Clustering:
Grouping Related Docs




CS229: Machine Learning

Carlos Guestrin
Stanford University

Slides include content developed by and co-developed with Emily Fox

Motivating clustering approaches

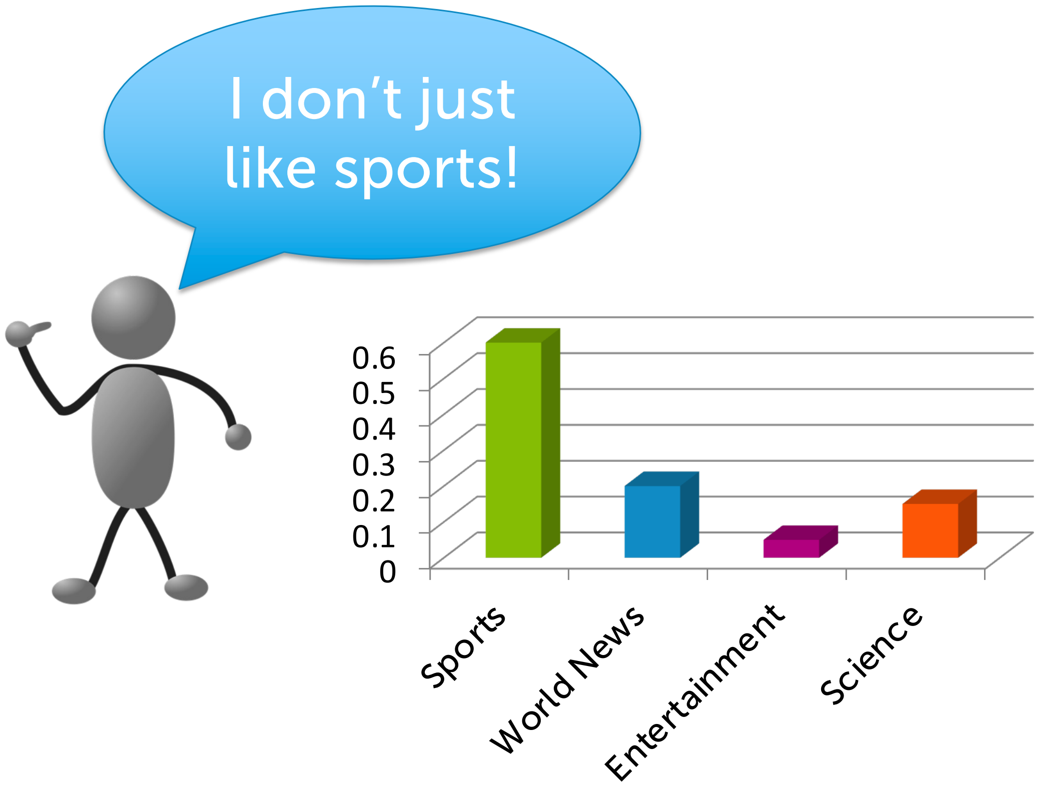
Goal: Structure documents by topic

Discover groups (*clusters*) of related articles

SPORTS

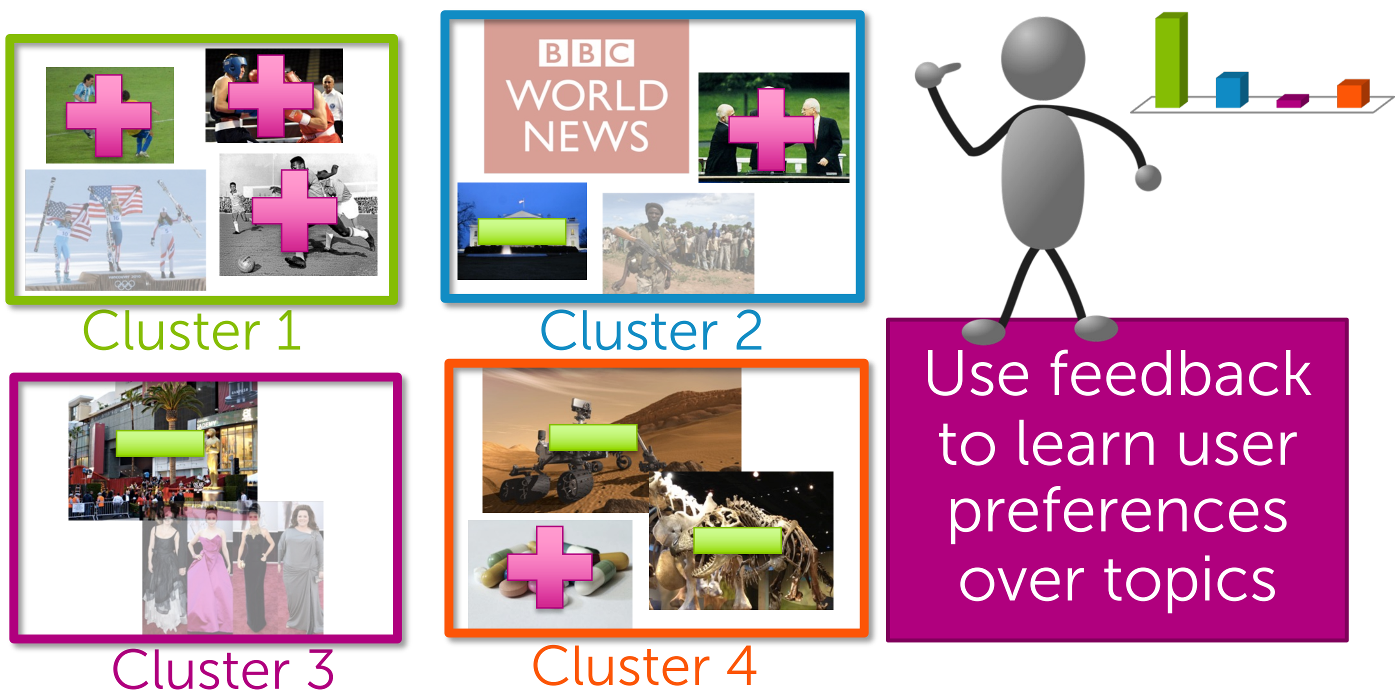
WORLD NEWS

Why might clustering be useful?



Learn user preferences

Set of clustered documents read by user



Clustering: An unsupervised learning task

What if some of the labels are known?

Training set of labeled docs

Clustering

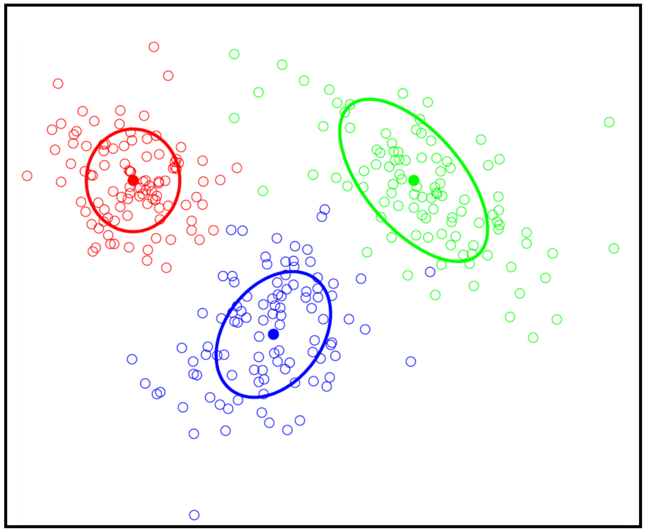
No labels provided

...uncover cluster structure from input alone

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

Output: cluster labels z_i

An unsupervised learning task

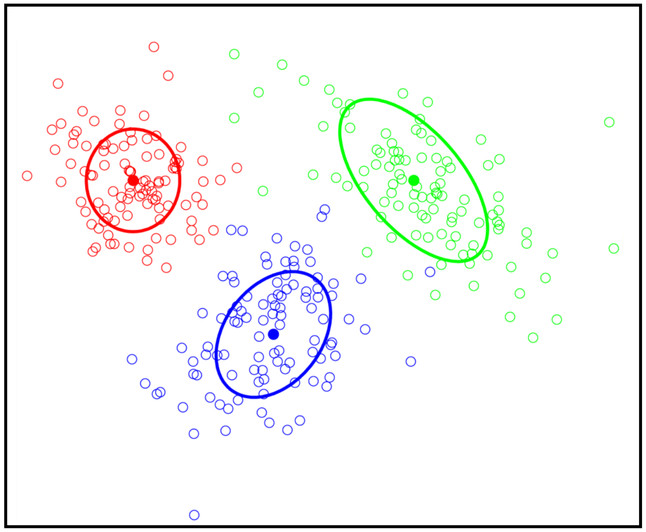


What defines a cluster?

Cluster defined by center & shape/spread

Assign observation $\mathbf{x}_i$ (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)

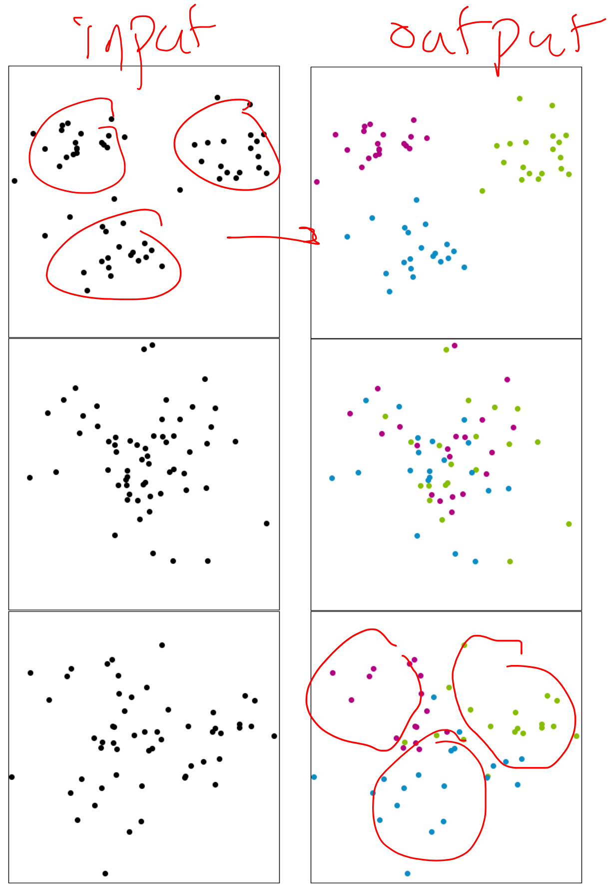


Hope for unsupervised learning

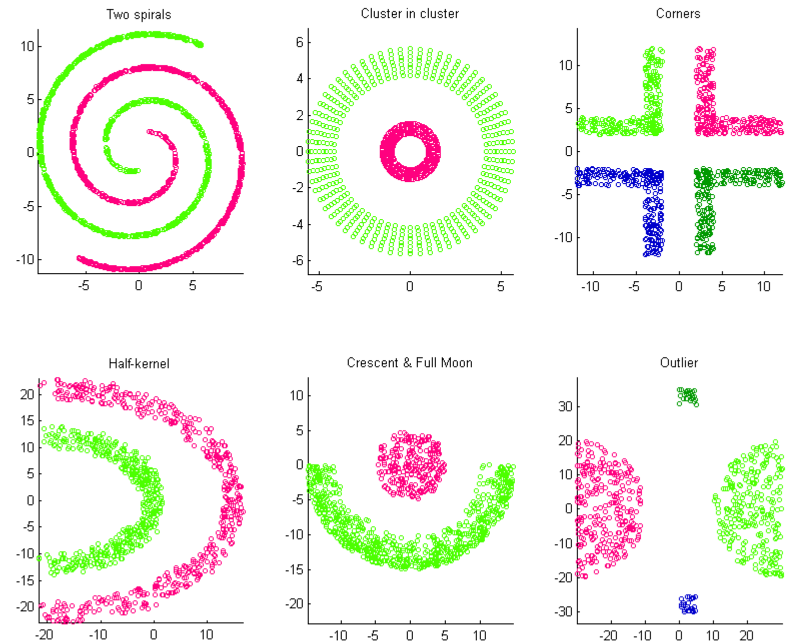
Easy

Impossible

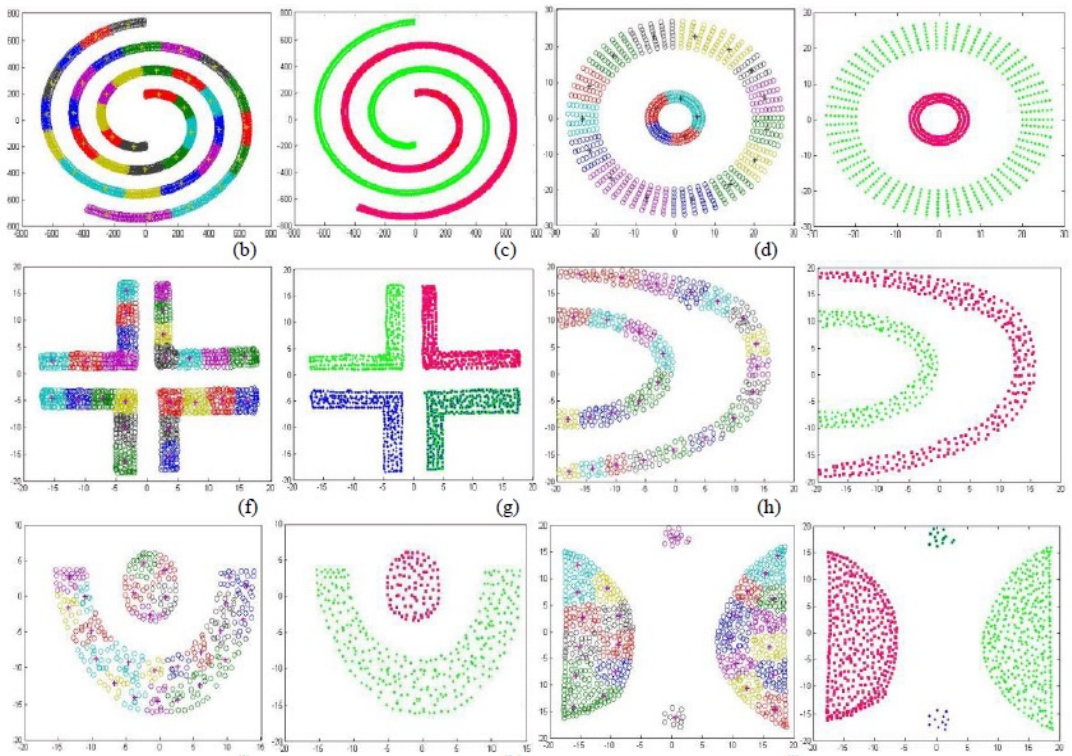
In between



Other (challenging!) clusters to discover…



Other (challenging!) clusters to discover…

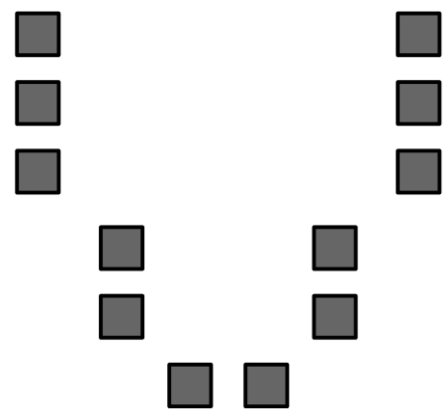


k-means: A clustering algorithm

k-means

Assume

- Score= distance to cluster center (smaller better)

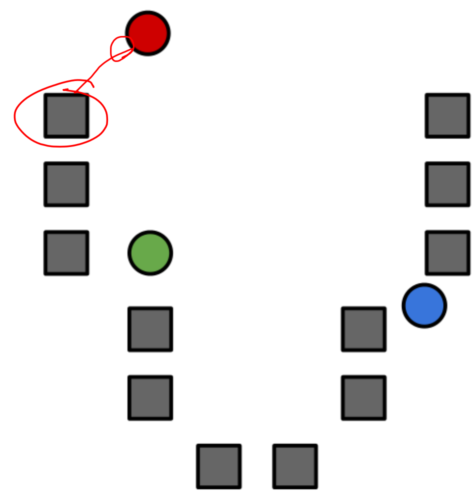


DATA
to
CLUSTER

k-means algorithm

0. Initialize cluster centers

$$\mu_1, \mu_2, \ldots, \mu_k$$



k-means algorithm

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

cluster center

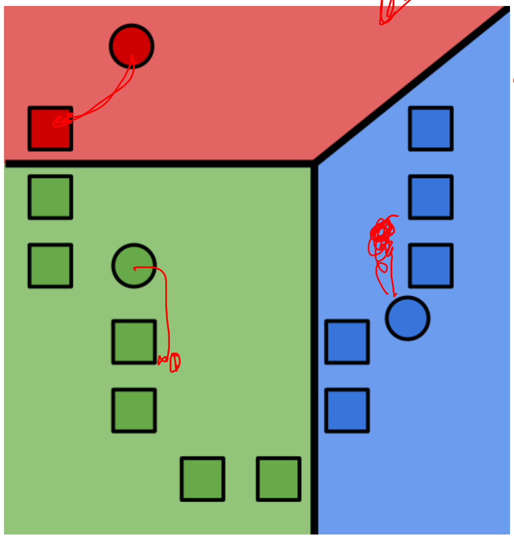
data point

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

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

" decision boundary"

Voronoi diagram



if this is the cluster, what is the best center?

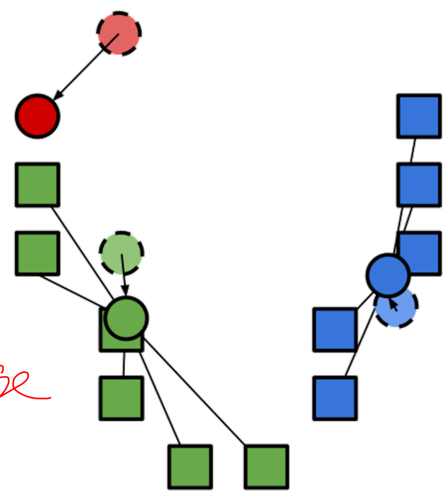
k-means algorithm

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

bari center

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

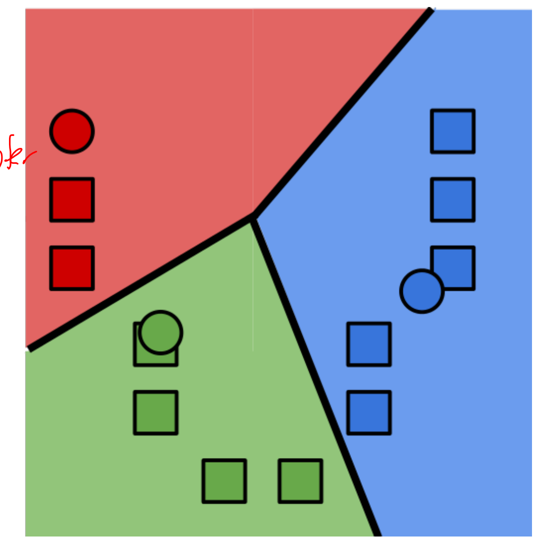
average



k-means algorithm

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

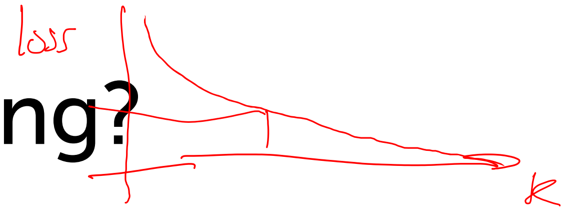
   no data point changes cluster assignment



Why does K-means work???

- What's k-means optimizing? ← What's the loss function
- Does it always converge? ← ??

What is k-means optimizing?



- Potential function F(**μ**,**z**) of centers **μ** and point allocations **z**:

$$F(\mu, z) = \sum_{i=1}^{N} \|\mu_{z_i} - x_i\|_2^2$$

loss function

cluster assignment

assigned cluster $z_i \in \{1, \dots, k\}$

input feature vector

- Optimal k-means:

$$\min_{\mu} \min_{z} = F(\mu, z)$$

Does K-means converge??? Part 1

- Optimize potential function:

$$\min_{\mu} \min_{\mathbf{z}} F(\mu, \mathbf{z}) = \min_{\mu} \min_{\mathbf{z}} \sum_{j=1}^{N} \left\| \mu_{z_i} - x_i \right\|_2^2$$

Classification step:

- Fix **μ** and minimize **z**:

$$\underbrace{\min_{z_1,\dots,z_N}}_{\text{no interaction terms between } z_i \,\&\, z_j} \sum_{i=1}^{N} \|\mu_{z_i} - x_i\|_2^2 = \sum_{i=1}^{N} \underbrace{\min_{z_i \in \{1,\dots,k\}} \|\mu_{z_i} - x_i\|_2^2}_{\text{independent min problem per data point}}$$

≡ "classification step" in k-means

Does K-means converge??? Part 2

n_j is number of data points in cluster j

- Optimize potential function:

$$\min_{\mu}\min_{z} F(\mu, z) = \min_{\mu}\min_{z} \sum_{j=1}^{N} \left\| \mu_{z_i} - x_i \right\|_2^2$$

Recenter step of K-means

- Fix **z** and minimize **μ**:

$$\min_{\mu_1,\ldots,\mu_K} \sum_{i=1}^{N} \| \mu_{z_i} - x_i \|_2^2 = \min_{\mu_1} \min_{\mu_2} \ldots \min_{\mu_K} \overbrace{\sum_{j=1}^{K} \sum_{i:z_i=j} \| \mu_j - x_i \|_2^2}^{\text{no interactions between } \mu_i \,\&\, \mu_j}$$

$$= \sum_{j=1}^{K} \underbrace{\min_{\mu_j} \sum_{i:z_i=j} \| \mu_j - x_i \|_2^2}_{\text{Find } \mu_j \text{ that is closest on avg to points in cluster } j}$$

take derivative, set to zero $\Rightarrow$ optimal μ_j is average:

$$\mu_j = \sum_{i:z_i=j} x_i / n_j$$

Coordinate descent algorithms

$$\min_{\mu} \min_{z} F(\mu, z) = \min_{\mu} \min_{z} \sum_{j=1}^{N} \|\mu_{z_i} - x_i\|_2^2$$

- Want: $\min_a \min_b F(a,b)$
- Coordinate descent:
  - fix a, minimize b
  - fix b, minimize a
  - repeat
- Converges!!!
  - if F is bounded
  - to a (often good) local optimum
    - as we saw in applet (play with it!)
      - (For LASSO it converged to the global optimum, because of convexity)

$F(a,b)$

a_0, b_0

a_1, b_0

a_1, b_1

$F(a_0, b_0) \geq F(a_1, b_0)$

$F(a_1, b_0) \geq F(a_1, b_1)$

non-increasing sequence

converges!!

lower bound on $F(a,b)$

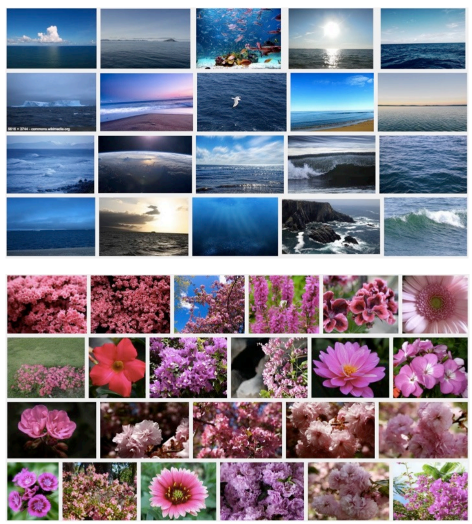
use random restarts or other tricks (k-means ++)

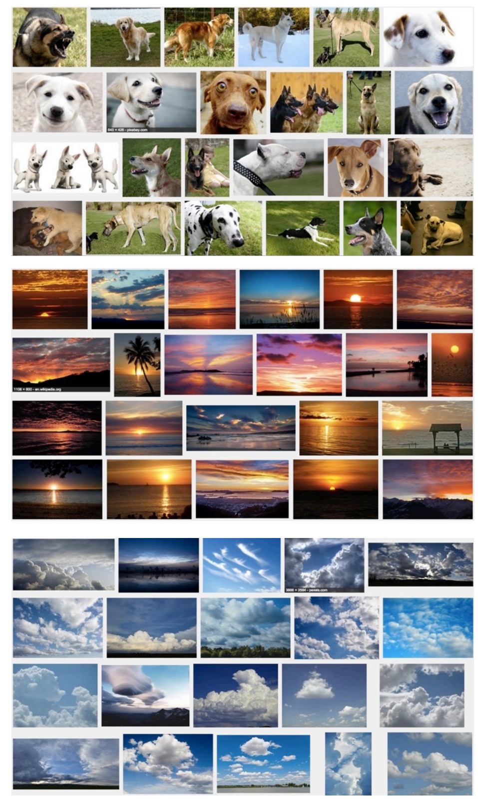
- K-means is a coordinate descent algorithm!

Summary for k-means

Clustering images

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





Limitations of k-means

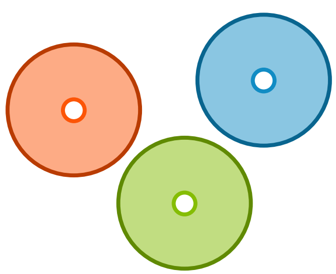
Assign observations to closest cluster center

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

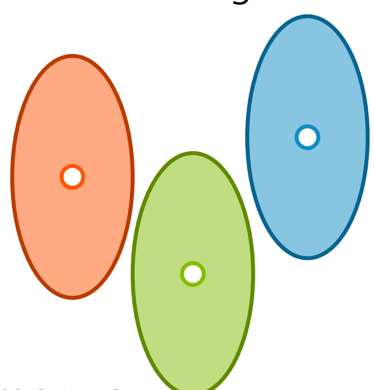
Only center matters

Can use weighted Euclidean,
but requires *known* weights

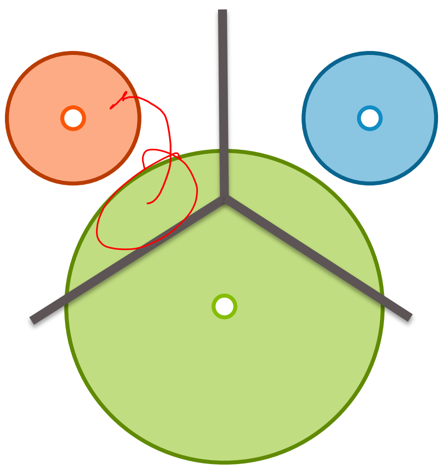
Equivalent to assuming
spherically symmetric clusters



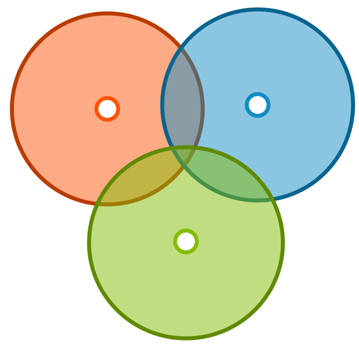
Still assumes all clusters have
the same axis-aligned ellipses



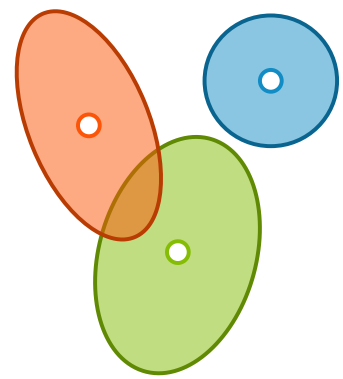
Failure modes of k-means



disparate cluster sizes



overlapping clusters



different
shaped/oriented
clusters

What you can do now…

- 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
- Describe potential applications of clustering