

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

dog Search SafeSearch moderate


About 5,400,000 results (0.65 seconds) Go to Google.com Advanced search

Sort by subject


german shepherd more like this



golden retriever



great dane



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, there are navigation tabs for 'Web', 'Images', 'Maps', 'Shopping', 'News', and 'More', along with a 'Search tools' button. The 'Web' tab is selected. Below the navigation, there are filters for 'Any time', 'Reading level', and 'Clear'. The 'Reading level' filter is expanded, showing a bar chart with three categories: 'Basic' at 58%, 'Intermediate' at 40%, and 'Advanced' at 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:

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)

The screenshot shows a Google search interface. At the top, the search bar contains the word "dog" and a "Search" button. Below the search bar, it indicates "About 5,400,000 results (0.65 seconds)" and provides links for "Go to Google.com" and "Advanced search". A "SafeSearch moderate" dropdown menu is visible on the right. Below the search results, there is a "Sort by subject" button. The first cluster is labeled "german shepherd" and contains four images: a German Shepherd lying on grass, a black and tan puppy, a white German Shepherd standing, and a group of dark-colored puppies. A "more like this" link is present to the right of the images. The second cluster is labeled "golden retriever" and contains four images: a golden retriever sitting, a golden retriever puppy with an orange, a golden retriever puppy's face, and a golden retriever standing in a wooded area.

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

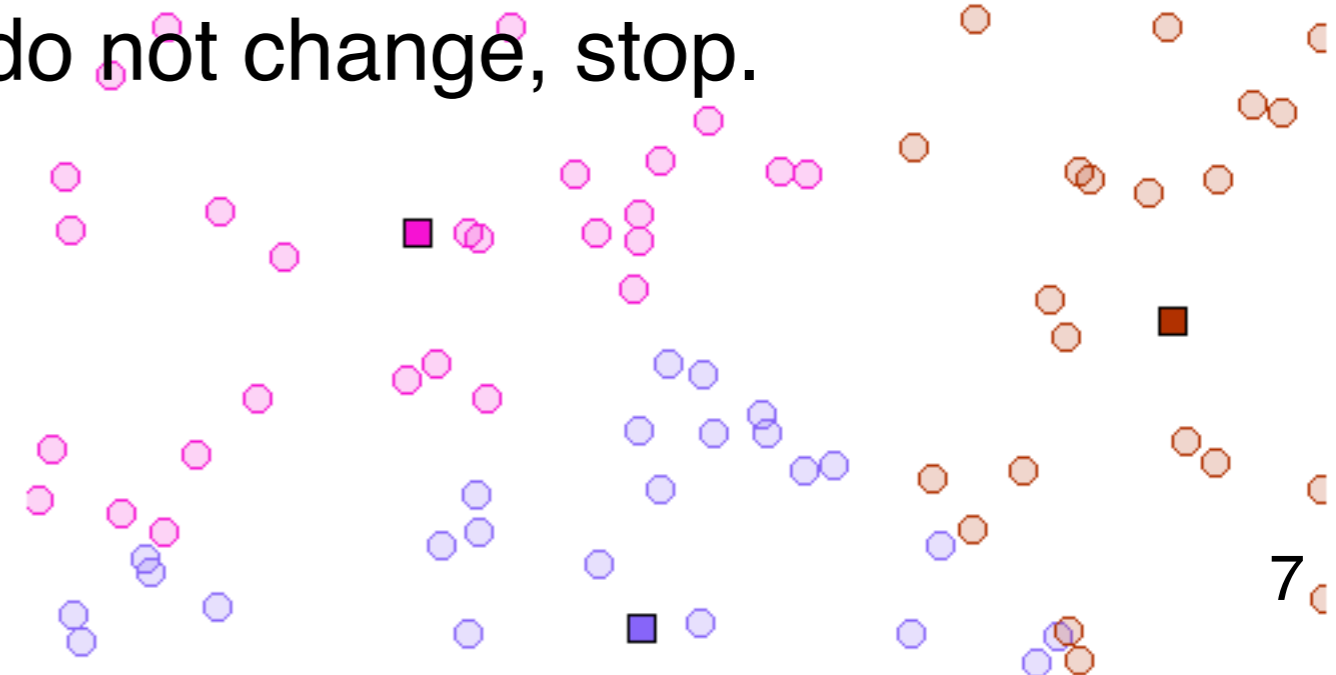
(DBSCAN)

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>