

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

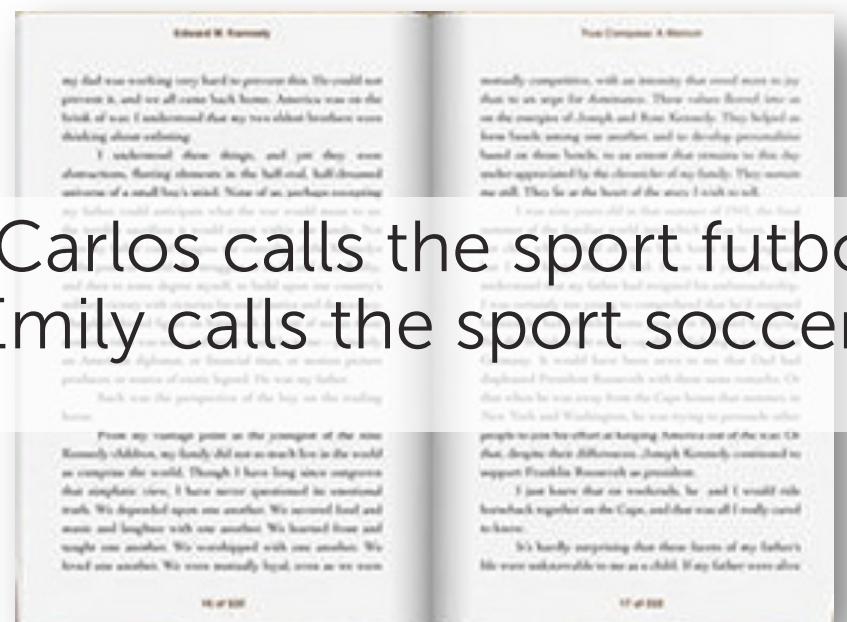
- How do we measure similarity?
- How do we search over articles?



Word count representation for measuring similarity

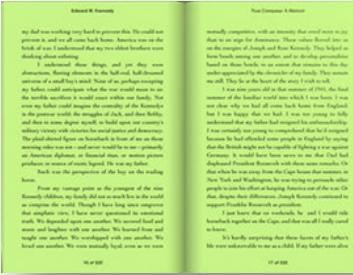
Word count document representation

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



"Carlos calls the sport futbol.
Emily calls the sport soccer."

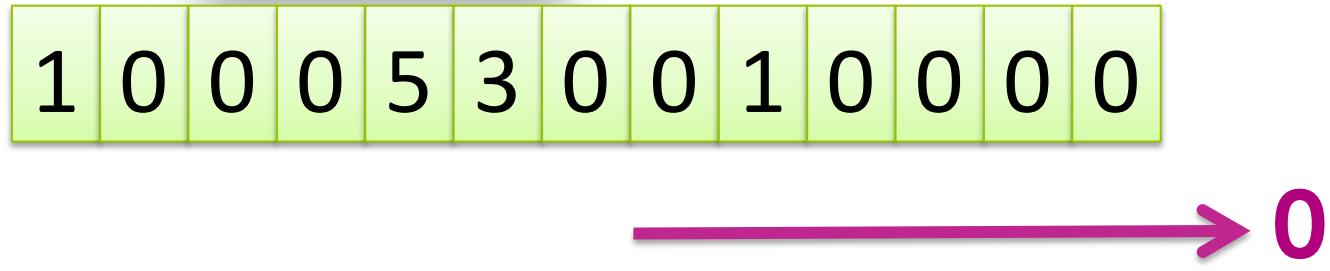
Measuring similarity



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



Measuring similarity



Issues with word counts – Doc length

1 0 0 0 5 3 0 0 1 0 0 0

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

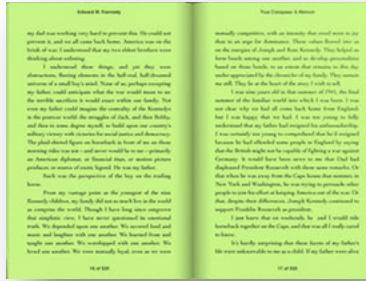
Edward W. Kennedy	True Confession A Memoir	Edward W. Kennedy
my dad was working very hard he'd get tired. He would not go out, or we all went home. America was the most important place in the world. I think that my dad was thinking about ending his life.		
When I was 10 years old, I had a summer job working in the salt field. Half-burdened because of a bad cold. Now I can perhaps recognize that my dad's attitude toward work was not unique. The terrible world he would create within our family. But it was the terrible world that plagued us, and that, and that's why we made things happen, to break open country's collective heart.		
The plot about Capo is based on one of all those terrible things that happened to us. It's a story of an American dilemma. It's a broken man, or an American dilemma. It's a broken man, or an American problem, or a sense of legal. Is my dad really sick? Or is he just a bad person? Or is he a bad father?		
From the moment I was born at the end of my mom's family line, my family was out here in the world to compete in the world. Though I have long since come to terms with the fact that I'm not a competitor in the world. I'd appreciate some others. We needed food and water, and we had to compete for both. And we had to compete against one another. We worked with one another. We loved one another. We were mostly landless as we were	mentally competitive, with an intensity that would rise to the sky as we age. That's where I come from. No one can compete with me. And we all come back home. America was the most important place in the world. I think that my dad was thinking about ending his life.	
	dramatically changing. Because in the salt field, half-burdened because of a bad cold. And that's why we made things happen, to break open country's collective heart. They're not bad people. They're not the best of the best. They're the best of the worst.	
	The terrible world he would create within our family. But it was the terrible world that plagued us, and that, and that's why we made things happen, to break open country's collective heart.	
	I think that my dad was out here in the world to compete in the world. Though I have long since come to terms with the fact that I'm not a competitor in the world. I'd appreciate some others. We needed food and water, and we had to compete for both. And we had to compete against one another. We worked with one another. We loved one another. We were mostly landless as we were	mentally competitive, with an intensity that would rise to the sky as we age. That's where I come from. No one can compete with me. And we all come back home. America was the most important place in the world. I think that my dad was thinking about ending his life.
	dramatically changing. Because in the salt field, half-burdened because of a bad cold. And that's why we made things happen, to break open country's collective heart. They're not bad people. They're not the best of the best. They're the best of the worst.	
	The terrible world he would create within our family. But it was the terrible world that plagued us, and that, and that's why we made things happen, to break open country's collective heart.	
	I think that my dad was out here in the world to compete in the world. Though I have long since come to terms with the fact that I'm not a competitor in the world. I'd appreciate some others. We needed food and water, and we had to compete for both. And we had to compete against one another. We worked with one another. We loved one another. We were mostly landless as we were	mentally competitive, with an intensity that would rise to the sky as we age. That's where I come from. No one can compete with me. And we all come back home. America was the most important place in the world. I think that my dad was thinking about ending his life.
	dramatically changing. Because in the salt field, half-burdened because of a bad cold. And that's why we made things happen, to break open country's collective heart. They're not bad people. They're not the best of the best. They're the best of the worst.	
	The terrible world he would create within our family. But it was the terrible world that plagued us, and that, and that's why we made things happen, to break open country's collective heart.	
	I think that my dad was out here in the world to compete in the world. Though I have long since come to terms with the fact that I'm not a competitor in the world. I'd appreciate some others. We needed food and water, and we had to compete for both. And we had to compete against one another. We worked with one another. We loved one another. We were mostly landless as we were	mentally competitive, with an intensity that would rise to the sky as we age. That's where I come from. No one can compete with me. And we all come back home. America was the most important place in the world. I think that my dad was thinking about ending his life.

2	0	0	0	10	6	0	0	2	0	0	0	0
---	---	---	---	----	---	---	---	---	---	---	---	---

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

Similarity = 52

Solution = normalize



1	0	0	0	5	3	0	0	1	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---

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

1					5	3		1				
/	0	0	0	/	/	0	0	/	0	0	0	0
6				6	6		6					

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



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

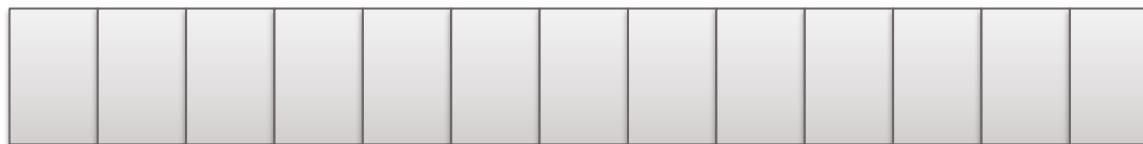


TF-IDF document representation

- 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 rare word

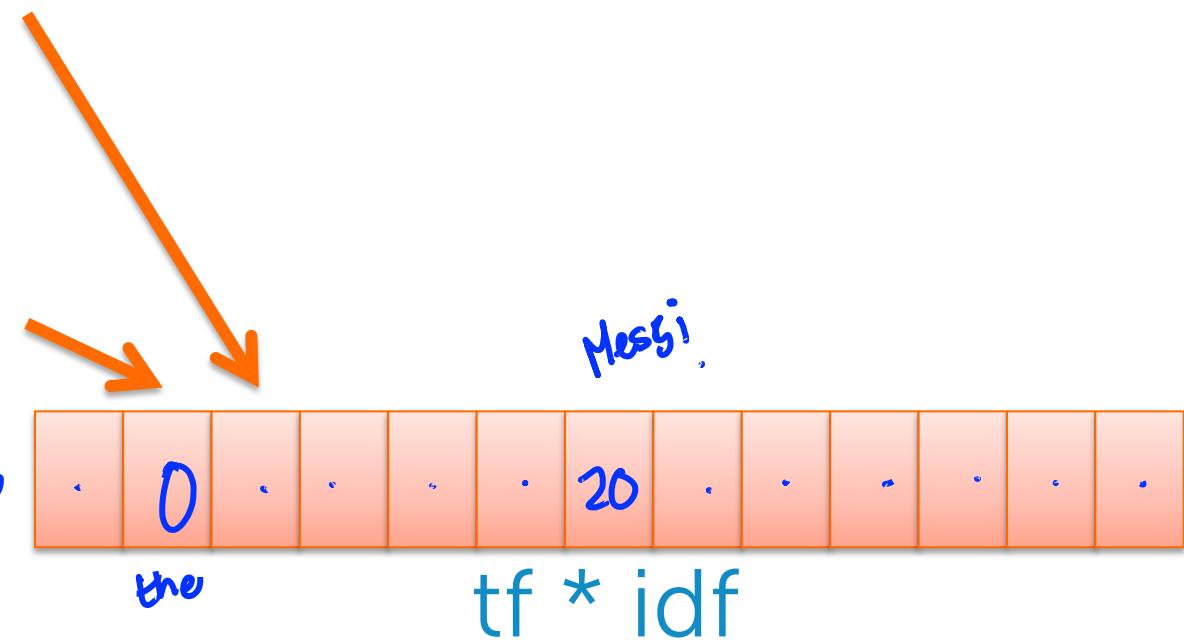
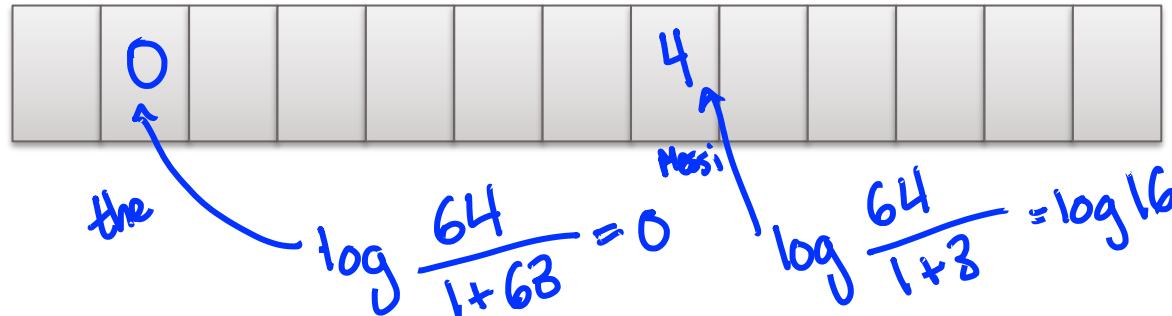
$$\log \frac{\text{large } \#}{1 + \text{large } \#} \approx \log 1 = 0$$
$$\log \frac{\text{large } \#}{1 + \text{small } \#} \rightarrow \text{large } \#$$

TF-IDF document representation

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



- Inverse document frequency



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

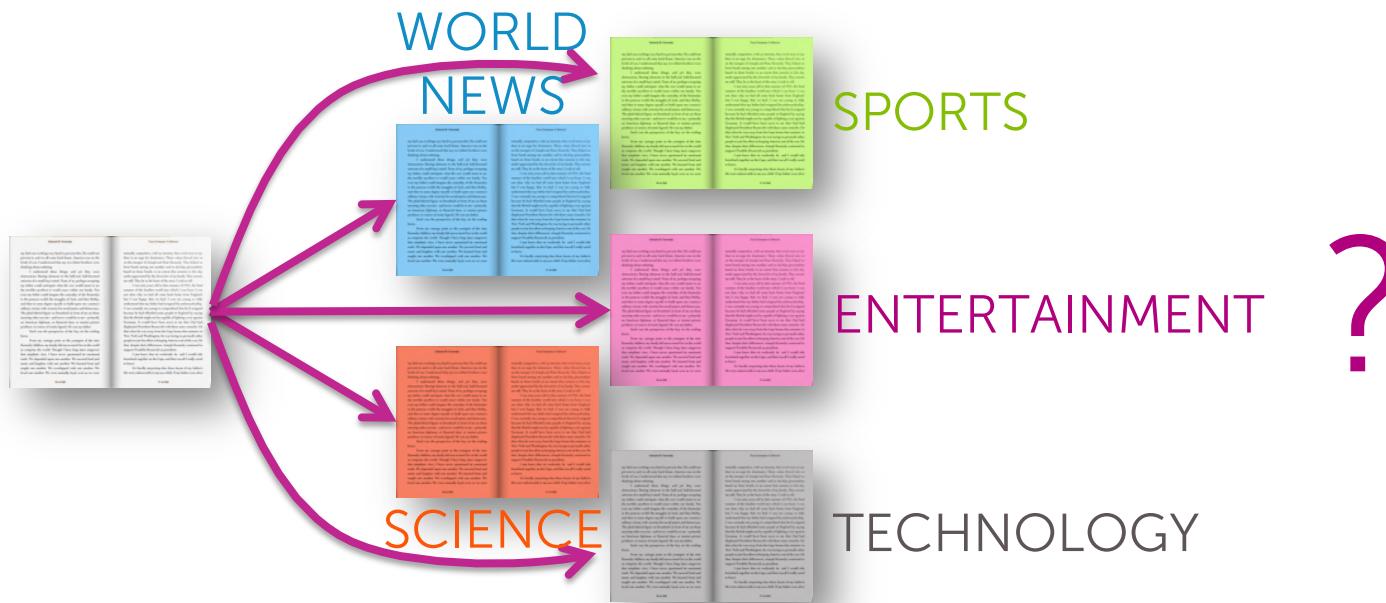


ENTERTAINMENT



SCIENCE

Multiclass classification problem

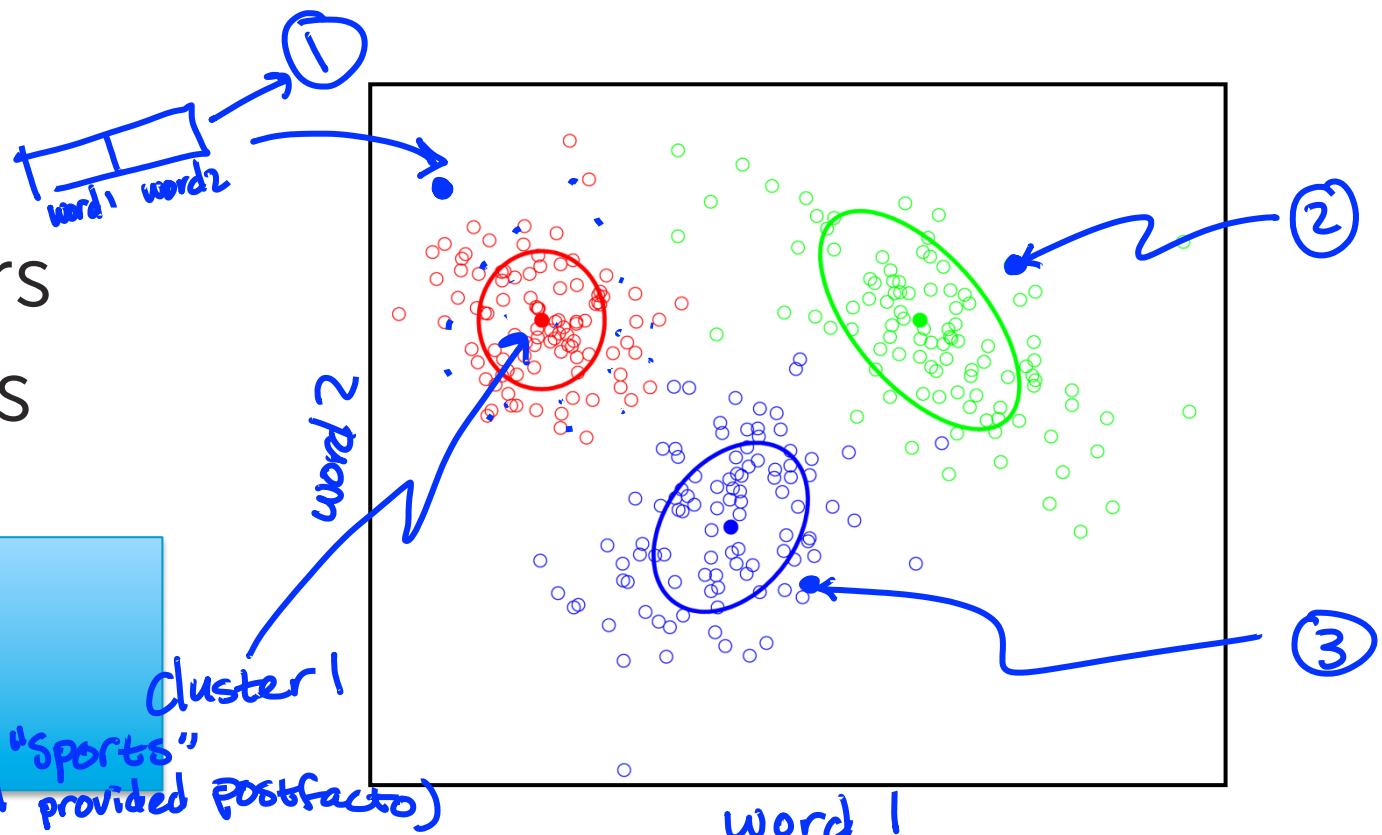


Example of
supervised learning

Clustering

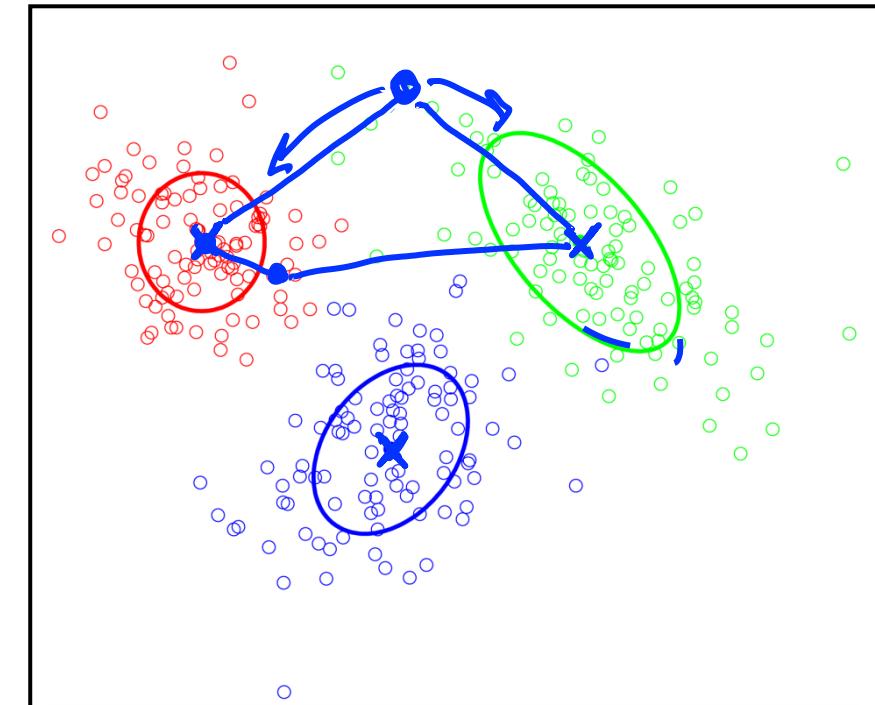
- No labels provided
 - Want to uncover cluster structure
-
- **Input:** docs as vectors
 - **Output:** cluster labels

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



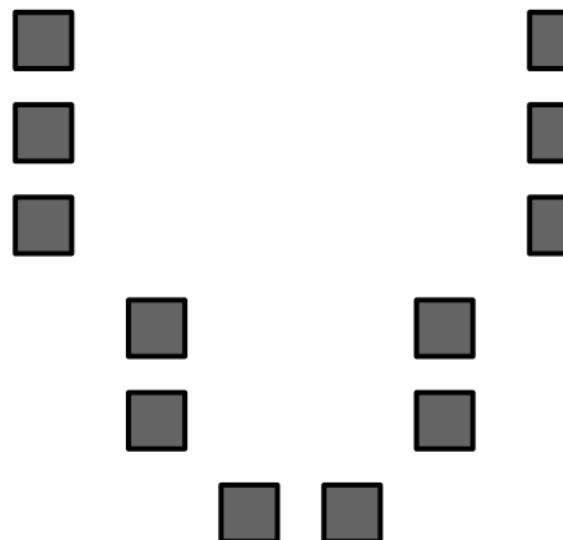
What defines a cluster?

- 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

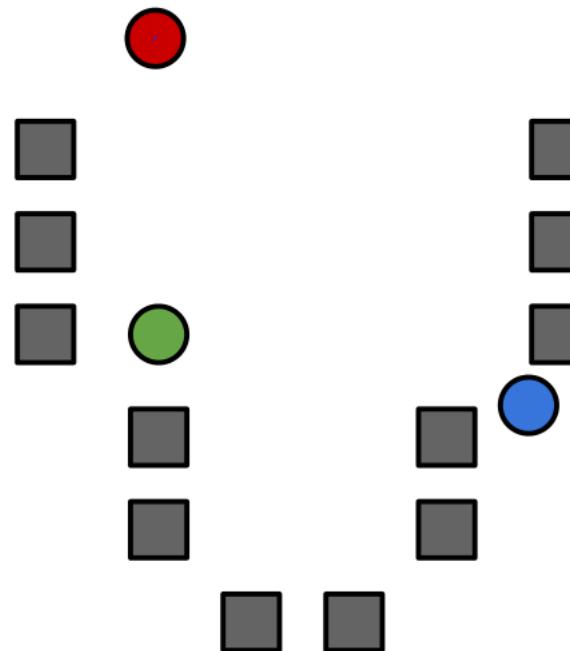
- Assume
 - Similarity metric =
distance to cluster center
(smaller better)



DATA
to
CLUSTER

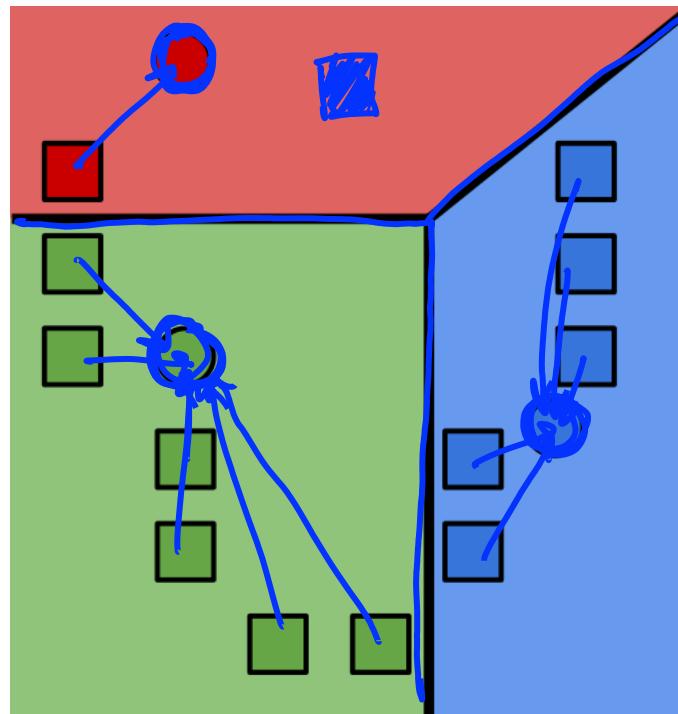
k-means algorithm

0. Initialize cluster centers



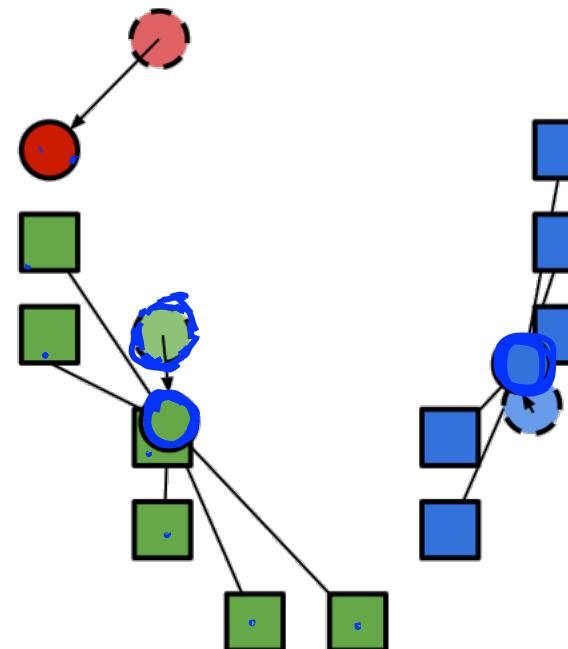
k-means algorithm

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



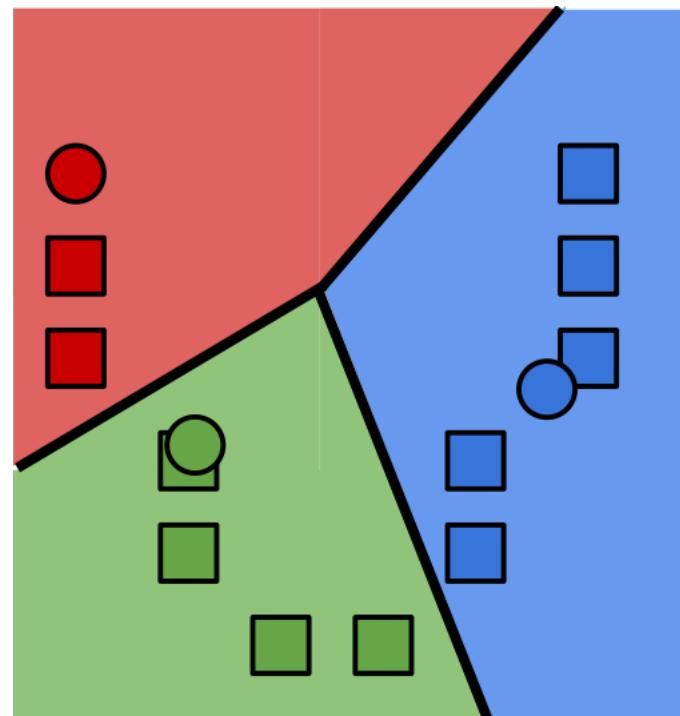
k-means algorithm

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



k-means algorithm

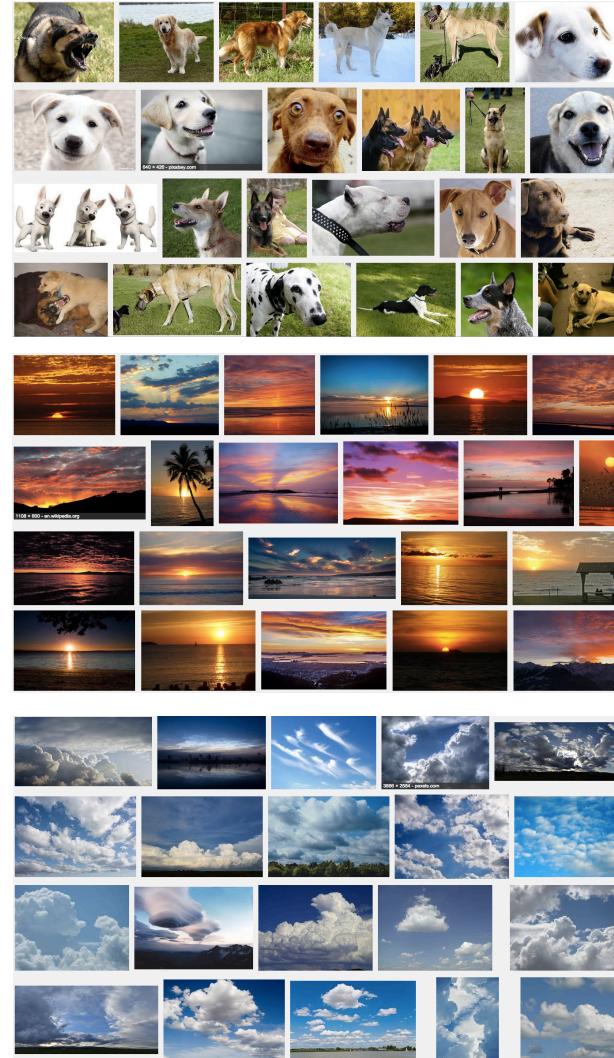
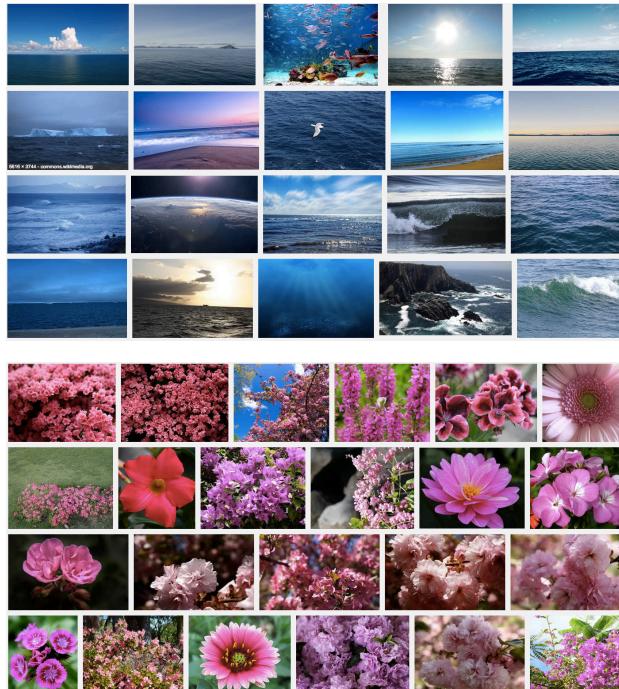
0. Initialize cluster centers
1. Assign observations to closest cluster center
2. Revise cluster centers as mean of assigned observations
3. Repeat 1.+2. until convergence



Other examples

Clustering images

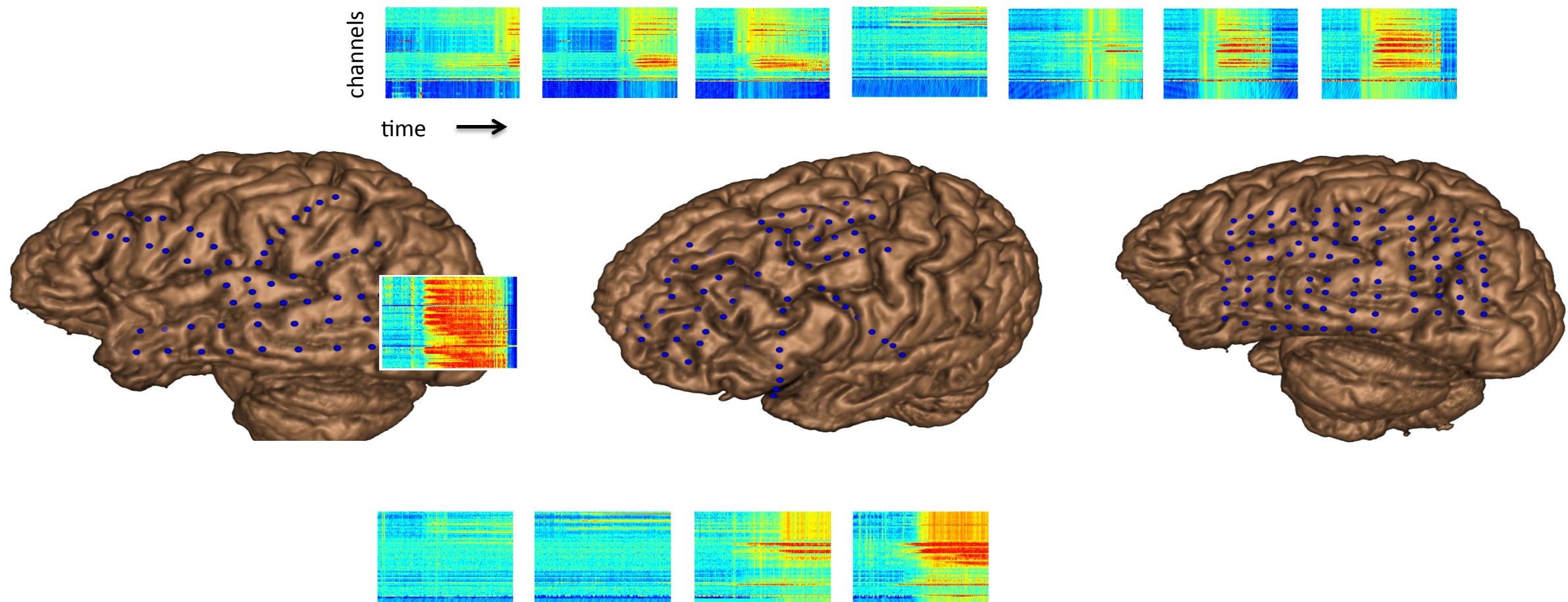
- For search, group as:
 - Ocean
 - Pink flower
 - Dog
 - Sunset
 - Clouds
 - ...



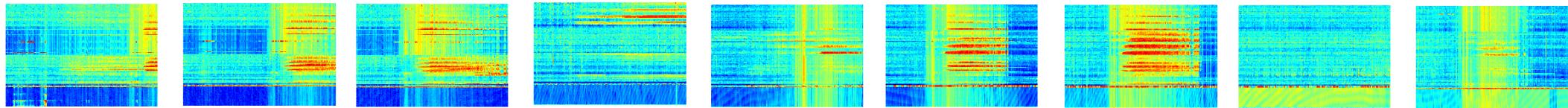
Grouping patients by medical condition

- 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

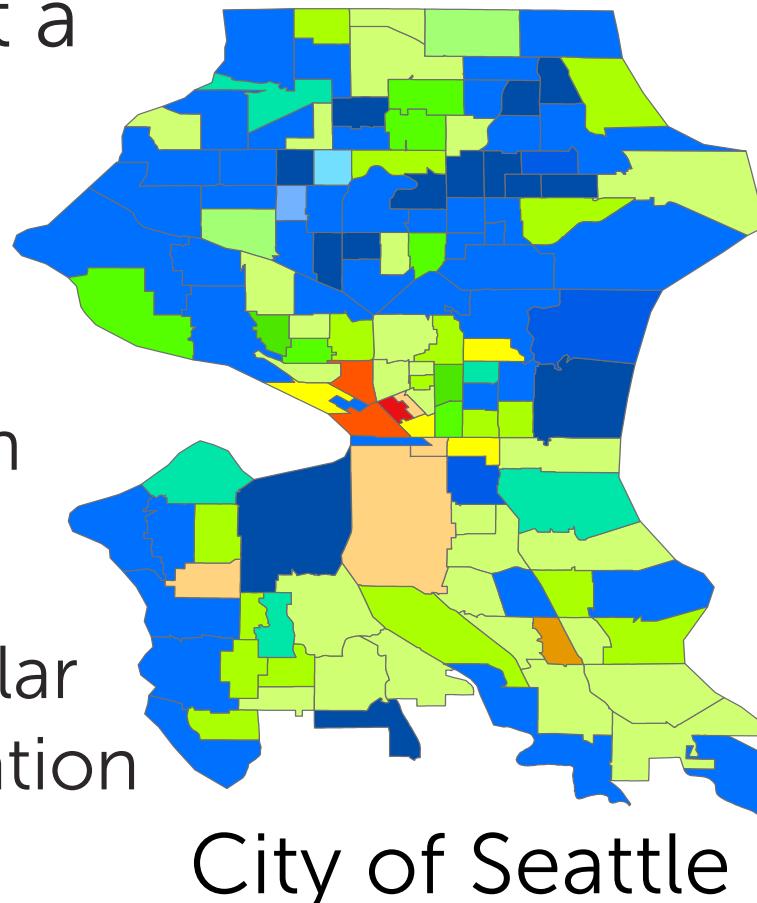
- Search terms can have multiple meanings
- Example: “**cardinal**”



- Use clustering to **structure output**

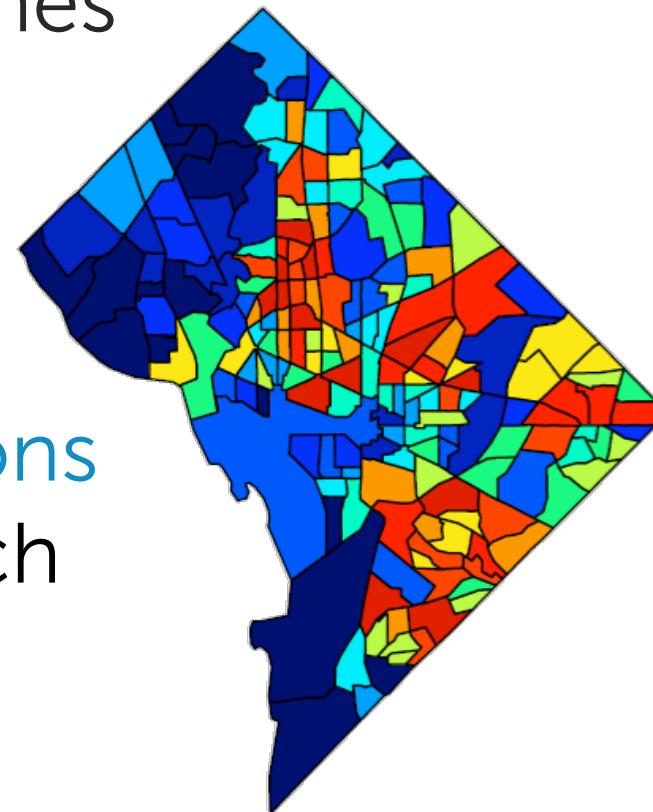
Discovering similar neighborhoods

- **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