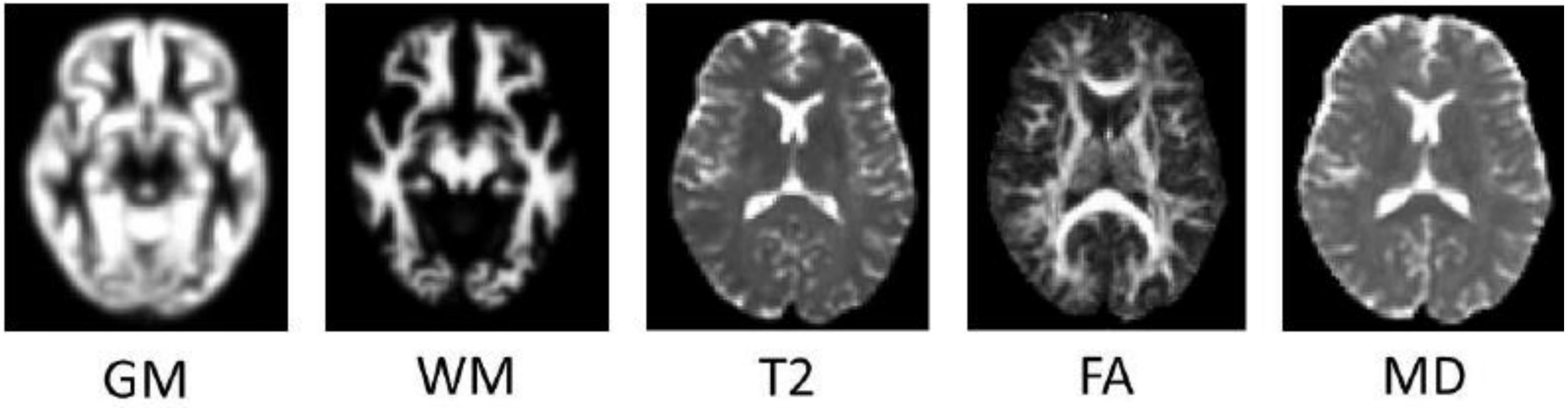
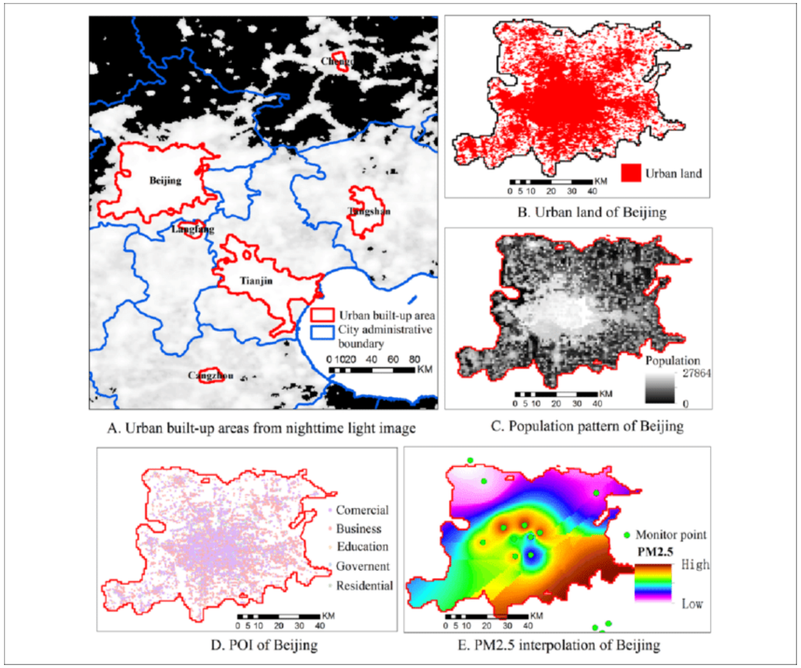
https://www.researchgate.net/publication/234113968_Probabilistic_prediction_of_neurological_disorders_with_a_statistical_assessment_of_neuroimaing_data_modalities/figures?lo=1&utm_source=yahoo&utm_medium=organic



3. Data preprocessing



A. Urban built-up areas from nighttime light image

B. Urban land of Beijing

C. Population pattern of Beijing

D. POI of Beijing

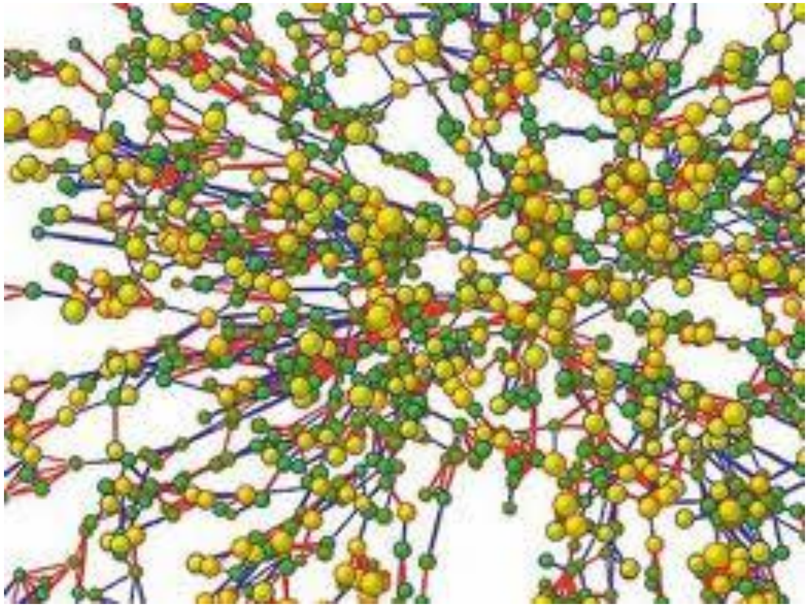
E. PM2.5 interpolation of Beijing

https://www.researchgate.net/publication/317160774_Exploring_the_Association_between_Urban_Form_and_Air_Quality_in_China/figures?lo=1

https://software.intel.com/

Data preprocessing

- Displaying raw data = precise, identification of outliers, missing data, …
- Sometimes preprocessing is required



blog.wpfwonderland.com

Preprocessing – techniques

- Metadata and statistics
- Missing values and data “cleaning“
- Normalization
- Segmentation
- Sampling and interpolation
- Dimension reduction
- Data aggregation
- Smoothing and filtration
- Raster to vector

Metadata and statistics

- Metadata – information for preprocessing
  - Reference point for measurement
  - Unit of measurement
  - Symbol for missing values
  - Resolution
- Statistical analysis
  - Detection of missing records
  - Cluster analysis
  - Correlation analysis

If you had two cans without labels, which would you eat?

Without a label, how would you know which was tuna and which was cat food?

Missing values and data “cleaning“

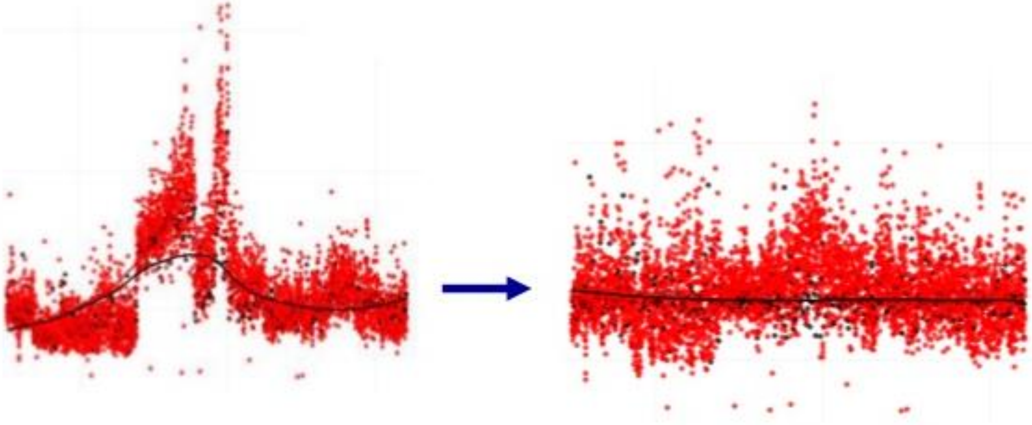
- Removing wrong records
- Assigning a given value
- Assigning an average value
- Assigning a value derived from the nearest neighbor value
- Calculating the value (imputation)

Normalization

- Transformation of the input dataset
- Adjusting values measured on different scales to a notionally common scale
- Normalization to interval [0.0, 1.0]:

$$d_{normalized} = (d_{original} - d_{min}) / (d_{max} - d_{min})$$

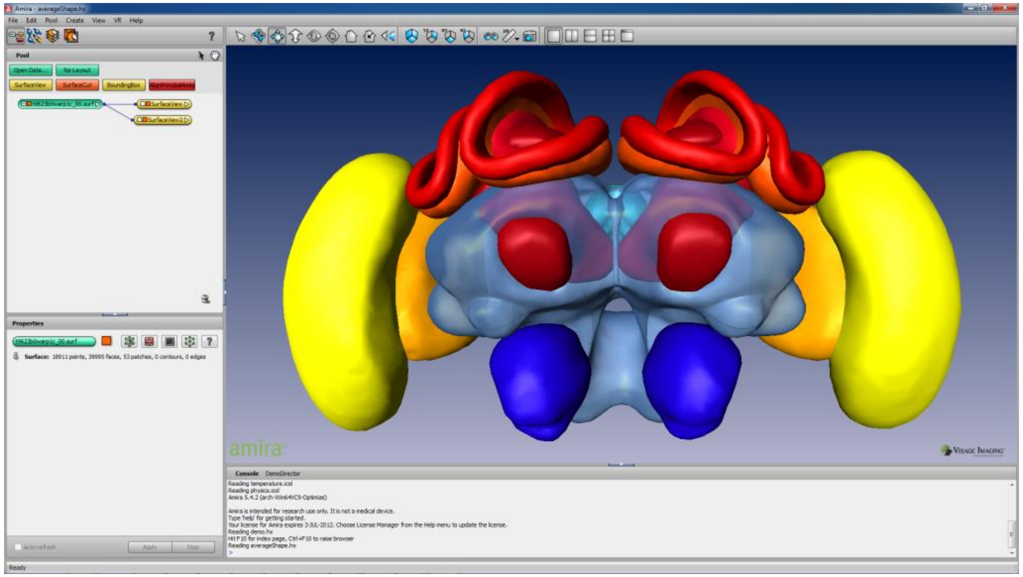
- Clamping according to the threshold values



Segmentation

- Classification of input data into given categories

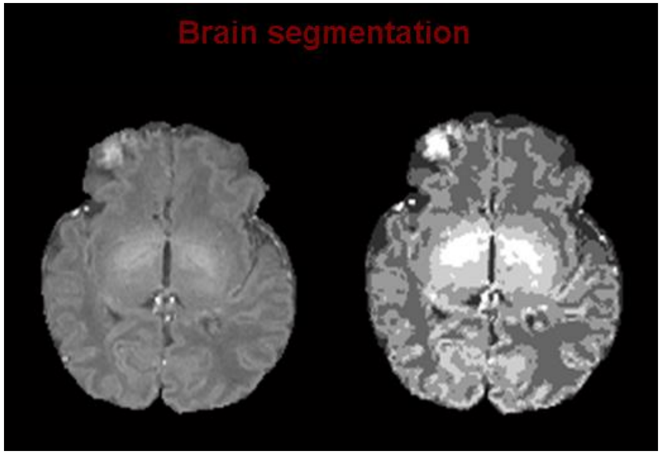
- Split-and-merge
  iterative algorithm



Surface: 18911 points, 39995 faces, 53 patches, 0 contours, 0 edges

```
Reading temperature.icol
Reading physics.icol
Amira 5.4.2 (arch-Win64VC9-Optimize)

Amira is intended for research use only. It is not a medical device.
Type 'help' for getting started.
Your license for Amira expires 3-JUL-2012. Choose License Manager from the Help menu to update the license.
Reading demo.hx
Hit F10 for index page, Ctrl+F10 to raise browser
Reading averageShape.hx
>
```



blog.campaigner.com

Split-and-merge

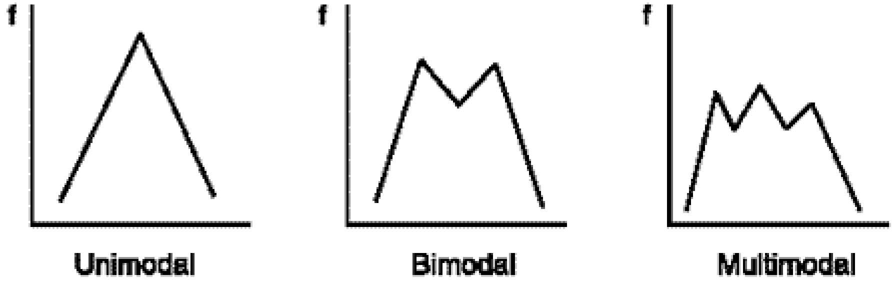
- similarThresh = defines the similarity of two regions with given characteristics
- homogeneousThresh = defines the region homogeneity (uniformity)

```
do {
        changeCount = 0;
        for each region {
                compare region with neighboring ones and find the most similar one;
                if the most similar one is within similarThresh of the current region {
                        connect these two regions;
                        changeCount++;
                }
                evaluate the homogeneity of the region;
                if homogeneity of region is smaller than homogeneousThresh  {
                        split the region to two parts;
                        changeCount++;

        }
} until changeCount  == 0
```

Complex parts of the algorithm

- Determining the similarity of two regions
- Evaluating the homogeneity of a region – histogram



www.statcan.gc.ca

- Splitting the region

Possible problem

- Infinite loop by repeating split and merge steps of the same region

- Solution:
  - Changing the threshold value for similarity or homogeneity
  - Taking into account other region properties (e.g., size and shape of regions)

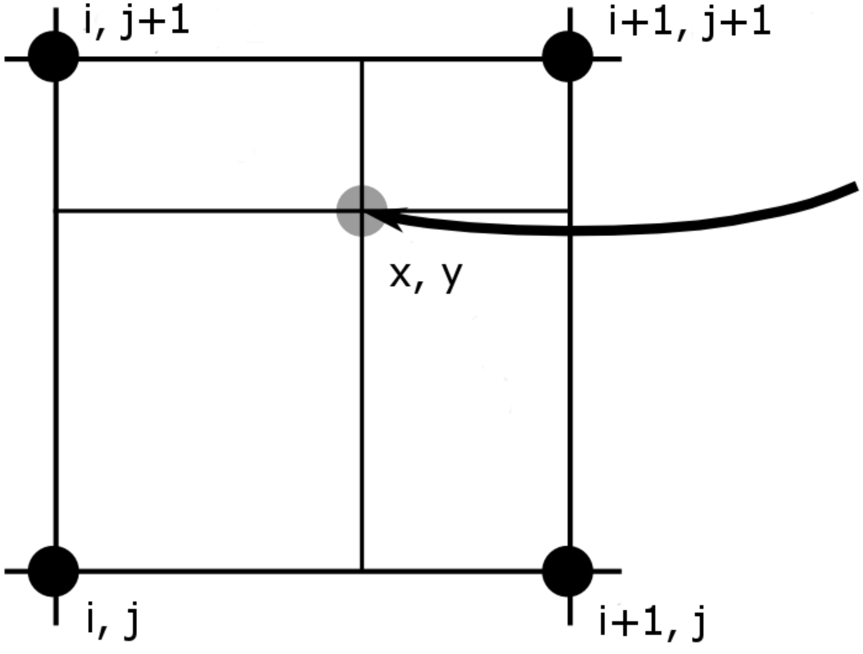
Sampling and interpolation

- Transformation of input data
- Interpolation = sampling method
  - Linear interpolation
  - Bilinear interpolation
  - Non-linear interpolation

inperc.com

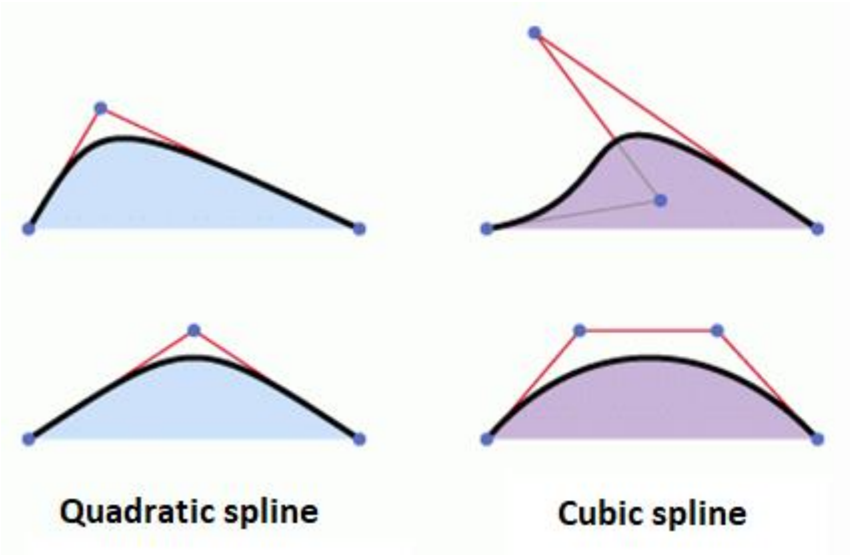
Bilinear interpolation

- Uniform grid
- Horizontal + vertical interpolation



Non-linear interpolation

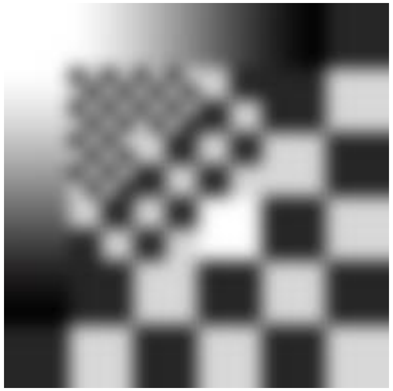
- Problems with linear interpolation – zero continuity in grid points
- Solution = using quadratic and cubic splines



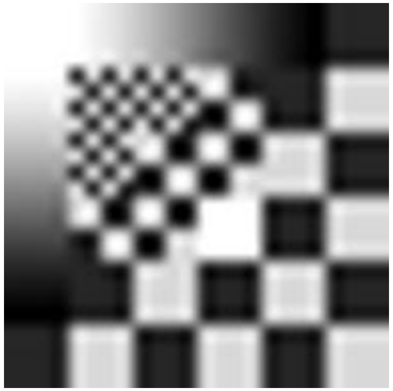
Result

- Original image (24x24 pixels)

cubic B-spline filter



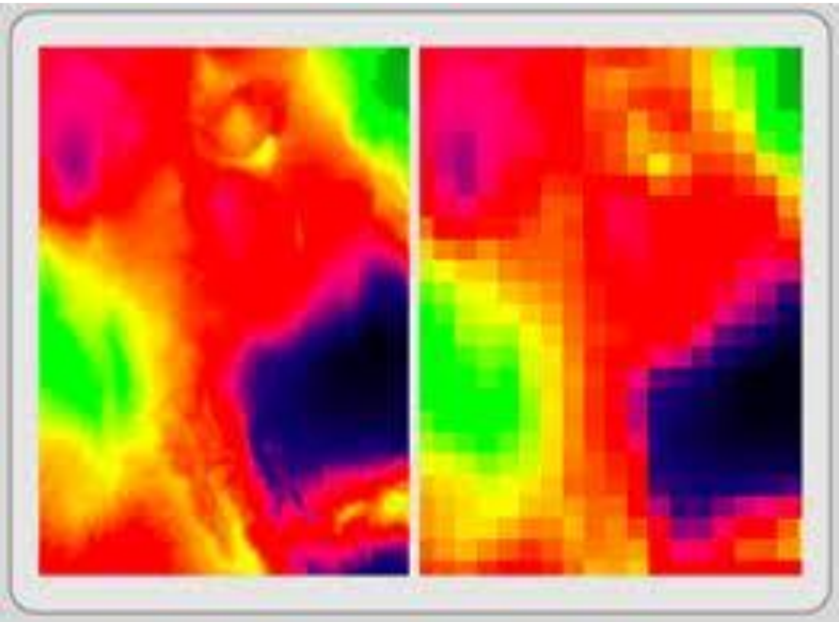
Catmull-Rom



research.cs.wisc.edu

Resampling

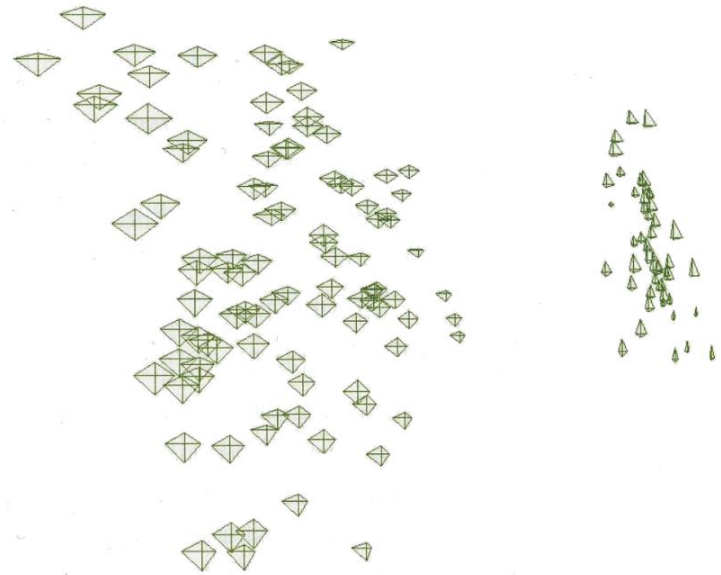
- Pixel replication
- Neighbor averaging
- Data subsetting



giscommons.org

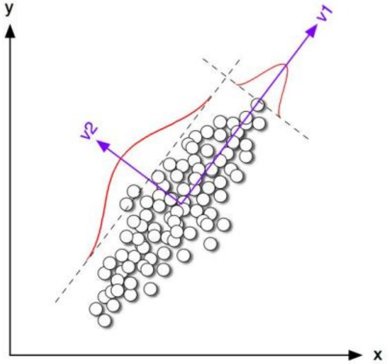
Dimension reduction

- Preparing multidimensional data for displaying
- Keep as much original information as possible
- Techniques:
  - **PCA** (principal component analysis)
  - **MDS** (multidimensional scaling)
  - **SOMs** (Kohonen self-organizing maps)



Interactive Data Visualization - Fondations, Techniques and Applications. Matthew Ward

PCA intuitively

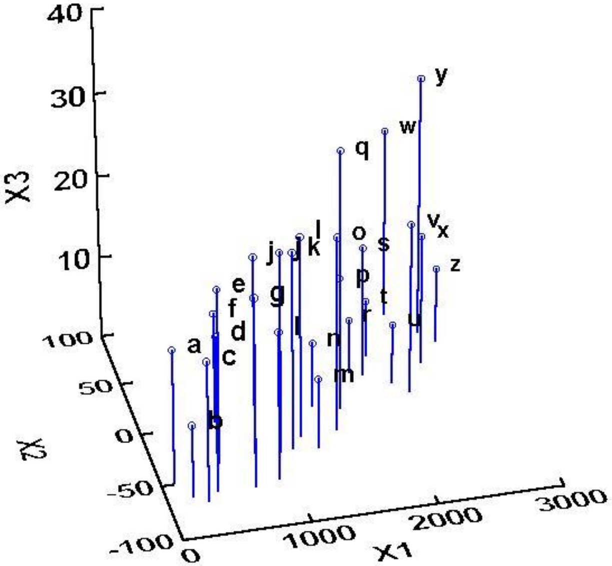


web.media.mit.edu

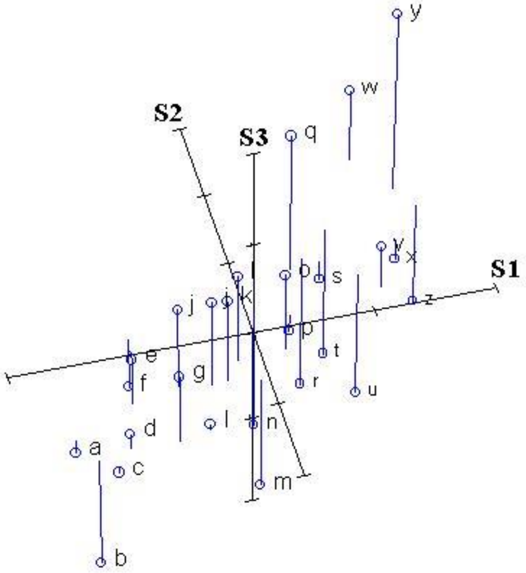
1. We select a line in space visualizing n-dimensional data. This line covers the most of the input data items and is called the first principal component (PC).
2. We select a second line perpendicular to the first PC, this forms the second PC.
3. We repeat this until we proces all PC dimensions or until we reach a desired number of principle components.

PCA – principal component analysis

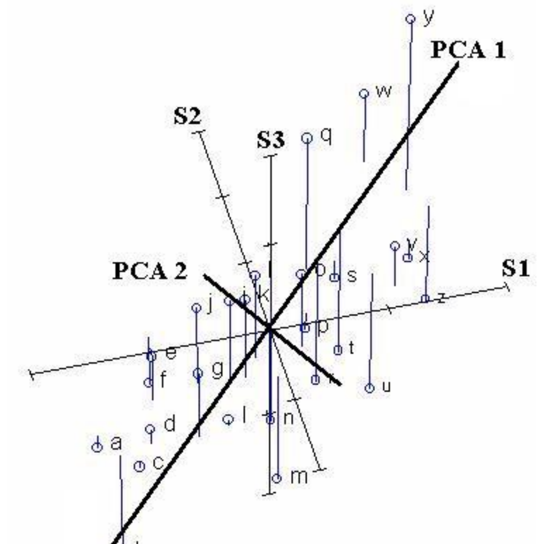
1)



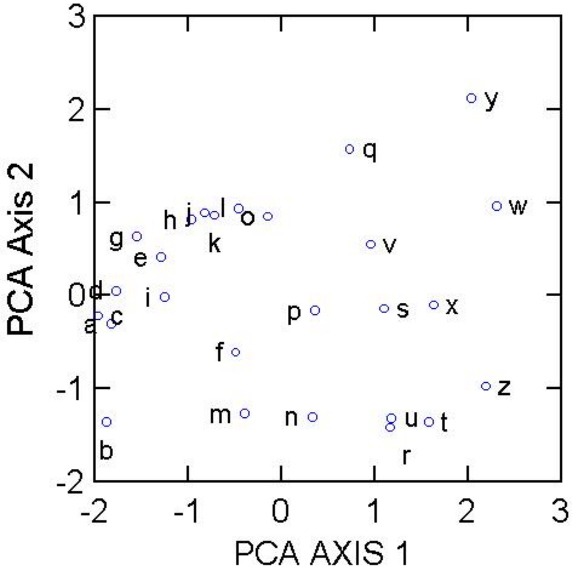
2)



3)



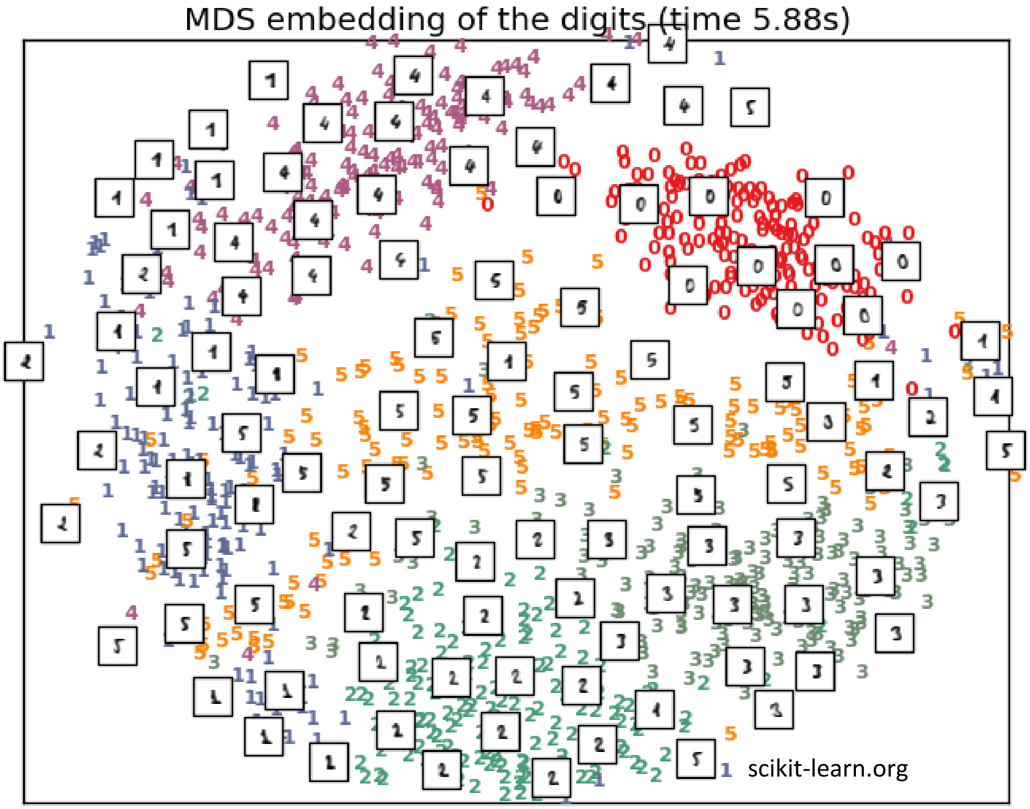
4)



http://ordination.okstate.edu/PCA.htm

MDS – multidimensional scaling

- Based on comparing the distances between individual data items in original and reduced space

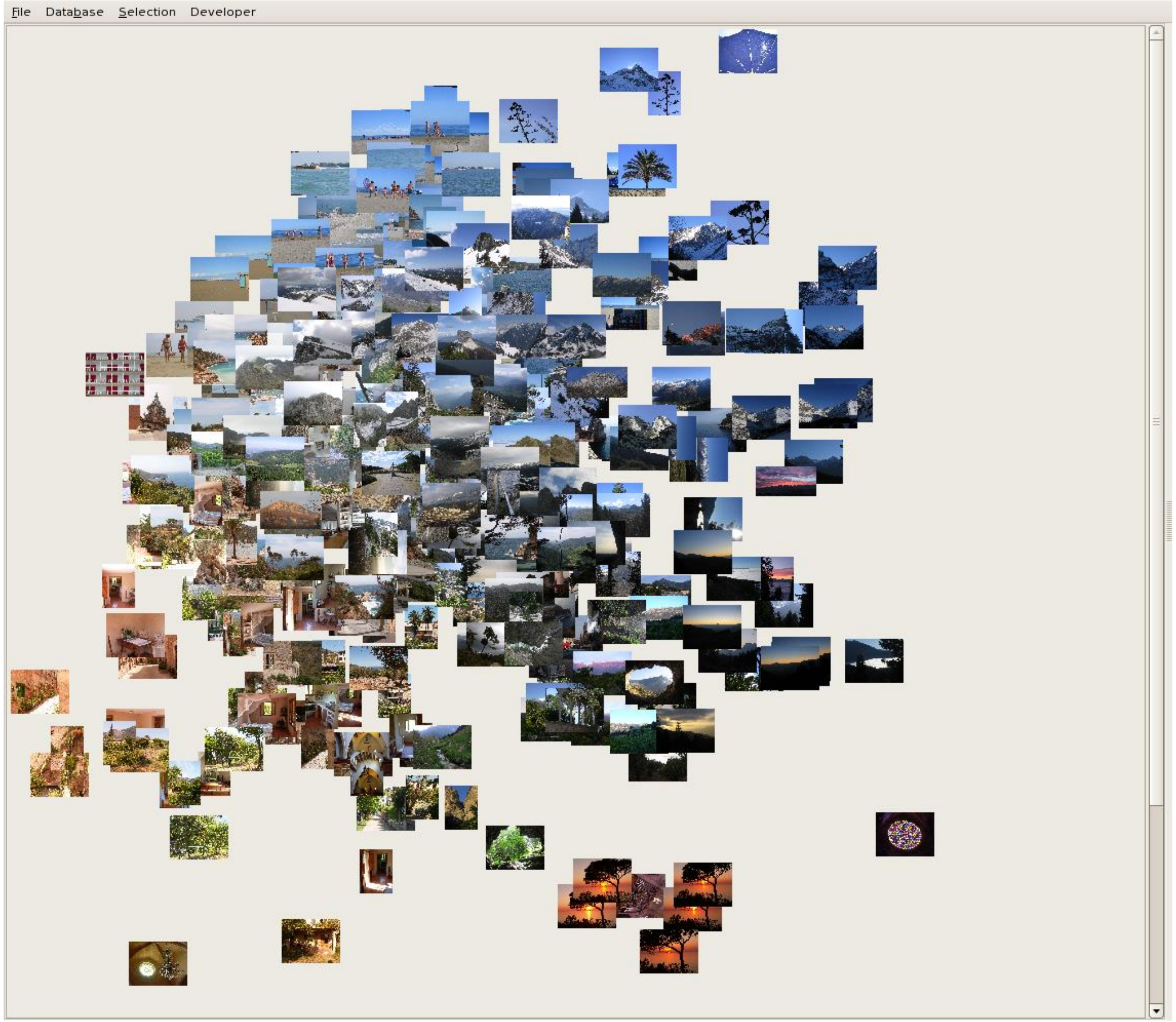

MDS embedding of the digits (time 5.88s)

scikit-learn.org

MDS – multidimensional scaling

1) We calculate the distances between all pairs of data points in the original space. If we have *n* points as an input, this step requires n(n – 1)/2 operations.
2) We transfer all input data points to points in the reduced dimension space (often randomly).
3) We calculate *stress*, i.e., difference in distance between points in the original and reduced space. This can be done using different approaches.
4) If the average and cummulated *stress* value is smaller than the user-defined threshold, the algorithm ends and returns the result..
5) If the *stress* value is higher than the threshold, for each point we calculate a directional vector pointing to the desired shift direction in order to reduce *stress* between this point and the other points. This is determined as the weighted average of vectors between this point and its neighbors and its weight is derived from *stress* value calculated between individual pairs. Positive *stress* value repulses the points, negative one attracts them. The higher the absolute value of *stress,* the bigger movement of point.
6) Based on these calculations we transform tha data points to the target reduced dimension, according to the calculated vectors. Return to step 3 of the algorithm.

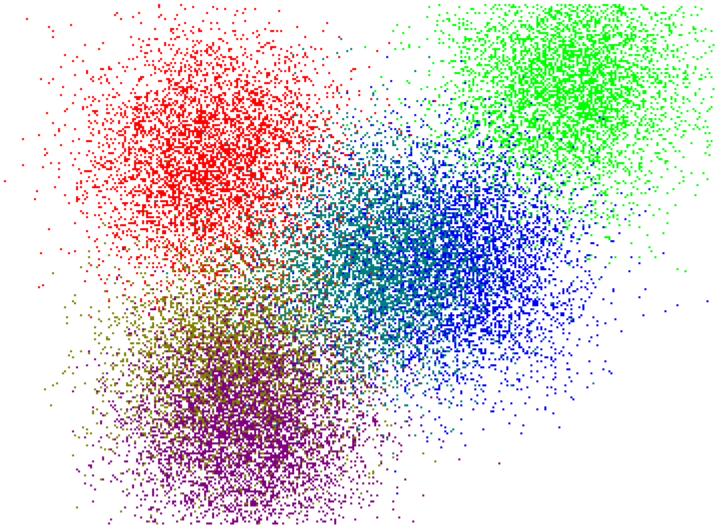
MDS – multidimensional scaling

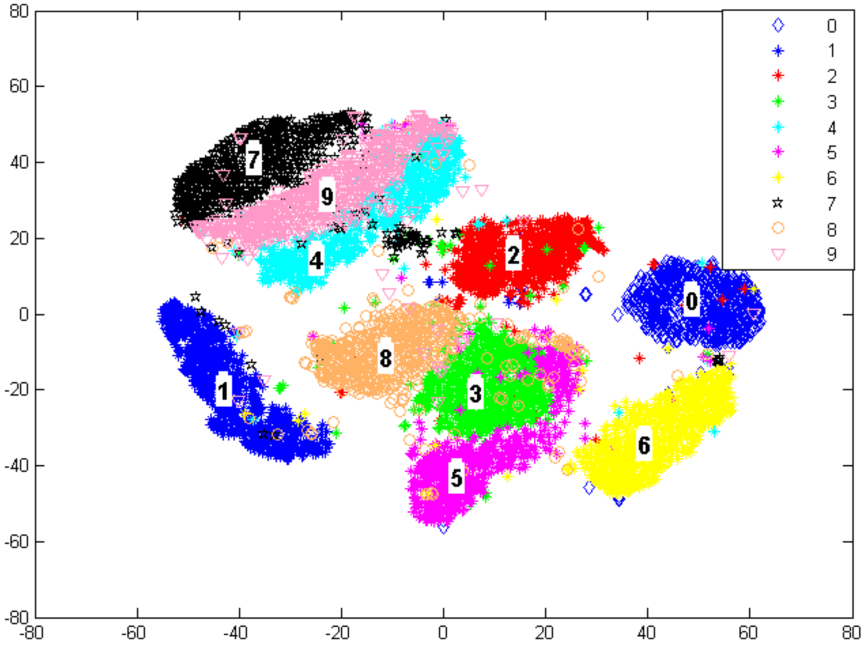


lear.inrialpes.fr/src/yorg/doc/index.html

Data aggregation

- Aggregation = clustering of similar data to groups.





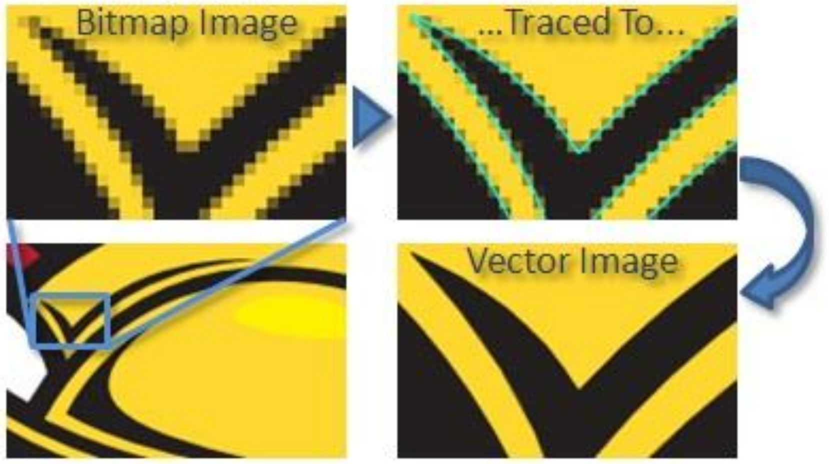
Smoothing and filtration

- Signal processing techniques – noise removal
- **Convolution** in 1D:

$$p_i = \frac{p_{i-1}}{4} + \frac{p_i}{2} + \frac{p_{i+1}}{4}$$

Converting rasters to vectors

- Used for:
  - Data compression
  - Image comparison
  - Data transformation
- Methods:
  - Thresholding
  - Region growing
  - Edge detection
  - …



Conclusion

- The techniques mentioned improve the efficiency of visualization
- We have to inform the user that the data has been transformed!!!