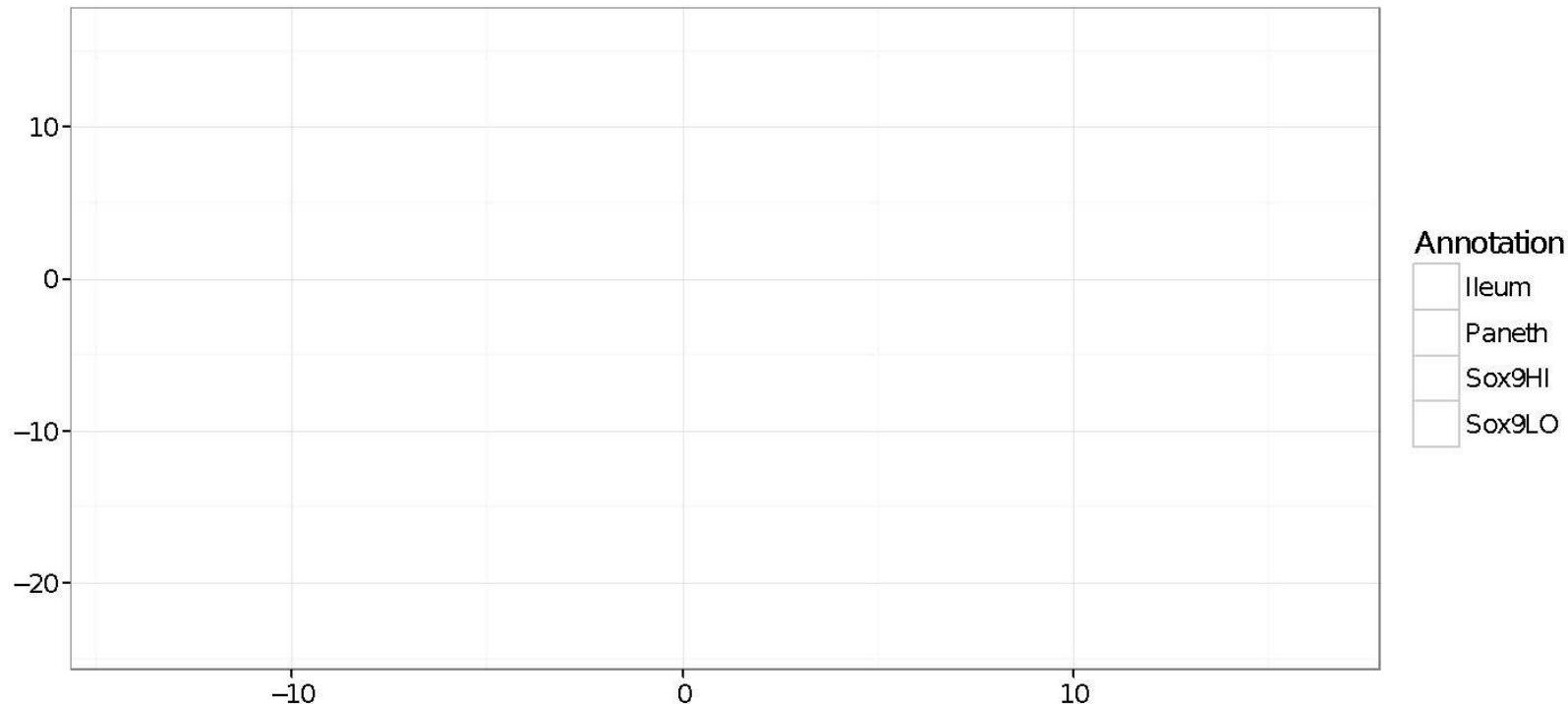
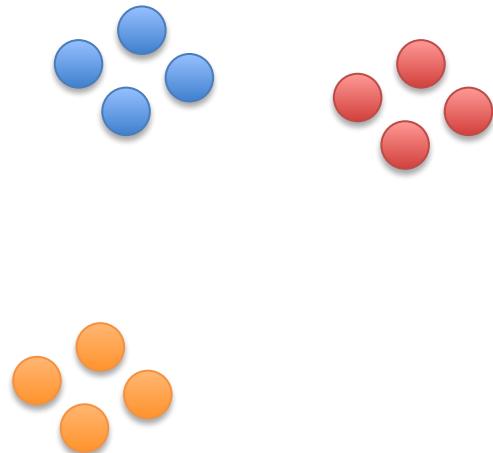


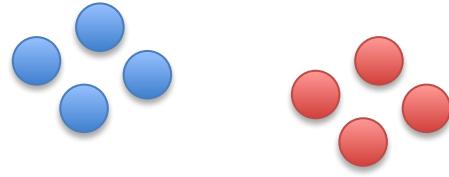
# t-SNE Explained



Here's a basic 2-D scatter plot.



Here's a basic 2-D scatter plot.



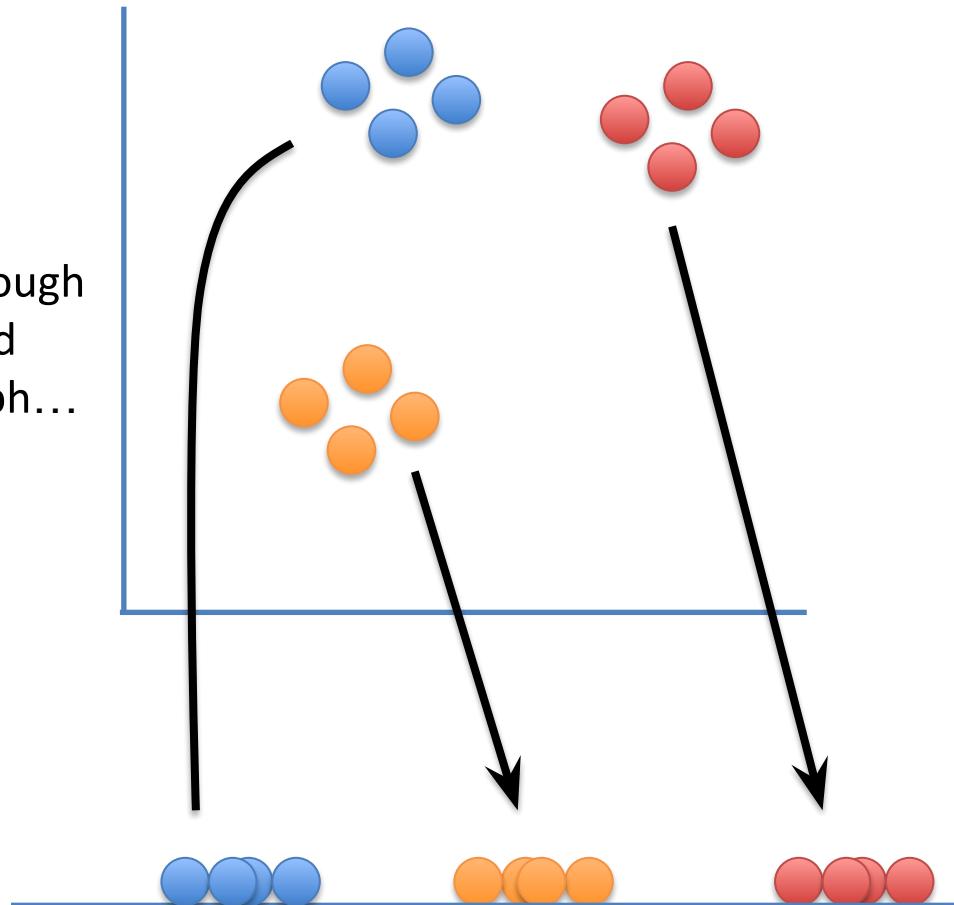
Let's do a walk through of how t-SNE would transform this graph...

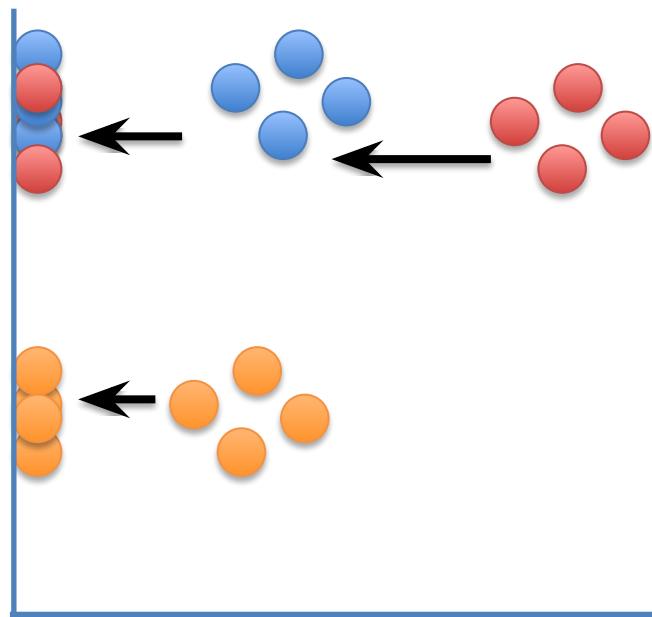


Here's a basic 2-D scatter plot.

Let's do a walk through of how t-SNE would transform this graph...

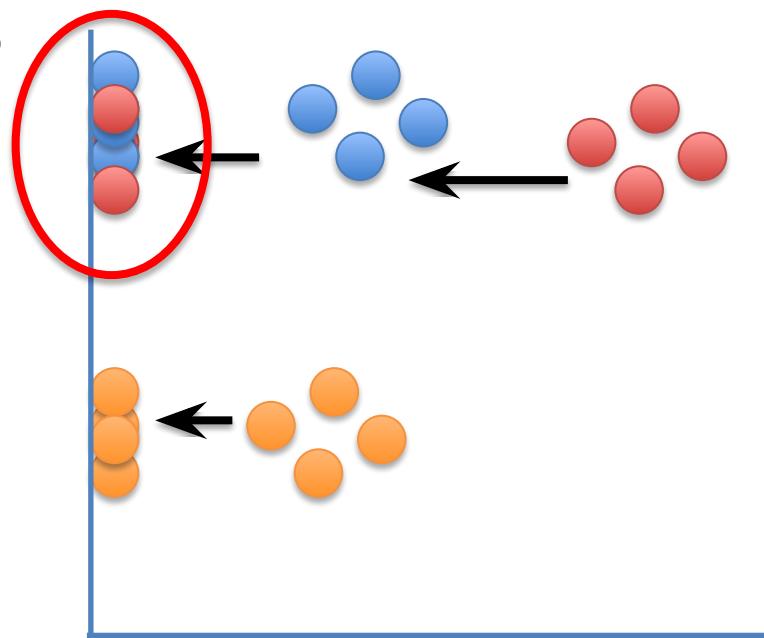
...into a flat, 1-D plot on a number line.

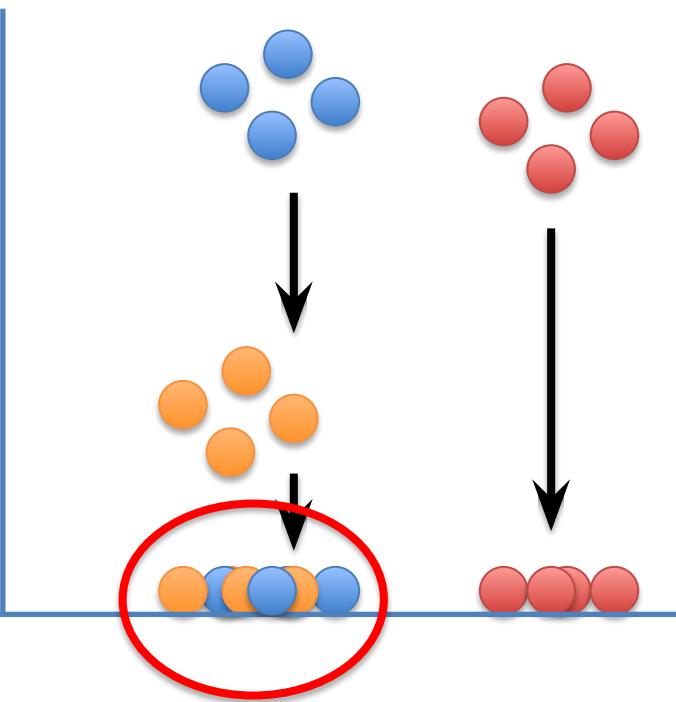


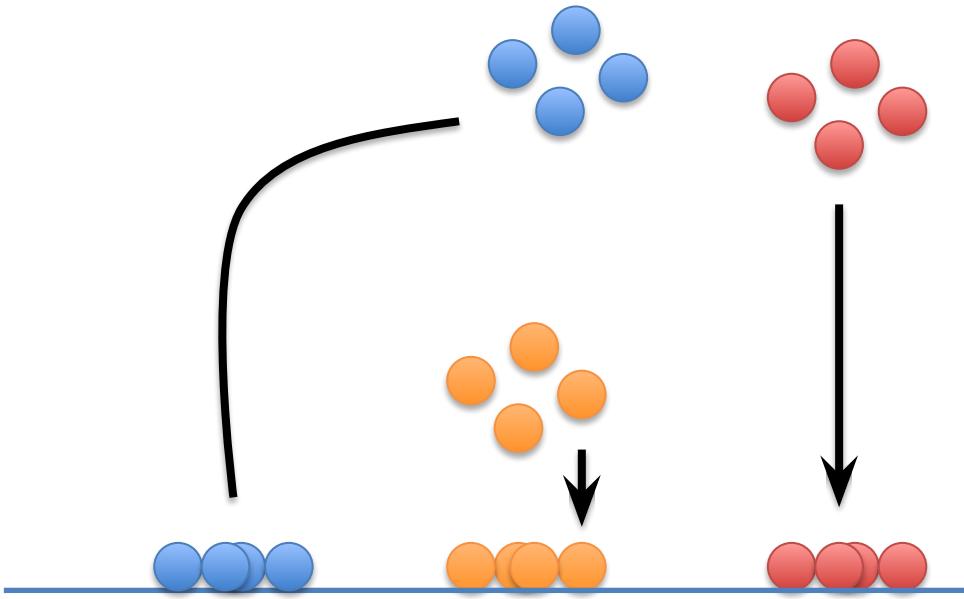


NOTE: If we just projected the data onto one of the axes, we'd just get a big mess that doesn't preserve the original clustering.

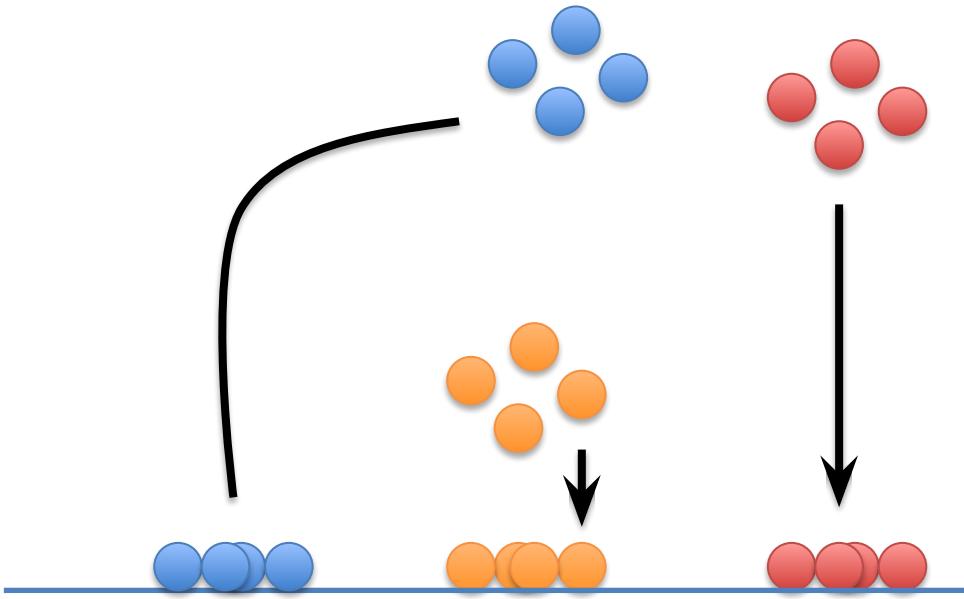
Instead of two distinct clusters, we just see a mishmash.





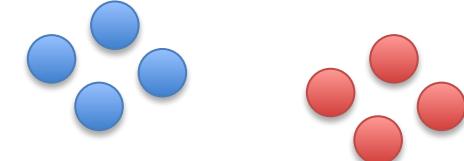


What t-SNE does is find a way to project data into a low dimensional space (in this case, the 1-D number line) so that the clustering in the high dimensional space (in this case, the 2-D scatter plot) is preserved.



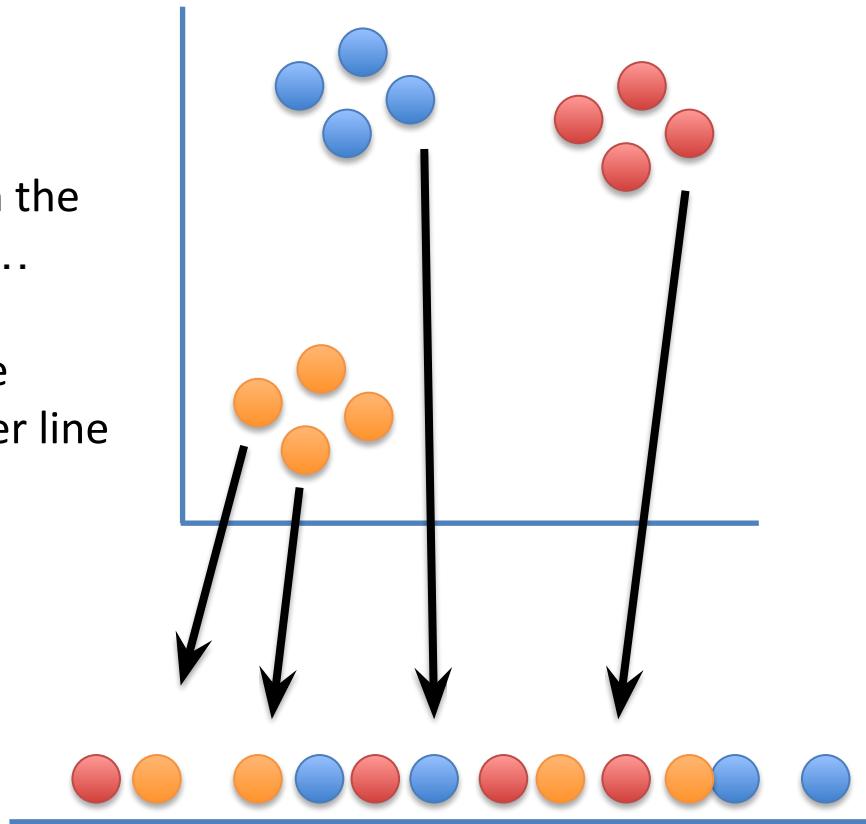
So let's step through the basic ideas of how t-SNE does this.

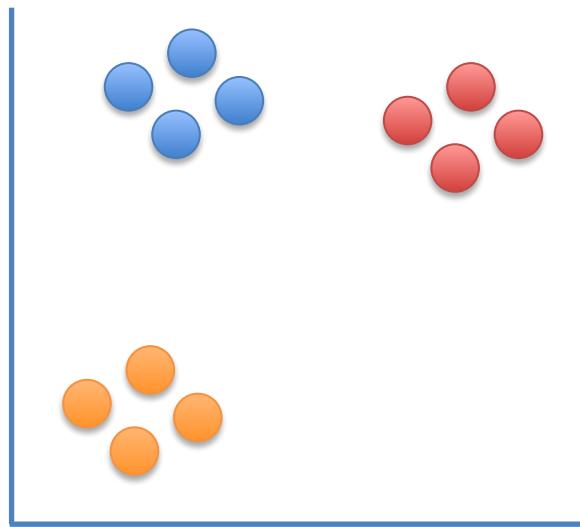
We'll start with the original scatter plot...



We'll start with the original scatter plot...

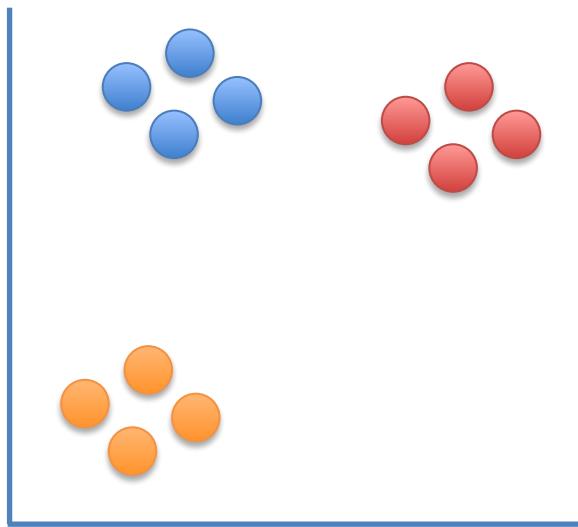
... then we'll put the points on the number line in a random order.



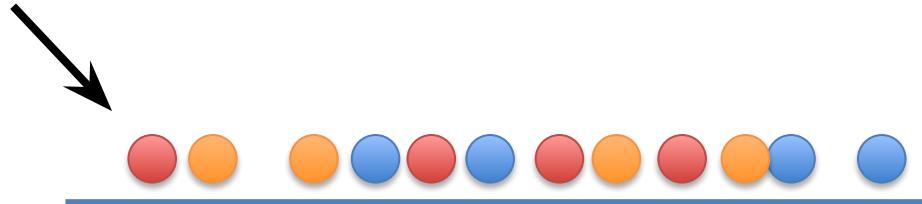


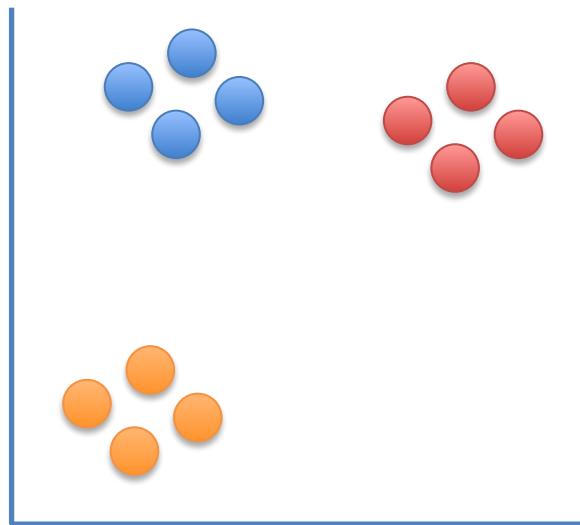
From here on out, t-SNE moves these points, a little bit at a time, until it has clustered them.





Let's figure out where to move this first point...



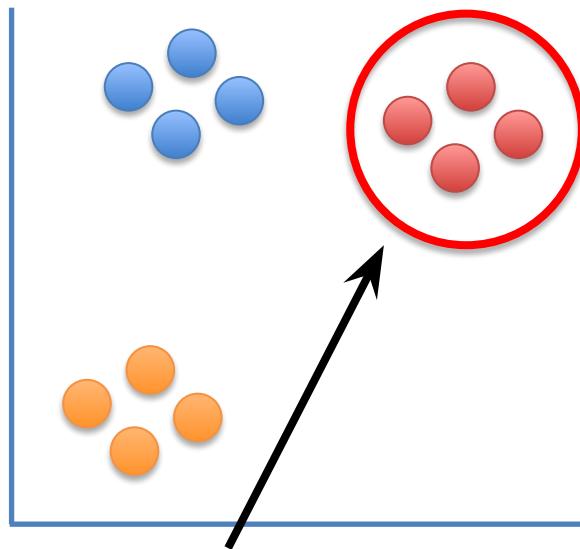


Let's figure out where to move this first point...



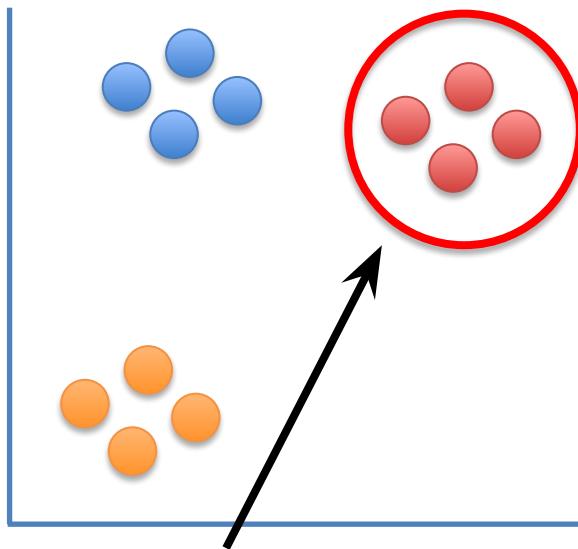
Should it move a little to the left or a little to the right?





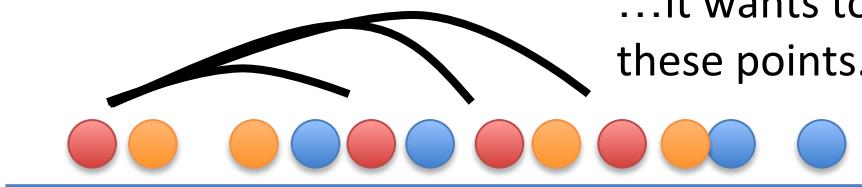
Because it is part of this  
cluster...

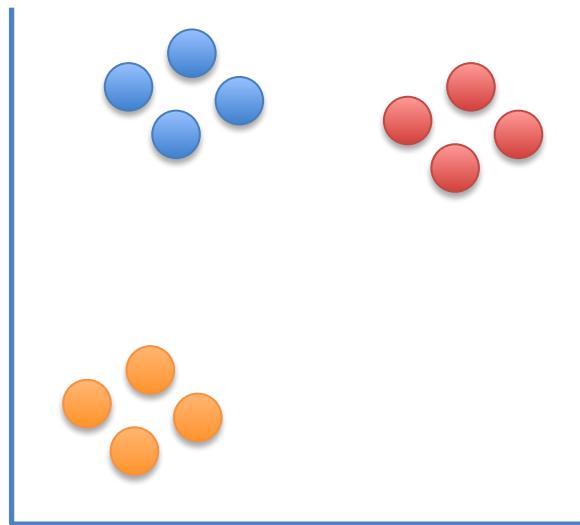




Because it is part of this  
cluster...

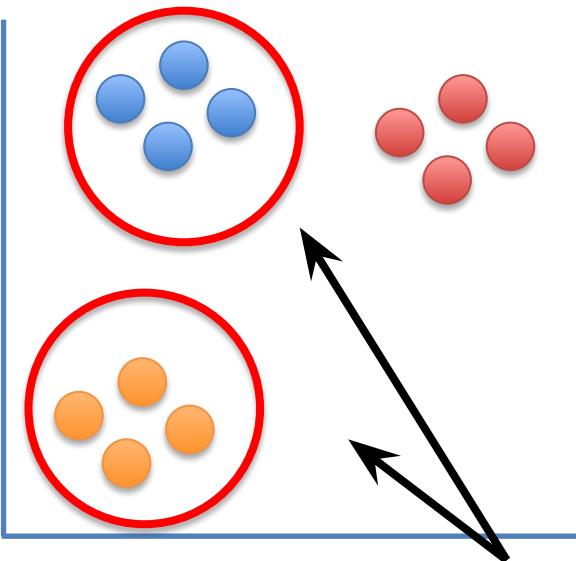
...it wants to move closer to  
these points.



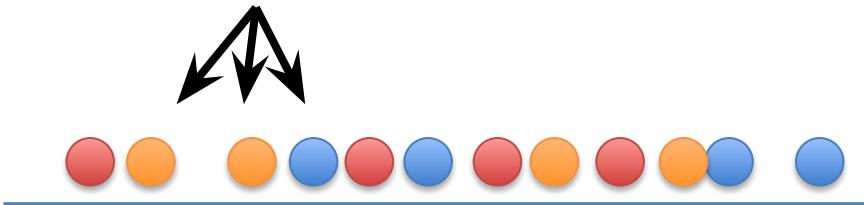


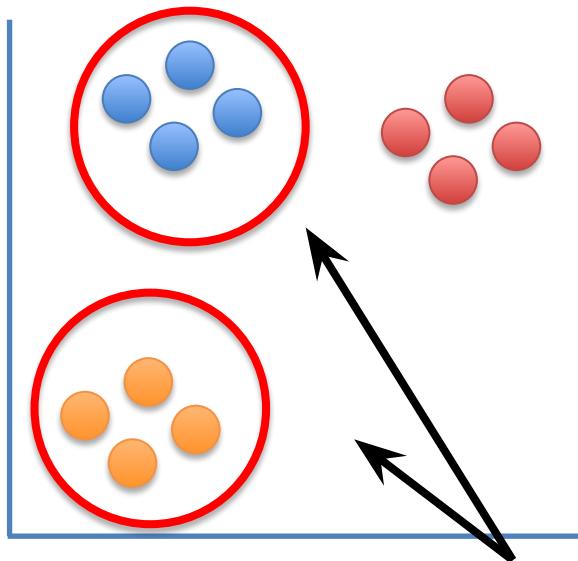
But at the same time, these  
points...



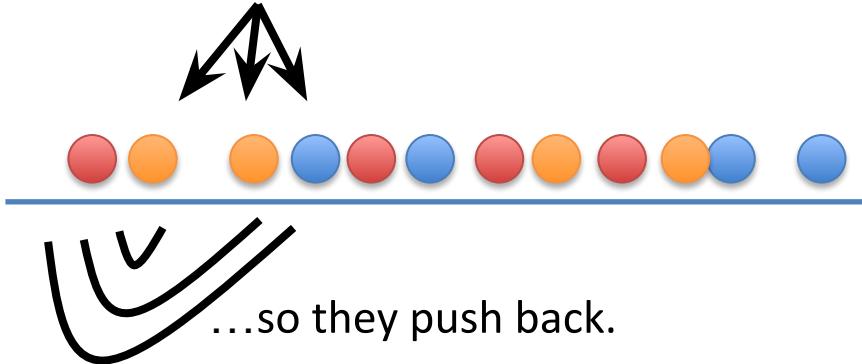


But at the same time, these points... ...are far away in the scatter plot...

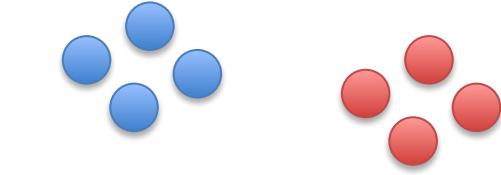




But at the same time, these points... ...are far away in the scatter plot...

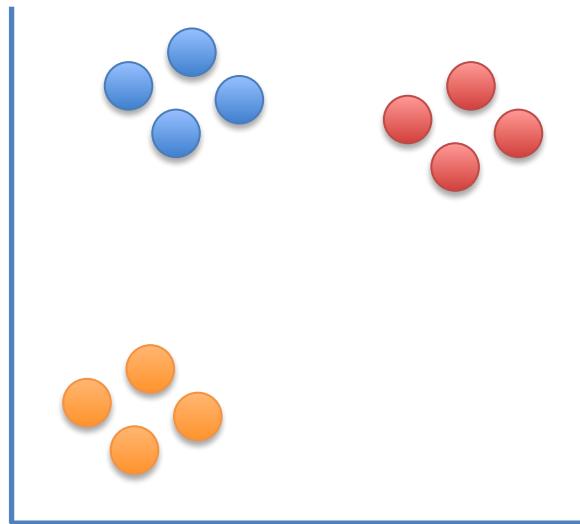


...so they push back.

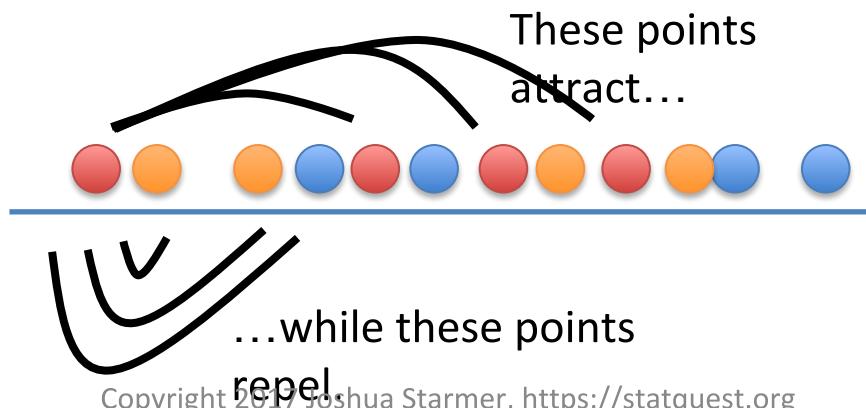


These points  
attract...

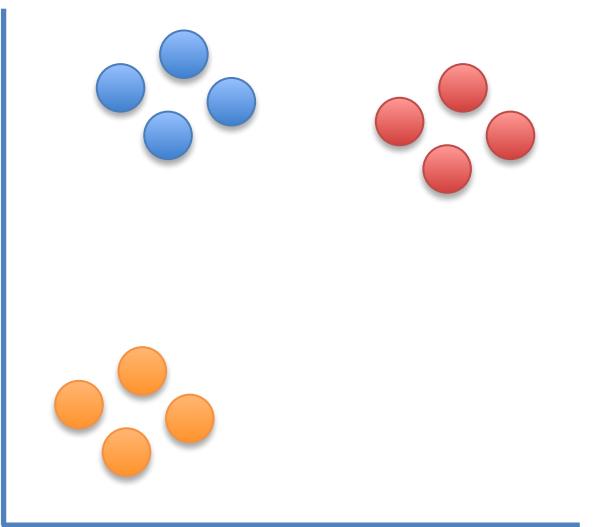




These points  
attract...

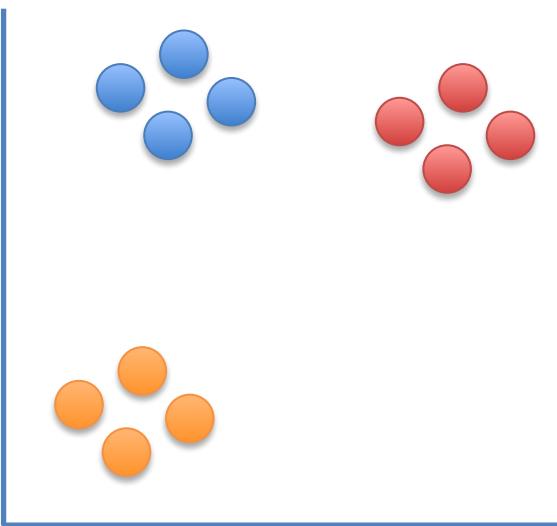


...while these points  
repel



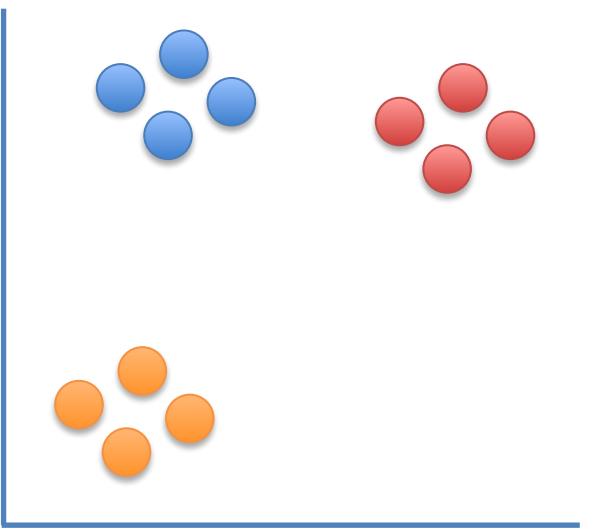
In this case, the attraction is strongest, so the point moves a little to the right.





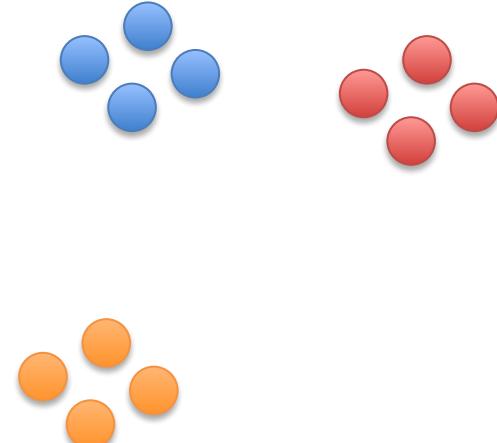
In this case, the attraction is strongest, so the point moves a little to the right.





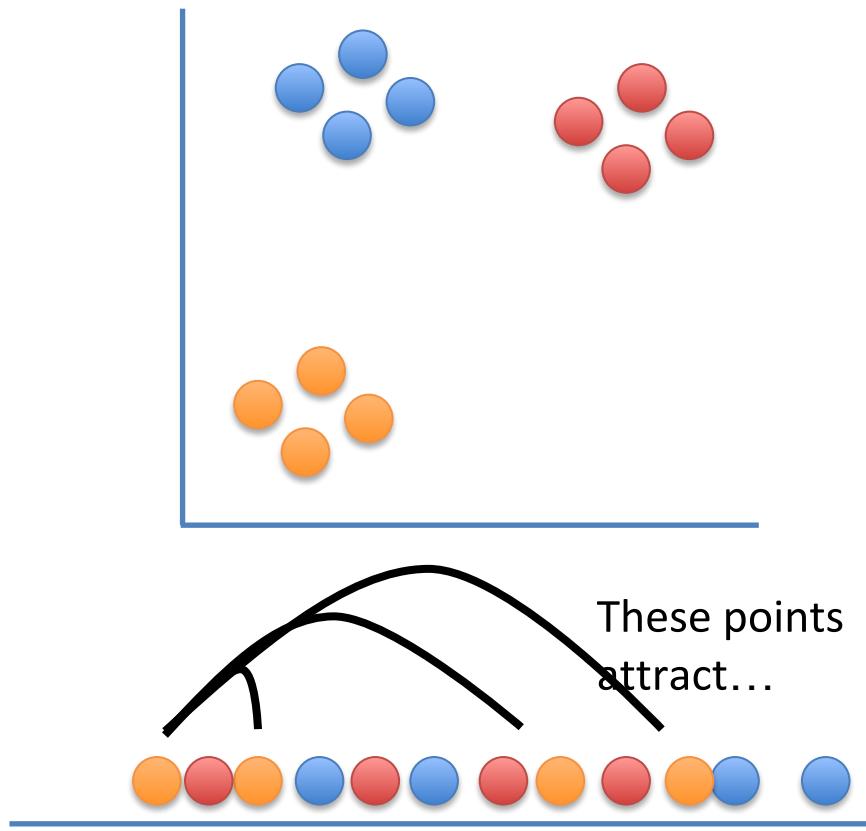
BAM!

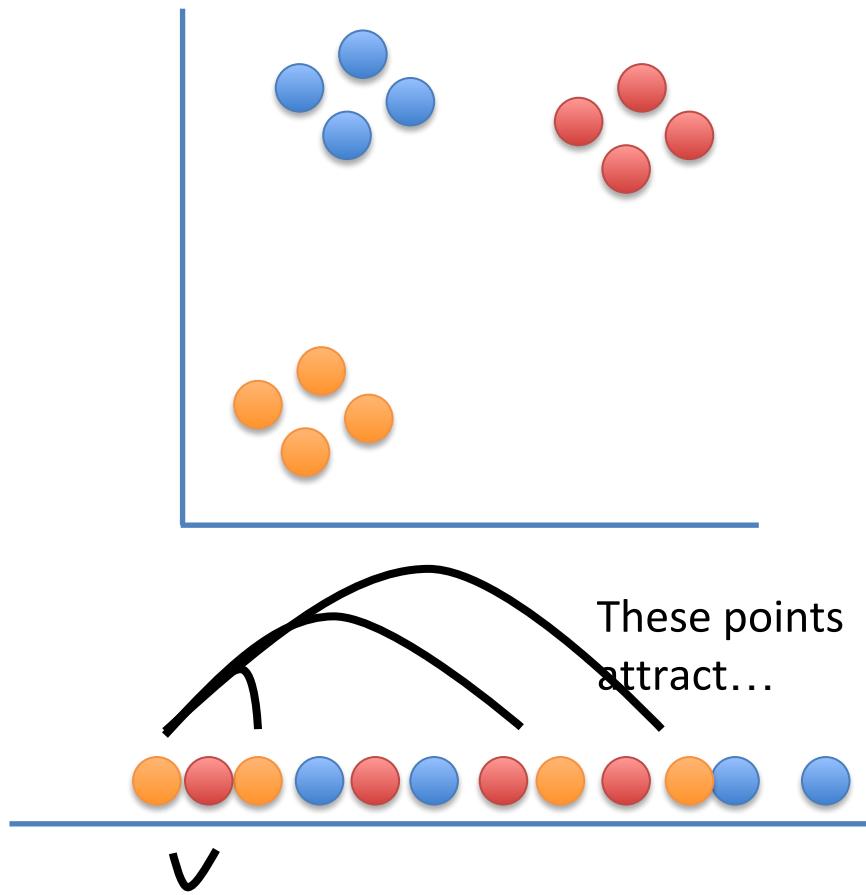




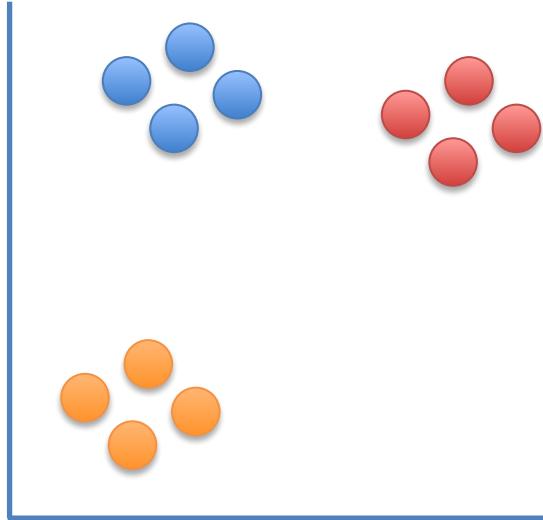
Now let's move this point a little  
bit...





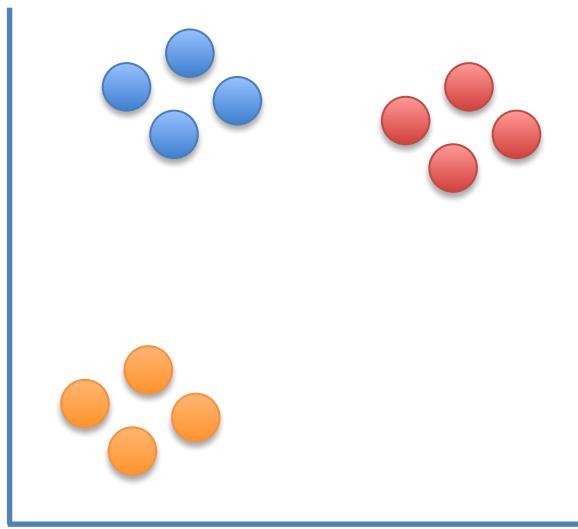


...and this point repels a little  
bit.  
Copyright 2017 Joshua Starmer, <https://statquest.org>



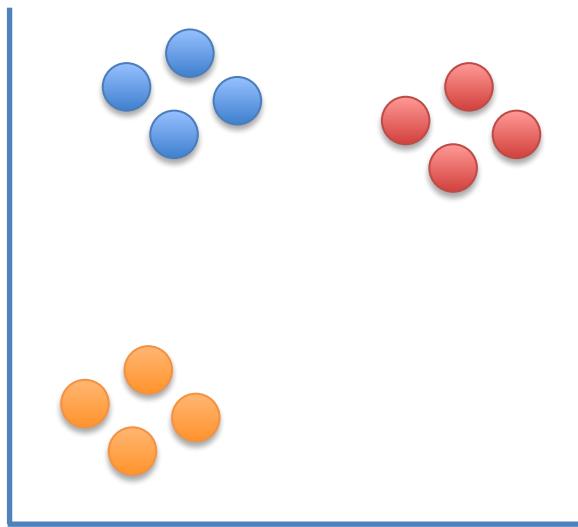
So it moves a little to closer to the other orange points.





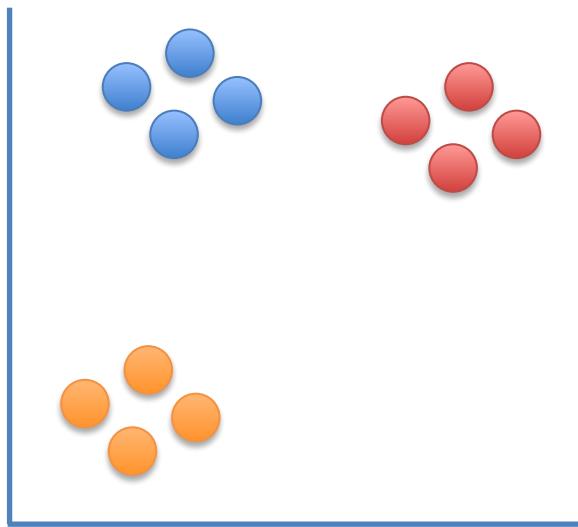
Double BAM!





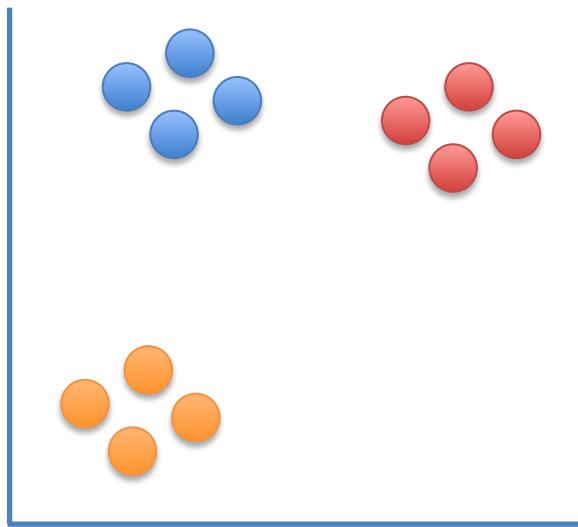
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





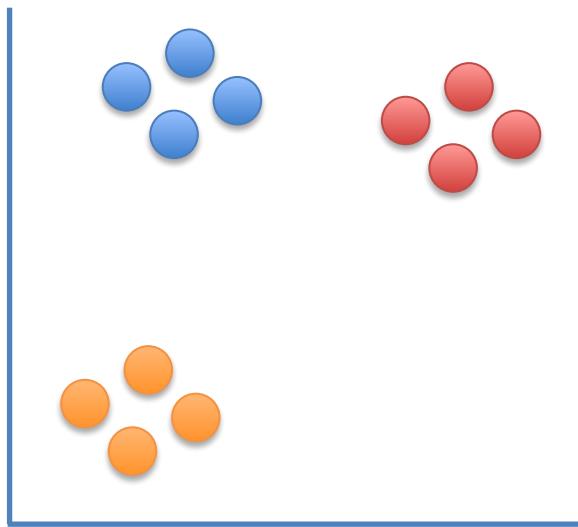
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





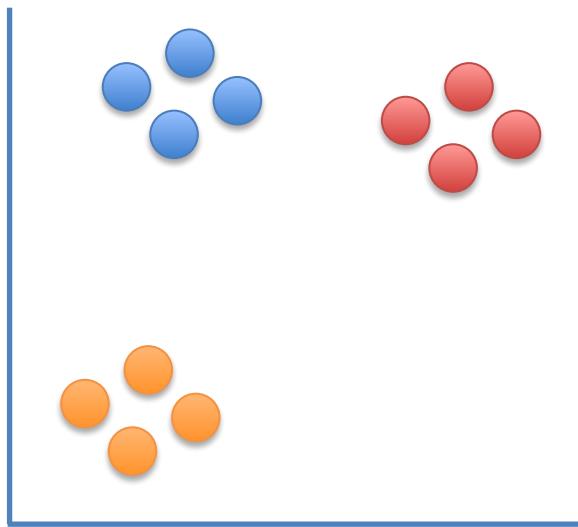
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





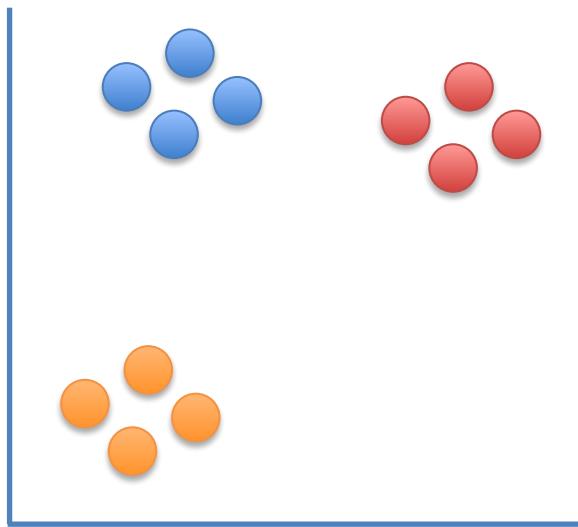
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





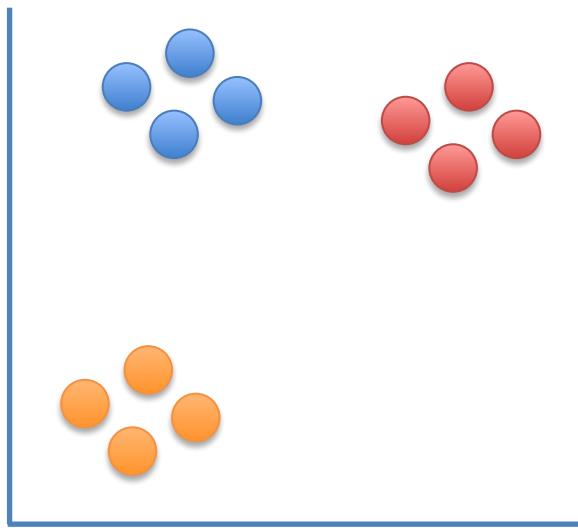
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





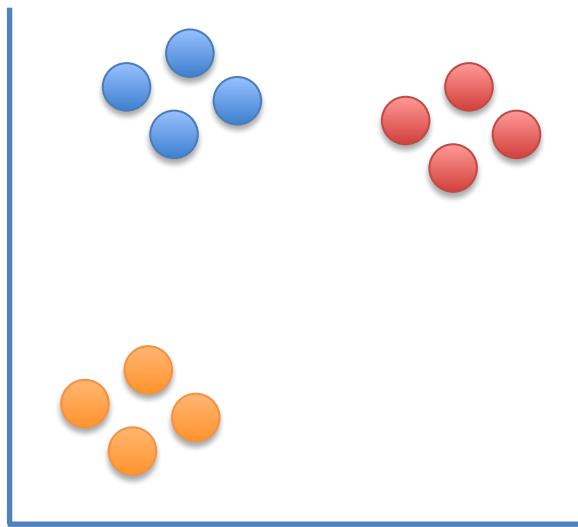
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





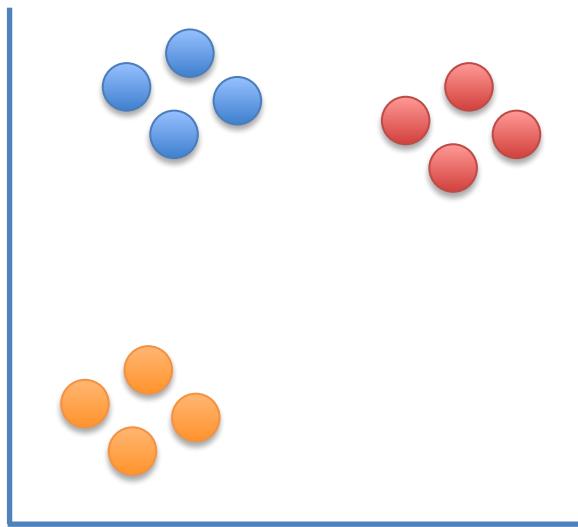
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





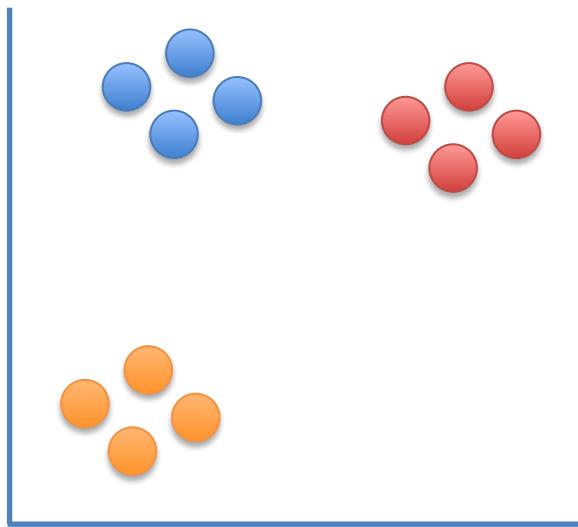
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





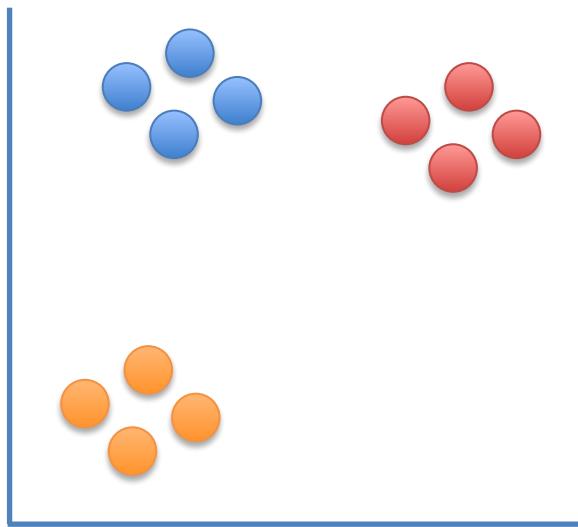
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





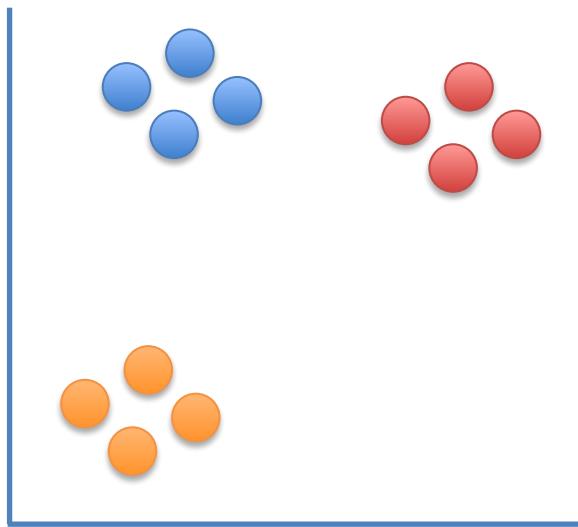
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





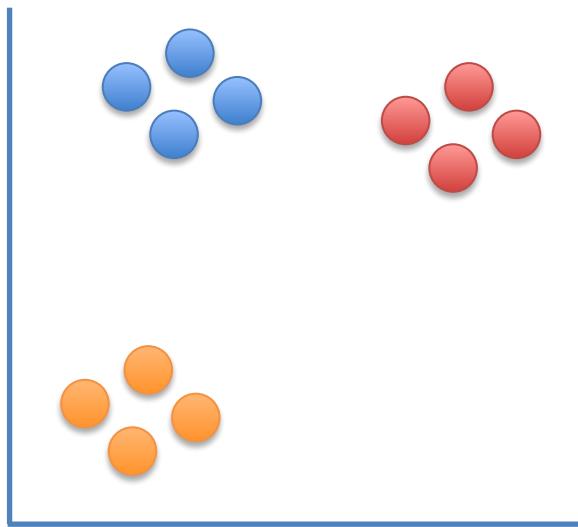
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





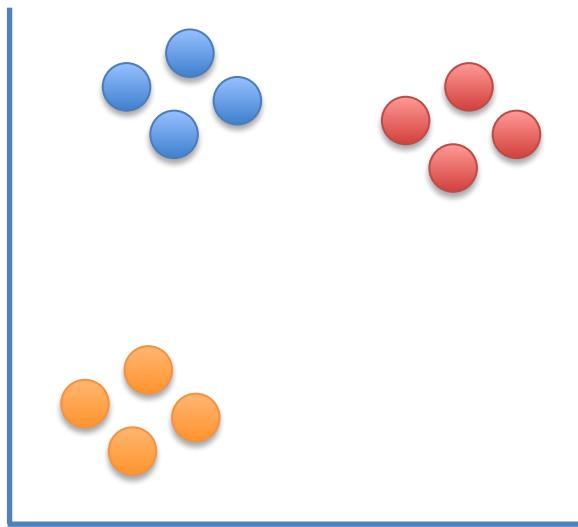
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





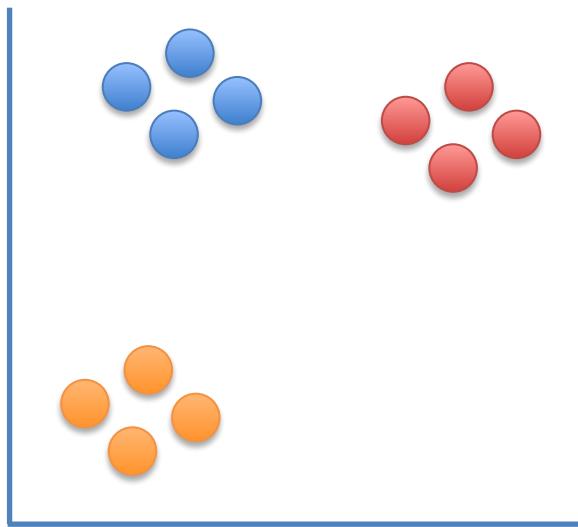
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





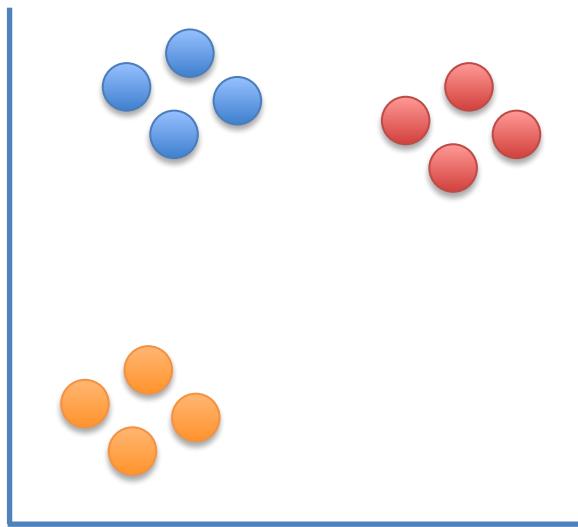
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





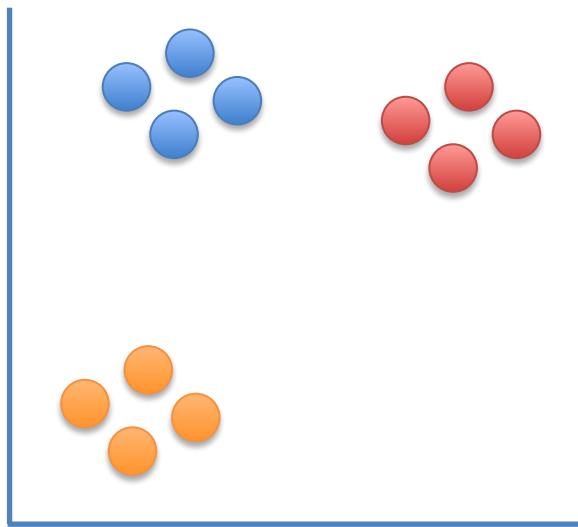
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





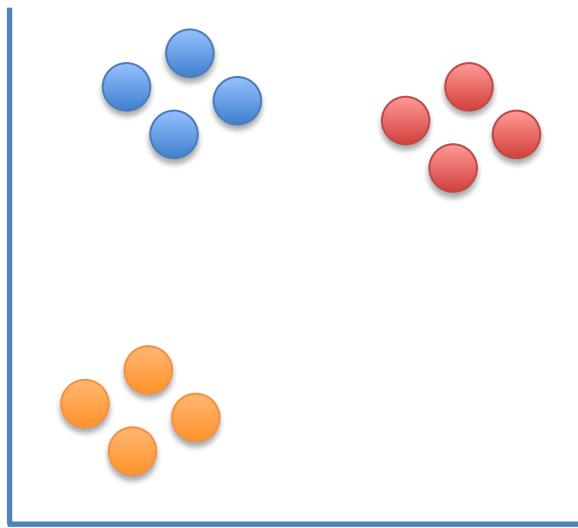
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





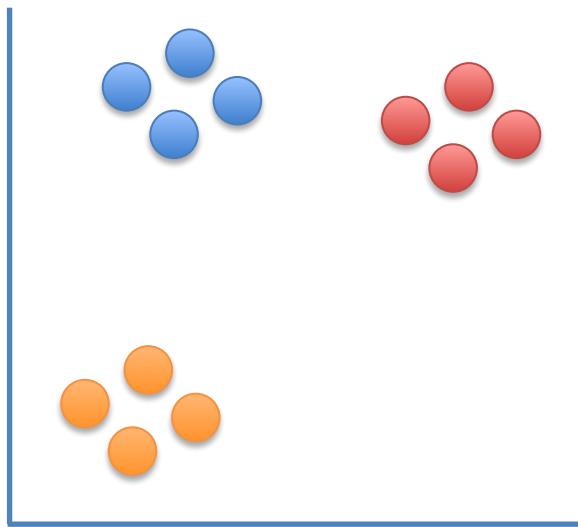
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





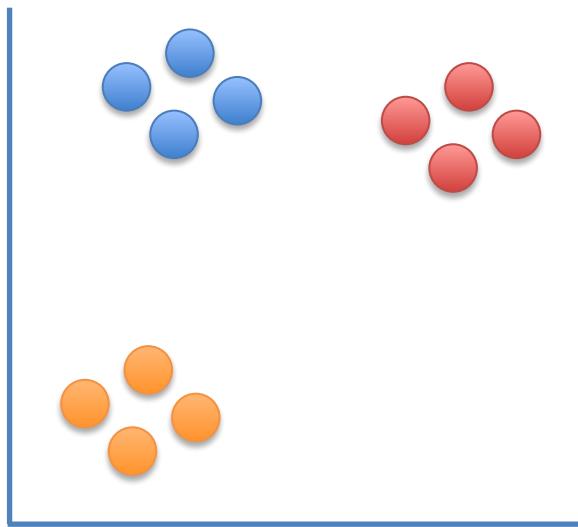
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





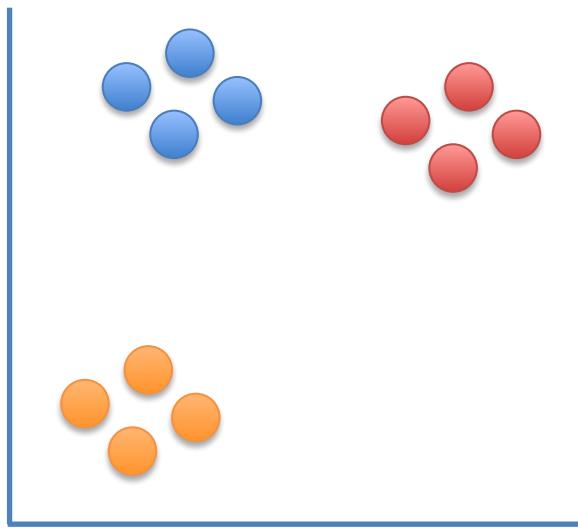
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...



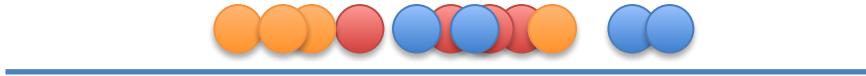


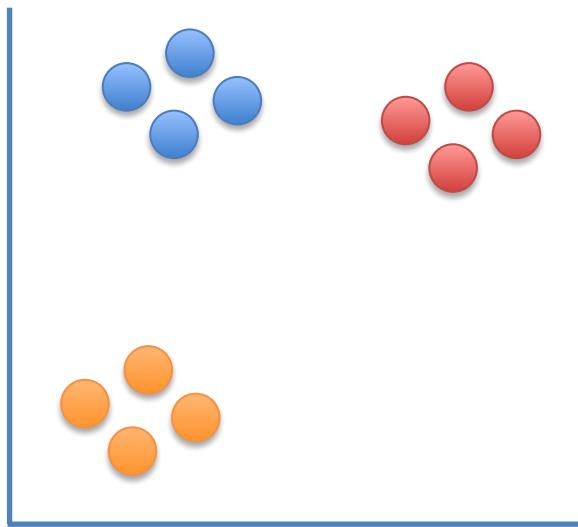
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





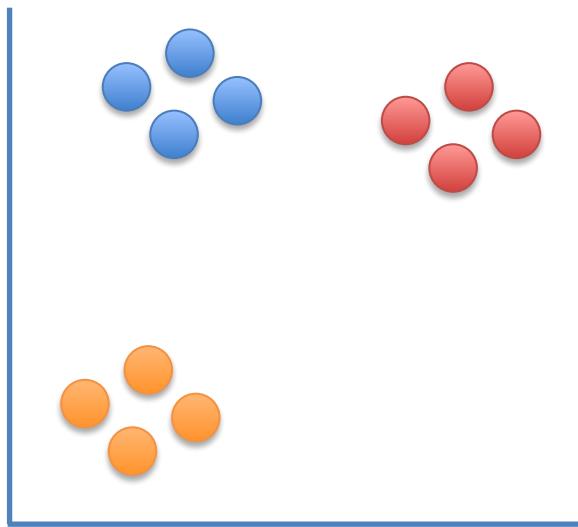
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





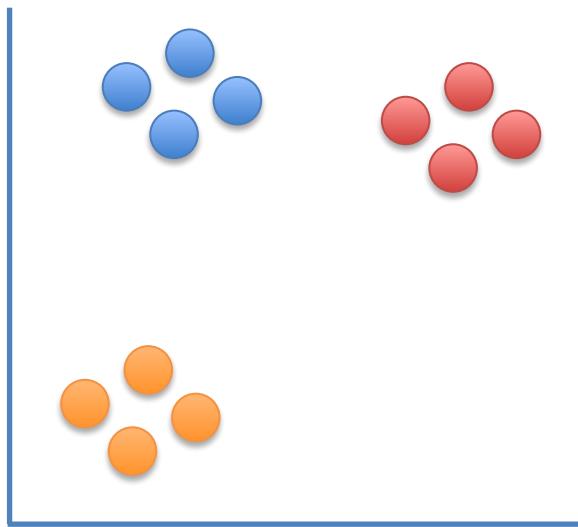
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





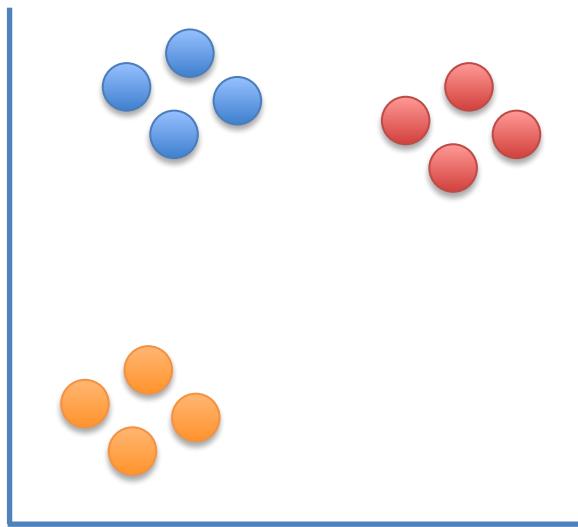
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





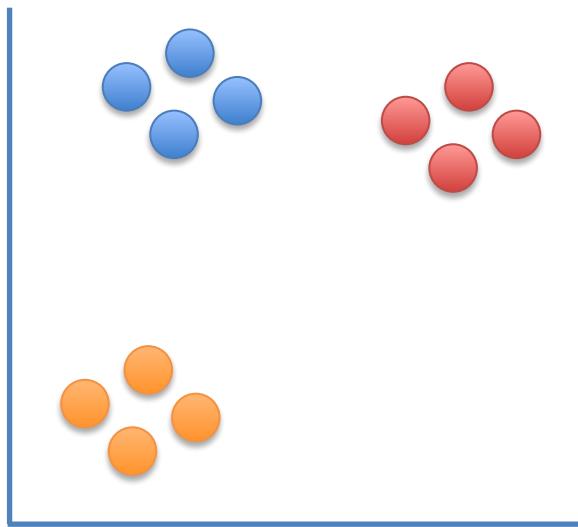
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





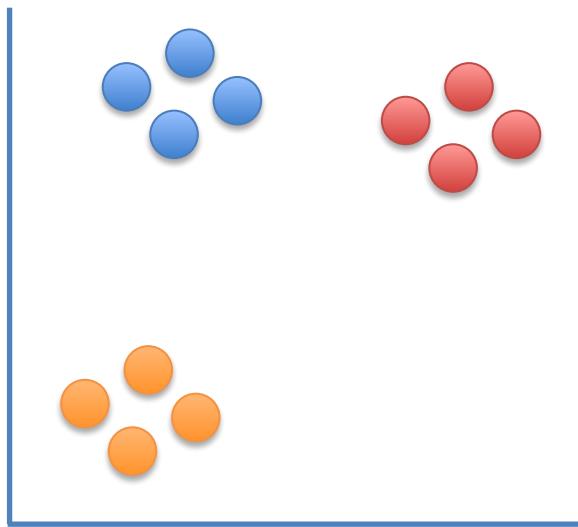
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





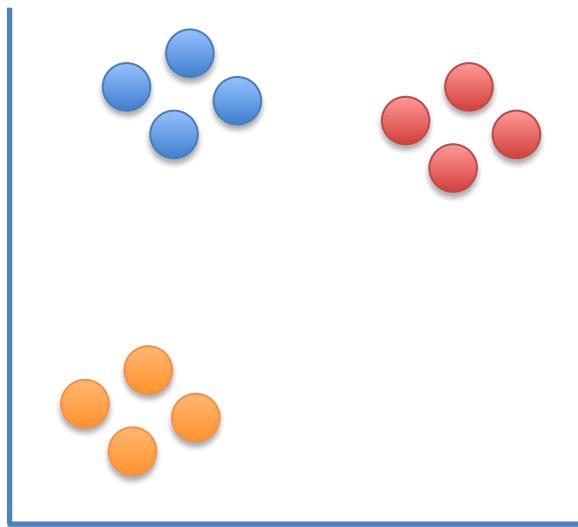
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





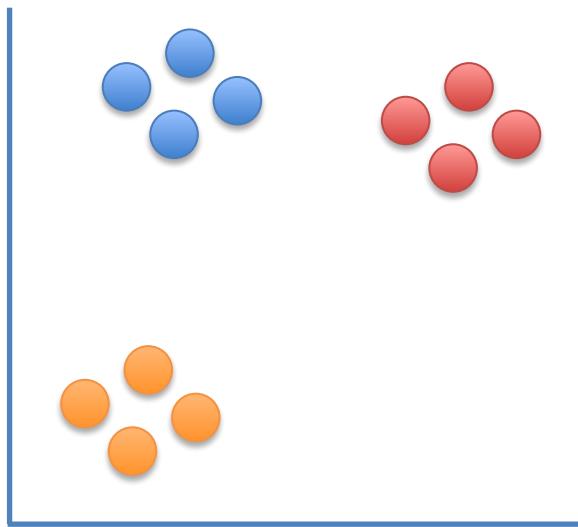
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





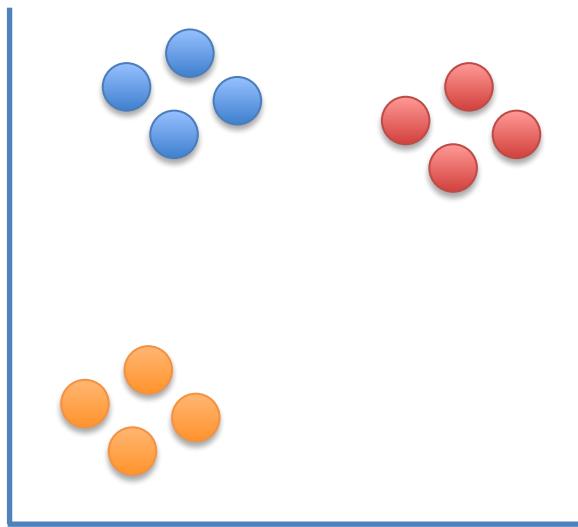
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





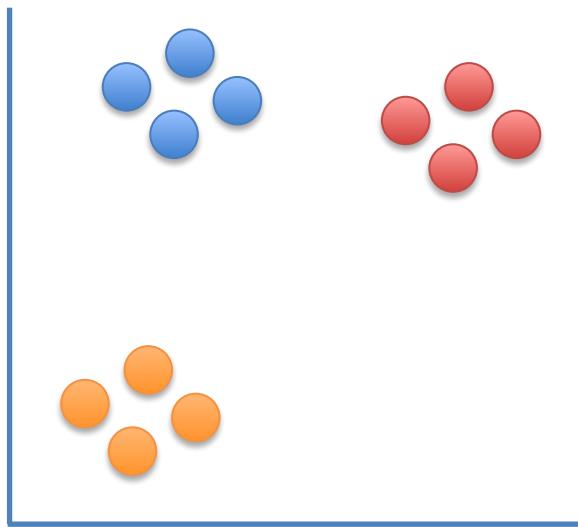
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





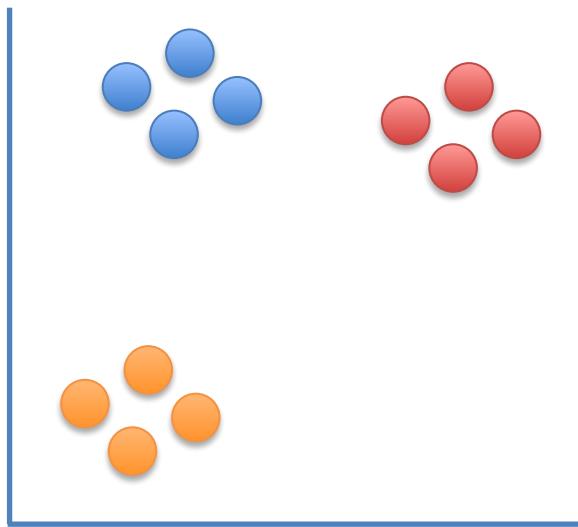
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





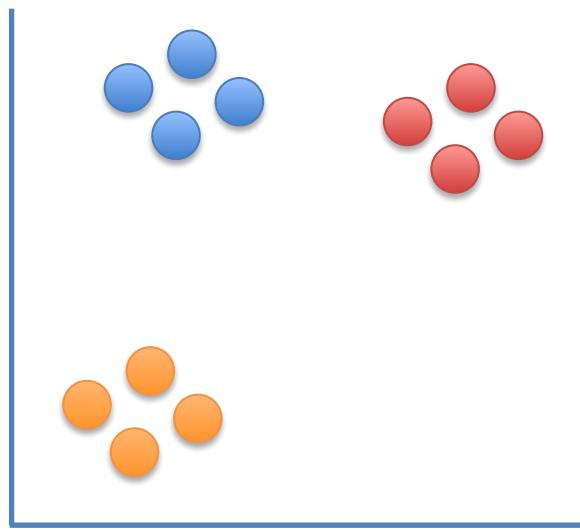
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...





At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from...

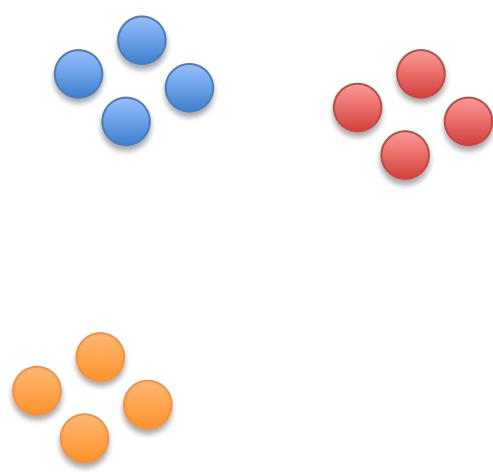




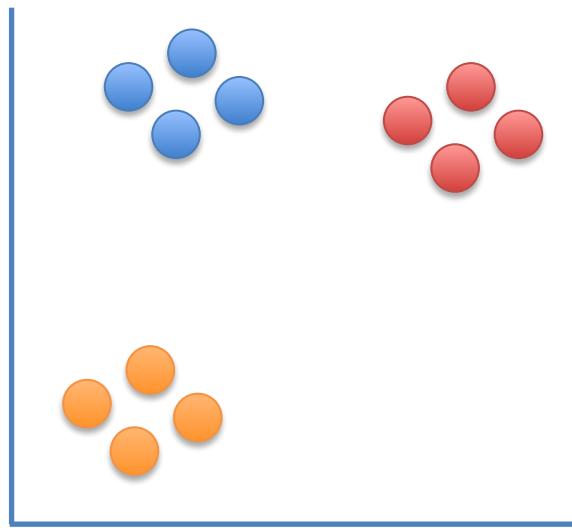
Triple BAM!!!!



Now that we've seen the what t-SNE tries to do, let's dive into the nitty-gritty details of how it does what it does.

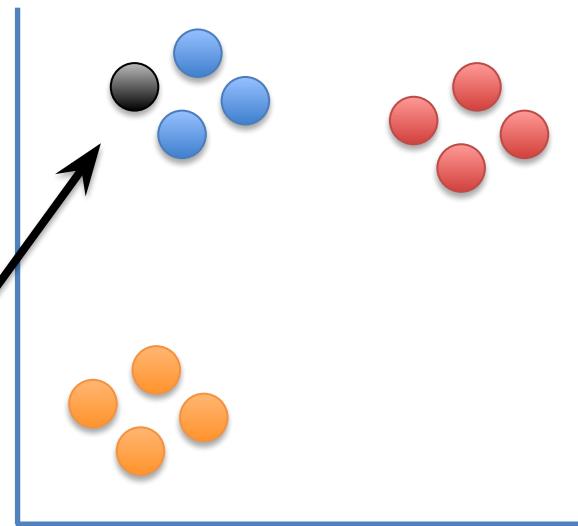


Step 1: Determine the “similarity” of all the points in the scatter plot.

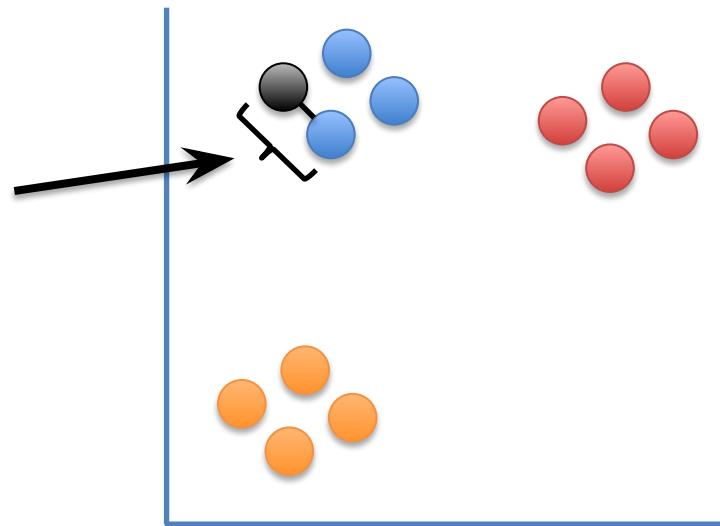


Step 1: Determine the “similarity” of all the points in the scatter plot.

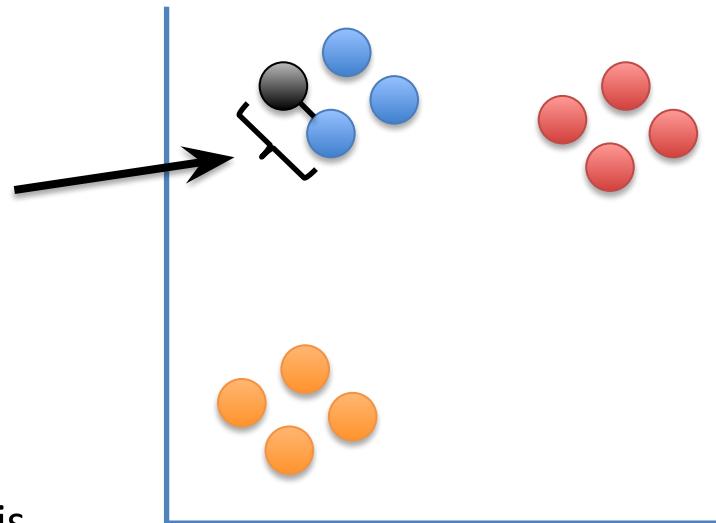
For this example, let's focus on determining the similarities between this point and all of the other points.



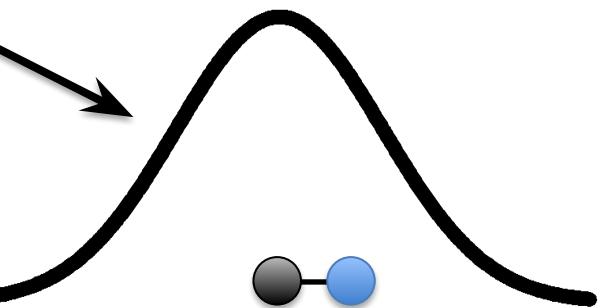
First, measure the distance between two points...



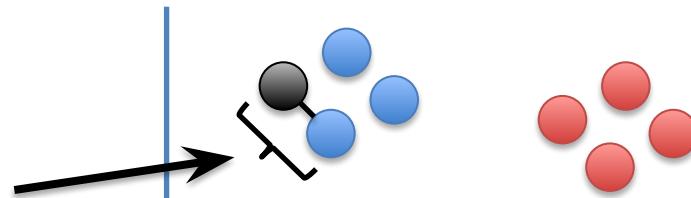
First, measure the distance between two points...



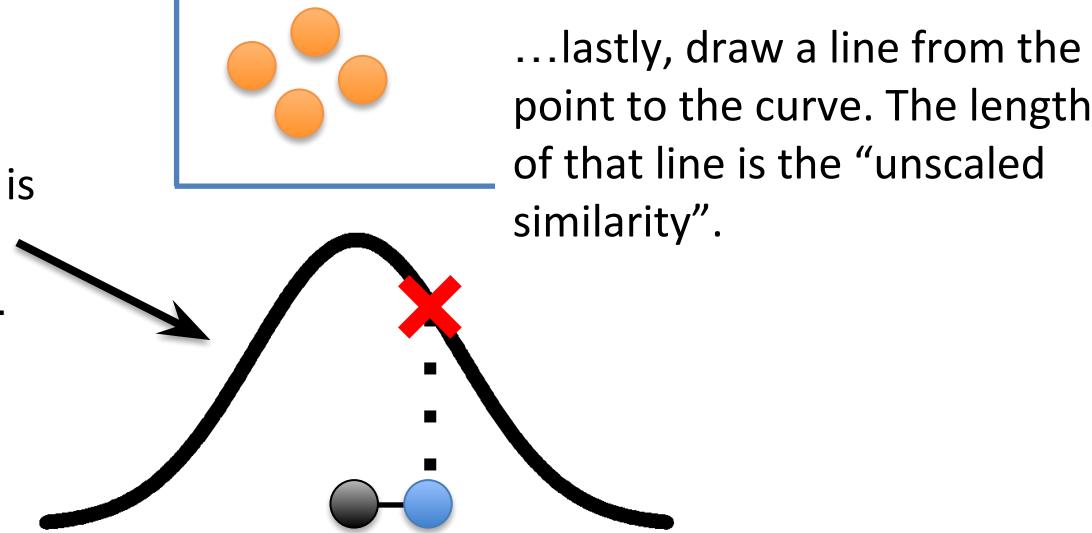
Then plot that distance on a normal curve that is centered on the point of interest...



First, measure the distance between two points...

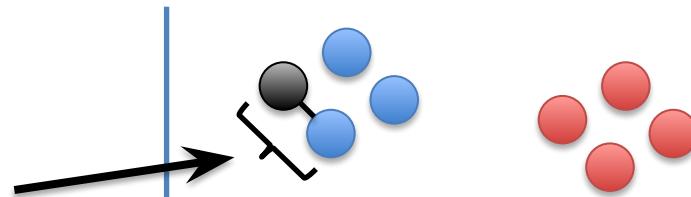


Then plot that distance on a normal curve that is centered on the point of interest...

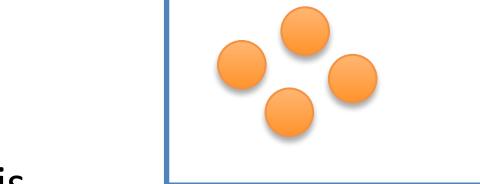


...lastly, draw a line from the point to the curve. The length of that line is the “unscaled similarity”.

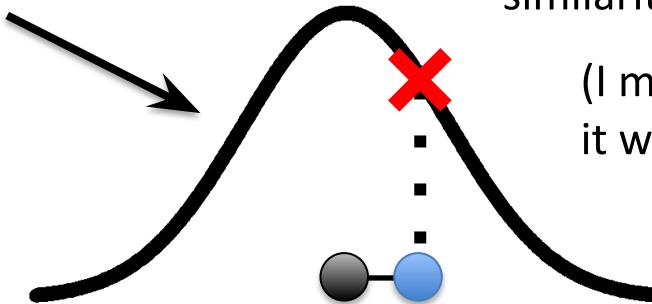
First, measure the distance between two points...



Then plot that distance on a normal curve that is centered on the point of interest...

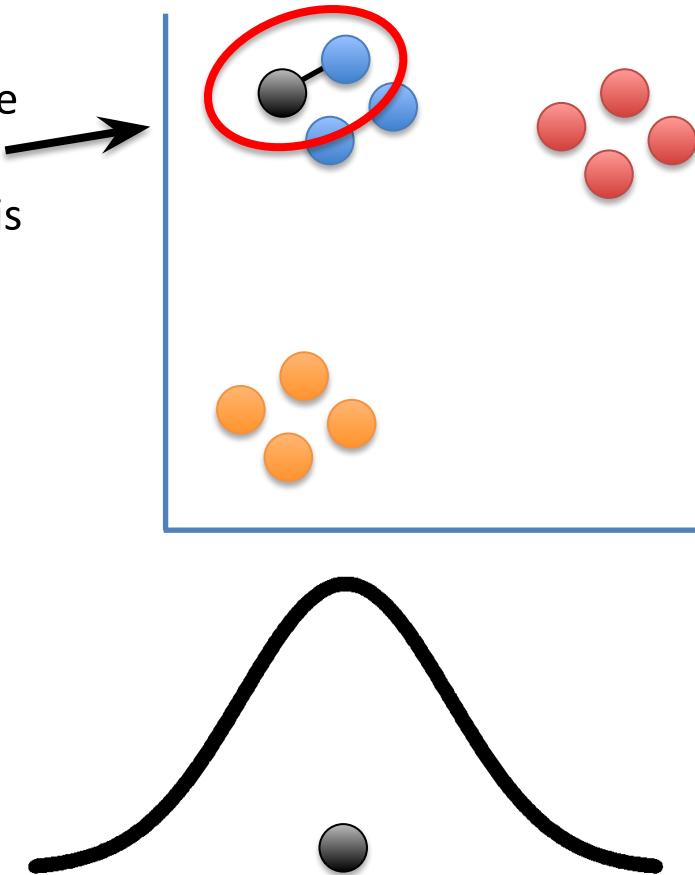


...lastly, draw a line from the point to the curve. The length of that line is the “unscaled similarity”.

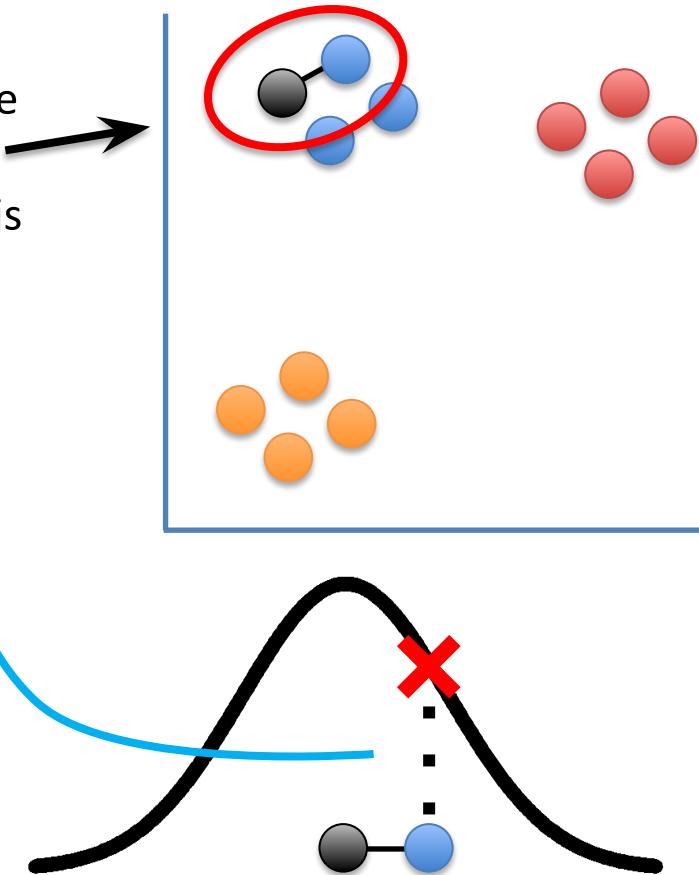


(I made that terminology up, but it will make sense in just a bit!)

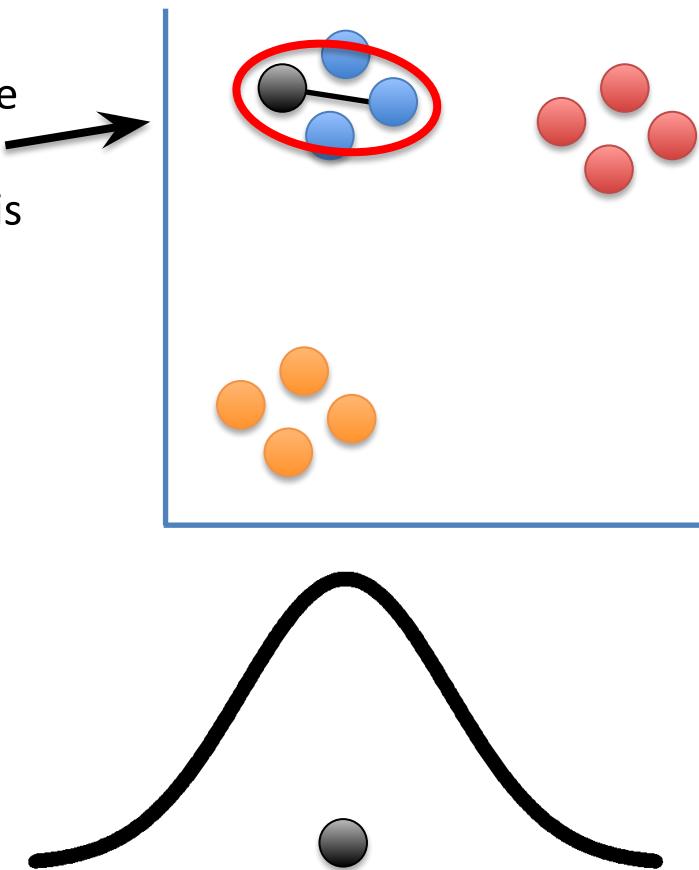
Now we calculate  
the “unscaled  
similarity” for this  
pair of points.



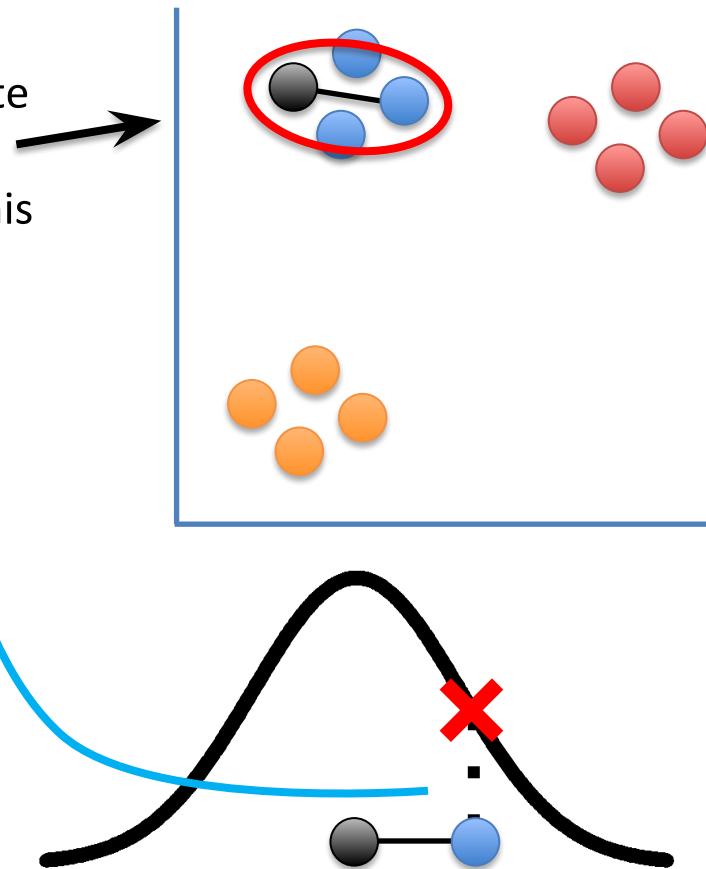
Now we calculate  
the “unscaled  
similarity” for this  
pair of points.



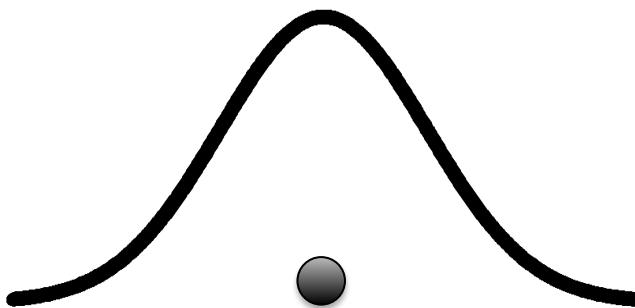
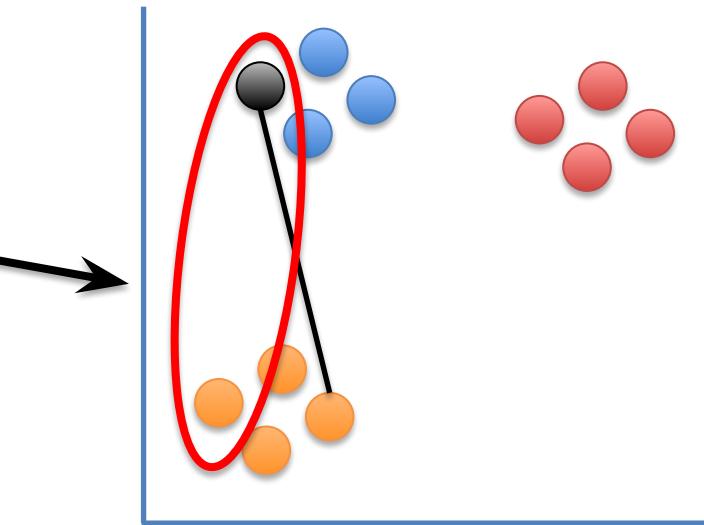
Now we calculate  
the “unscaled  
similarity” for this  
pair of points.



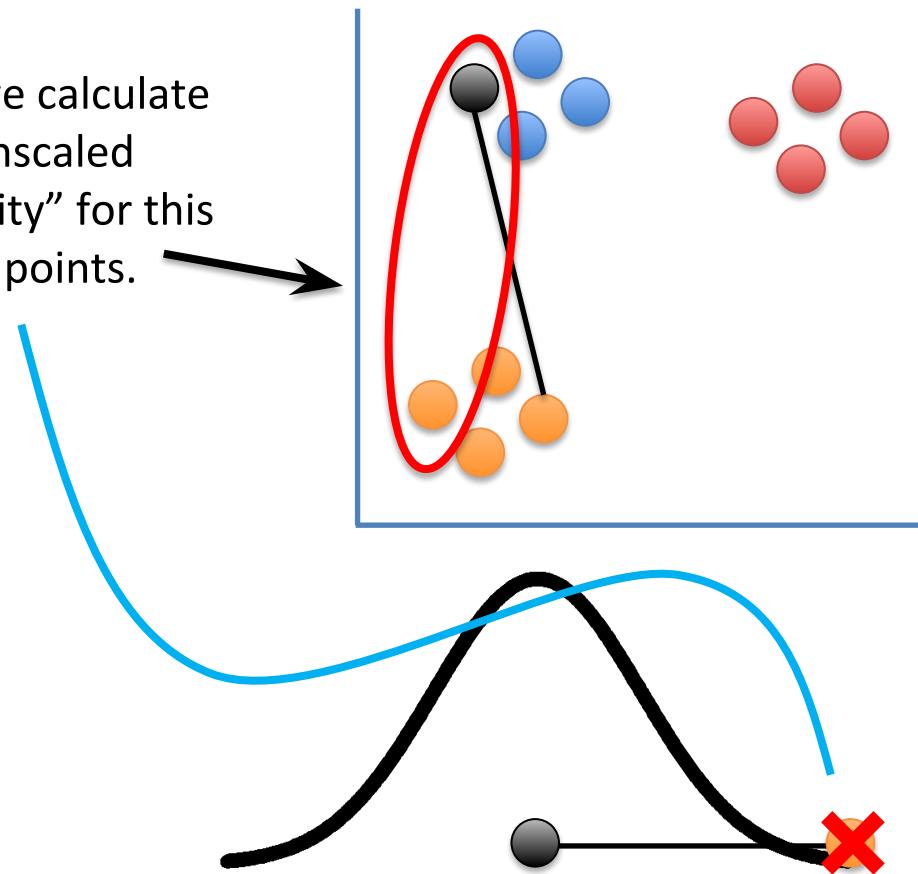
Now we calculate  
the “unscaled  
similarity” for this  
pair of points.



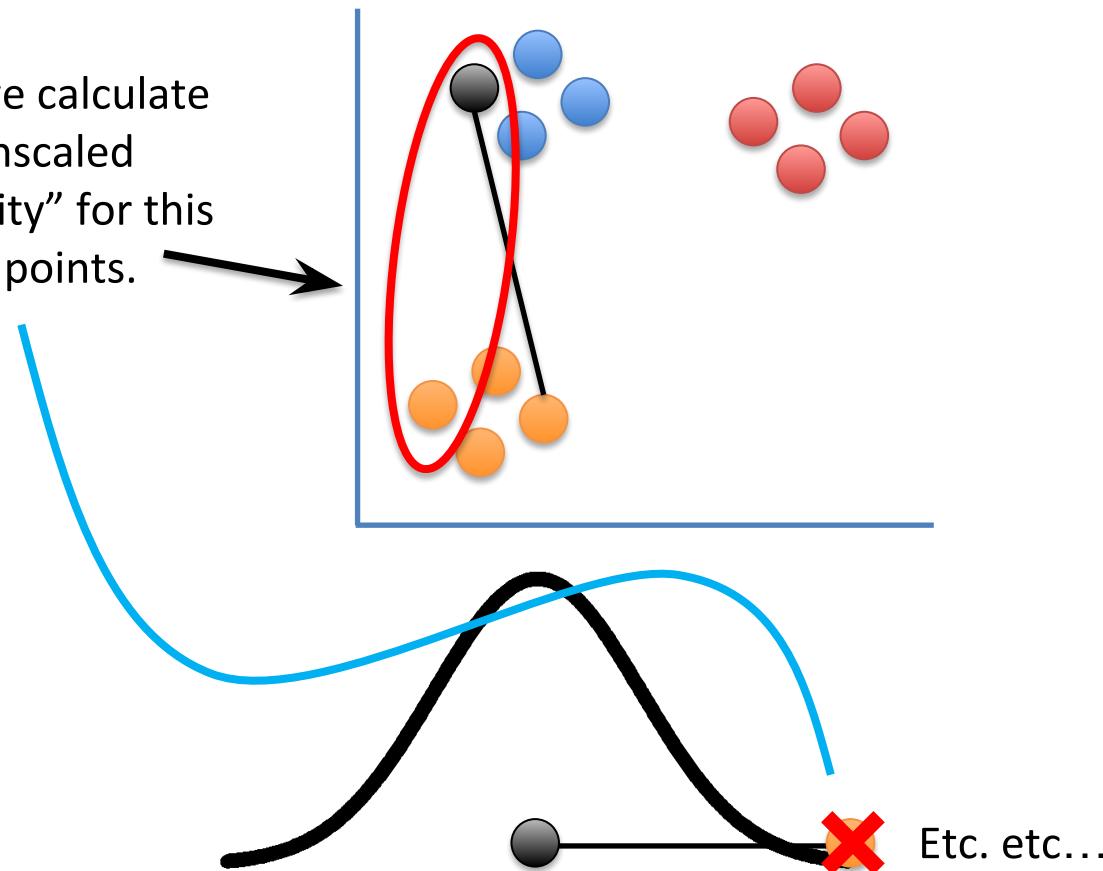
Now we calculate  
the “unscaled  
similarity” for this  
pair of points.

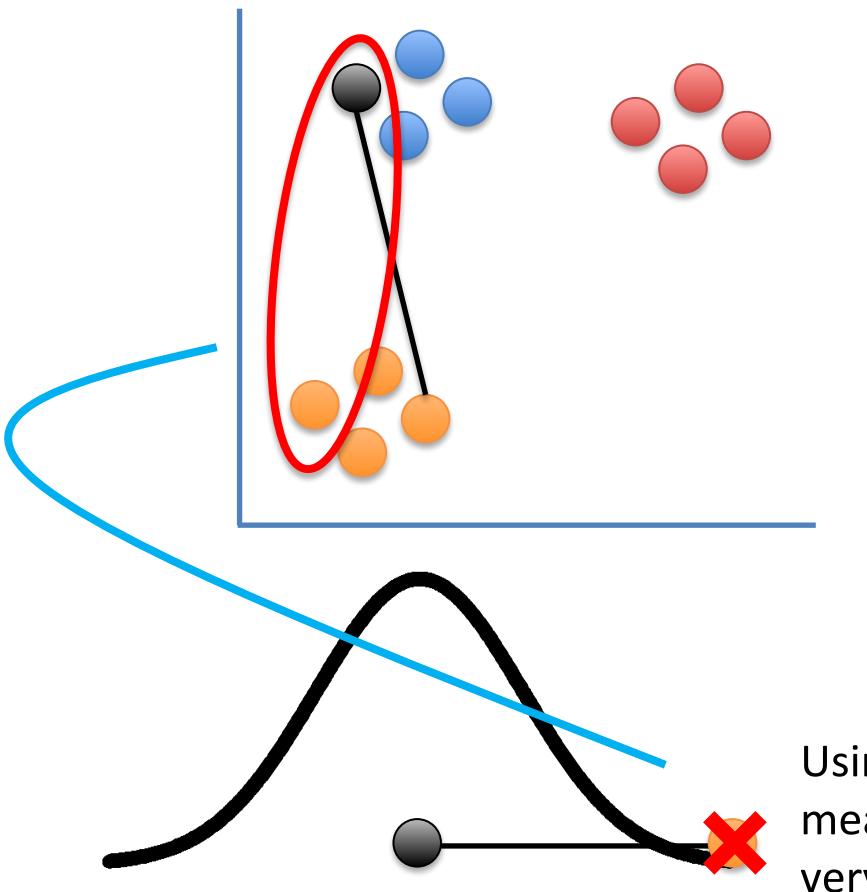


Now we calculate  
the “unscaled  
similarity” for this  
pair of points.

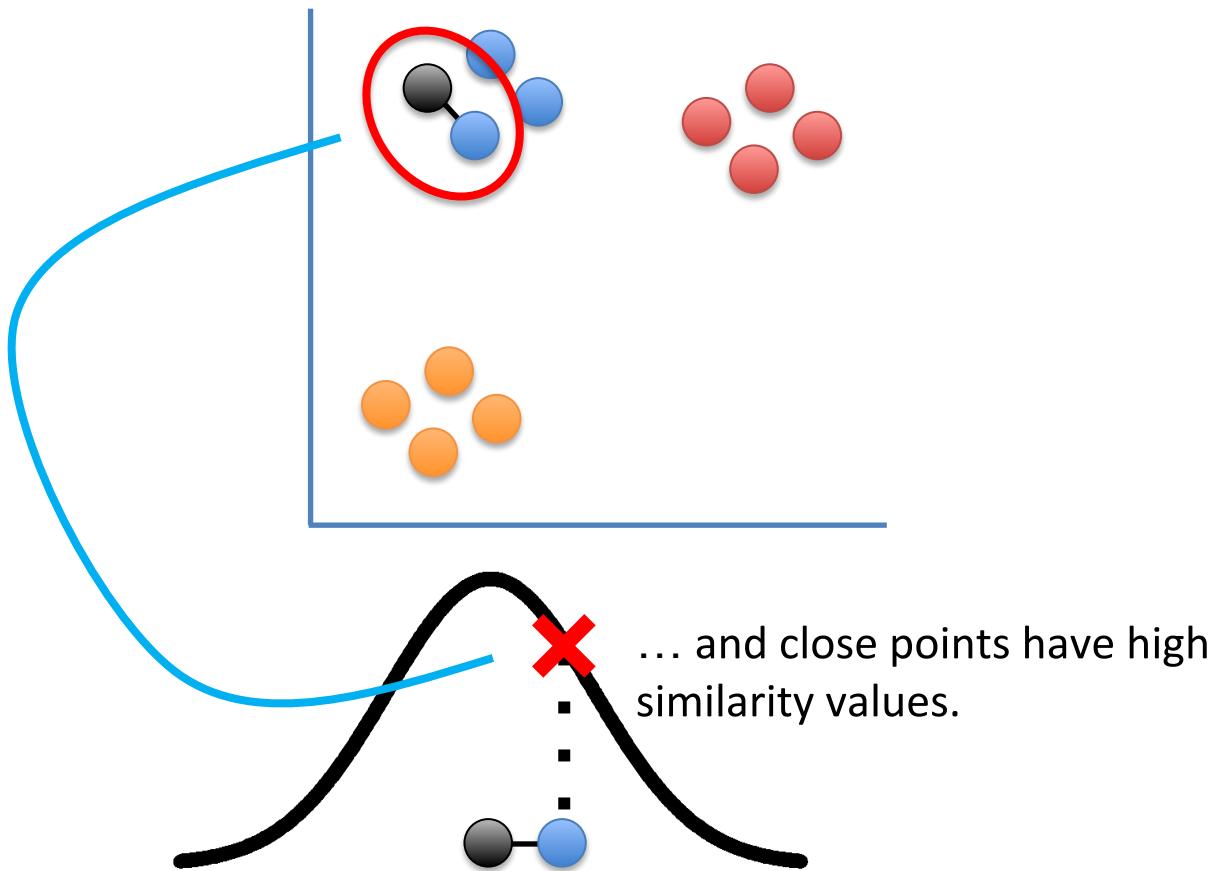


Now we calculate  
the “unscaled  
similarity” for this  
pair of points.

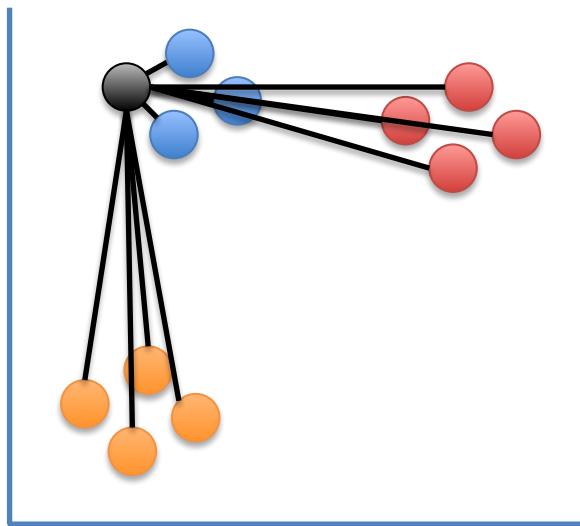




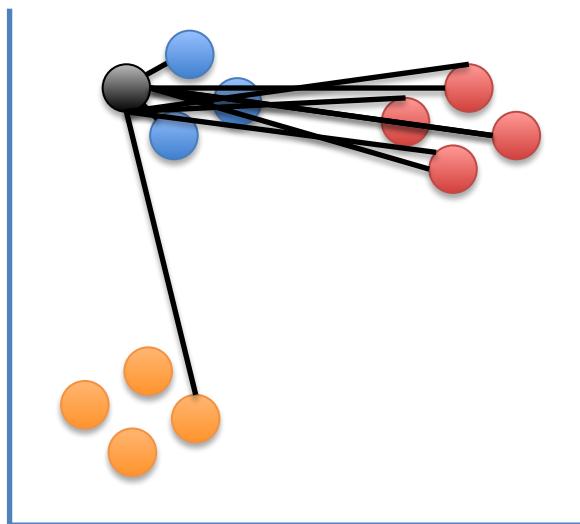
Using a normal distribution means that distant points have very low similarity values....



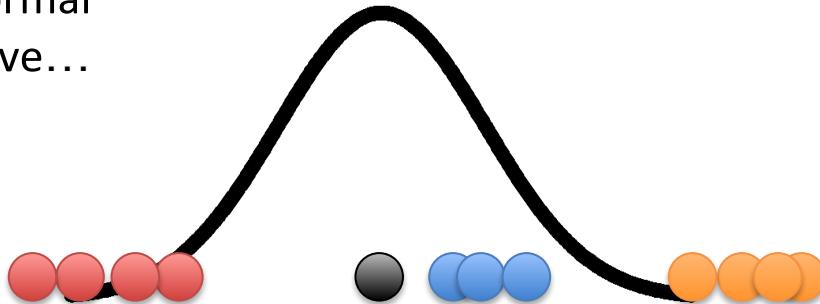
Ultimately, we measure  
the distances between  
all of the points and the  
point of interest...



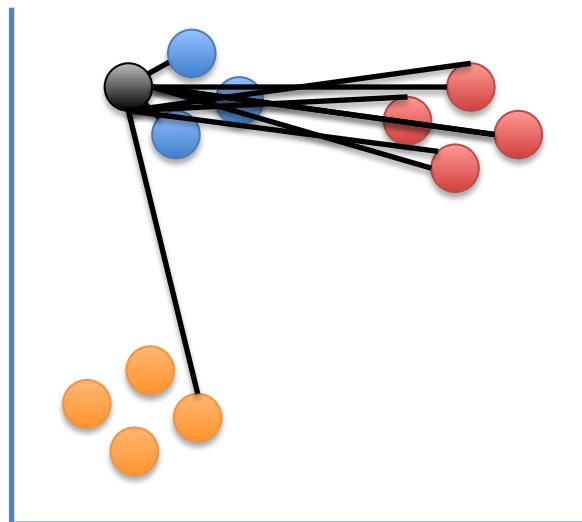
Ultimately, we measure  
the distances between  
all of the points and the  
point of interest...



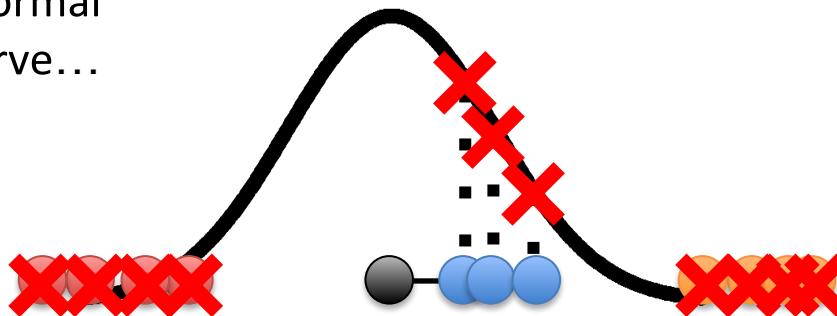
Plot them on the normal  
curve...



Ultimately, we measure  
the distances between  
all of the points and the  
point of interest...

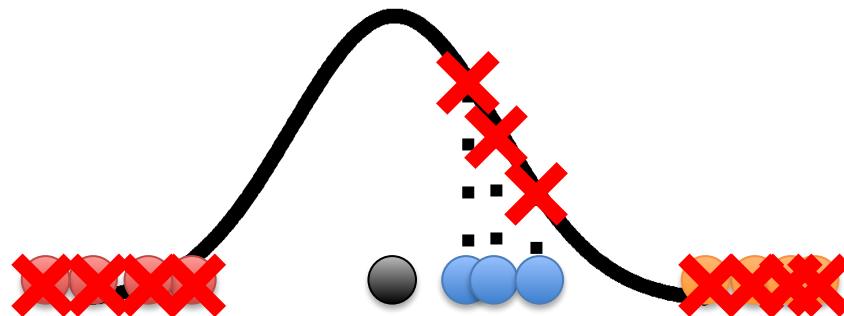


Plot them on the normal  
curve...



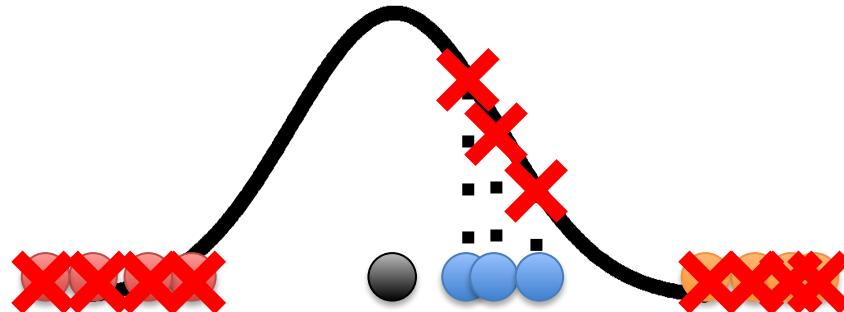
...and then measure the  
distances from the points  
to the curve to get the  
unscaled similarity scores  
with respect to the point  
of interest.

The next step is to scale the unscaled similarities so that they add up to 1.

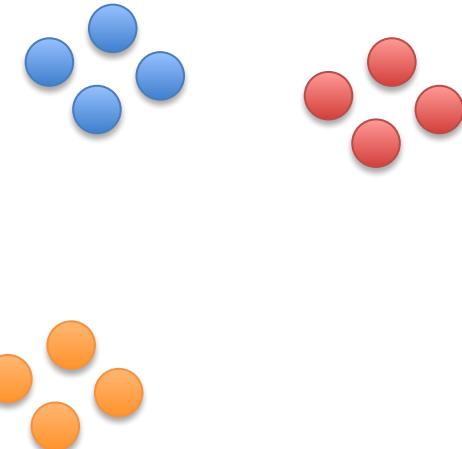


The next step is to scale the unscaled similarities so that they add up to 1.

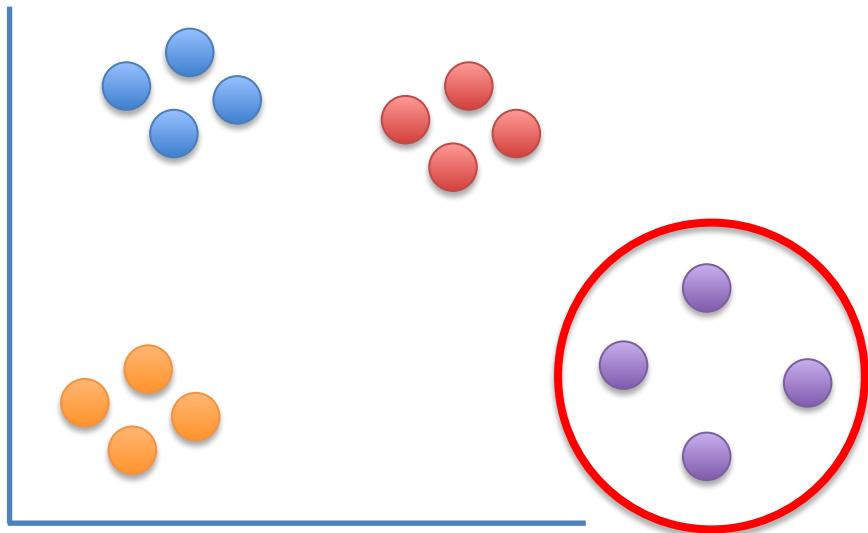
Umm... Why do the similarity scores need to add up to 1?



It has to do with something  
I didn't tell you earlier...

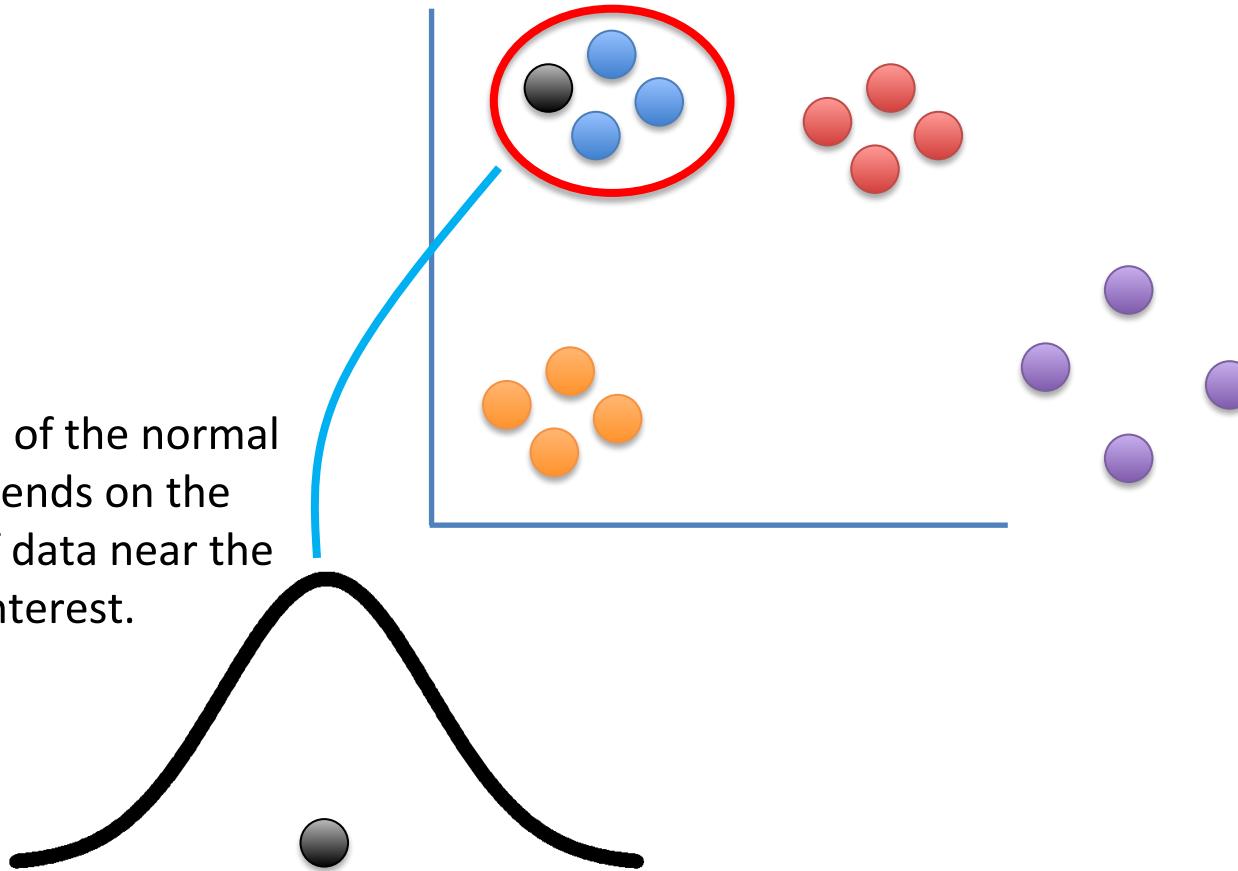


It has to do with something  
I didn't tell you earlier...

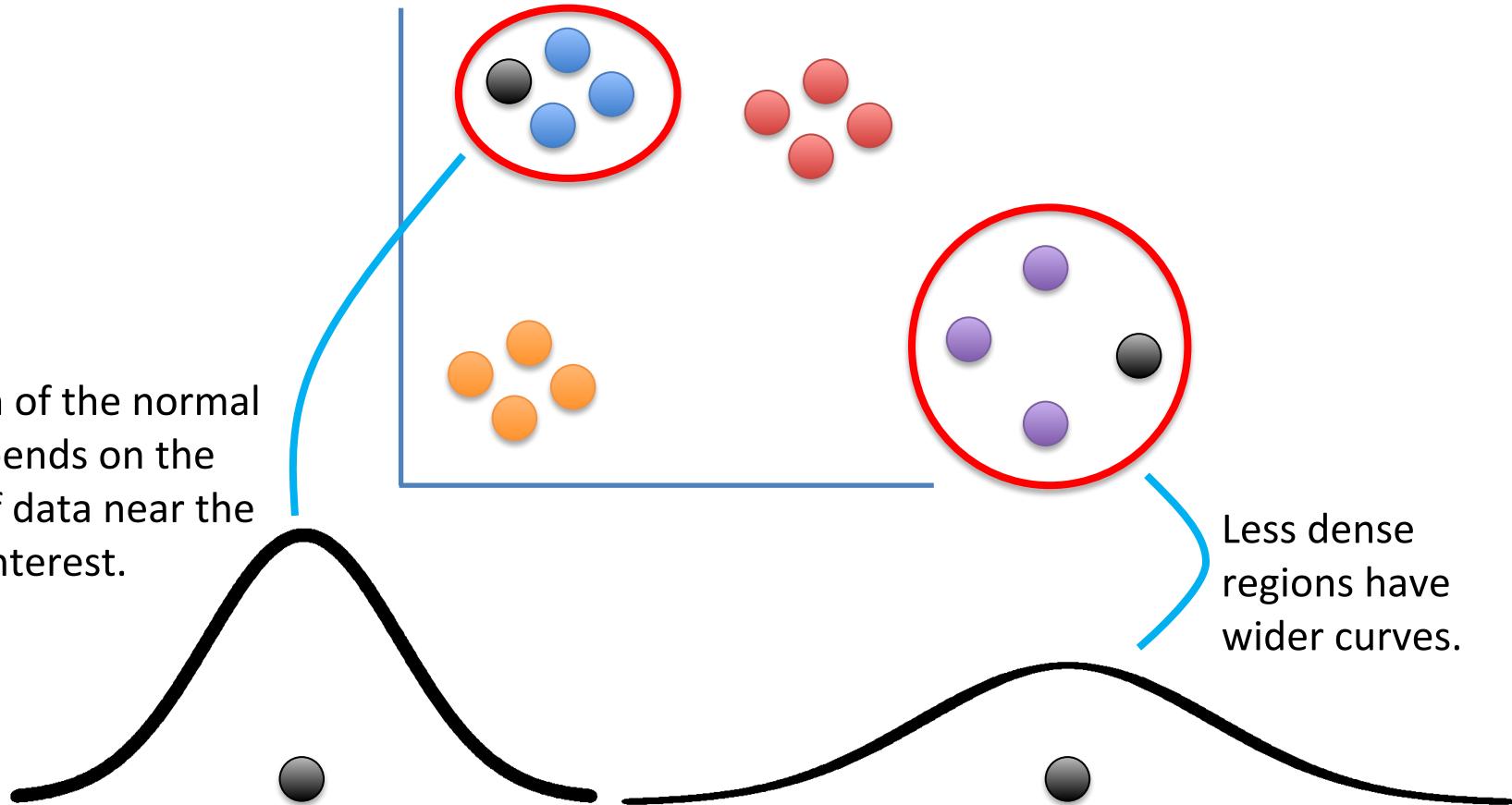


...and to illustrate the concept, I  
need to add a cluster that is half  
as dense as the others.

The width of the normal curve depends on the density of data near the point of interest.

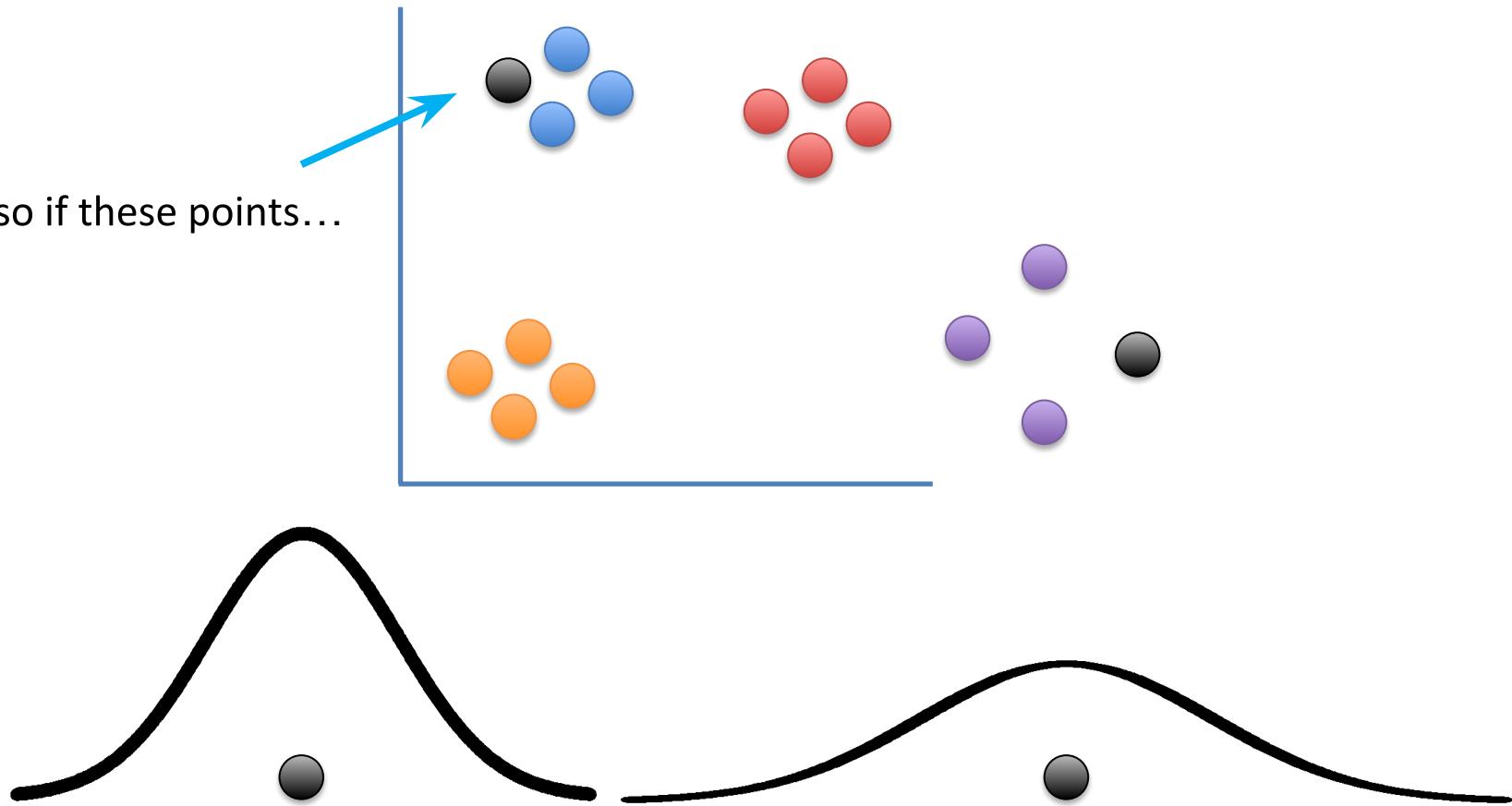


The width of the normal curve depends on the density of data near the point of interest.

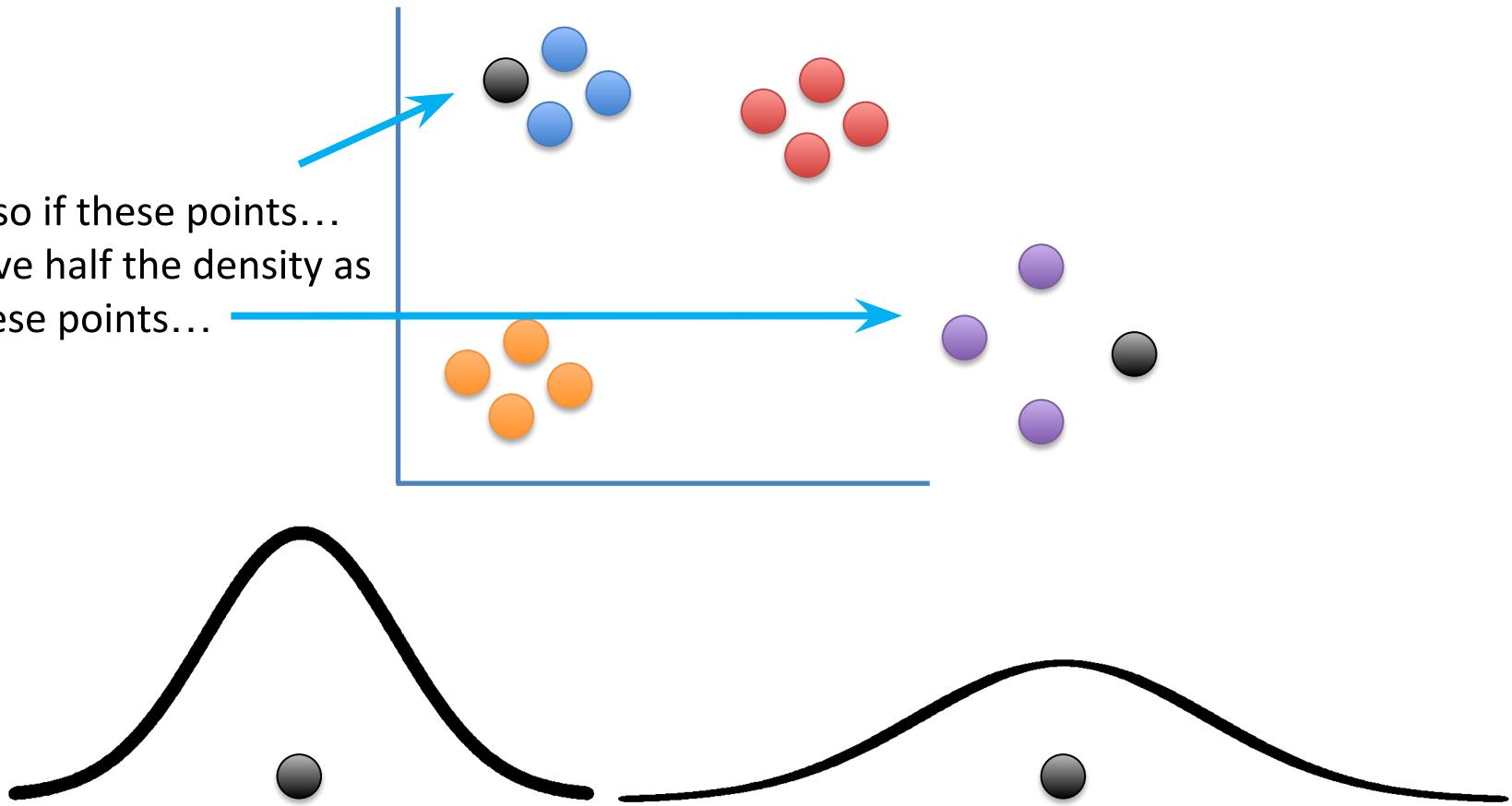


Less dense regions have wider curves.

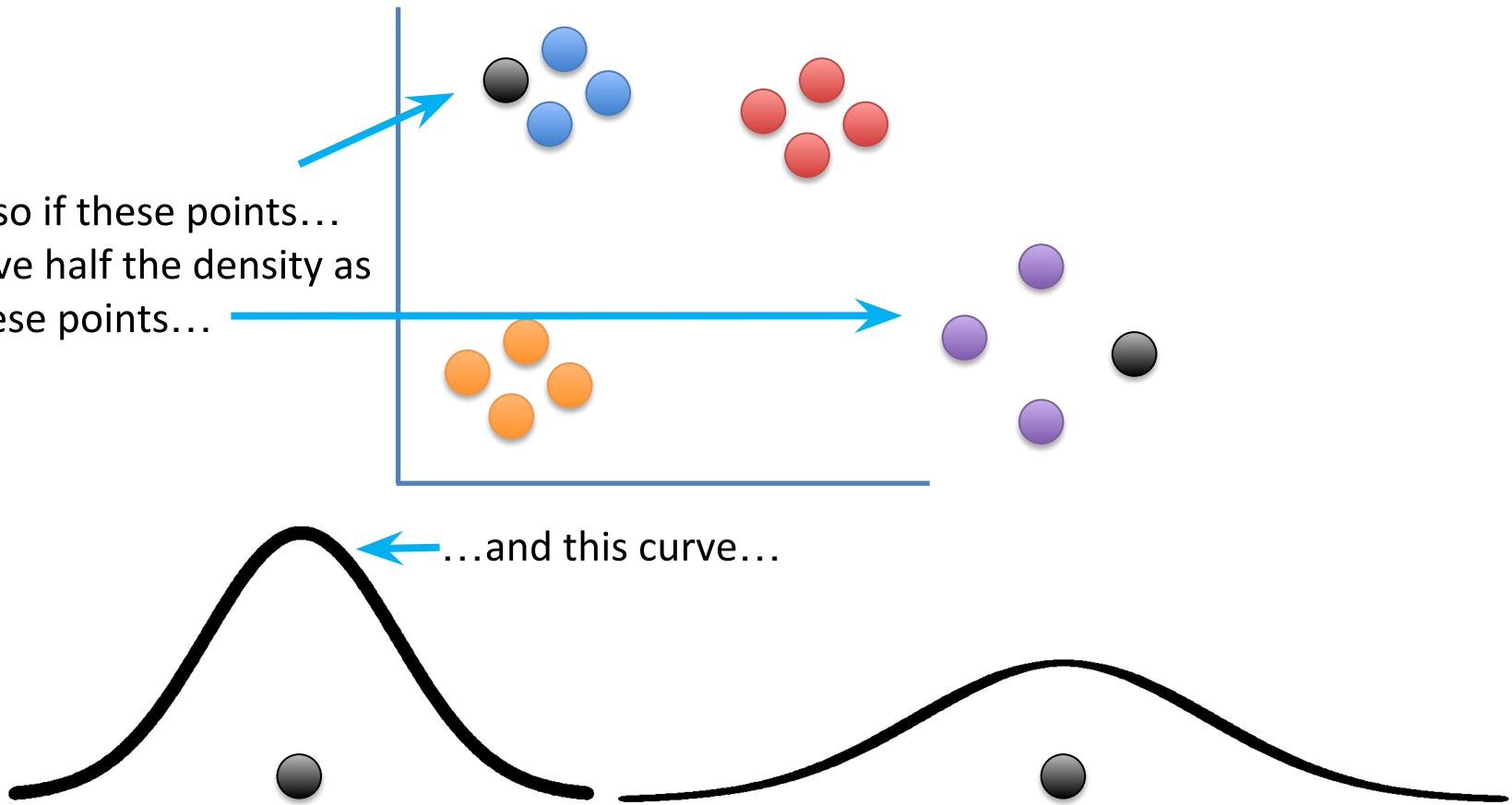
...so if these points...



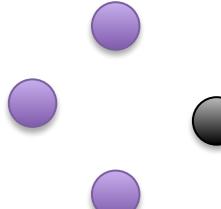
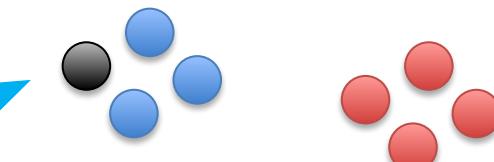
...so if these points...  
have half the density as  
these points...



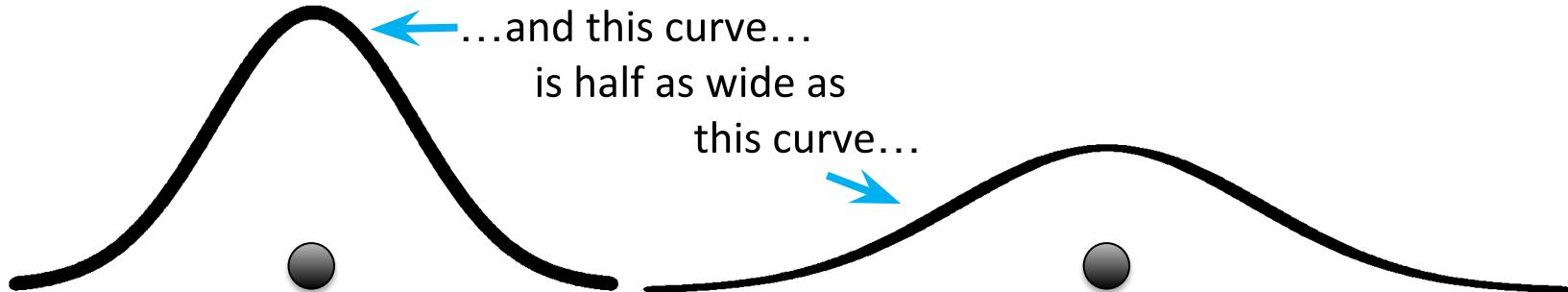
...so if these points...  
have half the density as  
these points...



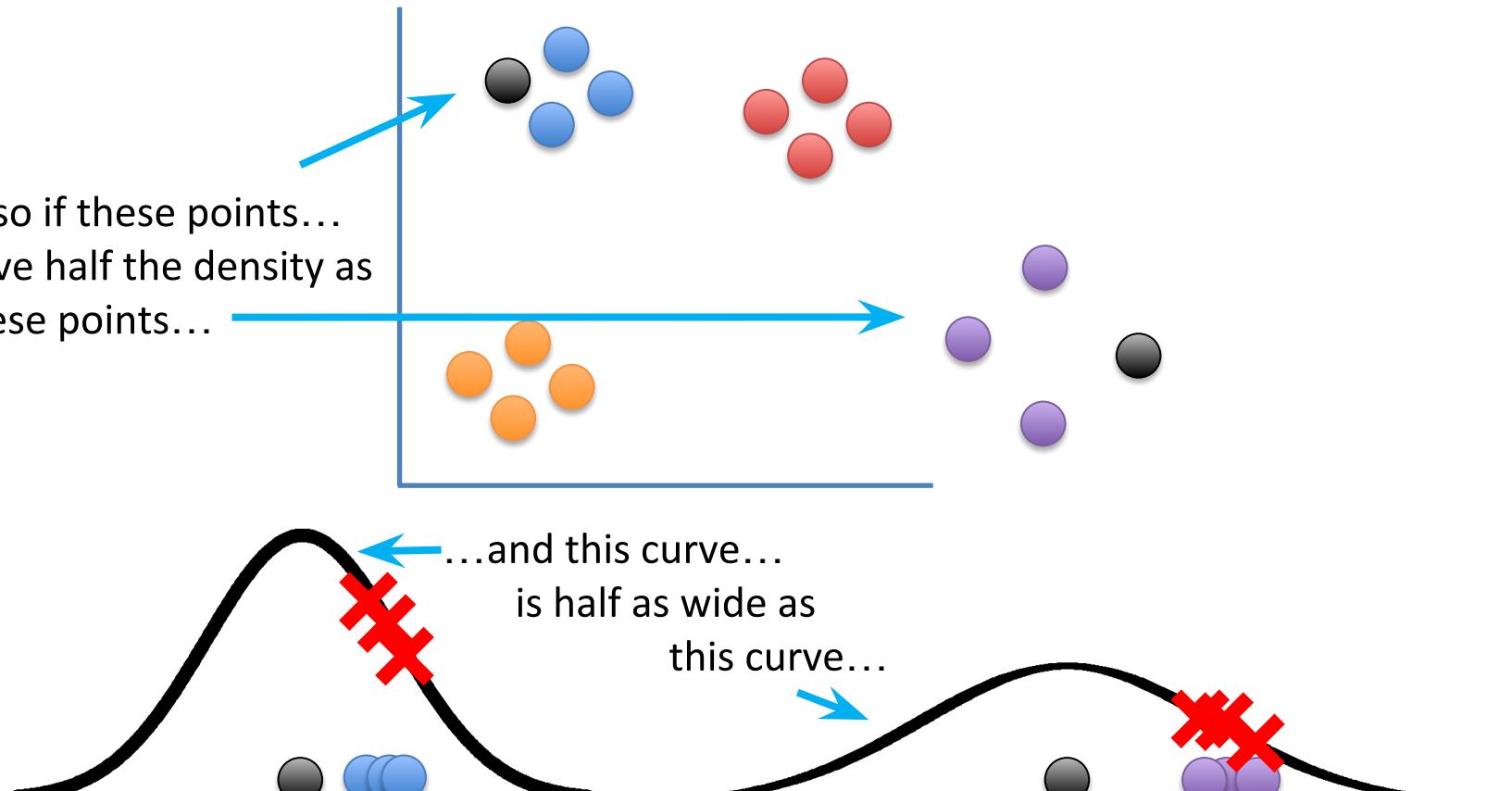
...so if these points...  
have half the density as  
these points...



...and this curve...  
is half as wide as  
this curve...



...so if these points...  
have half the density as  
these points...

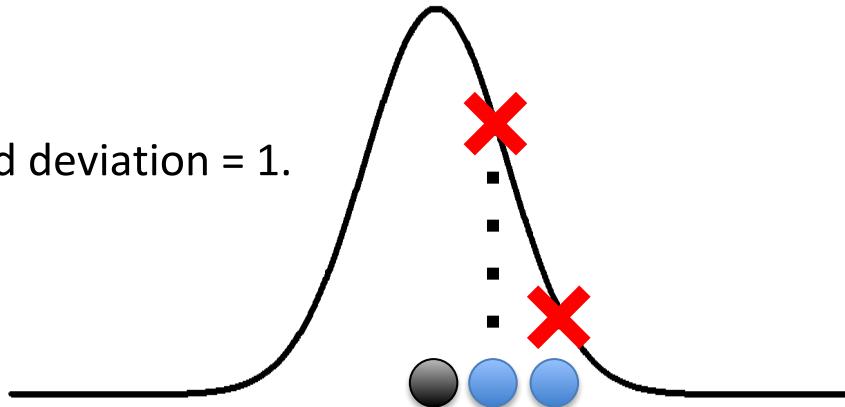


...then scaling the similarity scores will make them the same for both clusters.

Here's an  
example...

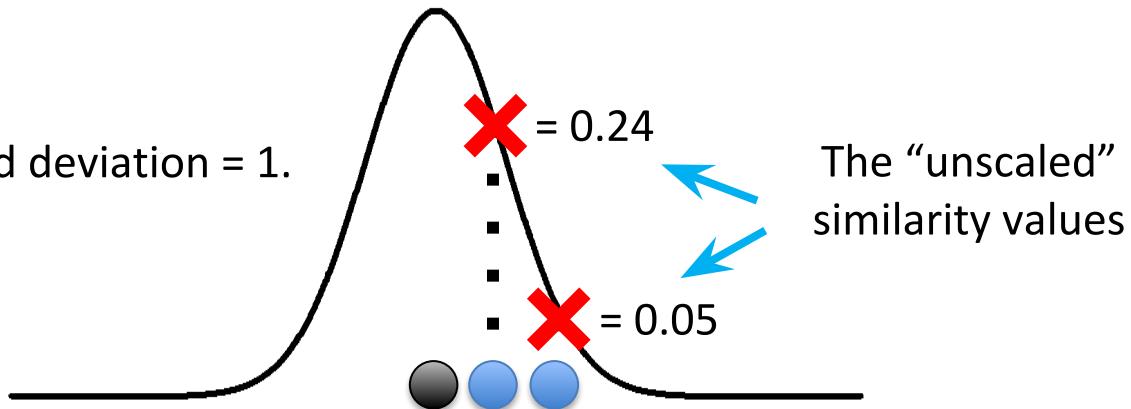
Here's an example...

This curve has a standard deviation = 1.



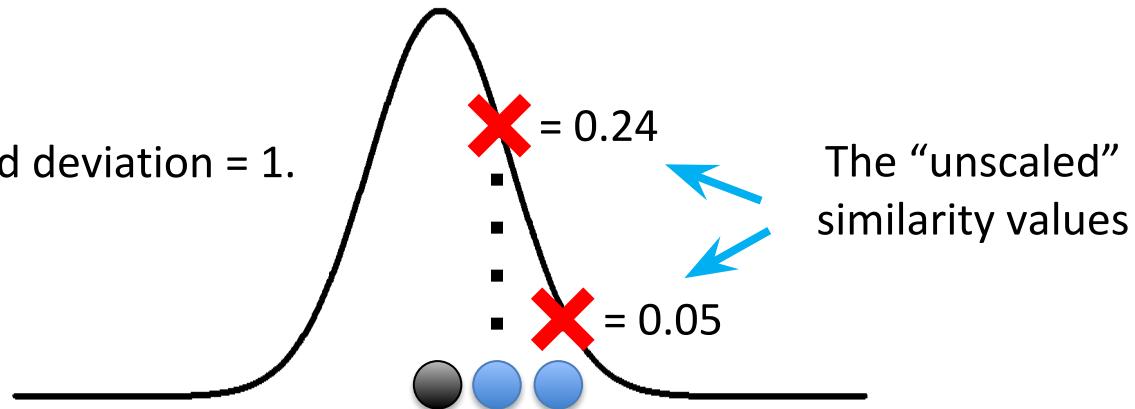
Here's an example...

This curve has a standard deviation = 1.

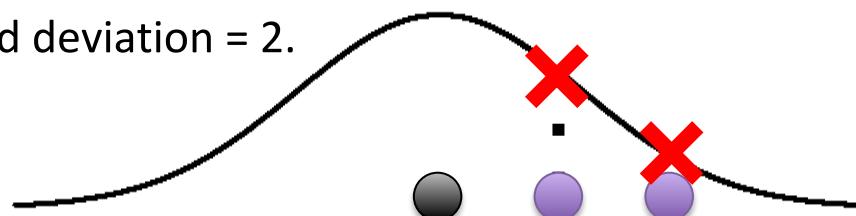


Here's an example...

This curve has a standard deviation = 1.

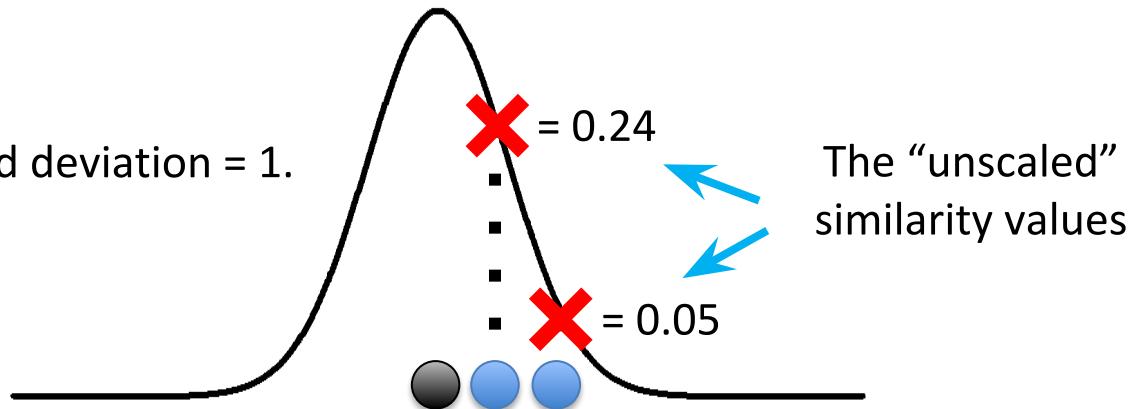


This curve has a standard deviation = 2.



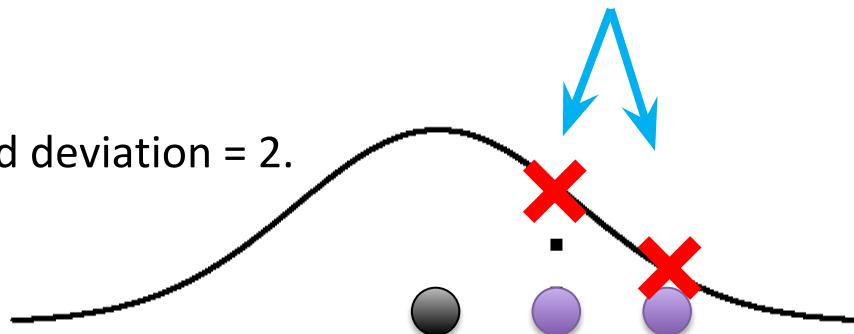
Here's an example...

This curve has a standard deviation = 1.



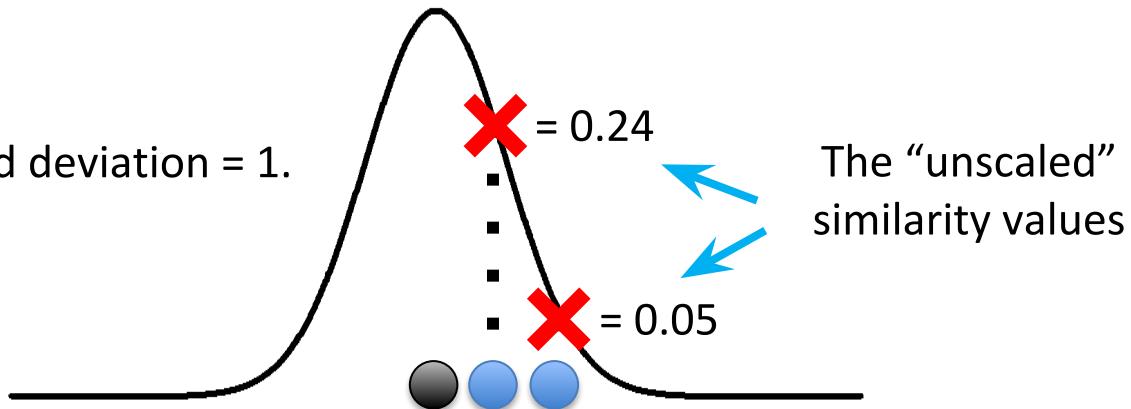
These points are twice as far from the middle.

This curve has a standard deviation = 2.



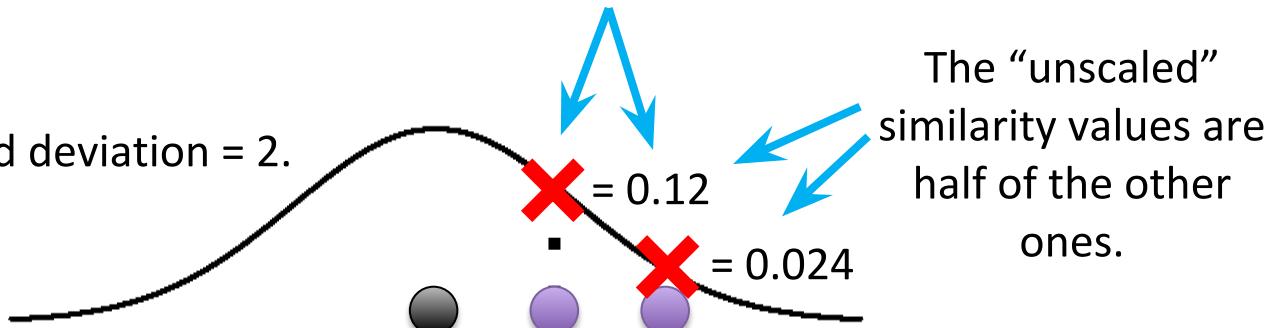
Here's an example...

This curve has a standard deviation = 1.



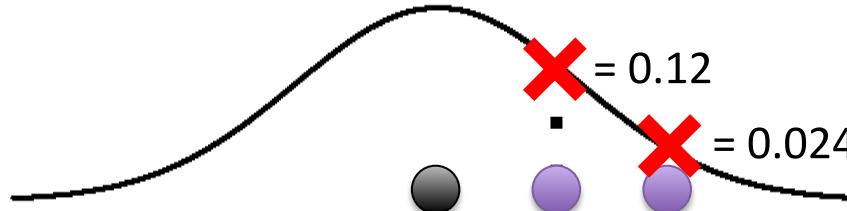
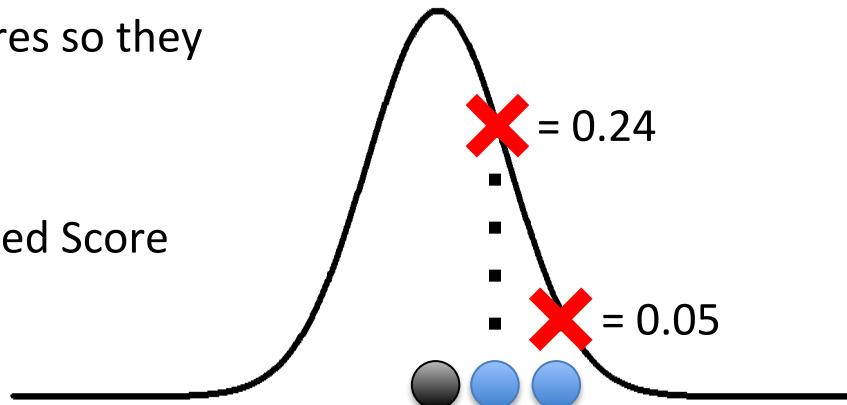
These points are twice as far from the middle.

This curve has a standard deviation = 2.



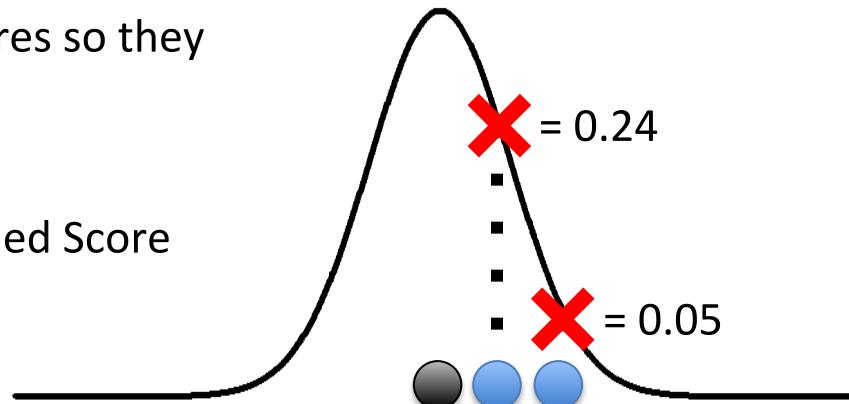
To scale the similarity scores so they sum to 1:

$$\frac{\text{Score}}{\text{Sum of all scores}} = \text{Scaled Score}$$



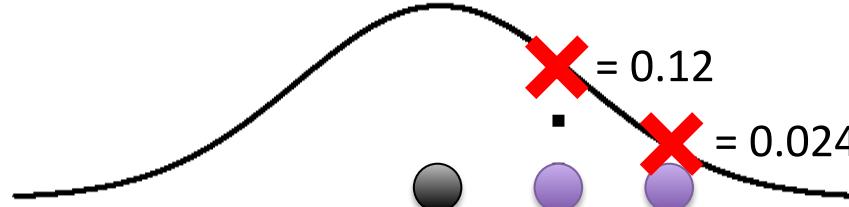
To scale the similarity scores so they sum to 1:

$$\frac{\text{Score}}{\text{Sum of all scores}} = \text{Scaled Score}$$



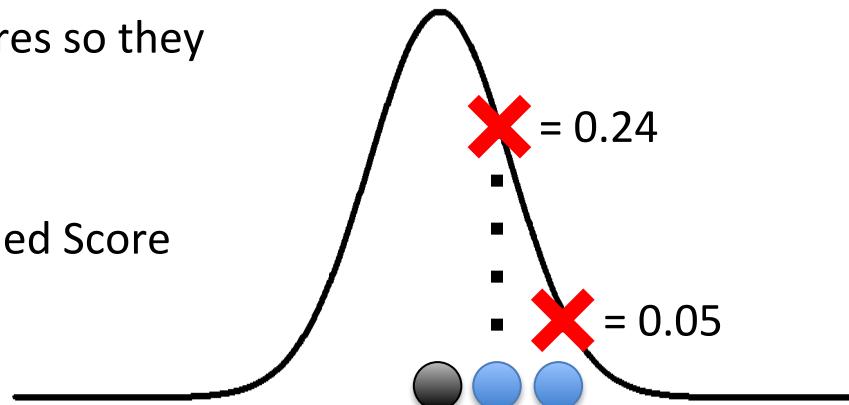
$$\frac{0.24}{0.24 + 0.05} = 0.82$$

$$\frac{0.05}{0.24 + 0.05} = 0.18$$



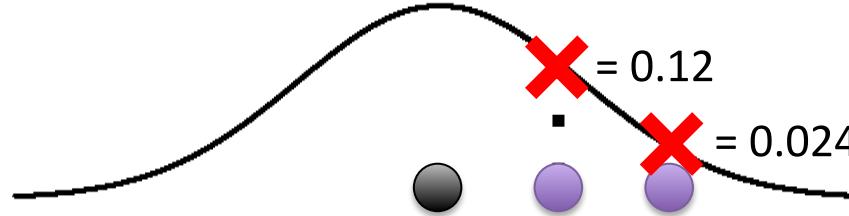
To scale the similarity scores so they sum to 1:

$$\frac{\text{Score}}{\text{Sum of all scores}} = \text{Scaled Score}$$



$$\frac{0.24}{0.24 + 0.05} = 0.82$$

$$\frac{0.05}{0.24 + 0.05} = 0.18$$

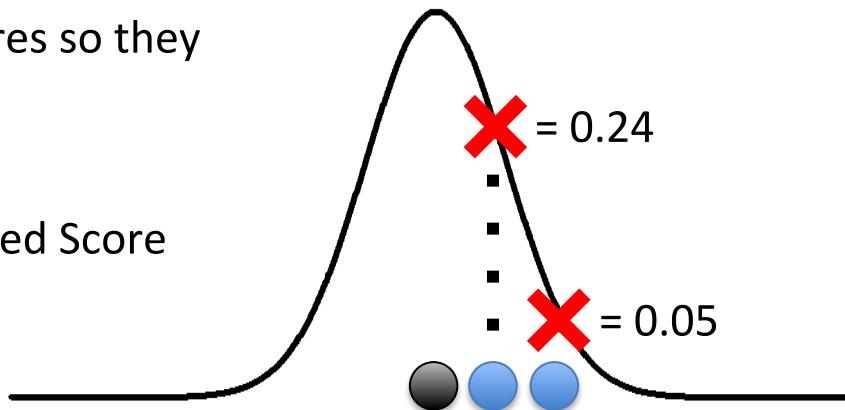


$$\frac{0.12}{0.12 + 0.024} = 0.82$$

$$\frac{0.024}{0.12 + 0.024} = 0.18$$

To scale the similarity scores so they sum to 1:

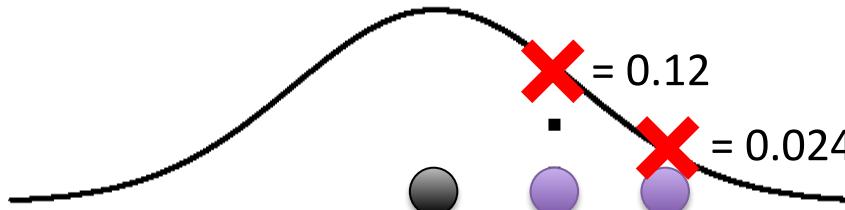
$$\frac{\text{Score}}{\text{Sum of all scores}} = \text{Scaled Score}$$



$$\frac{0.24}{0.24 + 0.05} = 0.82$$

$$\frac{0.05}{0.24 + 0.05} = 0.18$$

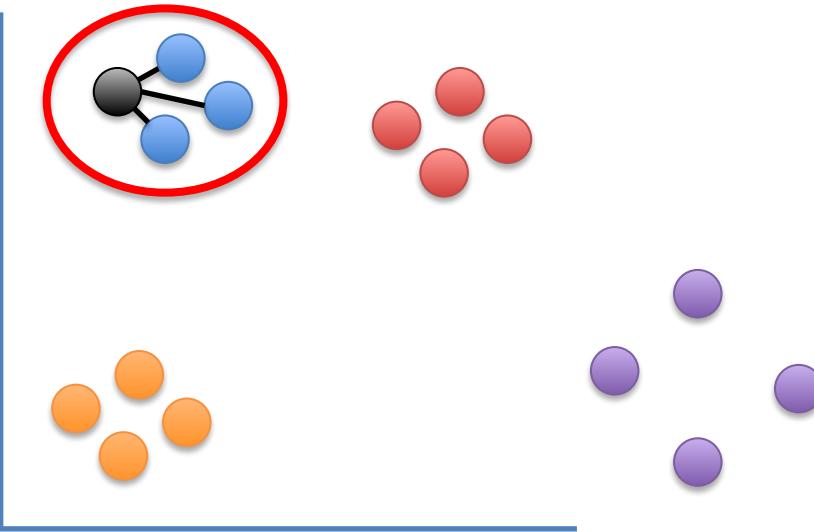
These are the same as these!



$$\frac{0.12}{0.12 + 0.024} = 0.82$$

$$\frac{0.024}{0.12 + 0.024} = 0.18$$

That implies that the scaled similarity scores for this relatively tight cluster...



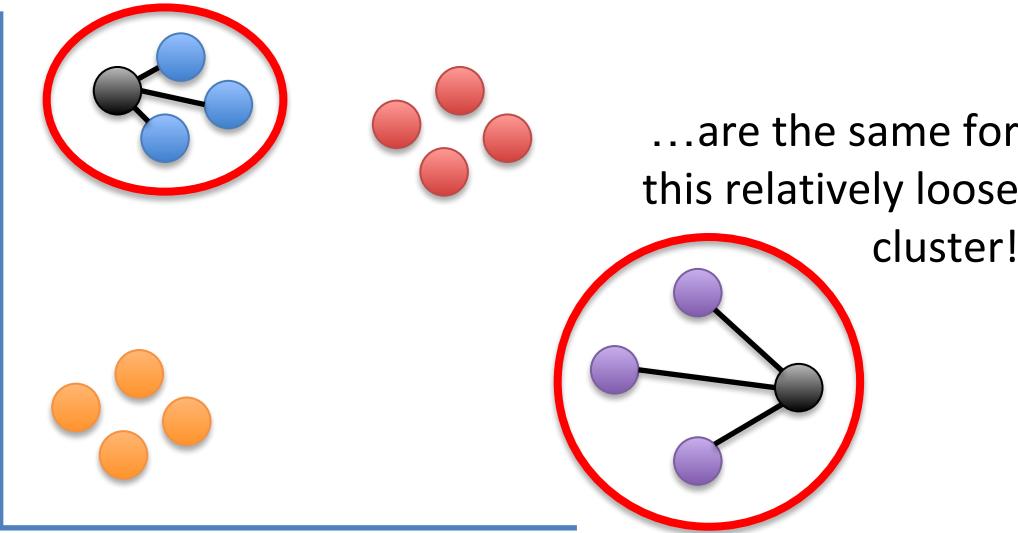
$$\frac{0.24}{0.24 + 0.05} = 0.82$$

$$\frac{0.05}{0.24 + 0.05} = 0.18$$

$$\frac{0.12}{0.12 + 0.024} = 0.82$$

$$\frac{0.024}{0.12 + 0.024} = 0.18$$

That implies that the scaled similarity scores for this relatively tight cluster...



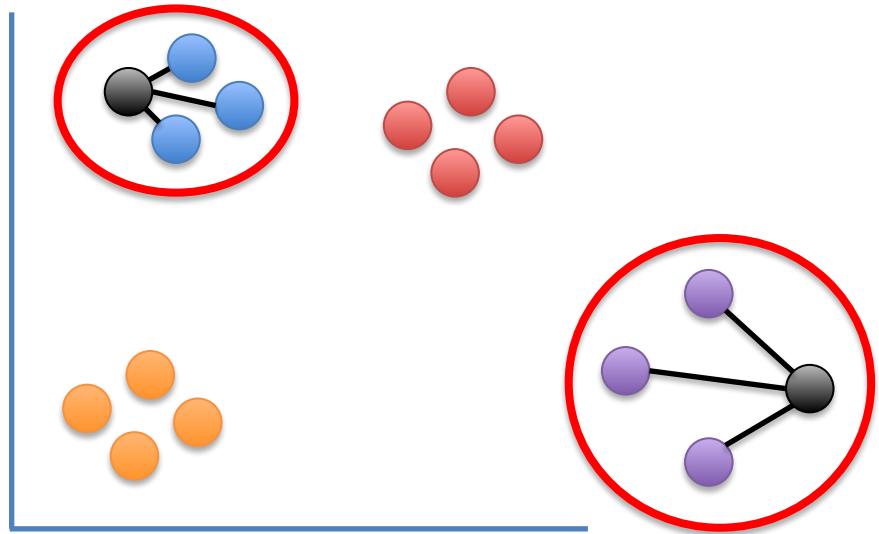
$$\frac{0.24}{0.24 + 0.05} = 0.82$$

$$\frac{0.05}{0.24 + 0.05} = 0.18$$

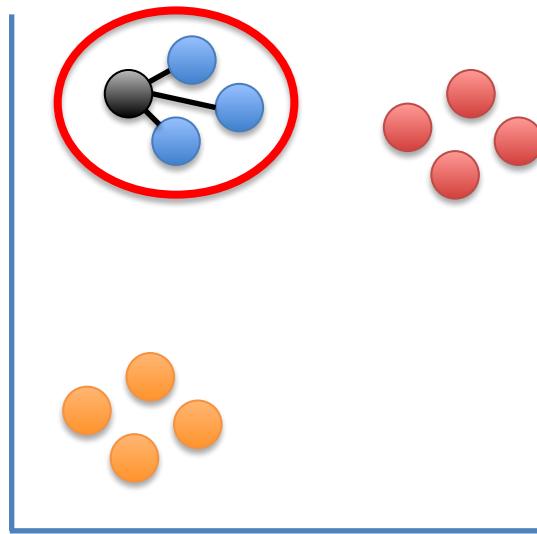
$$\frac{0.12}{0.12 + 0.024} = 0.82$$

$$\frac{0.024}{0.12 + 0.024} = 0.18$$

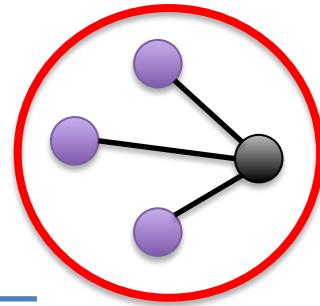
The reality is a little more complicated, but only slightly.



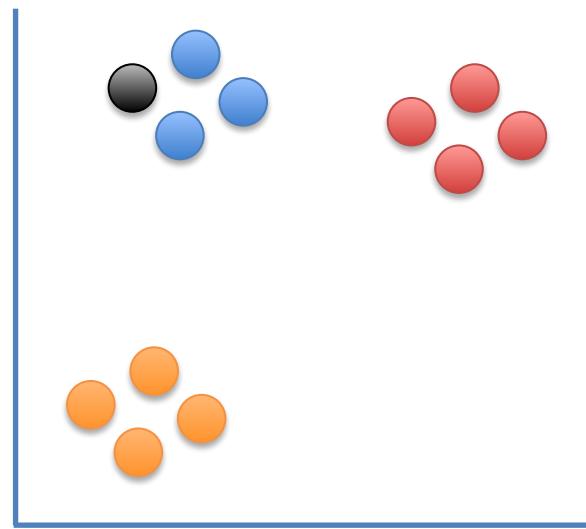
The reality is a little more complicated, but only slightly.



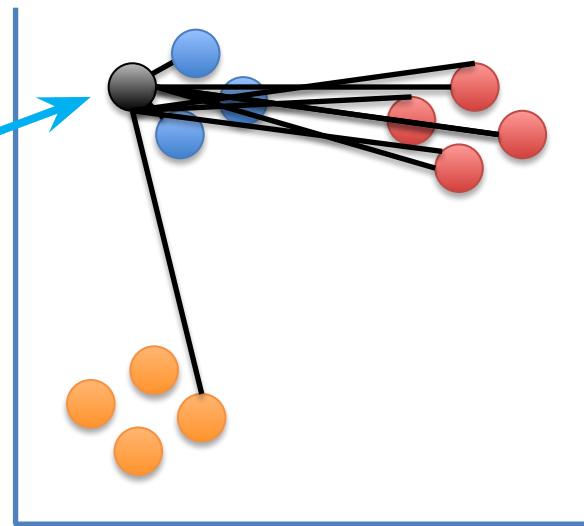
t-SNE has a “perplexity” parameter equal to the expected density, and that comes into play, but these clusters are still more “similar” than you might expect.



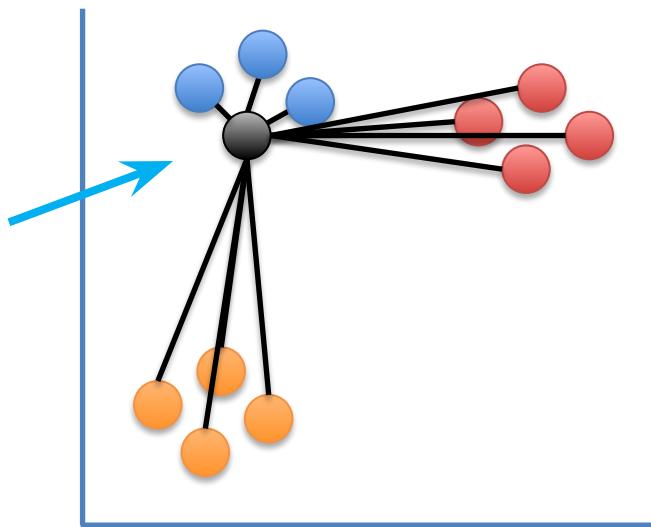
Now back to the original  
scatter plot...



We've calculated  
similarity scores for this  
point.

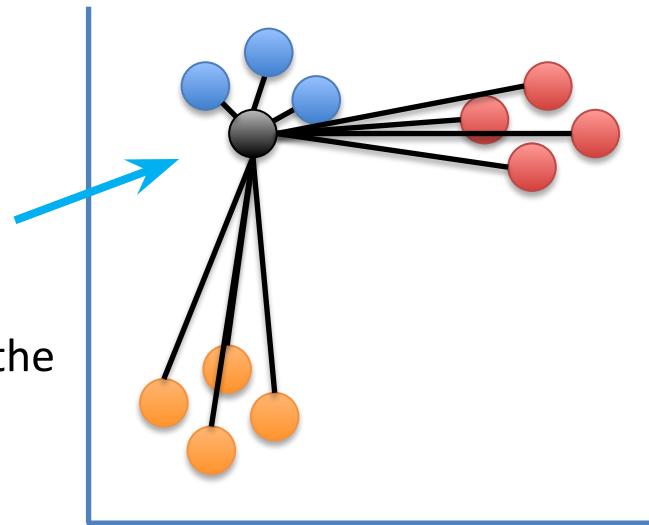


Now we do it for this point...

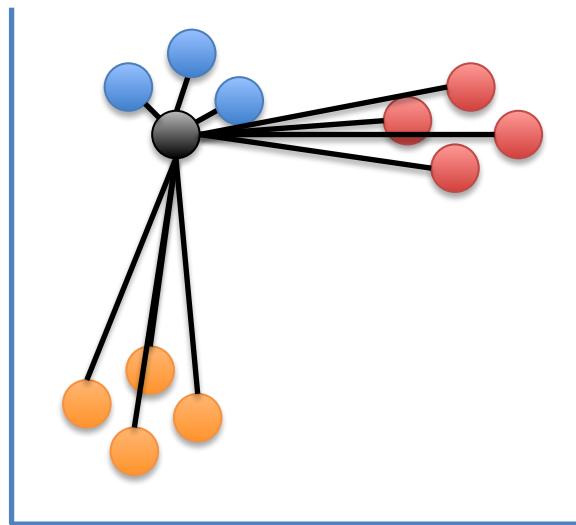


Now we do it for this  
point...

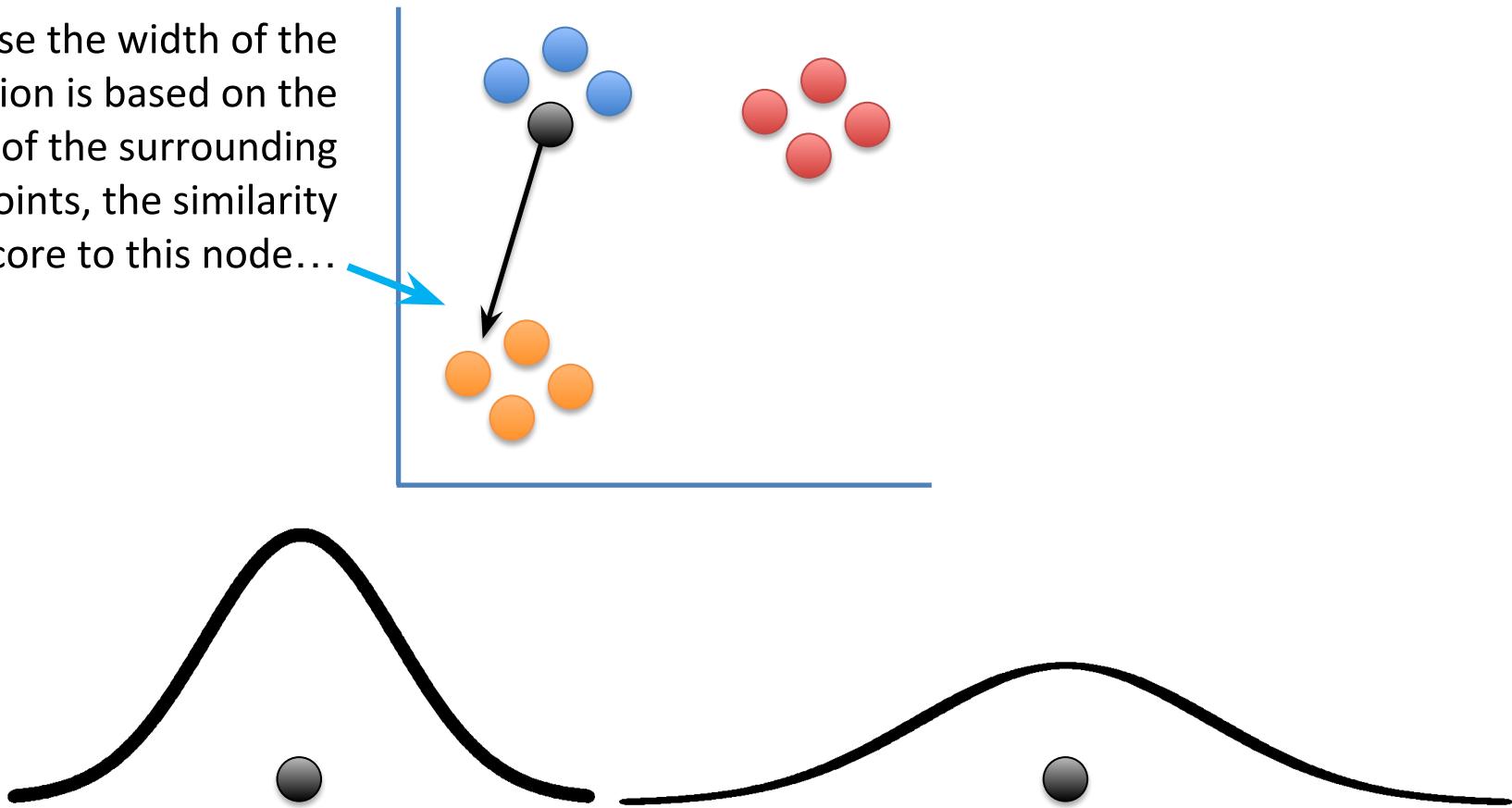
...and we do it for all the  
points.



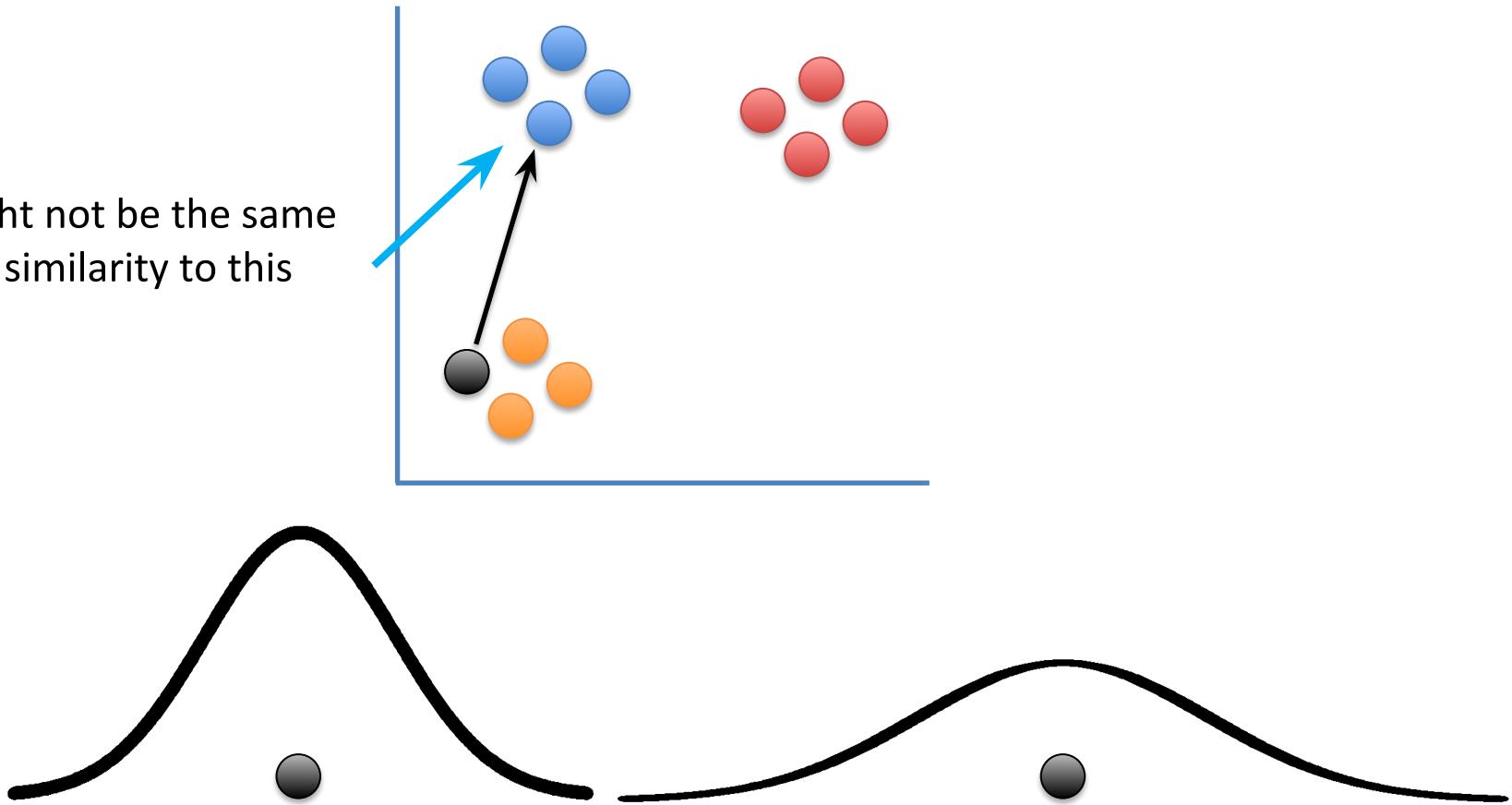
One last thing and the scatter plot will be all set with similarity scores!!!

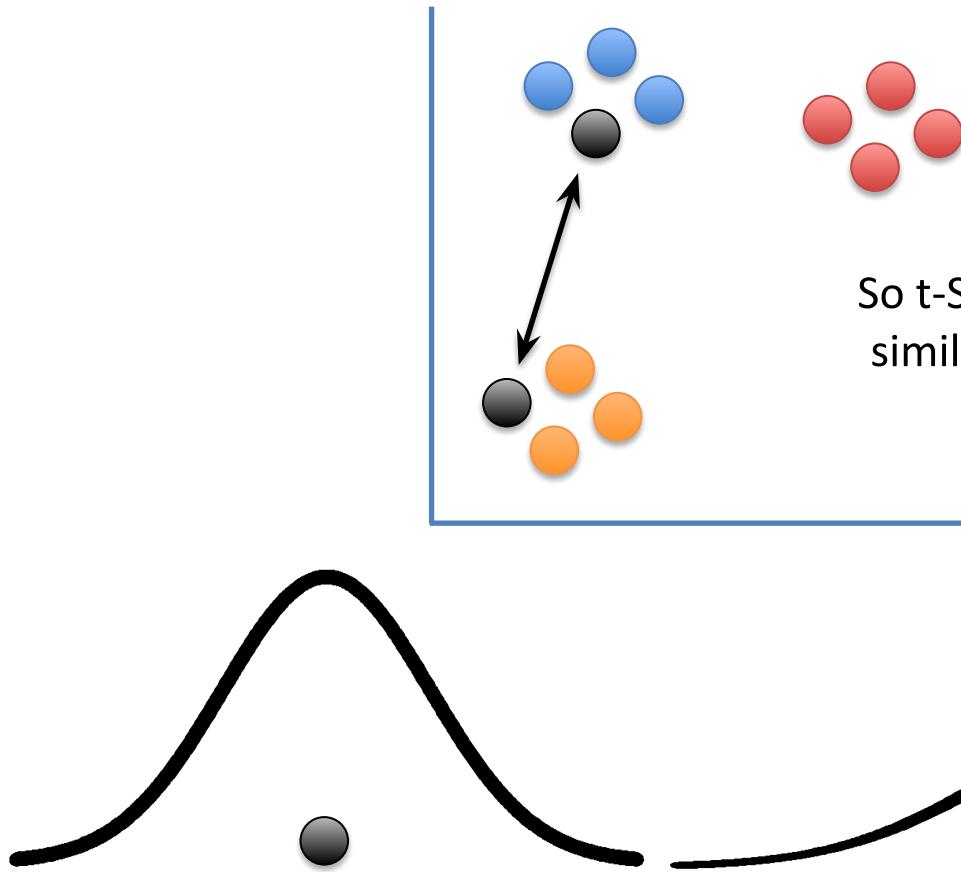


Because the width of the distribution is based on the density of the surrounding data points, the similarity score to this node...

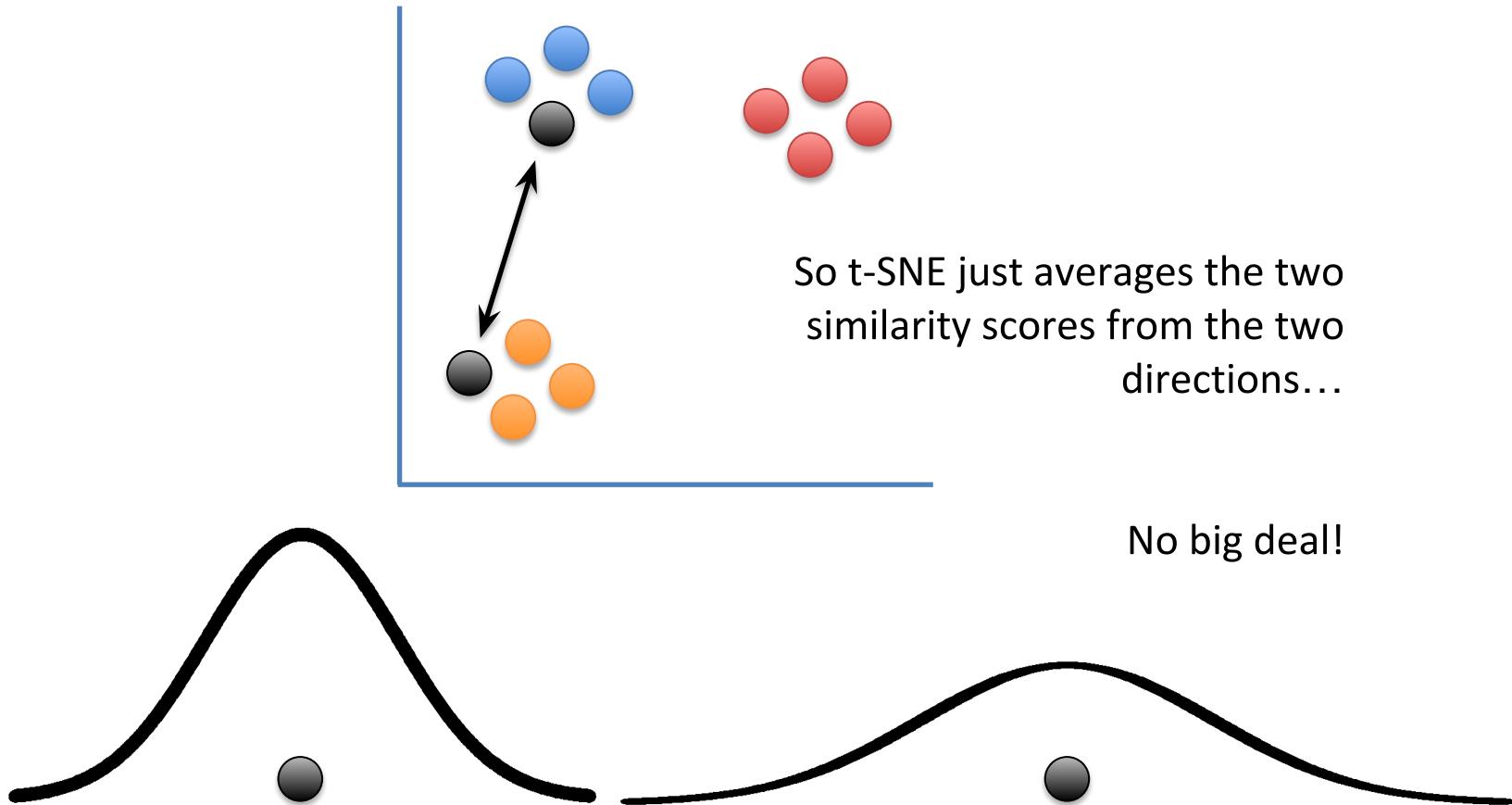


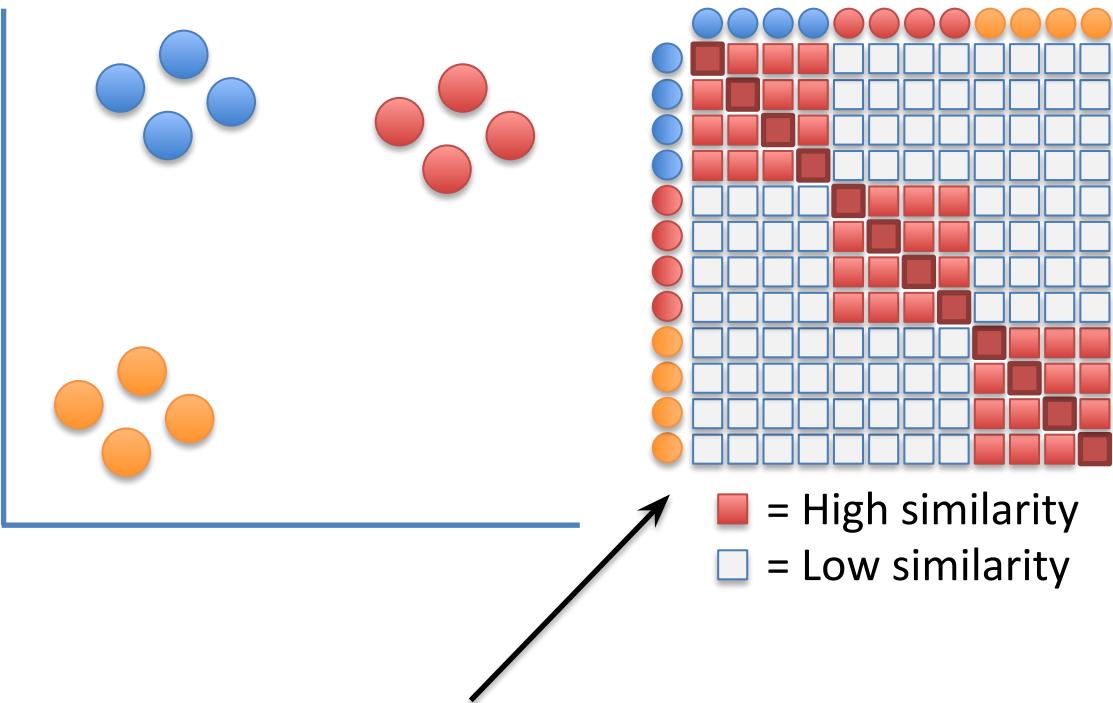
...might not be the same  
as the similarity to this  
node.



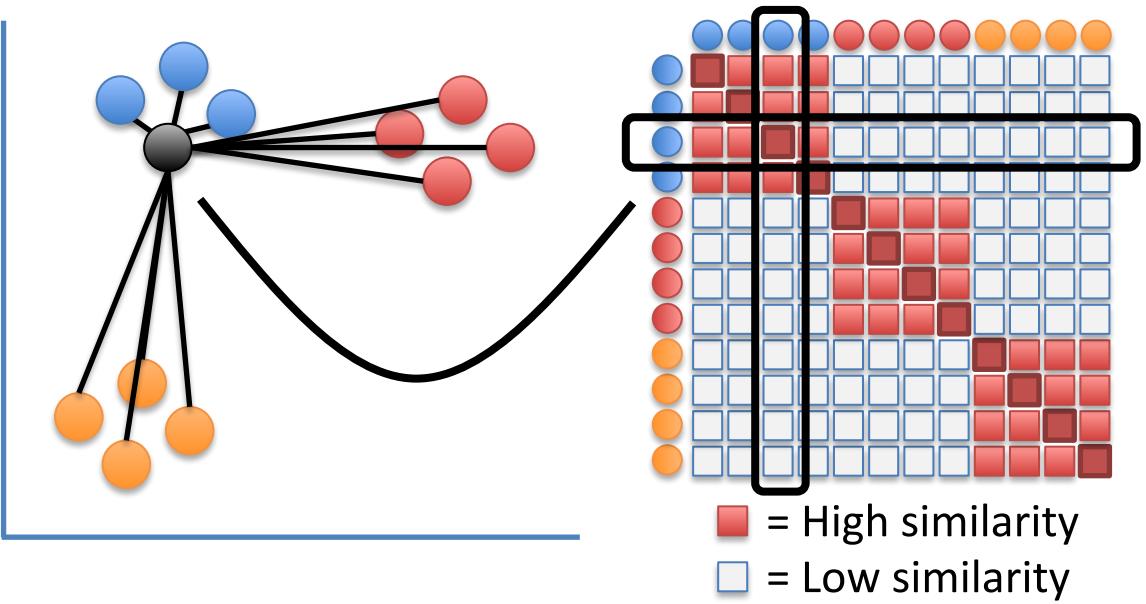


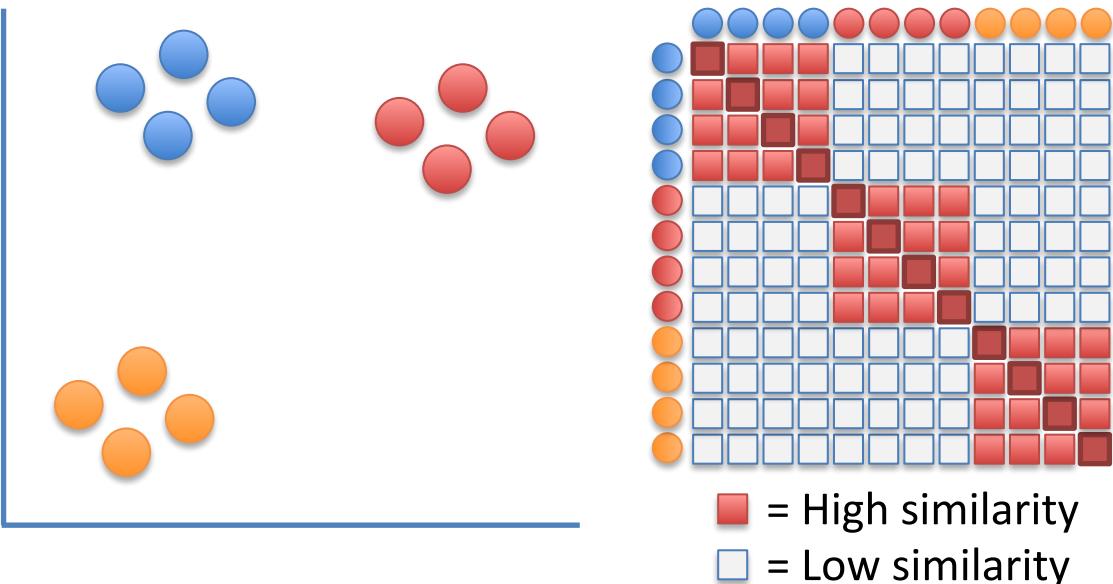
So t-SNE just averages the two similarity scores from the two directions...





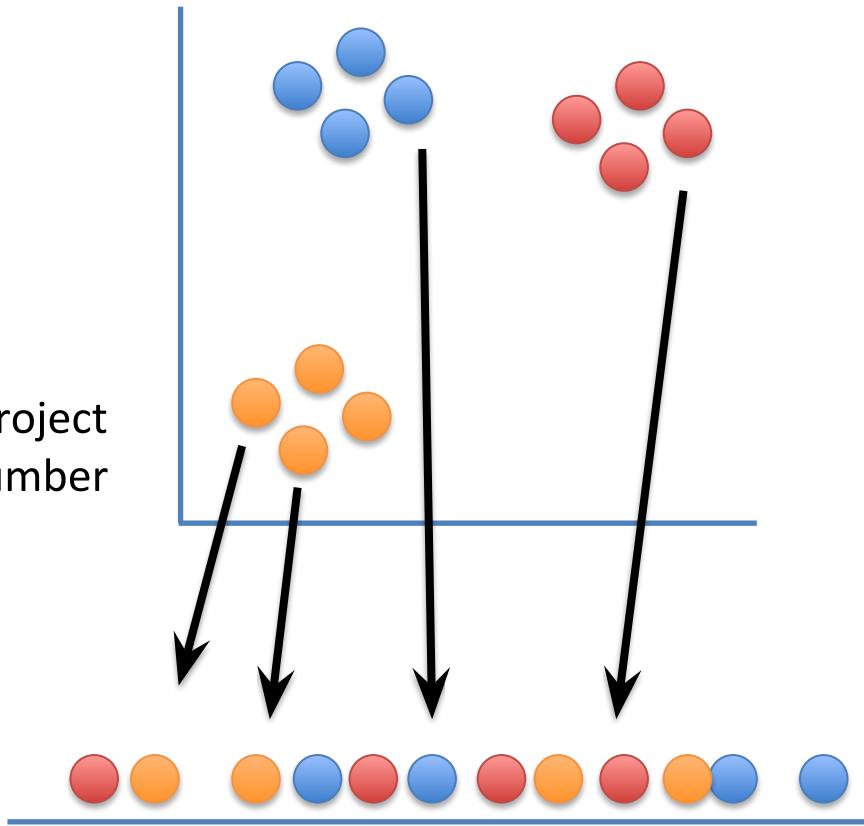
Ultimately, you end up with a matrix of similarity scores.



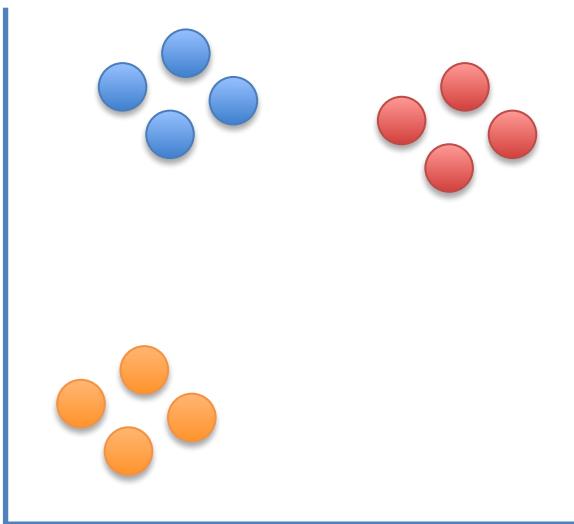


Hooray!!! We're done doing calculating similarity scores for the scatter plot!

Now we randomly project  
the data onto the number  
line...



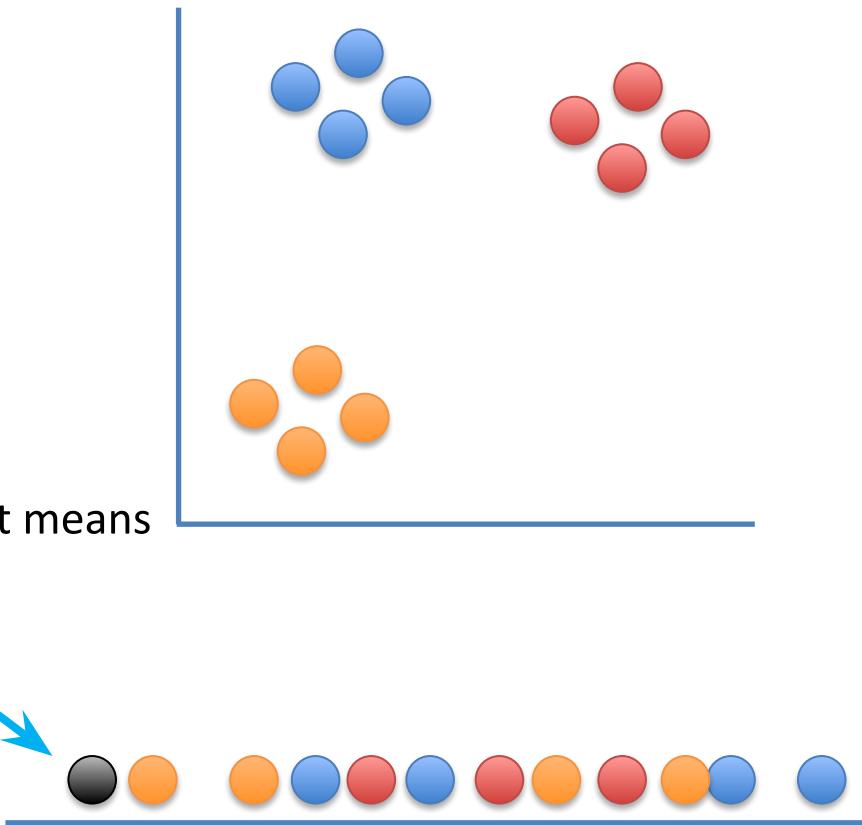
Now we randomly project  
the data onto the number  
line...



... and calculate  
similarity scores for  
the points on the  
number line.



Just like before, that means  
picking a point...



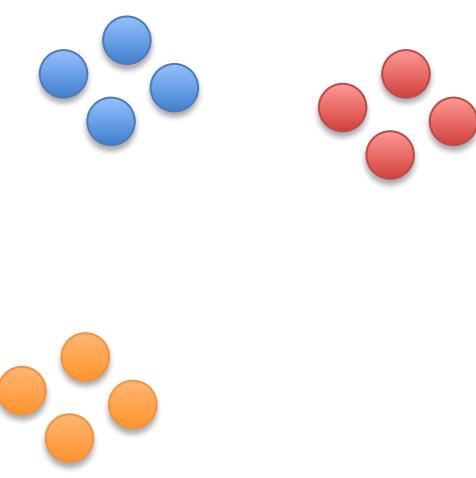
Just like before, that means  
picking a point...

...measuring a  
distance...

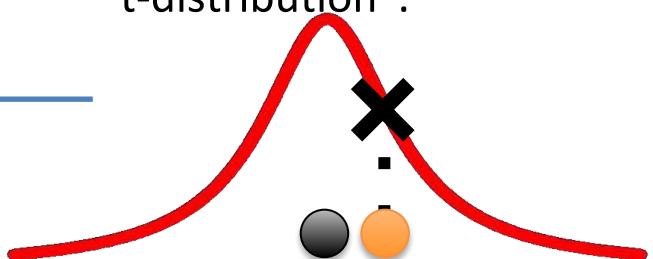


Just like before, that means picking a point...

...measuring a distance...

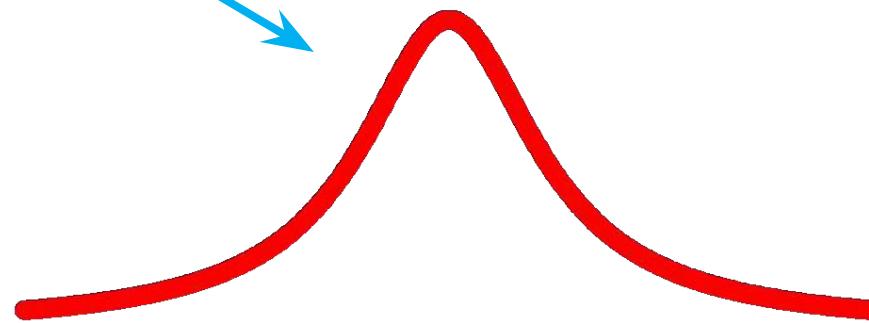


...and lastly, drawing a line from the point to a curve. However, this time we're using a “t-distribution”.



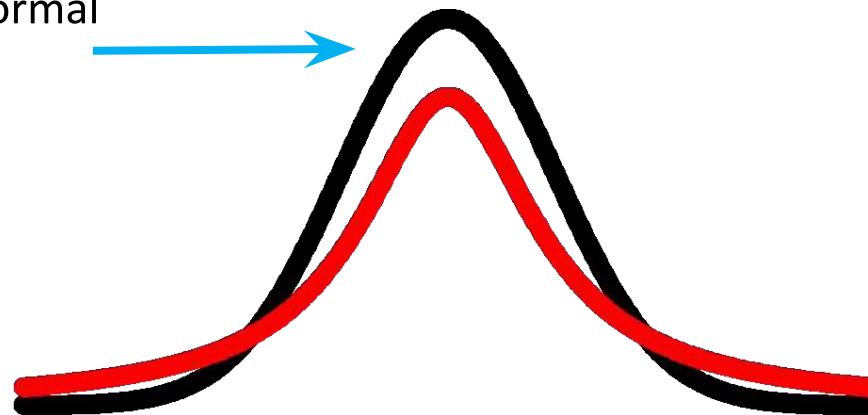
A

“t-distribution”...



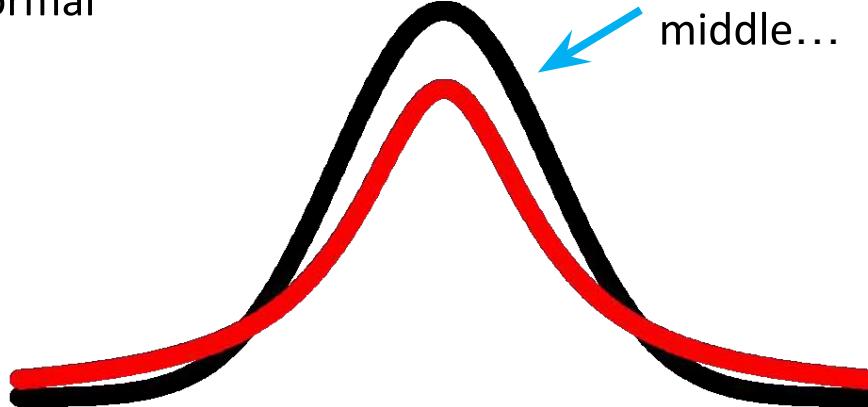
A “t-distribution”...

...is a lot like a normal distribution



A “t-distribution”...

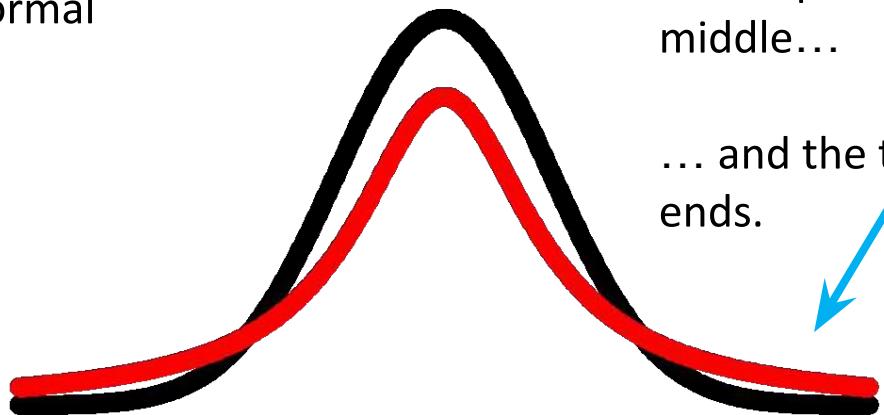
...is a lot like a normal distribution...



...except the “t” isn’t as tall in the middle...

A “t-distribution”...

...is a lot like a normal distribution...



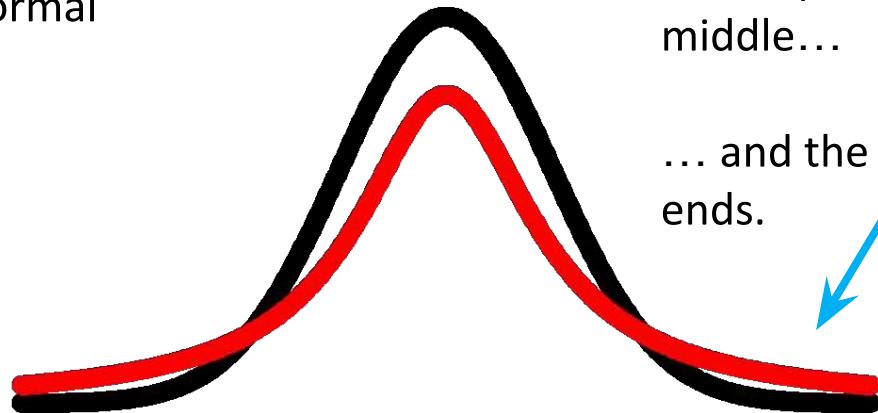
...except the “t” isn’t as tall in the middle...

... and the tails are taller on the ends.



A “t-distribution”...

...is a lot like a normal distribution...



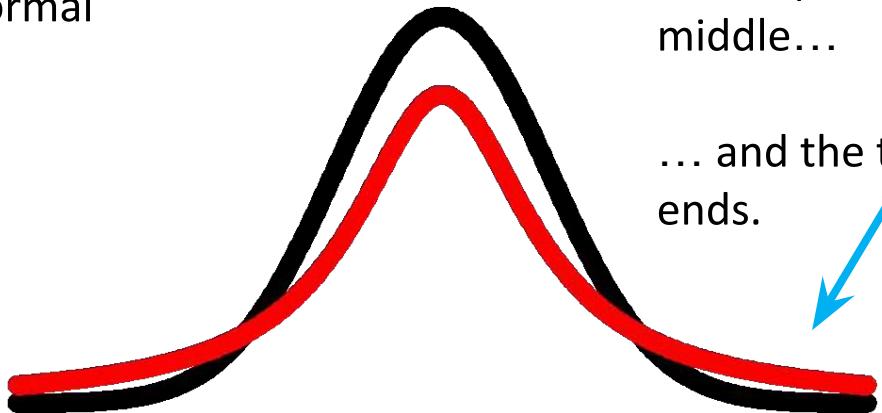
...except the “t” isn’t as tall in the middle...

... and the tails are taller on the ends.

The “t-distribution” is the “t” in t-SNE.

A “t-distribution”...

...is a lot like a normal distribution...

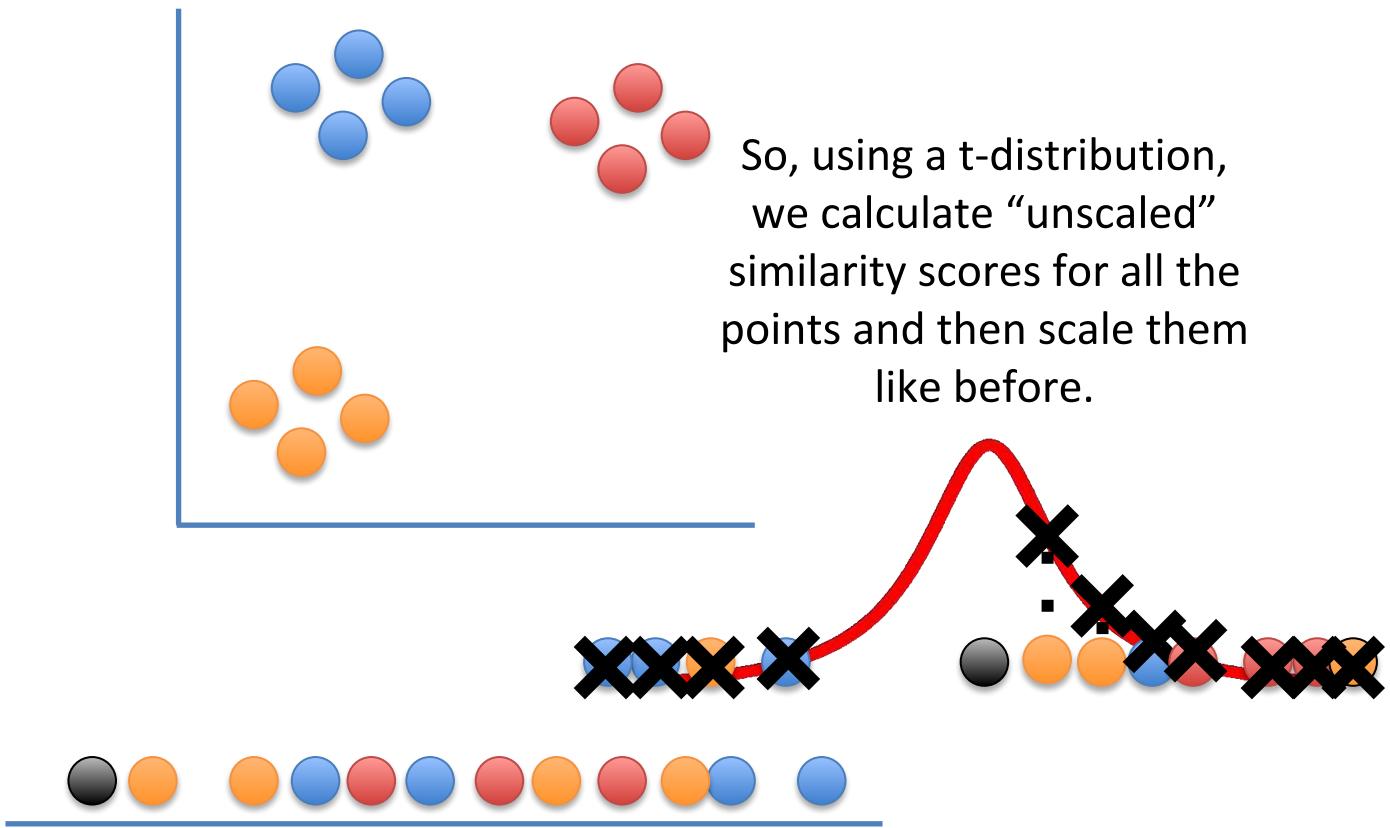


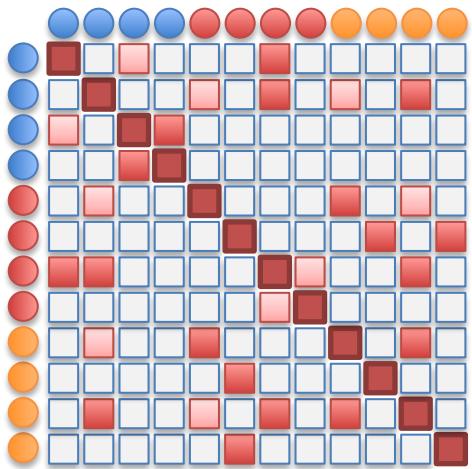
...except the “t” isn’t as tall in the middle...

... and the tails are taller on the ends.

The “t-distribution” is the “t” in t-SNE.

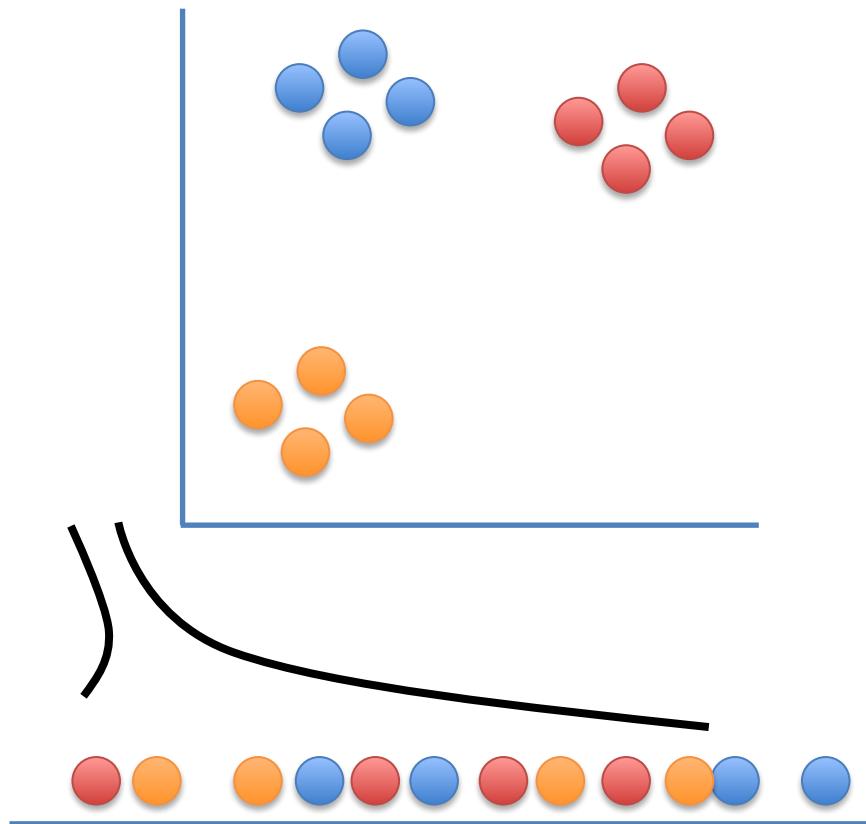
We'll talk about why the t-distribution is used in a bit...

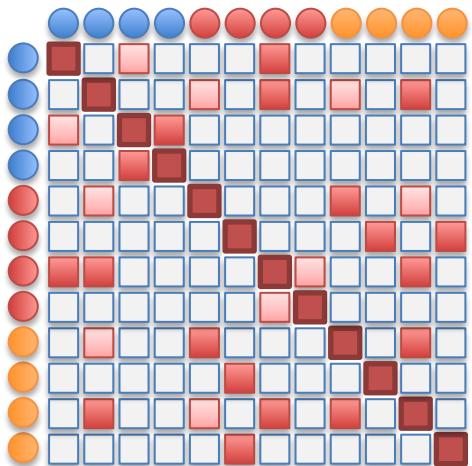




■ = High similarity  
□ = Low similarity

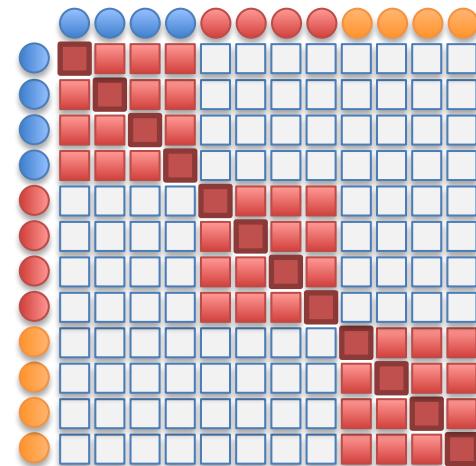
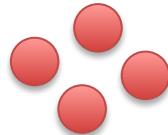
Like before, we end up with a matrix of similarity scores, but this matrix is a mess...





■ = High similarity  
□ = Low similarity

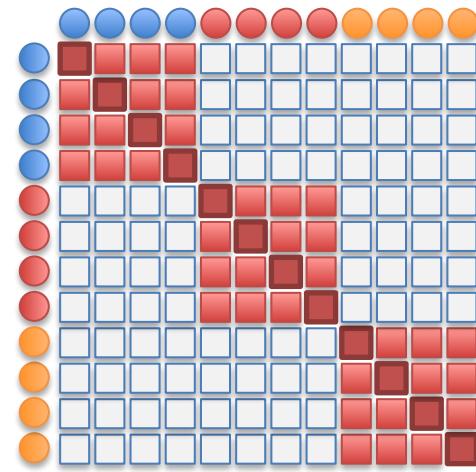
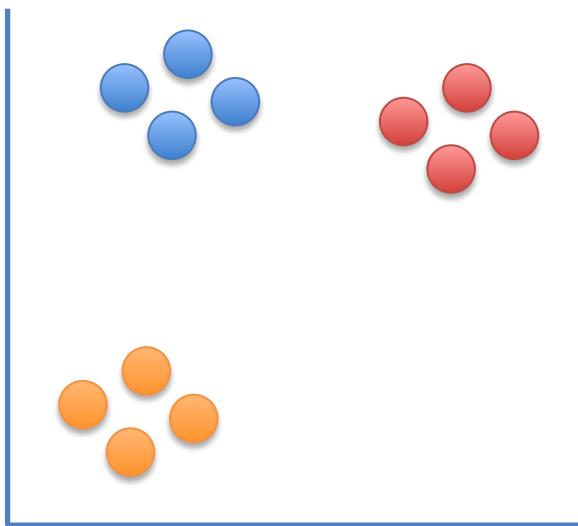
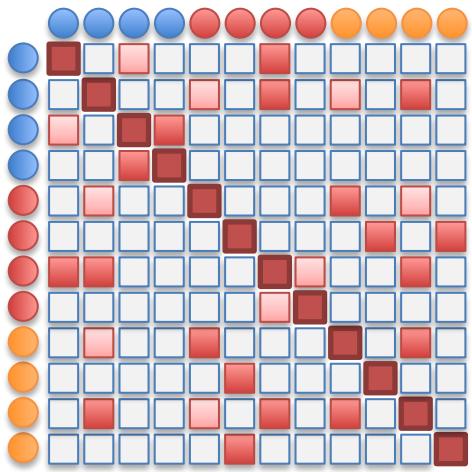
Like before, we end up with a matrix of similarity scores, but this matrix is a mess...



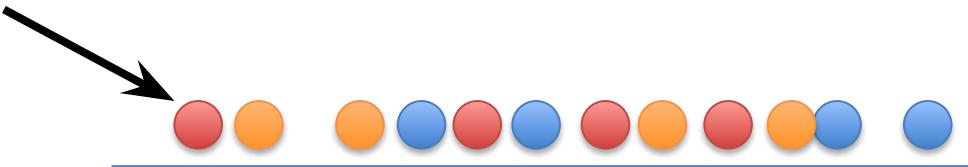
■ = High similarity  
□ = Low similarity

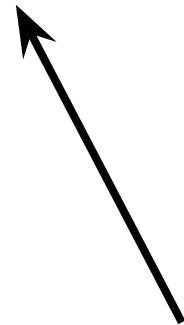
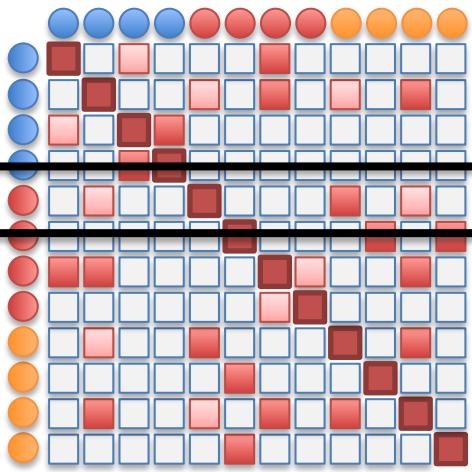
...compared to the original matrix.





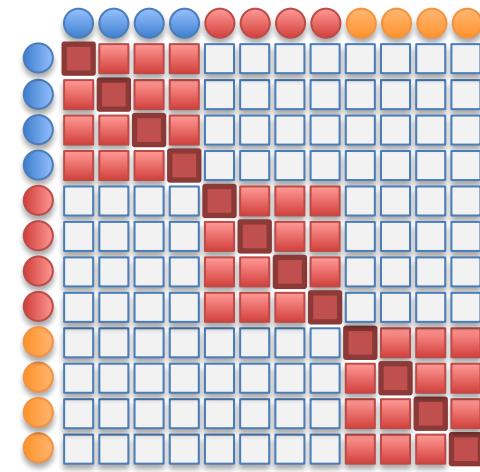
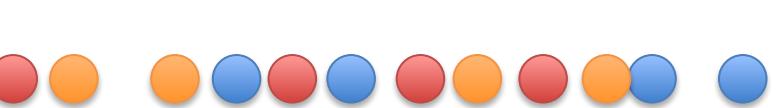
The goal of moving this point  
is...

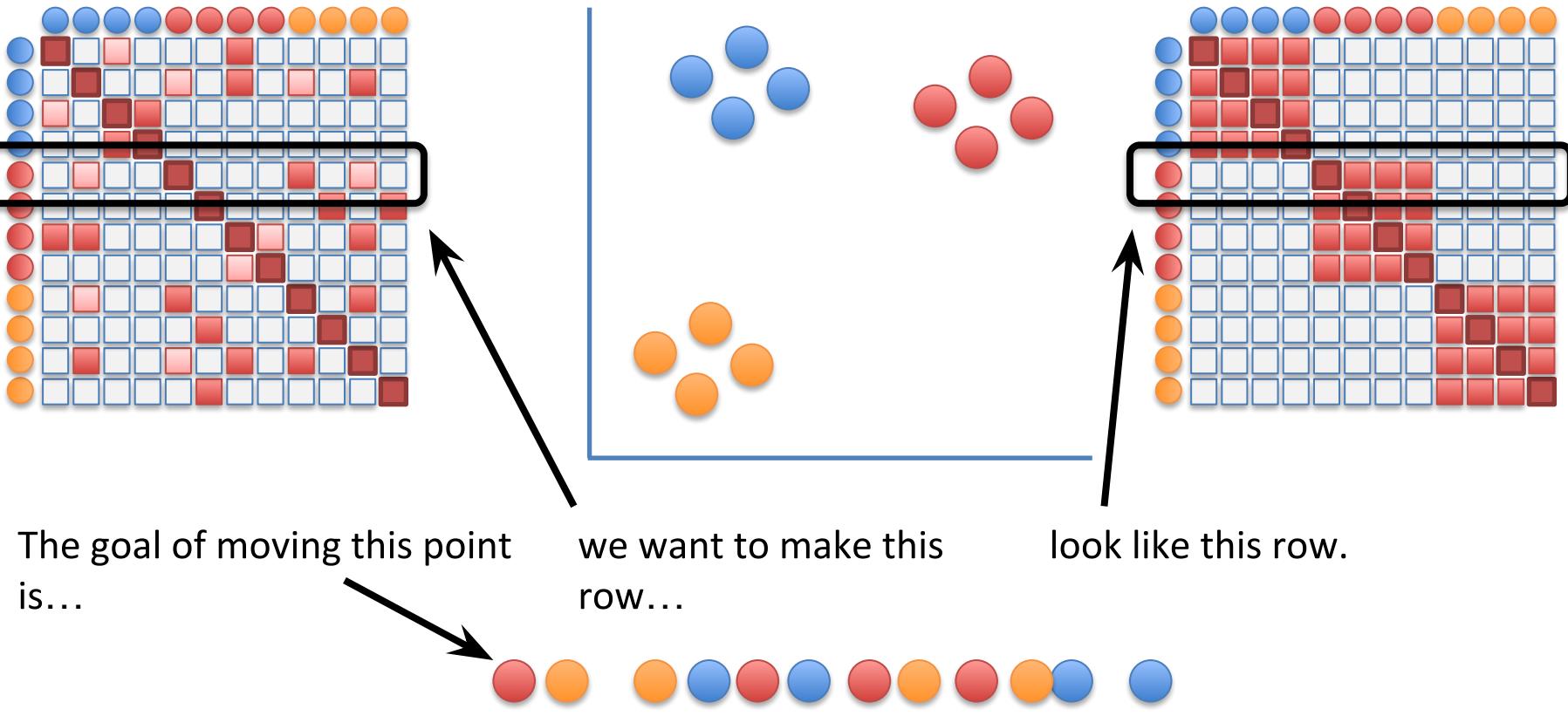


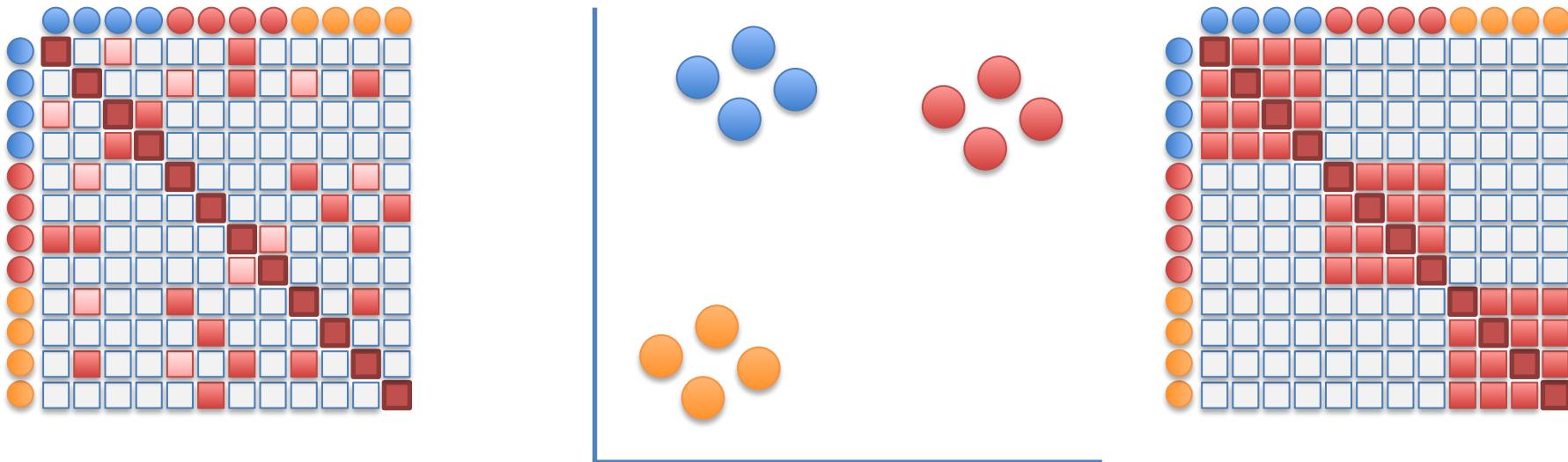


The goal of moving this point  
is...

we want to make this  
row...

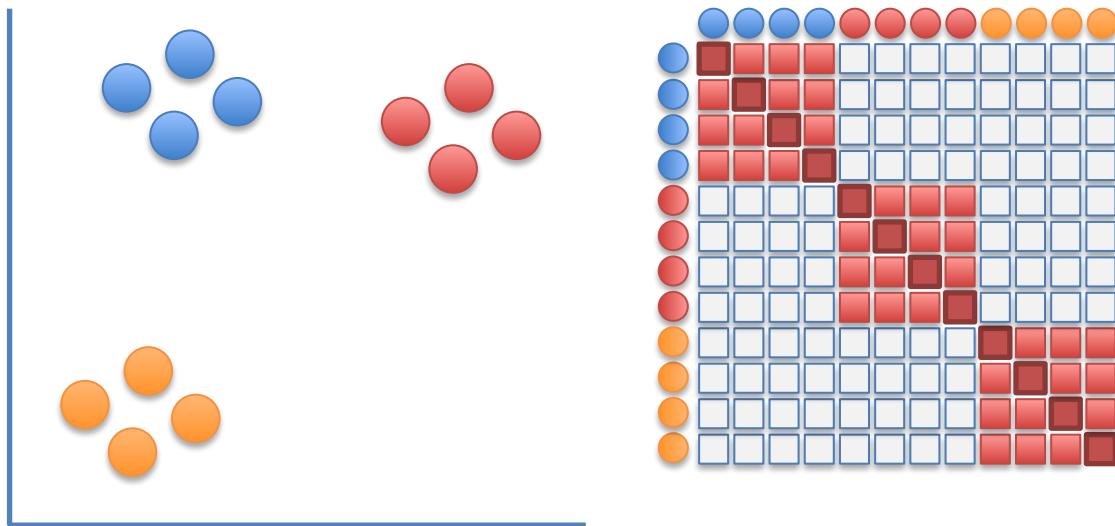
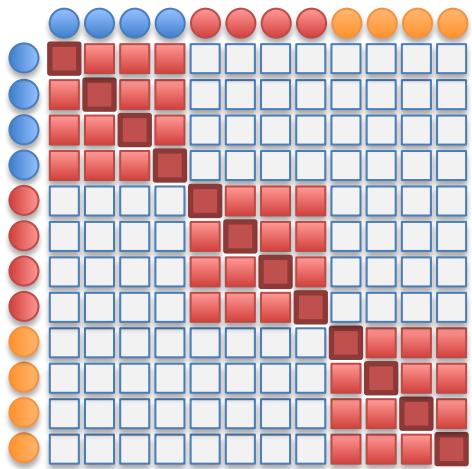






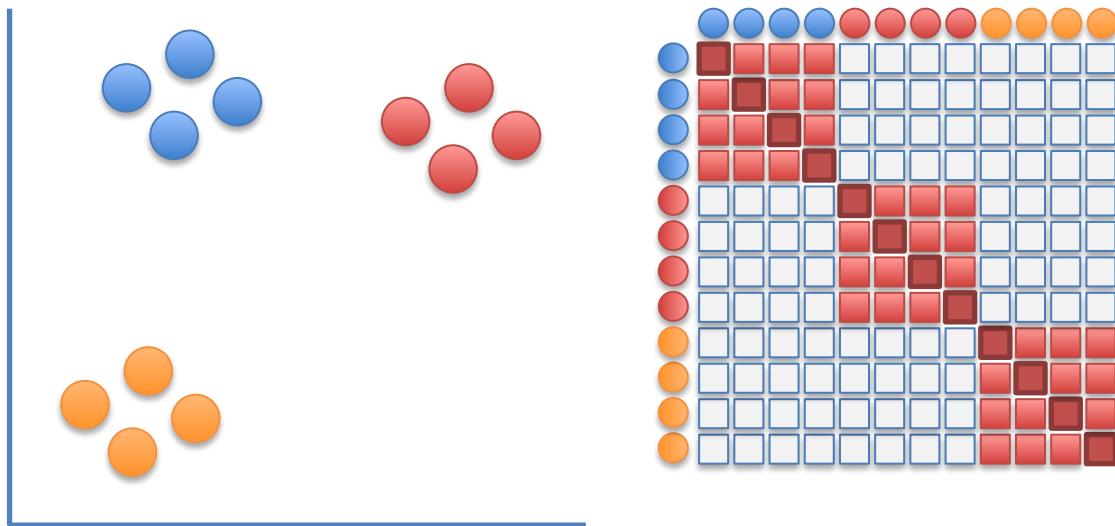
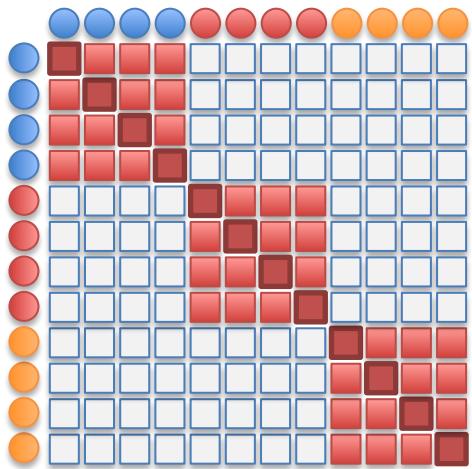
t-SNE moves the points a little bit at a time, and each step it chooses a direction that makes the matrix on the left more like the matrix on the right.





t-SNE moves the points a little bit at a time, and each step it chooses a direction that makes the matrix on the left more like the matrix on the right.

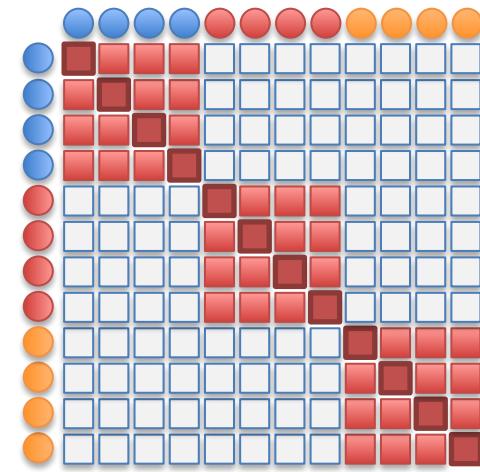
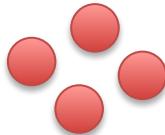
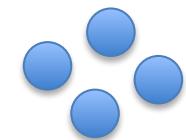
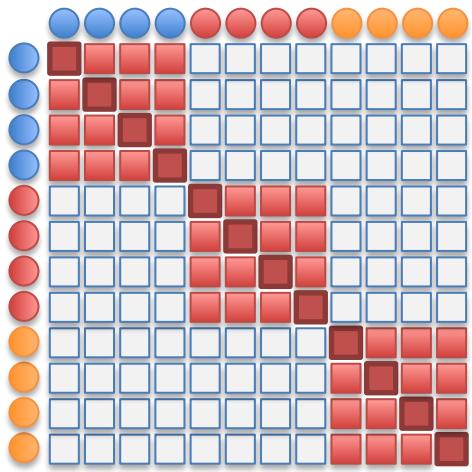




t-SNE moves the points a little bit at a time, and each step it chooses a direction that makes the matrix on the left more like the matrix on the right.

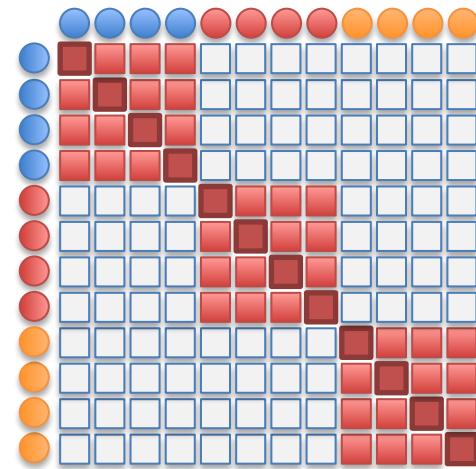
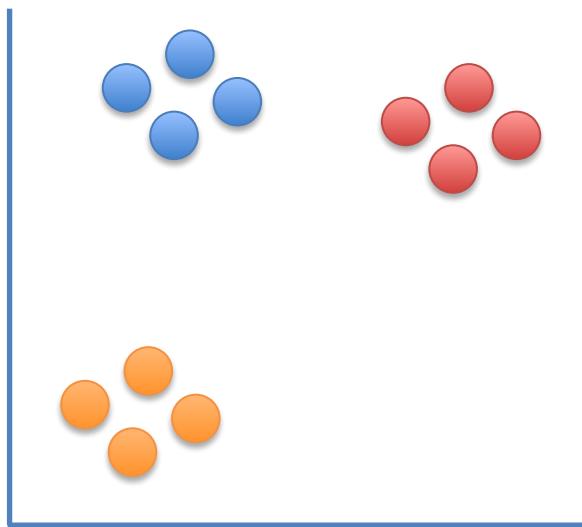
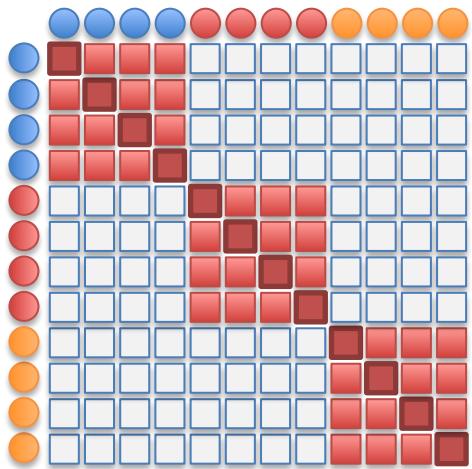


It uses small steps, because it's a little bit like a chess game and can't be solved all at once. Instead, it goes one move at a time.



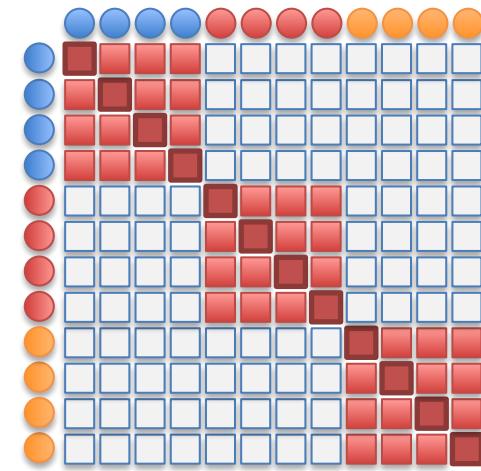
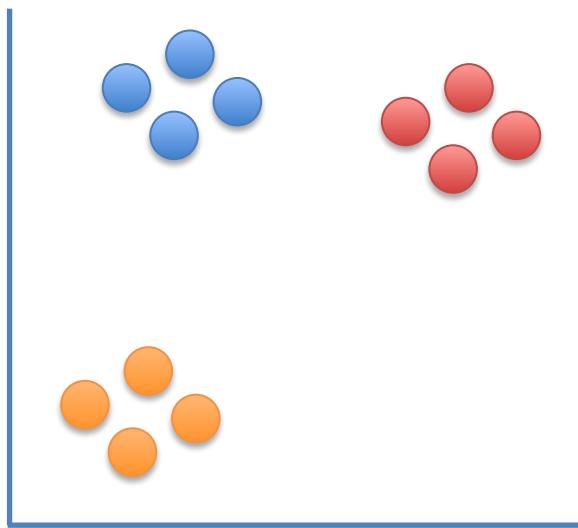
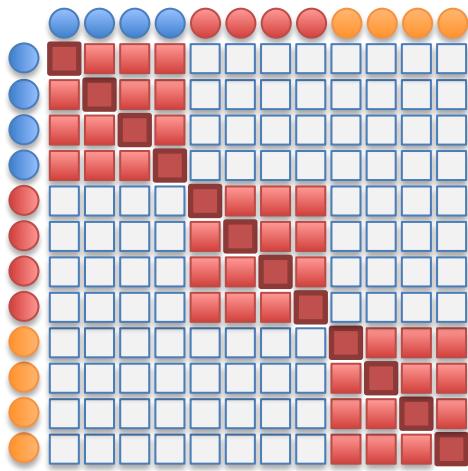
BAM!!!





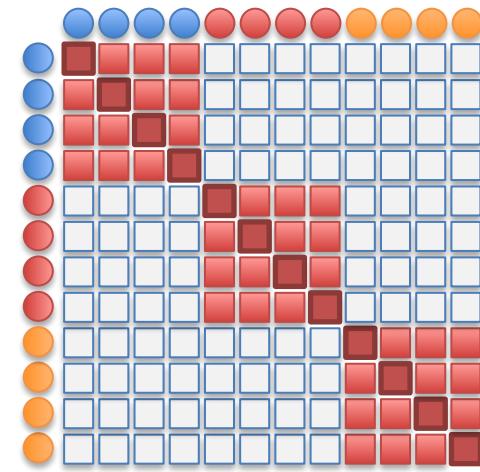
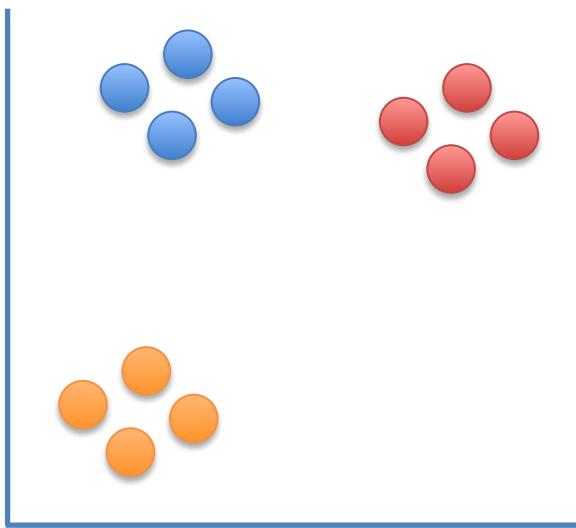
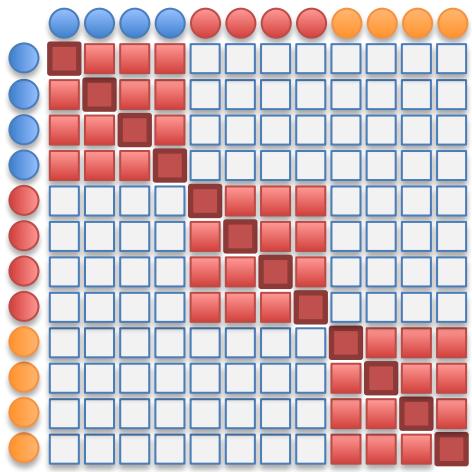
Now to finally tell you why the “t-distribution” is used...





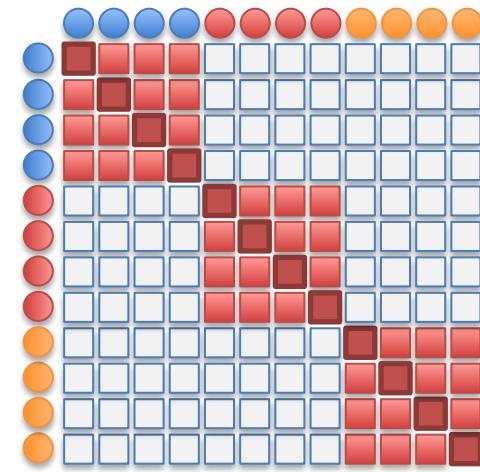
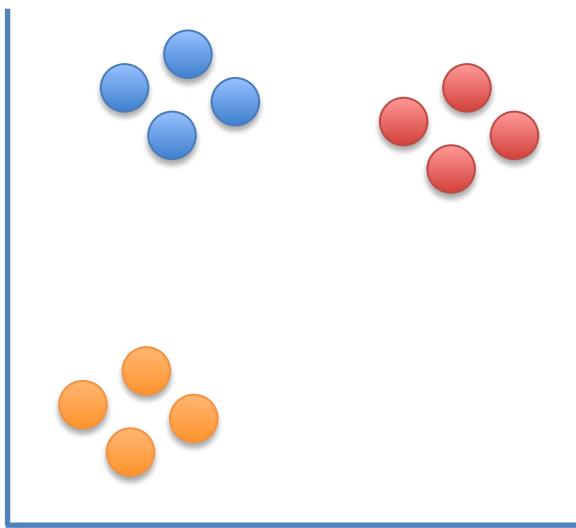
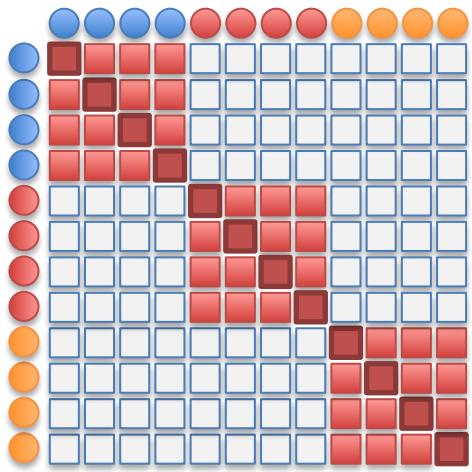
...originally, the “SNE” algorithm just used a normal distribution throughout and the clusters clumped up in the middle and were harder to see.





The t-distribution forces some space between the points.





# Triple Bam!!!



# The End !!!