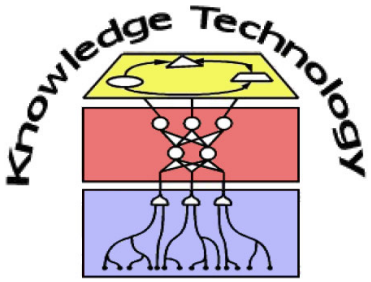
Data Mining

Lecture 2
From Data to Visualisation

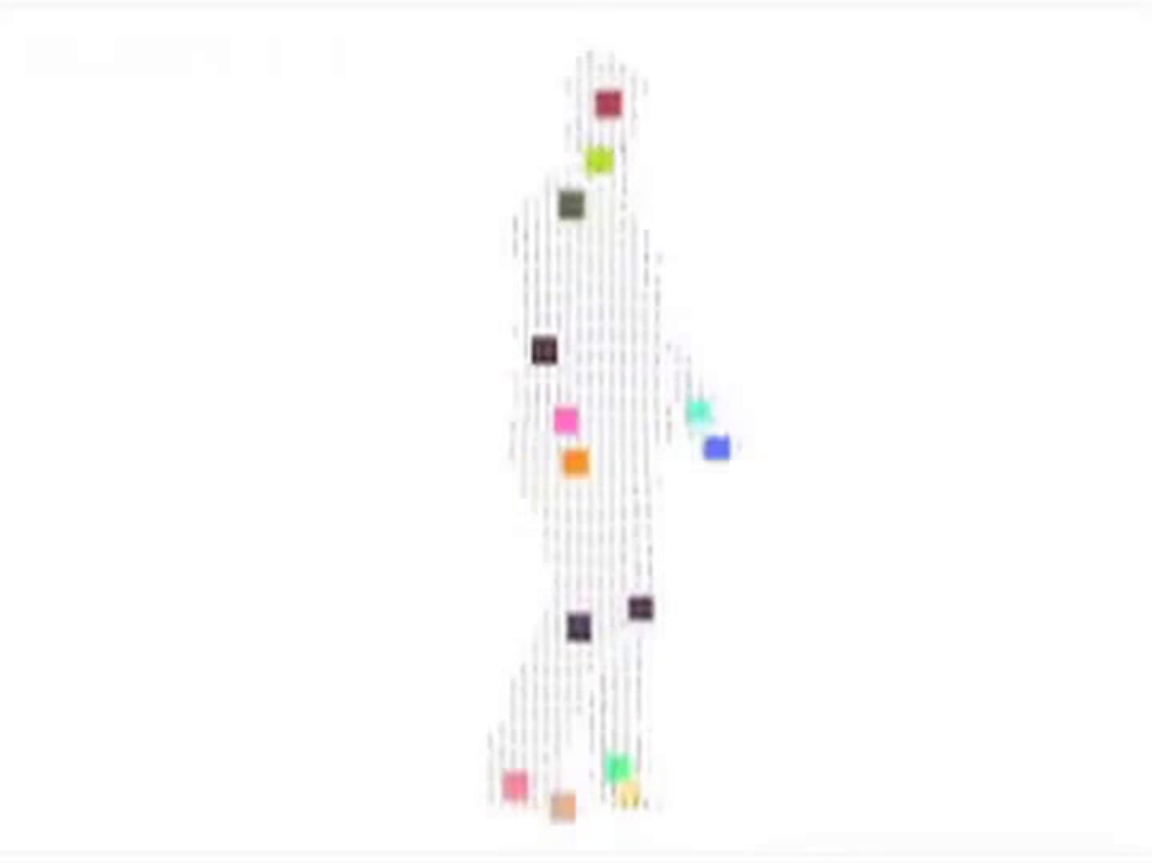


http://www.informatik.uni-hamburg.de/WTM/

Important Characteristics of structured Data

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

Motivating example: from Data to Visualisation
Similarity of Trajectories in the Kinect



http://www.newscientist.com/

Types of Data

- Structured Record
  - Relational records
  - Tables
  - Transaction data
- Sequential and semi-structured
  - Documents with text data
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures

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

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, tuples*.
- Data objects are described by ***attributes***.
- Database rows -> data objects; columns ->attributes.

Attributes

- ***Attribute*** **(**or ***dimensions***, ***features***, ***variables***): a data field, representing a characteristic or feature of a data object.
  - **E.g.**, customer _ID, name, address

- Types:
  - Nominal
  - Binary
  - Numeric: quantitative

Attribute Types

- ***Nominal***: categories, states, or "names of things"
  - *Hair_color* = {*black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- ***Binary***
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- ***Ordinal***
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings

Numeric Attribute Types

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

Attribute Types Overview

Type	Description	Examples	Operations
Nominal	Uses a label or name to distinguish one object from another.	ZIP-Code, ID, Gender	= or !=
Ordinal	Uses values to provide the ordering of objects.	Opinion, grades	< or >
Interval	Uses units of measurements, but the origin is arbitrary.	Celsius, Fahrenheit, calendar dates	+ or -
Ratio	Uses units of measurement, and the origin is not arbitrary.	Kelvin, length, counts, age, income	+, -, *, /

Discrete vs. Continuous Attributes

- ***Discrete Attribute***
  - Has only a finite or countable infinite set of values
    - **E.g.**, zip codes, profession, or set of words in collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- ***Continuous Attribute***
  - Has continuous 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

Static vs Temporal Attributes
(Another Dimension of Data Classification I)

- Some data are not changing with time and are considered as ***static data*.**
- On the other hand, there are attribute values that change with time, and this type of data we call ***dynamic*** or ***temporal data*.**
- The majority of the data mining methods and commercial data mining tools are more suitable for static data!

Experimental vs Observational Data
(Another Dimension of Data Classification II)

- **Experimental Data** (Primary, Prospective)
  - Hypothesis H
  - Design an experiment to test H
  - Collect data, infer how likely it is that H is true
  - **E.g.**, *clinical trials in medicine*

- **Observational Data** (Retrospective or Secondary)
  - Massive non-experimental data sets
    - **E.g.**, human genome, atmospheric data, retail data, etc.
  - Assumptions of experimental design no longer valid
  - Cheap compared to experimental data

Curse of Dimensionality
(Geometric Approach I)

The "*curse of dimensionality*" is due to the geometry of high-dimensional spaces.

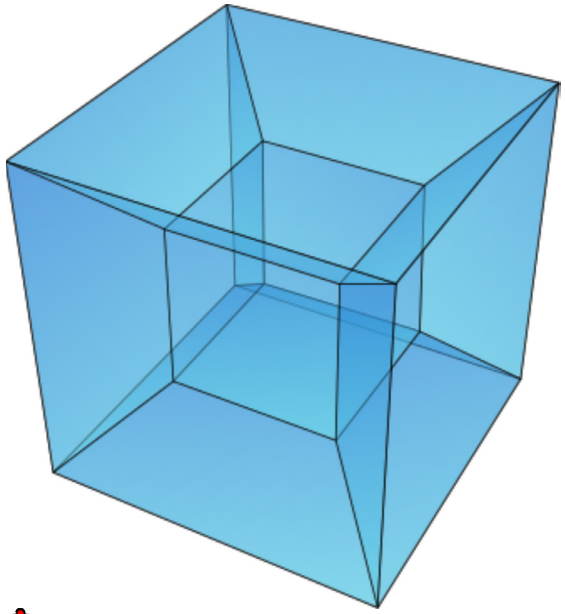
- The properties of high-dimensional spaces often appear ***counterintuitive*** because our experience with the physical world is in low-dimensional space such as space with two or three dimensions.

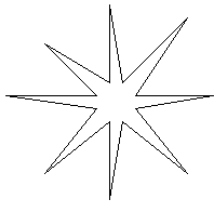
- Conceptually objects ***in high-dimensional spaces*** have a ***larger amount of surface*** area for a given volume than objects in low-dimensional spaces.

Curse of Dimensionality
(Geometric Approach II)

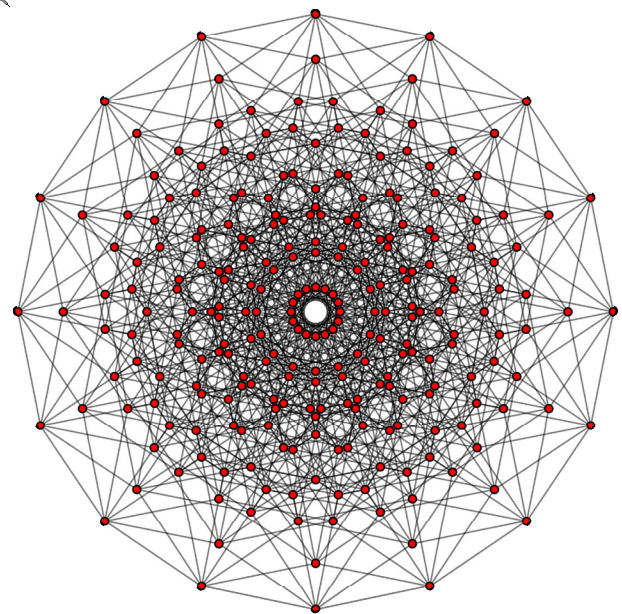
For example:

- A high-dimensional hypercube, if it could be visualized, would look like a porcupine





- As the dimensionality grows larger, the edges grow longer relative to the size of a central part of the hypercube.

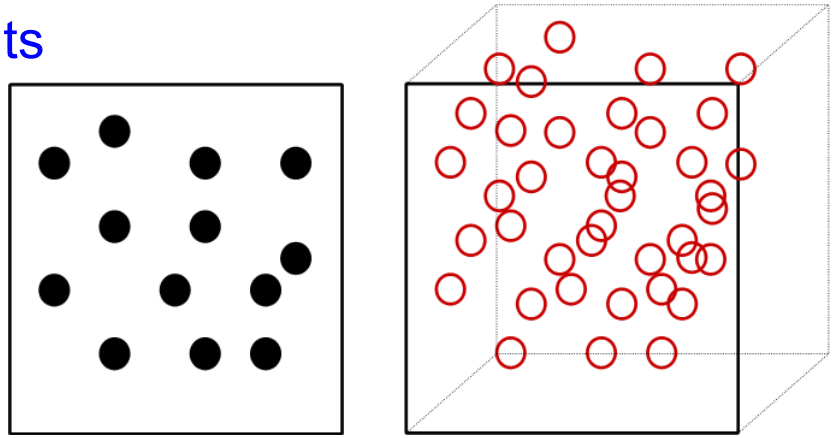


Curse of Dimensionality

- The size of a data set yielding the same density of data points in k-dimensional space, increases ***exponentially*** with dimensions

to achieve the same density of n points in k dimensions, we need n^k data points

Same density of data:



Low dimensions k dimensions

- **Example**
  - k = 1
    $\rightarrow$ n = 100 samples
  - k = 5
    $\rightarrow$ n = $100^5 = 10^{10}$ samples

Curse of Dimensionality (2)

- A ***larger radius is needed*** to enclose the same fraction of data points in a high-dimensional space.
  The ***edge length e*** of the hypercube:

$$e(p) = p^{1/d}$$

p: pre-specified fraction of samples
d: number of dimensions.

- **Example:**

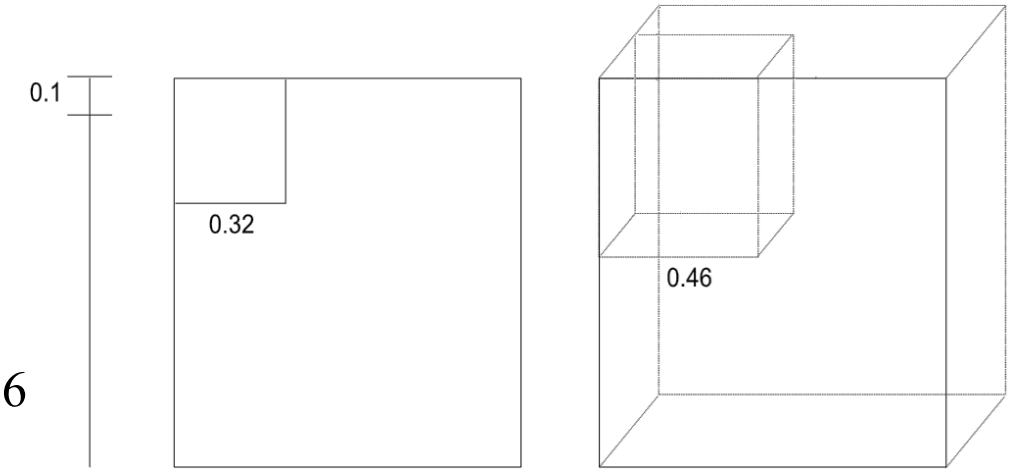
  10% of the sample (p=0.1):

  One Dimension: $e_1(0.1) = 0.1$

  Two dimensions: $e_2(0.1) = 0.32$

  Three Dimensions: $e_3(0.1) = 0.46$

  Ten Dimensions: $e_{10}(0.1) = 0.8$



Curse of Dimensionality (3)

- ***Almost every point is closer to an edge*** than to another sample point in a high-dimensional space:

For a sample ***size n***, the **expected distance *D* between normalized data points** in ***d***-dimensional space is:

$$D(d, n) = \frac{1}{2} \cdot \left(\frac{1}{n}\right)^{1/d}$$

- **Example:**

For a two-dimensional space with 10000 points

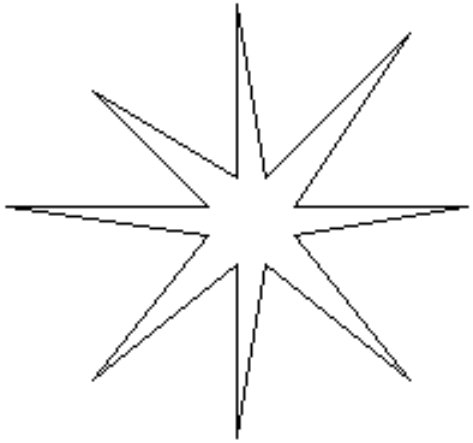
$\rightarrow\ D(2,10000) = 0.005$

For a 10-dimensional space with 10000 points

$\rightarrow\ D(10,10000) = 0.4$

Curse of Dimensionality (4)

- Almost every point is an ***outlier*** in high-dimensional spaces:
  - As the dimension of the input space increases, the distance between the prediction point and the center of data points increases.

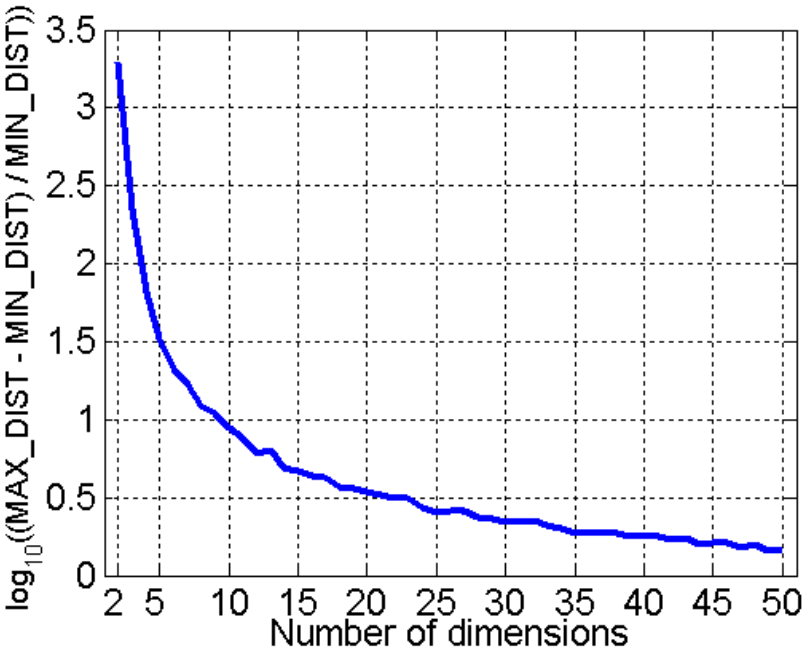


Curse of Dimensionality (5)

Experimental Confirmation:

- When dimensionality of data set increases, data becomes increasingly ***sparse*** with mostly ***outliers*** in the space that it occupies.

- Definitions of ***density*** and ***distance*** between points change the meaning:
  - difference between max & min distances become close to zero
  - Is critical for many data mining tasks



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

Preparing the Data for Data Mining

Two central tasks for the preparation of data:

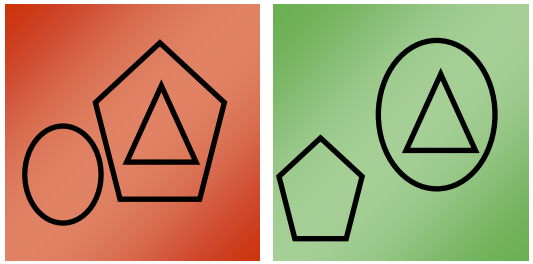
- To organize data into a standard form: typically, a standard form is a ***relational table*** (or tables)

- To prepare data sets by:
  - ***preprocessing and***
  - ***dimensionality reduction***

  … that will lead to the best data mining performances

Representing Data with Tables

Single Table Representation



Scene S1 Scene S2

Relational Representation

SCENE				
SceneID	Triangle	Square	Circle	Pentagon
S1	+	-	+	+
S2	+	-	+	+

SCENE		
SceneID	ObjectID	Shape
S1	O1	Triangle
S1	O2	Circle
S1	O3	Pentagon
S2	O1	Triangle
S2	O2	Circle
S2	O3	Pentagon

INSIDE		
SceneID	ObjectID	ObjectID
S1	O1	O3
S2	O1	O2

Representing Data with Tables
Market Baskets

TID: 100

TID: 200

TID:300

TID: 400

TID	Garlic	Milk	Detergent	Ketchup	Wine
100	Yes	No	Yes	Yes	No
200	No	Yes	Yes	No	Yes
300	Yes	Yes	Yes	No	Yes
400	No	Yes	No	No	Yes

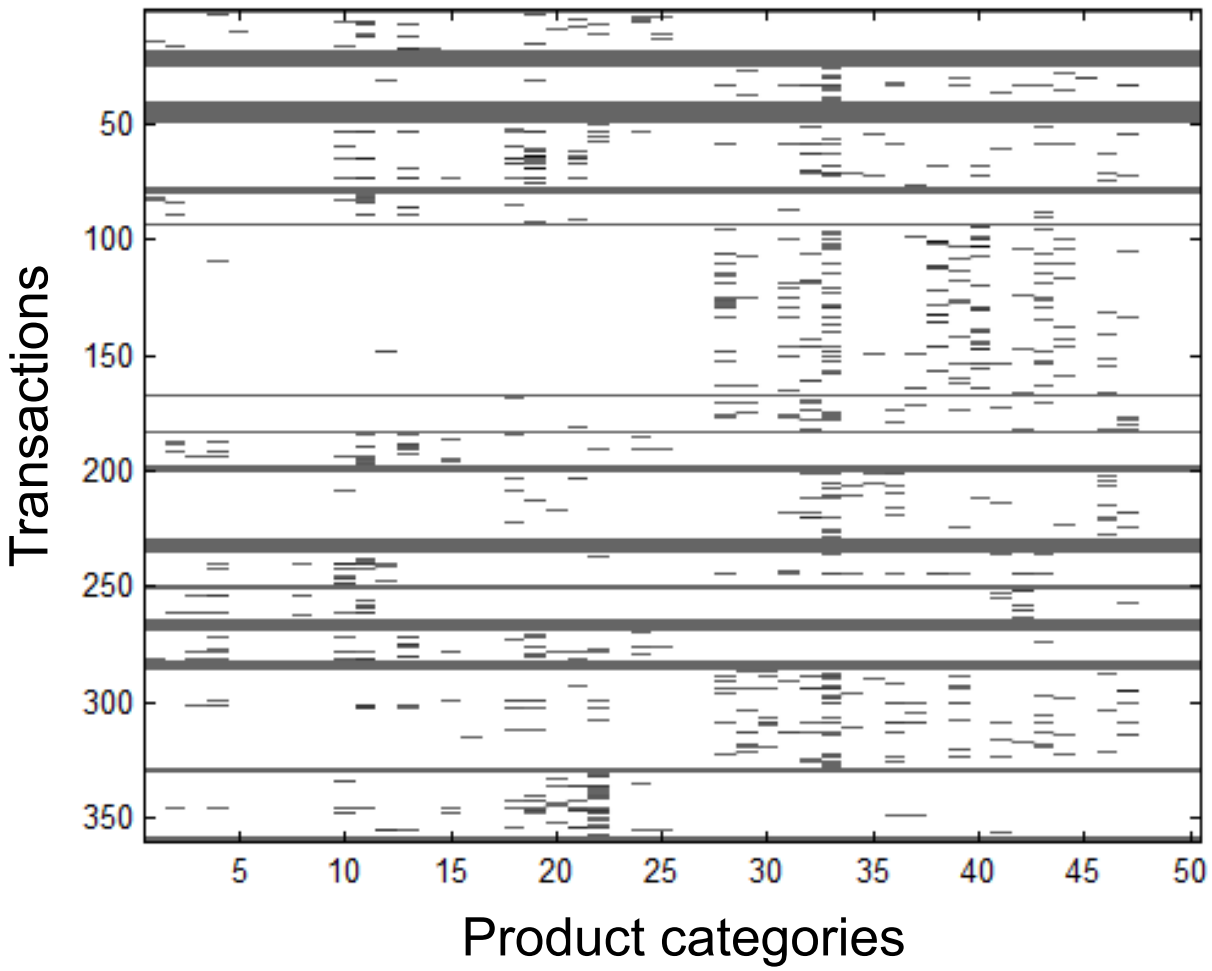
TID	Items
100	{Garlic, Detergent, Ketchup}
200	{Milk, Detergent, Wine}
300	{Garlic, Milk, Detergent, Wine}
400	{Milk, Wine}

Each basket represents one sample

Sparsity:
Eliminate „No‘s“

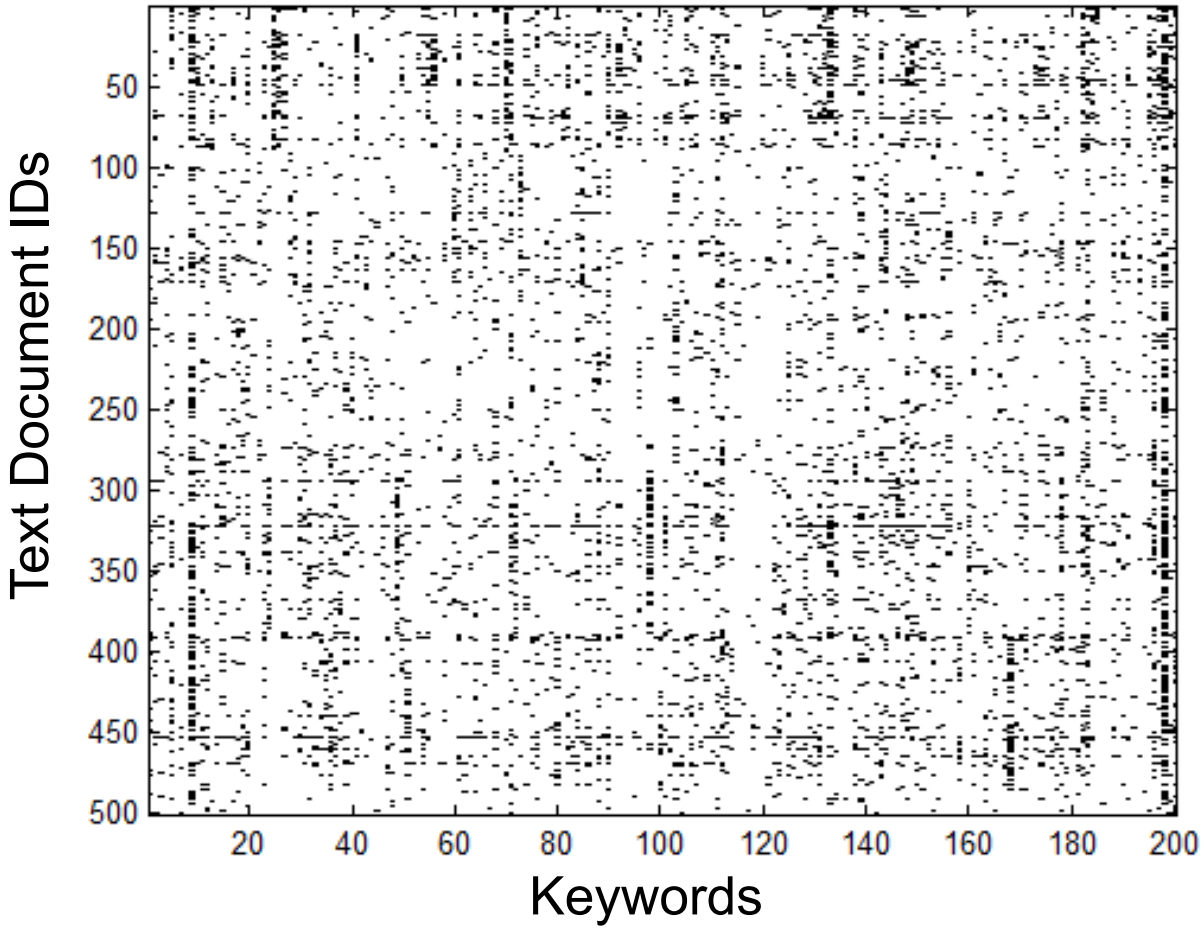
Market Basket Data

Trans. ID	Products
01	01, 03, 44, 76
02	22, 37, 76
...	...



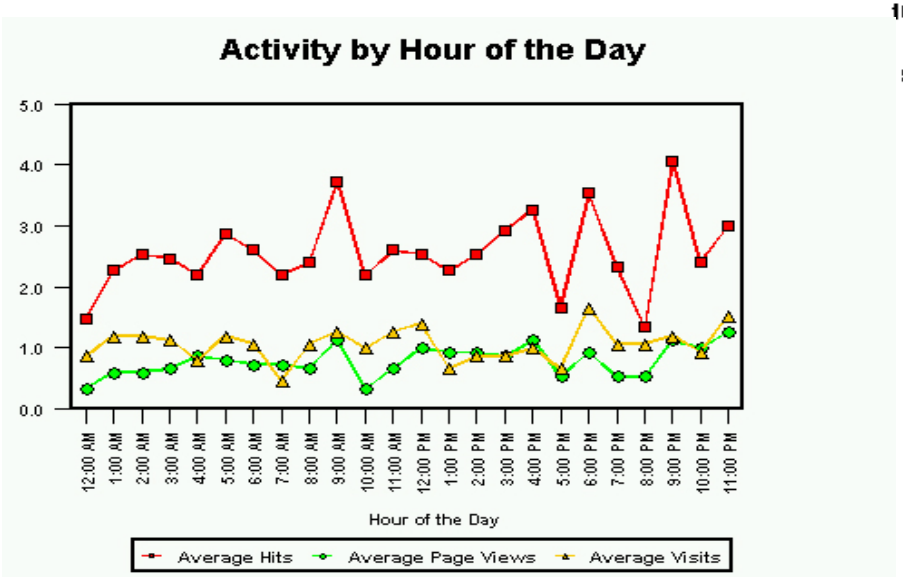
Representing Text with Tables

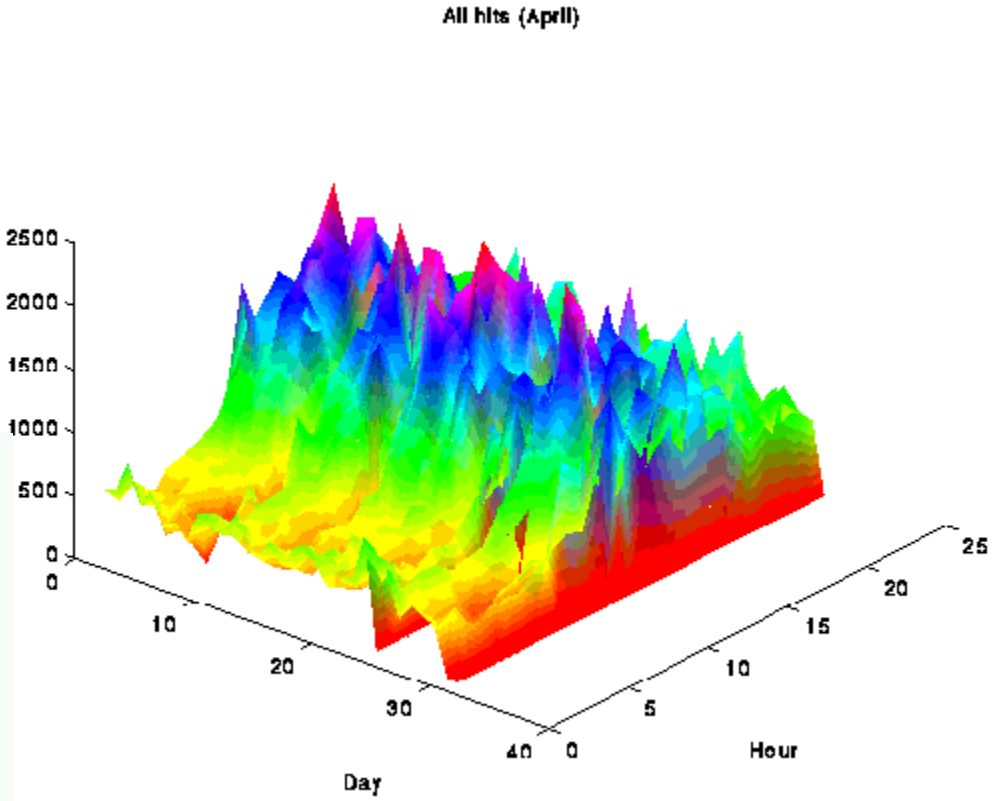
Text ID	Keywords
001	56, 34, 79
002	07, 122, 189
…	…



Web Log Data over Time – Table?

Day	Hour	# of hits
06/06/13	5 a.m.	58
06/07/13	6 a.m.	83
…	…	…





Time Series Data – Table?

Time	TS1	TS2		TSn
1	86	74	…	140
2	99	133	…	91
…	…	…	…	…

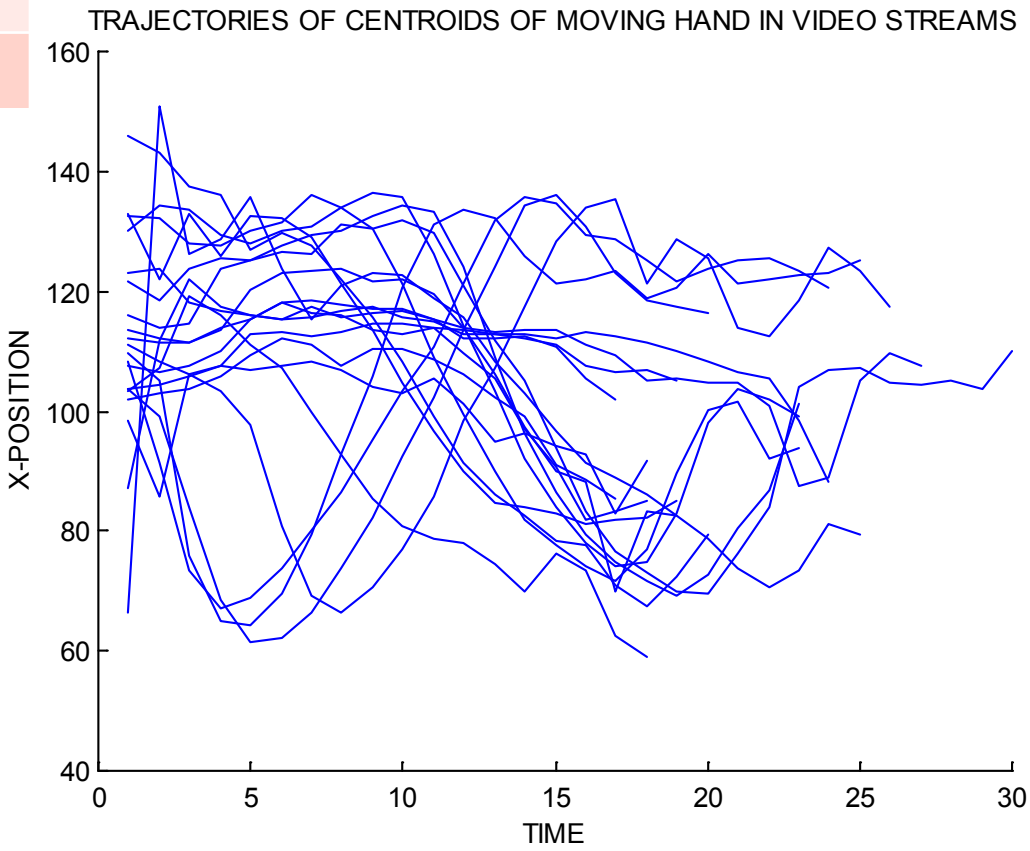
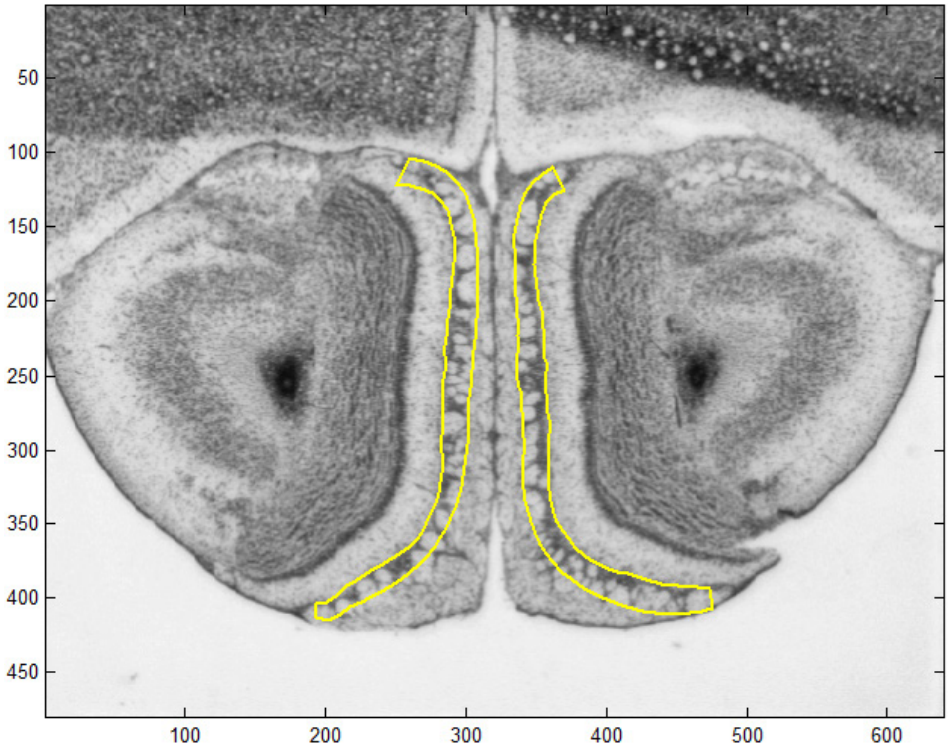


Image Data – Table?

X coor.	Y coor.	RGB value
100	250	87
100	255	85
…	…	…

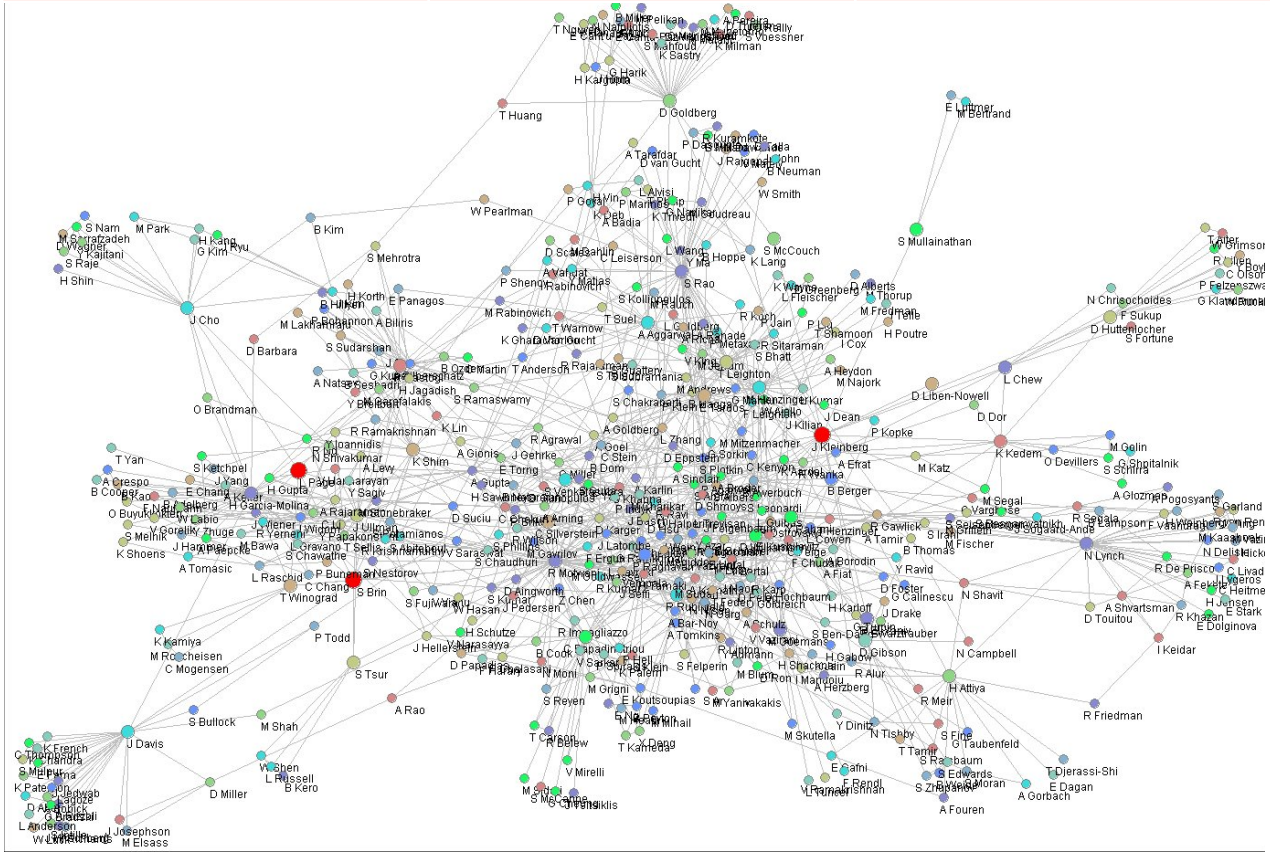
…or geographical images



Medical…

Relational Data (=*Graph*) – Table?

Beginning node	Ending node	Distance
Miller	Todd	134
Mile	Rao	78
…	…	…



Each row is defined with the beginning and ending node in one connection, and weight factor (or other factors like distance) connected with this link.

Basic Statistical Descriptions of Data

- Motivation
  - To better ***understand*** the data: central tendency, variation and spread
- Data ***dispersion characteristics***
  - median, max, min, quantiles, outliers, variance, etc.
- **Numerical dimensions** correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - ***Boxplot or quantile analysis*** on sorted intervals

Measuring the Central Tendency

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

- ***Mean*** (algebraic measure; sample vs. population):

  $$\mu = \frac{\sum x}{N}$$

  Note: *n* is sample size and *N* is population size.
  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values

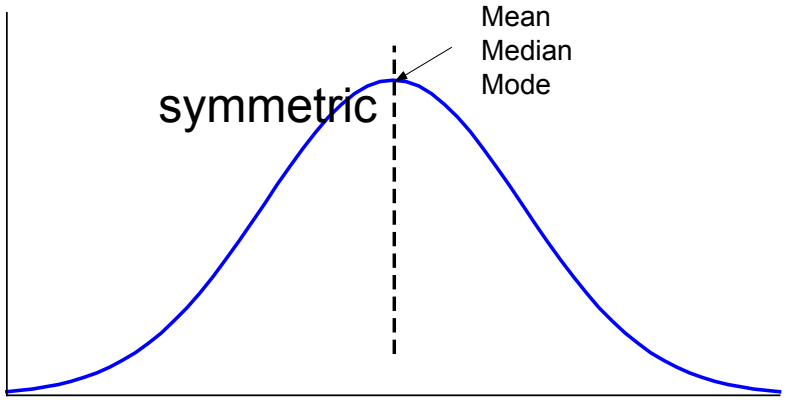
  $$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

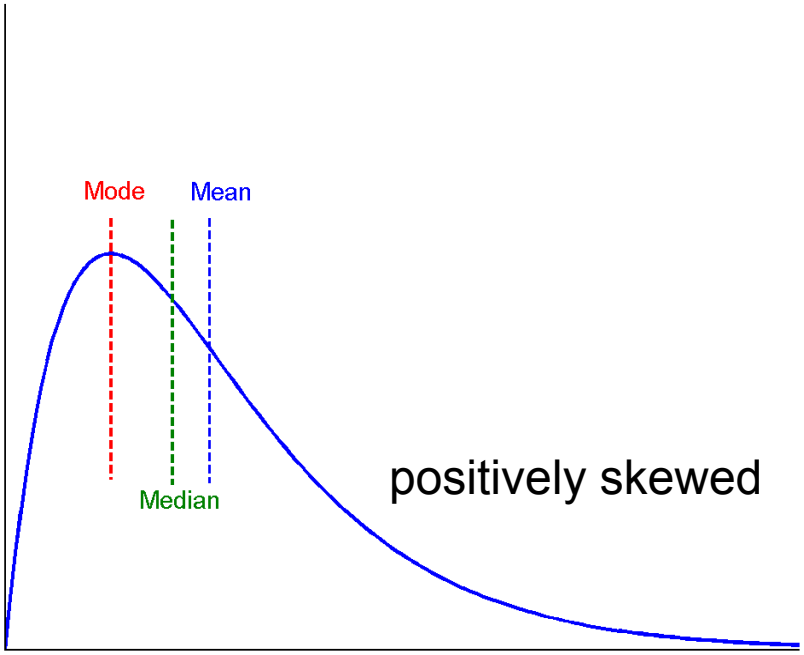
- ***Median***
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation for *grouped data*
- ***Mode***
  - Value that occurs most often in the data
  - Unimodal, bimodal, trimodal are data sets with 1, 2, 3 modes
  - Empirical formula for moderately asymmetrical curves

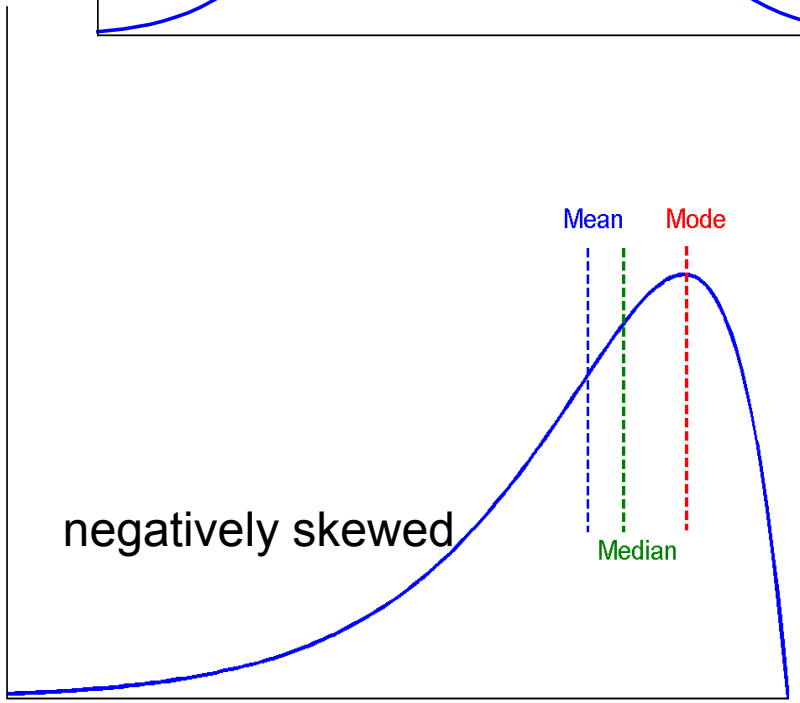
$$mean - mode = 3 \times (mean - median)$$

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data – always the same?





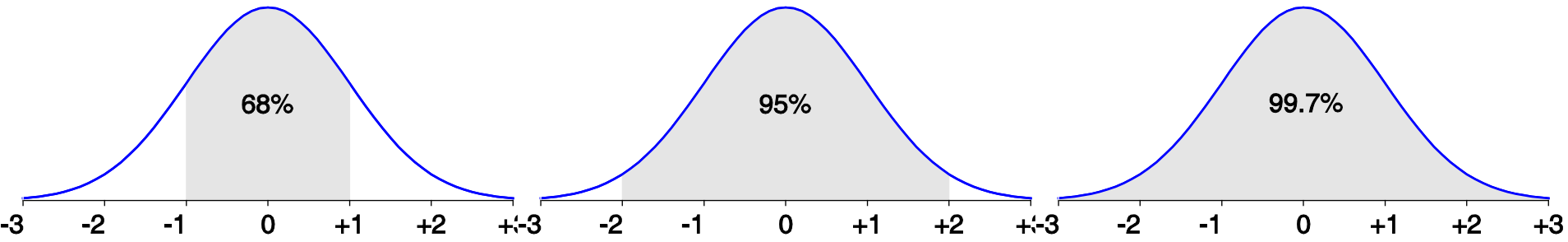


Measuring the Dispersion of Data

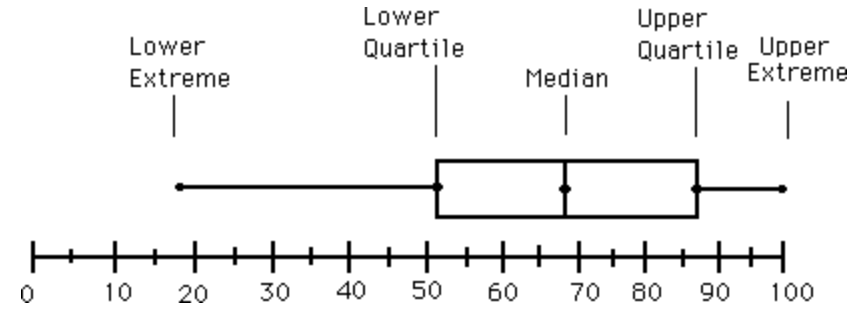
- Quartiles, outliers and boxplots
  - ***Quartiles***: Q1 (25th percentile), Q3 (75th percentile)
  - ***Inter-quartile range***: IQR = Q3 – Q1 (large? small?)
  - ***Five number summary***: min, Q1, median, Q3, max
  - ***Boxplot***: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - ***Outlier***: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation
  - ***Variance***: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$
  - ***Standard deviation*** σ is the square root of variance σ

Properties of normal Distribution Curve

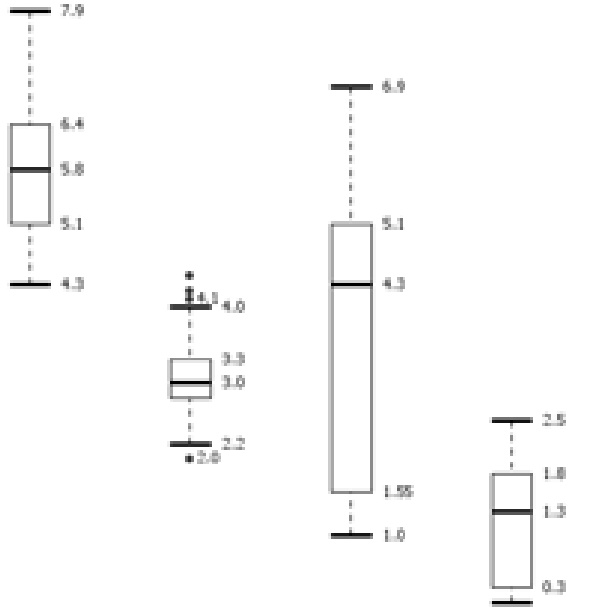
- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it



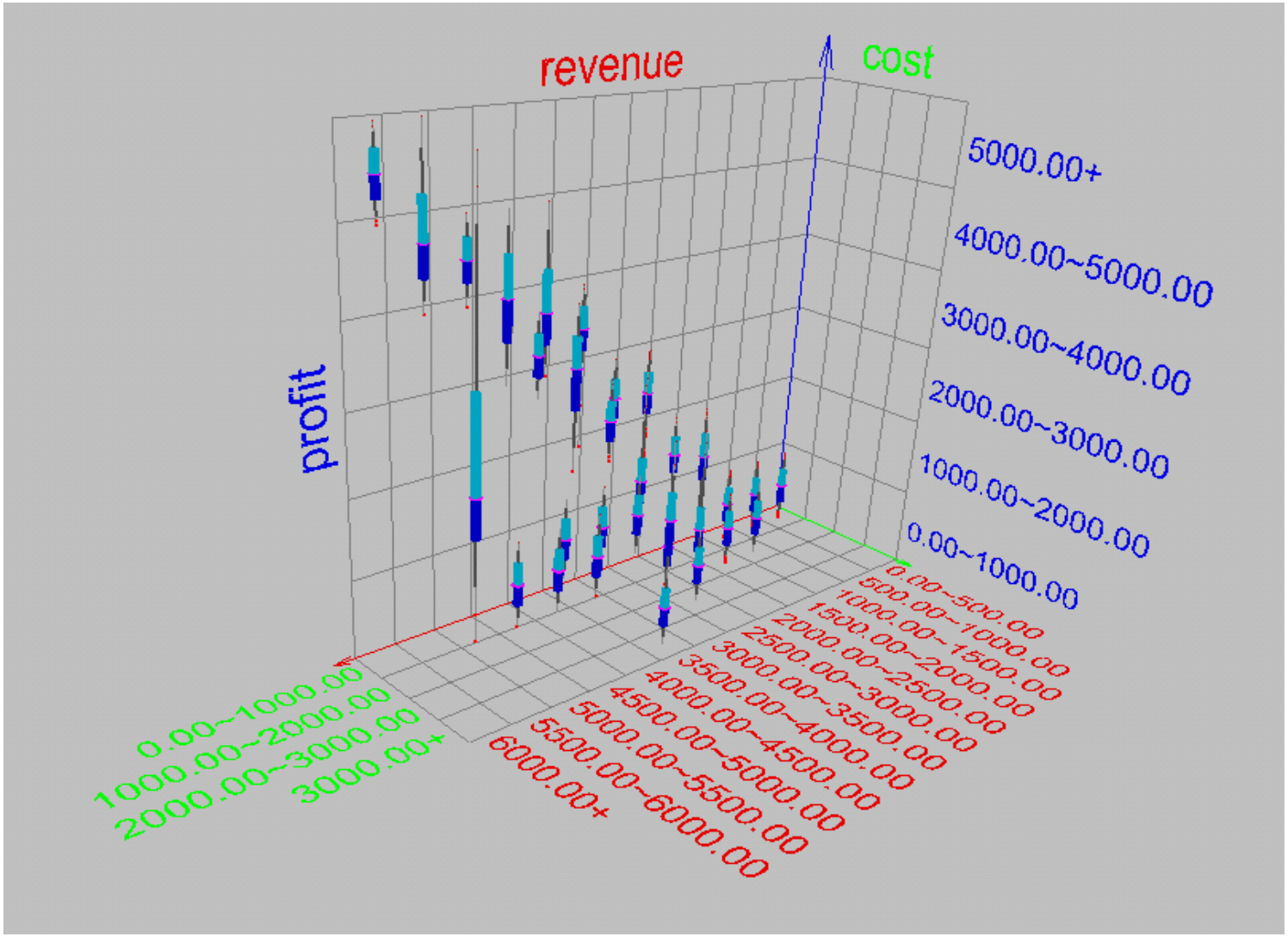
Boxplot Analysis



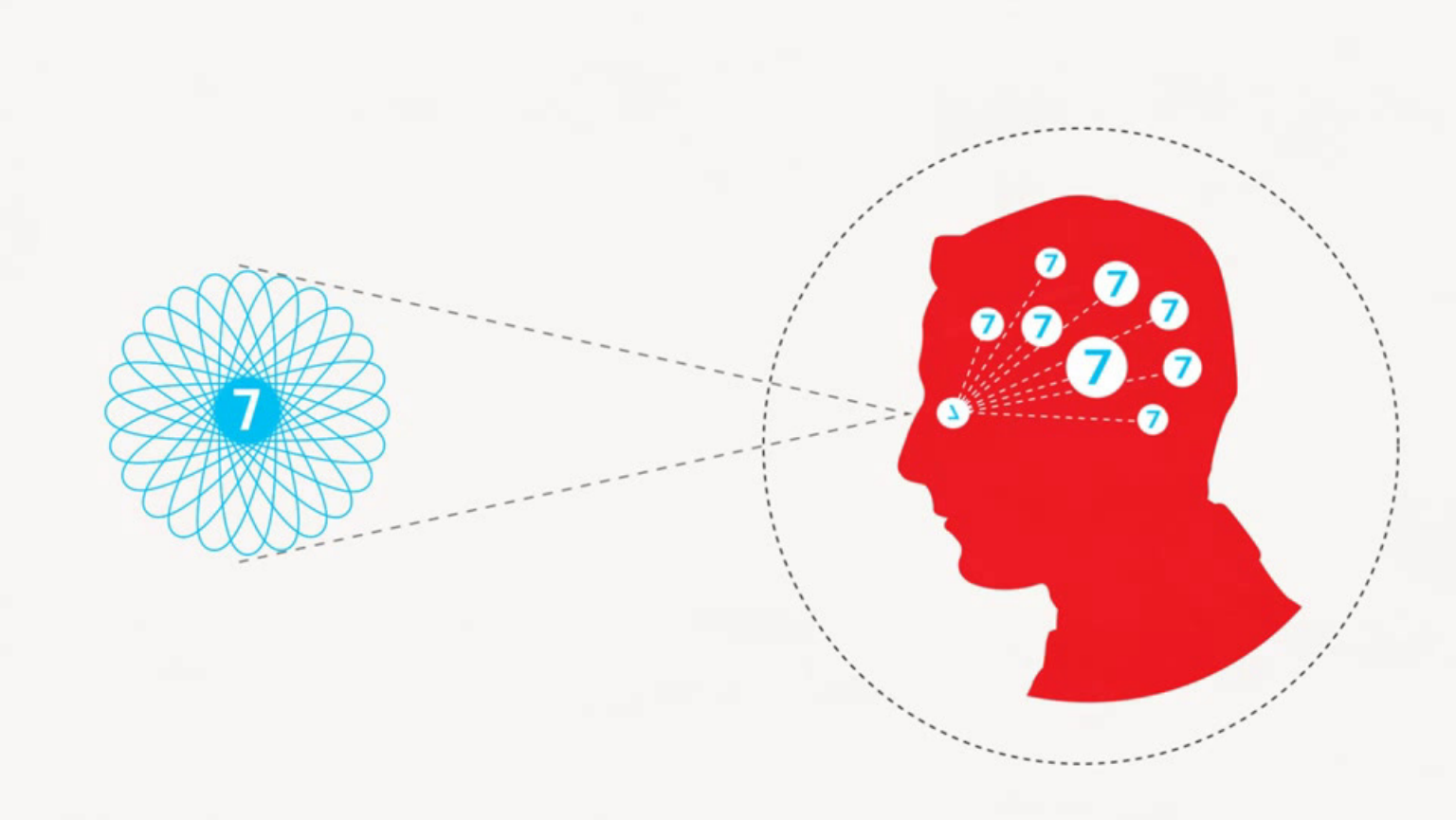
- ***Five-number summary*** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- ***Boxplot***
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines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



Visualization of Data Dispersion: 3-D Boxplots



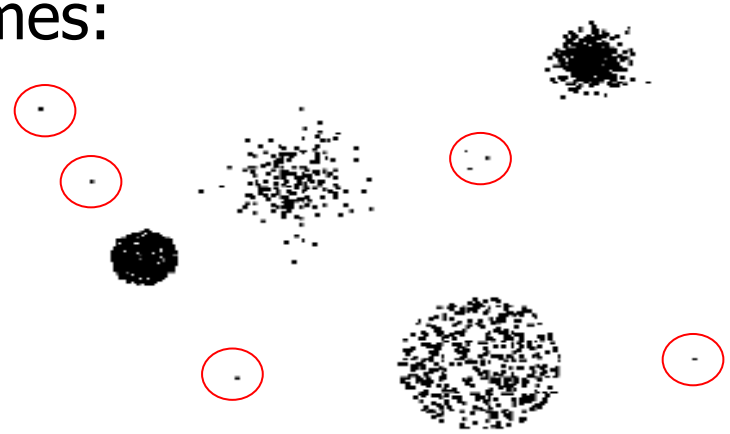
Data Visualization



Outlier Detection Schemes

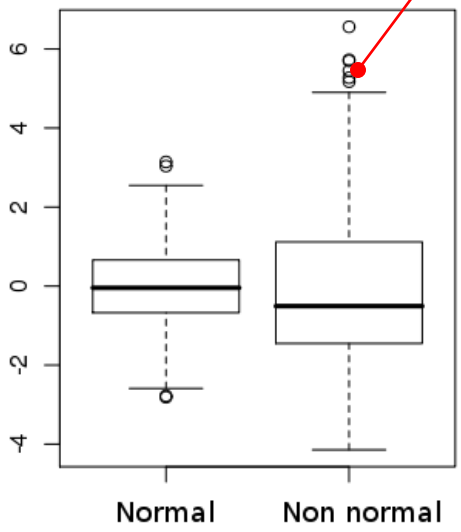
- General Steps:
  - Build a ***profile of the „normal“ behavior***.
    - Profile can be patterns or summary statistics for the overall population.
  - ***Use the “normal” profile to detect outliers***.
    - Outliers are observations whose characteristics differ significantly from the normal profile.

- Types of outlier detection schemes:
  - Graphical
  - Statistical-based
  - Distance-based
  - Model-based

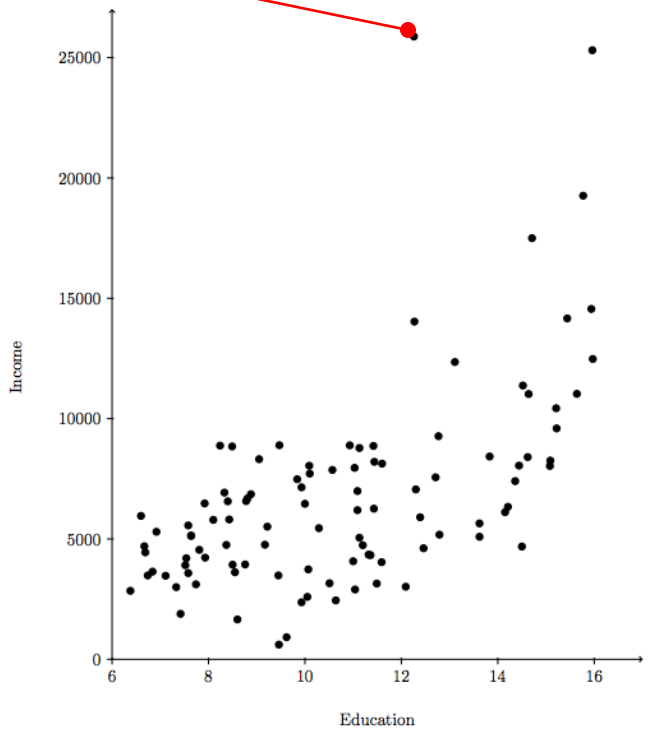


Outliers: Graphical Approaches

- Boxplot (1-D), Scatter plot (2-D), Spin plot (3-D),…
- Limitations:
  - Time consuming
  - Subjective



One-Dimensional

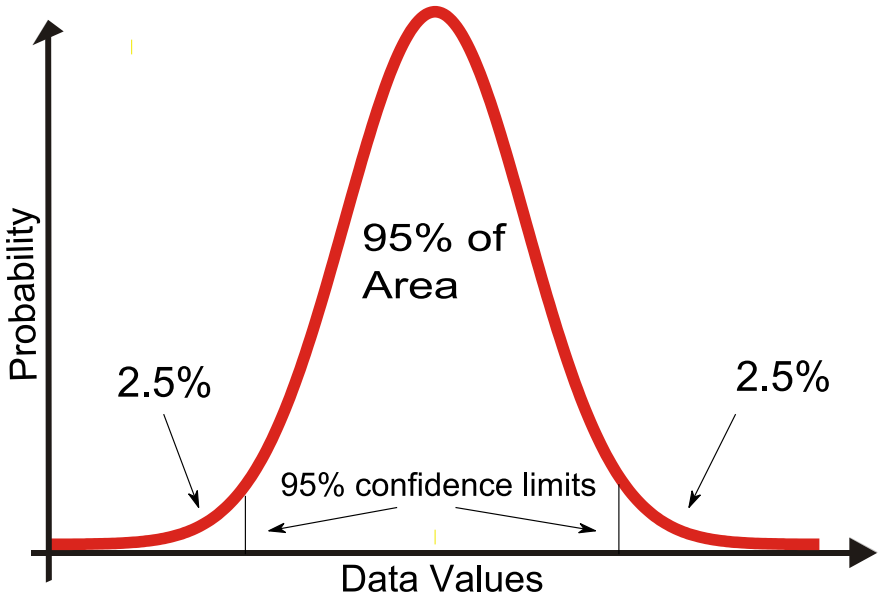


Two-Dimensional

Outliers: Statistical Approaches (1)

- Assume a ***parametric model*** describing the distribution of the data (e.g., normal distribution)

- Apply a ***statistical test*** that depends on:
  - Data distribution
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)



Outliers: Statistical Approaches (2)

Example: Outlier detection for one-dimensional samples:

Samples = {3,56,23,39,156,52,41,22,9,28,139,31,55,20, -67,37,11,55,45,37}

Statistical parameters are:

Mean = 39.9

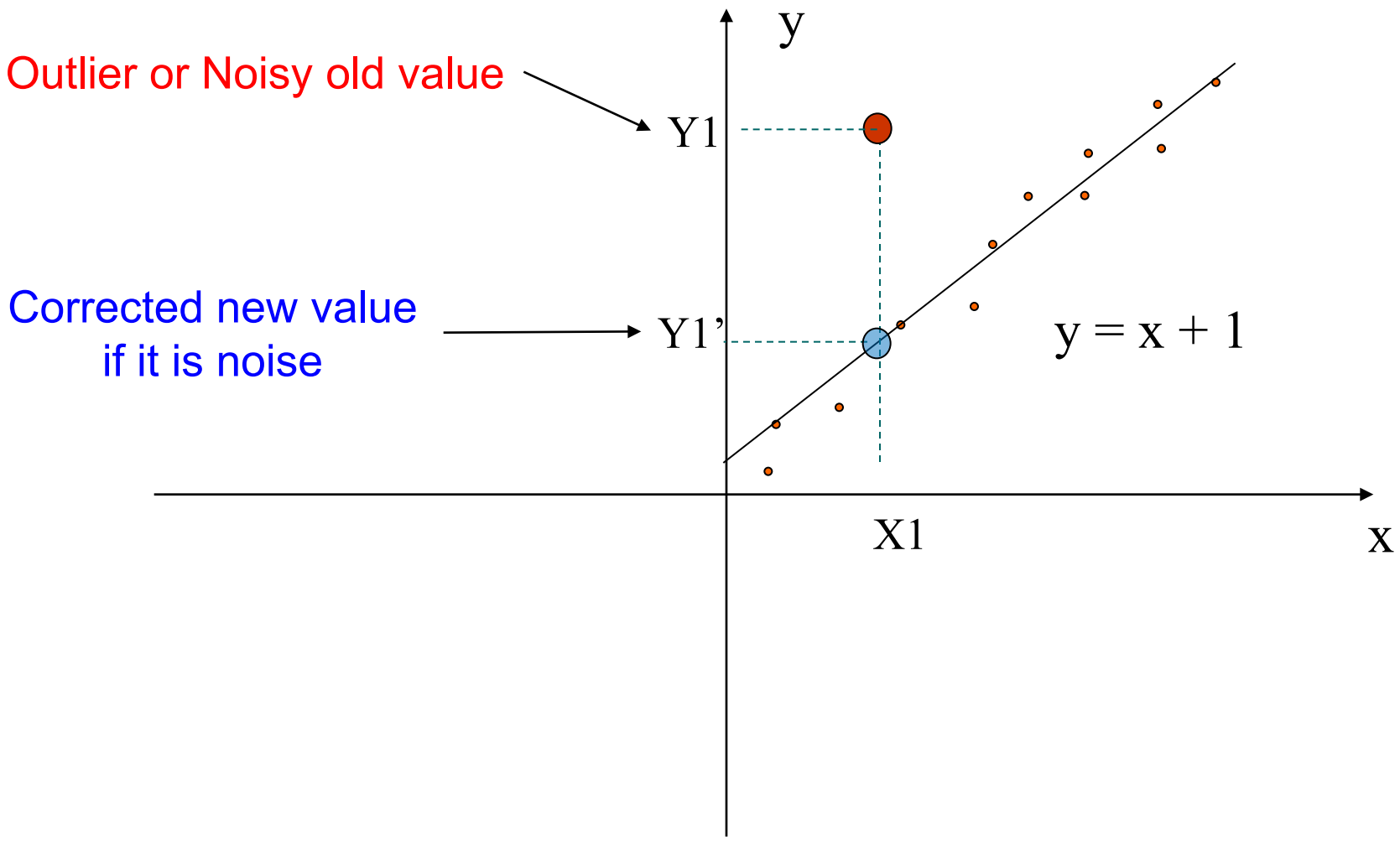
Standard deviation = 45.65

If we select that the threshold value for normal distribution of data is:

Threshold = Mean ± 2 × Standard deviation

…then all data out of range [-54.1, 131.2] will be potential outliers: {156, 139, -67}

Outliers or Noisy Data?
(Using Regression Analysis)



Limitations of Statistical Approaches

- Most of the tests are for a ***single attribute***
- In many cases, data ***distribution*** may ***not*** be ***known***
- For high dimensional data, it may be **difficult to estimate** the true distribution

Outliers: Distance-based Approaches

- Data is represented as a vector D of features
- Three major approaches:
  - ***Nearest neighbor-based***
  - ***Density-based***
  - ***Clustering-based***

Outliers: Nearest Neighbour Approach

- Outlier detection for *n*-dimensional samples:
  - Evaluate the distance measures between all samples in *n*-dimensional data set.

A sample s_i in a data set *S* is an outlier if at least a proportion ***p*** of the samples in *S* lies at a distance greater than ***d*** *from* s_i

In other words:

Distance-based outliers are those samples which do not have enough neighbors

Outliers: Distance-based Approach Example

- Data set: $S = \{ (2,4), (3,2), (1,1), (4,3), (1,6), (5,3), (4,2)\}$
- Requirements: $p > 4$, $d > 3.00$

	S2	S3	S4	S5	S6	S7
S1	2.236	**3.162**	2.236	2.236	**3.162**	2.828
S2		2.236	1.414	**4.472**	2.236	1.000
S3			**3.605**	**5.000**	**4.472**	**3.162**
S4				**4.242**	1.000	1.000
S5					**5.000**	**5.000**
S6						1.414

Table of distances

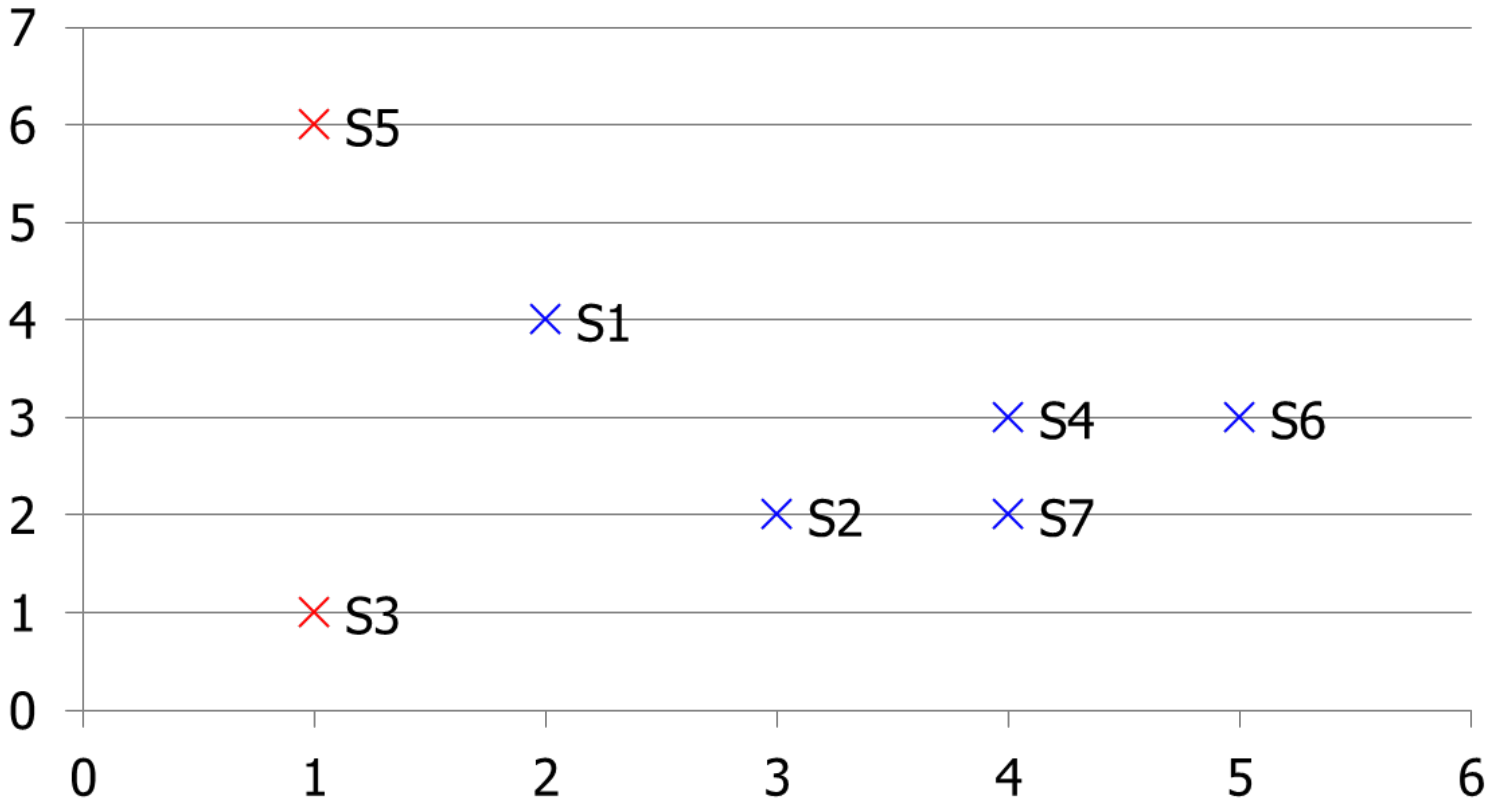
Outliers

Sample	p
S1	2
S2	1
S3	**5**
S4	2
S5	**5**
S6	3
S7	2

Computation p

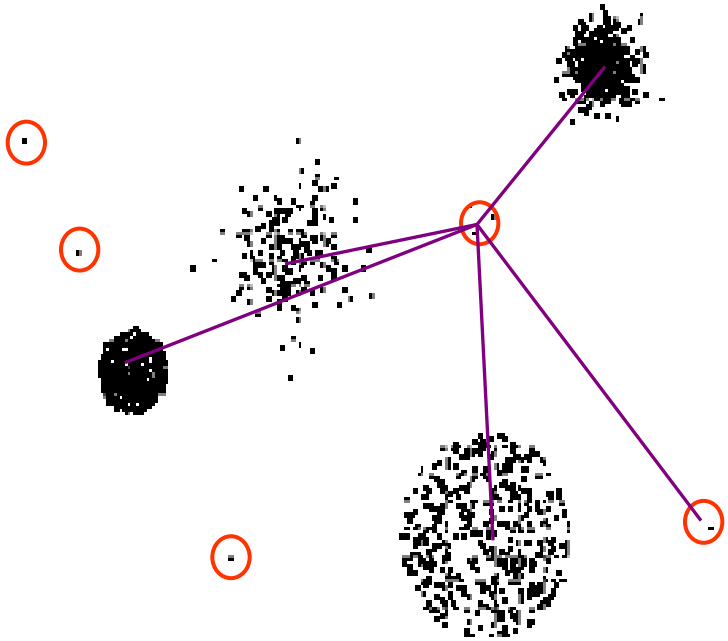
Outliers: Distance-based Approach Example Visual Inspection (2)

Data set: $S = \{ (2,4), (3,2), (1,1), (4,3), (1,6), (5,3), (4,2)\}$

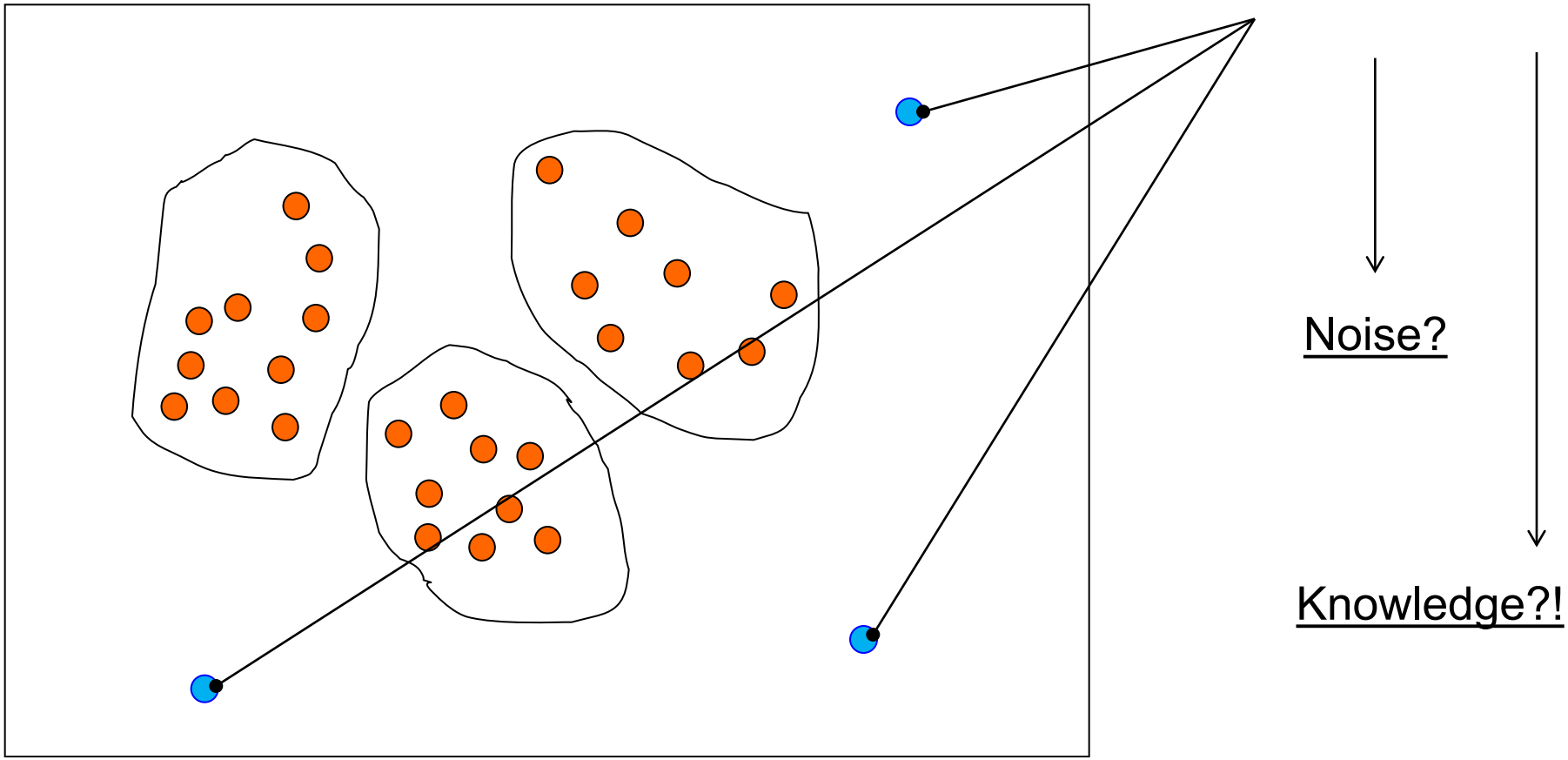


Outliers: Distance-based Approach Clustering

- Basic idea for large data sets - ***clustering based***:
  - Cluster the data into groups of different density
  - Choose points in small cluster as candidate outliers
  - Compute the distance between candidate points and non-candidate clusters:
    - ***If candidate points are far from all other non-candidate points, they are outliers***



Outliers or Noisy Data?
(Using Cluster Analysis)



Automatic removal of outliers is not recommended

Anomaly/Outlier Detection

- ***Variants of Anomaly/Outlier Detection Problems***
  - Given a database D, find all the data points **x** $\in$ D with anomaly scores greater than some threshold t
  - Given a database D, find all the data points **x** $\in$ D having the top-n largest anomaly scores f(**x**)
  - Given a database D, containing mostly normal (but unlabeled) data points, and a test point **x**, compute the anomaly score of **x** with respect to D

- **Applications:**
  - Credit card fraud detection, telecommunication fraud detection, network intrusion detection, fault detection, condition monitoring of machines

Further Advances in Data Visualization

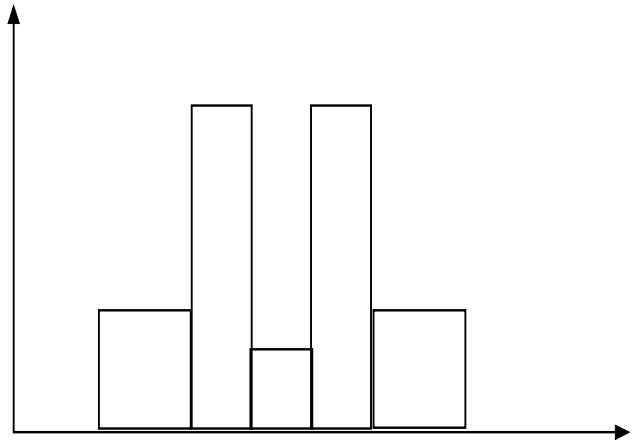
- Further forms of data visualization?
  - Gain insight into information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help to find interesting regions and suitable 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

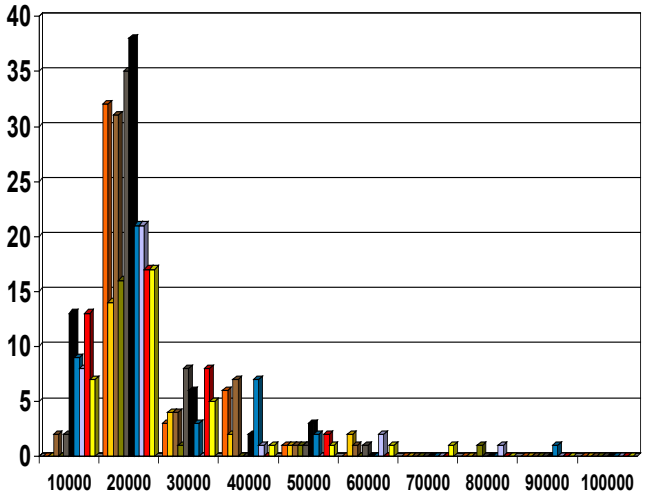
Further Displays of basic statistical Descriptions

- ***Histogram***: x-axis are values, y-axis represent frequencies
- ***Quantile plot***: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- ***Scatter plot***: each pair of values is a pair of coordinates and plotted as points in the plane

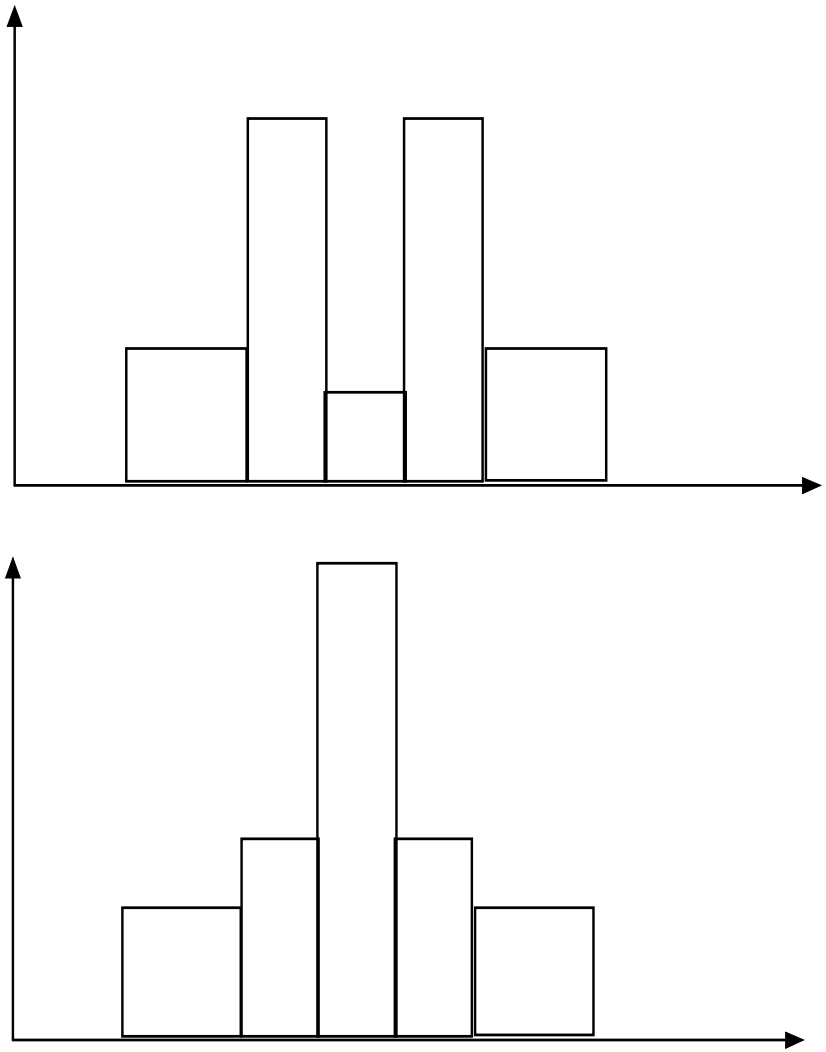
Bar Charts and Histogram Analysis

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





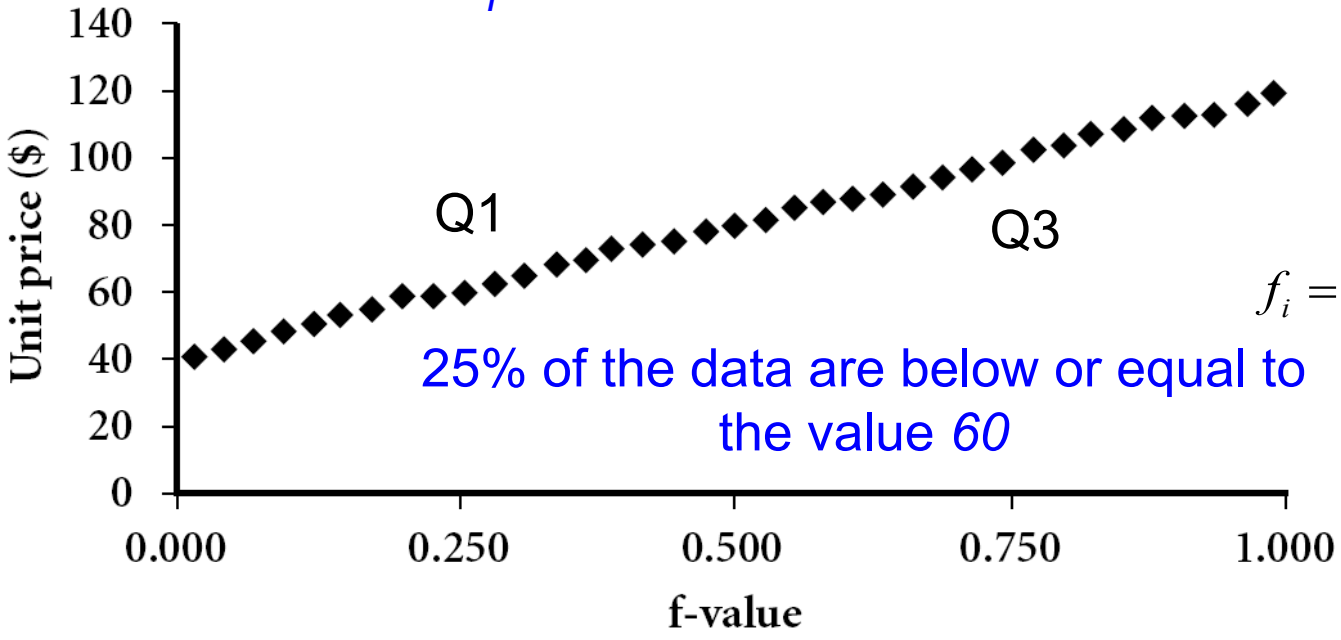
Do Histograms tell more than Boxplots?



- Yes
- 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

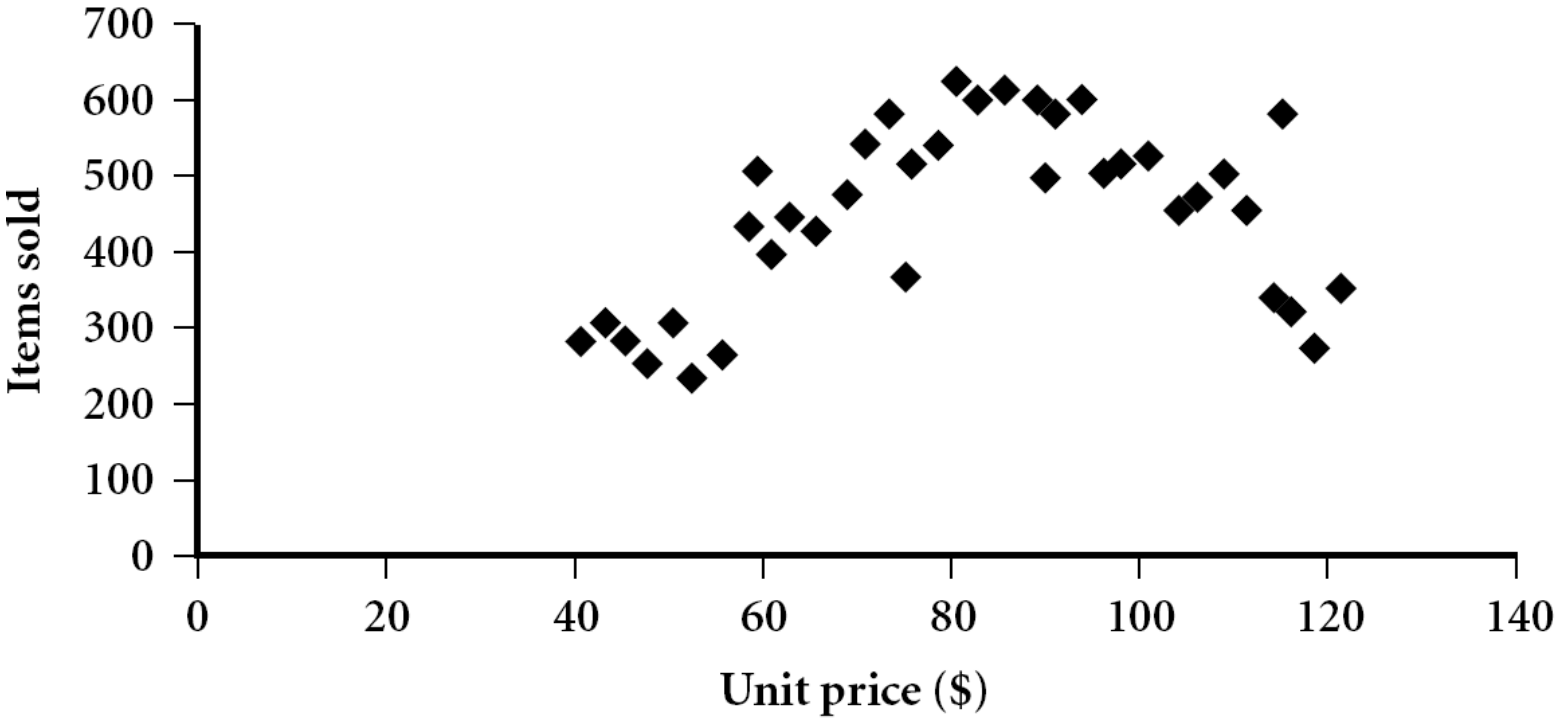
Quantile Plot

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

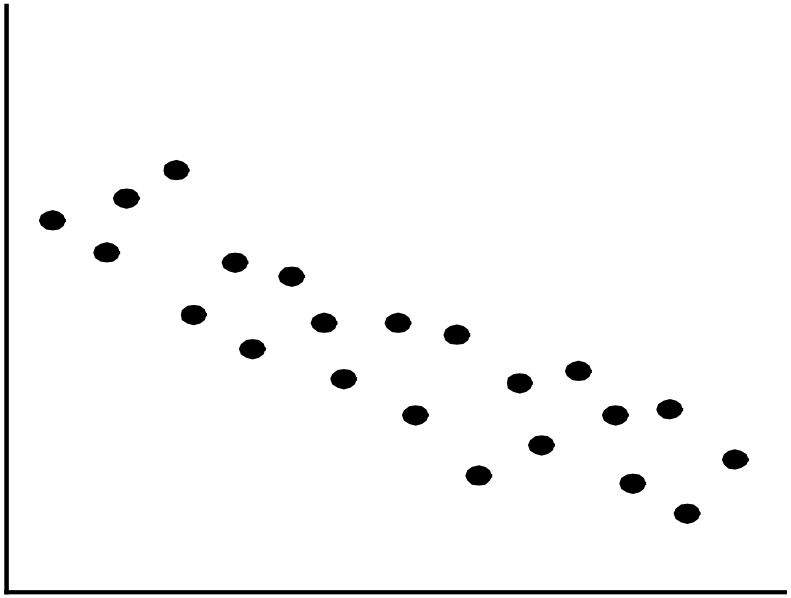


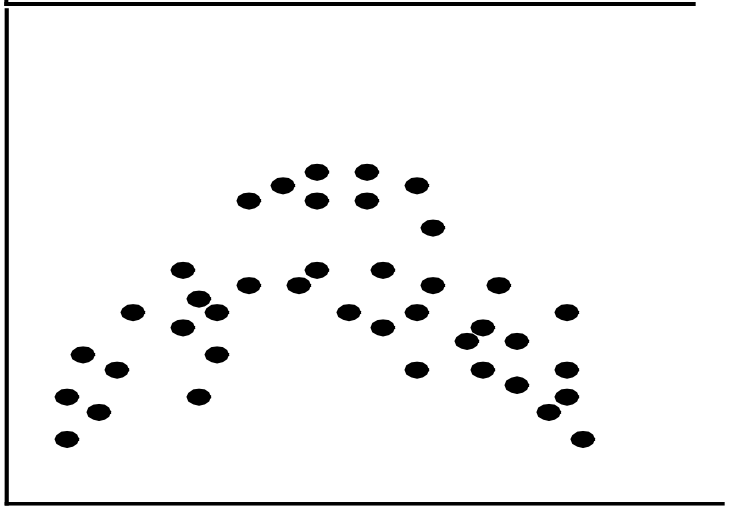
Scatter Plot

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



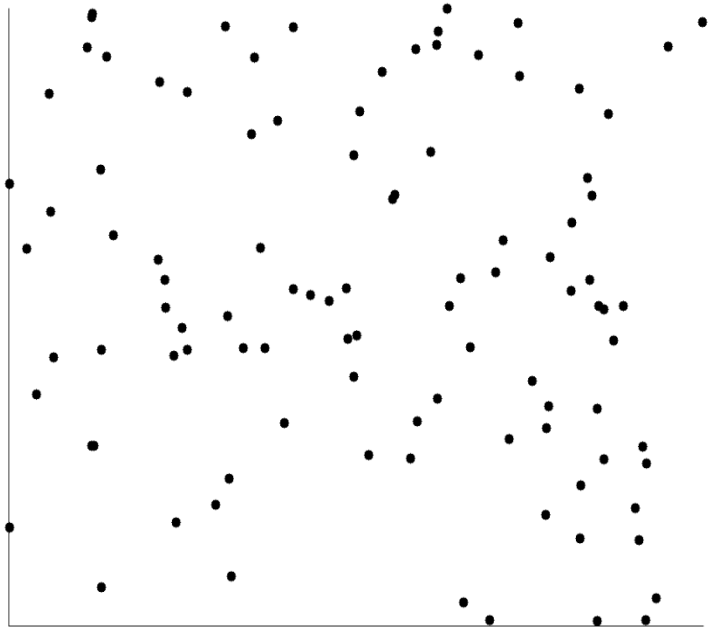
Positively and Negatively Correlated Data

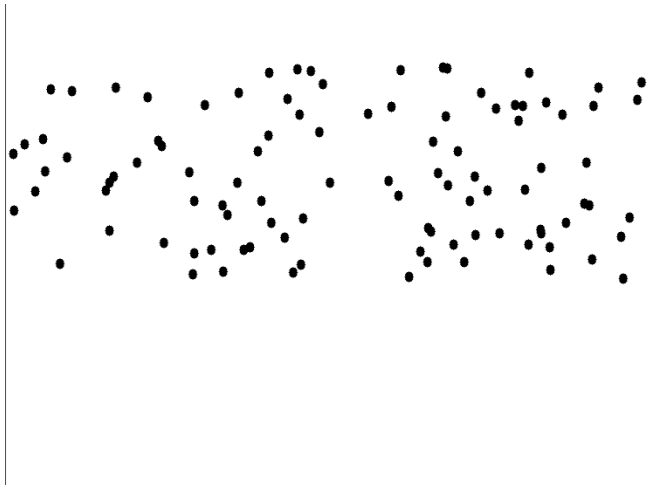


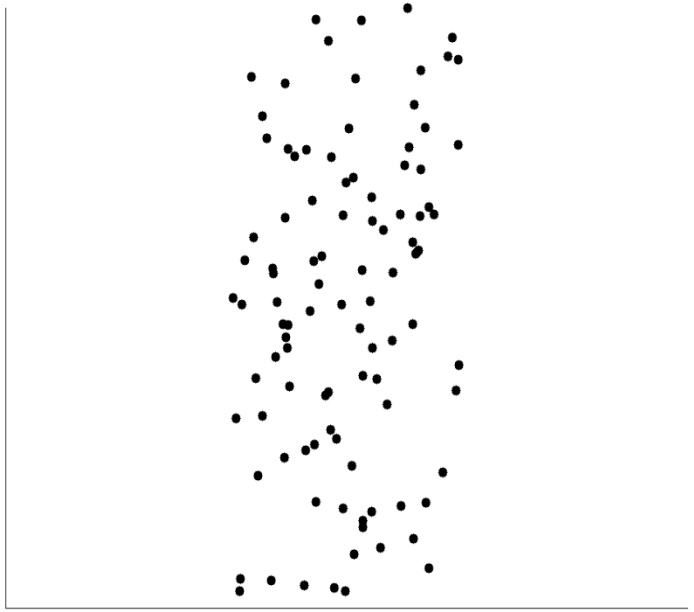


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data

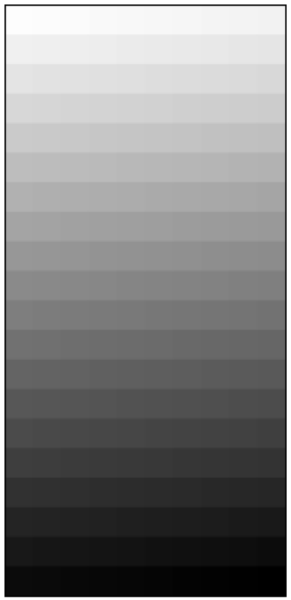




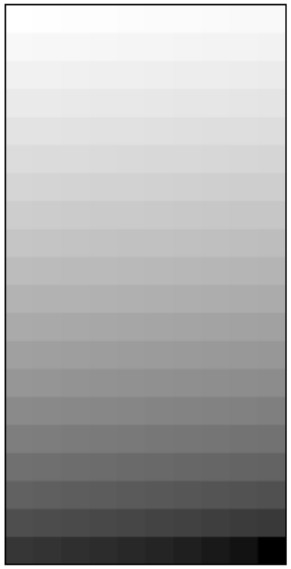


Pixel-oriented Visualization Techniques

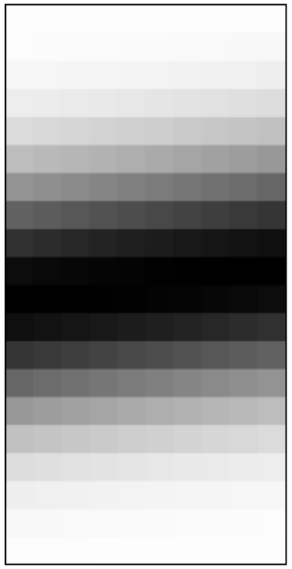
- For a data set of m dimensions, create m windows on the screen, one for each dimension
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values



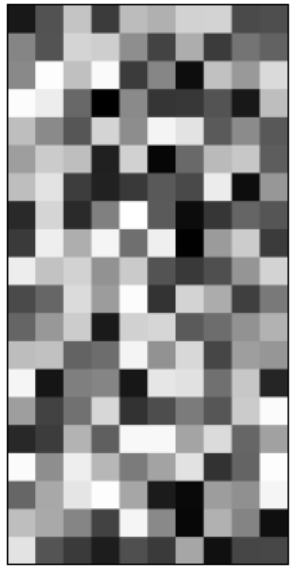
(a) Income



(b) Credit Limit



(c) transaction volume

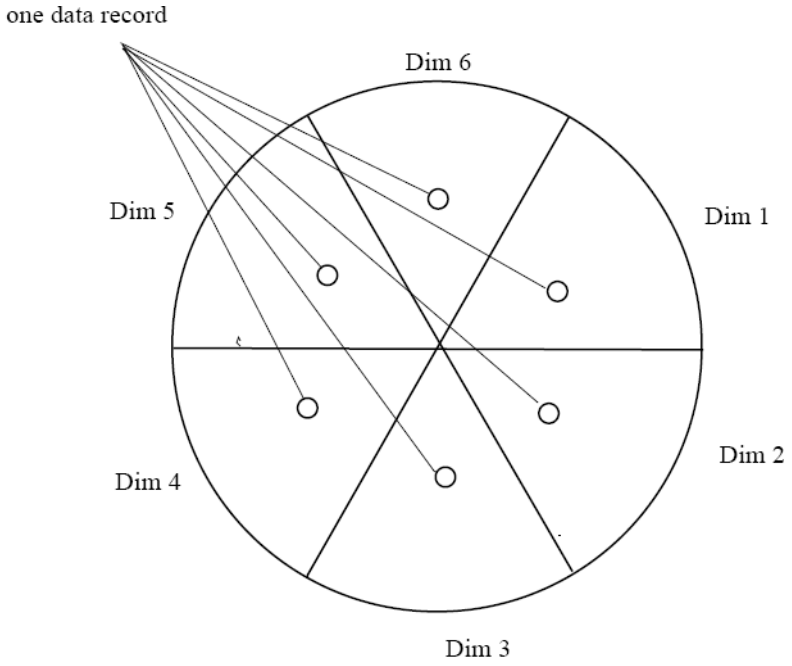


(d) age

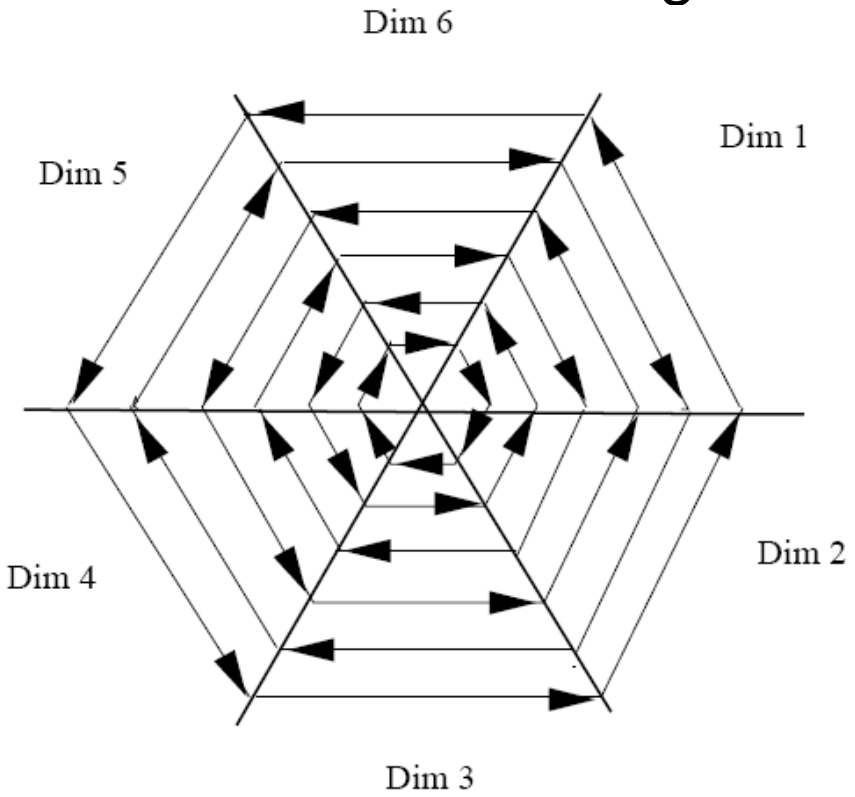
Income is correlated with credit limit but not with age

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

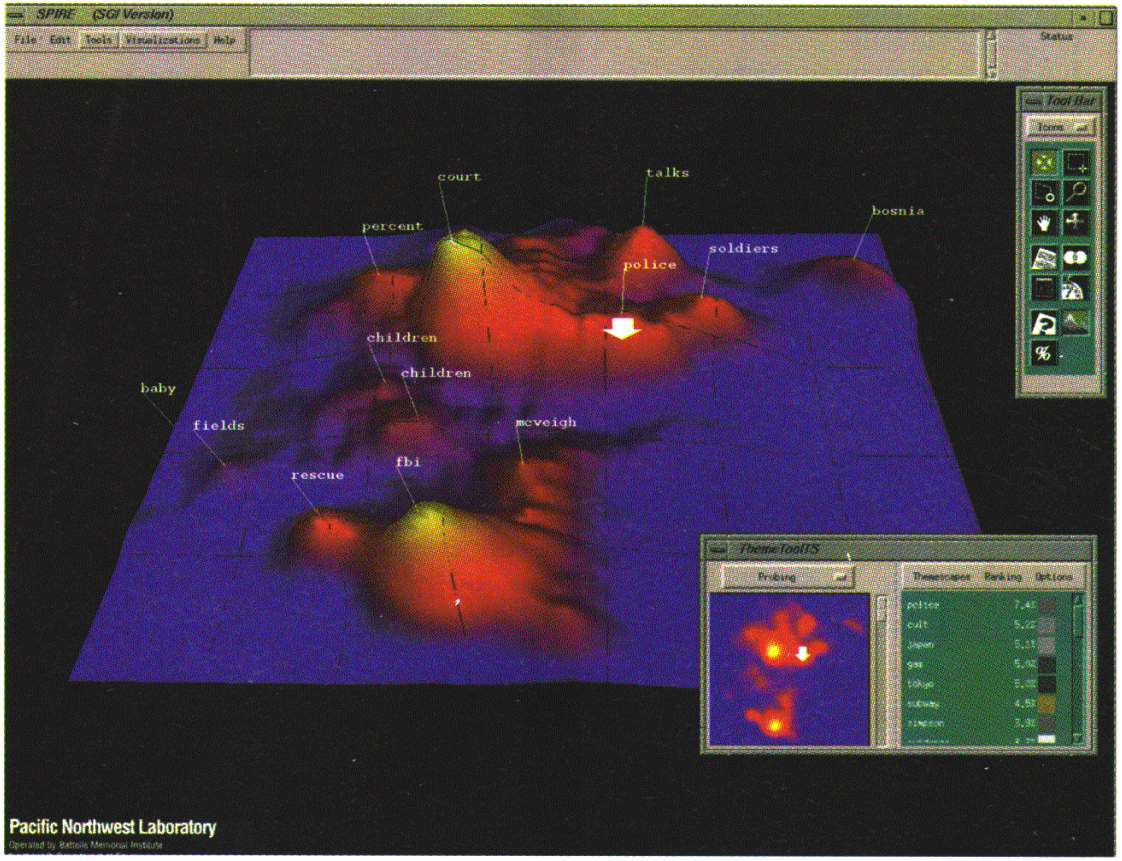


(a) Representing a data record in circle segment



(b) Laying out pixels in circle segment

Landscapes



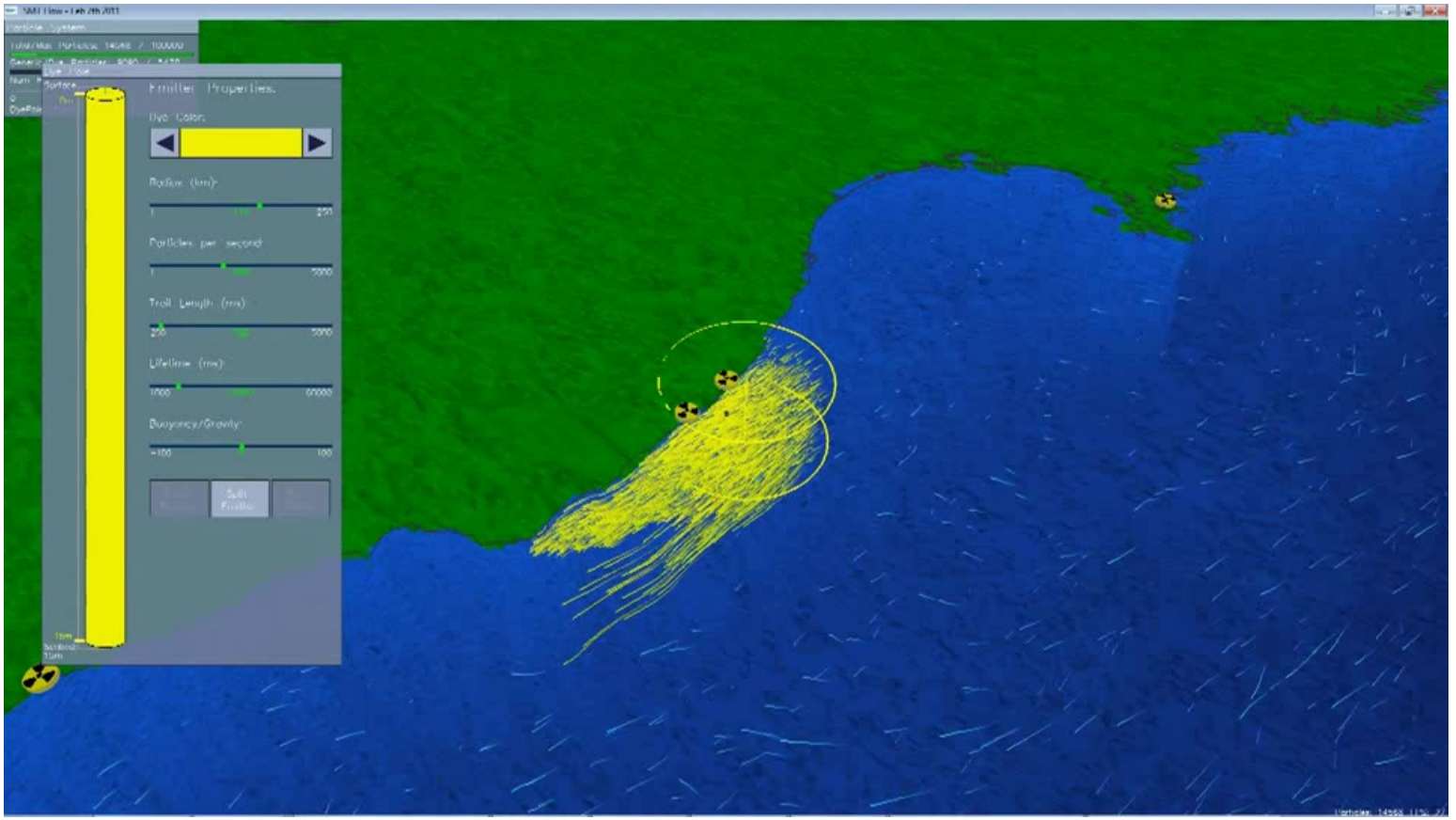
Used by permission of B. Wright, Visible Decisions Inc.

news articles visualised as a landscape

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

Example: Ocean Flow Analysis Visualization

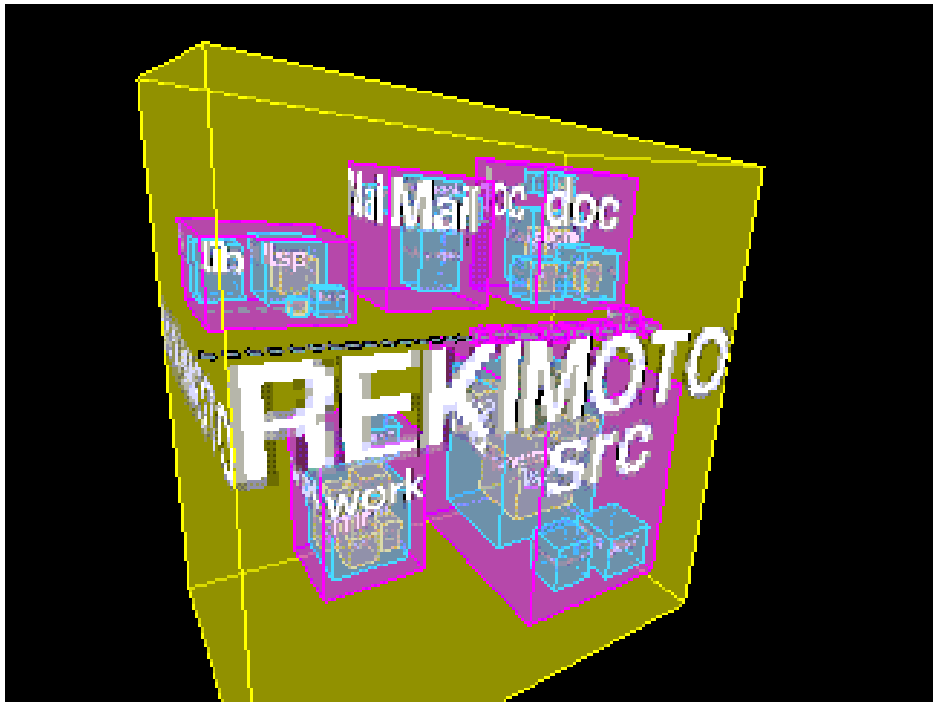
- A visualization of the ocean flow simulation software being run on the flow model



http://www.youtube.com/watch?v=YnRaq951mbs

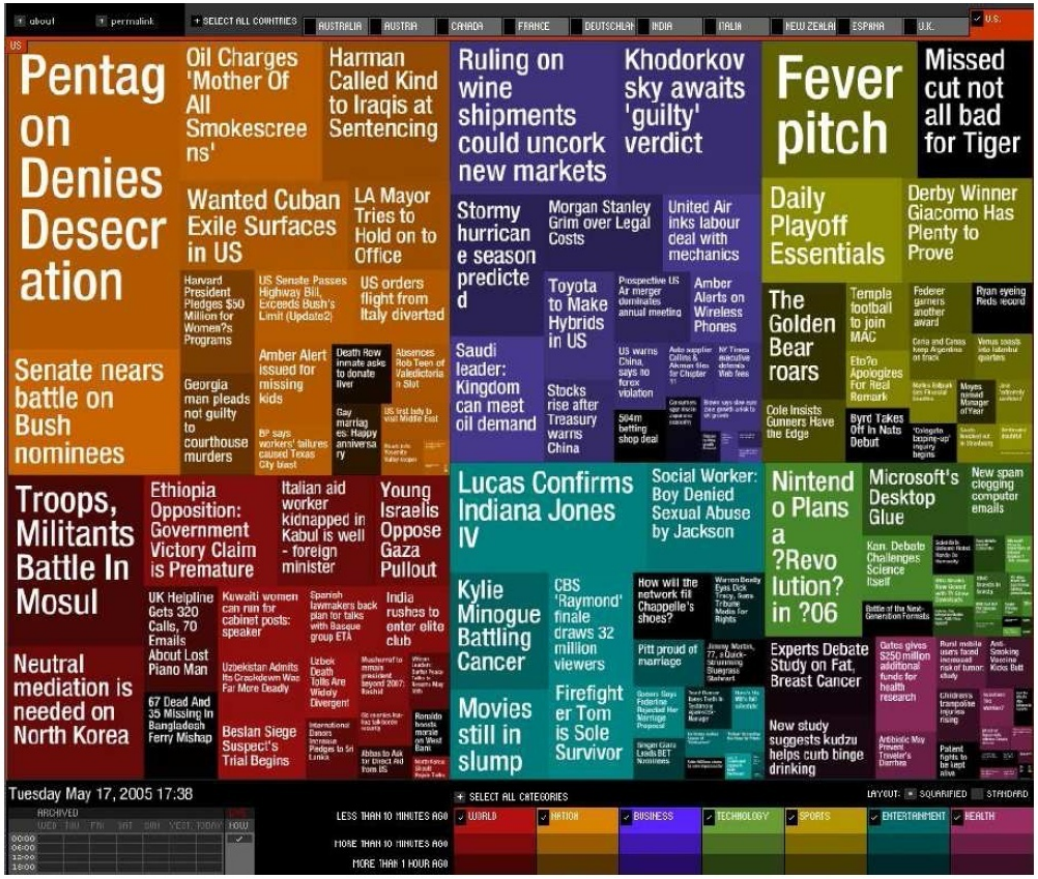
InfoCube

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



Visualizing Complex Data and Relations

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



Newsmap: Google News Stories

Similarity and Dissimilarity

- ***Similarity***
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- ***Dissimilarity*** (e.g., distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- ***Proximity*** refers to a similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

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

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

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

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

Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- **Method 1**: Simple matching
  - *m*: # of matches, *p*: total # of variables

$$d(i,j) = \frac{p - m}{p}$$

- **Method 2**: Use a large number of binary attributes
  - creating a new binary attribute for each of the *M* nominal states

Different 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: (negative matches not important)

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

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

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

Dissimilarity between Binary Variables

- **Example**

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

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

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

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

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

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

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 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

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

where

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

  - standardized measure (*z-score*):

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

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

Example:
Data Matrix and Dissimilarity Matrix

Data Matrix

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

Dissimilarity Matrix

(with Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3,61	0		
x3	2,24	5,1	0	
x4	4,24	1	5,39	0

Distance on Numeric Data: Minkowski Distance

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

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

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

- Properties
  - d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positive definiteness)
  - 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**

Special Cases of Minkowski Distance

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

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

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

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

- $h \to \infty$. "***supremum***" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component (attribute) of the vectors

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

Example: Minkowski Distance

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

Dissimilarity Matrices

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

Cosine Similarity

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

Document	team	coach	hockey	baseba	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

- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

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

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

Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\|$,
  where $\bullet$ indicates vector dot product, $\|d\|$: the length of vector d

- **Example**: Find the ***similarity*** between documents 1 and 2.

d_1 **=** (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)

d_2 **=** (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$

$\|d_1\| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\|d_2\| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$

$\cos(d_1, d_2) = 0.94$

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 knowledge discovery
- Many methods have been developed but still an active area of research

Knowledge Technology lab research