Data-driven Intelligent Systems

Lecture 3

Visual Interpretation of Data

KNOWLEDGE
TECHNOLOGY

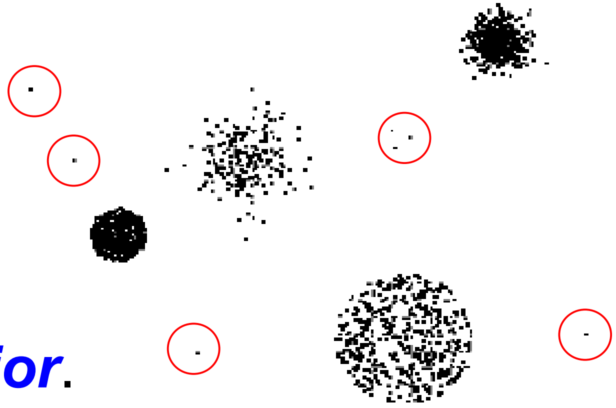
http://www.informatik.uni-hamburg.de/WTM/

Data Visualization

Overview

- Outliers
- Visualisation
- Similarities and Distance Measures

Outlier Detection Schemes

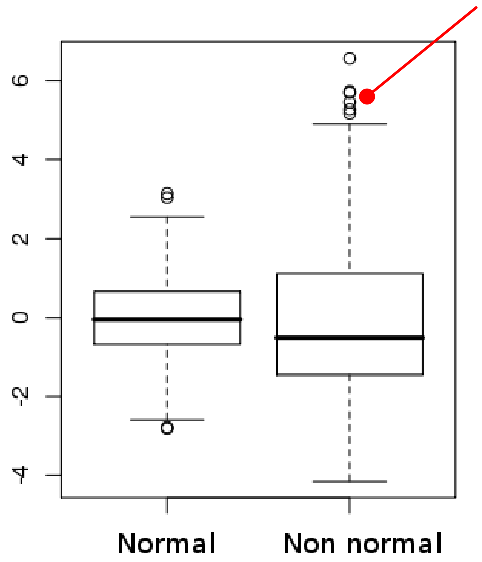


- General Steps:
  - Build a ***profile of the "normal" behavior***.
    - Profile can be patterns or summary statistics for the overall population.
  - ***Use the "normal" profile to detect outliers***.
    - Outliers are observations whose characteristics differ significantly from the normal profile.

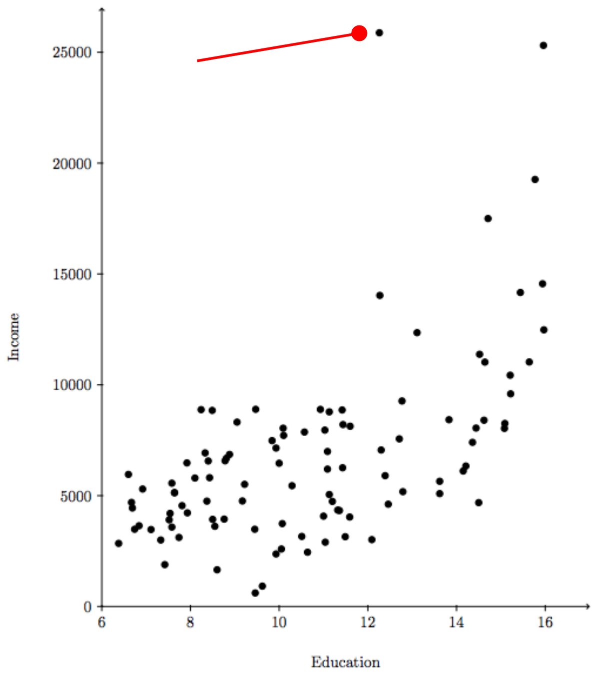
- Major types of outlier detection schemes:
  - **Graphical**
  - **Statistics-based**
    - Model-based
  - **Distance-based**

Outliers: Graphical Approaches

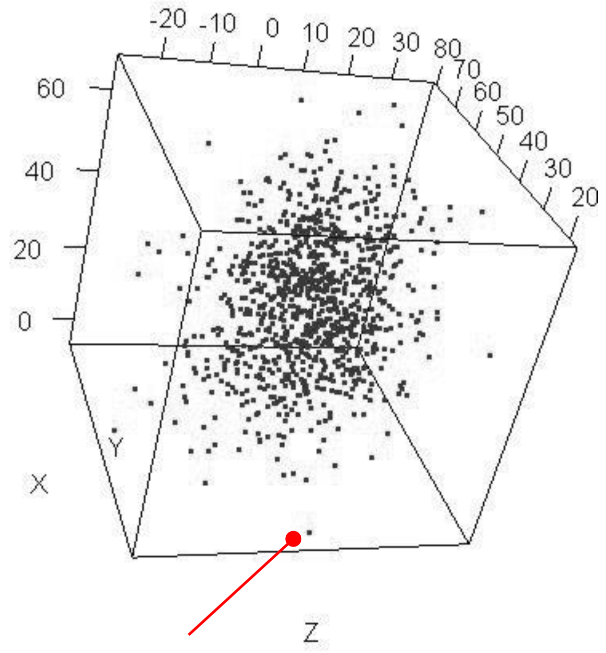
Boxplot (1-D)



Scatter plot (2-D)



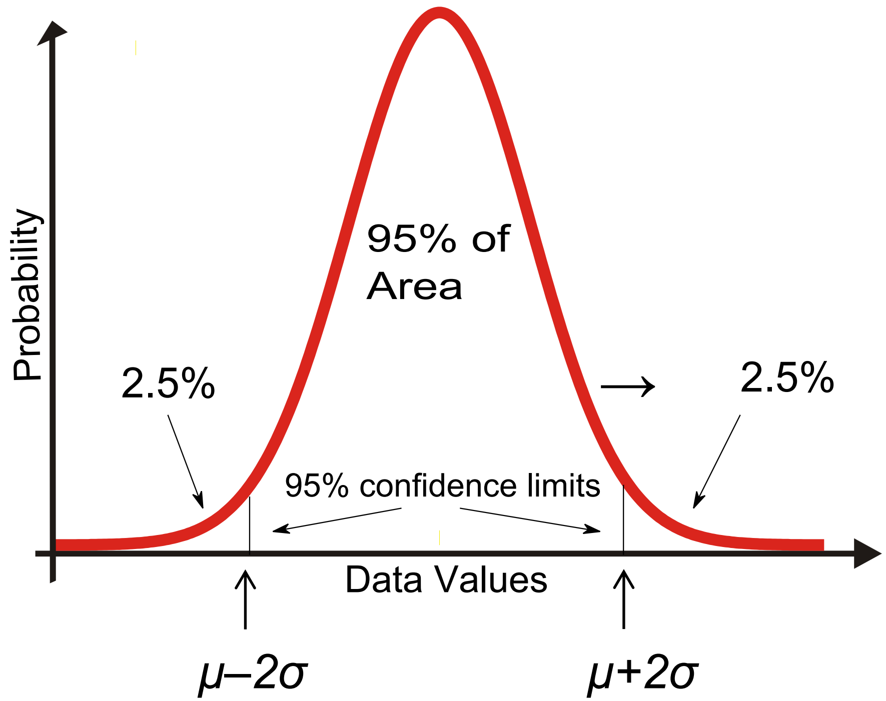
Spin plot (3-D)



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

Outliers: Statistical Approaches (1)

- Let a ***parametric model*** describe the distribution of the data
  - Example:
    normal distribution
    parameters are μ, σ



- Apply a ***statistical test*** that depends on:
  - Data distribution
  - Model parameters (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 $$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{N-1}} = 45.65$$

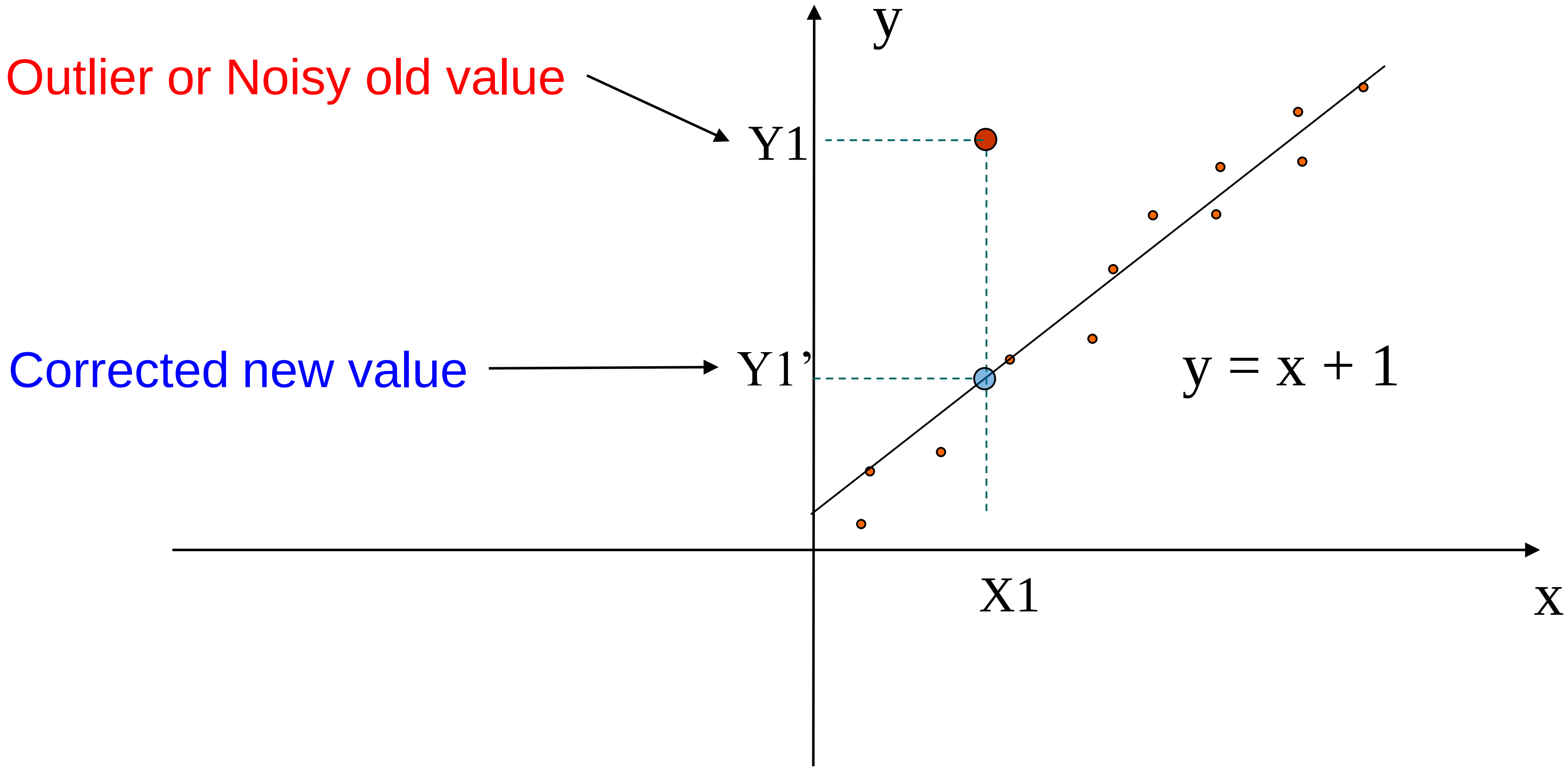
Select threshold value, e.g. 5% confidence for normal distribution:

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?

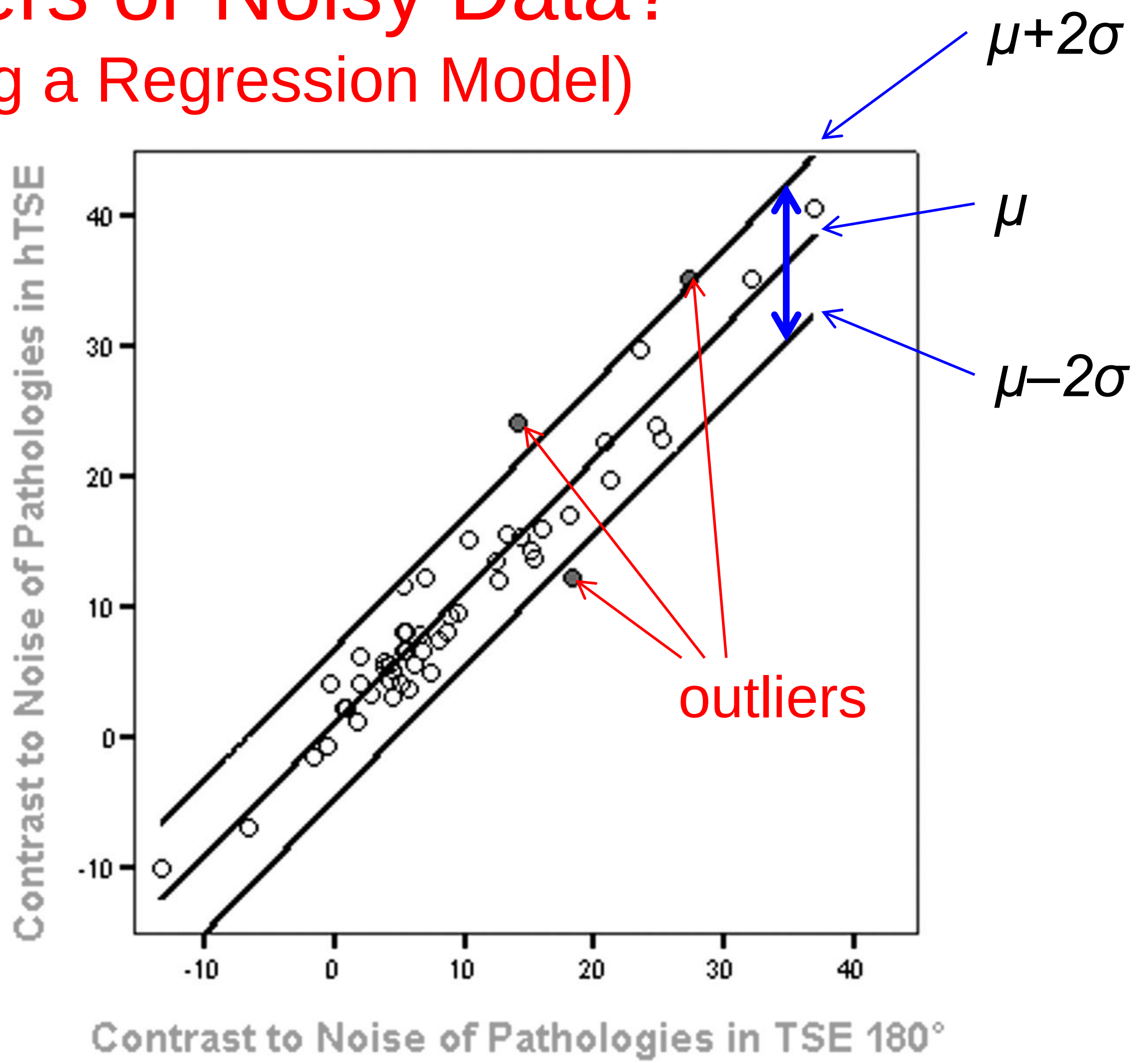
(Using a Regression Model)



- Model-based approach

Outliers or Noisy Data?
(Using a Regression Model)

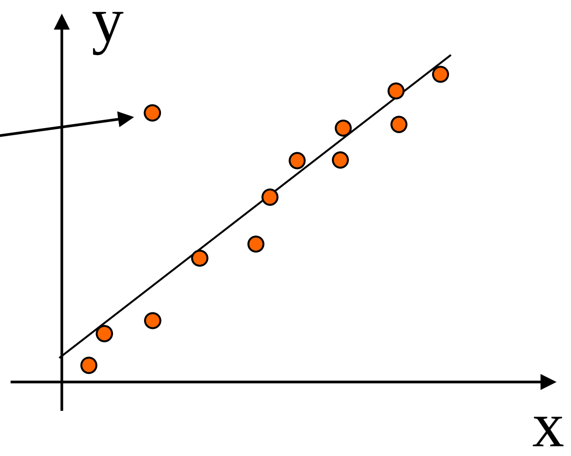
- Assumption:
  y-values Gauss-distributed around fitted curve (model)
- Model describes
  $\mu=\mu(x)$
- Recall:
  From $\mu-2\sigma$ to $\mu+2\sigma$ contains ~ 95%



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

- Determine 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

fraction *p*

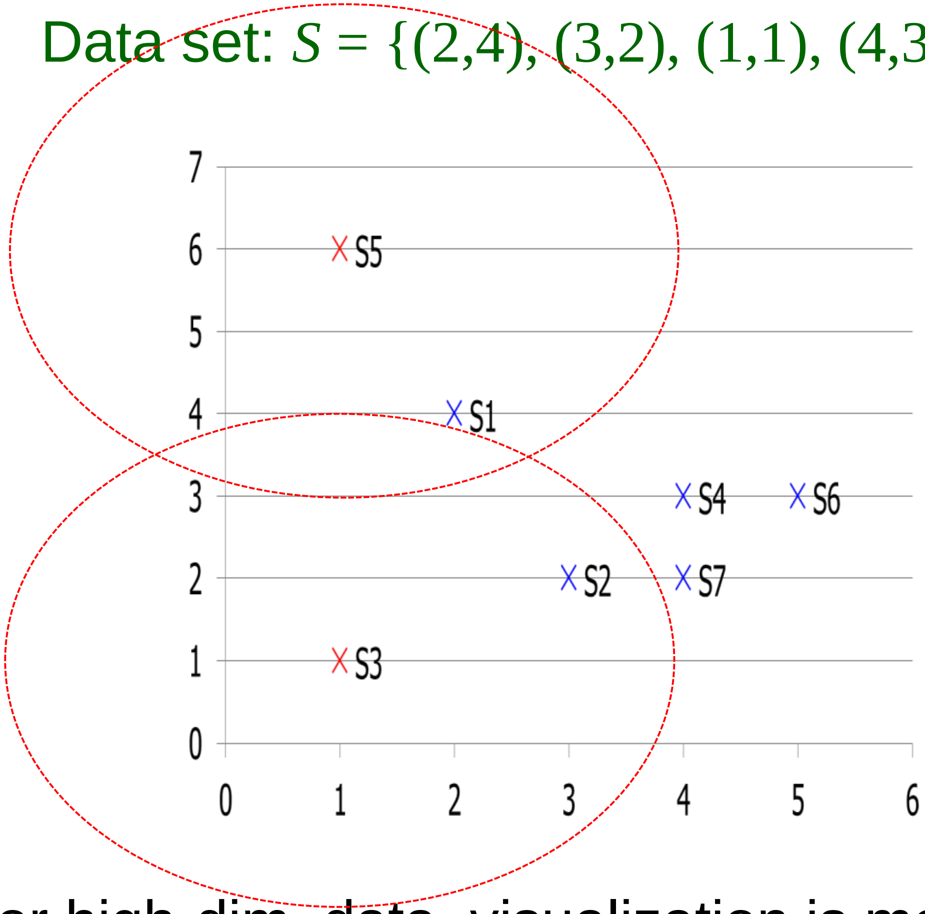
Sample	p
S1	2
S2	1
S3	**5**
S4	2
S5	**5**
S6	3
S7	2

Outliers

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-dim. data, visualization is more difficult
- For huge data sets, distance matrix gets large

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

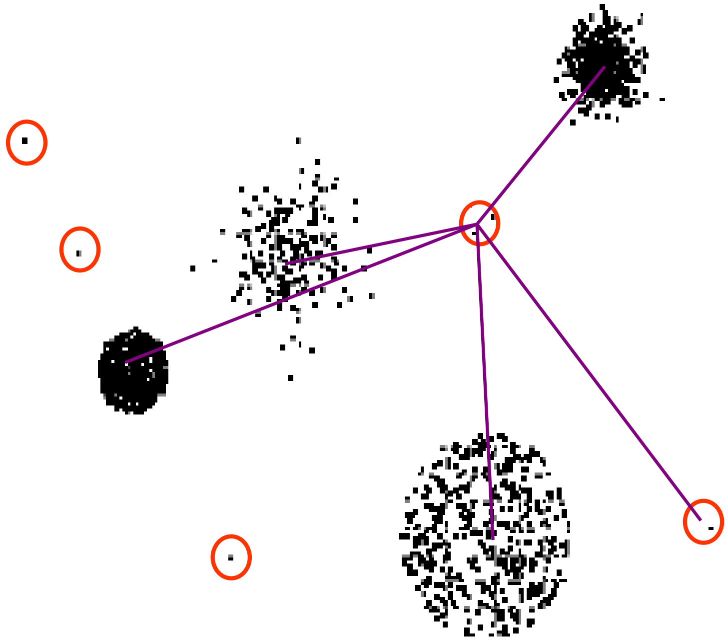
→ Distance-based outliers are those samples that do not have enough neighbors

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

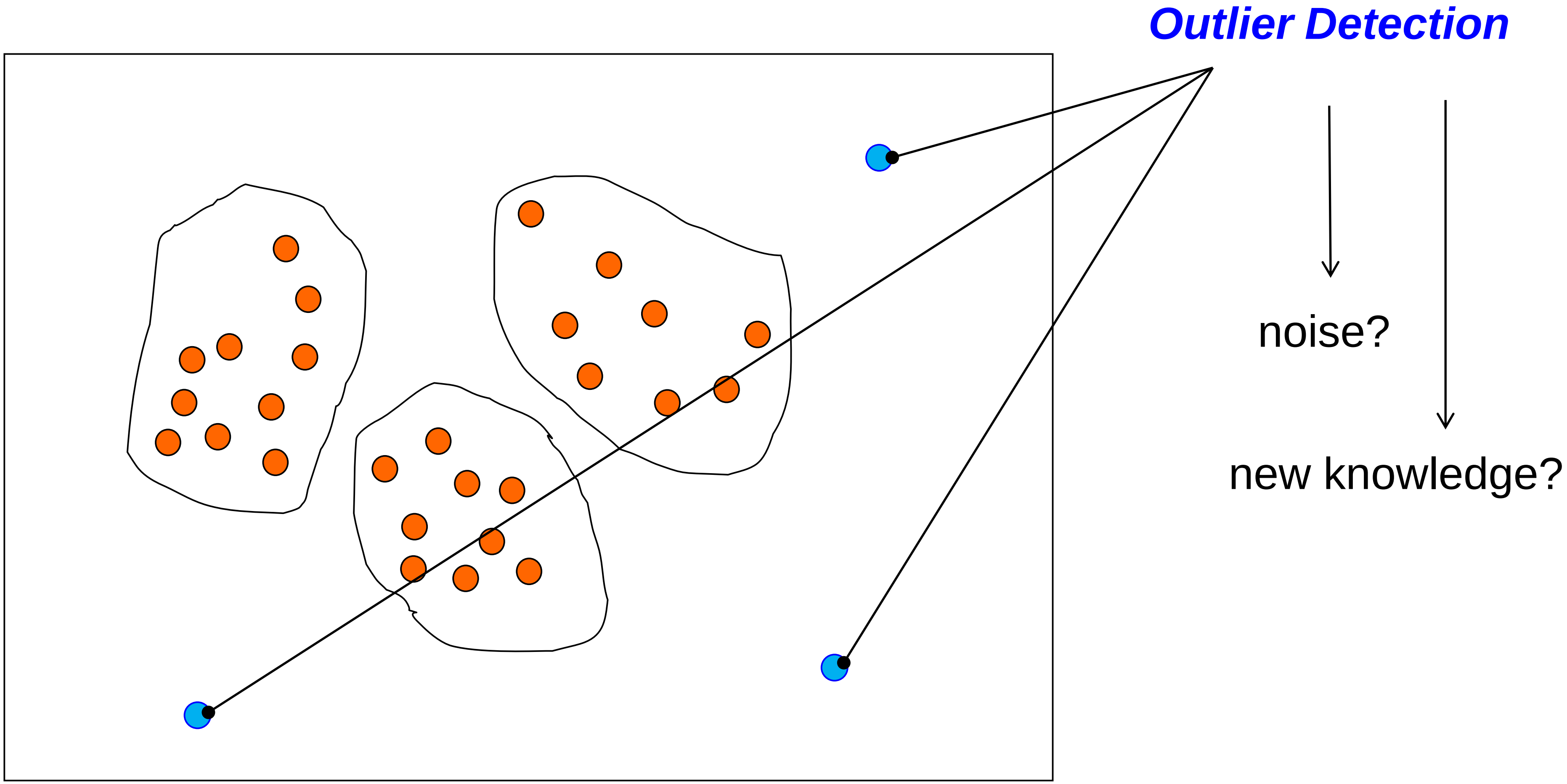
properties of the data determine useful settings of p and d

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)



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, …)

Overview

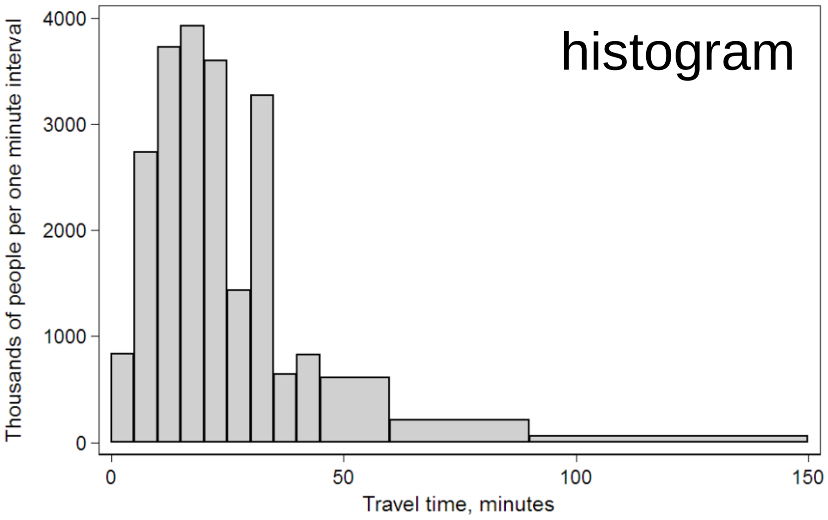
- Outliers
- Visualisation
- Similarities and Distance Measures

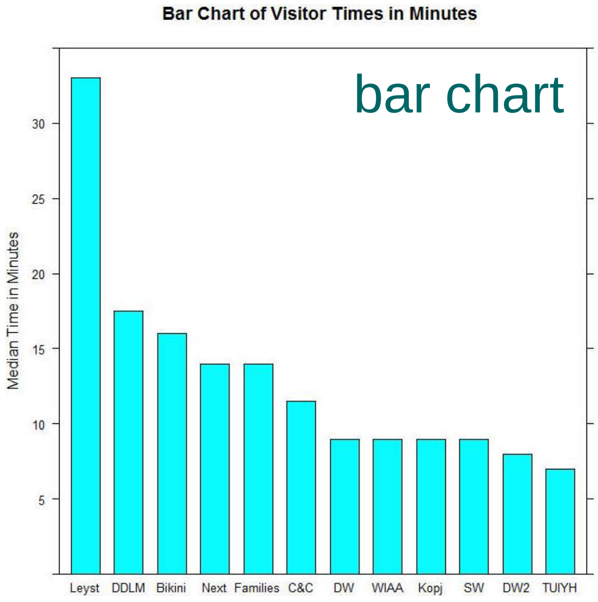
Displays of Basic Statistical Descriptions

- ***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

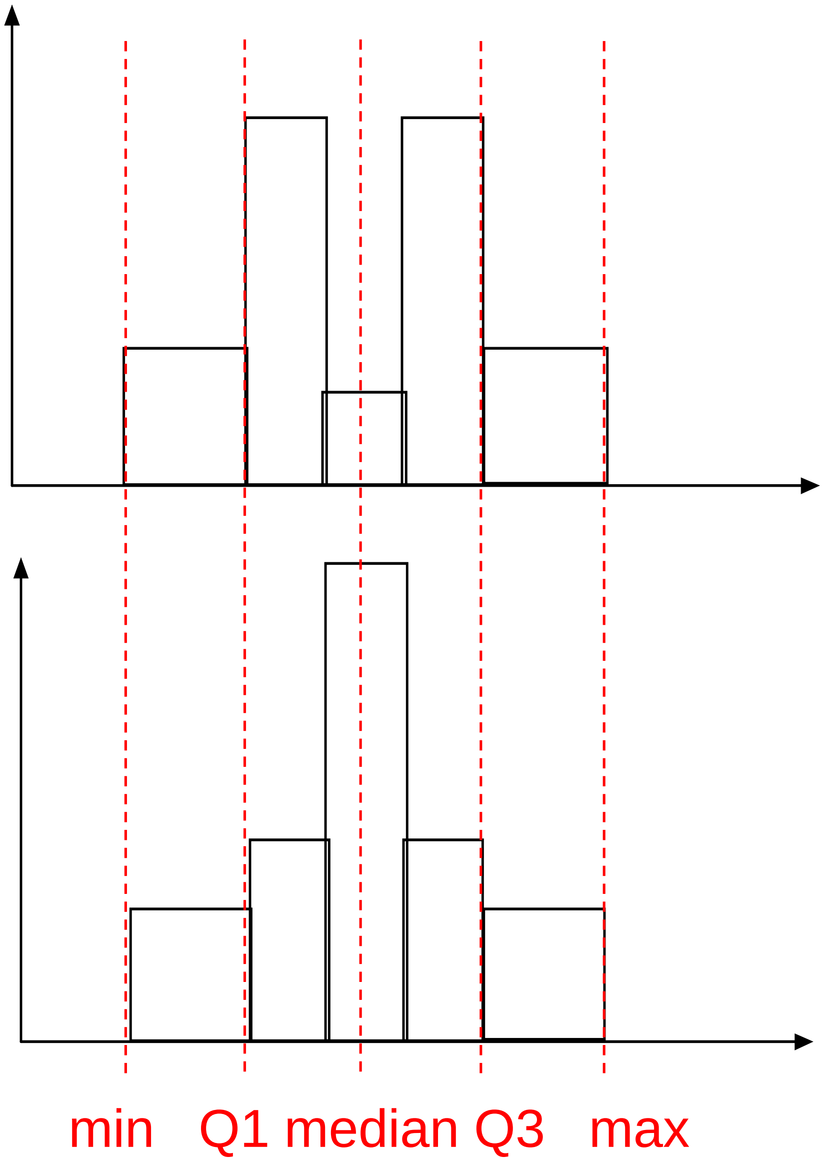
Histogram and Bar Chart Analysis

- Histograms are used to show *distributions* of a variable while bar charts are used to *compare* different 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.





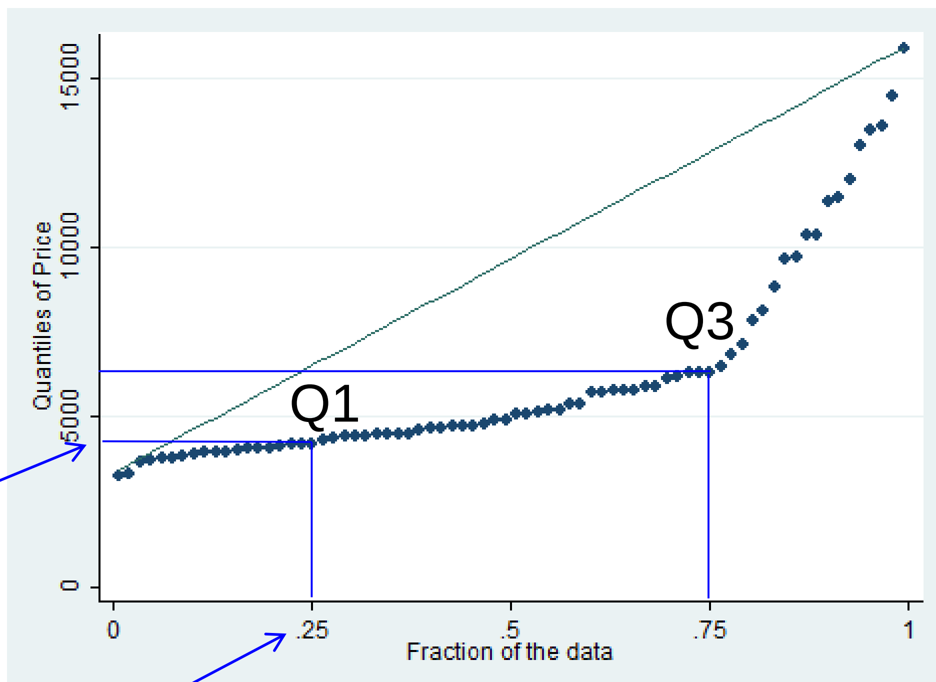
Histograms tell more than Boxplots



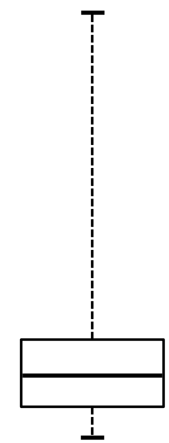
- The two histograms shown in the left may have the same boxplot representation
- The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

Quantile Plot

- Displays all of the data; 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



4200

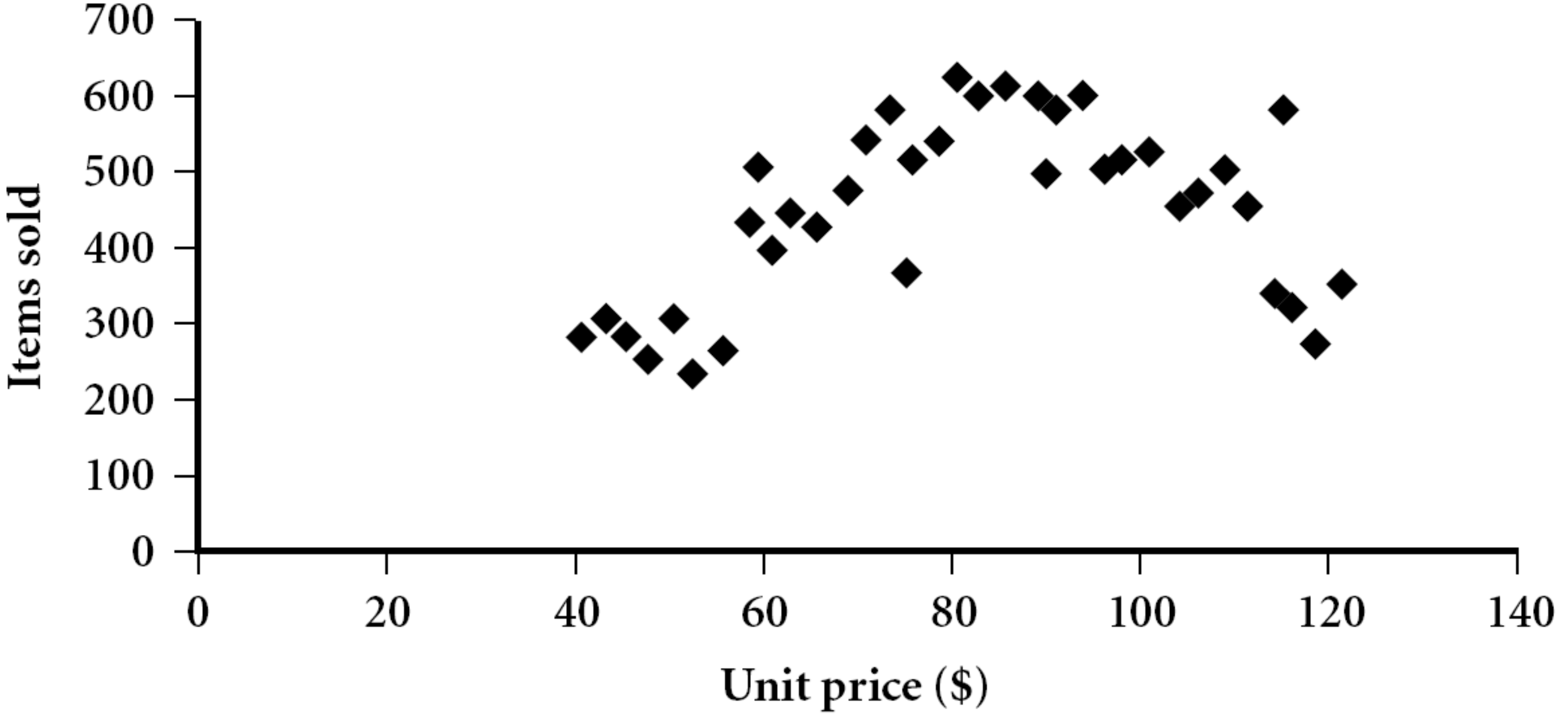


corresponding box plot

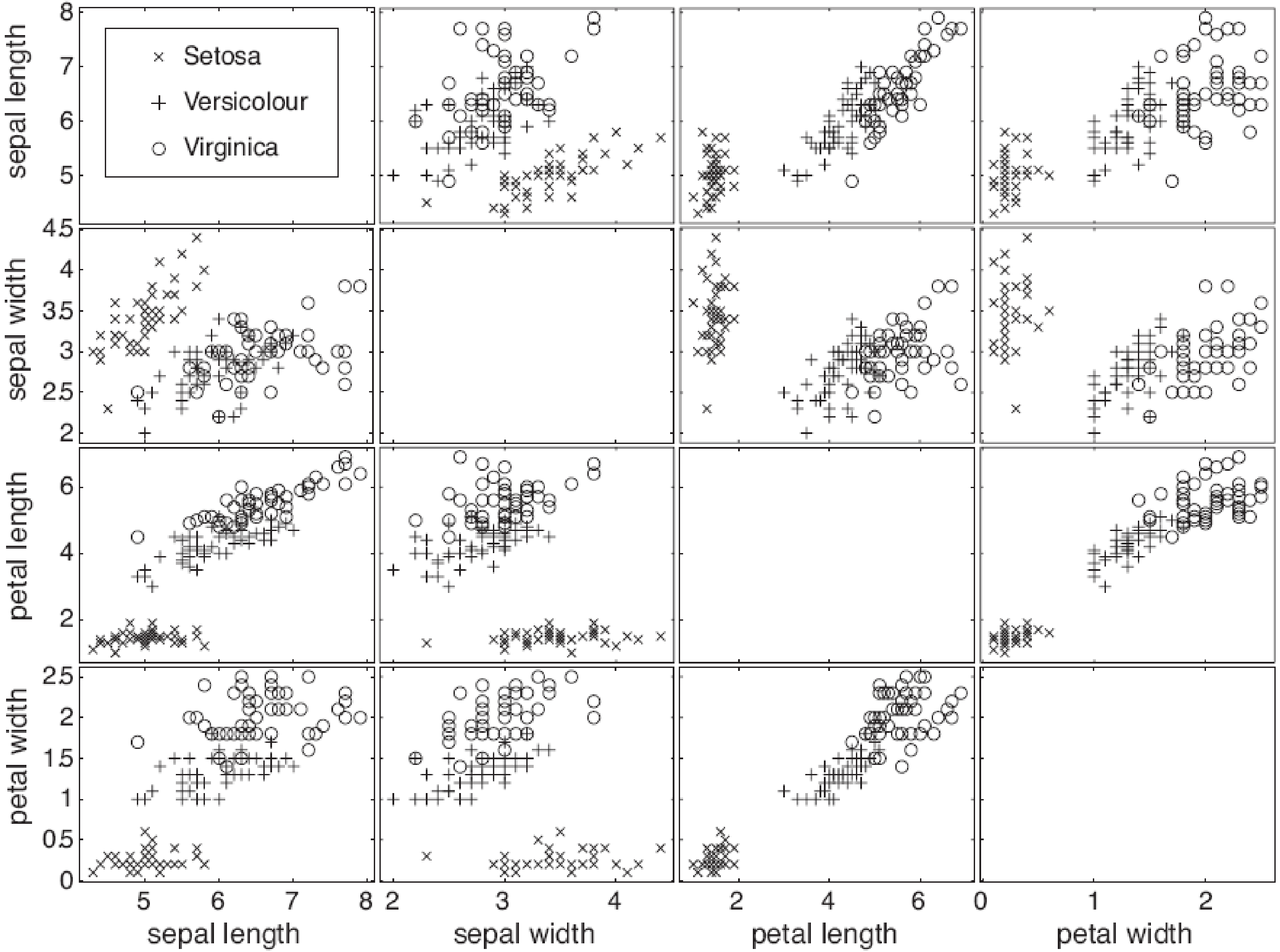
25% of the data are below or equal to the value 4200

Scatter Plot

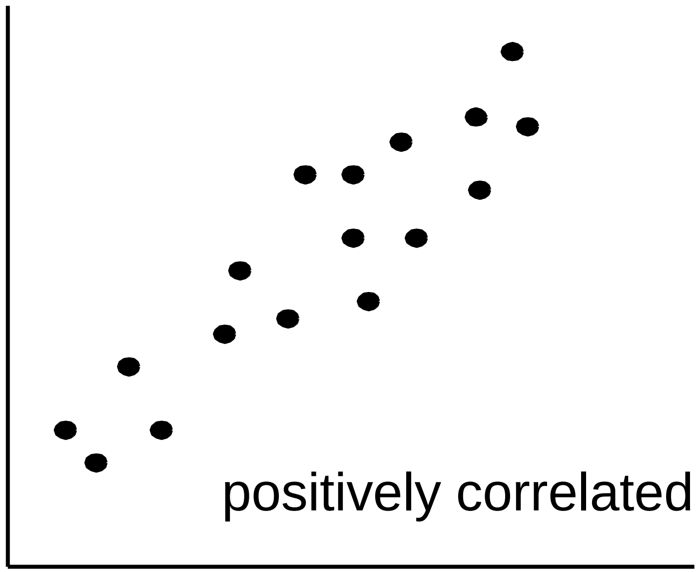
- Provides a ***first look*** at data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

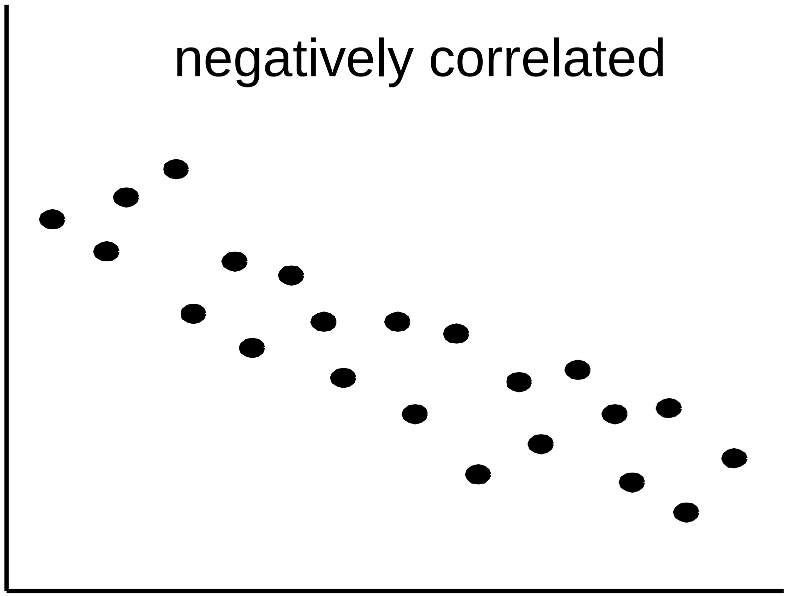


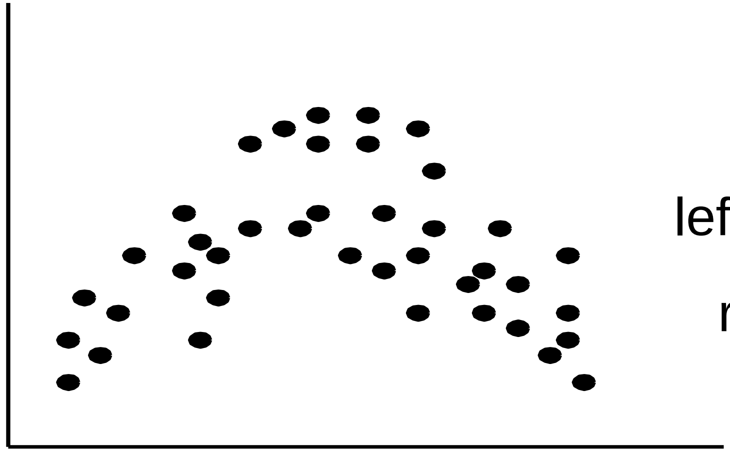
4x4 Matrix of Scatter Plots for 4-D Data



Positively and Negatively Correlated Data



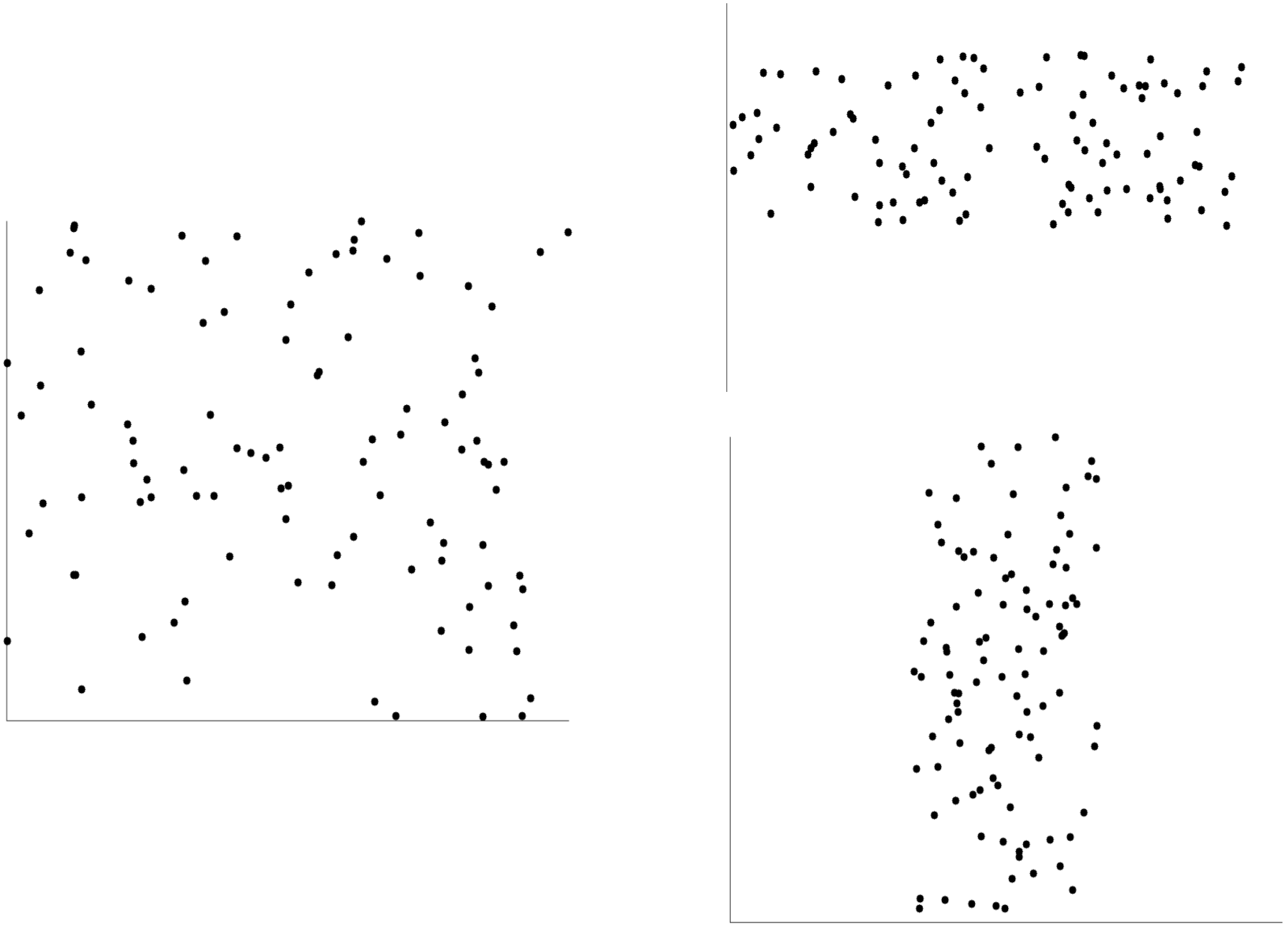




left half fragment: positively correlated

right half fragment: negatively correlated

Uncorrelated Data

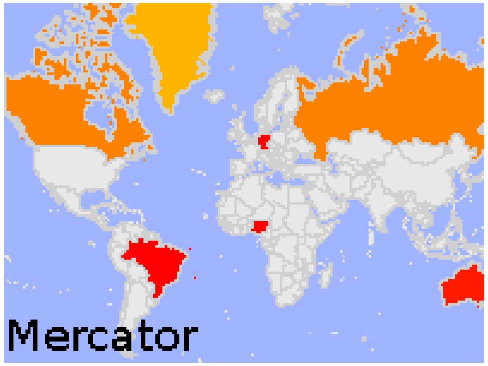


Further Forms of Data Visualization

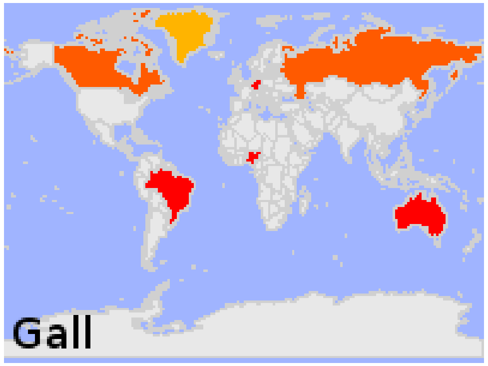
- Further aims of visualization:
  - Provide qualitative overview of large data sets
  - Gain insight into information space by mapping data onto graphical primitives
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help to find interesting regions
  - Identify suitable methods and parameters for further quantitative analysis
  - Provide a visual proof and sanity check of quantitative analyses

Geometric Projection – No One-size-fits-all

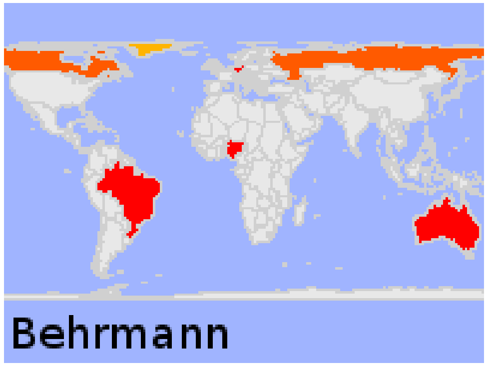
- Fitting a sphere onto a plane …



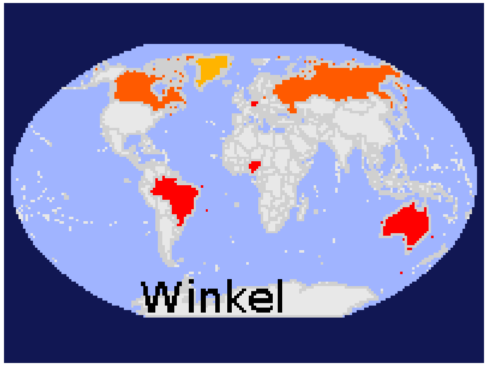
locally good shapes



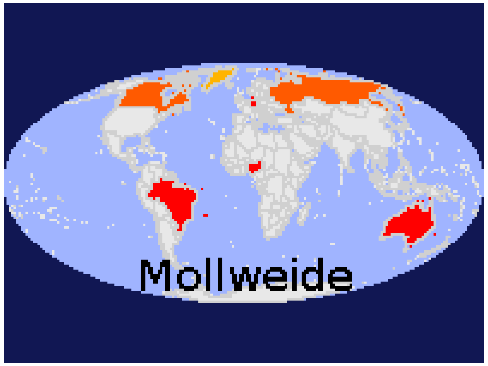
globally good shape



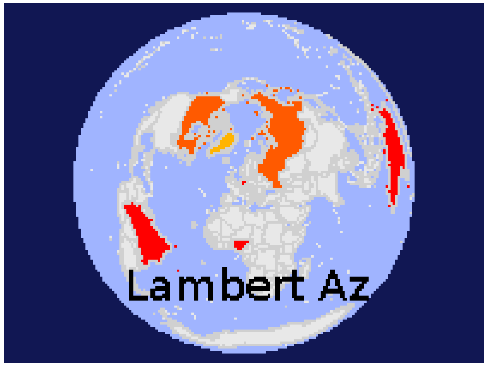
true area sizes



compromise between angle & area distortions

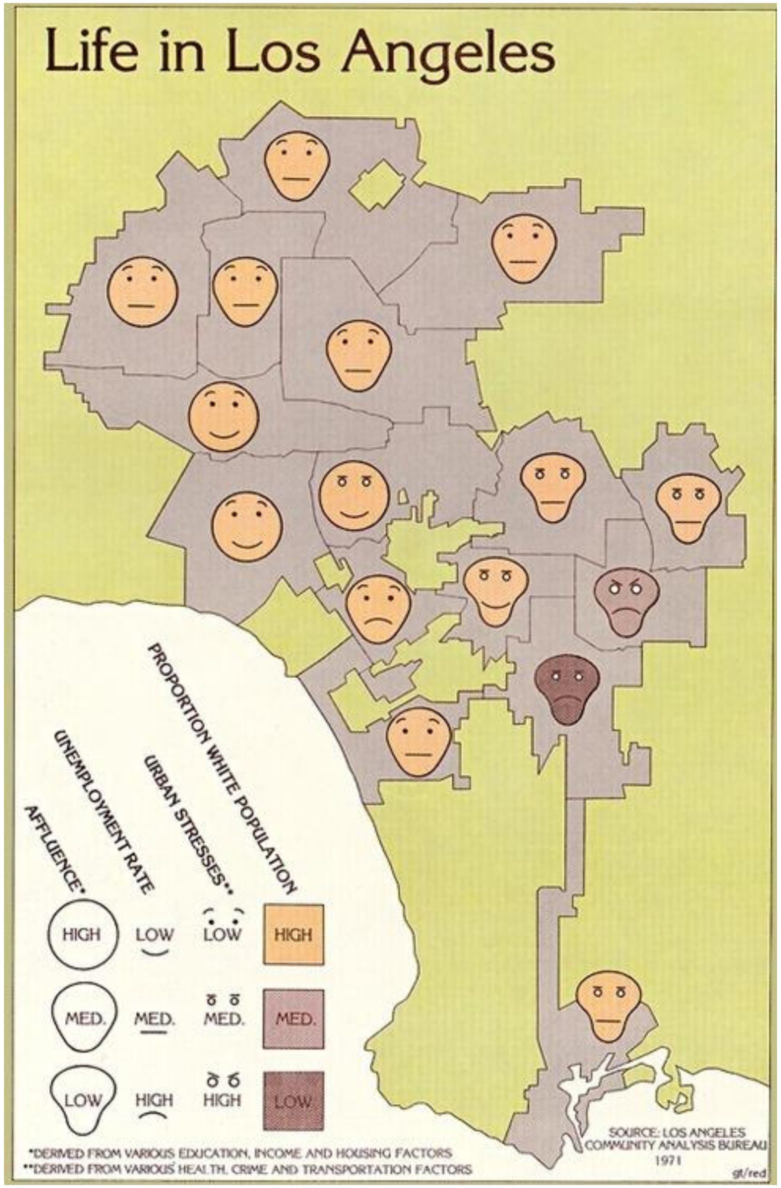


true area sizes, straight lines of latitude



true area sizes

Icon-based: Chernoff Faces

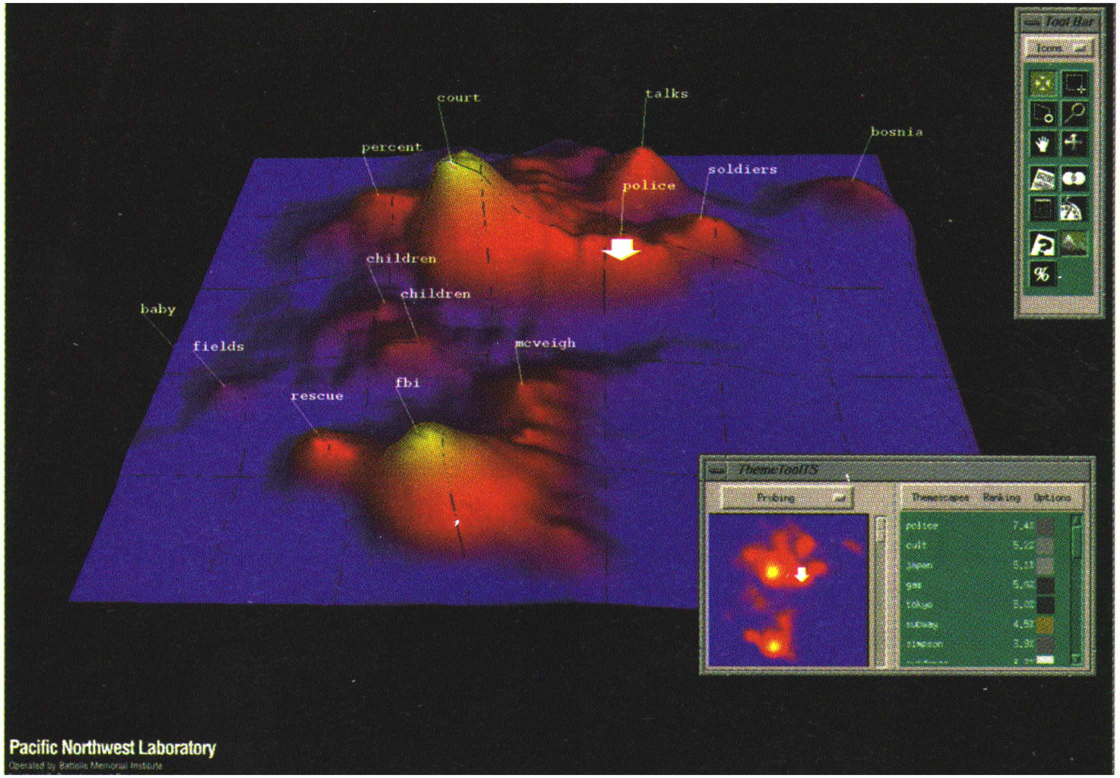


- Higher dimensionalities can be expressed – and easily perceived – in the parameters of cartoon faces
- Suitable for socio-economic data
- Yet, it is hard to be objective

http://maphugger.com/post/44499755749/the-trouble-with-chernoff

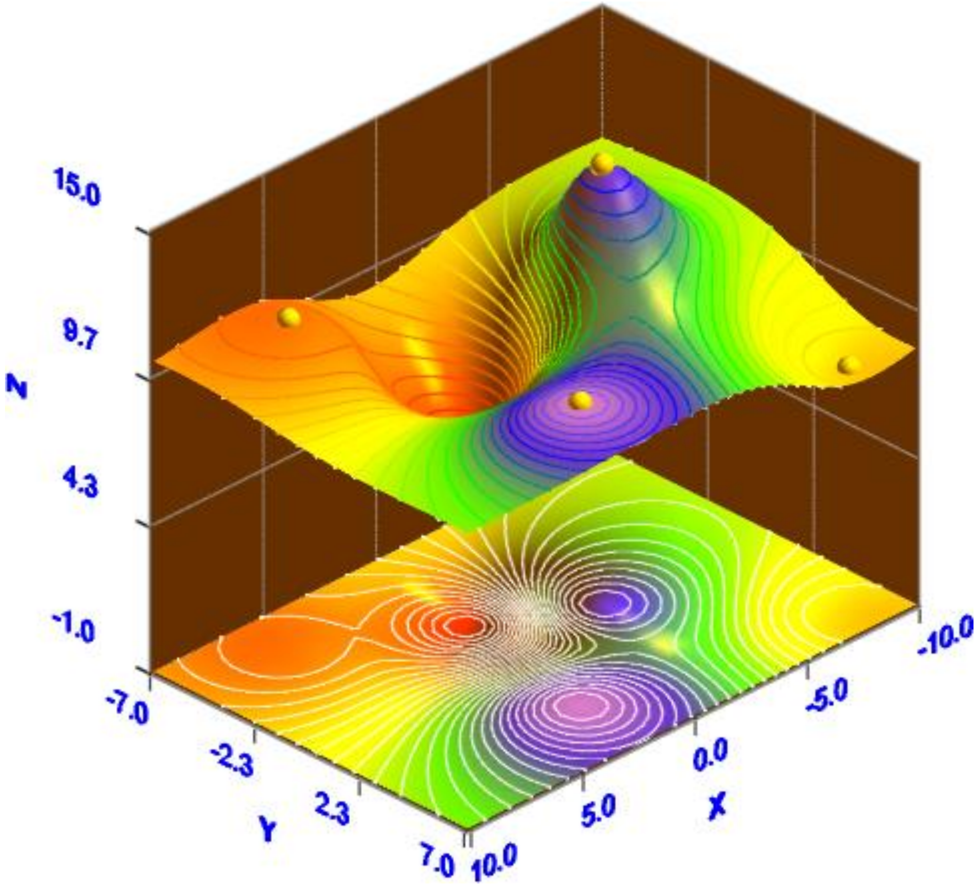
Visualization of Data as a Landscape

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



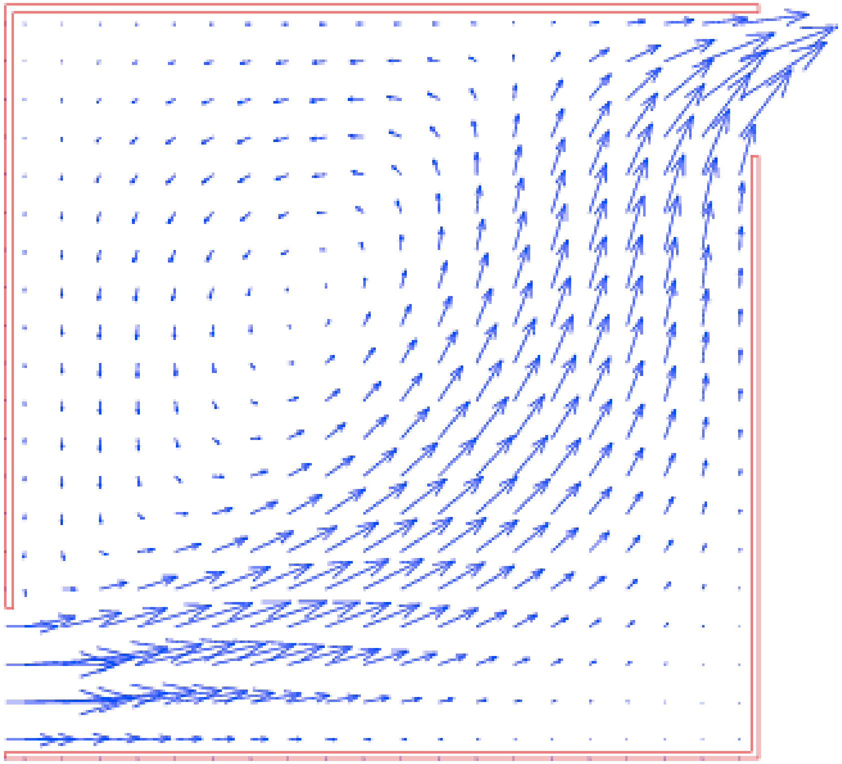
- Data first transformed into 2D
- Then, this shows a function 2D → 1D
- E.g. density of data over a 2D manifold

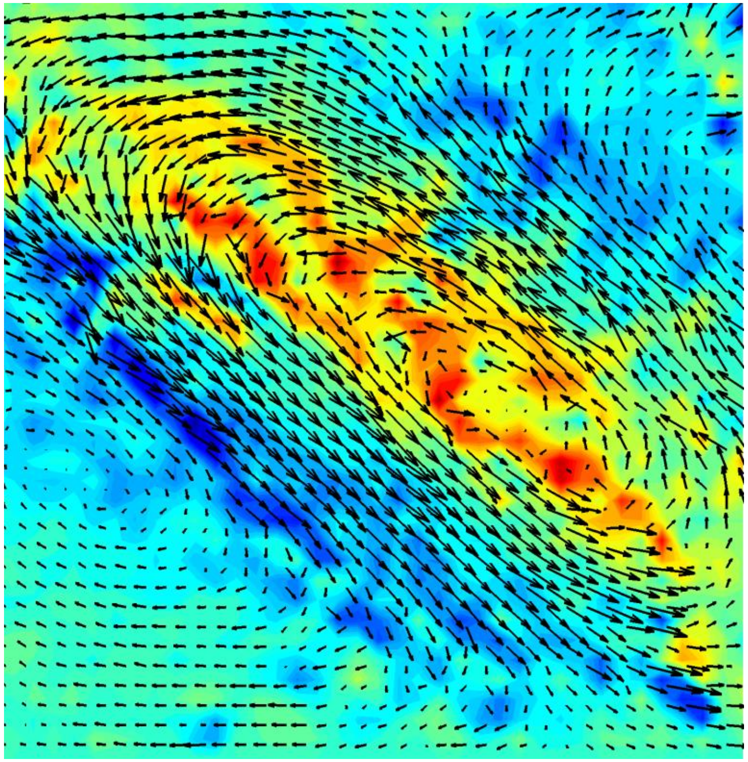
Visualization of Data as a Contour Plot



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

Visualization of Data as a Flow Field

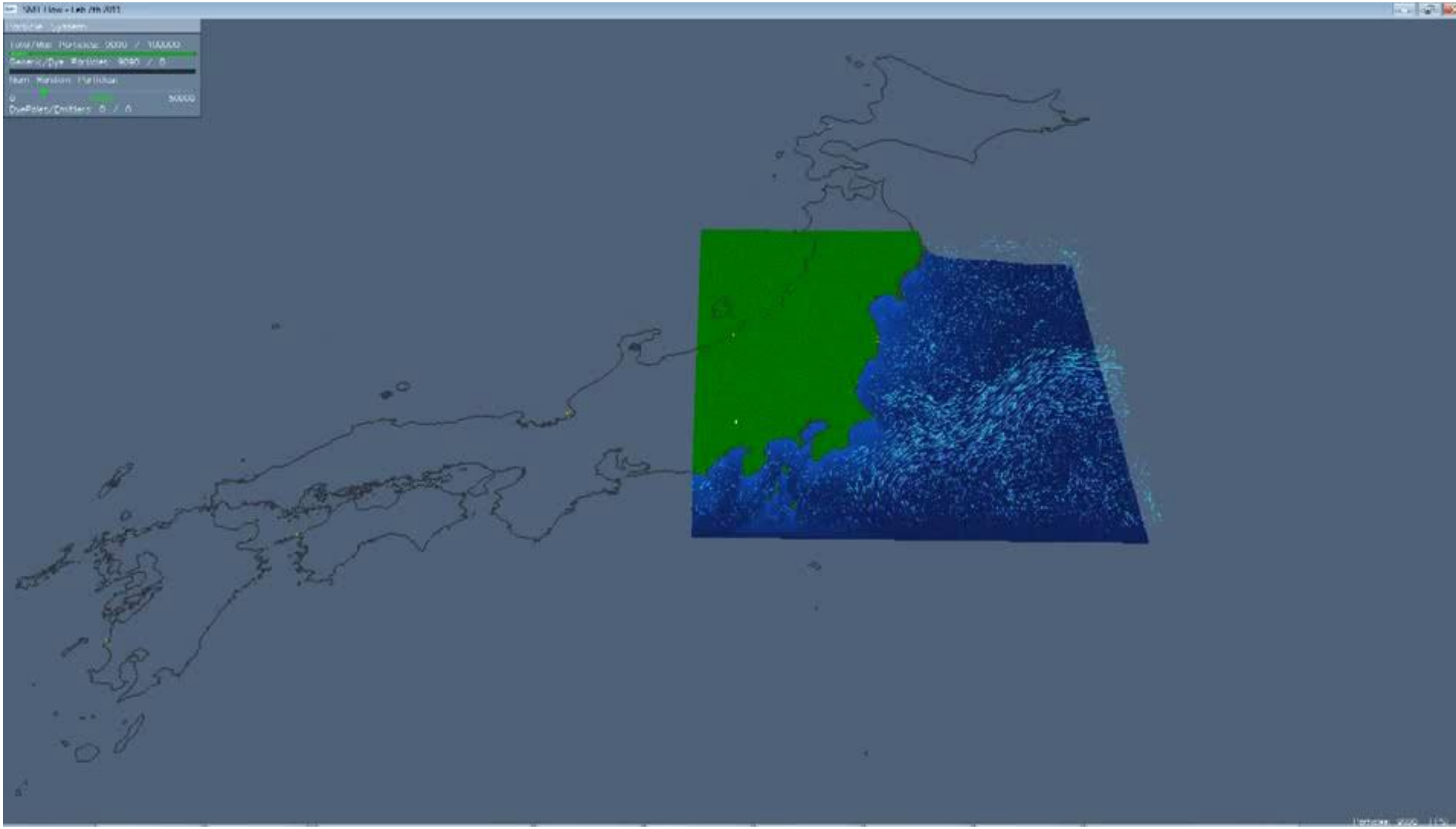




2D → 2D:
(x,y) → (dx,dy) or
(x,y) → (length,direction)

2D → 3D:
(x,y) → (dx,dy,color)

Example: Ocean Flow Analysis Visualization

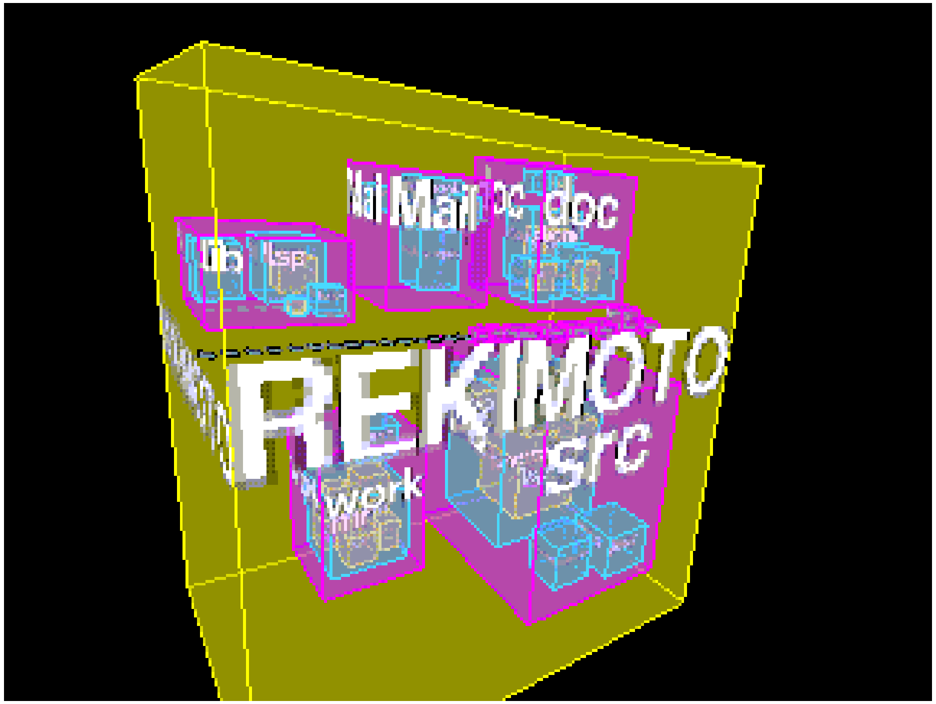


- Visualization: ocean flow simulation being run on the flow model

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

Hierarchical Visualisation: 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



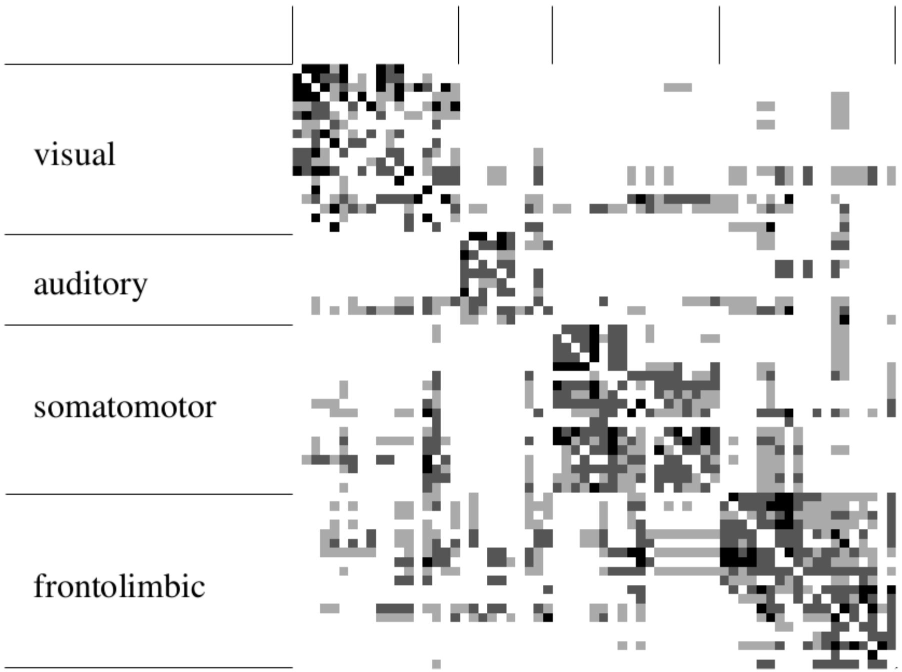
Visualizing Complex Data

- 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

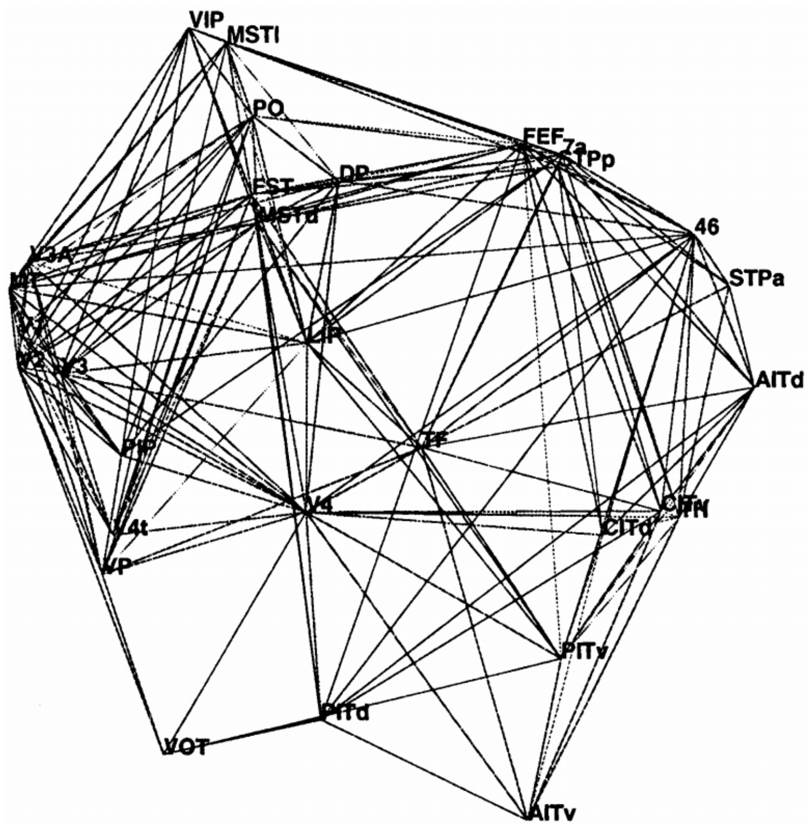
Newsmap: Google News Stories

Visualizing Relations

- Visualizing networks



Connections between 65 cortical areas in cat



Relations within the visual cluster
(non-metric multidimensional scaling)

Overview

- Outliers
- Visualisation
- Similarities and Distance Measures

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

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

  - A triangular matrix

Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Distance between vectors *i* and *j* of nominal attributes:
  - **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 *j*		
		1	0	sum
Object *i*	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 j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- *Distance* measure for symmetric binary variables:

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

- *Distance* measure for asymmetric binary variables: (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; we will neglect gender
- Use distance for asymmetric case:

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

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

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

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

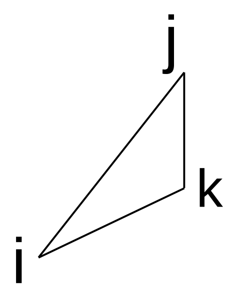
Distance on Numeric Data: Minkowski Distance

- ***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

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

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

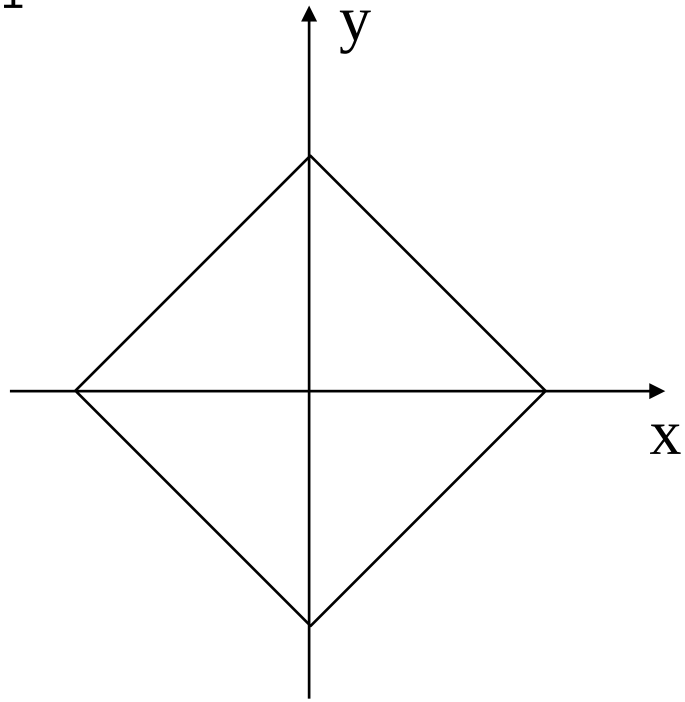
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

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

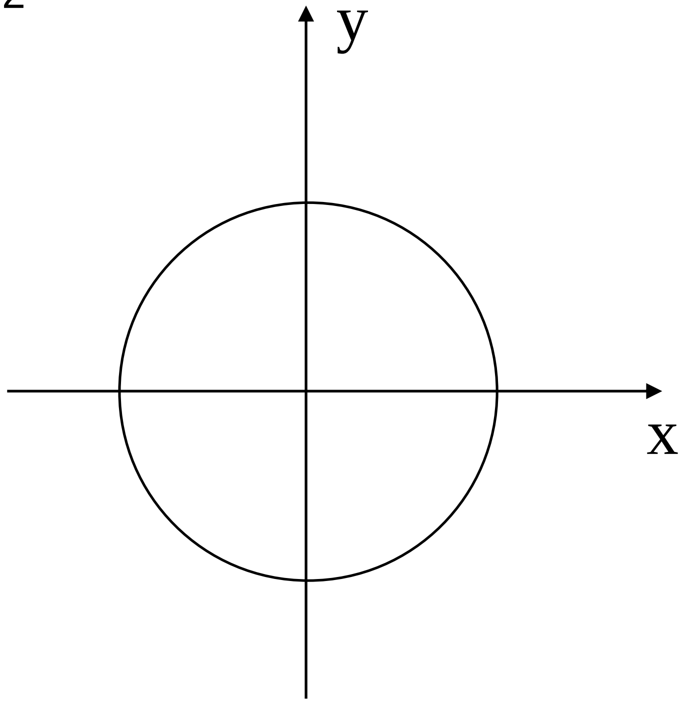
$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

Iso-contour Lines

L_1 norm

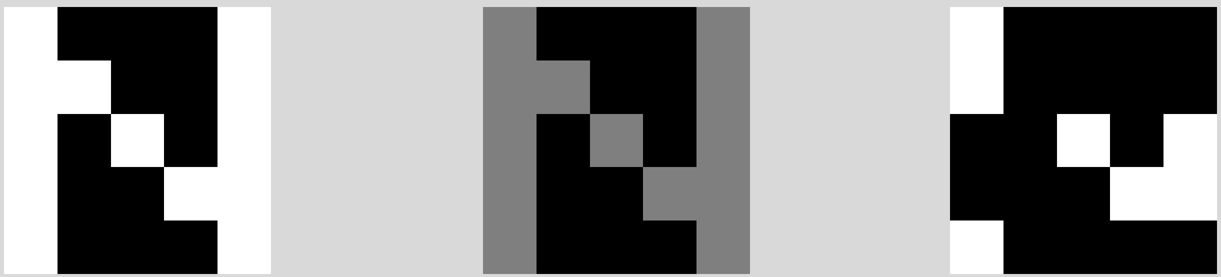


L_2 norm



Distances Between Image (Vector)s

$$\boldsymbol{v1} = \begin{matrix} 2&0&0&0&2 \\ 2&2&0&0&2 \\ 2&0&2&0&2 \\ 2&0&0&2&2 \\ 2&0&0&0&2 \end{matrix} \qquad \boldsymbol{v2} = \begin{matrix} 1&0&0&0&1 \\ 1&1&0&0&1 \\ 1&0&1&0&1 \\ 1&0&0&1&1 \\ 1&0&0&0&1 \end{matrix} \qquad \boldsymbol{v3} = \begin{matrix} 2&0&0&0&0 \\ 2&0&0&0&0 \\ 0&0&2&0&2 \\ 0&0&0&2&2 \\ 2&0&0&0&0 \end{matrix}$$



$$L_1(\boldsymbol{v1}-\boldsymbol{v2}) = \mathbf{13} \quad > \quad L_1(\boldsymbol{v1}-\boldsymbol{v3}) = \mathbf{12}$$

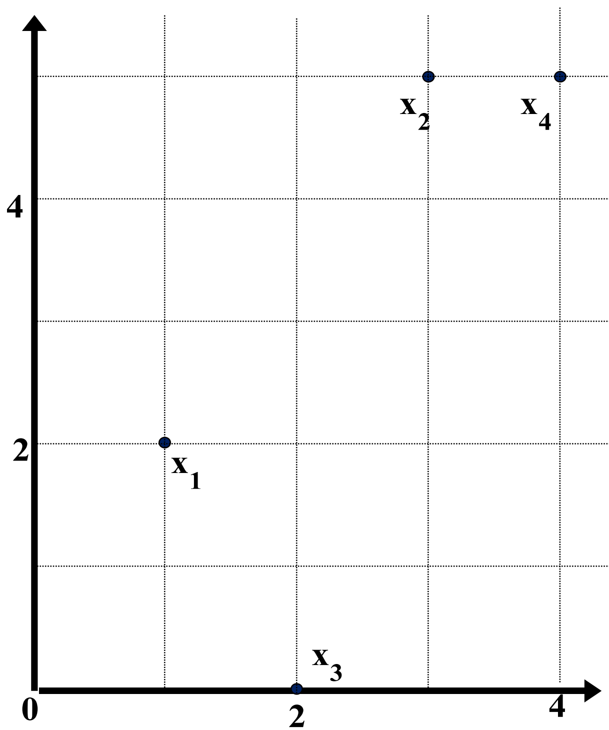
$$L_2(\boldsymbol{v1}-\boldsymbol{v2}) \approx \mathbf{3.6} \quad < \quad L_2(\boldsymbol{v1}-\boldsymbol{v3}) \approx \mathbf{4.9}$$

- 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)

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

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

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

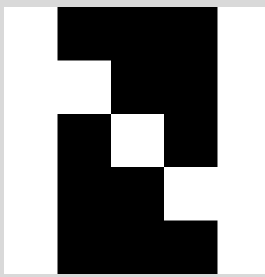
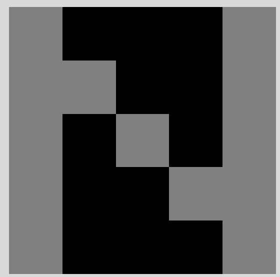
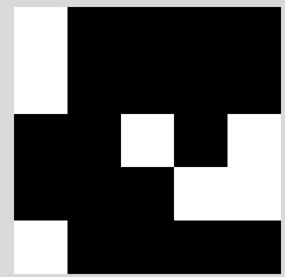
$||d_2|| = (3*3 + 0*0 + 2*2 + 0*0 + 1*1 + 1*1 + 0*0 + 1*1 + 0*0 + 1*1)^{\mathbf{0.5}}$

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

$\cos(d_1, d_2) = 0.94$

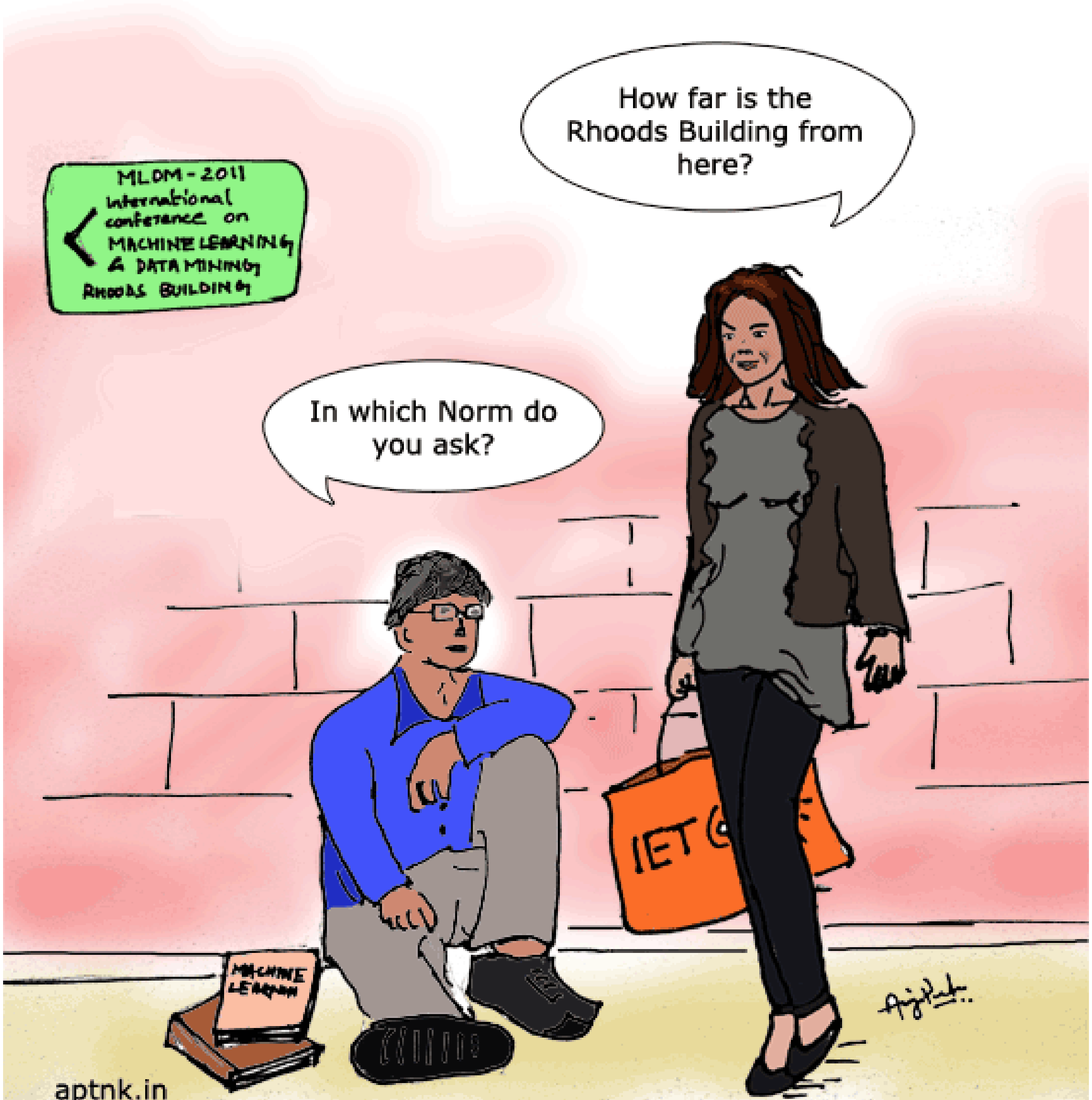
Cosine Similarity Between Image (Vector)s

$$\boldsymbol{v1} = \begin{matrix} 2 & 0 & 0 & 0 & 2 \\ 2 & 2 & 0 & 0 & 2 \\ 2 & 0 & 2 & 0 & 2 \\ 2 & 0 & 0 & 2 & 2 \\ 2 & 0 & 0 & 0 & 2 \end{matrix} \qquad \boldsymbol{v2} = \begin{matrix} 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{matrix} \qquad \boldsymbol{v3} = \begin{matrix} 2 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 2 \\ 0 & 0 & 0 & 2 & 2 \\ 2 & 0 & 0 & 0 & 0 \end{matrix}$$

$\cos(\boldsymbol{v1},\boldsymbol{v2}) = \mathbf{1}$ $\qquad$ $\cos(\boldsymbol{v1},\boldsymbol{v3}) \approx \mathbf{0.73}$

- Cosine similarity considers vector orientations but not vector lengths


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

Summary

- ***Outlier*** detection
  - graphical, statistics-based, distance-based, …
- Data ***visualization***: map data onto graphical primitives
  - Further descriptions: histograms, bar charts, quantile plots
- Measures for data ***(dis)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