

# Clustering & Retrieval: A machine learning perspective

# Nearest Neighbor Search: Retrieving Documents



# 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

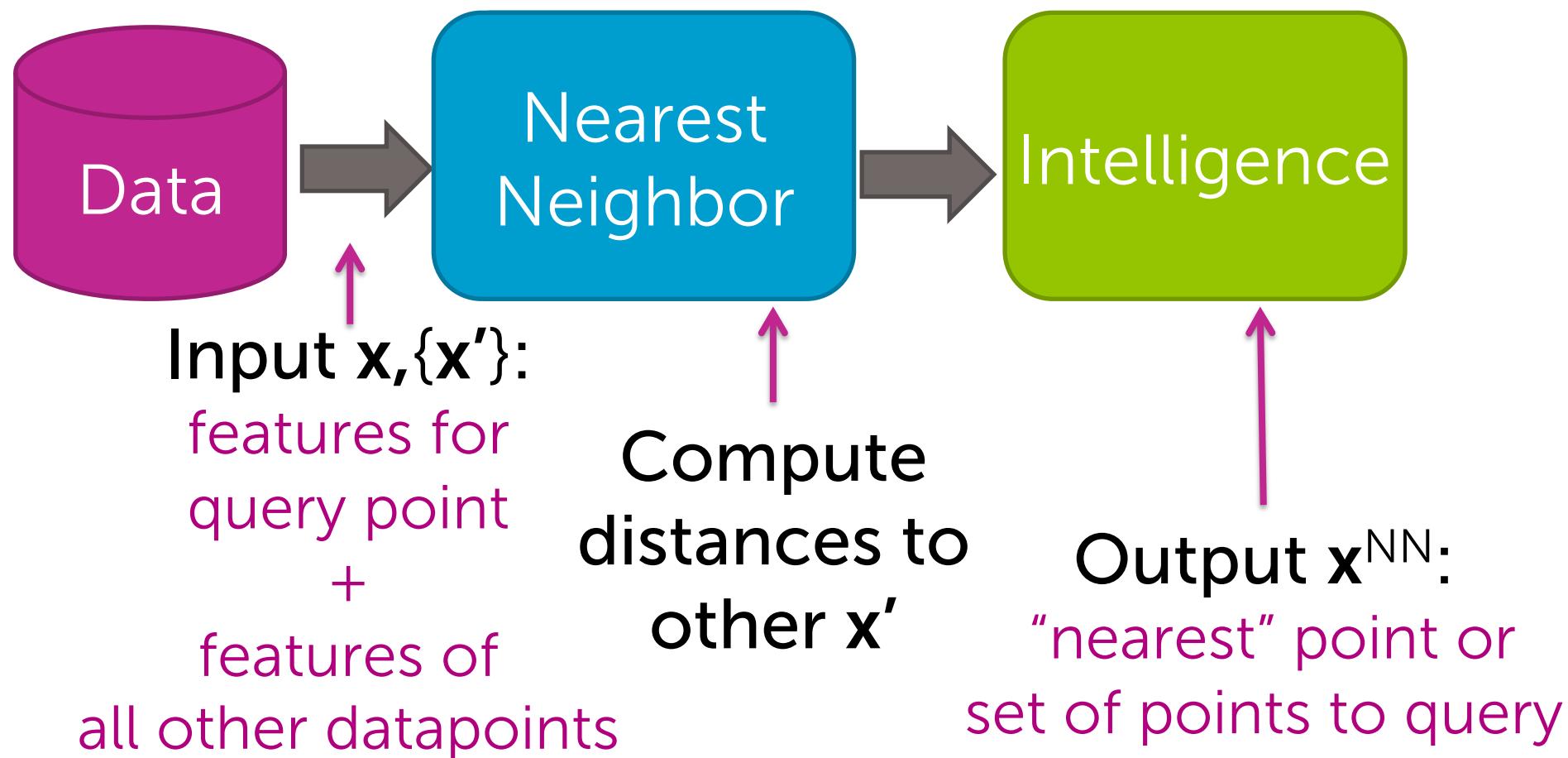
- How do we measure similarity?
  - How do we search over articles?



# Retrieval as k-nearest neighbor search

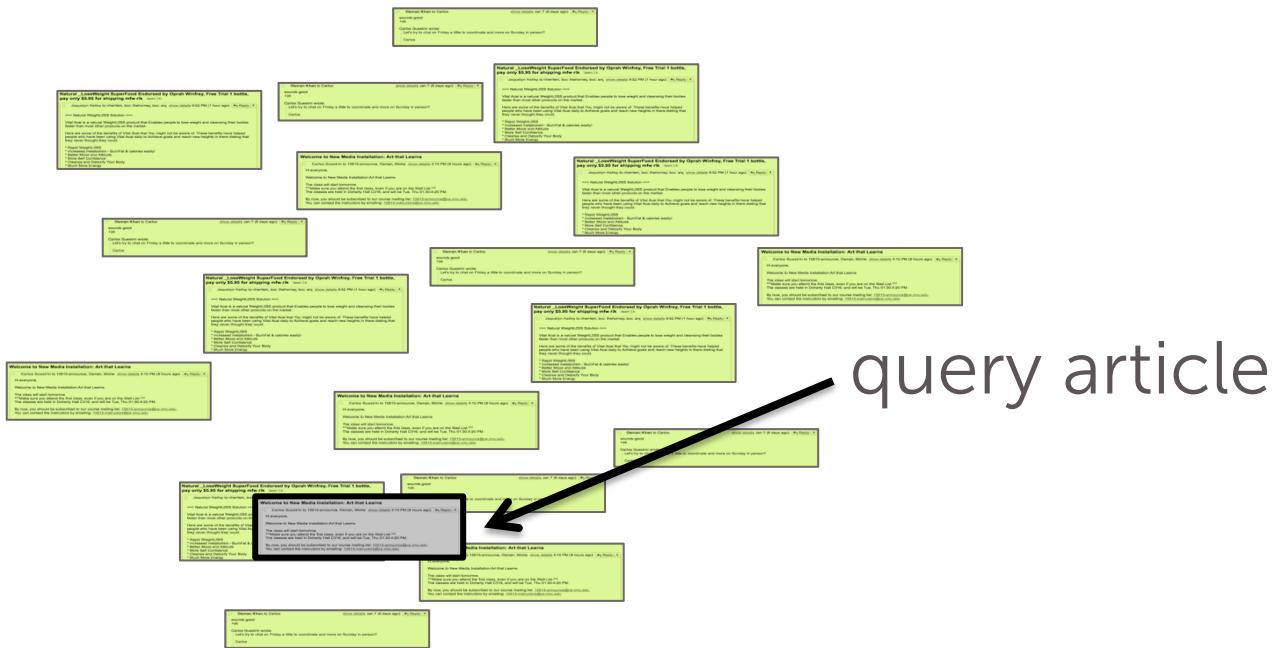
# What is retrieval?

Search for related items



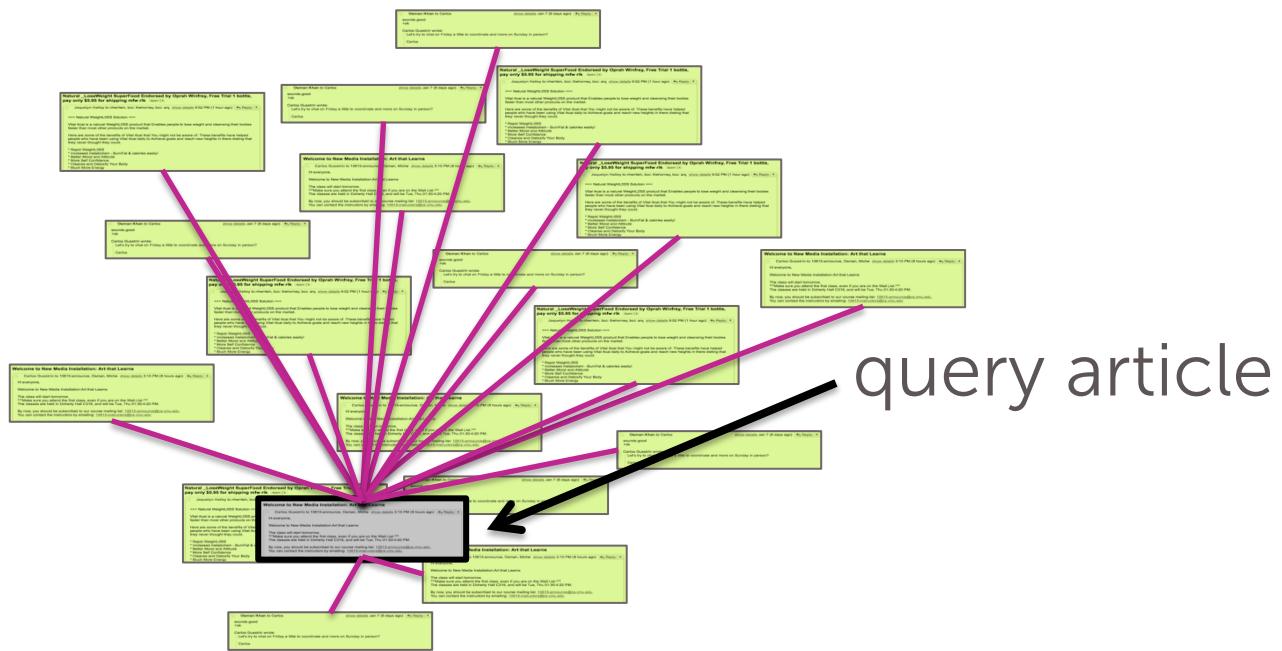
# 1-NN search for retrieval

Space of all articles,  
organized by similarity of text



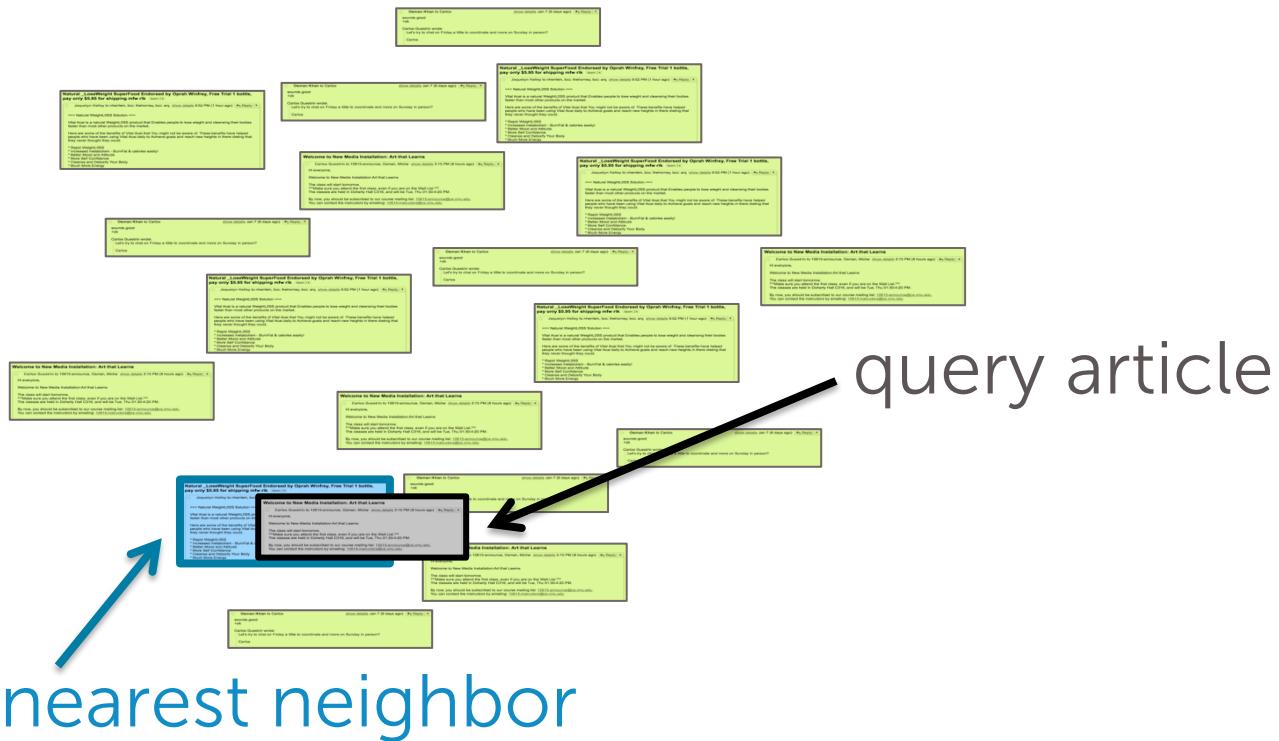
# Compute distances to all docs

Space of all articles,  
organized by similarity of text



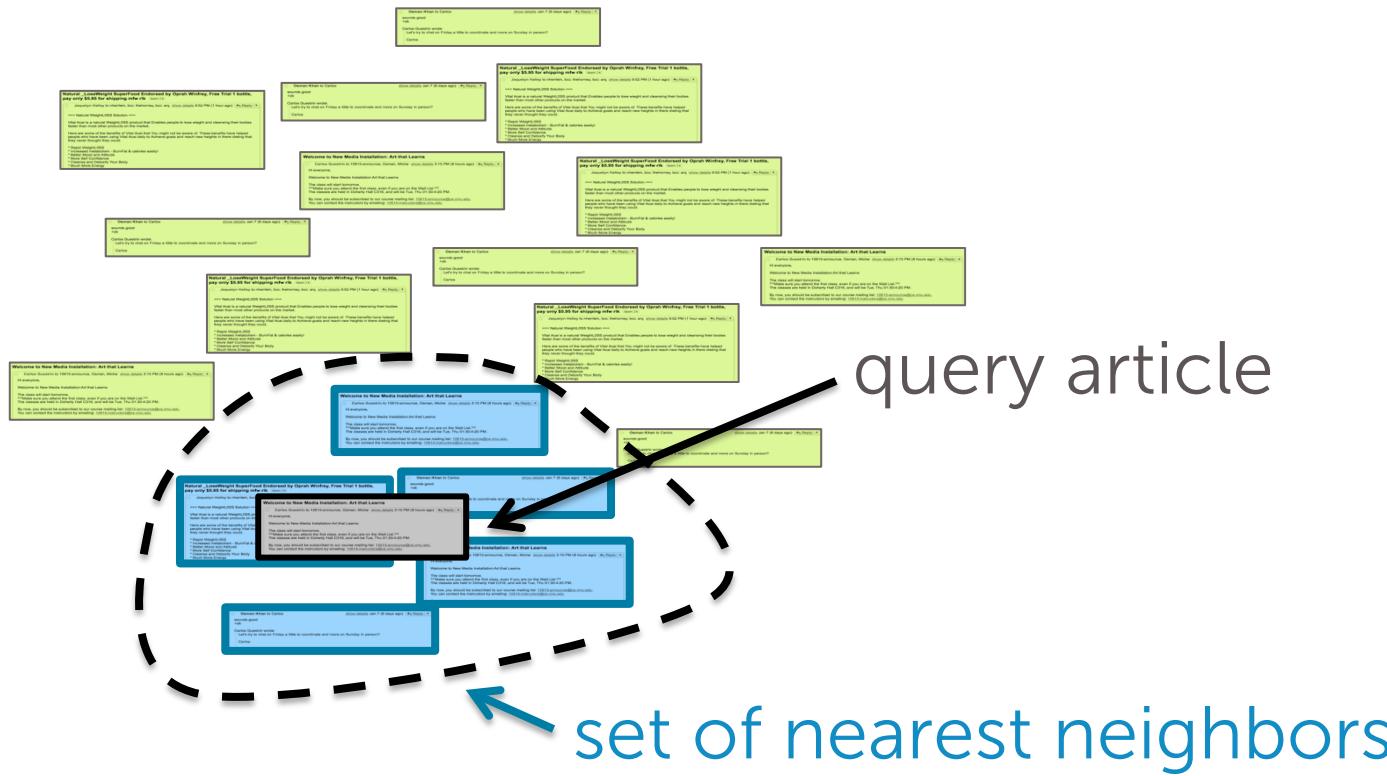
# Retrieve “nearest neighbor”

Space of all articles,  
organized by similarity of text



# Or set of nearest neighbors

Space of all articles,  
organized by similarity of text



# 1-NN algorithm

# 1 – Nearest neighbor

- **Input:** Query article  :  $\underline{\mathbf{x}}_q$   
Corpus of documents   $(N \text{ docs})$
- **Output:** *Most* similar article   $\leftarrow \mathbf{x}^{NN}$

Formally:

$$\mathbf{x}^{NN} = \min_{\mathbf{x}_i} \text{distance}(\mathbf{x}_q, \mathbf{x}_i)$$

# 1-NN algorithm

Initialize  $\text{Dist2NN} = \underline{\infty}$ ,  $= \emptyset$

For  $i=1,2,\dots,N$

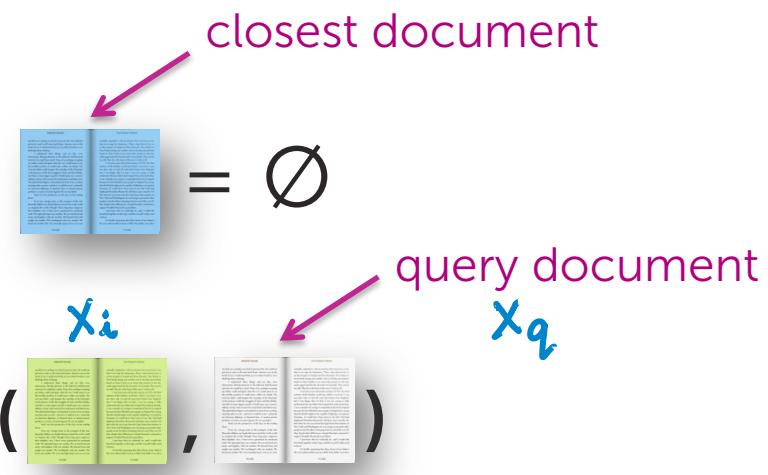
Compute:  $\delta = \text{distance}(x_i, x_q)$

If  $\delta < \text{Dist2NN}$

set  $x_i =$

set  $\text{Dist2NN} = \delta$

Return most similar document



# k-NN algorithm

# k – Nearest neighbor

- **Input:** Query article :  $\mathbf{x}_q$



Corpus of documents



:  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$

- **Output:** *List of k* similar articles



Formally:

$$X^{NN} = \{x^{NN_1}, \dots, x^{NN_k}\}$$

for all  $x_i$  not in  $X^{NN}$ ,  $\text{distance}(x_i, x_q) \geq \max_{x^{NN_j}, j=1..k} \text{distance}(x^{NN_j}, x_q)$

# k-NN algorithm

Initialize  $\text{Dist2kNN} = \text{sort}(\delta_1, \dots, \delta_k)$  ← list of sorted distances  
=   $\dots$    $\delta_1$    $\dots$    $\delta_k$  ← list of sorted docs

For  $i=k+1, \dots, N$

Compute:  $\delta = \text{distance}(\text{book}_i, \text{book}_q)$  ← query doc

If  $\delta < \text{Dist2kNN}[k]$  ← distance to  $k^{\text{th}}$  NN (furthest NN in set)

find  $j$  such that  $\delta > \text{Dist2kNN}[j-1]$  but  $\delta < \text{Dist2kNN}[j]$

remove furthest house and shift queue:

  $[1:k] =$    $-1$

$\text{Dist2kNN}[j+1:k] = \text{Dist2kNN}[j:k-1]$

set  $\text{Dist2kNN}[j] = \delta$  and   $[j] = \text{book}_i$  ← closest k docs to query doc

Return k most similar articles



# Critical elements of NN search

Item (e.g., doc) representation

$$\mathbf{x}_q \leftarrow$$



Measure of **distance** between items:

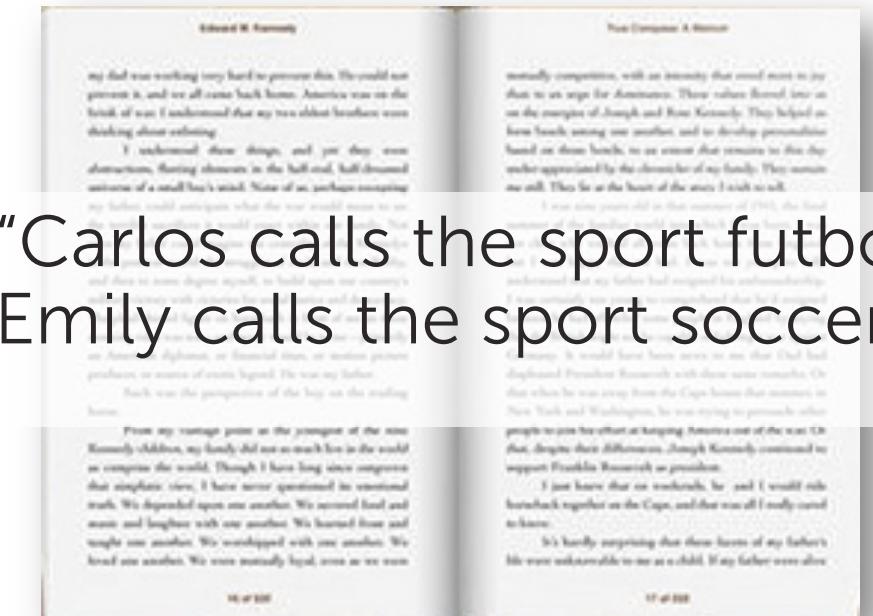
$$\delta = \text{distance}(\mathbf{x}_i, \mathbf{x}_q)$$

# Document representation

# Word count document representation

## Bag of words model

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



# Issues with word counts – Rare words



Common words in doc: "the", "player", "field", "goal"

**Dominante rare words** like: "futbol", "Messi"

# TF-IDF document representation

Emphasizes **important words**

- Appears frequently in document (**common locally**)
- Appears rarely in corpus (**rare globally**)

# TF-IDF document representation

Emphasizes **important words**

- Appears frequently in document (**common locally**)

Term frequency =  word counts

- Appears rarely in corpus (**rare globally**)



# TF-IDF document representation

Emphasizes **important words**

- Appears frequently in document (**common locally**)

Term frequency =  word counts

- Appears rarely in corpus (**rare globally**)

Inverse doc freq. =  $\log \frac{\# \text{ docs}}{1 + \# \text{ docs using word}}$



# TF-IDF document representation

Emphasizes **important words**

- Appears frequently in document (**common locally**)

Term frequency =  word counts

- Appears rarely in corpus (**rare globally**)

Inverse doc freq. =  $\log \frac{\# \text{ docs}}{1 + \# \text{ docs using word}}$

Trade off: **local frequency vs. global rarity**



**tf \* idf**

# Distance metrics

# Distance metrics: Defining notion of “closest”

In 1D, just Euclidean distance:

$$\text{distance}(x_i, x_q) = |x_i - x_q|$$

In multiple dimensions:

- can define many interesting distance functions
- most straightforwardly, might want to weight different dimensions differently

# Weighting different features

Reasons:

- Some features are more relevant than others



**# bedrooms**  
**# bathrooms**  
**sq.ft. living**  
sq.ft. lot  
floors  
**year built**  
year renovated  
**waterfront**



# Weighting different features

## Reasons:

- Some features are more relevant than others



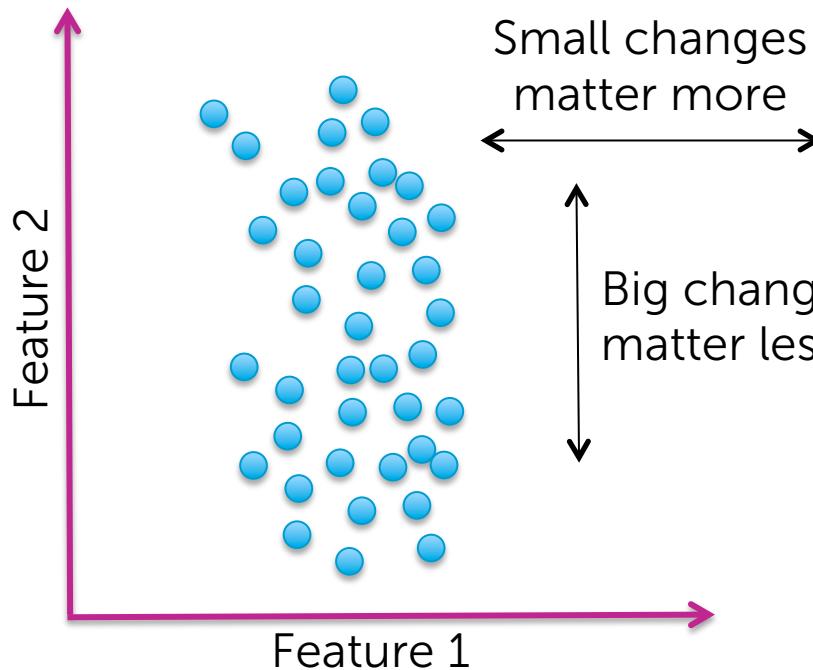
**title**  
**abstract**  
main body  
**conclusion**



# Weighting different features

Reasons:

- Some features are more relevant than others
- Some features vary more than others



Small changes  
matter more

Big changes  
matter less

Specify weights  
as a function of  
feature spread

For feature j:

$$\frac{1}{\max_i(\mathbf{x}_i[j]) - \min_i(\mathbf{x}_i[j])}$$

# Scaled Euclidean distance

Formally, this is achieved via

$$\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{a_1(\mathbf{x}_i[1]-\mathbf{x}_q[1])^2 + \dots + a_d(\mathbf{x}_i[d]-\mathbf{x}_q[d])^2}$$

weight on each feature  
(defining relative importance)

# Effect of binary weights

$$\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{a_1(\mathbf{x}_i[1]-\mathbf{x}_q[1])^2 + \dots + a_d(\mathbf{x}_i[d]-\mathbf{x}_q[d])^2}$$

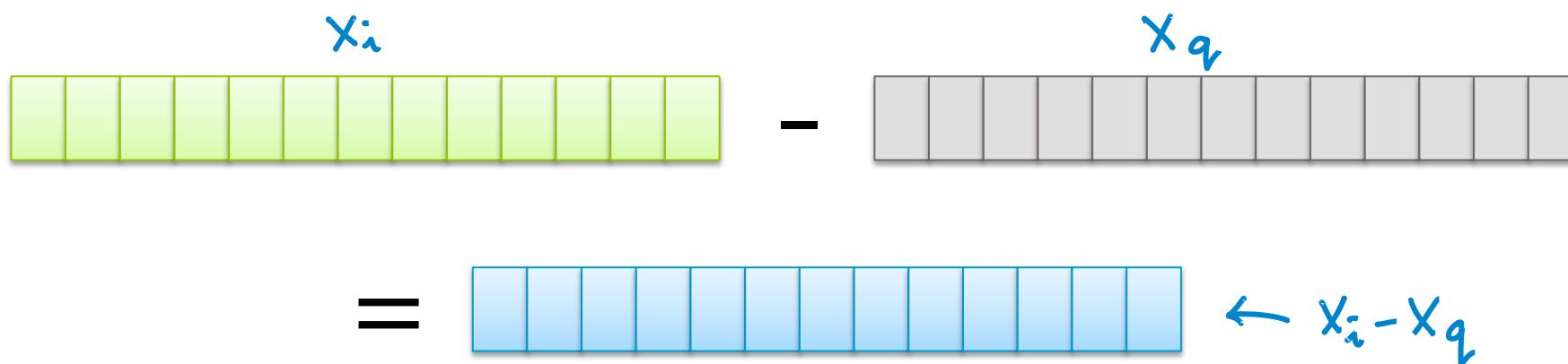
Setting weights as 0 or 1  
is equivalent to  
**feature selection**

Feature engineering/  
selection is  
**important, but hard**

# (non-scaled) Euclidean distance

Defined in terms of inner product

$$\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^T (\mathbf{x}_i - \mathbf{x}_q)}$$
$$\sqrt{(\mathbf{x}_i[1] - \mathbf{x}_q[1])^2 + \dots + (\mathbf{x}_i[d] - \mathbf{x}_q[d])^2}$$



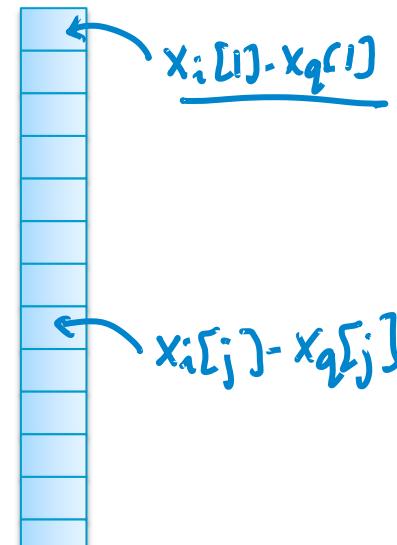
# (non-scaled) Euclidean distance

Defined in terms of inner product

$$\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^T (\mathbf{x}_i - \mathbf{x}_q)}$$

$$= \sqrt{(x_i[1] - x_q[1])^2 + \dots + (x_i[d] - x_q[d])^2}$$

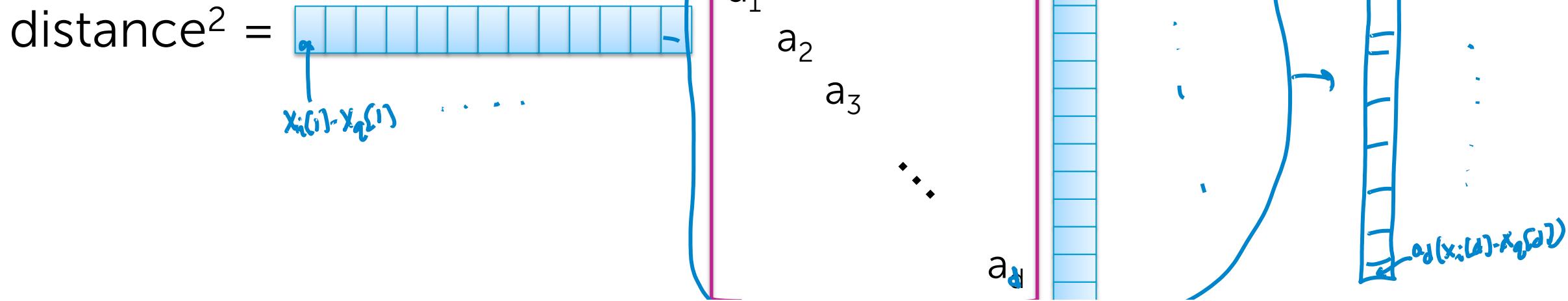
$$\text{distance}^2 = \underbrace{\begin{array}{ccccccccc} \text{---} & \text{---} \\ | & | & | & | & | & | & | & | & | \\ x_i[1] - x_q[1] & & x_i[2] - x_q[2] & & x_i[3] - x_q[3] & & x_i[4] - x_q[4] & & x_i[5] - x_q[5] \end{array}}$$



# Scaled Euclidean distance

Defined in terms of inner product

$$\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^T \mathbf{A} (\mathbf{x}_i - \mathbf{x}_q)}$$
$$= \sqrt{a_1(x_i[1] - x_q[1])^2 + \dots + a_d(x_i[d] - x_q[d])^2}$$



# Another natural inner product measure


 $x_q$ 

 $x_i$ 


## Similarity

$$= \mathbf{x}_i^T \mathbf{x}_q$$

$$= \sum_{j=1}^d \mathbf{x}_i[j] \mathbf{x}_q[j]$$

$$= 13$$

# Another natural inner product measure



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

# Similarity

= 0



# Cosine similarity – normalize

**Similarity** =

$$\frac{\sum_{j=1}^d \mathbf{x}_i[j] \mathbf{x}_q[j]}{\sqrt{\sum_{j=1}^d (\mathbf{x}_i[j])^2} \sqrt{\sum_{j=1}^d (\mathbf{x}_q[j])^2}}$$

$$\mathbf{x}_i^\top \mathbf{x}_q = \cos(\theta)$$

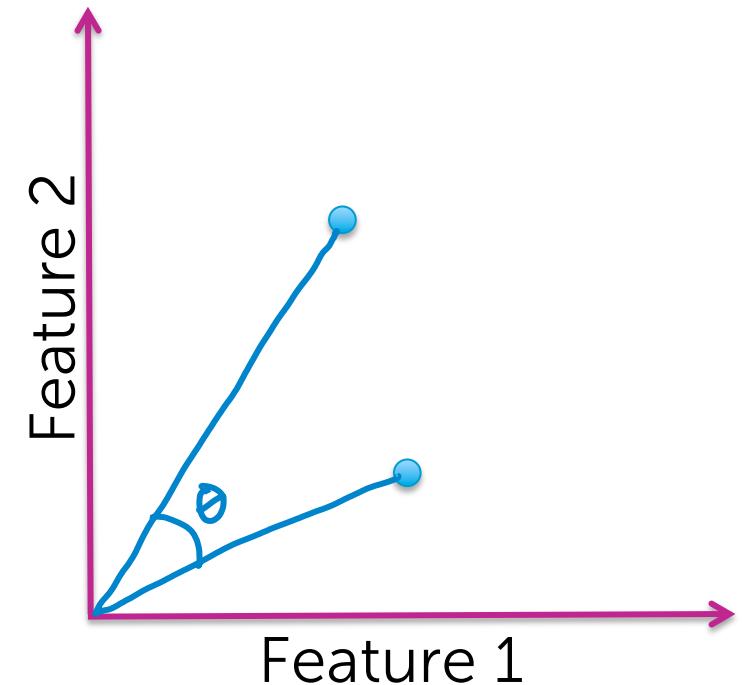
$$\|\mathbf{x}_i\| \|\mathbf{x}_q\|$$

$$= \left( \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|} \right)^\top \left( \frac{\mathbf{x}_q}{\|\mathbf{x}_q\|} \right)$$

first normalize

- Not a proper distance metric
- Efficient to compute for sparse vecs

$$\mathbf{a}^\top \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos(\theta)$$



# Normalize



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

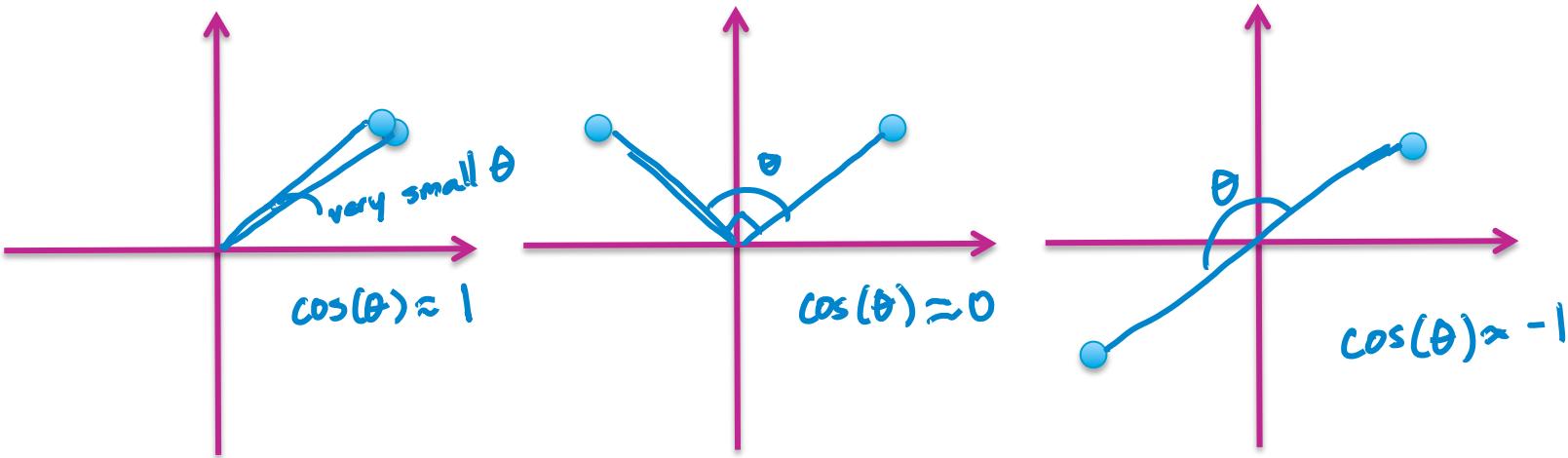
$\leftarrow x_i$

---


$$\sqrt{(1^2 + 5^2 + 3^2 + 1^2)} \leftarrow \|x_i\| = \sum_{j=1}^d x_i[j]^2$$

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

# Cosine similarity



In general,  $-1 < \text{similarity} < 1$

For positive features (like tf-idf)  
 $-1 < \text{similarity} < 1$

Define **distance** = **1-similarity**

# To normalize or not?

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

3	1	0	0	2	0	0	1	0	1	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---

Similarity = 13

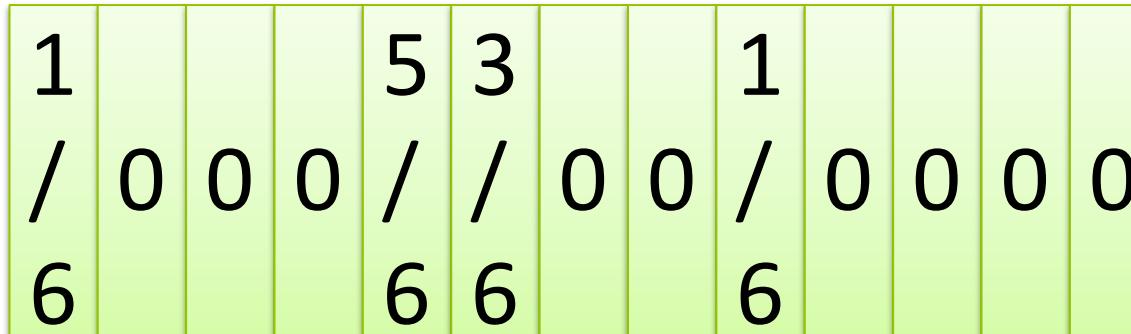
Edward W. Kennedy	You European A Mother	Edward W. Kennedy
my dad was working very hard to prove this. He could not prove it, and so he had to leave. America was on the losing side. I am not sure if my dad was right or wrong about thinking about colonizing.	mentally competing with an American. But could we not say in my case it is the step that makes the difference? I have lived in an American family all my life. I have been exposed to American ways from birth among my sisters and in developing personality. I have been exposed to American ways through my mother who was also American. She was not really American by the definition of an American. Her name was Sally. She is the heart of the story. In fact, she is the soul of the family.	my dad was working very hard to prove this. He could not prove it, and so he had to leave. America was on the losing side. I am not sure if my dad was right or wrong about thinking about colonizing.
It is interesting to note that there were two other Americans, the son of the general who was the strength of Zick, and then Bobbin, and there is some dispute as to whether he held open one country's colonies or another. I am not sure if my dad was right or wrong about this.	the final member of the Kennedy family who was still alive when I last saw him was his son, Ted. Ted was the son of the final member of the Kennedy family who was still alive when I last saw him. But I do not know if he was not. I am not sure if Sally understood that my father had brought his son to America to prove that he was right about the American way of life because he had offered some people in England for saying something that he did not like. I am not sure if my dad understood that he would have been one of the ones that Dad disapproved Franklin Roosevelt with some reason. Or maybe he did not understand that he would have been one of the ones that Dad disapproved Franklin Roosevelt with some reason.	It is interesting to note that there were two other Americans, the son of the general who was the strength of Zick, and then Bobbin, and there is some dispute as to whether he held open one country's colonies or another. I am not sure if my dad was right or wrong about this.
Family history. I am not sure if the son of the one Kennedy family, my dad did not go to America to the world as compared to the world. Though I have long since forgotten that he did not go to America to the world. I am not sure if he did not go to America to the world. We disputed over some another. We disputed and fought and fought and fought and fought and fought and won. We disputed with our another. We lived one another. We were mostly local, even as we were	mentally competing with an American. But could we not say in my case it is the step that makes the difference? I have lived in an American family all my life. I have been exposed to American ways from birth among my sisters and in developing personality. I have been exposed to American ways through my mother who was also American. She was not really American by the definition of an American. Her name was Sally. She is the heart of the story. In fact, she is the soul of the family.	mentally competing, with an interest that even made me think that I wanted to go to America to the world. I am not sure if my dad was right or wrong about thinking about colonizing.
It is interesting to note that there were two other Americans, the son of the general who was the strength of Zick, and then Bobbin, and there is some dispute as to whether he held open one country's colonies or another. I am not sure if my dad was right or wrong about this.	the final member of the Kennedy family who was still alive when I last saw him was his son, Ted. Ted was the son of the final member of the Kennedy family who was still alive when I last saw him. But I do not know if he was not. I am not sure if Sally understood that my father had brought his son to America to prove that he was right about the American way of life because he had offered some people in England for saying something that he did not like. I am not sure if my dad understood that he would have been one of the ones that Dad disapproved Franklin Roosevelt with some reason. Or maybe he did not understand that he would have been one of the ones that Dad disapproved Franklin Roosevelt with some reason.	It is interesting to note that there were two other Americans, the son of the general who was the strength of Zick, and then Bobbin, and there is some dispute as to whether he held open one country's colonies or another. I am not sure if my dad was right or wrong about this.
Family history. I am not sure if the son of the one Kennedy family, my dad did not go to America to the world as compared to the world. Though I have long since forgotten that he did not go to America to the world. We disputed over some another. We disputed and fought and fought and fought and fought and fought and won. We disputed with our another. We lived one another. We were mostly local, even as we were	mentally competing, with an interest that even made me think that I wanted to go to America to the world. I am not sure if my dad was right or wrong about thinking about colonizing.	mentally competing, with an interest that even made me think that I wanted to go to America to the world. I am not sure if my dad was right or wrong about thinking about colonizing.

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

# Normalize



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



# In the normalized case

my dad was working long hours to help procure this. He could not get away from his work, so we had to go to him. I think he was at the end of it and I believe that my two older brothers were shaking about thinking about it.

"After all these things, and yet you were sharing, floating in the half-odd, half-dream atmosphere, I think that if you had been asked to describe my belief, could anyone say that we would move on to the world outside? It would result over our family. But I think that the world outside was there, and I think that in the present world the strength of love, and then Hitler's influence, and the way that he was able to bring people together in military victory with the world outside and democracy."

The detailed plot of *Homeland* is based on a real life situation that occurred in 1945, when the United States and Britain sent agents to Germany to interrogate the Nazi leaders. The agents were American diplomats, or American citizens, or anyone present during the final days of the war in Europe.

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

From my vantage point as a visitor of the same Kennedy children, my family did not see us in the world as they did. We were not the ones who had to leave their comfortable home; we have never experienced the emotional pain of separation from our parents. We have not seen the last night of our childhood. We learned from and taught one another. We worked together as one. We lived one dream.

We were amazingly lucky, even as we were morally complicit, with an intensity that must stand in the shadow of the innocence of the young Joseph and Kennedy. They believed that they were doing the right thing. They believed in their basic mission, saving one another, and to develop the world around them. They believed that they were doing the right thing for the greater good, for the development of the world. They never imagined that they would be killed.

It was only after the fall of the Berlin Wall, the final stages of the hundred year old conflict that I saw many of the Kennedy children again. They were not the ones who had to leave their home, but they were the ones who had to leave their country. I was young, so I left my country to go to England for my education. I was young, so I became one person in England by the name of John F. Kennedy. I would have never known that that had happened to me if it were not for the fact that when I was being interviewed for the Cope House memoirs in New York and Washington, I was trying to prove to the interviewer that I was not the John F. Kennedy that she thought she was interviewing. Although Kennedy had died, I was still John F. Kennedy.

I just have to say that the methods, we could only consider the ways to the top, and we did not really consider to climb.

It's hardly surprising that three former of his children were welcomed into a club. A child. Father was also

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

3	1			1			1		1				
/	/	0	0	/	0	0	/	0	/	0	0	0	0
4	4			2			4		4				

Similarity  
= 13/24

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

3	1			1			1		1				
/	/	0	0	/	0	0	/	0	/	0	0	0	0
4	4			2			4		4				

Similarity  
= 13/24

# Other distance metrics

- Mahalanobis
- rank-based
- correlation-based
- Manhattan
- Jaccard
- Hamming
- ...

# Combining distance metrics

**Example of document features:**

1. Text of document
  - Distance metric: Cosine similarity
2. # of reads of doc
  - Distance metric: Euclidean distance

Add together with user-specified weights

# Clustering: Grouping Related Docs



# Motivating clustering approaches

# Goal: Structure documents by topic

Discover groups (*clusters*) of related articles



SPORTS

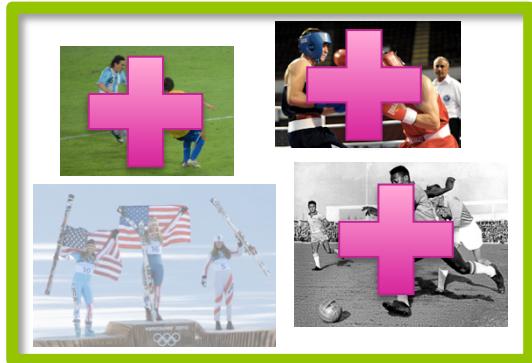
WORLD NEWS

# Why might clustering be useful?



# Learn user preferences

Set of clustered documents read by user



Cluster 1



Cluster 2



Cluster 3



Cluster 4

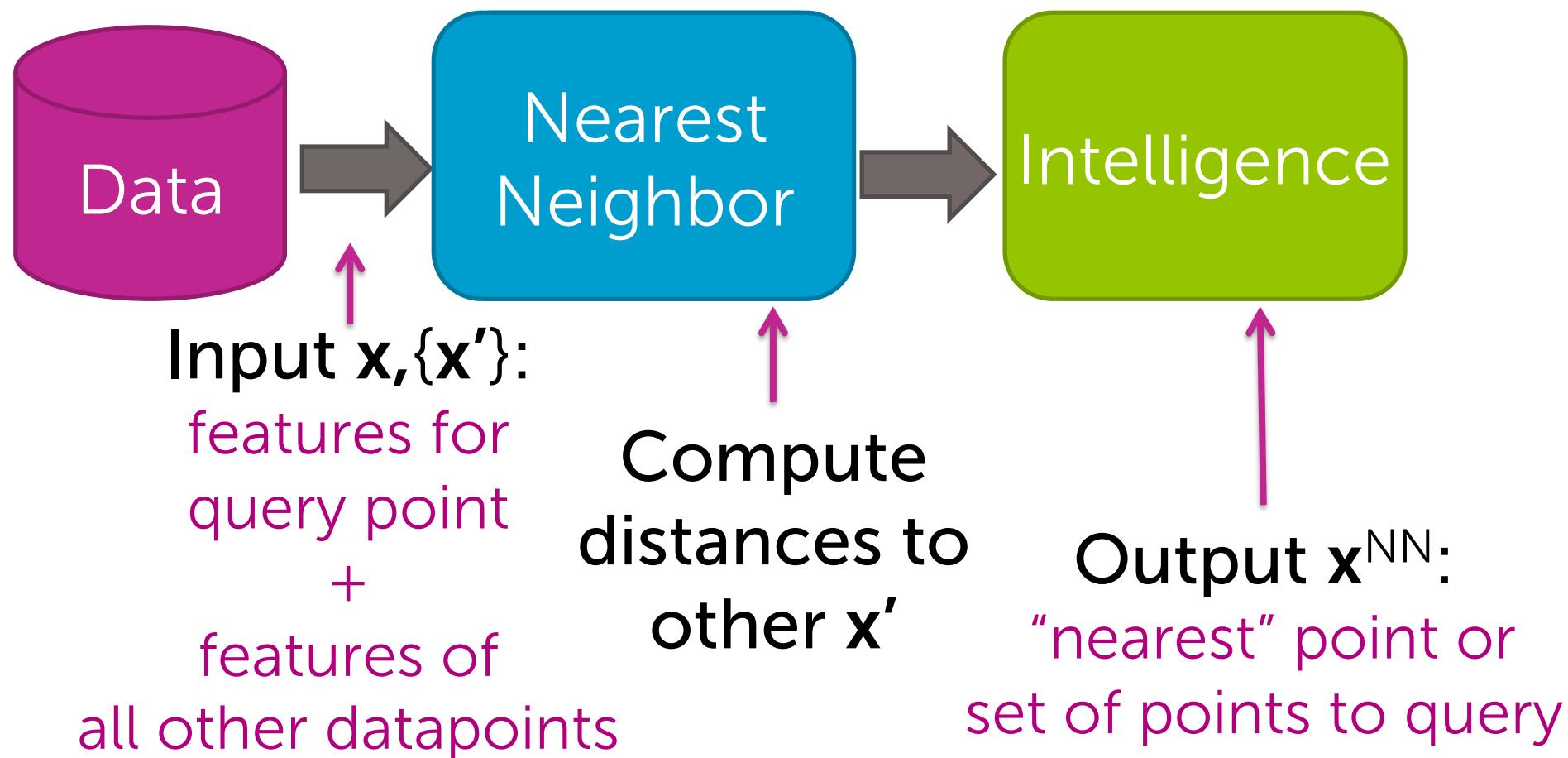


Use feedback  
to learn user  
preferences  
over topics

# Clustering: An unsupervised learning task

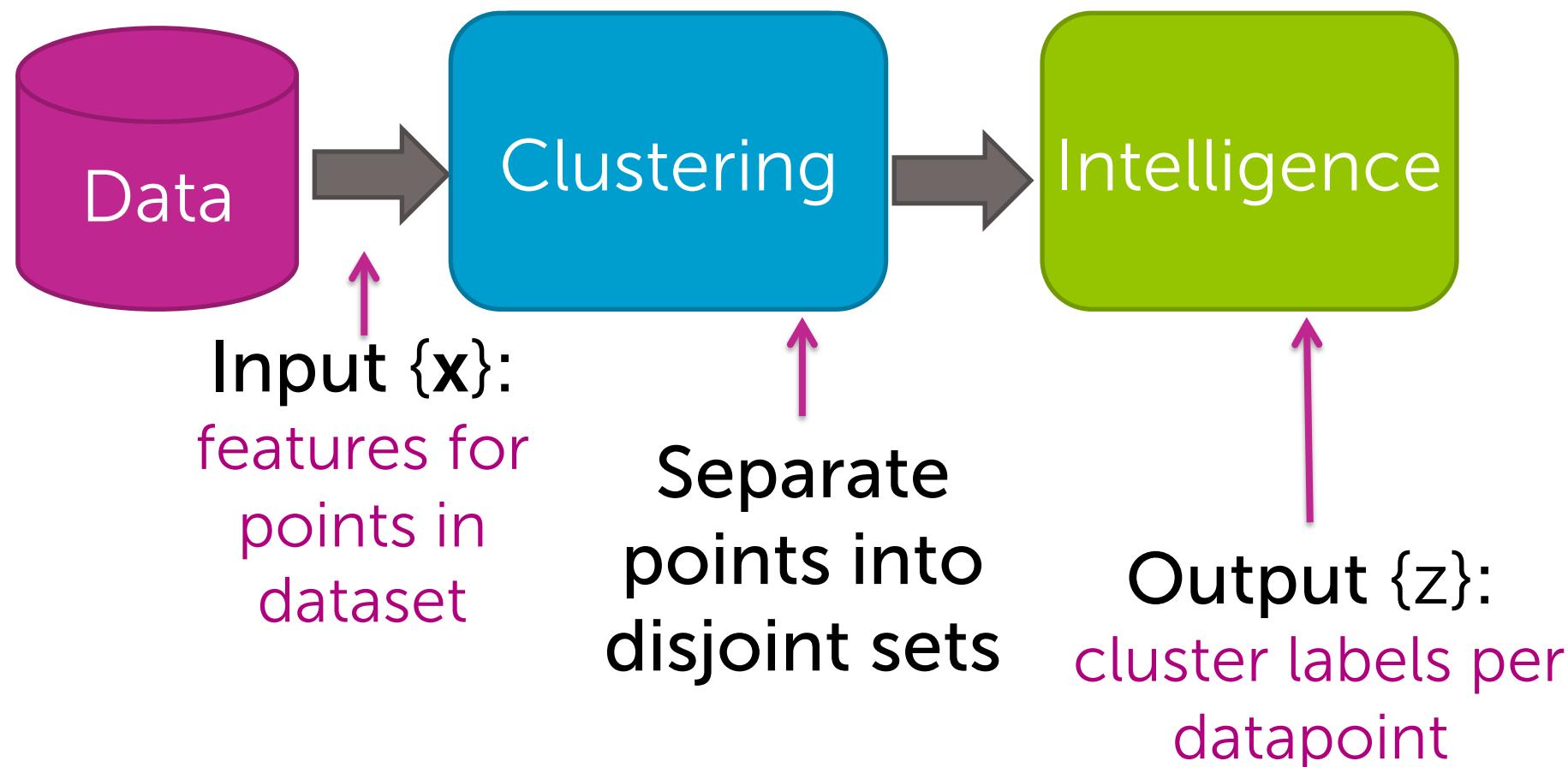
# What is retrieval?

Search for related items



# What is clustering?

Discover groups of similar inputs



# What if some of the labels are known?

Training set of labeled docs



SPORTS



WORLD NEWS

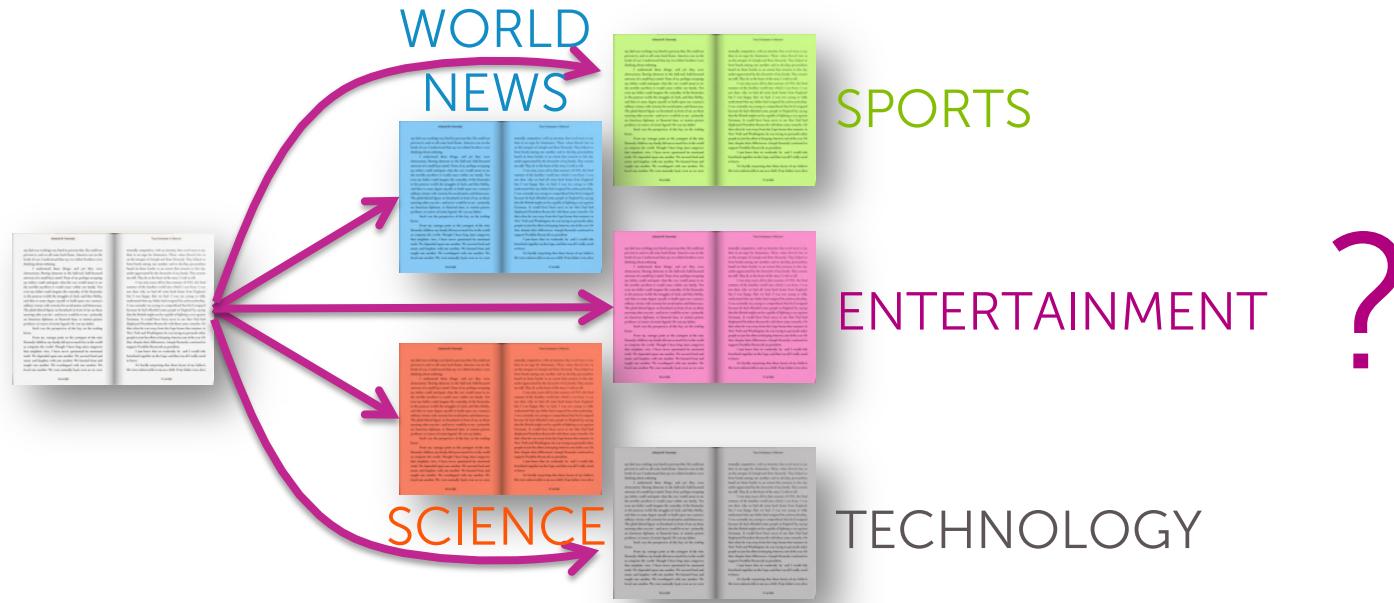


ENTERTAINMENT



SCIENCE

# Multiclass classification problem



Example of  
supervised learning

# Clustering

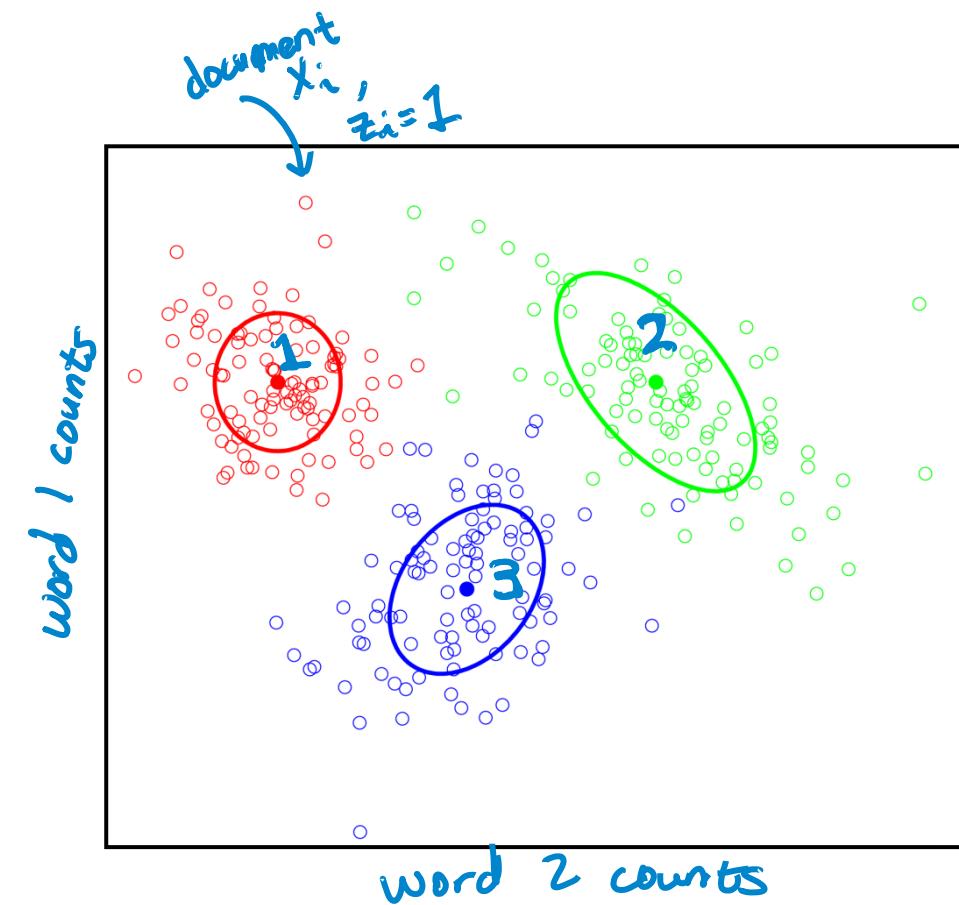
No labels provided

...uncover cluster structure  
from input alone

**Input:** docs as vectors  $\mathbf{x}_i$

**Output:** cluster labels  $z_i$

An unsupervised  
learning task

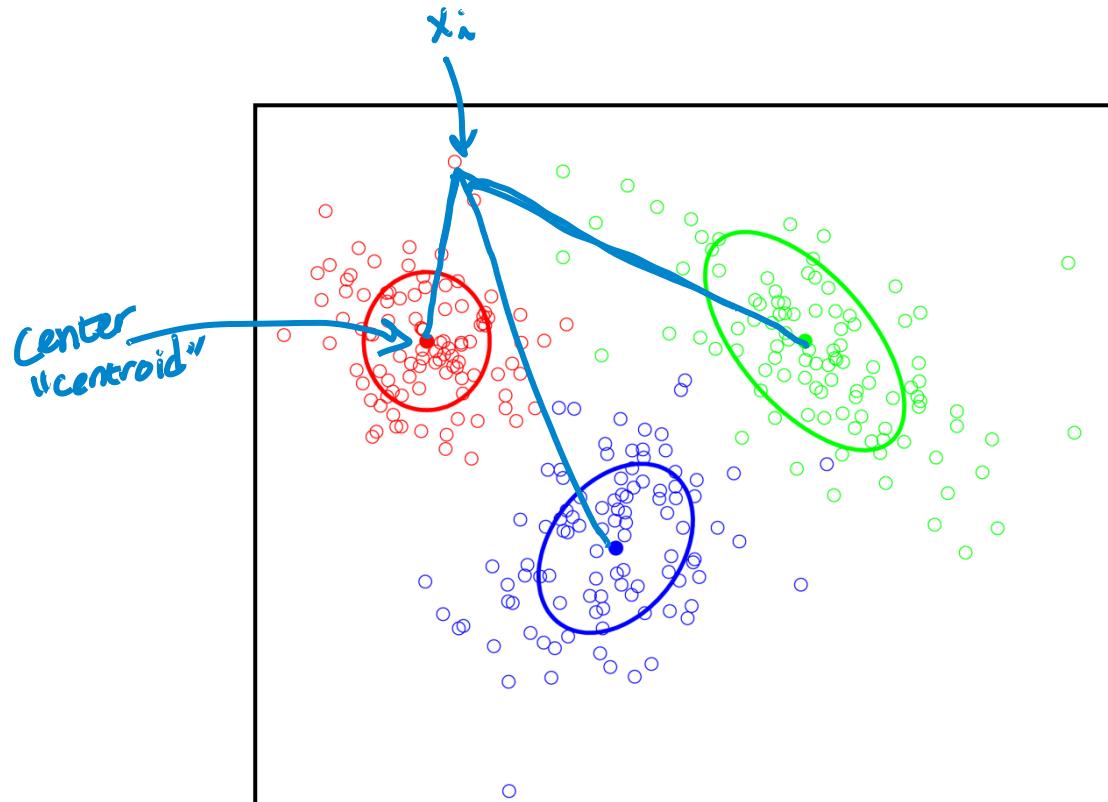


# What defines a cluster?

Cluster defined by  
center & shape/spread

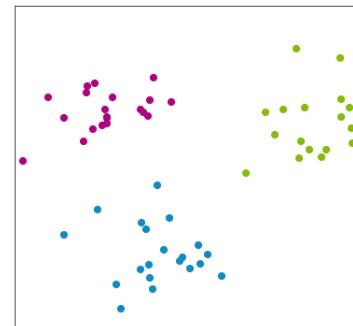
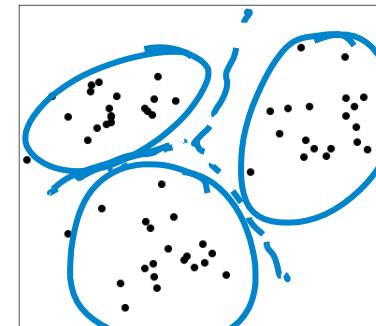
Assign observation  $x_i$  (doc)  
to cluster k (topic label) if

- Score under cluster k is higher than under others
- For simplicity, often define score as **distance to cluster center** (ignoring shape)

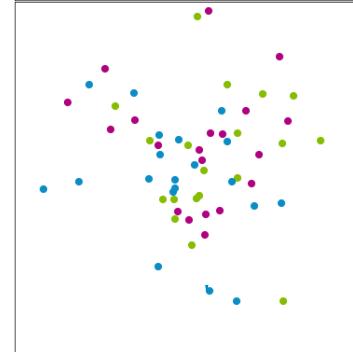
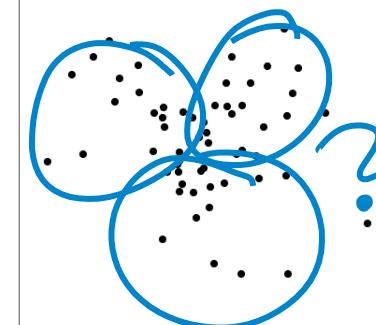


# Hope for unsupervised learning

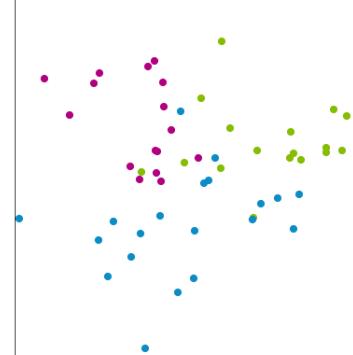
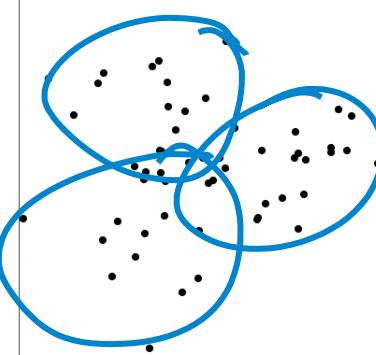
Easy



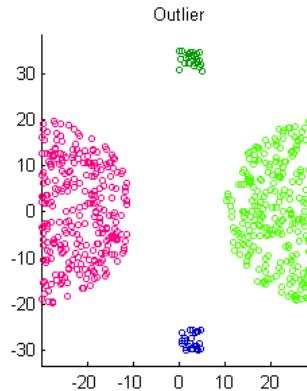
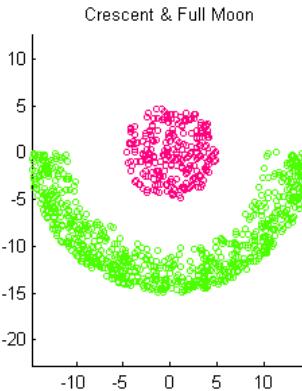
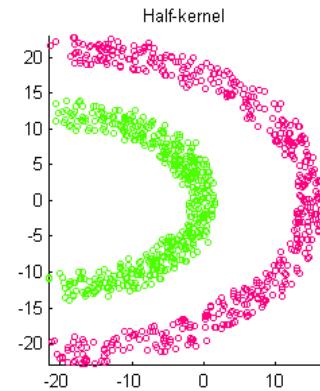
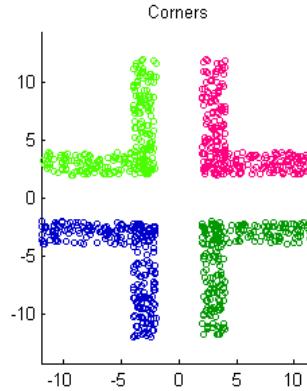
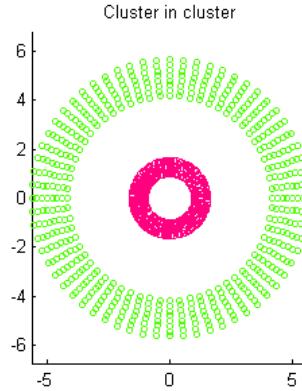
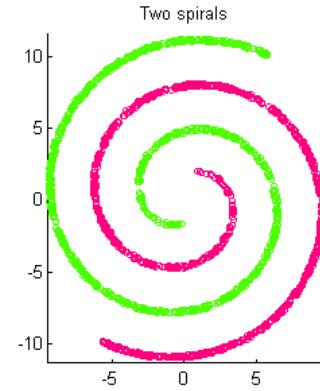
Impossible



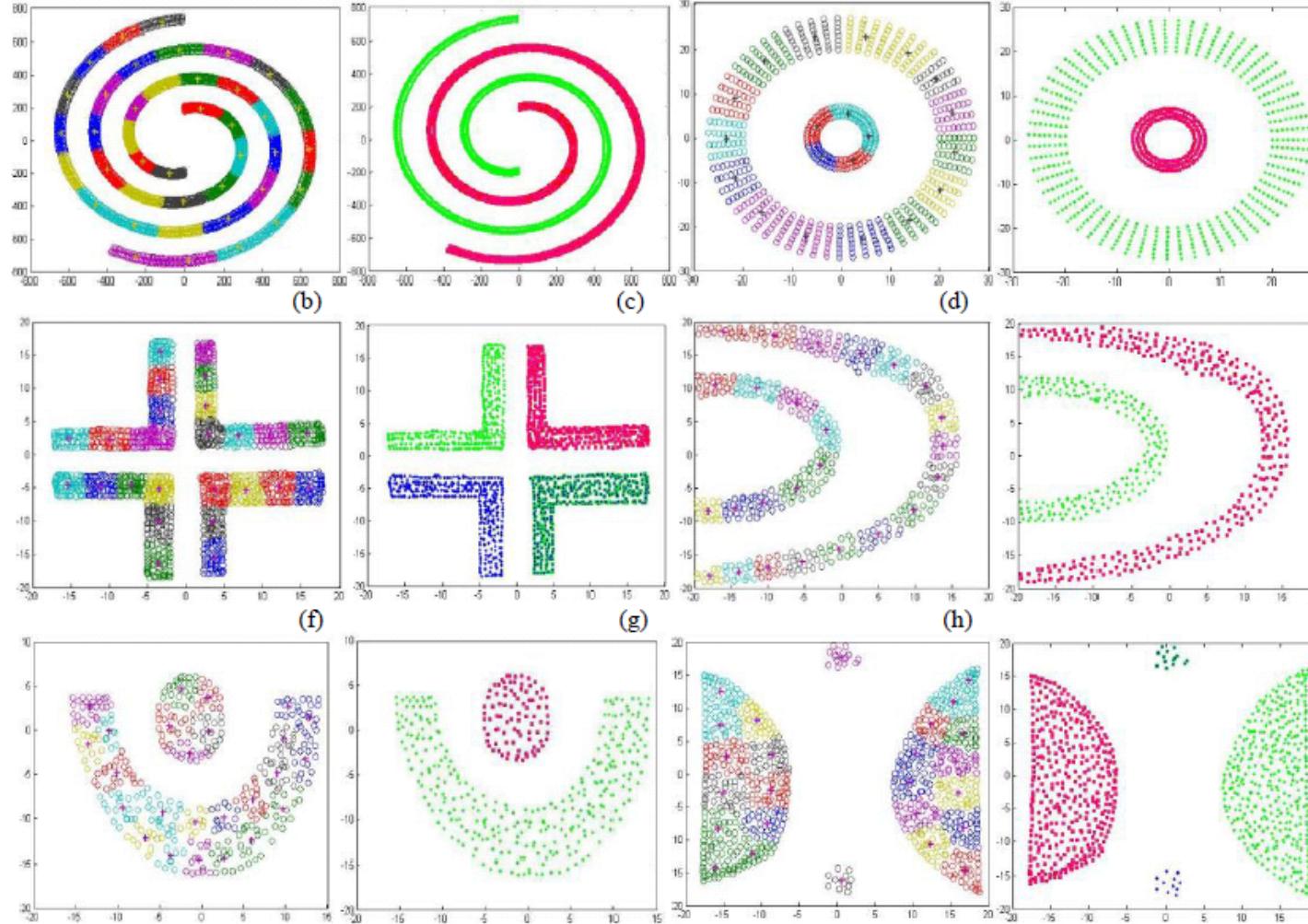
In between



# Other (challenging!) clusters to discover...



# Other (challenging!) clusters to discover...

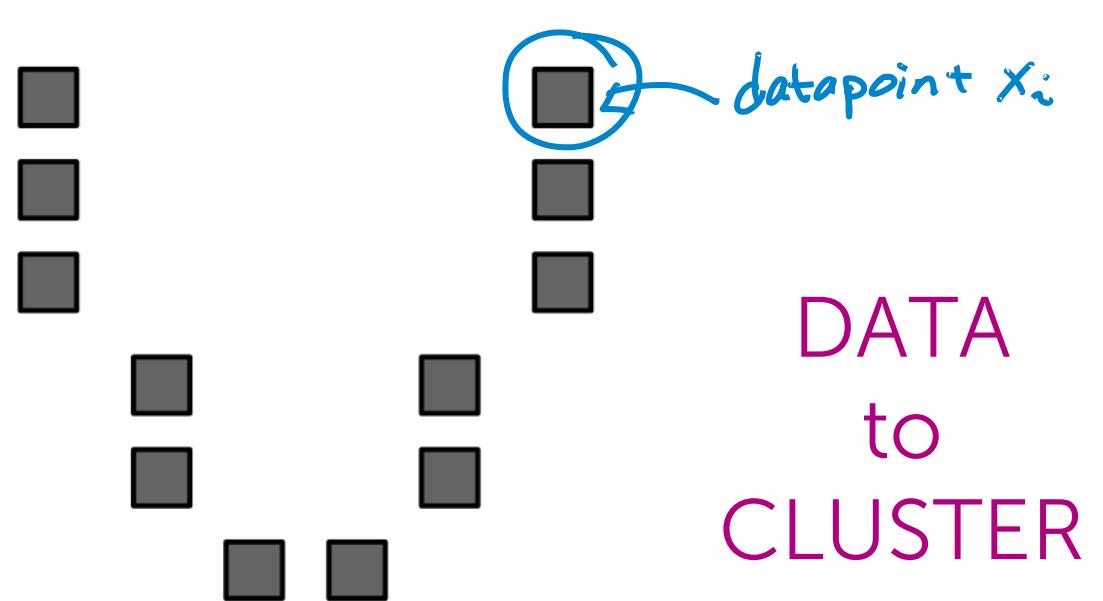


# k-means: A clustering algorithm

# k-means

Assume

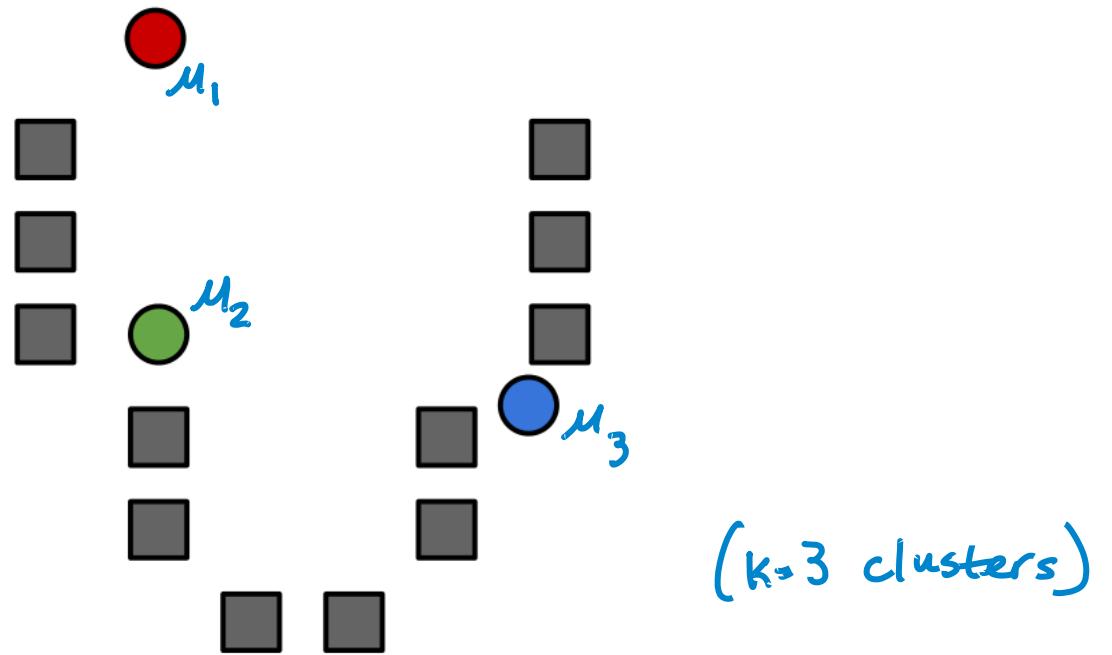
- Score = distance to cluster center  
(smaller better)



# k-means algorithm

0. Initialize cluster centers

$$\mu_1, \mu_2, \dots, \mu_k$$



# k-means algorithm

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

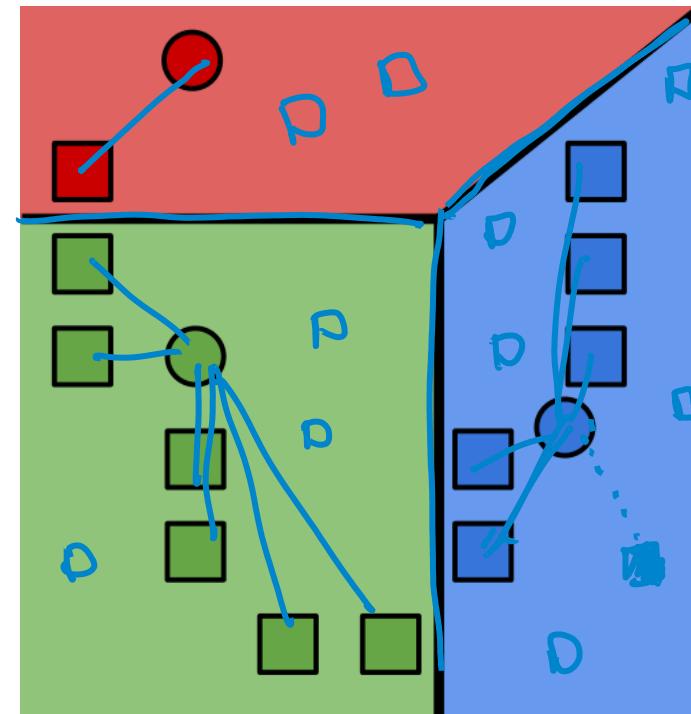
$$z_i \leftarrow \arg \min_j \|\mu_j - \mathbf{x}_i\|_2^2$$

Inferred label for obs i, whereas supervised learning has given label  $y_i$

return index  $j$  of the cluster whose center is closest to obs  $x_i$  (whereas min returning minimum value of  $\|\cdot\|_2^2$ )

*jth cluster center (varying)*

*i<sup>th</sup> obs. (fixed)*



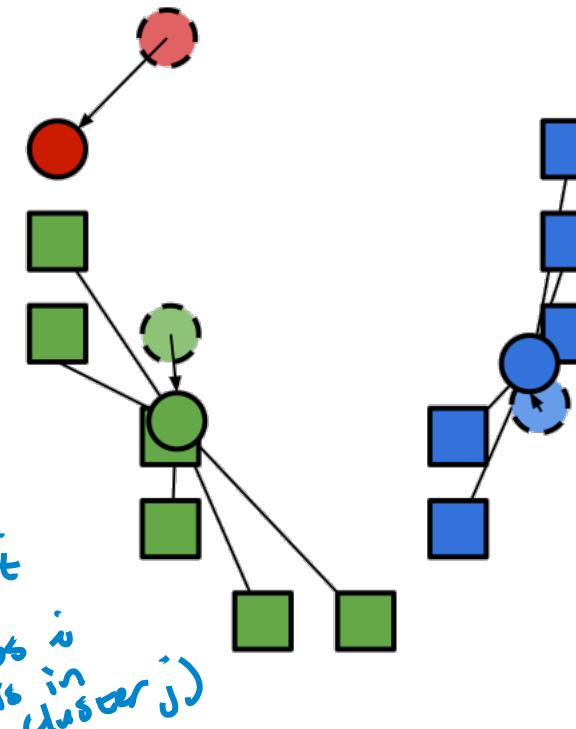
Voronoi tessellation  
(for visualization only ...  
you don't need to compute this)

# k-means algorithm

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

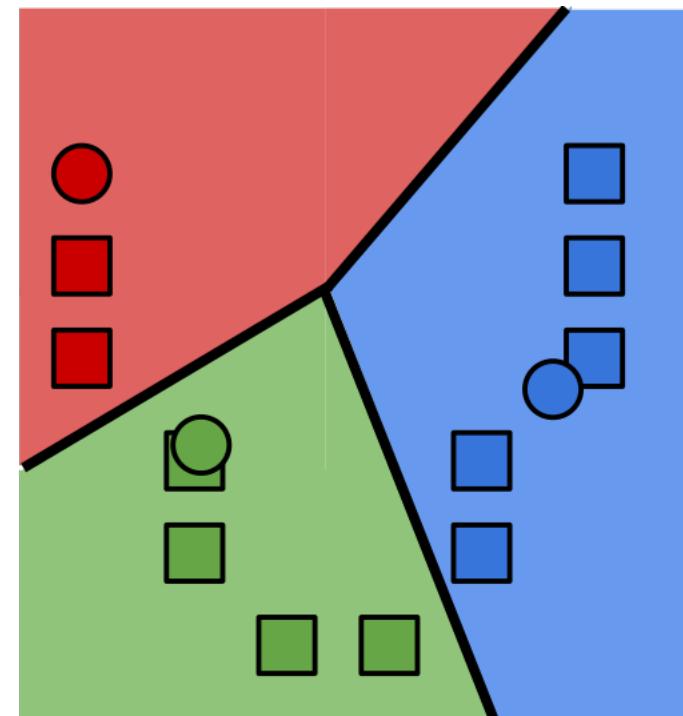
$$\underline{\underline{\mu_j}} = \frac{1}{n_j} \sum_{\substack{i: z_i=j \\ \text{\# of obs. in cluster } j}} \mathbf{x}_i$$

*all obs.  $i$  such that  $z_i=j$  (obs. in cluster  $j$ )*



# 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





# k-means as coordinate descent

# A coordinate descent algorithm

1. Assign observations to closest cluster center

$$z_i \leftarrow \arg \min_j \|\mu_j - \mathbf{x}_i\|_2^2$$

2. Revise cluster centers as mean of assigned observations

$$\mu_j = \frac{1}{n_j} \sum_{i:z_i=j} \mathbf{x}_i$$

equivalent to

$$\mu_j \leftarrow \arg \min_{\mu} \sum_{i:z_i=j} \|\mu - \mathbf{x}_i\|_2^2$$

# A coordinate descent algorithm

1. Assign observations to closest cluster center

$$z_i \leftarrow \arg \min_j \|\mu_j - \mathbf{x}_i\|_2^2$$

2. Revise cluster centers as mean of assigned observations

$$\mu_j \leftarrow \arg \min_{\mu} \sum_{i:z_i=j} \|\mu - \mathbf{x}_i\|_2^2$$

# A coordinate descent algorithm

1. Assign observations to closest cluster center

$$z_i \leftarrow \arg \min_j \|\mu_j - \mathbf{x}_i\|_2^2$$

2. Revise cluster centers as mean of assigned observations

$$\mu_j \leftarrow \arg \min_{\mu} \sum_{i:z_i=j} \|\mu - \mathbf{x}_i\|_2^2$$

Alternating minimization

1. (z given  $\mu$ ) and 2. ( $\mu$  given z)  
= **coordinate descent**

# Convergence of k-means

Converges to:

- Global optimum X
- Local optimum Local optimum
- neither X

# Smart initialization with k-means++

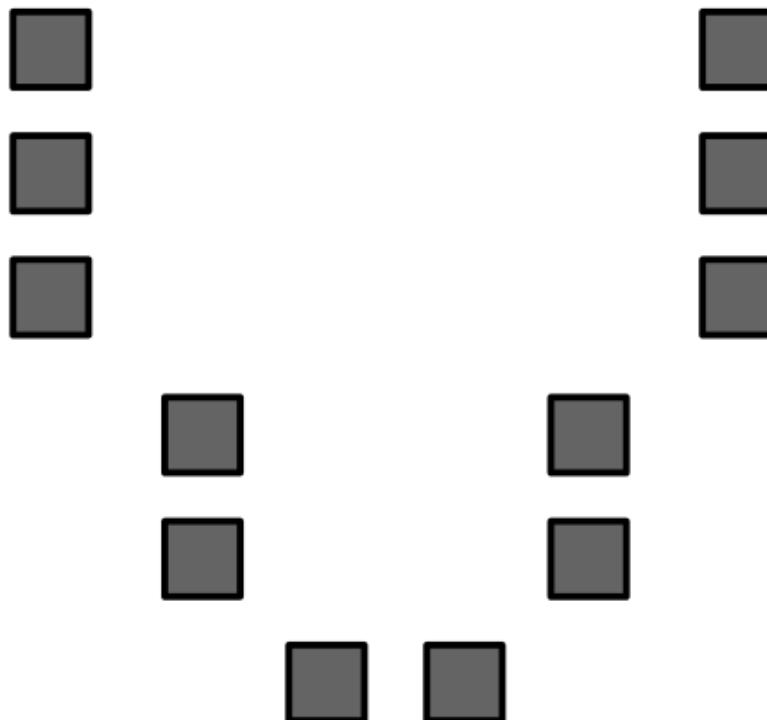
# k-means++ overview

Initialization of k-means algorithm is critical to quality of local optima found

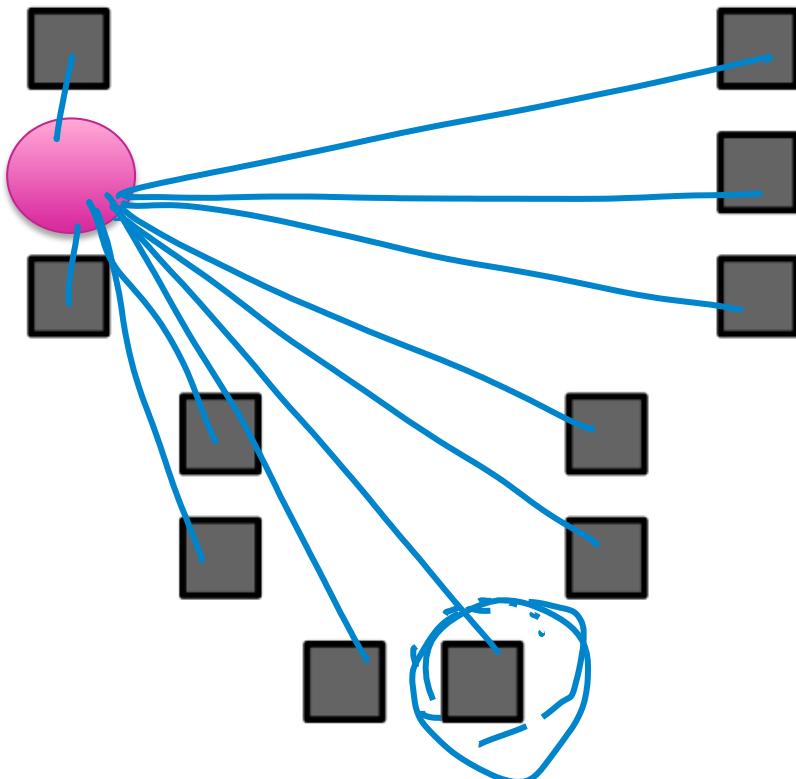
## Smart initialization:

1. Choose first cluster center uniformly at random from data points
2. For each obs  $\mathbf{x}$ , compute distance  $d(\mathbf{x})$  to nearest cluster center
3. Choose new cluster center from amongst data points, with probability of  $\mathbf{x}$  being chosen proportional to  $d(\mathbf{x})^2$
4. Repeat Steps 2 and 3 until  $k$  centers have been chosen

# k-means++ visualized

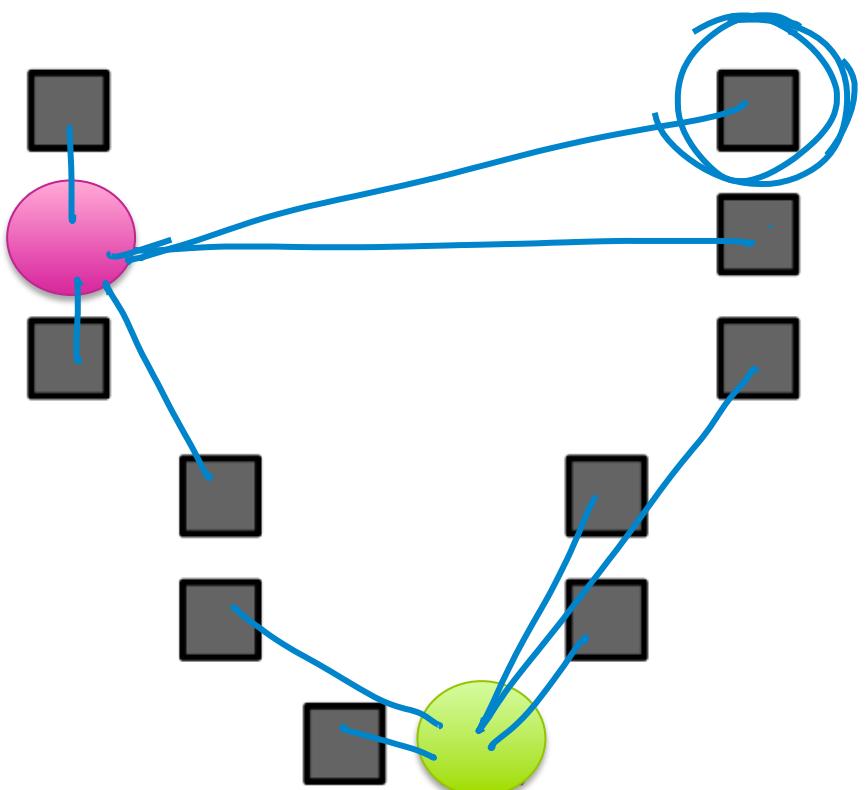


# k-means++ visualized

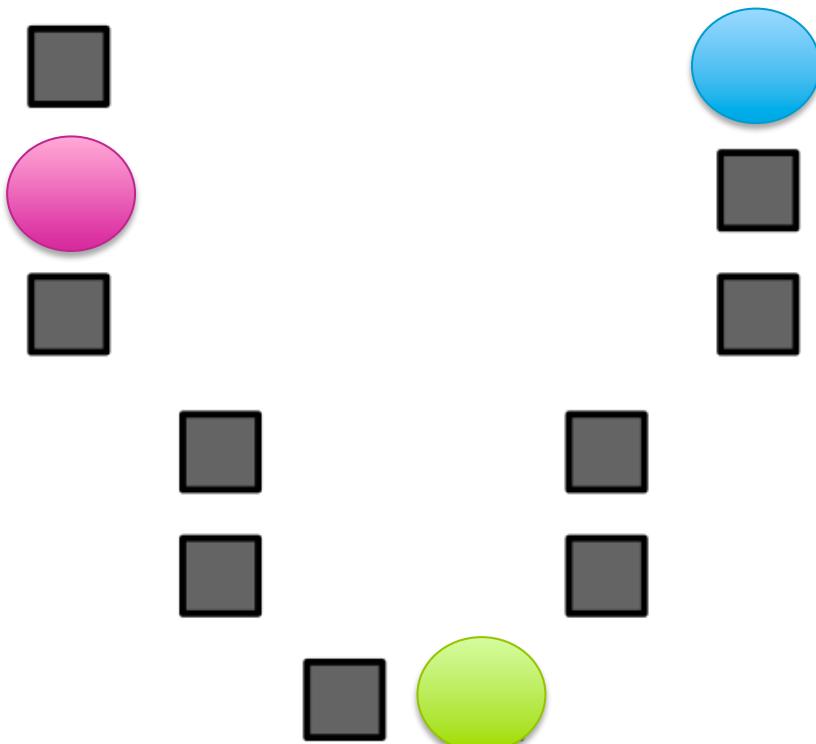


more likely to  
select a datapoint  
as a cluster center  
if that datapoint is  
far away  
( $dist^2$  increases  
this effect)

# k-means++ visualized



# k-means++ visualized



# k-means++ pros/cons

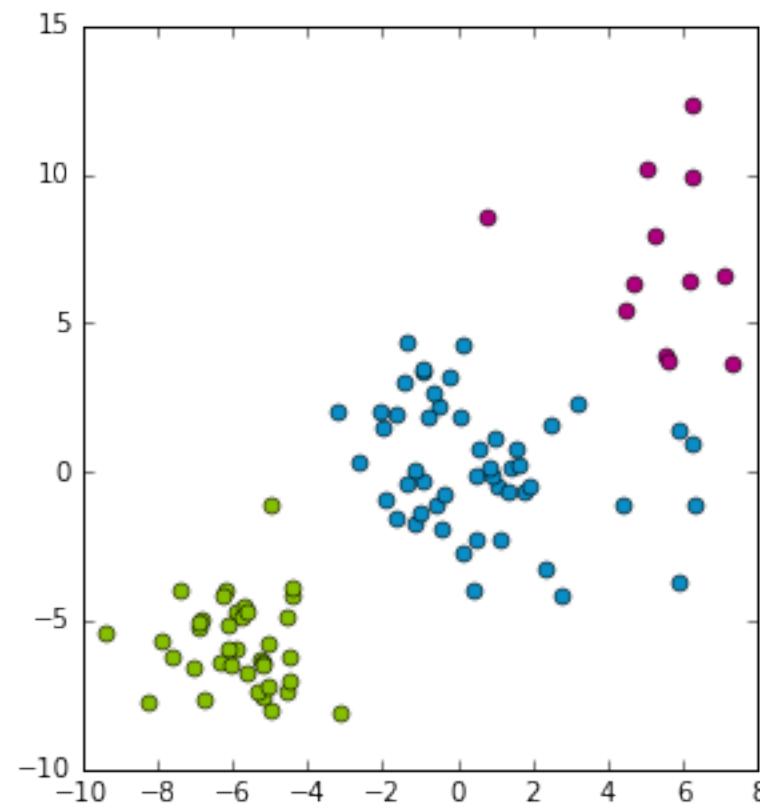
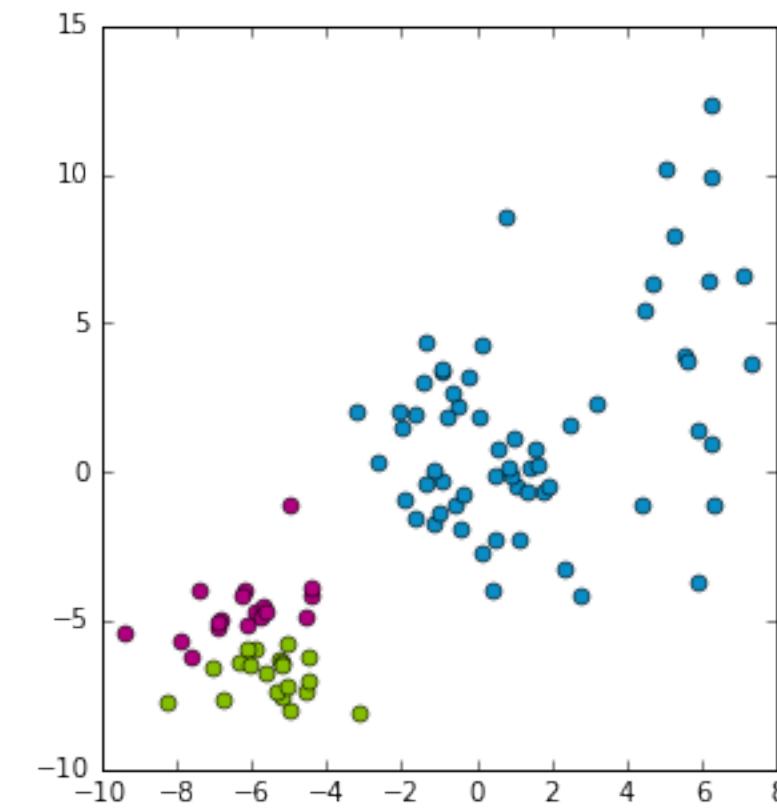
Computationally costly relative to random initialization, but the subsequent k-means often converges more rapidly

Tends to improve quality of local optimum and lower runtime

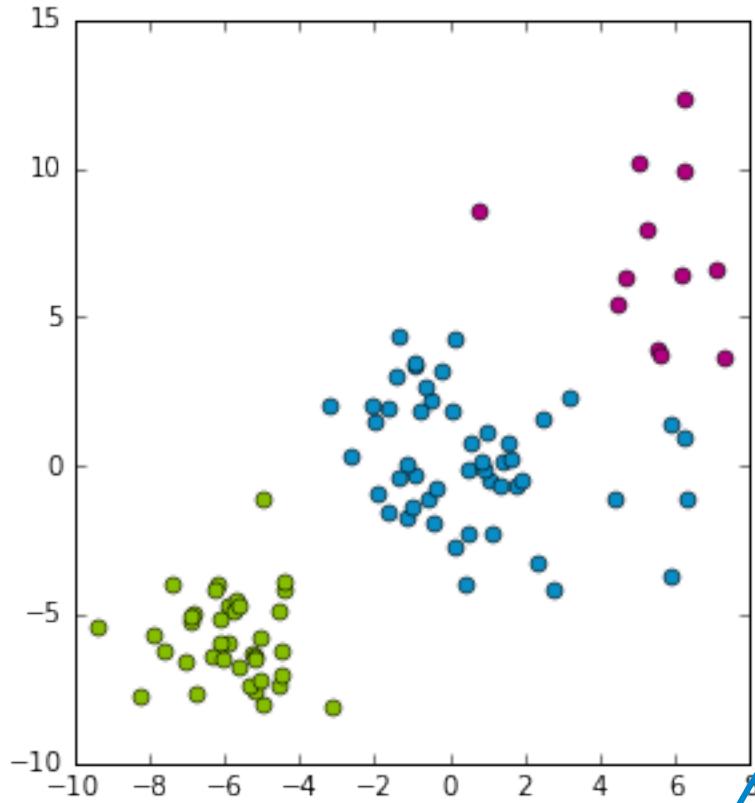
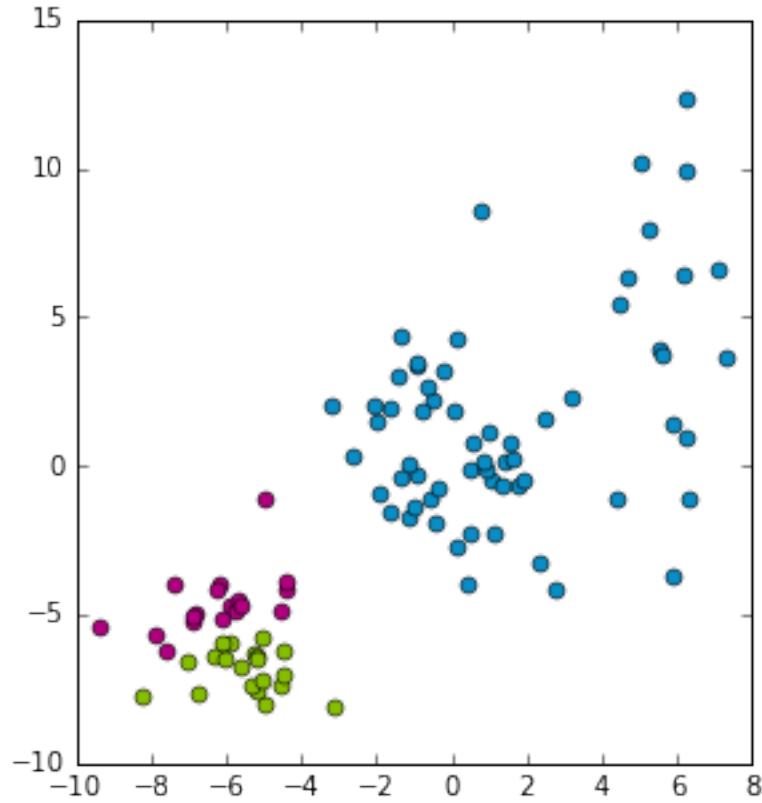


Assessing quality of the clustering  
and choosing the # of clusters

# Which clustering do I prefer?



# k-means objective



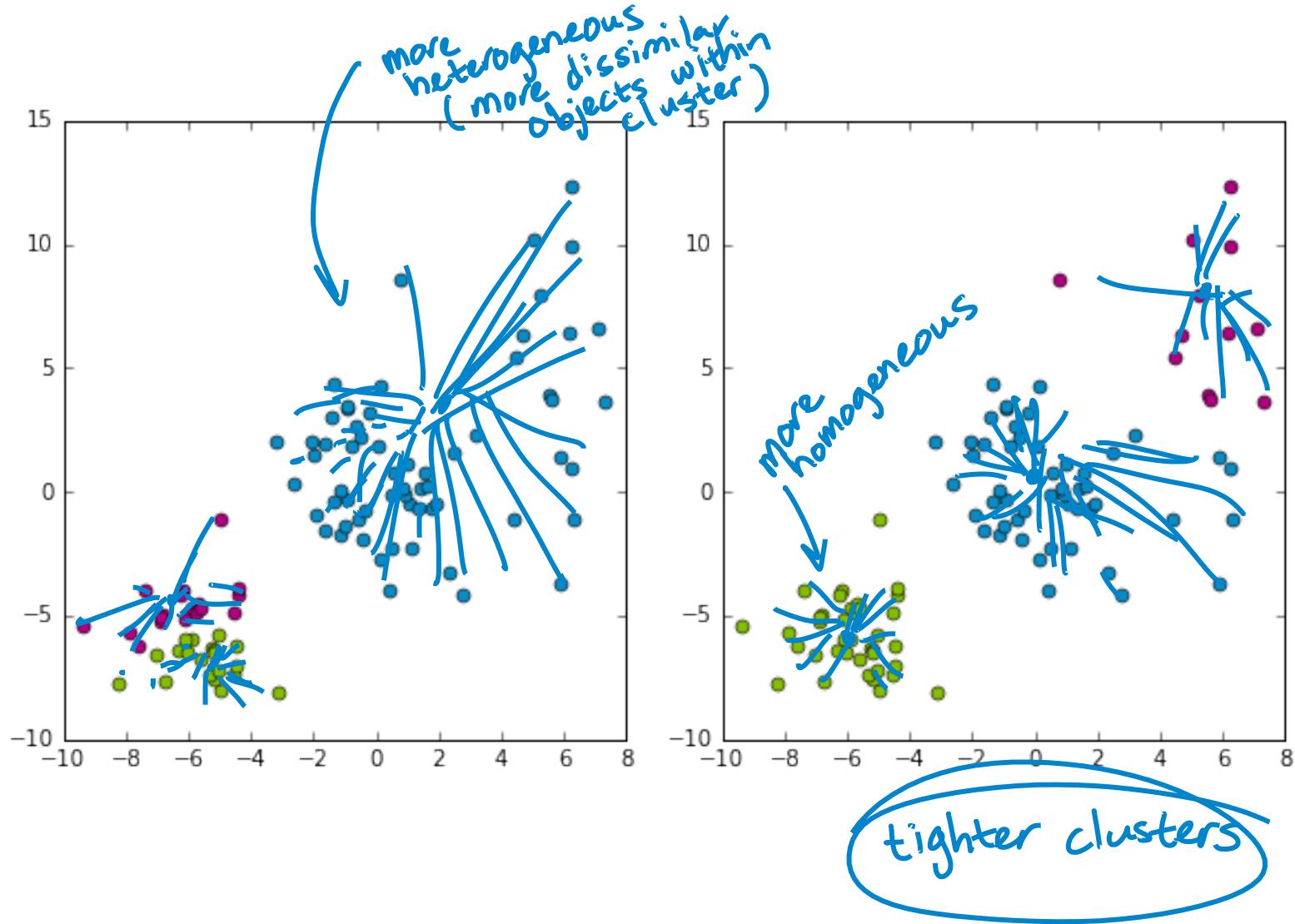
**k-means** is trying to minimize the **sum of squared distances**:

$$\min_{\{z_i\}, \{\mu_j\}} \sum_{j=1}^k \sum_{i:z_i=j} \|\mu_j - \mathbf{x}_i\|_2^2$$

Annotations:

- $k$  (blue arrow) → sum over all clusters
- $\sum_{i:z_i=j}$  (blue arrow) → sum of squared distances in cluster  $j$
- $\min_{\{z_i\}, \{\mu_j\}}$  (blue handwritten text)

# Cluster heterogeneity



Measure of quality of given clustering:

$$\sum_{j=1}^k \sum_{i:z_i=j} \|\mu_j - \mathbf{x}_i\|_2^2$$

Lower is better!

# What happens as k increases?

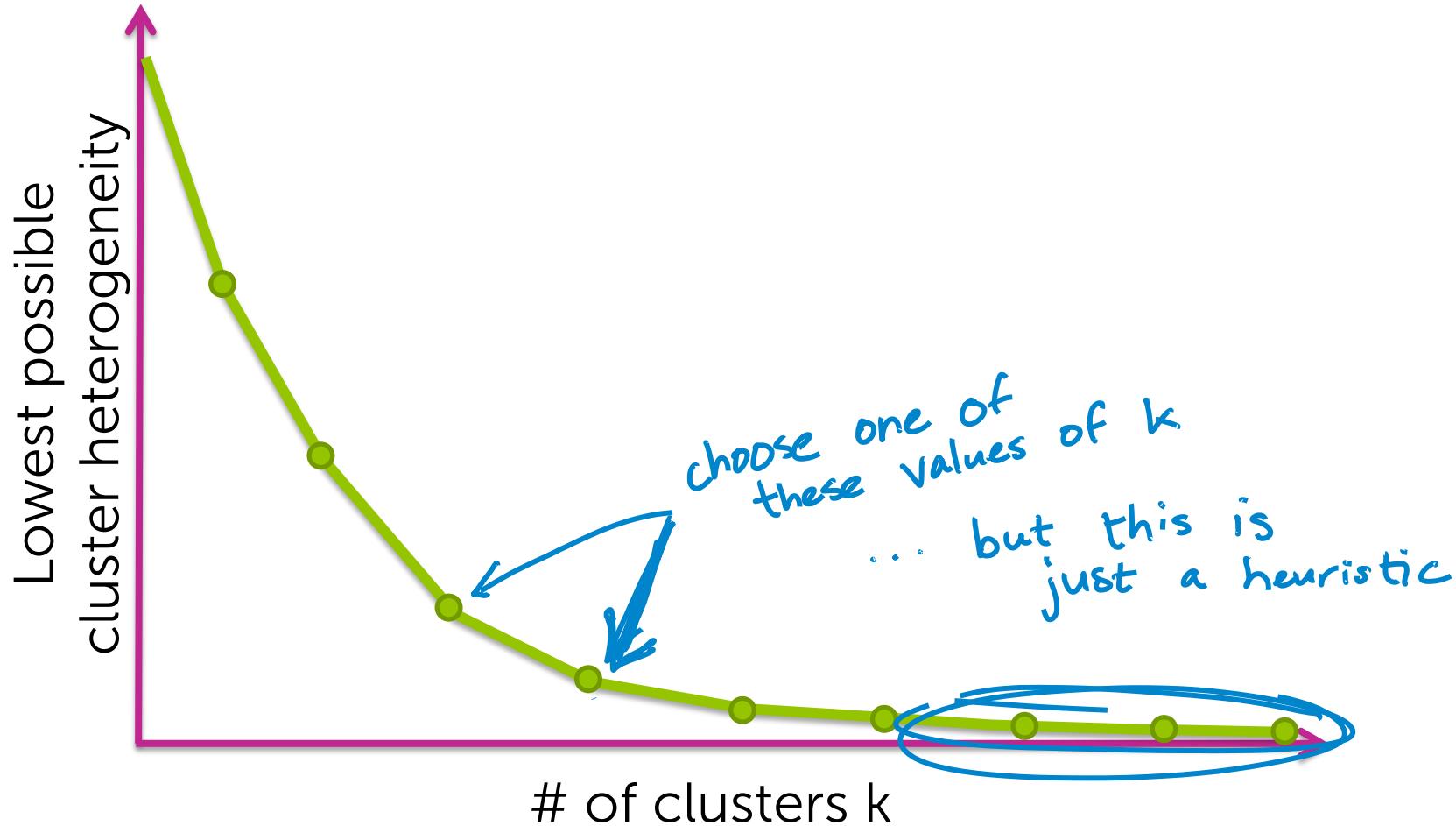
Can refine clusters more and more to the data  
→ overfitting!

**Extreme case** of  $k=N$ :

- can set each cluster center equal to datapoint
- heterogeneity =  $0$  ! *(all distances to cluster centers are 0)*

Lowest possible cluster heterogeneity  
decreases with increasing k

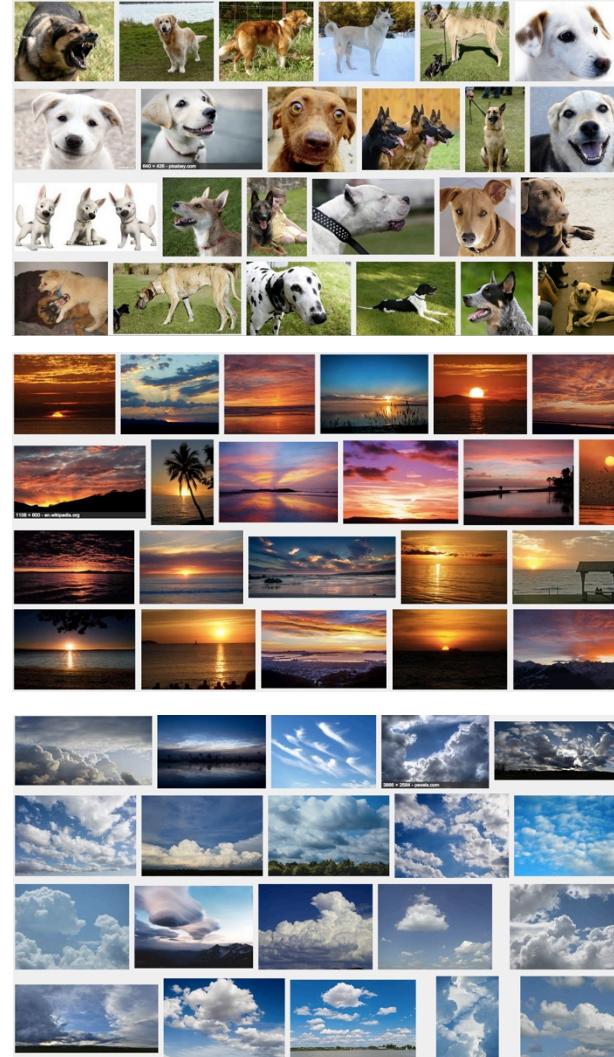
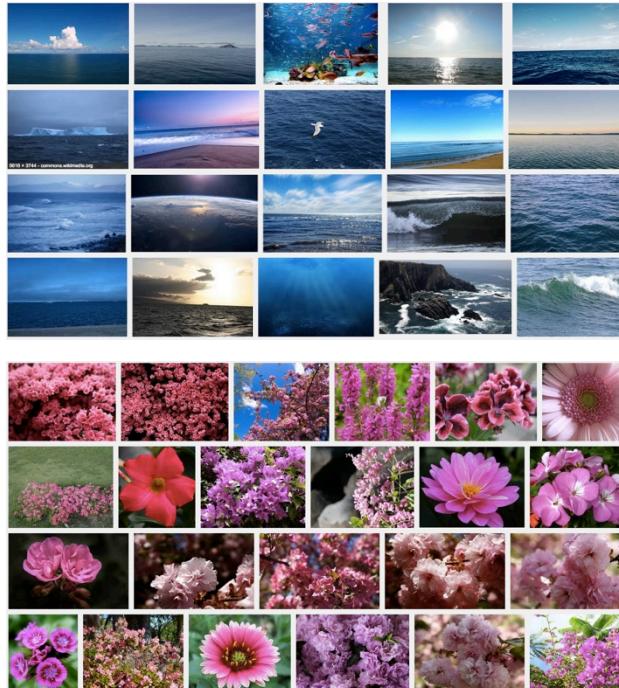
# How to choose k?



# Other examples

# Clustering images

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



# Structuring web search results

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



- Use clustering to **structure output**

# Grouping patients by medical condition

- Better characterize subpopulations and diseases

# Products on Amazon

- Discover product categories from purchase histories



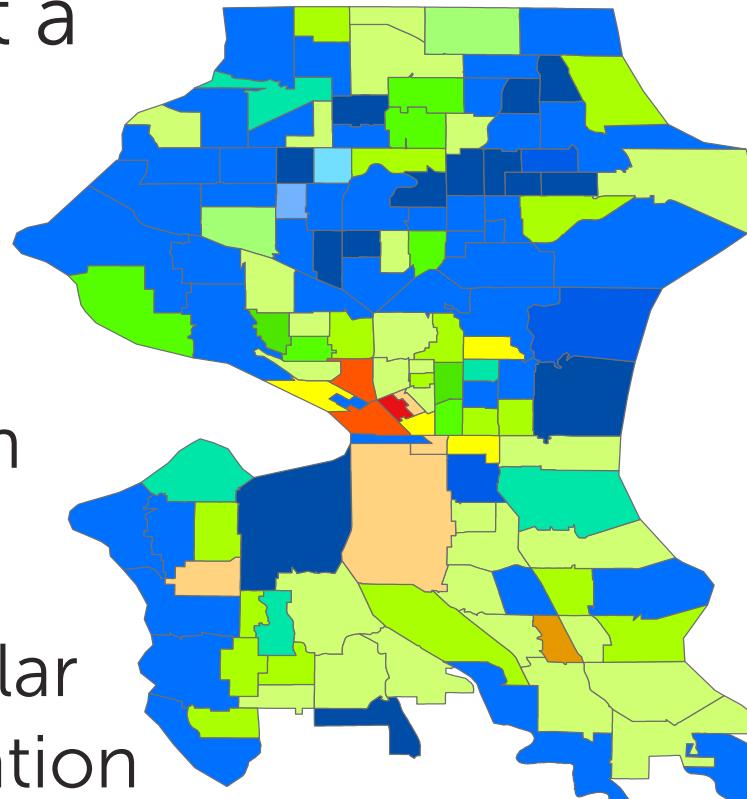
~~"furniture"~~  
**"baby"**



- Or discovering groups of **users**

# Discovering similar neighborhoods

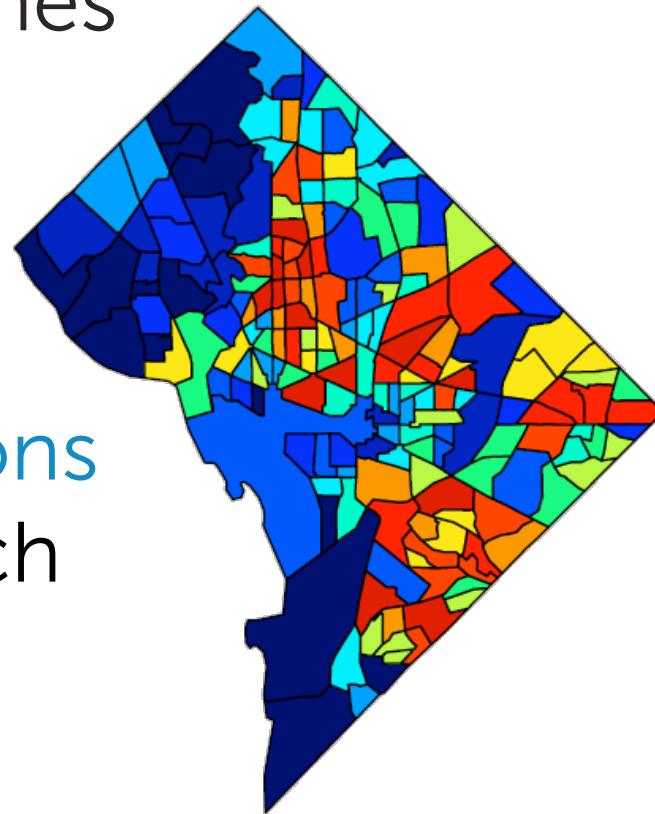
- **Task 1:** Estimate price at a small regional level
- **Challenge:**
  - Only a few (or no!) sales in each region per month
- **Solution:**
  - Cluster regions with similar trends and share information within a cluster



City of Seattle

# 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 k-means

# What you can do now...

- Describe potential applications of clustering
- 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
- Interpret k-means as a coordinate descent algorithm