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:

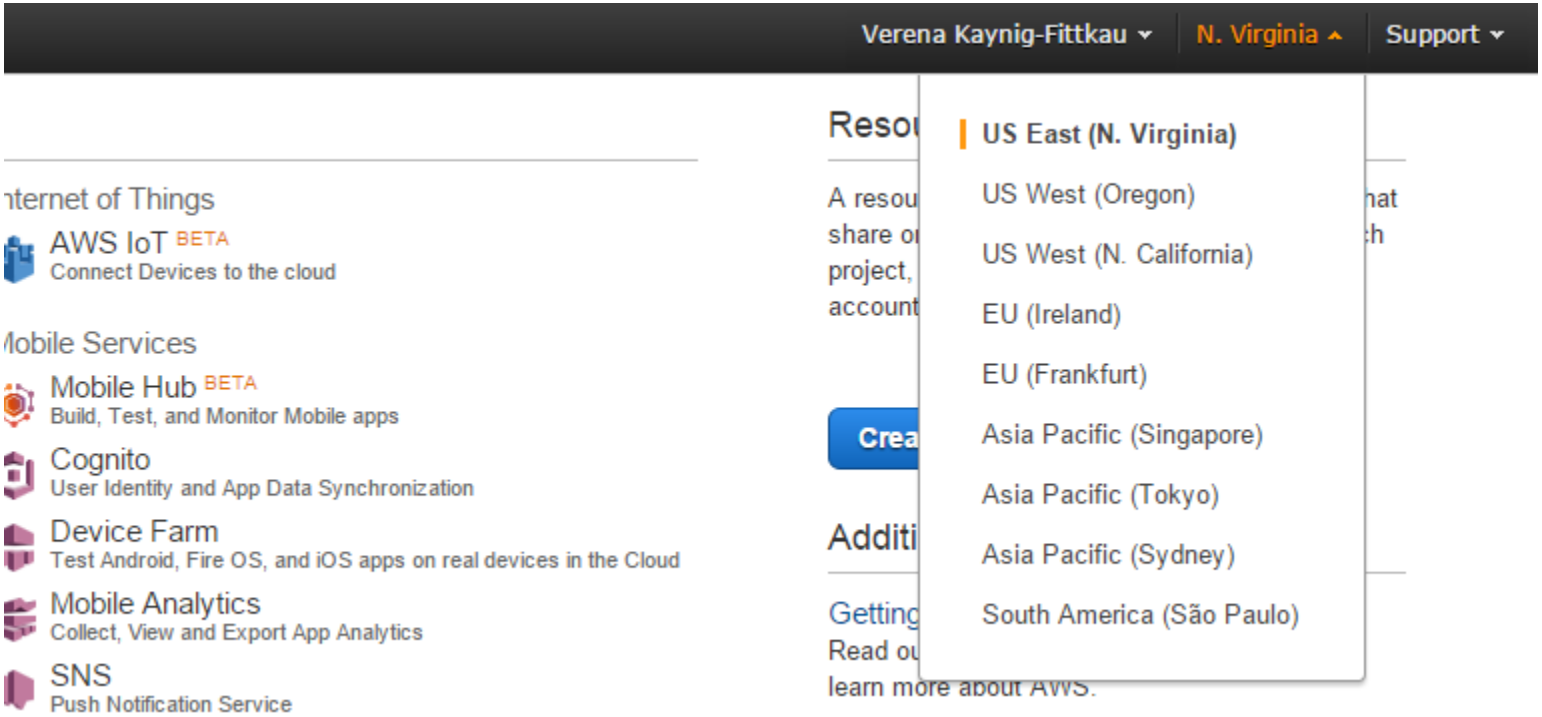
https://piazza.com/class/icf0cypdc3243c?cid=1369

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

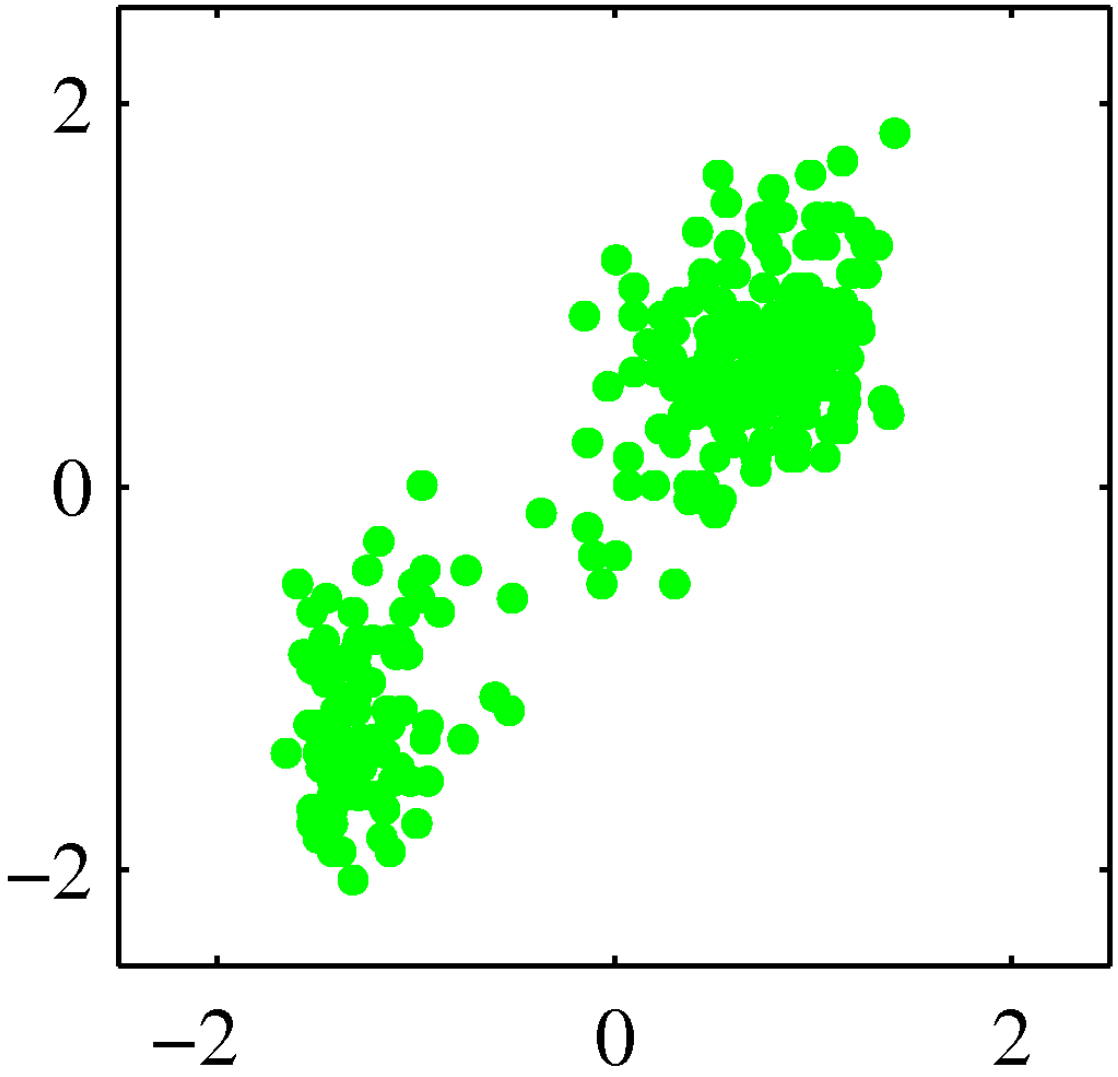
Supervised vs. Unsupervised

- 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.

Only have a matrix of data: X, and no list of labels/classes (ie, no y in y ~ X)

No labels: only have points. Can't use y to guide and find separating hyper-plane
We CAN see patterns: ie 2 clusters, can classify under 2 labels.

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.

Can use if patterns previously known, or to find NEW patterns.

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. ie, triangulation, where to put new cluster (area served by police stations)
- Outlier detection
  Focus on finding pattern, consider data that don't fit pattern as outliers.

Unsupervised Learning

- K-means
- Mean-shift
- Hierarchical Clustering

- Rand index, stability

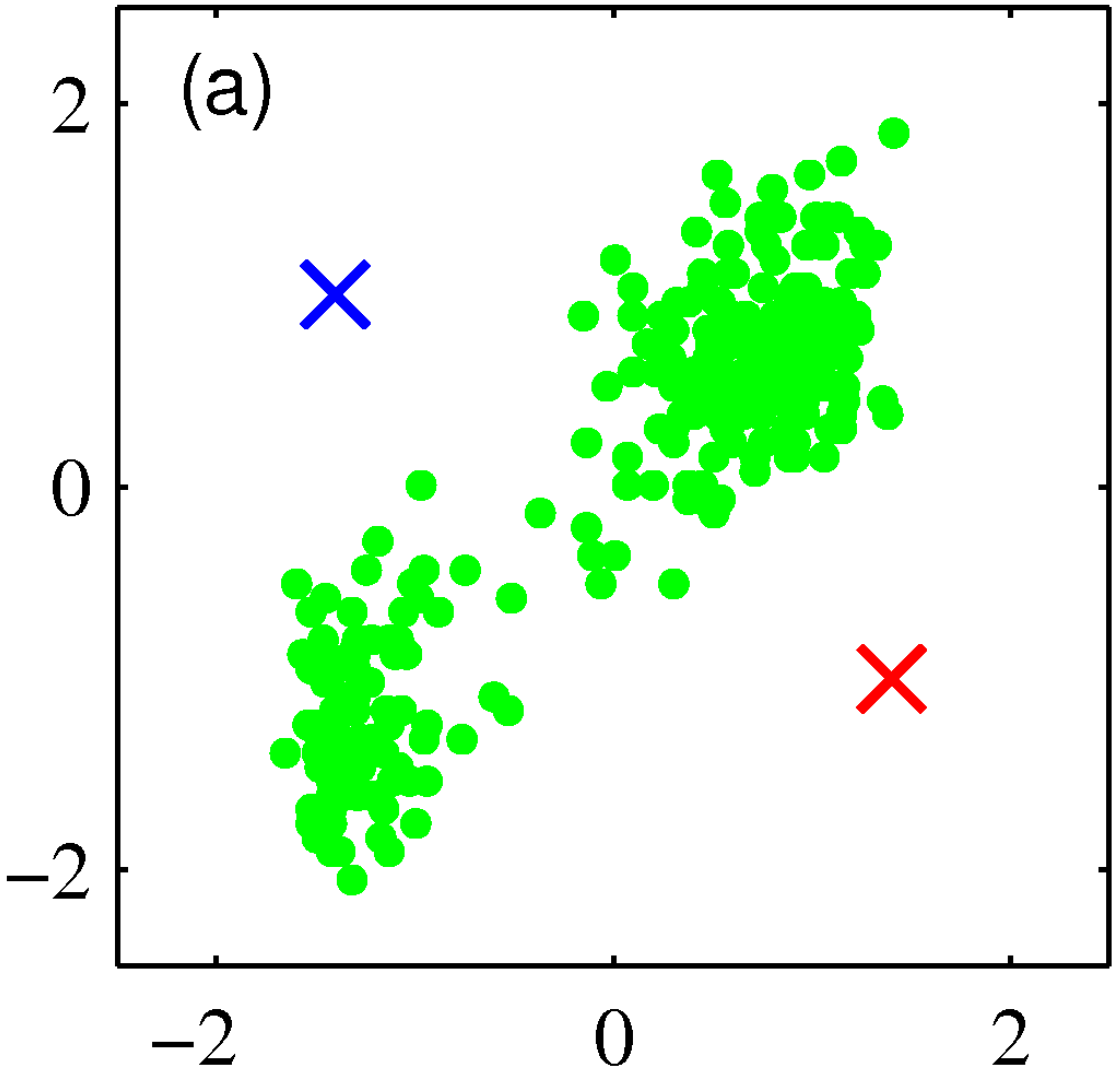
How do you evaluate cluster without prior labels?

K-means – Algorithm

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

initialize with two random cluster centers.

K-means



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

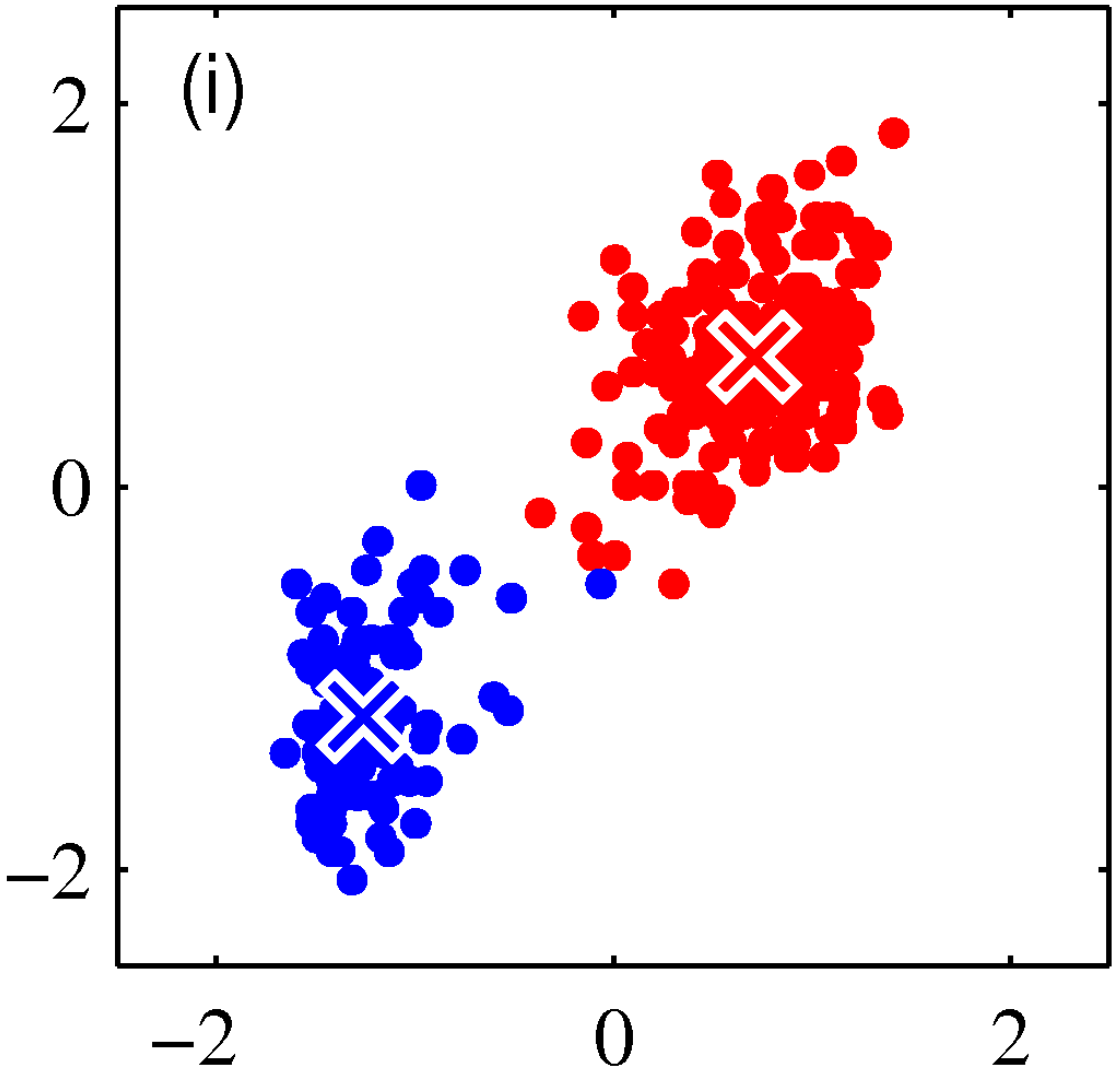
K-means

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

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

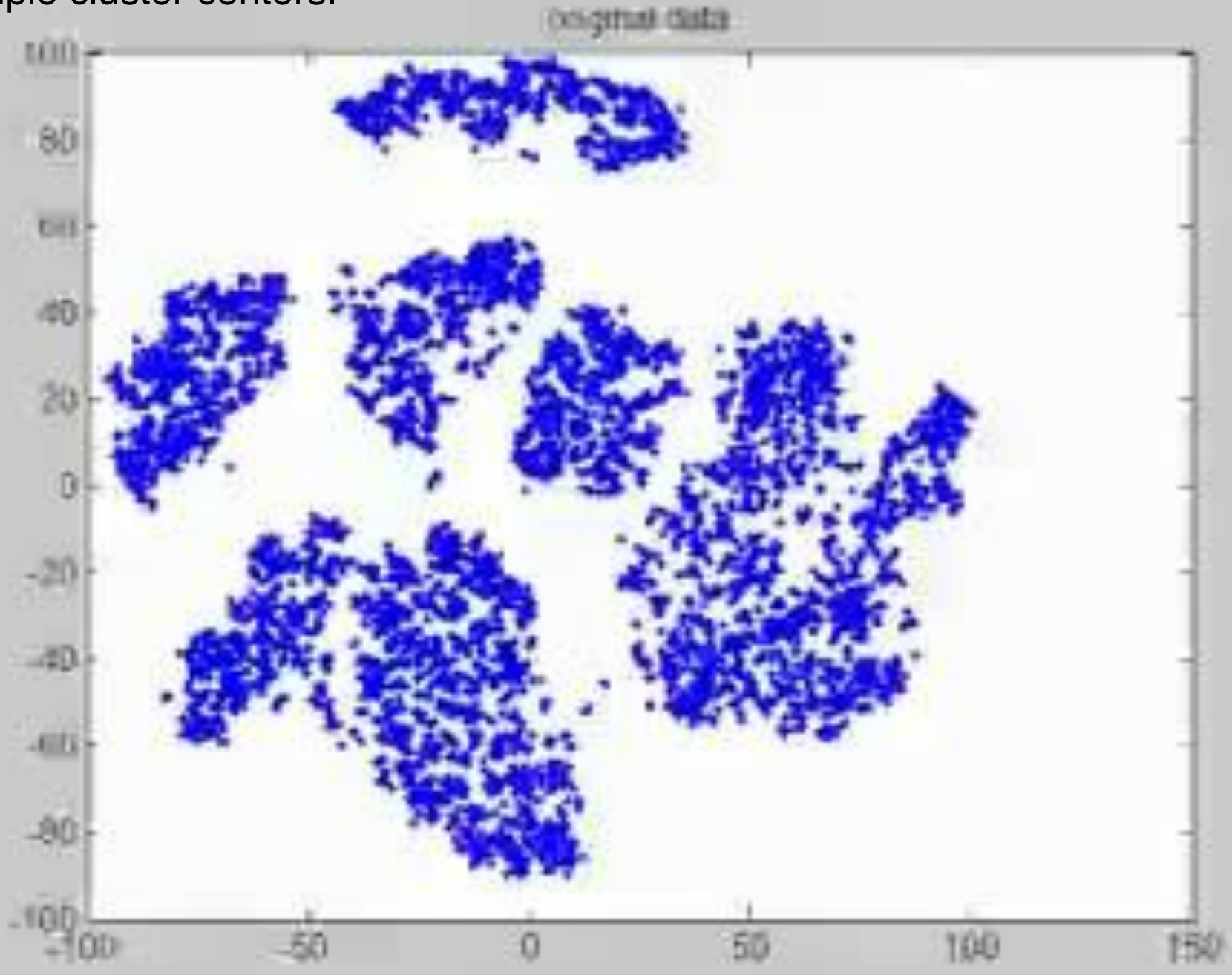
new center in center of prior classification scheme.

K-means



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

multiple cluster centers.



Color compression of images.
Each pixel is one observation with three features (RGB values) —> 3D

K-means Example

R

G

B

first step, color points with color of cluster center. Can increase number of clusters for more accuracy (but diminishing returns, won't need ALL unique colors

K-means Example

Both images have 10 clusters but different results. Why?
Randomized starts: guarantee convergence, but not convergence to same result

K-means Example

K-means Summary

- Guaranteed to converge
- Result depends on initialization

- Number of clusters is important
  May not be known beforehand.

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

how to check convergence: set a small epsilon as threshold for distance converge once sum of squares below that epsilon.

Initialization Methods

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

- Start several times

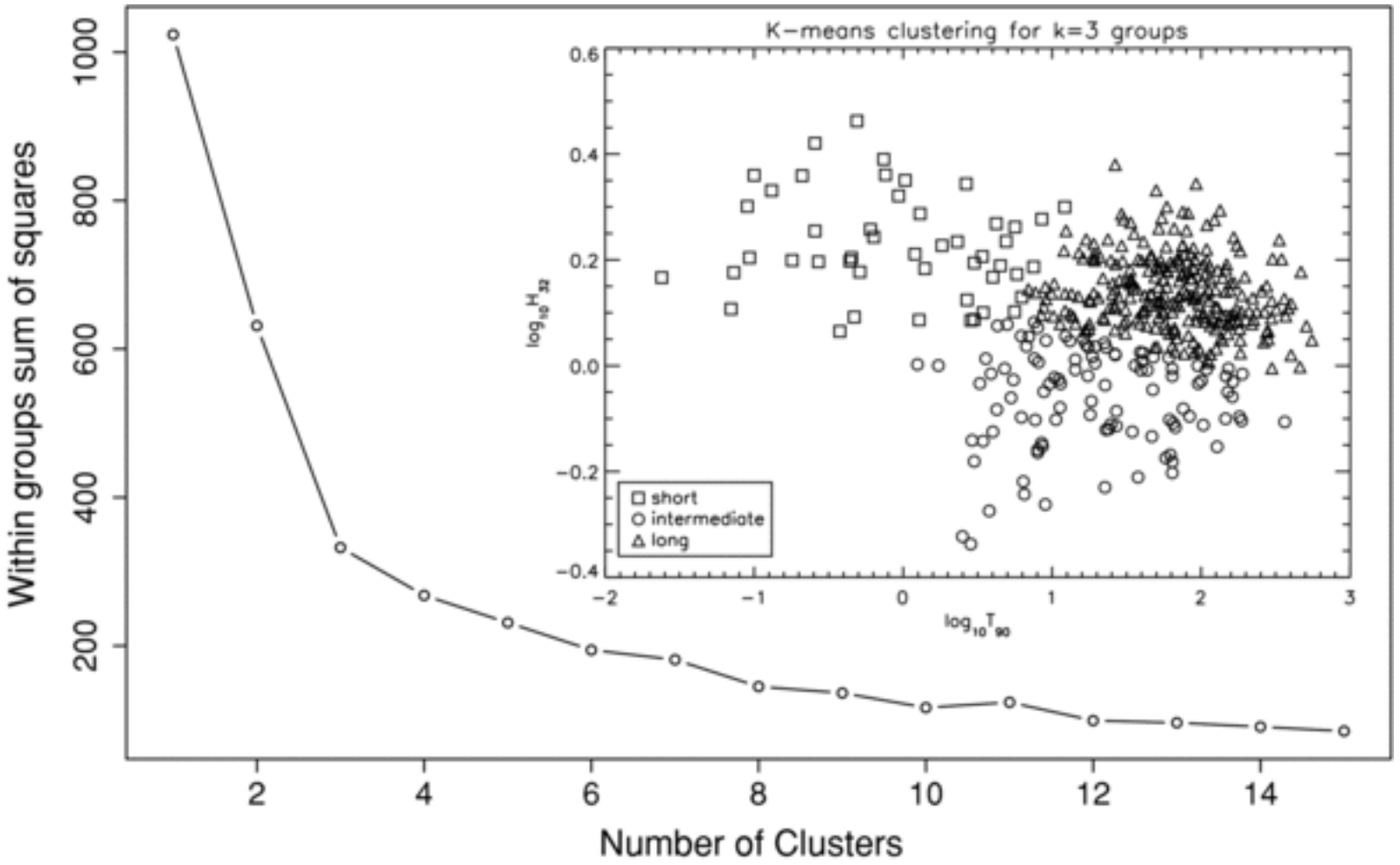
k = 10, 100 runs, one solution 90 times vs 10 times, 90 times seems better
ie, there is stability to solution outcome, a form of cross validation?

How to find K

- Extreme cases:
  - K=1 one cluster center: at points assigned to one pt: poor explanation
  - K=N each pt is a center, overfitting.
- Choose K such that increasing it does not model the data much better.

A cross-validations scheme: training data to solve k-means, validation data to determine how much variance explained (lower sum of sq)

"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?

conceptually simple, computational challenge (intensive)

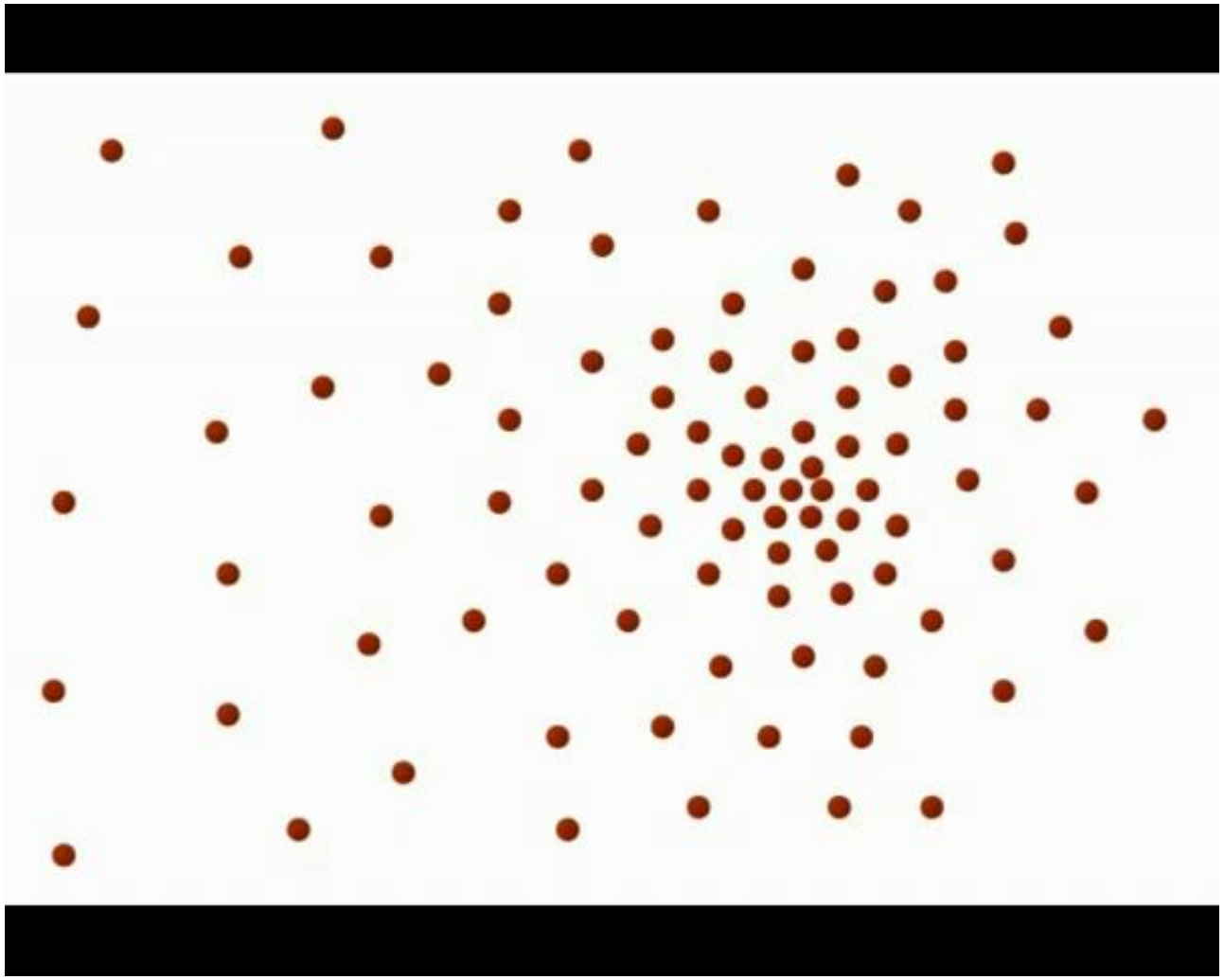
Mean Shift

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

have to do for every SINGLE data pt: window always shifts towards denser part (gradient in density) —> convergence and you get a cluster center.

do NOT specify number of clusters (k), just size of window.

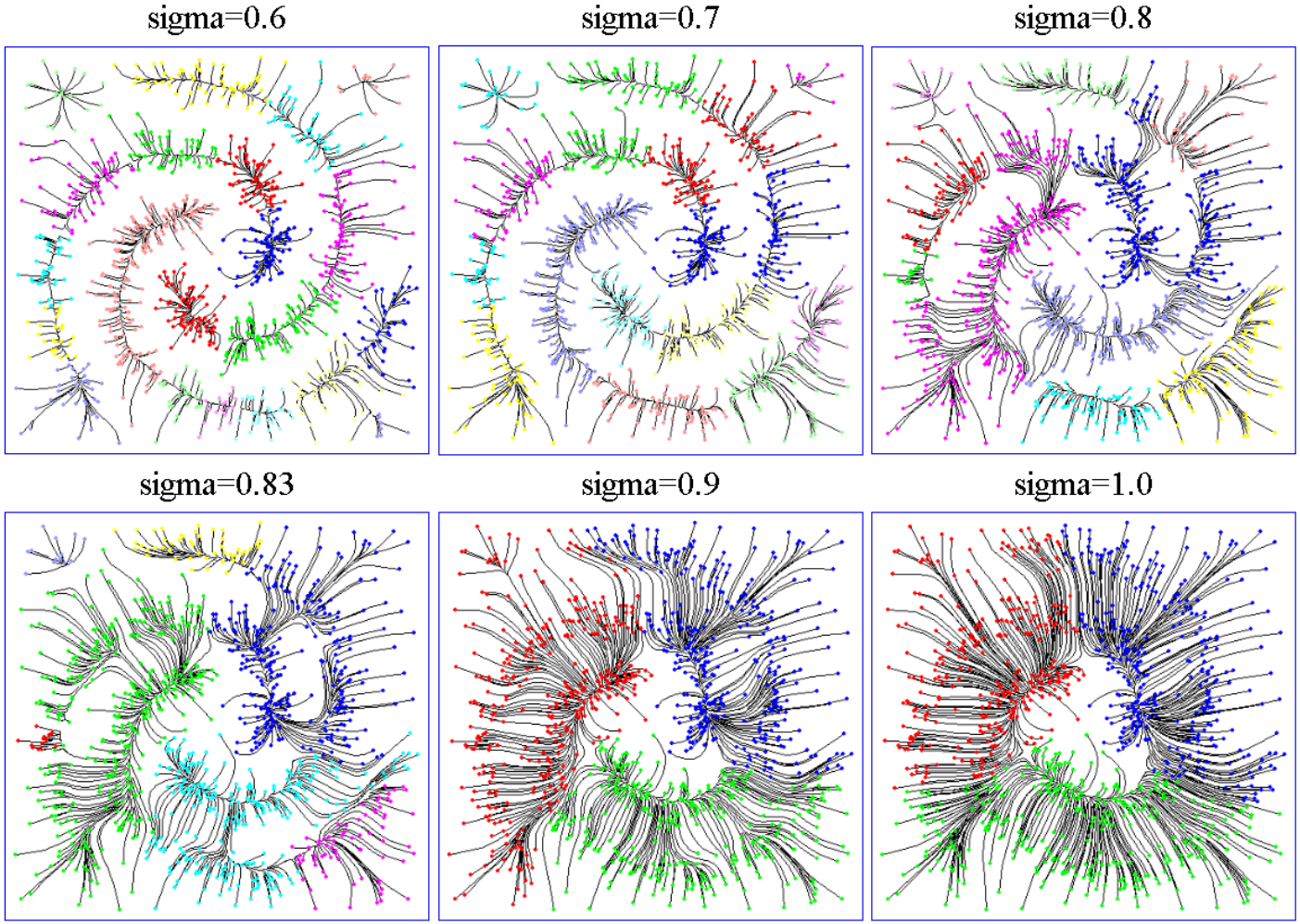
Mean Shift



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

size of window ~ number of clusters.

Mean Shift



lose spiral: window too lg

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

  embarrassingly parallel

- 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

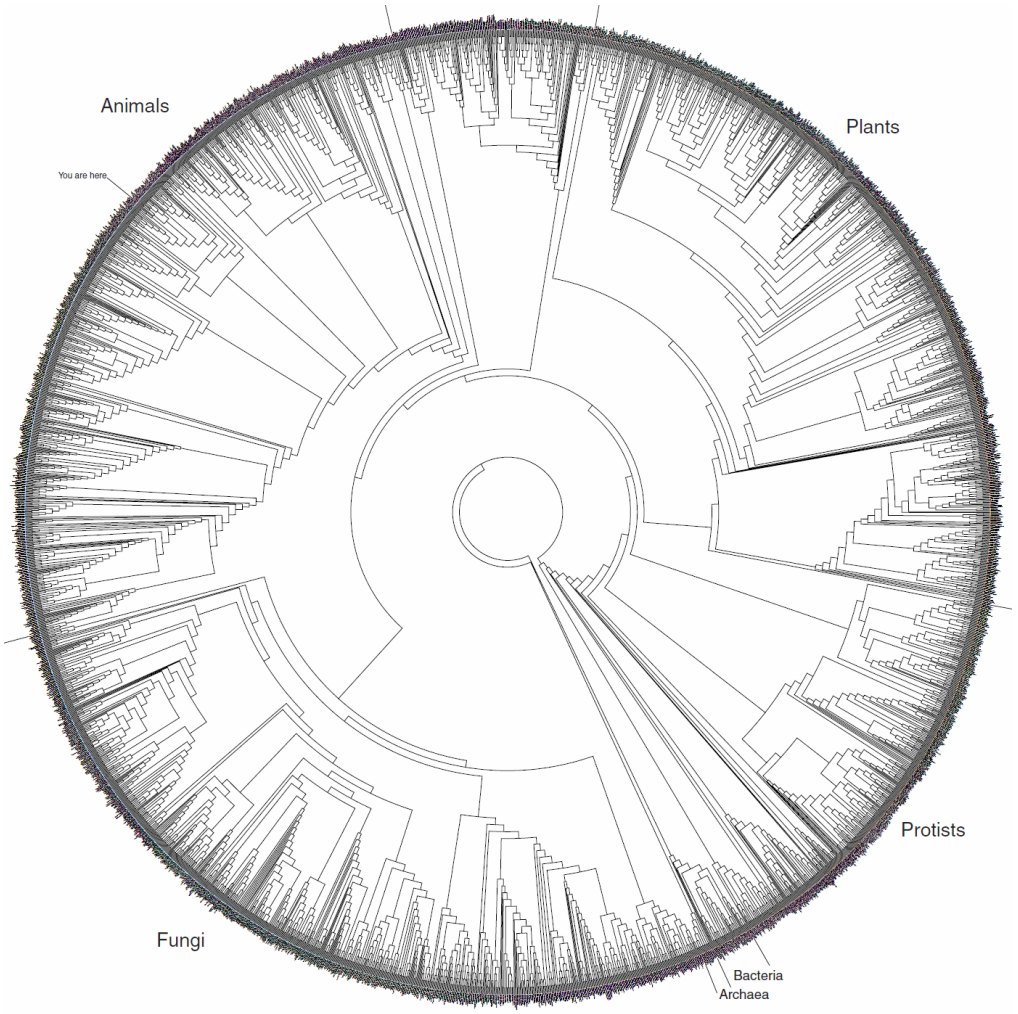
Parameters parameters

- 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

  no window-size or k.

Tree of Life

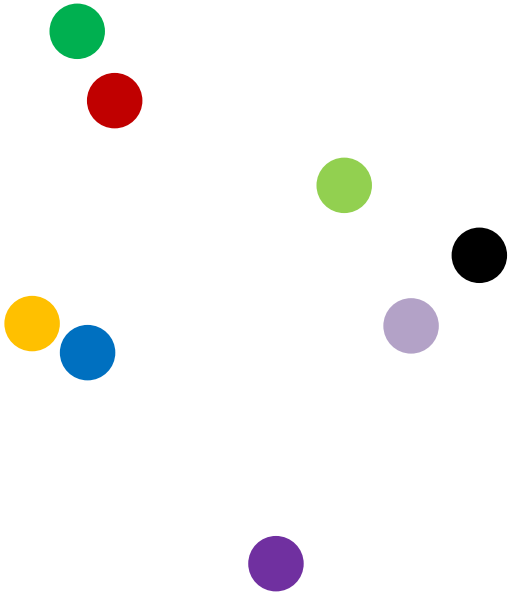


http://www.zo.utexas.edu/faculty/antisense/DownloadfilesToL.html

start: every pt is it's own cluster (k = N)

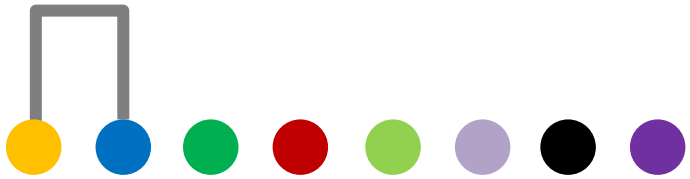
Hierarchical Clustering

computer shortest distance.



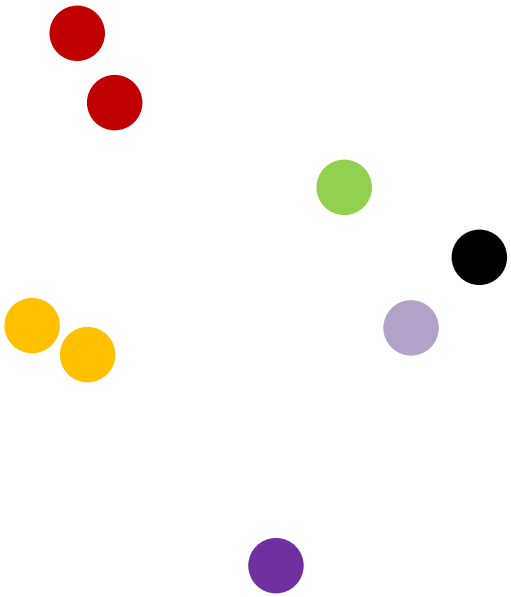
Hierarchical Clustering

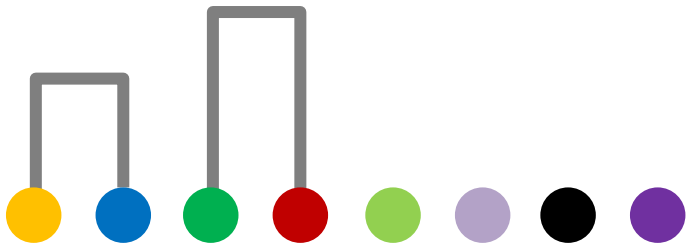
begin clustering the pairs.



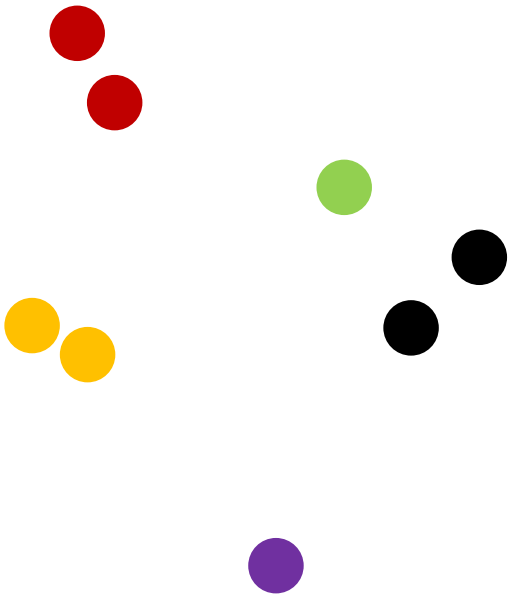
organized by shortest distance pairs

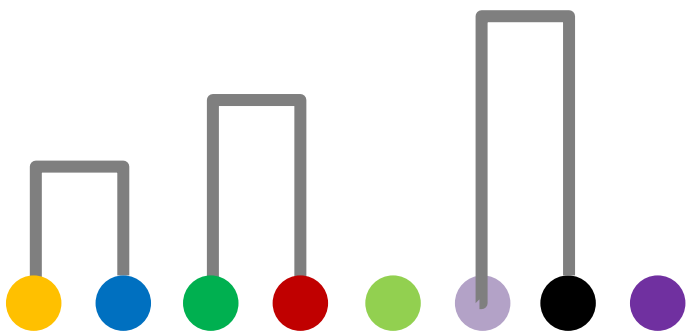
Hierarchical Clustering



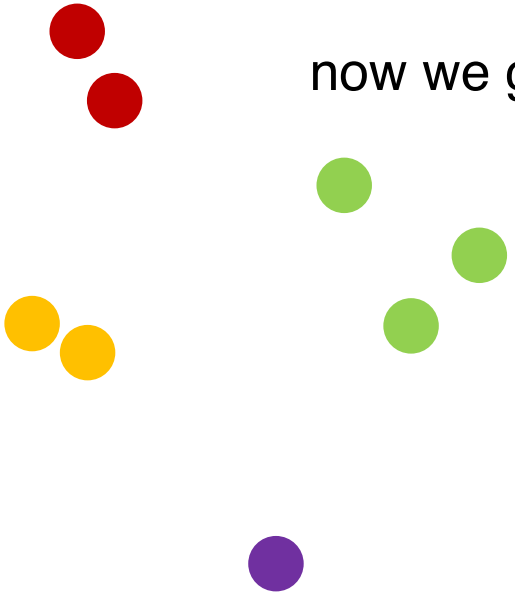


Hierarchical Clustering

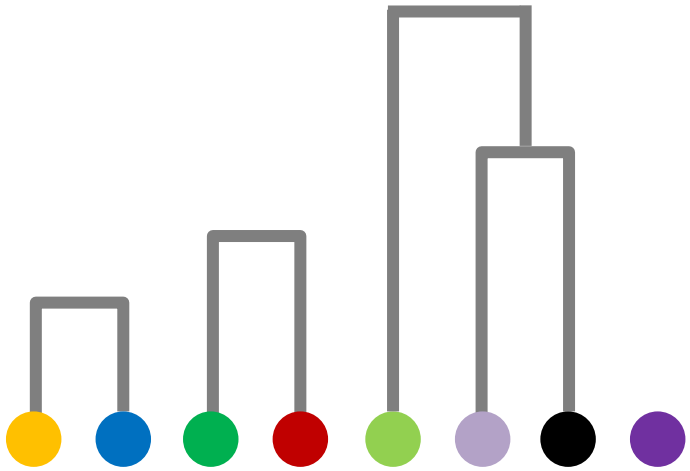




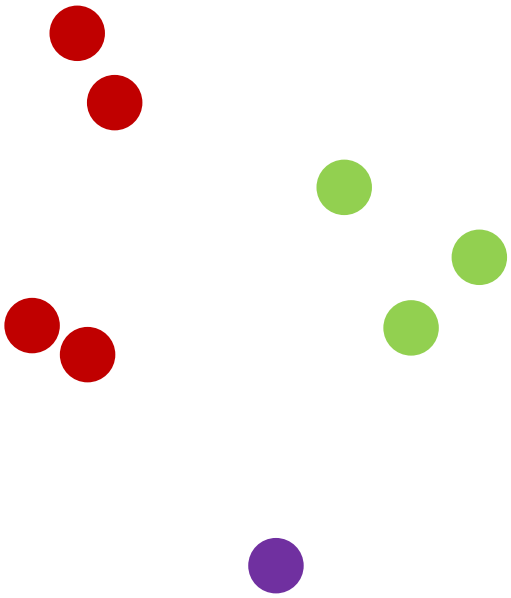
Hierarchical Clustering

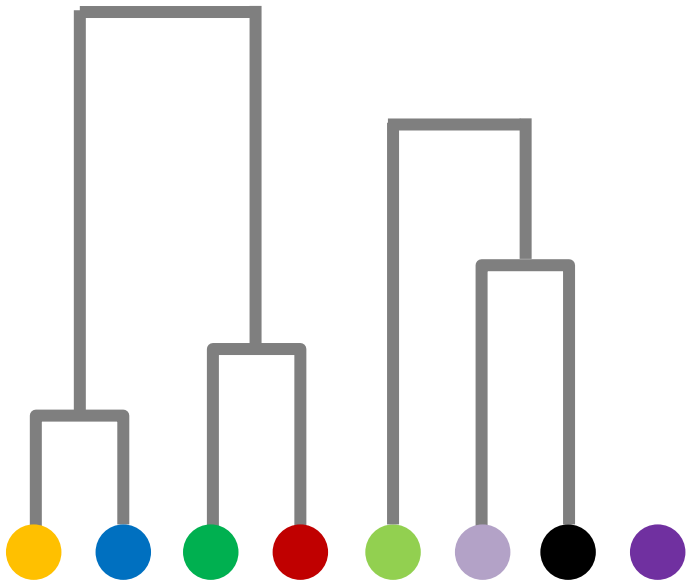


now we group clusters.

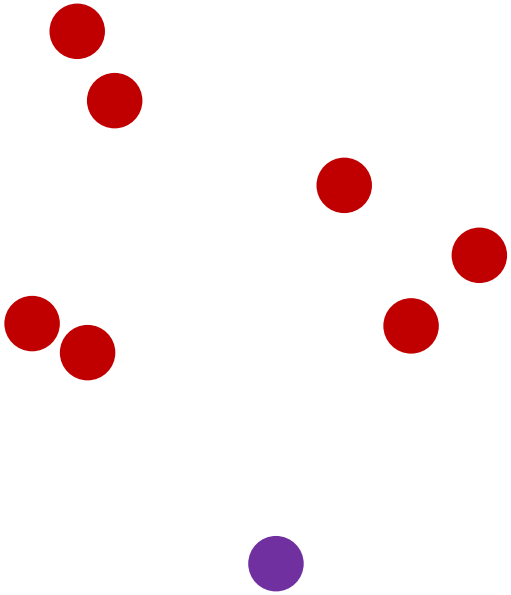


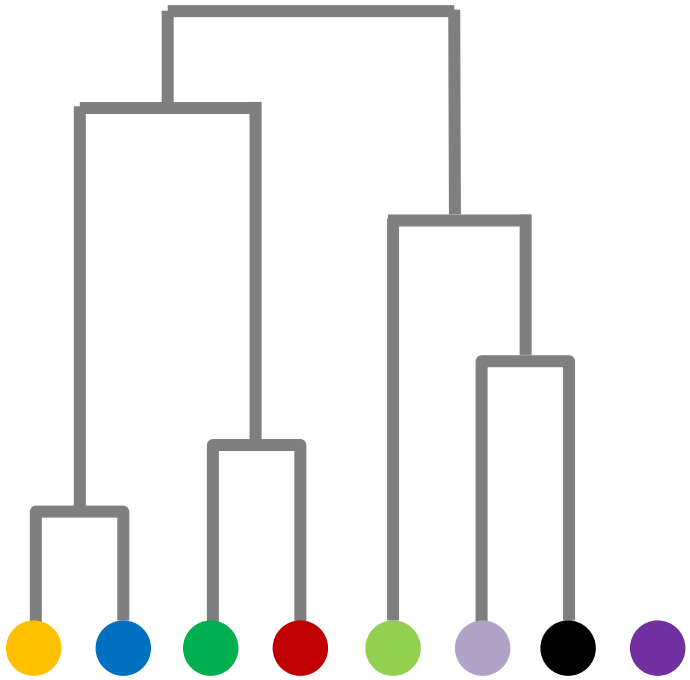
Hierarchical Clustering





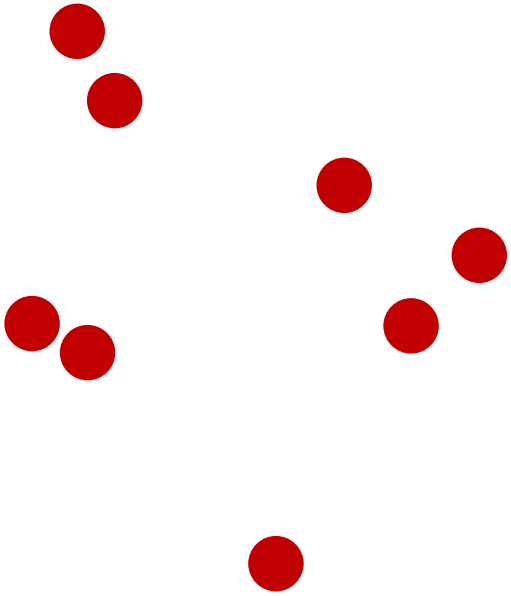
Hierarchical Clustering



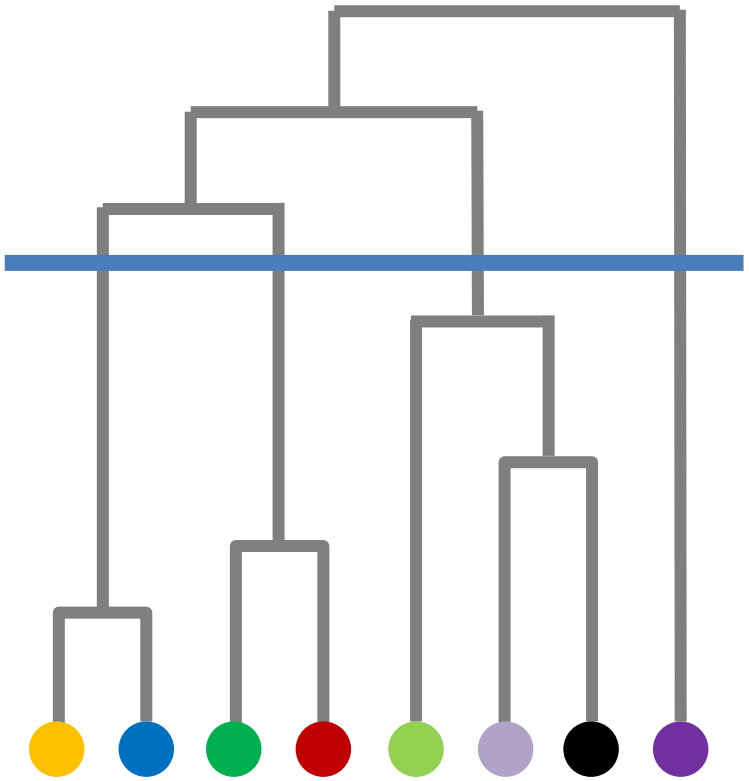


Hierarchical Clustering

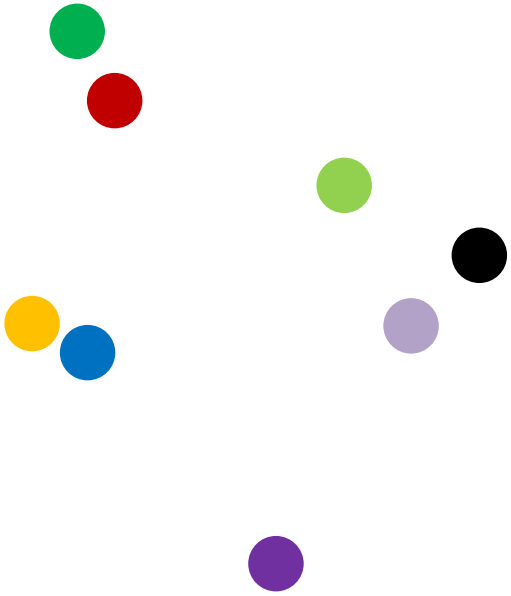
now we have k = 1
so we did extremes (k = N to k = 1)
AND everything BETWEEN.

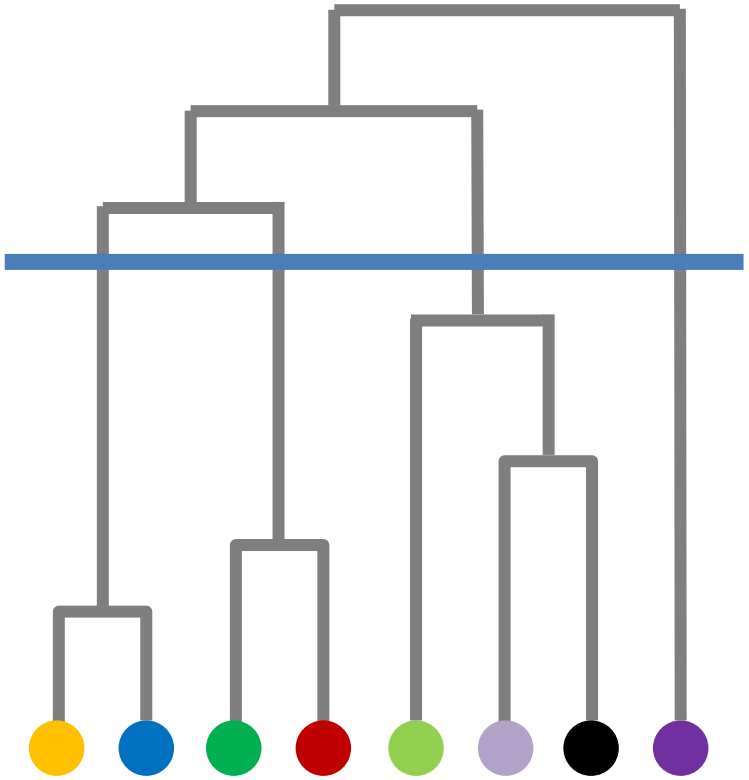


threshold.



Hierarchical Clustering



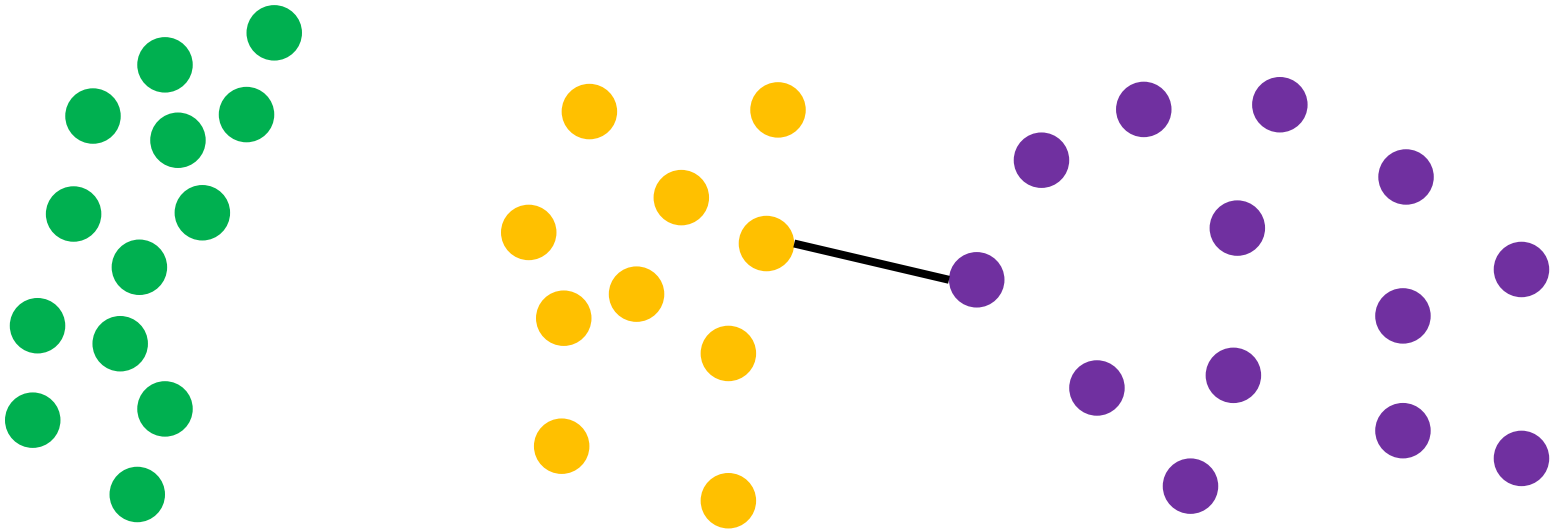


Hierarchical Clustering

- Produces complete structure
- No predefined number of clusters Do them all.

- 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

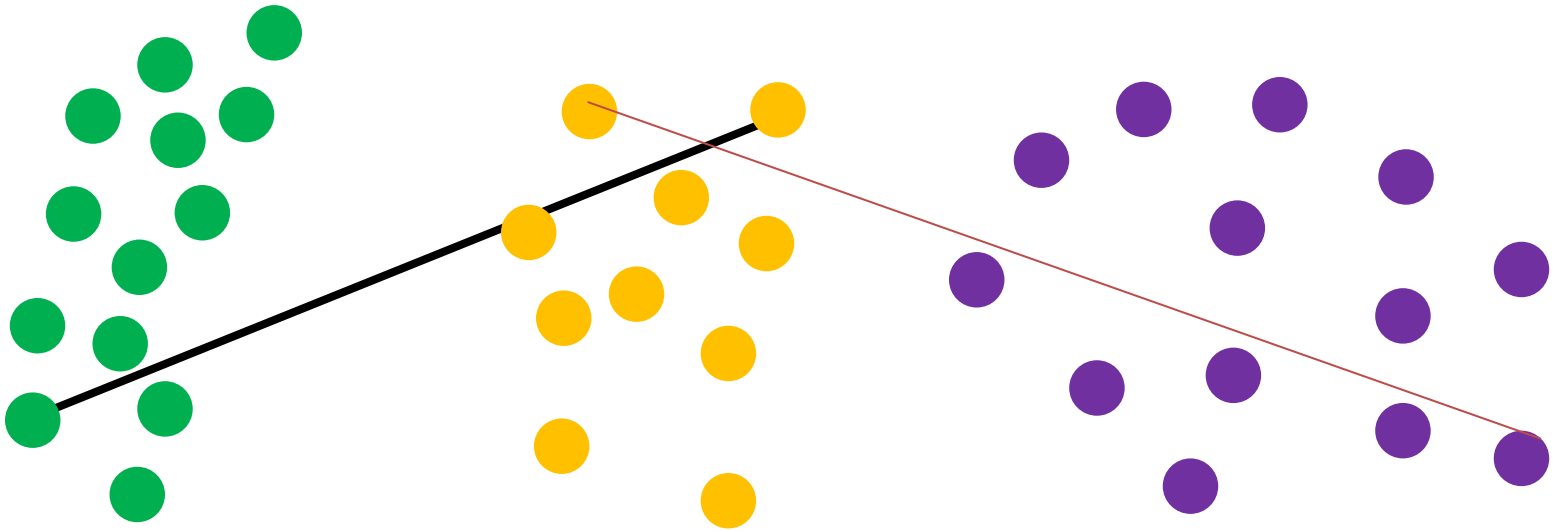


$$\min\{d(x,y) : x \in \mathcal{A}, y \in \mathcal{B}\}$$

two closest pts determine the WHOLE distance between clusters

ie, distance between clusters is ONE number.

Complete Linkage



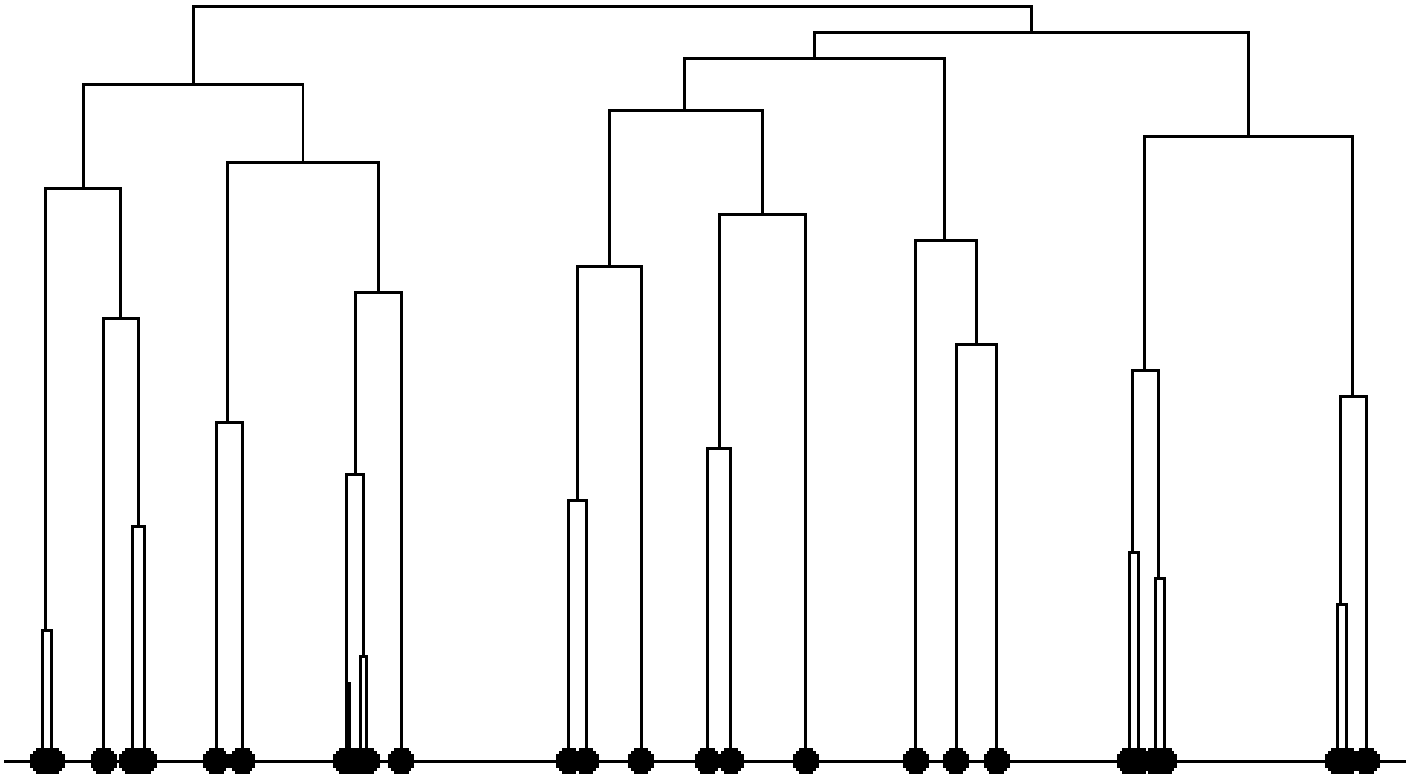
$$\max\{d(x,y) : x \in \mathcal{A}, y \in \mathcal{B}\}$$

Linkage Matters

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

Not balanced with single-linkage. If prefer balanced cluster use complete linkage

Chaining Phenomenon



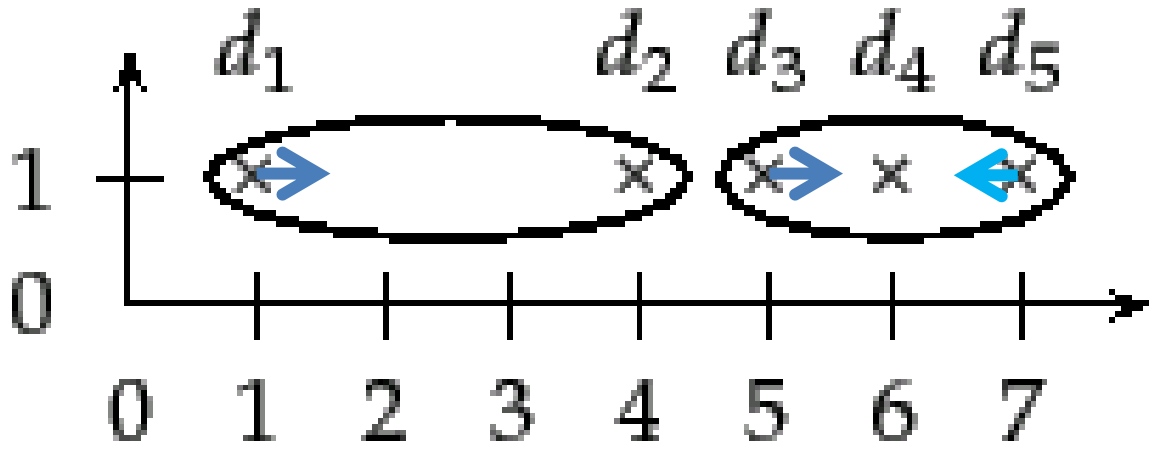
hierarchal clustering is FAST.

d1 should be alone as outlier.

Outlier Sensitivity

Single linkage is robust to outlier, but unbalanced

Complete linkage sensitive to outlier, but clustering is balanced



+ 2*epsilon

- 1*epsilon

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

Efficient Hierarchical Graph-Based Video Segmentation

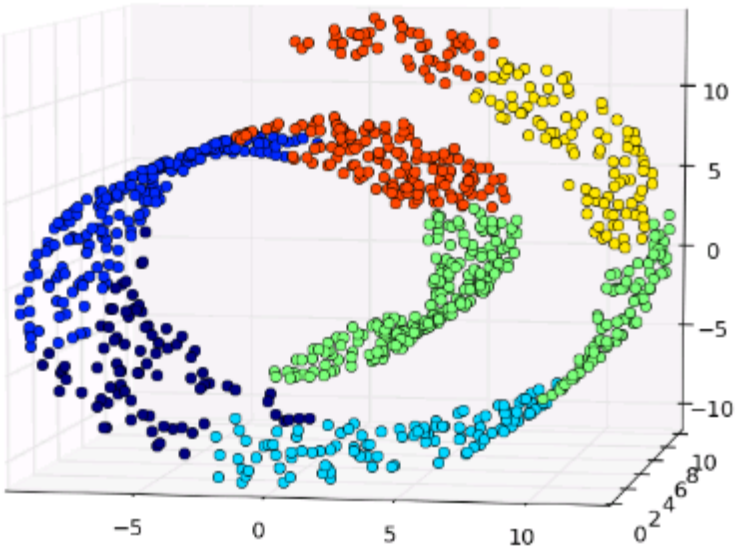
Matthias Grundmann[1,2], Vivek Kwatra[2],
Mei Han[2] and Irfan Essa[1]

[1]Georgia Tech [2]Google Research

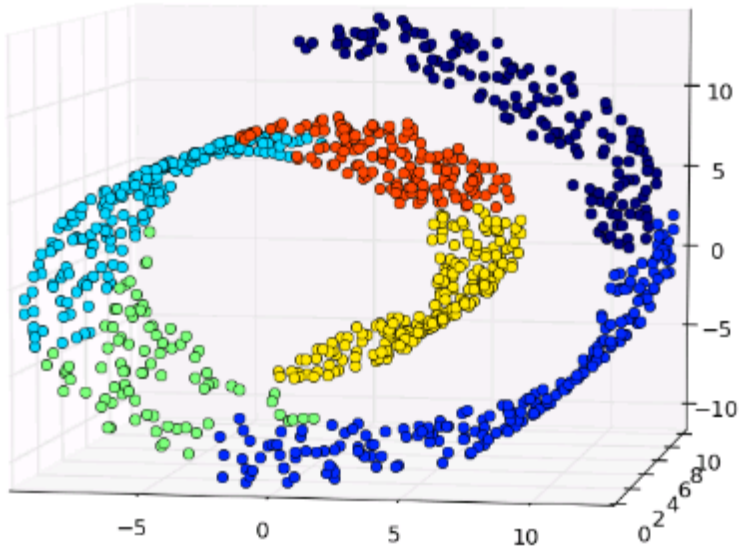
IEEE CVPR, San Francisco, USA, June 2010

k-means looks for blobs, other techniques do spirals b/c hard for k-means
scikit learn you can specify connectivity like in a spiral.

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

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

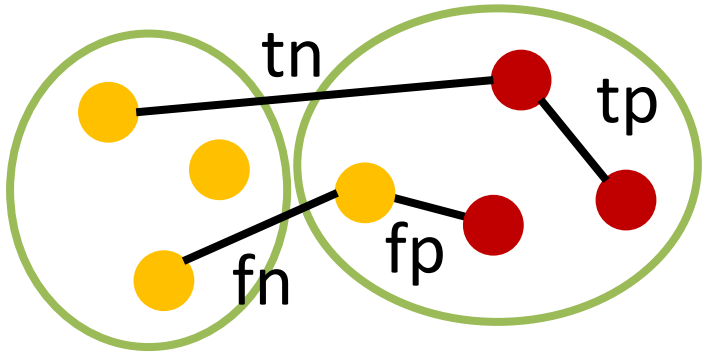
True positive: same clusters and should have been
True negative: diff clusters and should have been

Rand Index

False positive: same cluster, should've been diff
False negative: diff cluster, should've been the same

- Percentage of correct classifications
- Compare pairs of elements:

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

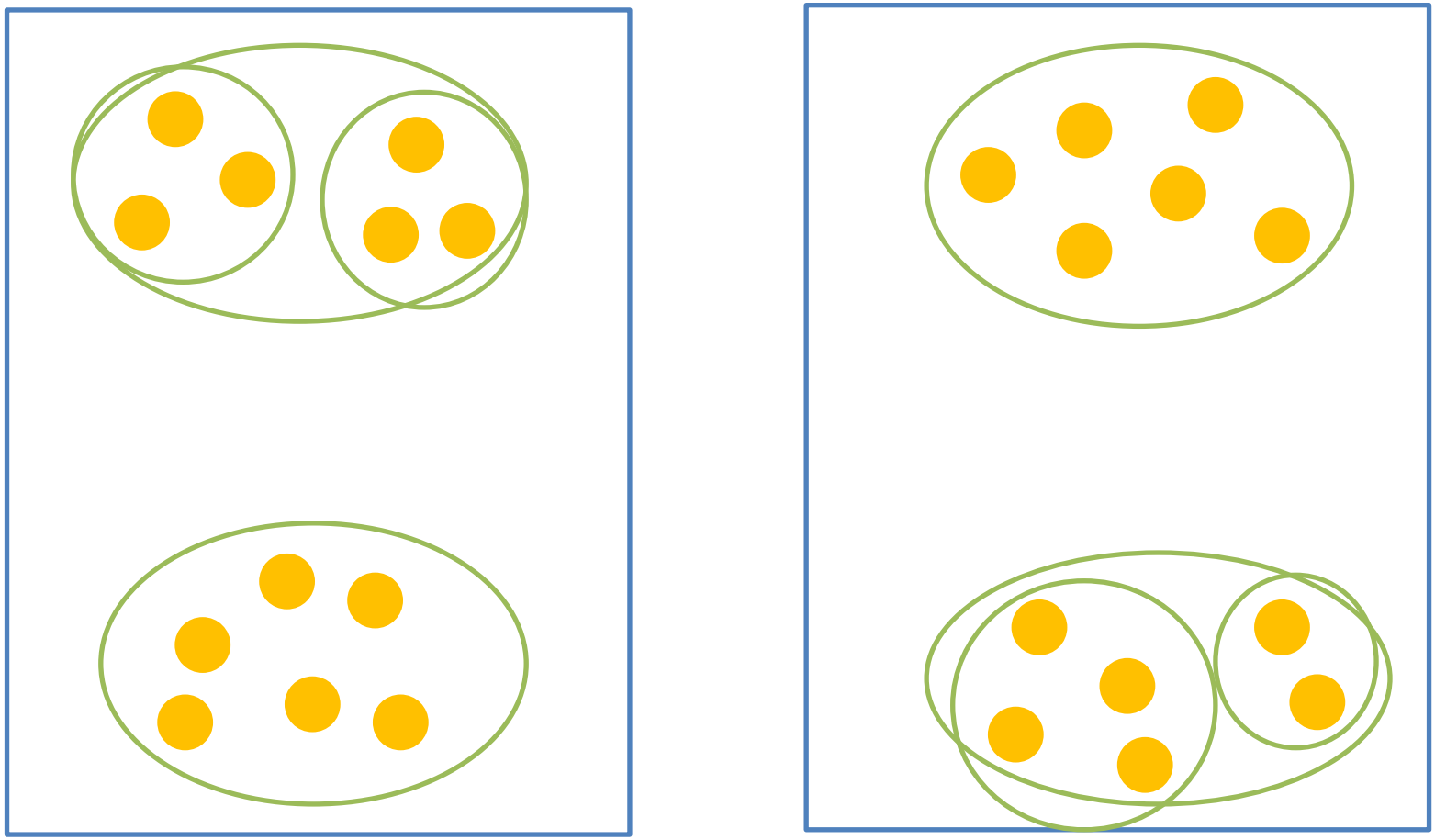


- Fp and fn are equally weighted

rand index: need labels to determine should have

split data: does clustering system (ie, k = 2 or 3) explain both train and validate sets well?

Stability



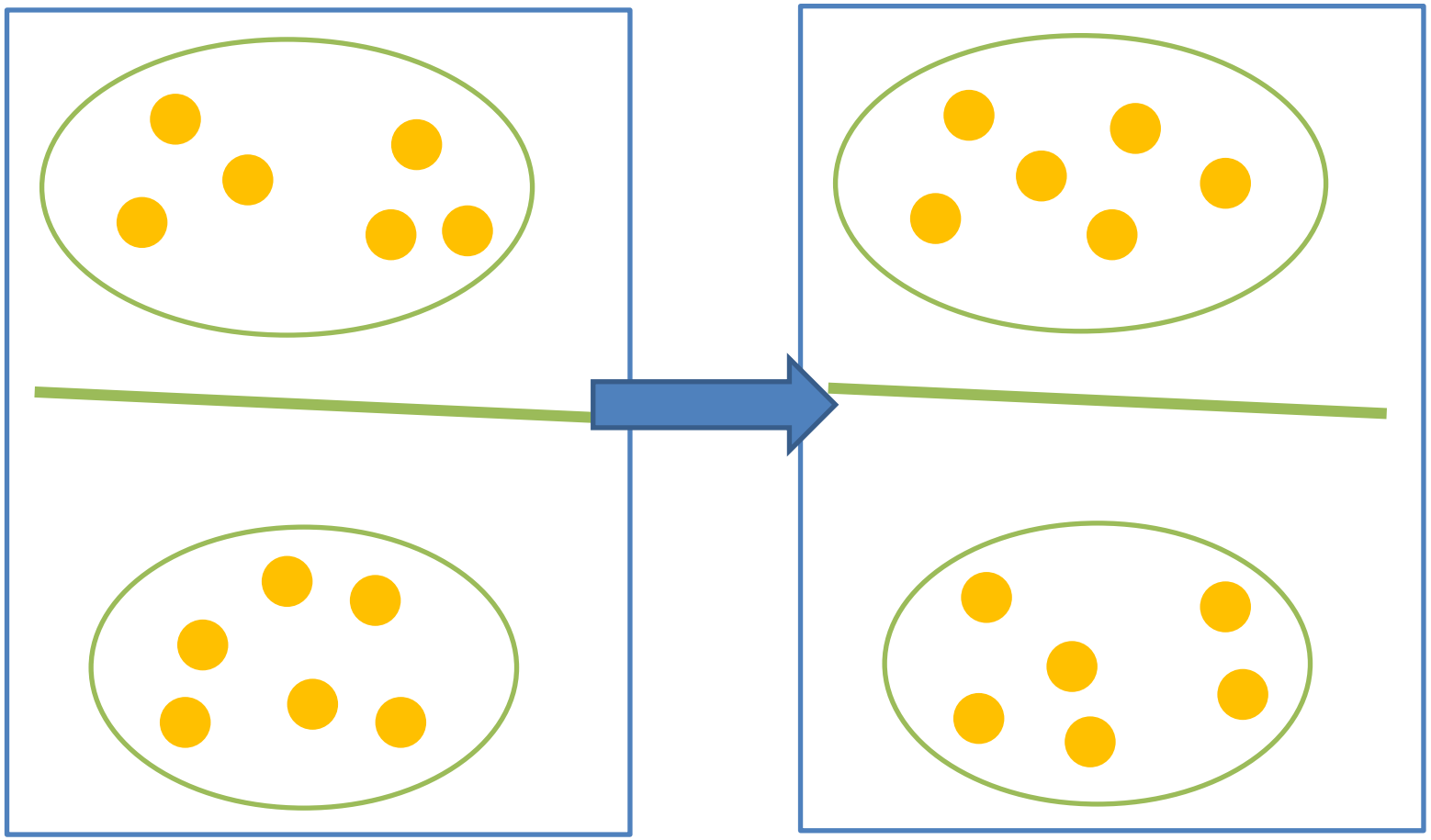
Stability

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

- Clustering should generalize!

Turn into supervised problem: make up a y (make up labels)
Apply labels to the validate/test set, compute an error, if low stability is high :)

Stability



Summary

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