CS109 – Data Science

Verena Kaynig-Fittkau

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

AWS Clusters

- New and updated instructions for Spark 1.5 are on Piazza:

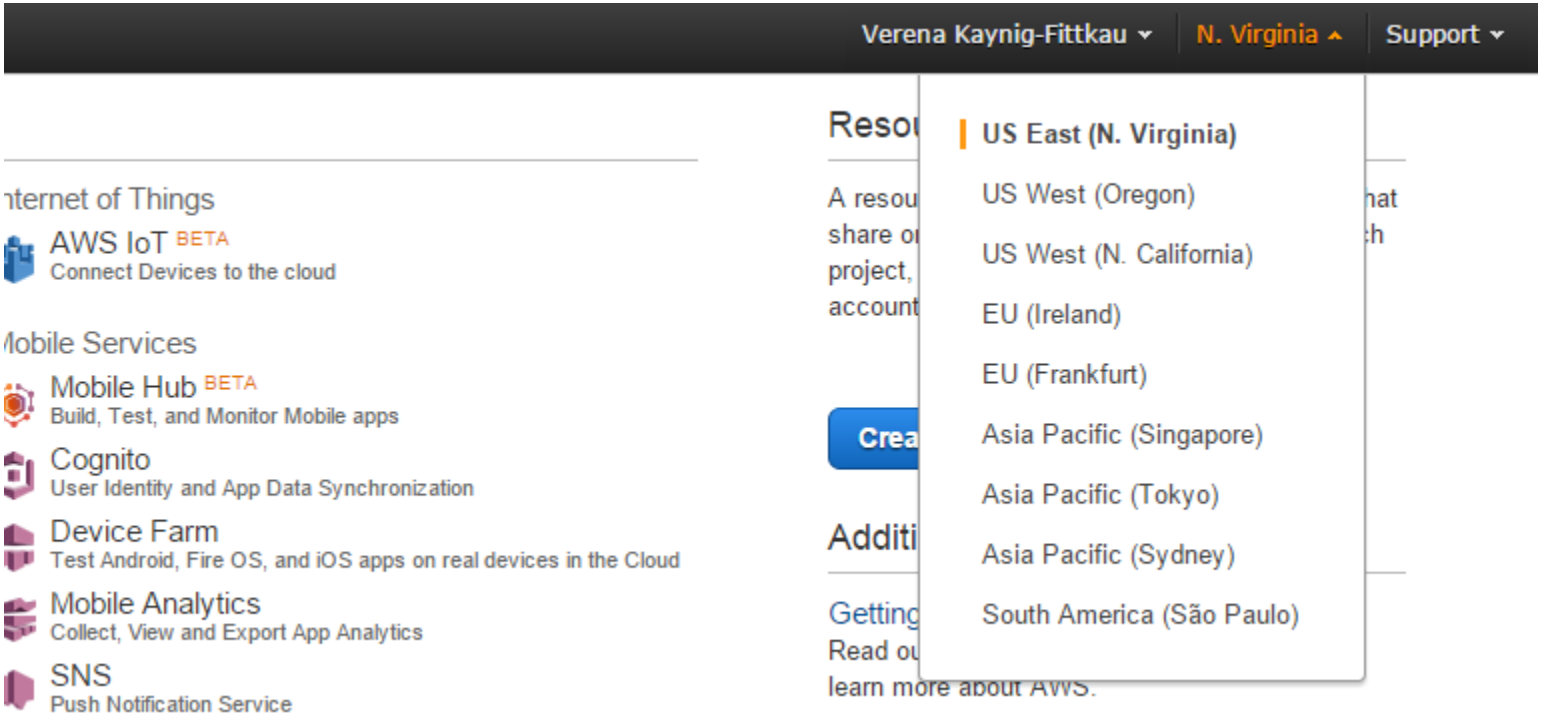
https://piazza.com/class/icf0cypdc3243c?cid=1369

Avoid Unnecessary Charges!

- Look at AWS console > Services > EMR
- There should be some terminated clusters there
- Check the region on the top right corner
- Make sure to change it to US East

https://piazza.com/class/icf0cypdc3243c?cid=1256

Region Setting in AWS



Announcements

- Final project
  - Team assignments have been posted to piazza
  - Make sure you are in a 3-4 person team
  - Try and date on the piazza thread
  - If you have problems write to staff@cs109.org

  - Project proposals are due on Thursday

  https://piazza.com/class/icf0cypdc3243c?cid=1317

Final Project Proposal

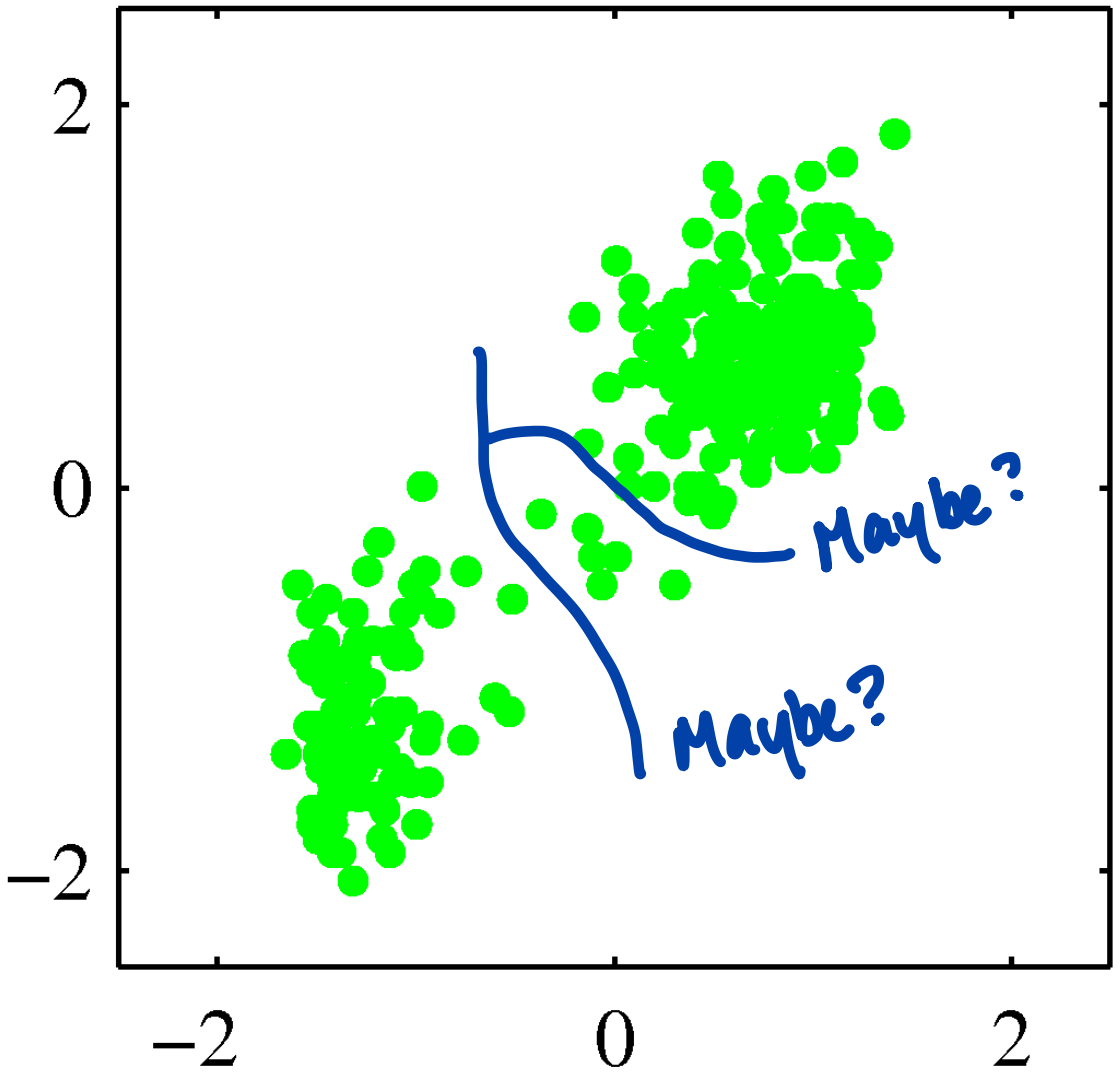
- Submit just **one form per team**.
- Do it as **early as possible**!
- No project approval until you meet your TF

https://piazza.com/class/icf0cypdc3243c?cid=1317

Supervised vs. Unsupervised

- We mainly talked about supervised learning so far
- Joe already moved to unsupervised with LDA
- In these settings we have **no labels** in our training data.

Unsupervised Setting



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

Unsupervised Learning

find existing and new patterns

- Find patterns in unlabeled data
- Sometimes used for a supervised setting in which labels are hard to get
- Can identify new patterns that you were not aware of.

Clustering Applications

- Google image search categories
- Author Clustering: http://academic.research.microsoft.com/VisualExplorer#1048044
- Opening a new location for a hospital, police station, etc.
- Outlier detection

Unsupervised Learning

- K-means
- Mean-shift
- Hierarchical Clustering

- Rand index, stability } Is it a good solution

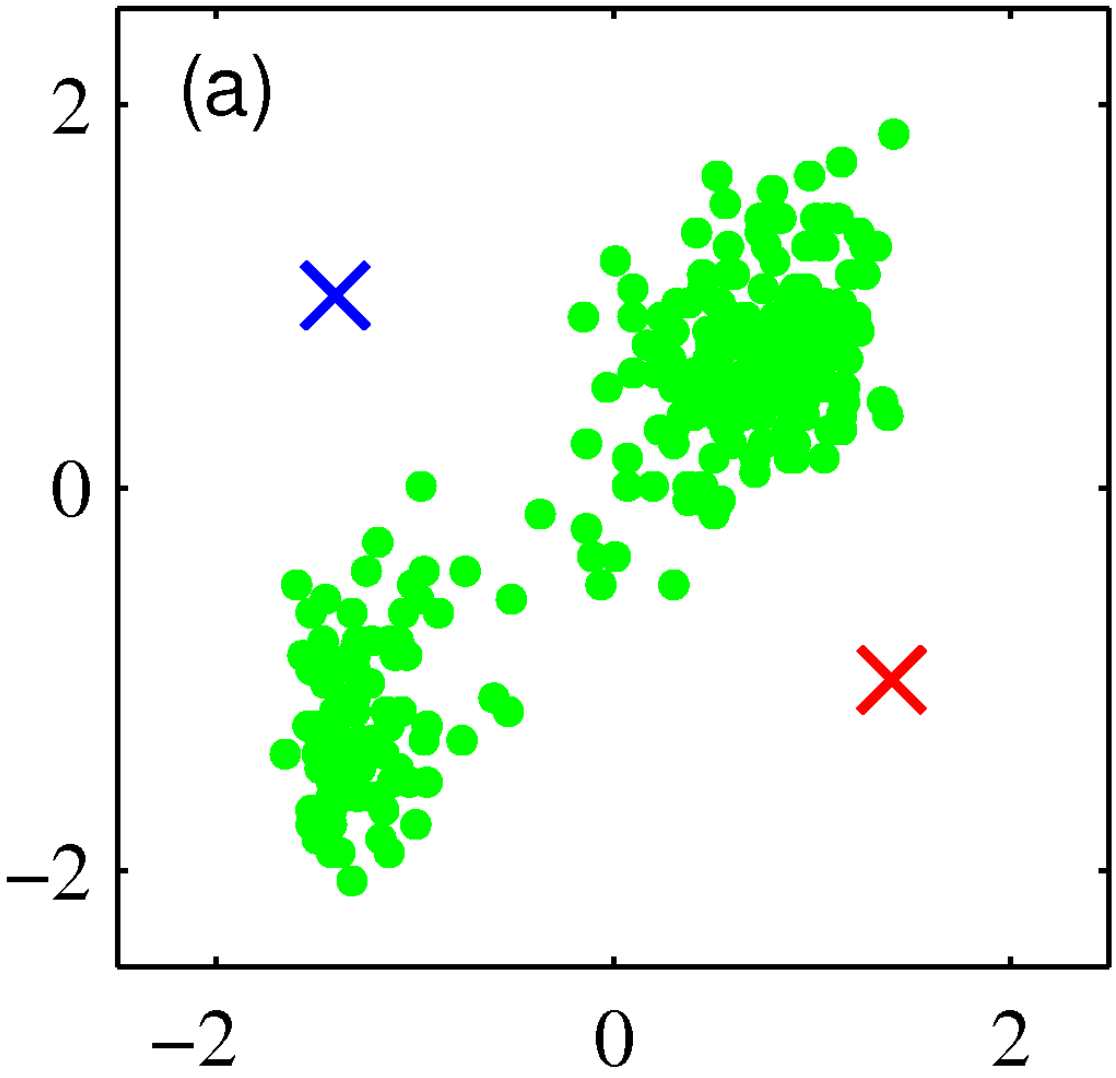
K-means – Algorithm

- Initialization:

  – choose k random positions

  – assign cluster centers $\mu^{(j)}$ to these positions

K-means



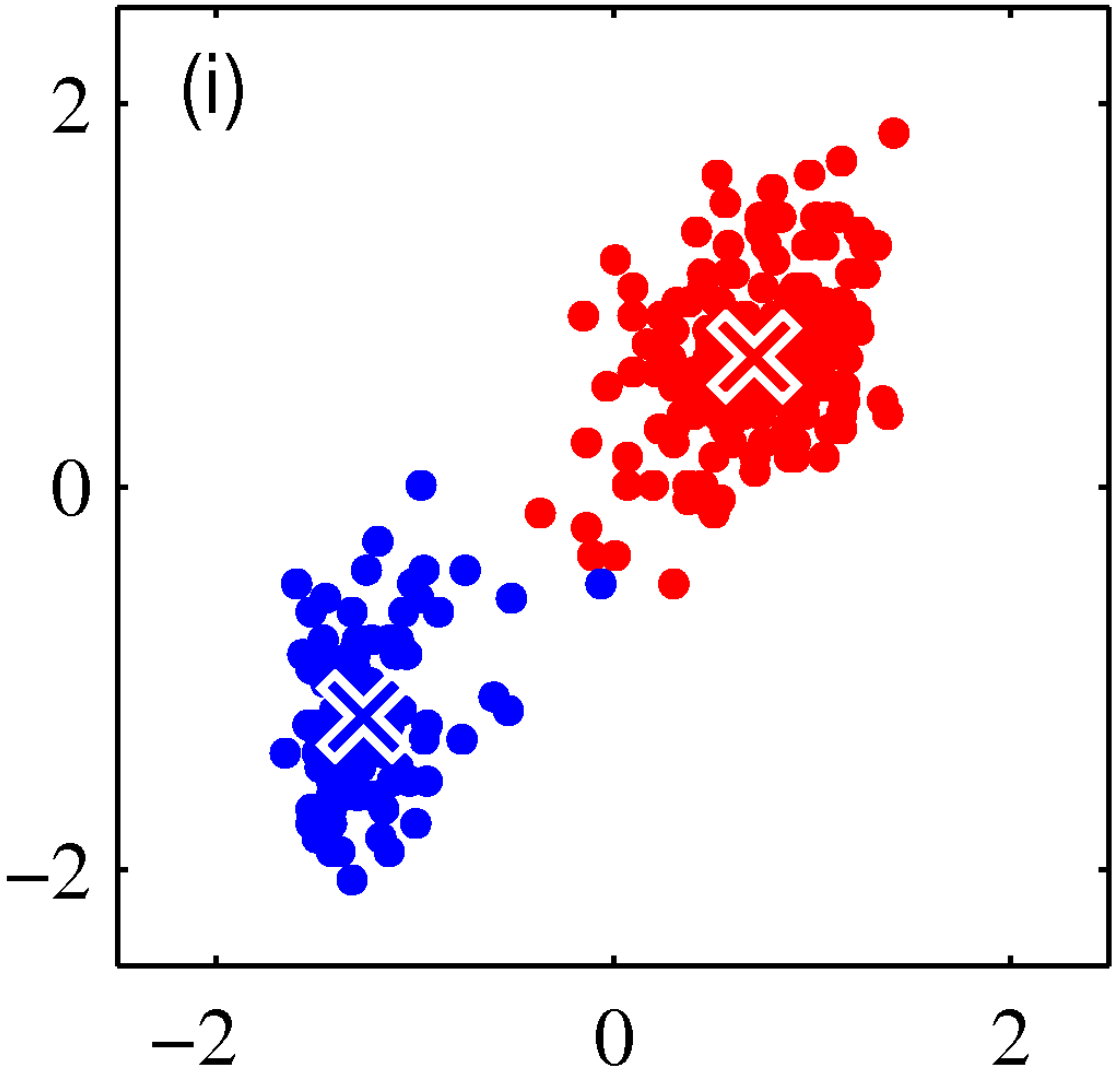
Bishop, “Pattern Recognition and Machine Learning”, Springer, 2006

K-means

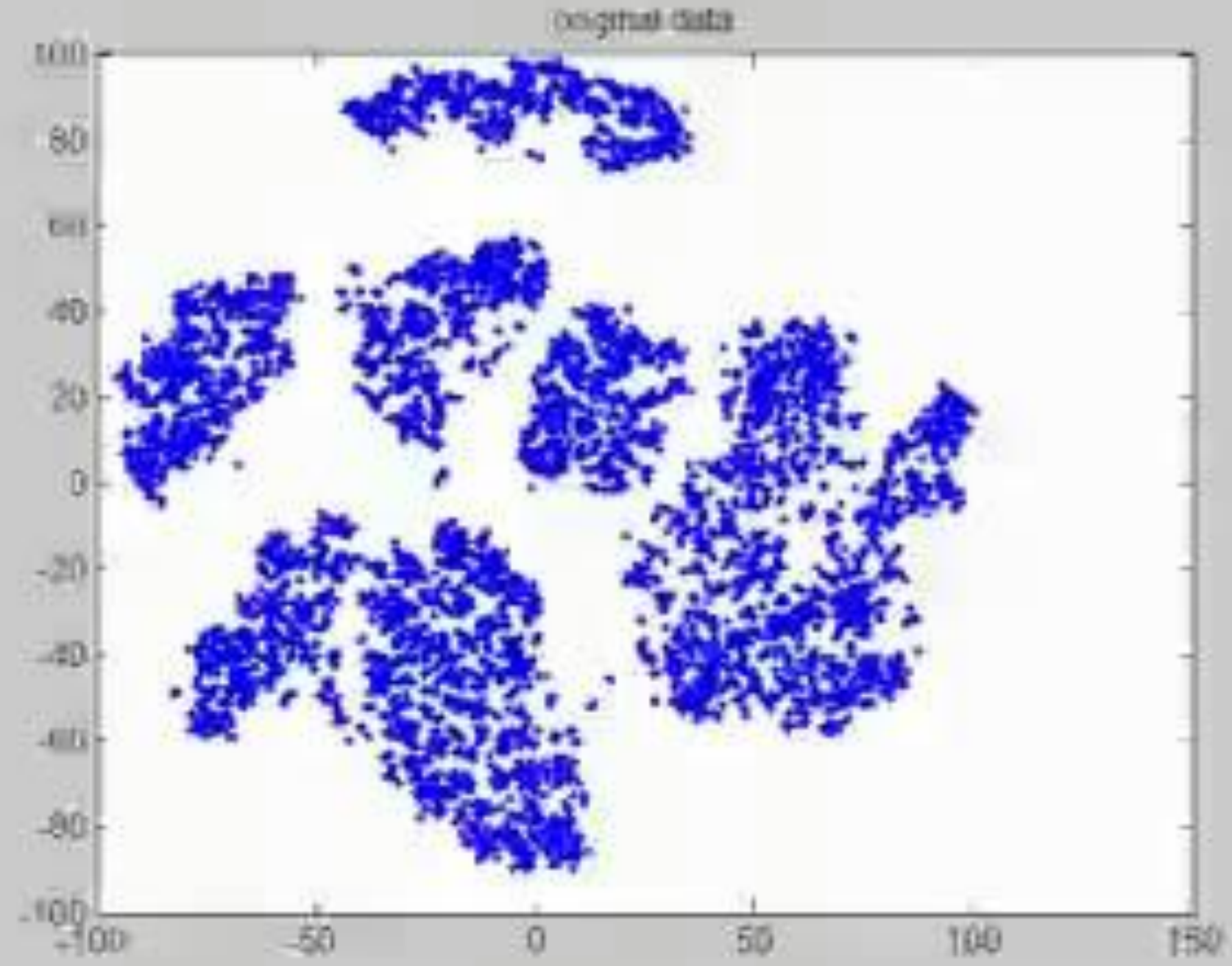
- Until Convergence:
  - Compute distances $||x^{(i)} - \mu^{(j)}||$
  - Assign points to nearest cluster center
  - Update Cluster centers:

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

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

(goal is 'image compression)

Same k = 10

K-means Summary

- Guaranteed to converge
- Result depends on initialization (which might be random).

- Number of clusters is important
  Do You have to specify. Binary is straightforward.
- 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 → see if one answer comes up multiple times.

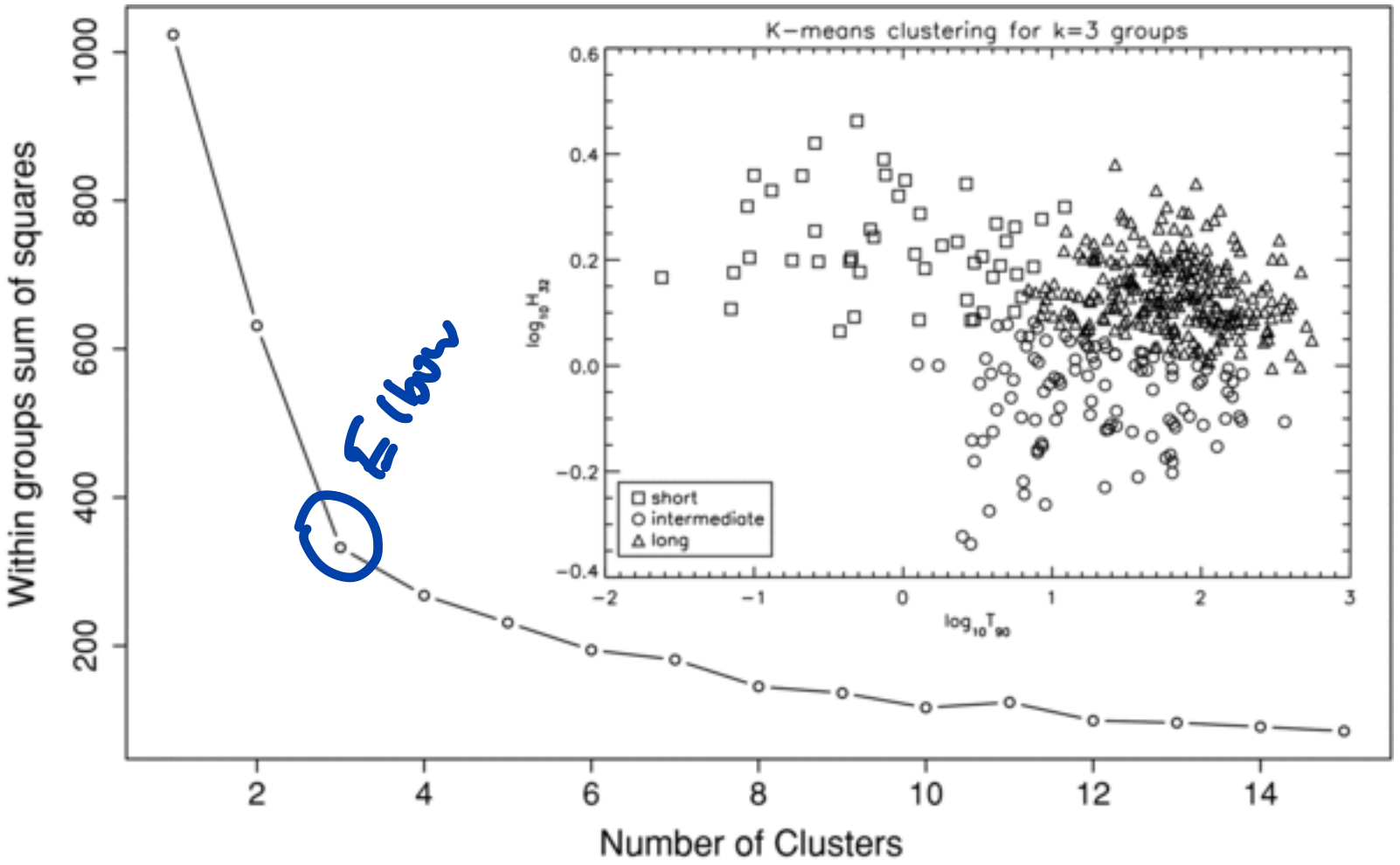
How to find K

- Extreme cases:
  - K=1 → likely won't explain.
  - K=N
- Choose K such that increasing it does not model the data much better.

→ will start to split clusters if K is too high

→ Also is overfitting

“Knee” or “Elbow” method



http://m.iopscience.iop.org/article/10.1088/0004-637X/725/2/1955;jsessionid=9B96FCAF94036AA723BA5F9C78FB33D6.c1.iopscience.cld.iop.org

Cross Validation

- Use this if you want to apply your clustering solution to new unseen data

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

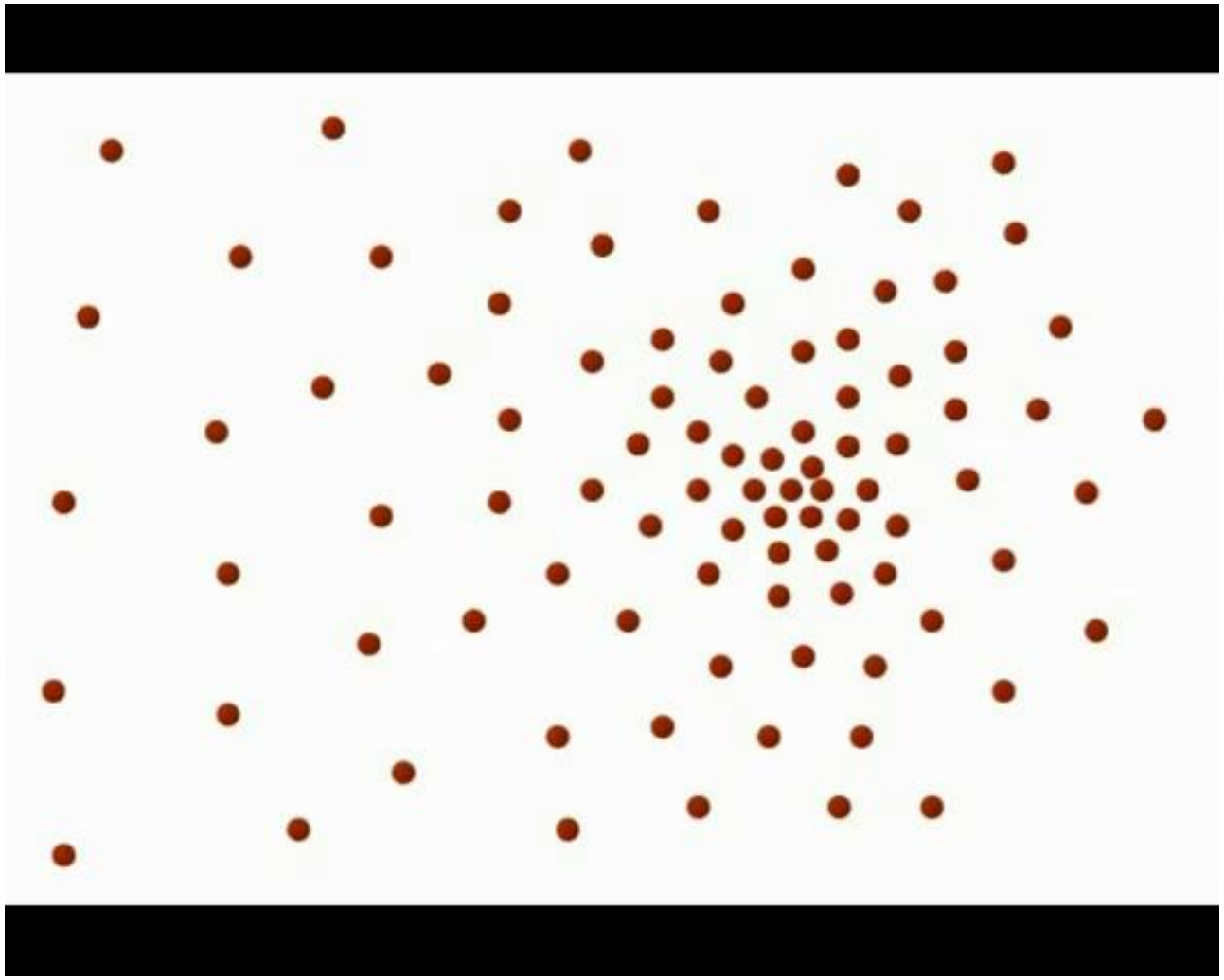
Getting Rid of K

- Having to specify K is annoying
- Can we do without?

Mean Shift

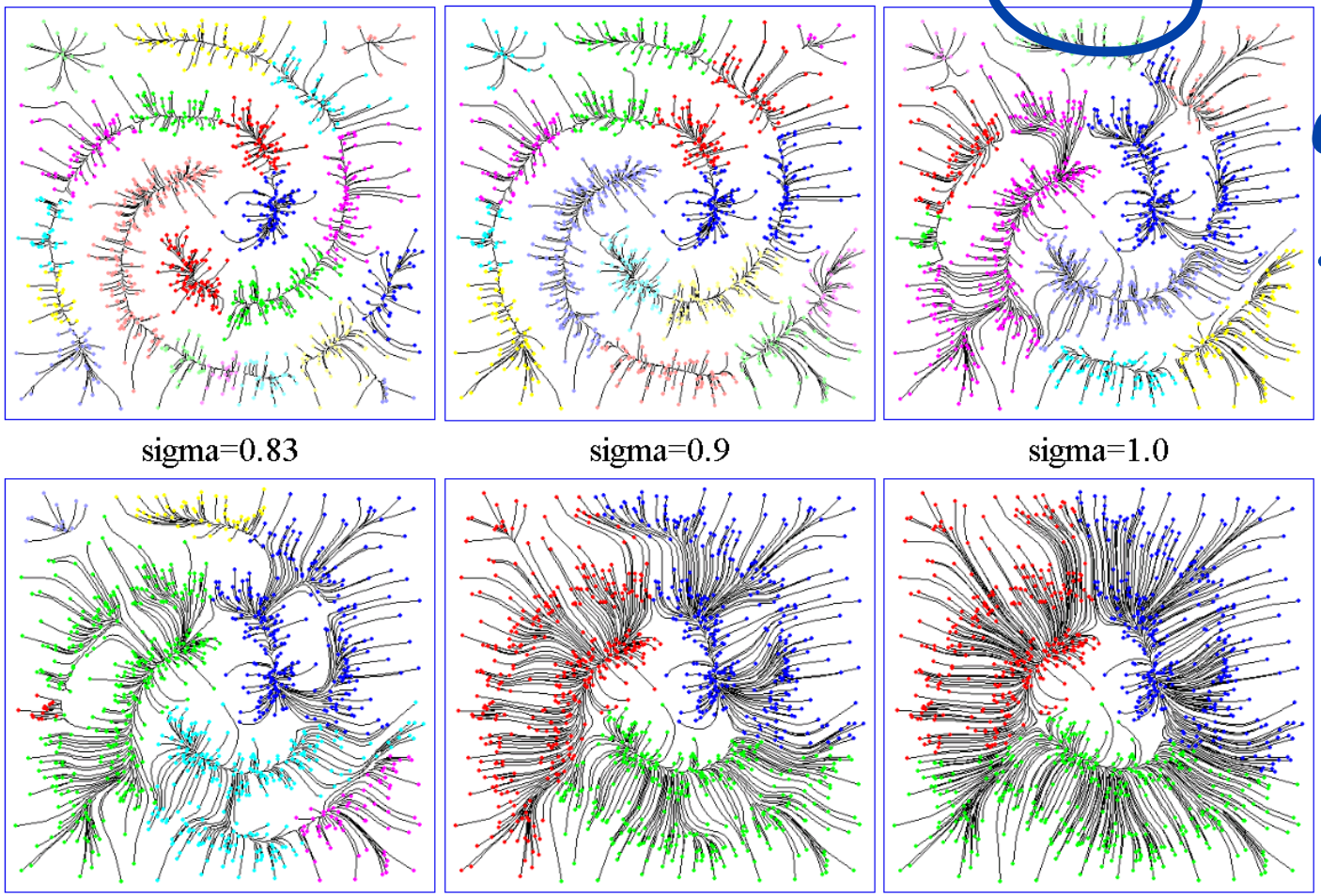
1. Put a window around each point
2. Compute mean of points in the frame.
3. Shift the window to the mean
4. Repeat until convergence

Mean Shift



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

Mean Shift



K=19

window size

# of clusters not specified.

K=3

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/

Removes choosing a random K value in KMeans.

Lot Blc done for every single data point.

Multi-feature object trajectory clustering for video analysis

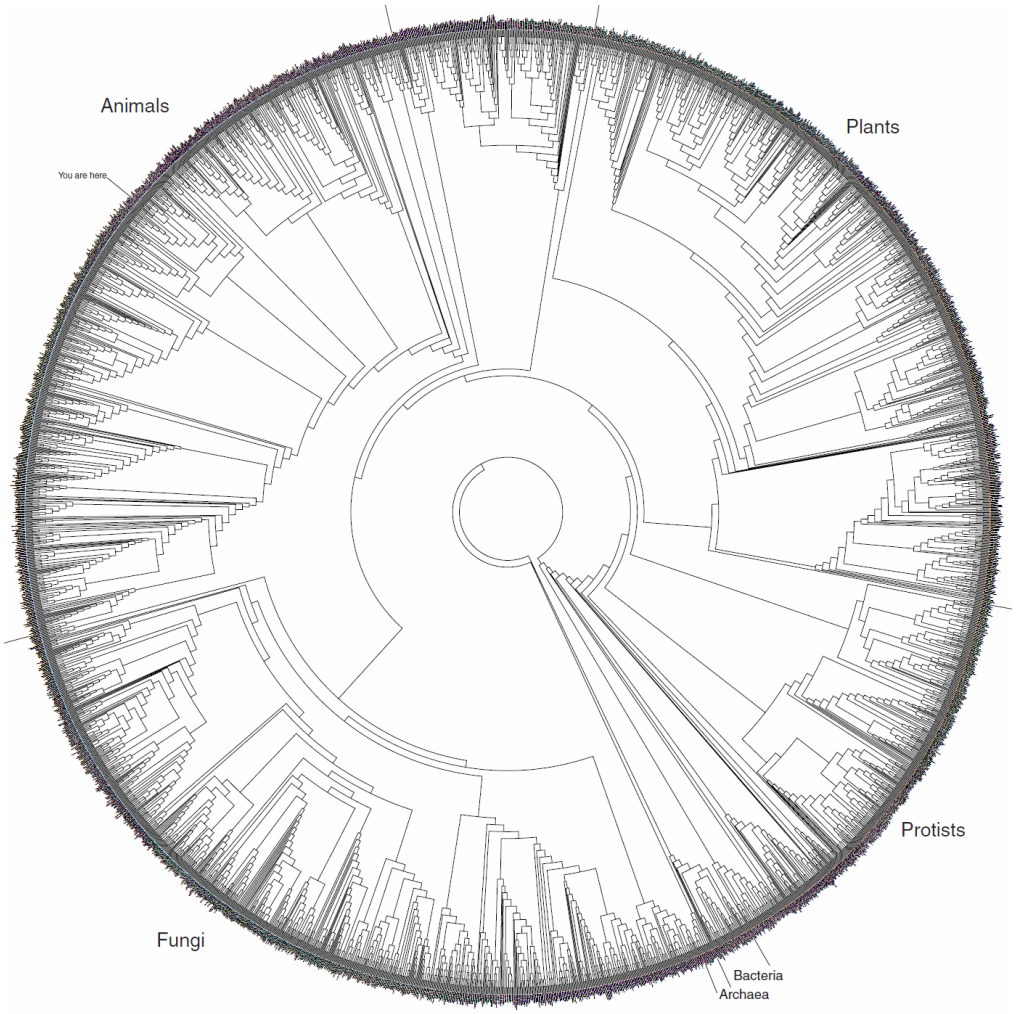
Nadeem Anjum Andrea Cavallaro

Parameters parameters

- For K means we need K and result depends on initialization
- For mean shift we need the window size and a lot of computation

- Hierarchical Clustering keeps a history of all possible cluster assignments

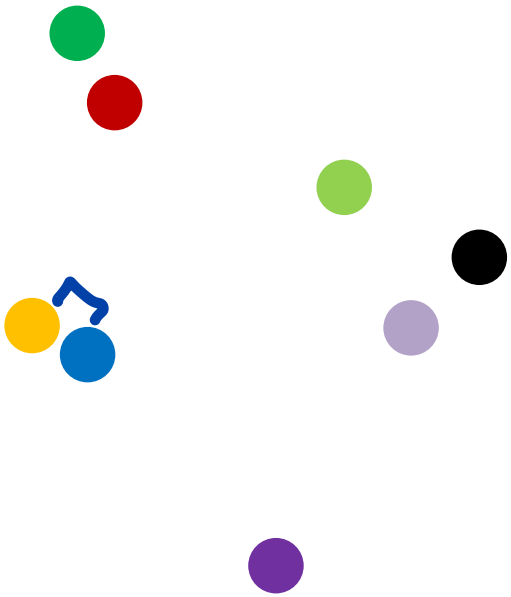
Tree of Life



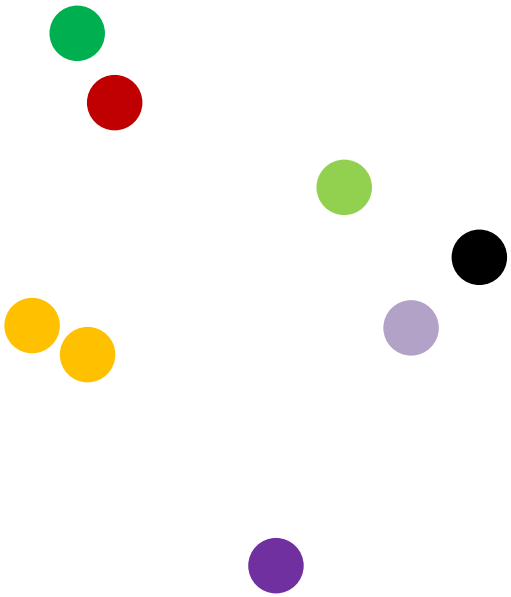
http://www.zo.utexas.edu/faculty/antisense/DownloadfilesToL.html

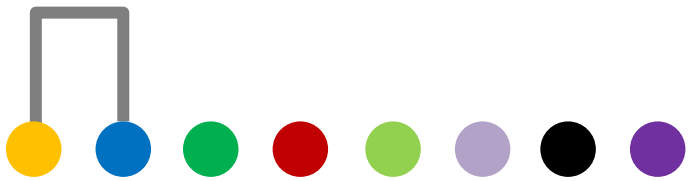
Hierarchical Clustering

Each point is its own cluster to start.

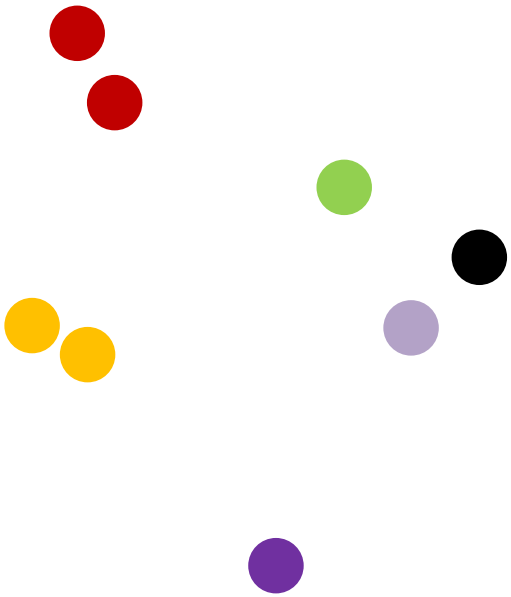


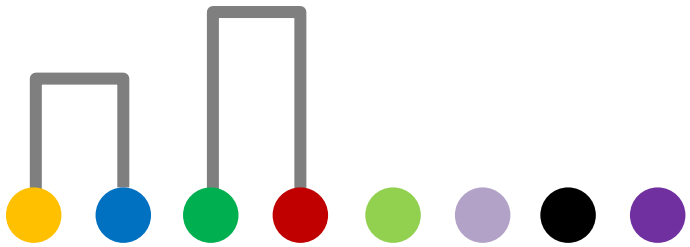
Hierarchical Clustering



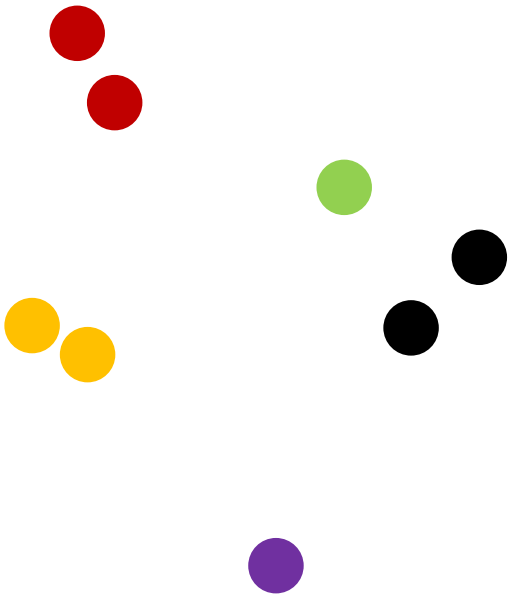


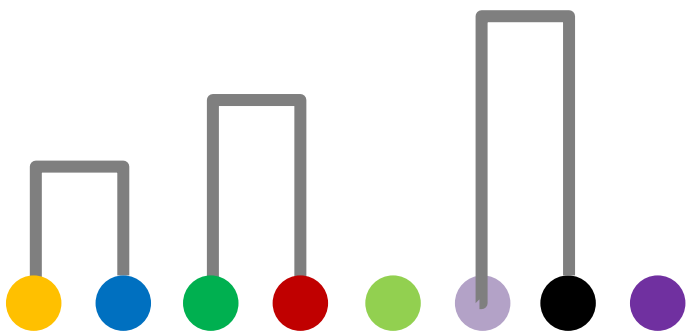
Hierarchical Clustering





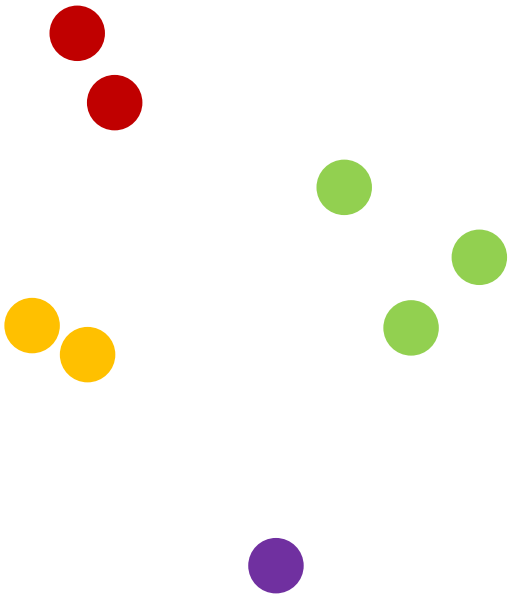
Hierarchical Clustering

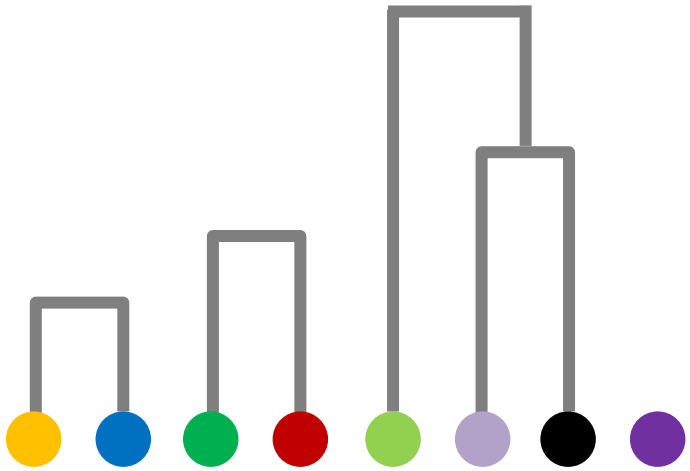




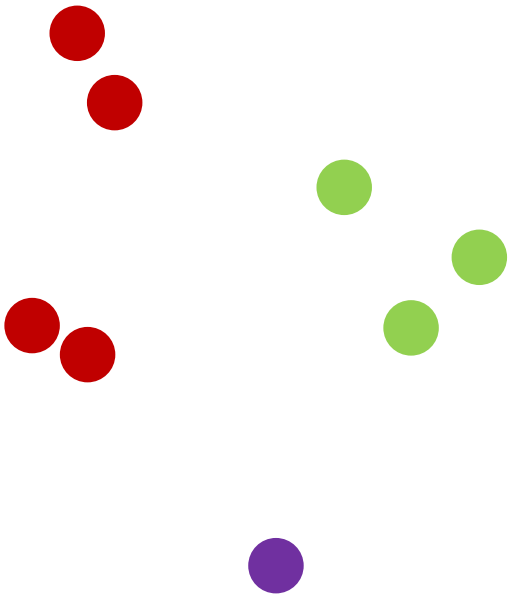
Hierarchical Clustering

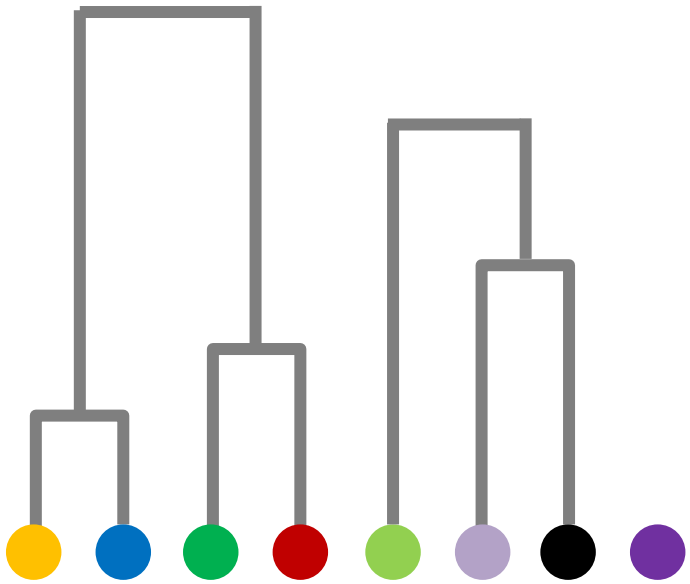
Growing clusters



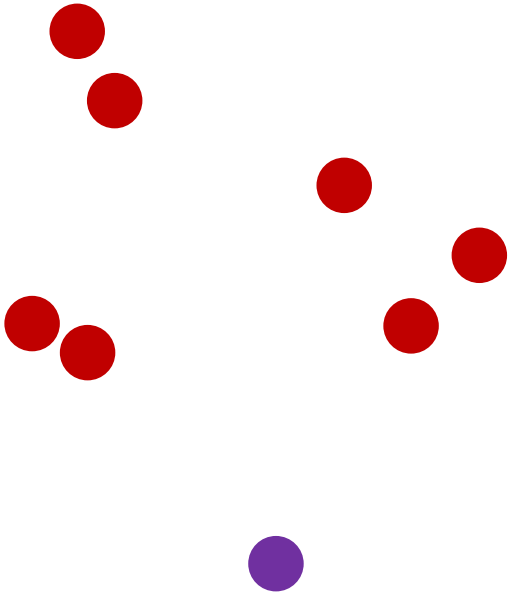


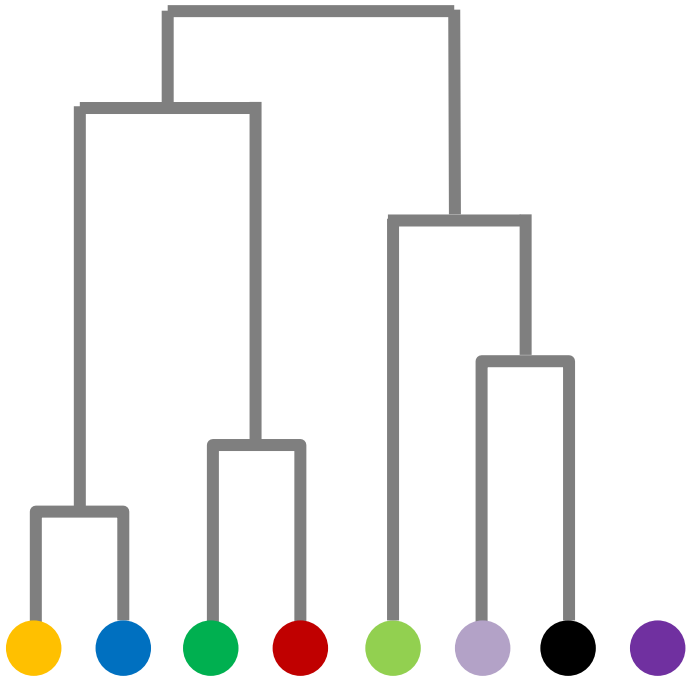
Hierarchical Clustering



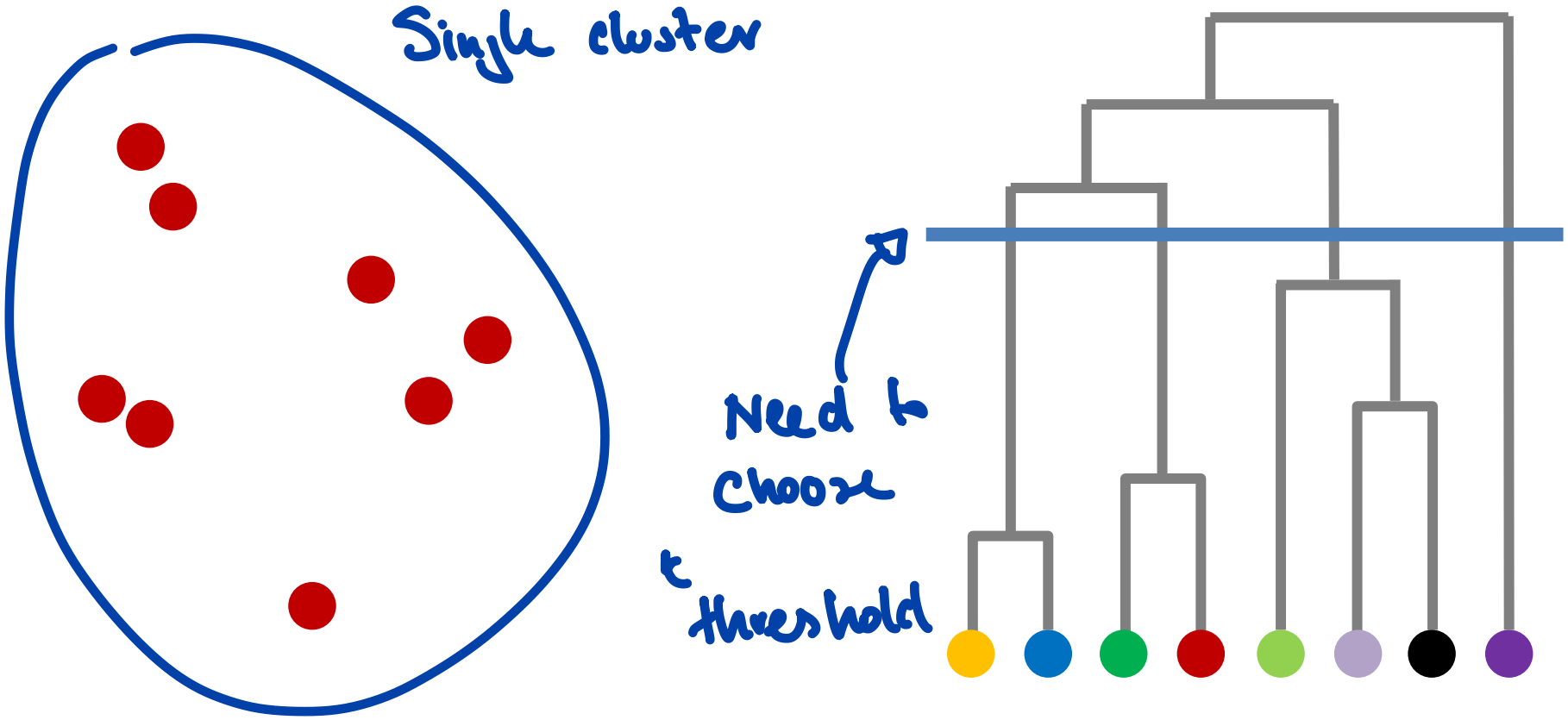


Hierarchical Clustering

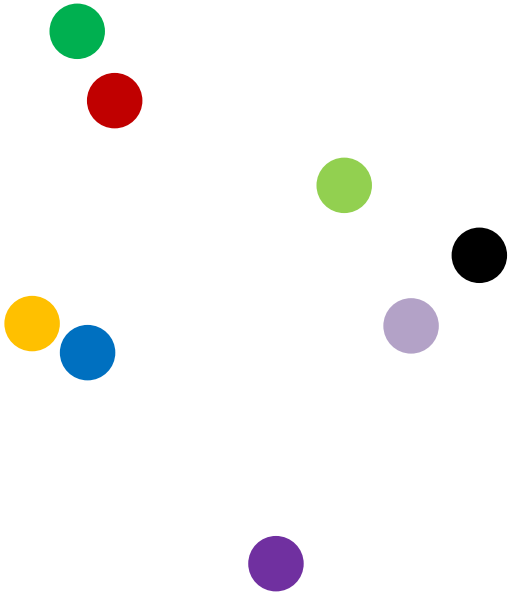


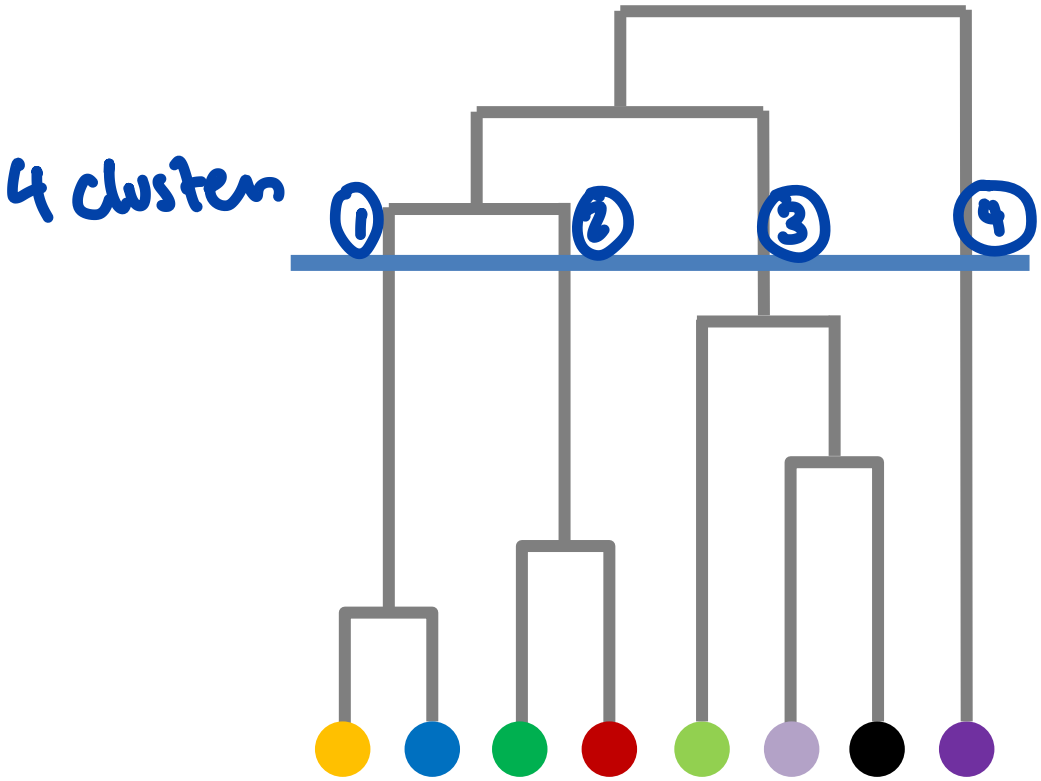


Hierarchical Clustering



Hierarchical Clustering





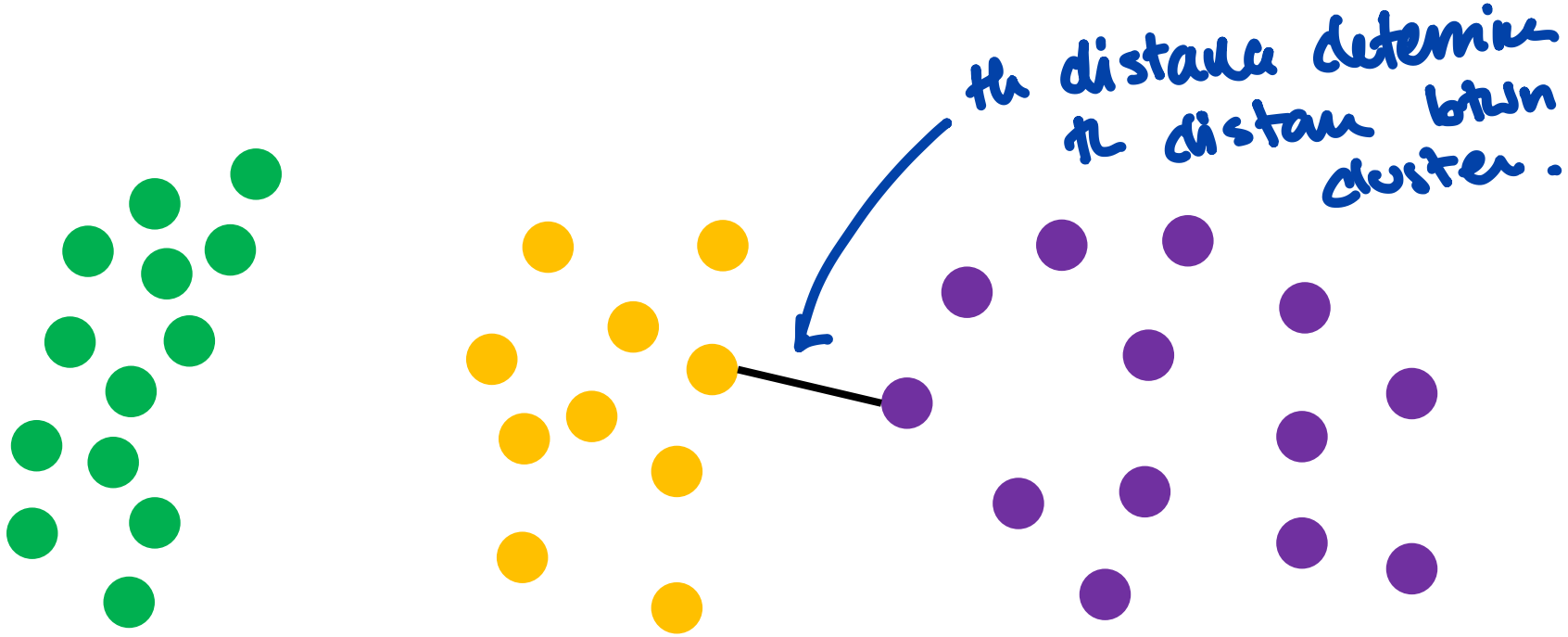
Hierarchical Clustering

- Produces complete structure
- No predefined number of clusters

Can be useful to have no threshold.
↳ like in tree of life.

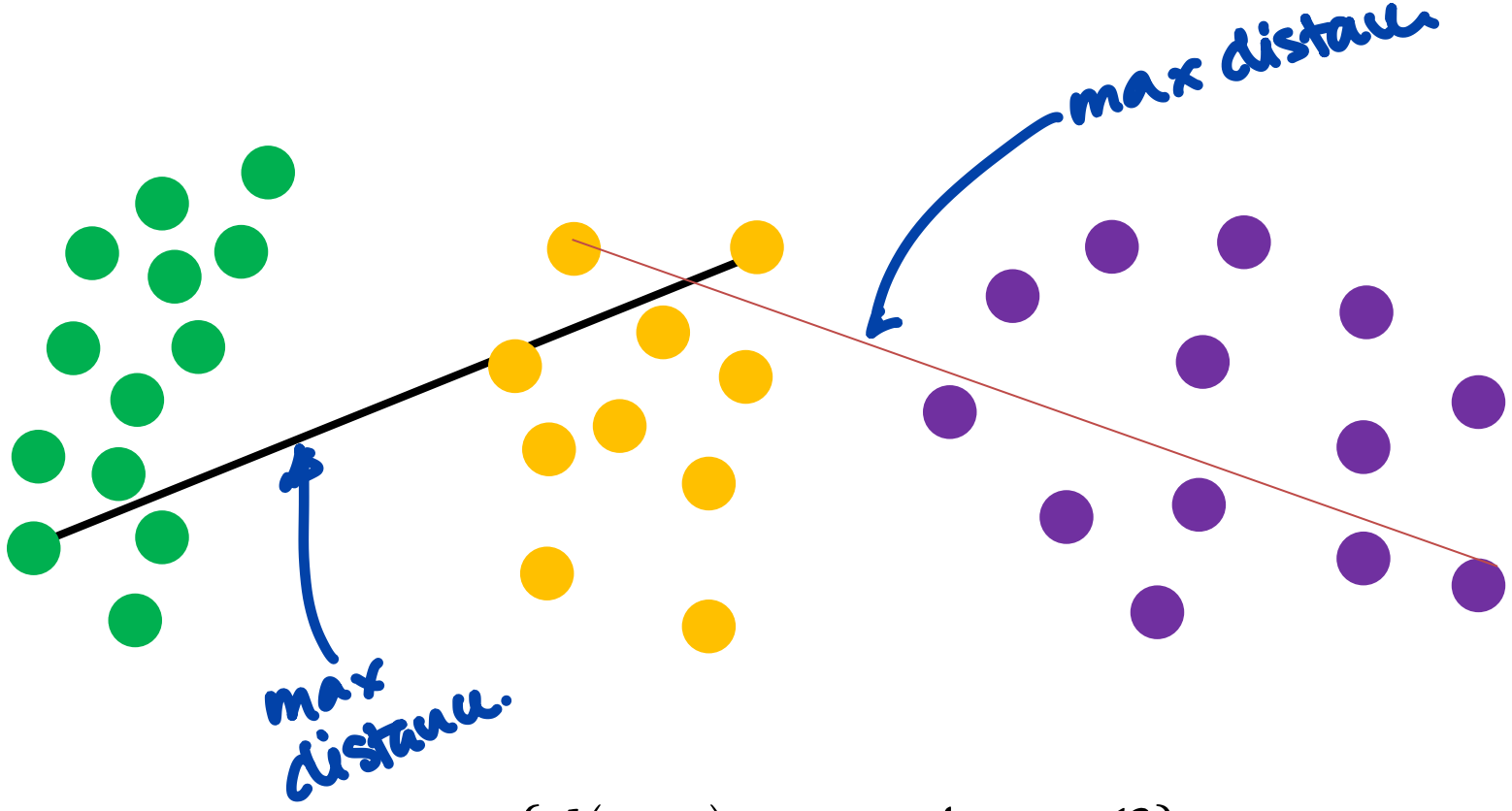
- Similarity between clusters:
  - single-linkage: $\min\{d(x,y) : x \in \mathcal{A}, y \in \mathcal{B}\}$ (As seen w/ dots)
  - complete-linkage: $\max\{d(x,y) : x \in \mathcal{A}, y \in \mathcal{B}\}$
  - average linkage: $\frac{1}{|\mathcal{A}|\cdot|\mathcal{B}|}\sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{B}}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

Single Linkage



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

Complete Linkage

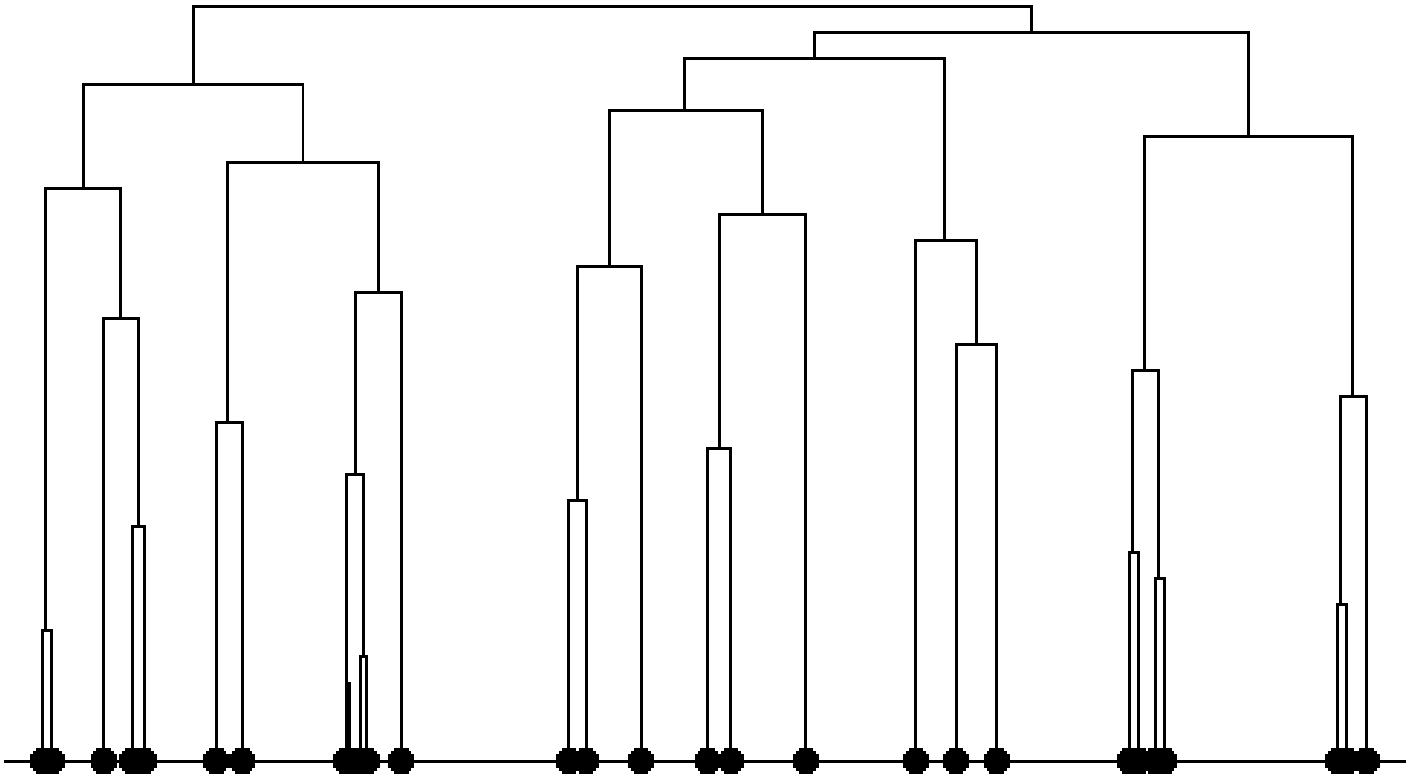


$$\max\{d(x,y) : x \in \mathcal{A}, y \in \mathcal{B}\}$$

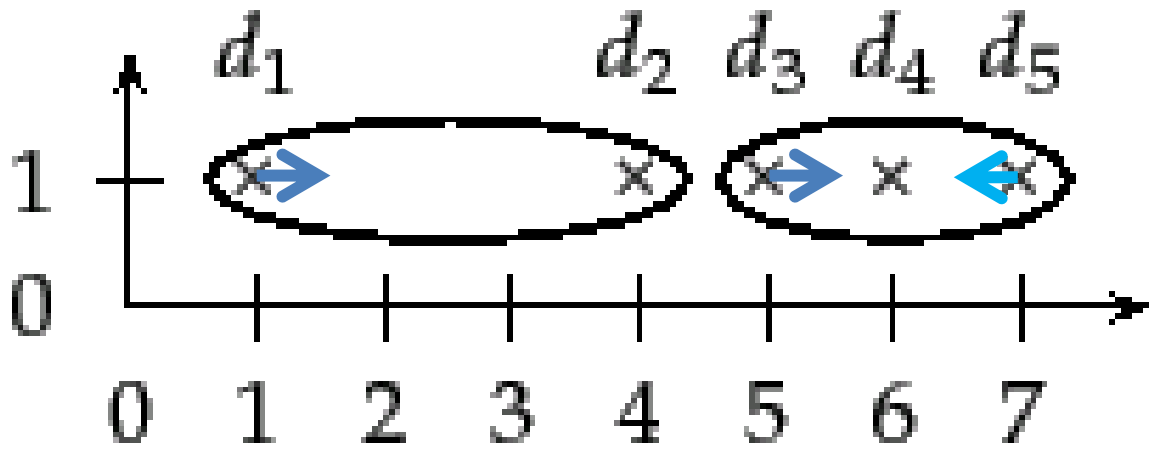
Linkage Matters

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

Chaining Phenomenon



Outlier Sensitivity



In this case, single linkage is better.

+ 2*epsilon

- 1*epsilon

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

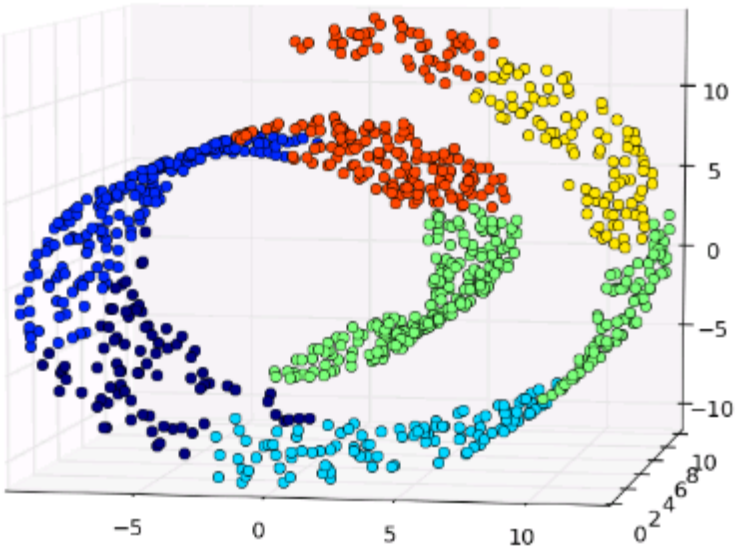
Efficient Hierarchical Graph-Based Video Segmentation

Matthias Grundmann[1,2], Vivek Kwatra[2],
Mei Han[2] and Irfan Essa[1]

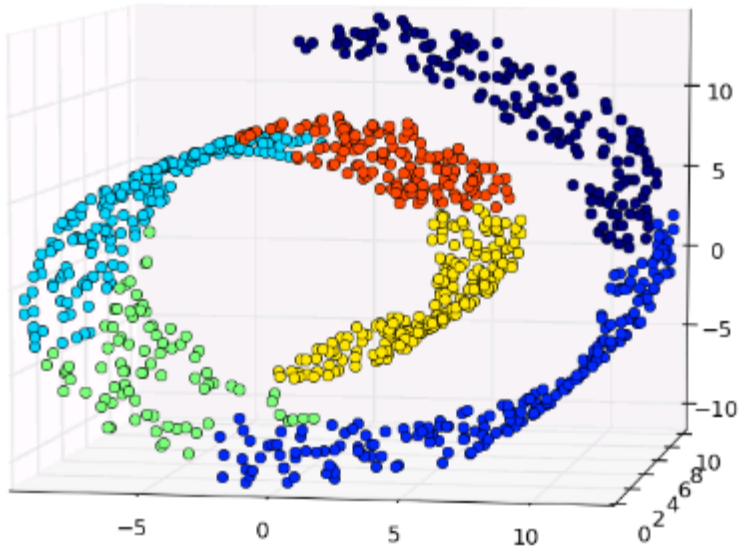
[1]Georgia Tech [2]Google Research

IEEE CVPR, San Francisco, USA, June 2010

Swiss Role Problem



without connectivity constraints



with connectivity constraints

only adjacent clusters can be merged together

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

Evaluation Criteria

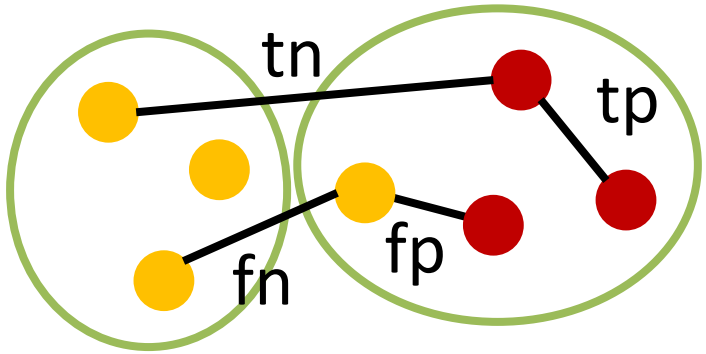
Very hard for unsupervised in general.

- Based on expert knowledge
- Debatable for real data
- Hidden Unknown structures could be present
- Do we even want to just reproduce known structure?

Rand Index

- Percentage of correct classifications
- Compare pairs of elements:

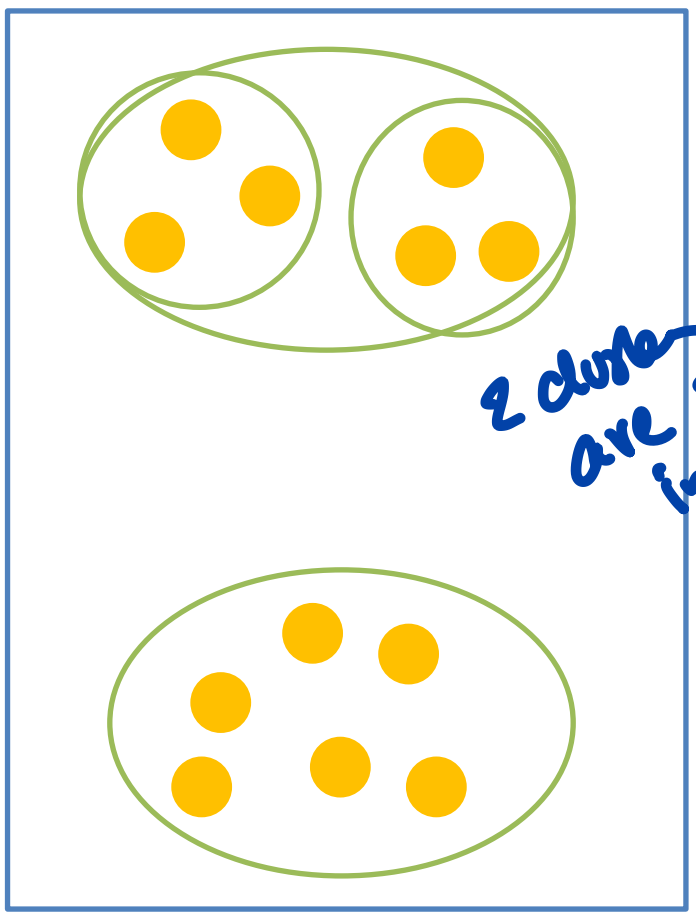
$$R = \frac{tp+tn}{tp+tn+fp+fn}$$



- Fp and fn are equally weighted

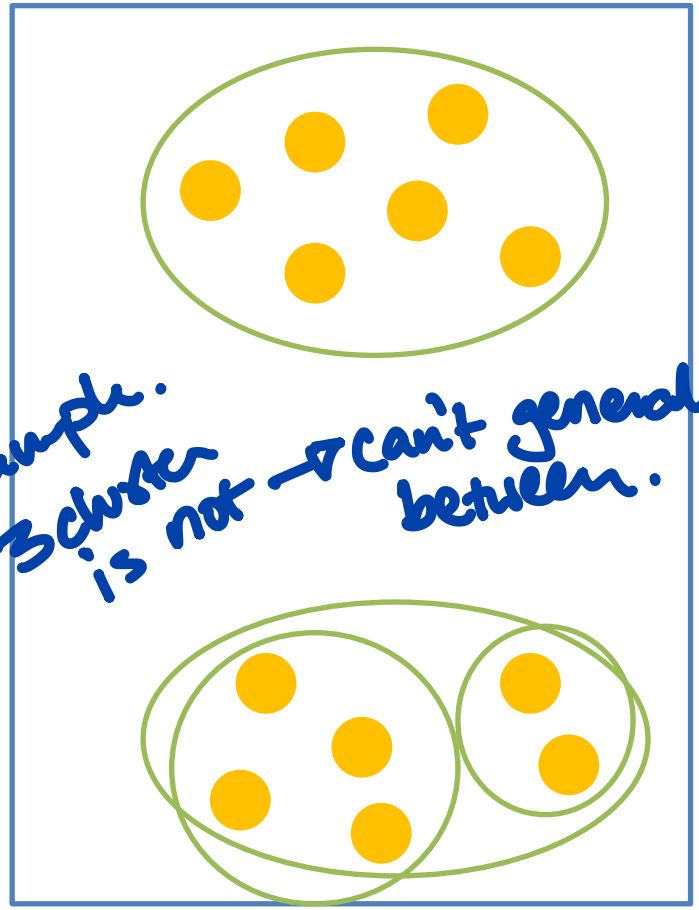
Stability

A good cluster solution should generalize



2 clusters are stable in this example.

3 cluster is not → can't generalize between.



Stability

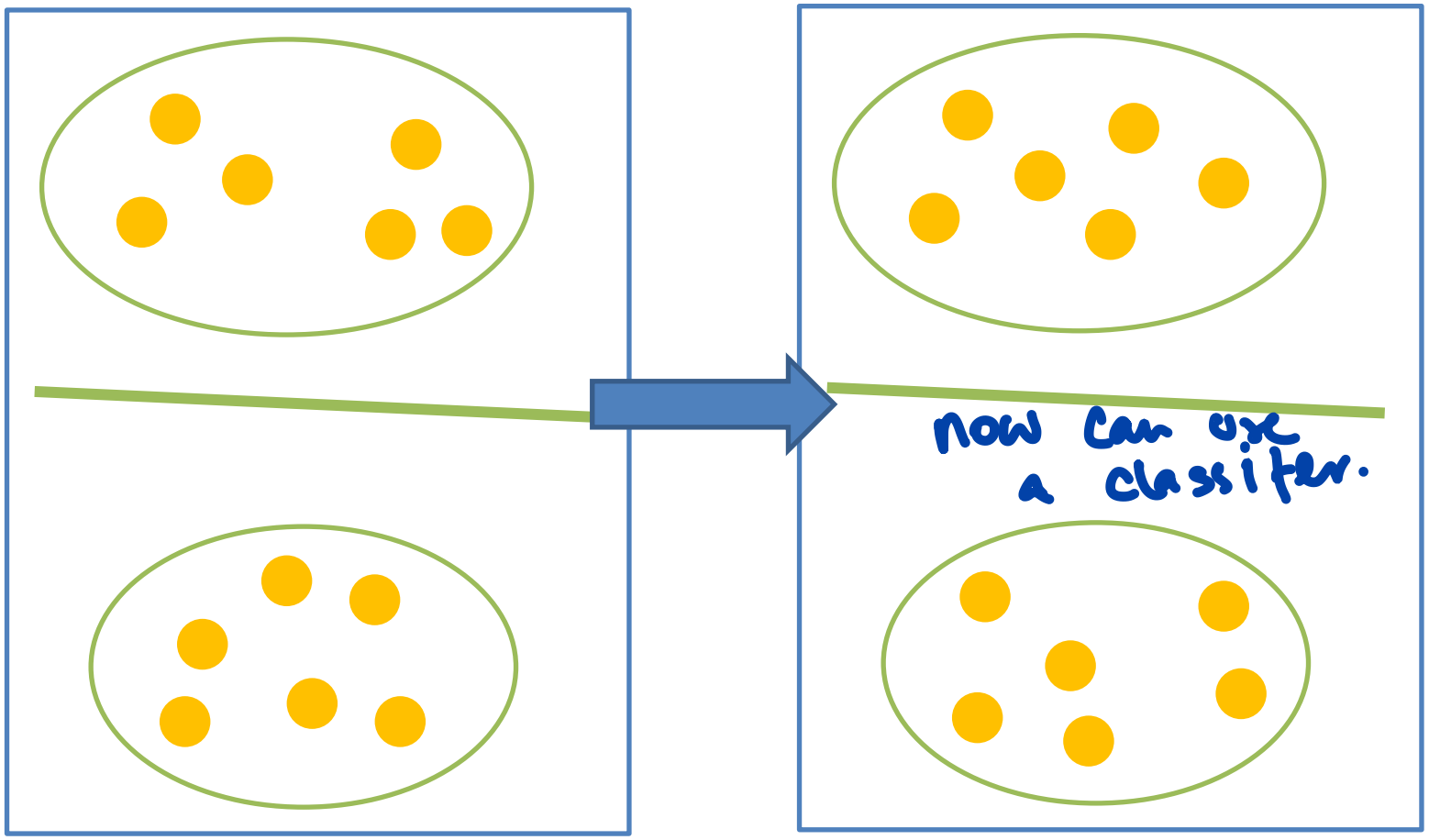
Can turn into a supervised set. If stable, the error should be small.

- What is the right number of clusters?
- What makes a good clustering solution?

Use clustering to turn an unsupervised to supervised.

- Clustering should generalize!

Stability



Summary

- We have covered a lot today
- Clustering
  - K-means
  - Mean-shift
  - Hierarchical clustering
- Evaluation criteria
  - Rand index
  - Stability