

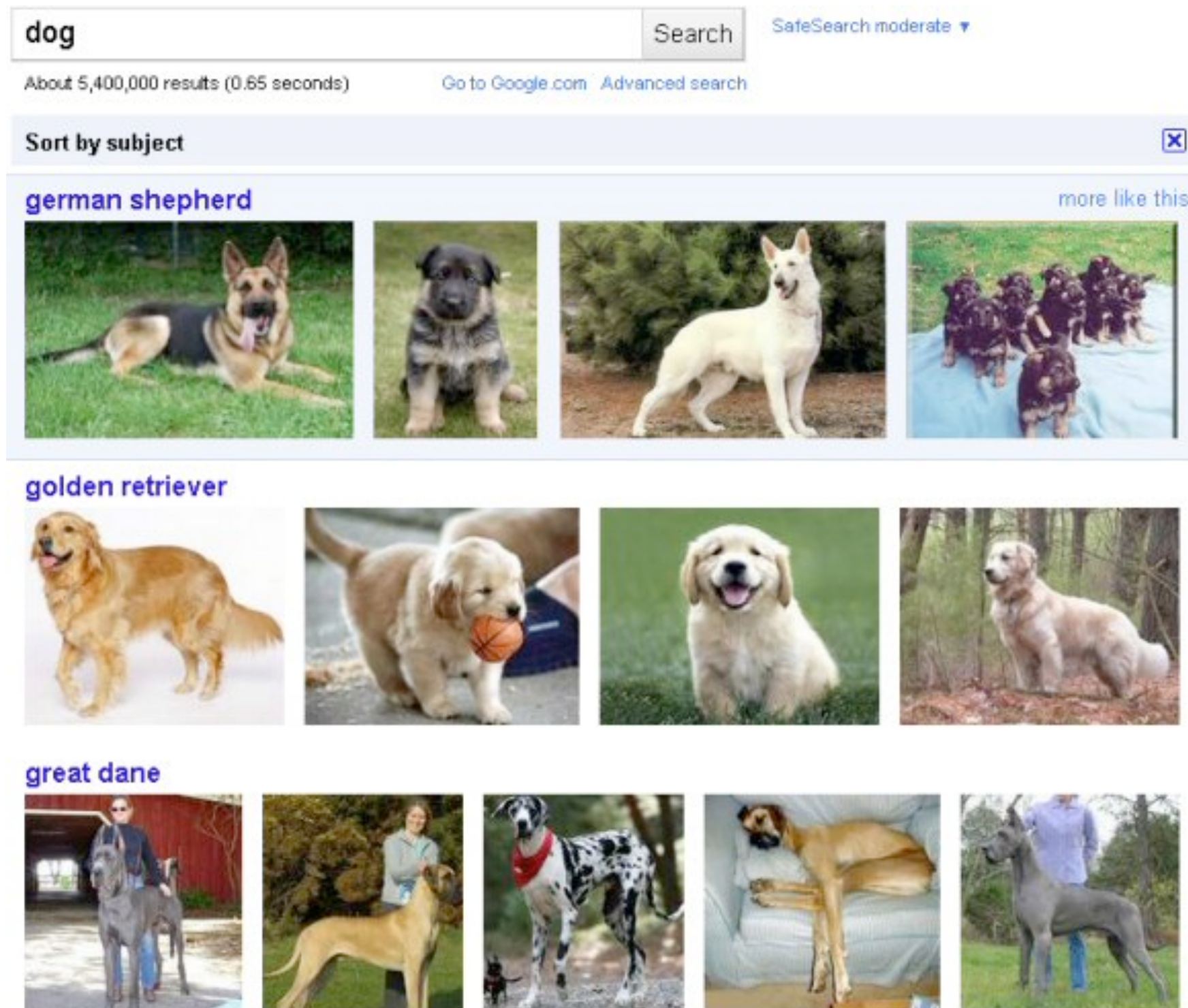
CSE 6242 / CX 4242

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

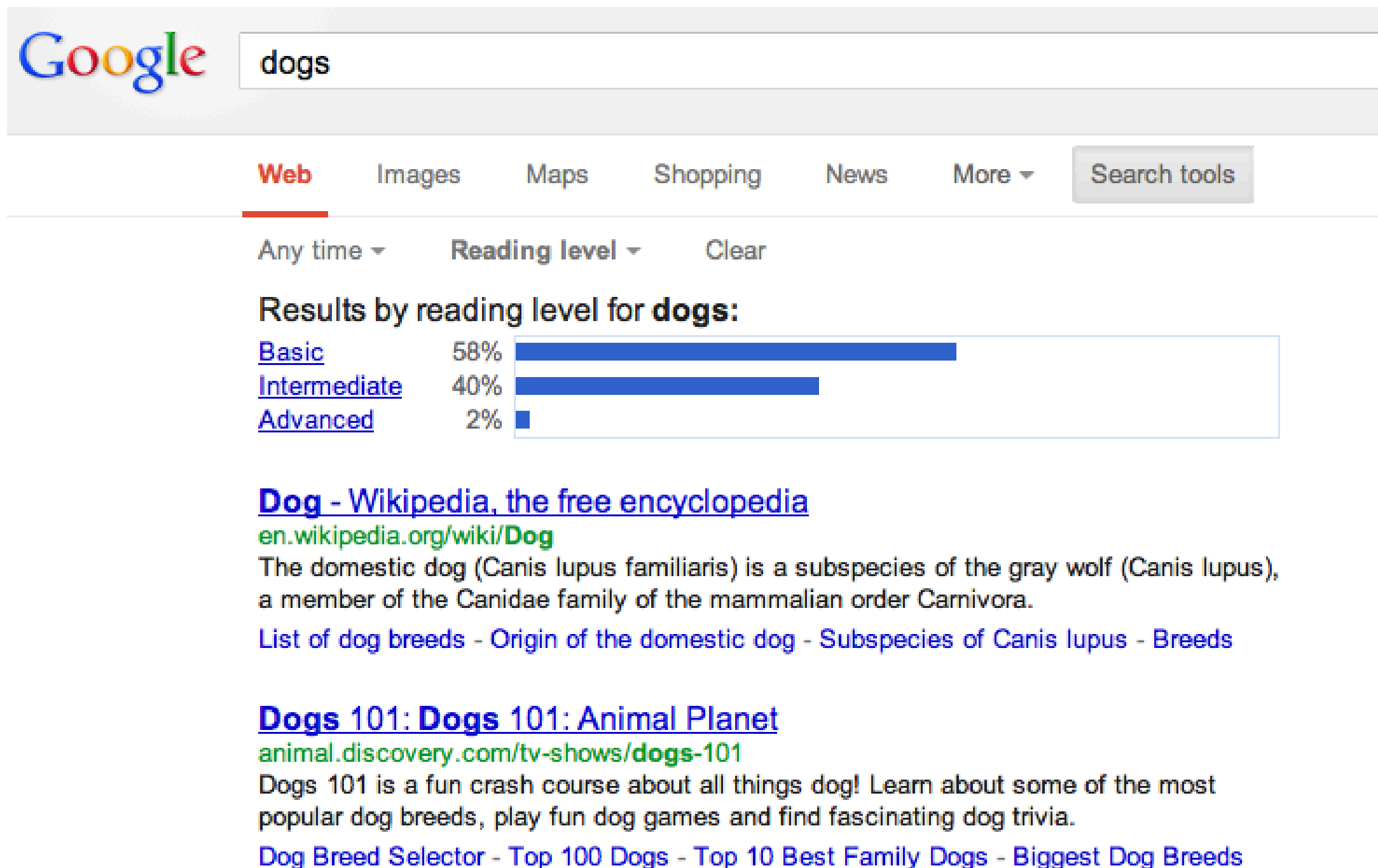


How would you build this?

Video: <http://youtu.be/WosBs0382SE>

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

Clustering in Google Search



The screenshot shows a Google search for "dogs". The search bar contains "dogs". Below the search bar, the "Web" tab is selected. There are filters for "Any time", "Reading level", and "Clear". The "Reading level" filter is expanded, showing a bar chart with three categories: Basic (58%), Intermediate (40%), and Advanced (2%). Below the chart, there are two search results. The first result is "Dog - Wikipedia, the free encyclopedia" with a link to "en.wikipedia.org/wiki/Dog". The second result is "Dogs 101: Dogs 101: Animal Planet" with a link to "animal.discovery.com/tv-shows/dogs-101".

Google dogs

Web Images Maps Shopping News More Search tools

Any time Reading level Clear

Results by reading level for **dogs**:

Reading Level	Percentage
Basic	58%
Intermediate	40%
Advanced	2%

[Dog - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Dog
The domestic dog (*Canis lupus familiaris*) is a subspecies of the gray wolf (*Canis lupus*), a member of the Canidae family of the mammalian order Carnivora.
[List of dog breeds](#) - [Origin of the domestic dog](#) - [Subspecies of Canis lupus](#) - [Breeds](#)

[Dogs 101: Dogs 101: Animal Planet](#)
animal.discovery.com/tv-shows/dogs-101
Dogs 101 is a fun crash course about all things dog! Learn about some of the most popular dog breeds, play fun dog games and find fascinating dog trivia.
[Dog Breed Selector](#) - [Top 100 Dogs](#) - [Top 10 Best Family Dogs](#) - [Biggest Dog Breeds](#)

How would you build this?

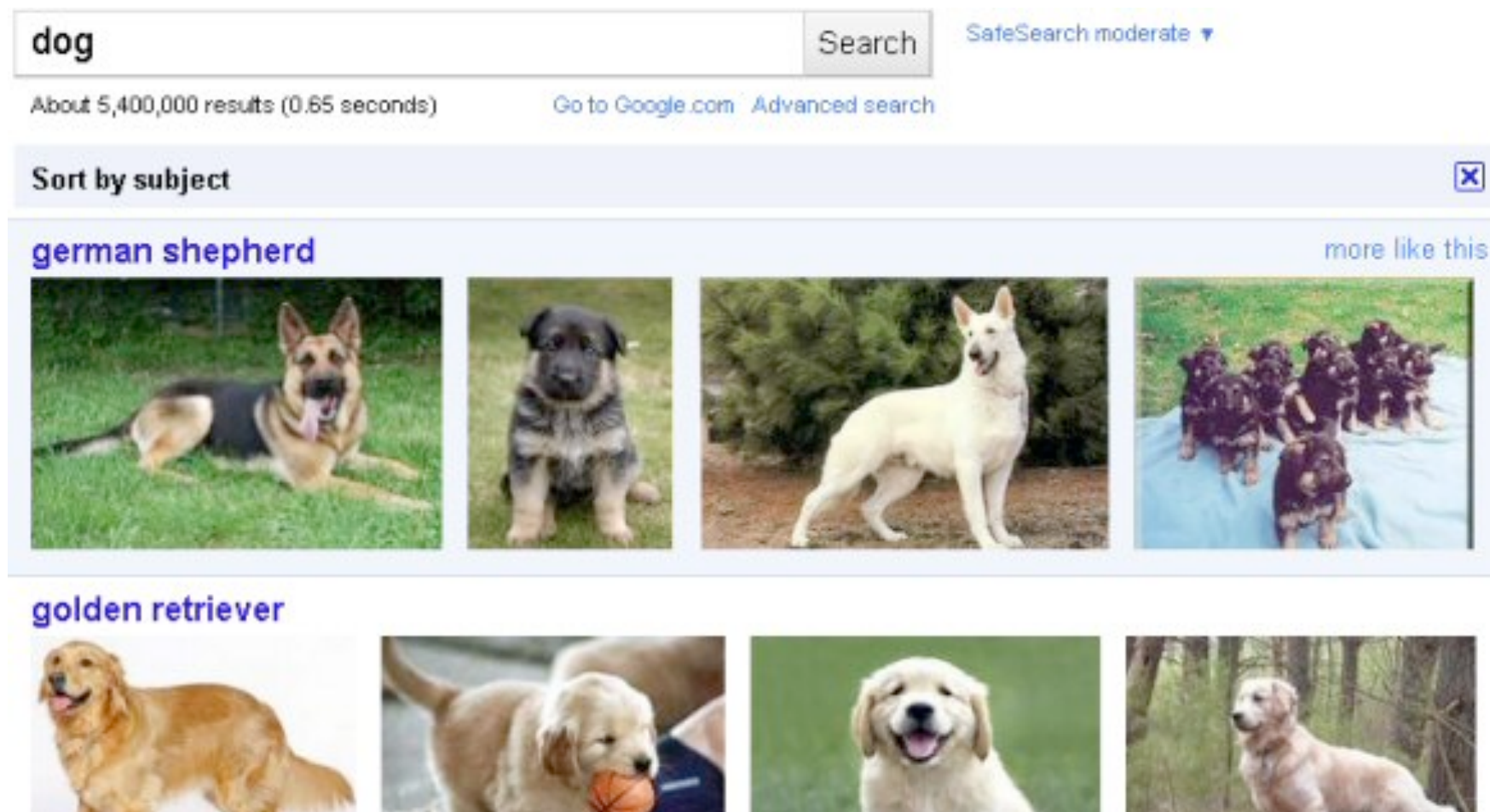
Clustering

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

- google news
- IMDB (movie sites)
- anomaly detection
- detecting population subgroups (community detection)
 - as in healthcare
- Twitter hashtags
 - text-based clustering
- (Age detection)

Clustering techniques you've got to know

K-means

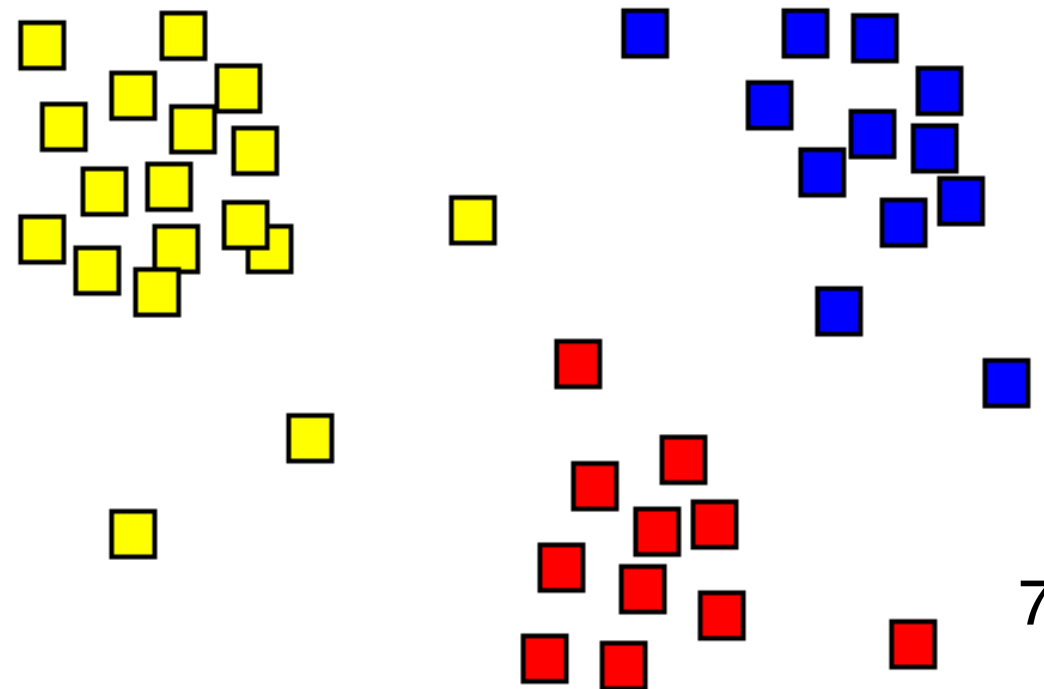
Hierarchical Clustering
(DBSCAN)

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 (so, we need a **similarity function**)
- **Update** the new “means” of all k clusters.
- If all items’ assignments do not change, stop.



K-means What's the catch?

<http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html>

Need to **decide k ourselves**.

- How to find the optimal k?

Only locally optimal (vs global)

- Different initialization gives different clusters
 - How to “fix” this?
- “Bad” starting points can cause algorithm to converge slowly
- Can work for relatively large dataset
 - Time complexity $O(n \log n)$

Hierarchical clustering

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

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

Agglomerative (bottom-up)

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



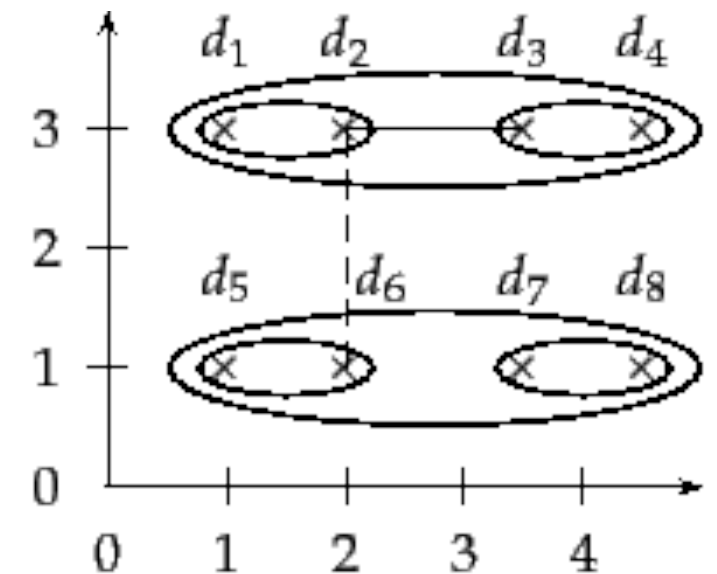
Divisive (top-down)

- Start with all items as *one cluster*
- Then iteratively divide into smaller clusters

Ways to calculate distances between two clusters

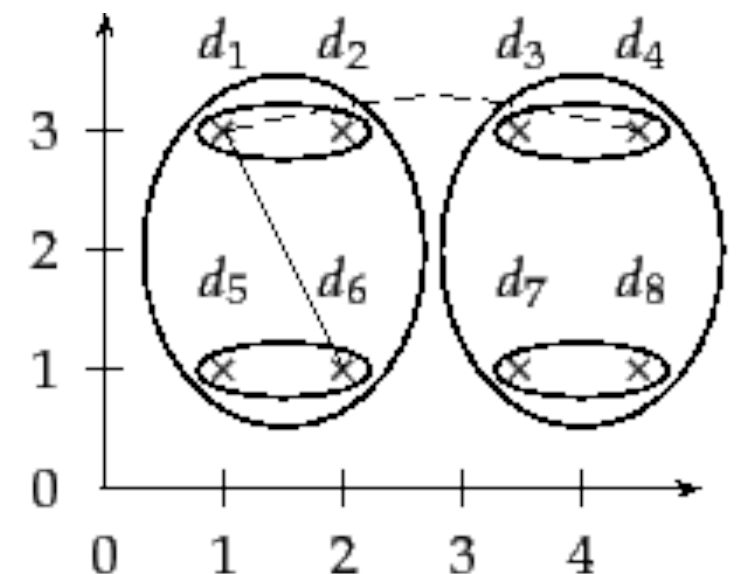
Single linkage

- minimum of distance between clusters
- similarity of two clusters = similarity of the clusters' **most similar** members



Complete linkage

- maximum of distance between clusters
- similarity of two clusters = similarity of the clusters' **most dissimilar** members



Average linkage

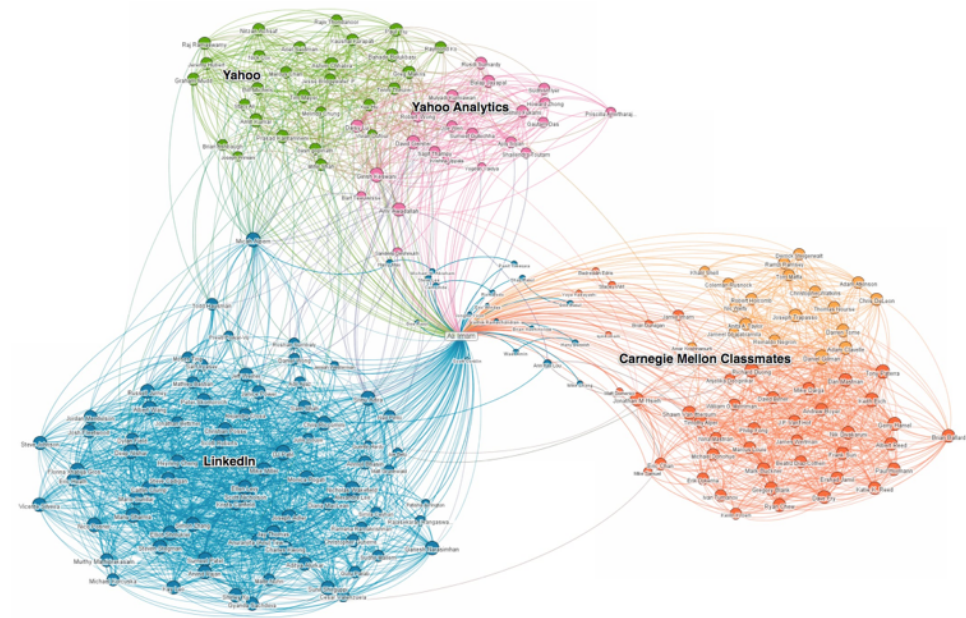
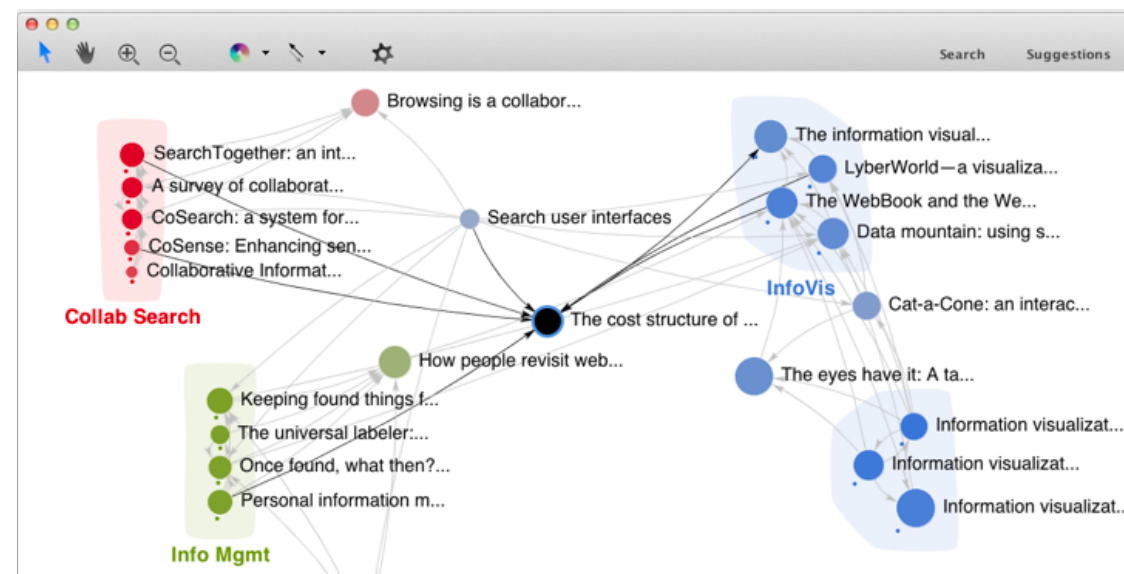
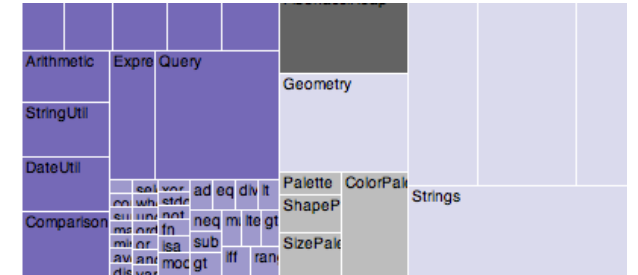
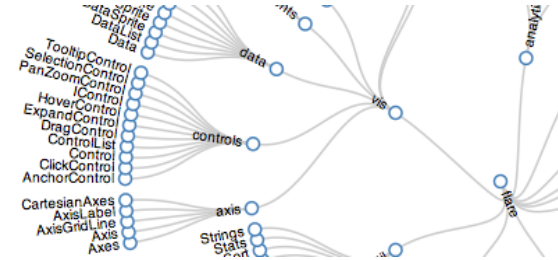
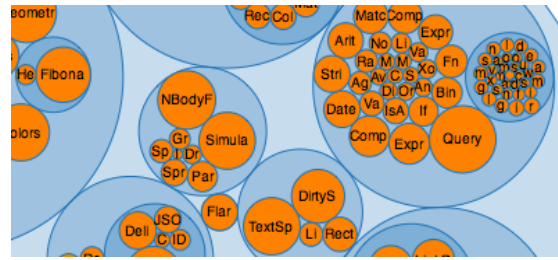
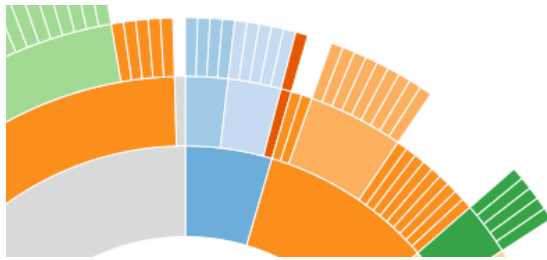
- distance between cluster centers

Hierarchical clustering for large datasets?

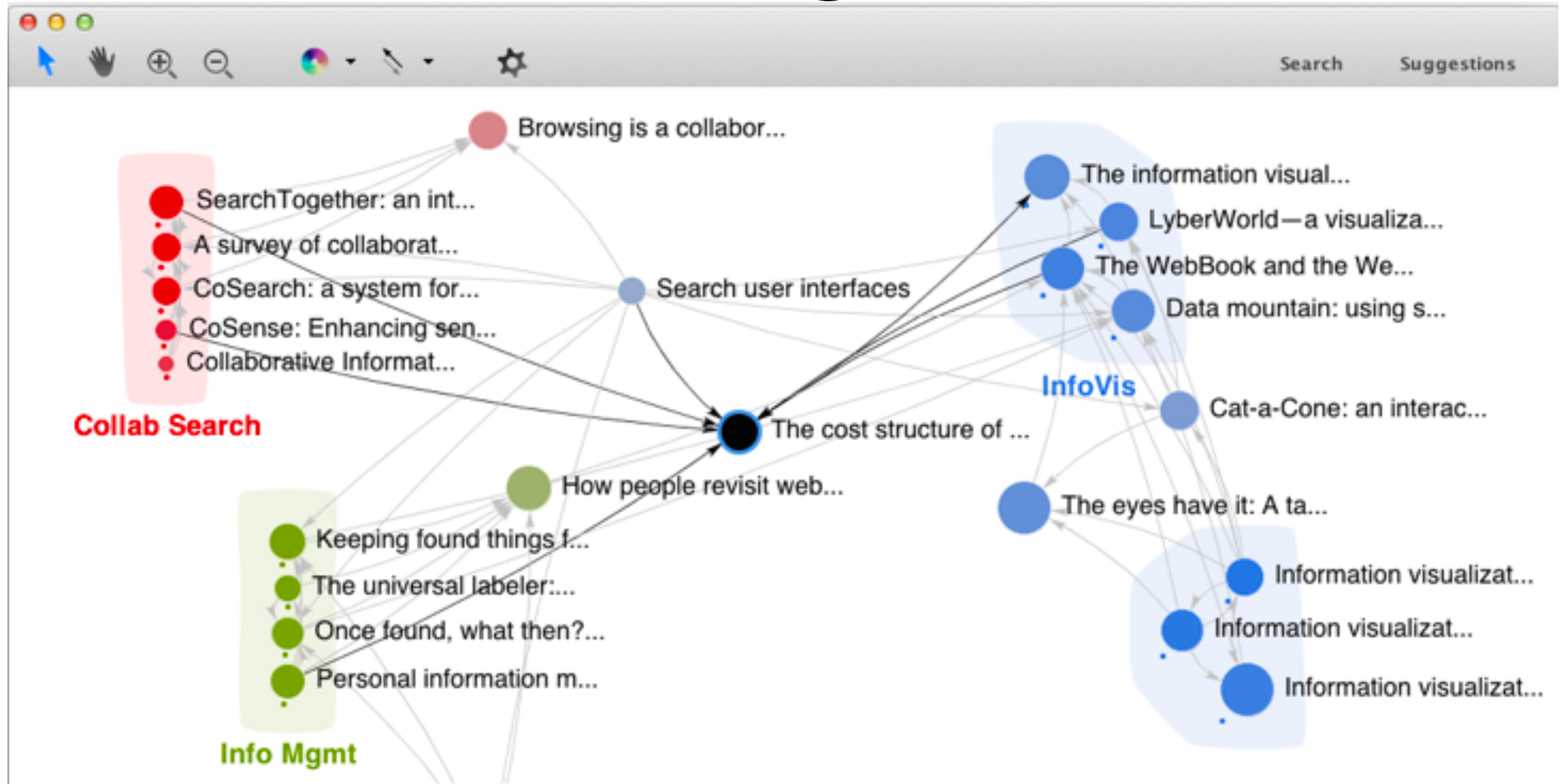
- OK for small datasets (e.g., <10K items)
- Time complexity between $O(n^2)$ to $O(n^3)$ where n is the number of data items
- Not good for millions of items or more
- But great for understanding concept of clustering

Visualizing Clusters

<https://github.com/mbostock/d3/wiki/Hierarchy-Layout>



Visualizing Clusters



Visualizing Clusters

