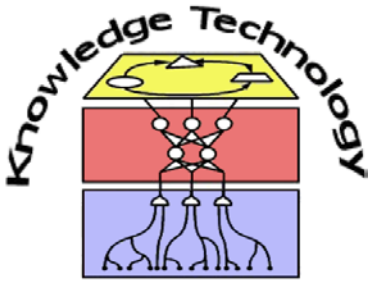
Data Mining

Lecture 2
From Data to Visualisation

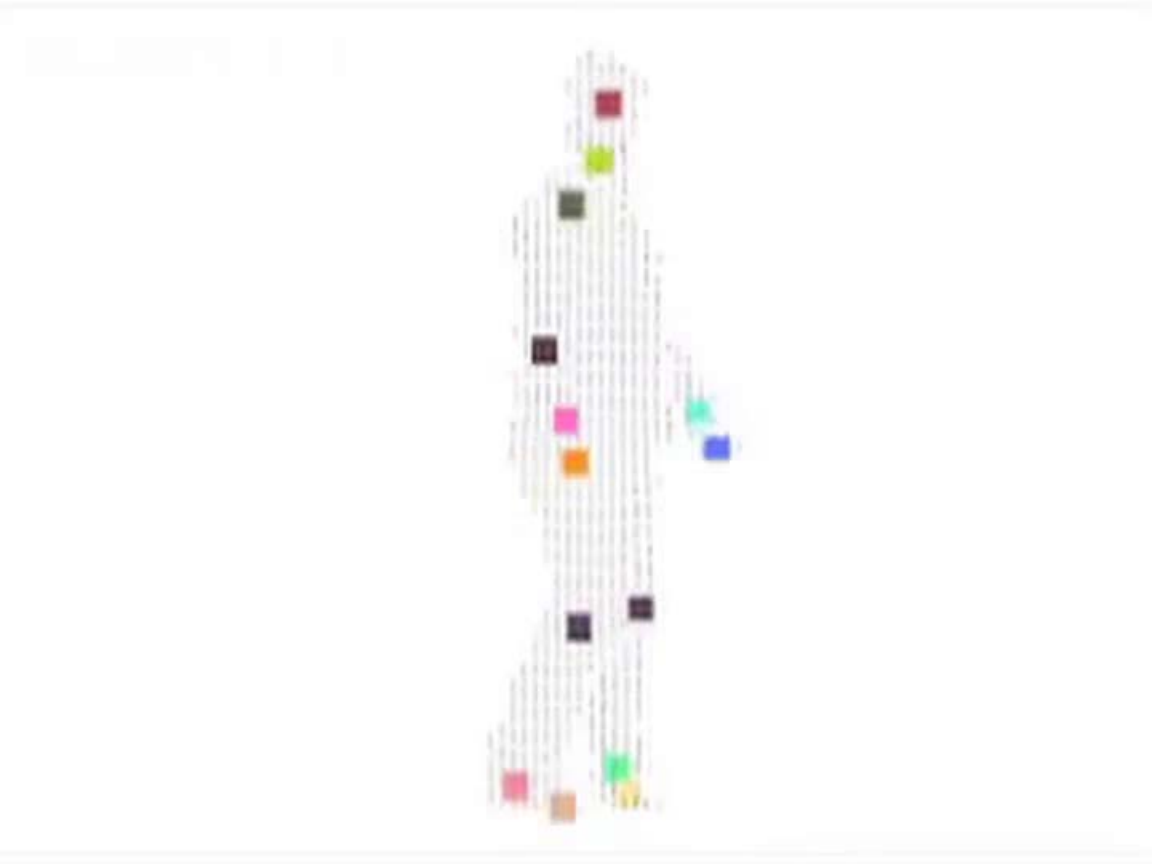


http://www.informatik.uni-hamburg.de/WTM/

Important Characteristics of structured Data

- Dimensionality
  - Curse of dimensionality
- Resolution
  - Patterns depend on the scale
- Sparsity
  - Few values are present
- Distribution
  - Centrality and dispersion
- Similarities
  - Find outliers

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



http://www.newscientist.com/

Types of Data

- Structured Records
  - Tables
  - Transaction data
  - Relational records
- 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. Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- A ***data object*** represents an entity
  - Also called *sample, example, instance, data point, tuple*
- Data objects are described by ***attributes***
- A data set as a ***matrix***:
  - rows -> data objects; columns ->attributes
- A ***database*** is an organised collection of data (sets)

Attributes

- ***Attribute*** (or ***dimensions***, ***features***, ***variables***):
  - a data field, representing a characteristic 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
  - However, there is no meaningful order
- ***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., cancer positive)
- ***Ordinal***
  - Values have a meaningful order (ranking)
  - However, magnitude between successive values is not known
  - *Size* = {*small, medium, large*}, army rankings, grades

Numeric Attribute Types

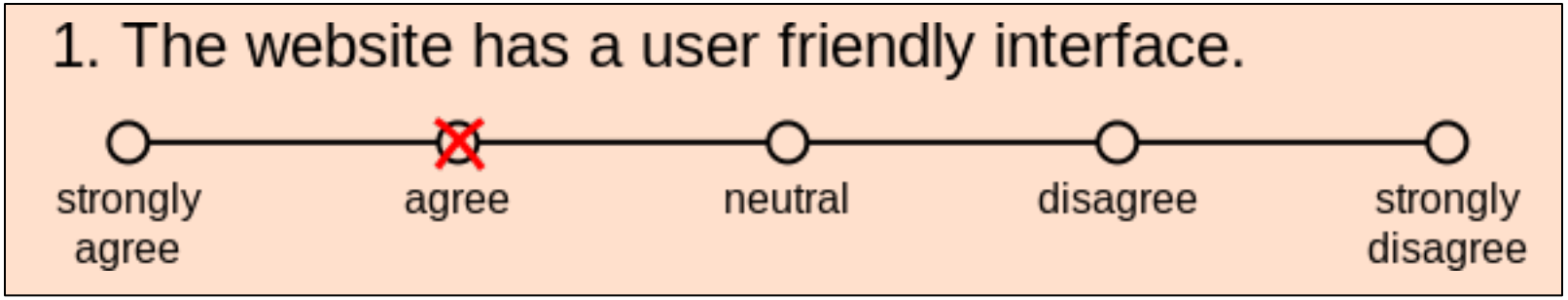
- ***Interval***
  - Measured on a scale of ***equal-sized units***
  - Values have ***order***
    - **Examples:** *temperature in C˚or F˚, calendar dates*
  - Differences between units can be ***quantified***
  - However, no true zero-point
- ***Ratio***
  - Inherent ***zero-point***
  - We can distinguish values by order of magnitude
    - "100 is 3 orders of magnitude larger than 0.1"
    - **Examples:** *temperature in Kelvin, durations of events, 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 with fixed origin	Kelvin, length, counts, age, income	+, -, *, /

Likert Scale

- **Example:**



- Of which type are the attributes of a Likert scale?
  - Nominal ✓
  - Ordinal ✓
  - Interval ✗ (not well-defined intervals)
  - Ratio ✗

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
  - Can all be mapped to integer values
  - Special case: Binary attributes
- ***Continuous Attribute***
  - Has continuous values
    - **E.g.**, temperature, height, or weight
  - One cannot list all possible values
  - Typically, real numbers represented as floating-point variables
    - Practically, represented using a finite number of digits

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

- Some data do not change with time and are considered as ***static data*.**
- Some attribute values do 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** (Secondary, Retrospective)
  - Massive non-experimental data sets
    - **E.g.**, human genome, atmospheric data, retail data, web logs for Amazon, Google, etc.
  - Not constrained by experimental design
  - 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.

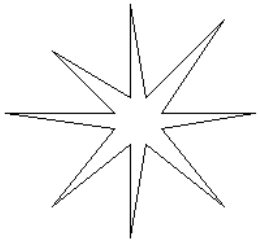
Curse of Dimensionality
(Geometric Approach II)

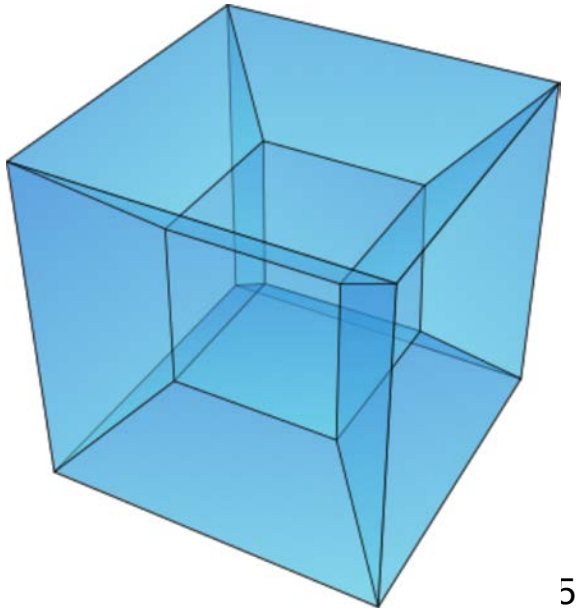
For example:

- A high-dimensional hypercube may be visualized as a porcupine (or even a hedgehog, as small 3D things have more surface per volume)

  surface~length^2
  volume~length^3

- As the dimensionality grows, the surface grows relative to the central part of the hypercube.



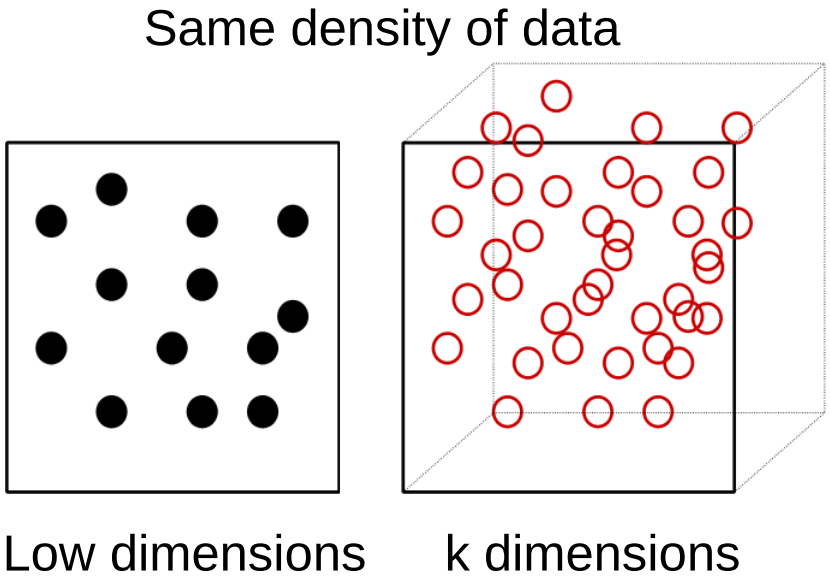


Curse of Dimensionality (1)

- 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.

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



Curse of Dimensionality (2)

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

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

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

- **Example:**

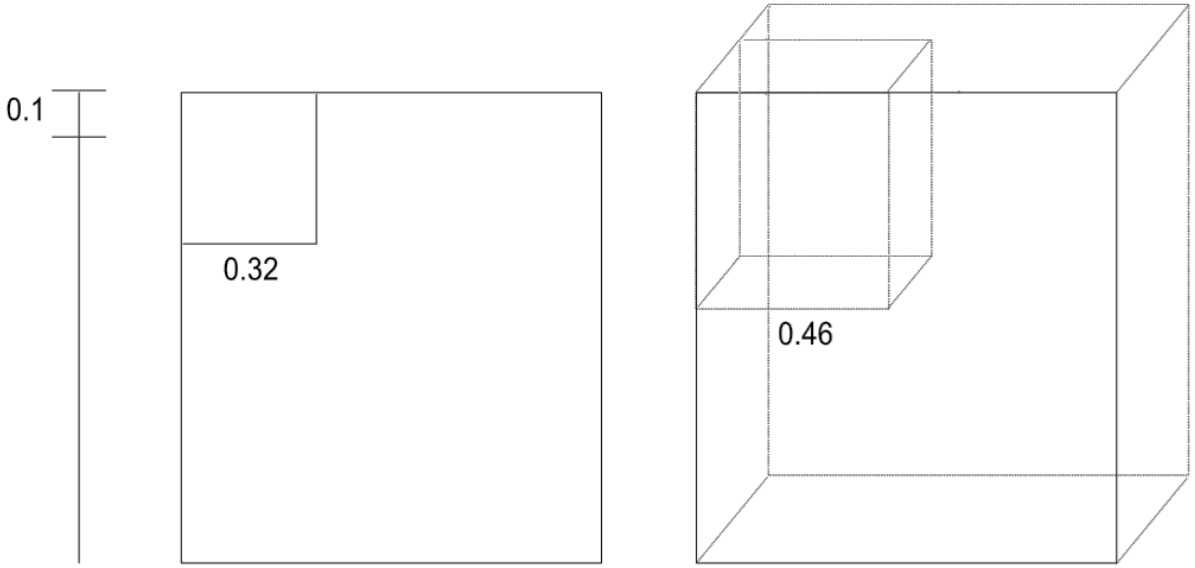
  10% of the samples (p=0.1):

  1-D: $e_1(0.1) = 0.1$

  2-D: $e_2(0.1) = 0.32$

  3-D: $e_3(0.1) = 0.46$

  10-D: $e_{10}(0.1) = 0.8$



Curse of Dimensionality (3)

- In a high-dimensional space
  - ***Almost every point is close to an edge***
  - The distance to another sample point gets large:

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

$$D(d, n) = \frac{1}{2} \cdot \left(\frac{1}{n}\right)^{1/d} = \frac{0.5}{\sqrt[d]{n}}$$

- **Example, expected distance between 10000 points:**

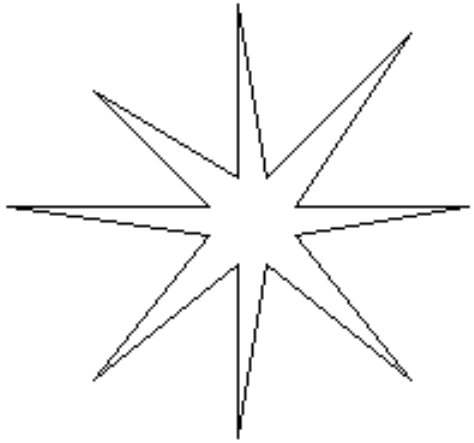
For a 2-D space: $\rightarrow D(2,10000) = 0.005$

For a 10-D space: $\rightarrow D(10,10000) \approx 0.2$

different in the Kantardzic book!

Curse of Dimensionality (4)

- Almost every point is an ***outlier*** in high-dimensional spaces:
  - With increasing input dimension, the distance between the prediction point and the center of some classified points increases.



  - d=10: expected value of prediction point is 3.1 SD away from the center of the data belonging to one class
  - d=20: the distance is 4.4 SD

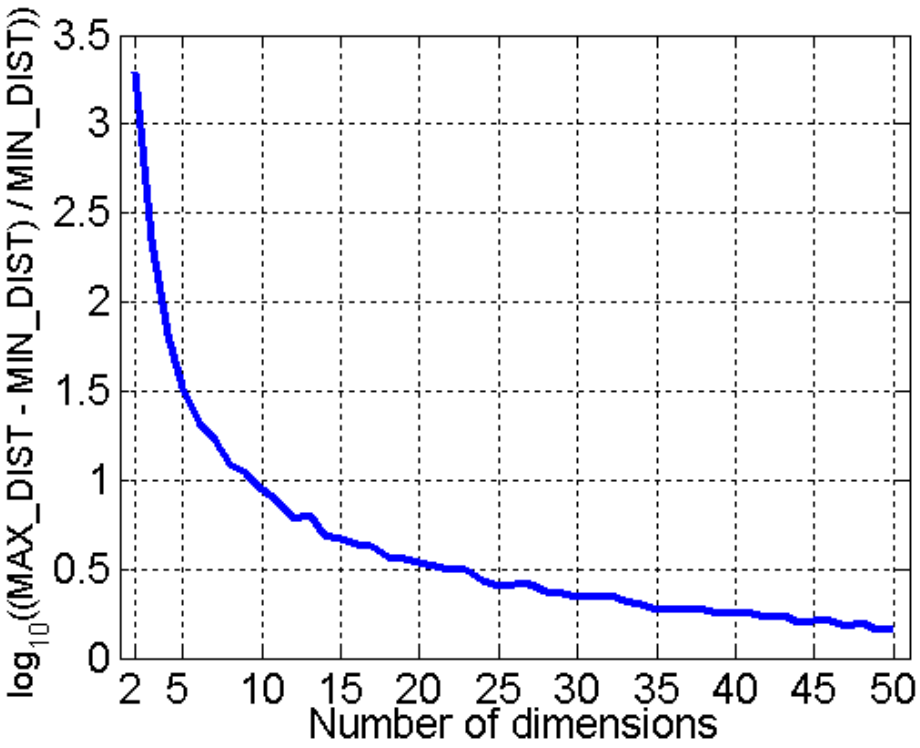
Curse of Dimensionality (5)

Experimental Confirmation:

With higher dimensionality:

- distances between data points become more similar

- data becomes increasingly ***sparse***
  → "density" looses its meaning, if not backed by sufficiently many data
- most are ***outliers***
  → "distance" less meaningful



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

Curse of Dimensionality (Summary)

As the dimension increases:

(1) we need exponentially more data for constant density,

(2) a hypercube of larger edge length covers same subspace,

(3) distance to an edge decreases,

(4) distance between points increases,

(5) every point becomes an outlier.

(1),(2) → difficulty in making local estimates; we need more and more samples to satisfy mining requirements.

(4),(5) → difficulty of predicting a response at a given point, since a new point will be far from the training examples.

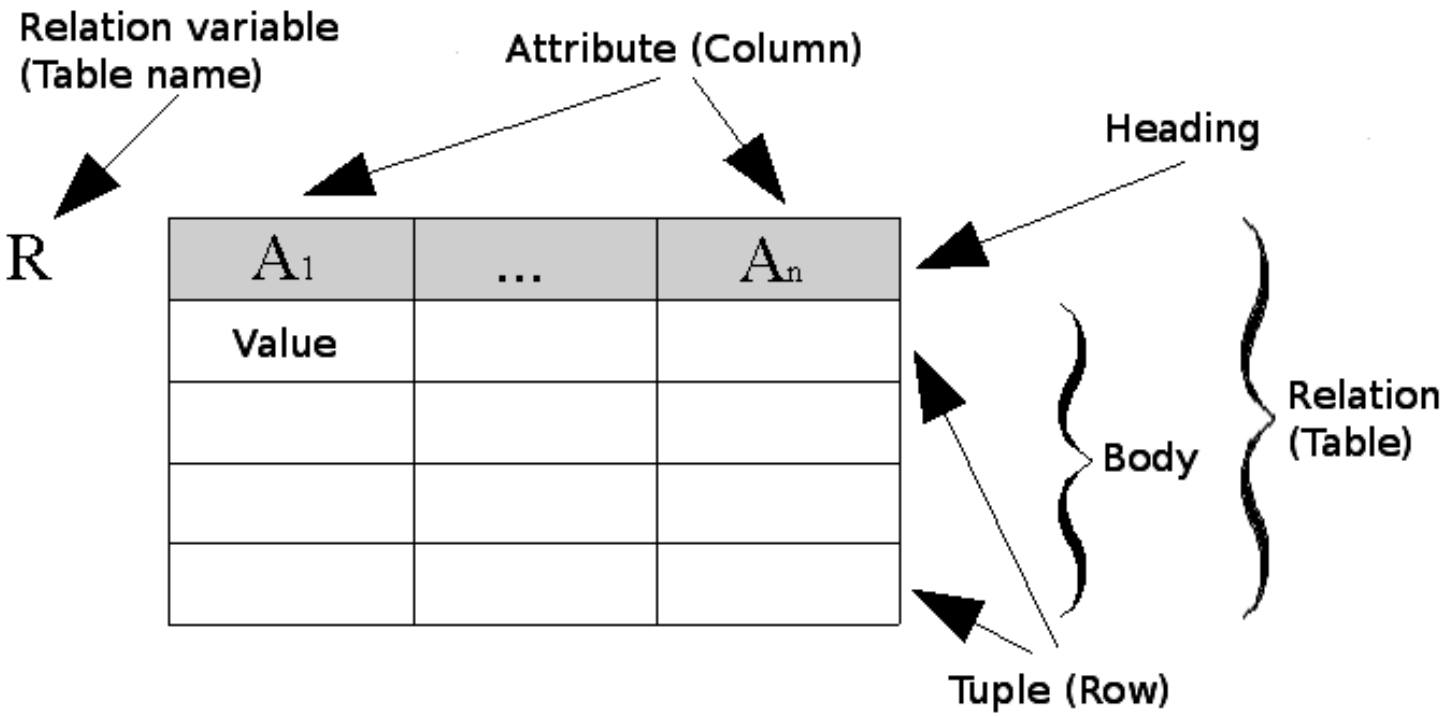
Preparing the Data for Data Mining

Two central tasks for the preparation of data:

- To organize data into a standard form, typically, a ***relational table*** (or tables)
- To prepare data sets by ***preprocessing***, such as dimensionality reduction

  … that will lead to the best data mining performance

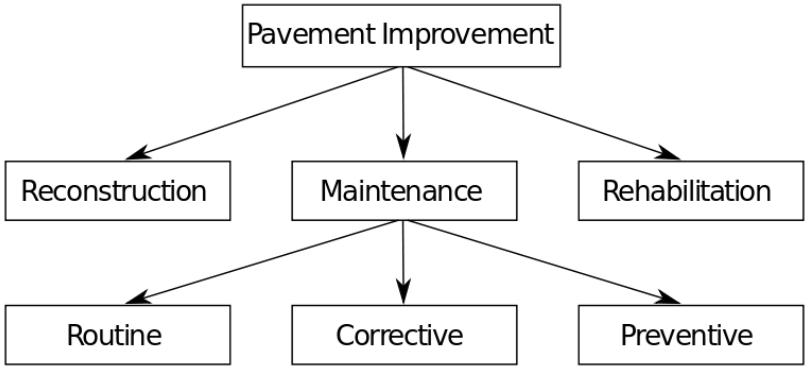
Relational Database Model



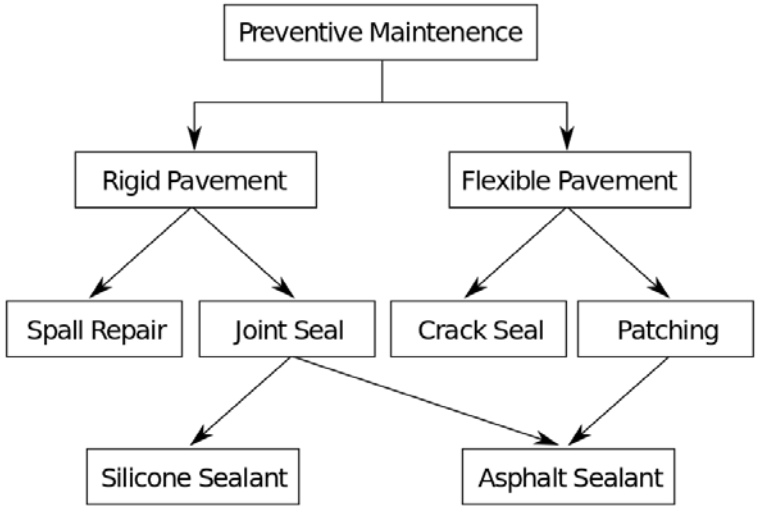
- Relation ~ set of n-tuples
- Tuples have no order, but attribute names are used instead
- An attribute may serve as a key to link to other tables
- Mostly use SQL data definition and query language

Other Database Models (Examples)

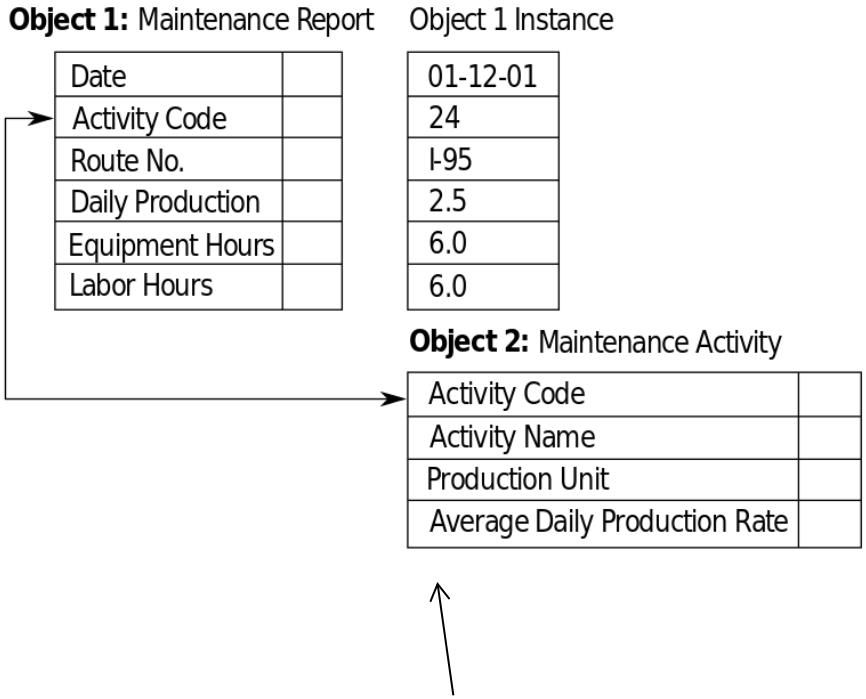
Hierarchical Model



Network Model

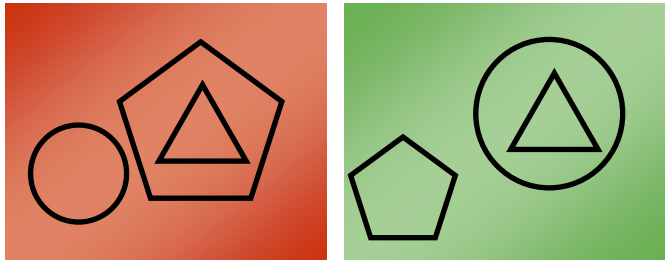


Object-Oriented Model



Information represented as objects in object oriented programming

Representing Data with Tables



Scene S1 Scene S2

Relational Representation

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

Single Table Representation

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

Representing Data with Tables: Market Baskets

Each basket represents one sample

TID: 100

TID: 200

TID:300

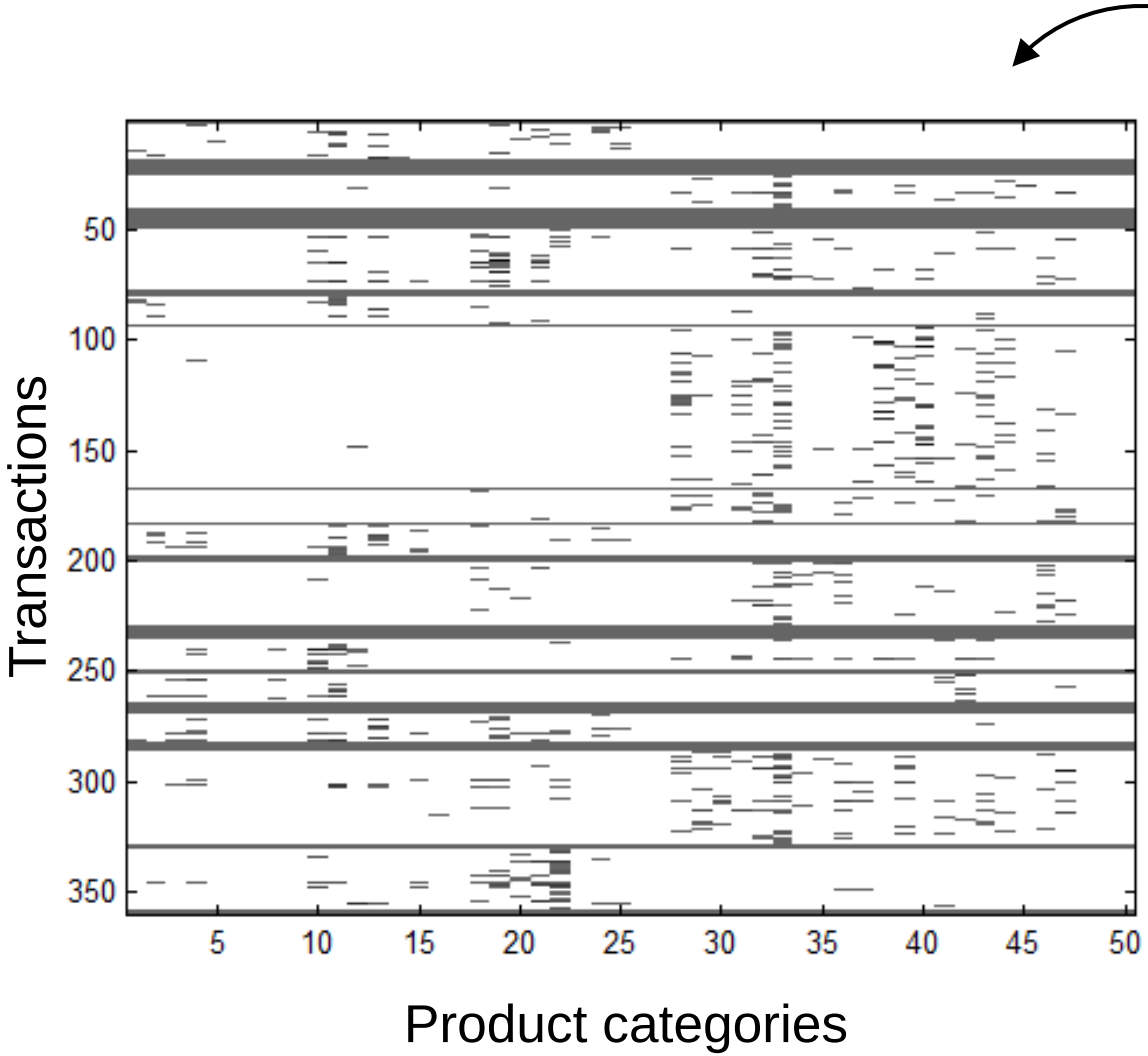
TID: 400

Sparsity:
eliminate “No‘s”

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

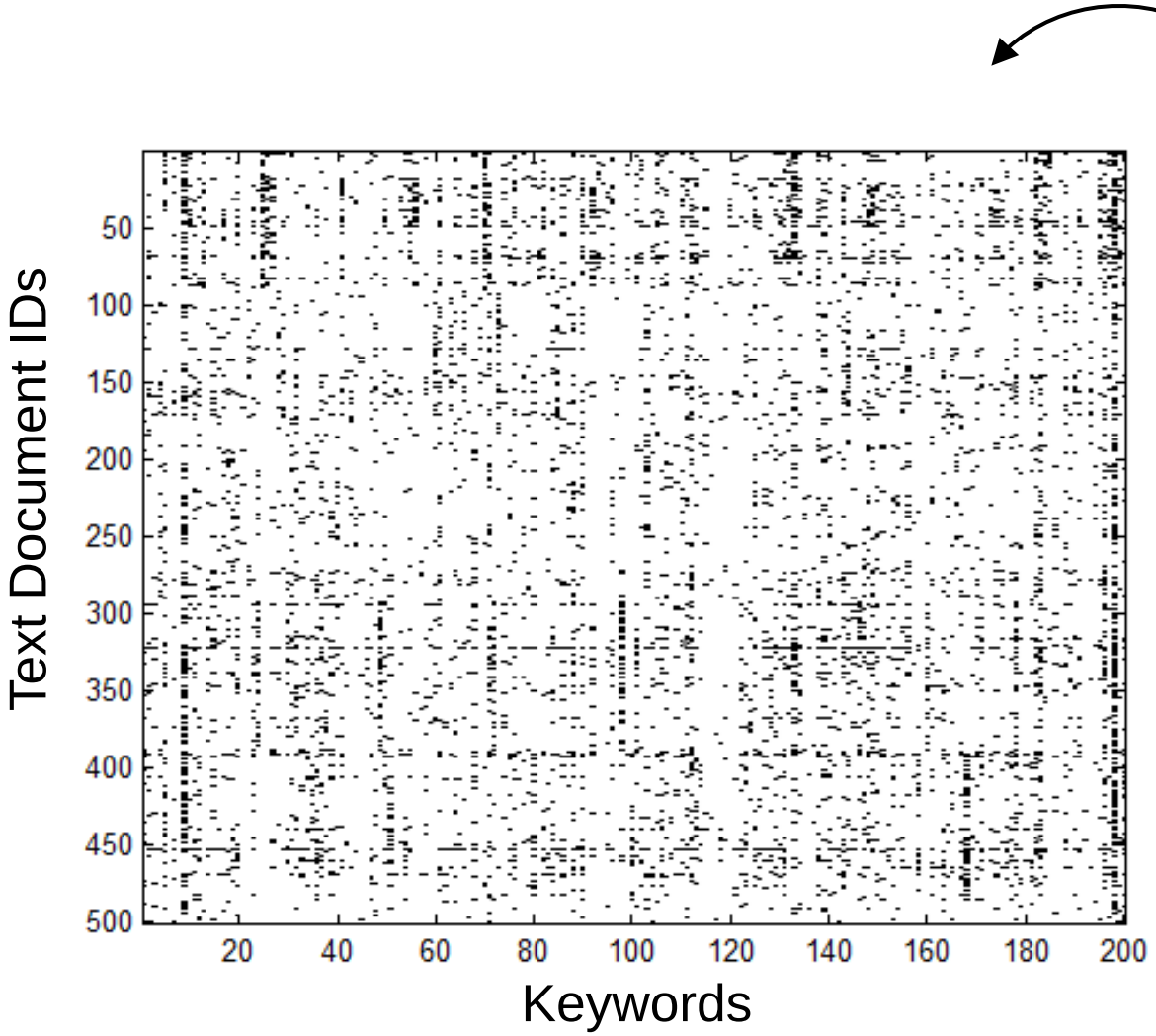
TID	Items
100	{Garlic, Detergent, Ketchup}
200	{Milk, Detergent, Wine}
300	{Garlic, Milk, Detergent, Wine}
400	{Milk, Wine}

Market Basket Data



TID	Items
01	01, 03, 44, 76
02	22, 37, 76
...	...

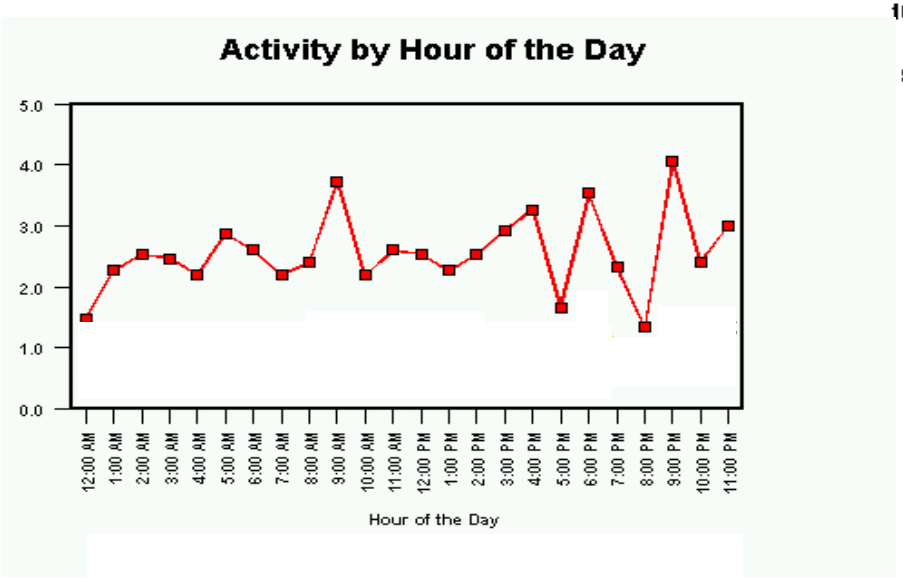
Representing Text as Tables

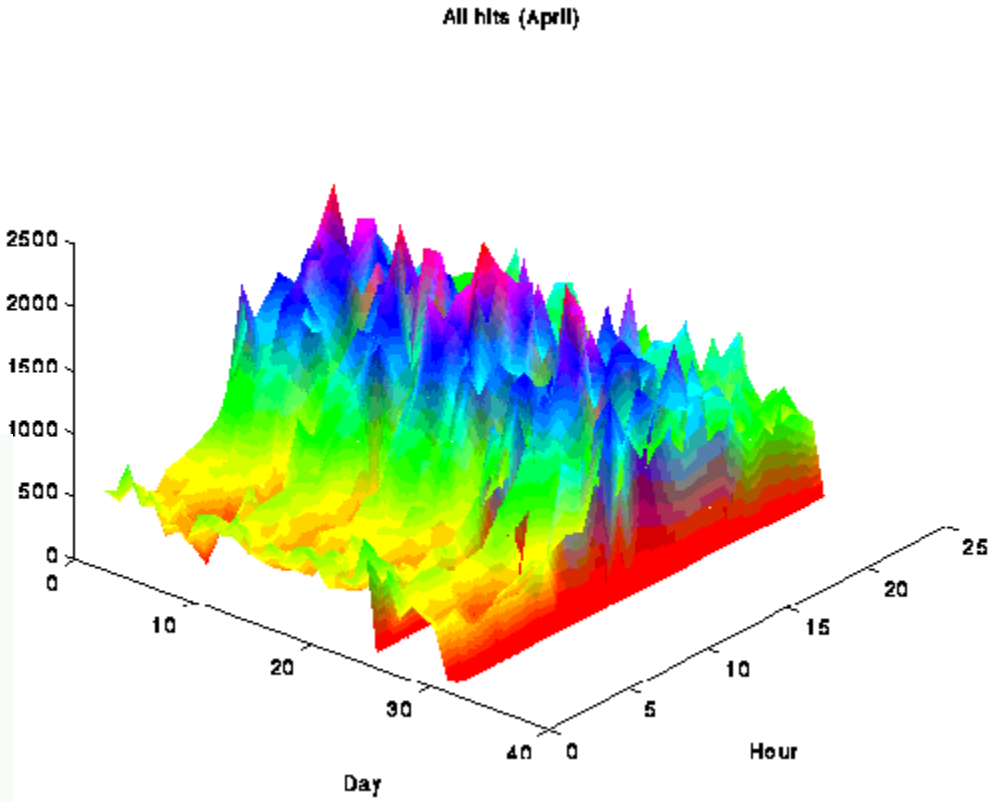


Text ID	Keywords
001	56, 34, 79
002	07, 122, 189
…	…

Web Log Data over Time as a Table

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





Time Series Data as a Table

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

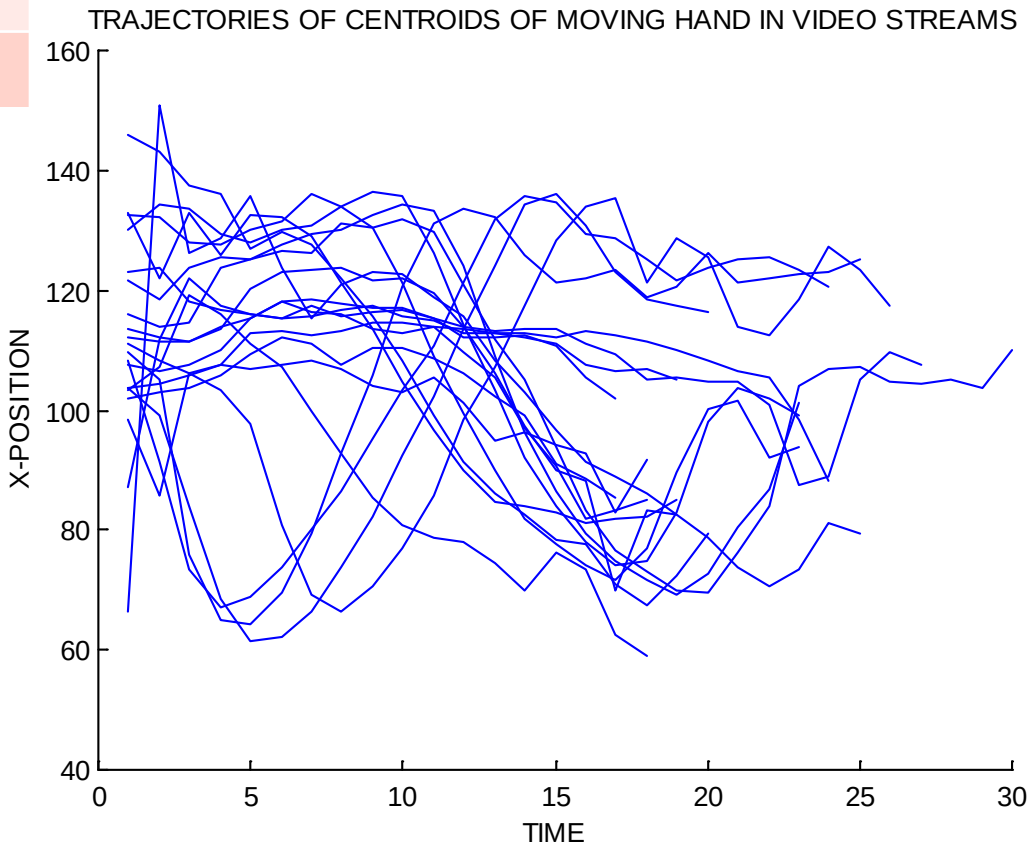
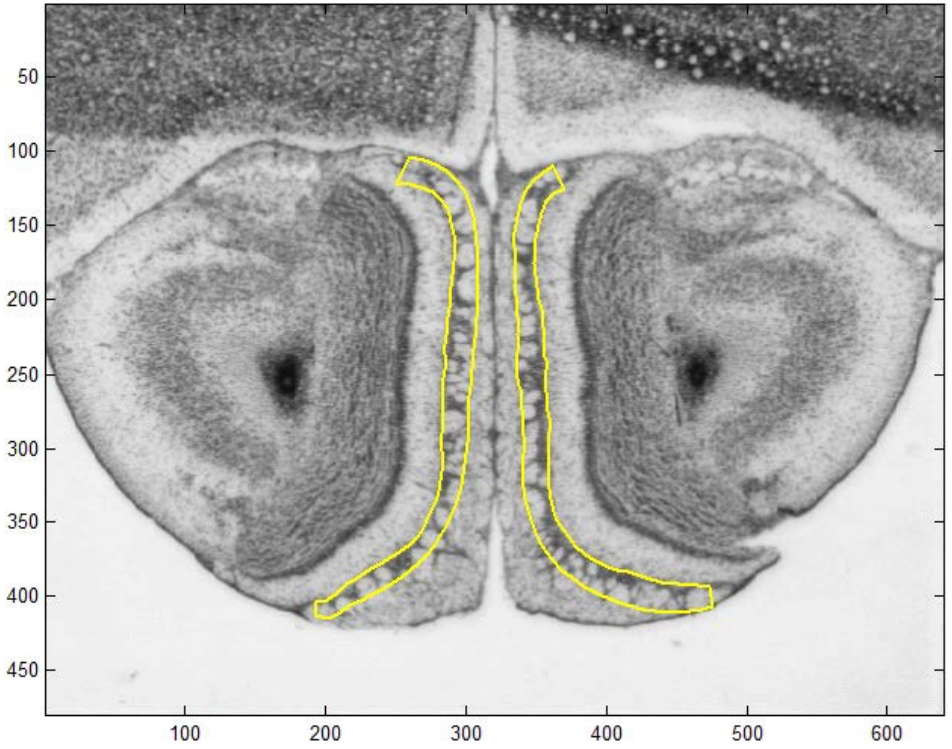


Image Data as a Table

X coor.	Y coor.	red	green	blue
100	250	87	107	43
100	251	85	104	39
…	…	…		

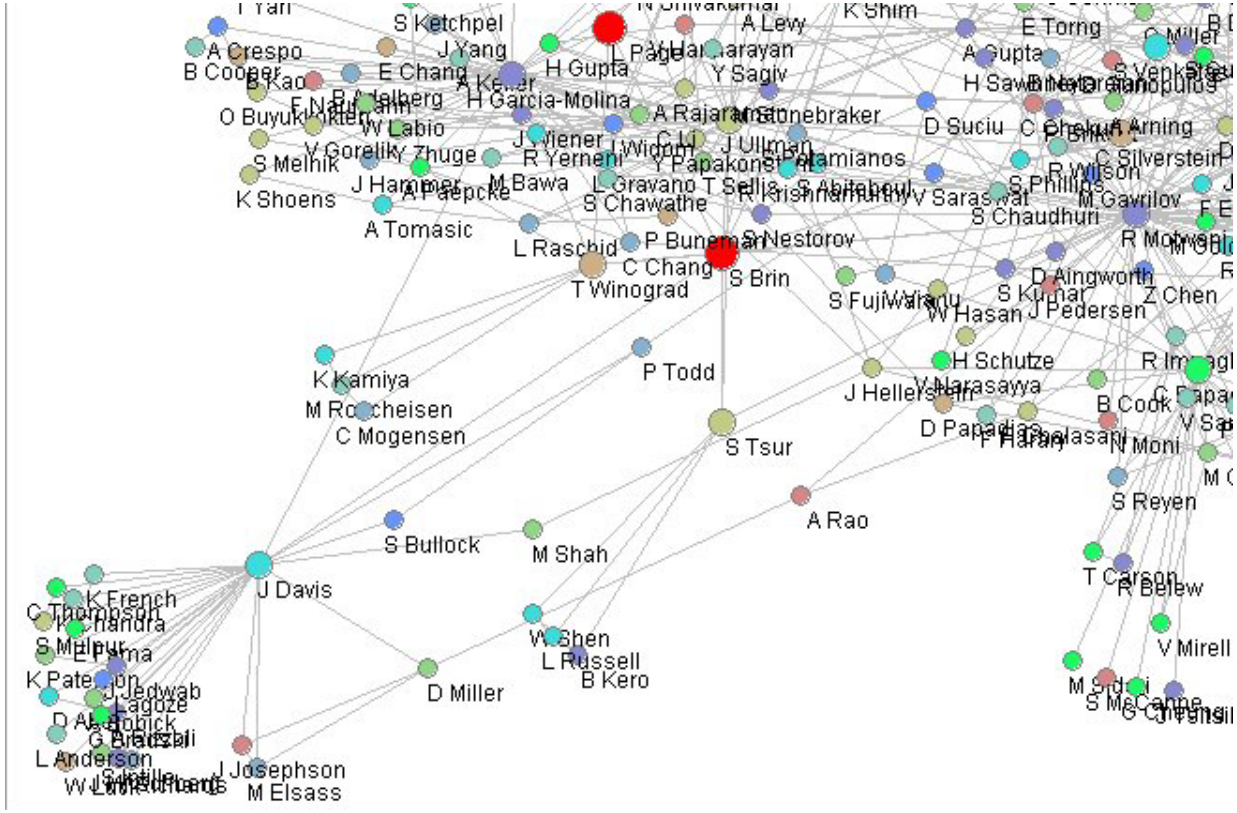
Geographical images …



… or medical

Relational Data (=*Graph*) as a Table

Beginning node	Ending node	Distance
Bullock	Todd	134
Miller	Davis	87
…	…	…



Each row contains the beginning and ending node in one connection, and weight factor (here distance) connected with this link.

Basic Statistical Descriptions of Data

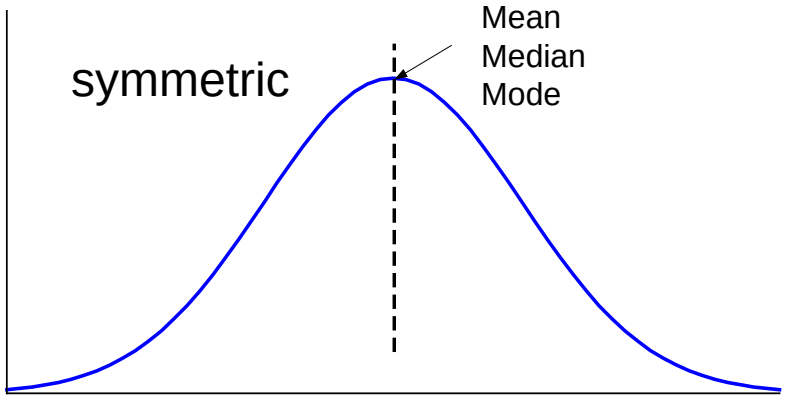
- Motivation
  - To better ***understand*** the data: central tendency, variation and spread
- Data ***dispersion characteristics***
  - median, max, min, quantiles, outliers, variance, etc.
- **Numerical dimensions** correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities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

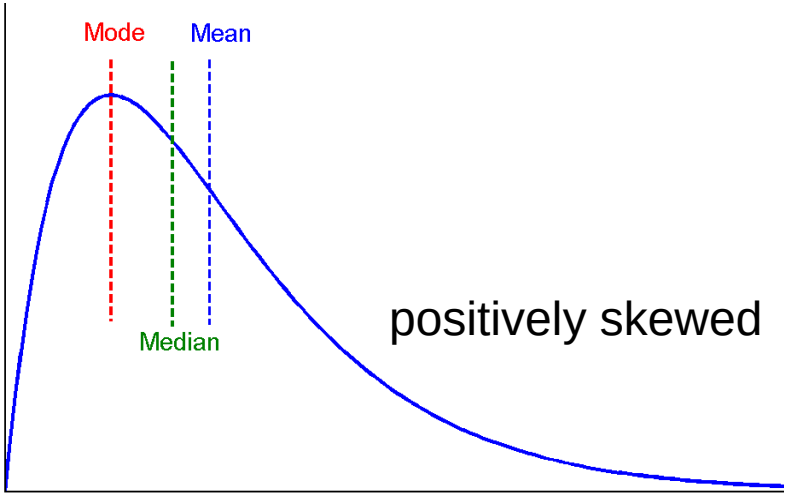
Measures of the Central Tendency

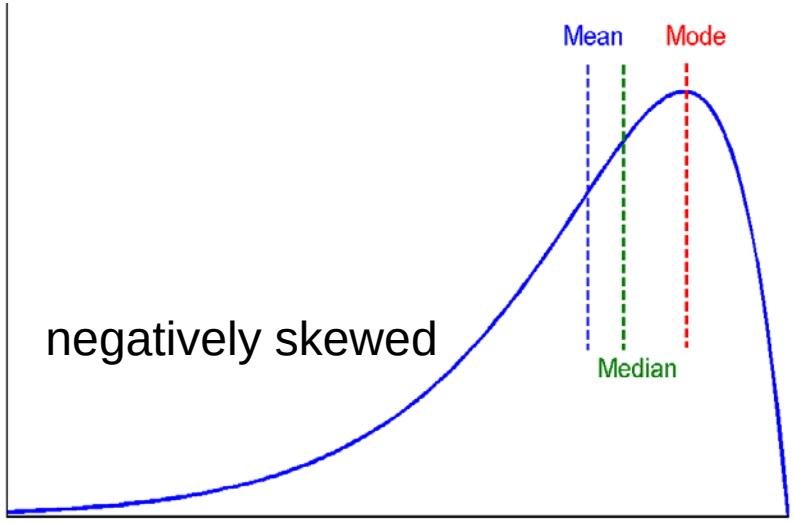
- ***Mean*** (algebraic measure):
  - Population mean (*N* = population size): $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - Mean estimated from samples (*n* = sample size): $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Mostly n << N
  - Weighted arithmetic mean: $\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- ***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

Symmetric vs. Skewed Data

- Symmetric data:
  - Median = mean = mode
- Skewed data:
  - Median ≠ mean ≠ mode







Empirical formula for moderately asymmetrical curves:

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

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - ***Quartiles***: Q1 (25th percentile), Q3 (75th percentile)
  - ***Inter-quartile range***: IQR = Q3 – Q1
  - ***Five number summary***: min, Q1, median, Q3, max
    → Visualised as a ***Boxplot***
- Variance and standard deviation
  - ***Variance***:

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

Variance estimated from sample:

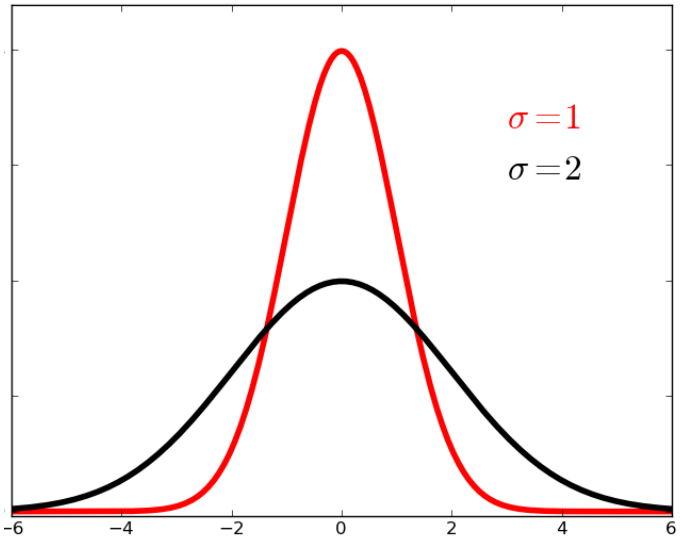
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  - ***Standard deviation*** σ is the square root of variance σ^2
  - Outliers contribute over-proportionally to the variance, due to the square

Symmetric Distribution: Normal (Gaussian)

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

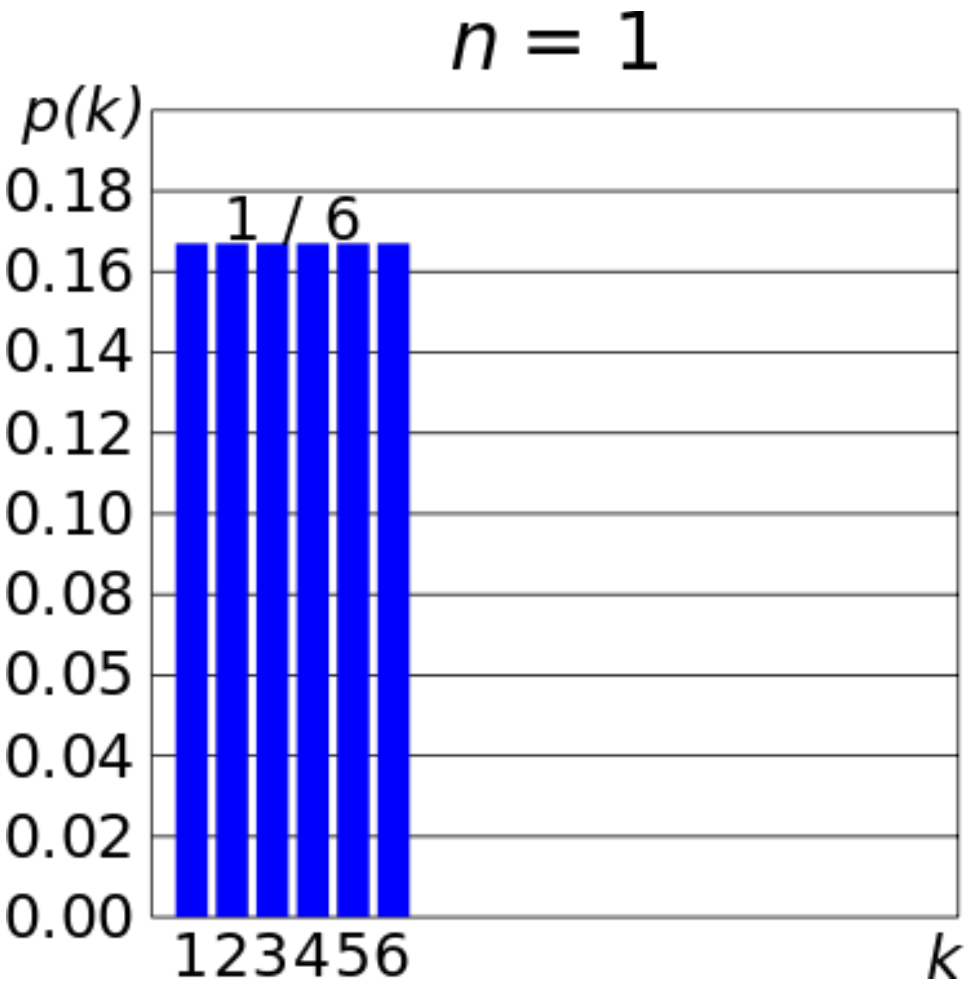
normalizer (not exact on discrete space!)



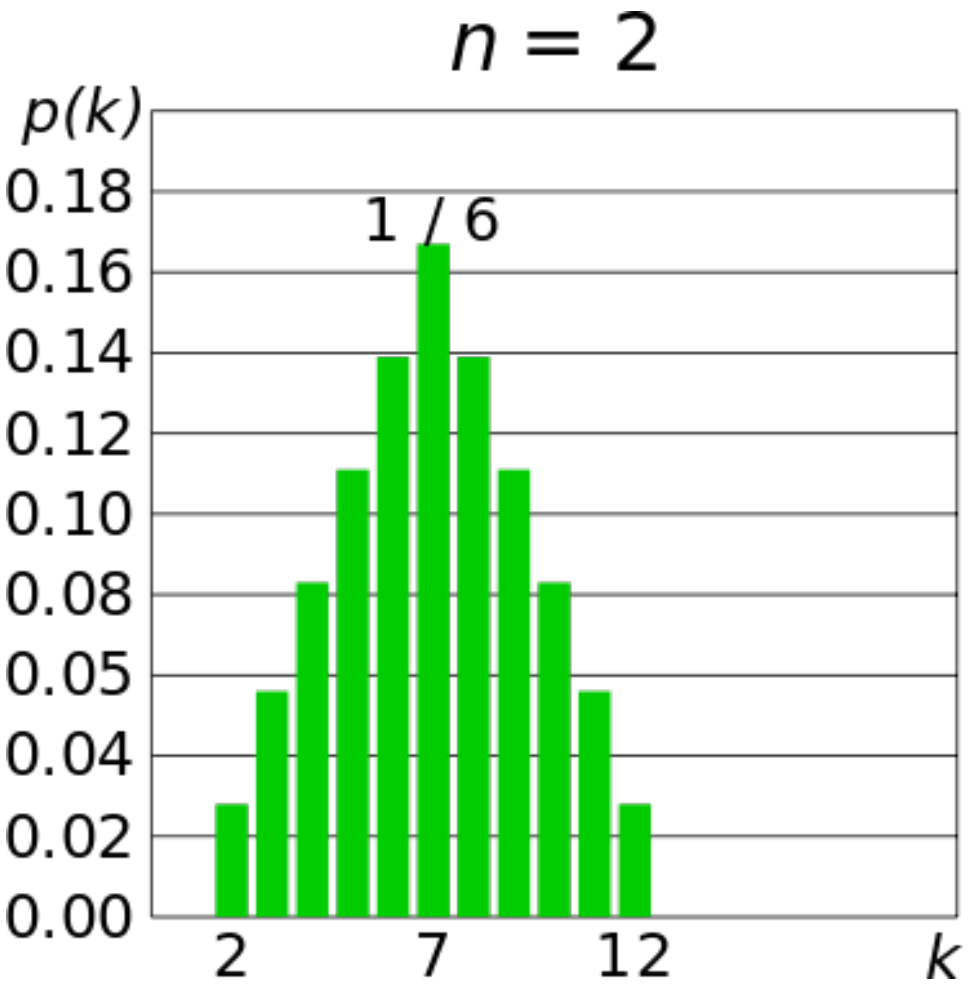
- Central Limit Theorem: (under certain conditions …) the sum of many random variables converges to a Gaussian

- **Example:** sum of n fair 6-sided dice

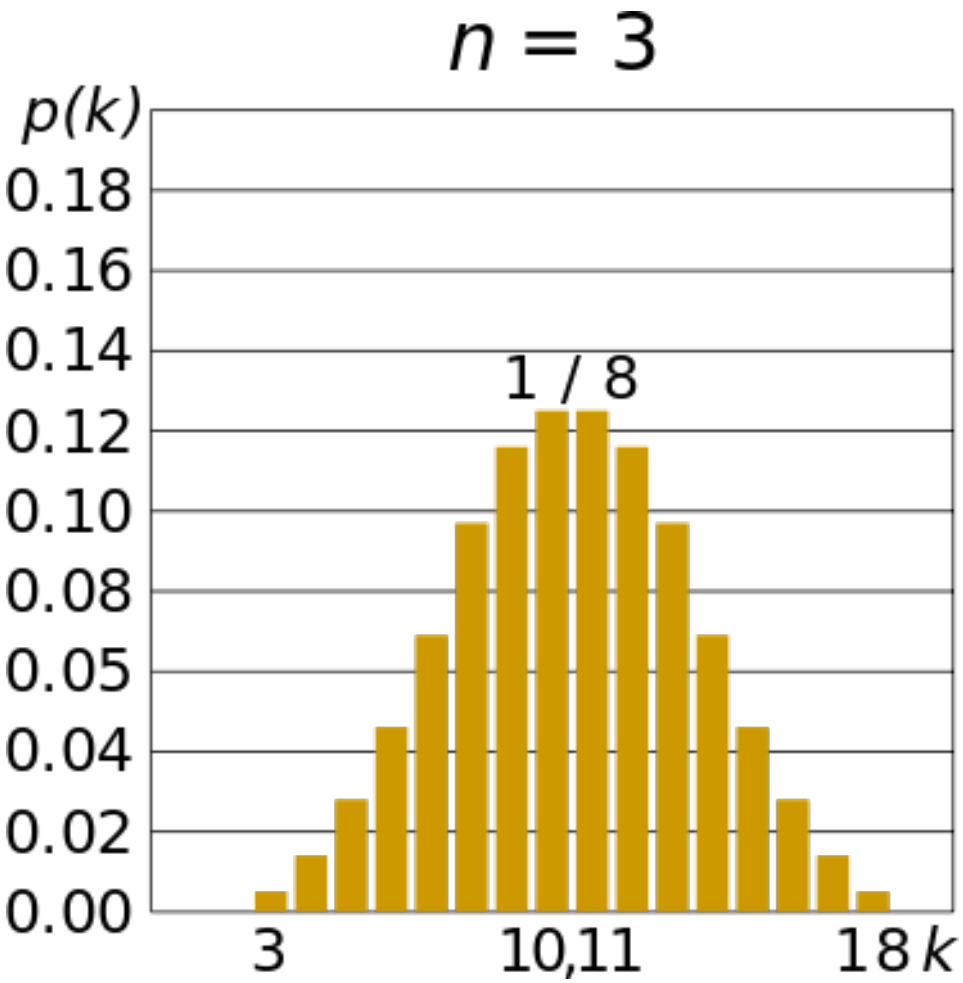
Sum of random variables: 6-sided dice



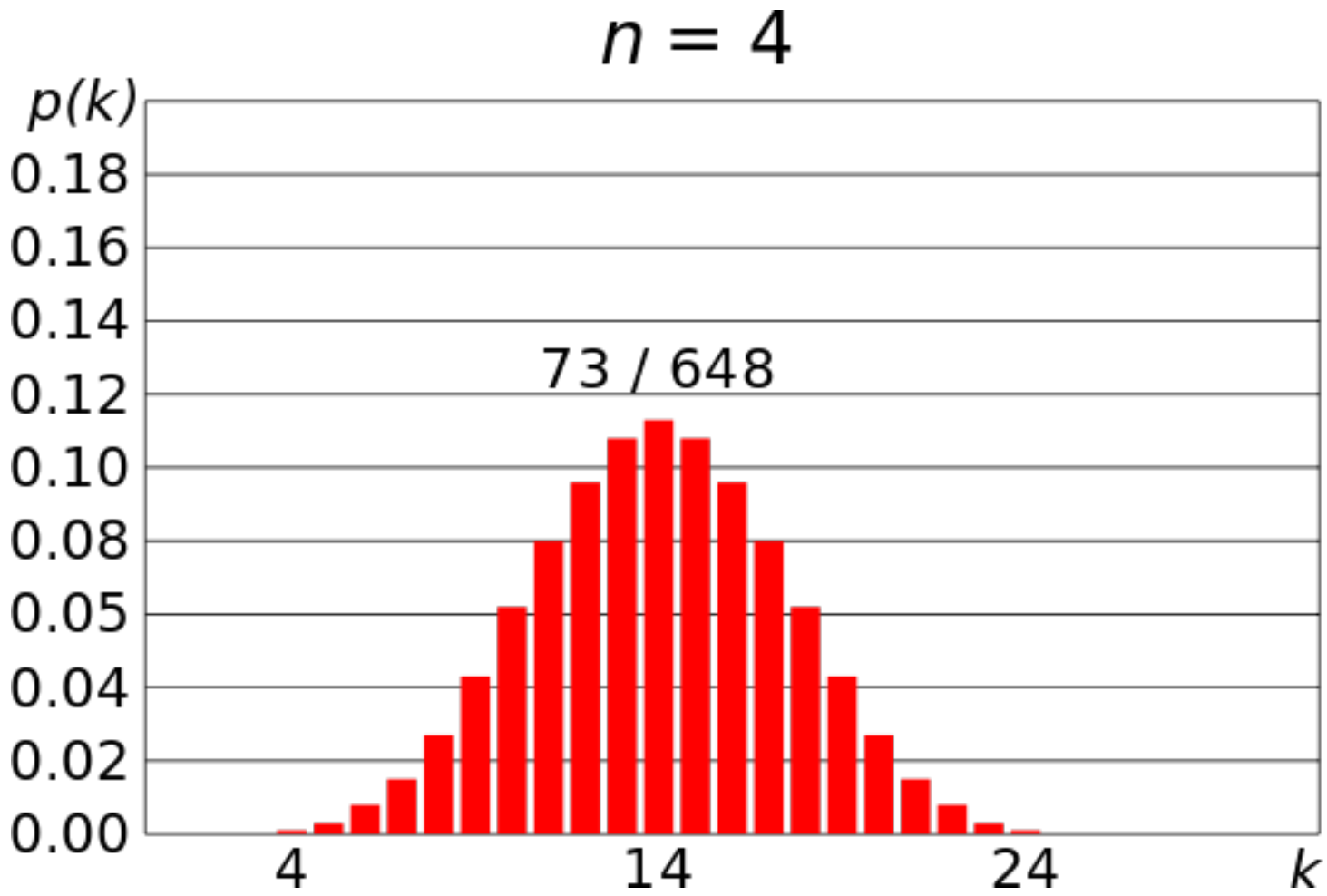
Sum of random variables: 6-sided dice



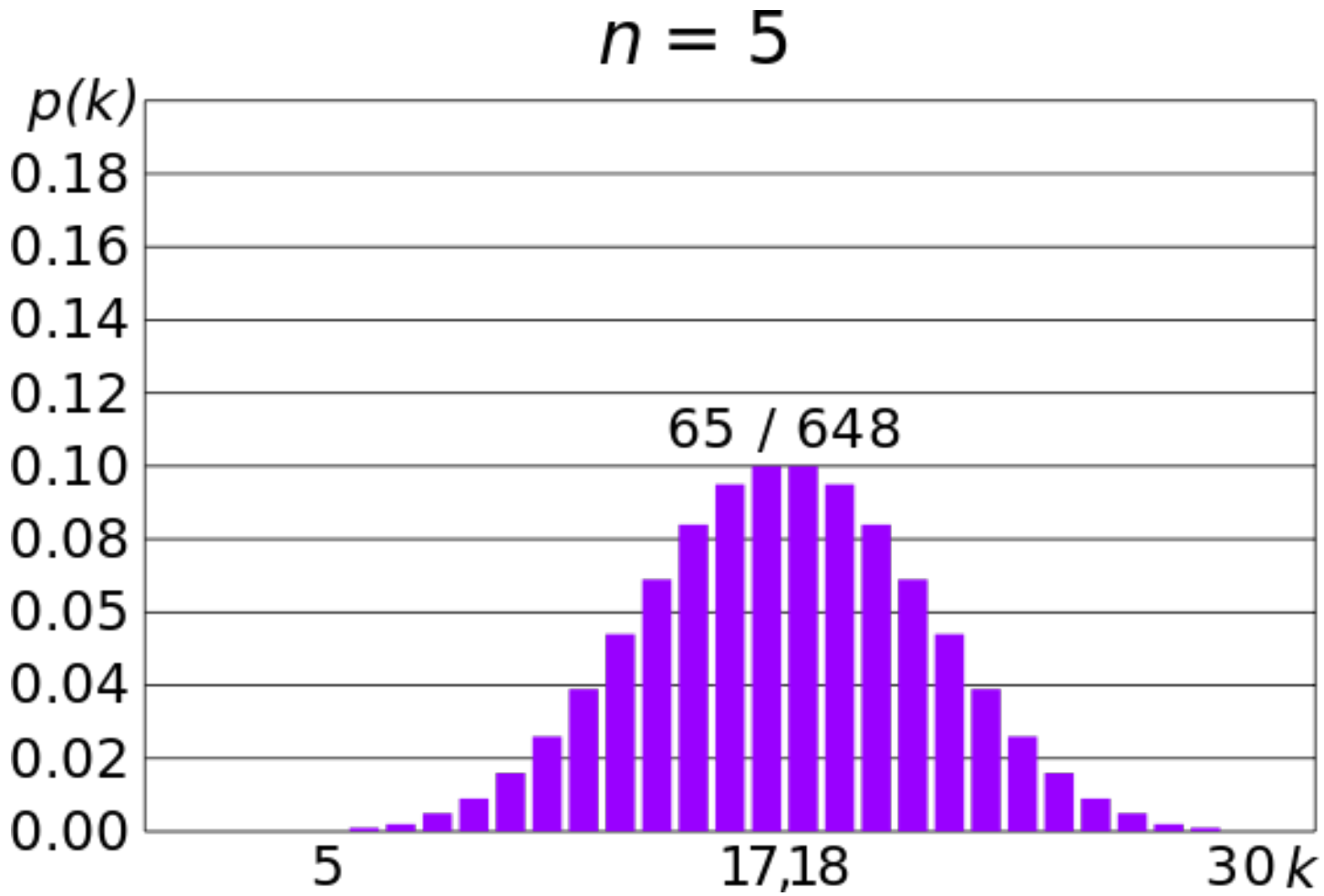
Sum of random variables: 6-sided dice



Sum of random variables: 6-sided dice



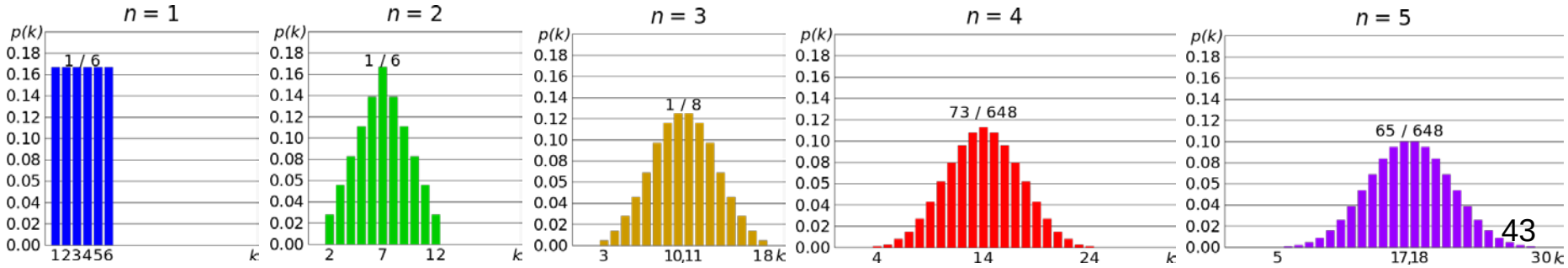
Sum of random variables: 6-sided dice



Symmetric Distribution: Normal (Gaussian)

- Central Limit Theorem: the sum of many random variables converges to a Gaussian

- **Example:** sum of n fair 6-sided dice



Symmetric Distribution: Normal (Gaussian)

- From $\mu-\sigma$ to $\mu+\sigma$:

  contains ~ 68%
  of the measurements

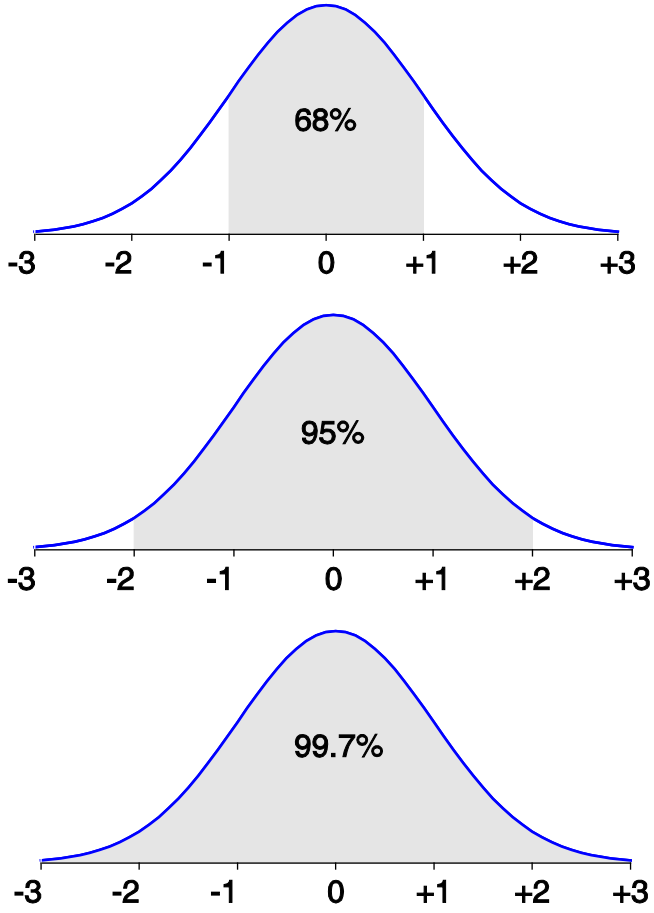
  (μ: mean, σ: standard deviation)

- From $\mu-2\sigma$ to $\mu+2\sigma$:

  contains ~ 95%

- From $\mu-3\sigma$ to $\mu+3\sigma$:

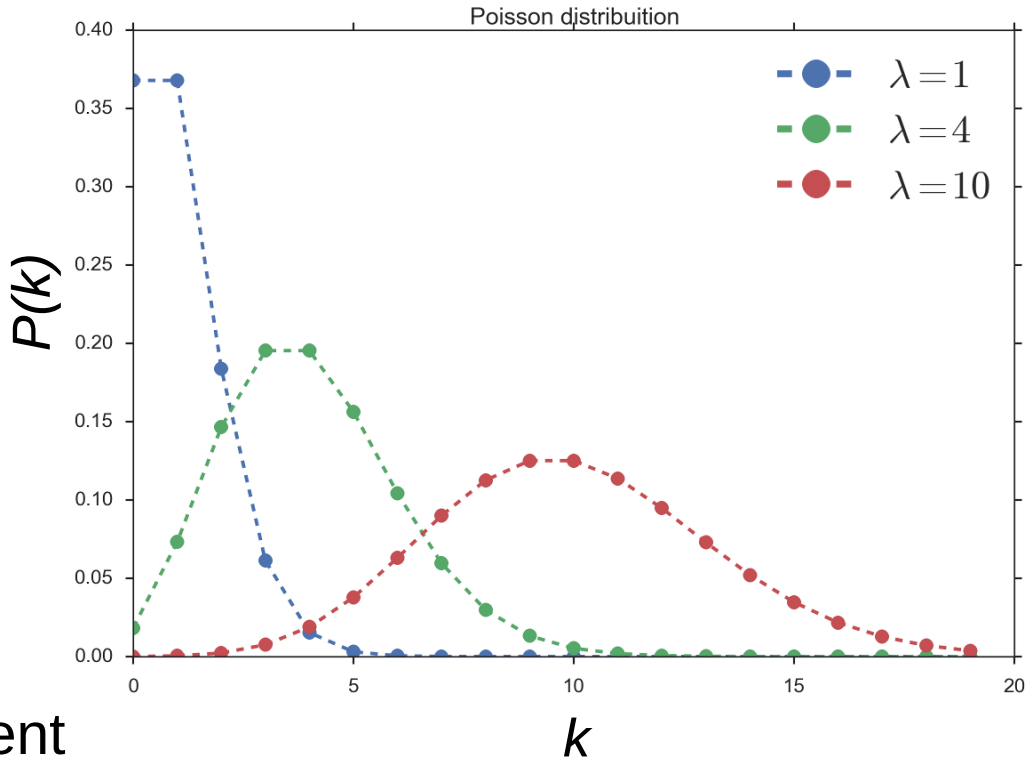
  contains ~ 99.7%



- Of all distributions with given mean and variance, the Gaussian maximises the entropy

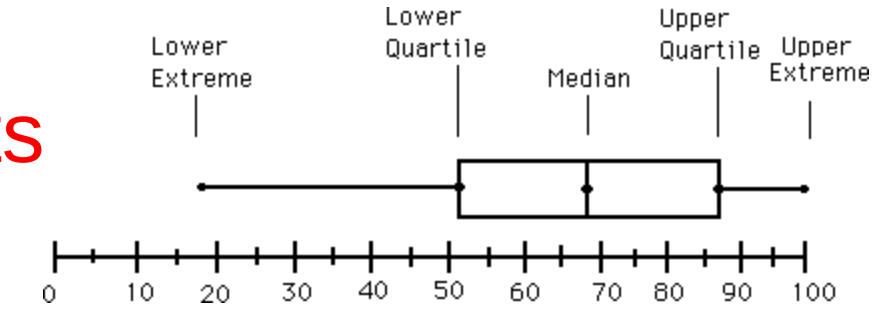
Skewed Distribution: Poisson

$$P(k \mid \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

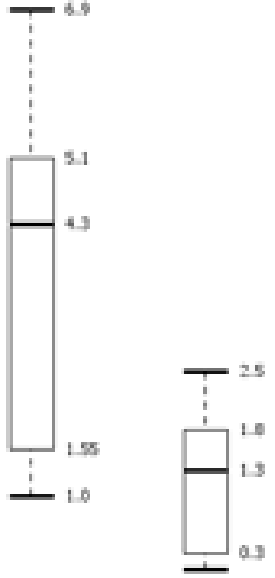


- Probablity that *k* events happen in a given interval
- λ = average number of events in an interval
- Events must be independent
- Large $\lambda \rightarrow$ ≈ Gaussian-like
- **E.g.:**
  - # meteors that hit earth per year
  - # patients arriving at an emergency room at a given hour
  - # neural spikes per second (model)

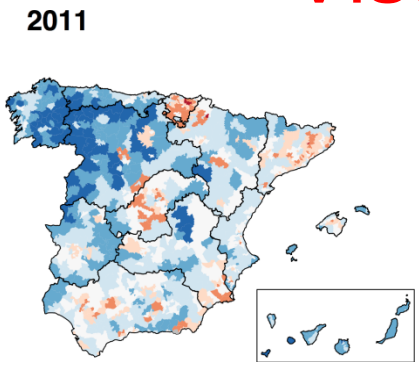
Box (-and-Whisker) Plots

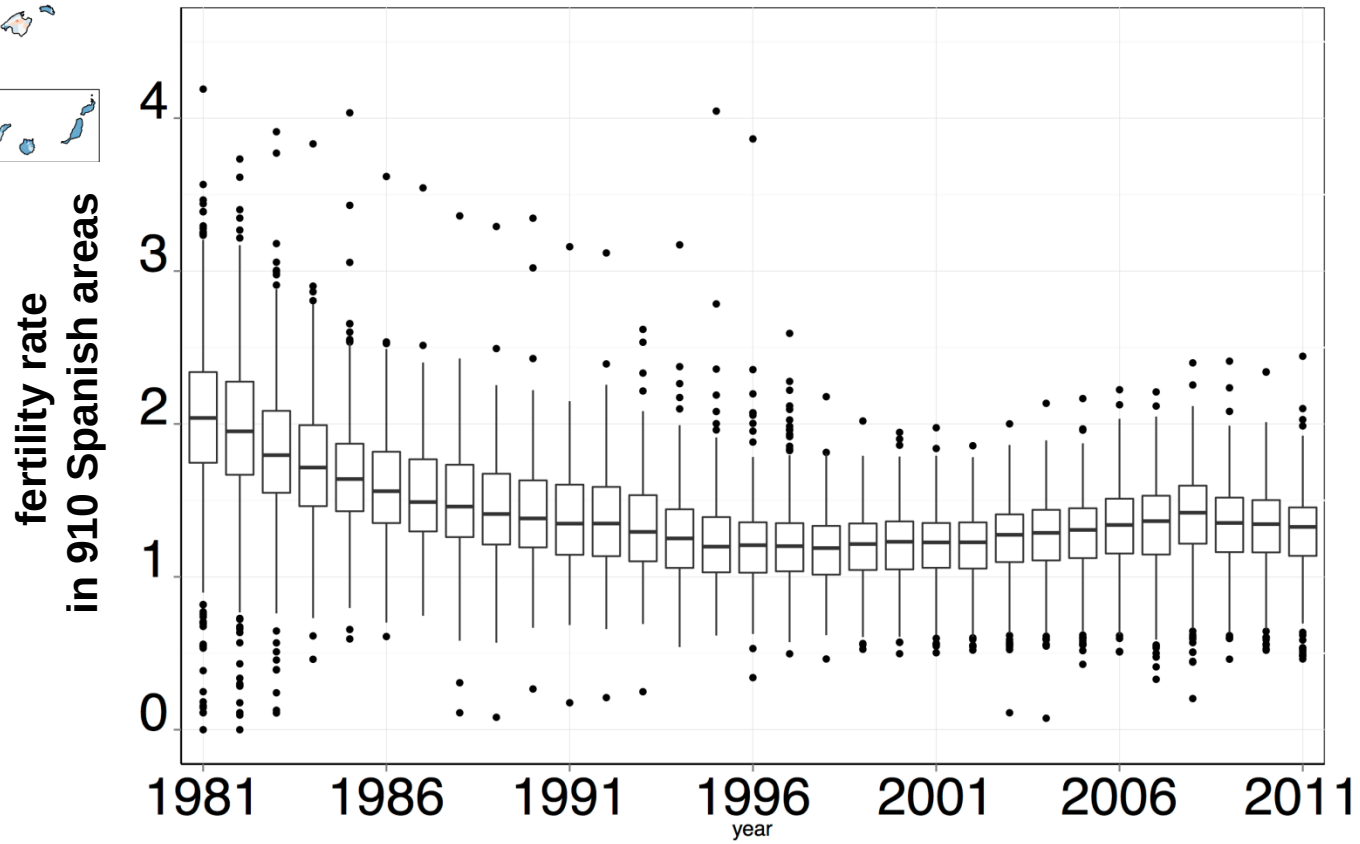


- ***Five-number summary*** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- ***Boxplot***
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is the interquartile range (IQR)
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to minimum and maximum
  - If outliers: points beyond specified thresholds, plotted individually e.g. value lower than Q1 - 1.5·IQR or higher than Q3 + 1.5·IQR. Whiskers extend only to the non-outlier data.



Visualization of Data Dispersion: Boxplot Time Series

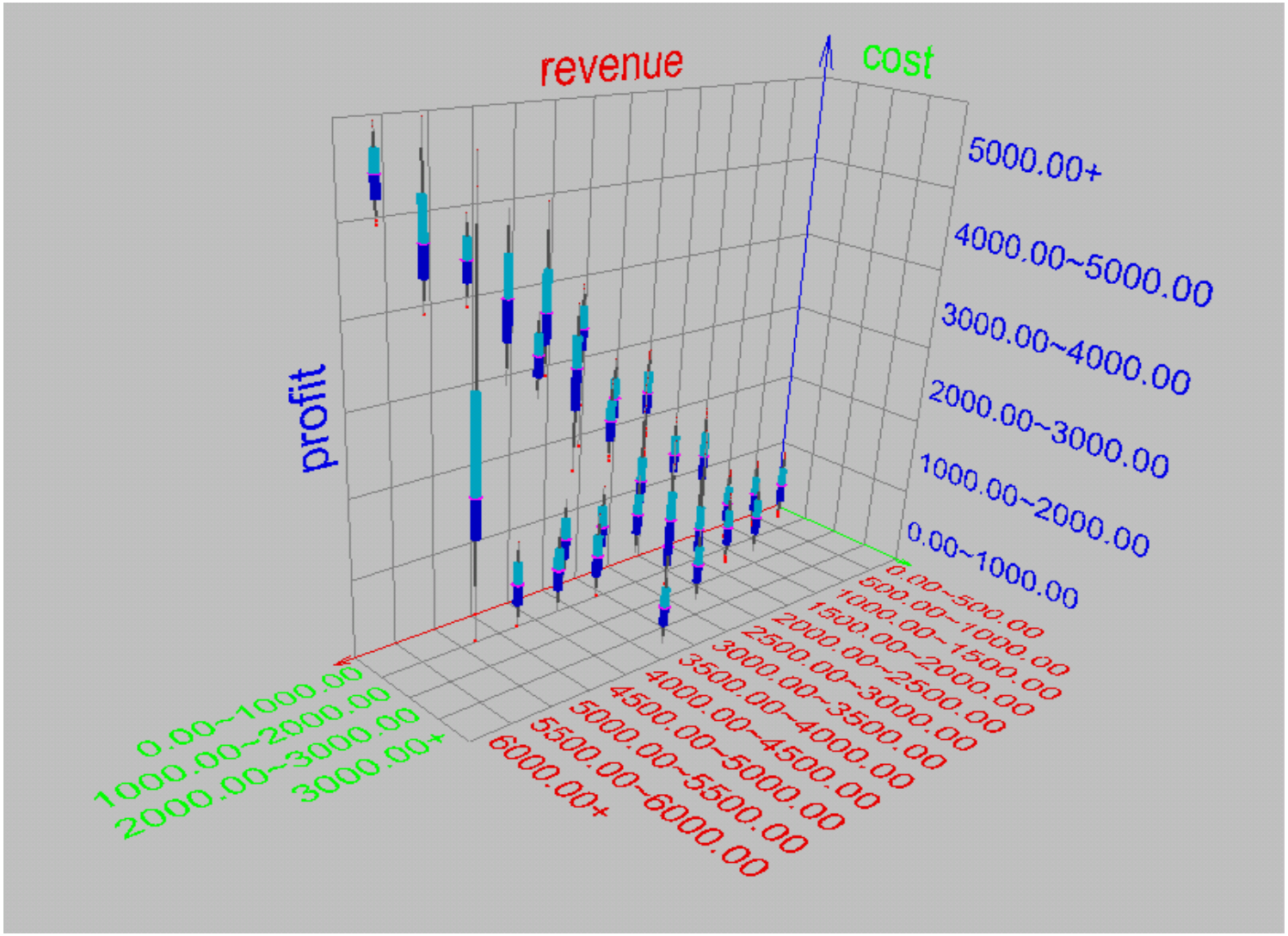




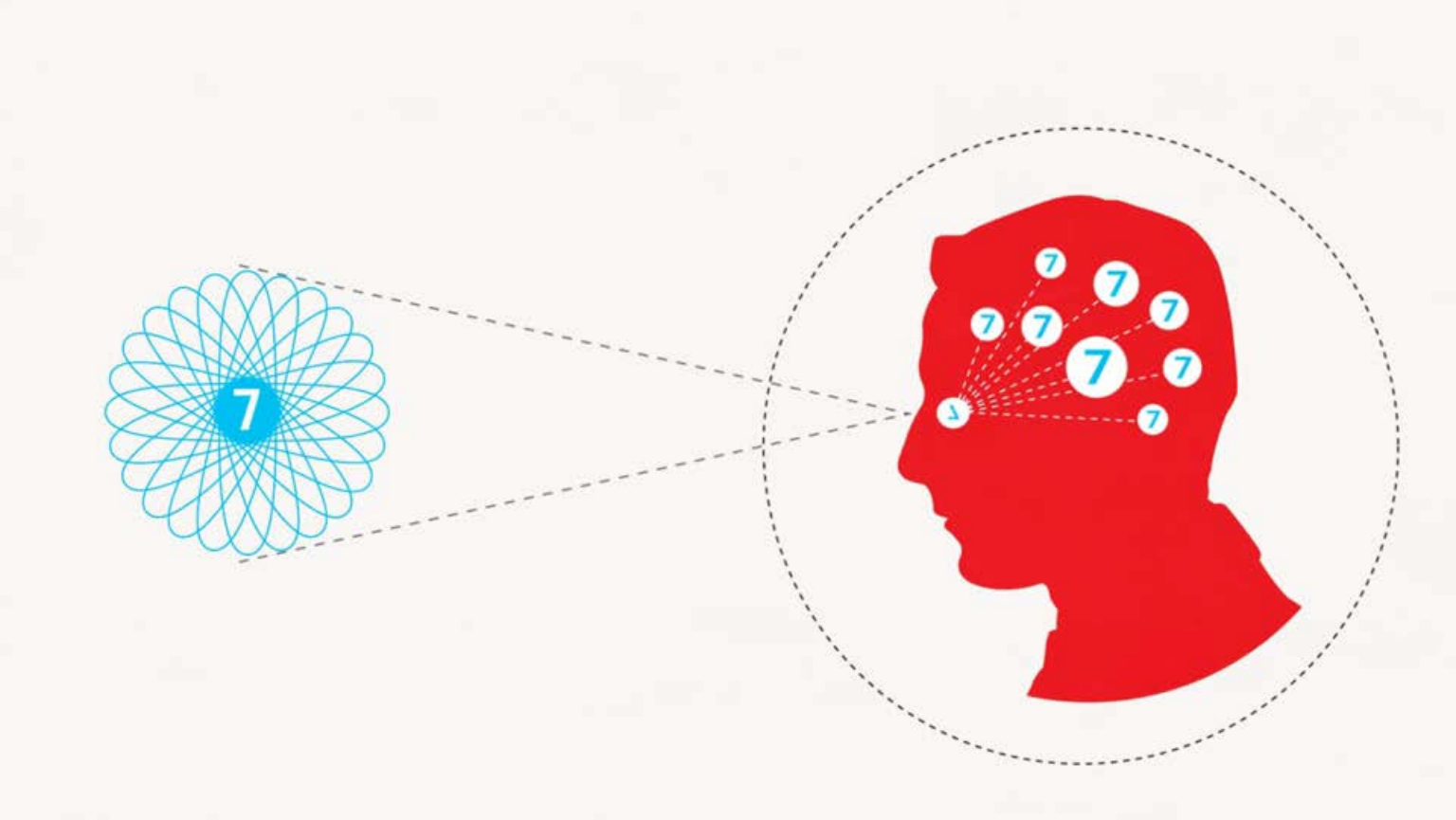
Here:

- Lines in the boxes show national *average* value (instead of *median*)

Visualization of Data Dispersion: 3-D Boxplots



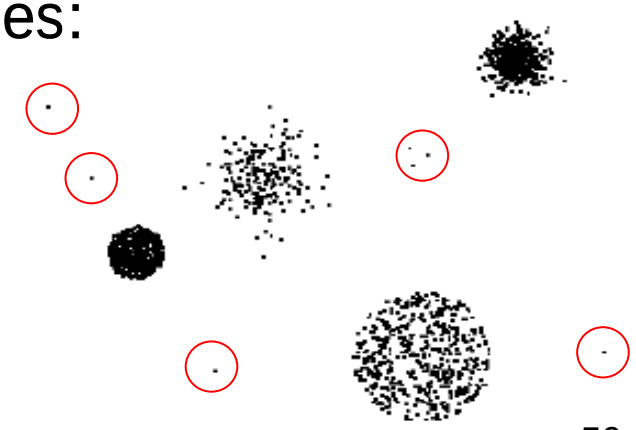
Data Visualization



Outlier Detection Schemes

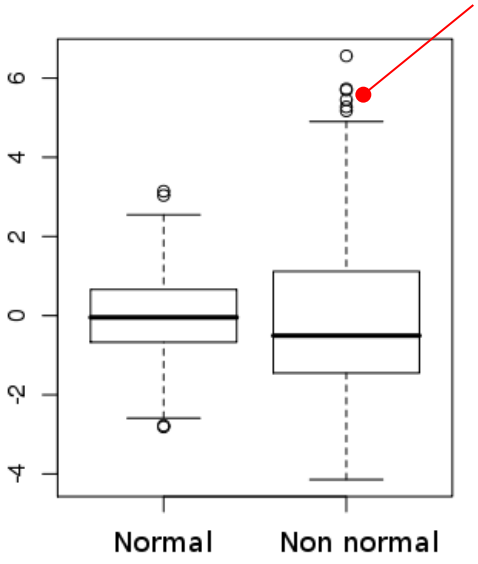
- General Steps:
  - Build a ***profile of the "normal" behavior***.
    - Profile can be patterns or summary statistics for the overall population.
  - ***Use the "normal" profile to detect outliers***.
    - Outliers are observations whose characteristics differ significantly from the normal profile.

- Major types of outlier detection schemes:
  - **Graphical**
  - **Statistics-based**
    - (Model-based)
  - **Distance-based**

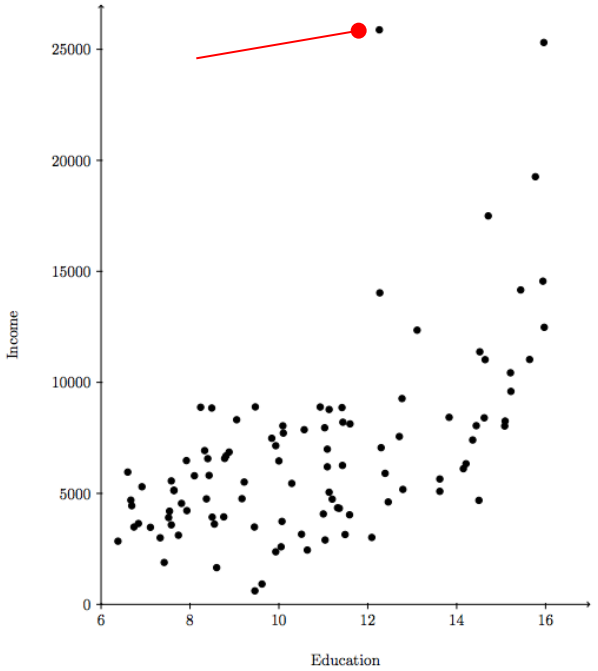


Outliers: Graphical Approaches

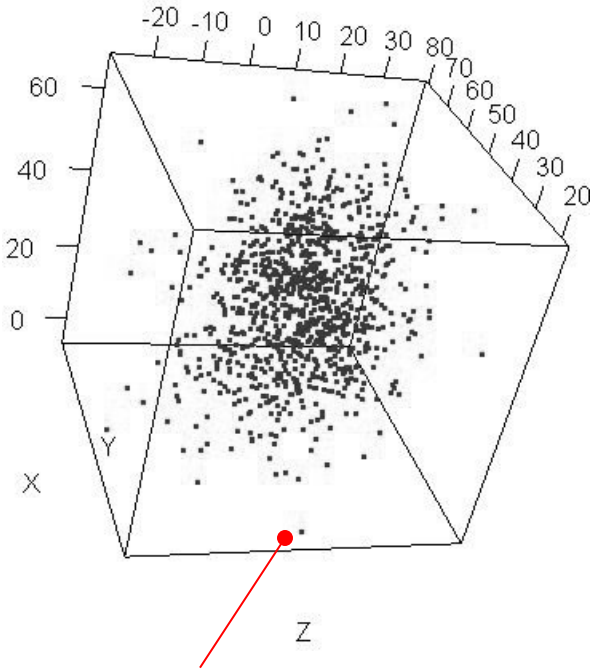
Boxplot (1-D)



Scatter plot (2-D)



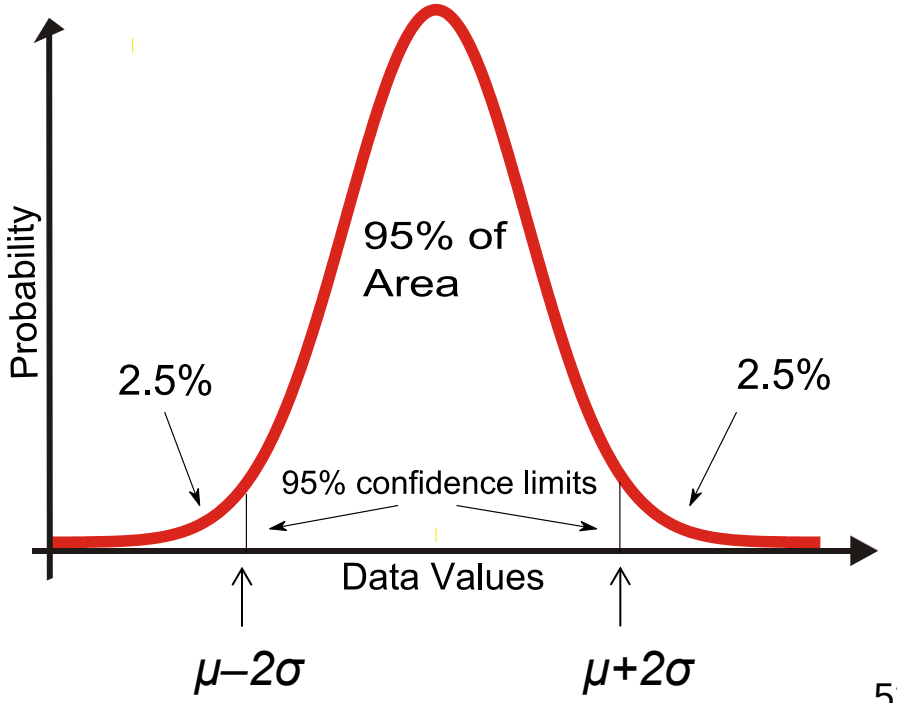
Spin plot (3-D)



- Limitations:
  - Time consuming
  - High-dimensional data
  - Subjective

Outliers: Statistical Approaches (1)

- Assume a ***parametric model*** describing the distribution of the data
  - Example: normal distribution → parameters are μ, σ
- 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:

Mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = 39.9$

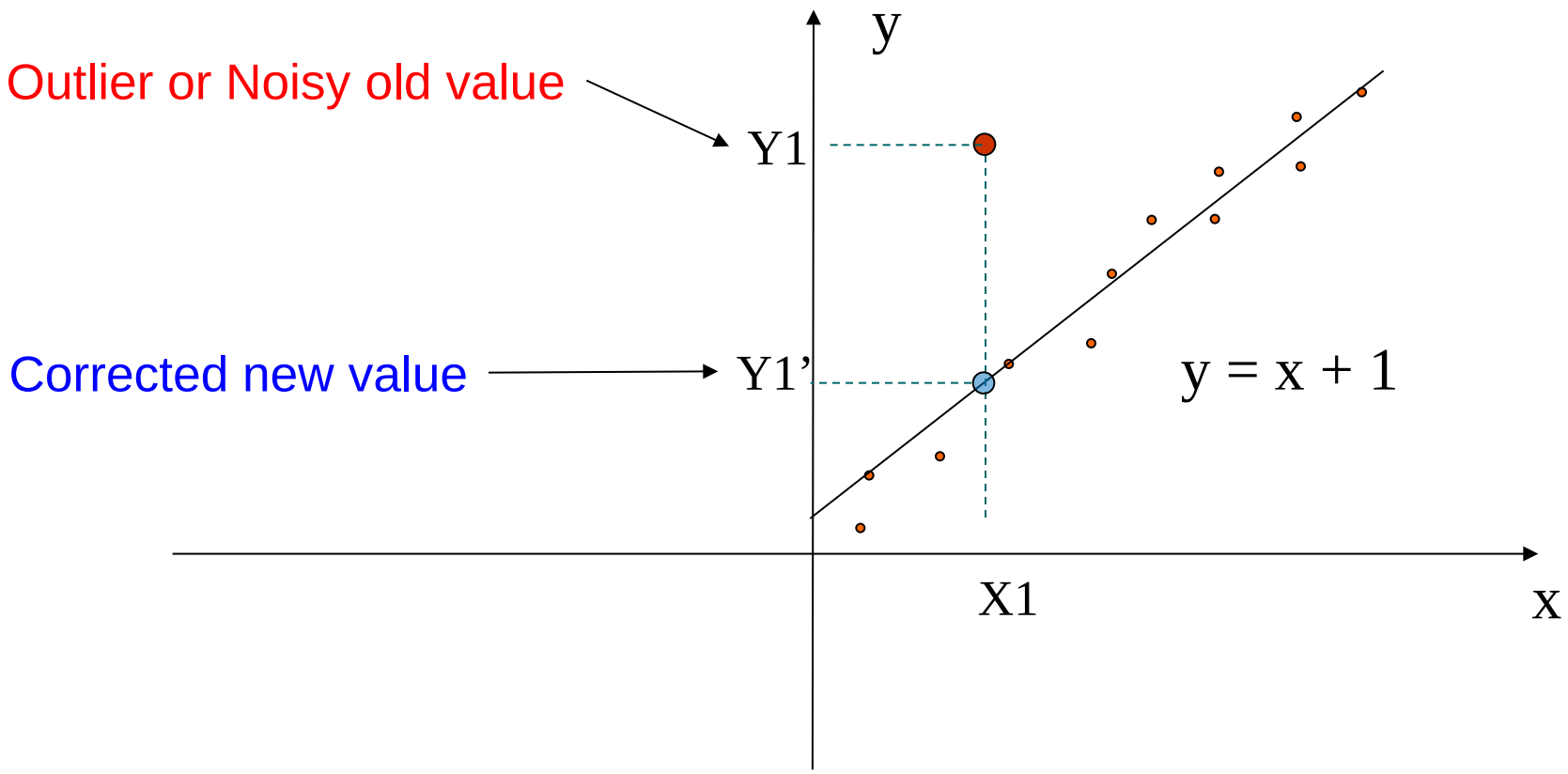
Standard deviation $\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{N-1}} = 45.65$

Select threshold value, e.g. 5% confidence for normal distribution:

Threshold = Mean $\pm 2 \times$ *Standard deviation*

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

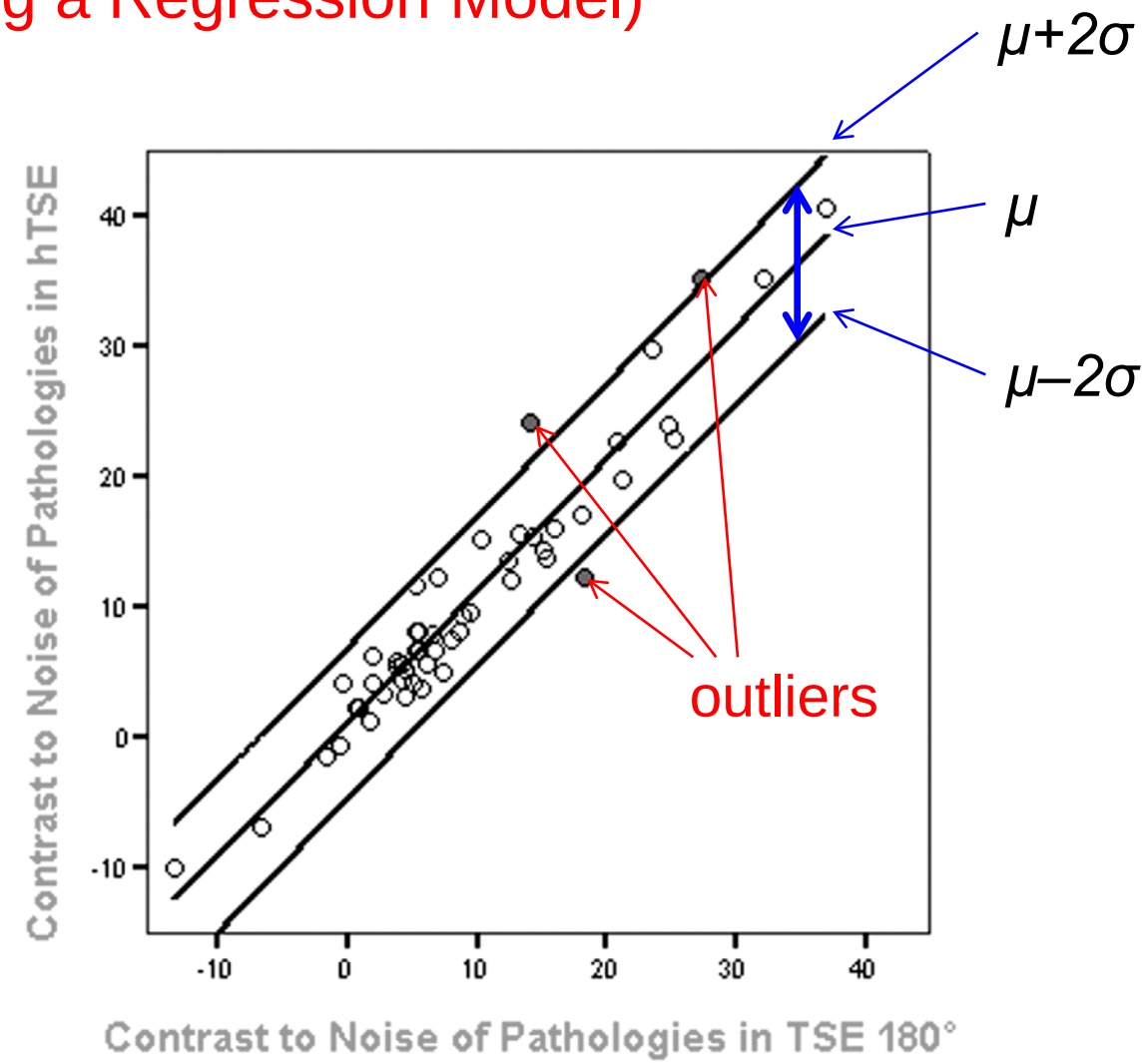
Outliers or Noisy Data?
(Using a Regression Model)



- Model-based approach

Outliers or Noisy Data?
(Using a Regression Model)

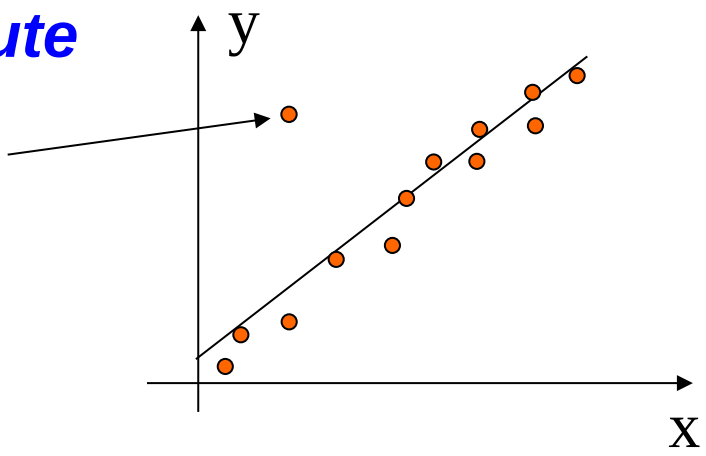
- Assumption:
  y-values Gauss-distributed around fitted curve (model)
- Model describes
  $\mu=\mu(x)$
- Recall:
  From $\mu-2\sigma$ to $\mu+2\sigma$ contains ~ 95%



Limitations of Statistical Approaches

- Tests are often for a ***single attribute***

Not an outlier if y- or x-value considered alone



- Often, assumption of normal distribution is made
  - But 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

- Three major sub-classes of distance-based approaches:
  - ***Nearest neighbor-based***
  - ***Density-based***
  - ***Clustering-based***

Outliers: Nearest Neighbour Approach

- Outlier detection for *n*-dimensional samples:
  - Evaluate the distances between all sample pairs in an *n*-dimensional data set.

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

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

- Estimate parameters p and d:
  - using prior knowledge or
  - by trial-and error

Outliers: Nearest Neighbour Approach Example

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

$$d = [(x1 - x2)^2 + (y1 - y2)^2]^{1/2}$$

	S2	S3	S4	S5	S6	S7
S1	2.236	**3.162**	2.236	2.236	**3.162**	2.828
S2	0	2.236	1.414	**4.472**	2.236	1.000
S3		0	**3.605**	**5.000**	**4.472**	**3.162**
S4			0	**4.242**	1.000	1.000
S5				0	**5.000**	**5.000**
S6					0	1.414

Table of distances

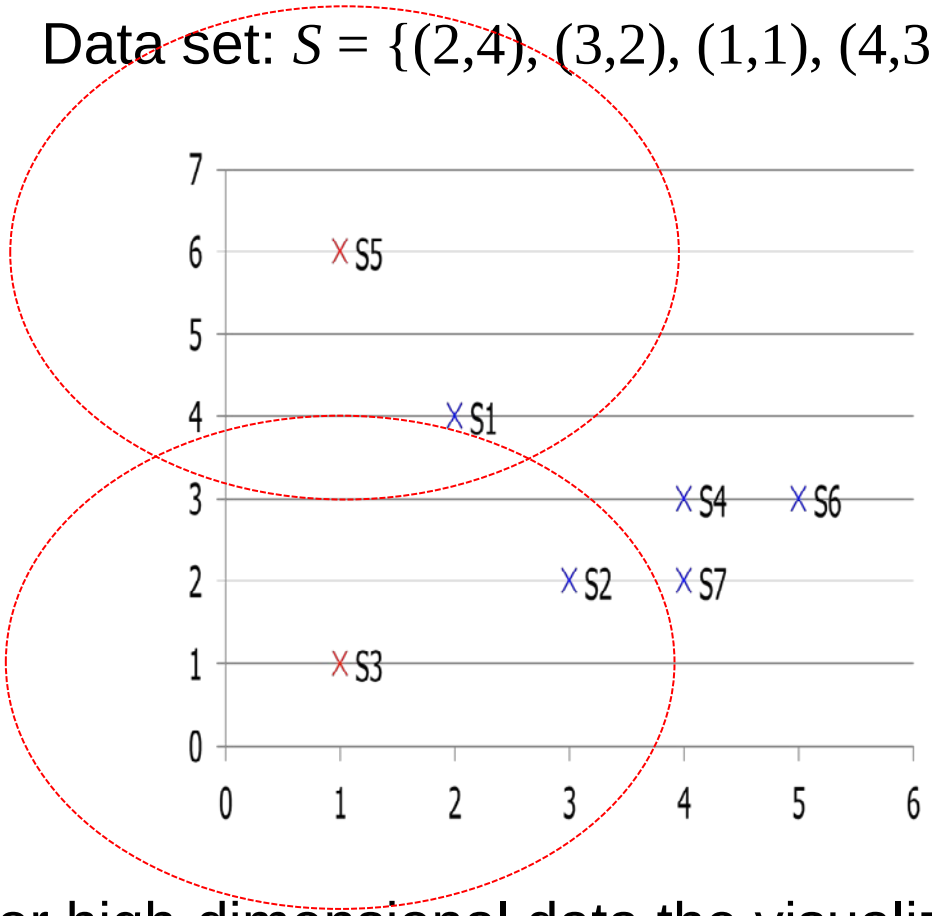
Outliers

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

fraction *p*

Outliers: Nearest Neighbour Approach Example, Visual Inspection

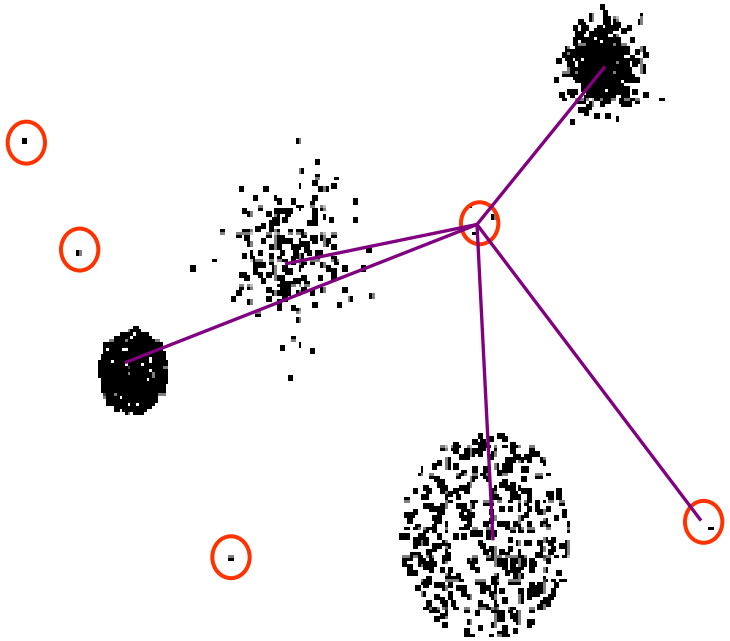
Data set: $S = \{(2,4), (3,2), (1,1), (4,3), (1,6), (5,3), (4,2)\}$



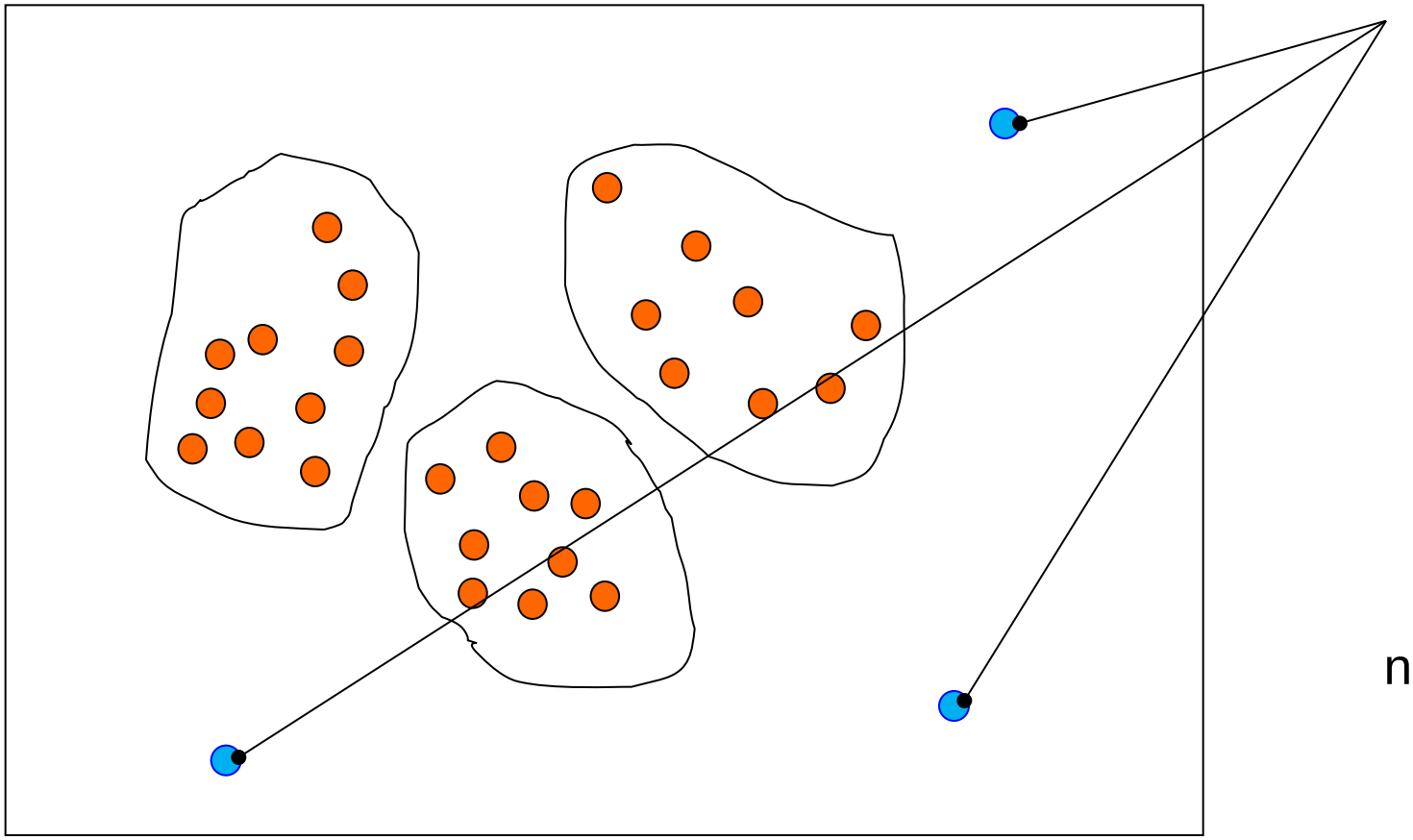
- For high-dimensional data the visualization is more difficult
- For large data sets, the distance matrix becomes too large

Outliers: Distance-based Approach Clustering

- Basic idea for large data sets - ***clustering based***:
  - Cluster the data into a finite number of groups
  - Choose points in small clusters 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)



Outlier Detection

noise?

new knowledge?

Automatic removal of outliers is not recommended

Variants of Anomaly/Outlier Detection

(1) Given a database *D*, find all the data points $\boldsymbol{x} \in D$ with anomaly scores greater than some threshold *t*

(2) Given a database *D*, find all the data points $\boldsymbol{x} \in D$ having the top-*n* largest anomaly scores $f(\boldsymbol{x})$

(3) 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*

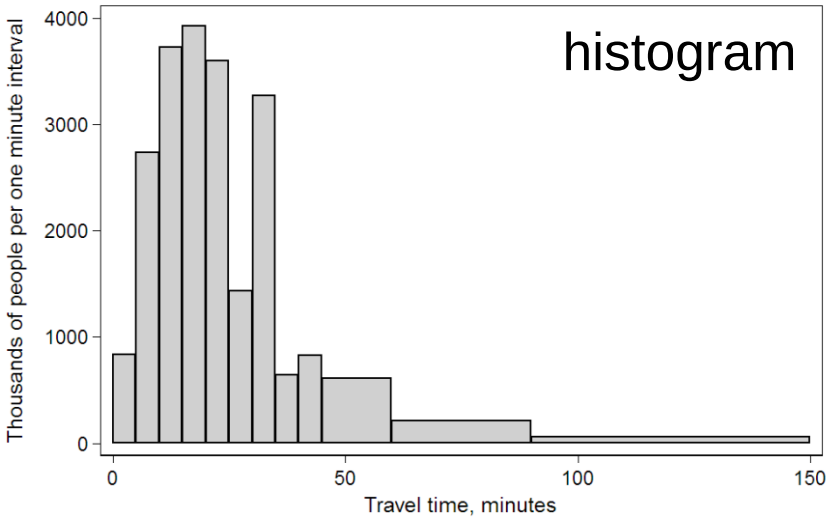
- **Applications**
  - fraud detection (credit card, telecommunication, …)
  - network intrusion detection
  - fault detection & condition monitoring of machines (trains, oil platforms, …)

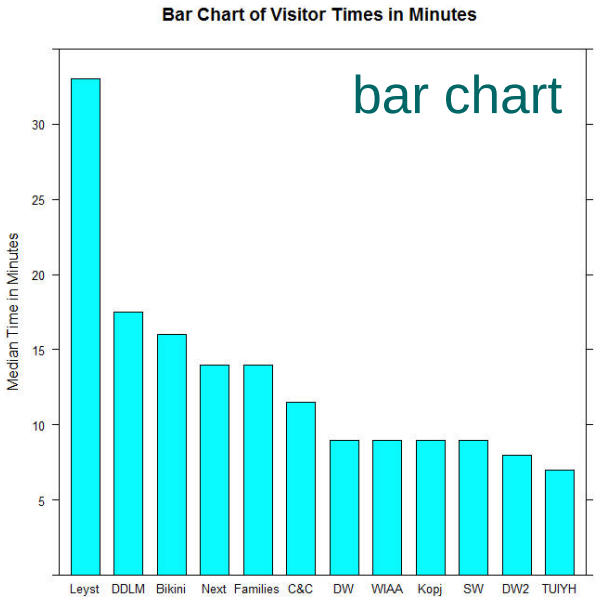
Further Displays of basic statistical Descriptions

- ***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

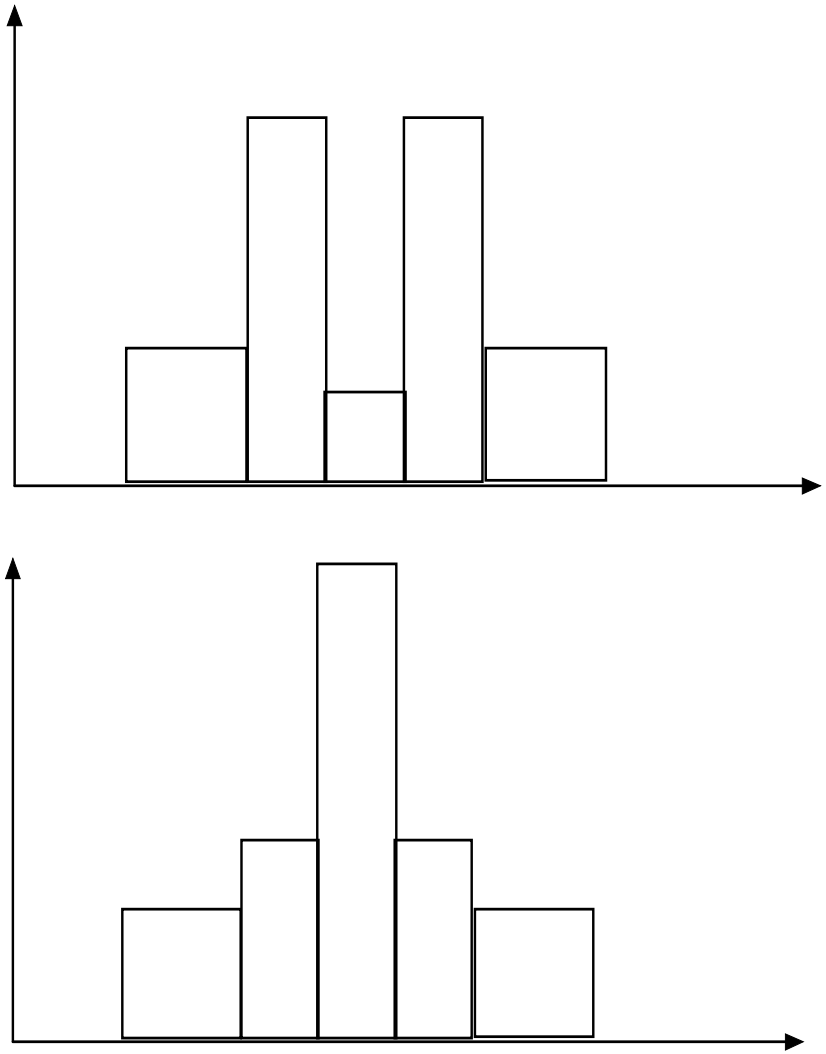
Histogram and Bar Chart Analysis

- Histograms are used to show *distributions* of variables while bar charts are used to *compare* variables.
- Histograms plot quantitative data with ranges grouped into bins while bar charts plot categorical (nominal) data.
- Bars can be reordered in bar charts but not in histograms.
- Bar charts are plotted with gaps between the bars; histograms not.
- Histograms may have bars of different widths (area is important) while bar charts denote their values by the lengths of the bars.





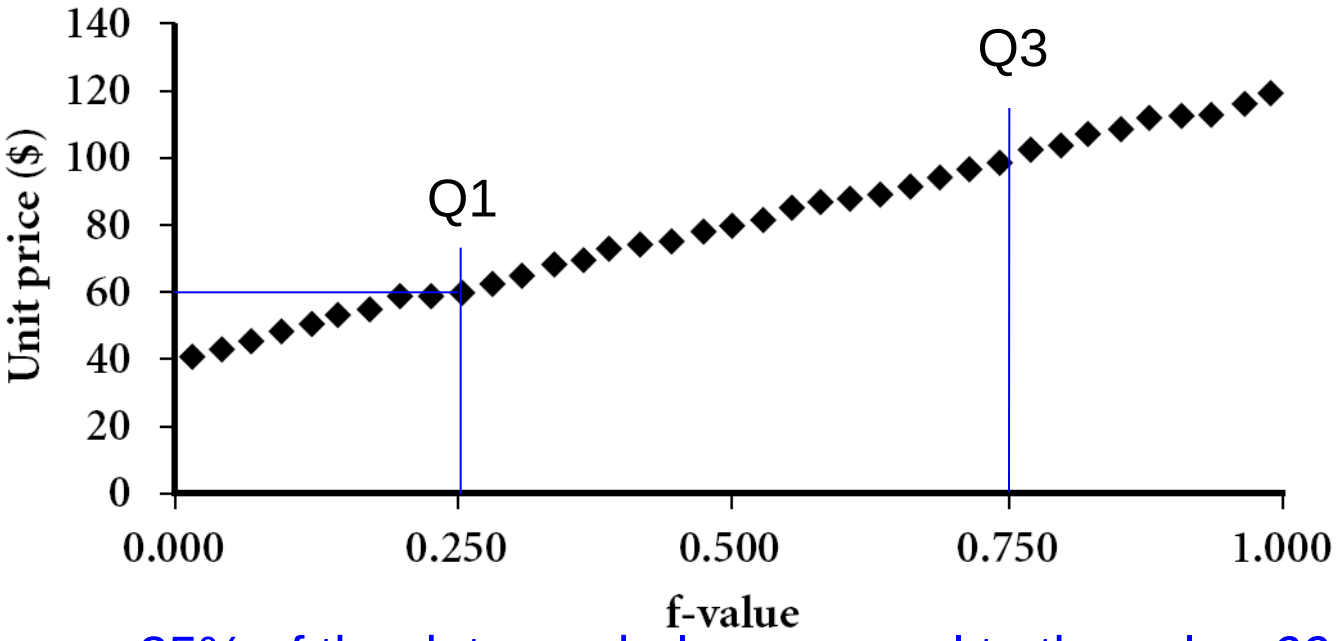
Do Histograms tell more than Boxplots?



- 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

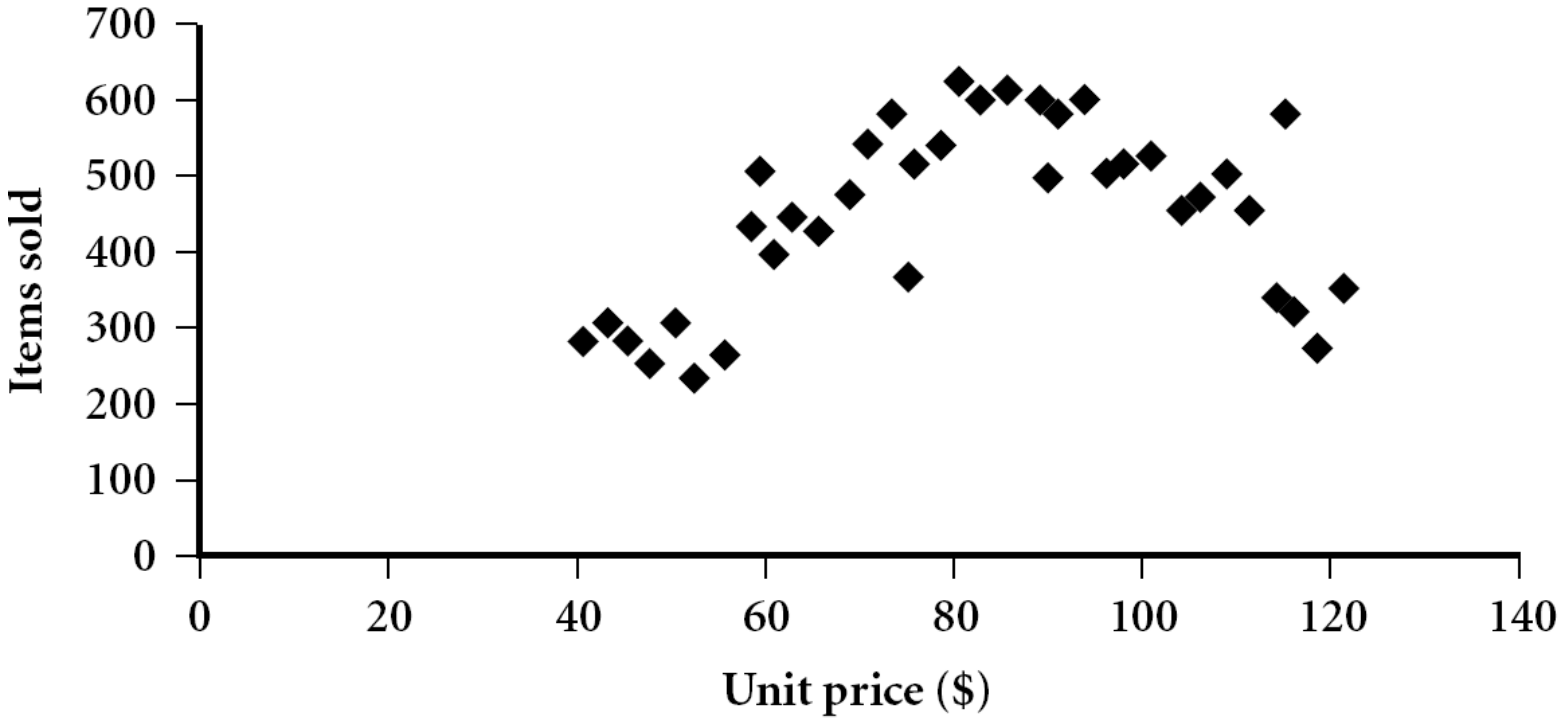
- Displays all of the data
- Plots ***quantile*** information
  - For data x_i sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



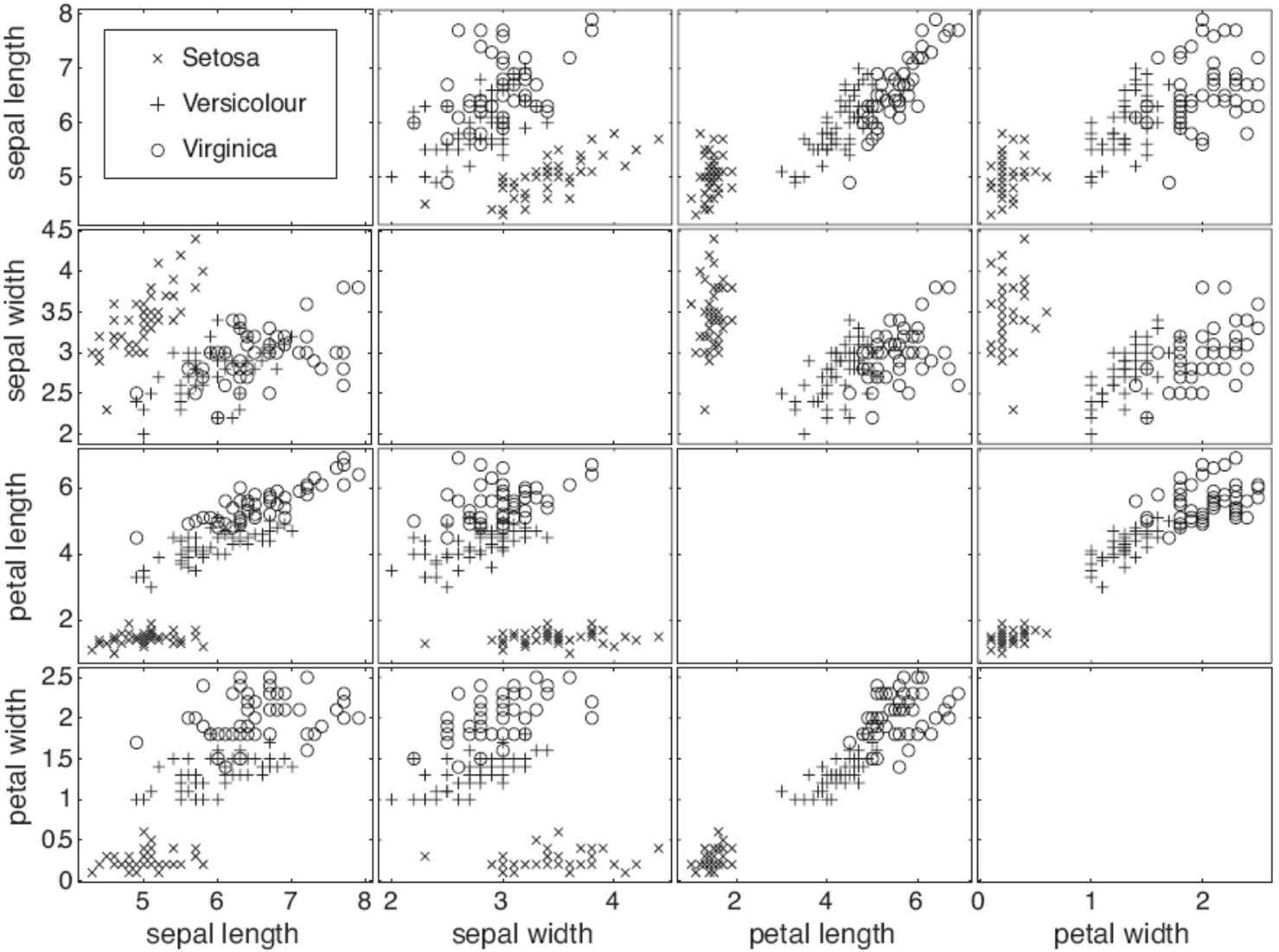
25% of the data are below or equal to the value *60*

Scatter Plot

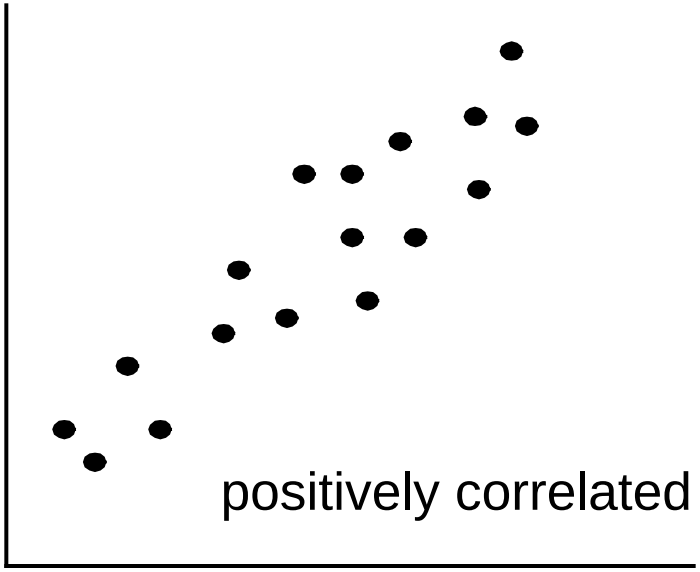
- Provides a ***first look*** at data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

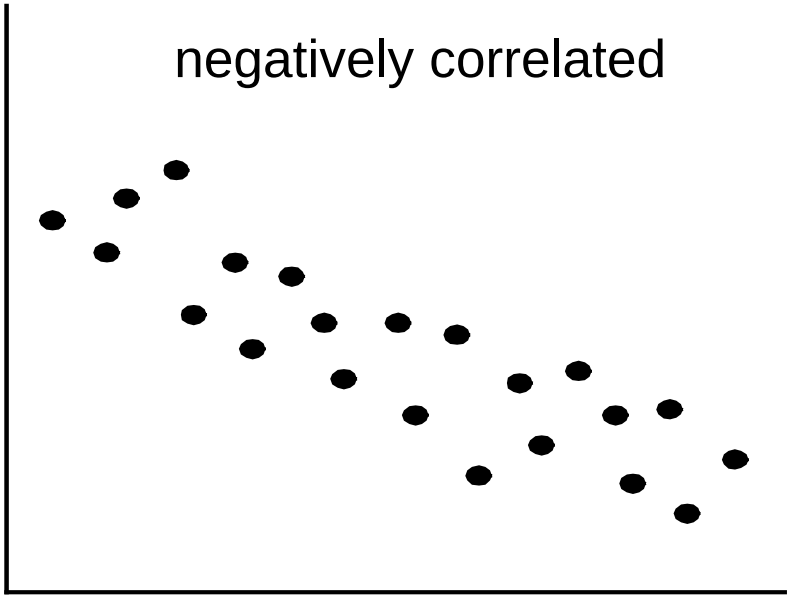


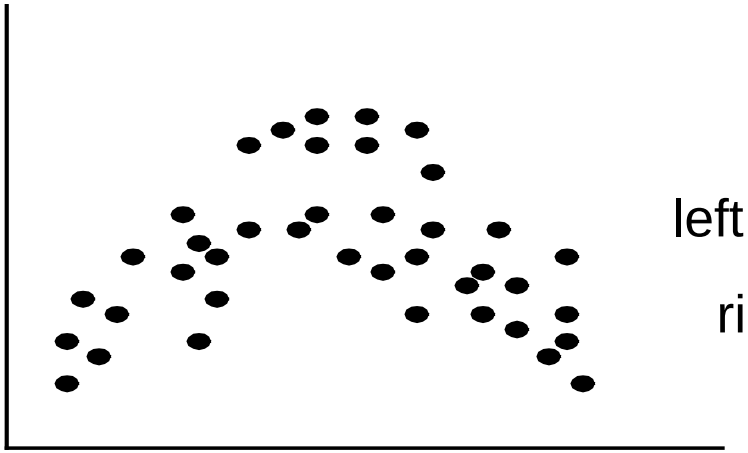
4x4 Matrix of Scatter Plots for 4-D Data



Positively and Negatively Correlated Data



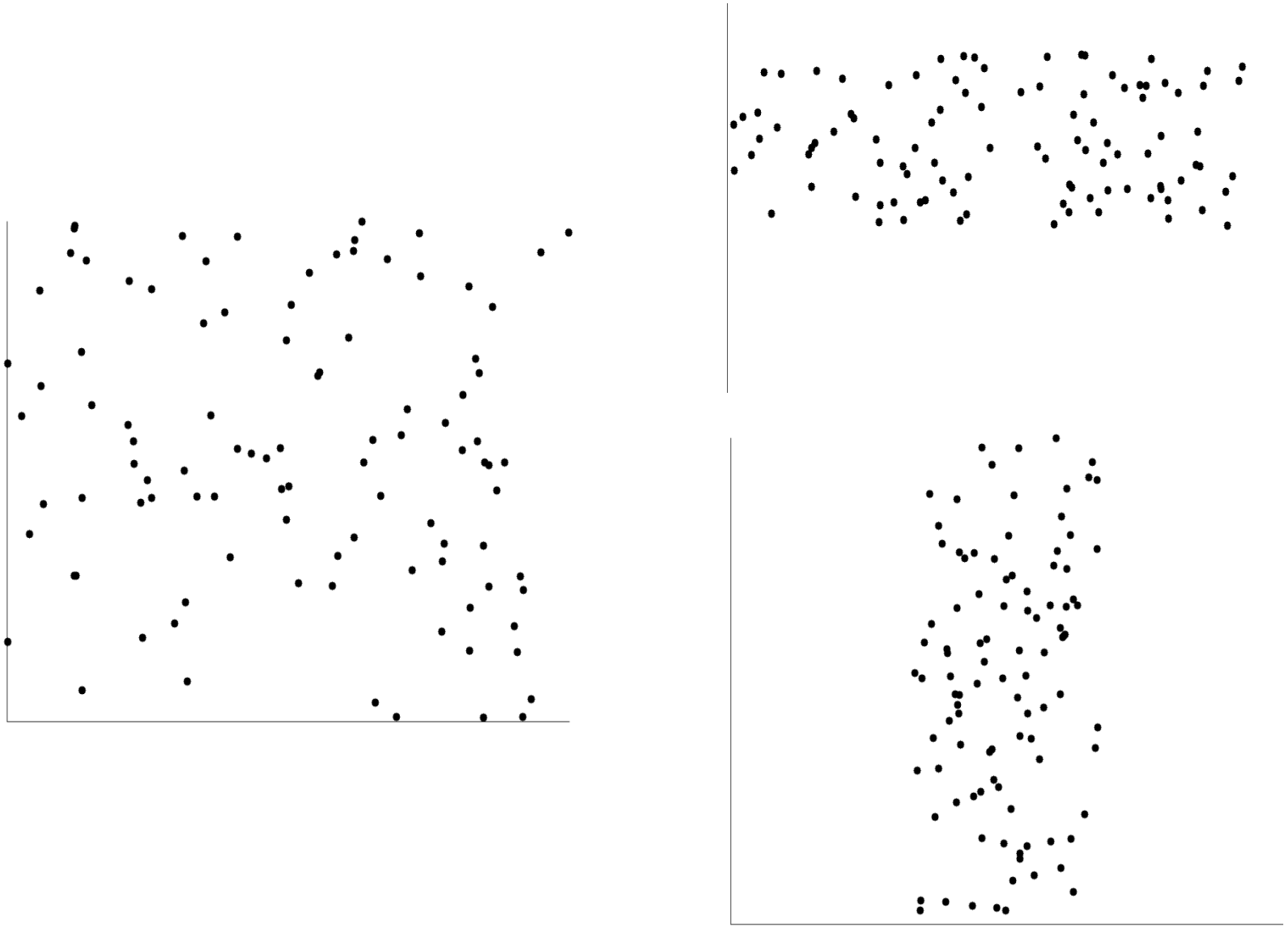




left half fragment: positively correlated

right half fragment: negatively correlated

Uncorrelated Data



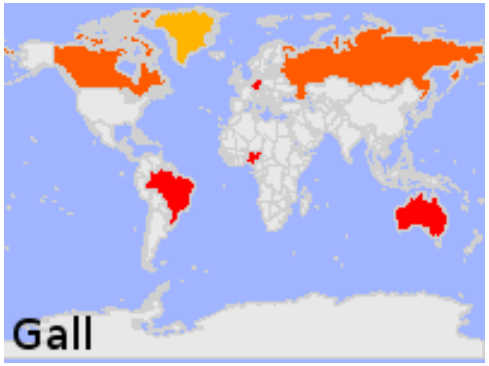
Further Advances in Data Visualization

- 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

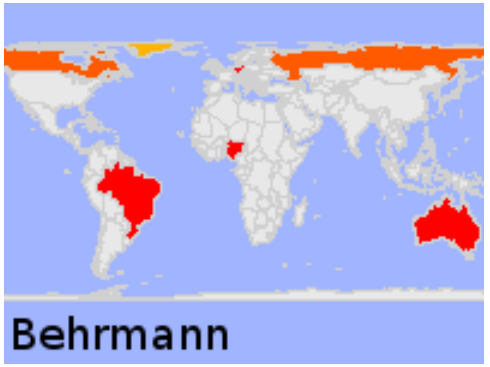
Data Visualization – No One-size-fits-all

- Fitting a sphere onto a plane …

locally good shapes

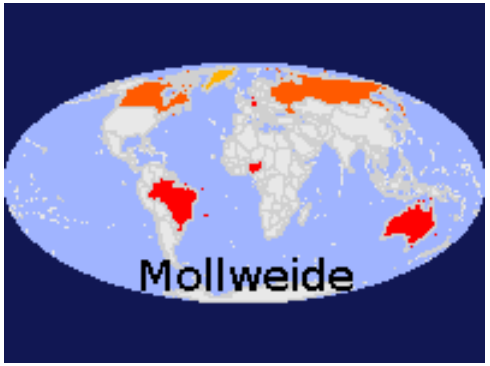


globally good shape



true area sizes

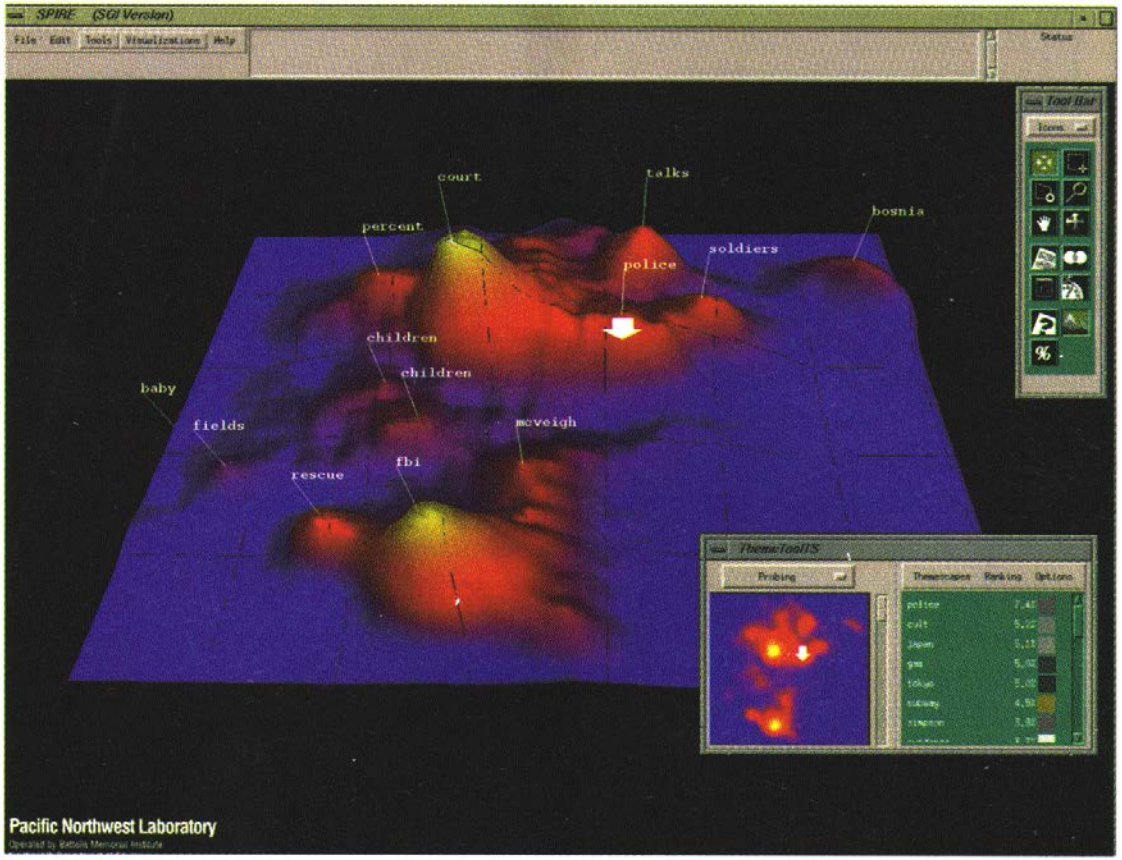
compromise between
angle & area distortions



true area sizes,
straight lines of latitude

true area sizes

Visualization of Data as a Landscape

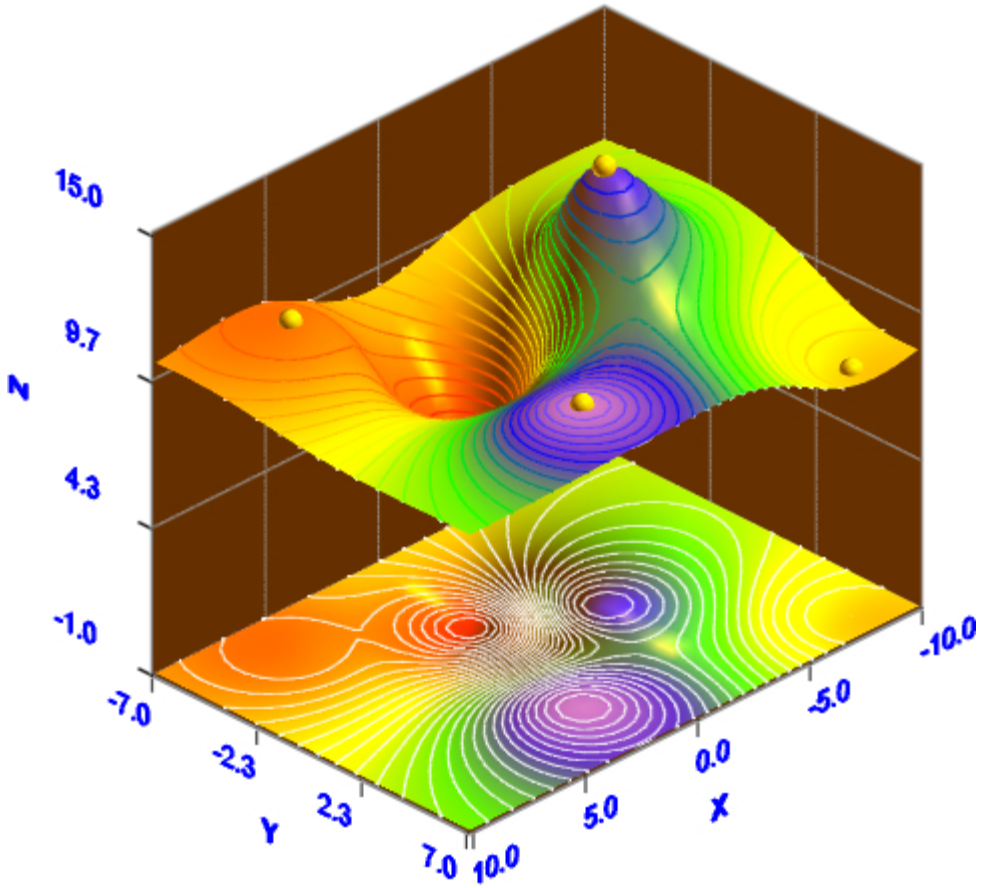


Used by permission of B. Wright, Visible Decisions Inc.

news articles visualised as a landscape

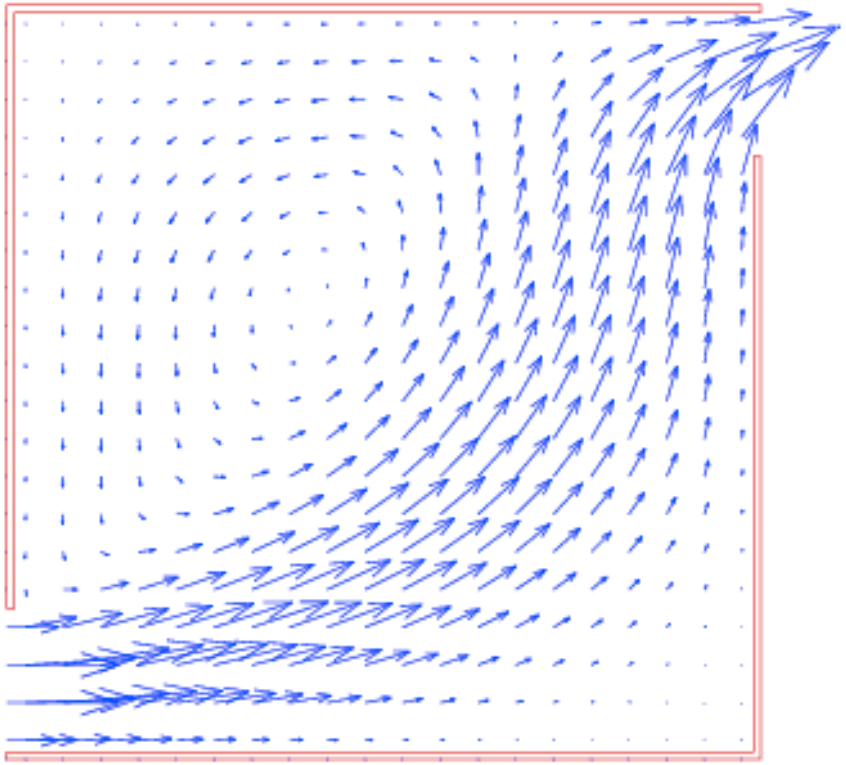
- The data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

Visualization of Data as a Contour Plot



- Added contour lines
- This plots a function of 2D into 1D – but what about 2D into 2D?

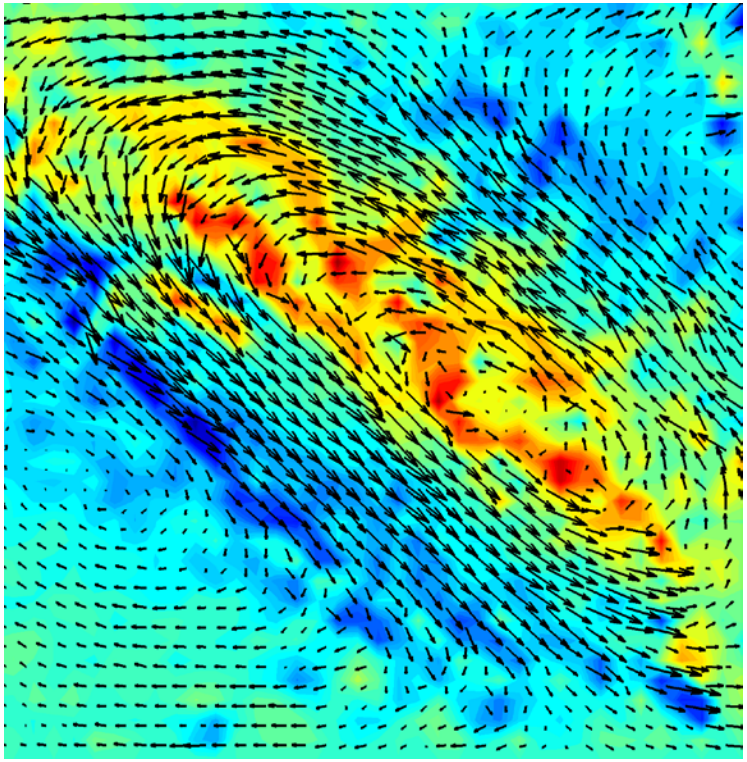
Visualization of Data as a Flow Field



2D → 2D:

(x,y) → (dx,dy) or

(x,y) → (length,direction)

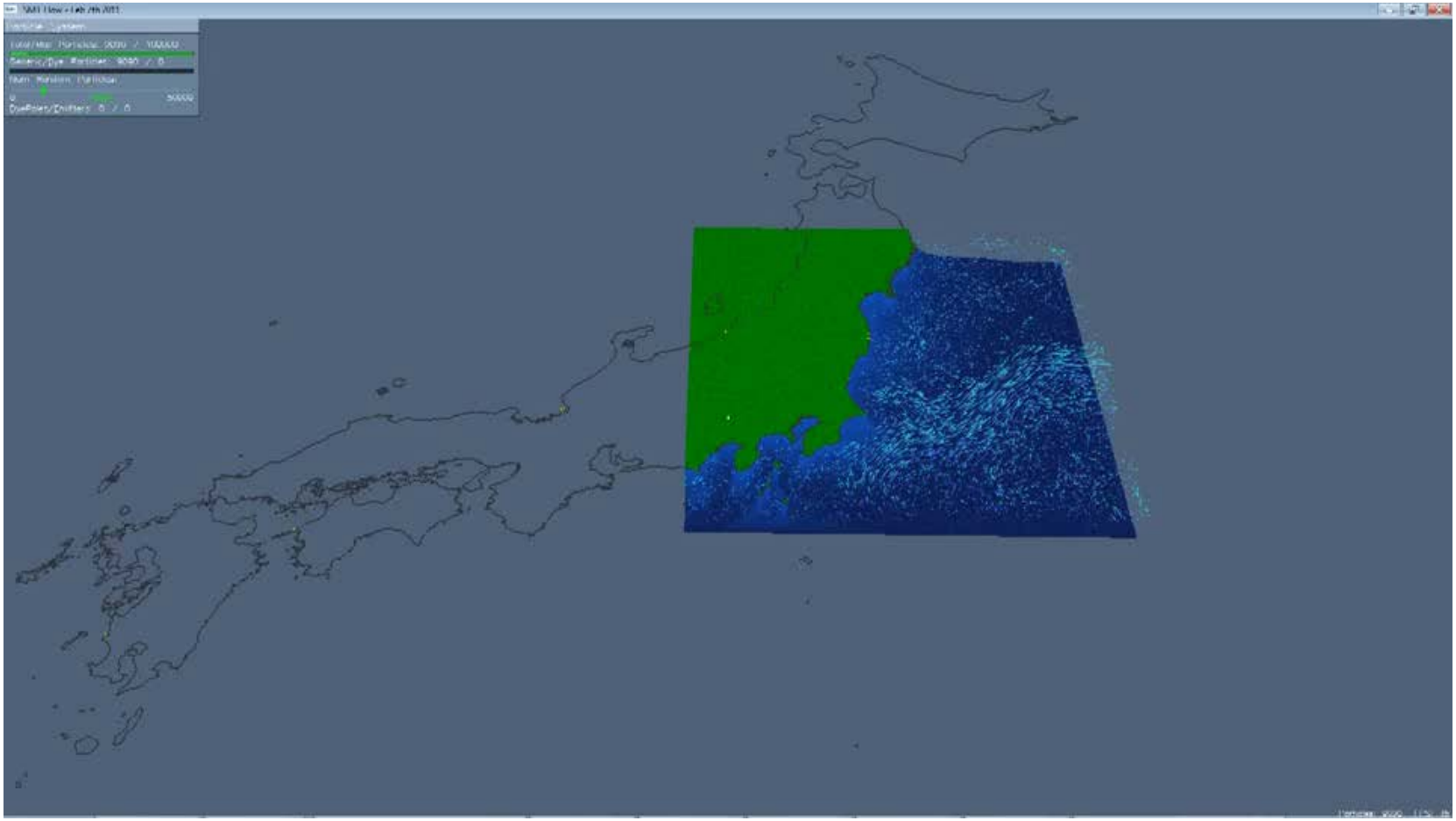


2D → 3D:

(x,y) → (dx,dy,color)

Example: Ocean Flow Analysis Visualization

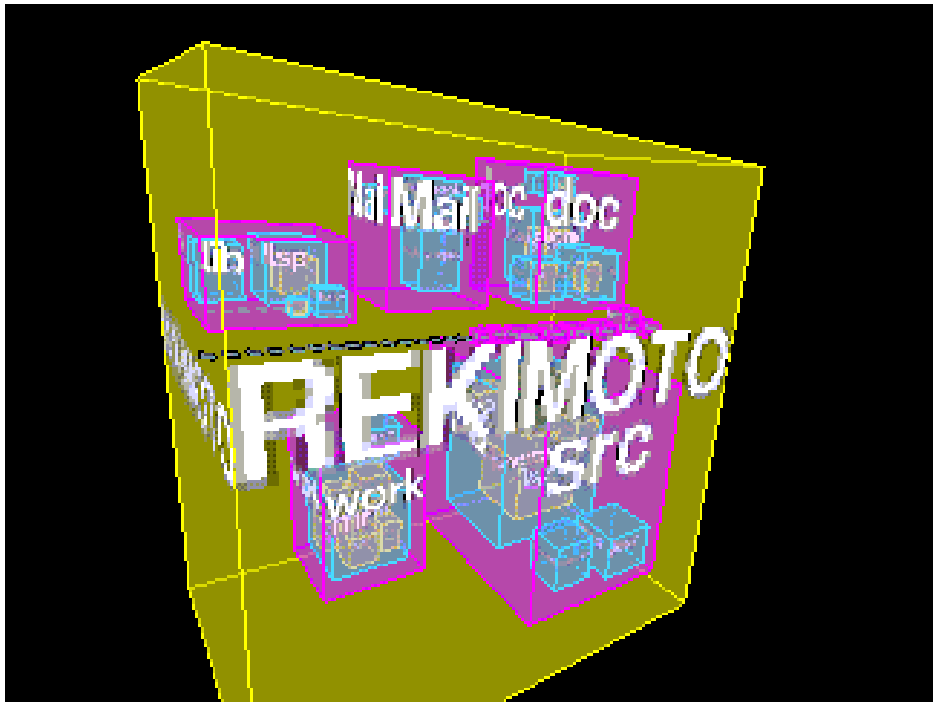
- A visualization of the ocean flow simulation being run on the flow model



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



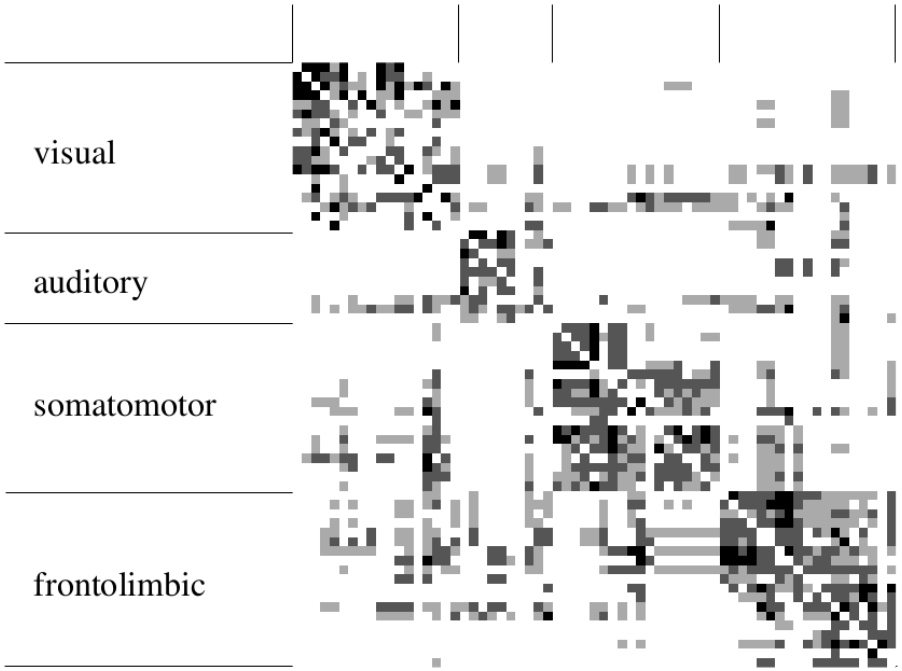
Visualizing Complex Data

- Visualizing non-numerical data: text
- ***Tag cloud***: visualizing user-generated tags
- The importance of tag is represented by font size/color
- Similar data are placed nearby

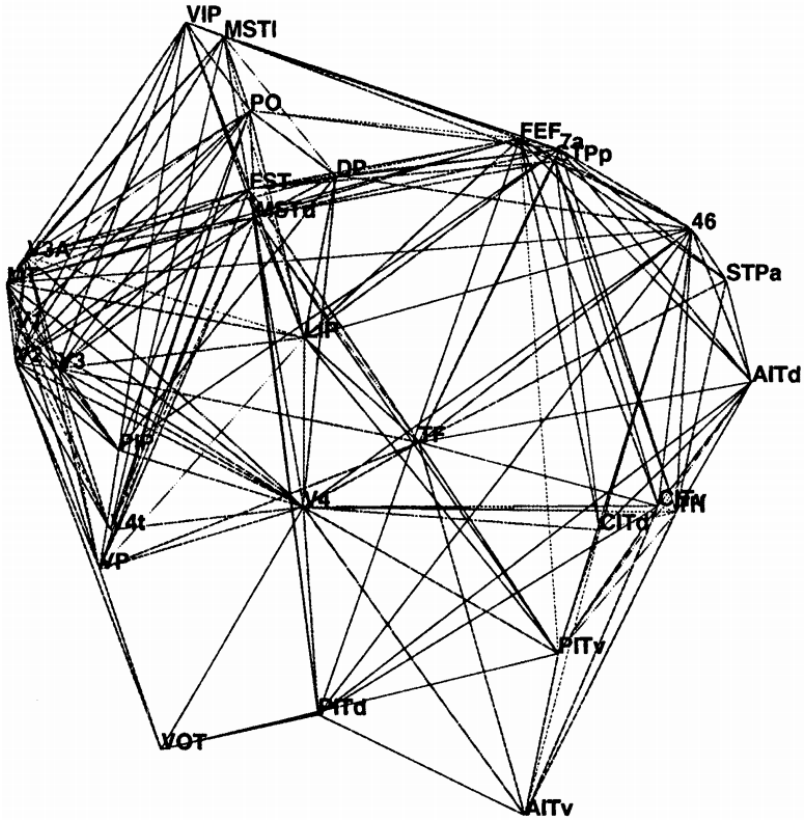
Newsmap: Google News Stories

Visualizing Relations

- Visualizing networks



Connections between 65 cortical areas in cat



Relations within the visual cluster

(non-metric multidimensional scaling)

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]
  - Sometimes referred to as *proximity*
- ***Dissimilarity***
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
  - E.g. *distance*

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

Proximity Measures for Binary Attributes

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

A contingency table for binary data, where

- q = # variables that equal 1 for both objects *i* and *j*,
- r = # variables that equal 1 for object *i* but equal 0 for object *j*,
- s = # variables that equal 0 for object *i* but equal 1 for object *j*,
- t = # variables that equal 0 for both objects *i* and *j*.

Proximity Measures for Binary Attributes

Contingency table

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 be 0
- Use asymmetric distance:

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

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

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

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

Standardizing Numeric Data

- z-score: $z = \frac{x - \mu}{\sigma}$
  - *x*: value ("raw score") to be standardized
  - μ: mean of the population
  - σ: standard deviation
  - the distance between the value and the mean in units of the std.dev.
  - negative if the raw score is below the mean, positive if above
- Alternative: Use the mean absolute deviation instead of std.dev.

$$z = \frac{x - m}{s} \qquad \text{where:} \quad m = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) \approx \mu$$

$$s = \frac{1}{n}(|x_1 - m| + |x_2 - m| + \ldots + |x_n - m|)$$

  - standardized measure (*z-score*)
  - more robust against outliers than using standard deviation

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

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}|$$

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

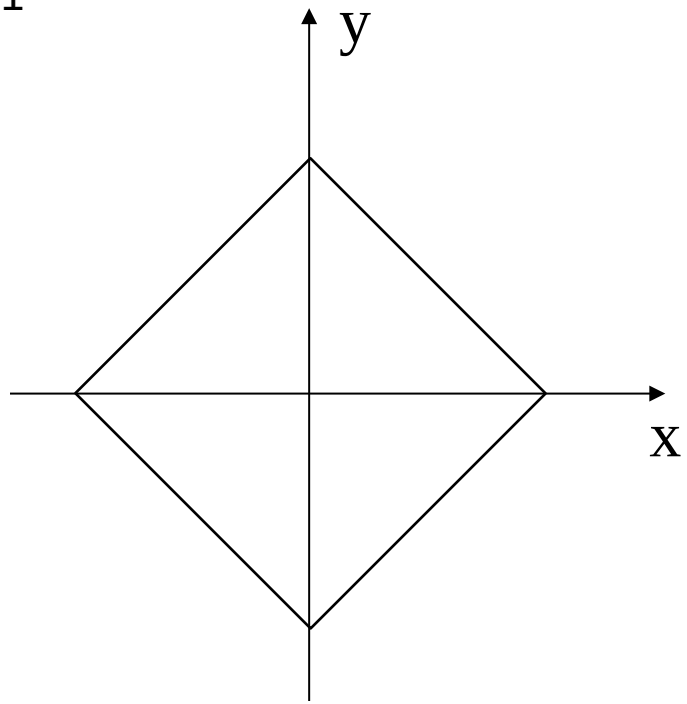
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2)}$$

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

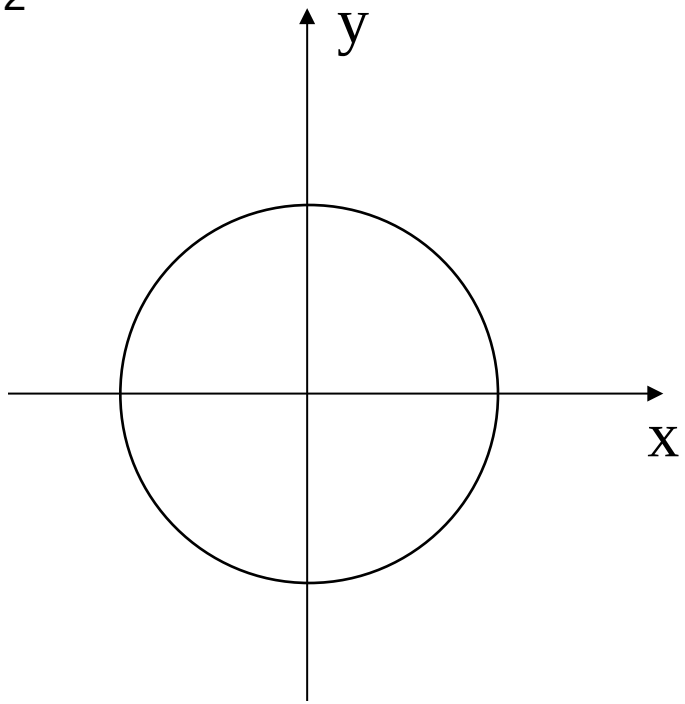
$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

Contour Lines

L_1 norm

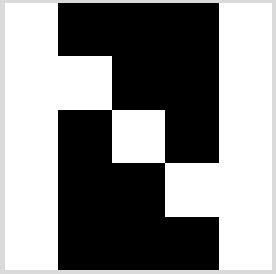
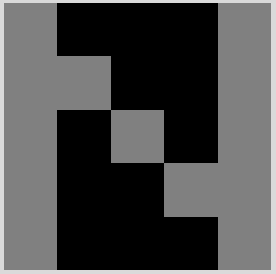
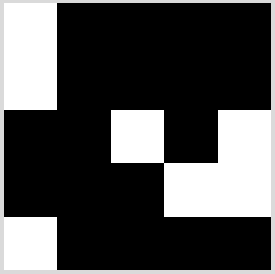


L_2 norm



Distances Between Image (Vector)s

v1 =	***v2*** =	***v3*** =
2 0 0 0 2	1 0 0 0 1	2 0 0 0 0
2 2 0 0 2	1 1 0 0 1	2 0 0 0 0
2 0 2 0 2	1 0 1 0 1	0 0 2 0 2
2 0 0 2 2	1 0 0 1 1	0 0 0 2 2
2 0 0 0 2	1 0 0 0 1	2 0 0 0 0

L_1 (***v1-v2***) = **13** > L_1 (***v1-v3***) = **12**

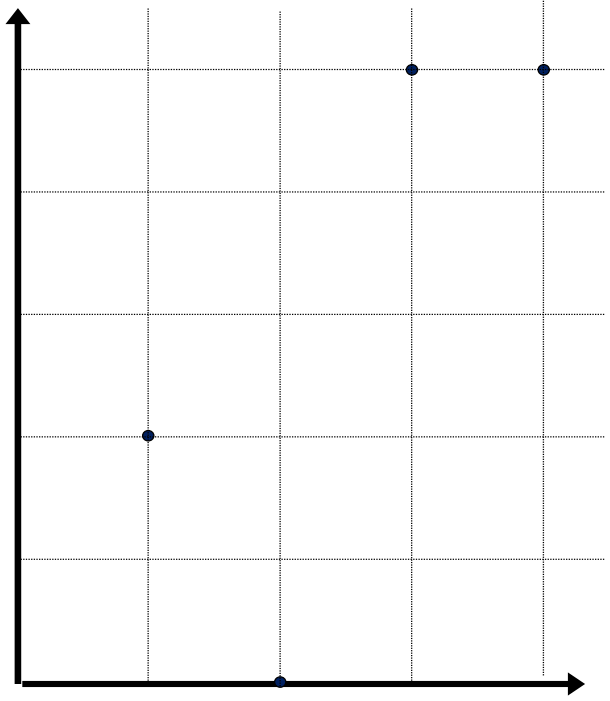
L_2 (***v1-v2***) ≈ **3.6** < L_2 (***v1-v3***) ≈ **4.9**

- L_1 and L_2 norms can lead to different similarity relations
- L_2 large when individual differences large due to square

Example: Minkowski Distance

Data Matrix

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_∞)

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

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

  - $\bullet$ indicates vector dot product
  - $||d||$ is the length (norm) of vector d

Example: Cosine Similarity

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

- **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

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

$= (42)^{0.5} = 6.481$

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

$= (17)^{0.5} = 4.12$

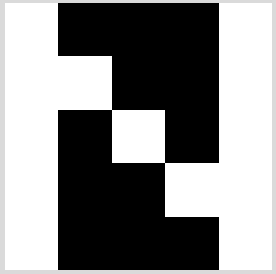
$\cos(d_1, d_2) = 0.94$

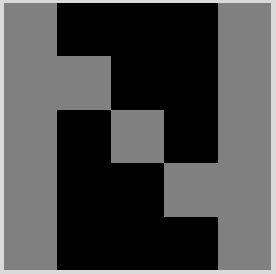
Cosine Similarity Between Image (Vector)s

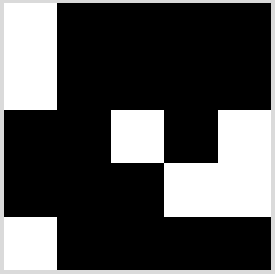
v1 **=**
2 0 0 0 2
2 2 0 0 2
2 0 2 0 2
2 0 0 2 2
2 0 0 0 2

v2 **=**
1 0 0 0 1
1 1 0 0 1
1 0 1 0 1
1 0 0 1 1
1 0 0 0 1

v3 **=**
2 0 0 0 0
2 0 0 0 0
0 0 2 0 2
0 0 0 2 2
2 0 0 0 0



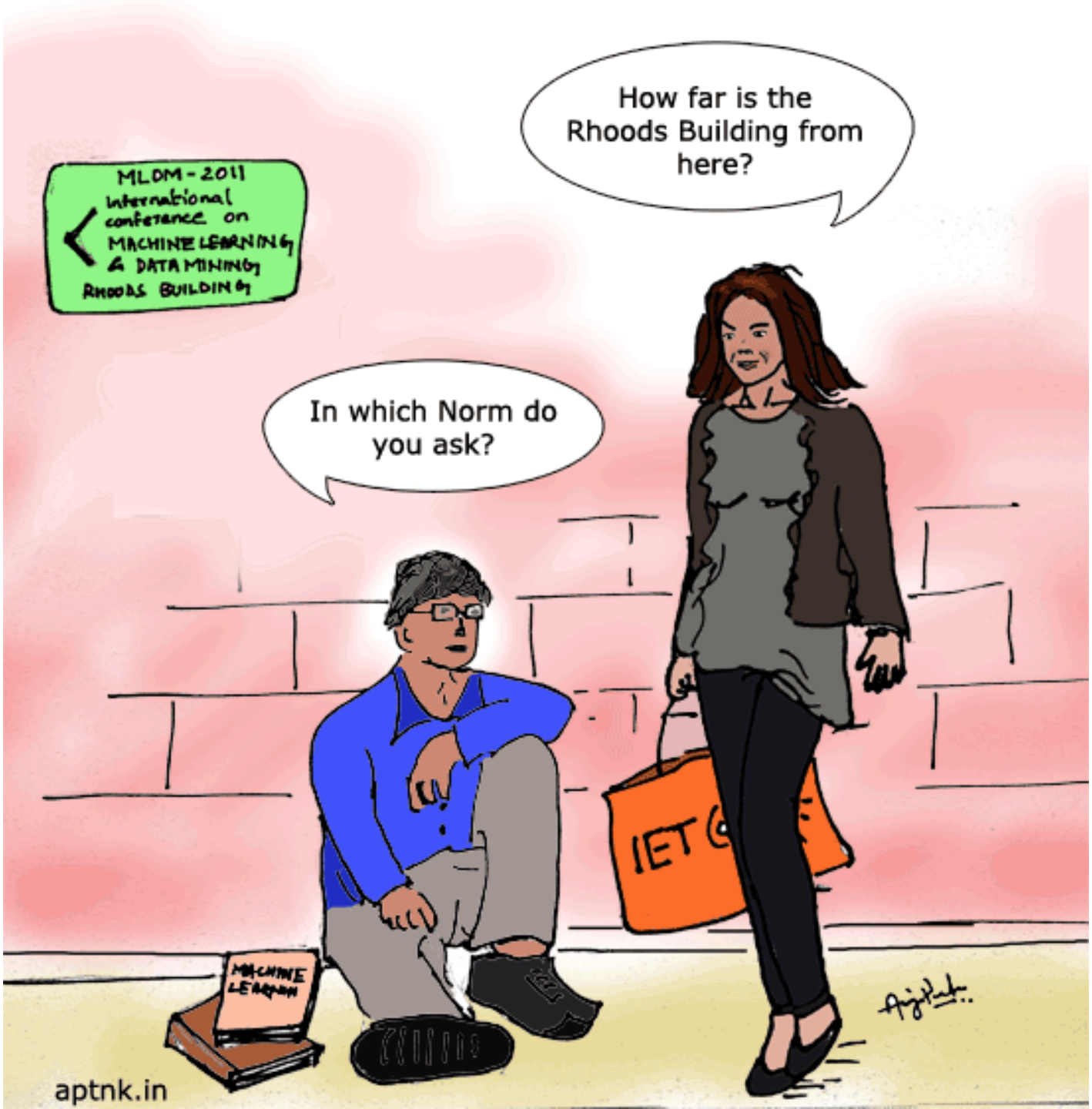




cos(***v1***,***v2***) = **1**

cos(***v1***,***v3***) ≈ **0.73**

- Cosine similarity considers vector orientations but not vector lengths


How far is the Rhoods Building from here?
MLDM - 2011 International conference on MACHINE LEARNING & DATA MINING Rhoods Building
In which Norm do you ask?
aptnk.in

Summary

- Data attribute types: nominal, ordinal, interval-, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, web, image
- If high-dimensional, we need more data for density estimation
- Gain insight into the data by:
  - Basic ***statistical*** data ***description***: central tendency, dispersion, graphical displays
  - ***Outlier*** detection (graphical, statistics-based, distance-based, …)
  - Data ***visualization***: map data onto graphical primitives
  - ***Measure*** data similarity
- Above steps are the beginning of knowledge discovery
- Many methods have been developed but currently a very active area of research due to novel dimensions of data collections

Knowledge Technology Lab research: Cognitive Data Mining…