Lecture 2:

Data, measurements, and visualization

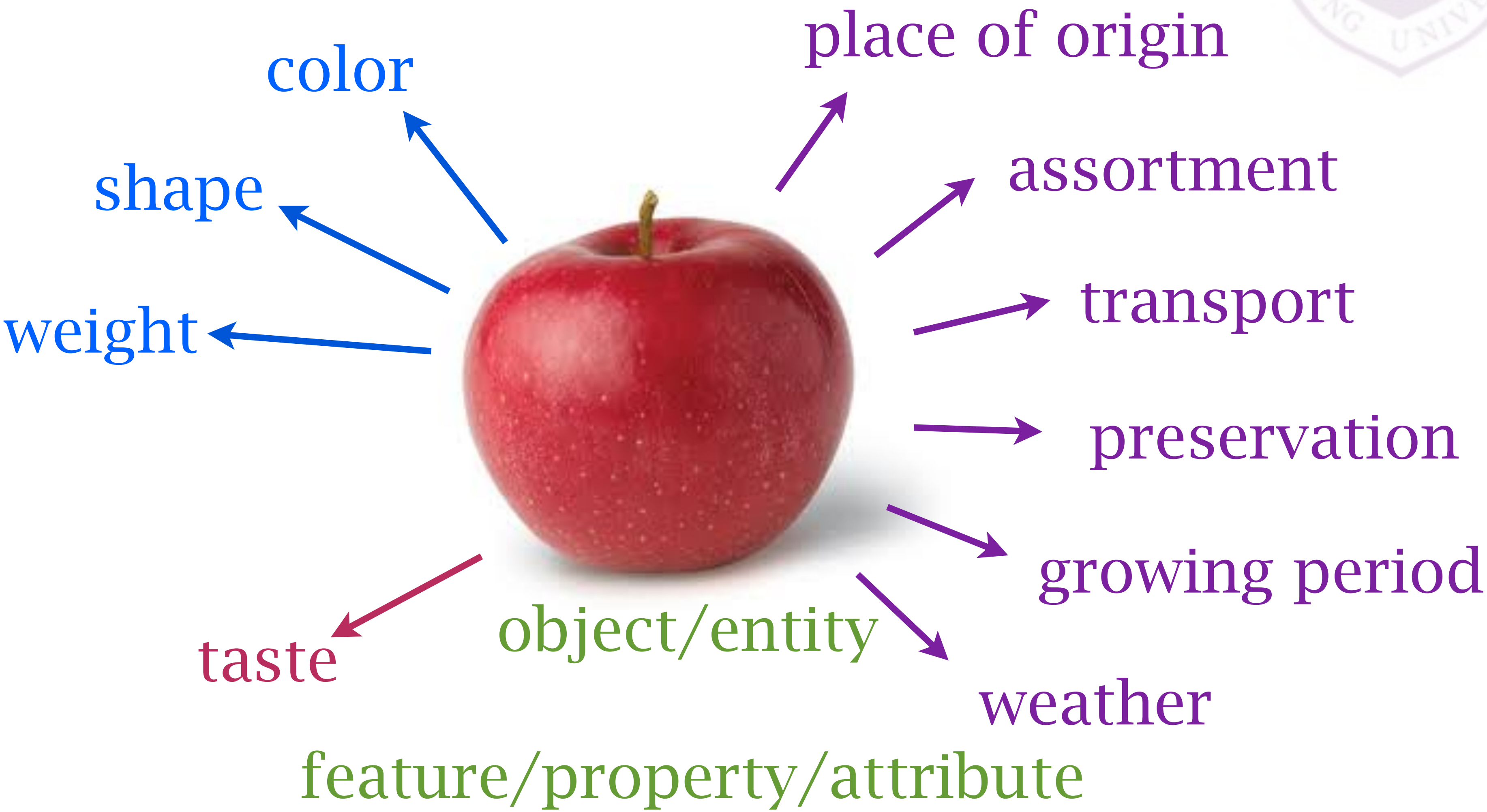
http://cs.nju.edu.cn/yuy/course_dm12.ashx

What is data

Data are collected by mapping entities in the domain of interest to symbolic representation by means of some **measurement** procedure, which associates the value of a variable with a **given property** of an entity.

[D. Hand et al. , Principles of Data Mining]

Object and attribute



name	color	shape	weight	PoO	assortment	transport	preservation	growing	weather	taste
A1	red	round	200	Yantai	H	express	frozen	150	sunny	sweet

Data quality

sufficient features

a good data set= sufficient amount of unbiased sampled data

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

noise free

garbage in garbage out

data from http://www.alistapart.com/articles/zebrastripingdoesithelp/

Types of attribute

- Nominal
- Ordinal
- Numerical

why should we care about the type
  proper description
  proper approach

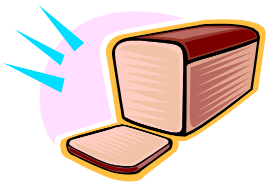
Types of attribute

Nominal / categorical / discrete:

The values of the attribute are only **symbols**, which is used to distinguish each other.

- Finite number of candidates
- No order information
- No algebraic operation can be conducted

e.g., {1, 2, 3}
~ {Red, Green, Blue}
~ {Milk, Bread, Coffee}



Types of attribute

Ordinal:

The values of the attribute is to indicate certain **ordering relationship** resided in the attribute.

- Order is more important than value!
- No algebraic operation can be conducted except those related to sorting.

e.g., {1, 2, 3}
~ {Fair, Good, Excellent}
~ {Irrelevant, Relevant, Highly relevant}

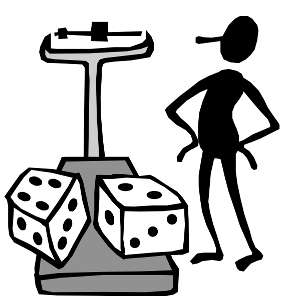
Types of attribute

Numerical / real:

The values of the attribute is to indicate the **quantity** of some predefined unit.

- There should be a basic unit.
- The value is how many copies of the basic unit
- Some algebraic operation can be conducted w.r.t the meaning of the attribute

e.g., 4 km = 4 * 1km
4 km is twice as longer as 2 km



Data transformation

- Legitimate transformation
- Normalization
- Transformation of attribute type

why should we care about transformation

Legitimate transformation

- Nominal scale:
  Bijective mapping (=)   **e.g., 1 → 4**
- Ordinal scale:
  Monotonic increasing (<)   **e.g., {1,2, 3} → {2,6,10}**
- Ratio scale:
  Multiplication (*)   **e.g., 2 → 20**
- Interval scale:
  Affine (*, +)   **e.g., 2 → 21**

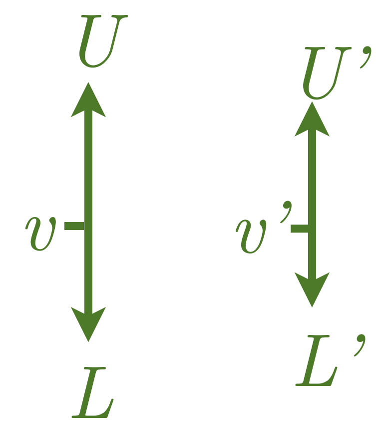
Normalization

Normalization is to scale the (numerical) attribute values to some specified range

- min-max normalization

$$v' = \frac{v - L}{U - L}(U' - L') + L'$$

out of bound risk

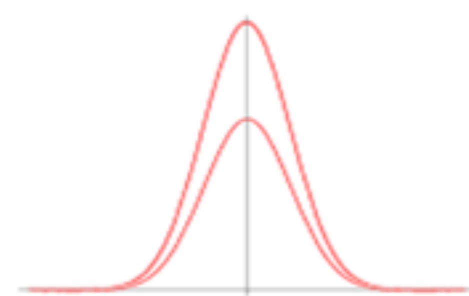


- z-score normalization

$$v' = \frac{v - \mu}{\sigma}$$

μ -- mean

σ^2 -- variance



- decimal scaling normalization

$$v' = \frac{v}{10^j}$$

j is the smallest integer such that $\max\{|v'|\} \leq 1$

Transformation of attribute type

discretization:
numerical --> nominal/ordinal

Natural partitioning (unsupervised):

The 3-4-5 rule: For the most significant digit,

- if it covers {3,6,7,9} distinct values then divide it into 3 equi-width interval;
- if it covers {2,4,8} distinct values then divide it into 4 equi-width interval;
- if it covers {1,5,10} distinct values then divide it into 5 equi-width interval

(0,500)

(0,100)	[100,200)	[200,300)	[300,400)	[400,500)
0	1	2	3	4

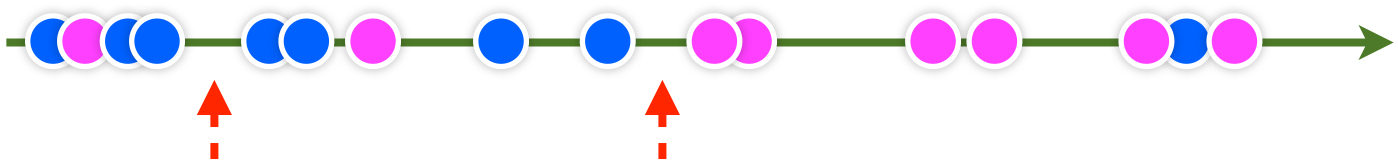
(300,1000)

(300,533)	[533,766)	[766,1000)
low	moderate	high

Transformation of attribute type

discretization:
numerical --> nominal/ordinal

Entropy-based discretization (supervised):



Entropy: $H(X) = -\sum_i p_i \ln(p_i)$ $\quad p_1 = \frac{\#\text{blue}}{\#\text{all}}$

Entropy after split:

$$I(X;\text{split}) = \frac{\#\text{left}}{\#\text{all}} H(\text{left}) + \frac{\#\text{right}}{\#\text{all}} H(\text{right})$$

Information gain:

$$Gain(X; split) = H(X) - I(X;\text{split}) > \theta$$

Transformation of attribute type

continuous-lization:
nominal --> continuous/ordinal

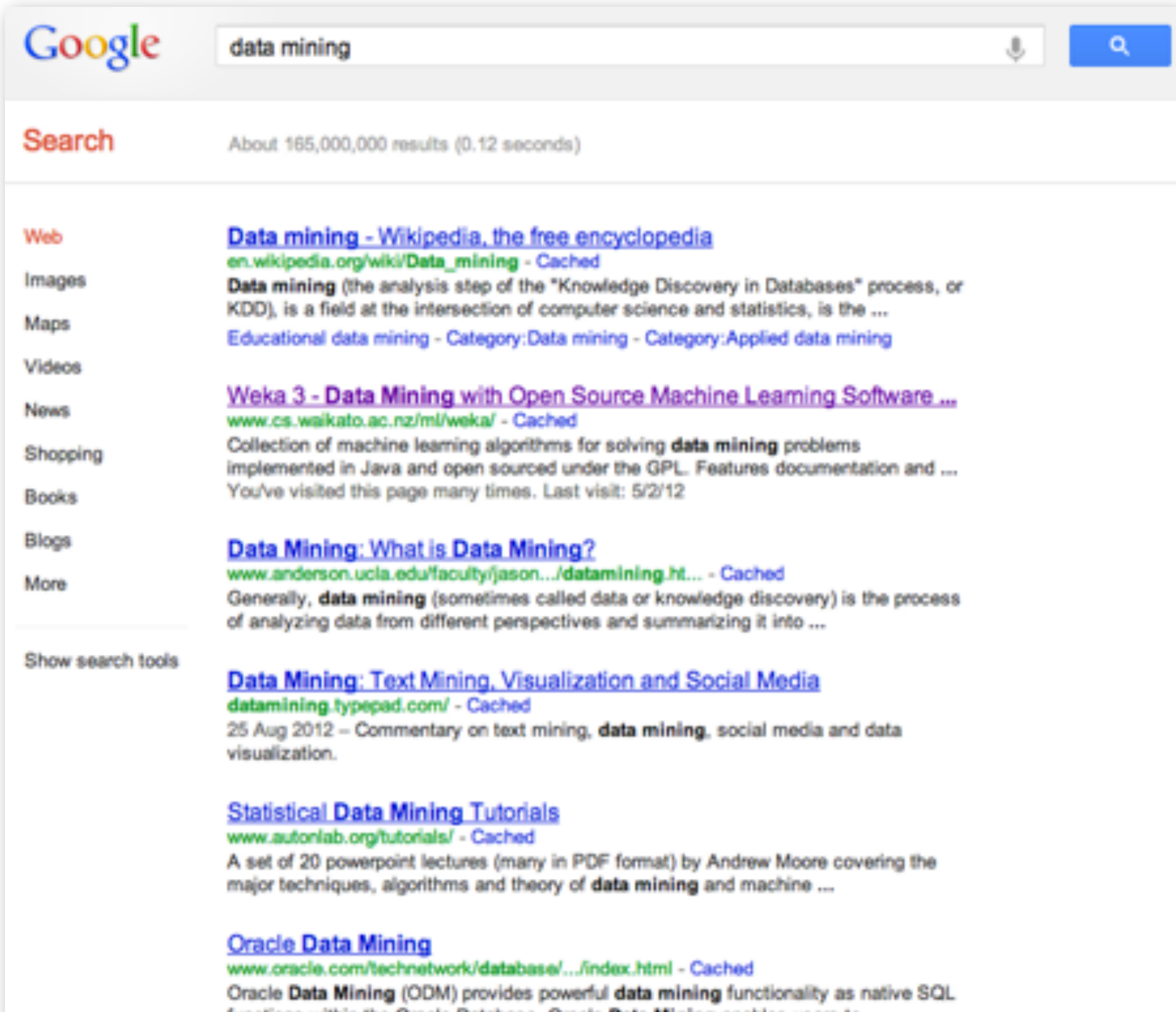
How to assign values to nominal symbols?

red	-> 1
orange	-> 2
green	-> 8
blue	-> 10

Similarity and distance

Similarity is an essential concept in DM

distance is a commonly used similarity



What is distance

distance is a function of two objects satisfying

- Non-negativity: $d(i,j) \geq 0, d(i,i) = 0$

- Symmetry: $d(i,j) = d(j,i)$

- Triangle inequality: $d(i,j) \leq d(i,k) + d(k,j)$

Common similarity functions

Minkowski distance:

order p (p-norm) $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^{p} \right)^{\frac{1}{p}}$$

special cases:

p=2: Euclidean distance $\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

p=1: Manhattan distance $\sum_{i=1}^{n} |x_i - y_i|$

p->+∞ : $\max_{i=1,2,\ldots,n} |x_i - y_i|$

Questions: what is the effect of normalization? what if p<1?

Common similarity functions

weighted Minkowski distance:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{n} w_i \left| x_i - y_i \right|^p \right)^{\frac{1}{p}}$$

Mahalanobis distance:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left((\boldsymbol{x} - \boldsymbol{y})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{y}) \right)^{\frac{1}{2}}$$

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}.$$

$\Sigma = I$: Euclidean distance

Σ is diagonal: normalized Euclidean $\sqrt{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\sigma_i^2}}$

Common similarity functions

Distances/similarities for binary strings:

- Hamming distance

$$d(\ 01010, 01001) = 2$$

- Matching coefficient

$$Sim = \frac{n_{1,1} + n_{0,0}}{n_{1,1} + n_{0,0} + n_{1,0} + n_{0,1}}$$

- Jaccard coefficient

$$J = \frac{n_{1,1}}{n_{1,1} + n_{1,0} + n_{0,1}}$$

- Dice coefficient

$$D = \frac{2n_{1,1}}{2n_{1,1} + n_{1,0} + n_{0,1}}$$

$n_{0,0}$	$n_{0,1}$
$n_{1,0}$	$n_{1,1}$

Common similarity functions

Dealing with nominal attributes

- convert to binary attributes

apple	(0,0,1)
orange	(0,1,0)
banana	(1,0,0)

- VDM (value difference metric)

#instances having value *x* in class *c*

#instances having value *x*

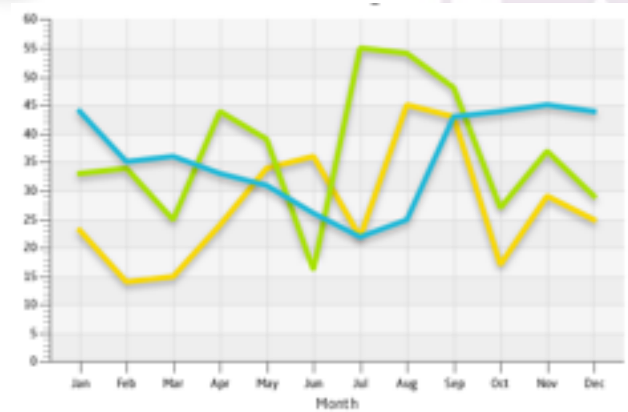
$$VDM(x,y) = \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q$$

[Wilson & Martines, JAIR'97]

"China is like India more than Australia, since they both have large population."

Common similarity functions

Similarity for time series data:

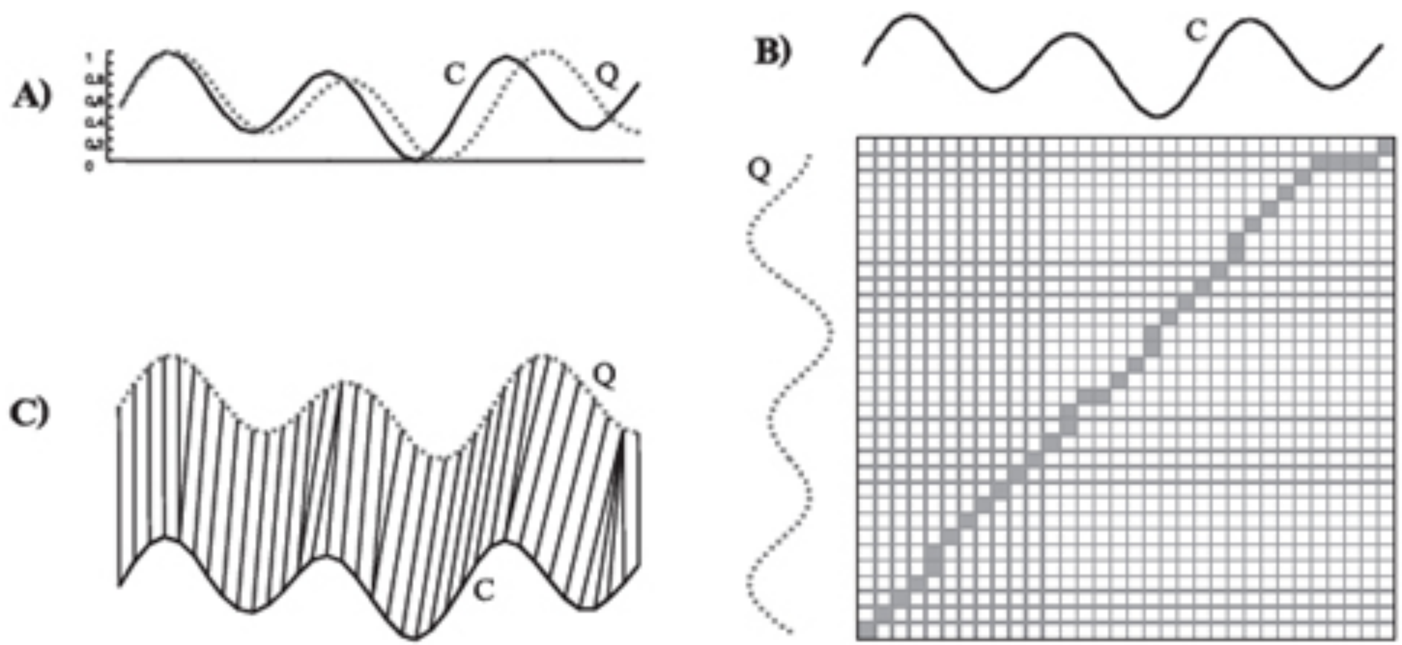


Dynamic Time Wrapping (DTW): minimize the sum of distances of the matched points

$x_1, x_2, \ldots, x_n$

$y_1, y_2, \ldots, y_m$

$d(x_i, y_j)$



$$d(X,Y) = \sum_{i=1}^{T} d(x_{\phi_{i,x}}, y_{\phi_{i,y}})$$ minimize -> dynamic programming

pic from http://www.ibrahimkivanc.com/post/Dynamic-Time-Warping.aspx

Why visualization

Data visualization is an important way for identifying deep relationship

- Pros
  - straight-forward
  - usually interactive
  - ideal for sifting through data to find unexpected relation
- Cons
  - requires special people to read the results to find unexpected relation
  - might not be good for large data sets, too many details may shade the interesting patterns

- The brain processes visual information 60,000 times faster than text.
- 90 percent of information that comes to the brain is visual.
- 40 percent of all nerve fibers connected to the brain are linked to the retina.

@DATA

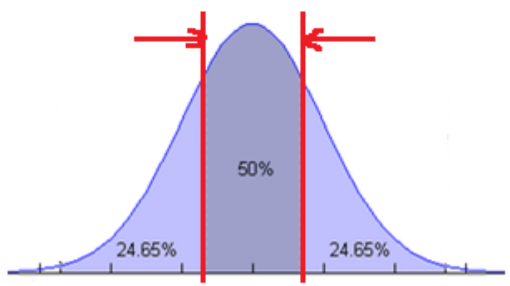
october, normal, gt-norm, norm, yes, same-lst-yr, low-areas, pot-severe, none, 90-100, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, no, above-sec-nde, brown, present, firm-and-dry, absent, none, absent, norm, dna, norm, absent, absent, norm, absent, norm, diaporthe-stem-canker
august, normal, gt-norm, norm, yes, same-lst-two-yrs, scattered, severe, fungicide, 80-89, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, yes, above-sec-nde, brown, present, firm-and-dry, absent, none, absent, norm, dna, norm, absent, absent, norm, absent, norm, diaporthe-stem-canker
july, normal, gt-norm, norm, yes, same-lst-yr, scattered, severe, fungicide, lt-80, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, yes, above-sec-nde, dna, present, firm-and-dry, absent, none, absent, norm, dna, norm, absent, absent, norm, absent, norm, diaporthe-stem-canker
july, normal, gt-norm, norm, yes, same-lst-yr, scattered, severe, none, 80-89, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, yes, above-sec-nde, dna, present, firm-and-dry, absent, none, absent, norm, dna, norm, absent, absent, norm, absent, norm, diaporthe-stem-canker
october, normal, gt-norm, norm, yes, same-lst-two-yrs, scattered, pot-severe, none, lt-80, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, yes, above-sec-nde, brown, present, firm-and-dry, absent, none, absent, norm, dna, norm, absent, absent, norm, absent, norm, diaporthe-stem-canker
september, normal, gt-norm, norm, yes, same-lst-sev-yrs, scattered, pot-severe, none, 80-89, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, yes, above-sec-nde, dna, present, firm-and-dry, absent [illegible]one, absent, norm, d[illegible] norm, absent, absent, norm, absent, norm, diaporthe-stem-canker
september, normal, gt-norm, norm, yes, same-lst-two-y[illegible]attered, pot-sev[illegible]ngicide, 90-100, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, no, above-sec-nde, brown, present, firm-and-dr[illegible]nt, none, [illegible], norm, dna, norm, absent, absent, norm, absent, norm, diaporthe-stem-canker
august, normal, gt-norm, norm, no, same-lst-yr, scattered, p[illegible]ere, [illegible]t-80, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, yes, above-sec-nde, brow[illegible]t, fir[illegible]-and-dry, [illegible]bsent, no[illegible]orm, a[illegible]ent, absent, norm, a[illegible]ent, norm, d[illegible]the-stem-canker
october, normal, [illegible]ent, absent, absent, abnorm, yes, above-sec-[illegible]he-stem-canker
august, normal, gt-norm, norm, yes, same-lst-two-yrs, scat[illegible]severe, [illegible]lt-80, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, yes, above-sec-nde, brown, present, firm-and-dry, absent, n[illegible]sent, norm, [illegible]orm, absent, absent, norm, absent, norm, diaporthe-stem-canker
october, normal, lt-norm, gt-norm, yes, same-lst-yr, w[illegible]eld, pot-severe, f[illegible]de, 90-100, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, yes, absent, tan, absent, absent, absent, black, present, [illegible]orm, dna, norm, absen[illegible] absent, norm, absent, norm, charcoal-rot
august, normal, lt-norm, norm, no, same-lst-yr, whole-field, pot-severe, fungicide, 80-89, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, no, absent, tan, absent, absent, absent, black, present, norm, dna, norm, absent, absent, norm, absent, norm, charcoal-rot
july, normal, lt-norm, norm, yes, same-lst-yr, upper-areas, pot-severe, none, 90-100, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, yes, absent, tan, absent, absent, absent, black, present, norm, dna, norm, absent, absent, norm, absent, norm, charcoal-rot
october, normal, lt-norm, norm, no, same-lst-sev-yrs, whole-field, pot-severe, fungicide, 90-100, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, yes, absent, tan, absent, absent, absent, black, present, norm, dna, norm, absent, absent, norm, absent, norm, charcoal-rot
october, normal, lt-norm, gt-norm, yes, same-lst-yr, whole-field, pot-severe, fungicide, 80-89, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, yes, absent, tan, absent, absent, absent, black, present, norm, dna, norm, absent, absent, norm, absent, norm, charcoal-rot
september, normal, lt-norm, gt-norm, no, same-lst-sev-yrs, whole-field, pot-severe, fungicide, lt-80, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, yes, absent, tan, absent, absent, absent, black, present, norm, dna, norm, absent, absent, norm, absent, norm, charcoal-rot
october, normal, lt-norm, gt-norm, no, diff-lst-year, upper-areas, pot-severe, none, 90-100, abnorm, abnorm, absent, dna, dna, absent, absent, absent, abnorm, no,

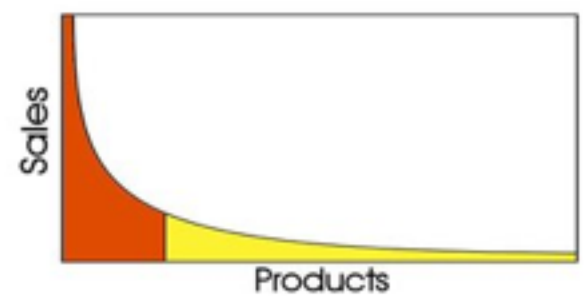
This is NOT visualization

What to visualize

- Displaying single attribute/property

  mean, median, quartile, percentile, mode, variance, interquartile range, skewness



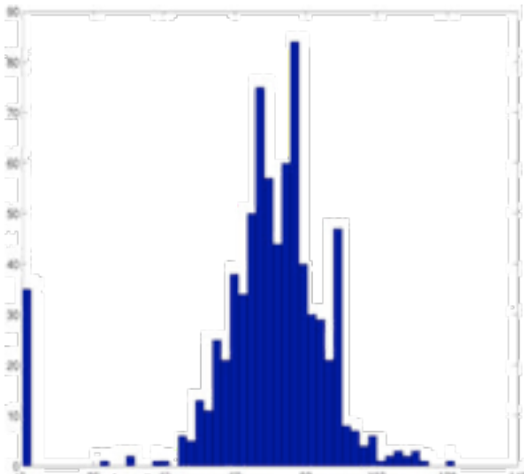


- Displaying the relationships between two attributes

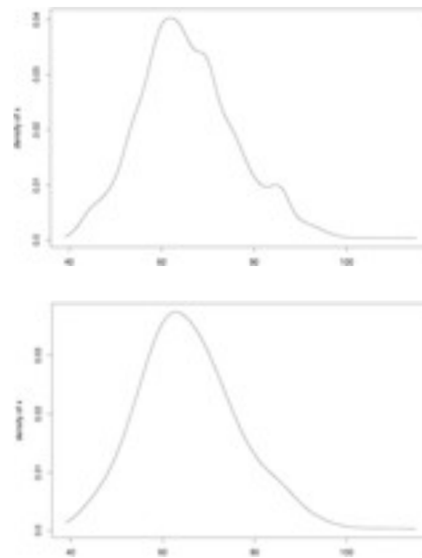
- Displaying the relationships between multiple attributes

- Displaying important structure of data in a reduced number of dimensions

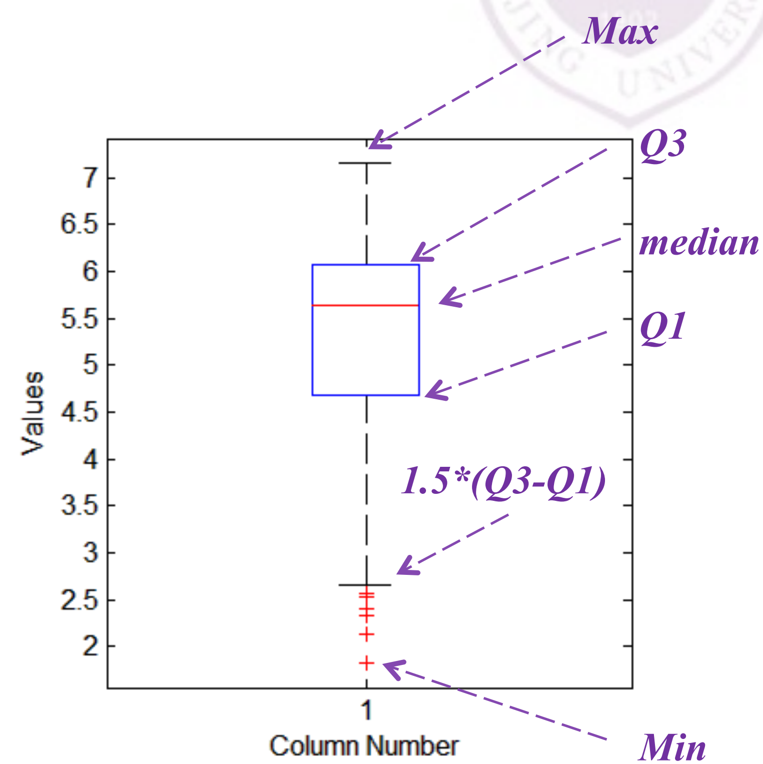
Displaying single attribute



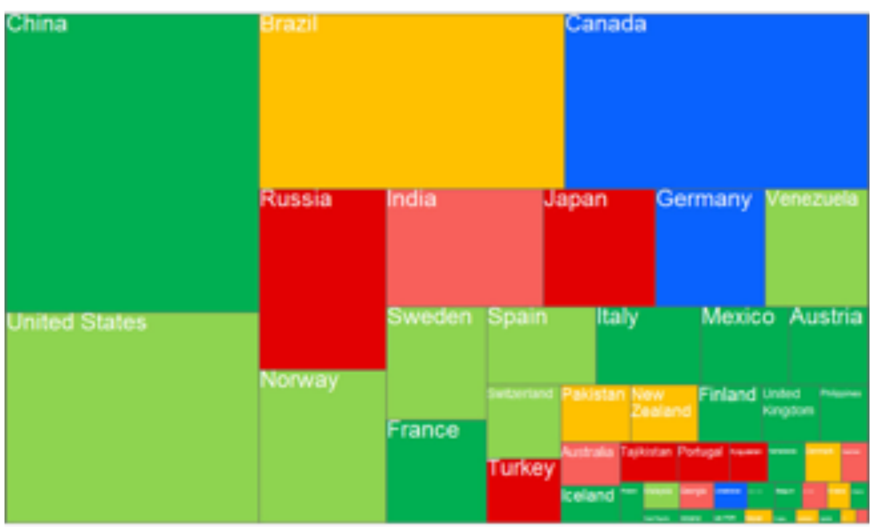
histogram



density

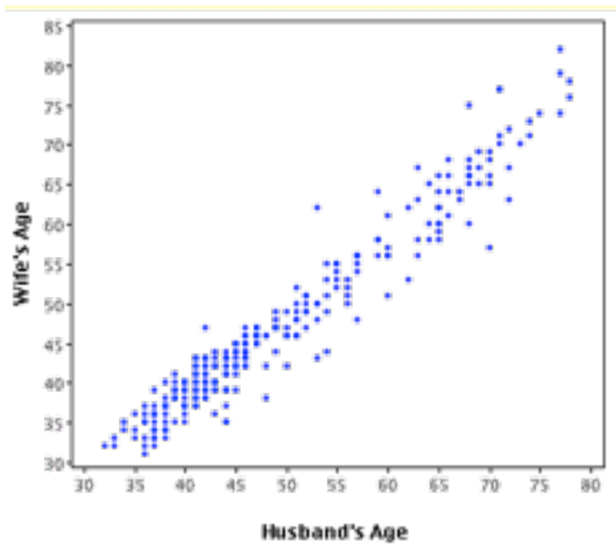


box plots

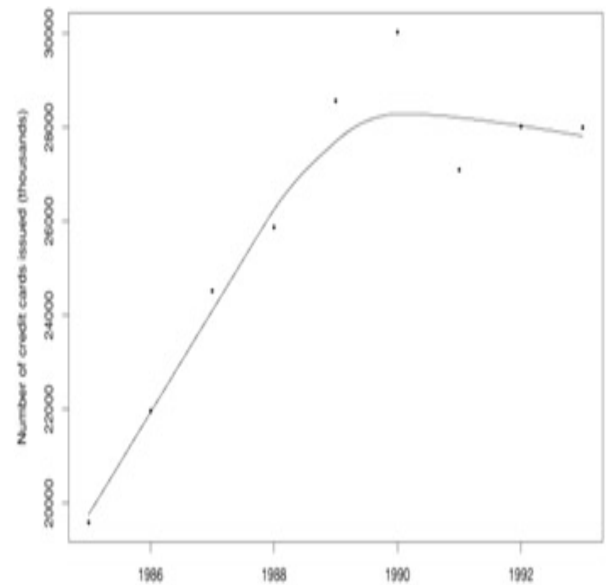


treemap

Displaying pair of attributes

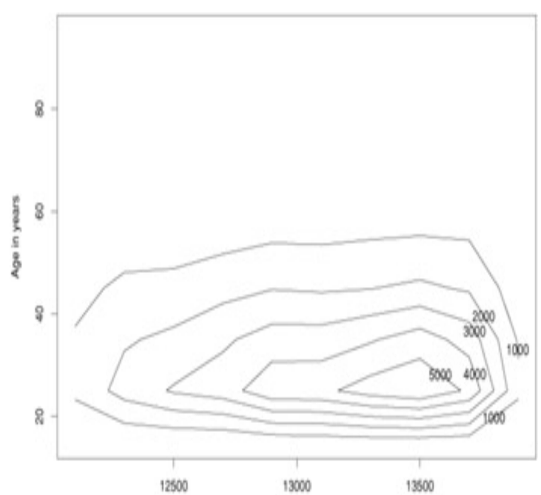


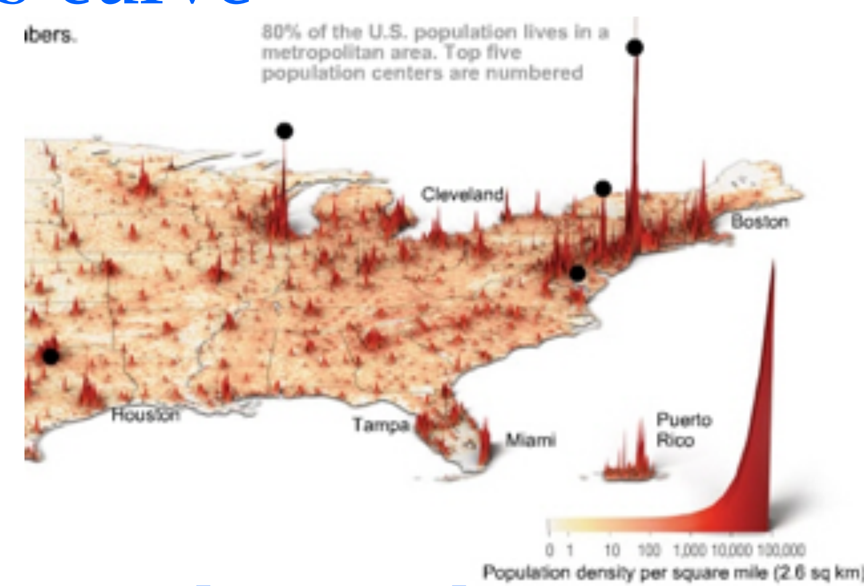
Scatter plot



loess curve

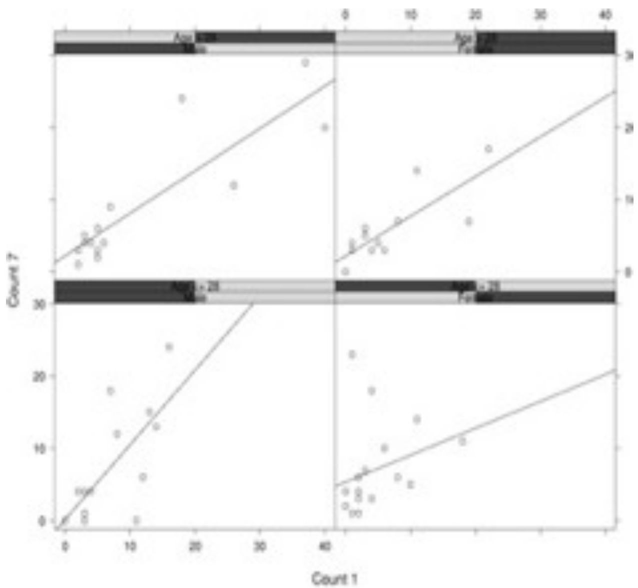
contour plot



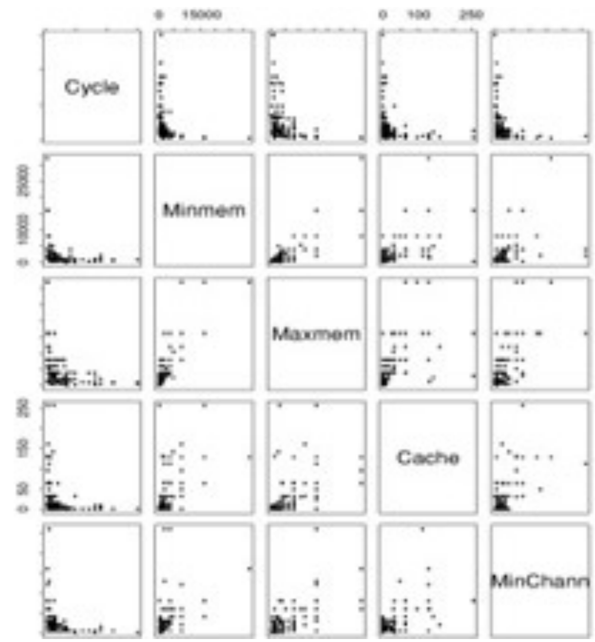


particular application

Displaying multiple attributes

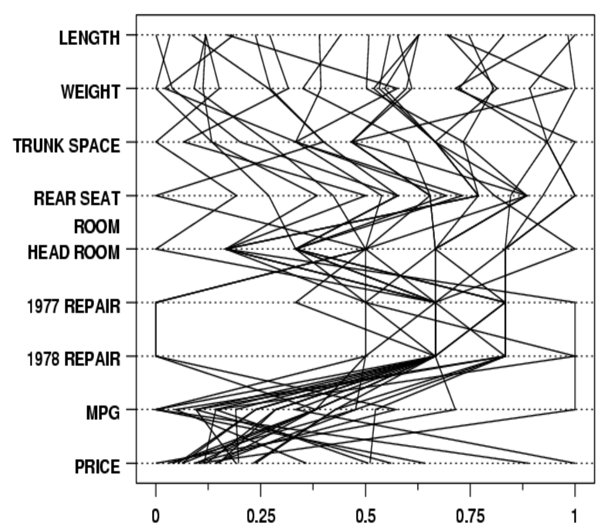


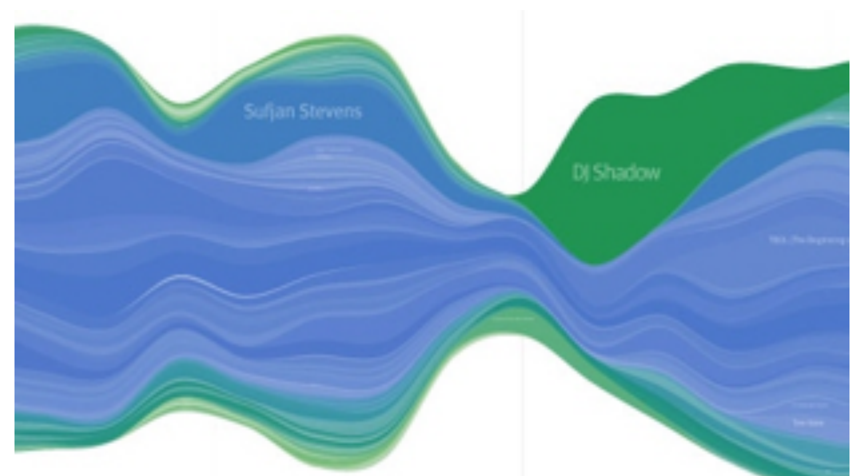
trellis plot (conditional scatter plot)



scatterplot matrix

parallel coordinates plot



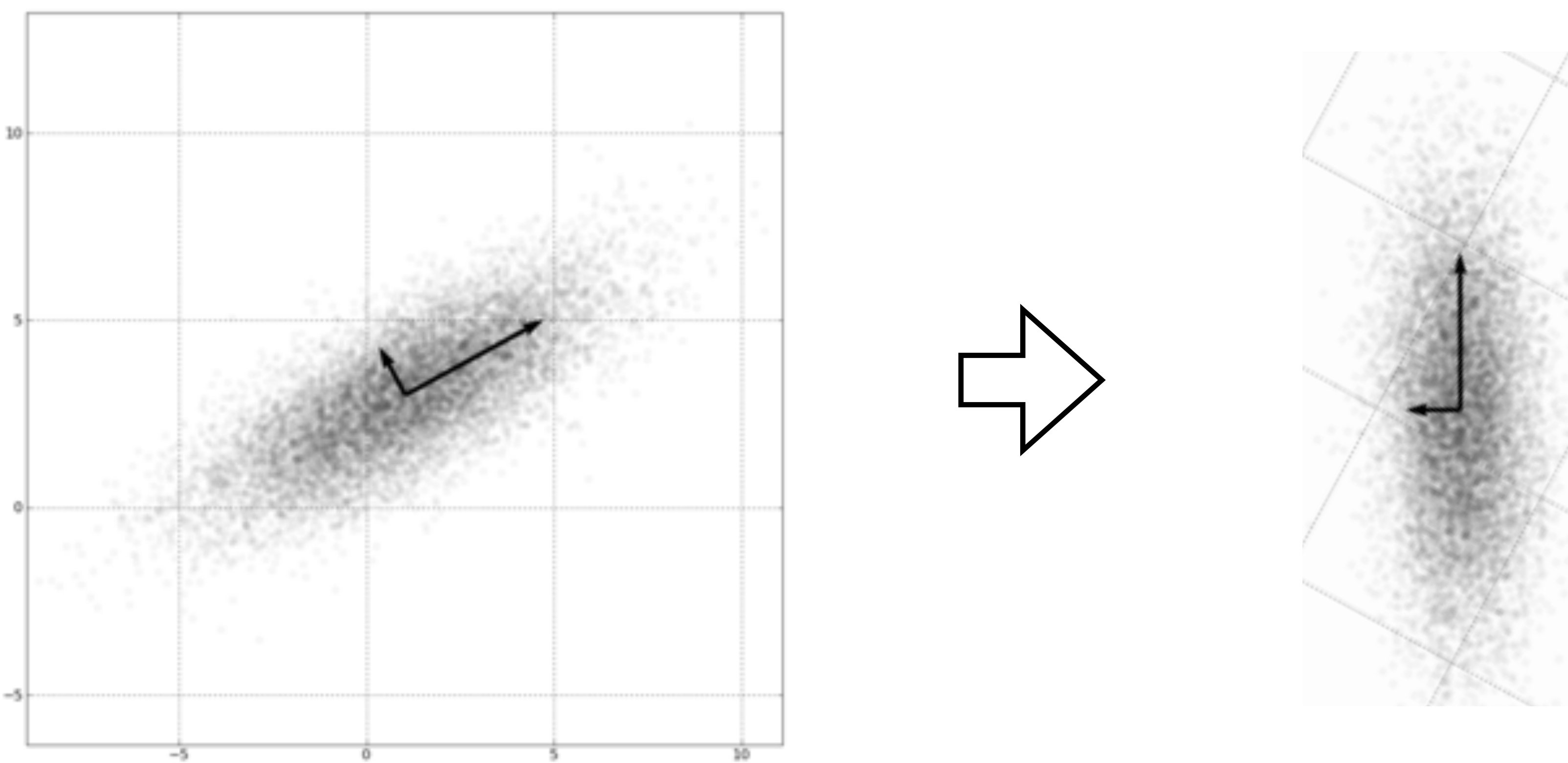


time series

Displaying multiple attributes

Dimension reduction

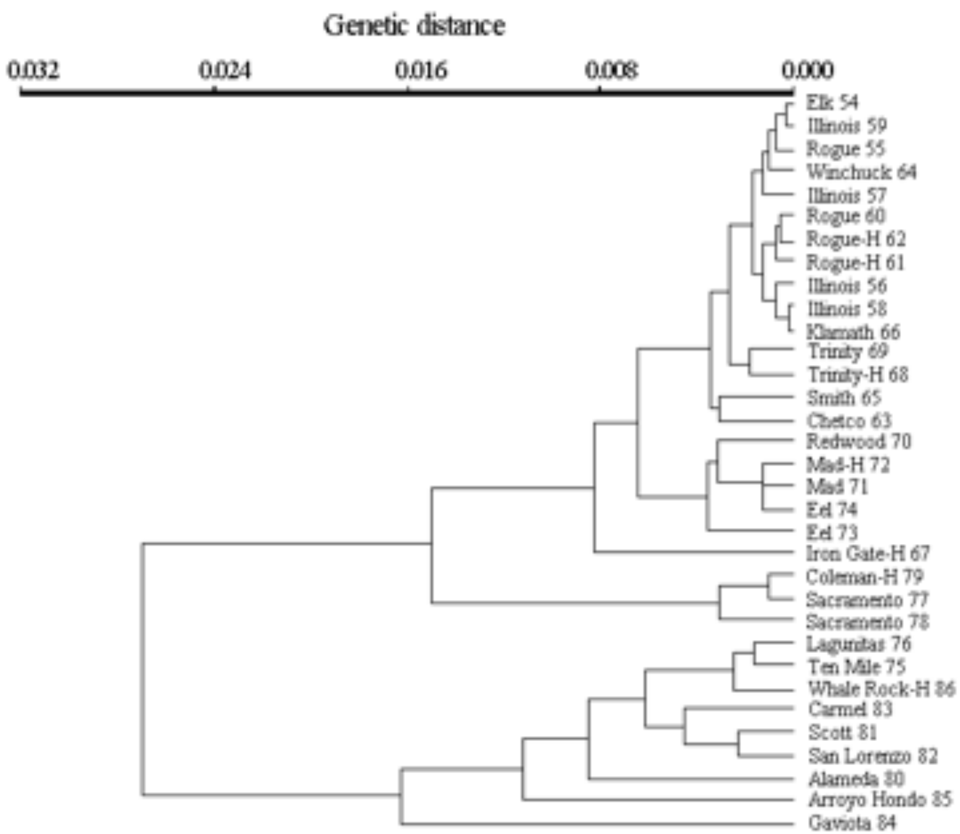
- Principle Component Analysis (PCA)

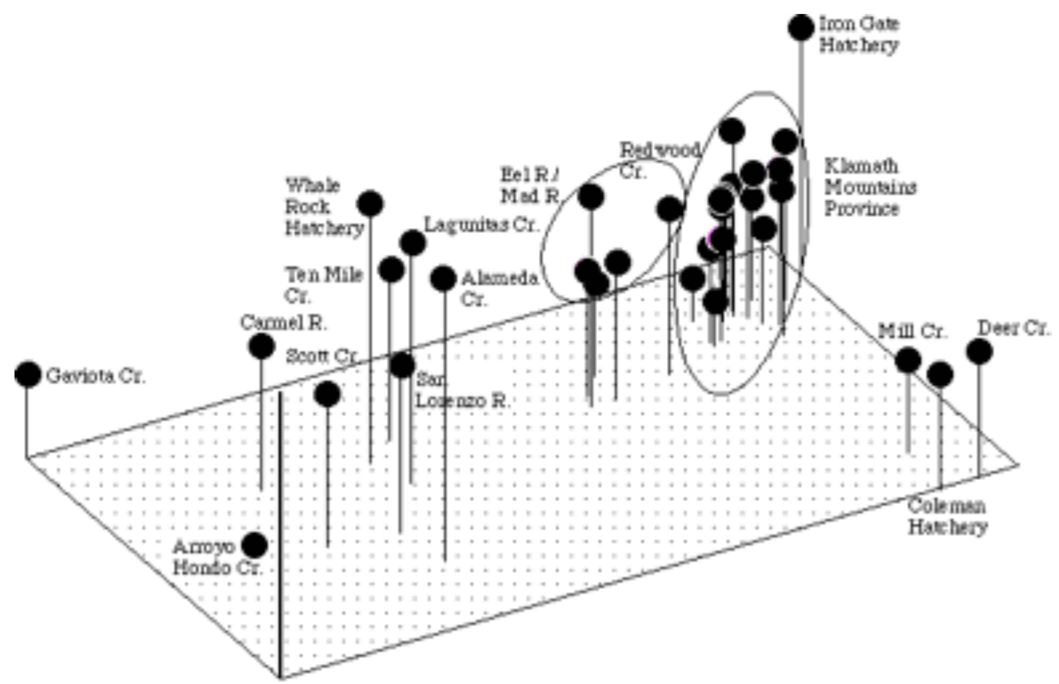


Displaying multiple attributes

Dimension reduction

- Multi-dimensional Scaling (MDS)





pic from http://www.nwfsc.noaa.gov/publications/techmemos/

Displaying multiple attributes

Dimension reduction

- Manifold learning

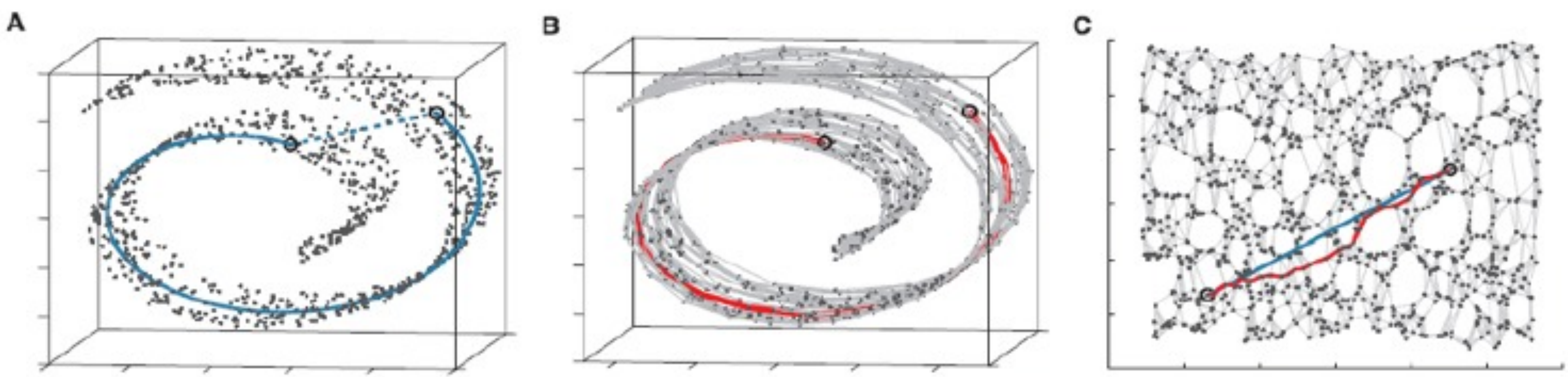
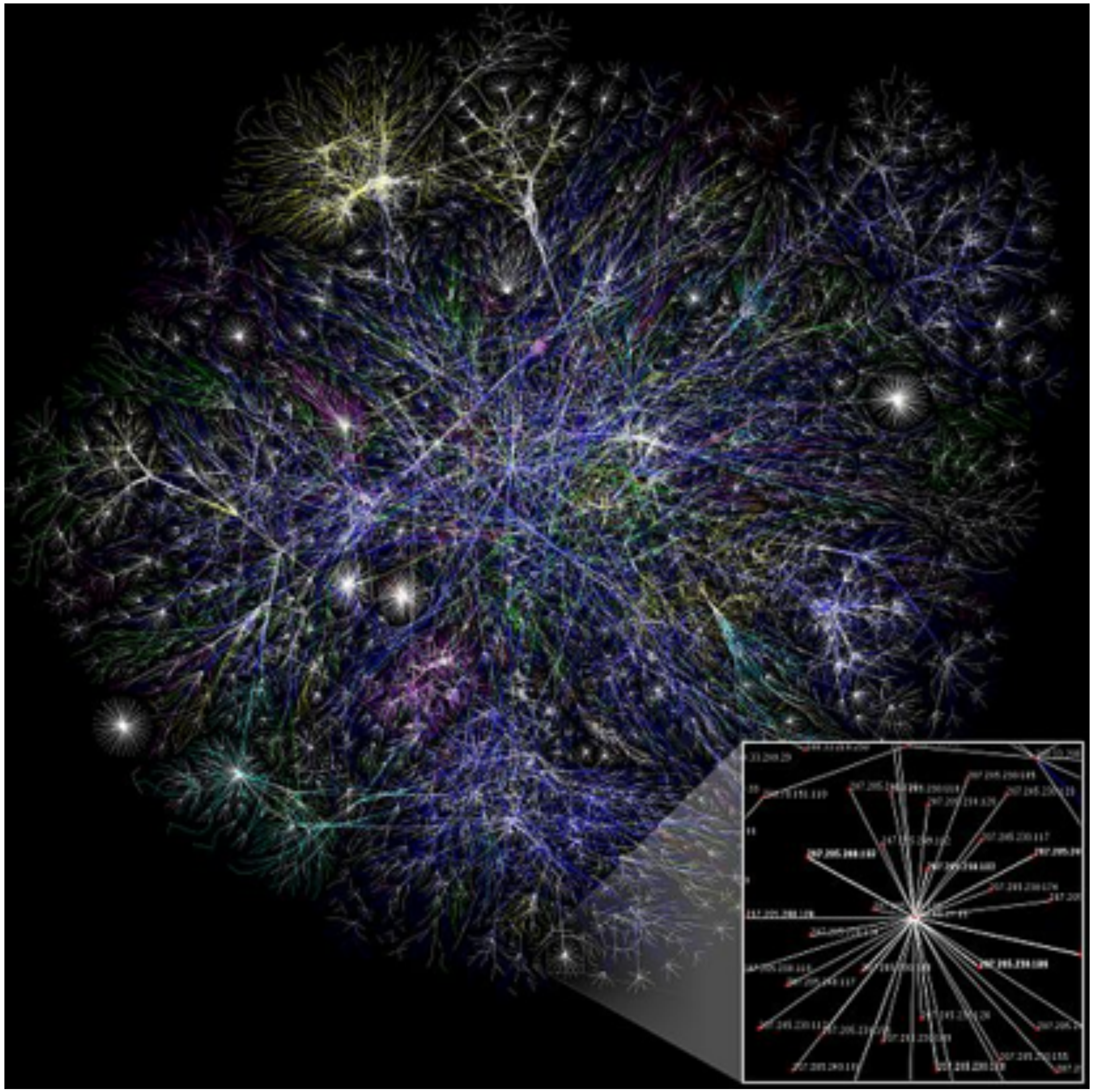
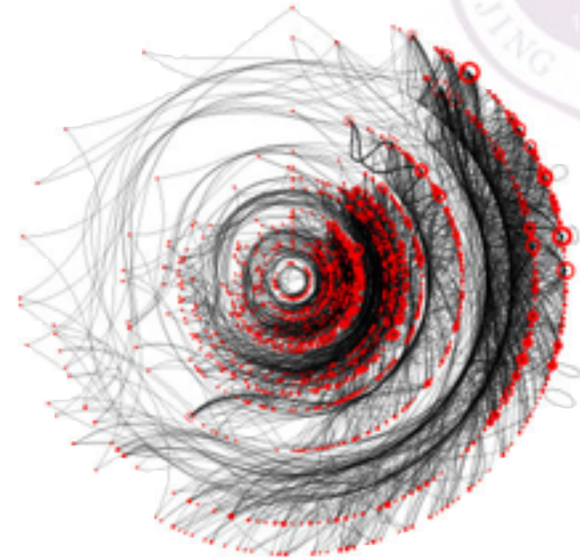


Fig. 3. The "Swiss roll" data set, illustrating how Isomap exploits geodesic paths for nonlinear dimensionality reduction. (**A**) For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high-dimensional input space (length of dashed line) may not accurately reflect their intrinsic similarity, as measured by geodesic distance along the low-dimensional manifold (length of solid curve). (**B**) The neighborhood graph G constructed in step one of Isomap (with $K = 7$ and $N =$ 1000 data points) allows an approximation (red segments) to the true geodesic path to be computed efficiently in step two, as the shortest path in G. (**C**) The two-dimensional embedding recovered by Isomap in step three, which best preserves the shortest path distances in the neighborhood graph (overlaid). Straight lines in the embedding (blue) now represent simpler and cleaner approximations to the true geodesic paths than do the corresponding graph paths (red).

www.sciencemag.org SCIENCE VOL 290 22 DECEMBER 2000

Displaying link relationship





pic from http://www.smashingmagazine.com/2007/08/02/data-visualization-modern-approaches/

习题

min-max和z-score规范化谁会有数据出界的风险?

基于信息熵(entropy)的离散化方法是否需要监督信息(supervised or unsupervised)?

当p=0.5时Minkowski距离 $\left(\sum_{i=1}^{n}|x_i-y_i|^{0.5}\right)^2$ 是否仍然是距离(distance)?