

95-865 Unstructured Data Analytics

Lecture 8: Clustering (cont'd)

Slides by George H. Chen

News Flash: Quiz 1 is Happening This Friday 🤖

during the recitation slot!

Reminder regarding the format:

- In person, on paper
- Quiz length: 80 minutes
- No electronics may be used during the exam
(e.g., do not use a laptop, tablet, phone, calculator)
- Open notes (must be on paper and not electronic)
 - There is no limit on how many pages you bring
- Late exams will *not* be accepted

Topic coverage: up to and including end of lecture last Friday Nov 10

Other Things

- I've already released many past Quiz 1s
(see my Nov 8 & Nov 14 Canvas announcements for details)
- The Quiz 1 review session is tomorrow (Wednesday Nov 15),
7:30pm-8:30pm over Zoom (the link is in Canvas)
- **To get the most out of this review session, please attempt the
Spring 2023 Quiz 1 beforehand!**
- This review session will be recorded so if you can't make or want to
watch the recording afterward, it will be available

Today

1. Wrap of code demo from last Friday's lecture
2. A simple strategy for choosing the number of clusters for *k*-means and GMMs
3. Clustering on unstructured data

Automatically Choosing the Number of Clusters k

For $k = 2, 3, \dots$ up to some user-specified max value:

Fit model (k -means or GMM) using k

Compute a score for the model

But what score function should we use?

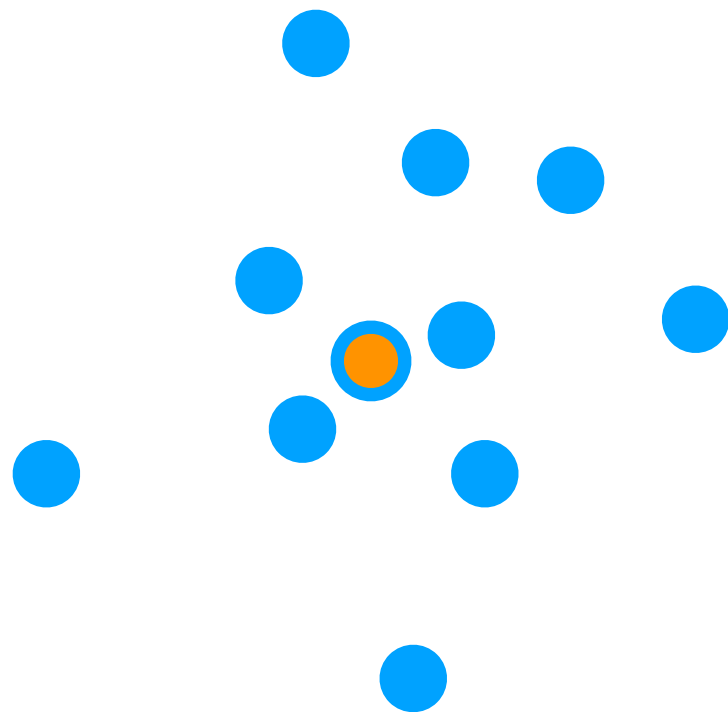
Use whichever k has the best score

No single way of choosing k is the “best” way

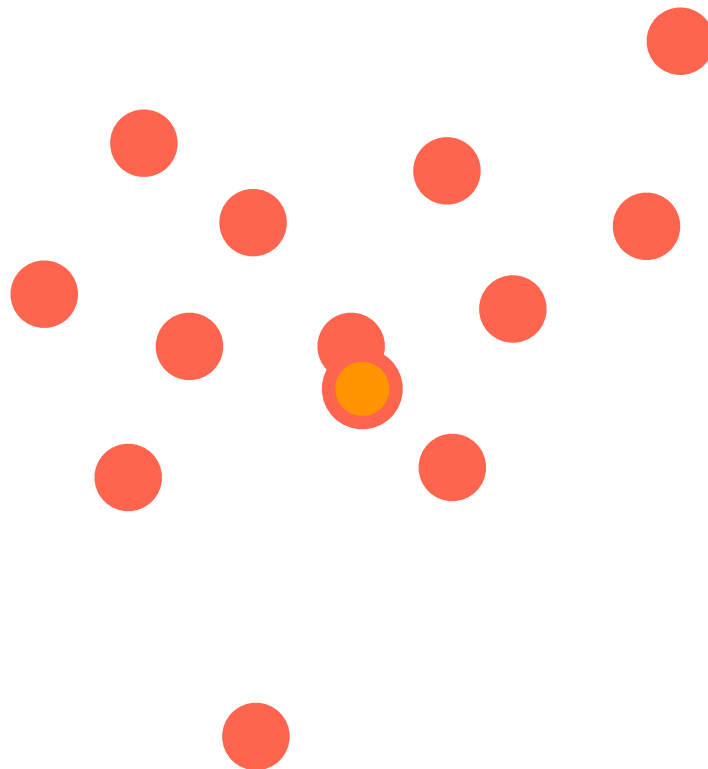
Here's an example of a score
function you don't want to use

Residual Sum of Squares

Look at one cluster at a time



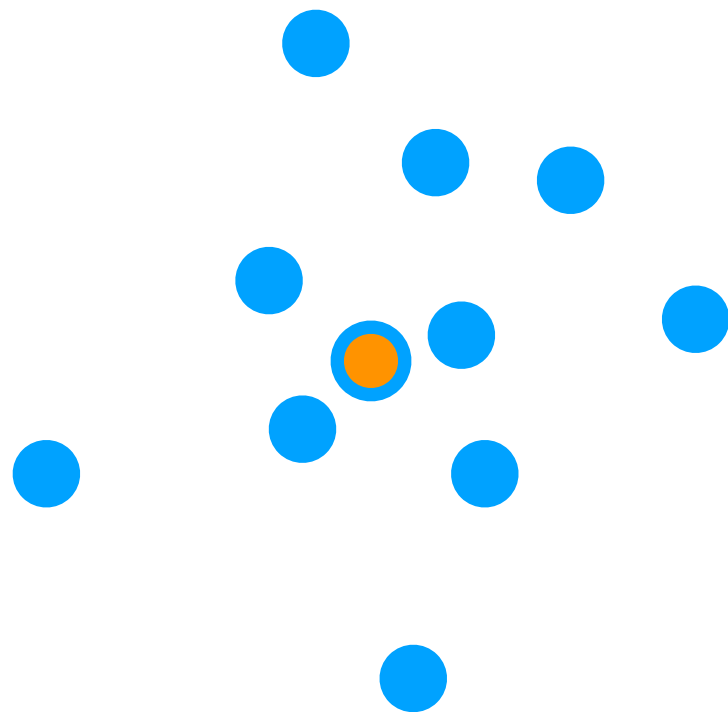
Cluster 1



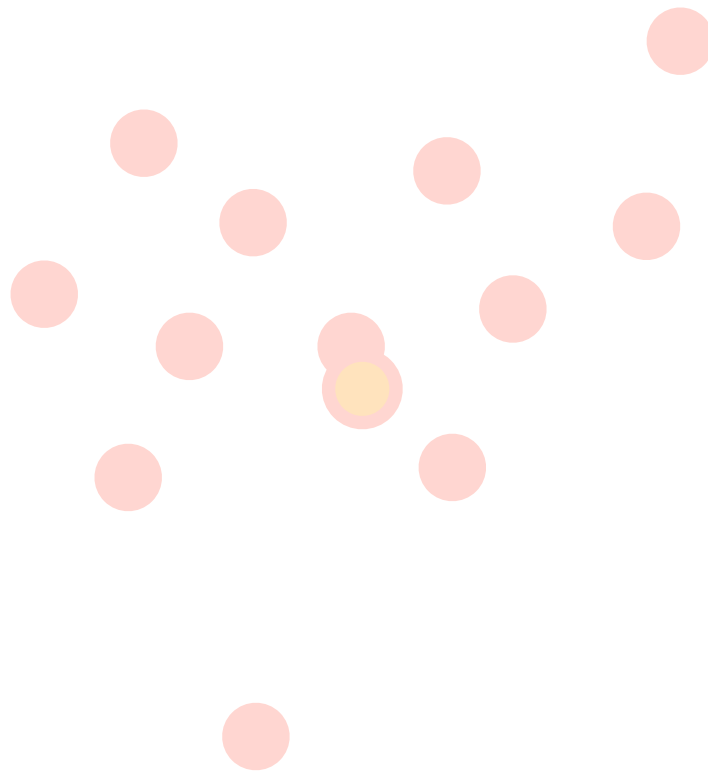
Cluster 2

Residual Sum of Squares

Look at one cluster at a time



Cluster 1

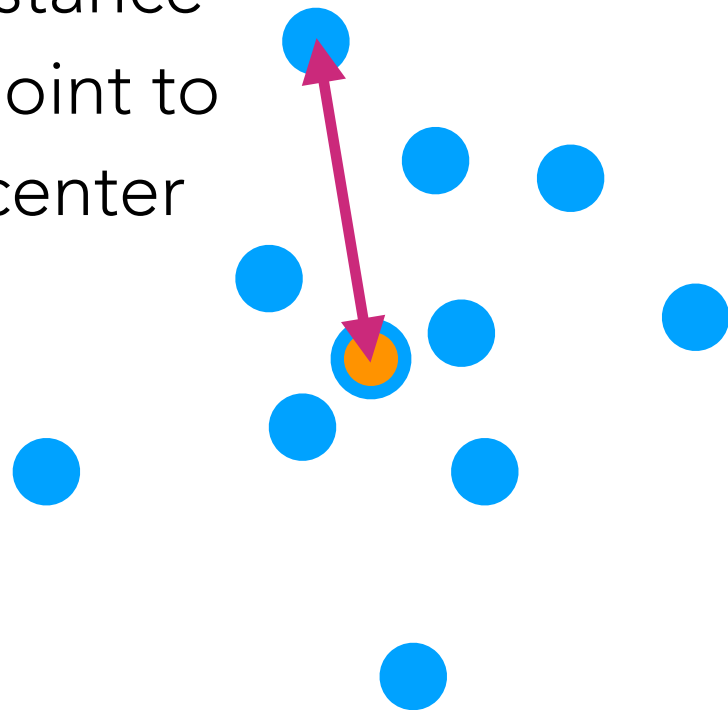


Cluster 2

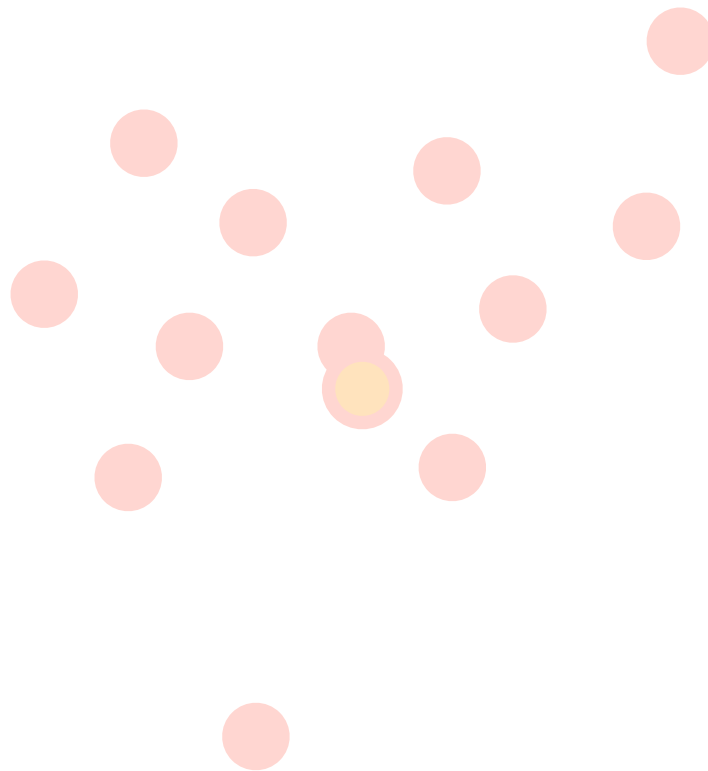
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

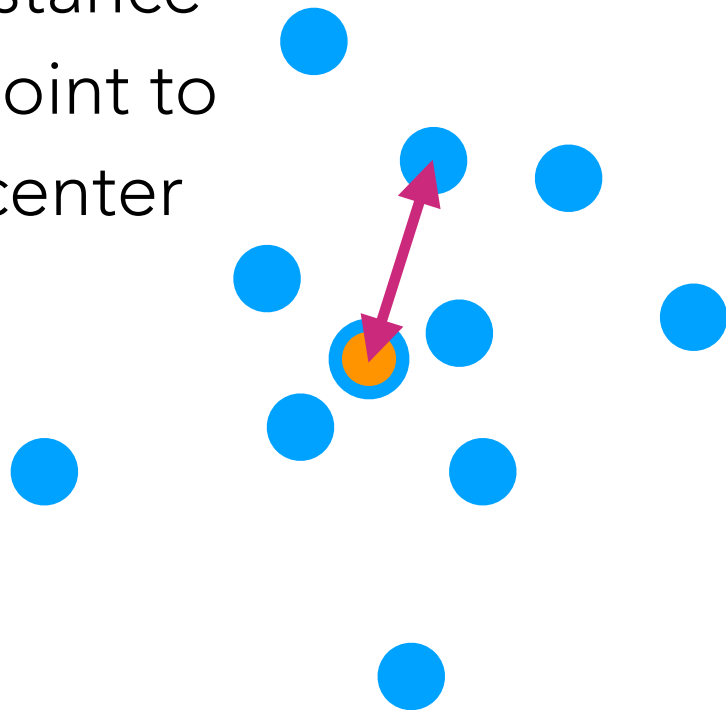


Cluster 2

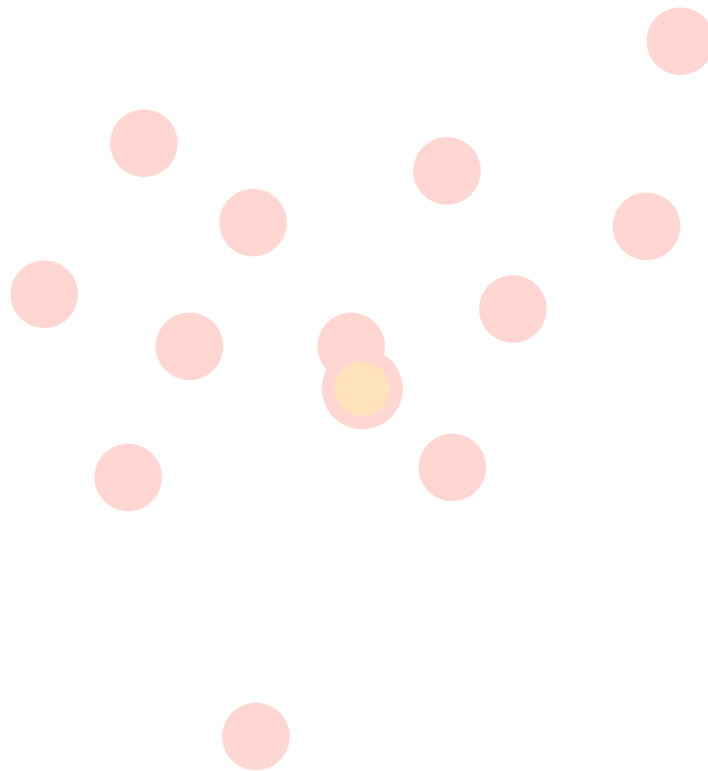
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

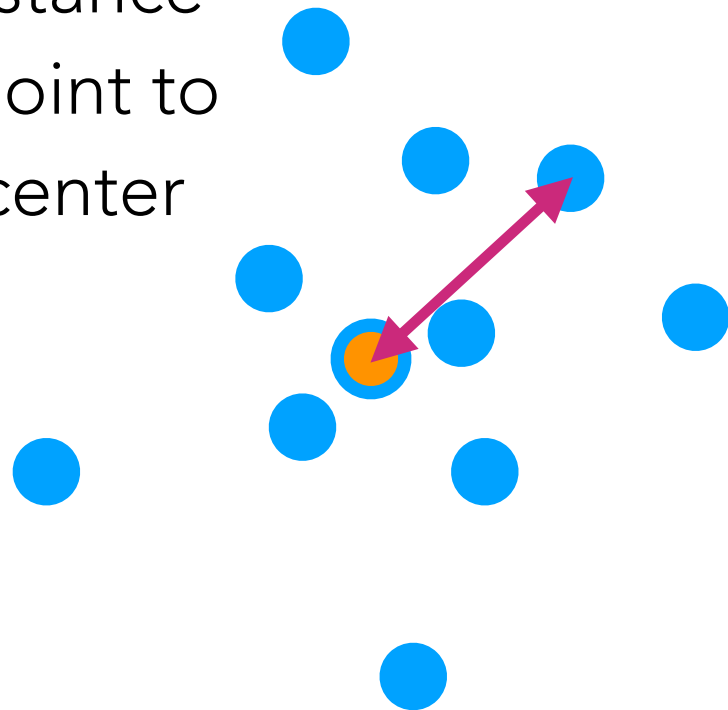


Cluster 2

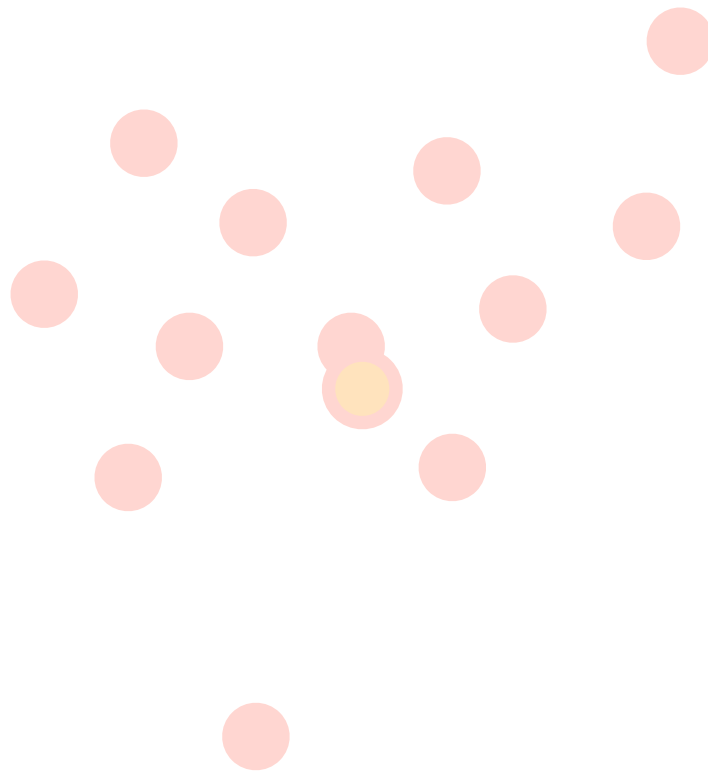
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

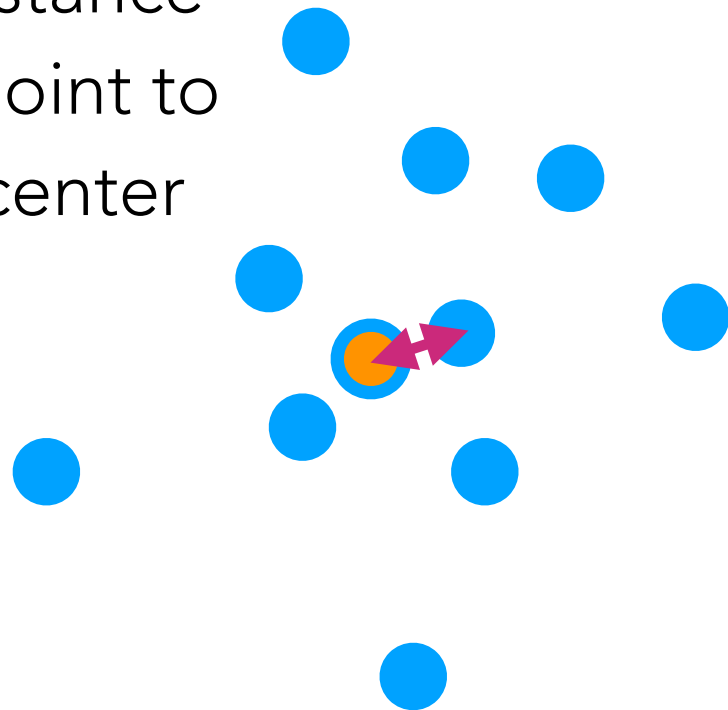


Cluster 2

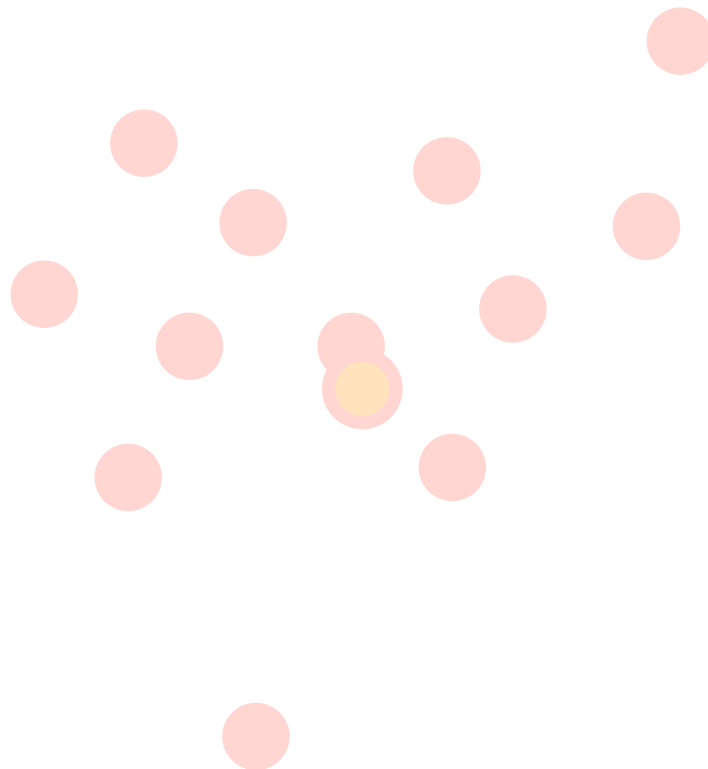
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

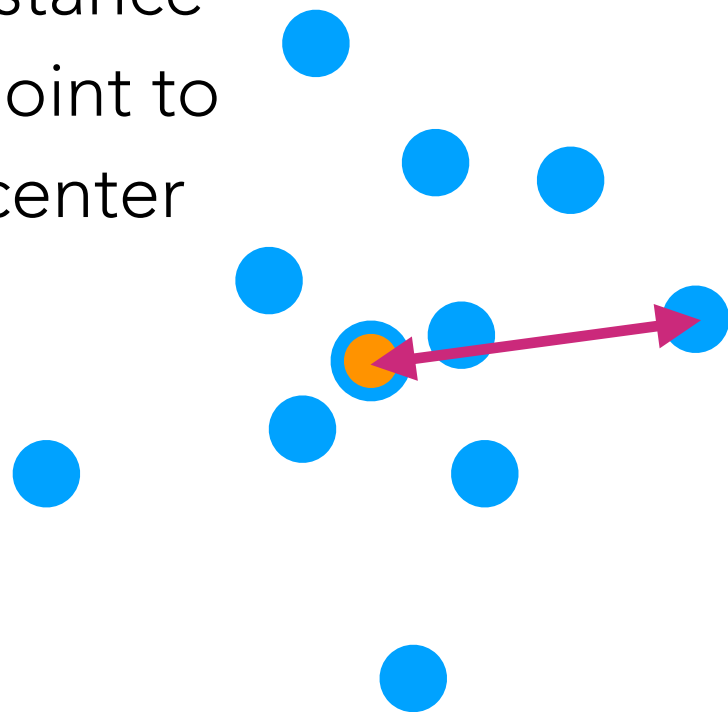


Cluster 2

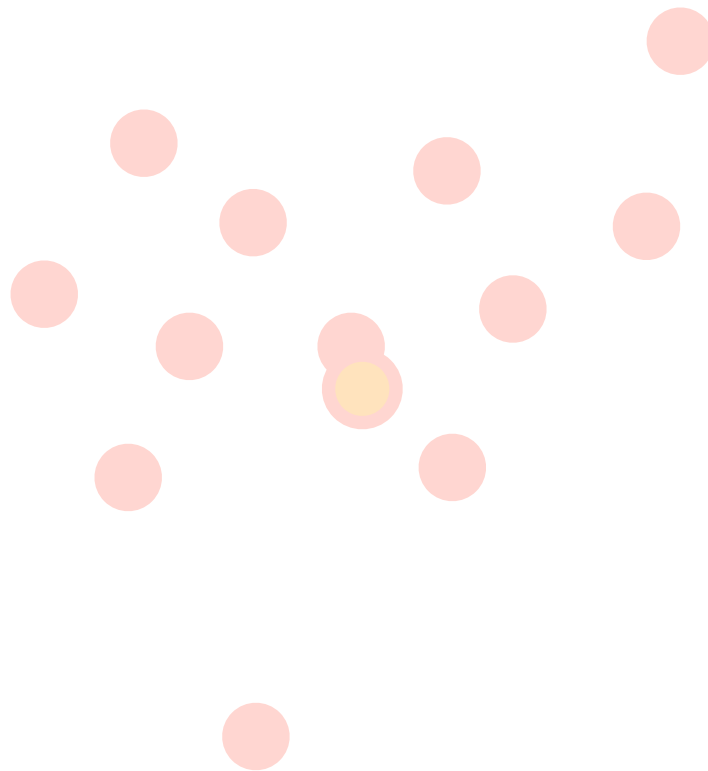
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

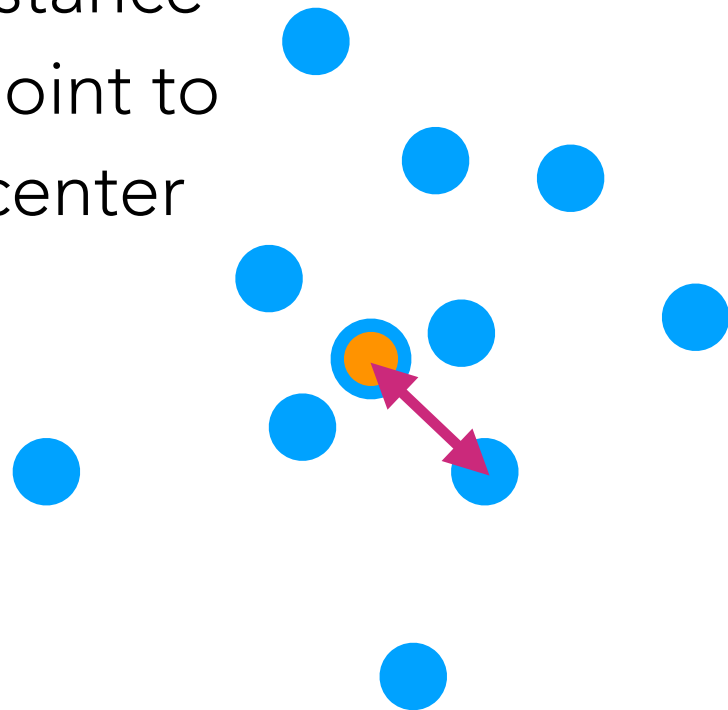


Cluster 2

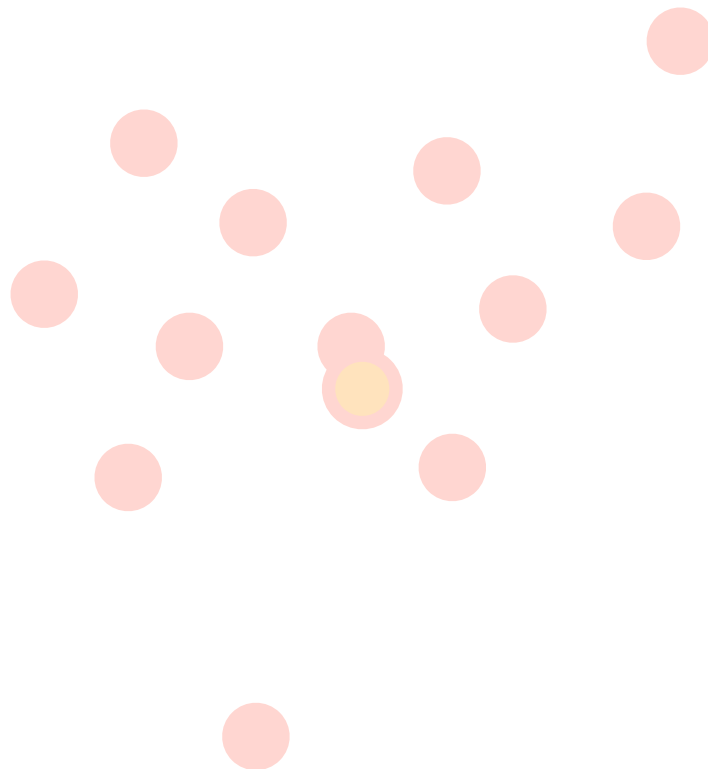
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

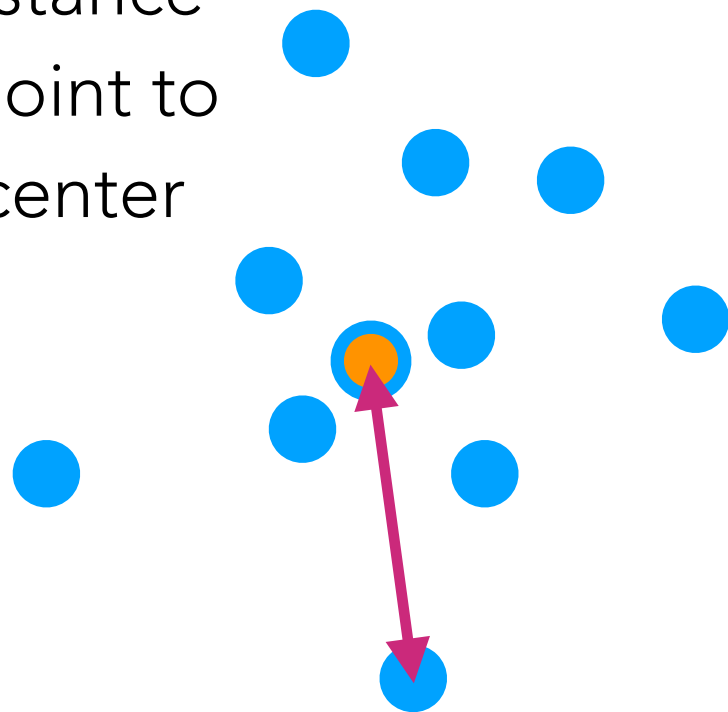


Cluster 2

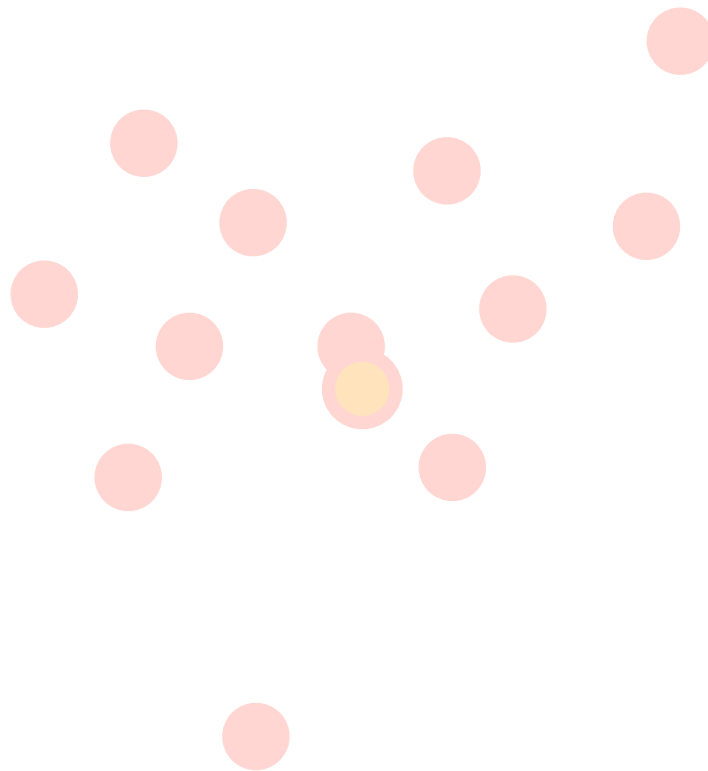
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

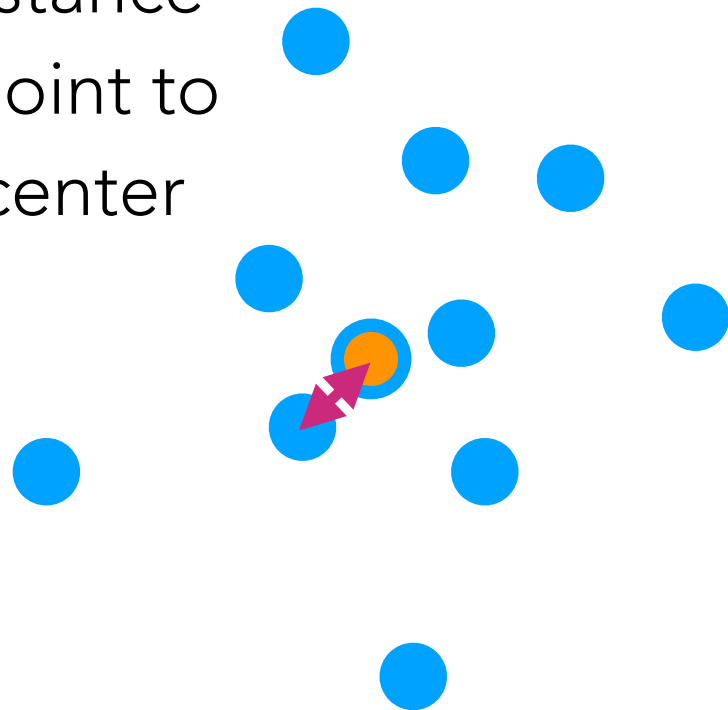


Cluster 2

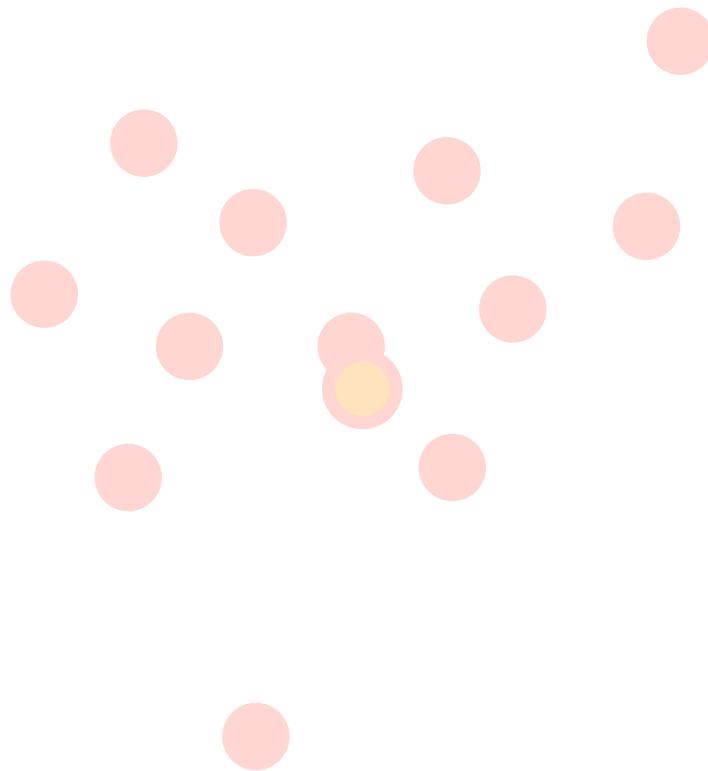
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

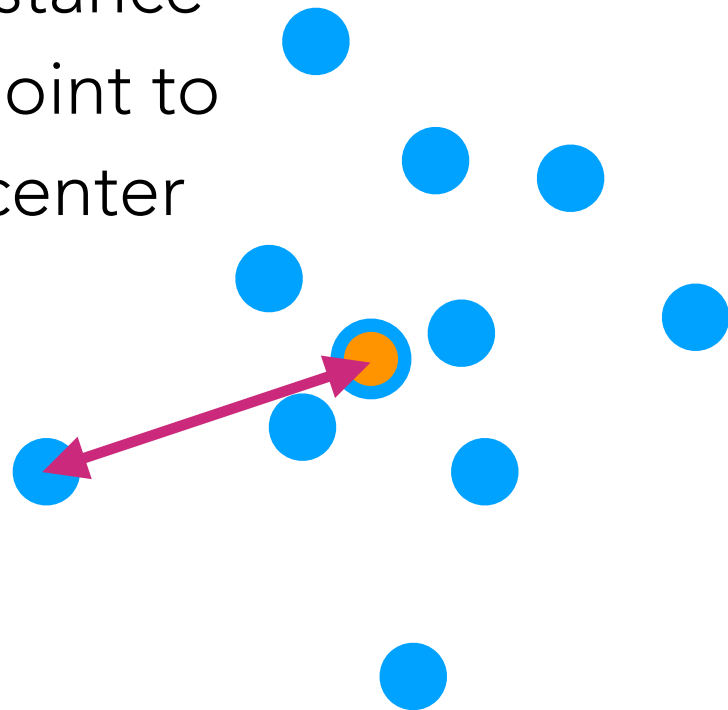


Cluster 2

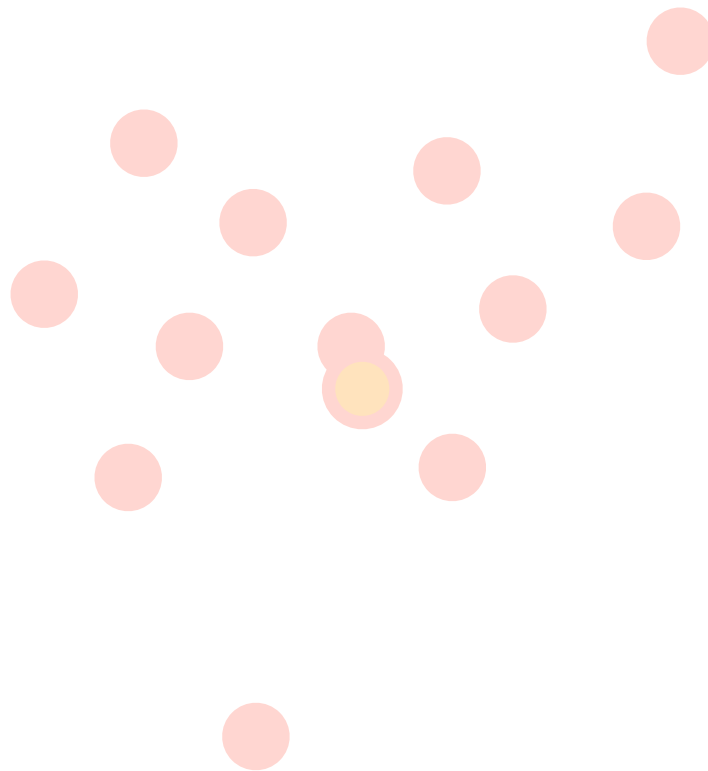
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

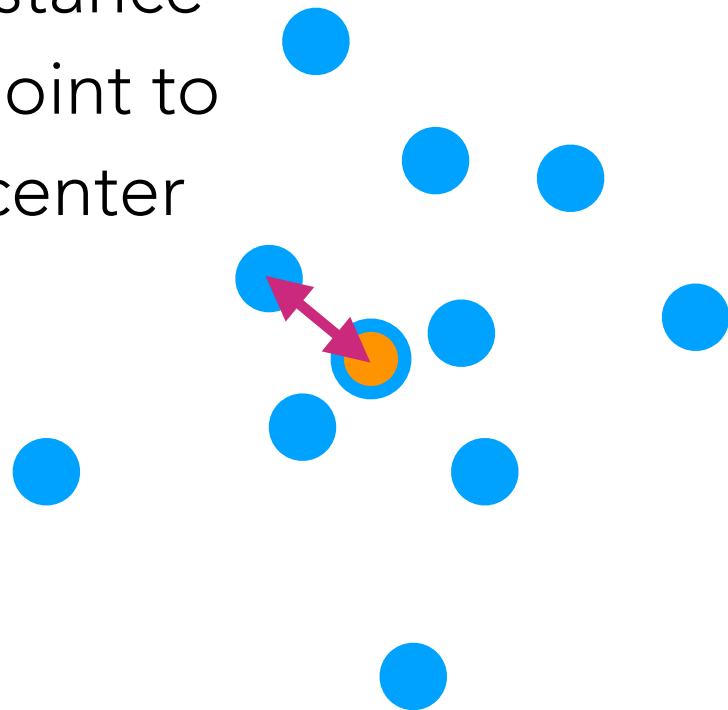


Cluster 2

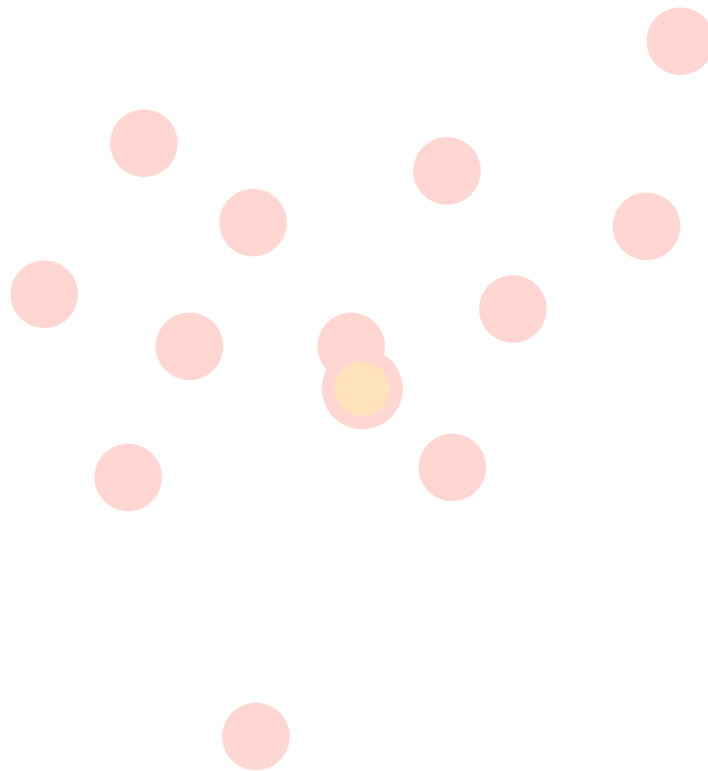
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

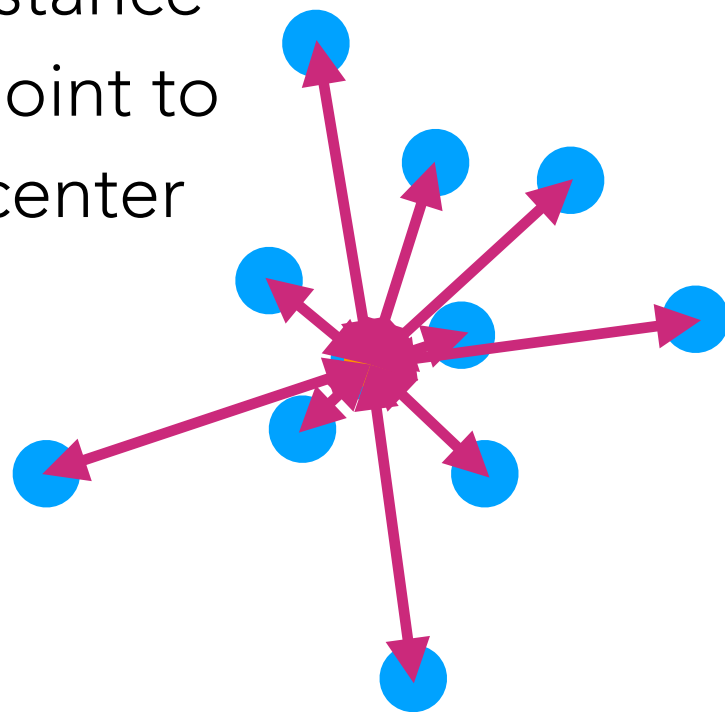


Cluster 2

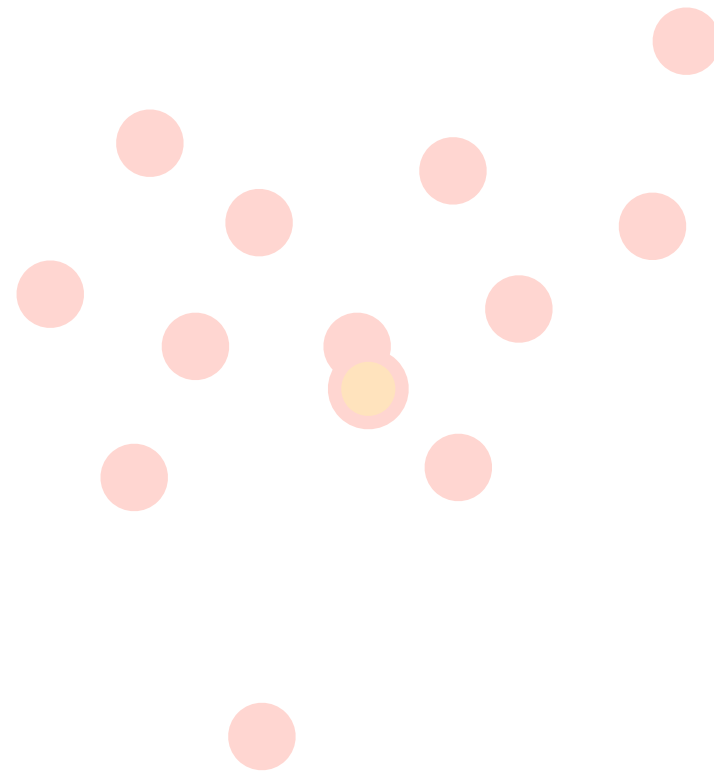
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1



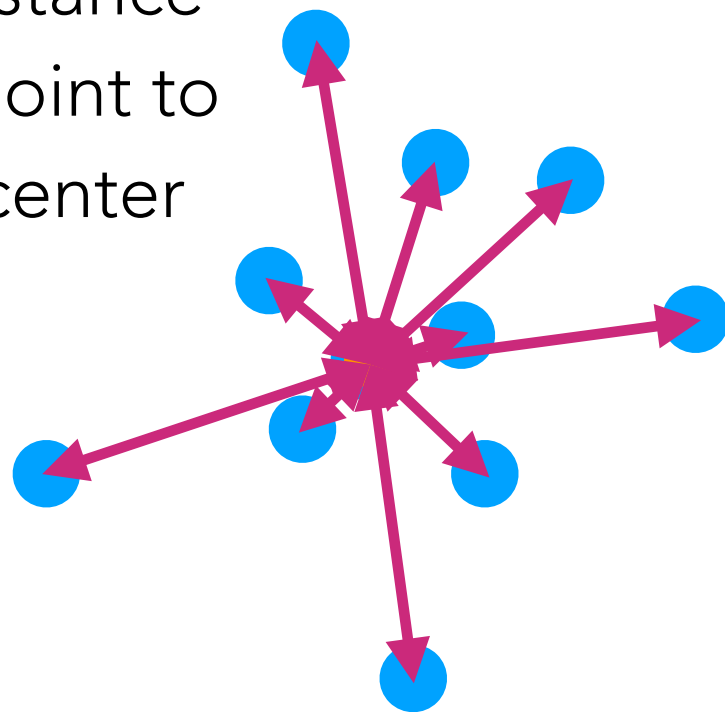
Cluster 2

Residual sum of squares for cluster 1:
sum of *squared* purple lengths

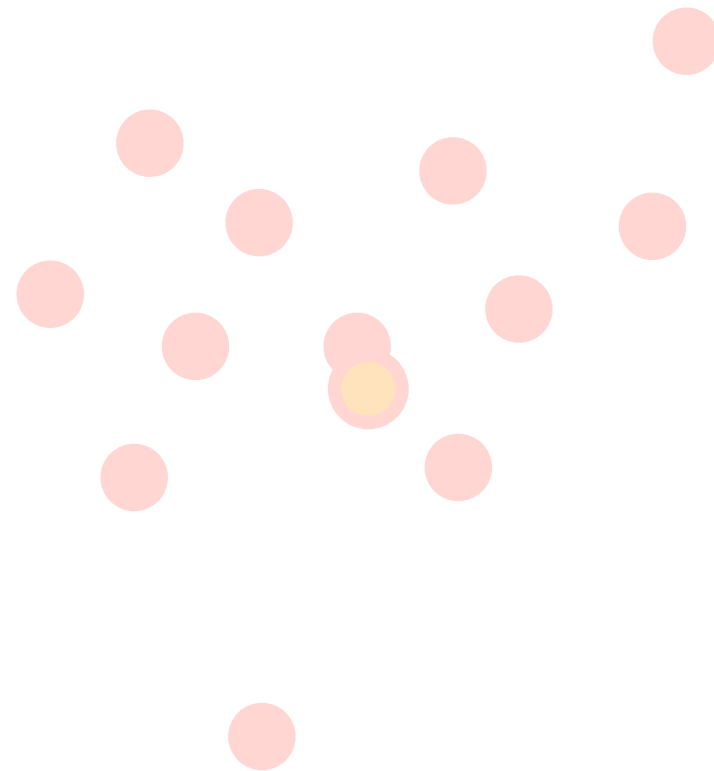
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1



Cluster 2

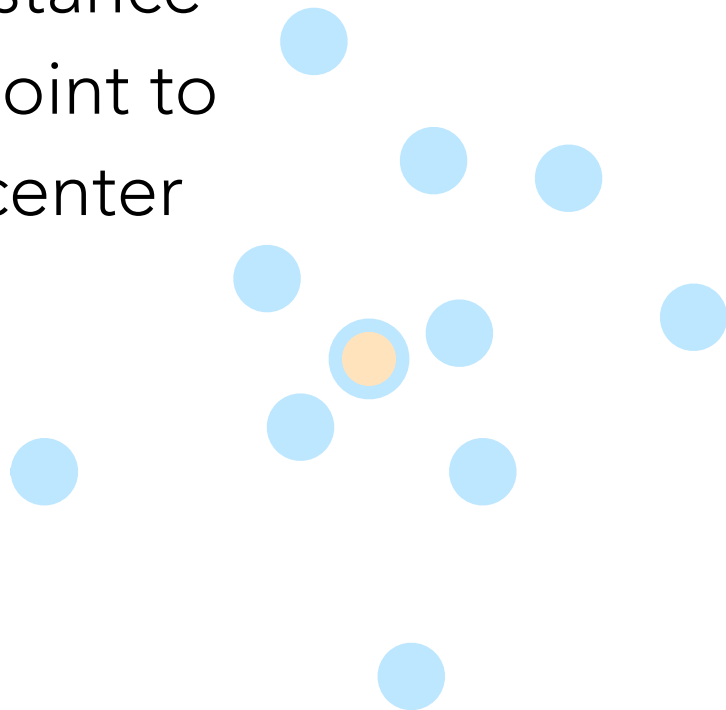
Residual sum of squares for cluster 1:

$$\text{RSS}_1 = \sum_{x \in \text{cluster 1}} \|x - \mu_1\|^2$$

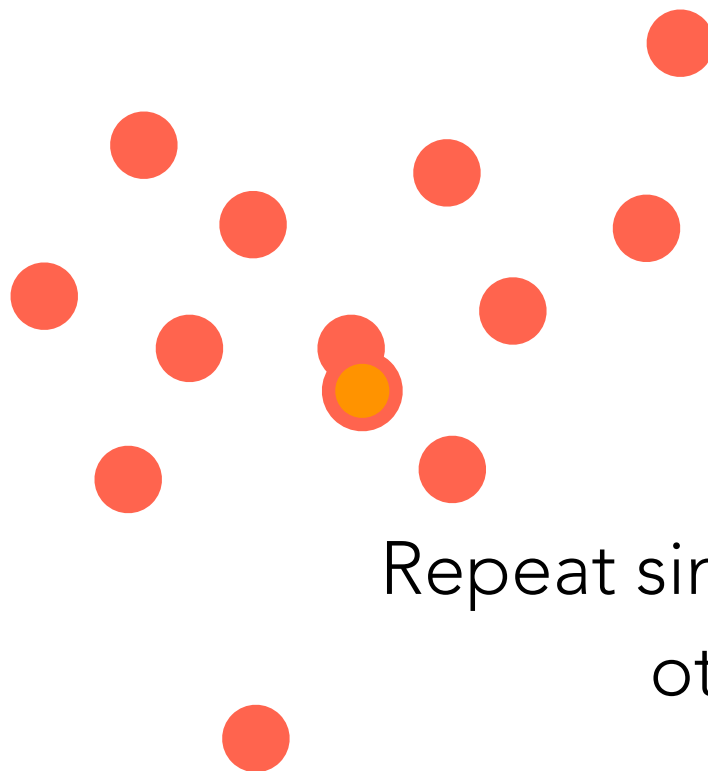
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1



Repeat similar calculation for
other cluster

Cluster 2

Residual sum of squares for cluster 2:

$$\text{RSS}_2 = \sum_{x \in \text{cluster 2}} \|x - \mu_2\|^2$$

Residual Sum of Squares

Look at one cluster at a time

$$\text{RSS} = \text{RSS}_1 + \text{RSS}_2 = \sum_{x \in \text{cluster 1}} \|x - \mu_1\|^2 + \sum_{x \in \text{cluster 2}} \|x - \mu_2\|^2$$

Measure distance
from each point to
its cluster center

In general if there are k clusters:

$$\text{RSS} = \sum_{g=1}^k \text{RSS}_g = \sum_{g=1}^k \sum_{x \in \text{cluster } g} \|x - \mu_g\|^2$$

Remark: k -means *tries* to minimize RSS for a fixed value of k
(it does so *approximately*, with no guarantee of optimality)

RSS does not account for clusters having, for instance, ellipse shapes

Why is minimizing RSS a bad way to choose k ?

What happens when k is equal to the number of data points?

A Good Way to Choose k

RSS measures *within-cluster variation*

$$W = \text{RSS} = \sum_{g=1}^k \text{RSS}_g = \sum_{g=1}^k \sum_{x \in \text{cluster } g} \|x - \mu_g\|^2$$

Want to also measure *between-cluster variation*

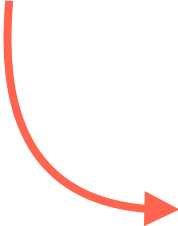
$$B = \sum_{g=1}^k (\# \text{ points in cluster } g) \|\mu_g - \mu\|^2$$

Called the CH index

[Calinski and Harabasz 1974]

mean of *all* points

A good score function to use for choosing k :


$$\text{CH}(k) = \frac{B \cdot (n - k)}{W \cdot (k - 1)}$$

n = total # points

Pick k with highest $\text{CH}(k)$

(Choose k among 2, 3, ... up to pre-specified max)

Automatically Choosing the Number of Clusters k

Demo

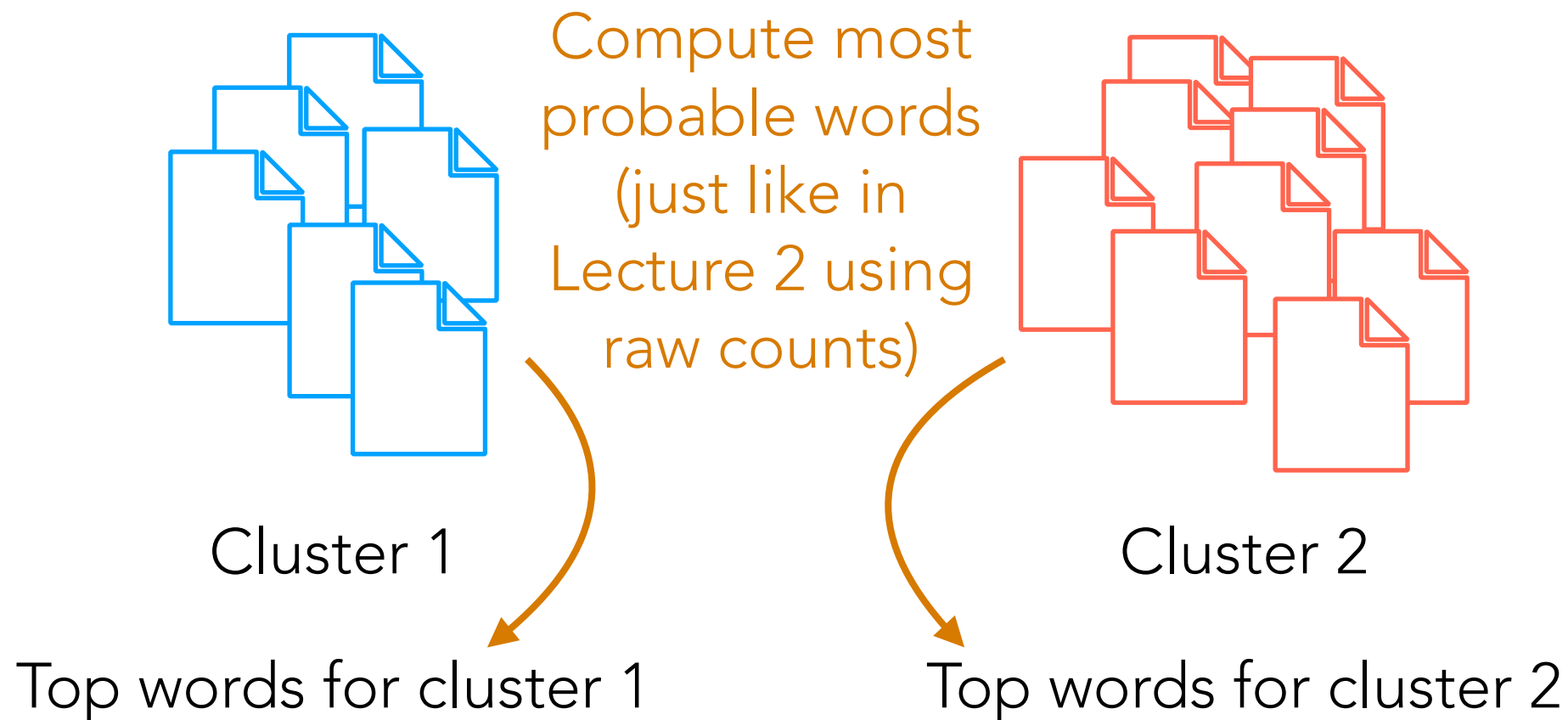
What about unstructured data?

Clustering on Images

Demo

Clustering on Text

Basic visualization strategy



⇒ We can then compare top words across clusters

Per cluster, can compute other information aside from top words (such as the distribution of document lengths per cluster, or some co-occurrence information per cluster—an example of these ideas is in the Spring 2023 Quiz 1)

Last Remarks on Clustering

- We only saw two clustering methods (k -means, GMM)
- We only saw one general strategy to automatically choose # of clusters
 - You must specify a score function — no score function is perfect
- There are *lots* of clustering methods out there!
 - Many do not require specifying # of clusters (DP-means, DP-GMM, many variants of hierarchical clustering, DBSCAN, OPTICS, ...)
- Ultimately, you have to decide on which clustering method and number of clusters make sense for your data
 - After you run a clustering algorithm, make visualizations to interpret the clusters *in the context of your application!*
 - Do not just blindly rely on numerical metrics (e.g., CH index)