

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:

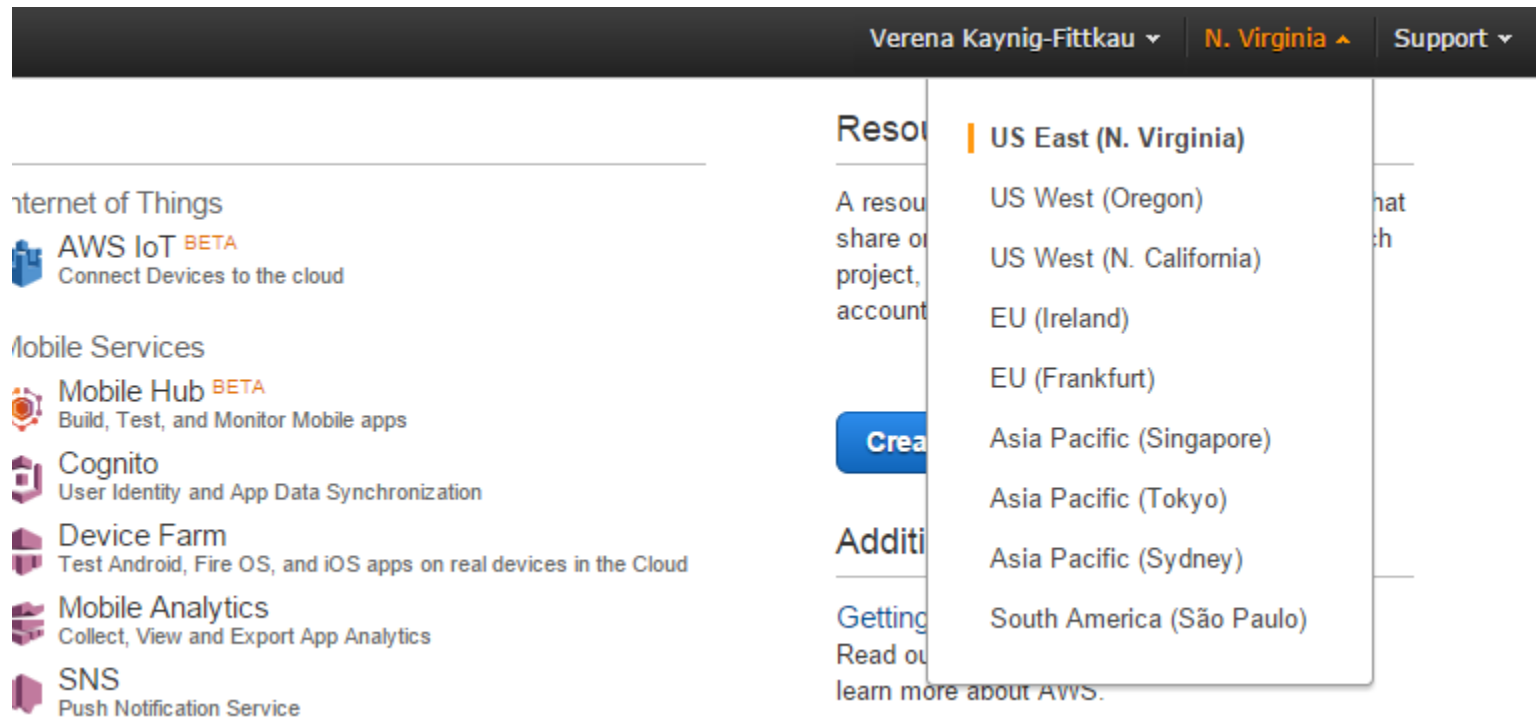
<https://piazza.com/class/icf0cypdc3243c?cid=1369>

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




The screenshot shows the AWS Management Console interface. At the top, a dark navigation bar contains the user name 'Verena Kaynig-Fittkau', the current region 'N. Virginia' (highlighted in orange), and a 'Support' link. On the left, a sidebar lists various AWS services under categories like 'Internet of Things' and 'Mobile Services'. The main content area is partially visible, showing a 'Resources' section. A dropdown menu is open, displaying a list of AWS regions: 'US East (N. Virginia)' (selected with an orange bar), 'US West (Oregon)', 'US West (N. California)', 'EU (Ireland)', 'EU (Frankfurt)', 'Asia Pacific (Singapore)', 'Asia Pacific (Tokyo)', 'Asia Pacific (Sydney)', and 'South America (São Paulo)'. Below the region list, there is a 'Getting started' link and a note to 'Read on to learn more about AWS.'.


Verena Kaynig-Fittkau ▾ N. Virginia ▲ Support ▾


Internet of Things


 **AWS IoT** BETA
Connect Devices to the cloud


Mobile Services

 **Mobile Hub** BETA
Build, Test, and Monitor Mobile apps

 **Cognito**
User Identity and App Data Synchronization

 **Device Farm**
Test Android, Fire OS, and iOS apps on real devices in the Cloud

 **Mobile Analytics**
Collect, View and Export App Analytics

 **SNS**
Push Notification Service

Resources

A resource that you can share or use in your project, account, or organization.

Create new resource

Additional resources

Getting started

Read on to learn more about AWS.

- US East (N. Virginia)
- US West (Oregon)
- US West (N. California)
- EU (Ireland)
- EU (Frankfurt)
- Asia Pacific (Singapore)
- Asia Pacific (Tokyo)
- Asia Pacific (Sydney)
- South America (São Paulo)

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>

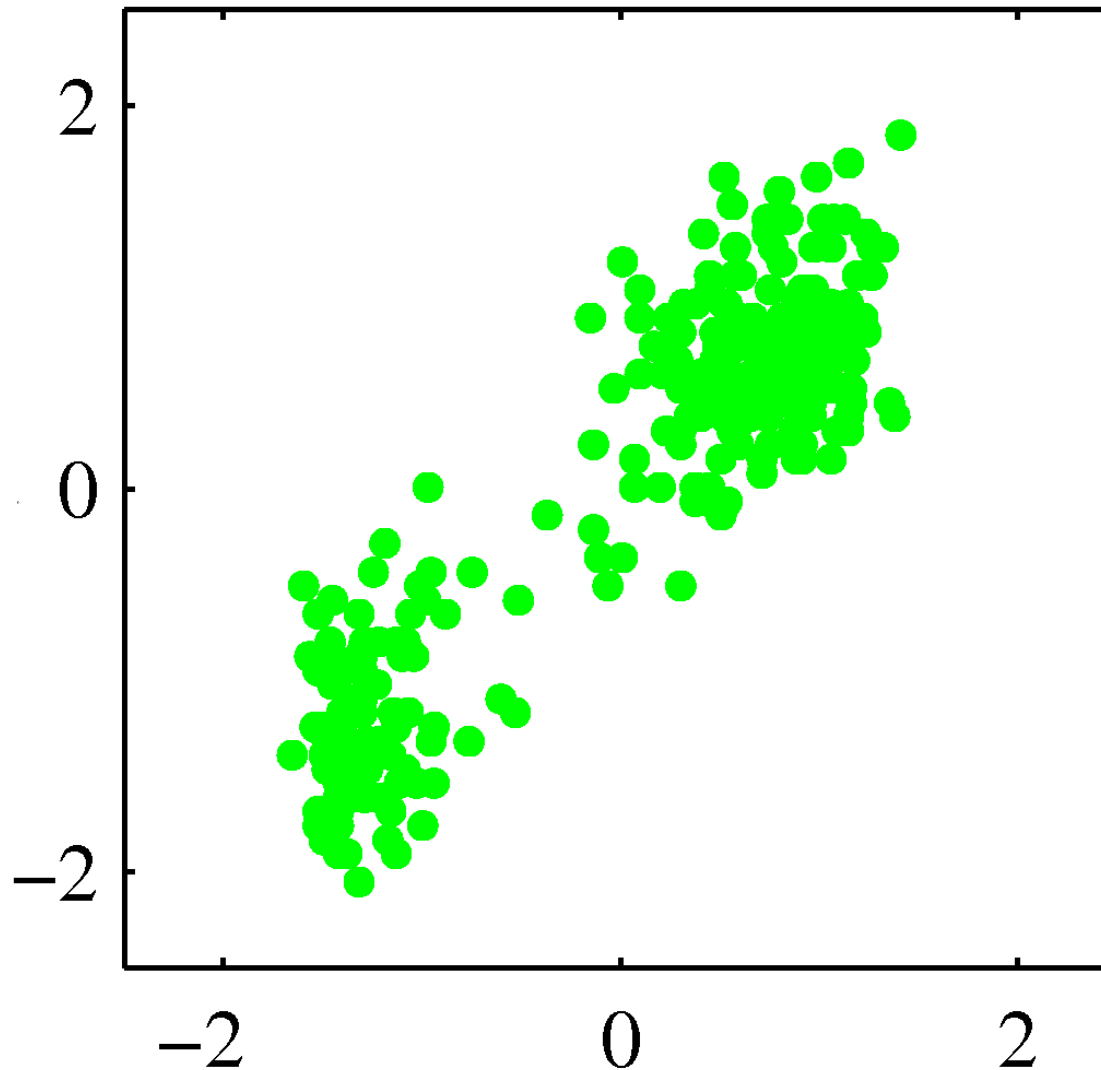
Final Project Proposal

- 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>

Where before we had the y , or 'labels', now we don't have any of that and the task becomes much more difficult, because we cannot use the y /label to guide us defining the hyperplane.

Unsupervised Setting



Bishop, "Pattern Recognition and Machine Learning", Springer, 2006

Unsupervised Learning

- 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

In clustering, you're trying to find a pattern that you don't already know ahead of time.

- 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

In this scenario, some institutions throw out nearly all of their information and only keep the outlier data or only the significant events data.

Unsupervised Learning

- K-means
- Mean-shift
- Hierarchical Clustering

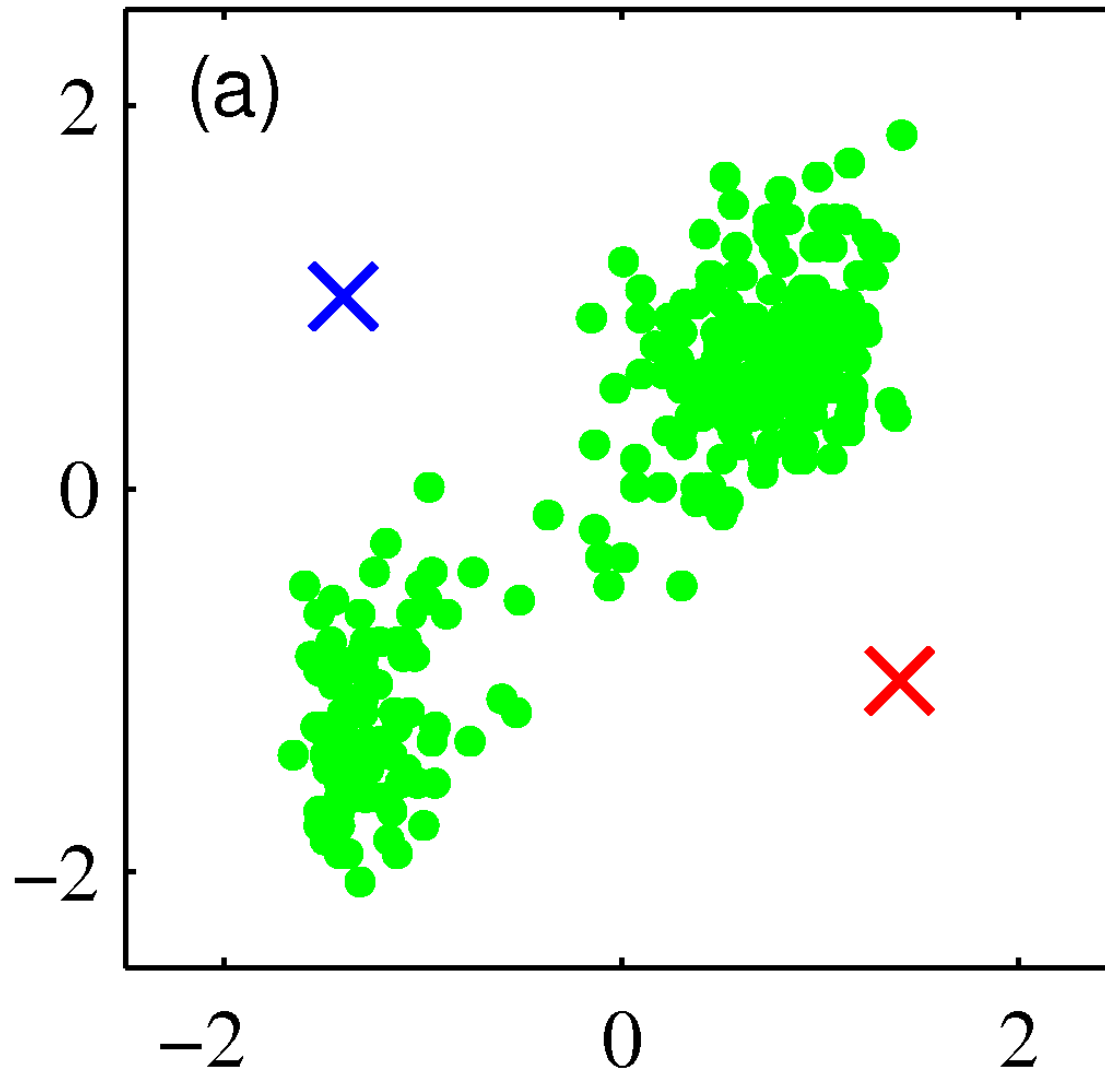
- Rand index, stability

Because we don't have y labels, this is how to evaluate how well the above methods performed.

K-means – Algorithm

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

K-means



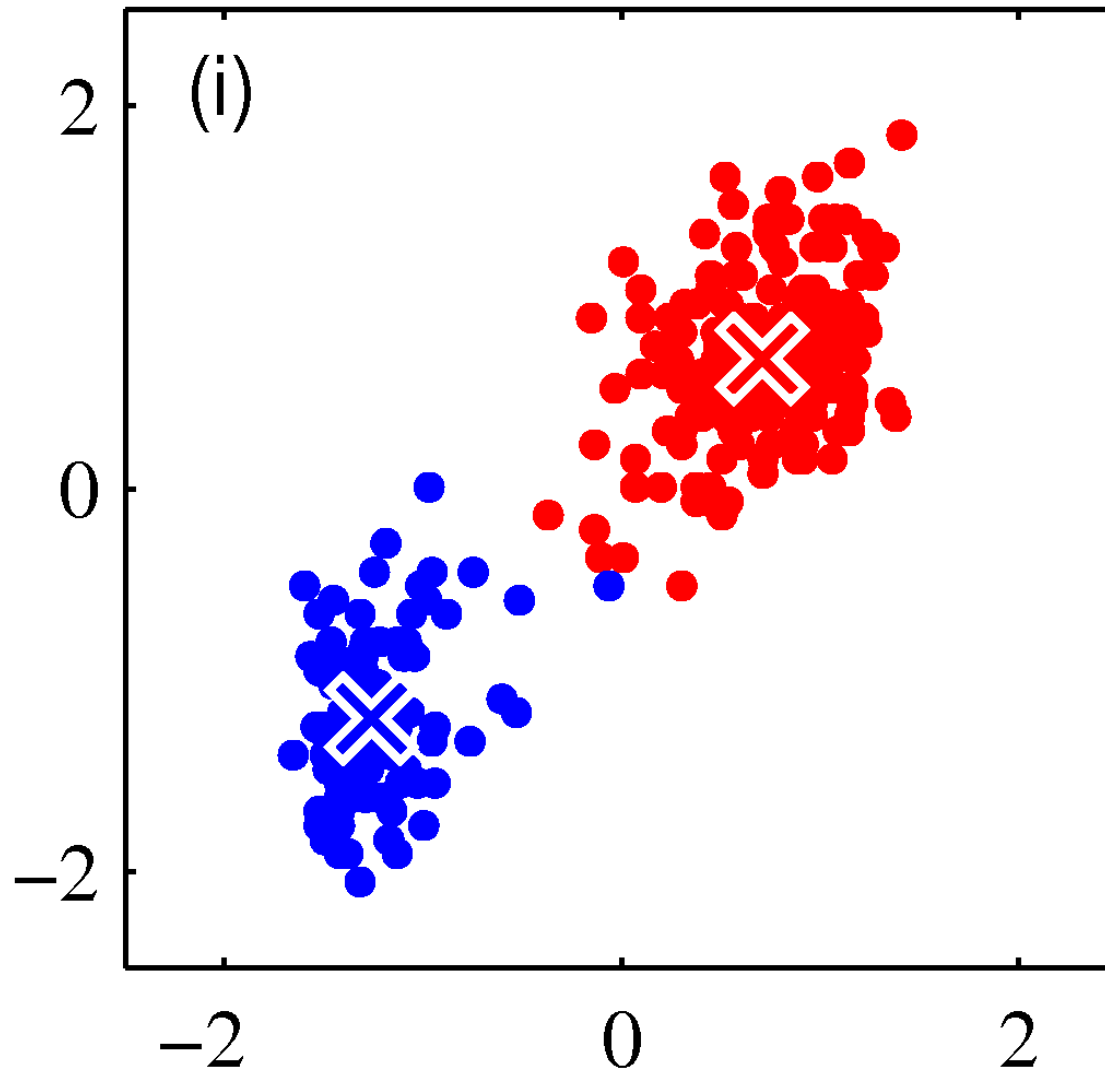
Bishop, "Pattern
Recognition and
Machine
Learning",
Springer, 2006

K-means

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

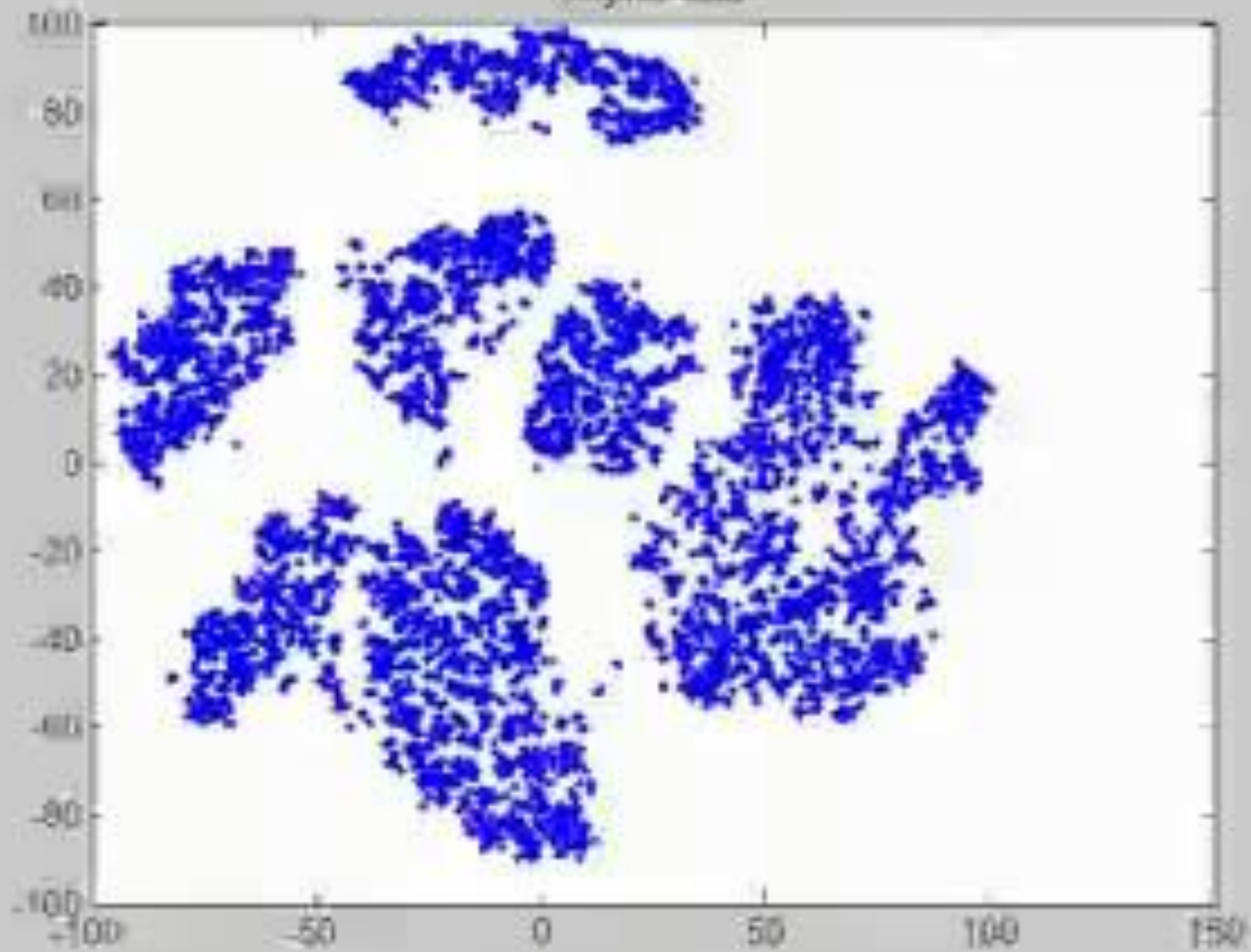
$$\mu^{(j)} = \frac{1}{N_j} \sum_{x_i \in C_j} x_i$$

K-means



Bishop, "Pattern
Recognition and
Machine
Learning",
Springer, 2006

(original data)



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

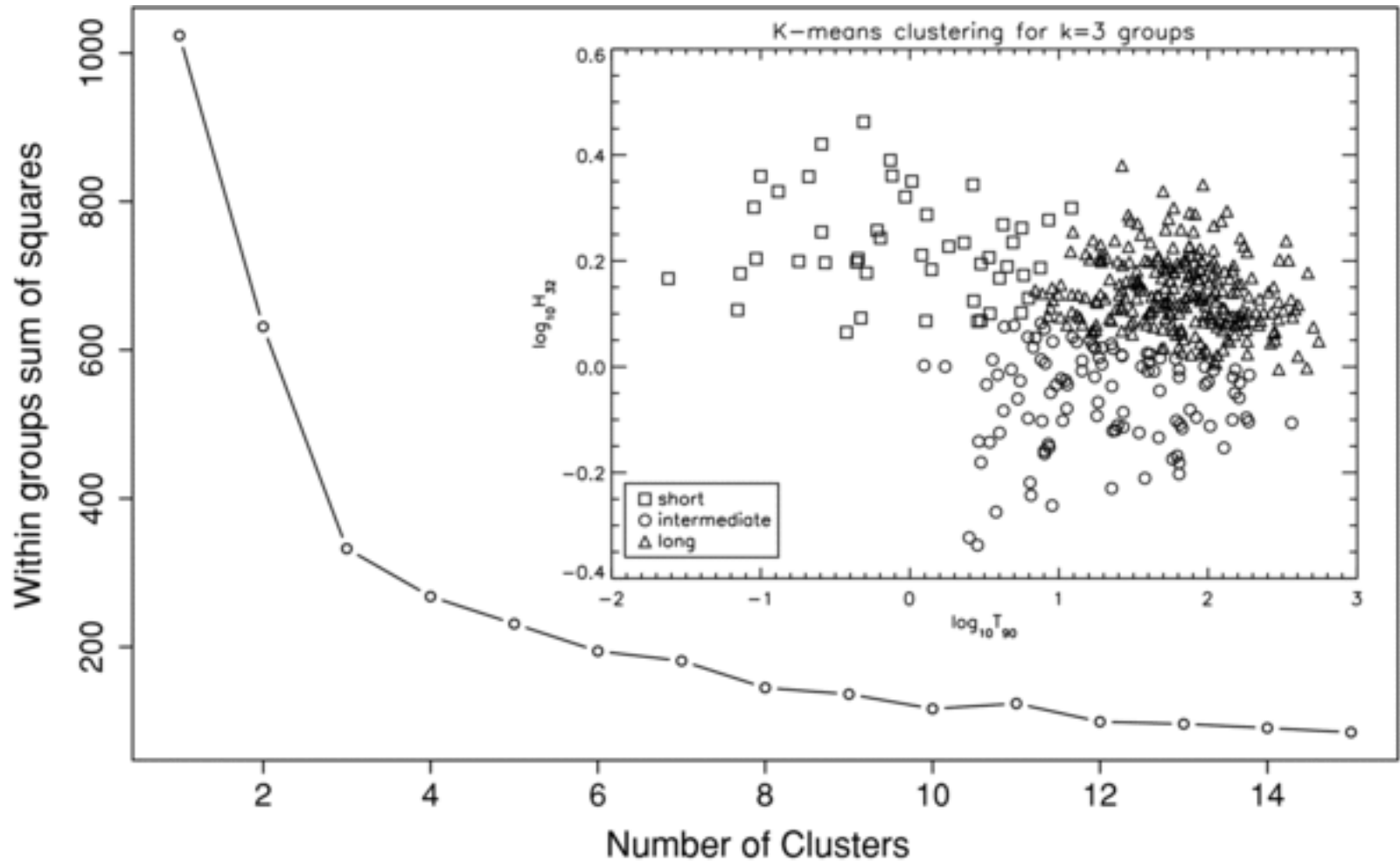
Initialization Methods

- Random Positions
- Random data points as Centers
- Random Cluster assignment to data points
- Start several times

How to find K

- Extreme cases:
 - $K=1$
 - $K=N$
- Choose K such that increasing it does not model the data much better.

“Knee” or “Elbow” method



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

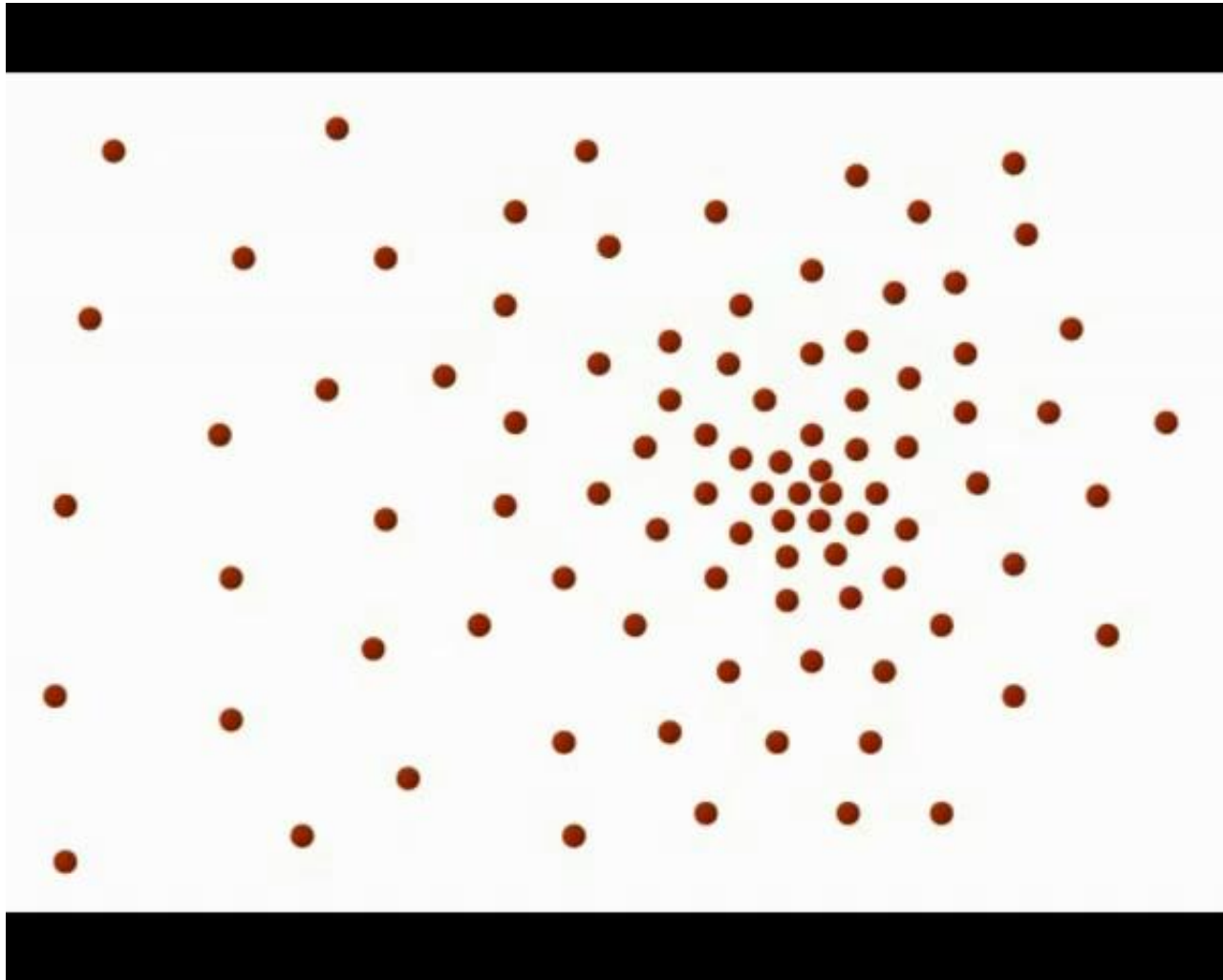
Getting Rid of K

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

Mean Shift

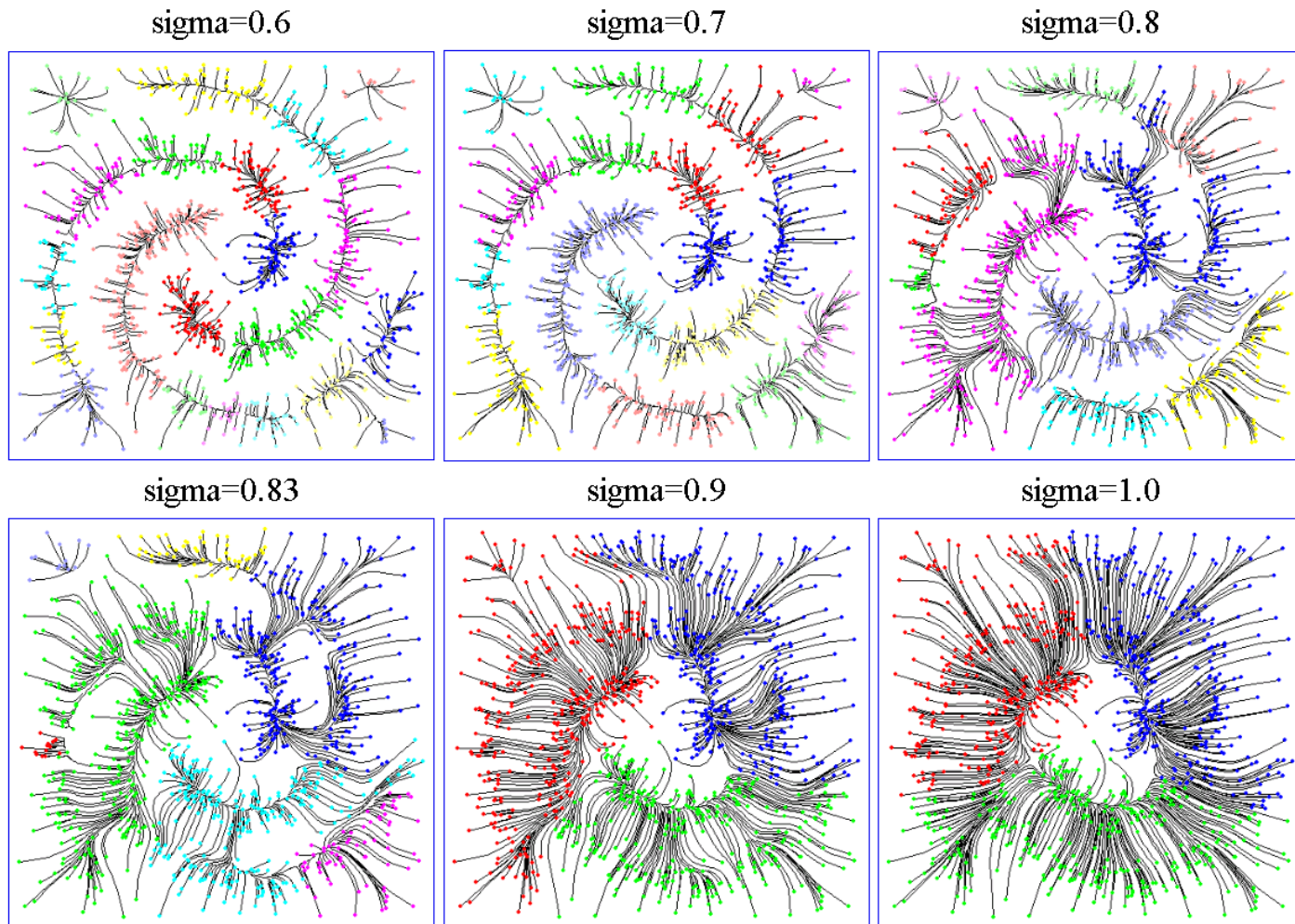
1. Put a window around each point
2. Compute mean of points in the frame.
3. Shift the window to the mean
4. Repeat until convergence

Mean Shift



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

Mean Shift



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/>

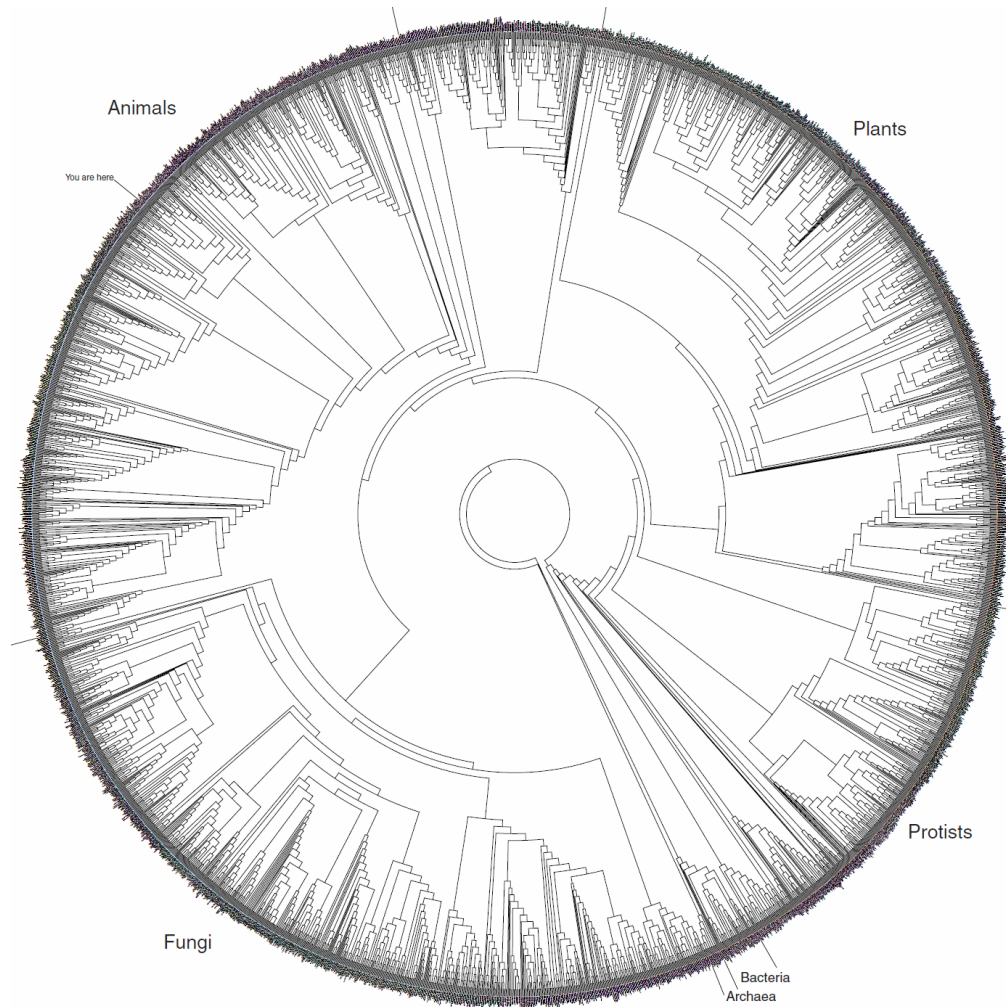
Multi-feature object trajectory clustering for video analysis

Nadeem Anjum Andrea Cavallaro

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

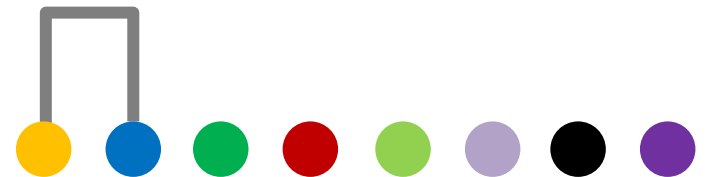
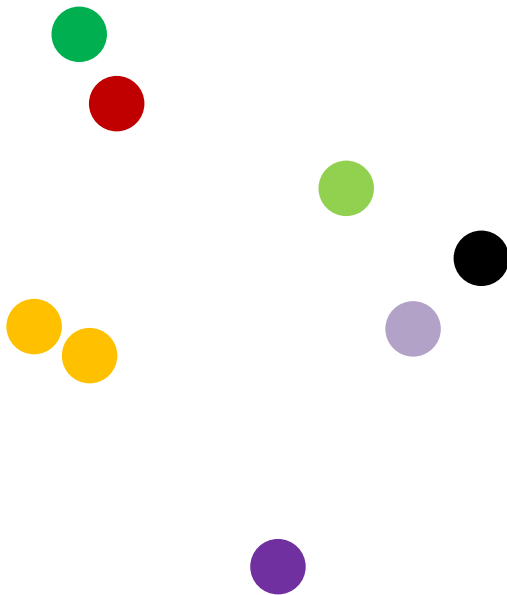


<http://www.zo.utexas.edu/faculty/antisense/DownloadfilesToL.html>

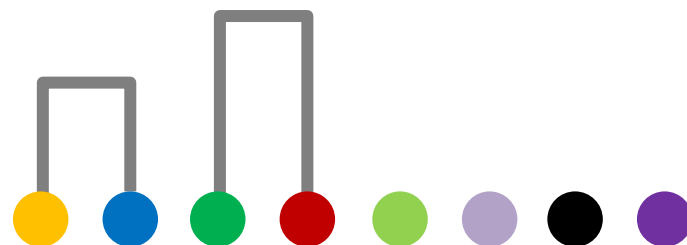
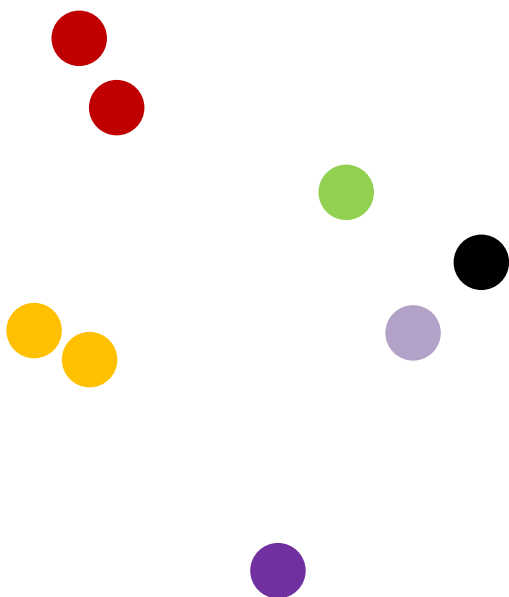
Hierarchical Clustering



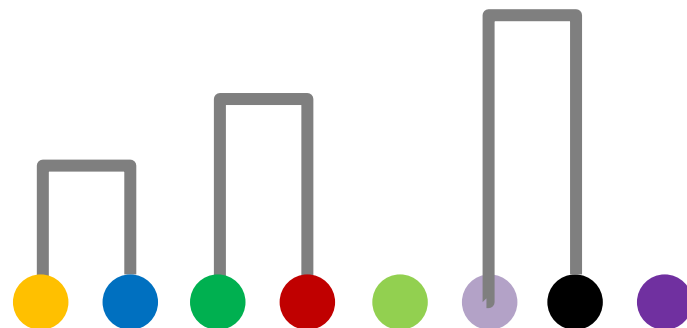
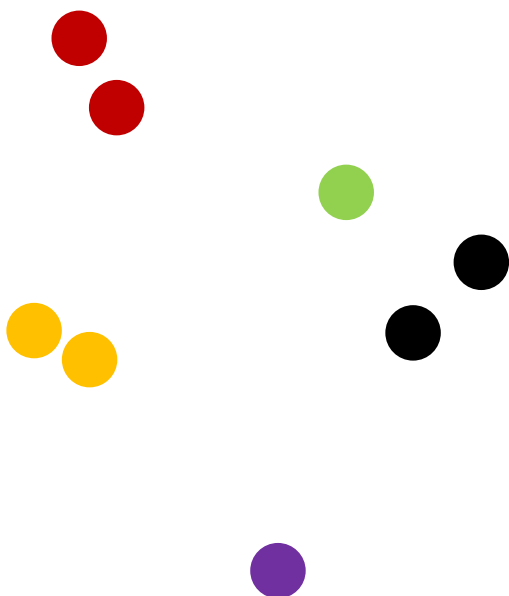
Hierarchical Clustering



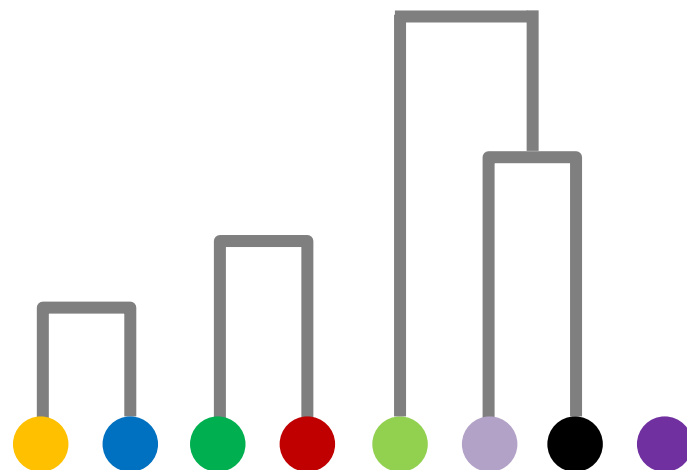
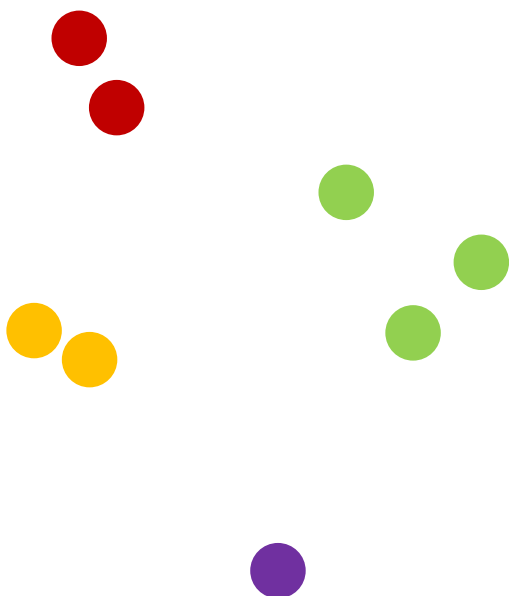
Hierarchical Clustering



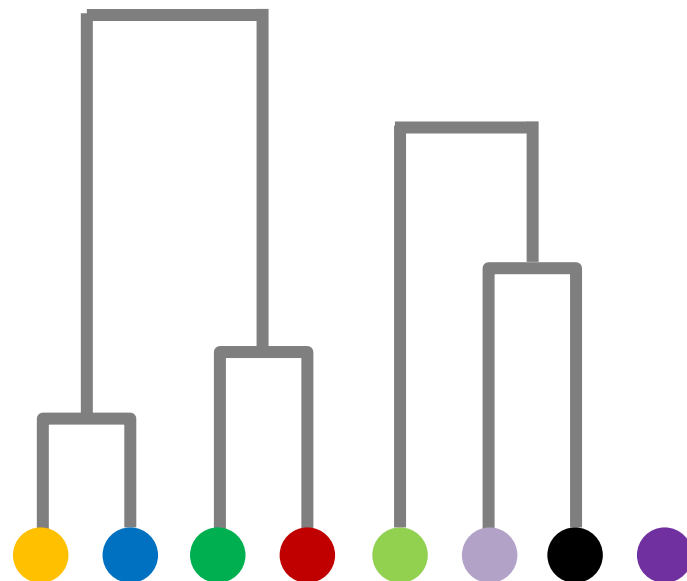
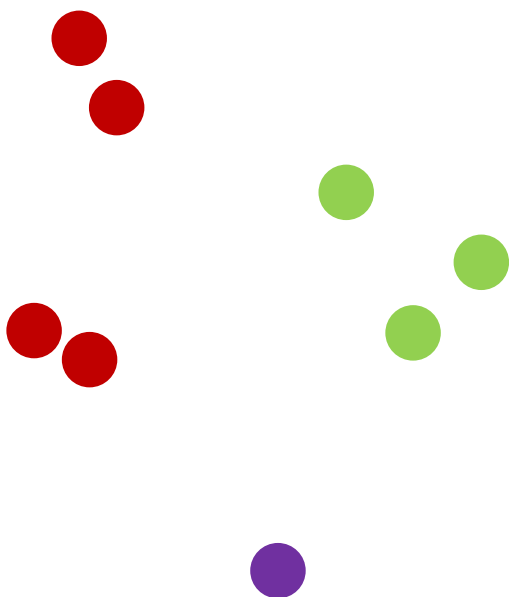
Hierarchical Clustering



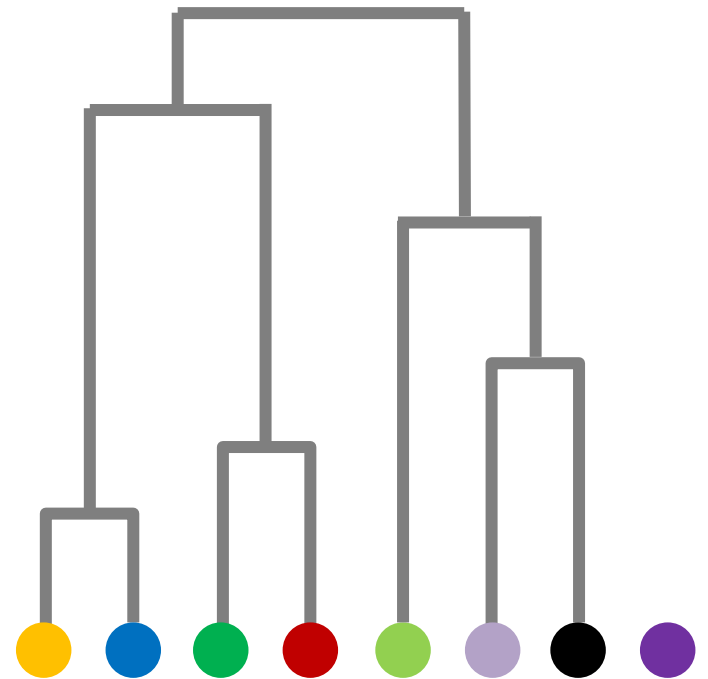
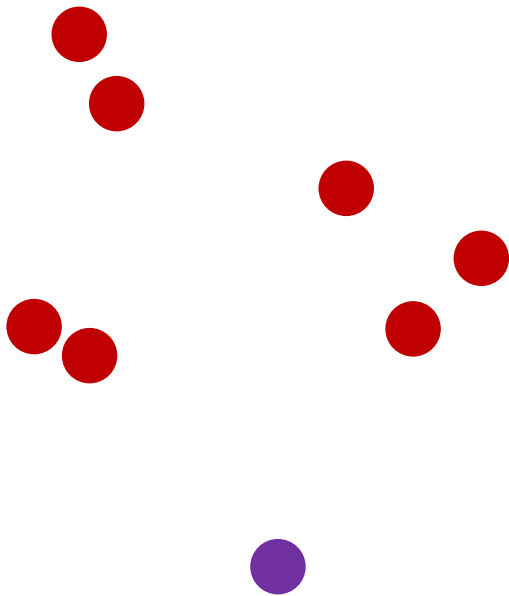
Hierarchical Clustering



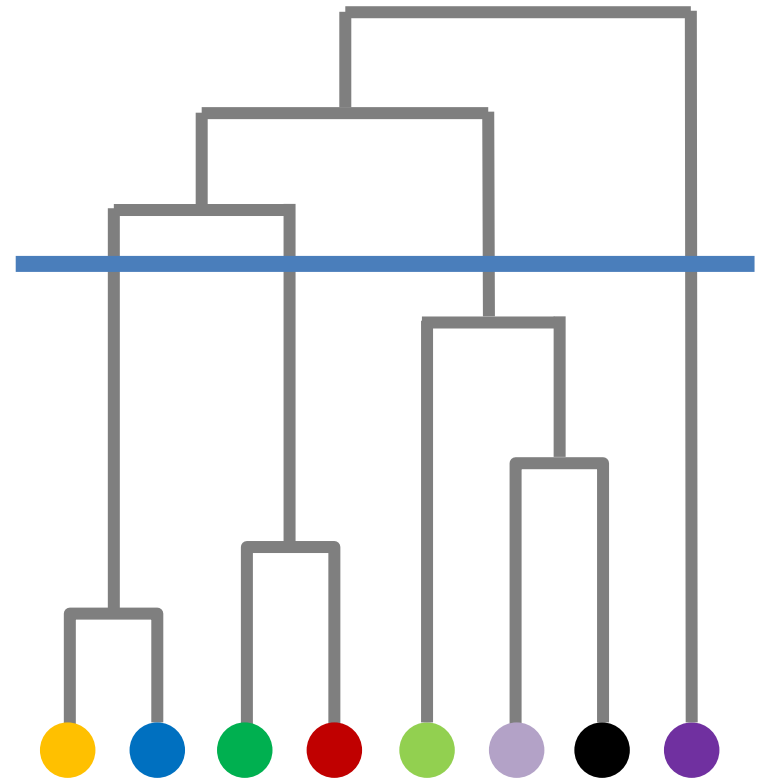
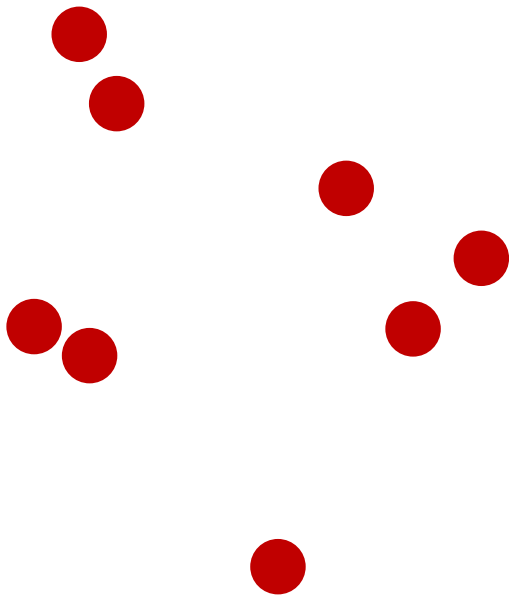
Hierarchical Clustering



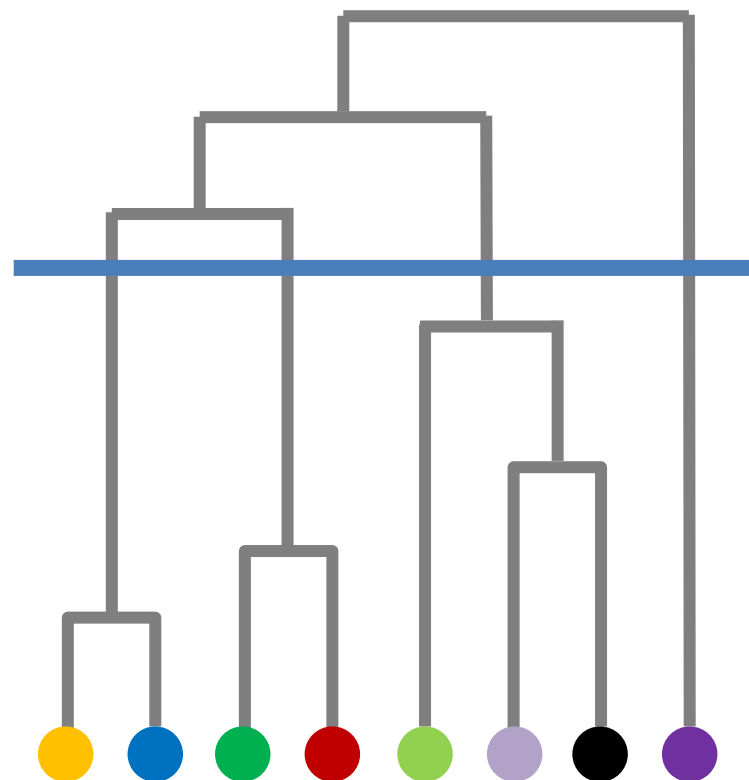
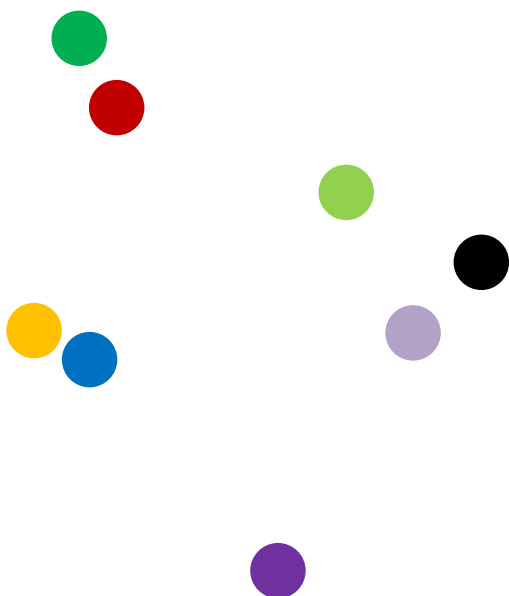
Hierarchical Clustering



Hierarchical Clustering



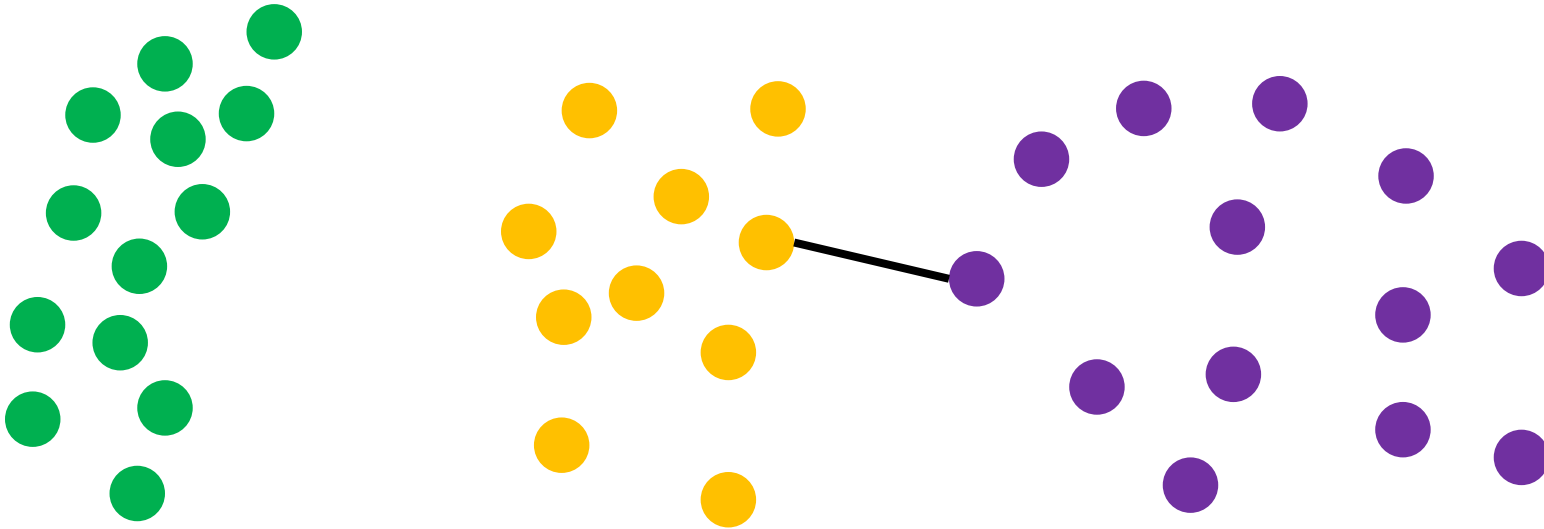
Hierarchical Clustering



Hierarchical Clustering

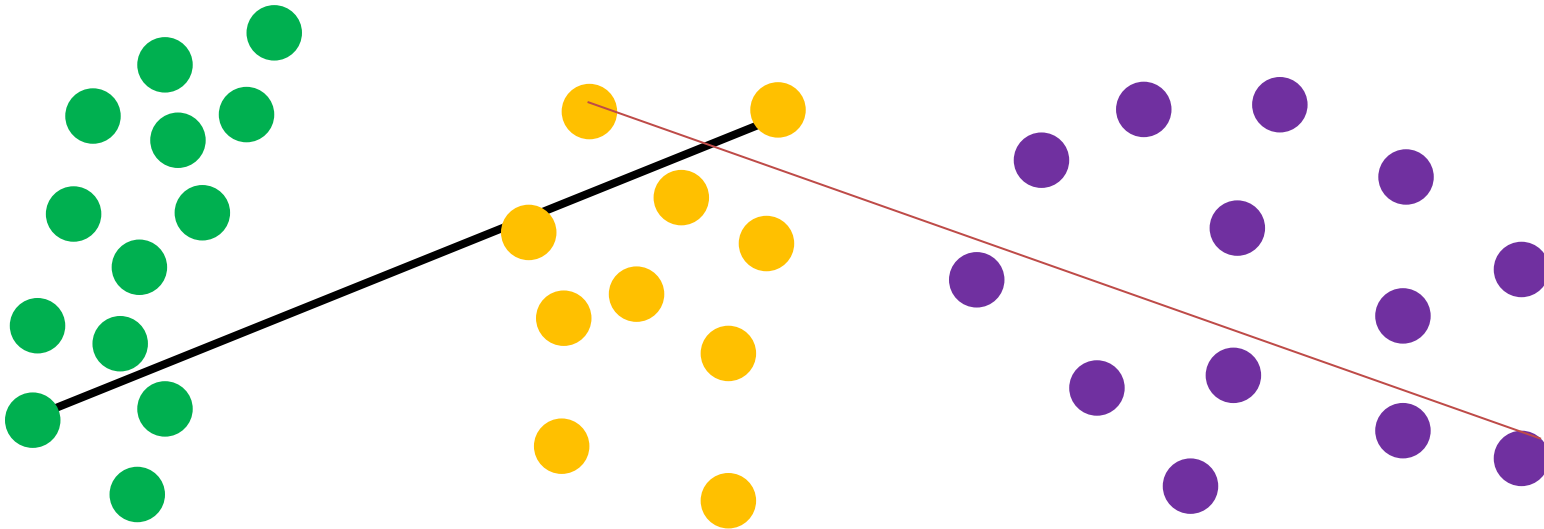
- Produces complete structure
- No predefined number of clusters
- Similarity between clusters:
 - single-linkage: $\min\{d(x,y) : x \in \mathcal{A}, y \in \mathcal{B}\}$
 - complete-linkage: $\max\{d(x,y) : x \in \mathcal{A}, y \in \mathcal{B}\}$
 - average linkage: $\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x,y)$

Single Linkage



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

Complete Linkage

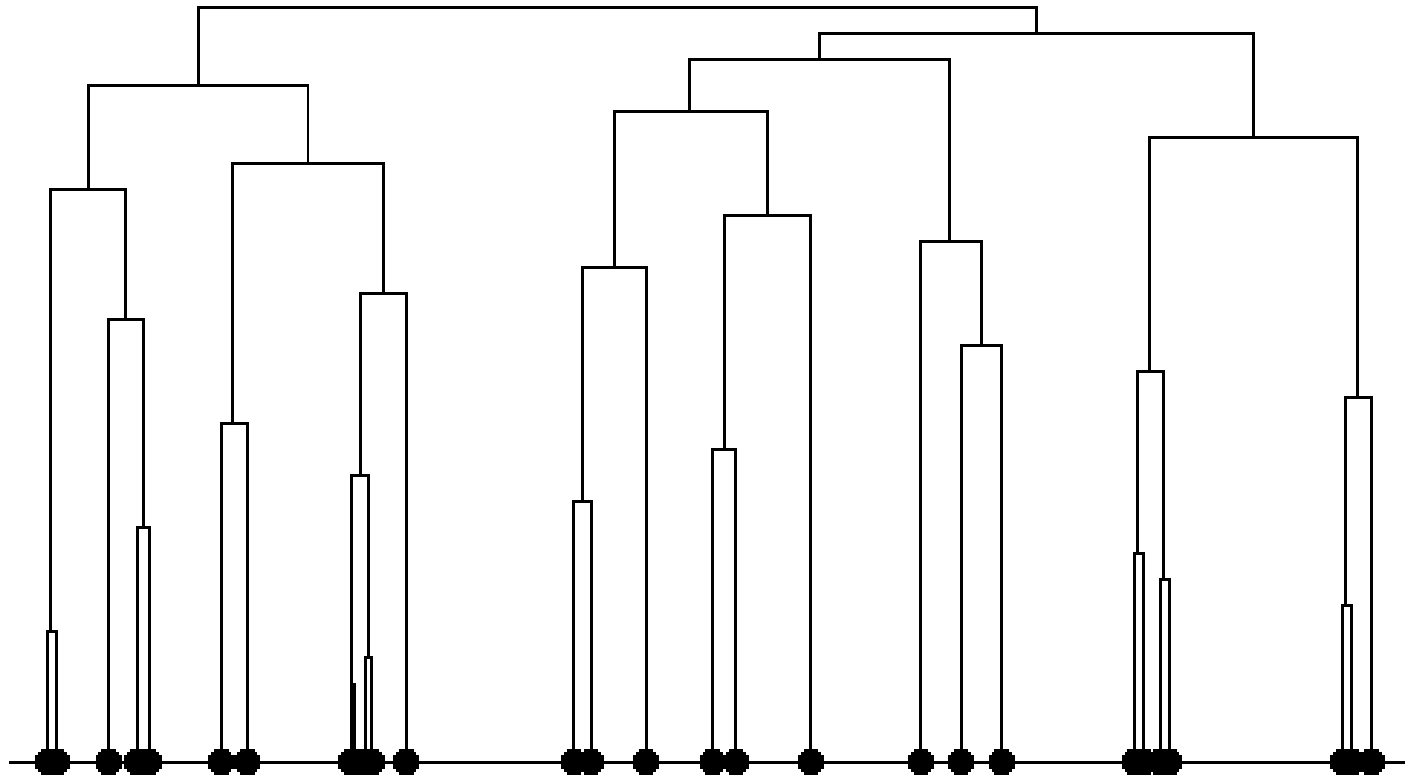


$$\max\{d(x, y) : x \in \mathcal{A}, y \in \mathcal{B}\}$$

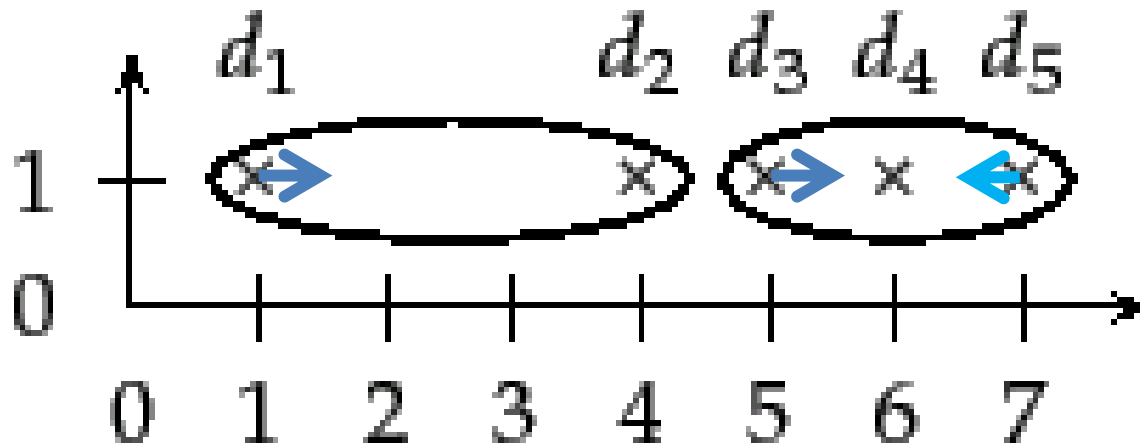
Linkage Matters

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

Chaining Phenomenon



Outlier Sensitivity



➡ + 2*epsilon

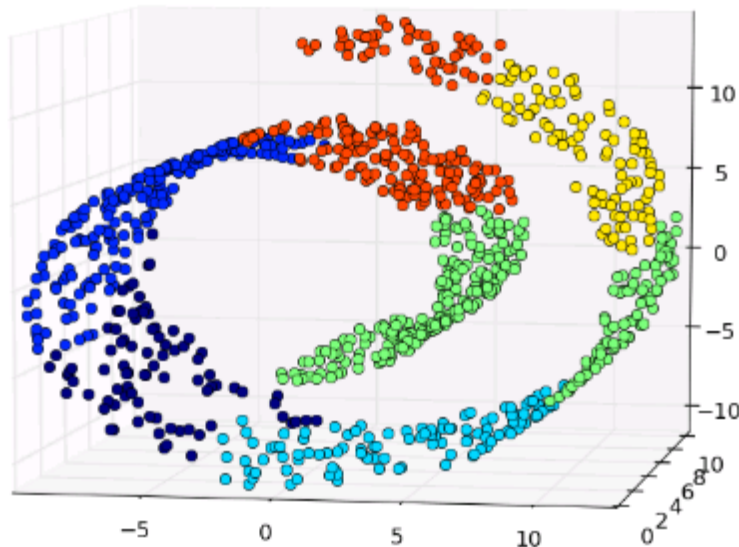
➡ - 1*epsilon

Efficient Hierarchical Graph-Based Video Segmentation

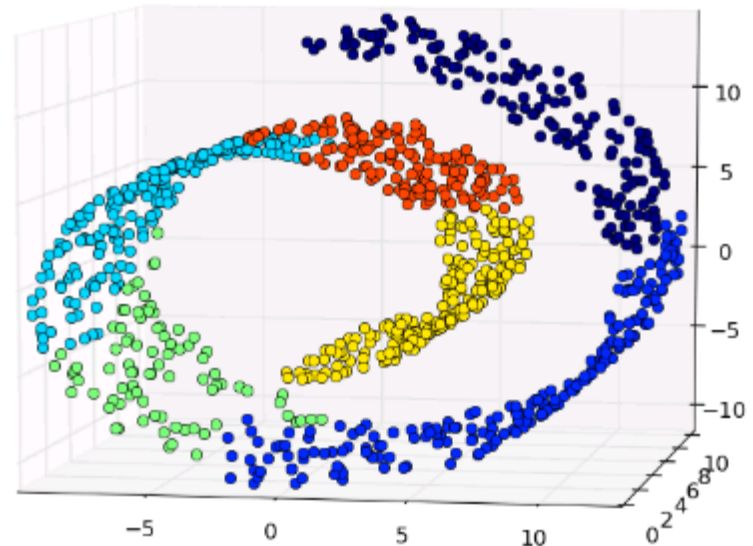
Matthias Grundmann^{1,2}, Vivek Kwatra²,
Mei Han² and Irfan Essa¹
¹Georgia Tech ²Google Research

IEEE CVPR, San Francisco, USA, June 2010

Swiss Role Problem



without connectivity
constraints



with connectivity
constraints

only adjacent clusters can be merged together

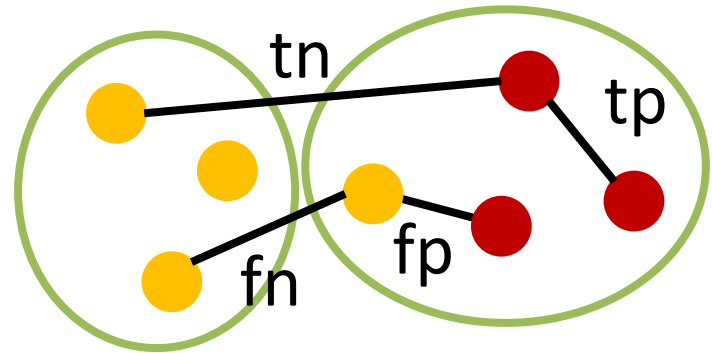
Evaluation Criteria

- 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

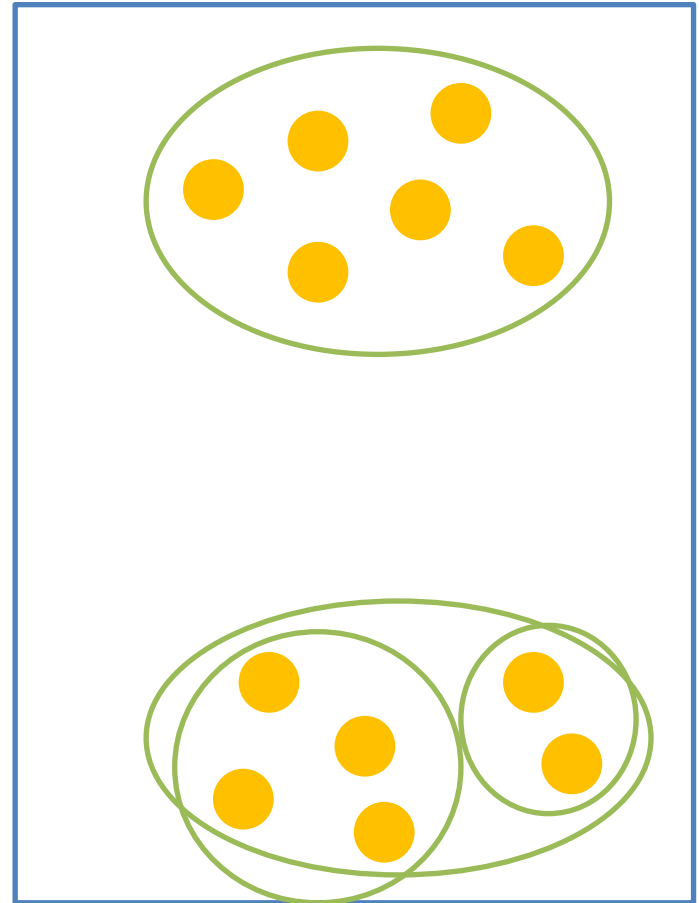
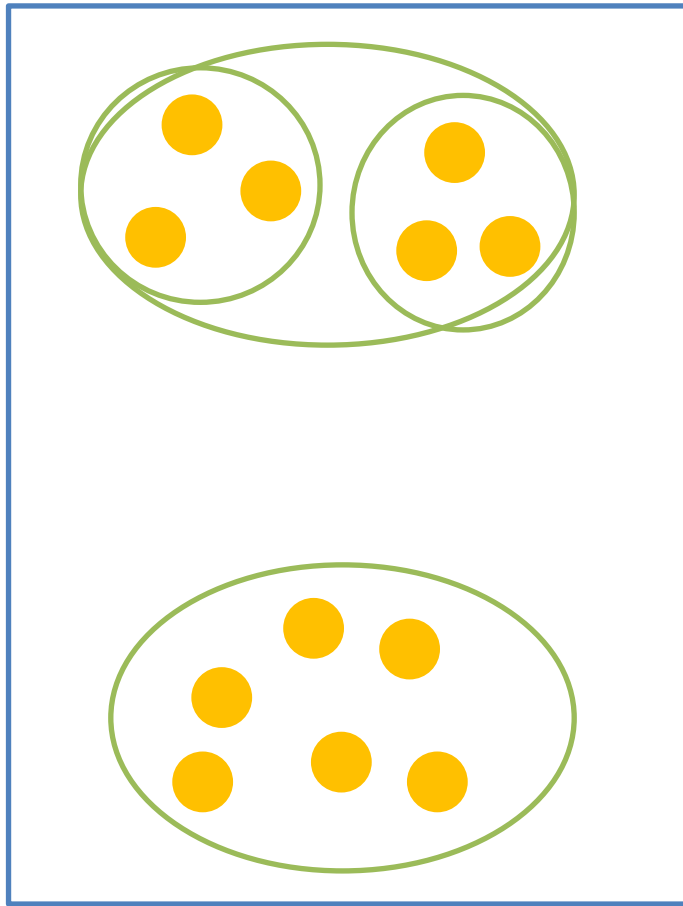
- Percentage of correct classifications
- Compare pairs of elements:

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



- Fp and fn are equally weighted

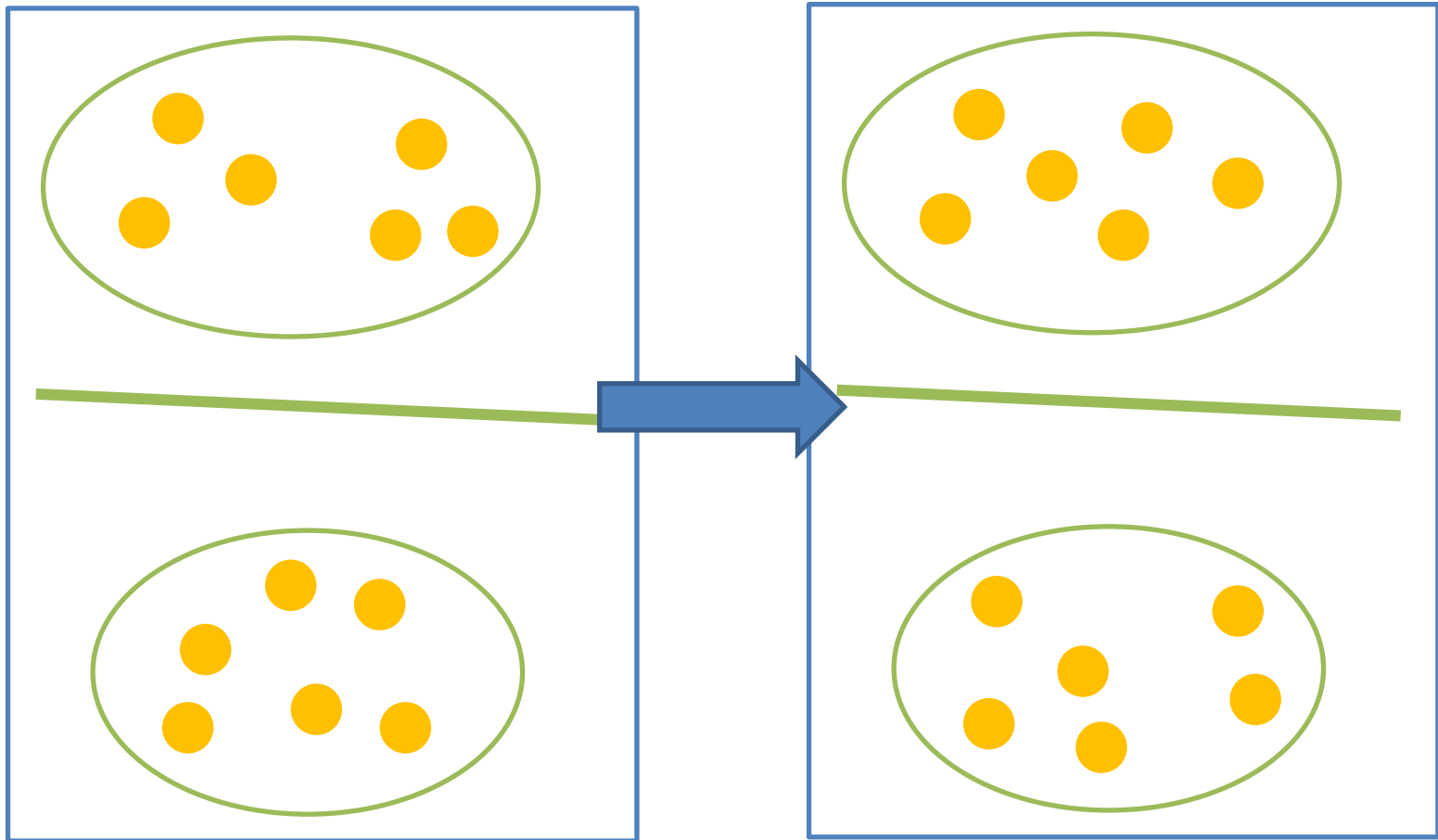
Stability



Stability

- What is the right number of clusters?
- What makes a good clustering solution?
- Clustering should generalize!

Stability



Summary

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