CS109 – Data Science

Verena Kaynig-Fittkau

vkaynig@seas.harvard.edu
staff@cs109.org

Announcements

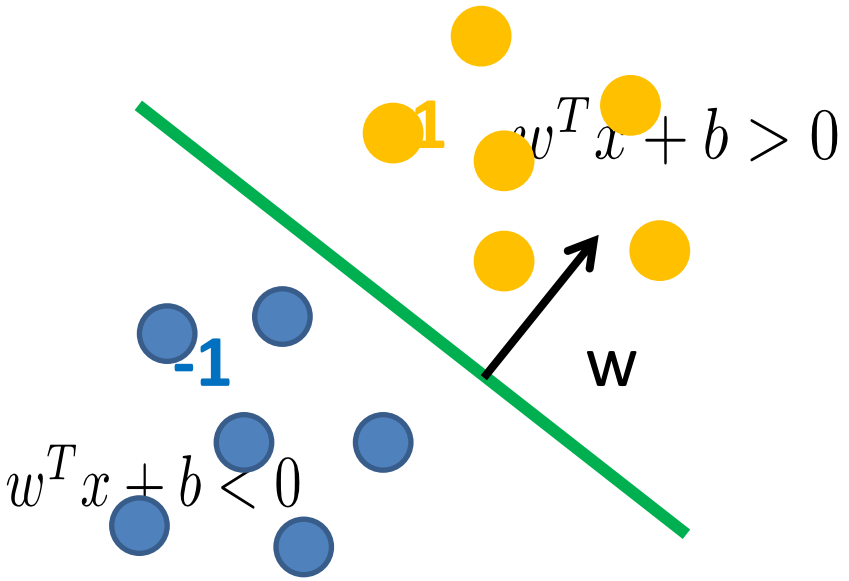
- Due date for HW4 has been changed!
- Now due Monday 11/03
- No late days, just one dropbox

Announcements

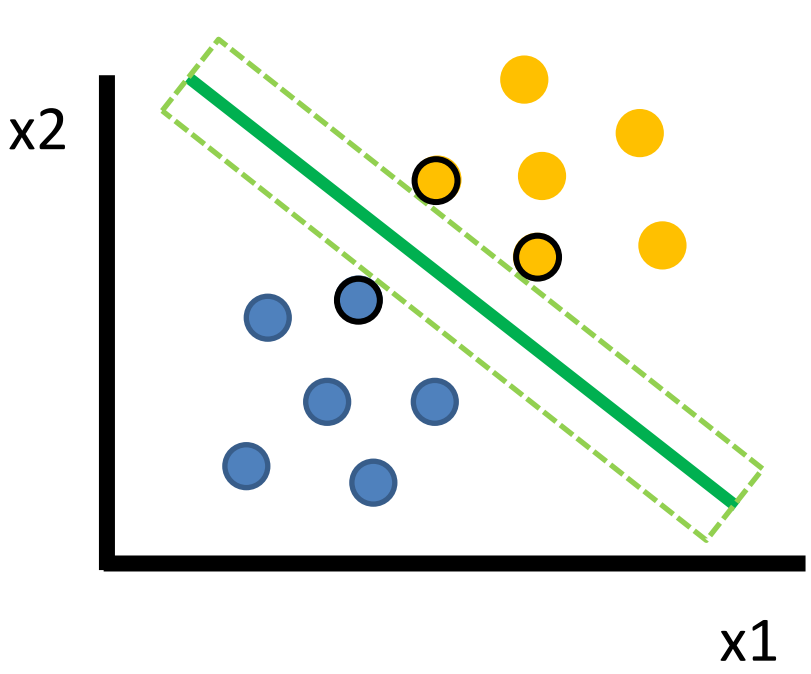
- Chris Wiggins, the Chief Data Scientist at the New York Times, is presenting in the IACS seminar **tomorrow**
- Lunch at 12:30, seminar starts at 1 pm
- MD G115

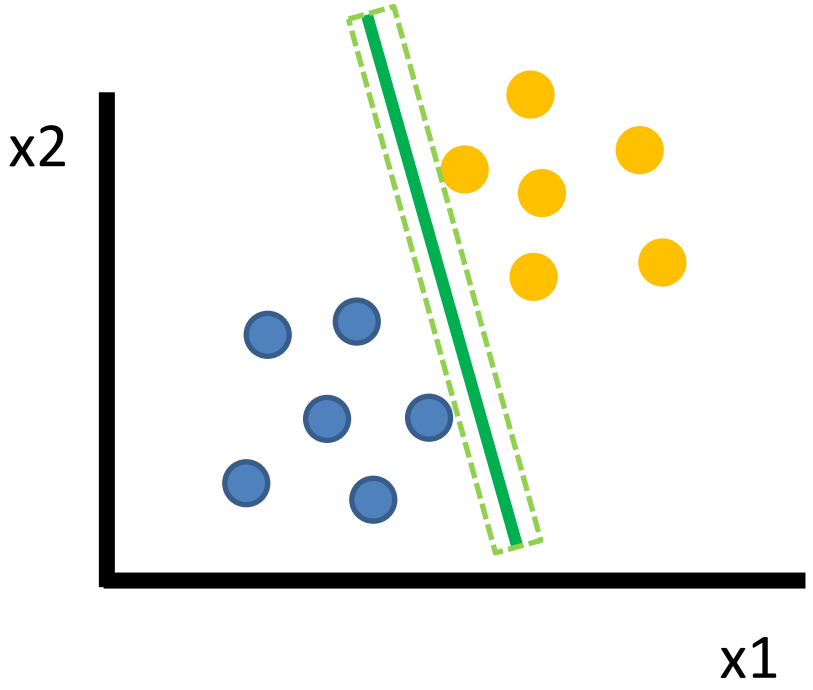
Separating Hyperplane

- x: data point
- y: label $\in \{-1, +1\}$
- w: weight vector
- b: bias



Maximum Margin Classification

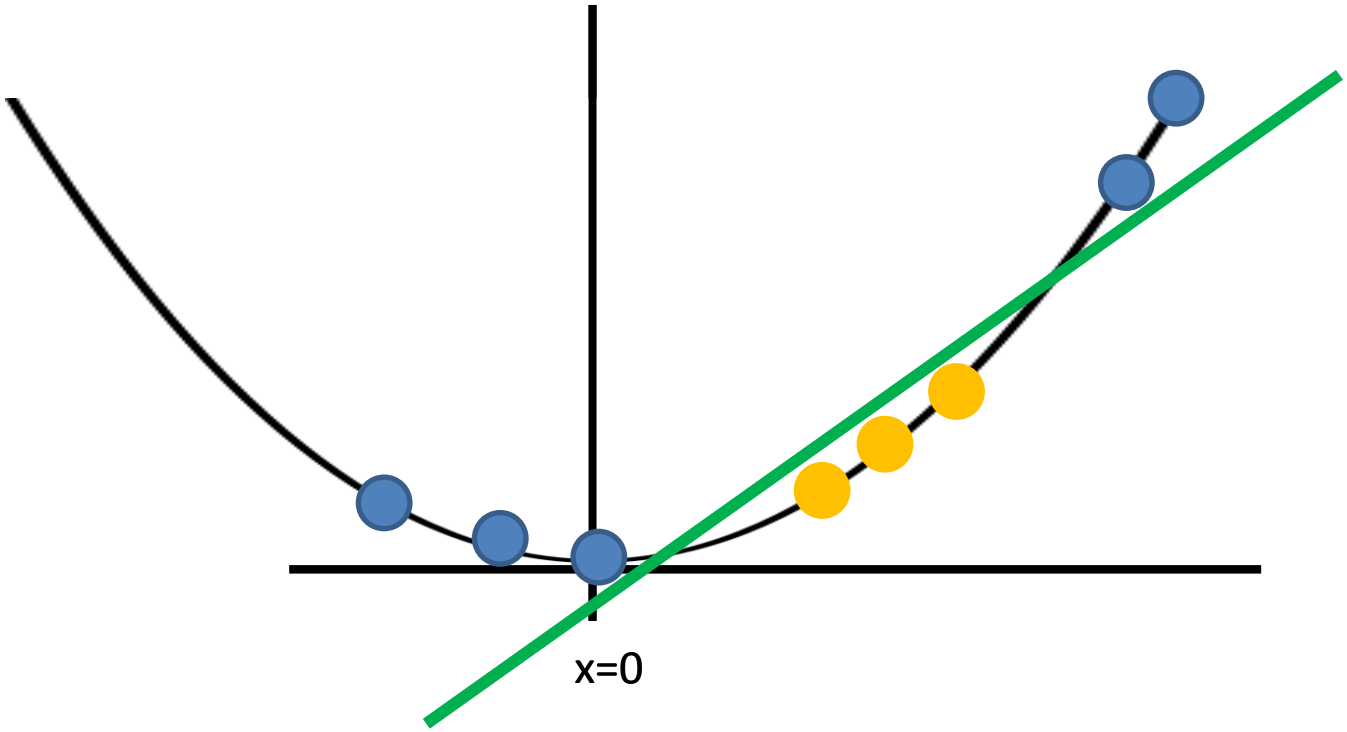




Tips and Tricks

- SVMs are not scale invariant
- Check if your library normalizes by default
- Normalize your data
  - mean: 0 , std: 1
  - map to [0,1] or [-1,1]
- Normalize test set in same way!

XOR problem revised



Did we add information to make the problem seperable?

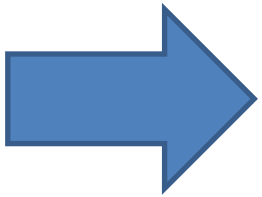
SVM Applet, Part 2

http://www.ml.inf.ethz.ch/education/lectures_and_seminars/annex_estat/Classifier/JSupportVectorApplet.html

Parameter Tuning

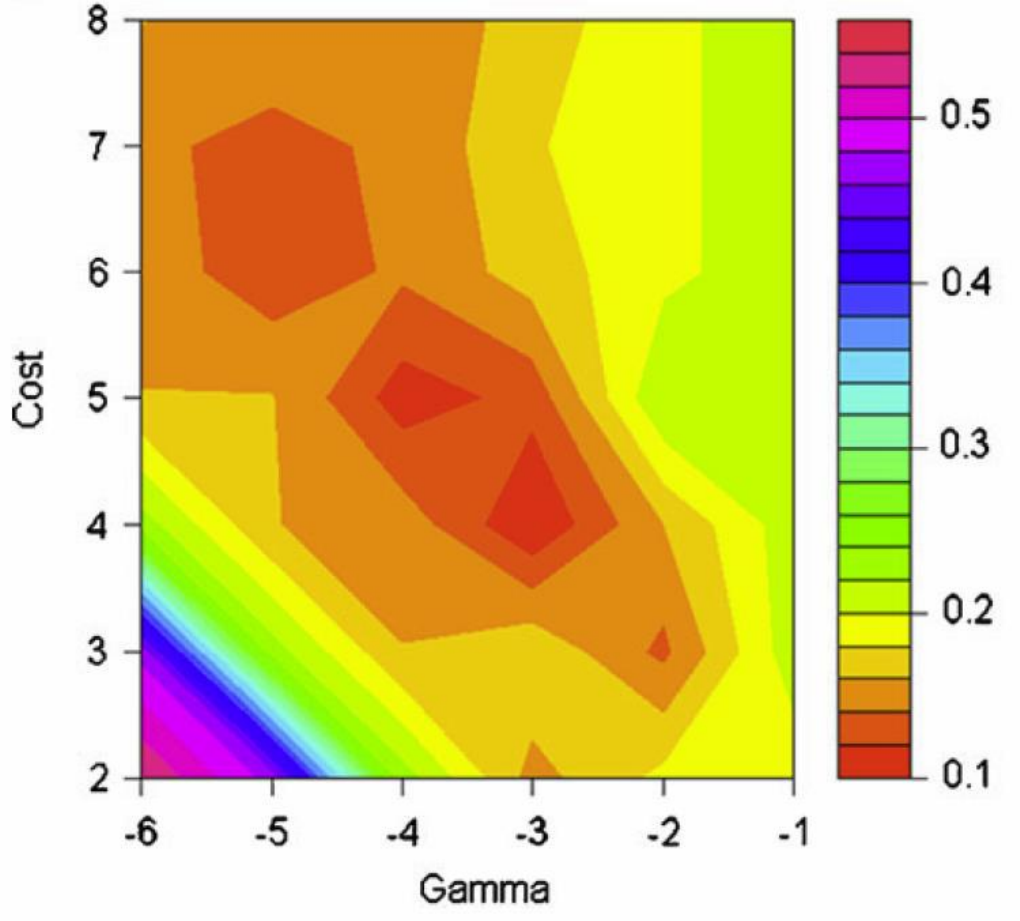
- Given a classification task

- Which kernel ?
- Which kernel parameter values?
- Which value for **C**?



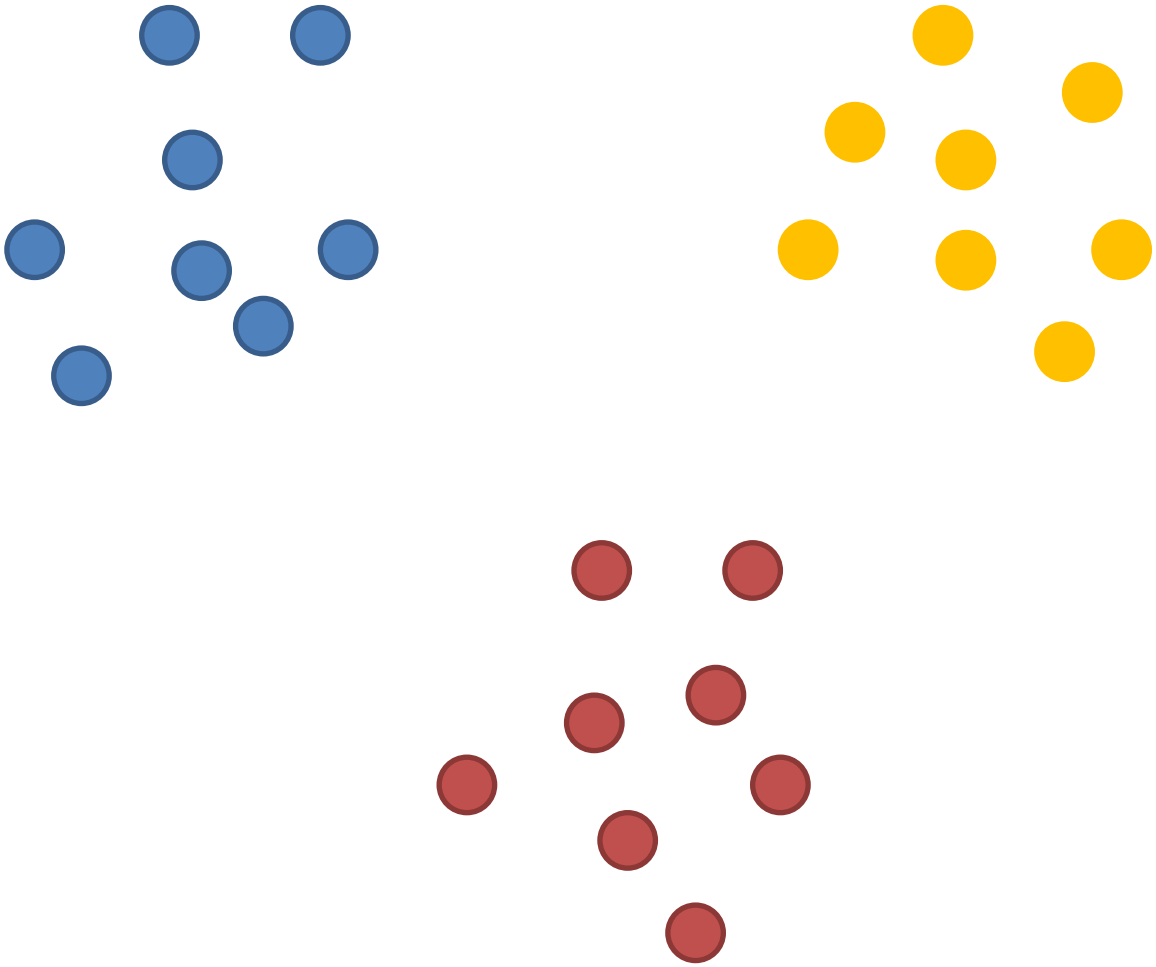
Try different combinations
and take the best.

Grid Search



Zang et al., “Identification of heparin samples that contain impurities or contaminants by chemometric pattern recognition analysis of proton NMR spectral data”, Anal Bioanal Chem (2011)

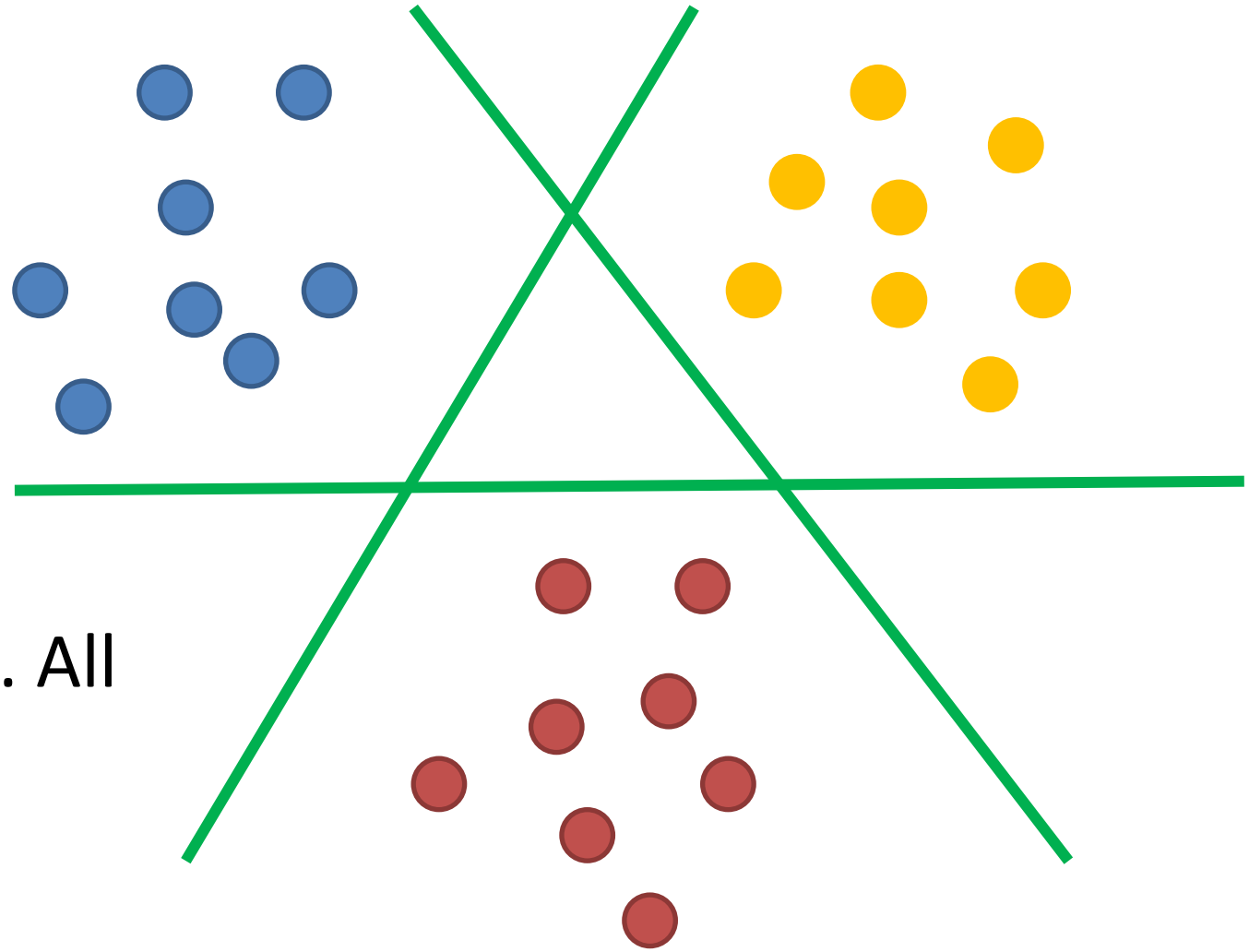
Multi Class



One vs All

- Train n classifier for n classes
- Take classification with greatest positive margin
- Slow training

Multi Class

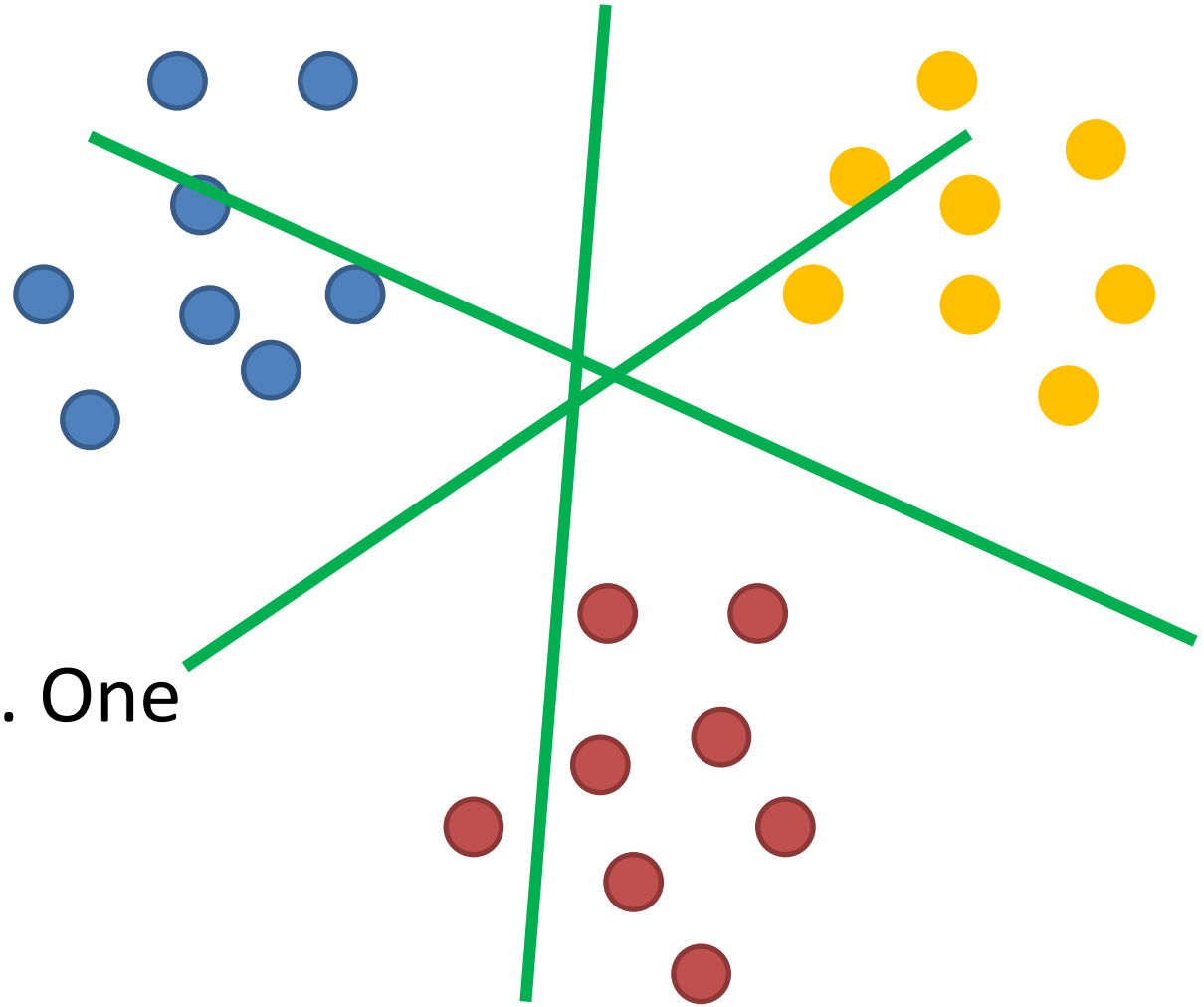


One vs. All

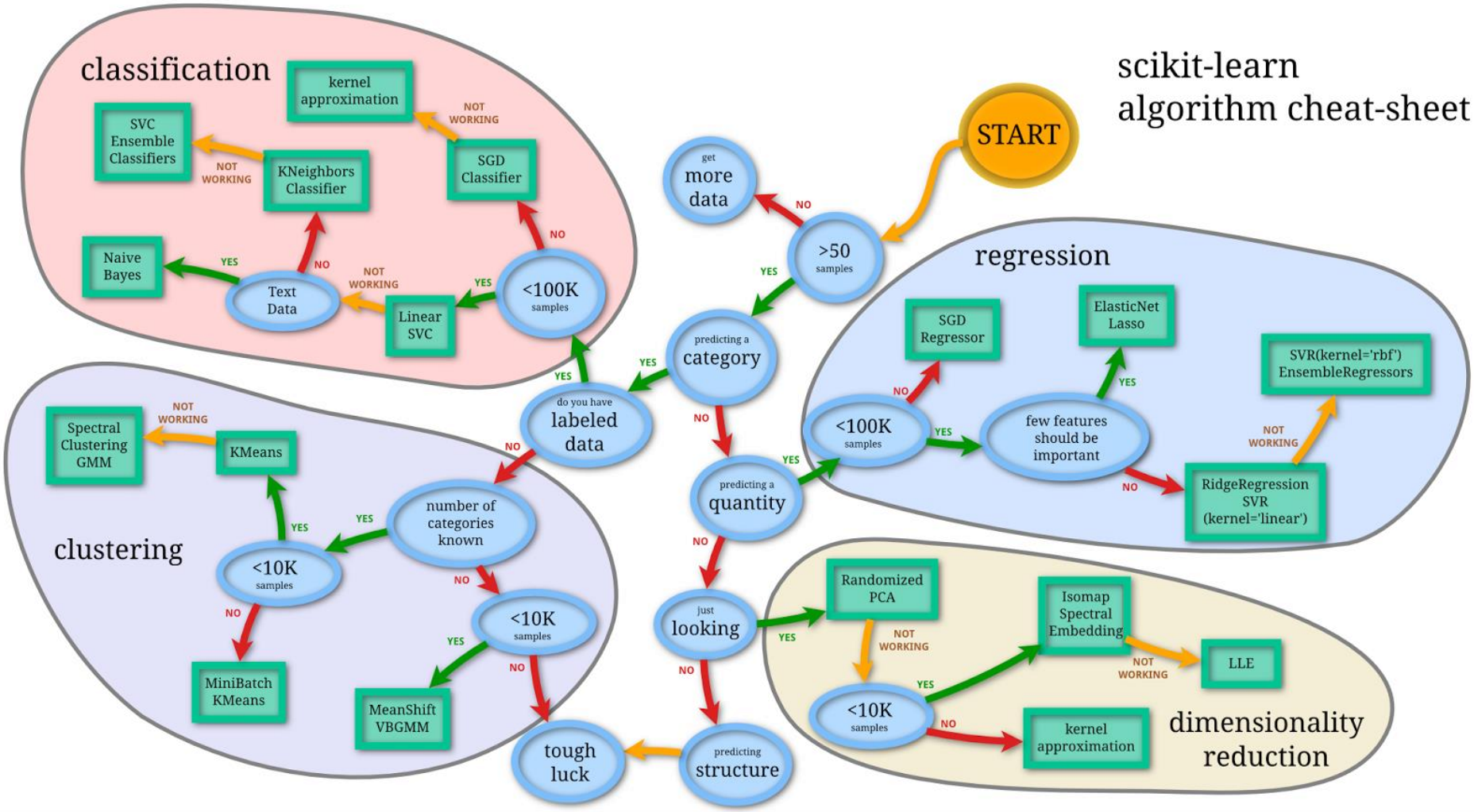
One vs One

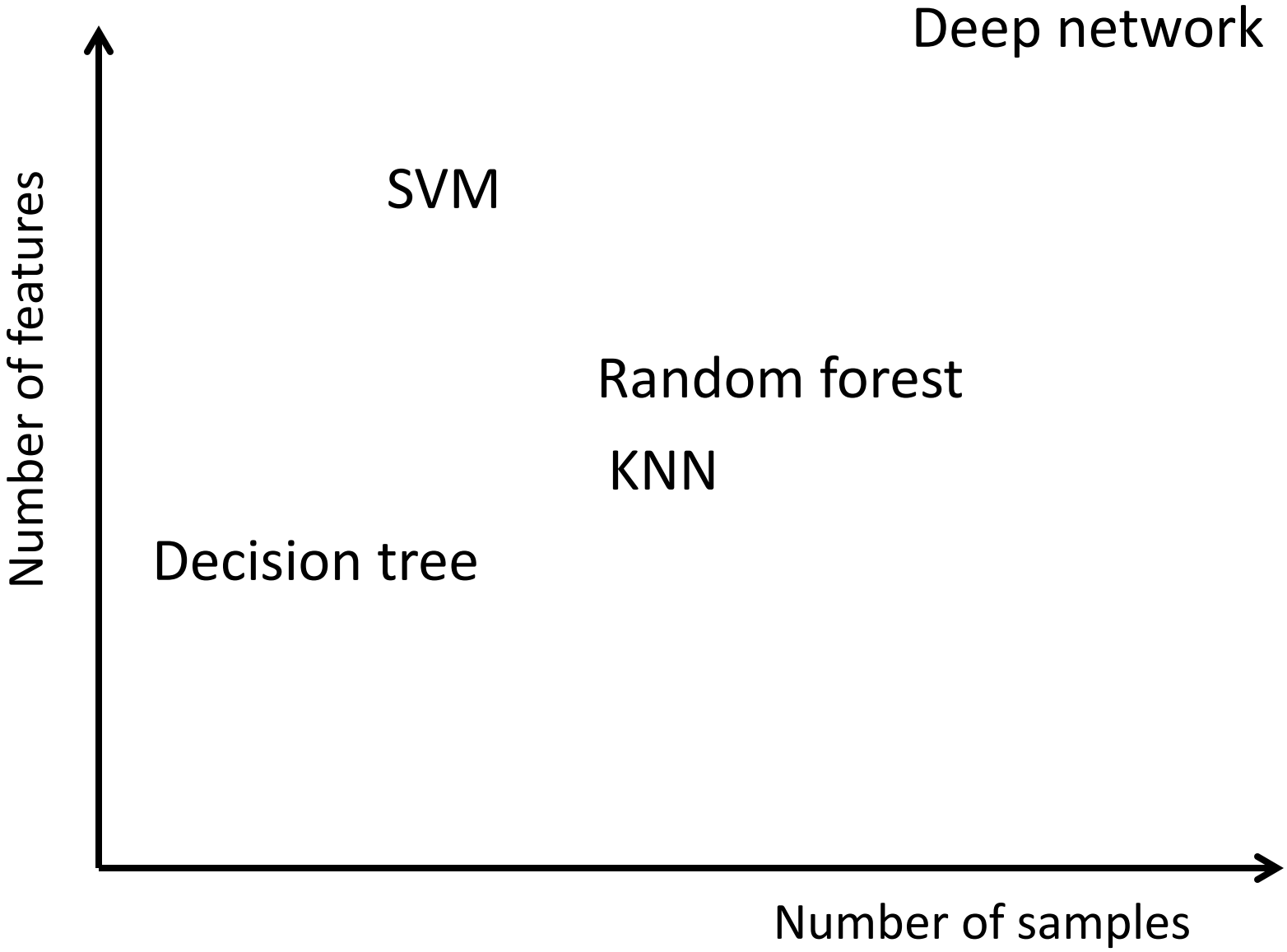
- Train n(n-1)/2 classifiers
- Take majority vote
- Fast training

Multi Class



One vs. One


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
NO
NOT WORKING
YES
Text Data
Linear SVC
<100K samples
get more data
NO
>50 samples
YES
regression
SGD Regressor
ElasticNet Lasso
SVR(kernel='rbf') EnsembleRegressors
NO
YES
<100K samples
YES
few features should be important
NOT WORKING
NO
RidgeRegression SVR (kernel='linear')
predicting a category
YES
NO
do you have labeled data
YES
NO
predicting a quantity
YES
NO
clustering
Spectral Clustering GMM
NOT WORKING
KMeans
YES
number of categories known
YES
NO
<10K samples
NO
<10K samples
YES
NO
MiniBatch KMeans
MeanShift VBGMM
just looking
YES
Randomized PCA
NOT WORKING
Isomap Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck


Deep network
SVM
Random forest
KNN
Decision tree
Number of features
Number of samples

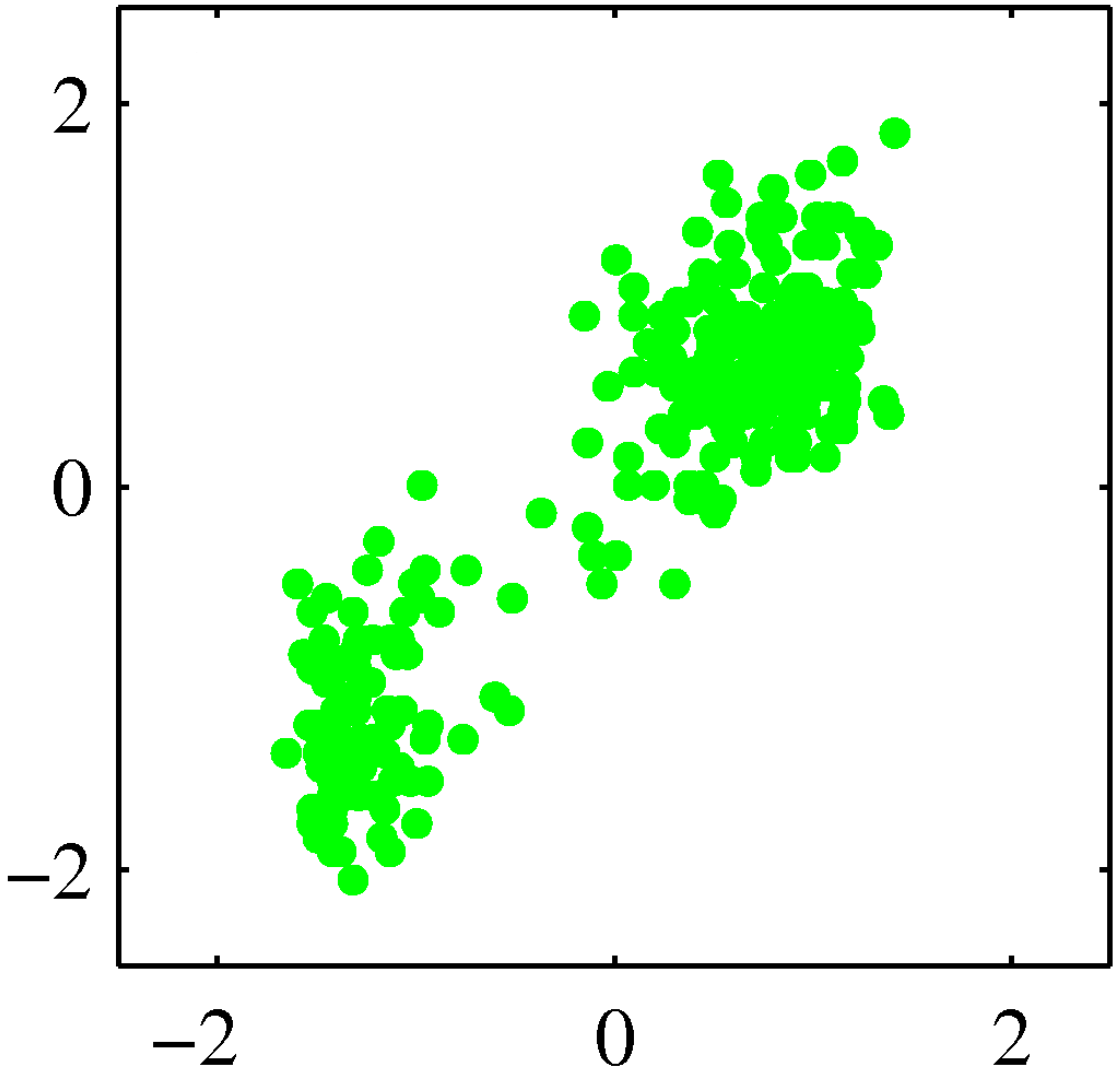
Unsupervised Learning

- K-means
- Hierarchical Clustering
- Mean-shift

- Rand index, stability

- Applications

Unsupervised Setting

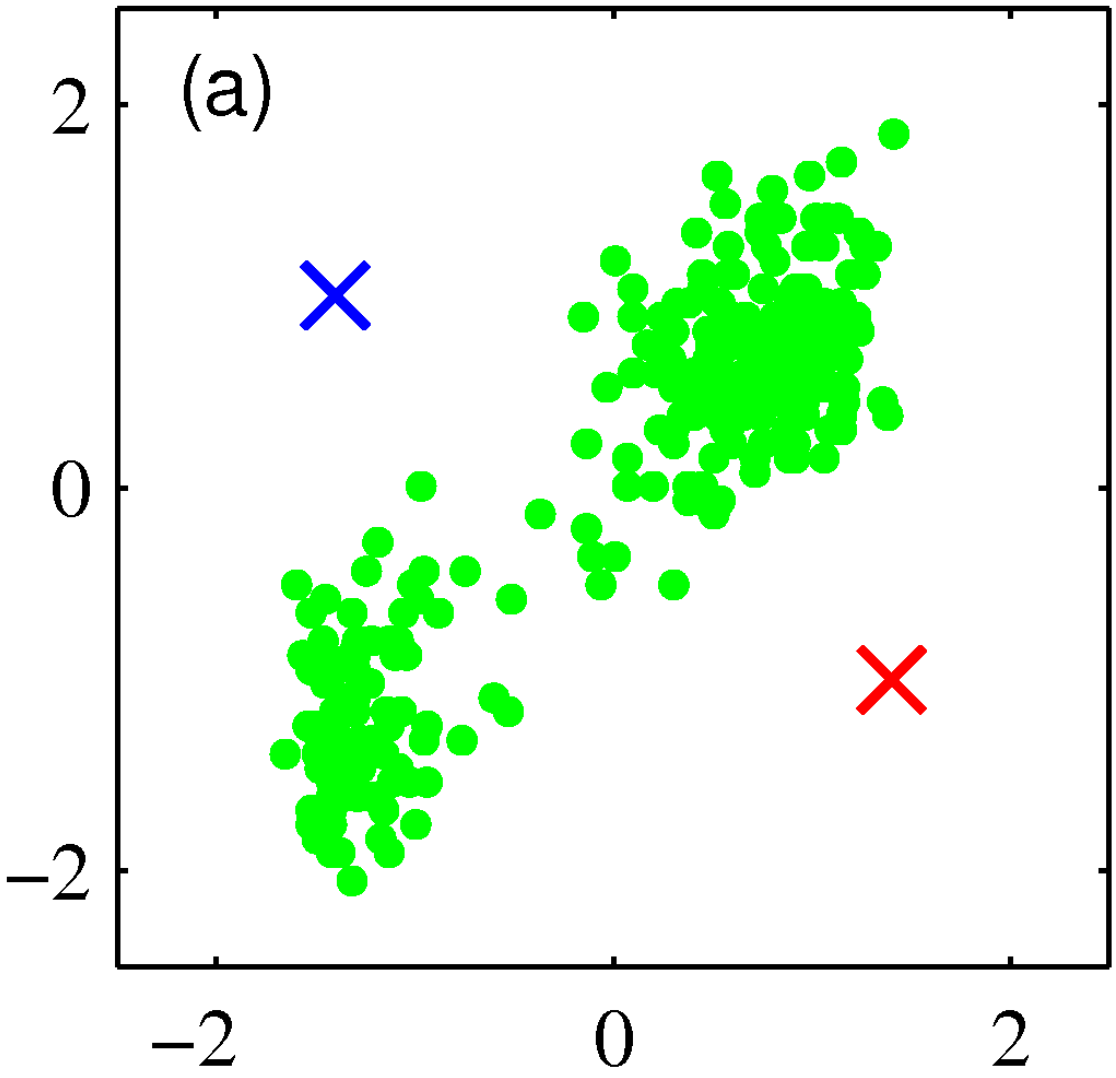


Bishop, “Pattern Recognition and Machine Learning”, Springer, 2006

K-means – Algorithm

- Initialization:
  - choose k random positions
  - assign cluster centers $\mu^{(j)}$ to these positions

K-means



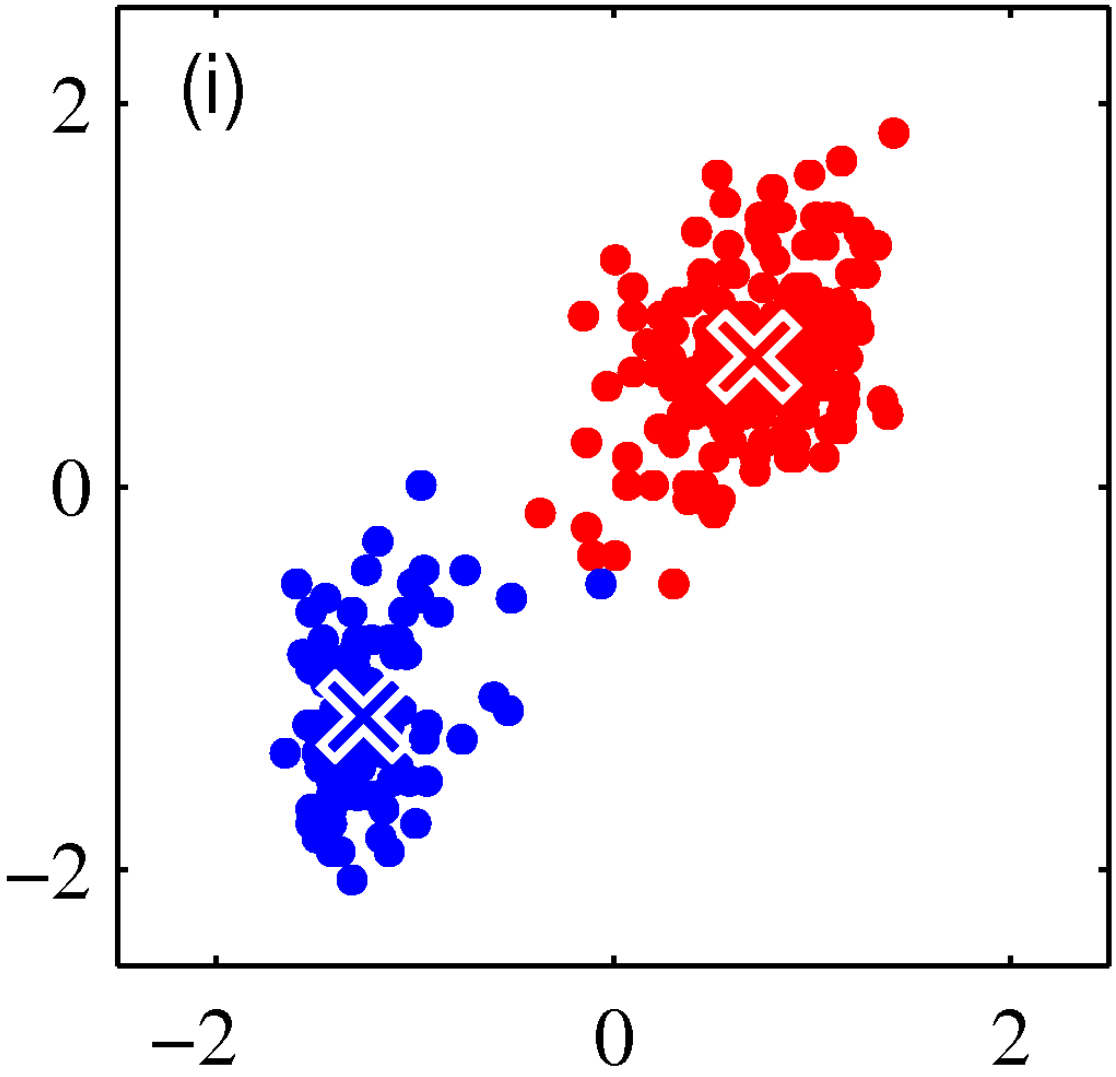
Bishop, “Pattern Recognition and Machine Learning”, Springer, 2006

K-means

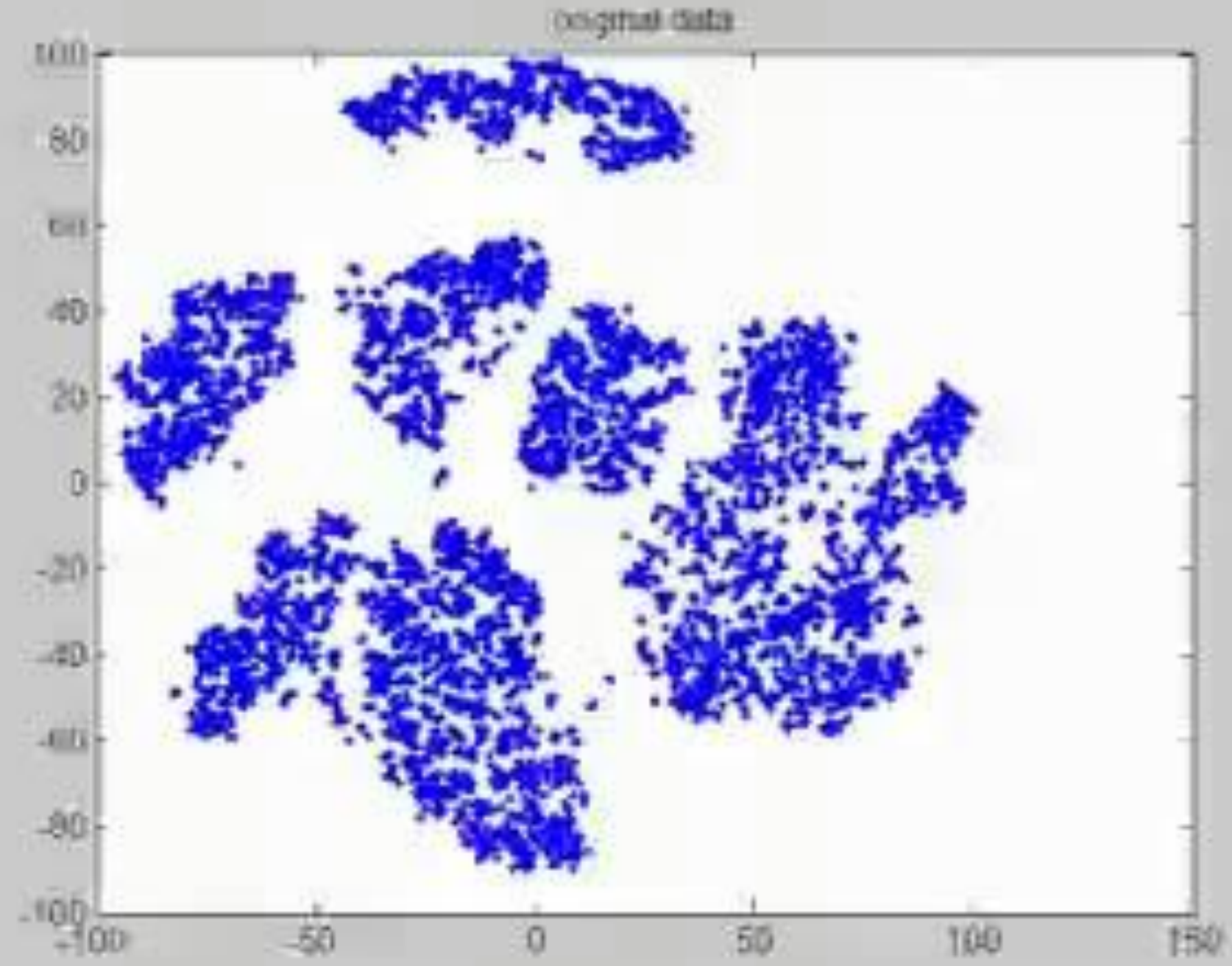
- Until Convergence:
  - Compute distances $||x^{(i)} - \mu^{(j)}||$
  - Assign points to nearest cluster center
  - Update Cluster centers:

$$\mu^{(j)} = \frac{1}{N_j} \sum_{x_i \in C_j} x_i$$

K-means



Bishop, “Pattern Recognition and Machine Learning”, Springer, 2006


original data
100
80
60
40
20
0
-20
-40
-60
-80
-100
-100
-50
0
50
100
150

K-means Example

R

G

B

K-means Example

K-means Example

K-means Summary

- Guaranteed to converge
- Result depends on initialization

- Number of clusters is important

- Sensitive to outliers
  - Use median instead of mean for updates

Initialization Methods

- Random Positions
- Random data points as Centers
- Random Cluster assignment to data points

- Start several times

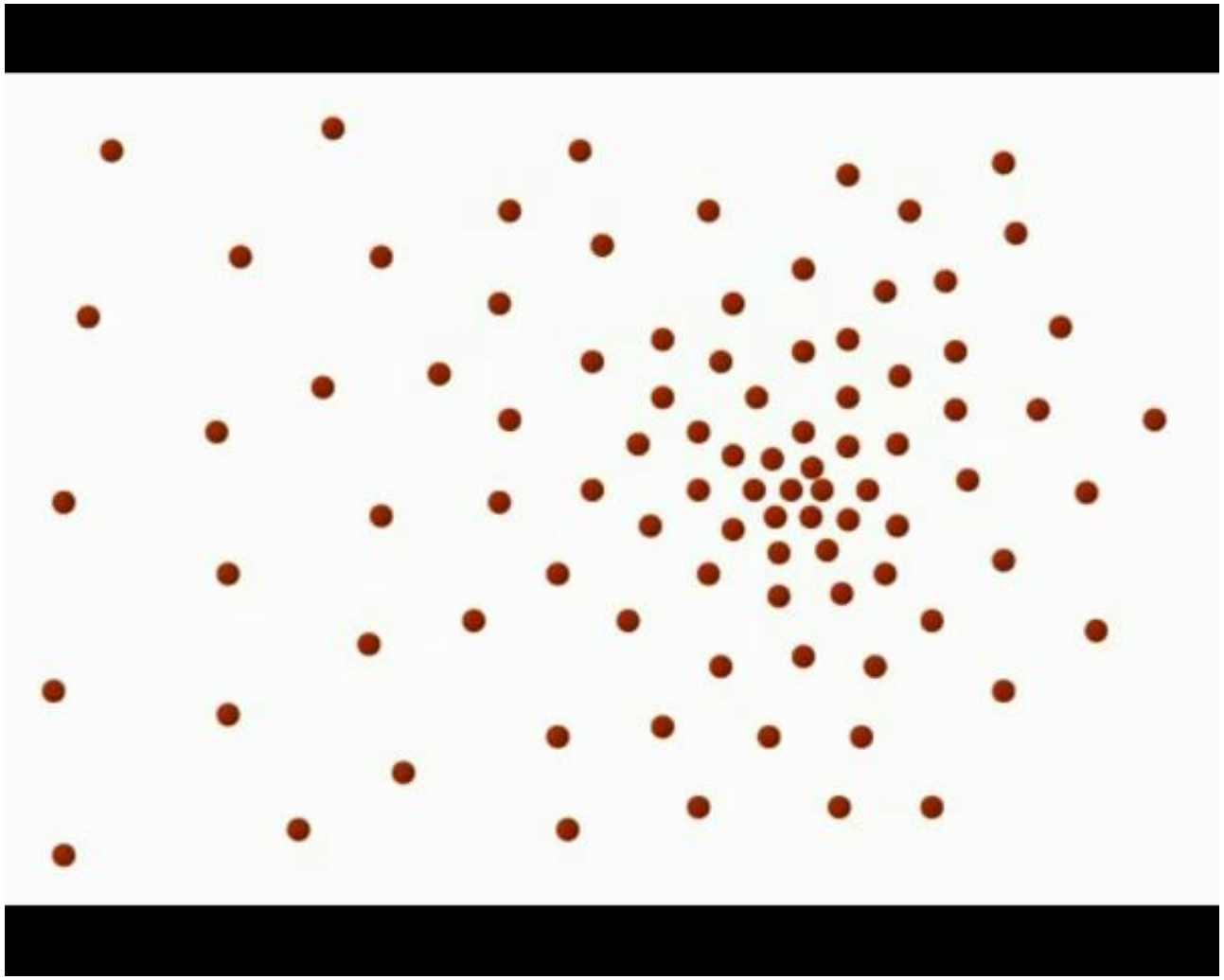
How to find k

- Cross Validation
- Partition data into n folds
- Cluster on n-1 folds
- Compute sum of squared distances to centroids for validation set

Mean Shift

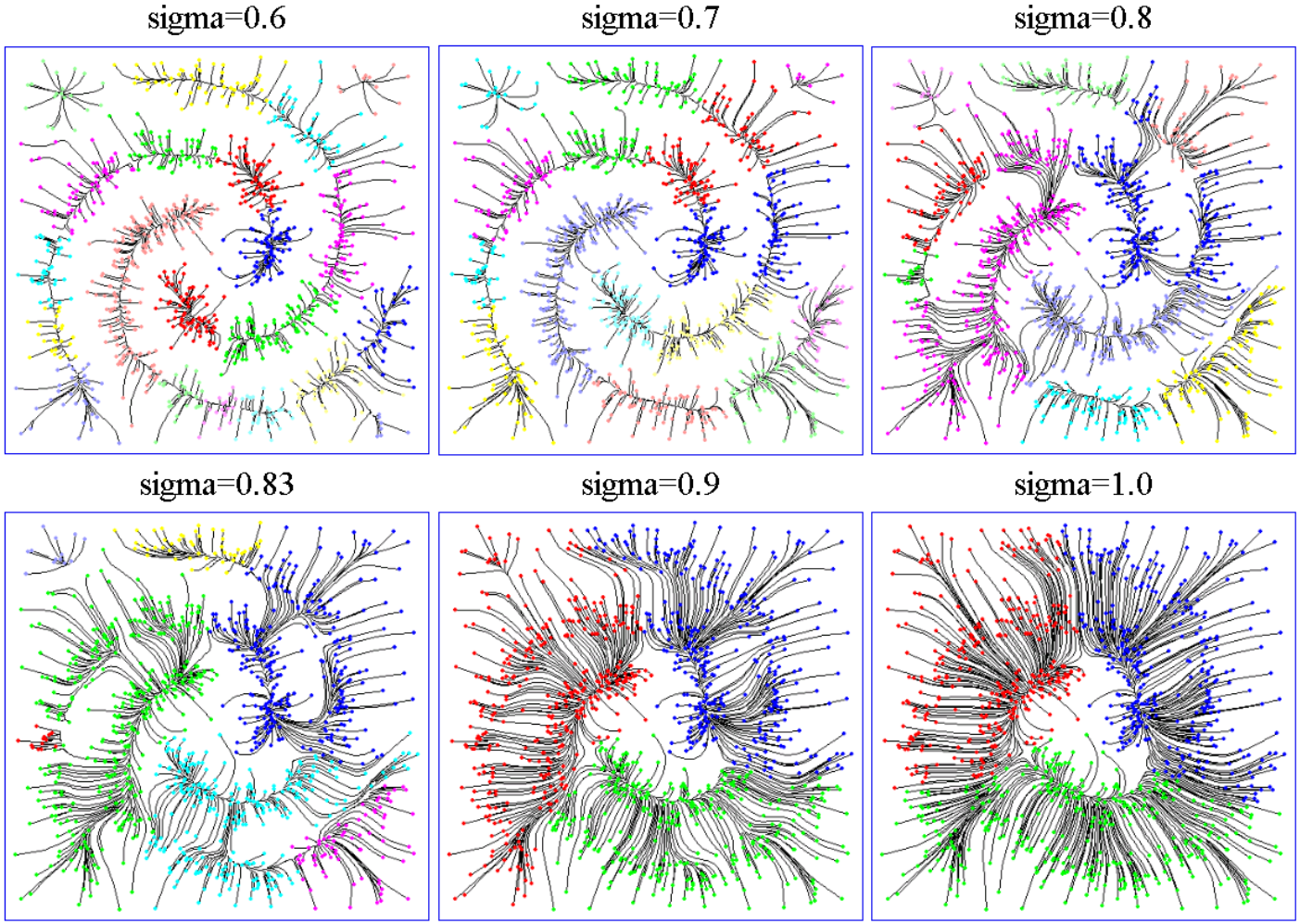
1. Put a window around each point
2. Compute mean of points in the frame.
3. Shift the window to the mean
4. Repeat until convergence

Mean Shift



http://www.youtube.com/watch?v=kmaQAsotT9s

Mean Shift



Fischer et al., “Clustering with the Connectivity Kernel“, NIPS (2003)

Mean Shift Summary

- Does not need to know number of clusters
- Can handle arbitrary shaped clusters
- Robust to initialization
- Needs bandwidth parameter (window size)
- Computationally expensive

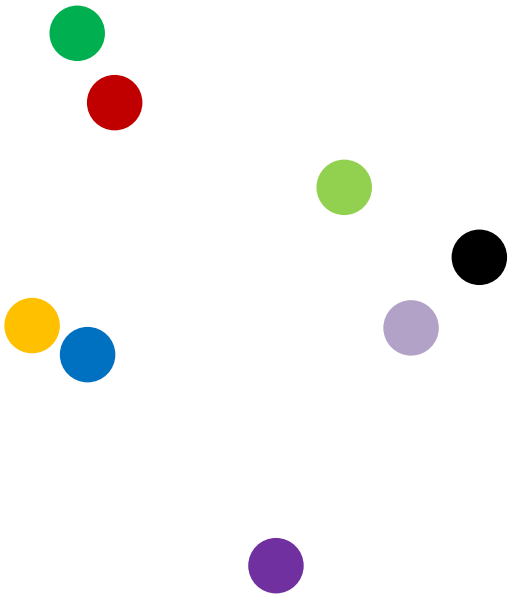
- Very good article:

http://saravananthirumuruganathan.wordpress.com/2010/04/01/introduction-to-mean-shift-algorithm/

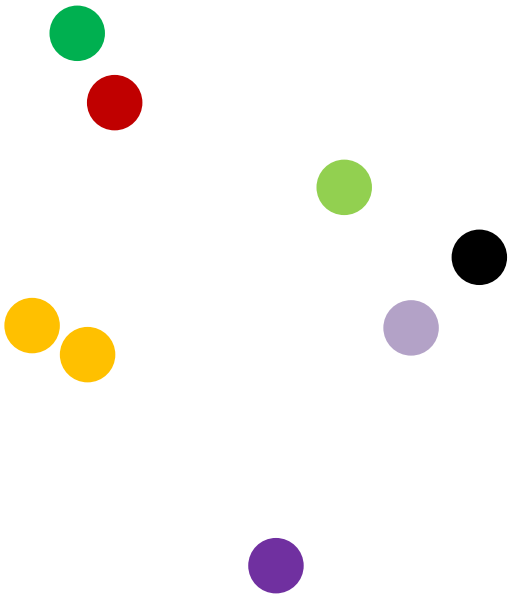
Multi-feature object trajectory clustering for video analysis

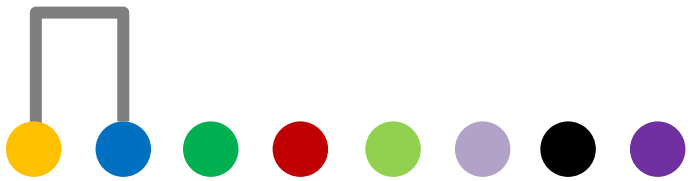
Nadeem Anjum Andrea Cavallaro

Hierarchical Clustering

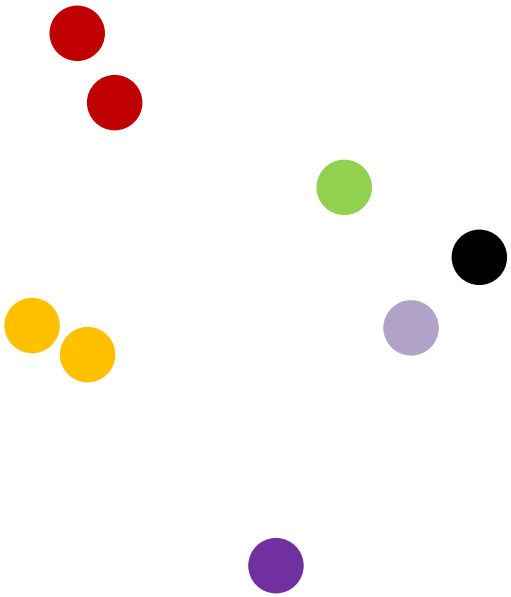


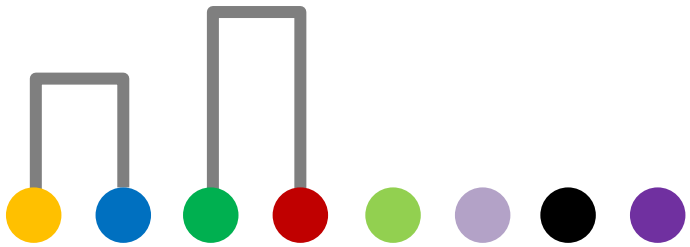
Hierarchical Clustering



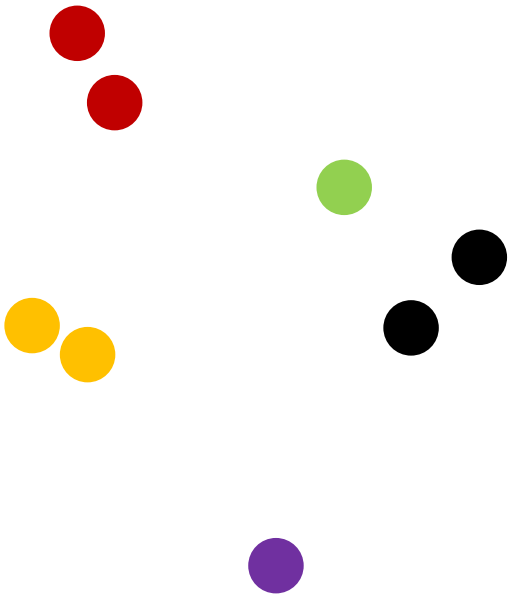


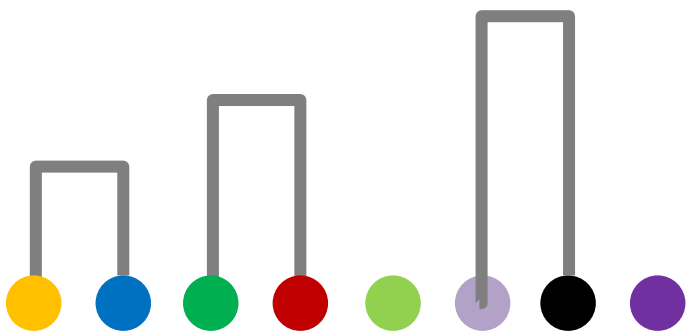
Hierarchical Clustering



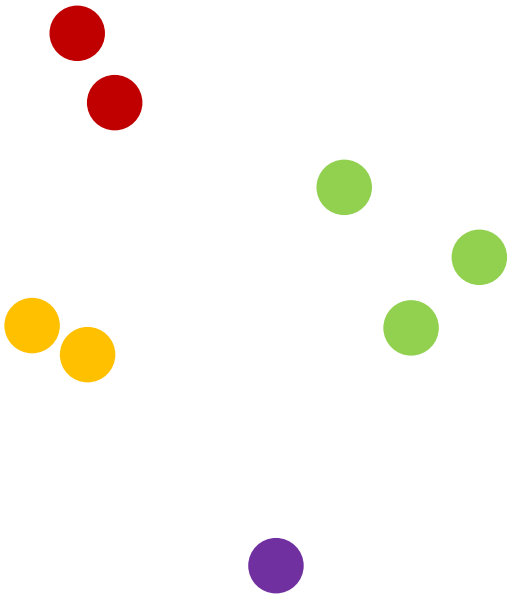


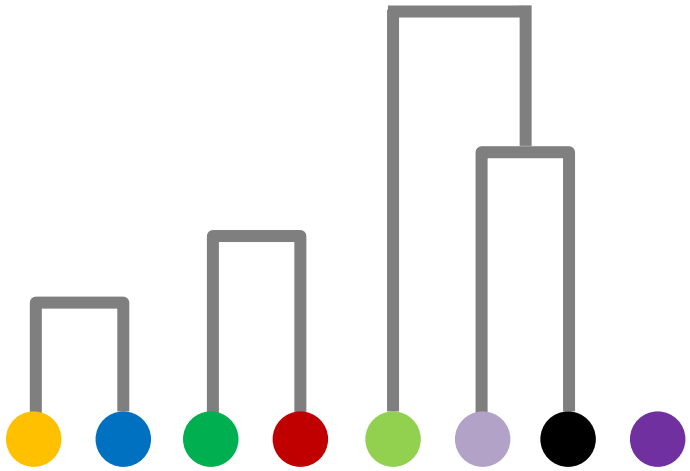
Hierarchical Clustering



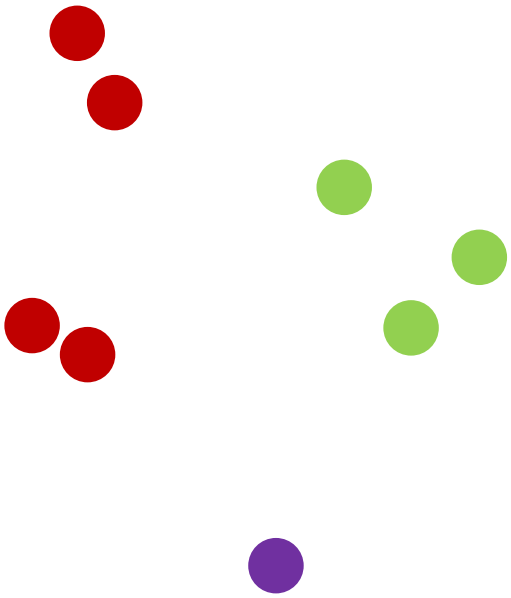


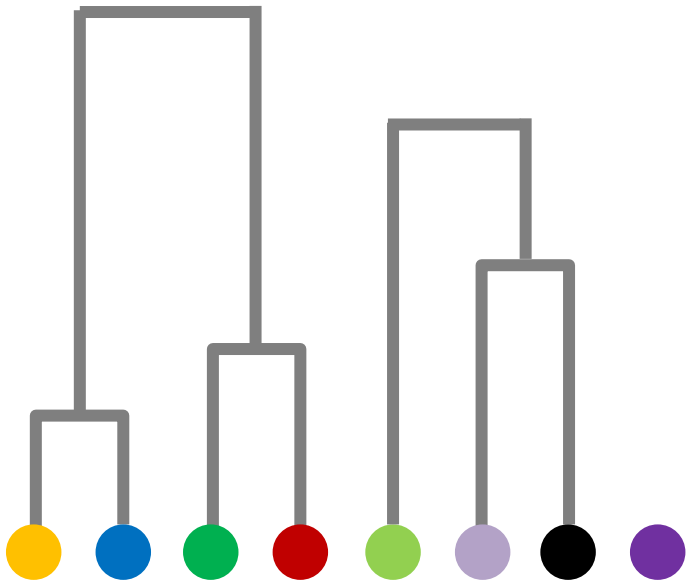
Hierarchical Clustering



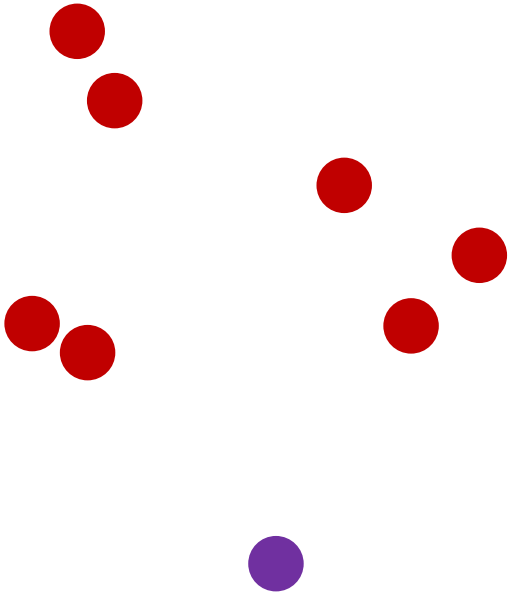


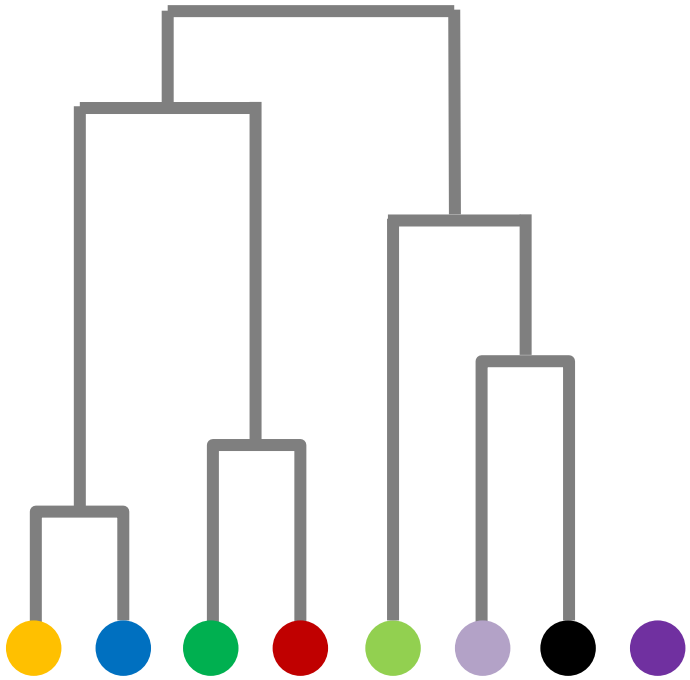
Hierarchical Clustering



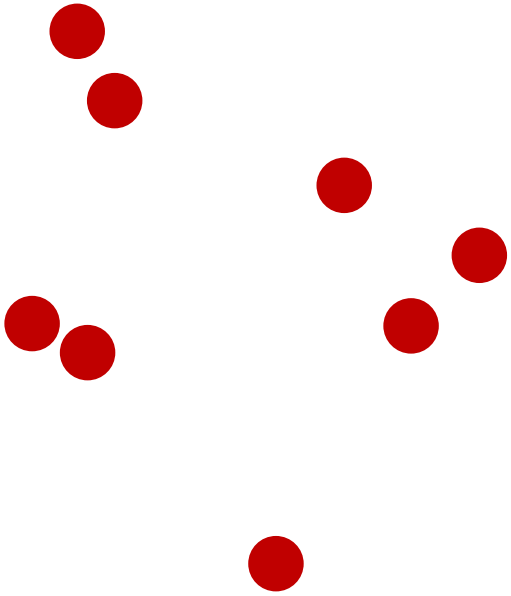


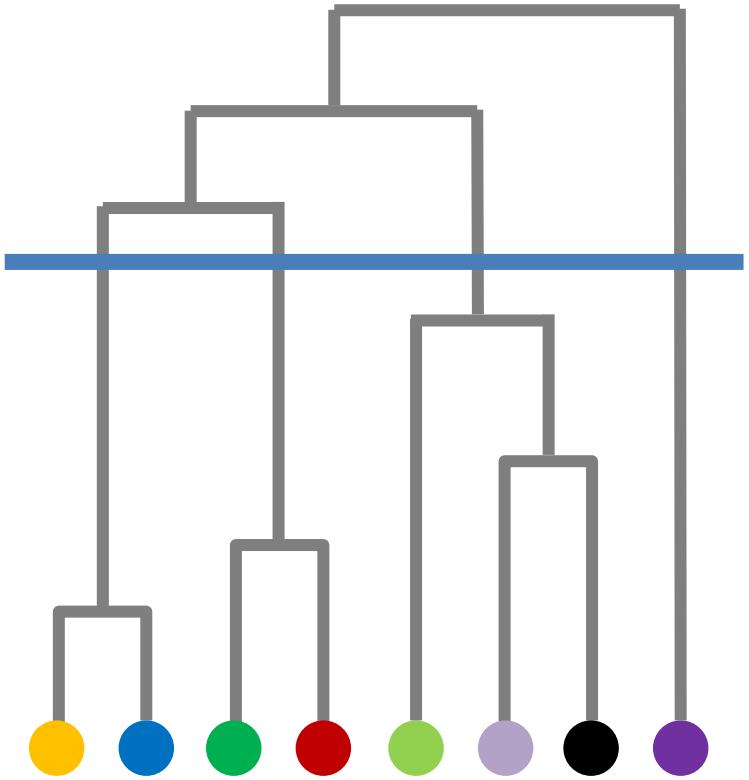
Hierarchical Clustering



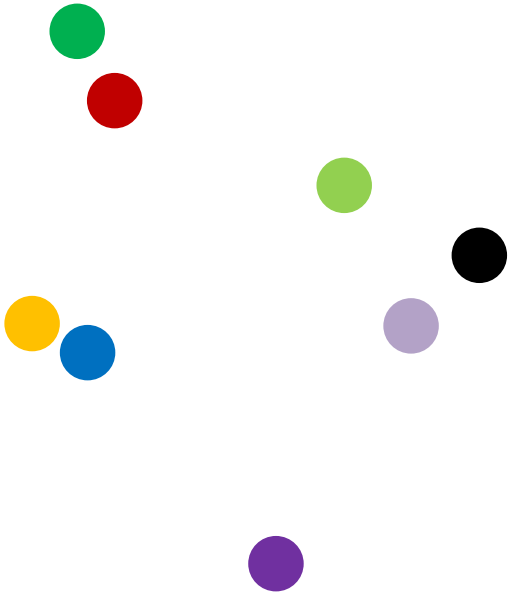


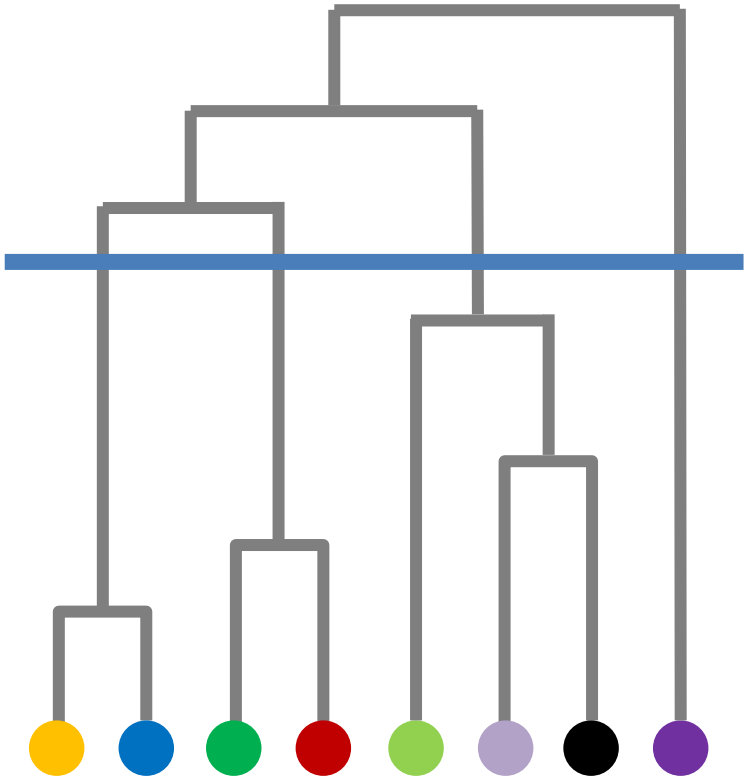
Hierarchical Clustering





Hierarchical Clustering



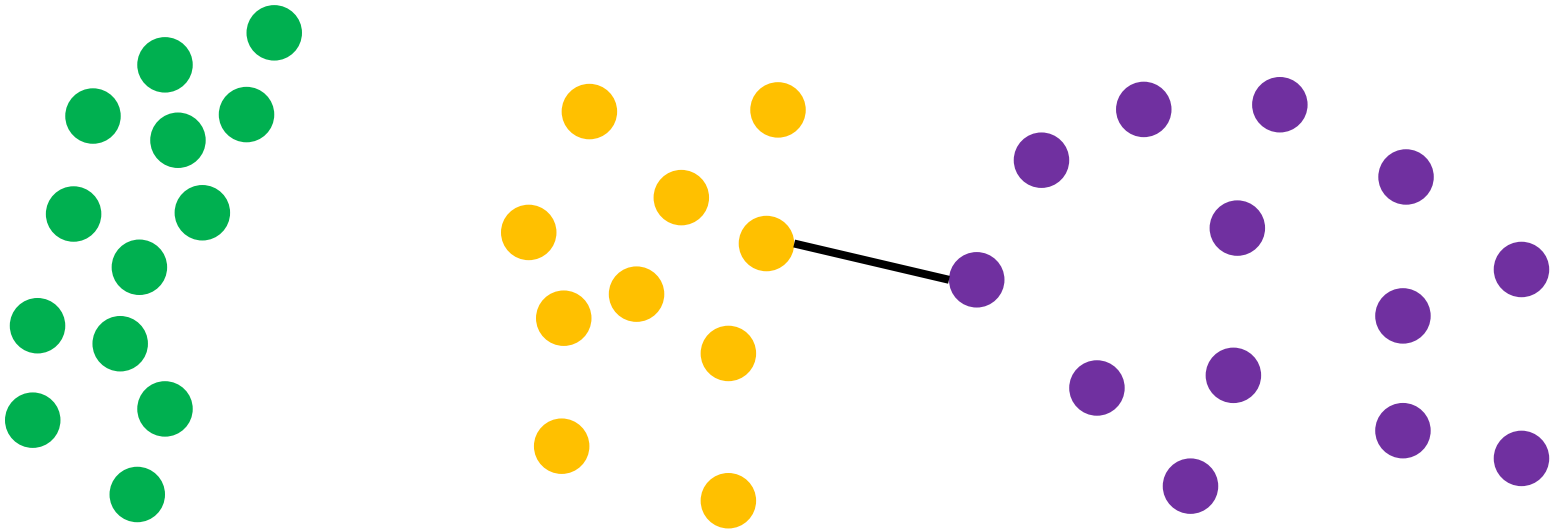


Hierarchical Clustering

- Produces complete structure
- No predefined number of clusters

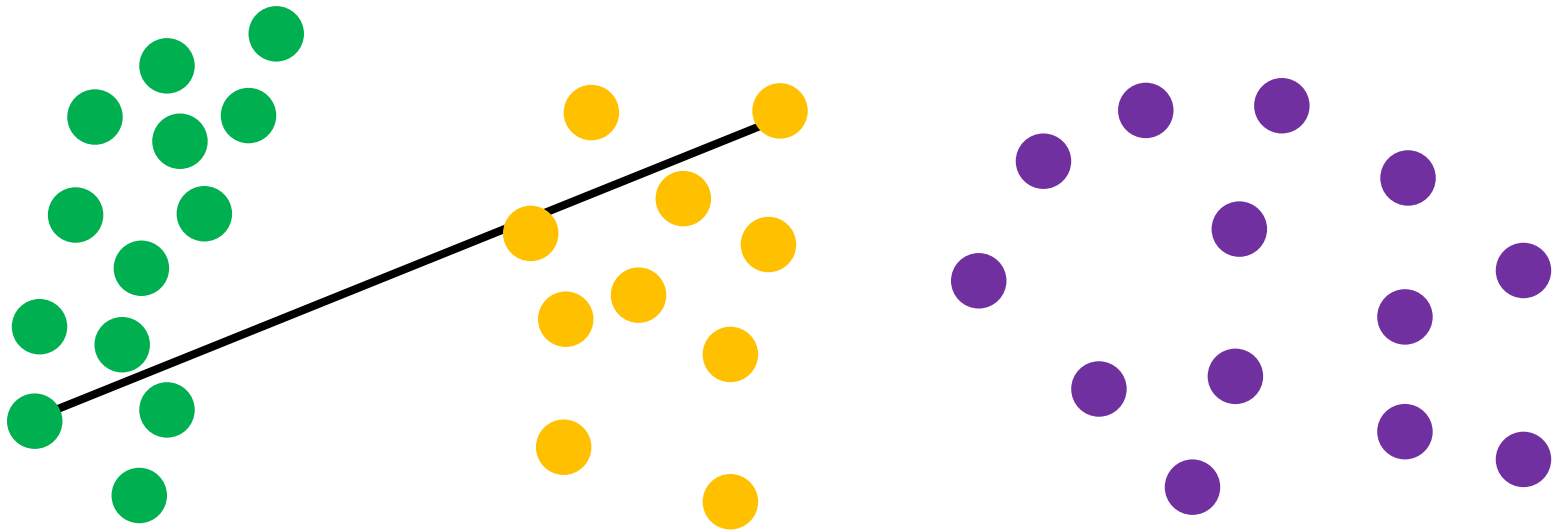
- Similarity between clusters:
  - single-linkage: $\min\{d(x,y) : x \in \mathcal{A}, y \in \mathcal{B}\}$
  - complete-linkage: $\max\{d(x,y) : x \in \mathcal{A}, y \in \mathcal{B}\}$
  - average linkage: $\frac{1}{|\mathcal{A}|\cdot|\mathcal{B}|} \sum_{x\in\mathcal{A}} \sum_{y\in\mathcal{B}} d(x,y)$

Single Linkage



$$\min\{d(x,y) : x \in \mathcal{A}, y \in \mathcal{B}\}$$

Complete Linkage



$$\max\{d(x,y) : x \in \mathcal{A}, y \in \mathcal{B}\}$$

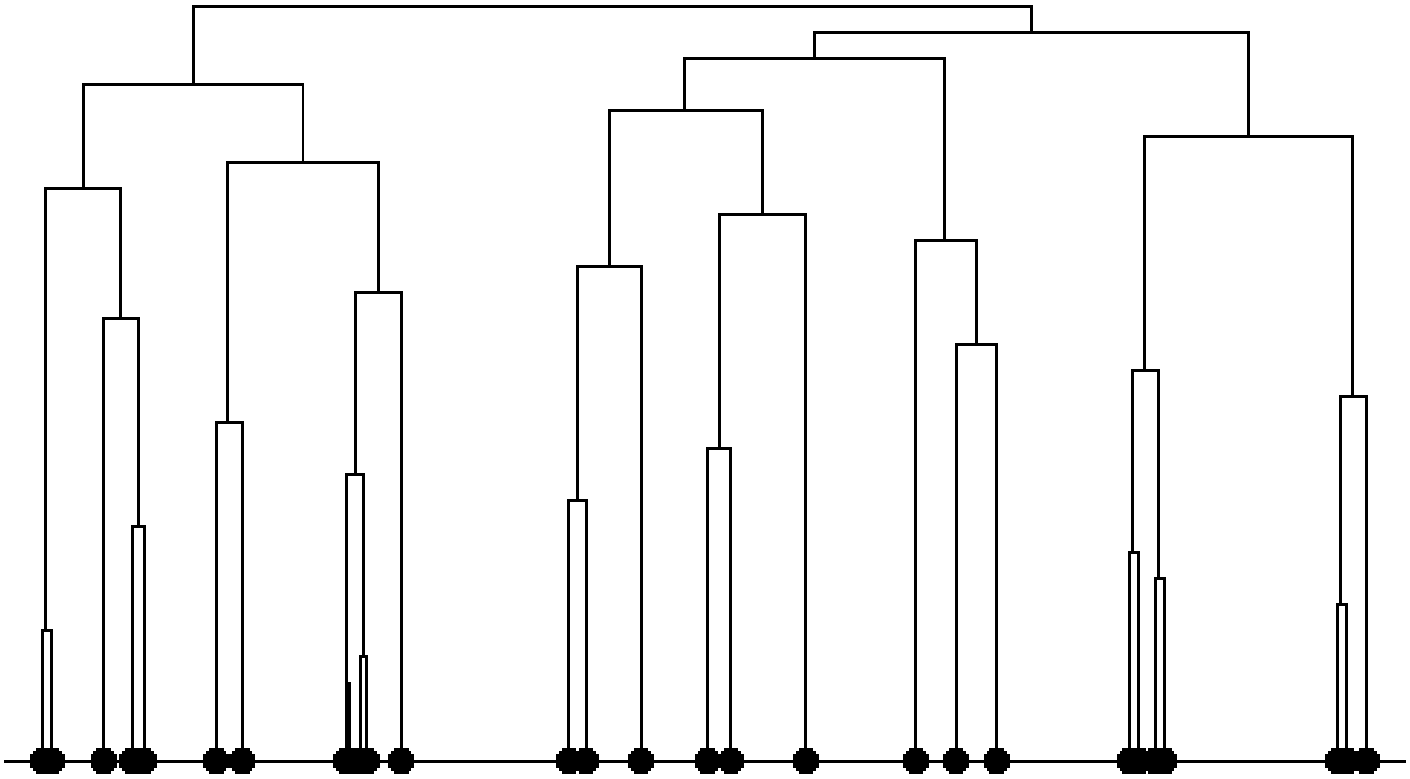
Linkage Matters

- Single linkage: tendency to form long chains
- Complete linkage: Sensitive to outliers
- Average-link: Trying to compromise between the two

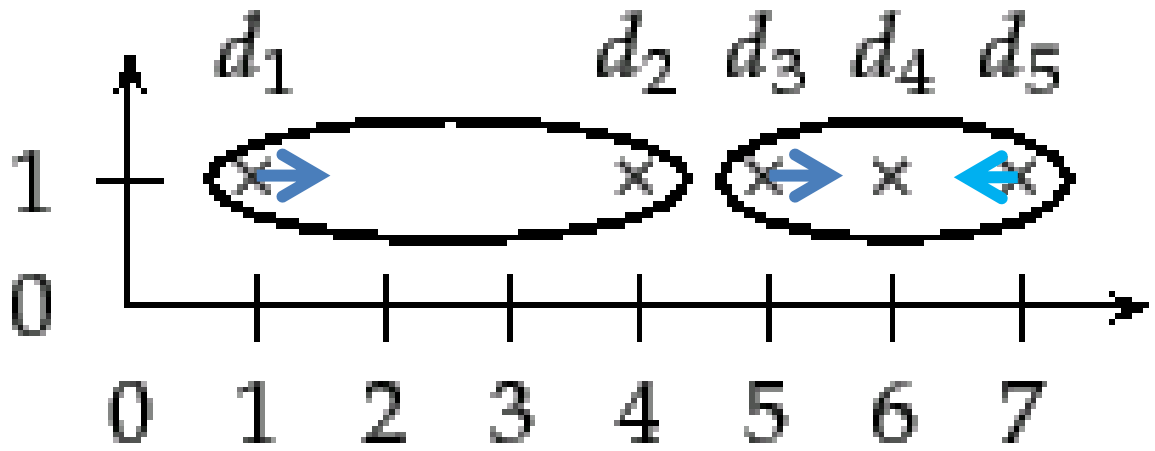
Hierarchical Clustering

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

Chaining Phenomenon



Outlier Sensitivity

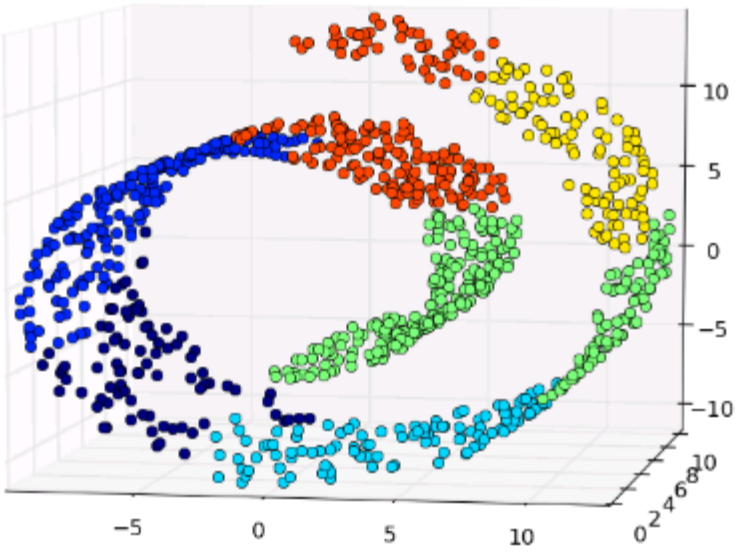


+ 2*epsilon

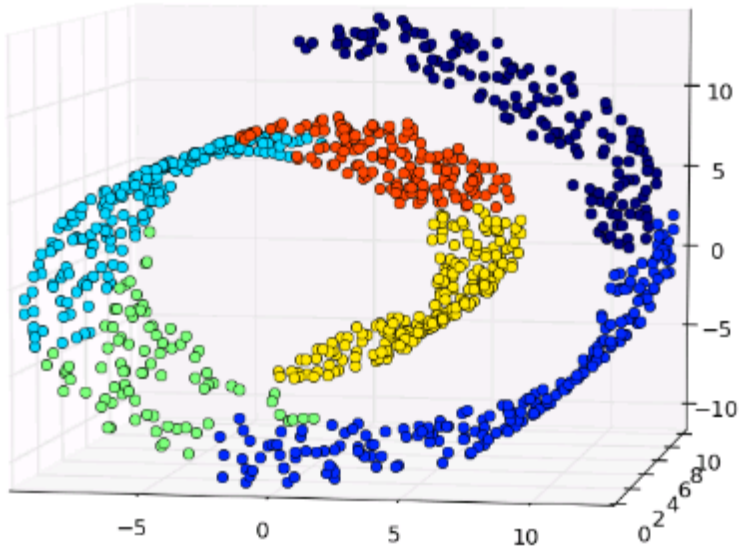
- 1*epsilon

http://nlp.stanford.edu/IR-book/html/htmledition/img1569.png

Swiss Role Problem



without connectivity constraints

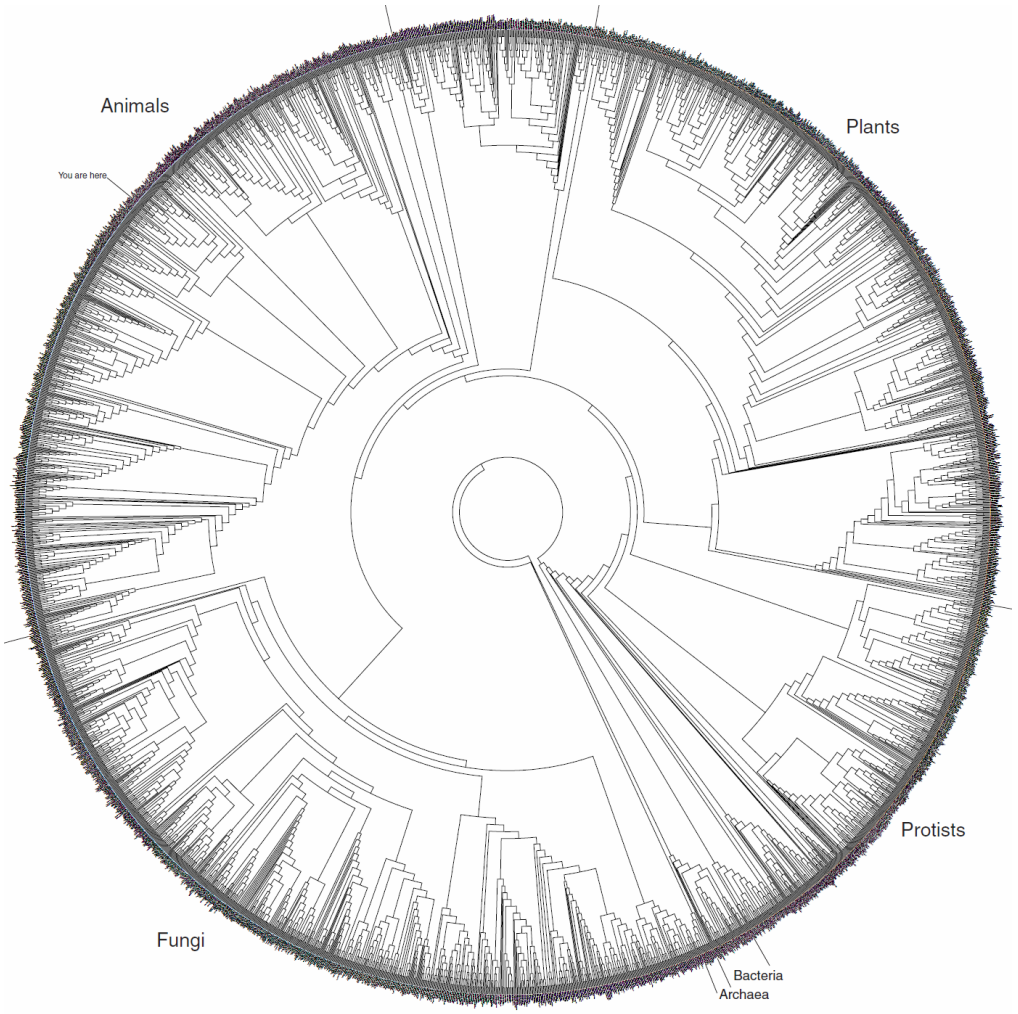


with connectivity constraints

only adjacent clusters can be merged together

http://scikit-learn.sourceforge.net/modules/clustering.html

Tree of Life



http://www.zo.utexas.edu/faculty/antisense/DownloadfilesToL.html