

95-865 Pittsburgh Lecture 6: Clustering Part II

George Chen

Learning a GMM

Demo

Automatically Choosing k

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

Fit model using k

Compute a score for the model

But what score function should we use?

Use whichever k has the best score

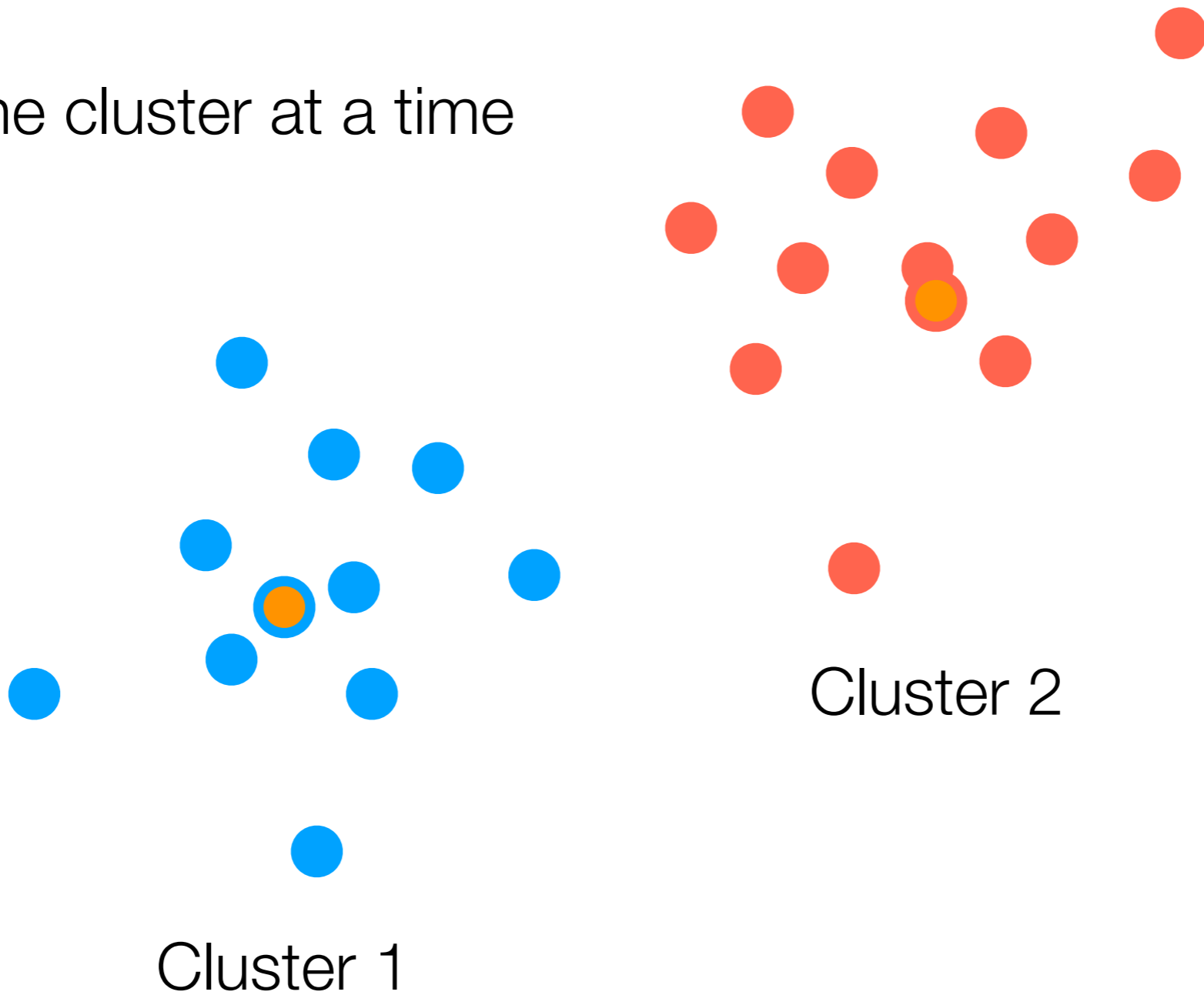
No single way of choosing k is the “best” way

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

But hey it's worth a shot

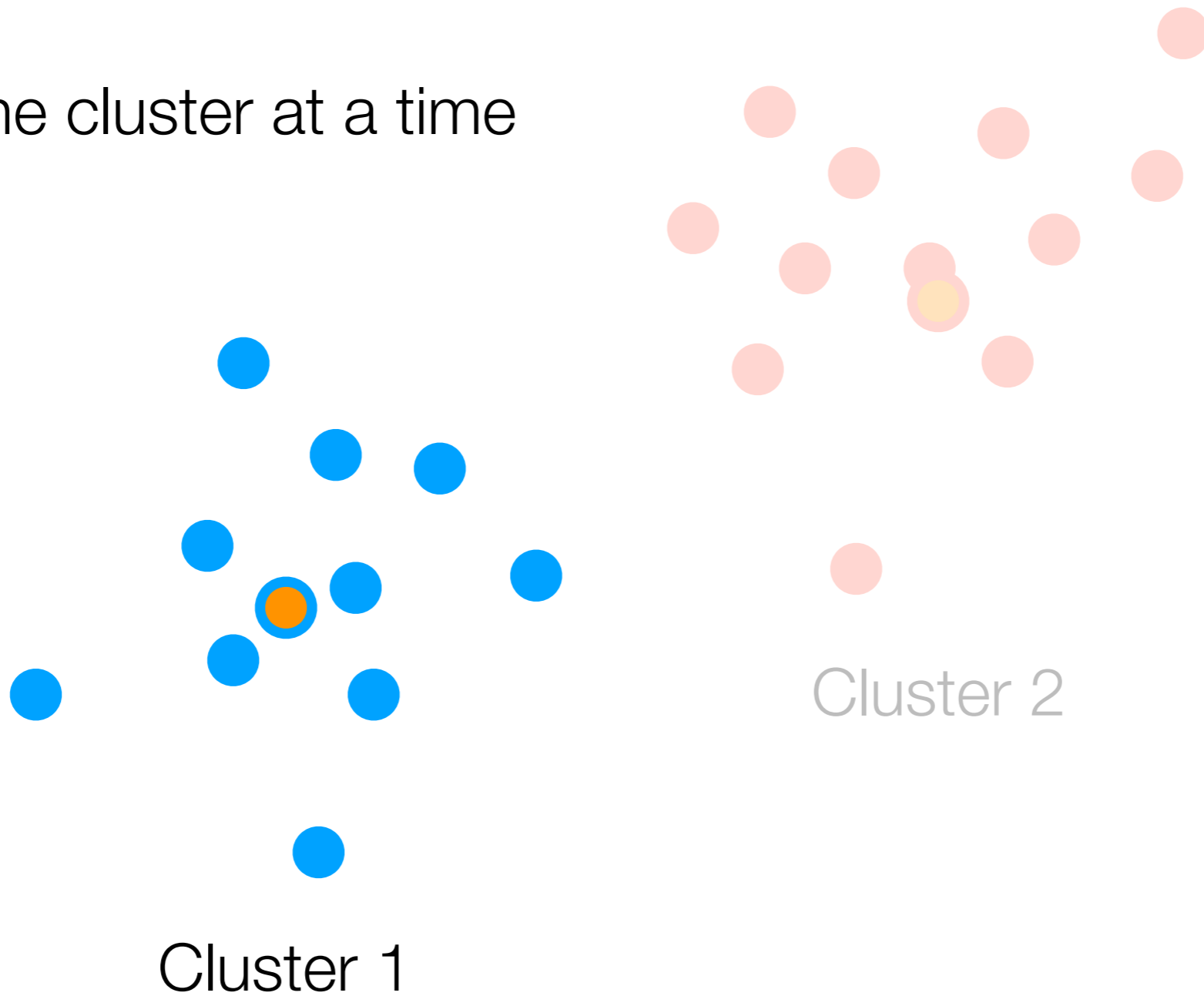
Residual Sum of Squares

Look at one cluster at a time



Residual Sum of Squares

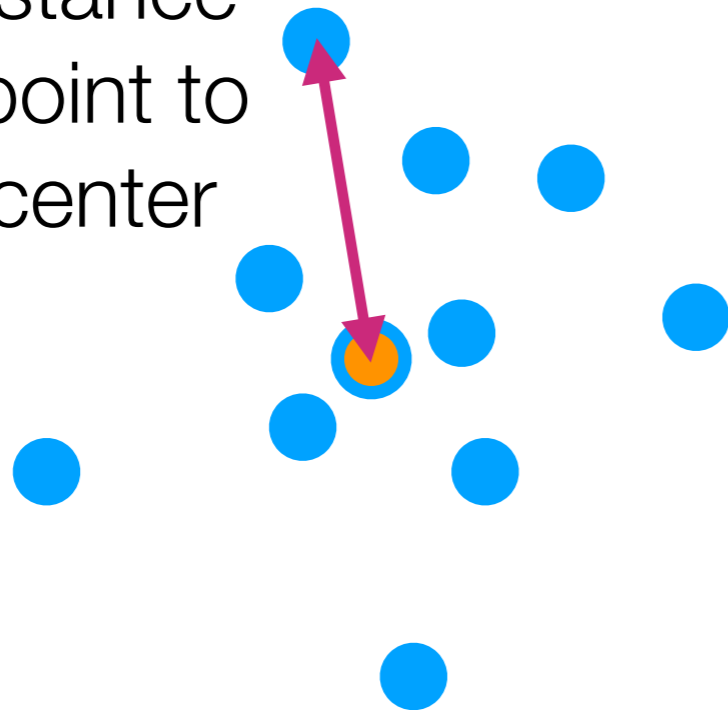
Look at one cluster at a time



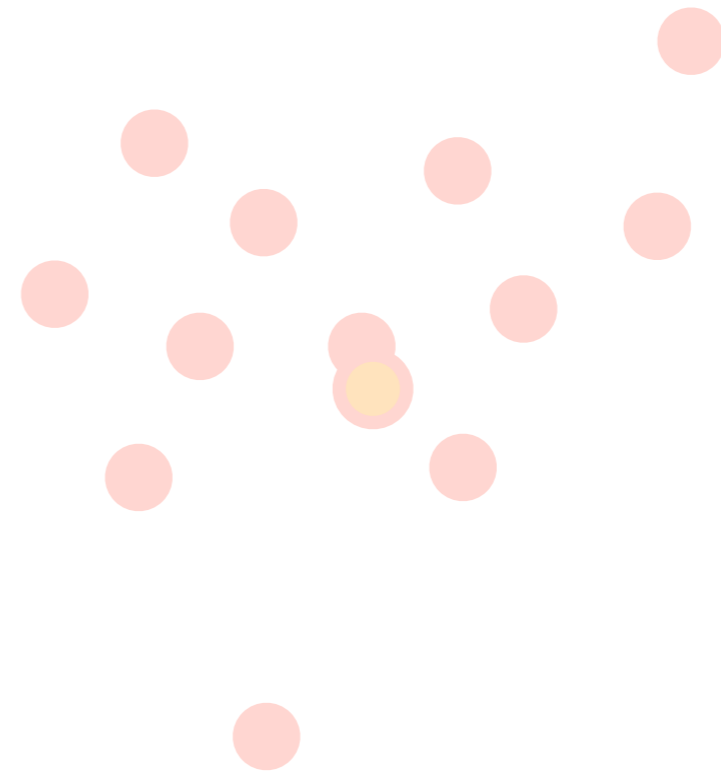
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

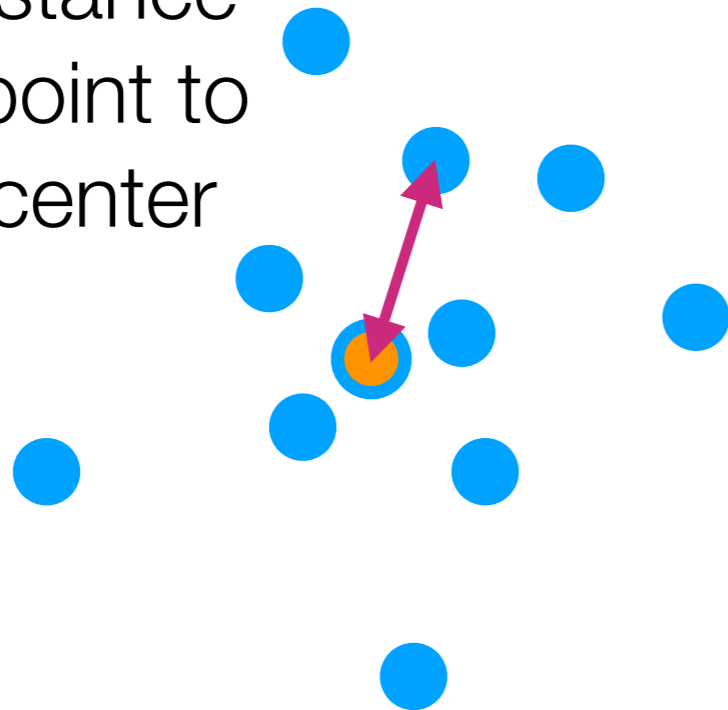


Cluster 2

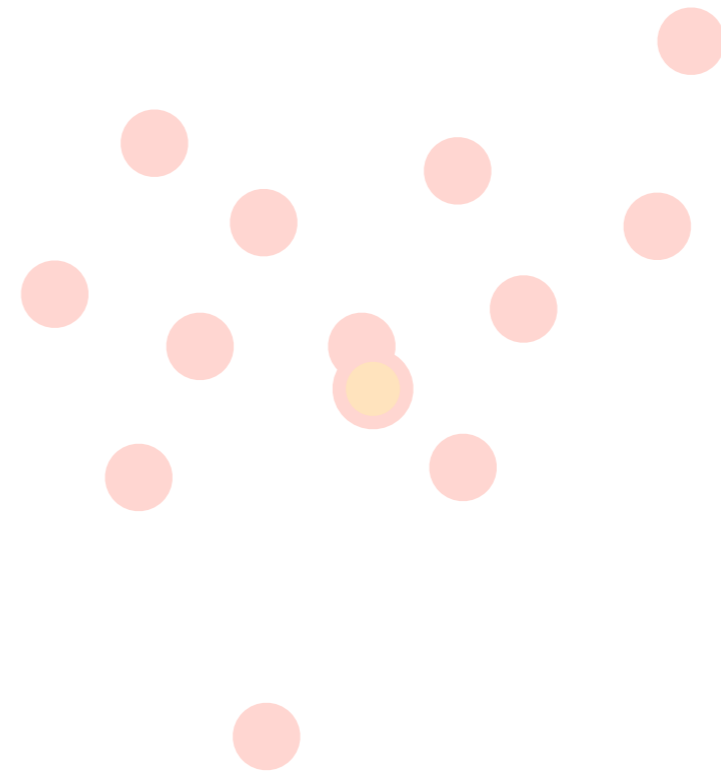
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

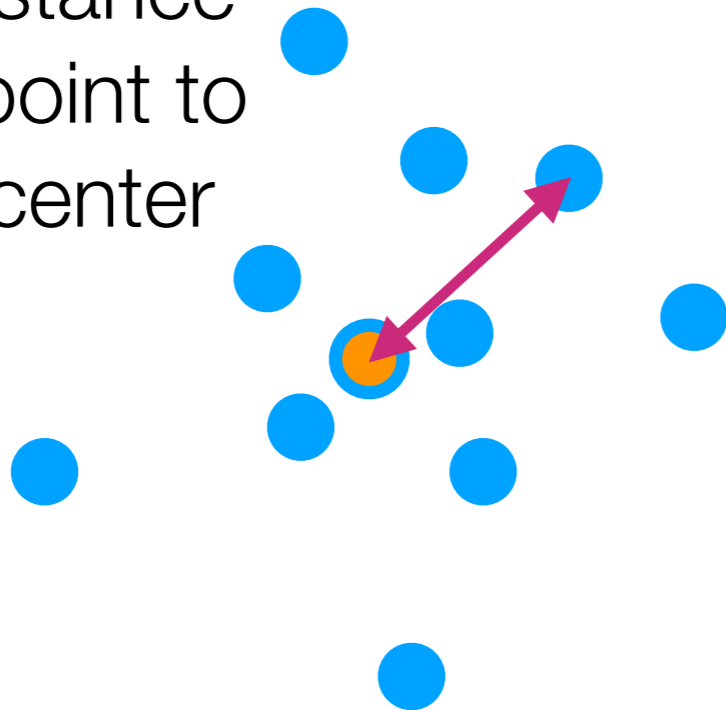


Cluster 2

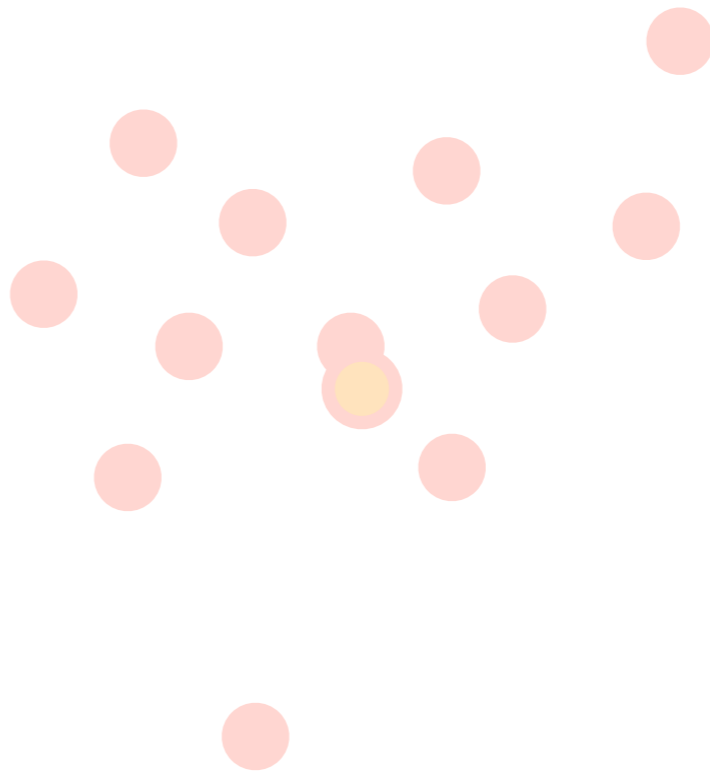
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

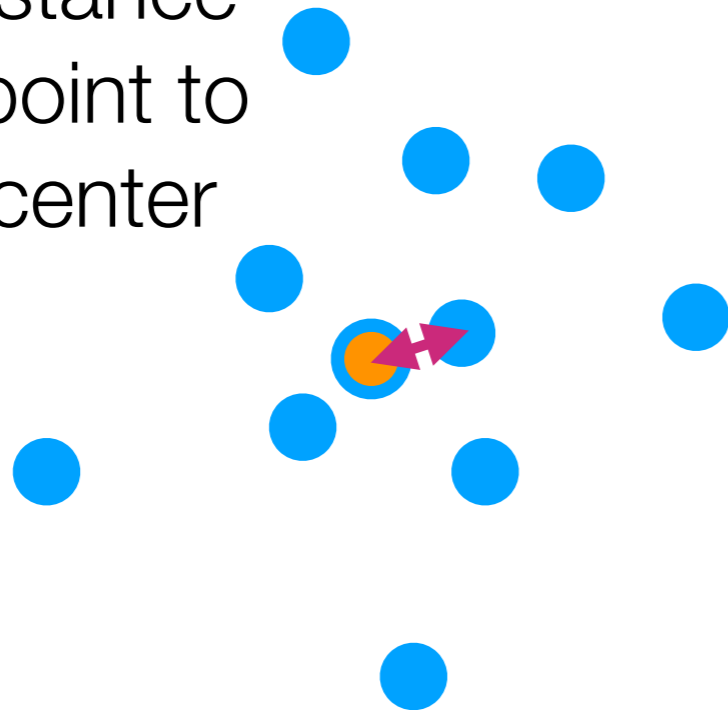


Cluster 2

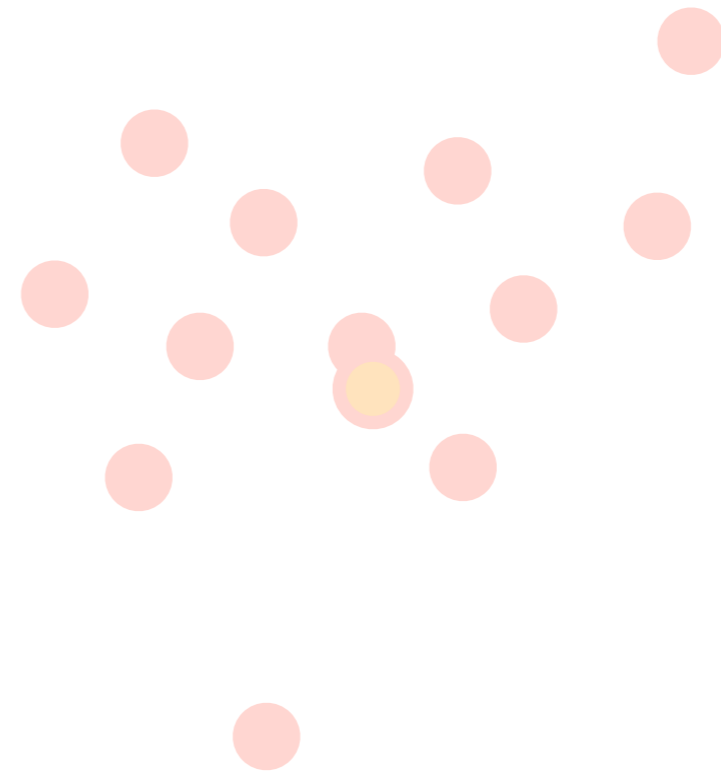
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

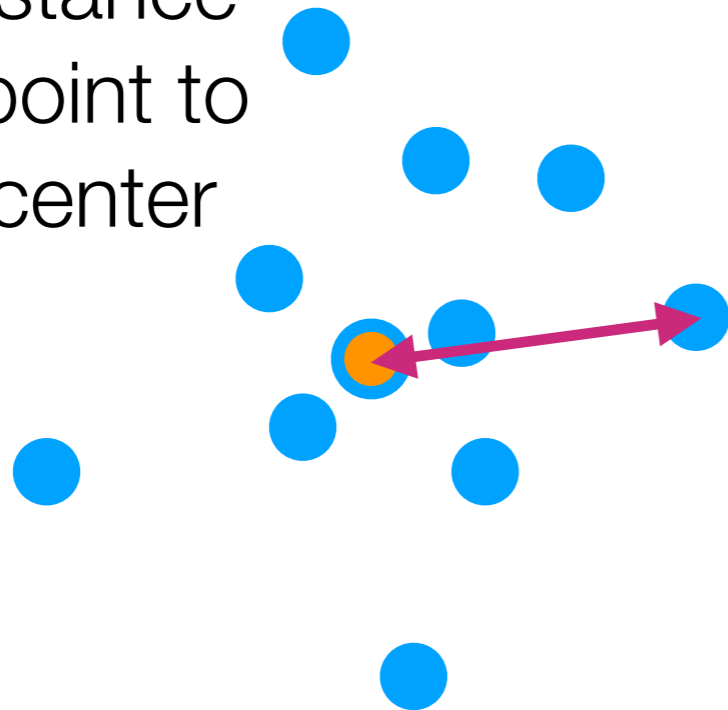


Cluster 2

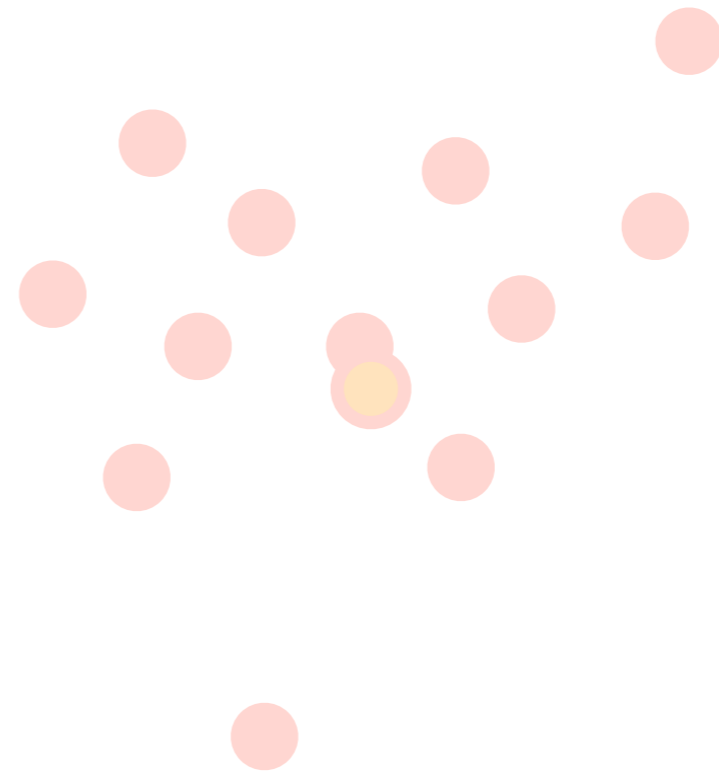
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

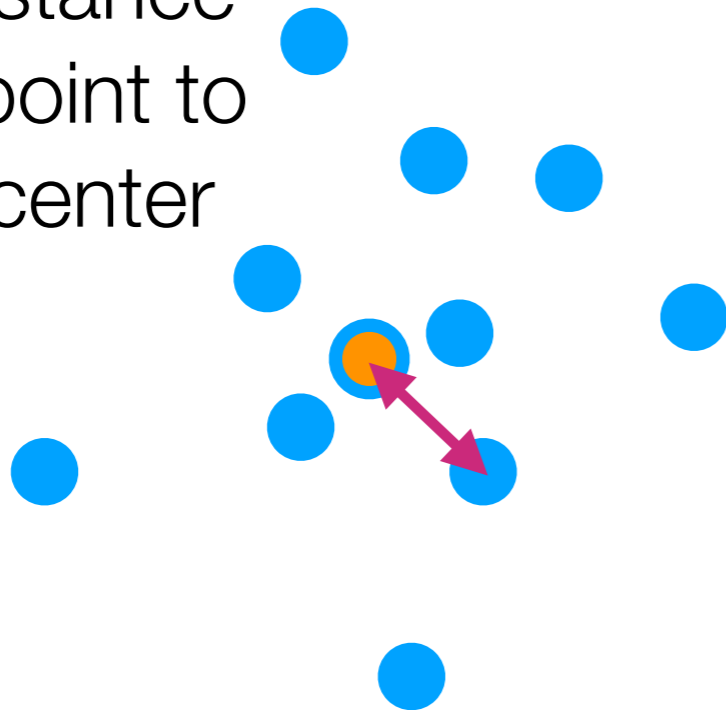


Cluster 2

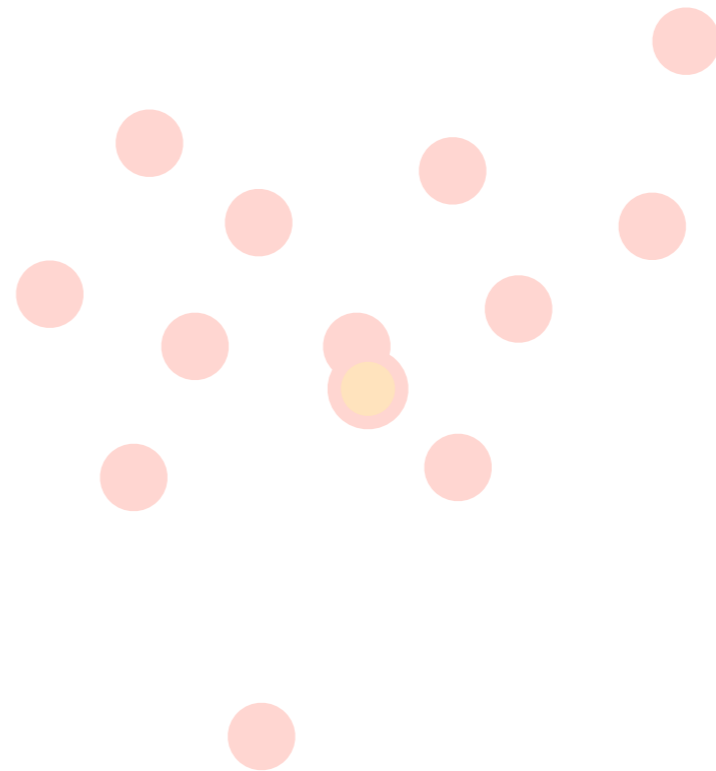
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

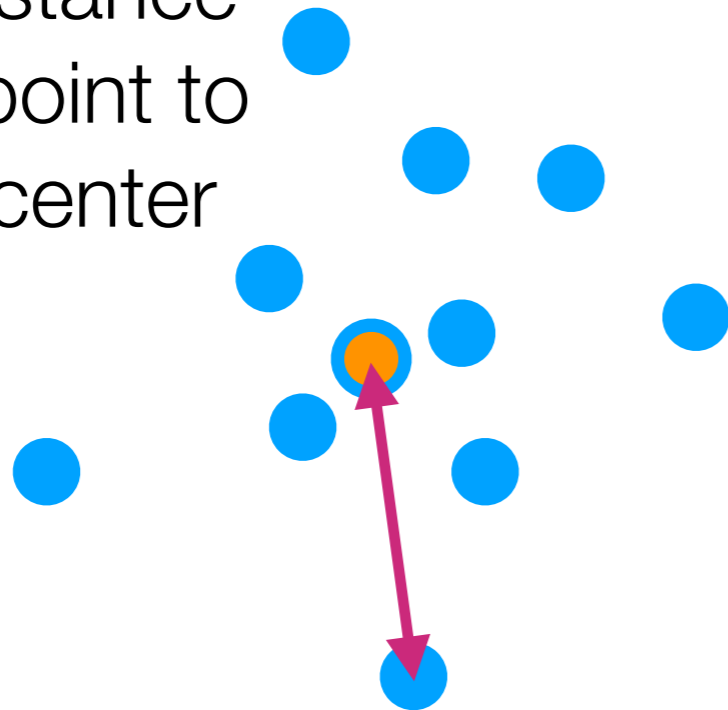


Cluster 2

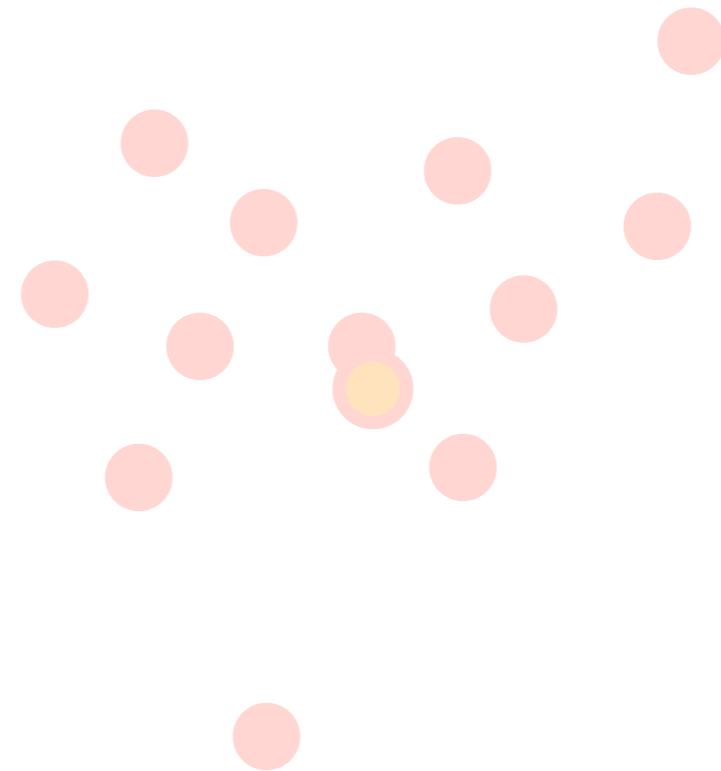
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

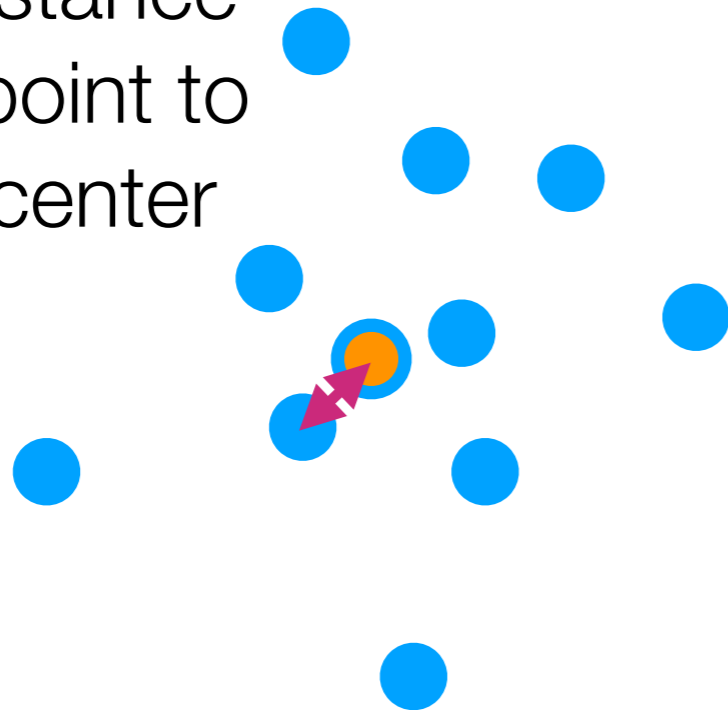


Cluster 2

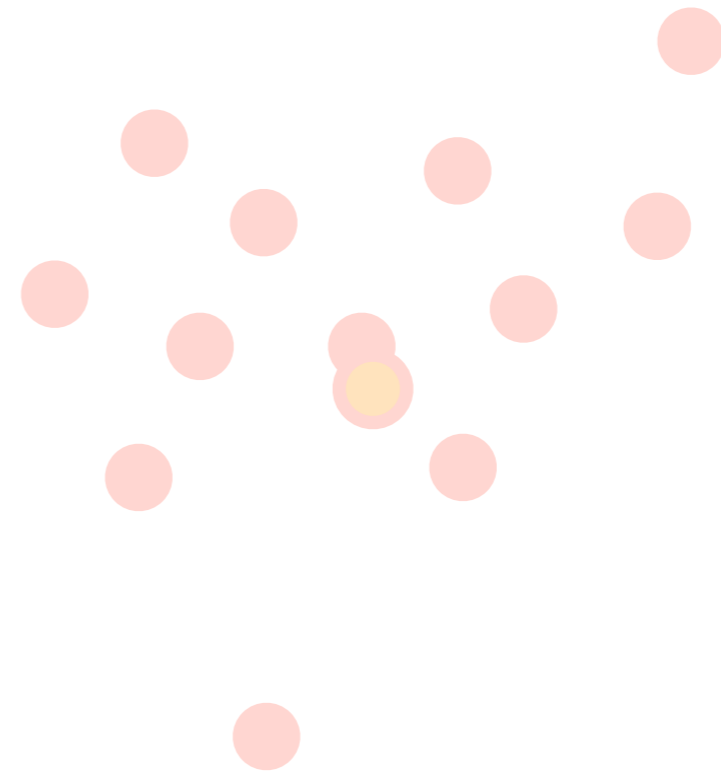
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

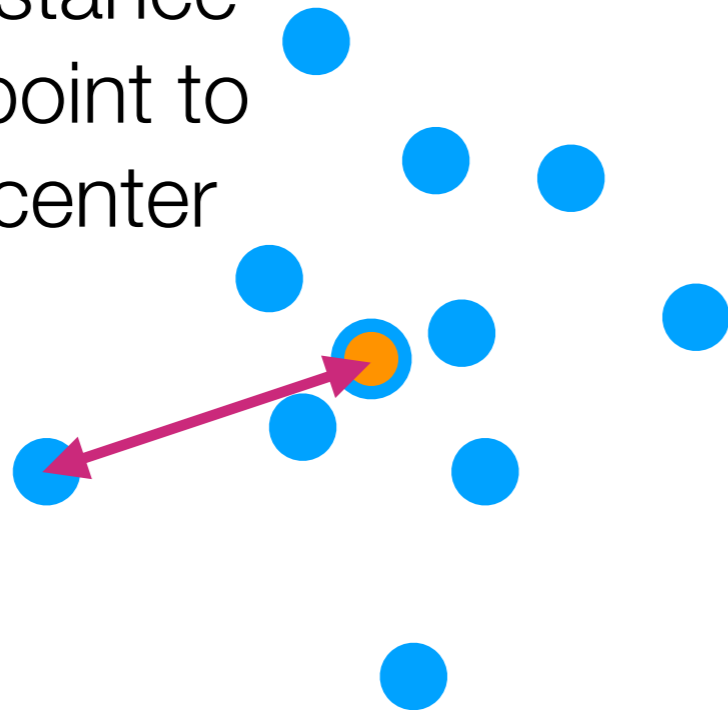


Cluster 2

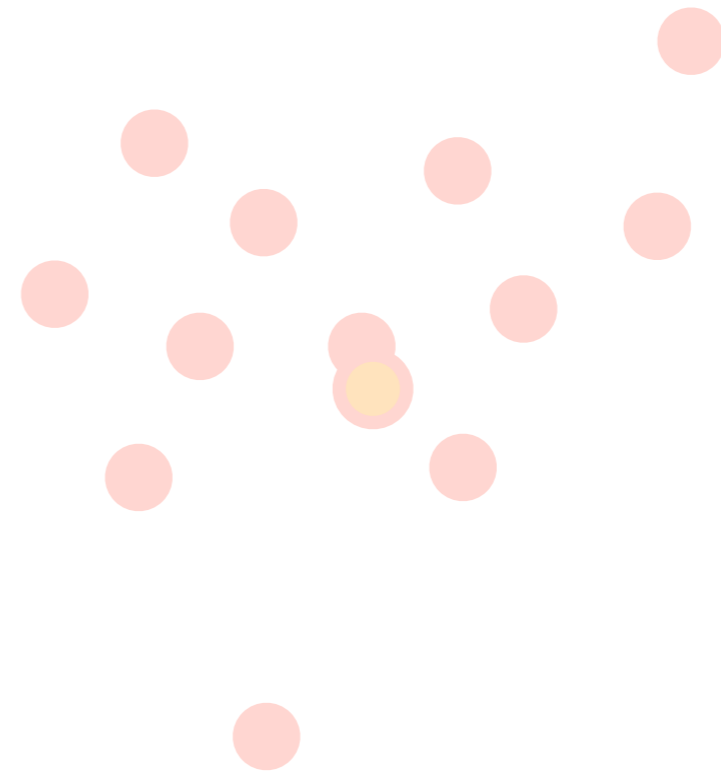
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

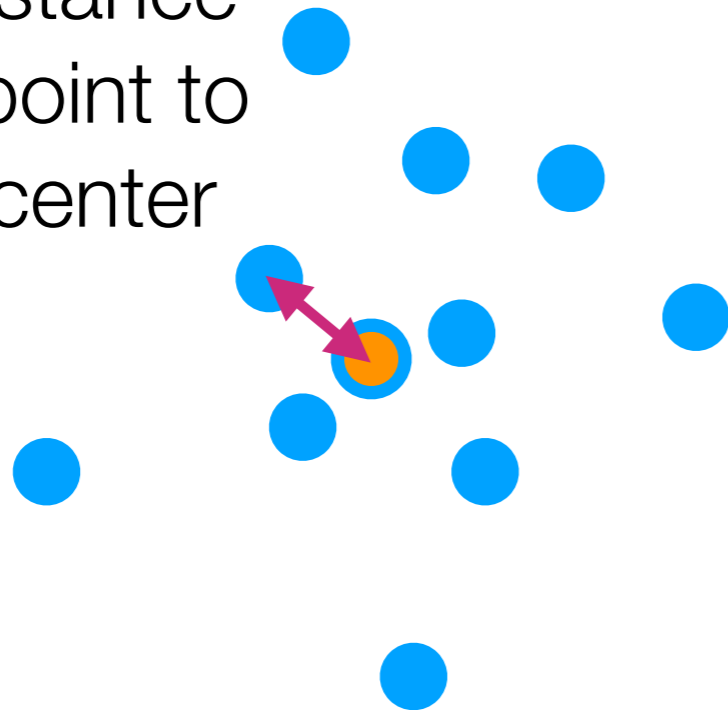


Cluster 2

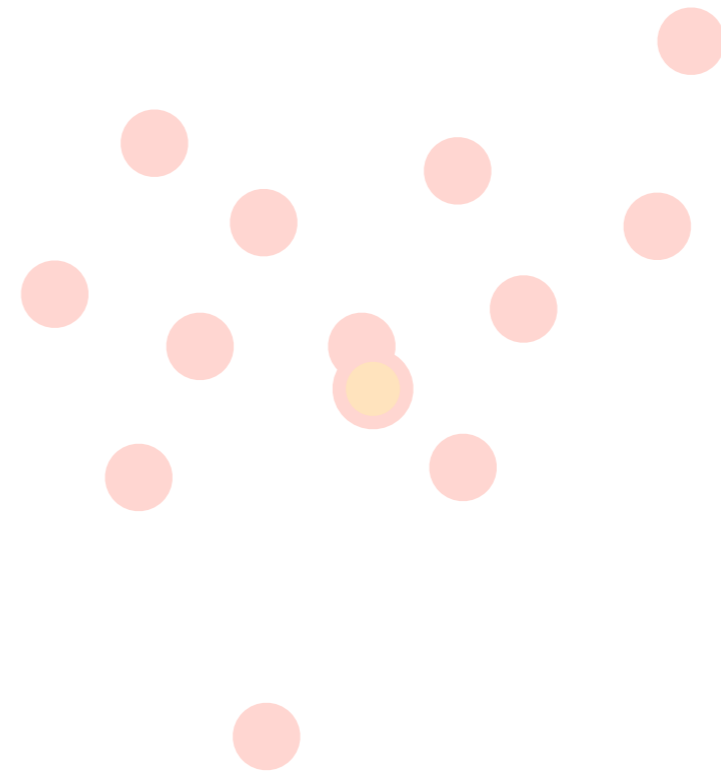
Residual Sum of Squares

Look at one cluster at a time

Measure distance
from each point to
its cluster center



Cluster 1

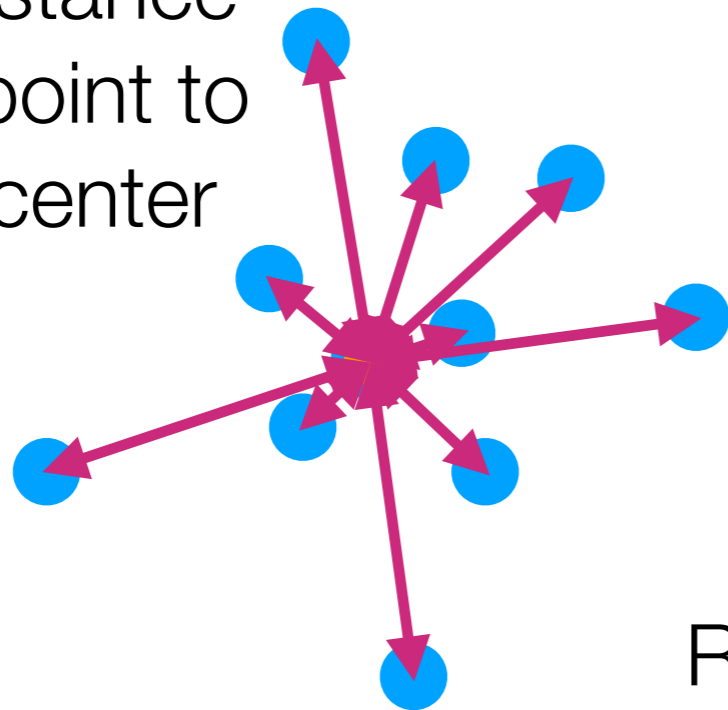


Cluster 2

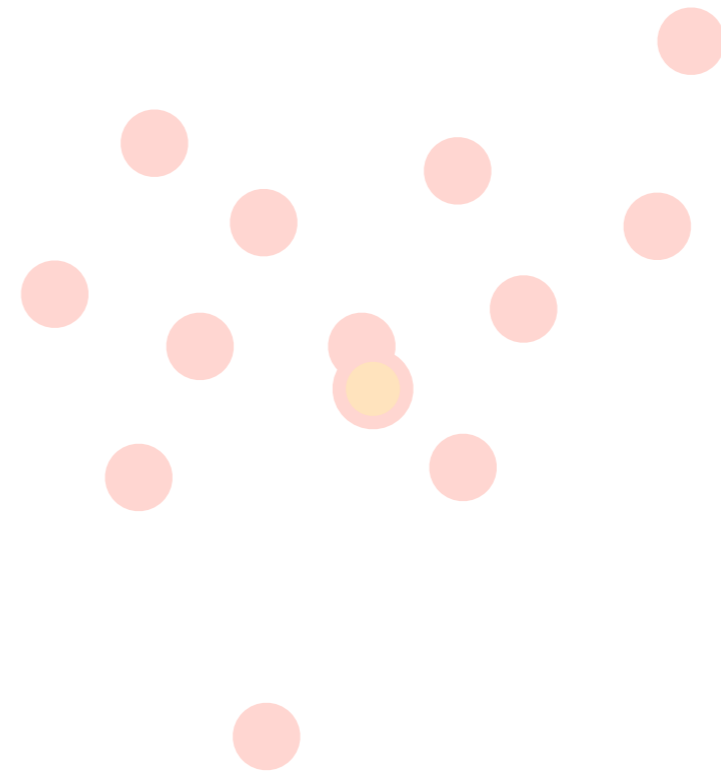
Residual Sum of Squares

Look at one cluster at a time

Measure distance from each point to its cluster center



Cluster 1



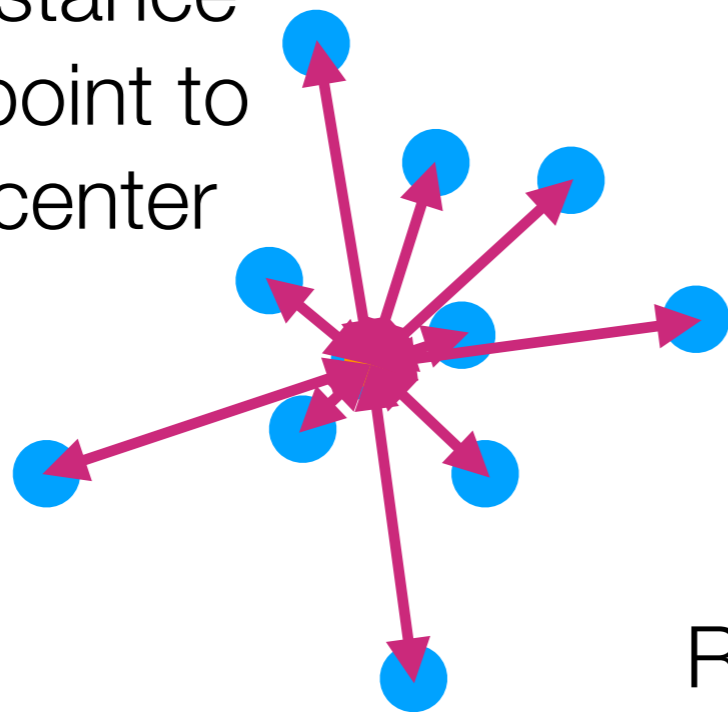
Cluster 2

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

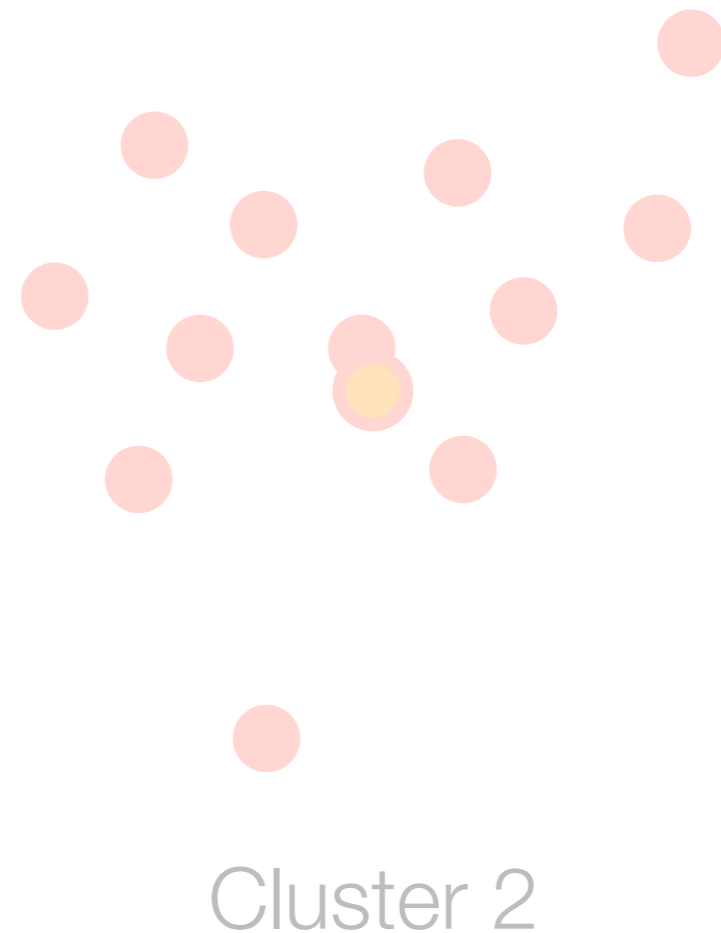
Residual Sum of Squares

Look at one cluster at a time

Measure distance from each point to its cluster center



Cluster 1



Residual sum of squares for cluster 1:

$$RSS_1 = \sum_{x \in \text{cluster 1}} \|x - \mu_1\|^2$$

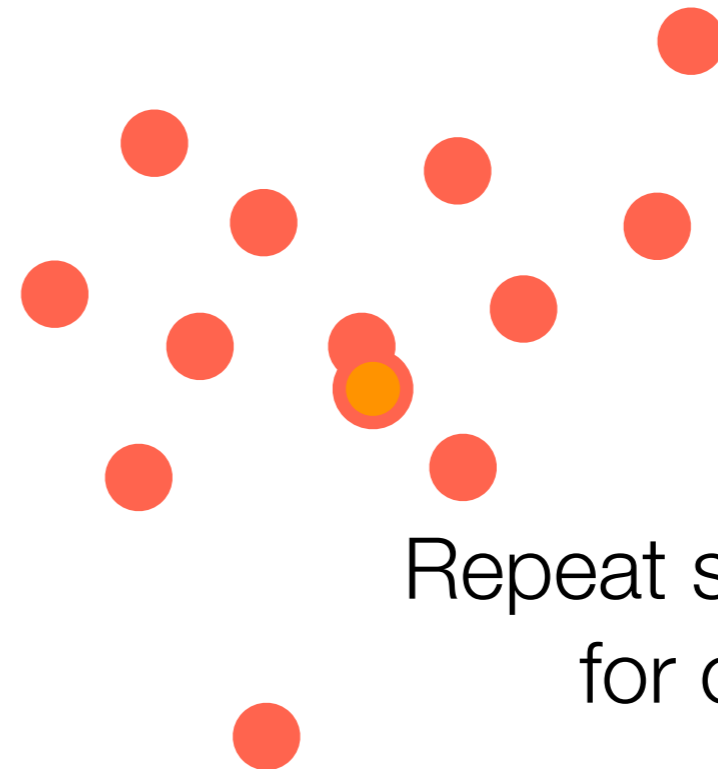
Residual Sum of Squares

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:

$$RSS_2 = \sum_{x \in \text{cluster 2}} \|x - \mu_2\|^2$$

Residual Sum of Squares

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

Measure distance
from each point to
its cluster center

In general if there are k clusters:

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

repeat similar calculation
for other cluster

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

Cluster 1

RSS only really makes sense for clusters that look like circles

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**

mean of *all* points

[Calinski and Harabasz 1974]

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 k

Demo