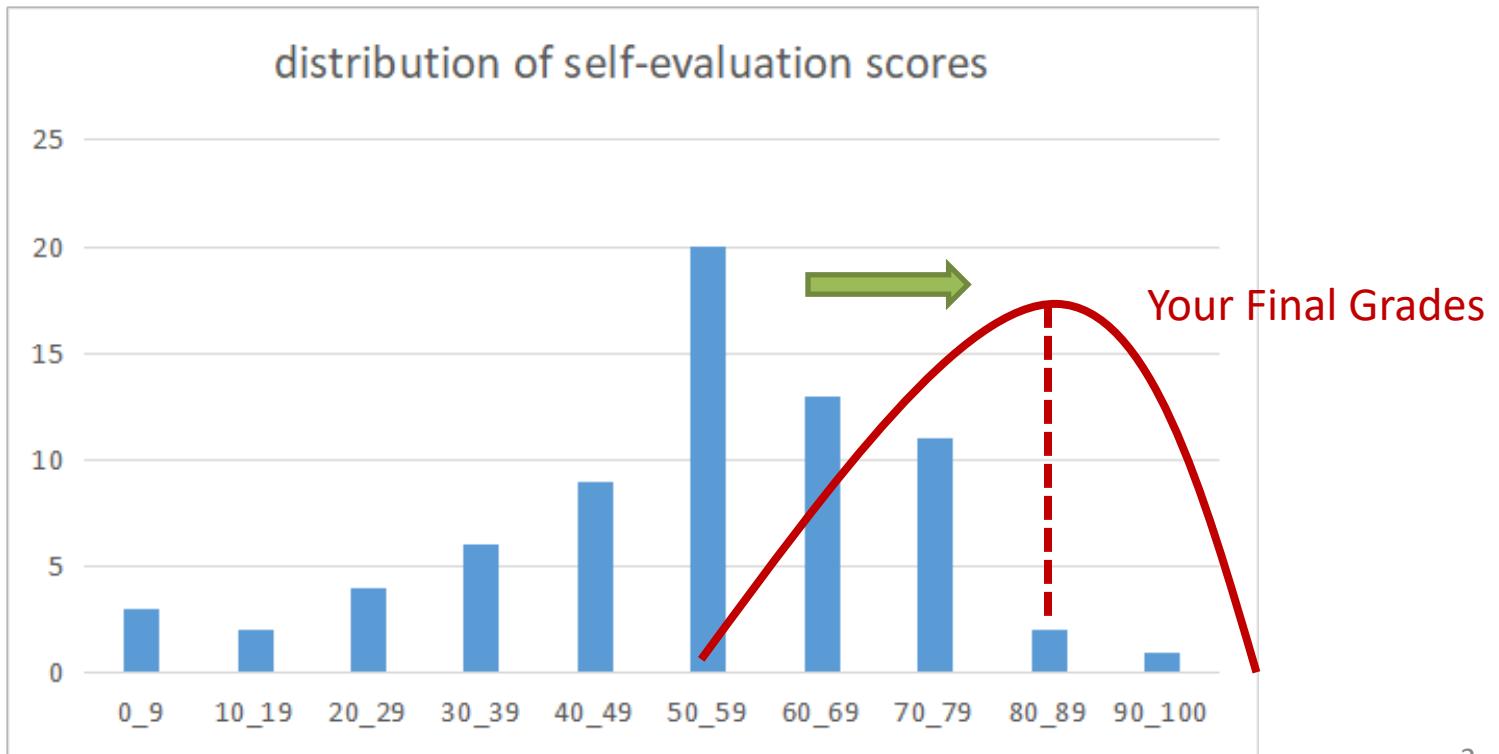


Quiz Summarization

Score Distribution

- Uncertainty
 - Noise on measurements (e.g., online exam)
 - Limited observation (e.g., failing to join the first class)





CS5824: Advanced Machine Learning

Dawei Zhou
CS, Virginia Tech

Please keep your face covering on!

Data Objects and Attribute Types

Types of Data Sets: (1) Record Data

- Relational records
 - Relational tables, highly structured
- Data matrix, e.g., numerical matrix

	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Lycra Glove		10.00	6.00		323.00	339.00
InFlux Crochet Glove	3.00	6.00	8.00		132.00	149.00
InFlux Lycra Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

- Transaction data

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

- Document data: Term-frequency vector (matrix) of documents

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

— no relation

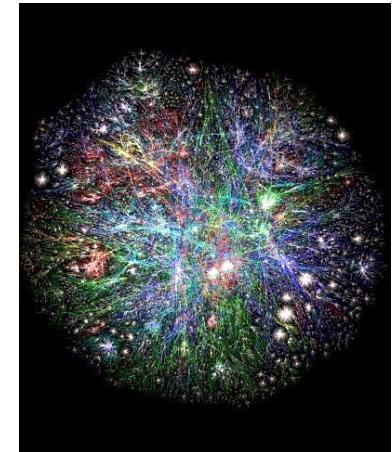
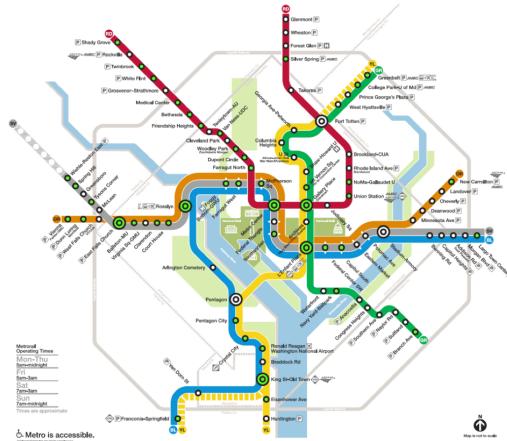
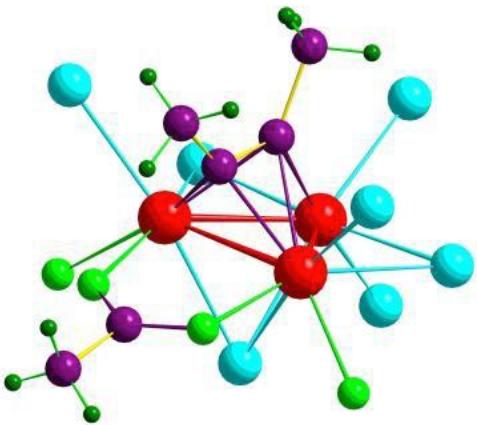
Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

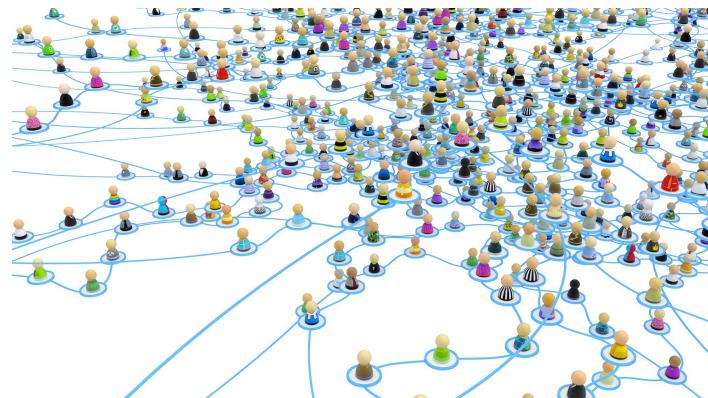
	team	coach	Y d a	ball	score	game	u	w	l	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2		
Document 2	0	7	0	2	1	0	0	3	0	0		
Document 3	0	1	0	0	1	2	2	0	3	0		

Types of Data Sets: (2) Graphs

- Transportation network
- World Wide Web

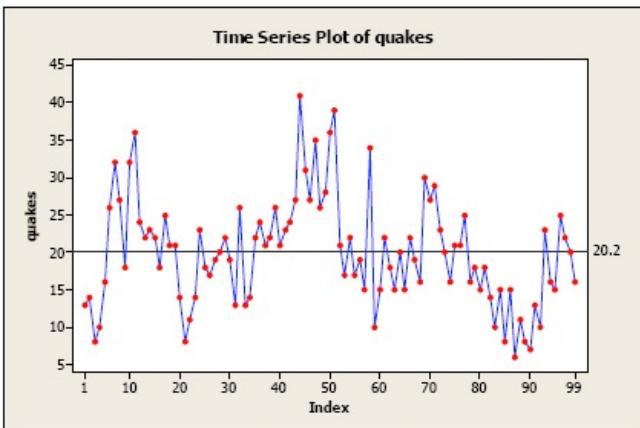


- Molecular Structures
- Social or information networks



Types of Data Sets: (3) Ordered Data

- Video data: sequence of images
- Temporal data: time-series



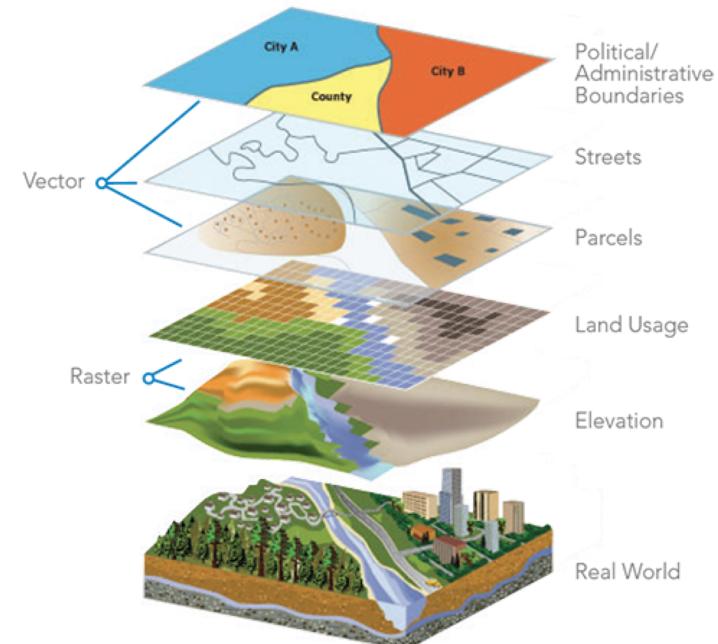
Start

	Human	Chimpanzee	Macaque
1	GTTTTGAGG - - -	ATGTTCAACAAATGCTCCCTTCATTTCTCTTATTTACAGACCTGCCGCA	
2	GTTTTGAGG - - -	ATGTTCAATAAAATGCTGCTTTCATCTCTTATTTACAGACCTGCCGCA	
3	GTTTTGAGG - - -	ATGCTCAATAAAATGCTCCCTTCATTCCTCCATTACAAAGTTGCCGCA	
4	GACATTTCCTGCTAGCAGCCTTGCTATTATCTGTTTCTAAACTTGTAGTAATTGAGGT		
5	GACATTTCCTGCTAGCAGCCTTGCTATTATCTGTTTCTAAACTTGTAGTAATTGAGGT		
6	GATCTGGAGACTAA-CTCTGAAATAAAAGCTGATTATTTTTTTCTCAAACAA		
7	GATCTGGAGACTAA-CTCTGAAATAAAAGCTGATTATTTTTCTCAAACAA		
8	TATCTGGAGACTAAACTCTGAAATAAAAGCTGATTATTTTTCTCAAACAA		
9	CAGAAATCGATTAGCAAAATTAGCTTAAAGATAATTATTTACATTTCATAATTCCTCA		
10	CAGAAATCGATTAGCAAAATTAGCTTAAAGATAATTATTTACATTTCATAATTCCTCA		
11	CAGAAATCGATTAGCAAAATTAGCTTAAAGATAATTATTTACATTTCATAATTCCTCA		
12	CCCTGAAGTTGATGTTGAGCGTATGTCACCTTCATAAAAGCCAGGTATACAC- - -TTATG		
13	CCCTGAAGTTGATGTTGAGCGTATGTCACCTTCATAAAAGCCAGGTATACAC- - -TTATG		
14	GACAGGTAAAGTAAAAACATATTTATCTACGTTTGTCCAAATTTAAATTTC		
15	GACAGGTAAAGTAAAAACATATTTATCTACGTTTGTCCAAATTTAAATTTC		
16	GACAGGTAAAGTAAAAACATATTTATCTACGTTTGTCCAAATTTAAATTTC		
17	AACCTGTTGCGCGTGTGGTAA - - - TGTAAAAACAAACTCAGTACAA		
18	AACCTGTTGCGCGTGTGGTAA - - - TGTAAAAACAAACTCAGTACAA		
19	AACCTGTTGCGCGTGTGGTAA - - - CATAAAACAAATTCAGTACAA		

- Sequential Data: transaction sequences
- Genetic sequence data

Types of Data Sets: (4) Spatial, image and multimedia Data

- Spatial data: maps



- Image data:
- Video data:

Characteristics of Structured Data

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

Data Objects

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

Attributes

- **Attribute (or dimensions, features, variables)**
 - A data field, representing a characteristic or feature of a data object.
 - E.g., customer _ID, name, address
- Types:
 - Nominal (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., Covid test (positive vs. negative)
 - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
 - Values have a meaningful order (ranking) but magnitude between successive values is not known
 - Size = {small, medium, large}, grades, army rankings

Numeric Attribute Types

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

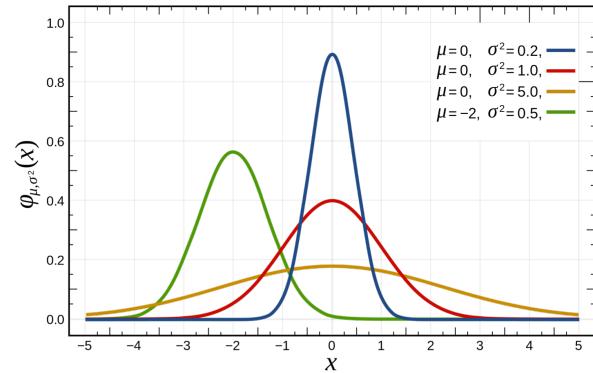
Discrete vs. Continuous Attributes

- **Discrete Attribute**
 - Has only a finite or countably infinite set of values
 - E.g., zip codes, profession, or the set of words in a collection of documents
 - Sometimes, represented as integer variables
 - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
 - Has real numbers as attribute values
 - E.g., temperature, height, or weight
 - Practically, real values can only be measured and represented using a **finite number of digits**
 - Continuous attributes are typically represented as floating-point variables

Statistical Descriptions of Data

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, ...
- 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.

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

- 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

Approximate
median



Sum before the median interval

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

Low interval limit

Interval width ($L_2 - L_1$)

Measuring the Central Tendency: (3) Mode

- Mode: Value that occurs most frequently in the data

- Unimodal

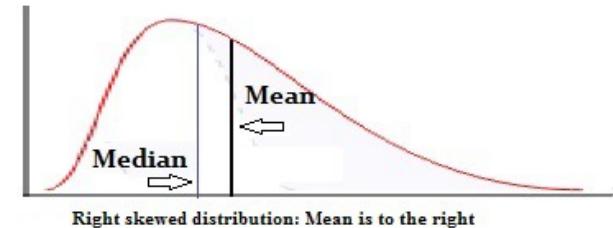
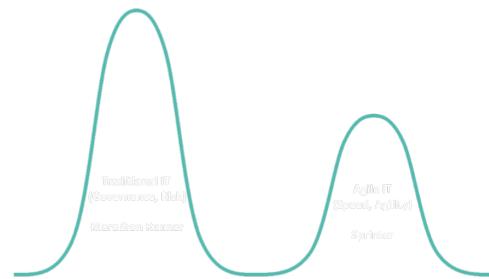
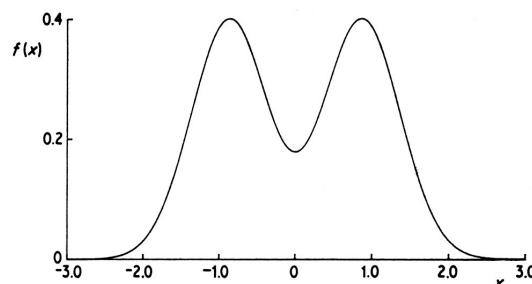
- Empirical formula:

$$\text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median})$$

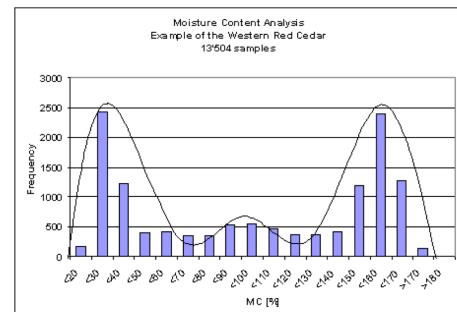
- Multi-modal

- Bimodal

- Trimodal

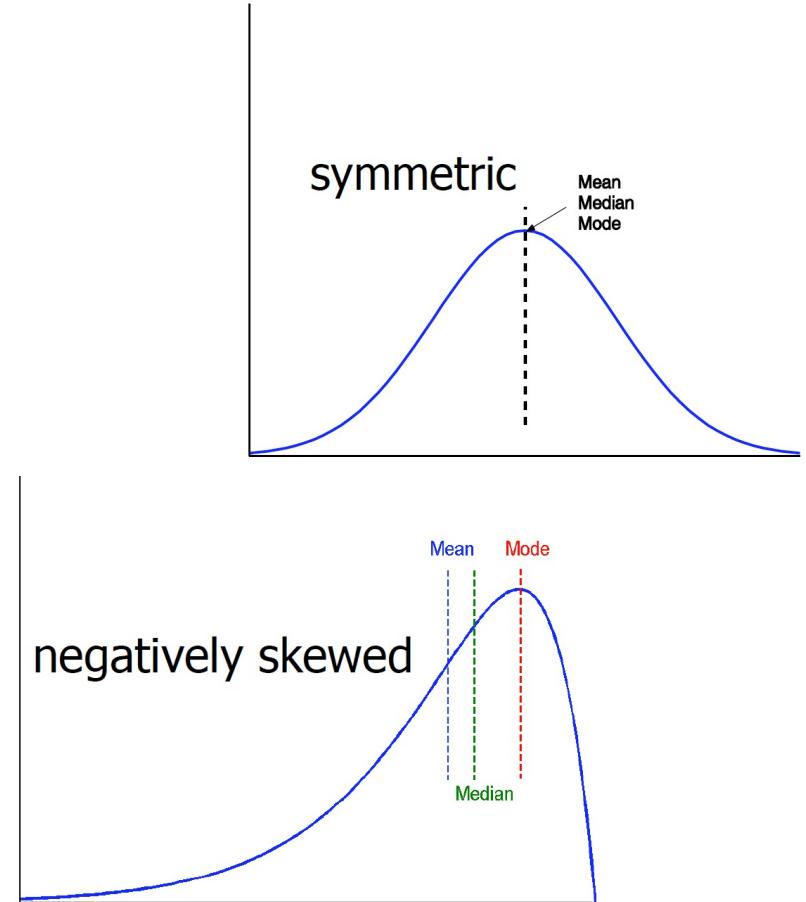
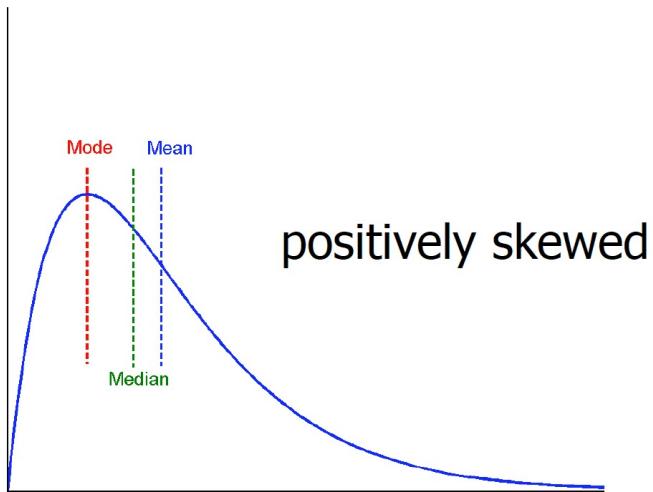


Right skewed distribution: Mean is to the right



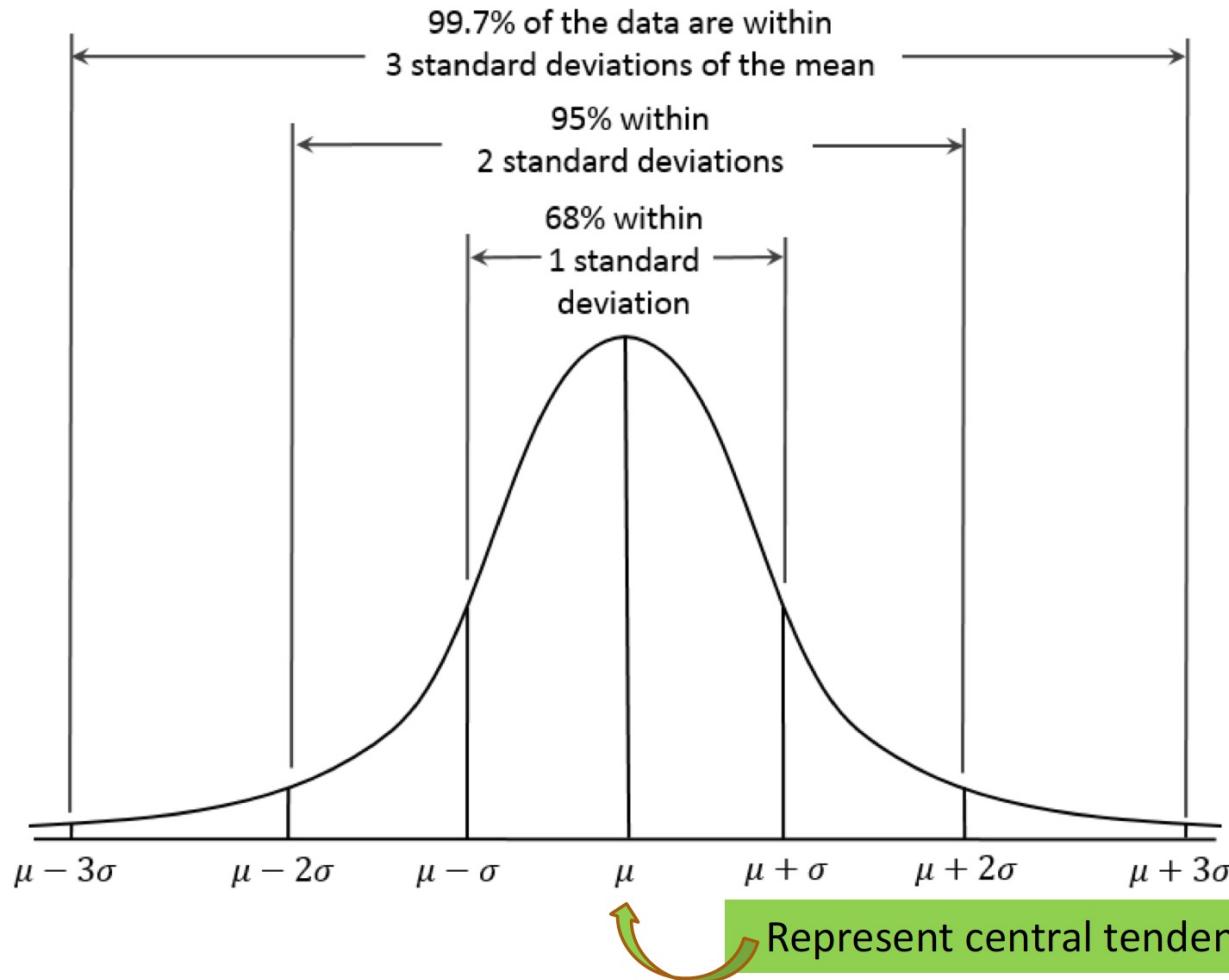
Symmetric vs. Skewed Data

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



Properties of Normal Distribution Curve

← ----- Represent data dispersion, spread ----- →



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?

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

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

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

Note: The subtle difference of formulae for sample vs. population

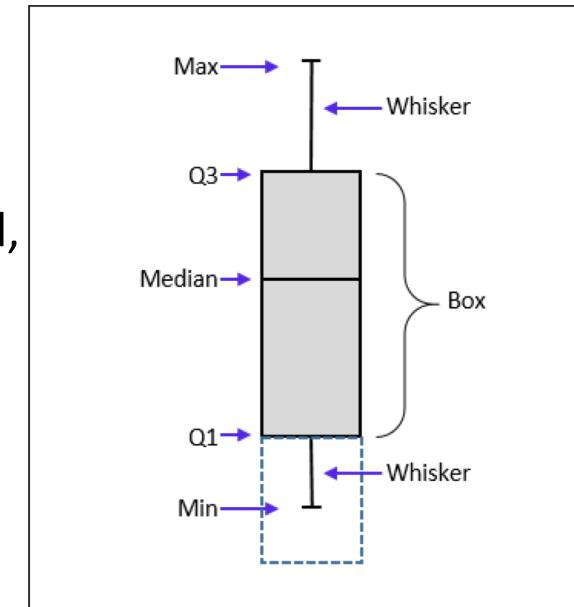
- n : the size of the sample
- N : the size of the population

Graphic Displays of Basic Statistical Descriptions

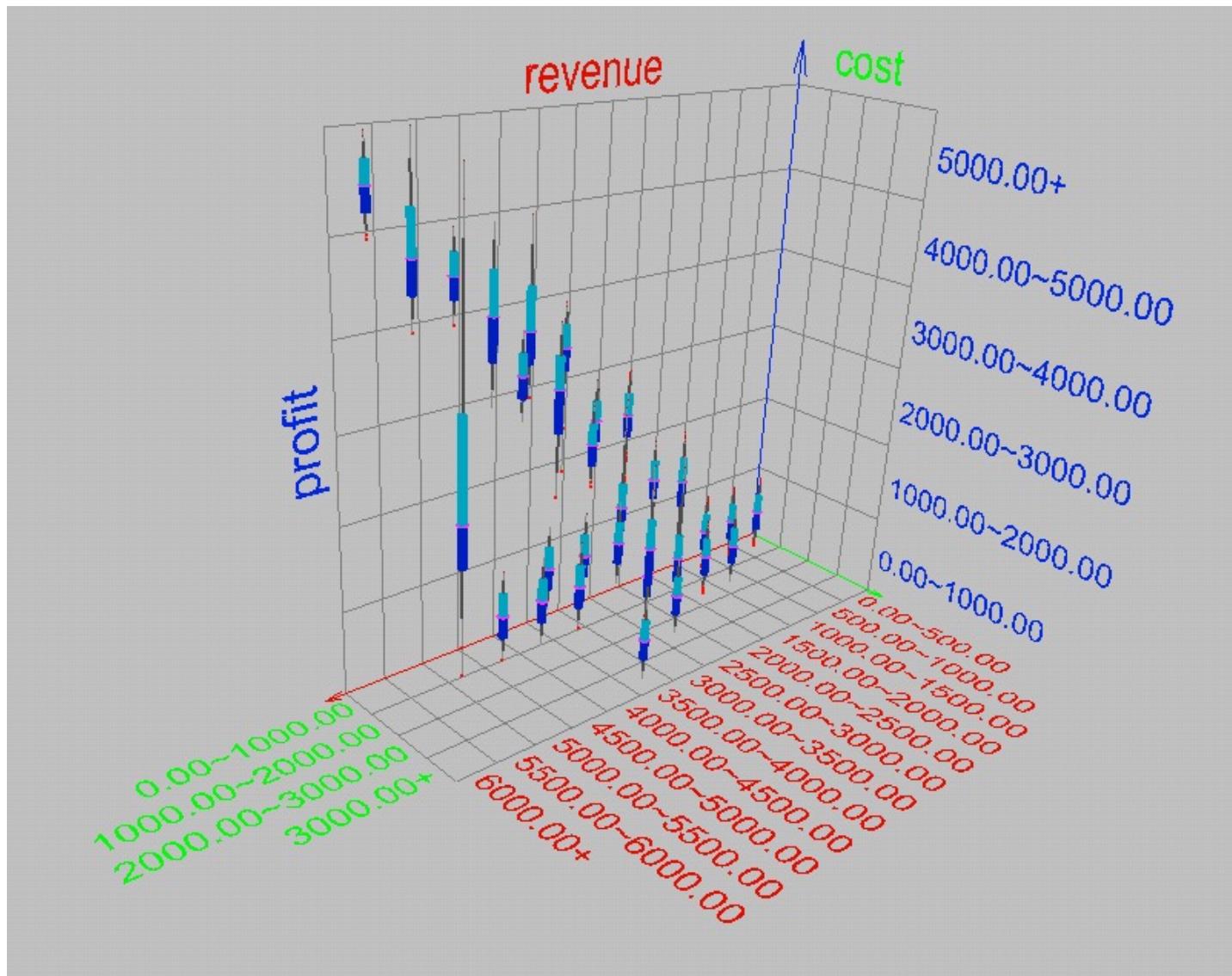
- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately $100 f_i$ % of data are $\leq x_i$
- **Quantile-quantile plot**: graphs the quantiles of one univariant 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: Quartiles & Boxplots

- **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
- **Inter-quartile range**: $IQR = Q_3 - Q_1$
- **Five number summary**: min, Q_1 , median, Q_3 , max
- **Boxplot**: 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 $1.5 \times IQR$

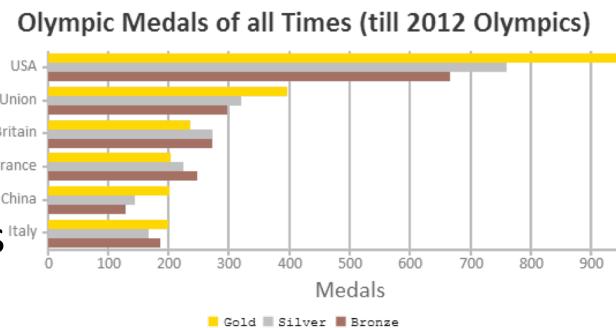
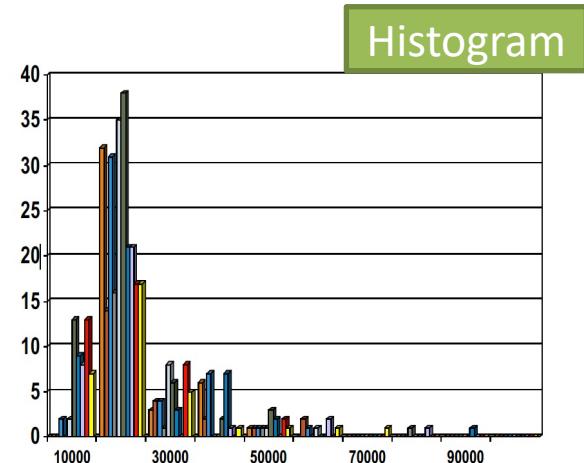


Visualization of Data Dispersion: 3-D Boxplots



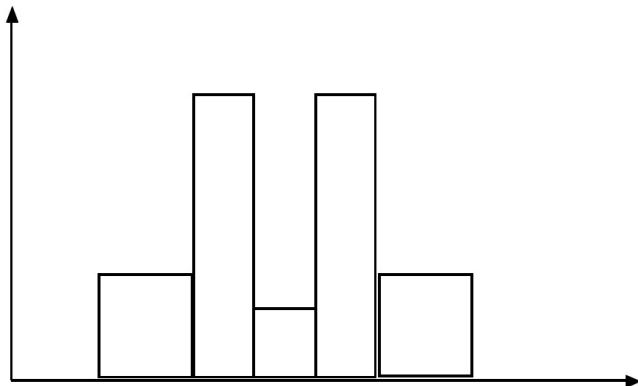
Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- Differences 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

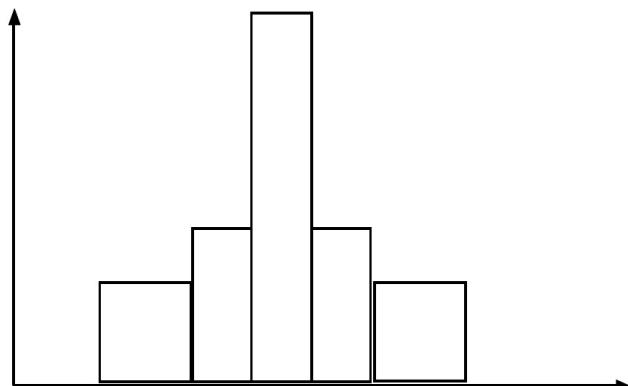


Bar Chart

Histograms Often Tell More than Boxplots



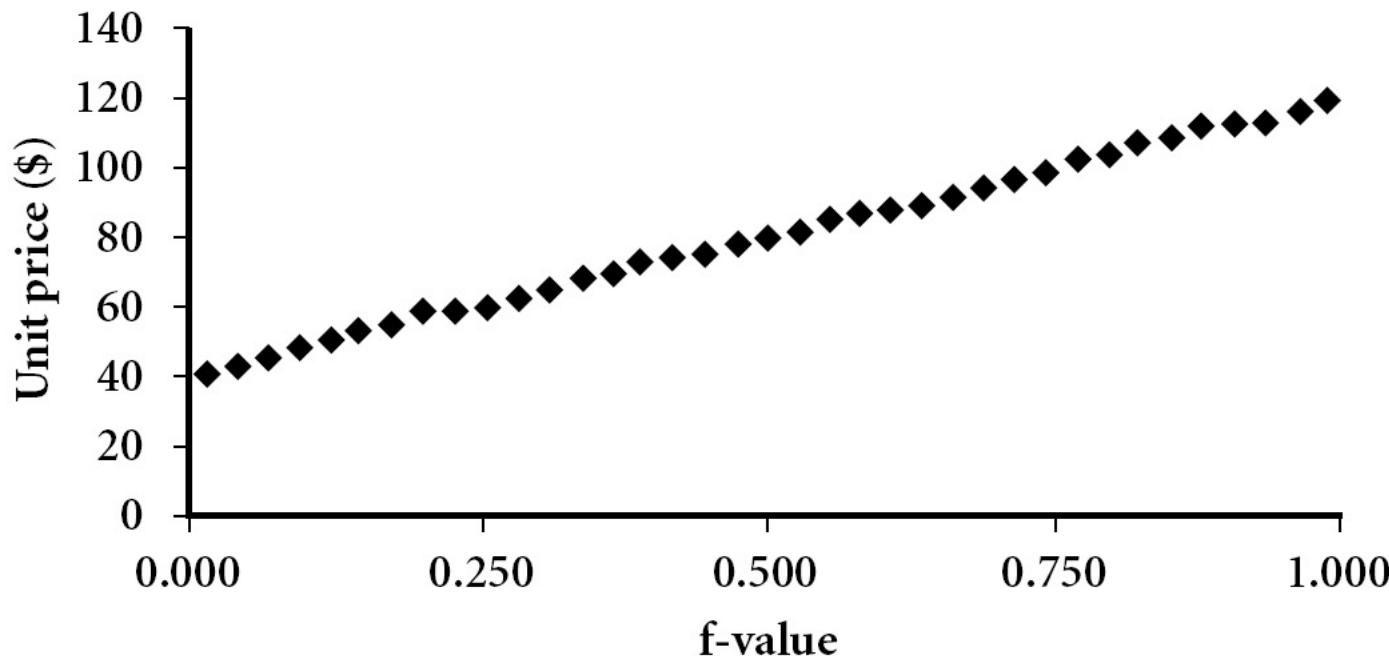
- The two histograms shown in the left may have the same boxplot representation
 - The same values for: min, Q1, median, Q3, max



- But they have rather different data distributions

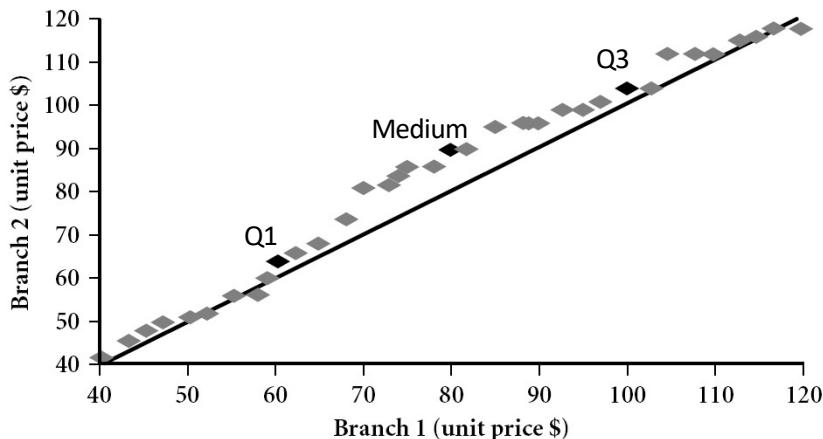
Quantile Plot

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



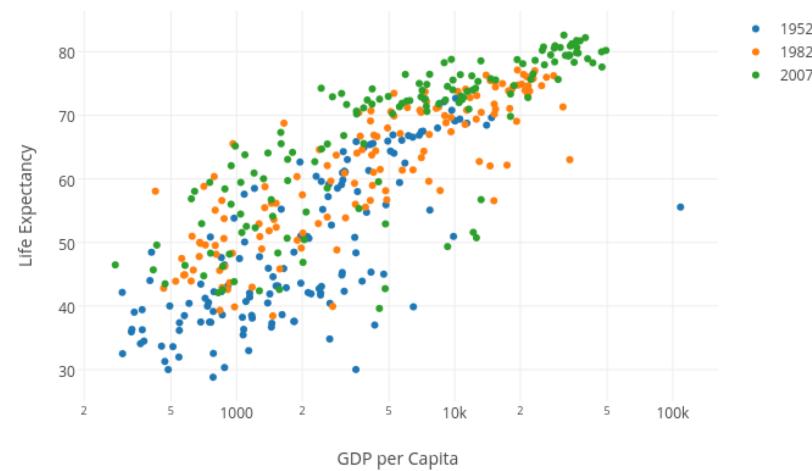
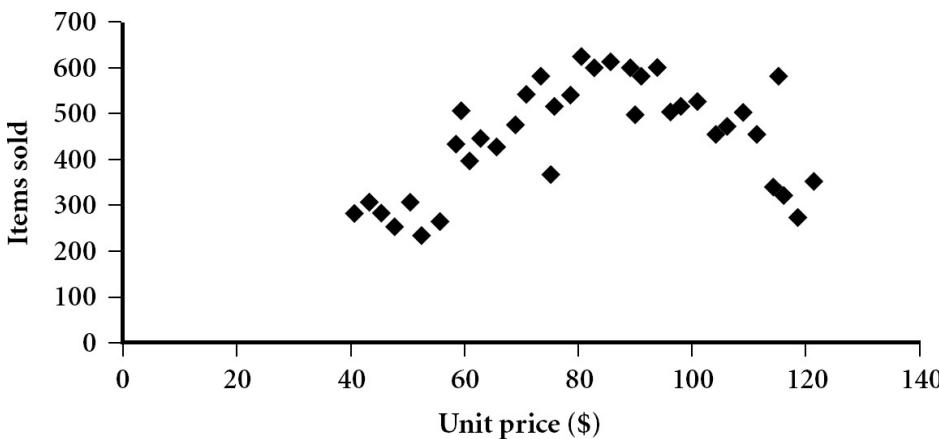
Quantile-Quantile (Q-Q) Plot

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

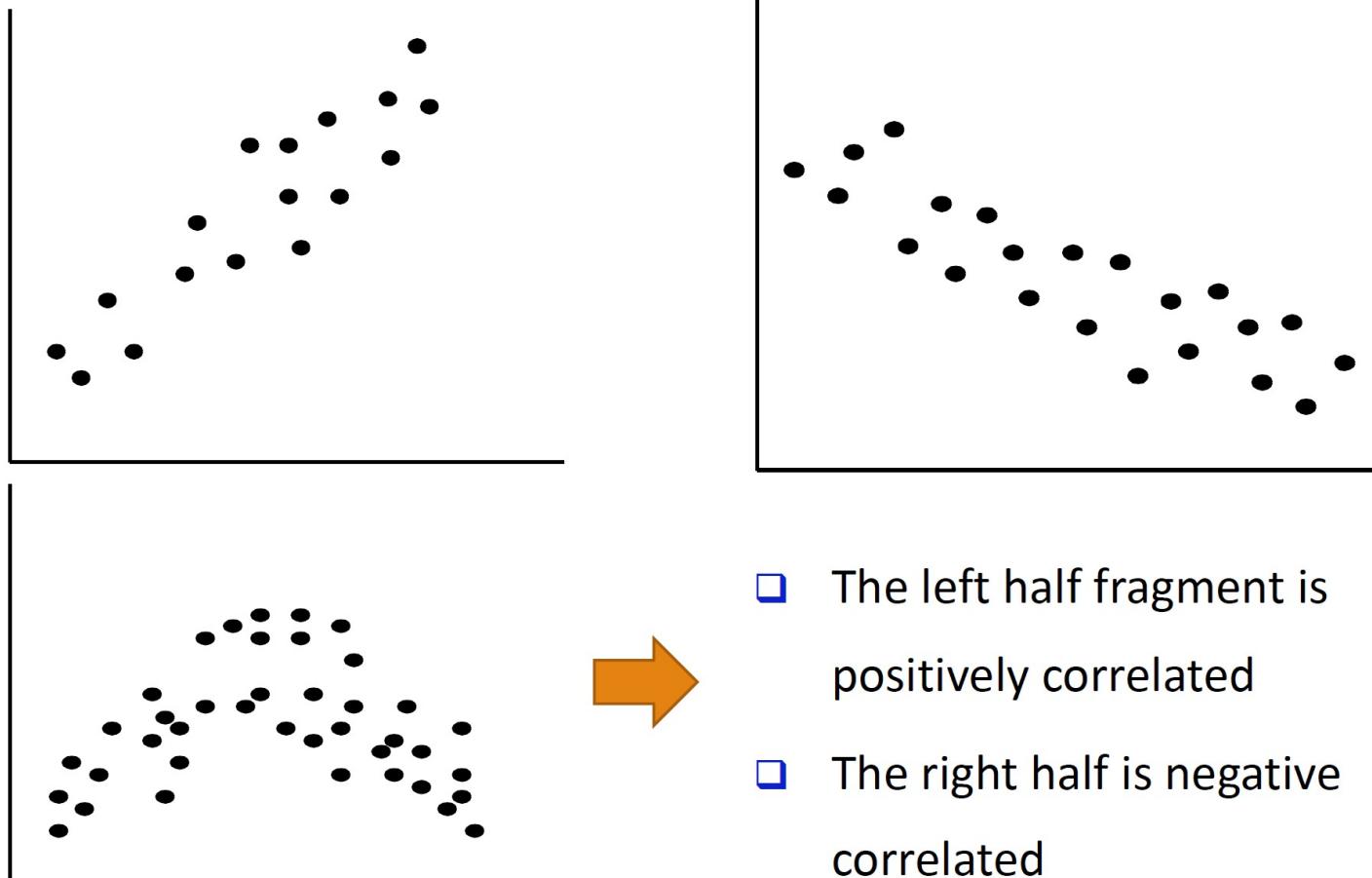


Scatter plot

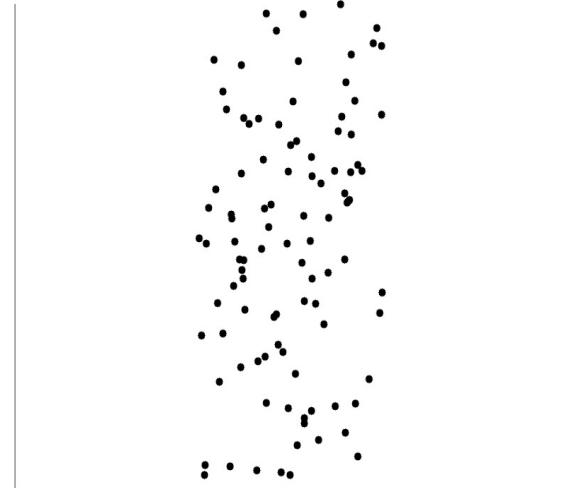
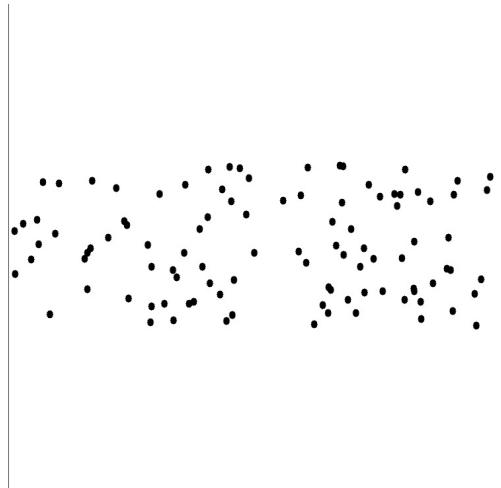
- Provides a first look at bivariate data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data



Uncorrelated Data



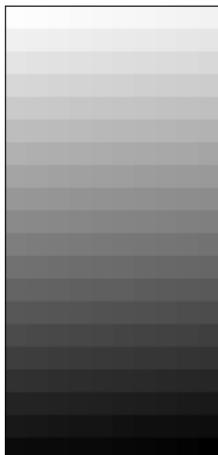
Data Visualization

Data Visualization

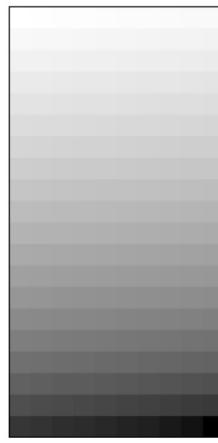
- Why data visualization?
 - Gain insight into an information space by mapping data onto graphical primitives
 - Provide qualitative overview of large data sets
 - Search for patterns, trends, structure, irregularities, relationships among data
 - Help find interesting regions and suitable parameters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

Pixel-Oriented Visualization Techniques

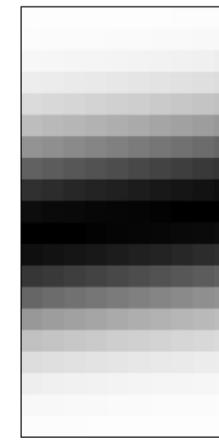
- For a data set of m dimensions, create m windows on the screen, one for each dimension
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values



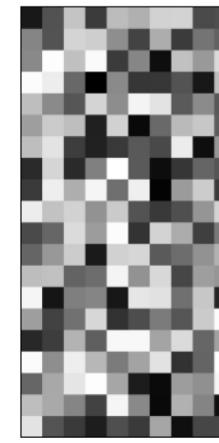
(a) Income



(b) Credit Limit



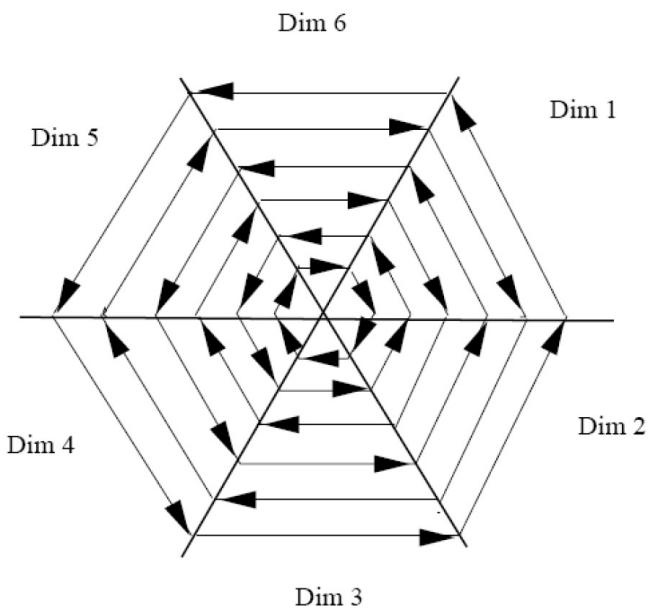
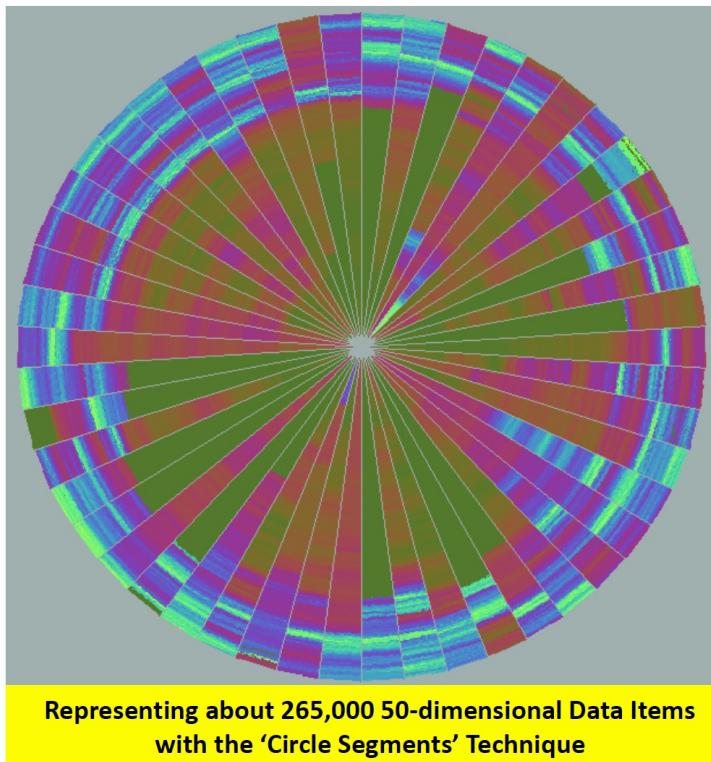
(c) transaction volume



(d) age

Laying Out Pixels in Circle Segments

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



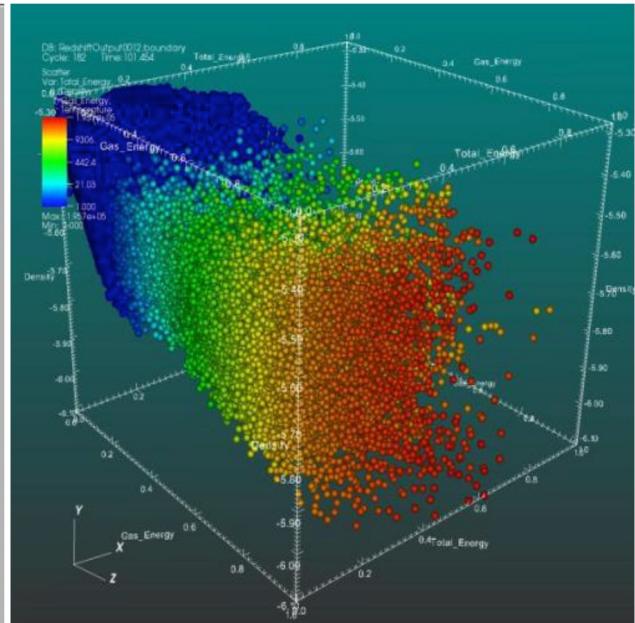
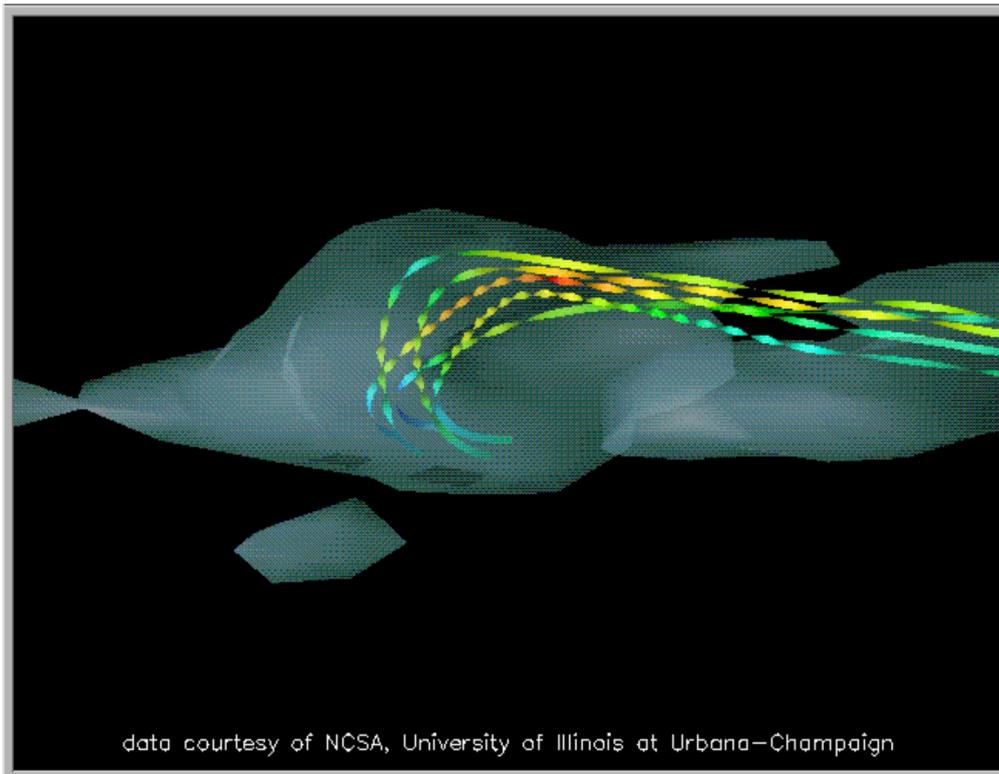
(b) Laying out pixels in circle segment

Geometric Projection Visualization Techniques

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

Direct Data Visualization

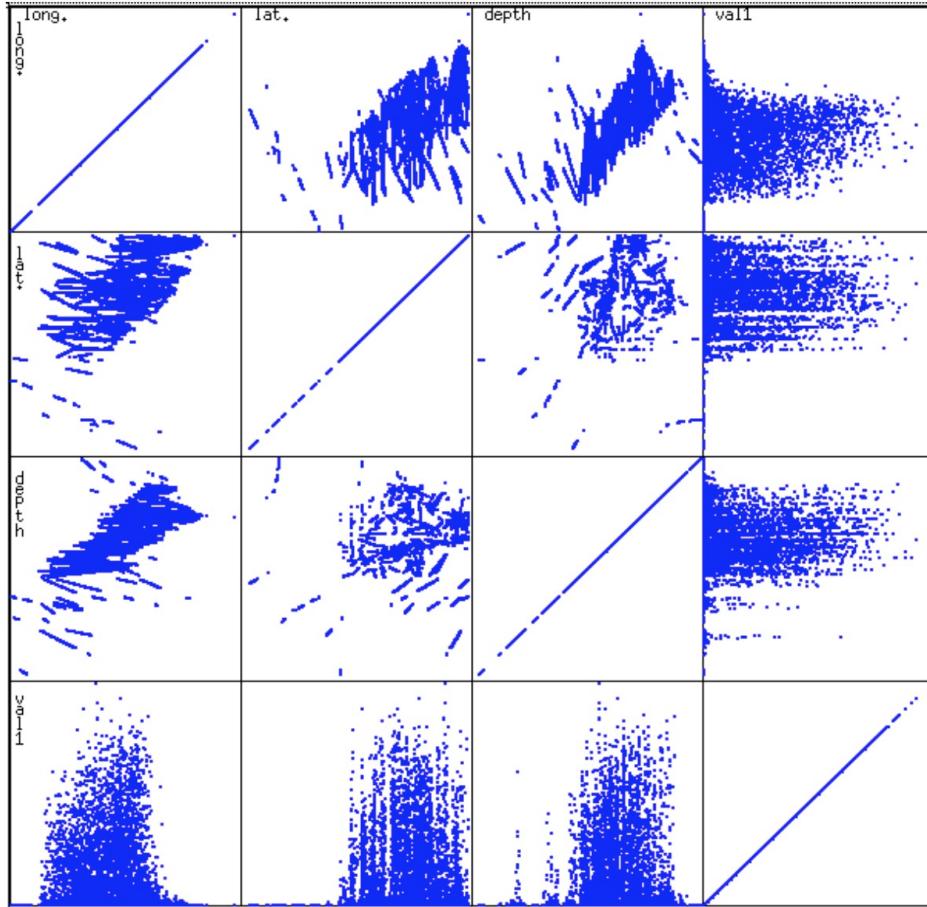
Ribbons with Twists Based on Vorticity



From Wiki: Scatter plot: A 3D scatter plot to visualize multivariate data

Scatterplot Matrices

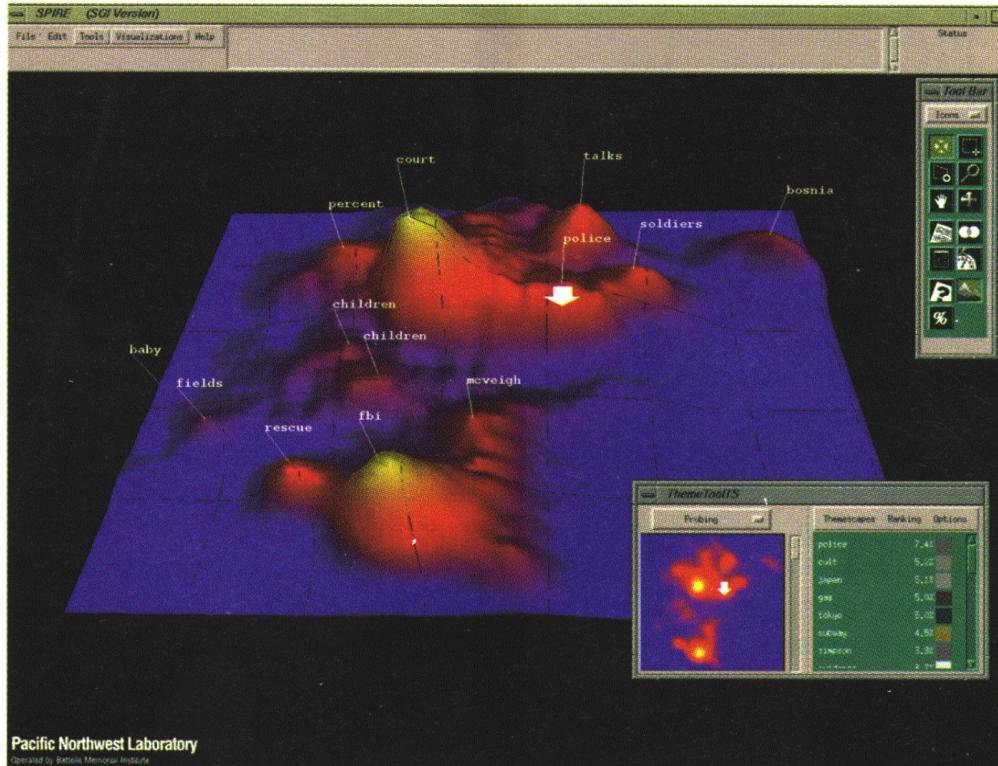
Used by permission of M. Ward, Worcester Polytechnic Institute



- Matrix of scatterplots (x-y-diagrams) of the k-dim. data
- A total of $k(k-1)/2$ distinct scatterplots

Landscapes

Used by permission of B. Wright, Visible Decisions Inc.

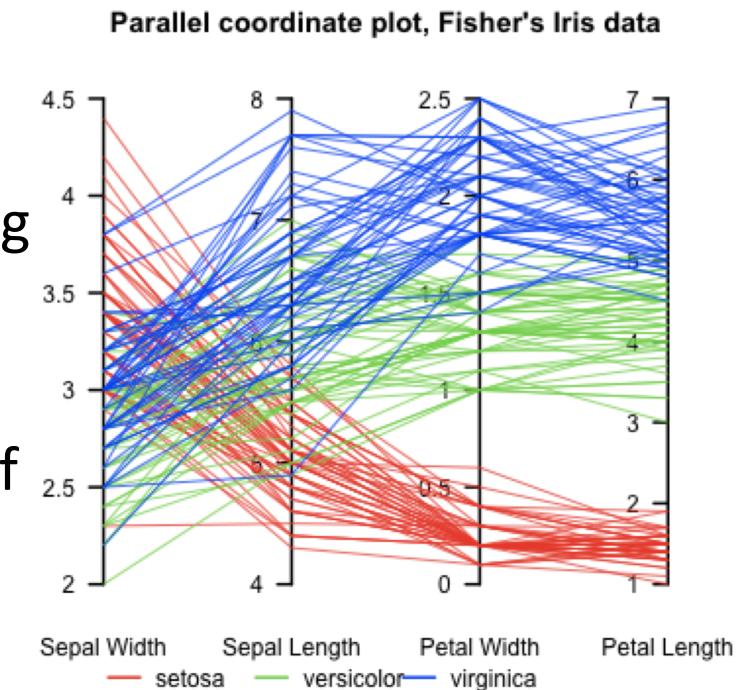


news articles visualized as a landscape

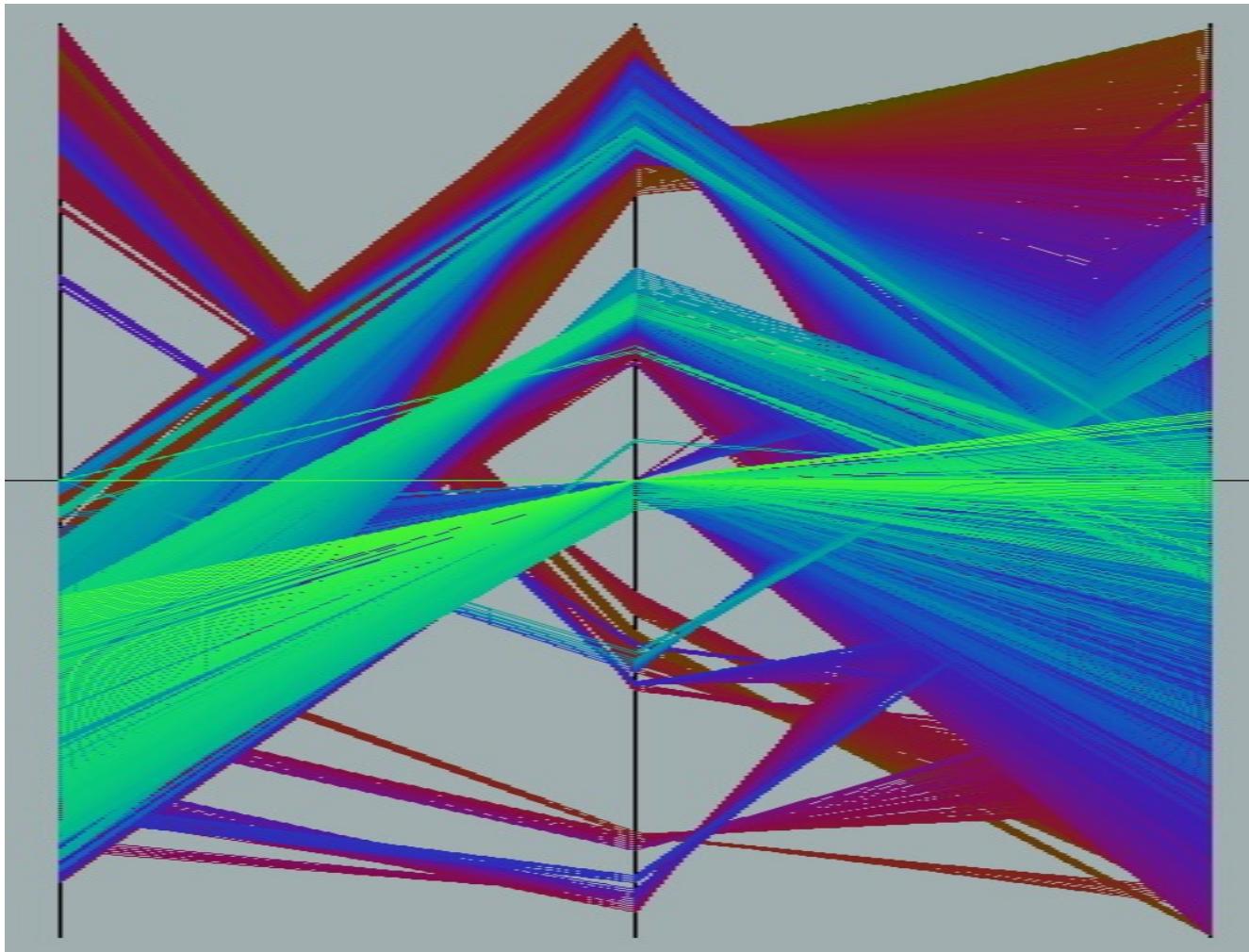
- Visualization of the data as perspective landscape
- The data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

Parallel Coordinates

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



Parallel Coordinates of a Data Set

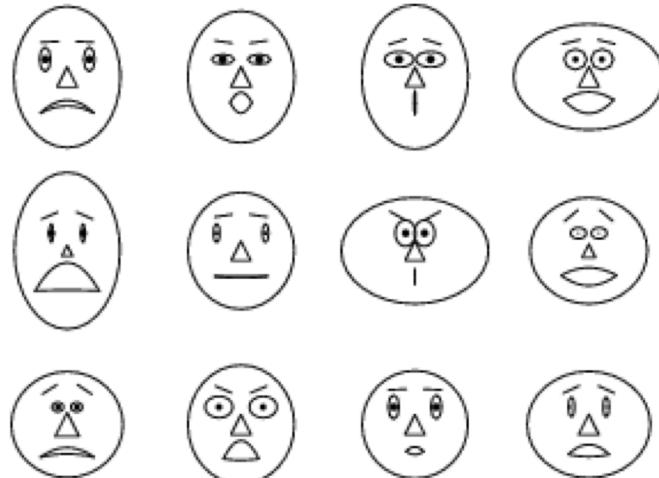


Icon-Based Visualization Techniques

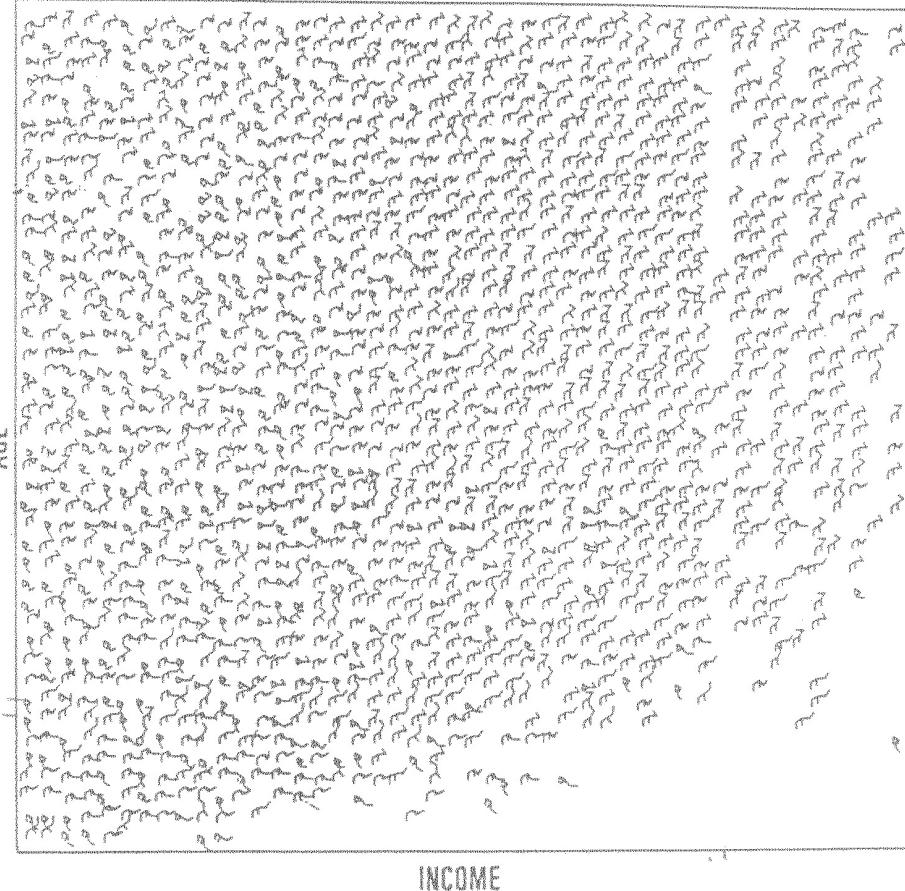
- Visualization of the data values as features of icons
- Typical visualization methods
 - Chernoff Faces
 - Stick Figures
- General techniques
 - Shape coding: Use shape to represent certain information encoding
 - Color icons: Use color icons to encode more information
 - Tile bars: Use small icons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

Chernoff Faces

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



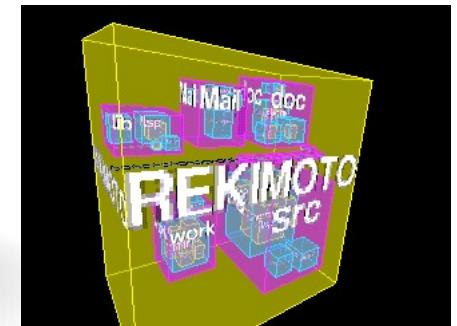
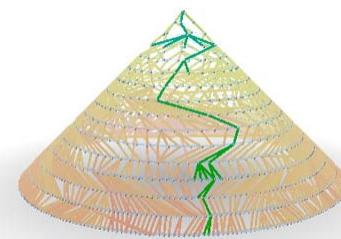
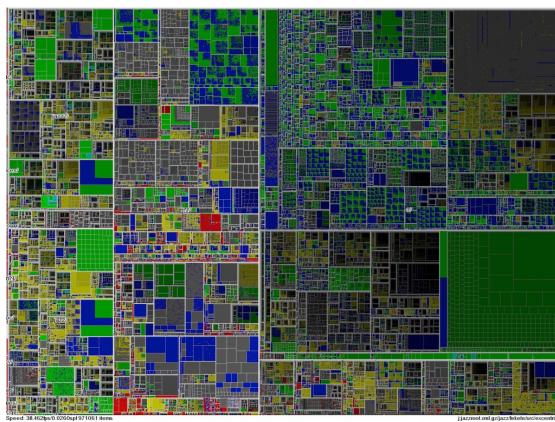
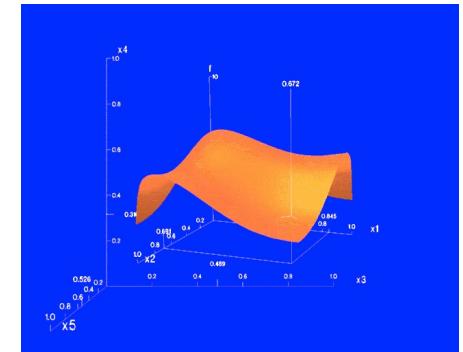
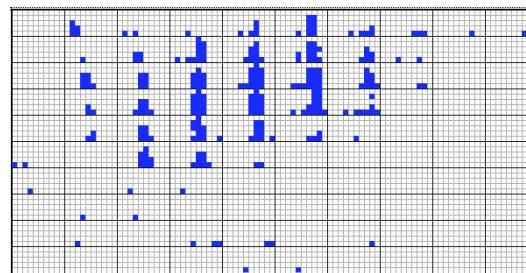
Stick Figure



- A census data figure showing age, income, gender, education, etc.
- A 5-piece stick figure (1 body and 4 limbs w. different angle/length)

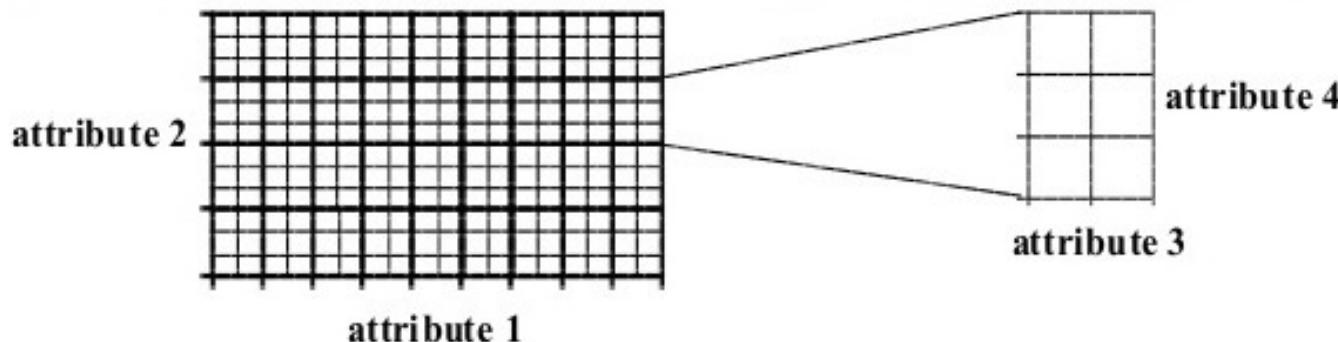
Hierarchical Visualization Techniques

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



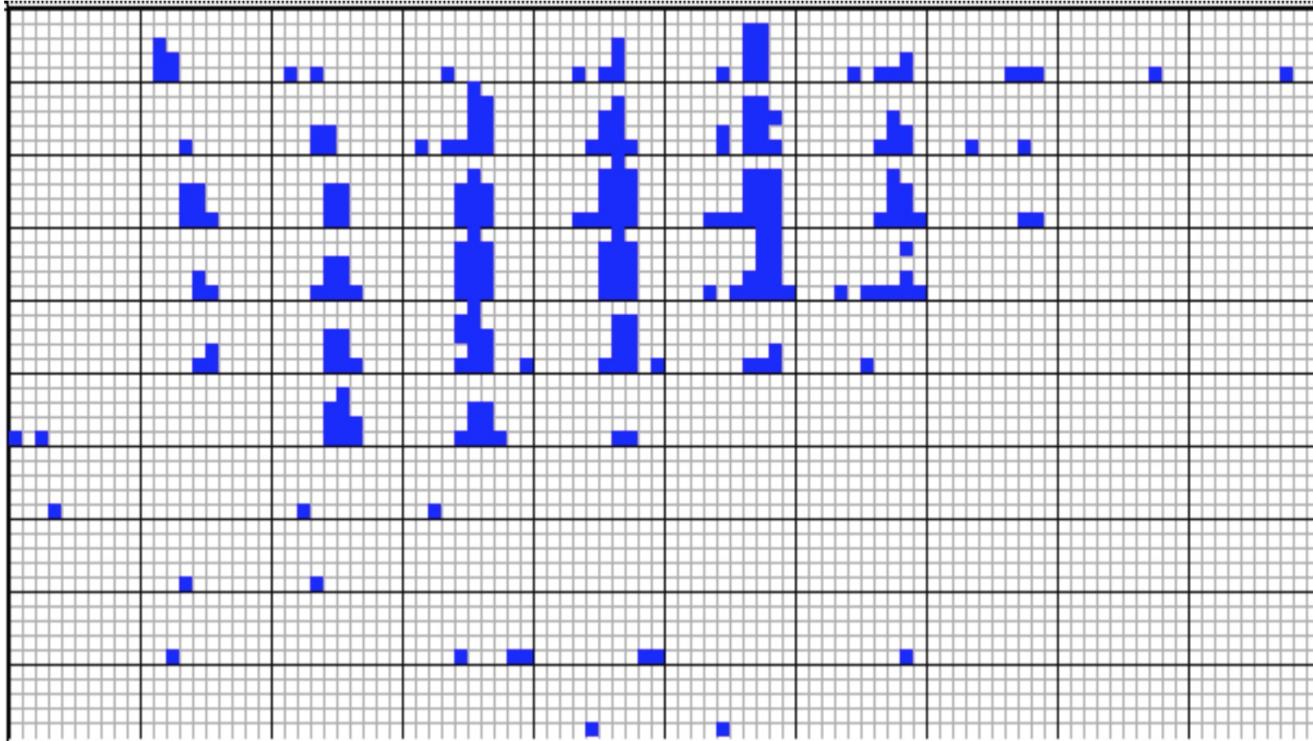
Dimensional Stacking

- Partitioning of the n-dimensional attribute space in 2-D subspaces, which are ‘stacked’ into each other
- Partitioning of the attribute value ranges into classes. The important attributes should be used on the outer levels.
- Adequate for data with ordinal attributes of low cardinality
- But, difficult to display more than nine dimensions
- Important to map dimensions appropriately



Dimensional Stacking

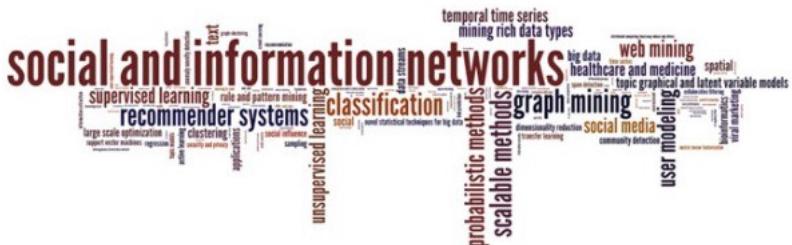
Used by permission of M. Ward, Worcester Polytechnic Institute



Visualization of oil mining data with longitude and latitude mapped to the outer x-, y-axes and ore grade and depth mapped to the inner x-, y-axes

Visualizing Complex Data and Relations: Tag Cloud

- Tag cloud: Visualizing user-generated tags
 - The importance of tag is represented by font size/color
 - Popularly used to visualize

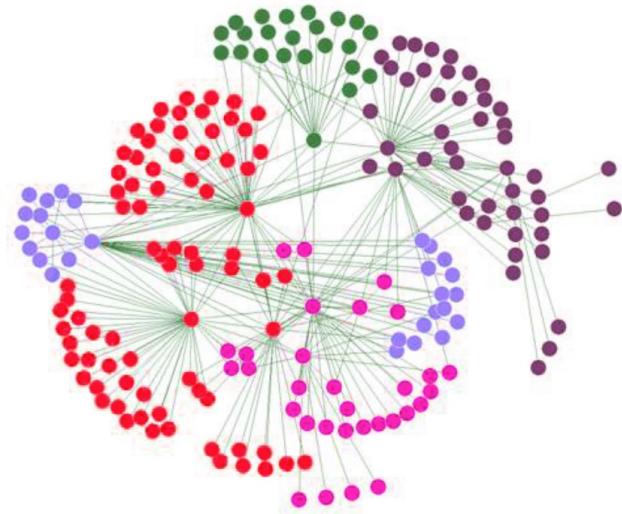


KDD 2013 Research Paper Title Tag Cloud

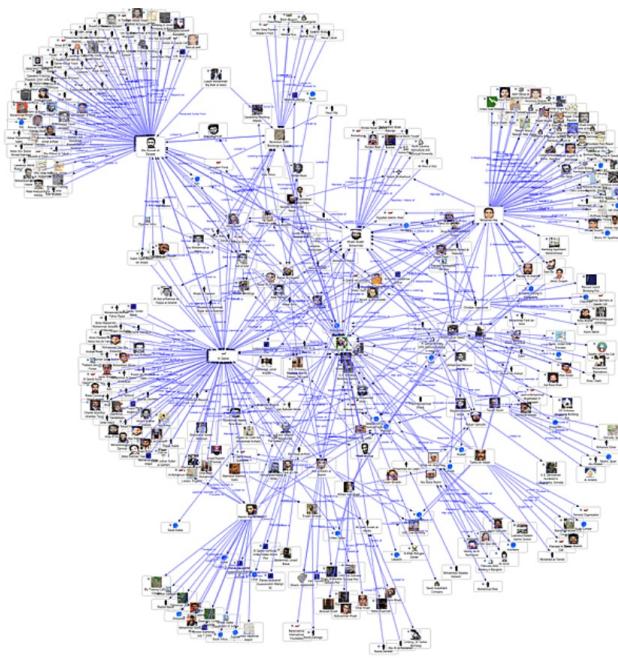


Newsmap: Google News Stories in 2005

-
- Visualizing non-numerical data: social networks



A typical network structure



A social network

Measuring Data Similarity and Correlation

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 $d(i, j)$
- (typically metric)
 - 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

$$D = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1l} \\ x_{21} & x_{22} & \dots & x_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nl} \end{pmatrix}$$

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

Standardizing Numeric Data

- Z-score:
$$z = \frac{x - \mu}{\sigma}$$
 - X: raw score to be standardized, μ : mean of the population, σ : 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

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

where

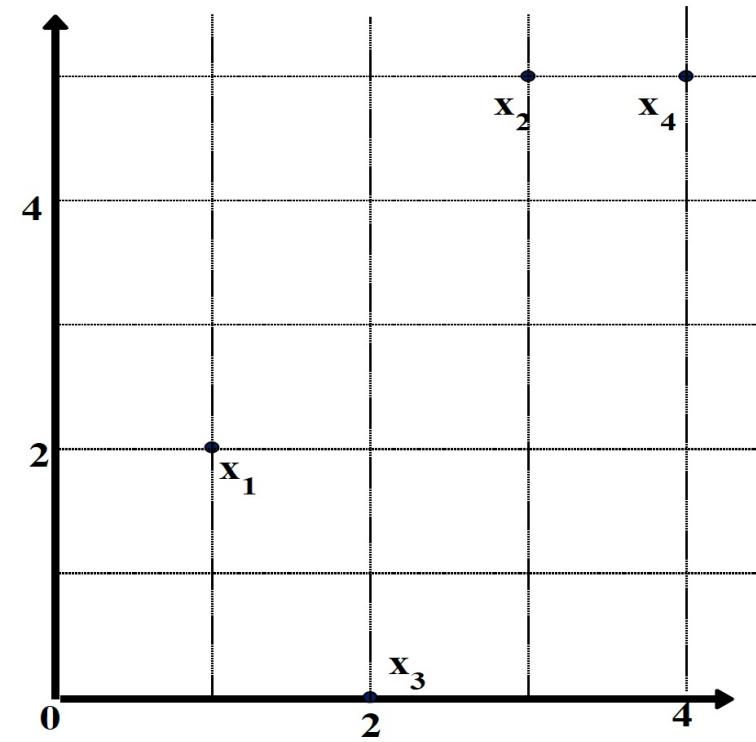
$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf})$$

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- standardized measure (z-score):

- Using mean absolute deviation is more robust than using standard deviation

Example: Data Matrix and Dissimilarity Matrix



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			5.1	0
x_4				0

Distance on Numeric Data: 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 also 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**

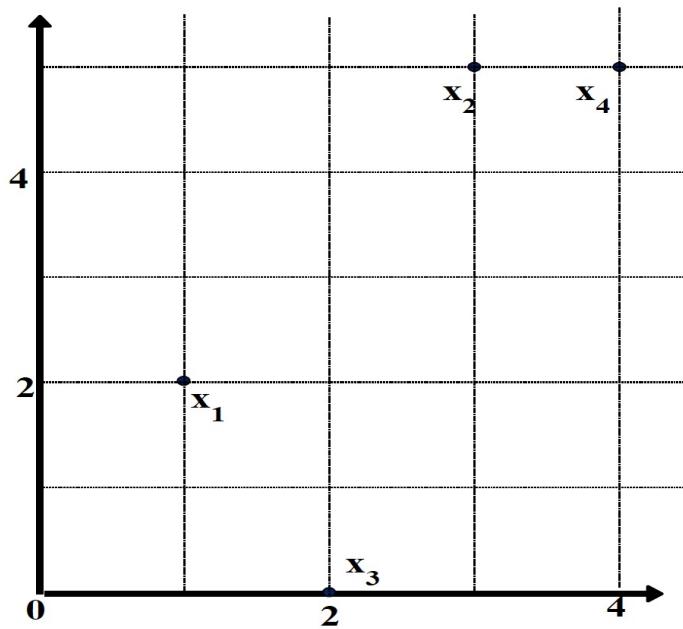
$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{il} - x_{jl}|^2}$$

- $p \rightarrow \infty$: (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	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):

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

Example: Dissimilarity between Asymmetric Binary Variables

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

		Mary		
		1	0	Σ_{row}
Jack		1	2	2
		0	1	4
Σ_{col}		3	3	6

- Gender is a symmetric attribute (not counted in)
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0
- Distance: $d(i, j) = \frac{r + s}{q + r + s}$

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

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

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

		Jim		
		1	0	Σ_{row}
Jack		1	1	2
		0	1	4
Σ_{col}		2	4	6

		Mary		
		1	0	Σ_{row}
Jim		1	1	2
		0	2	4
Σ_{col}		3	3	6

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 p 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: $r_{if} \in \{1, \dots, M_f\}$
 - Map the range of each variable onto $[0, 1]$ by replacing i -th object in the f -th variable by
$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$
 - 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 Type

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

Example: Calculating Cosine Similarity

- Calculating Cosine Similarity:

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|}$$
$$sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

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

- Ex: Find the similarity between documents 1 and 2.

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

- First, calculate vector dot product

$$d_1 \cdot d_2 = 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 1 + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25$$

- Then, calculate $\|d_1\|$ and $\|d_2\|$

$$\|d_1\| = \sqrt{5 \times 5 + 0 \times 0 + 3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0} = 6.481$$

$$\|d_2\| = \sqrt{3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1} = 4.12$$

- Calculate cosine similarity: $\cos(d_1, d_2) = 25 / (6.481 \times 4.12) = 0.94$

Correlation Analysis (for Categorical Data)

- χ^2 (chi-square) test:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

observed
↓
expected

- Null hypothesis: The two distributions are independent
- The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count
 - The larger the χ^2 value, the more likely the variables are related
- Note: Correlation does not imply causality
 - # of hospitals and # of car-theft in a city are correlated
 - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (X1)	200 (X2)	450
Not like science fiction	50 (X3)	1000 (X4)	1050
Sum(col.)	300	1200	1500

- Null hypothesis: The two distributions are independent
 - What does that mean?
 - The ratio between people who play chess vs not play chess is the same for both groups of like science fiction and not like science fiction
 - $X1:X2=X3:X4=300:1200$
 - $X1:X3=X2:X4=450:1050$
 - $X1+X2=450 \quad X3+X4=1050$
 - $X1+X3=300 \quad X2+X4=1200$

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

How to derive 90?
450/1500 * 300 = 90

We can reject the null hypothesis of independence at a confidence level of 0.001

- X² (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Variance for Single Variable (Numerical Data)

- The variance of a random variable X provides a measure of how much the value of X deviates from the mean or expected value of X :

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

- where σ^2 is the variance of X , σ is called standard deviation
- μ is the mean, and $\mu = E[X]$ is the expected value of X
- That is, variance is the expected value of the square deviation from the mean
- It can also be written as: $\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = E[X^2] - \mu^2 = E[X^2] - [E(x)]^2$
- Sample variance

$$s^2 = \frac{1}{n} \sum_i^n (x_i - \hat{\mu})^2$$

$$s^2 = \frac{1}{n-1} \sum_i^n (x_i - \hat{\mu})^2$$

Covariance for Two Variables

- Covariance between two variables X_1 and X_2

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

where $\mu_1 = E[X_1]$ is the respective mean or **expected value** of X_1 ;
similarly for μ_2

- Sample covariance between X_1 and X_2 : $\hat{\sigma}_{12} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$
- Sample covariance is a generalization of the sample variance:

$$\hat{\sigma}_{11} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i1} - \hat{\mu}_1)$$

- **Positive covariance**: If $\sigma_{12} > 0$
- **Negative covariance**: If $\sigma_{12} < 0$

Covariance for Two Variables

- **Independence:** If X_1 and X_2 are independent, $\sigma_{12} = 0$ but the reverse is not true
- Some pairs of random variables may have a covariance 0 but are not independent
- Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:
 - (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
- Covariance formula
$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1X_2] - \mu_1\mu_2 = E[X_1X_2] - E[X_1]E[X_2]$$
- Its computation can be simplified as: $\sigma_{12} = E[X_1X_2] - E[X_1]E[X_2]$
 - $E(X_1) = (2 + 3 + 5 + 4 + 6)/5 = 20/5 = 4$
 - $E(X_2) = (5 + 8 + 10 + 11 + 14)/5 = 48/5 = 9.6$
 - $\sigma_{12} = (2*5 + 3*8 + 5*10 + 4*11 + 6*14)/5 - 4 * 9.6 = 4$
- Thus, X_1 and X_2 rise together since $\sigma_{12} > 0$

Correlation between Two Numerical Variables

- Correlation between two variables X_1 and X_2 is the standard covariance, obtained by normalizing the covariance with the standard deviation of each variable

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

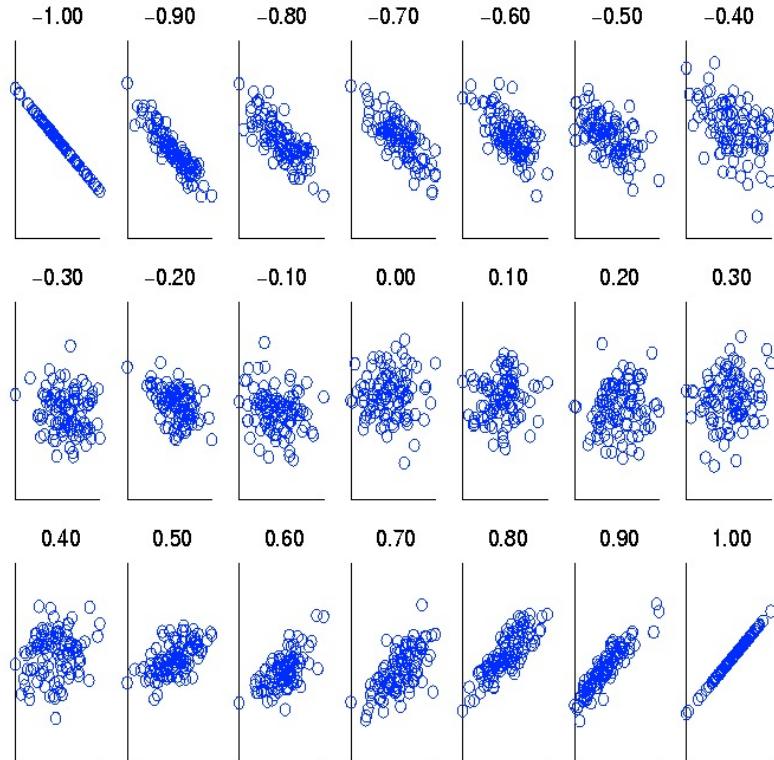
- Sample correlation for two attributes X_1 and X_2 :

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^n (x_{i2} - \hat{\mu}_2)^2}}$$

where n is the number of tuples, μ_1 and μ_2 are the respective means of X_1 and X_2 ,

- σ_1 and σ_2 are the respective standard deviation of X_1 and X_2
- If $\rho_{12} > 0$: A and B are positively correlated (X1's values increase as X2's)
- The higher, the stronger correlation
- If $\rho_{12} = 0$: independent (under the same assumption as discussed in covariance)
- If $\rho_{12} < 0$: negatively correlated

Visualizing Changes of Correlation Coefficient



- Correlation coefficient value range: $[-1, 1]$
- A set of scatter plots shows sets of points and their correlation coefficients changing from -1 to 1

Figure 5.11. Scatter plots illustrating correlations from -1 to 1 .

Covariance Matrix

- The variance and covariance information for the two variables X_1 and X_2 can be summarized as 2 X 2 covariance matrix as

$$\begin{aligned}\Sigma &= E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = E\left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} (X_1 - \mu_1 \quad X_2 - \mu_2) \right] \\ &= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\end{aligned}$$

- Generalizing it to d dimensions, we have,

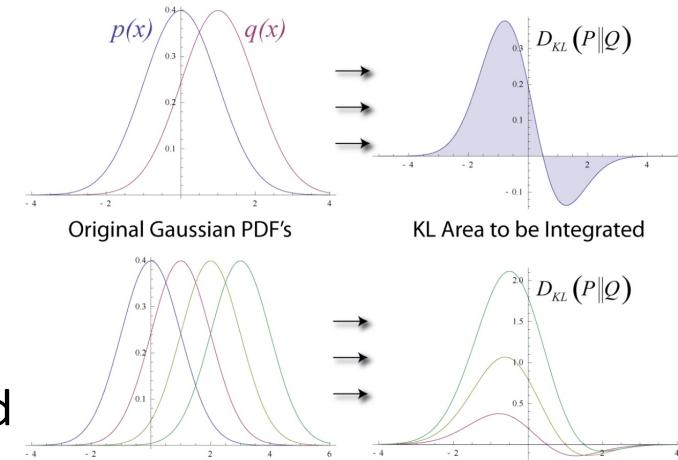
$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dd} \end{pmatrix} \quad \Sigma = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

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 $p(x)$

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

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



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 and correlation
- Above steps are the beginning of data preprocessing
- Many methods have been developed but still an active area of research