

Clustering and Similarity: Retrieving Documents



Emily Fox & Carlos Guestrin

Machine Learning Specialization

University of Washington

Retrieving documents of interest

Document retrieval

- Currently reading article you like



Document retrieval

- Currently reading article you like
- **Goal:** Want to find similar article



Document retrieval



Challenges

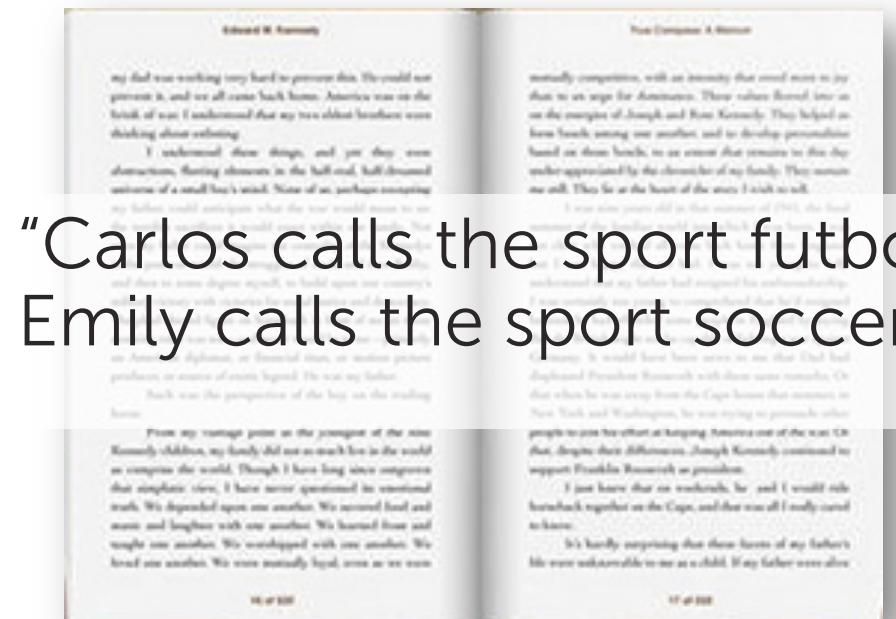
- How do we measure similarity?
- How do we search over articles?



Word count representation for measuring similarity

Word count document representation

- Bag of words model
 - Ignore order of words
 - Count # of instances of each word in vocabulary



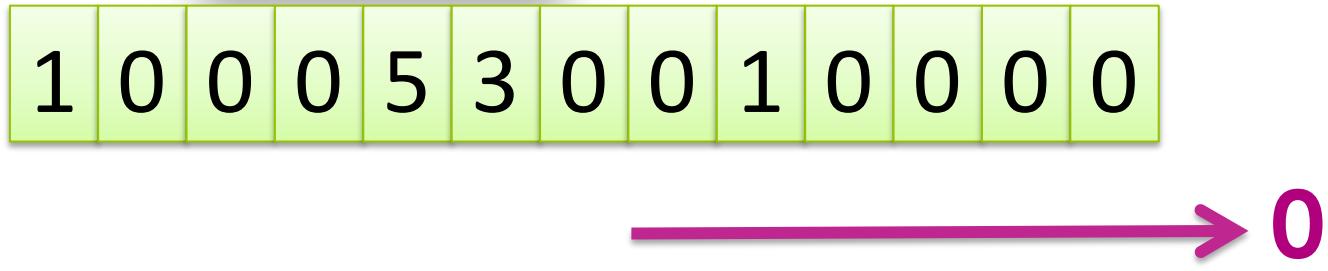
Measuring similarity



$$\begin{array}{r} 1*3 \\ + \\ 5*2 \\ \hline = 13 \end{array}$$



Measuring similarity



Issues with word counts – Doc length

But that was taking my best to generate this. The world can never be as we all can see it because America is run by the brain of us. And I understand that my two other brothers were thinking the same.

I submitted these things, and yet they were dismissed, then I went to the Senate, and I had discussed them with Senator John Kennedy. He said, "John, my father could encourage what the world would want to see in the United States." So I went to the White House, and my father could imagine the kind of knowledge the Senator is the person with the strength of it, and, I think, he did. Then I went to the White House again, and I had a meeting with Secretary of State and Defense. I told them that the world would want to see that the American people will stand up and be ready to be used as American diplomats or financial ones, or nation protectors, or whatever.

That was the prognosis of the man on the reading from.

From my point come as the progeny of the one family children, my family can't work much in the world because we have no money. We have to depend on our own strength. I have some strength to understand math. We depended upon me to understand math. We learned how to do math. We learned how to do science. We worked with one another. We lived one another. We must always live one another as we are naturally competitive, but I think the first time is to let the world see for us. Those values reflect more on the strength of Joseph and Rose Kennedy. They had a home, and they had a house, and they had a car, and they had a bank account on hand, so in that sense it's not the wealth but the strength of the family. That's where the world would want to see.

I am now approaching at the moment of 65, the head of the family, and I have a son, and I have a daughter, and I am not sure who he had all come back from Europe. I am not happy that I am going to be a young old father, but I am not too worried about it. I am not too worried about it.

I am actually so young, I am probably that I am going to be a good dad. I have had a lot of experience with children, and I am probably of higher age than most parents. It would have been nice to see as the old dad had a son, and I am not too worried about it. I am not too worried about it. The way it was even from the Captains home in New York and Washington, we are trying to support the family, and we are trying to support the family. I think that the differences, though Kennedy concerned apparently more about the family, and the Captains less.

I know that on board, and I would like to be back there again. And the Captains, and the Captains less.

It's really surprising that three of my babies are still here, and we are so close to it. As if my father was still



3 0 0 0 2 0 0 1 0 1 0 0 0
Similarity = 13

6	0	0	0	4	0	0	2	0	2	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---

Similarity = 52

Solution = normalize

my flat was destroyed - by an explosion there. I could see a person's arm & a few other body parts. There was no one at the front of it & I understood that my two brothers were missing.

I remembered those things, and yet they were disconcerting, though I had been told, half-dozed off, stories about them. I could not understand how my father could envisage what he would move to or leave behind if he had to leave us. I could not understand how my father could imagine the character of the Committee in the particular world of strength & lack, of boldness & the need for caution, of the need for a certain kind of military strategy in our own affairs and decisions. I could not understand how my father could envisage the strong-willed men who would be called to the task of American diplomats, or financial men, or nation builders, or revolutionaries, or revolutionaries who had been革命派.

Such was the perspective of the boy in the reading room.

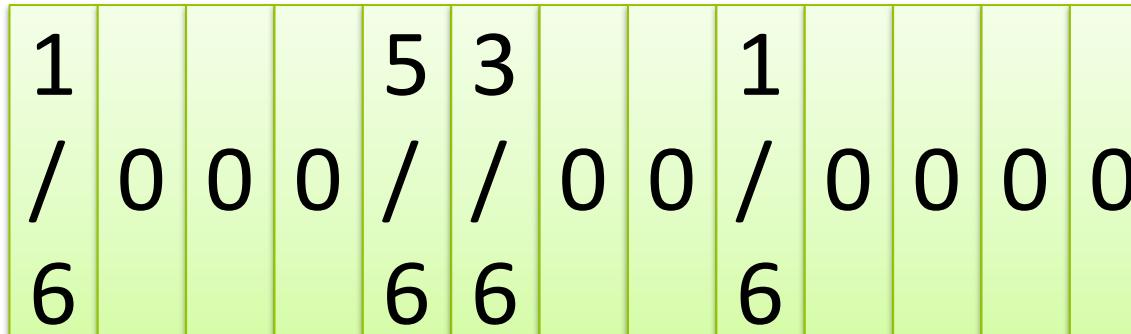
From my vantage point as the son of the man himself, I could see that he had been right to be in the world as he was, and that he had been right to do what he did. But that simple care, I have never questioned or remained unaffected by. I have never been able to understand how we stood and fought with one another. We learned from one another. We lived on another. We had to live by each other as we were

- 10 -

New Communist in Britain



$$\sqrt{1^2 + 5^2 + 3^2 + 1^2}$$



Prioritizing important words with tf-idf

Issues with word counts – Rare words



Common words in doc: “the”, “player”, “field”, “goal”

Dominate rare words like: “futbol”, “Messi”

Document frequency

- What characterizes a **rare word**?
 - Appears **infrequently** in the corpus
- Emphasize words appearing in **few docs**
 - Equivalently, discount word **w** based on
of docs containing w in corpus

Important words

- Do we want only rare words to dominate???
- What characterizes an **important word**?
 - Appears frequently in document (**common locally**)
 - Appears rarely in corpus (**rare globally**)
- Trade off between **local frequency** and **global rarity**

TF-IDF document representation

- Term frequency – inverse document frequency (tf-idf)



TF-IDF document representation

- Term frequency – inverse document frequency (tf-idf)
 - Term frequency



- Same as word counts



TF-IDF document representation

- Term frequency – inverse document frequency (tf-idf)
- Term frequency



- Inverse document frequency



$$\log \frac{\# \text{ docs}}{1 + \# \text{ docs using word}}$$

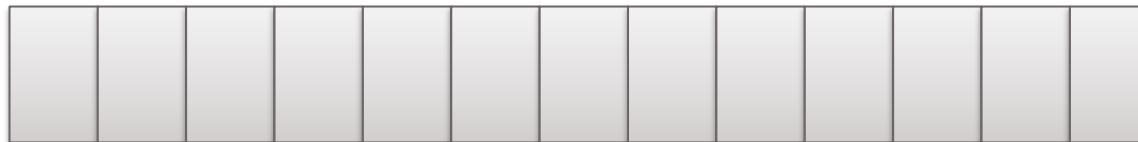


TF-IDF document representation

- Term frequency – inverse document frequency (tf-idf)
- Term frequency



- Inverse document frequency



$$\log \frac{\# \text{ docs}}{1 + \# \text{ docs using word}}$$

word in many docs → $\log \frac{\text{large } \#}{1 + \text{large } \#} \approx \log 1 = 0$

rare word → $\log \frac{\text{large } \#}{1 + \text{small } \#} \rightarrow \text{large } \#$

TF-IDF document representation

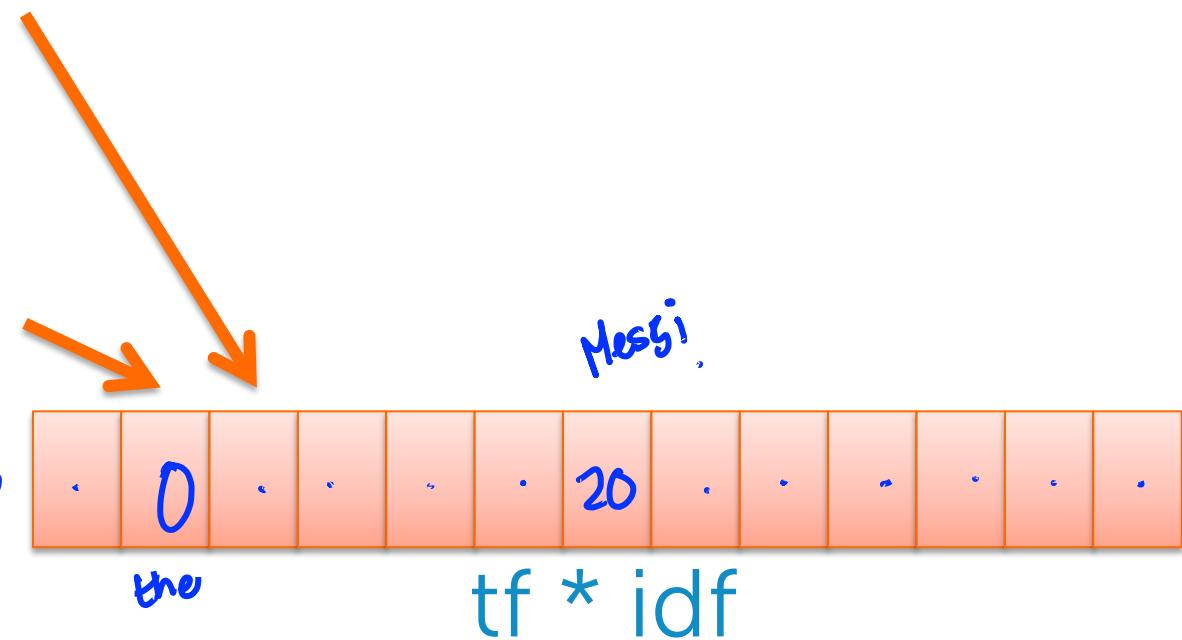
- Term frequency – inverse document frequency (tf-idf)
- Term frequency



- Inverse document frequency



$$\log \frac{64}{1+63} = 0$$
$$\log \frac{64}{1+3} = \log 16$$



Retrieving similar documents

Nearest neighbor search

- Query article:



- Corpus:



- **Specify:** Distance metric
- **Output:** Set of most similar articles



1 – Nearest neighbor

- **Input:** Query article 
- **Output:** *Most* similar article

- Algorithm:
 - Search over each article  in corpus
 - Compute $s = \text{similarity}(\text{query}, \text{article})$
 - If $s > \text{Best_s}$, record  = and set $\text{Best_s} = s$
 - Return 

k – Nearest neighbor

- **Input:** Query article
- **Output:** *List of k* similar articles



Clustering documents

Structure documents by topic

- Discover groups (*clusters*) of related articles



SPORTS

WORLD NEWS

What if some of the labels are known?

- Training set of labeled docs



SPORTS



WORLD NEWS

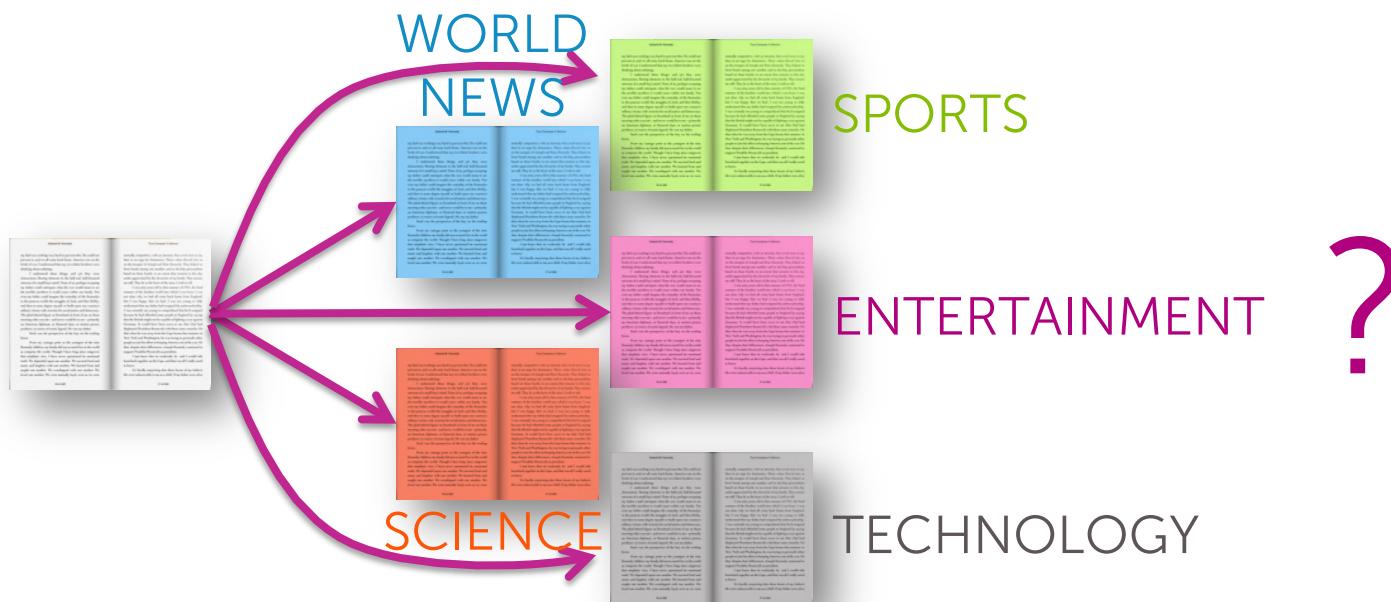


ENTERTAINMENT



SCIENCE

Multiclass classification problem

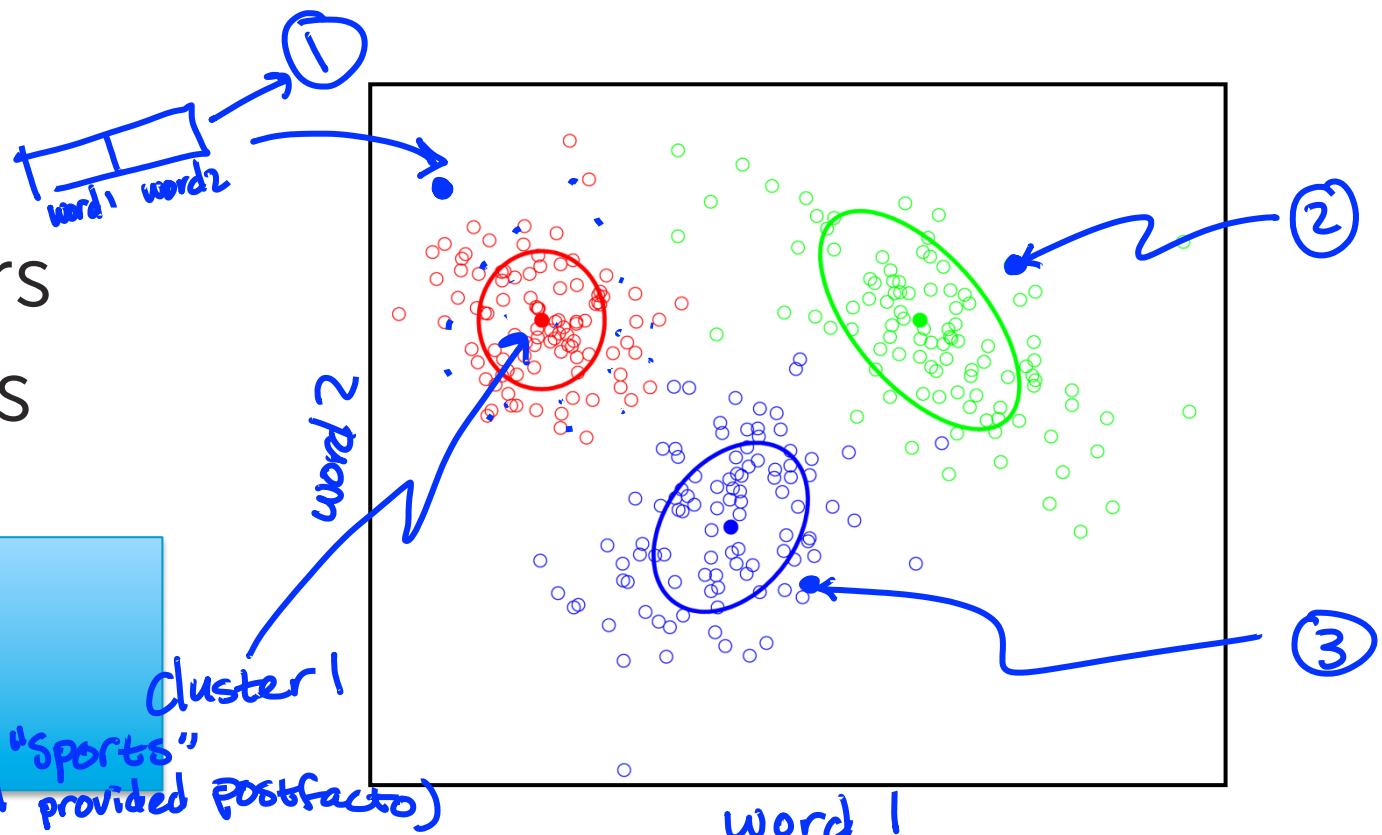


Example of
supervised learning

Clustering

- No labels provided
 - Want to uncover cluster structure
-
- **Input:** docs as vectors
 - **Output:** cluster labels

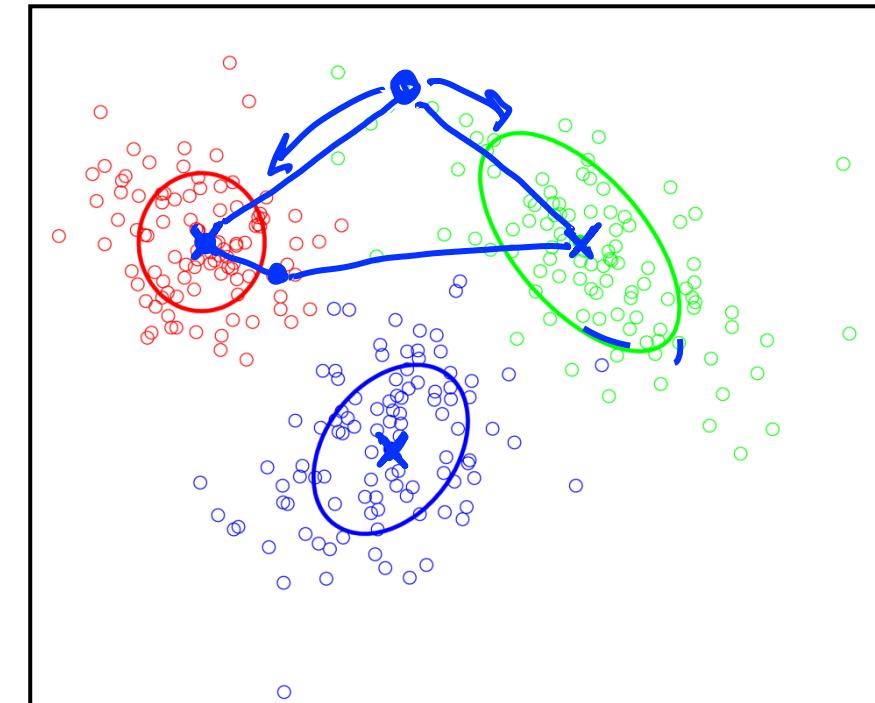
An unsupervised learning task
("Sports" (label provided postfacto))



What defines a cluster?

- Cluster defined by center & shape/spread

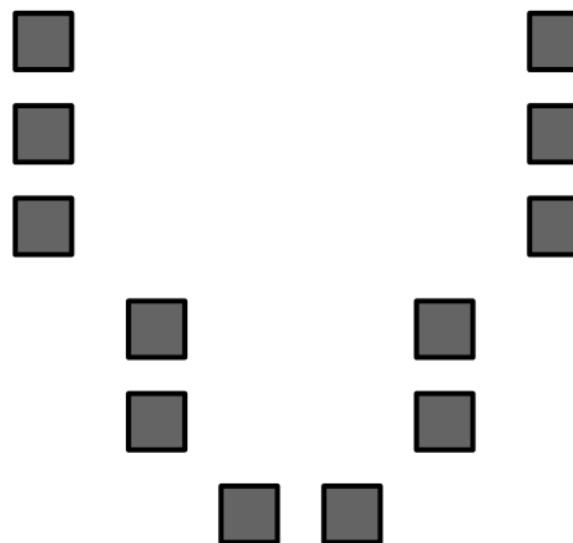
- Assign observation (doc) to cluster (topic label)
 - Score under cluster is higher than others
 - Often, just more similar to assigned cluster center than other cluster centers



k-means

- Assume

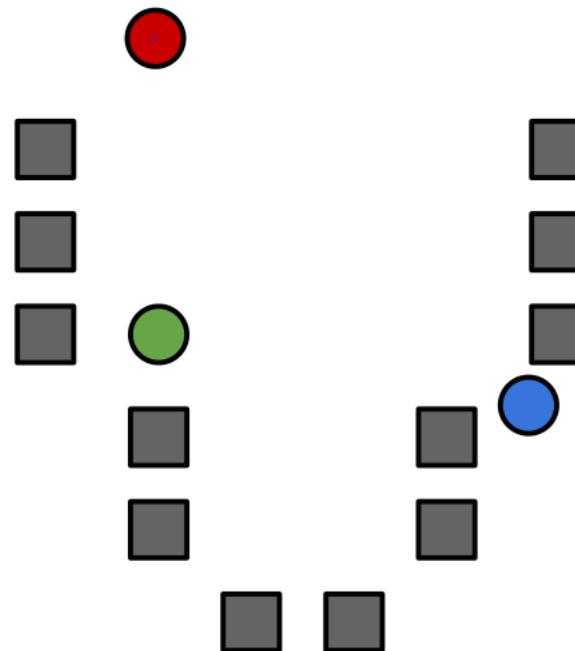
- Similarity metric =
distance to cluster
center
(smaller better)



DATA
to
CLUSTER

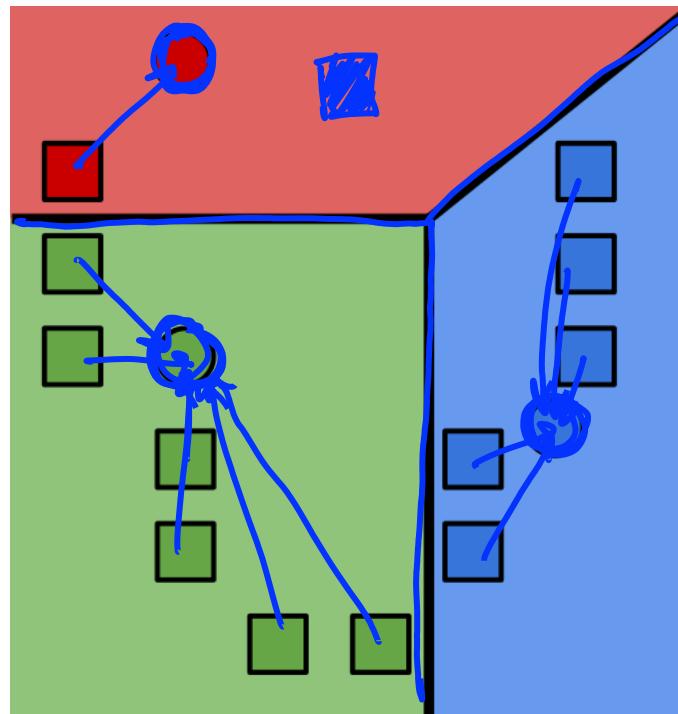
k-means algorithm

0. Initialize cluster centers



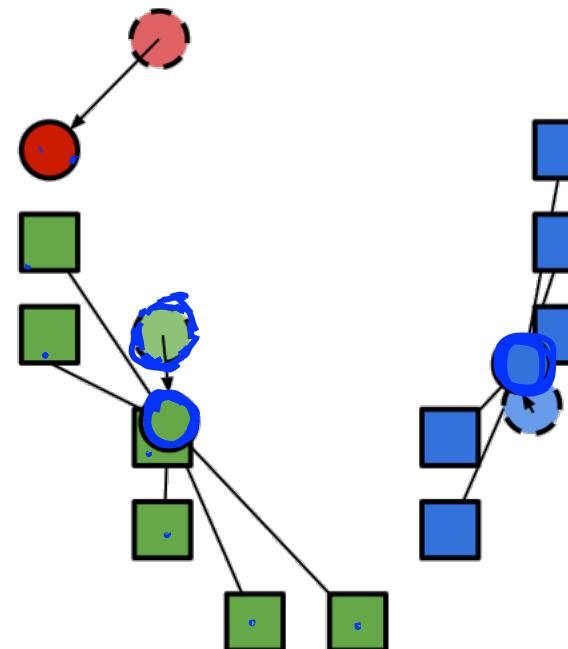
k-means algorithm

0. Initialize cluster centers
1. Assign observations to closest cluster center



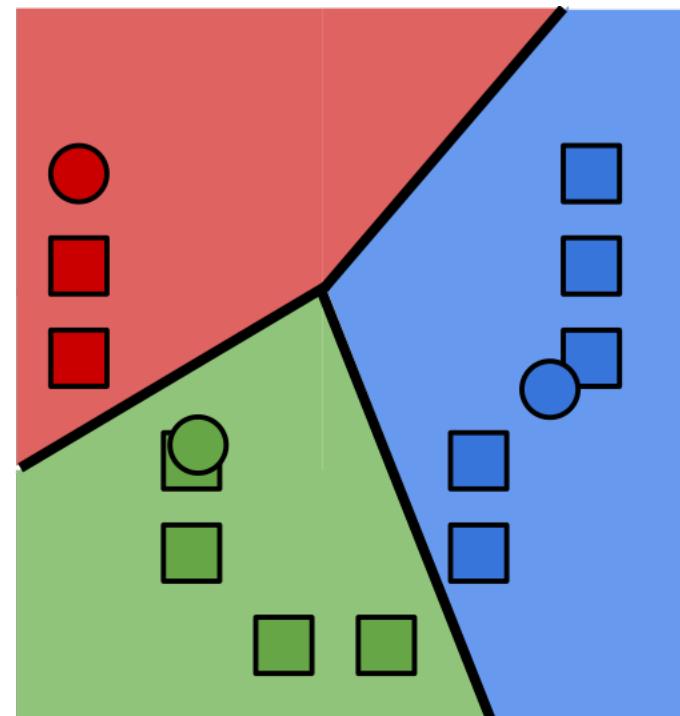
k-means algorithm

0. Initialize cluster centers
1. Assign observations to closest cluster center
2. Revise cluster centers as mean of assigned observations



k-means algorithm

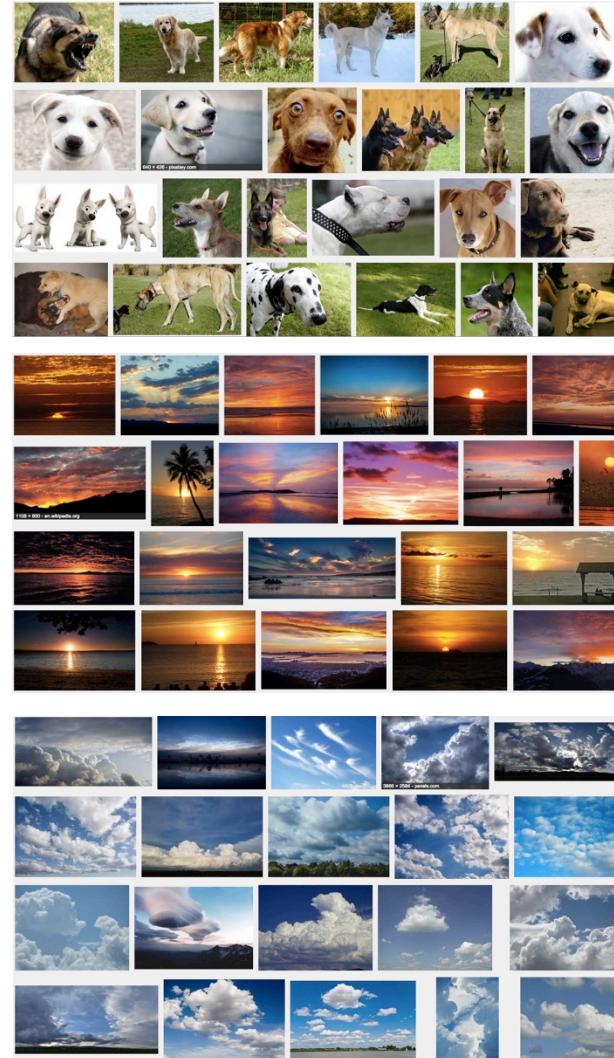
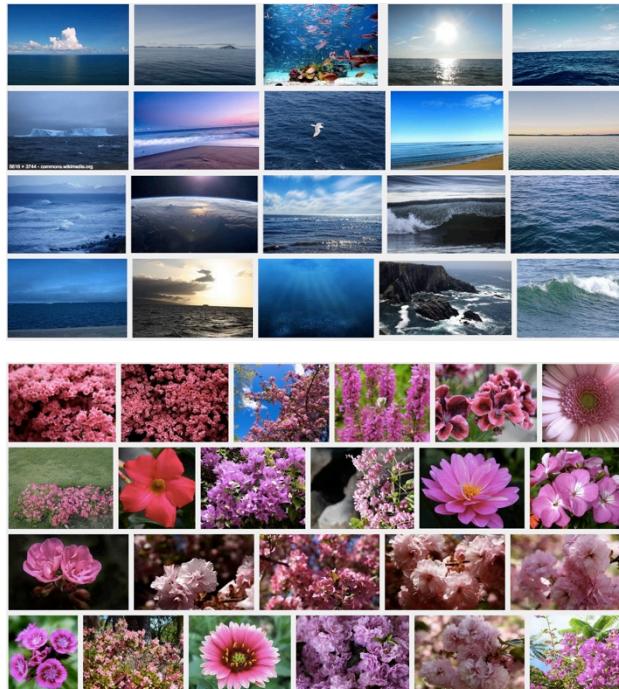
0. Initialize cluster centers
1. Assign observations to closest cluster center
2. Revise cluster centers as mean of assigned observations
3. Repeat 1.+2. until convergence



Other examples

Clustering images

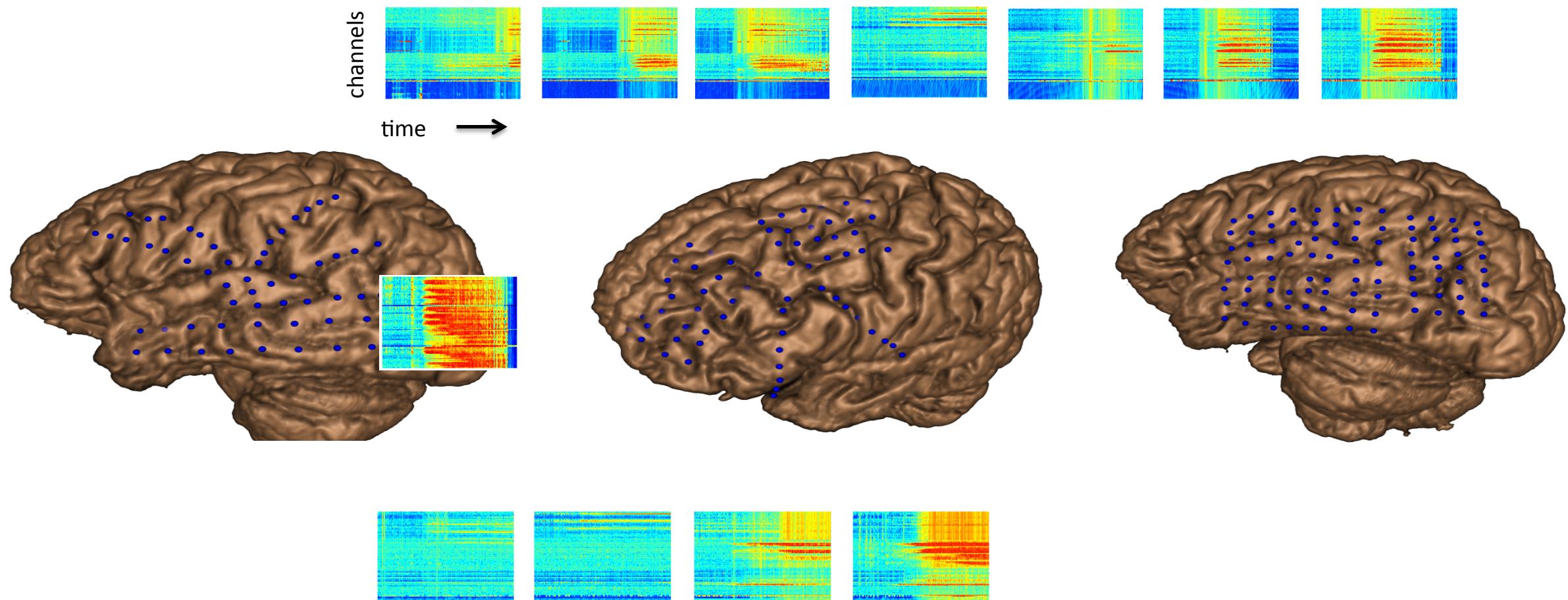
- For search, group as:
 - Ocean
 - Pink flower
 - Dog
 - Sunset
 - Clouds
 - ...



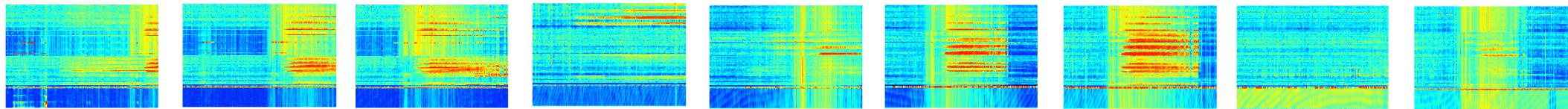
Grouping patients by medical condition

- Better characterize subpopulations and diseases

Example: Patients and seizures are diverse



Cluster seizures by observed time courses



Products on Amazon

- Discover product categories from purchase histories



~~"furniture"~~
"baby"



- Or discovering groups of **users**

Structuring web search results

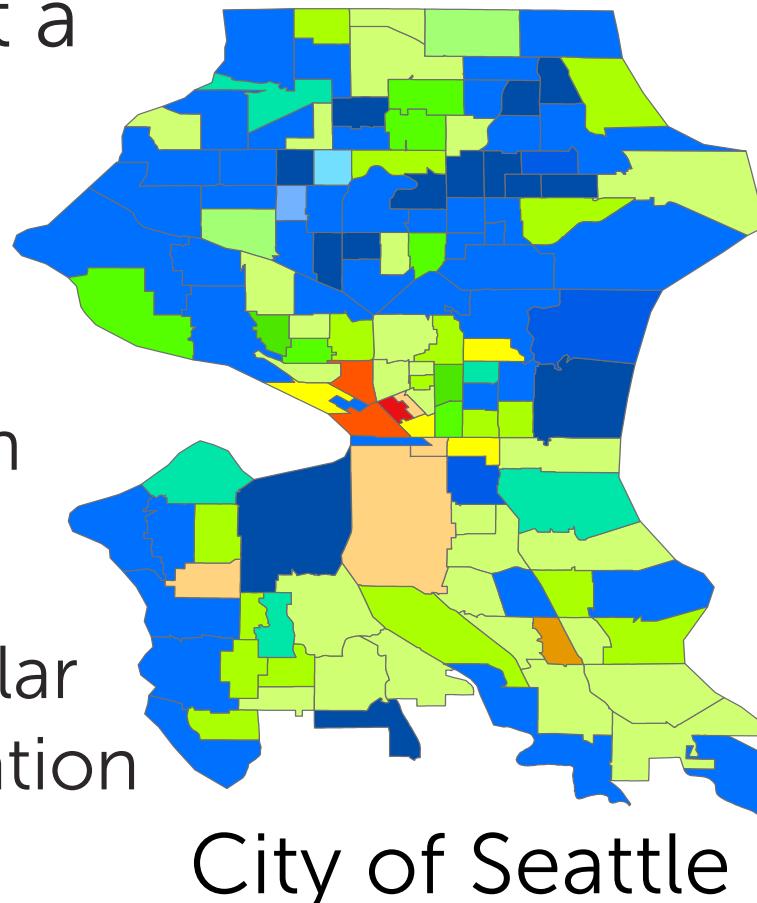
- Search terms can have multiple meanings
- Example: “**cardinal**”



- Use clustering to **structure output**

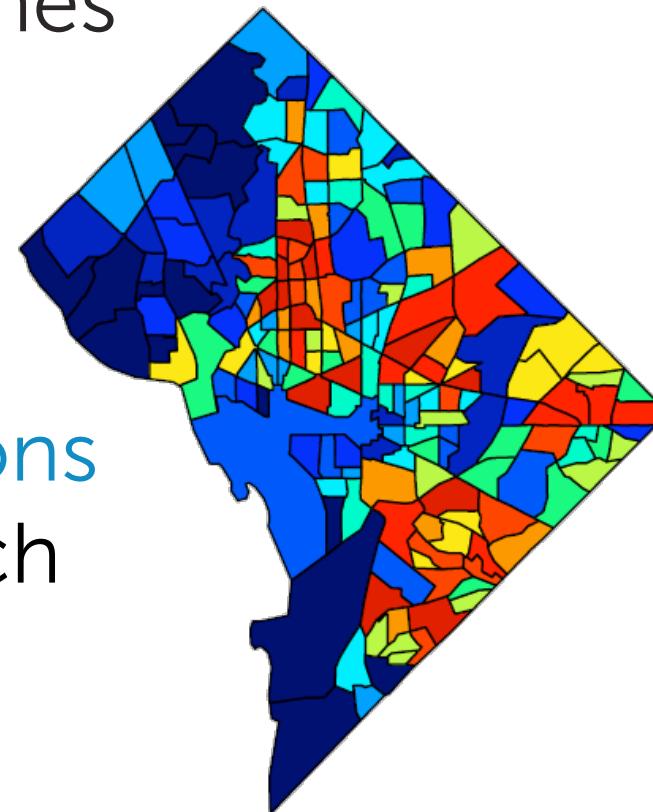
Discovering similar neighborhoods

- **Task 1:** Estimate price at a small regional level
- **Challenge:**
 - Only a few (or no!) sales in each region per month
- **Solution:**
 - Cluster regions with similar trends and share information within a cluster



Discovering similar neighborhoods

- **Task 2:** Forecast violent crimes to better task police
- Again, **cluster regions** and **share information!**
- Leads to **improved predictions** compared to examining each region independently



Washington, DC

Summary for clustering and similarity

What you can do now...

- Describe ways to represent a document (e.g., raw word counts, tf-idf,...)
- Measure the similarity between two documents
- Discuss issues related to using raw word counts
 - Normalize counts to adjust for document length
 - Emphasize important words using tf-idf
- Implement a nearest neighbor search for document retrieval
- Describe the input (unlabeled observations) and output (labels) of a clustering algorithm
- Determine whether a task is supervised or unsupervised
- Cluster documents using k-means (algorithmic details to come...)
- Describe other applications of clustering