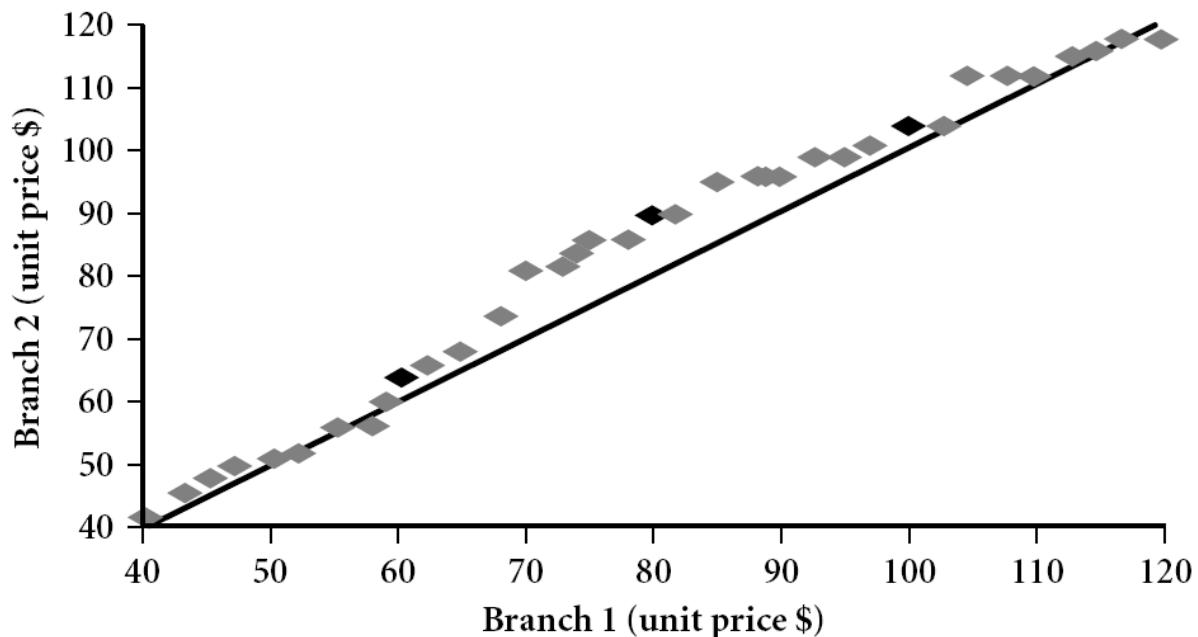
(3) Quantile Plot

- 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



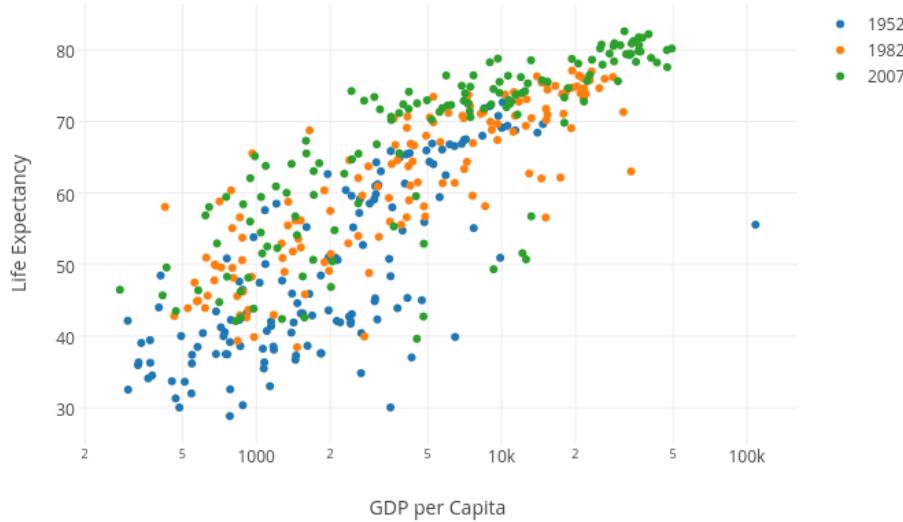
(4) Quantile-Quantile (Q-Q) Plot

- 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

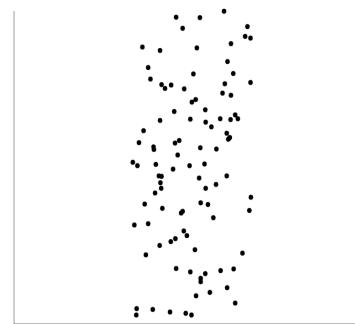
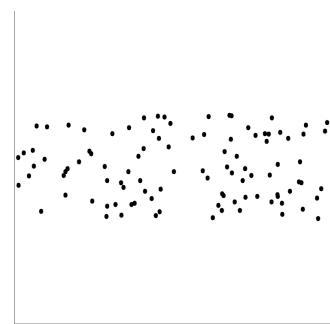
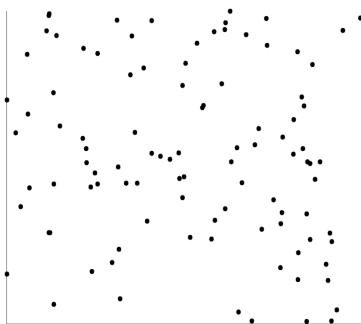
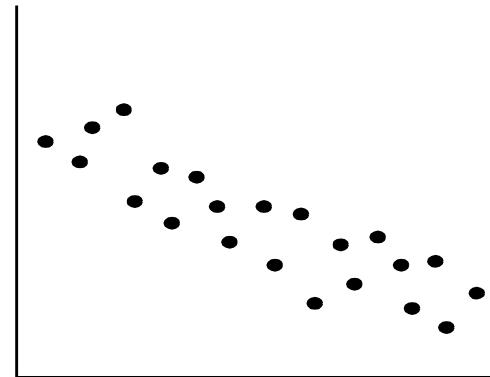
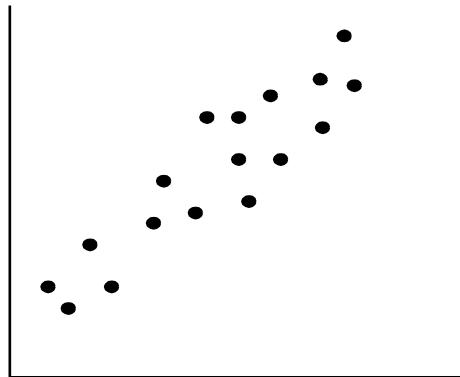


(5) Scatter plot

- 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



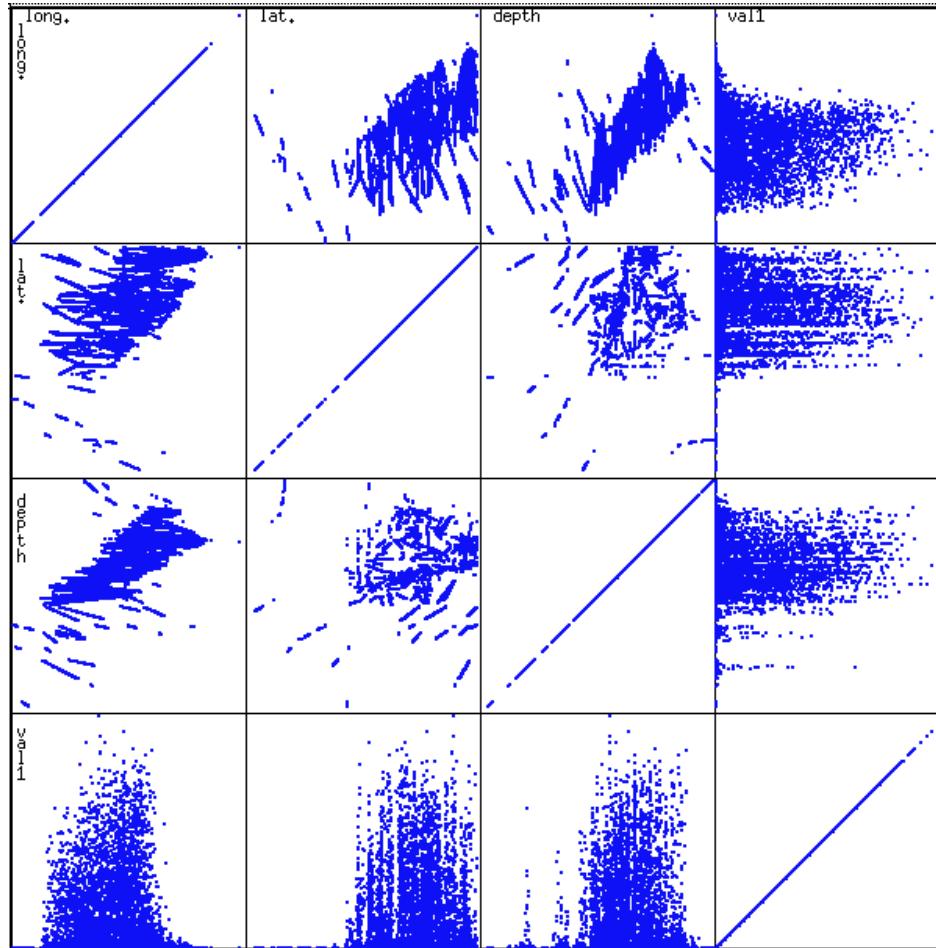
Positively, Negatively Correlated, and Uncorrelated Data



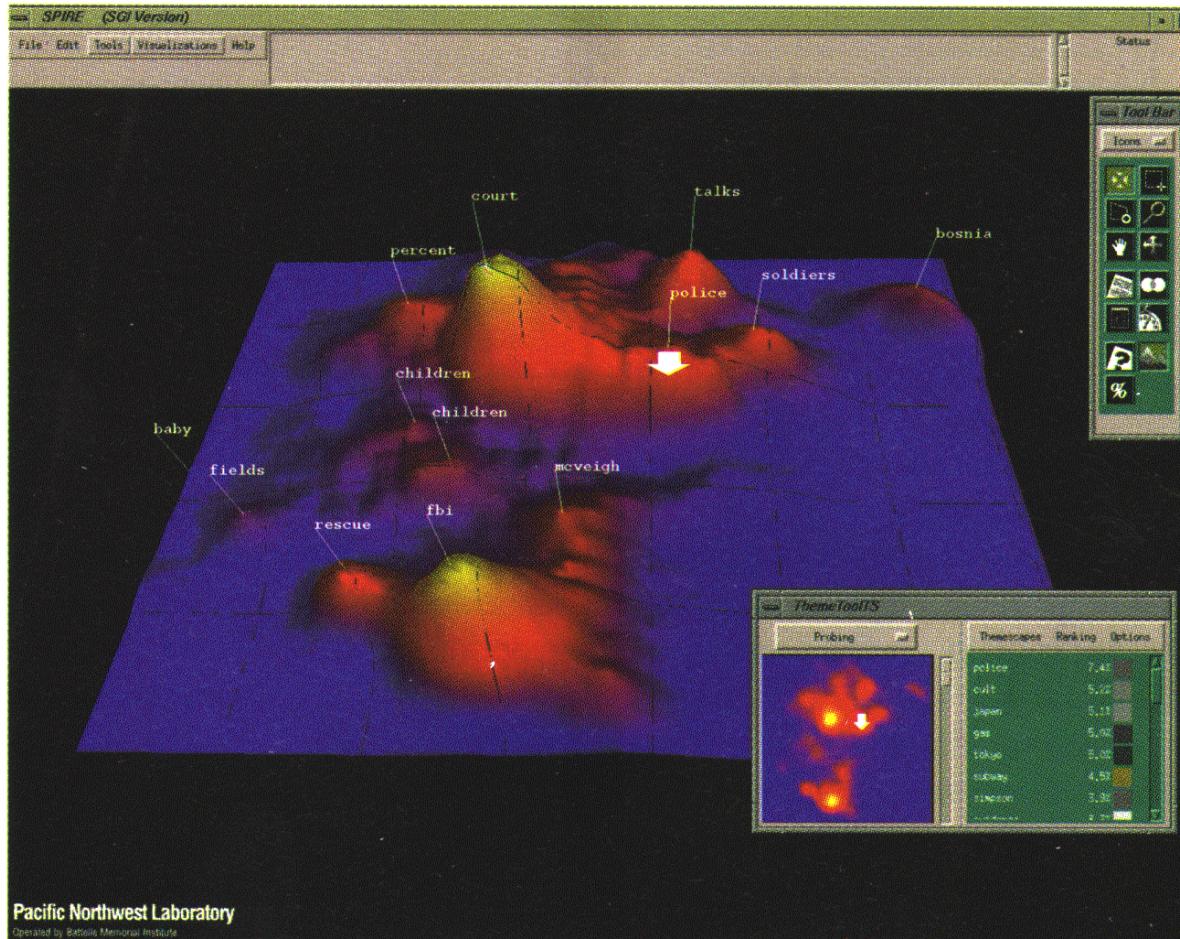
Chapter 2. Getting to Know Your Data

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

Scatterplot Matrices



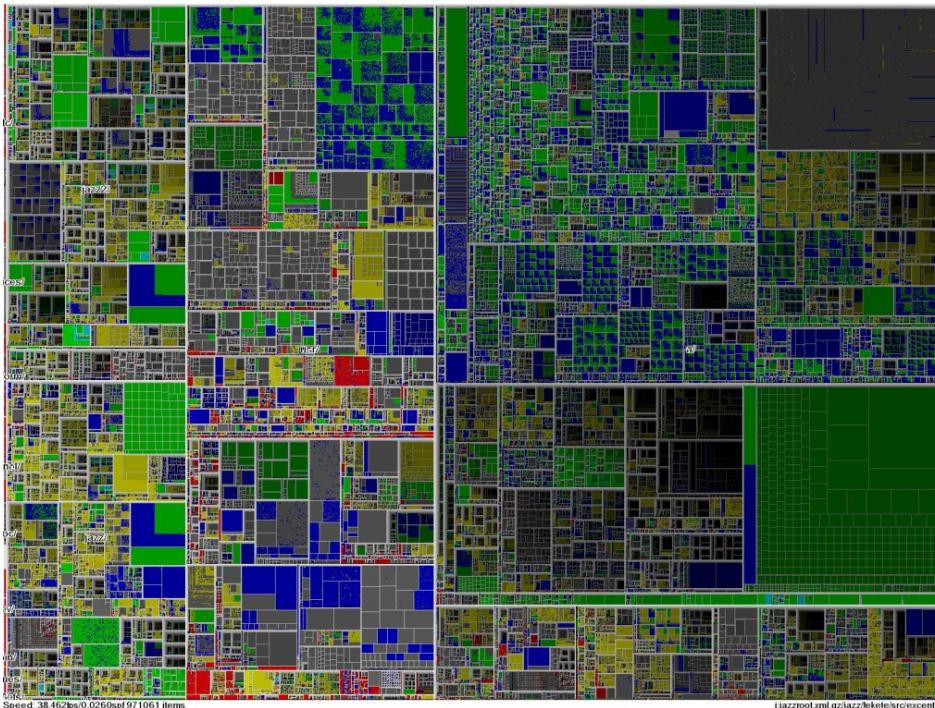
Landscapes



news articles visualized as a landscape

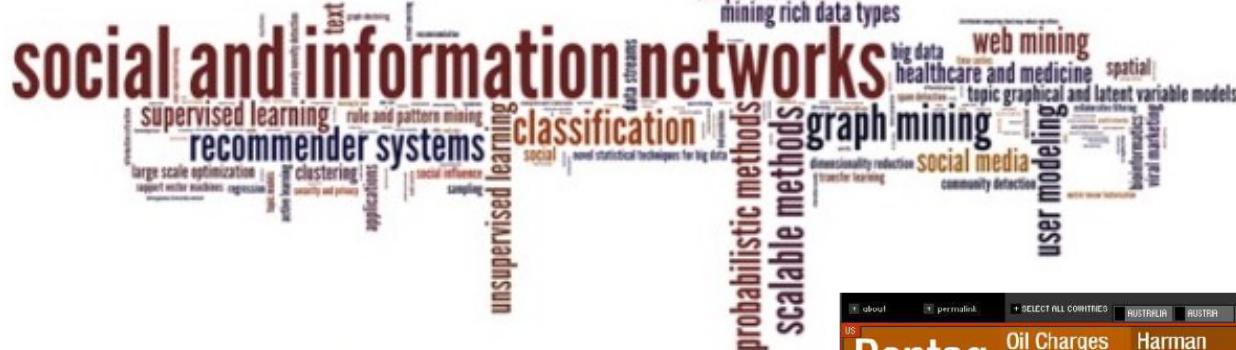
Tree-Map

- 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)



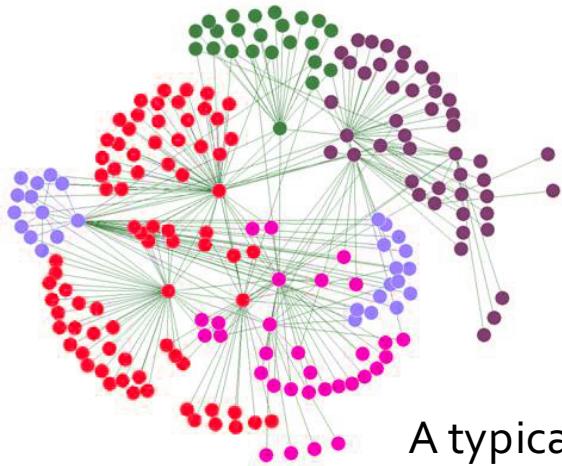
Tag Cloud

KDD 2013 Research Paper Title

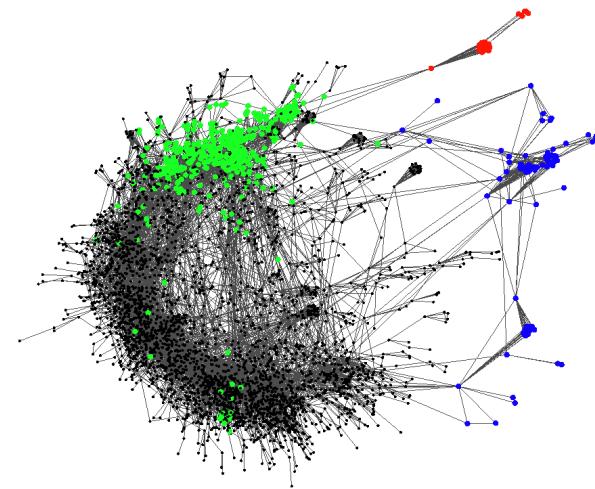


Newsmap: Google News Stories in 2005

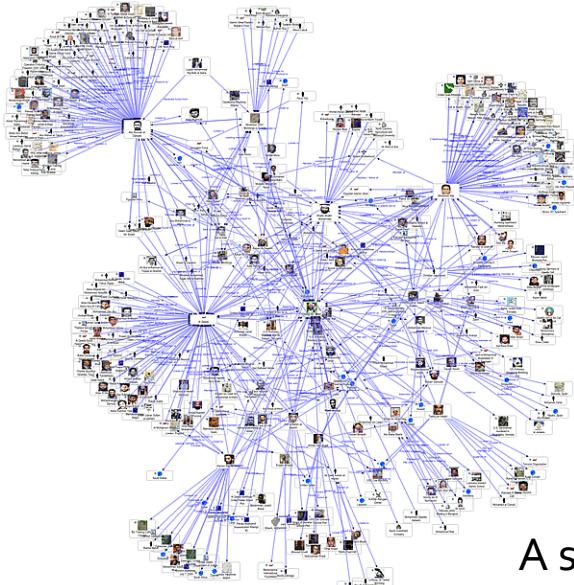
Networks



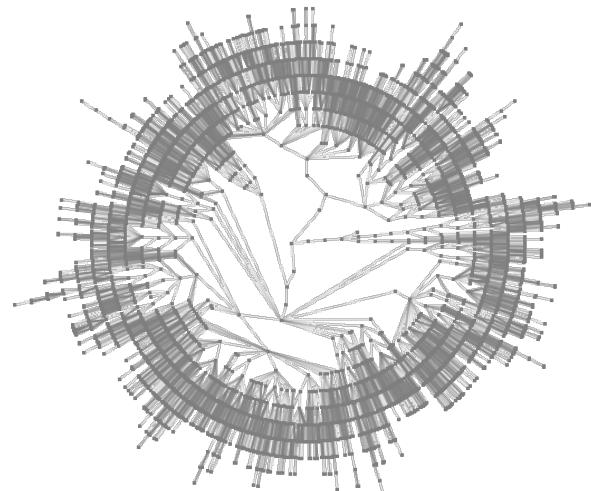
A typical network structure



organizing information networks



A social network



Chapter 2. Getting to Know Your Data

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

Similarity, Dissimilarity, and Proximity

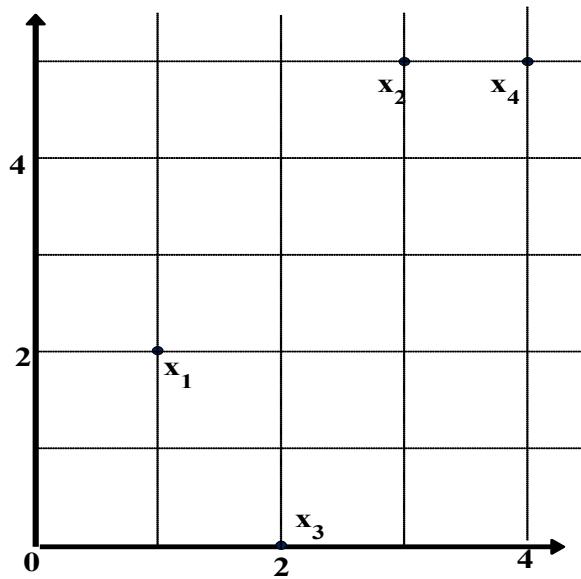
- **Similarity measure** or **similarity function**
 - A real-valued function that quantifies the similarity between two objects
 - Measure how two data objects are alike: The higher value, the more alike
 - Often falls in the range $[0,1]$: 0: no similarity; 1: completely similar
- **Dissimilarity** (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 $[0, \infty)$, 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

$$D = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1l} \\ x_{21} & x_{22} & \dots & x_{2l} \\ M & M & O & M \\ x_{n1} & x_{n2} & \dots & x_{nl} \end{pmatrix}$$
$$\begin{pmatrix} 0 \\ d(2,1) & 0 \\ M & M & O \\ d(n,1) & d(n,2) & \dots & 0 \end{pmatrix}$$

Example: Euclidean Distance



Data Matrix

point	attribute1	attribute2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5

Dissimilarity Matrix (by Euclidean Distance)

	x_1	x_2	x_3	x_4
x_1	0			
x_2	3.61	0		
x_3	2.24	5.1	0	
x_4	4.24	1	5.39	0

Minkowski Distance

- **Minkowski distance:** A popular distance measure

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

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

- Properties
 - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positivity)
 - $d(i, j) = d(j, i)$ (Symmetry)
 - $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)
- A distance that satisfies these properties is a **metric**
- 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

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

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

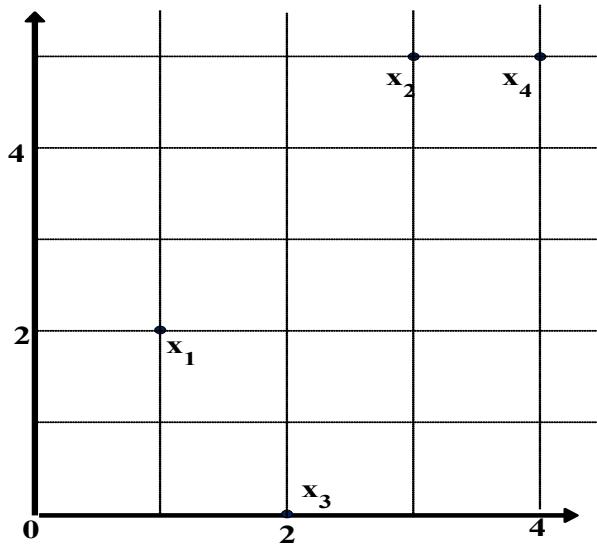
$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{il} - x_{jl}|^2}$$

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

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

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>j</i>		sum
		1	0	
Object <i>i</i>	1	<i>q</i>	<i>r</i>	<i>q+r</i>
	0	<i>s</i>	<i>t</i>	<i>s+t</i>
sum	<i>q+s</i>	<i>r+t</i>	<i>p</i>	

- 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):

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

- Note: Jaccard coefficient is the same as “coherence”:

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

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: $d(\text{freshman}, \text{senior}) = 1$, $d(\text{junior}, \text{senior}) = 1/3$
 - 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

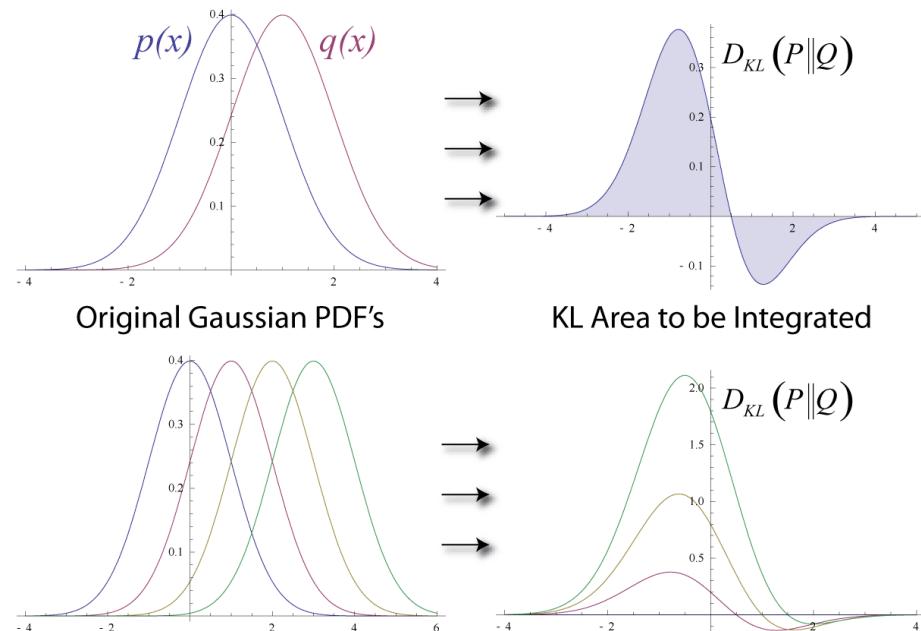
- Other vector objects: Gene features in micro-arrays
- Applications: Information retrieval, biological taxonomy, gene feature mapping, etc.
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

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

where \bullet 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*
- $D_{KL}(p(x) \parallel q(x))$: divergence of $q(x)$ from $p(x)$, measuring the information lost when $q(x)$ is used to approximate $p(x)$



Discrete form

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

Continuous form

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

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