CSE 6242 A / CS 4803 DVA

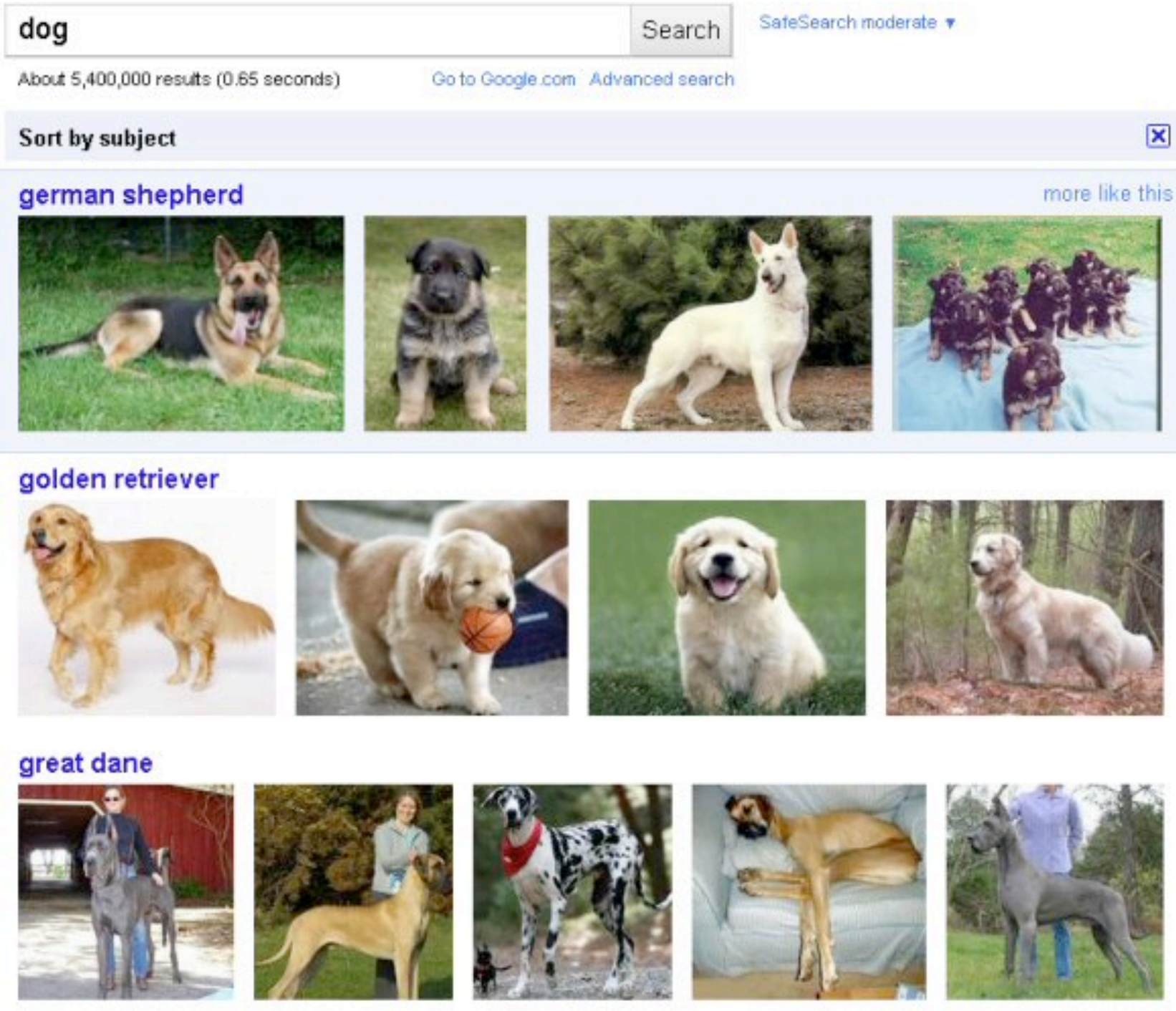
Feb 7, 2013

Clustering

Duen Horng (Polo) Chau
Georgia Tech

Partly based on materials by
Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos, Le Song

Clustering in Google Image Search

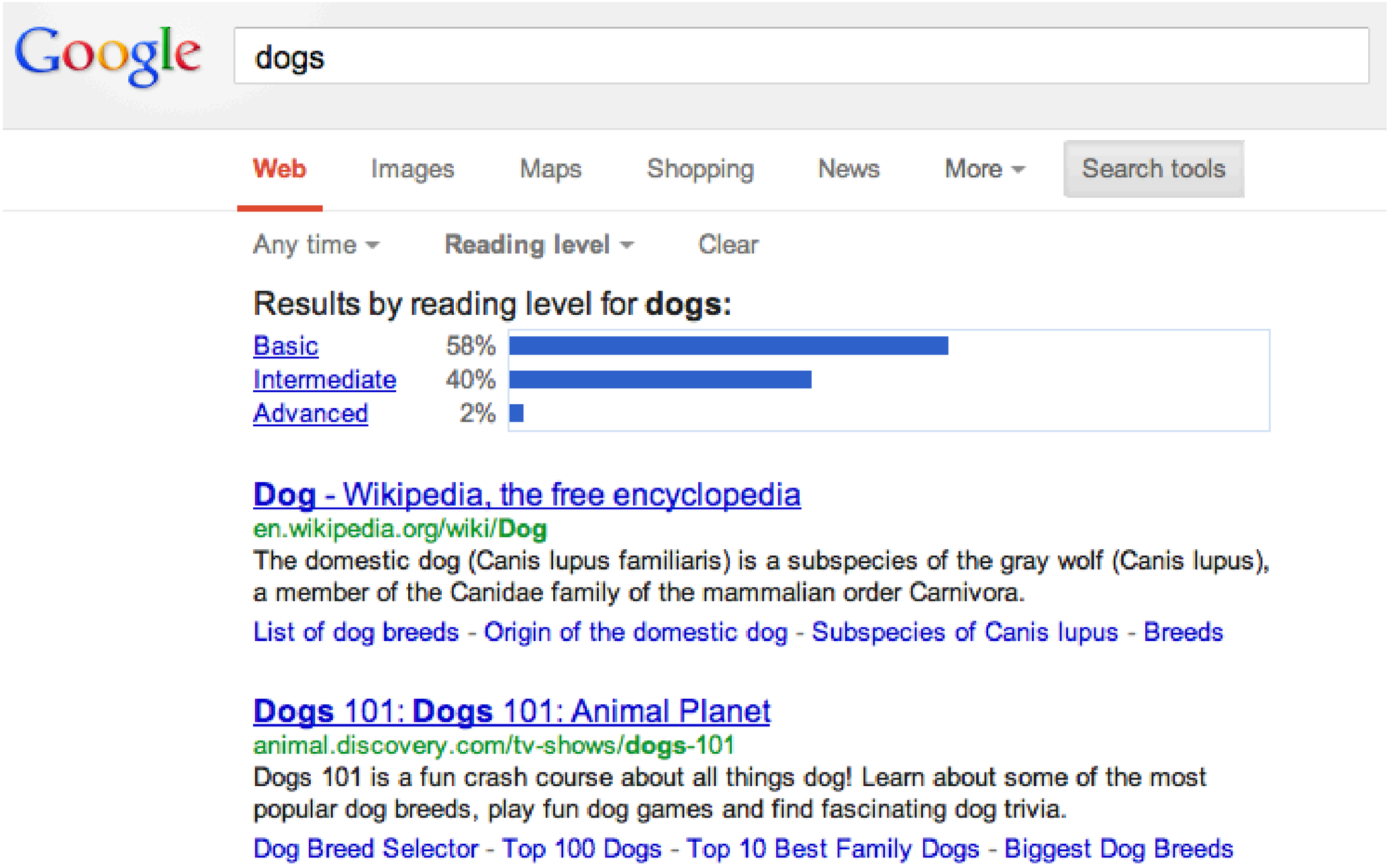


How would you build this?

Video: http://youtu.be/WosBs0382SE

http://googlesystem.blogspot.com/2011/05/google-image-search-clustering.html

Clustering in Google Search



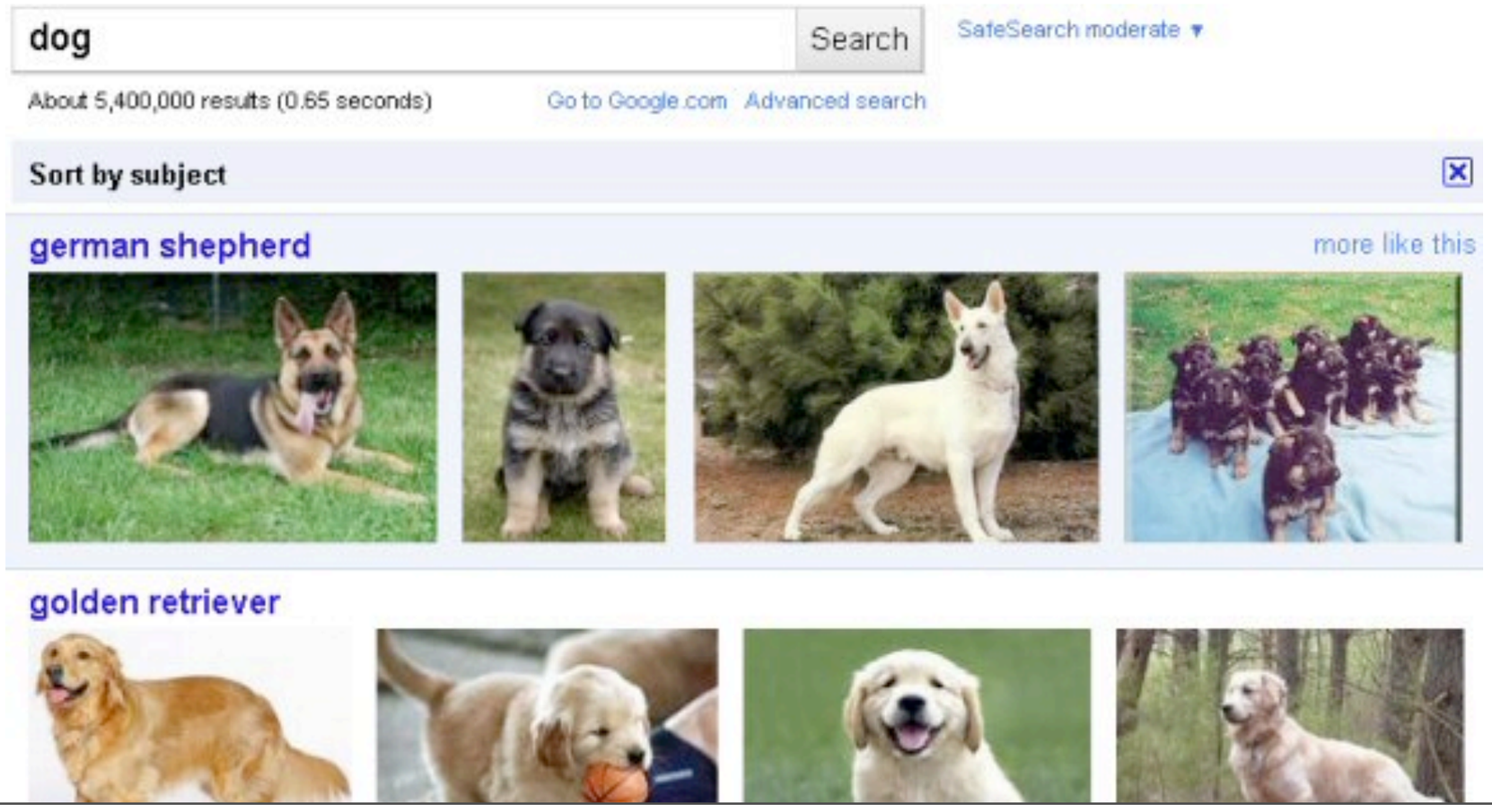
How would you build this?

Clustering

The most common type of **unsupervised** learning

High-level idea: group **similar** things together

"**Unsupervised**" because clustering model is learned without any labeled examples
(e.g., here are some pictures of dog, group them by their breed)



Applications of Clustering

Group pictures by subjects (as in Google image search)

figure out if there are clusterings in the data

- for exploratory analysis

grouping related research papers (e.g., by research areas)

grouping songs

- apple genius

similar users

- for recommendation
- fraud detection (outlier detection)

Clustering techniques you've got to know

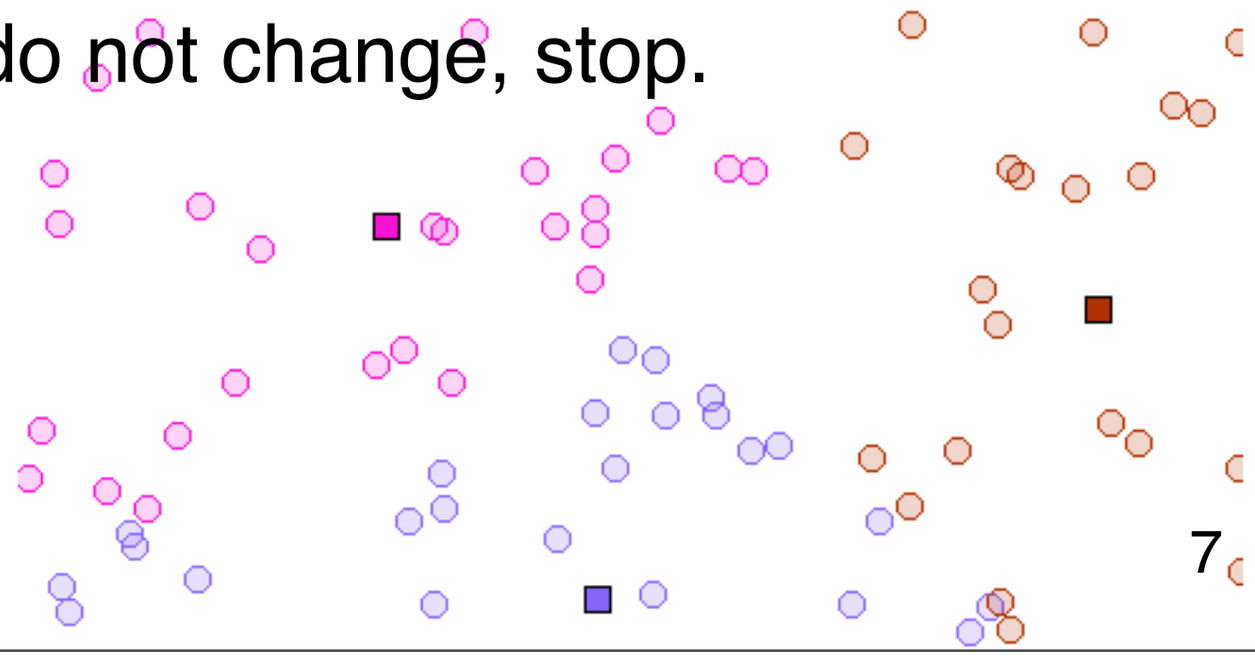
K-means
Hierarchical Clustering
(DBSCAN)

K-means (the “simplest” technique)

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

Summary

- We tell K-means the value of **k** (#clusters we want)
- **Randomly** initialize the k cluster “means” (“centroids”)
- **Assign** each item to the the cluster whose mean the item is closest to
- **Update** the new “means” of all k clusters.
- If all items’ assignments do not change, stop.



K-means What's the catch?

Need to **decide k ourselves**.

- How to find the optimal k?
  (more on this a few slides down)

Only locally optimal

- Different initialization gives different clusters
  - How to "fix" this?
- Bad starting points can cause algorithm to converge slowly

Hierarchical clustering

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

High-level idea: build a tree (hierarchy) of clusters

Divisive (top-down)

- Start with all items as *one cluster*
- Then iteratively divide into smaller clusters
- Too slow
  (why? need to consider all cut, to choose the best cut)

Agglomerative (bottom-up)

- Start with individual items
- Then iteratively group into larger clusters

Ways to calculate distances between two clusters

Single linkage

- minimum of distance between clusters

Complete linkage

- maximum of distance between clusters

Average linkage

- distance between cluster centers

Scatter Gather

http://people.ischool.berkeley.edu/~hearst/research/scattergather.html

Clustering in interactive application

Projects

You can choose your topics.

- Must contain data, algorithm, UI (can be visual, gesture-controlled, etc.)

Can be fun...

- Gmail Motion http://youtu.be/Bu927_ul_X0
- Gmail Tap http://youtu.be/1KhZKNZO8mQ