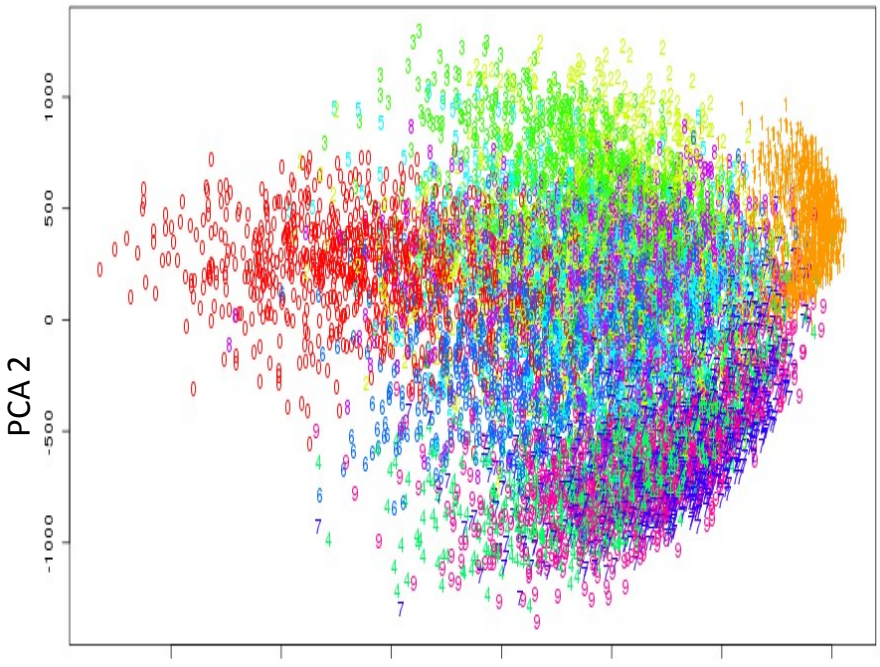
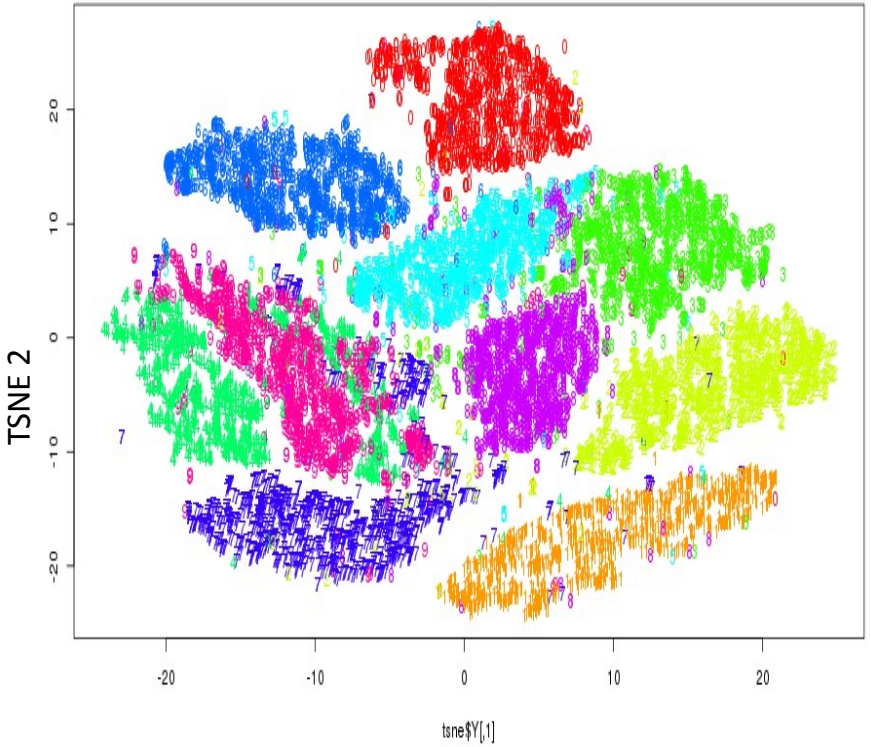
PCA



PCA 1

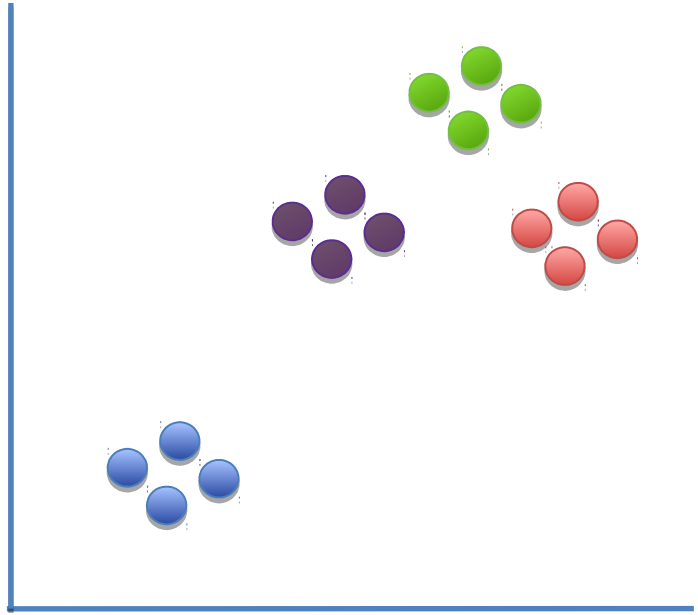
TSNE



TSNE 1

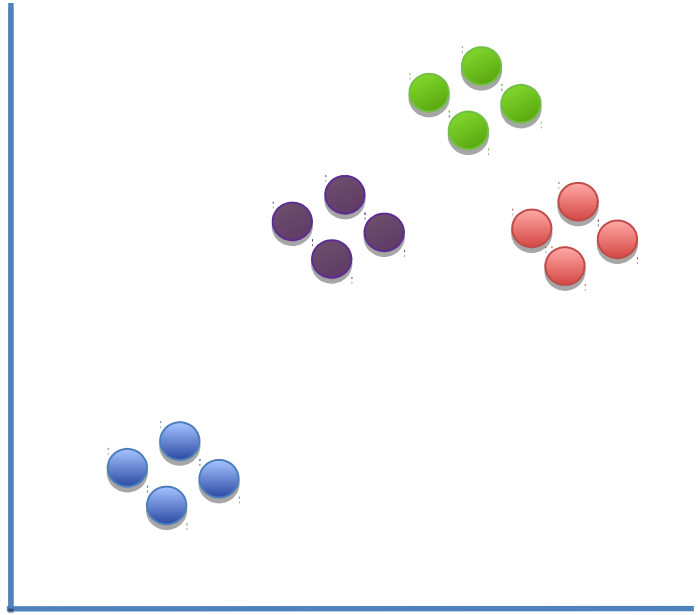
Image from https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/unsupervised_learning.html

Here's a basic 2-D scatter plot.



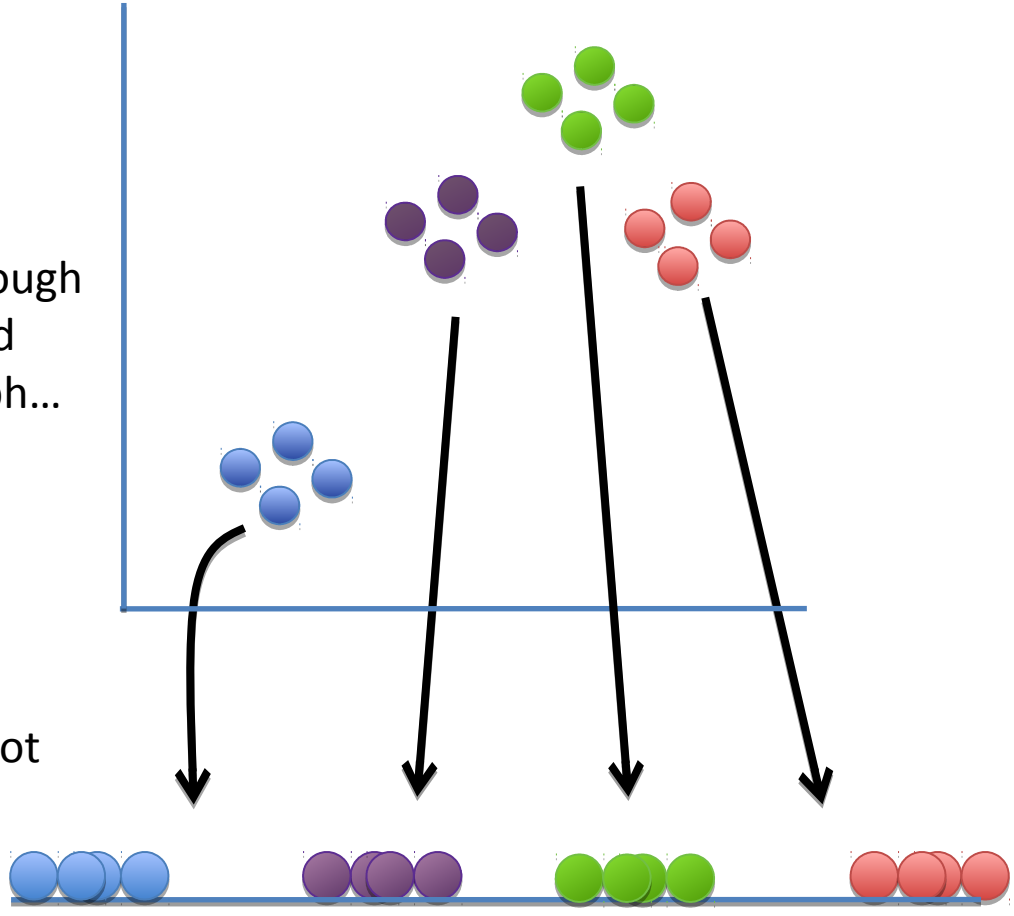
Here's a basic 2-D scatter plot.

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

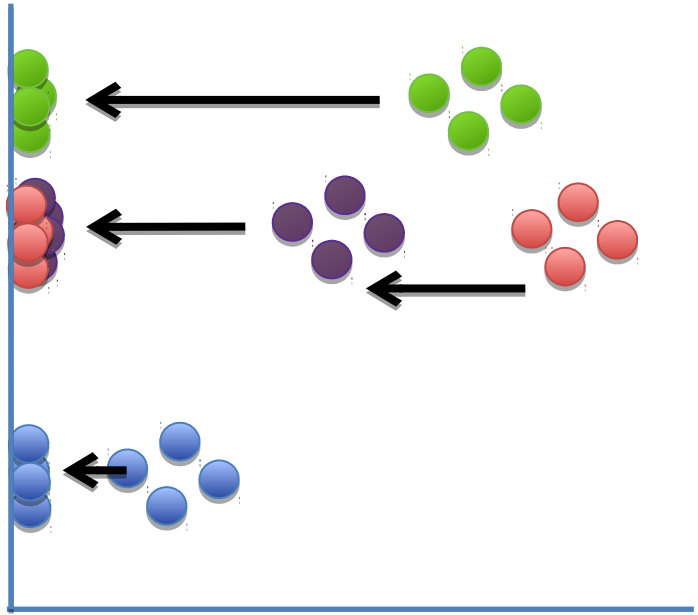


Here's a basic 2-D scatter plot.

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

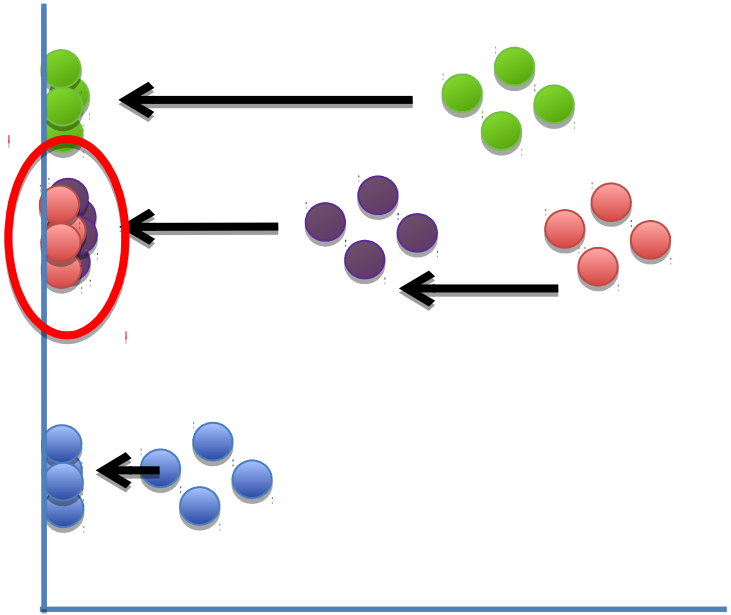


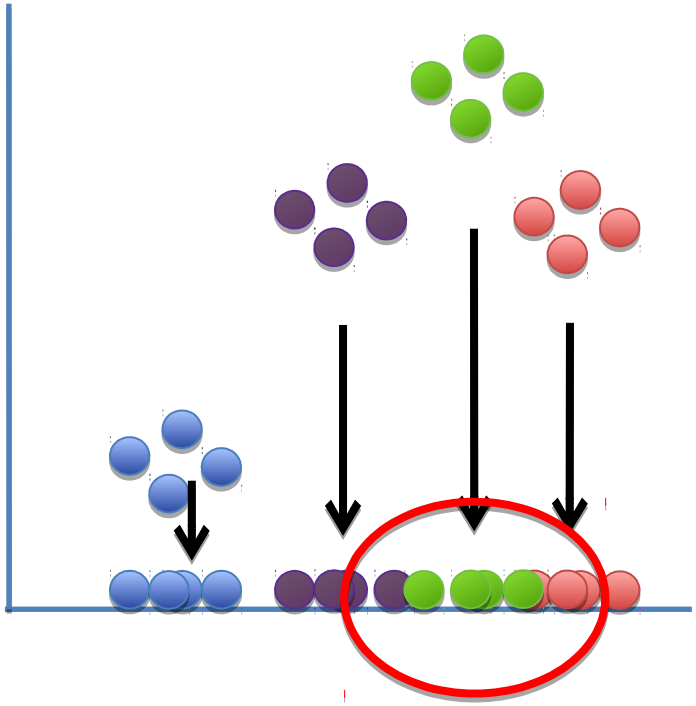
…into a flat, 1-D plot on a number line.



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.

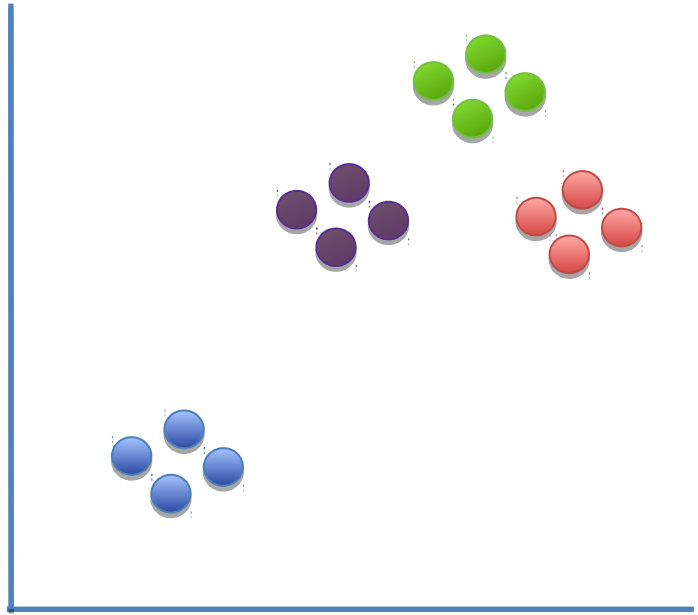
Instead of two distinct clusters, we just see a mess.

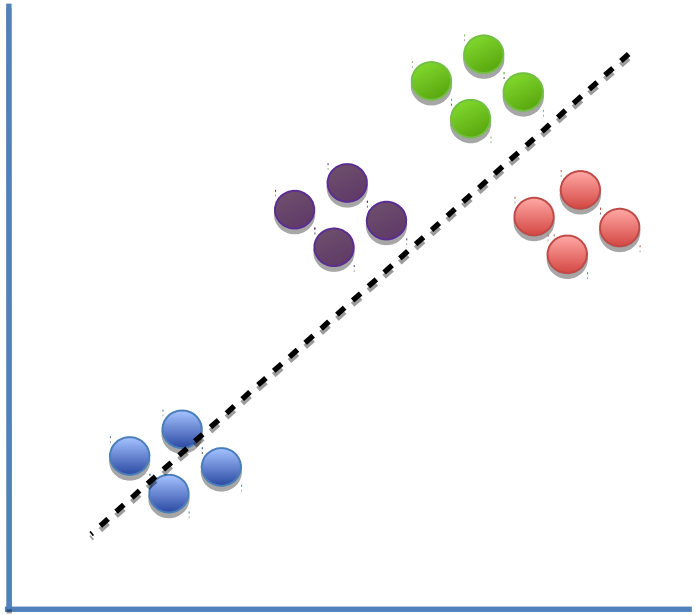


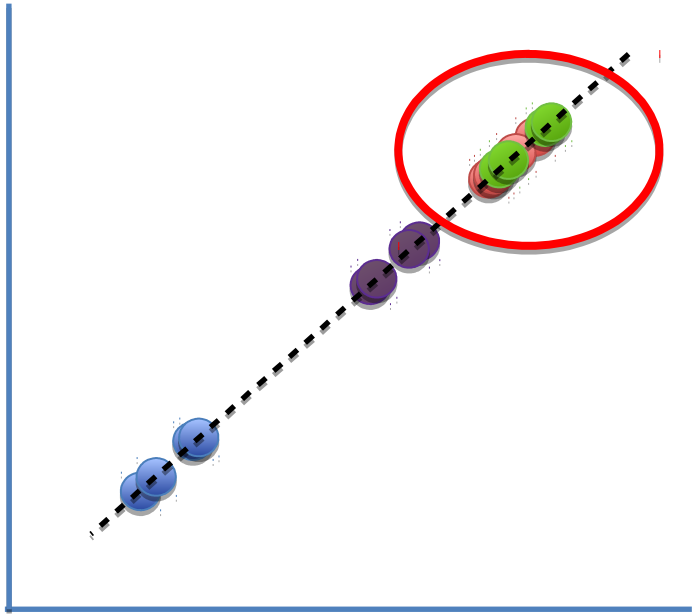


Same here...

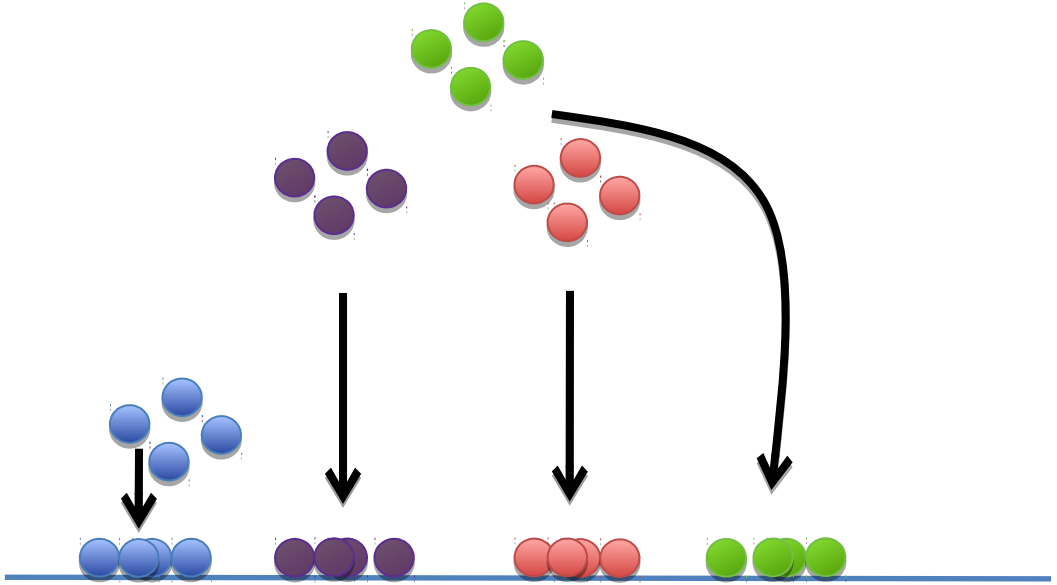
What about PCA?



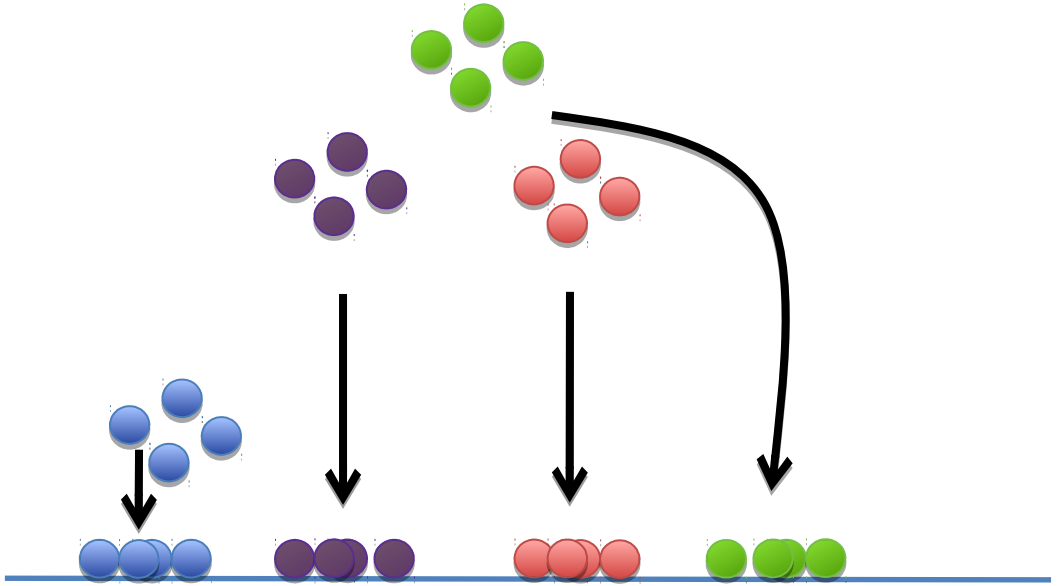




Same
again..!

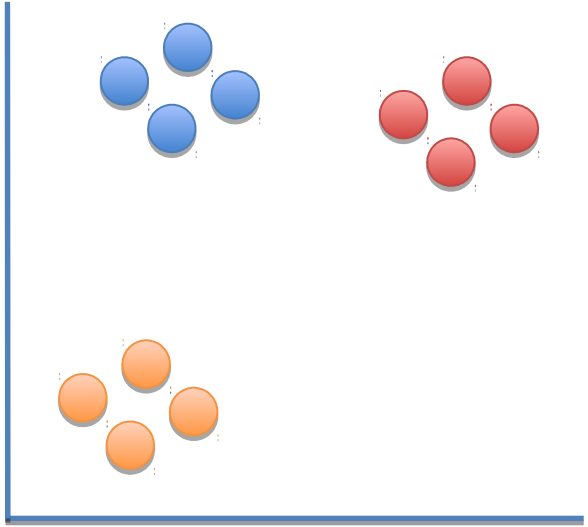


What t-SNE does is find a way to project data into a low dimensional space (in this case, the 1-D number line) so that the clustering structure in the high dimensional space (in this case, the 2-D scatter plot) is preserved.



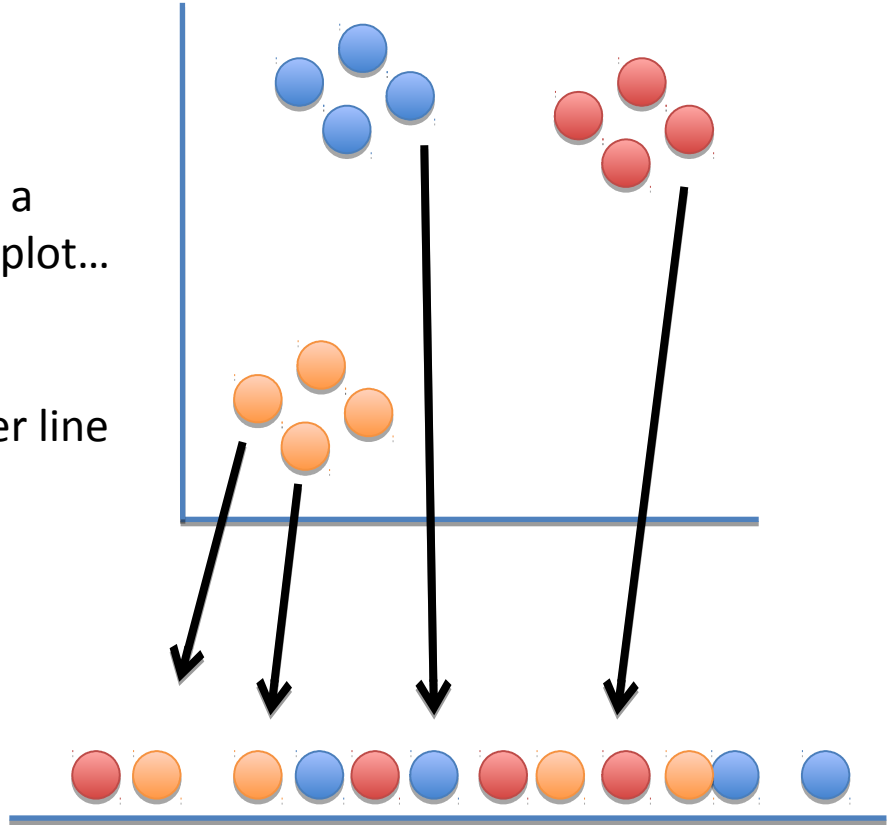
So let’s step through the basic ideas of how t-SNE does this.

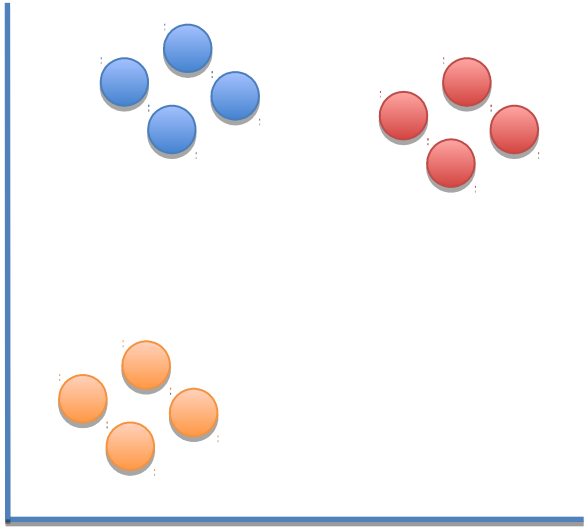
We'll start start with a more simple scatter plot…



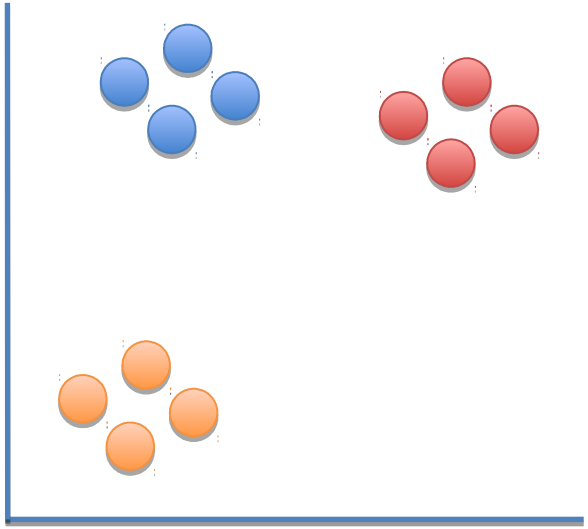
We'll start start with a more simple scatter plot…

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

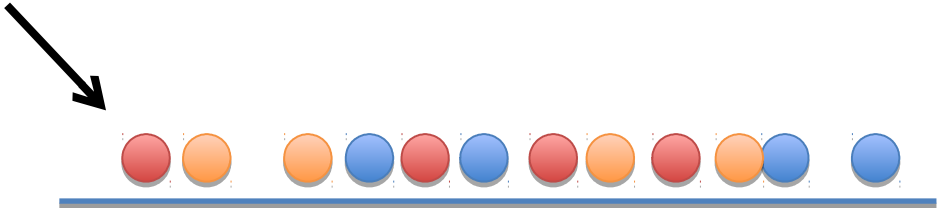


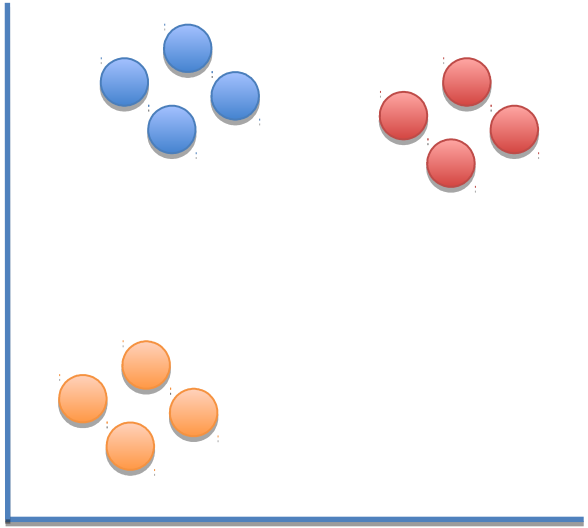


From here on out, t-SNE moves these points, a little bit at a time, until it has clustered them.

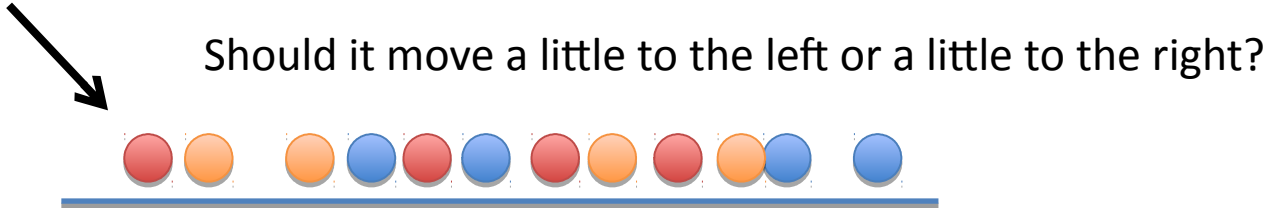


Let’s figure out where to move this first point…

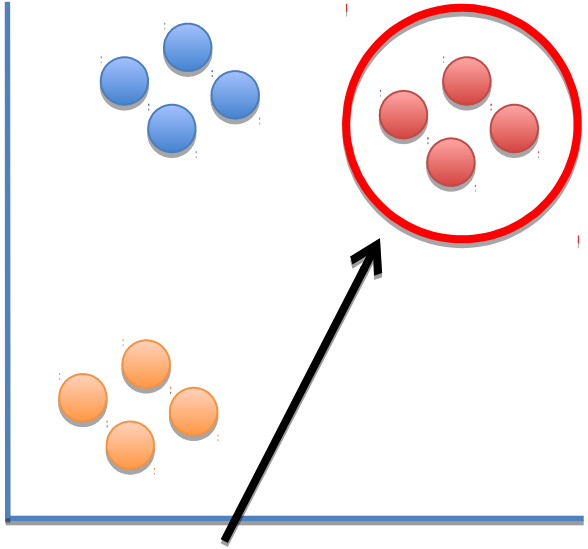




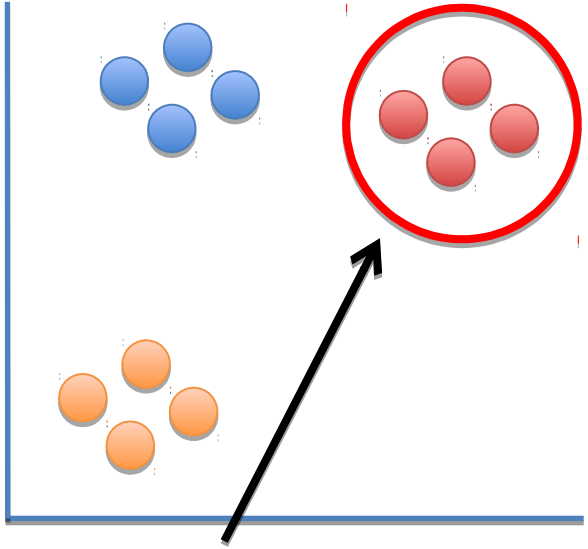
Let’s figure out where to move this first point…



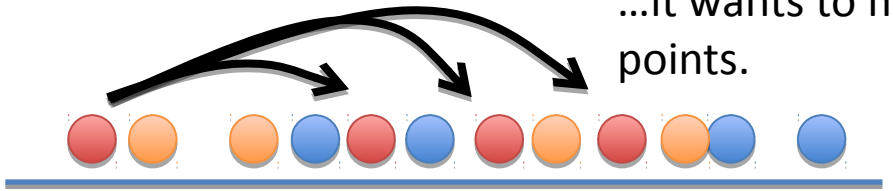
Should it move a little to the left or a little to the right?



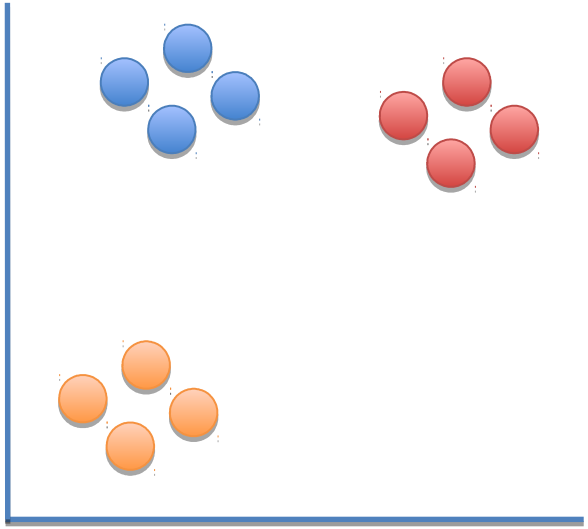
Because it is part of this cluster…



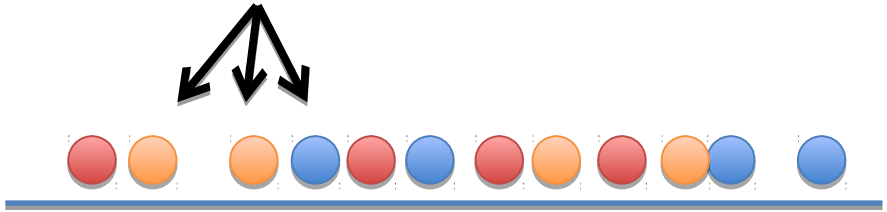
Because it is part of this cluster…

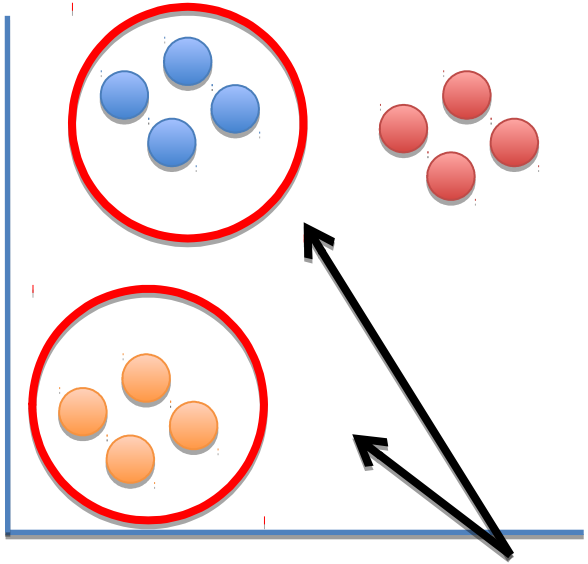


…it wants to move closer to these points.

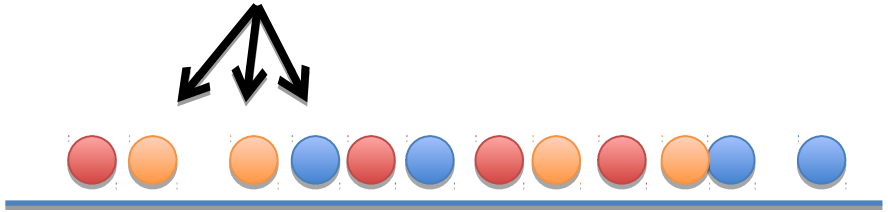


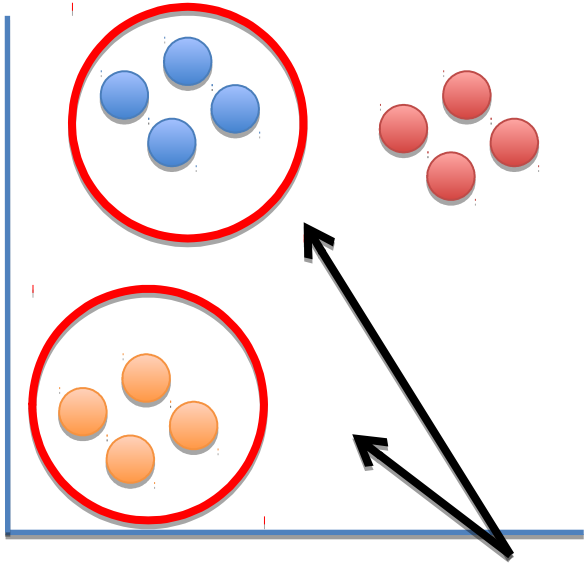
But at the same time, these points…



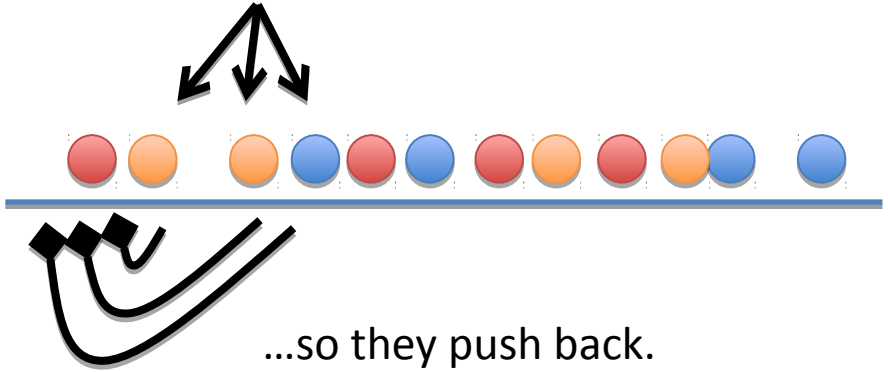


But at the same time, these points… …are far away in the scatter plot…

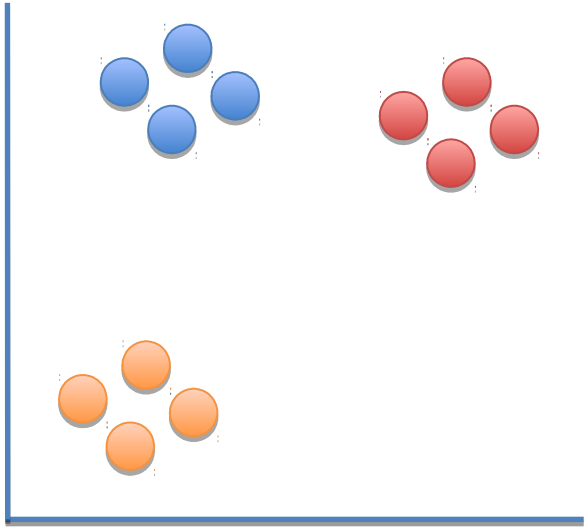


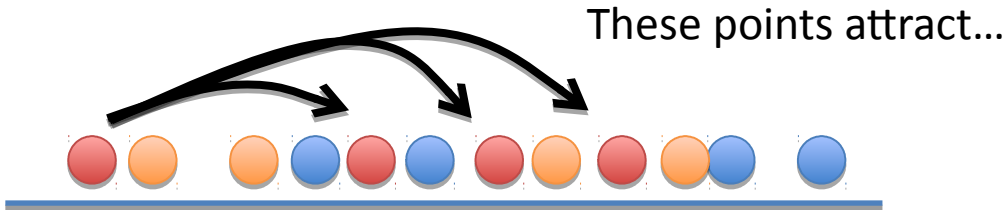


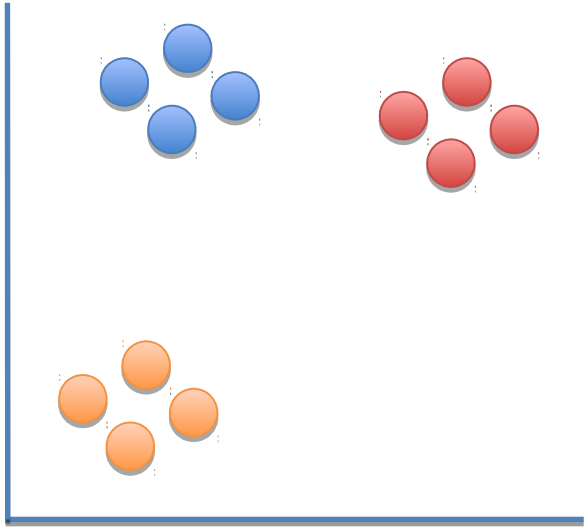
But at the same time, these points… …are far away in the scatter plot…

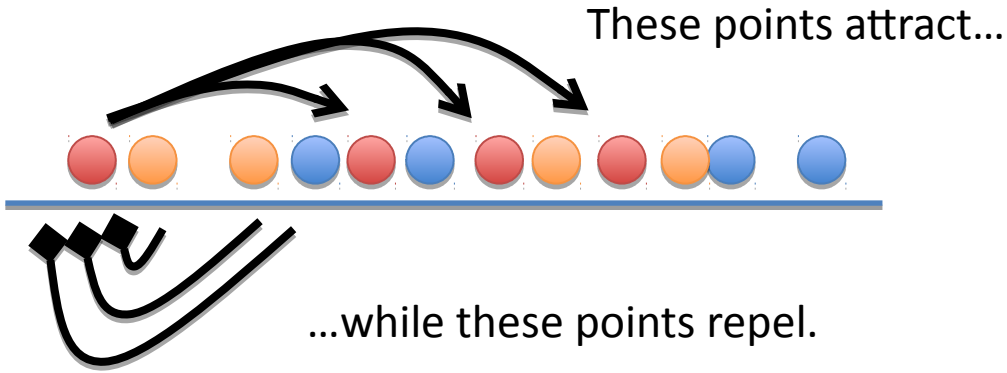


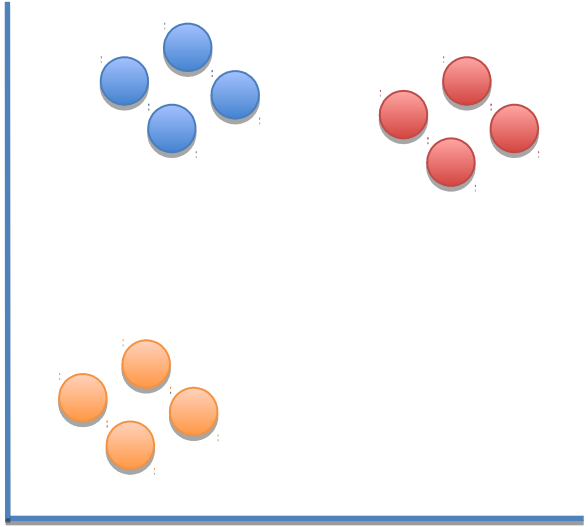
…so they push back.



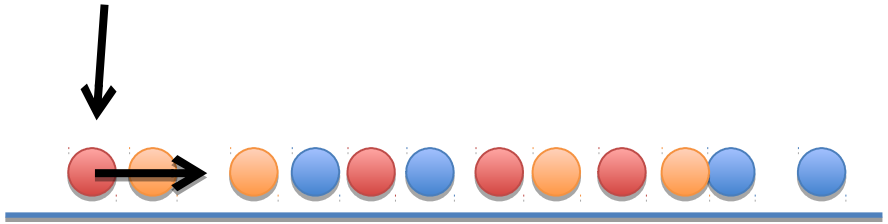

These points attract…

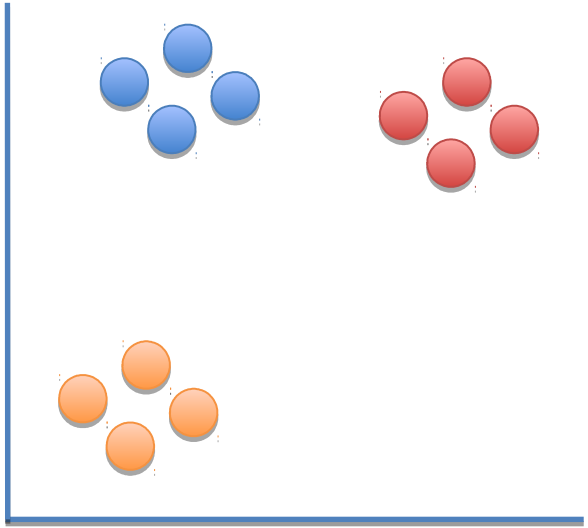



These points attract…
…while these points repel.

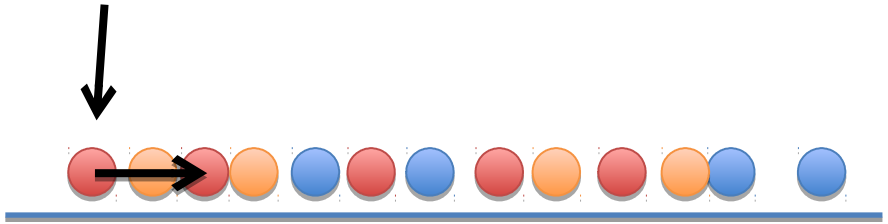


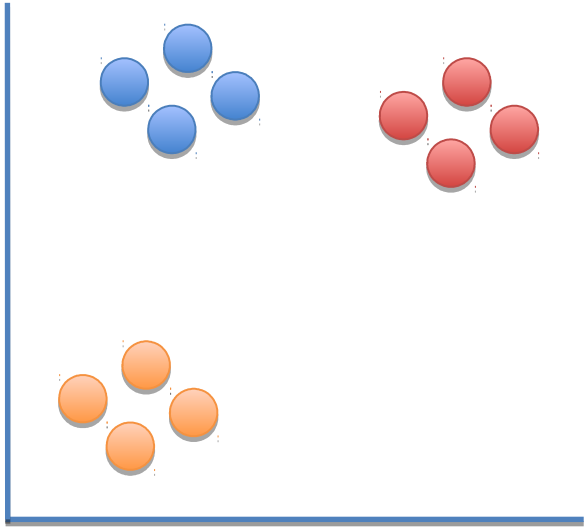
In this case, the attraction is strongest, so the point moves a little to the right.

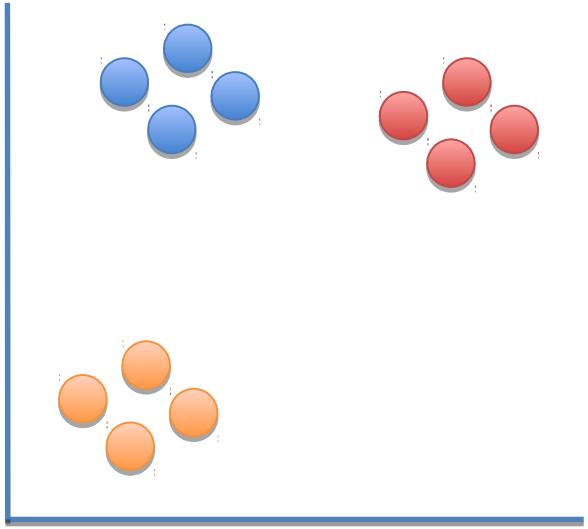




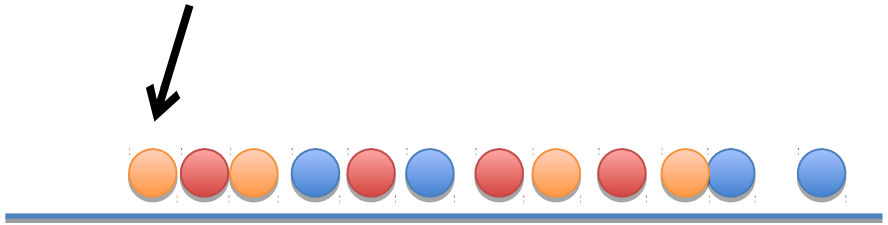
In this case, the attraction is strongest, so the point moves a little to the right.

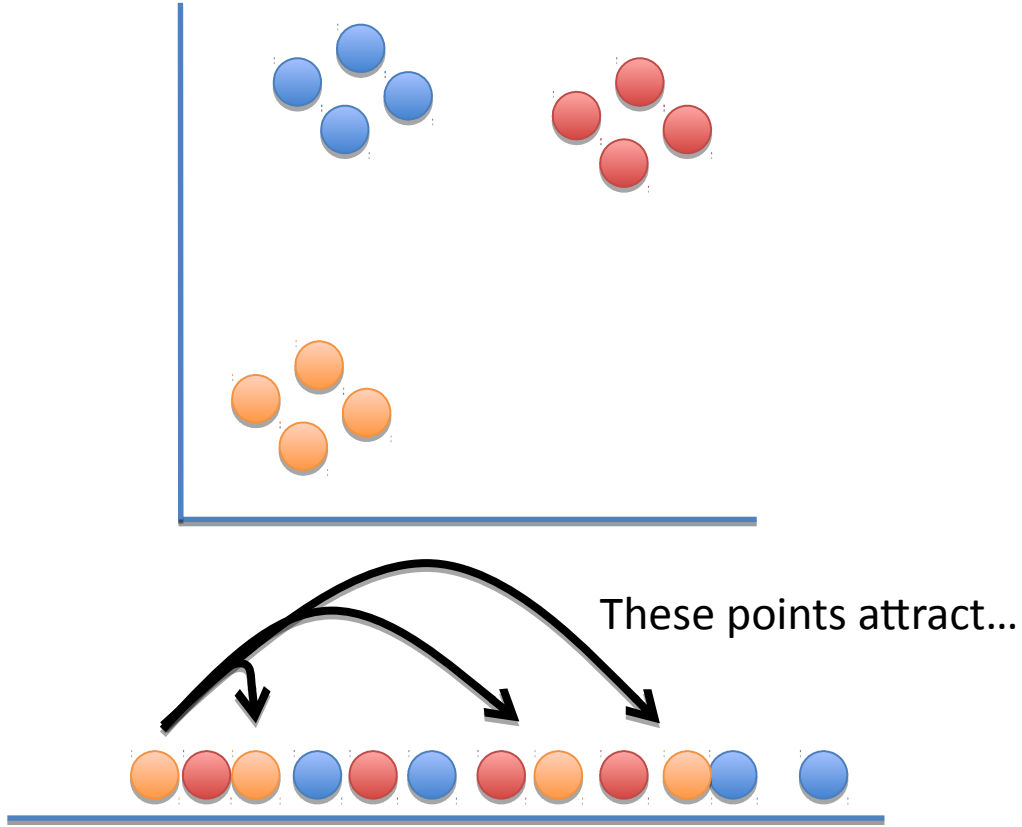


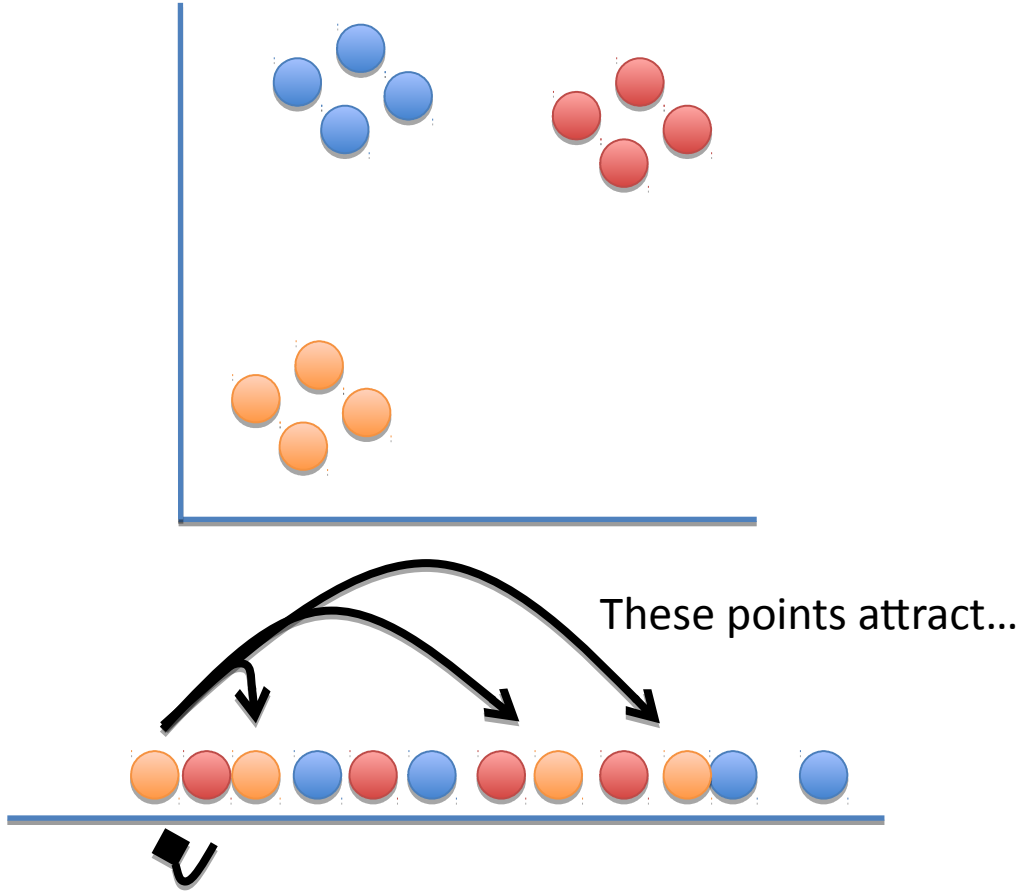


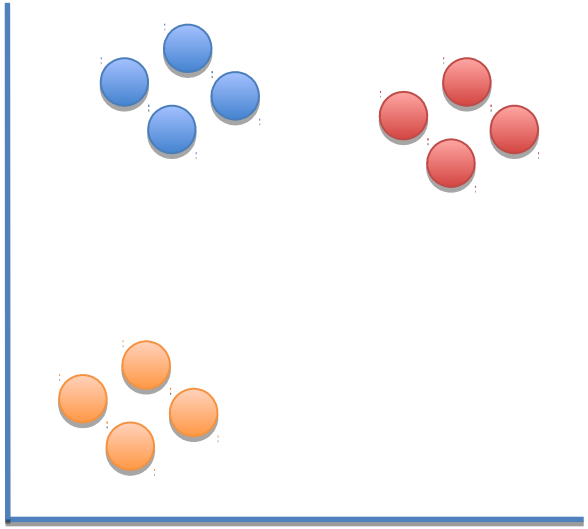


Now let’s move this point a little bit…

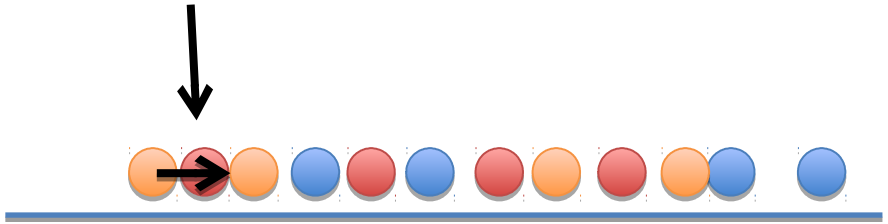


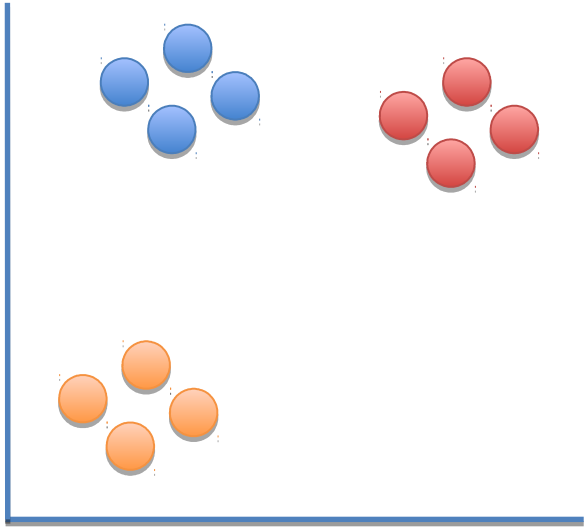

These points attract…

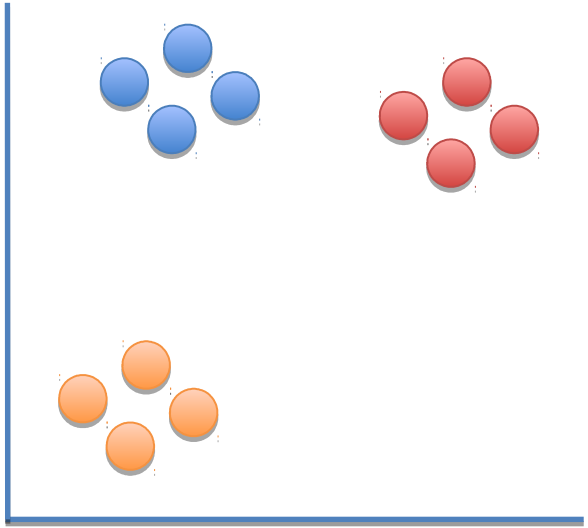

These points attract…
…and this point repels a little bit.



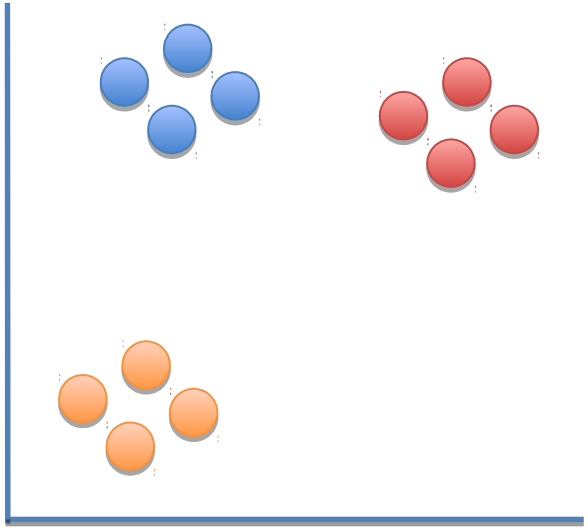
So it moves a little to closer to the other orange points.



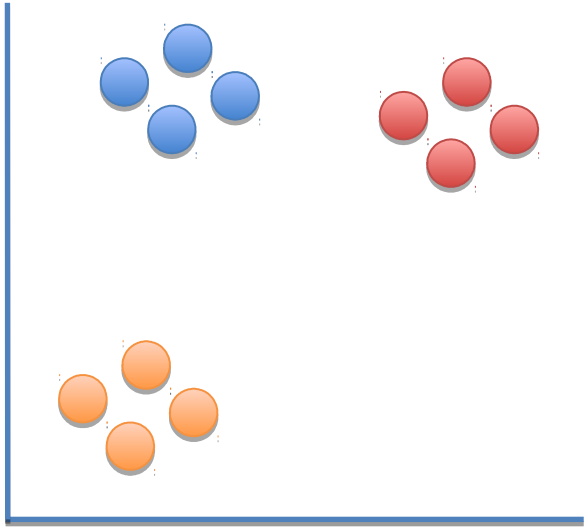




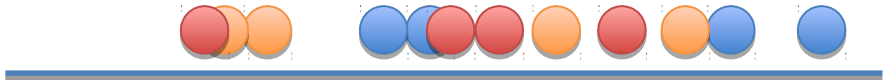
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…

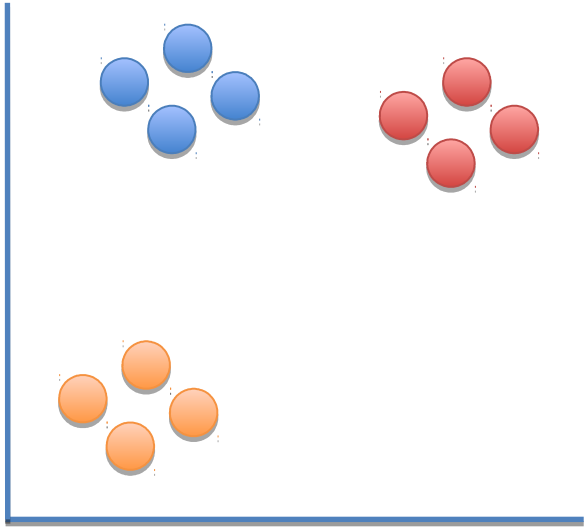


At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…

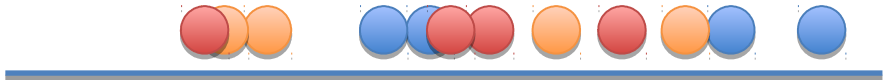


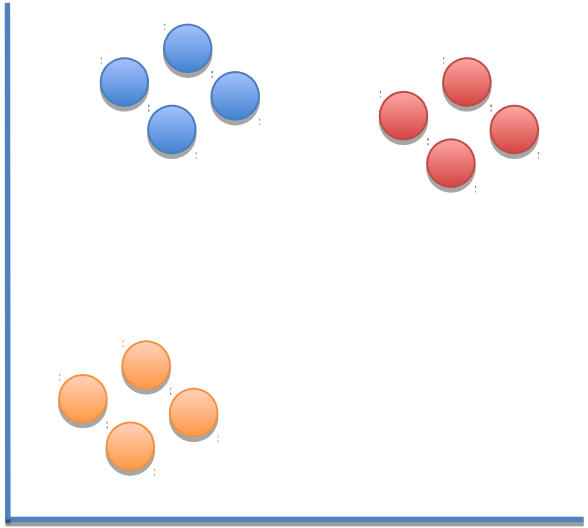
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



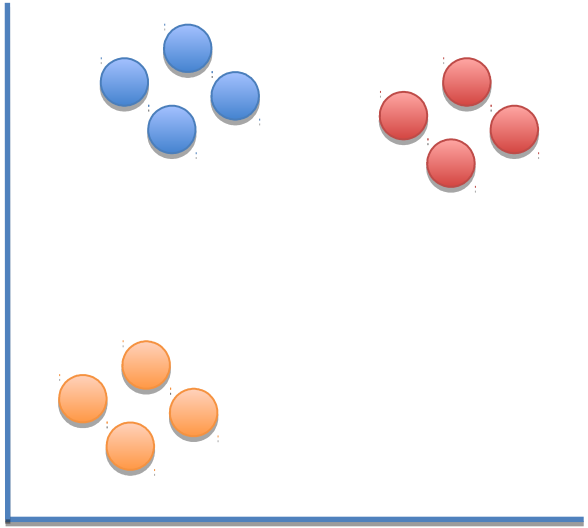


At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…

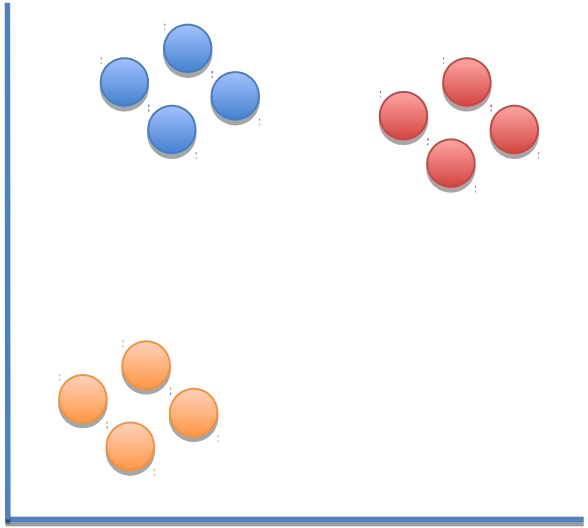




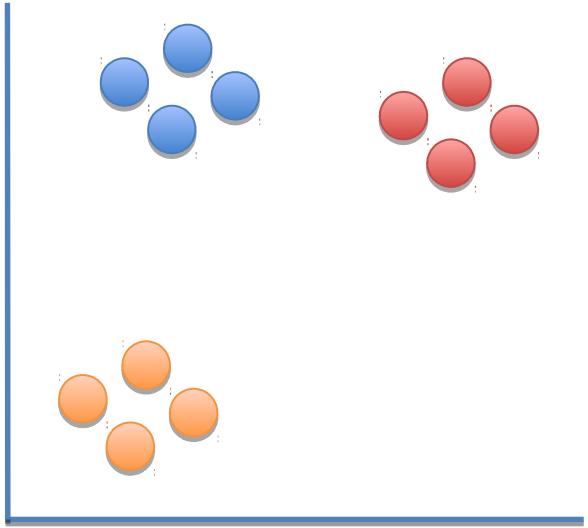
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



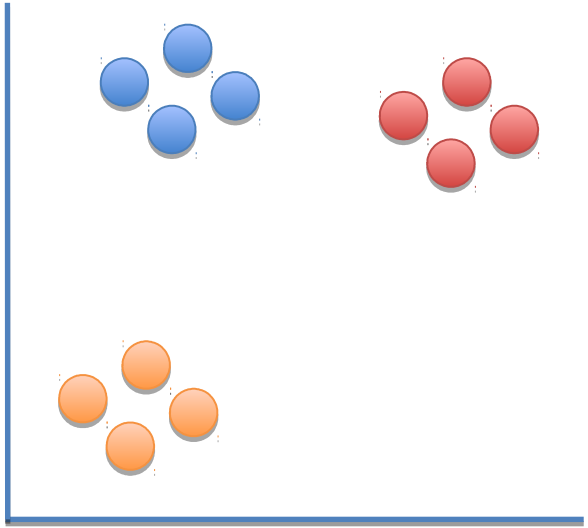
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



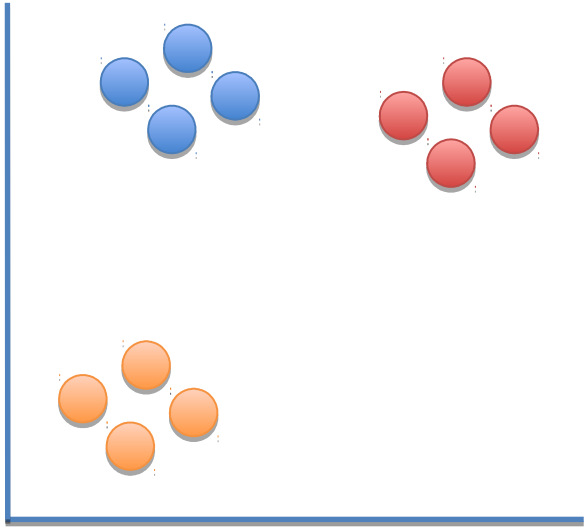
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



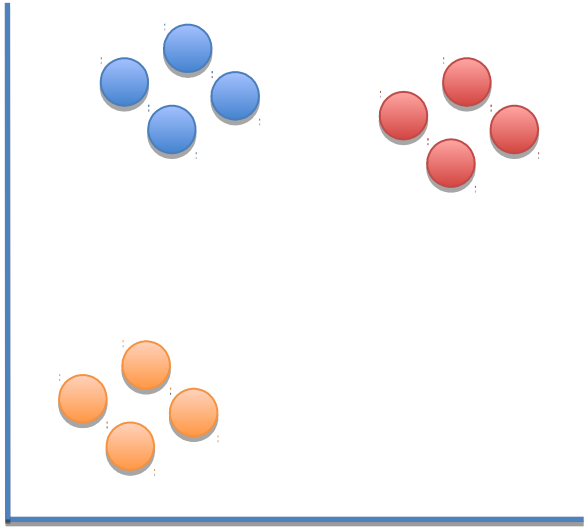
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



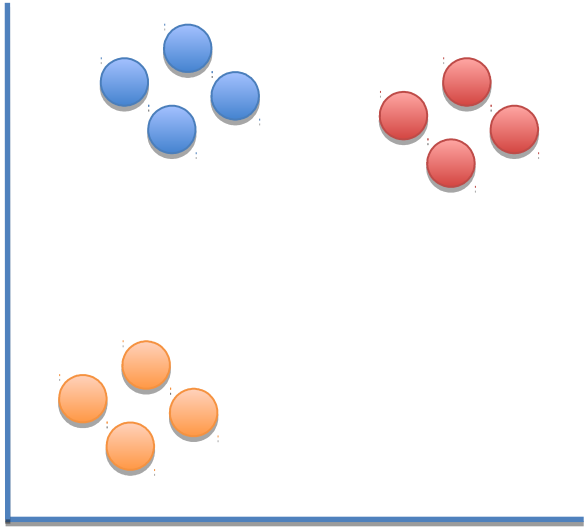
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



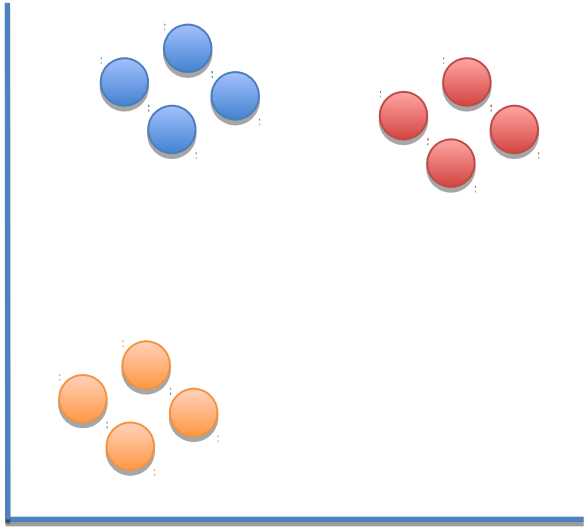
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



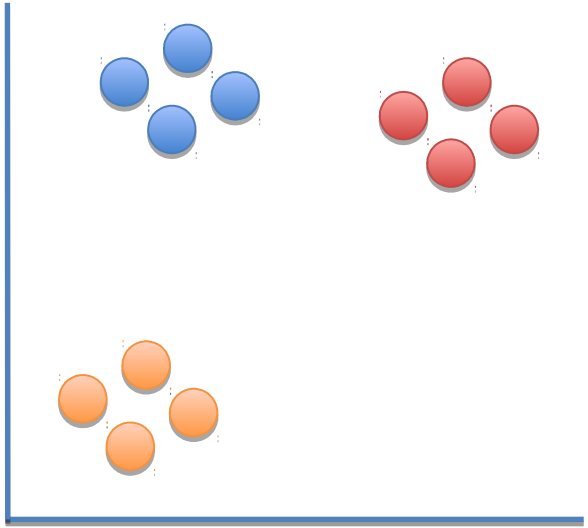
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



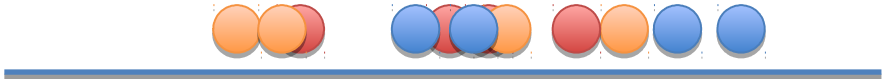
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…

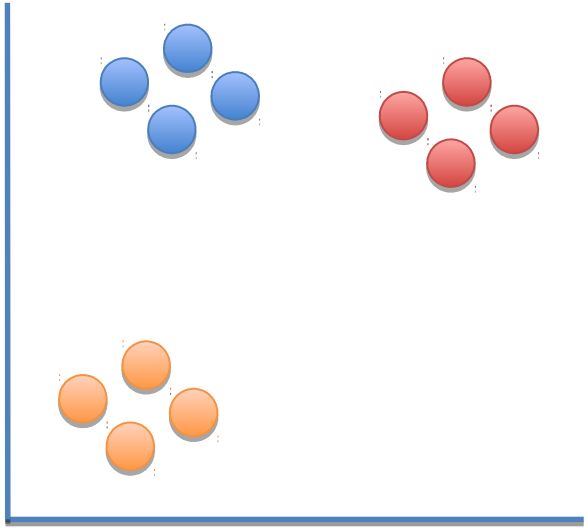


At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…

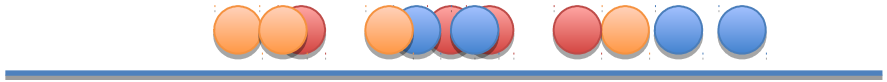


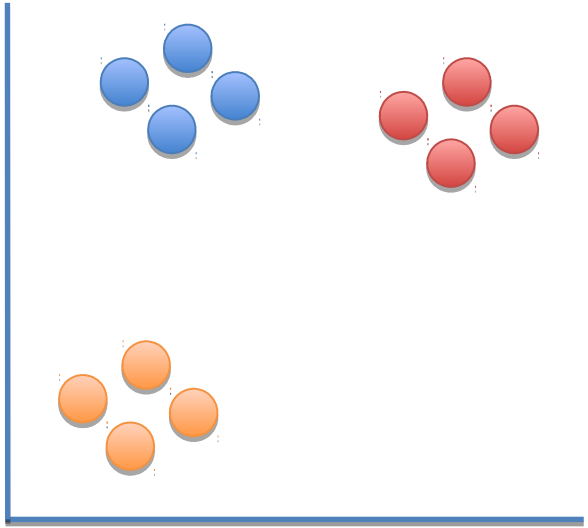
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



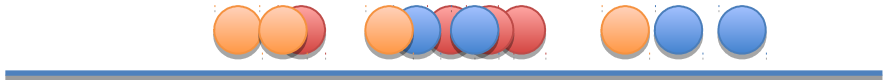


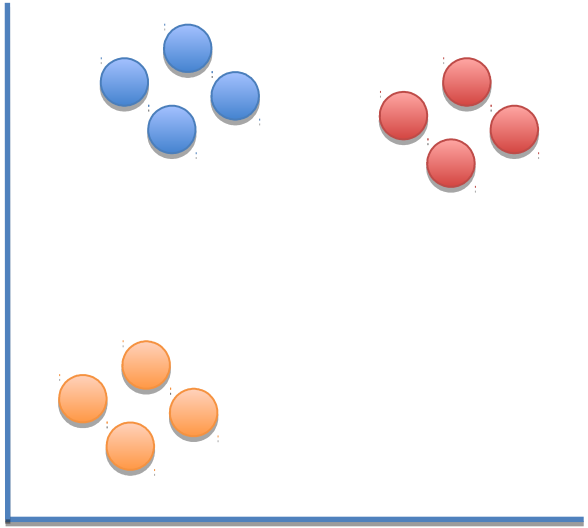
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



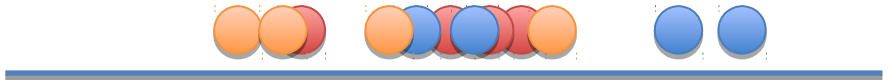


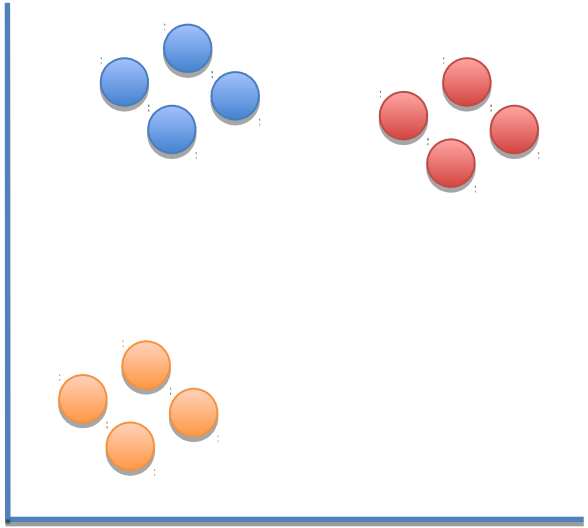
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



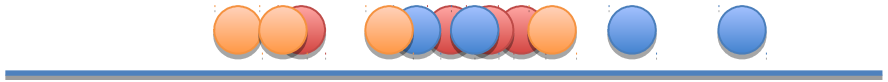


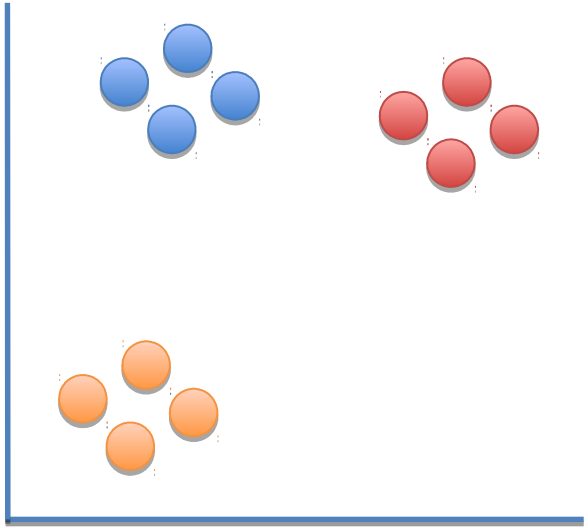
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



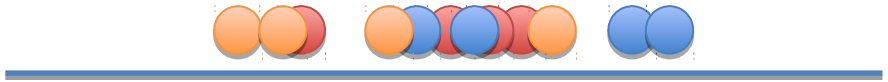


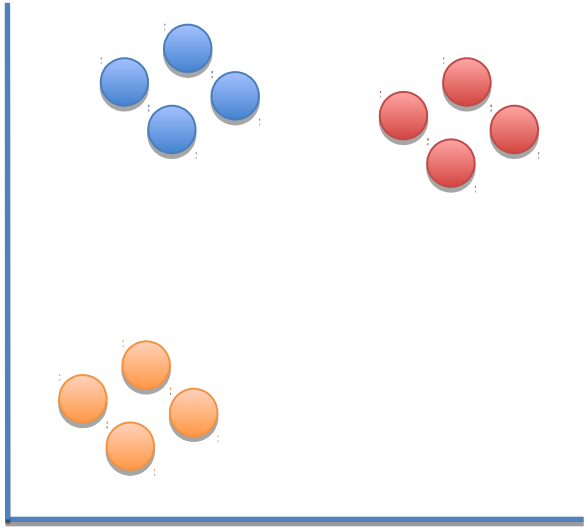
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



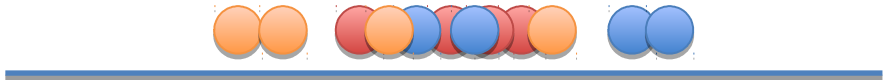


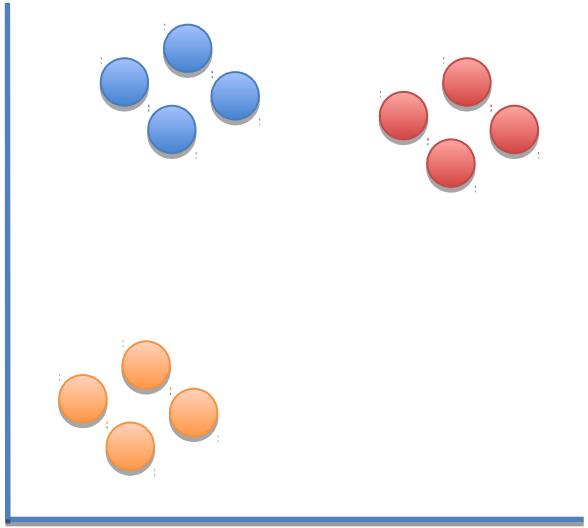
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



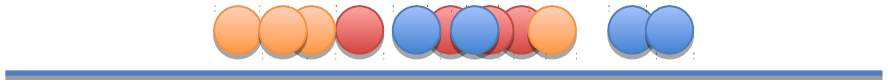


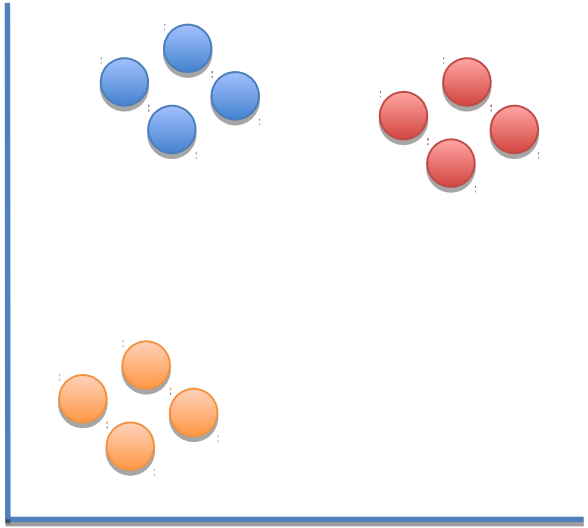
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



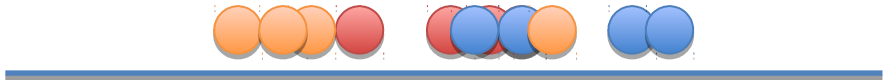


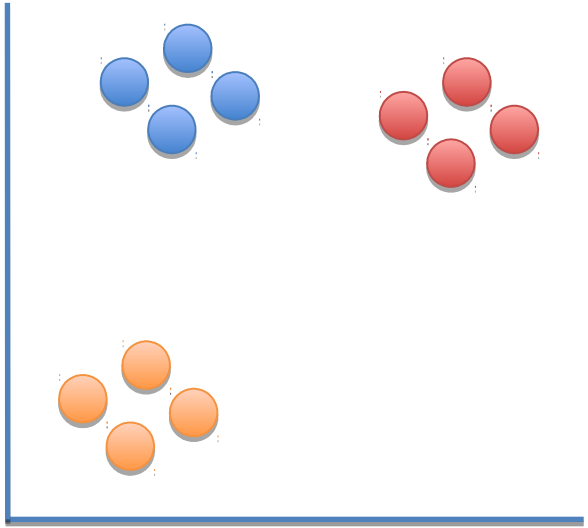
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



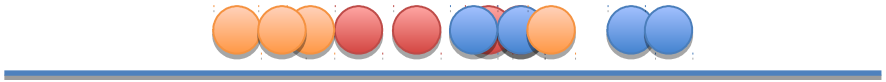


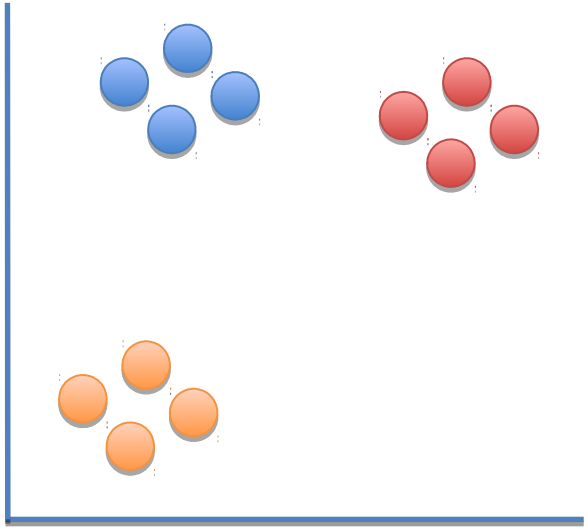
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



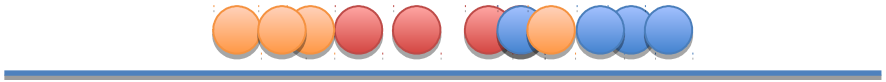


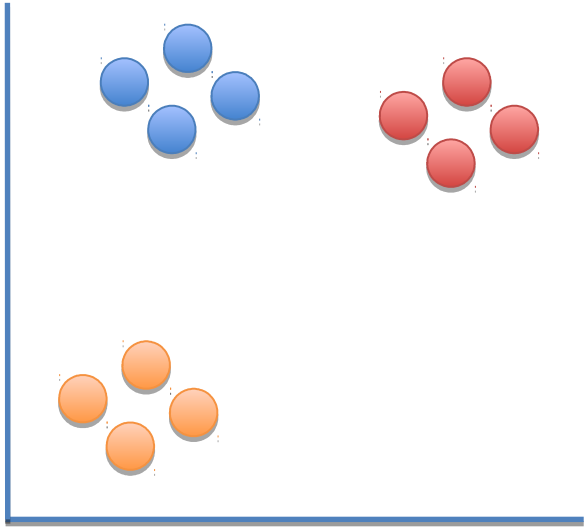
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



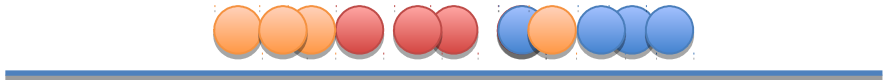


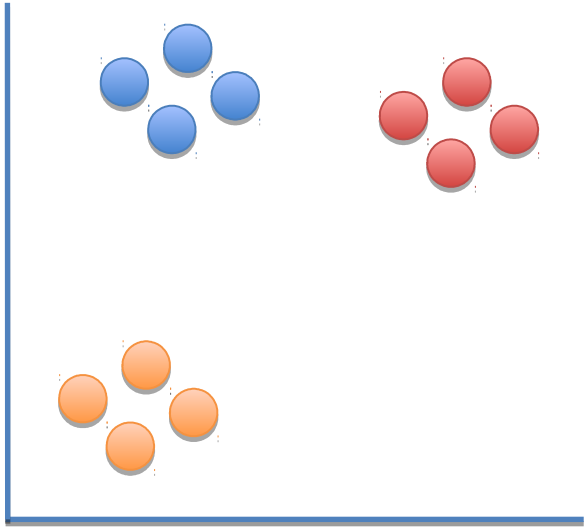
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



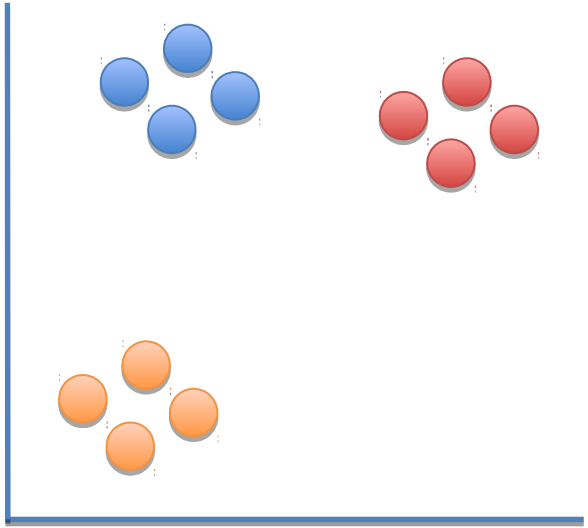


At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…

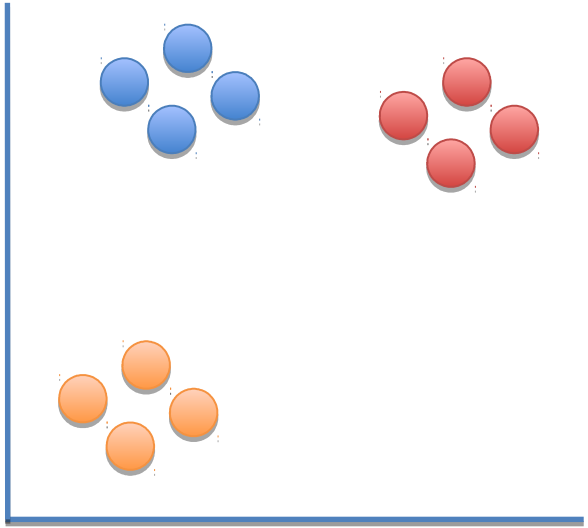




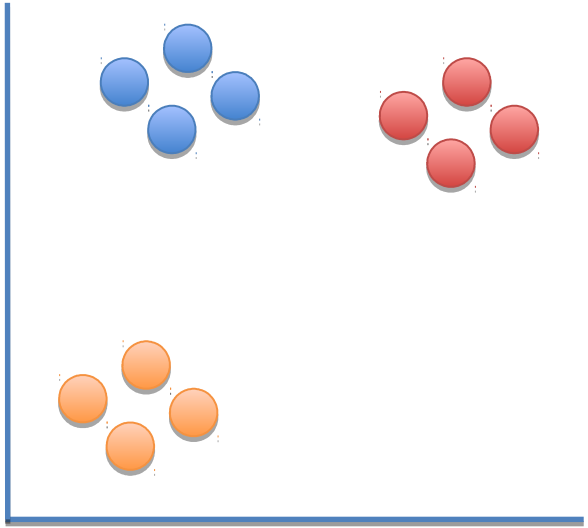
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



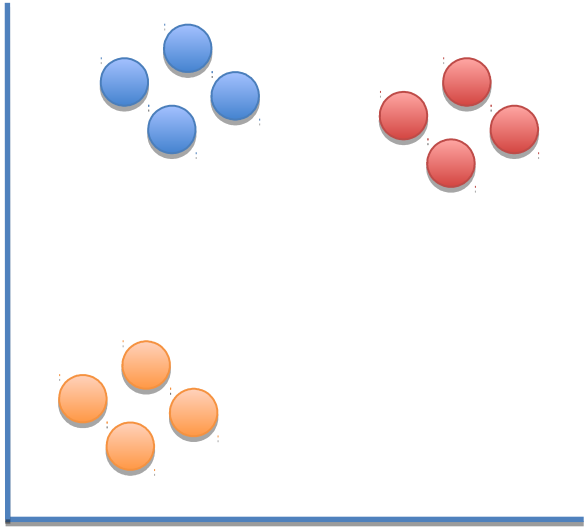
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



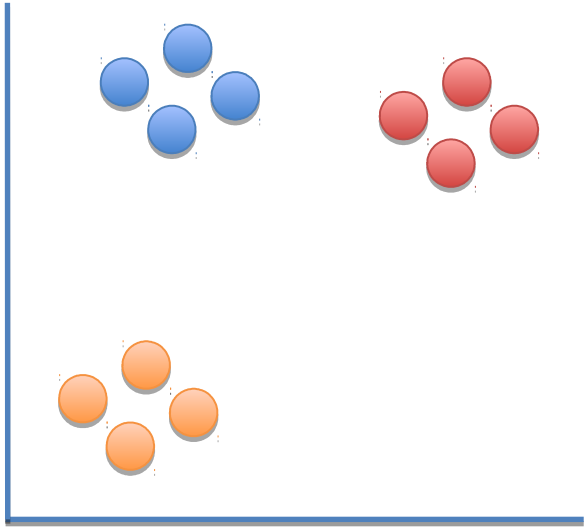
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…



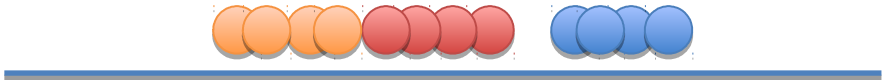
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…

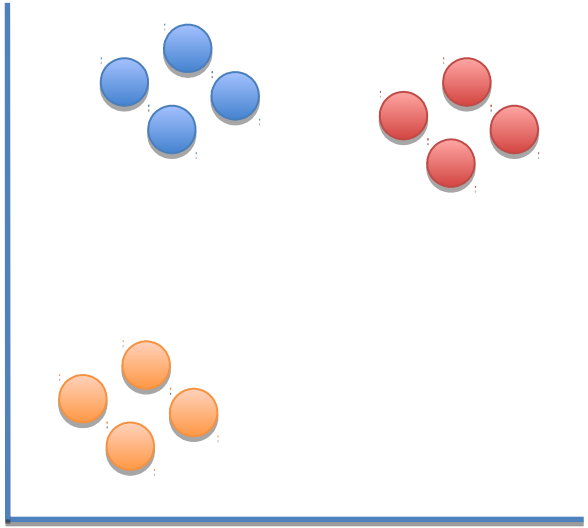


At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…

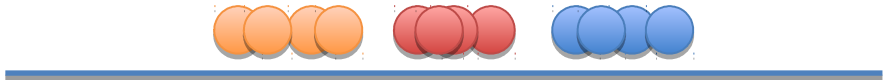


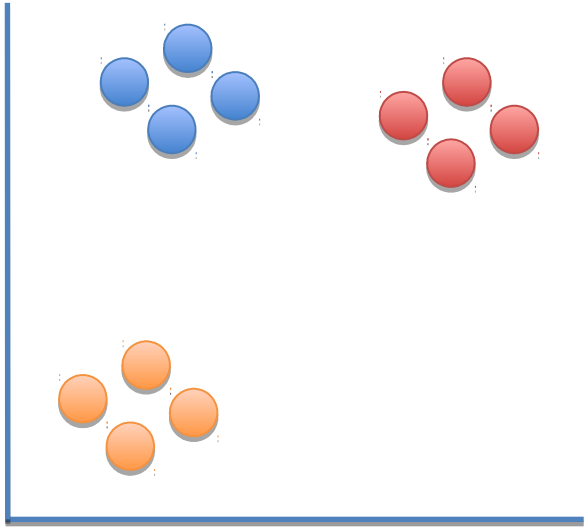
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…

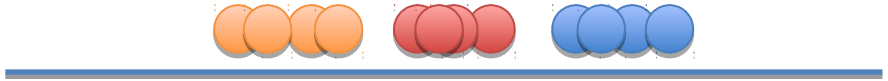




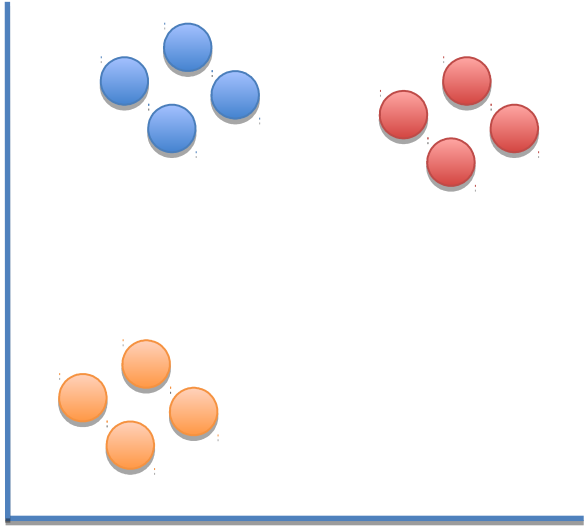
At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…

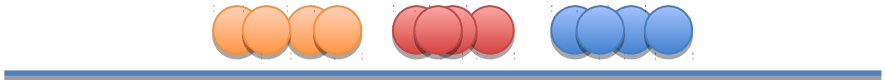




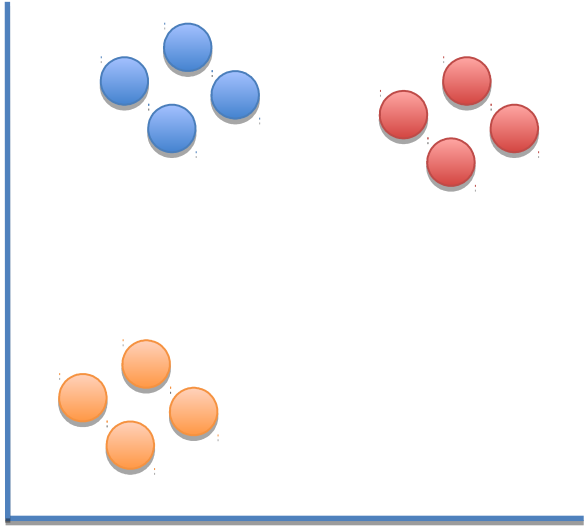


Now that we've seen the what t-SNE tries to do, let's dive into the nitty-gritty details of how it does what it does.



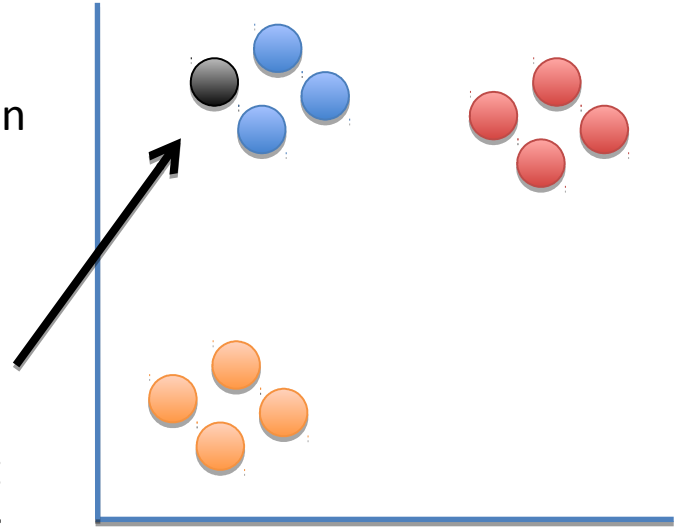


Step 1: Determine the “similarity” of all the points in the scatter plot.

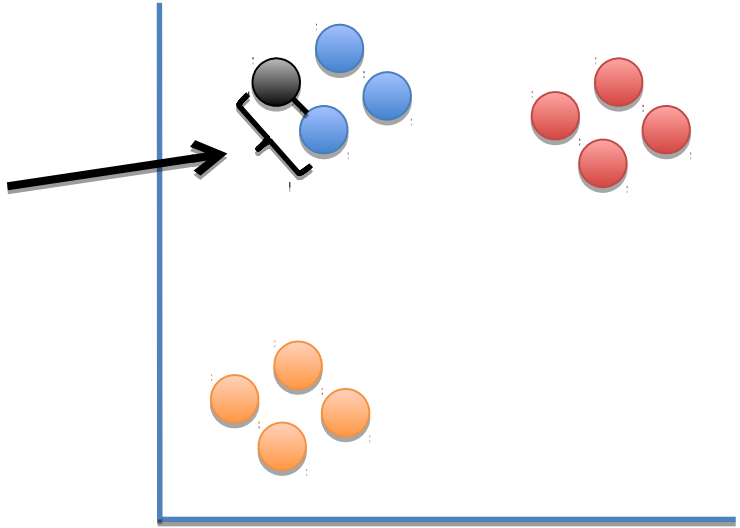


Step 1: Determine the "similarity" of all the points in the scatter plot.

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

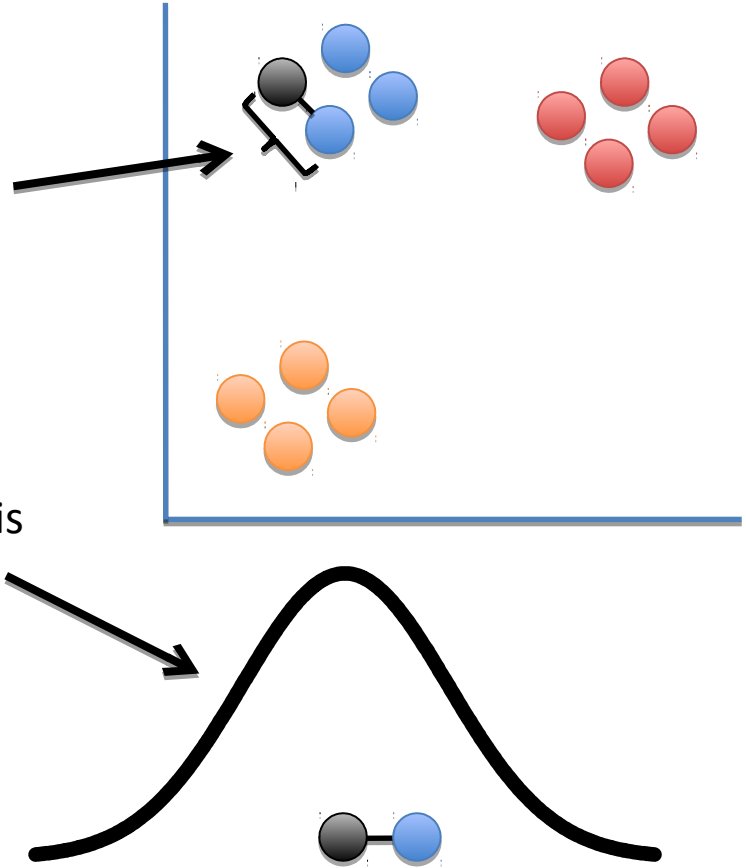


First, measure the distance between two points…

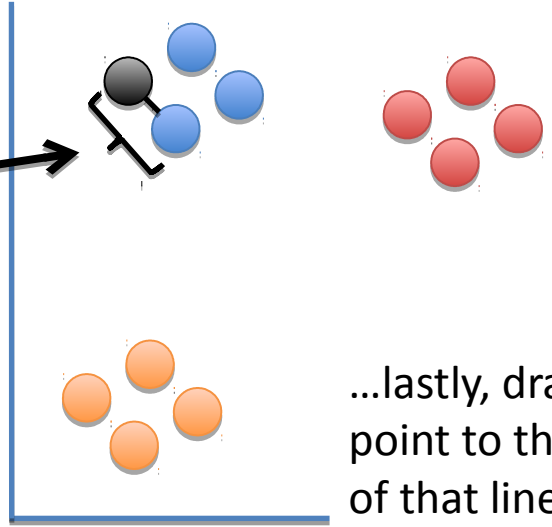


First, measure the distance between two points…

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

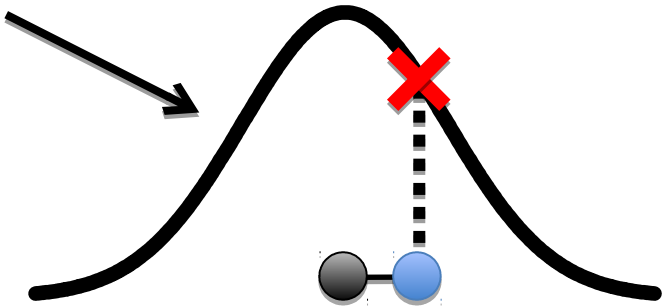


First, measure the distance between two points…

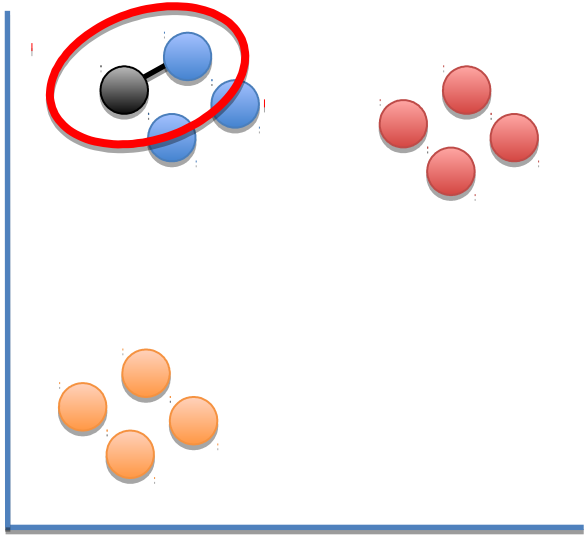


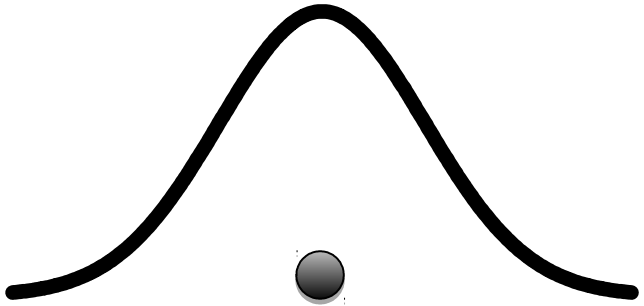
…lastly, draw a line from the point to the curve. The length of that line is the “similarity”.

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

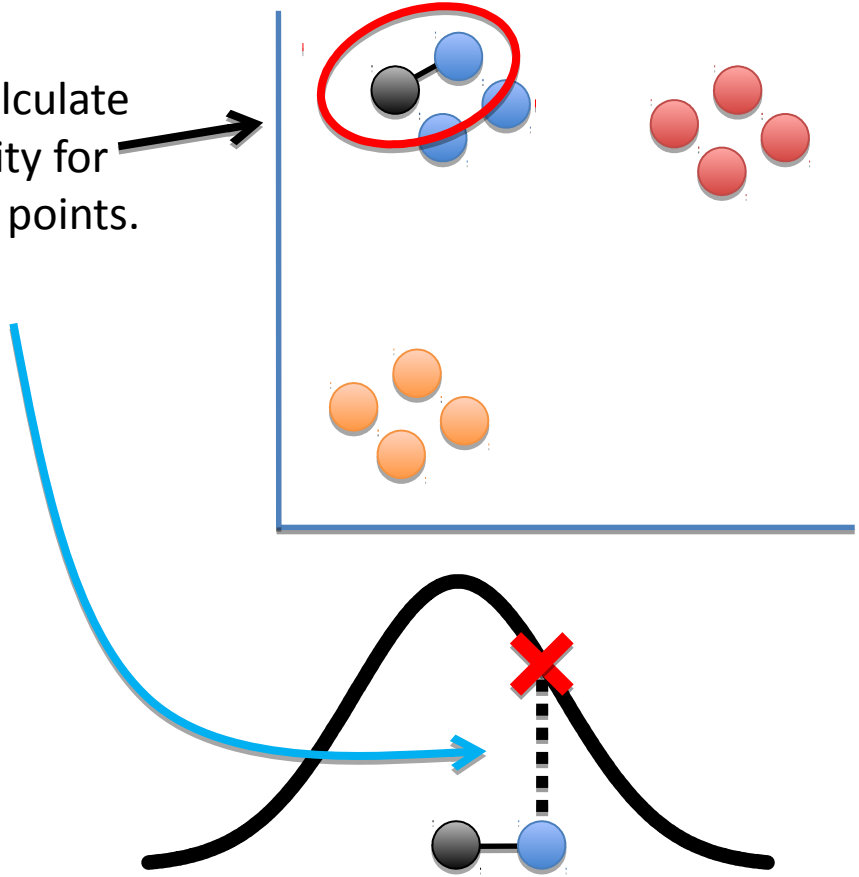


Now we calculate the similarity for this pair of points.

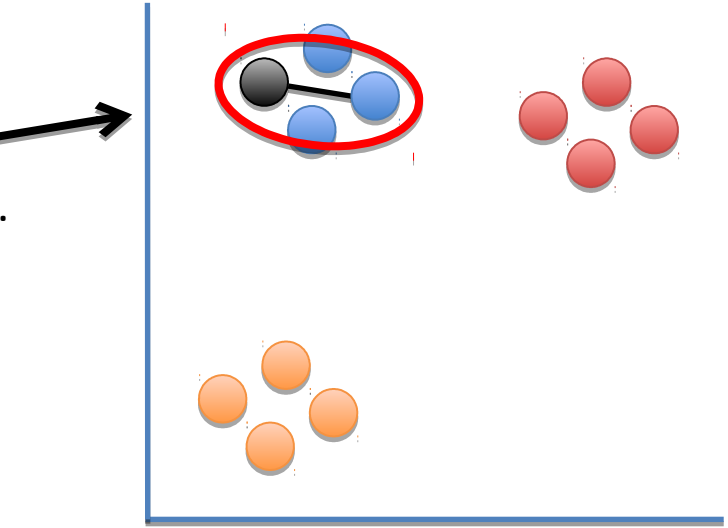


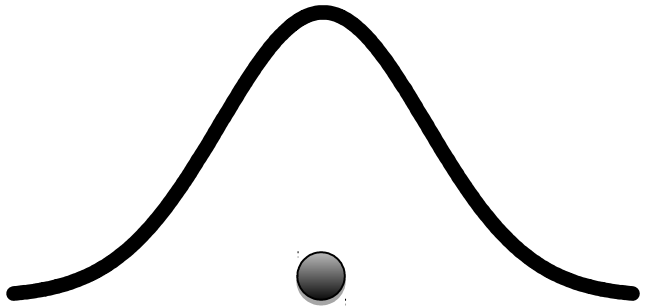


Now we calculate the similarity for this pair of points.

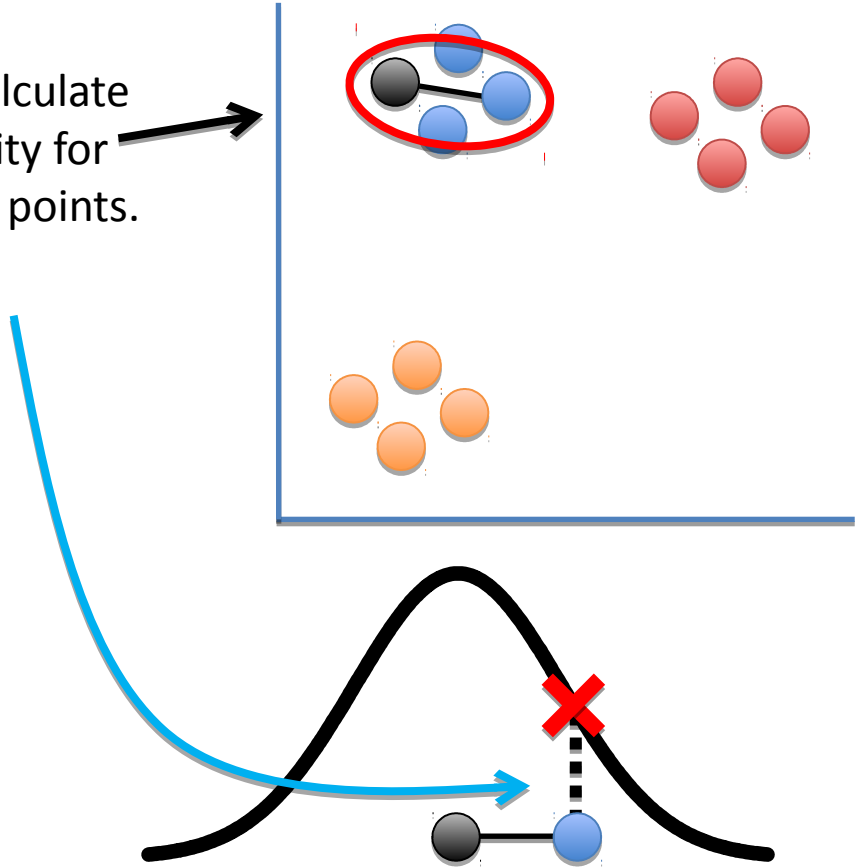


Now we calculate the similarity for this pair of points.

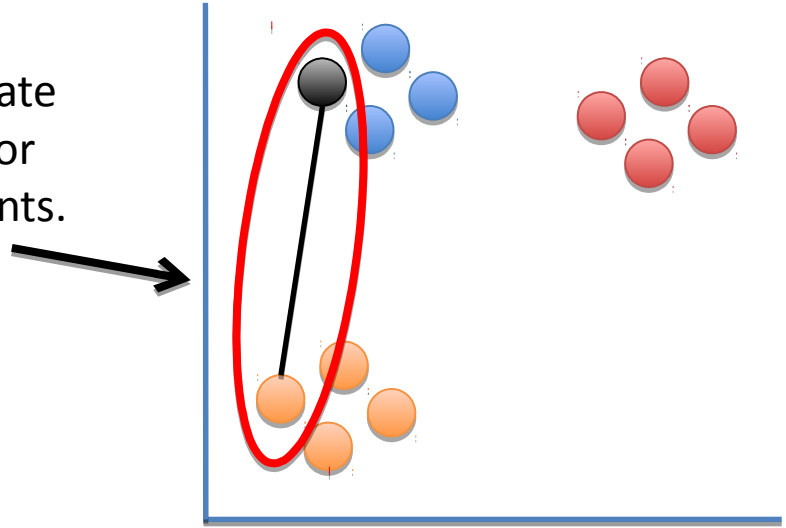


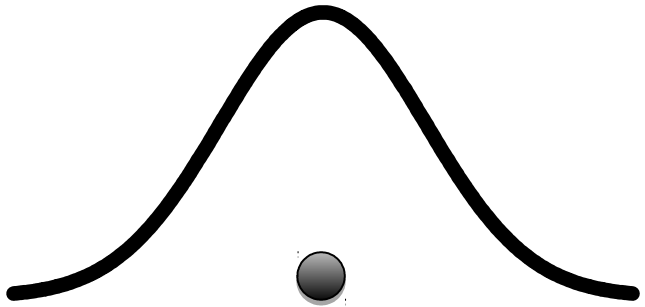


Now we calculate the similarity for this pair of points.

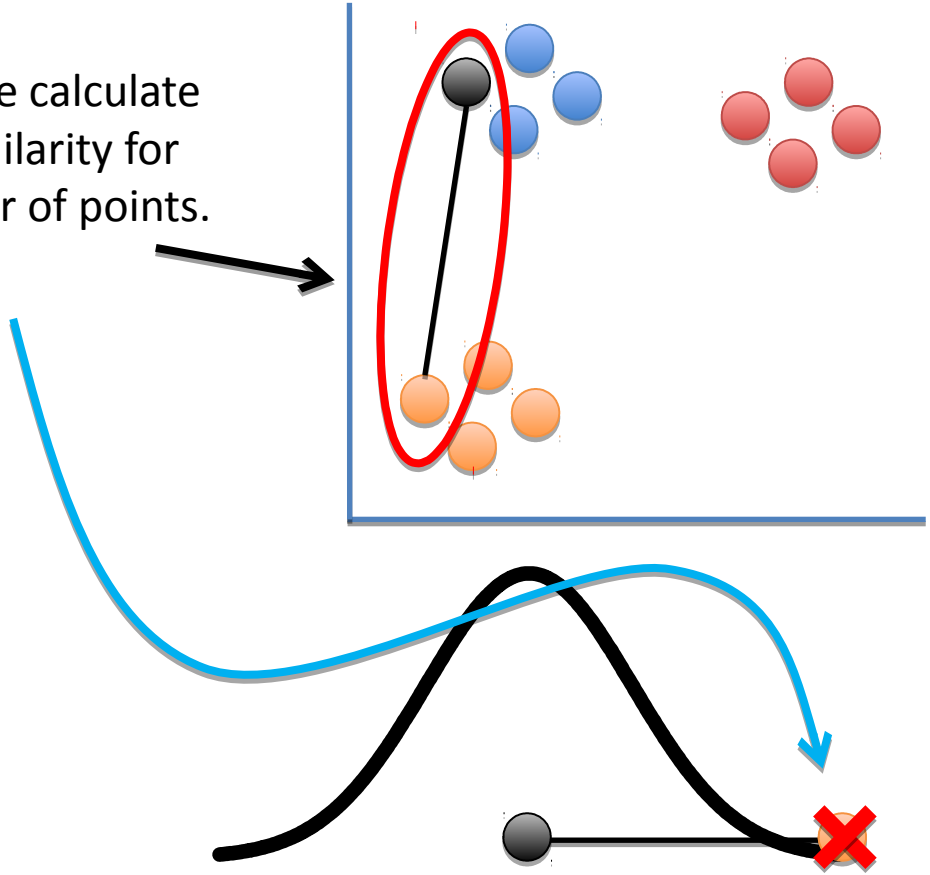


Now we calculate the similarity for this pair of points.

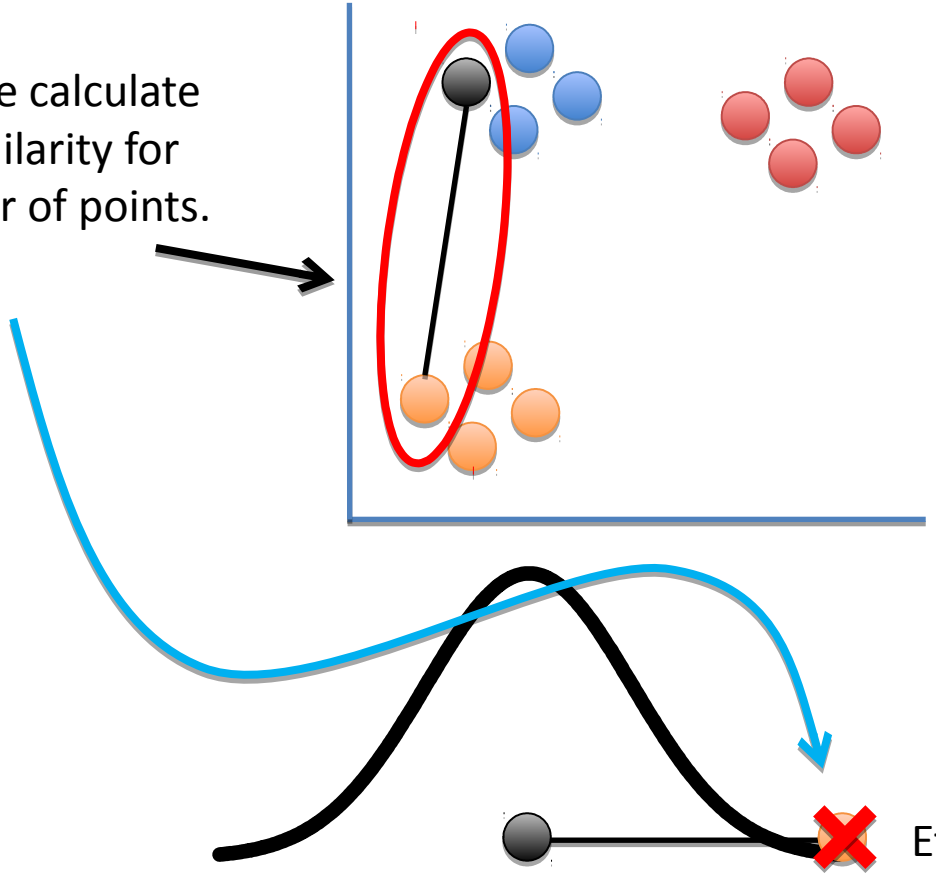




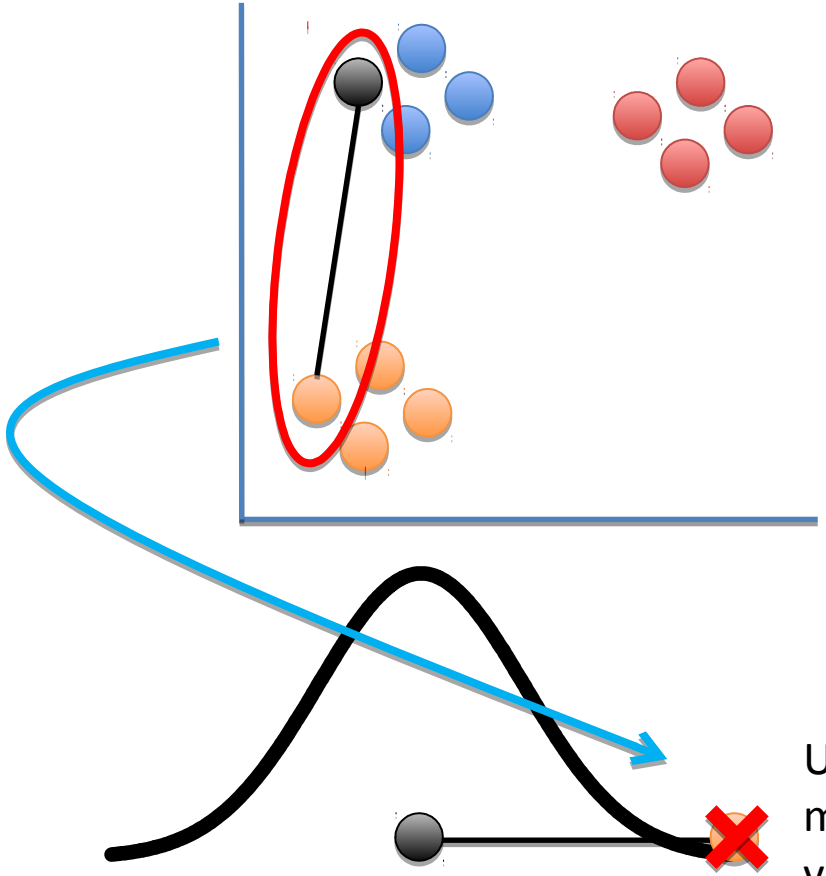
Now we calculate the similarity for this pair of points.



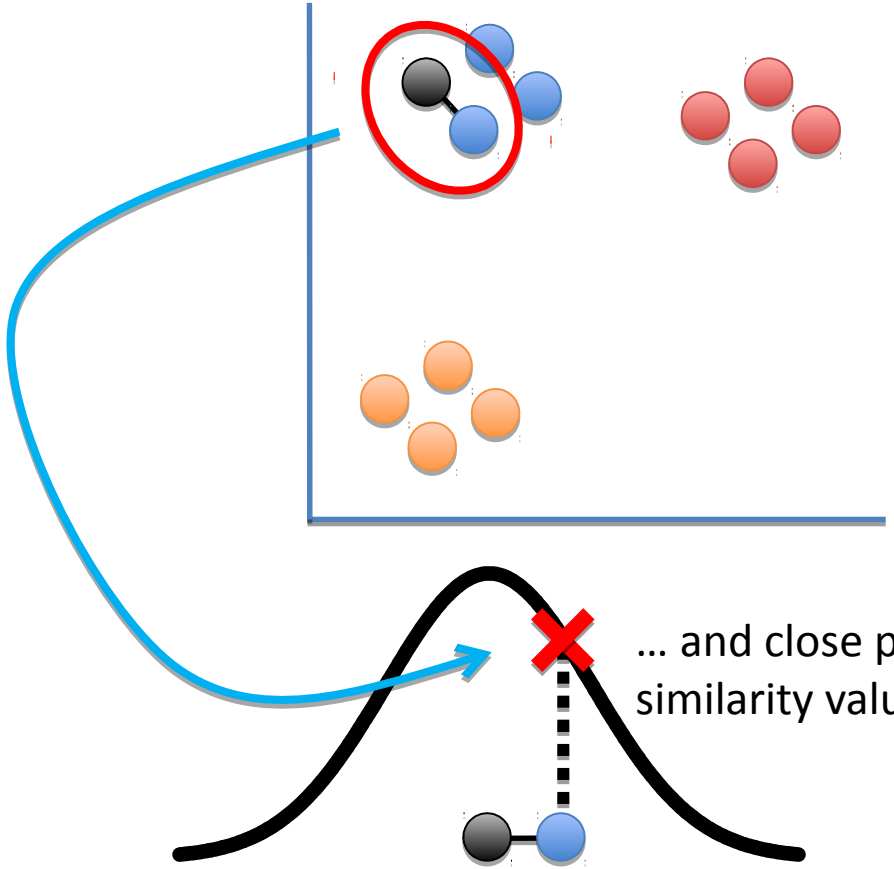
Now we calculate the similarity for this pair of points.



Etc. etc…

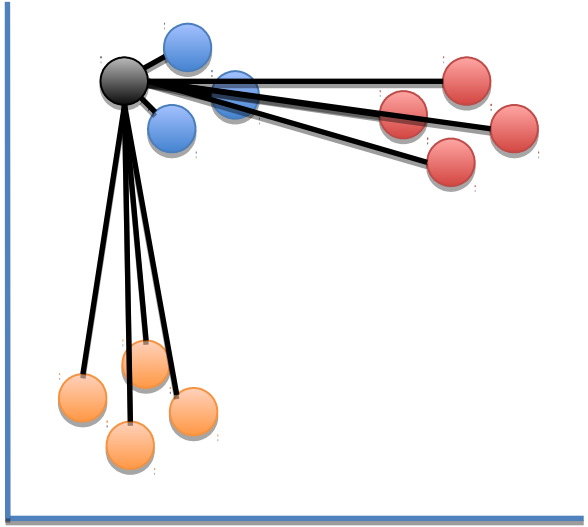


Using a normal distribution means that distant points have very low similarity values….

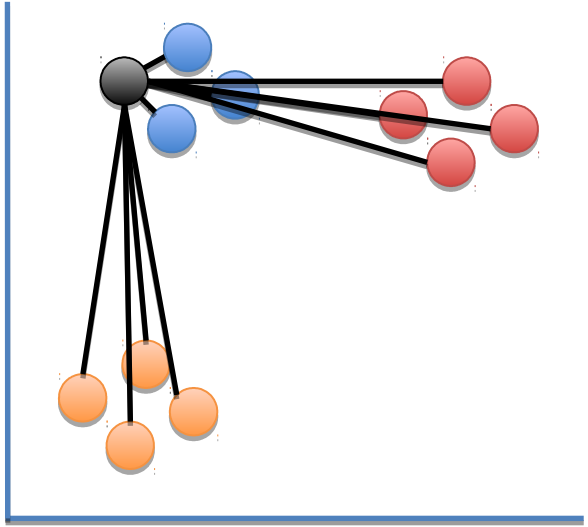


… and close points have high similarity values.

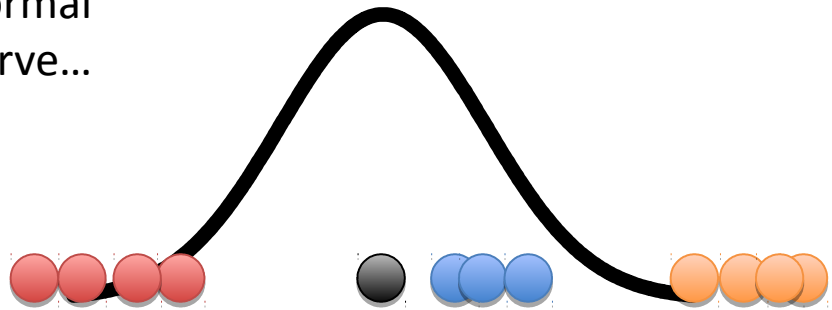
Ultimately, we measure the distances between all of the points and the point of interest…



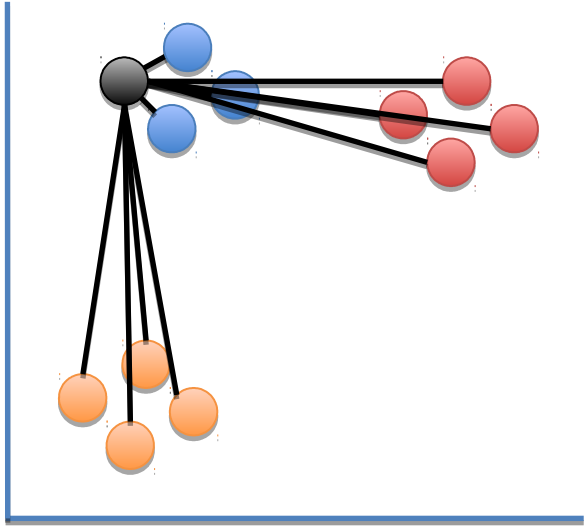
Ultimately, we measure the distances between all of the points and the point of interest…



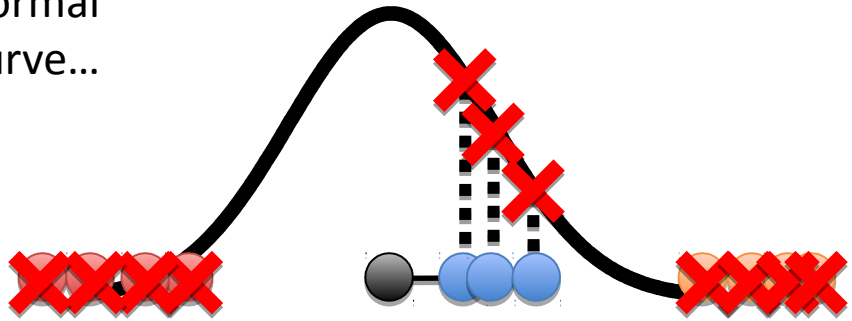
Plot them on the normal curve…



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

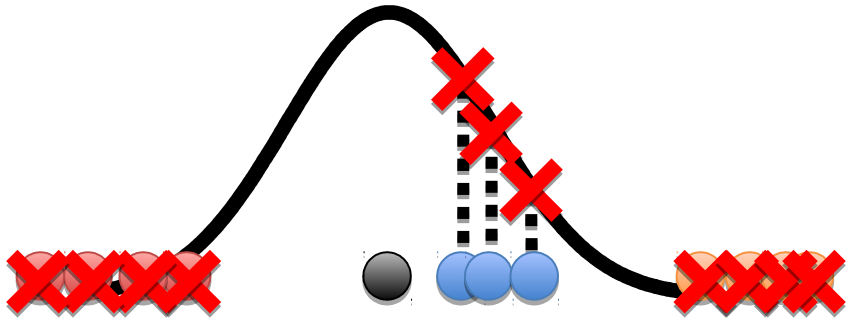


Plot them on the normal curve…



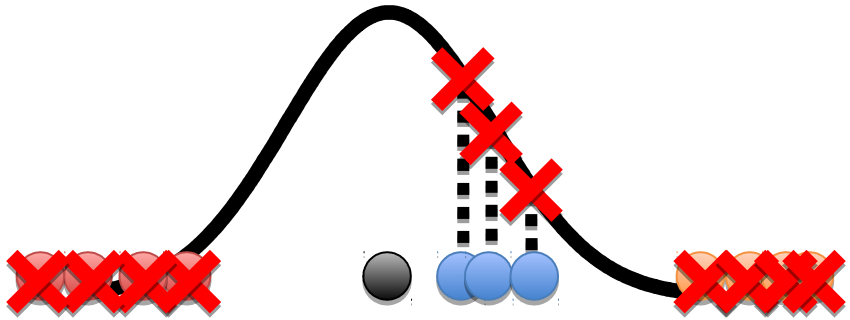
…and then measure the distances from the points to the curve to get the similarity scores with respect to the point of interest.

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

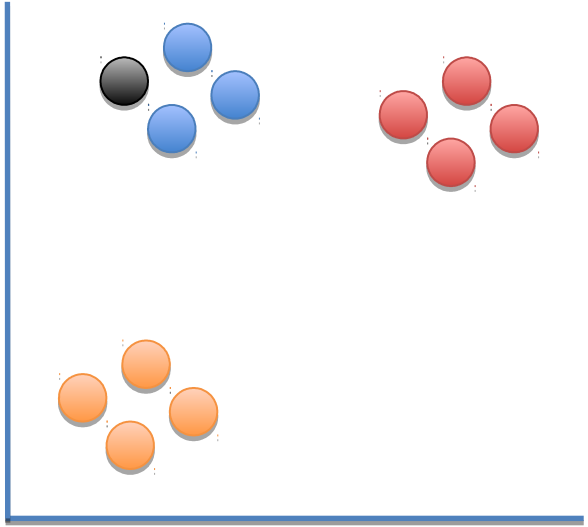


The next step is to scale the similarities so that they add up to 1.

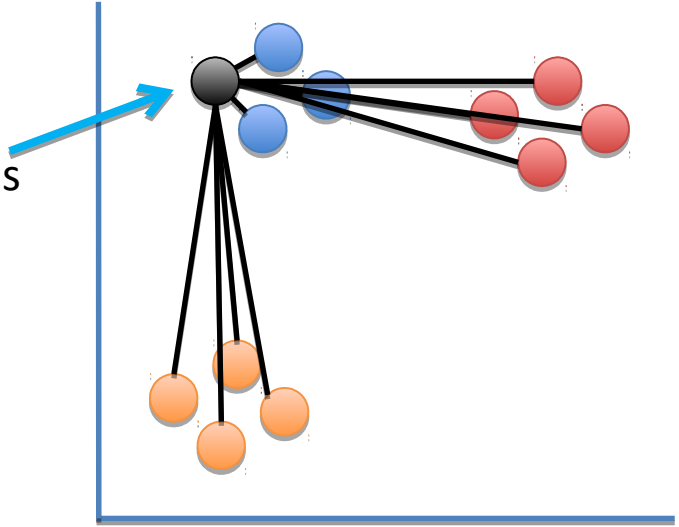
Umm… Why do the similarity scores need to add up to 1?



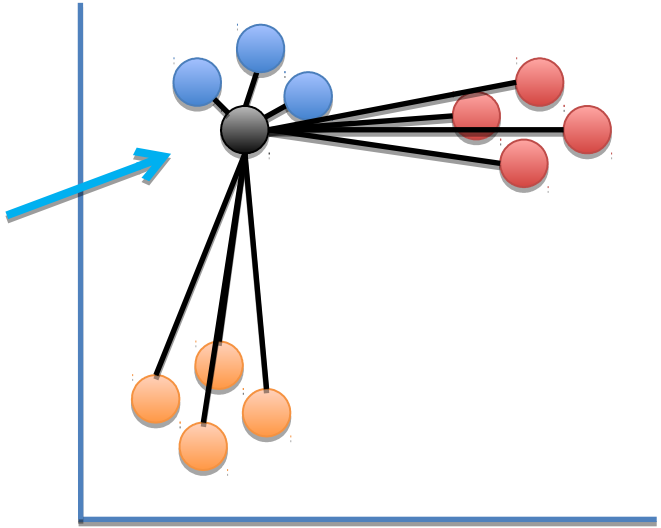
Now back to the simple scatter plot…



We've calculated similarity scores for this point.

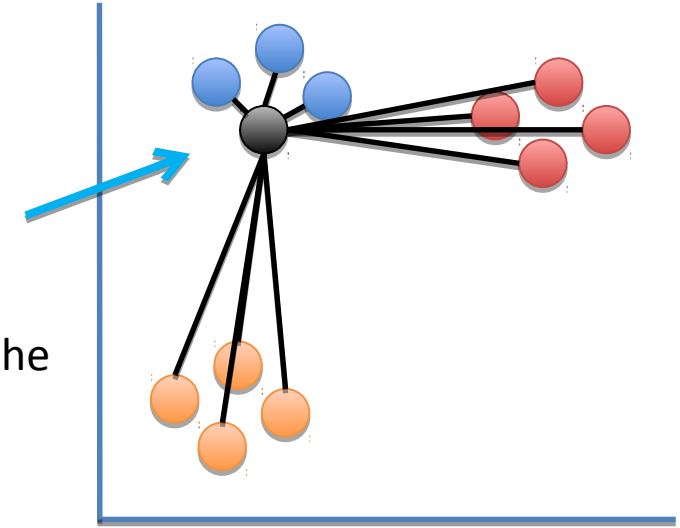


Now we do it for this point…

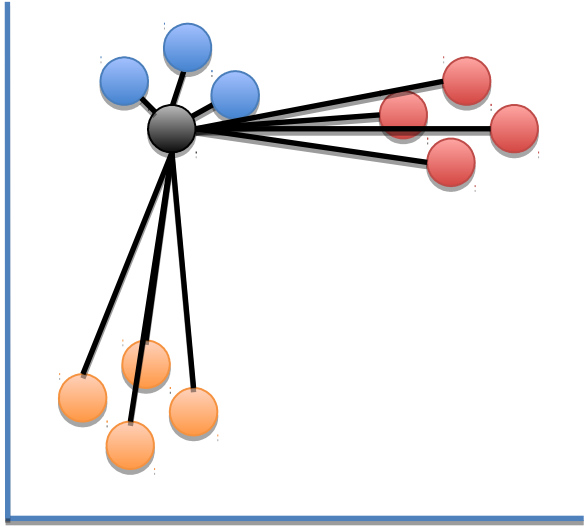


Now we do it for this point…

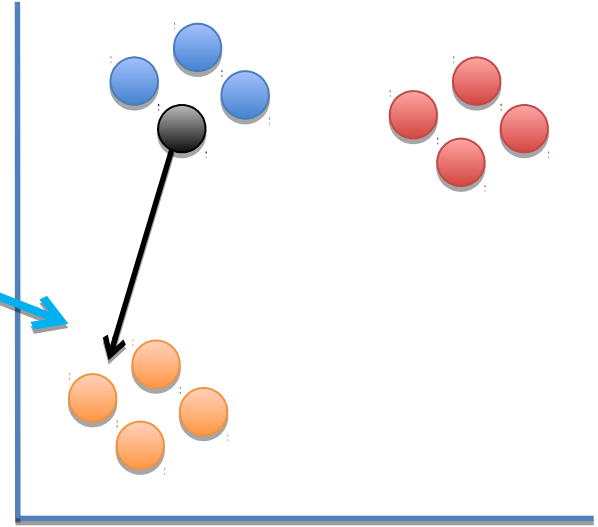
…and we do it for all the points.

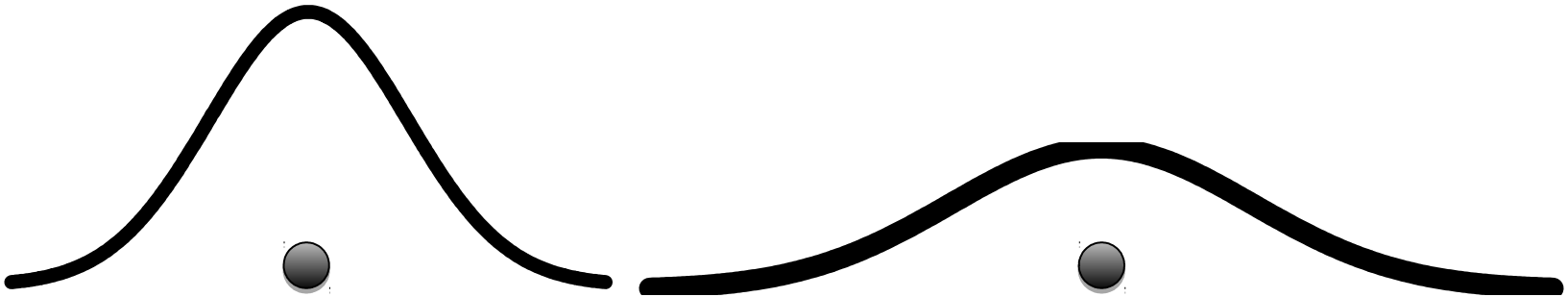


One last thing and the scatter plot will be all set with similarity scores!!!

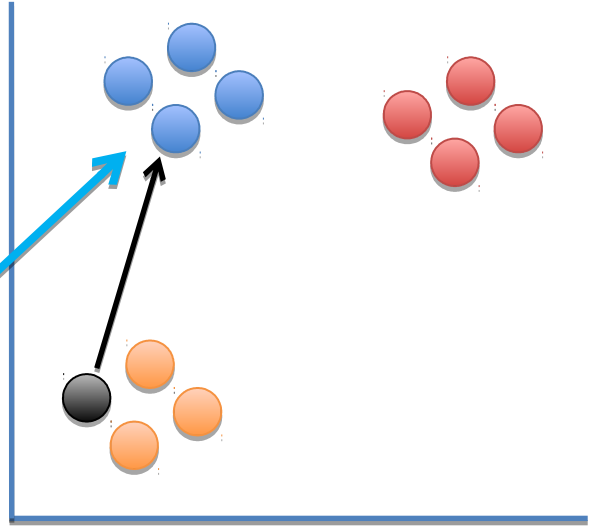


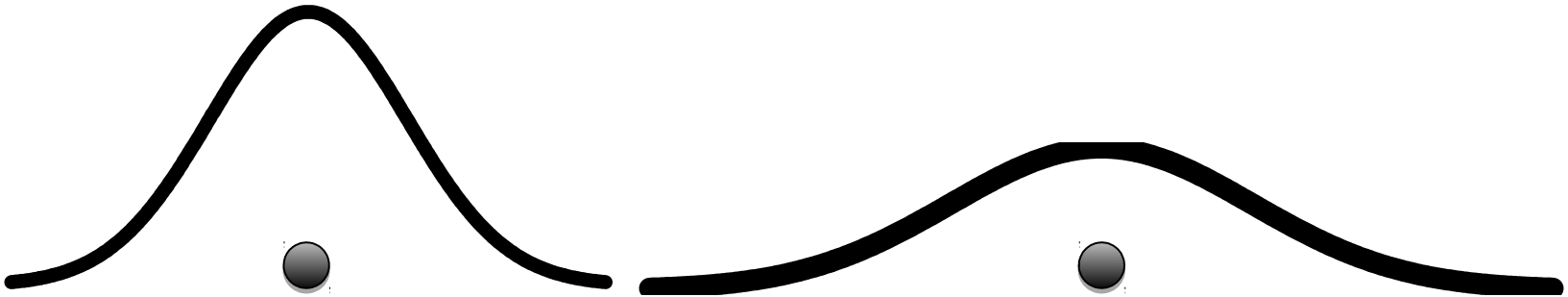
Because the width of the distribution is based on the density of the surrounding data points, the similarity score to this node…

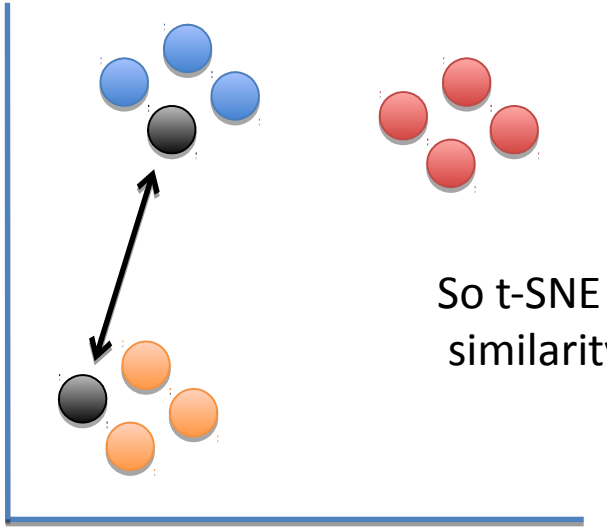




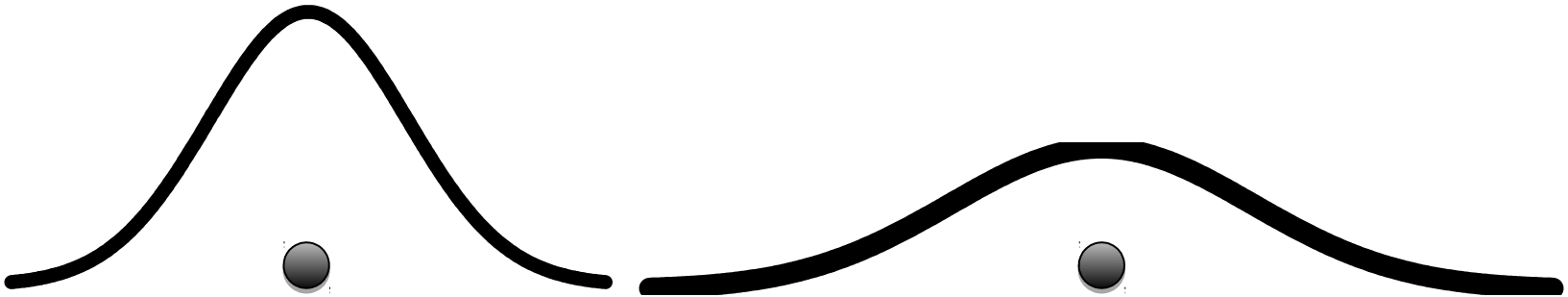
…might not be the same as the similarity to this node.

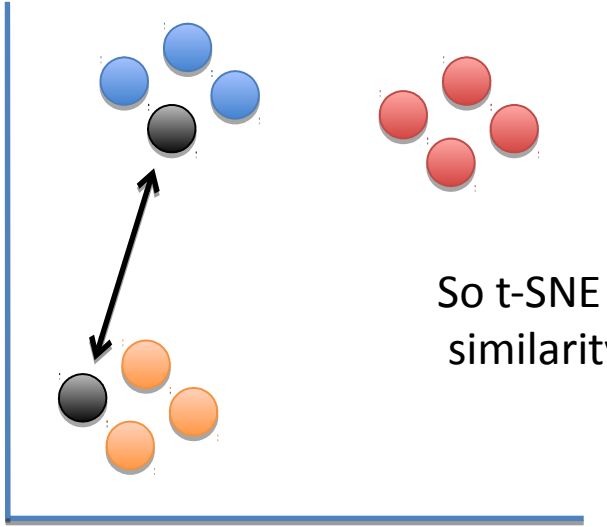






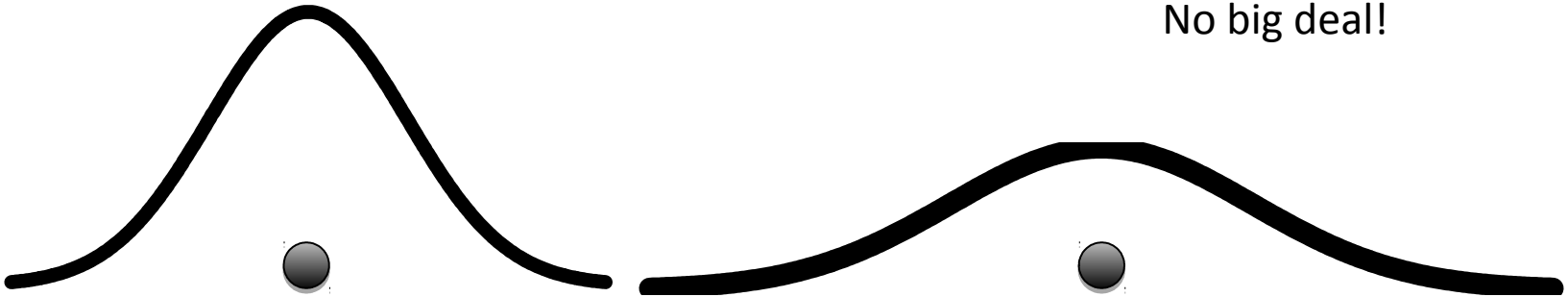
So t-SNE just averages the two similarity scores from the two directions…

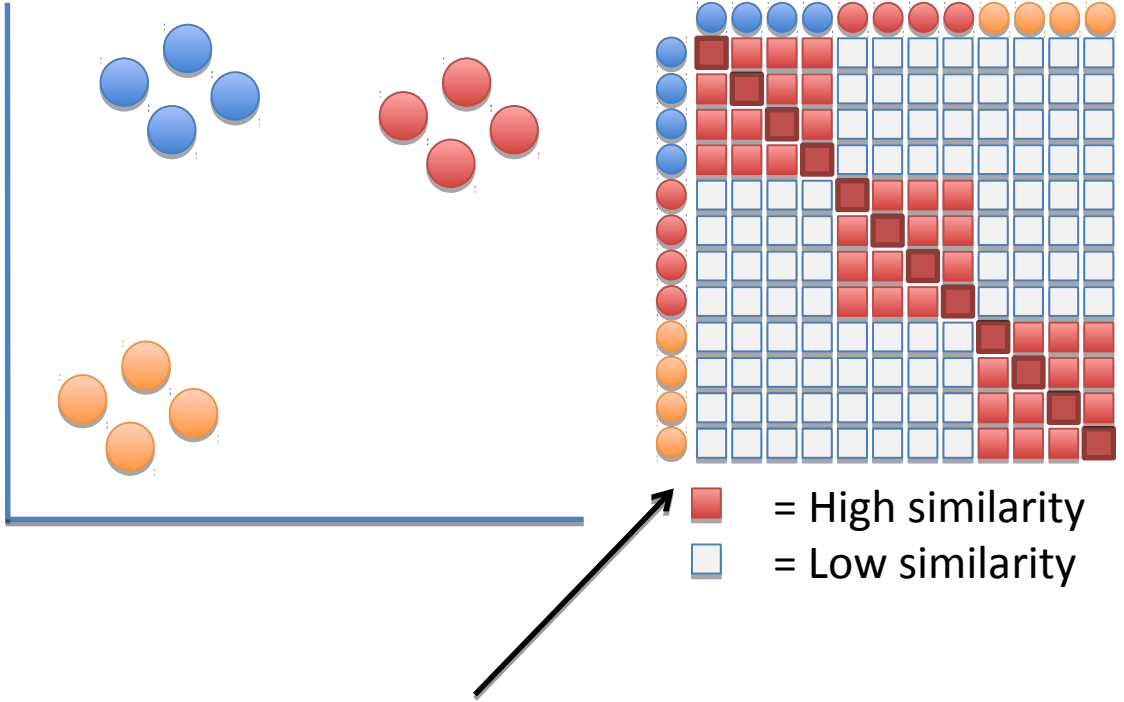




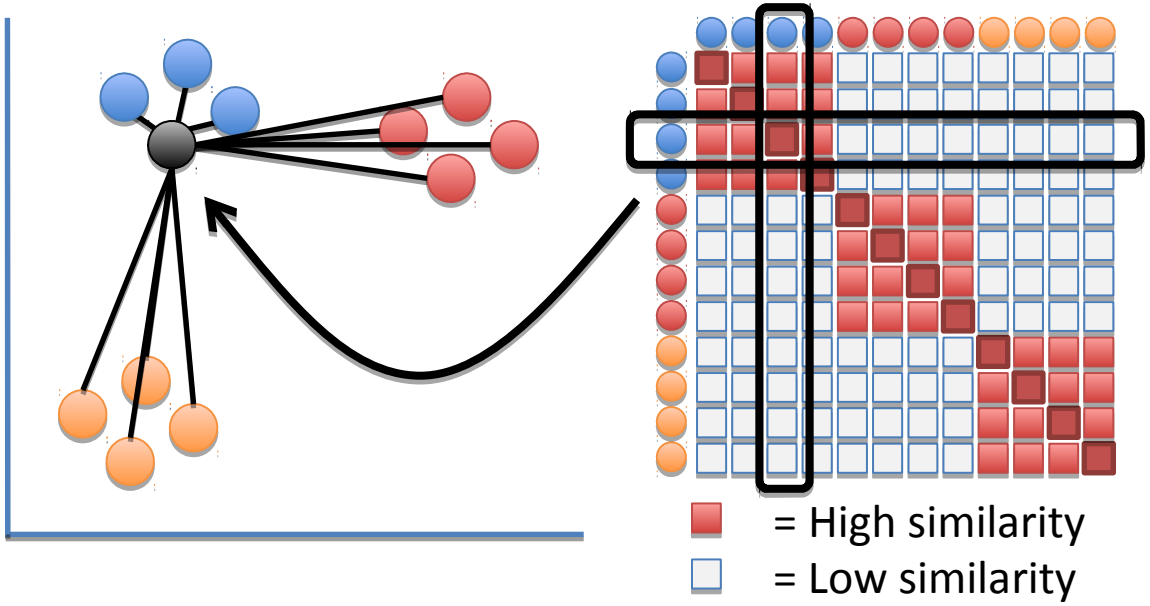
So t-SNE just averages the two similarity scores from the two directions…

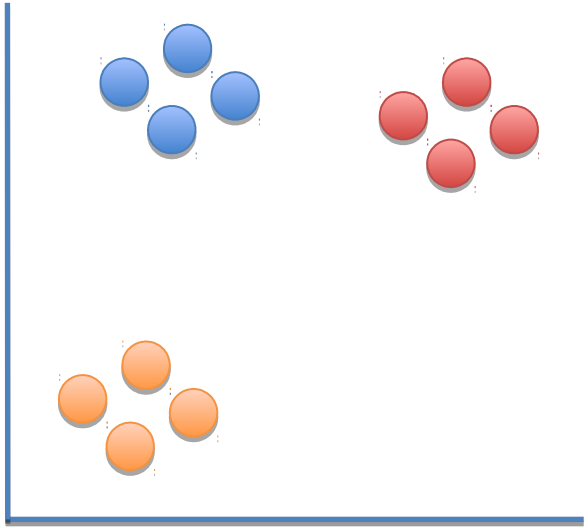
No big deal!

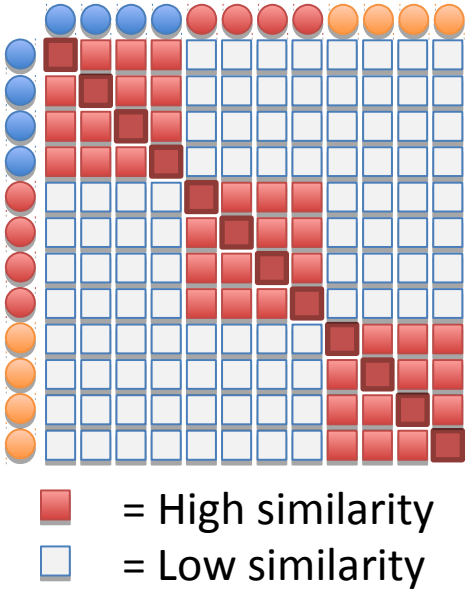




Ultimately, you end up with a matrix of similarity scores.

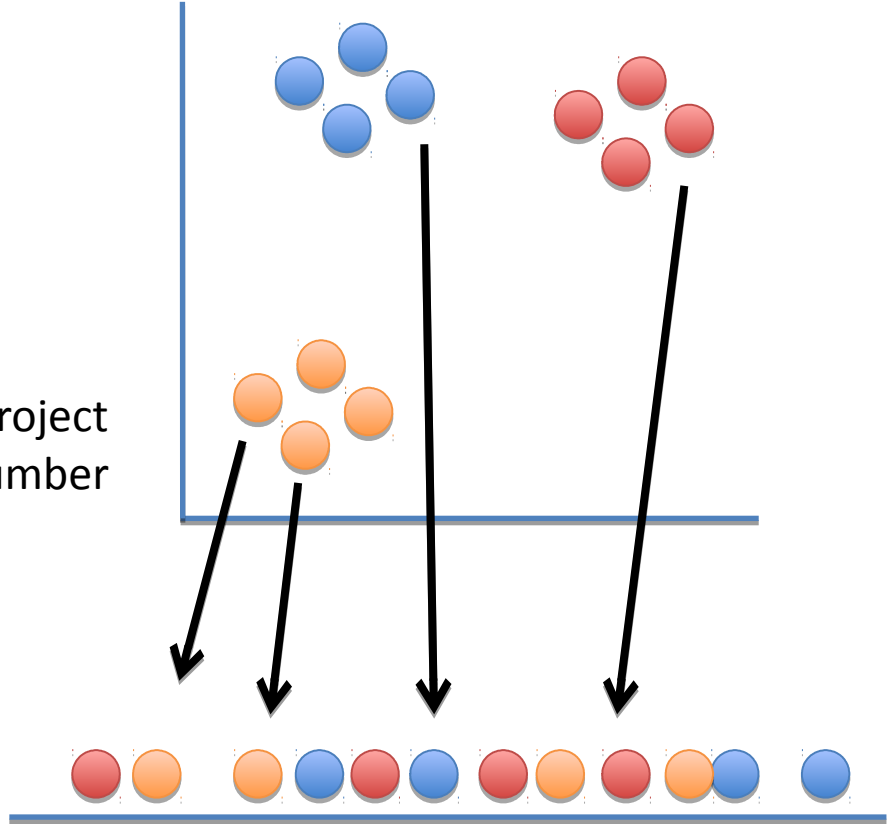

= High similarity
= Low similarity



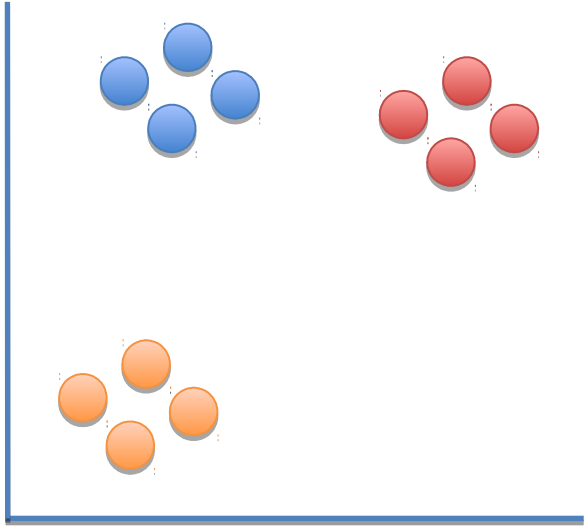


Hooray!!! We're done doing calculating similarity scores for the scatter plot!

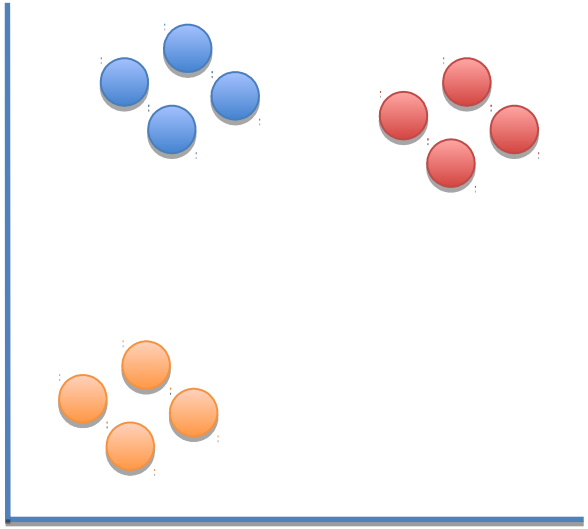
Now we randomly project the data onto the number line…



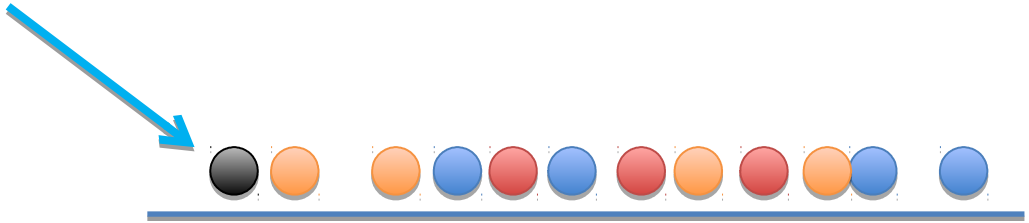
Now we randomly project the data onto the number line…

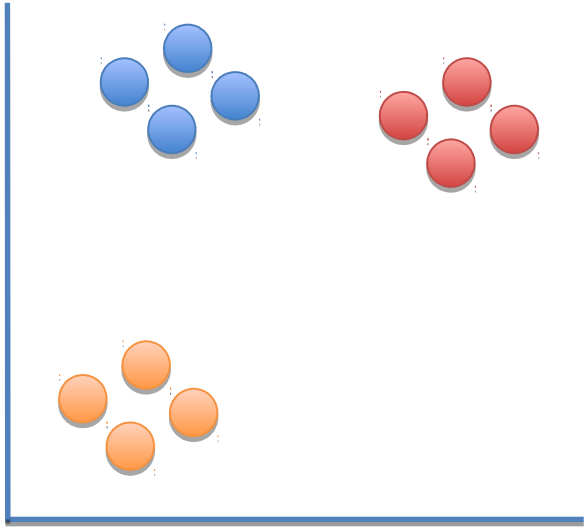


… and calculate similarity scores for the points on the number line.



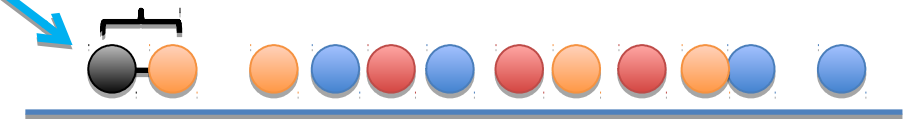
Just like before, that means picking a point…

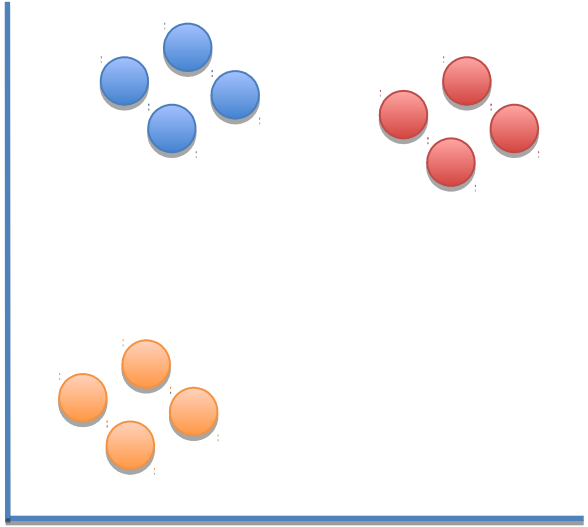




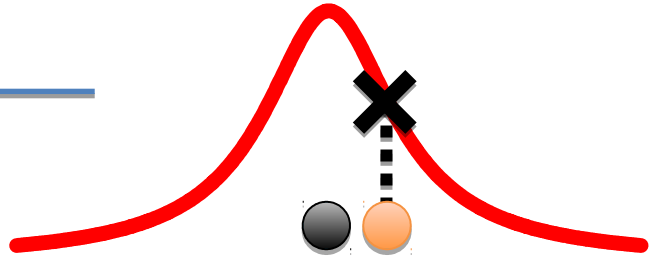
Just like before, that means picking a point…

…measuring a distance…



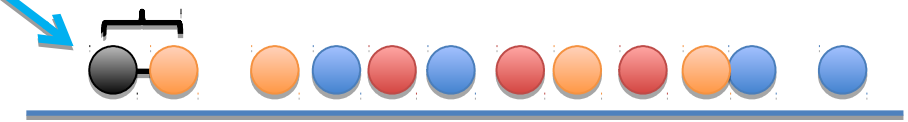


…and lastly, drawing a line from the point to a curve. However, this time we're using a "t-distribution".

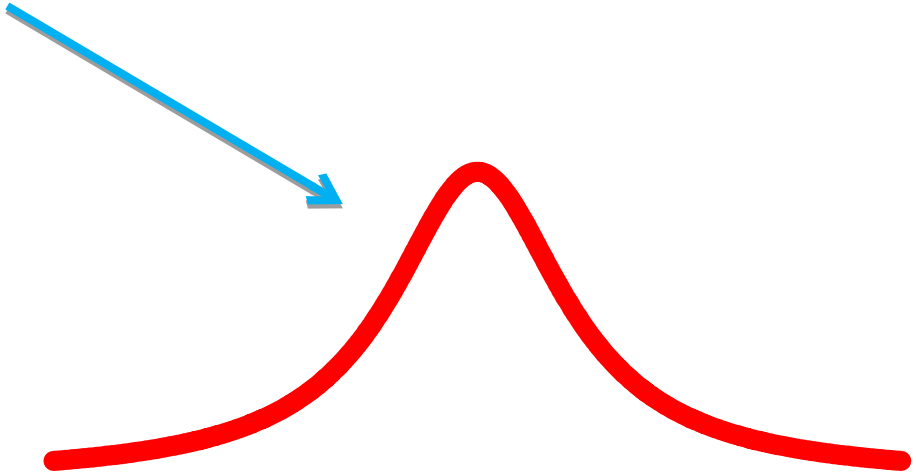


Just like before, that means picking a point…

…measuring a distance…

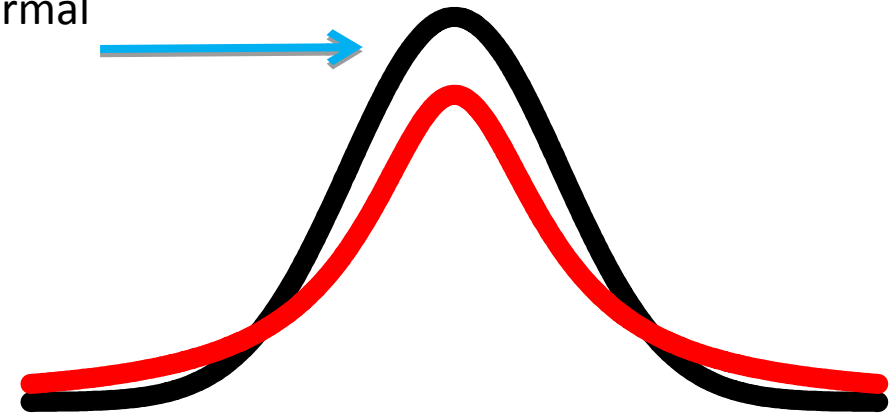


A “t-distribution”…



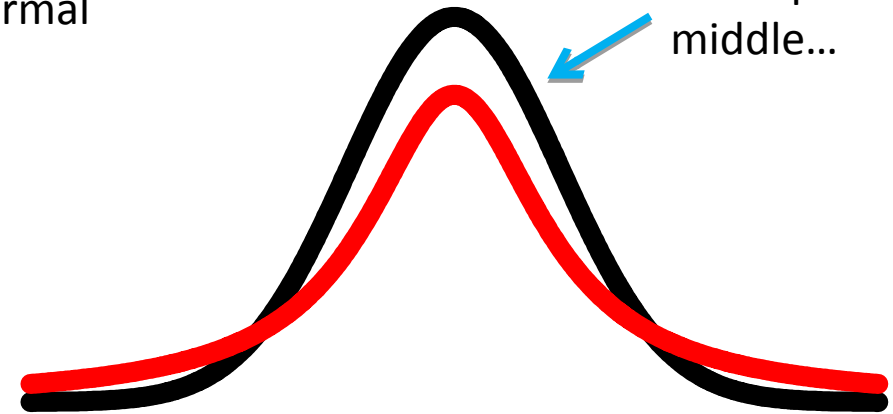
A “t-distribution”…

…is a lot like a normal distribution



A “t-distribution”…

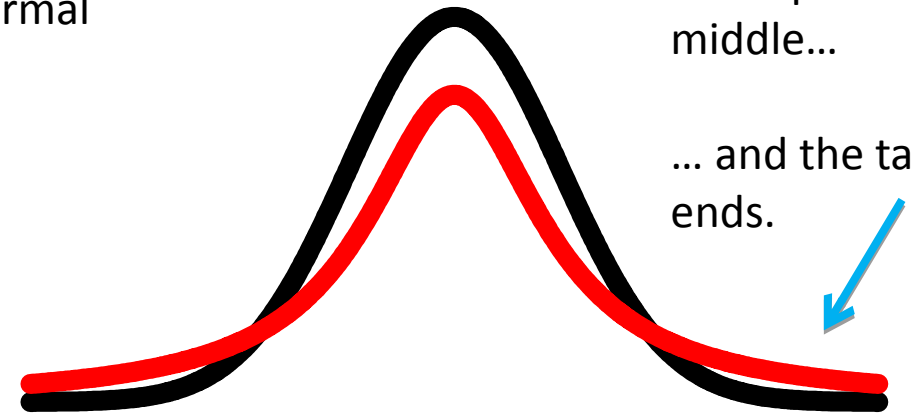
…is a lot like a normal distribution…



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

A “t-distribution”…

…is a lot like a normal distribution…

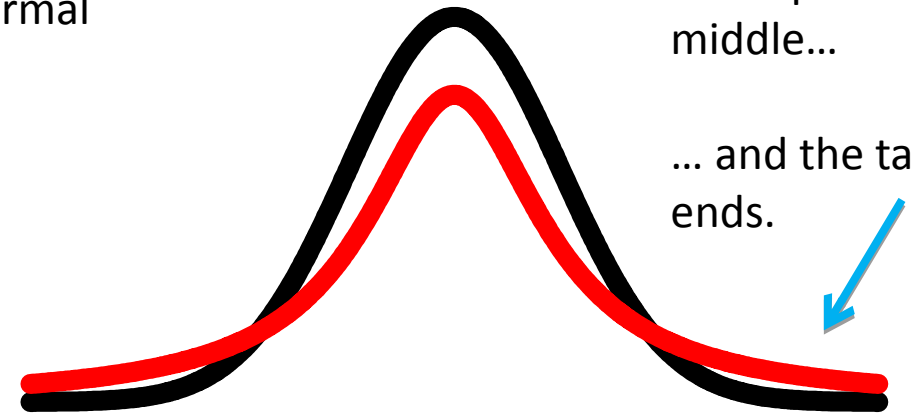


…except the “t” isn’t as tall in the middle…

… and the tails are taller on the ends.

A “t-distribution”…

…is a lot like a normal distribution…



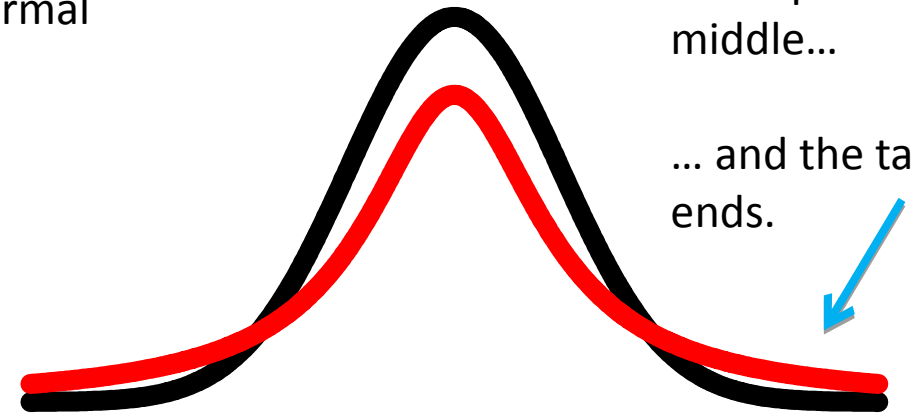
…except the “t” isn’t as tall in the middle…

… 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…



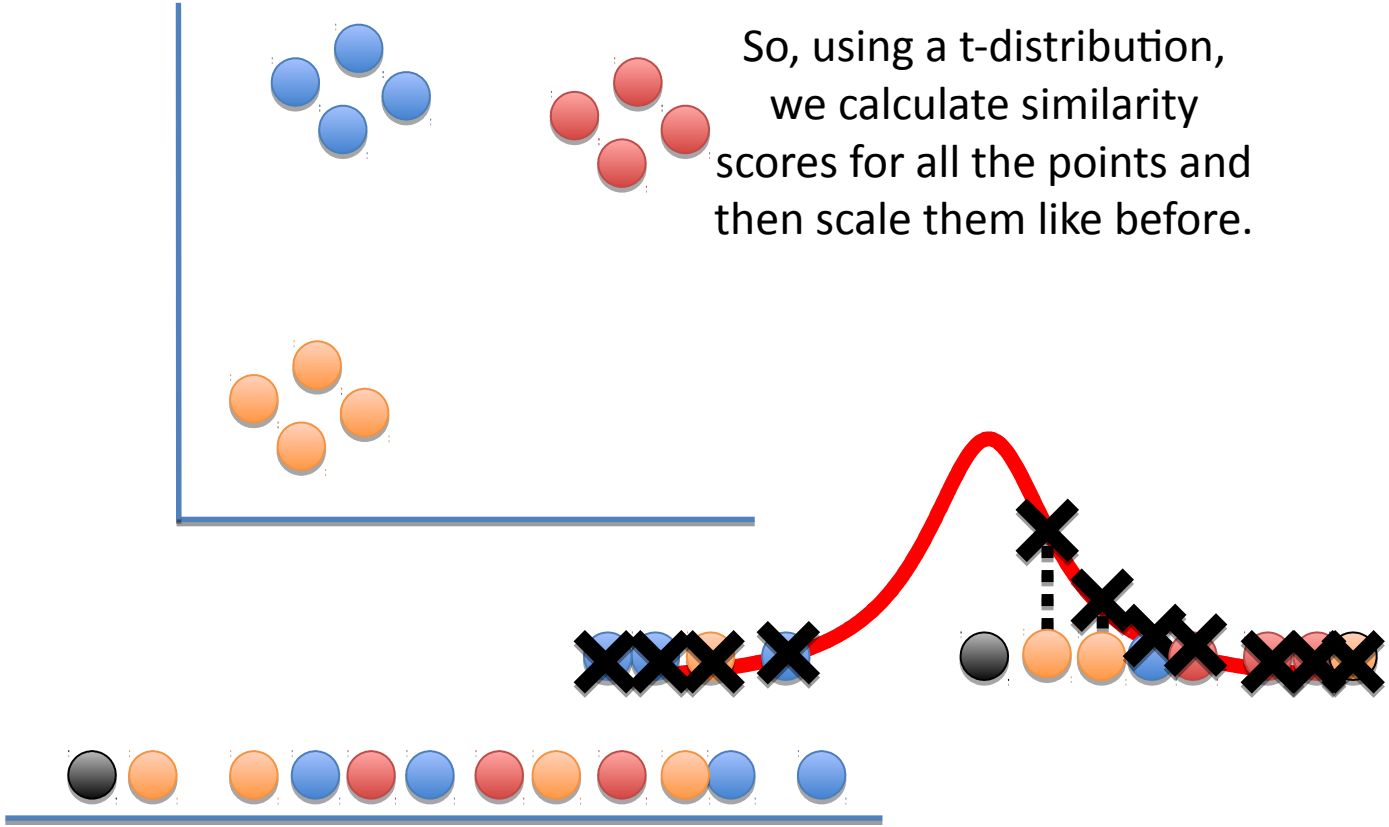
…except the “t” isn’t as tall in the middle…

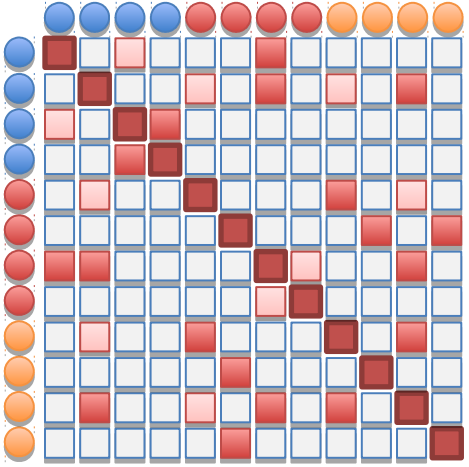
… and the tails are taller on the ends.

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

We’ll talk about why the t-distribution is used in a bit…

So, using a t-distribution, we calculate similarity scores for all the points and then scale them like before.

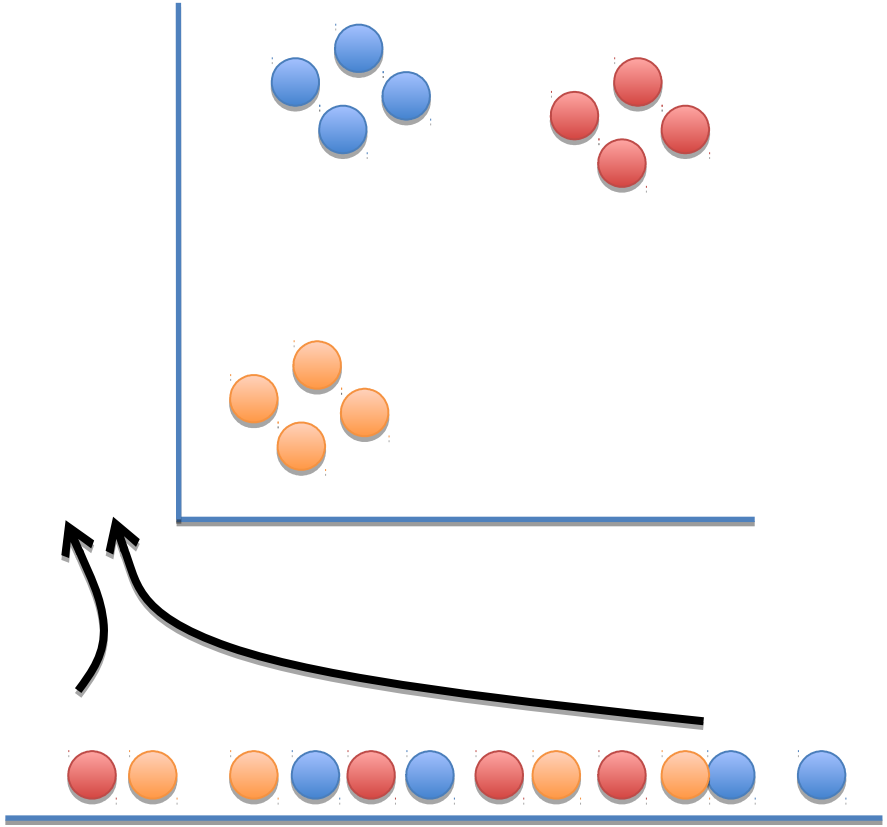


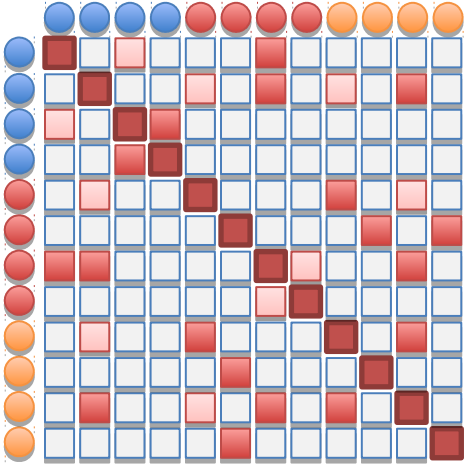


= High similarity

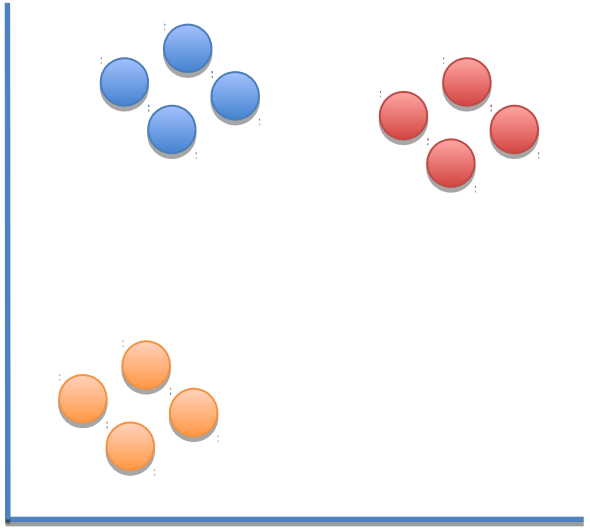
= Low similarity

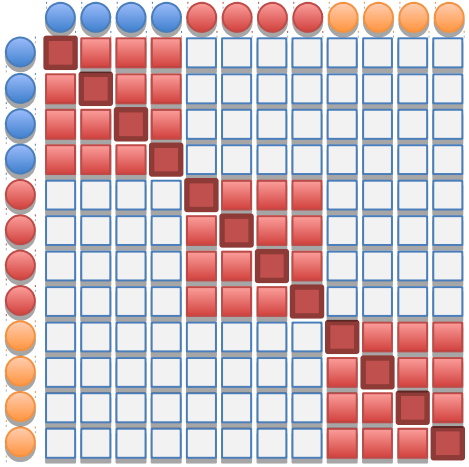
Like before, we end up with a matrix of similarity scores, but this matrix is a mess…





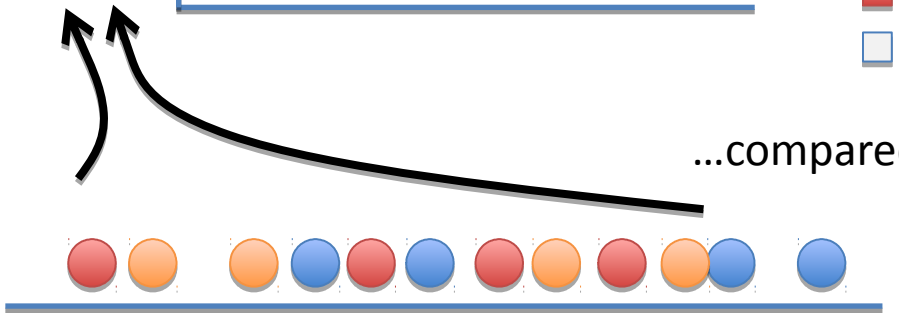
= High similarity
= Low similarity



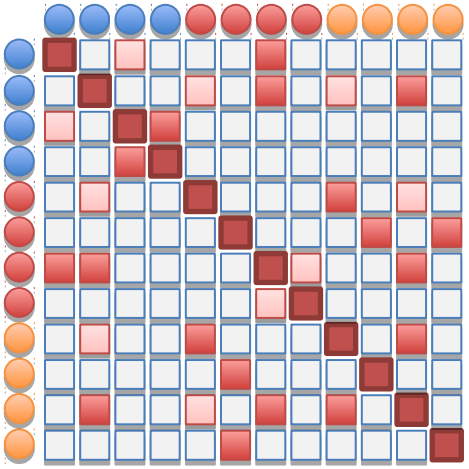


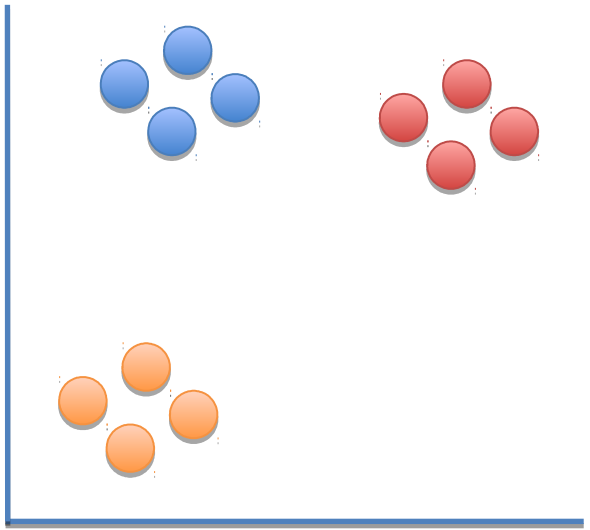
= High similarity
= Low similarity

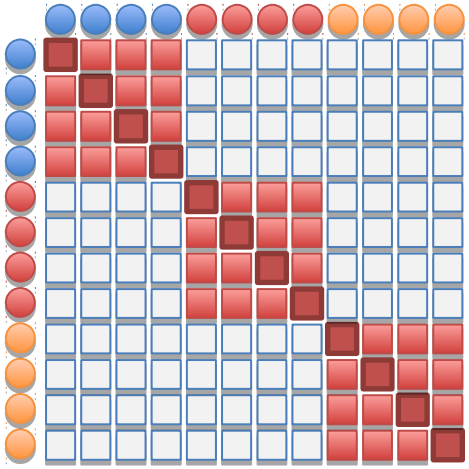
Like before, we end up with a matrix of similarity scores, but this matrix is a mess…



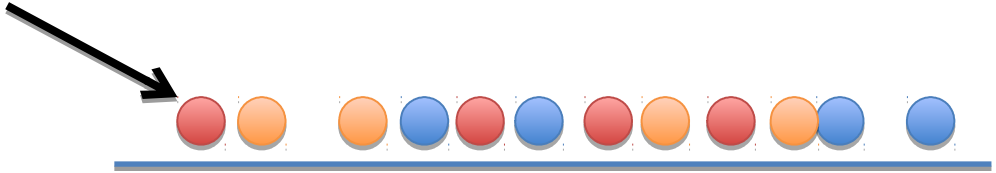
…compared to the original matrix.

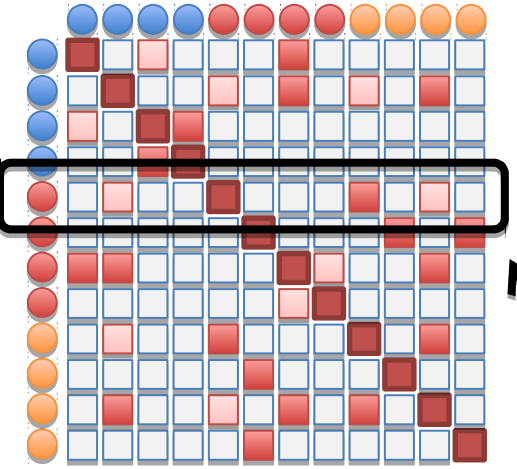


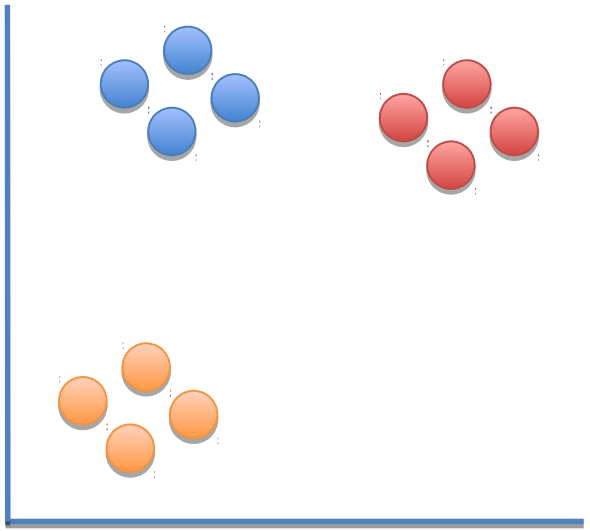


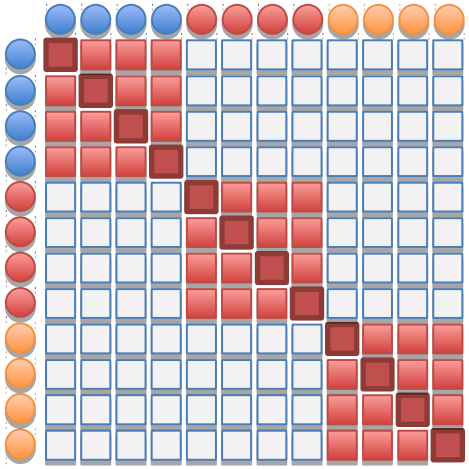


The goal of moving this point is…



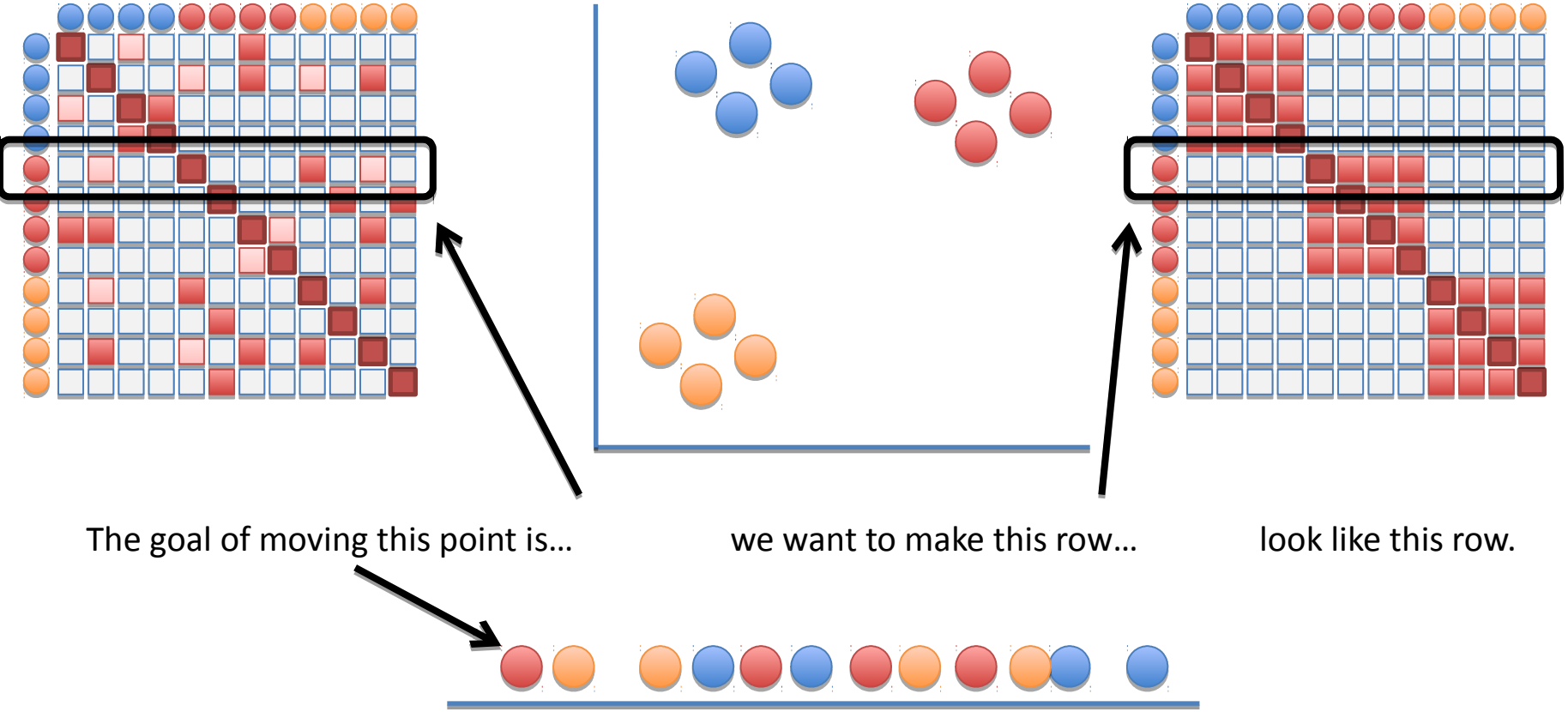


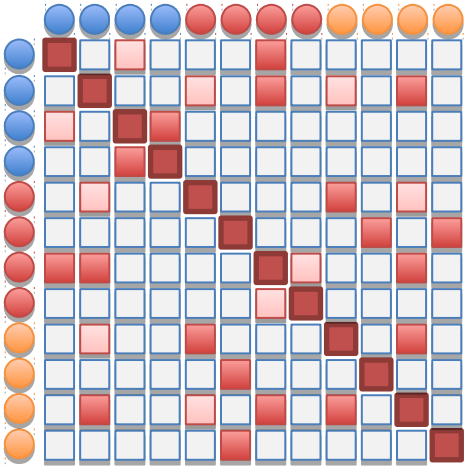


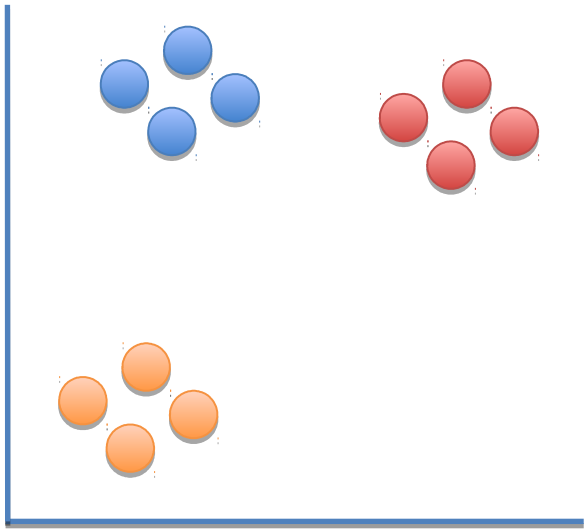


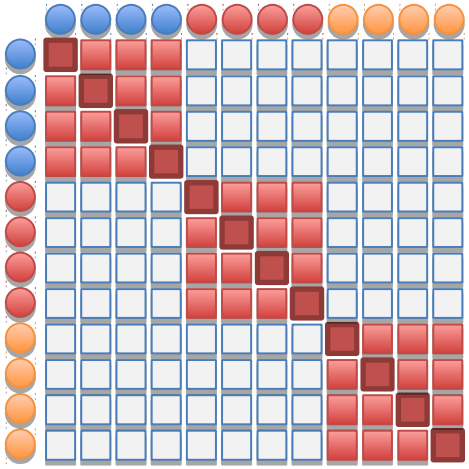
The goal of moving this point is…

we want to make this row…

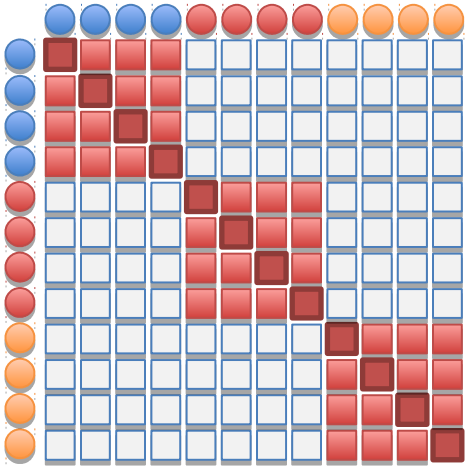

The goal of moving this point is…
we want to make this row…
look like this row.

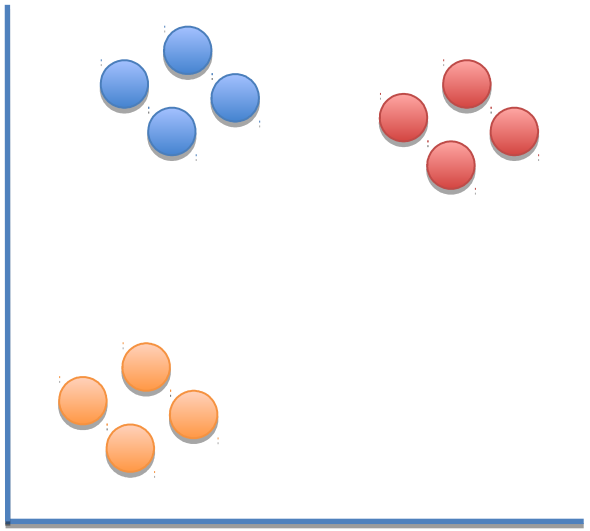


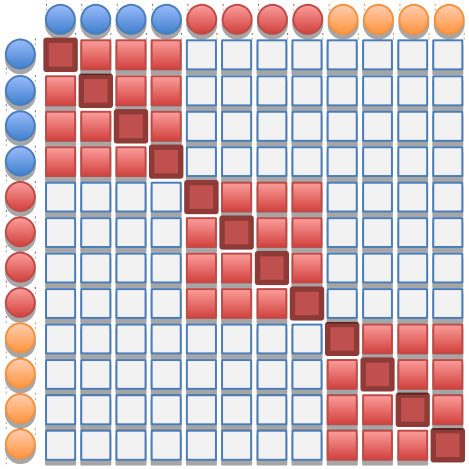




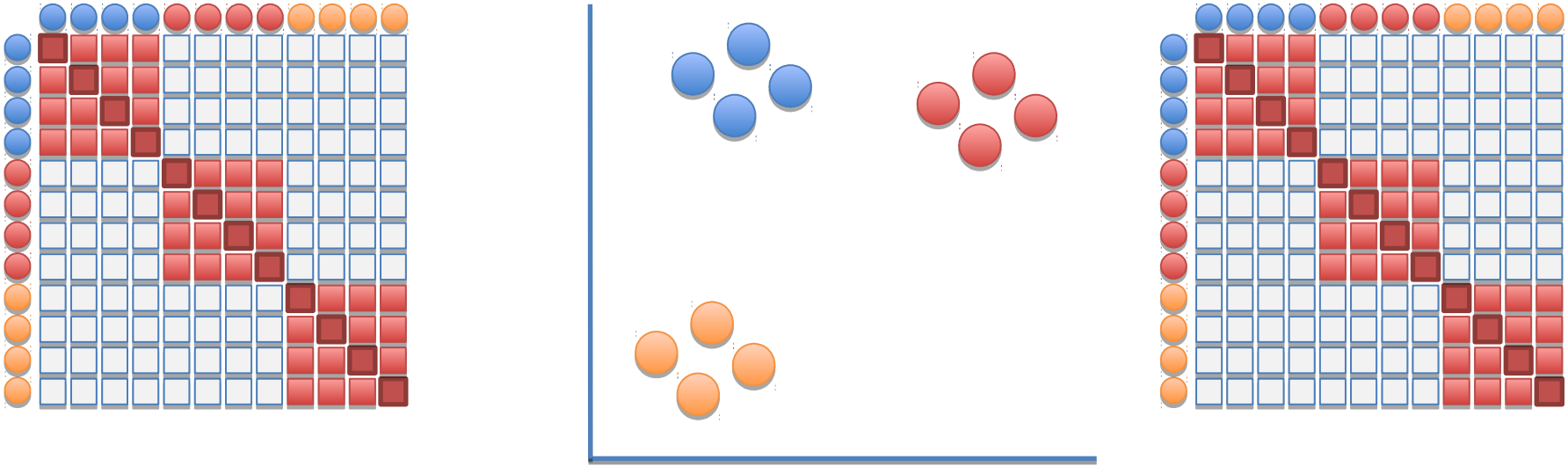
t-SNE moves the points a little bit at a time, and each step it chooses a direction that makes the matrix on the left more like the matrix on the right.



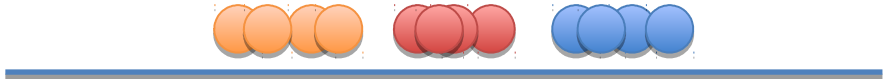




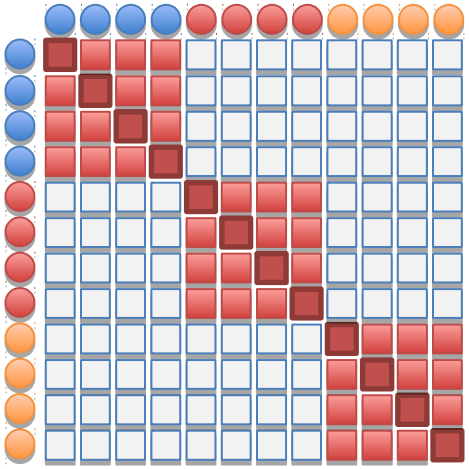
t-SNE moves the points a little bit at a time, and each step it chooses a direction that makes the matrix on the left more like the matrix on the right.

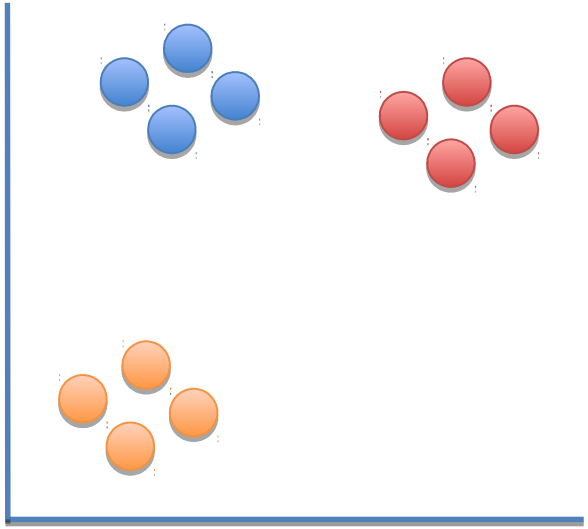


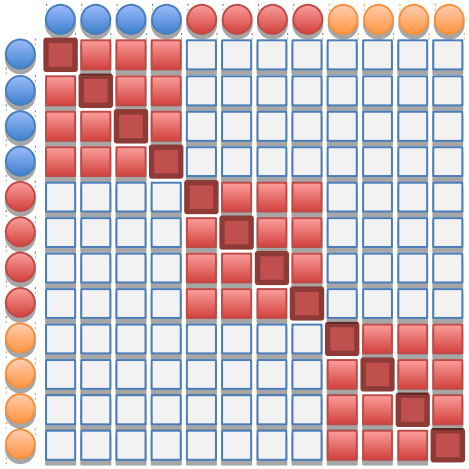
t-SNE moves the points a little bit at a time, and each step it chooses a direction that makes the matrix on the left more like the matrix on the right.

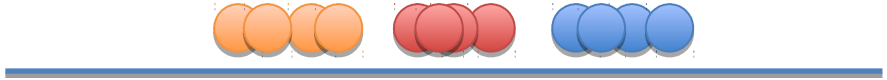


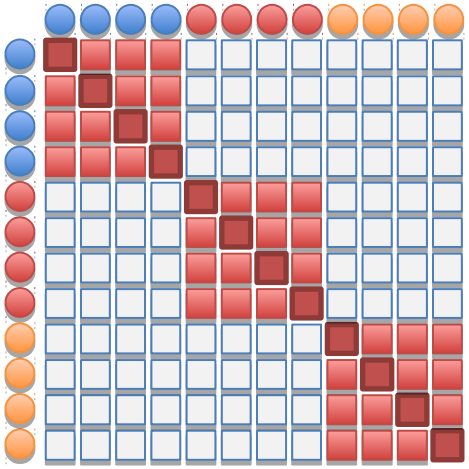
It uses small steps, because it's a little bit like a chess game and can't be solved all at once. Instead, it goes one move at at time.

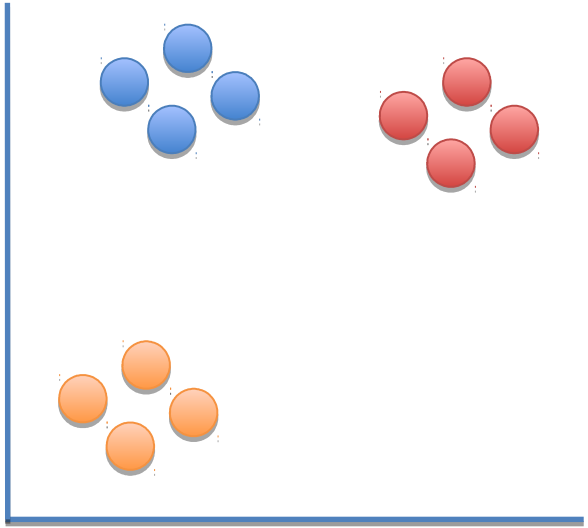


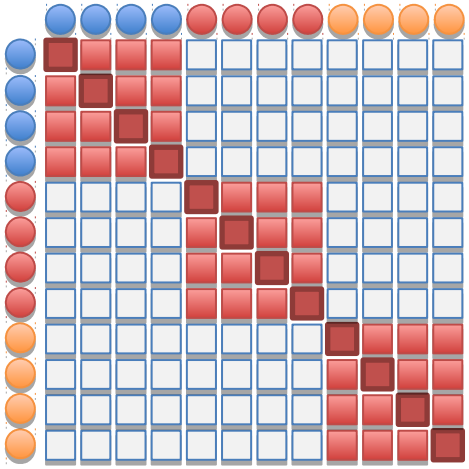




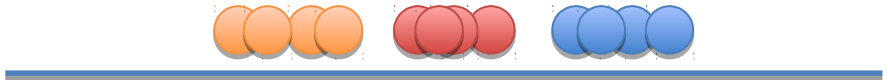


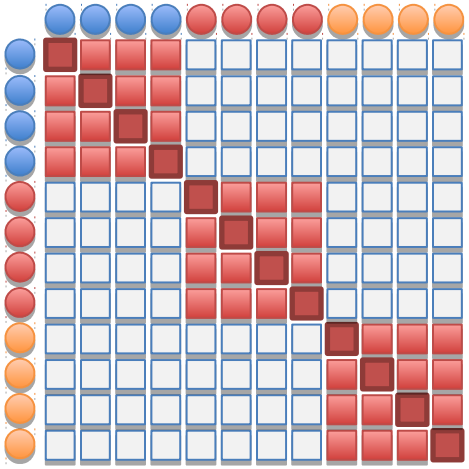


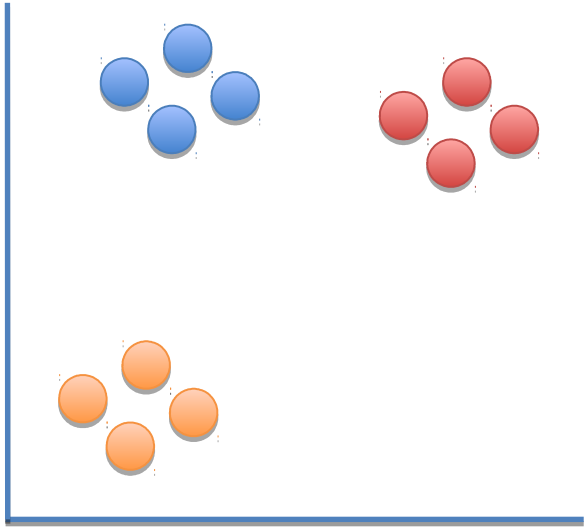


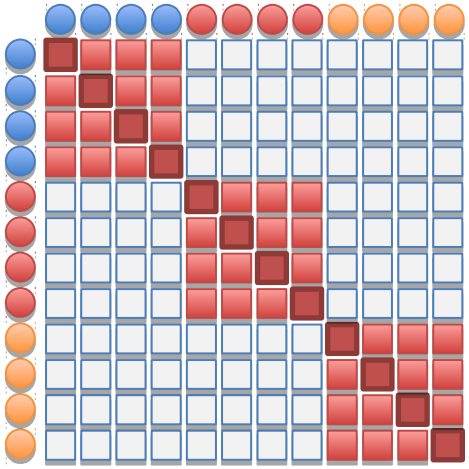


Now to finally tell you why the “t-distribution” is used…

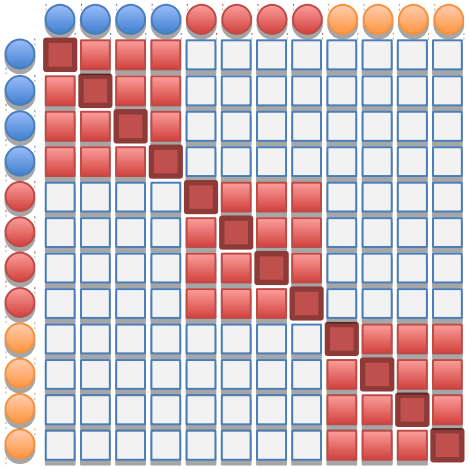


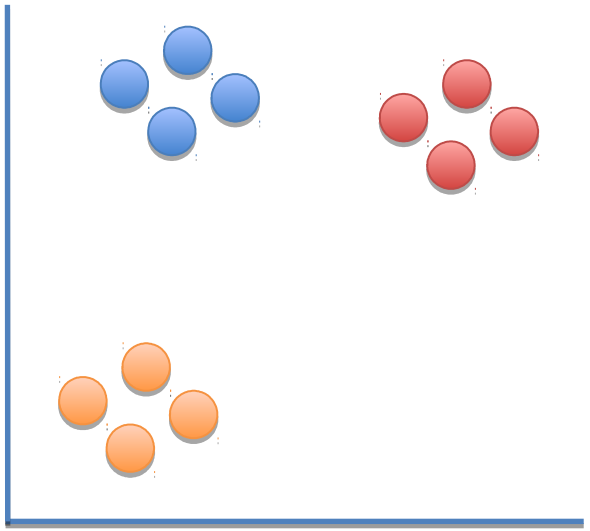


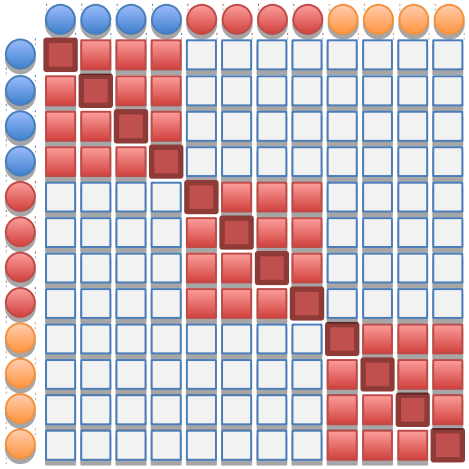




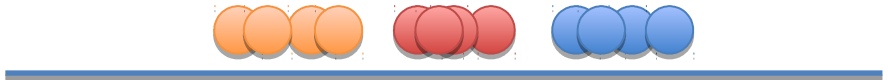
…originally, the “SNE” algorithm just used a normal distribution throughout and the clusters clumped up in the middle and were harder to see.

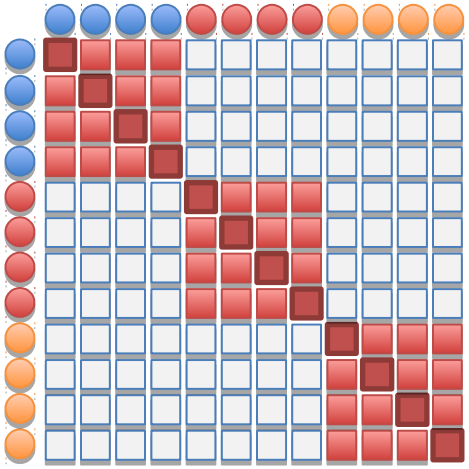


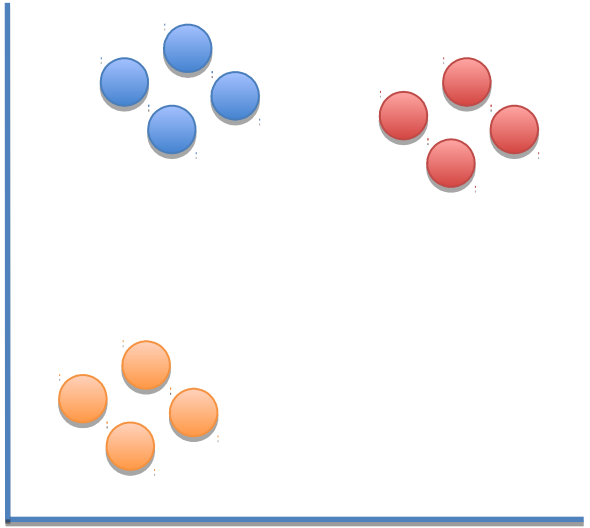


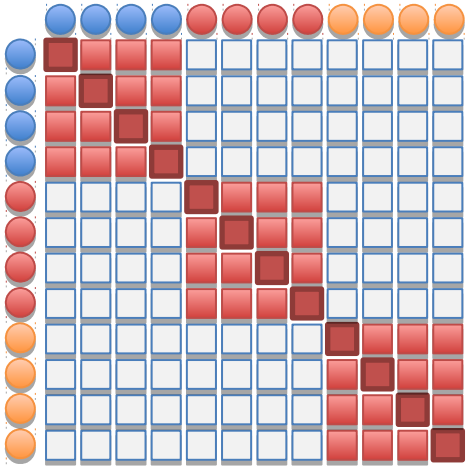


The t-distribution forces some space between the points.

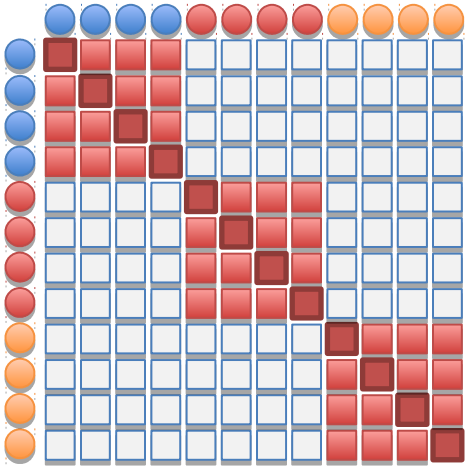


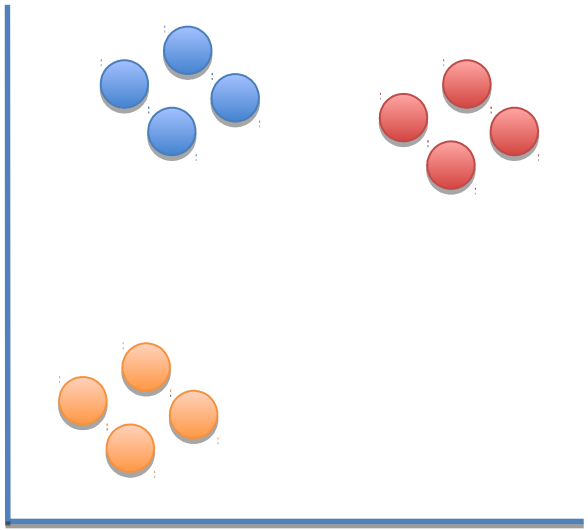


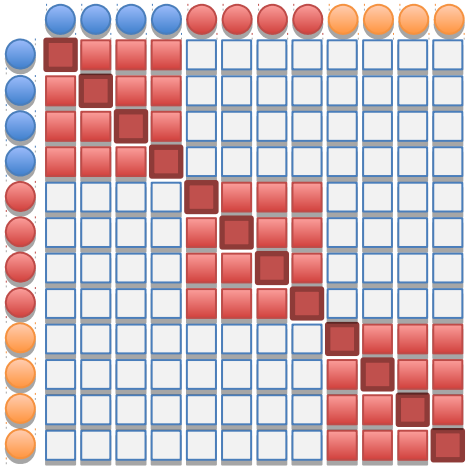




You should look up “**the CURSE of dimensionality**”, to understand the need for this additiontional ‘space’ provided by the T-distribution. It is really useful to know in general
https://en.wikipedia.org/wiki/Curse_of_dimensionality

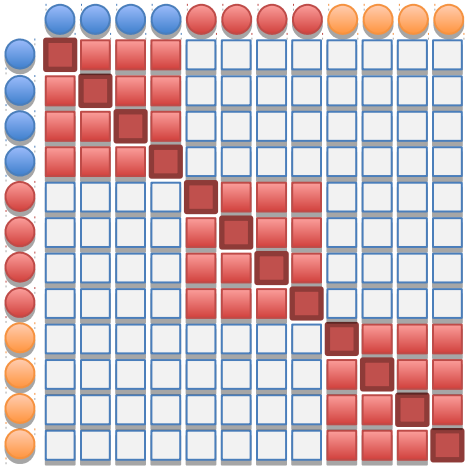


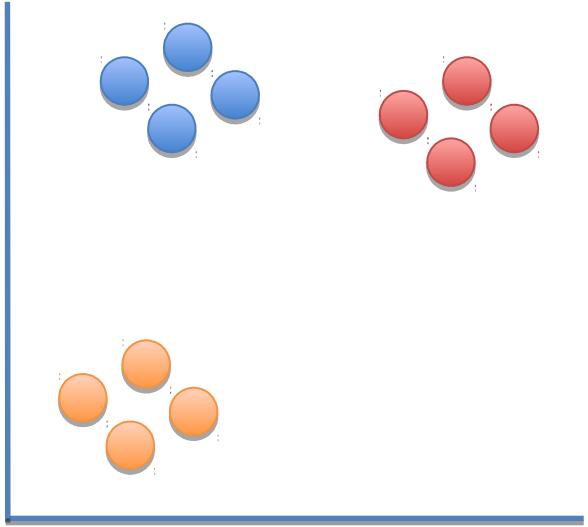


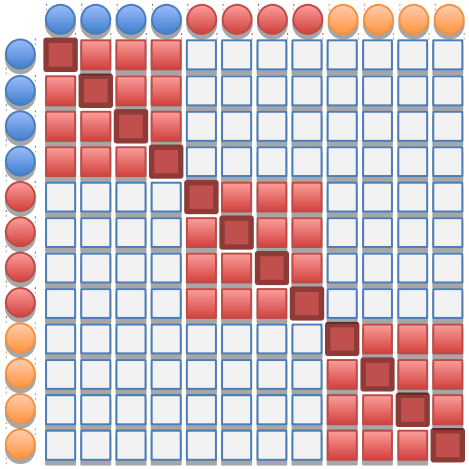


To understand how the two matrices are compared, read about kullback leibler divergence :)

$$D_{\mathrm{KL}}(P \parallel Q) = -\sum_{i} P(i) \log\left(\frac{Q(i)}{P(i)}\right),$$



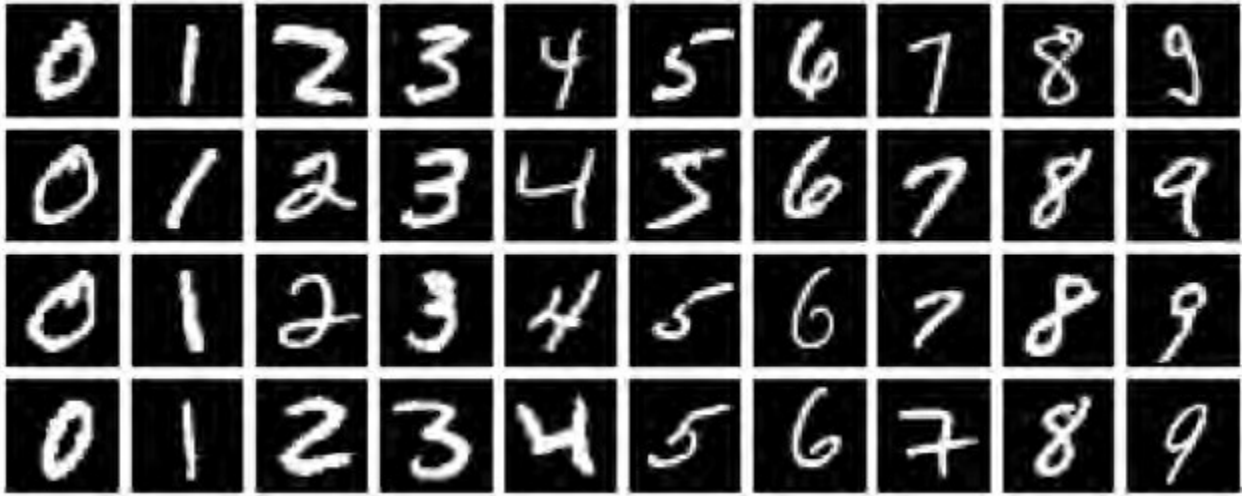




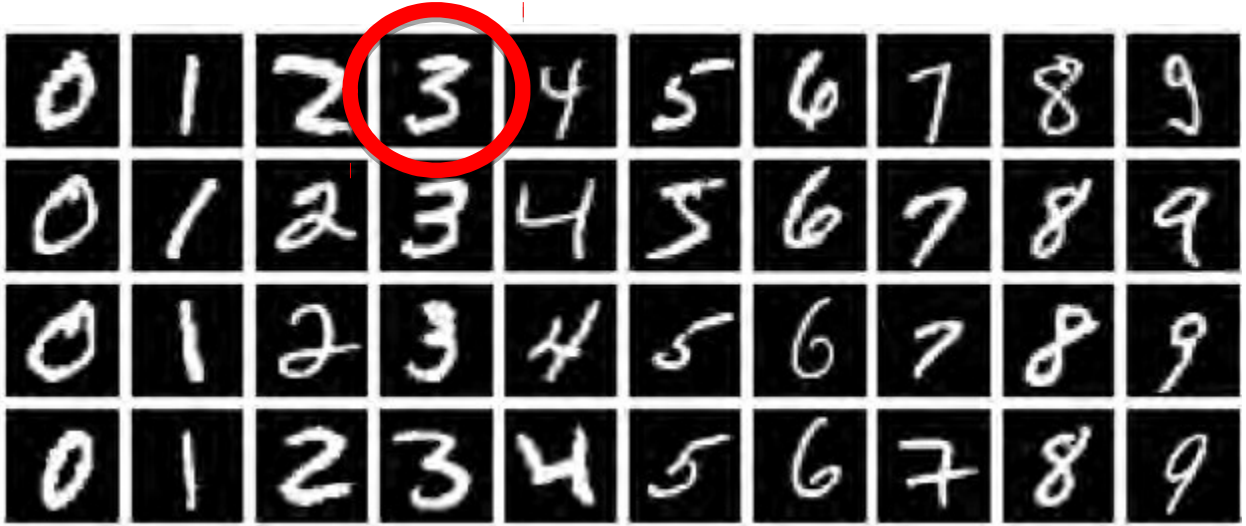
You can also read about the most optimal way to randomly optimize the arrangement of the points, by looking into the original paper on T-SNE

https://lvdmaaten.github.io/tsne/

Now some actual TSNE examples

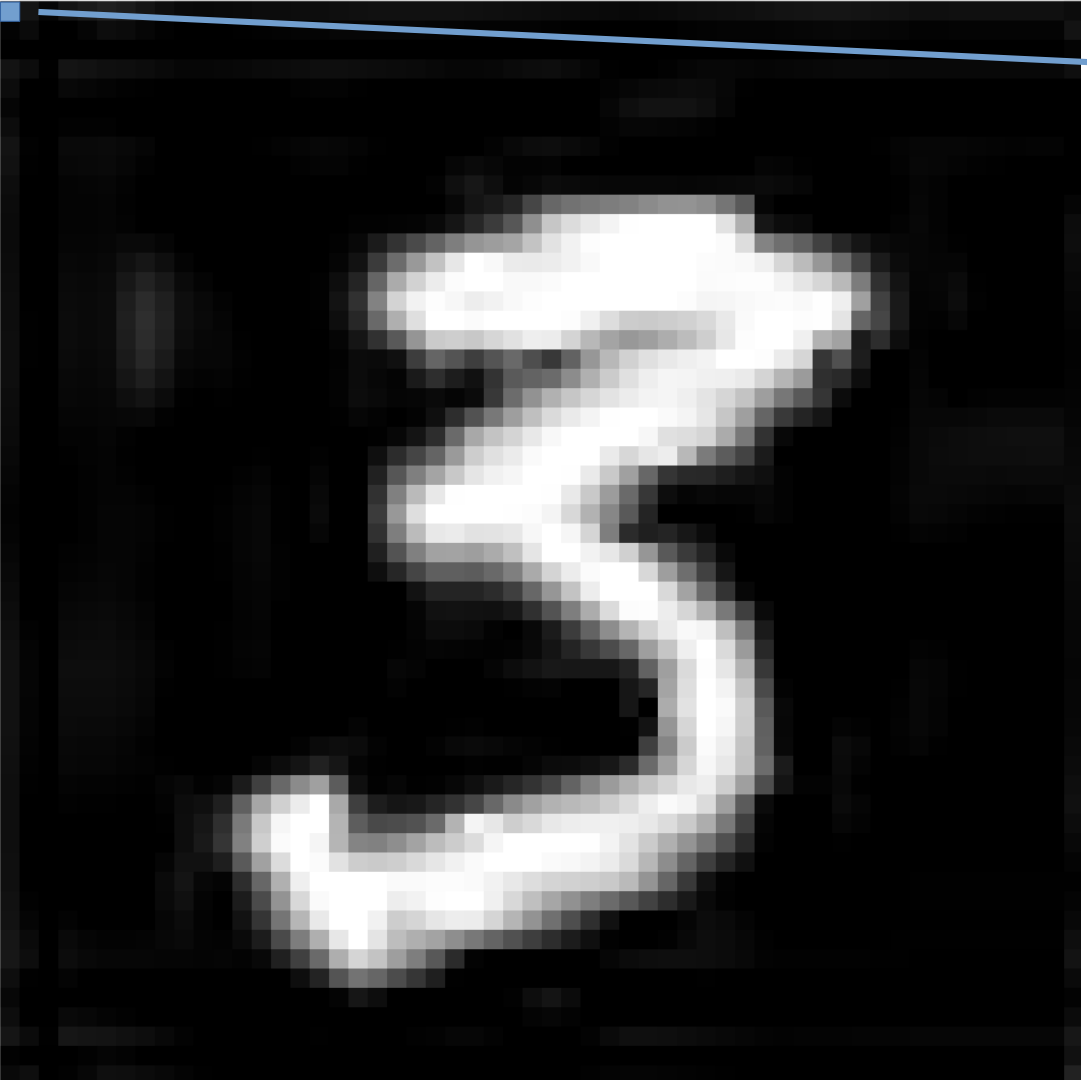


MNIST digits dataset

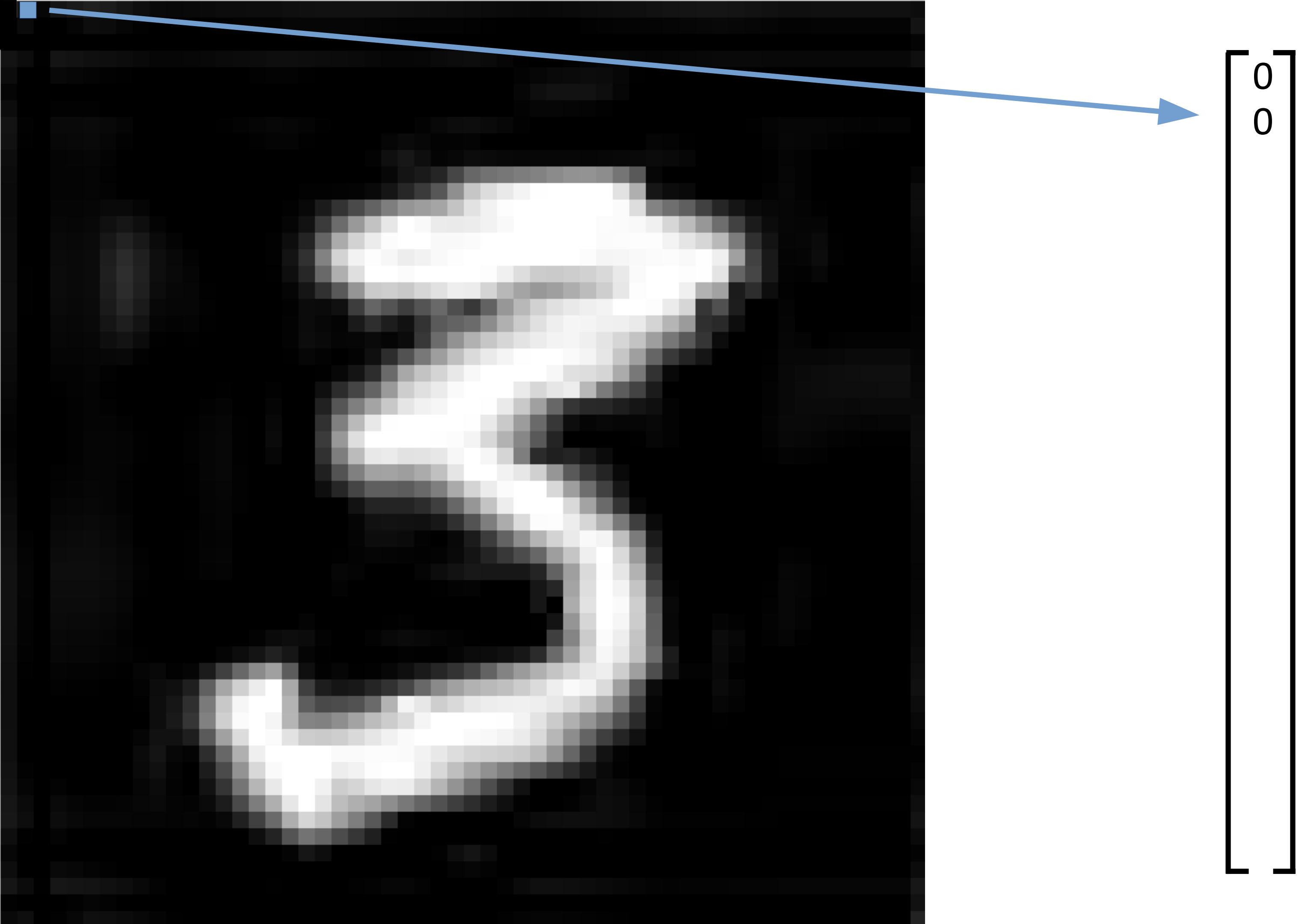


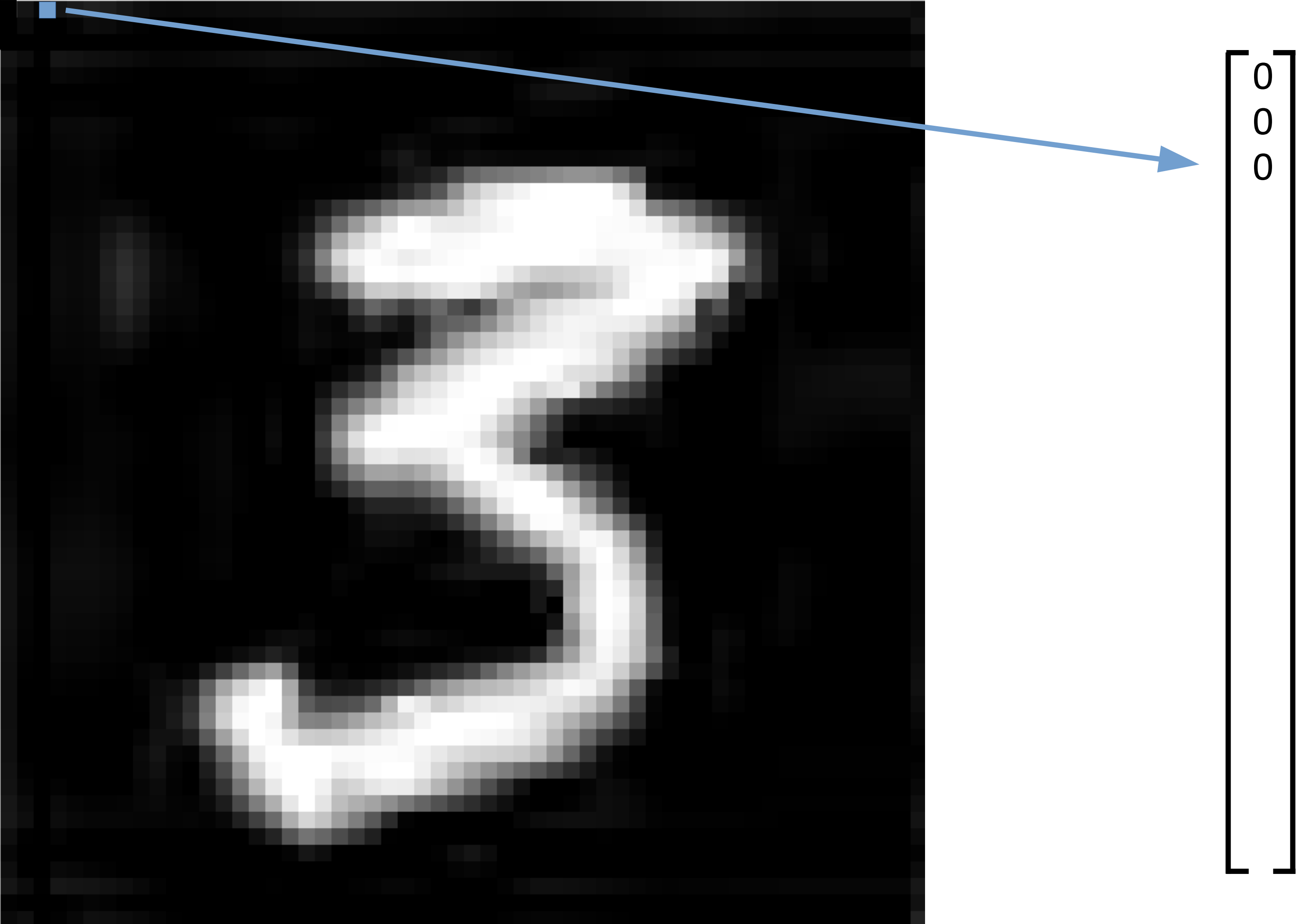
MNIST digits dataset

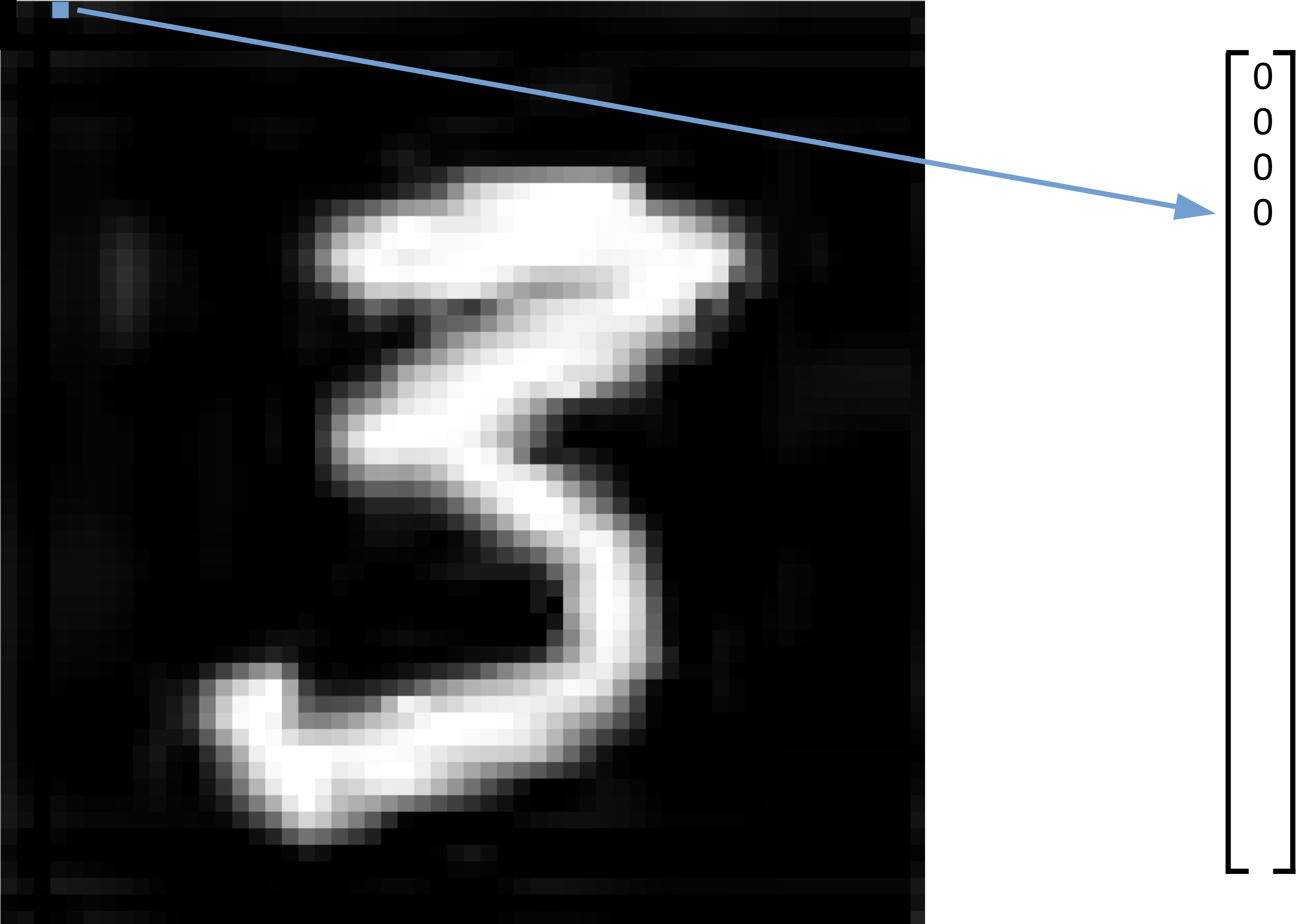
0



0

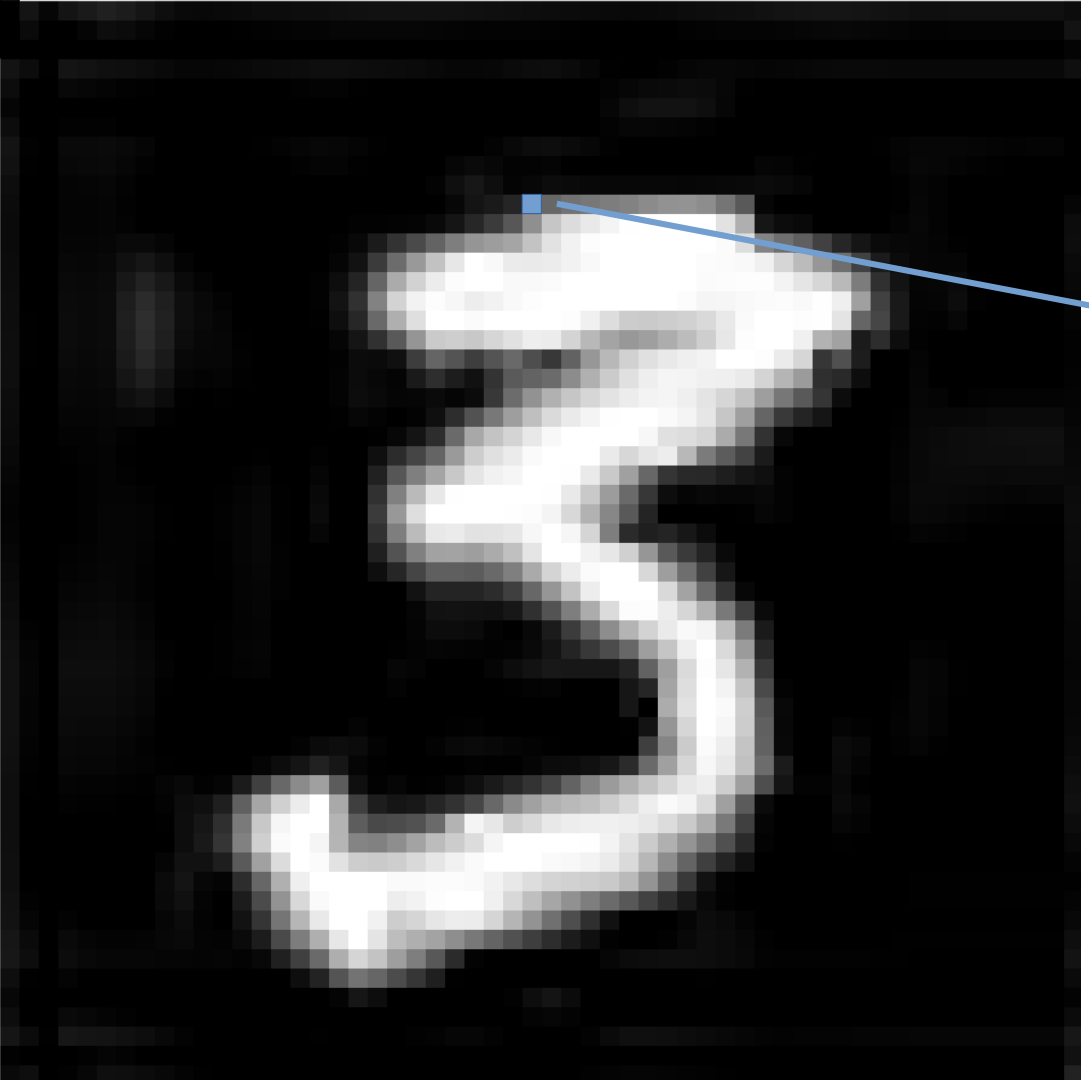

0
0


0
0
0

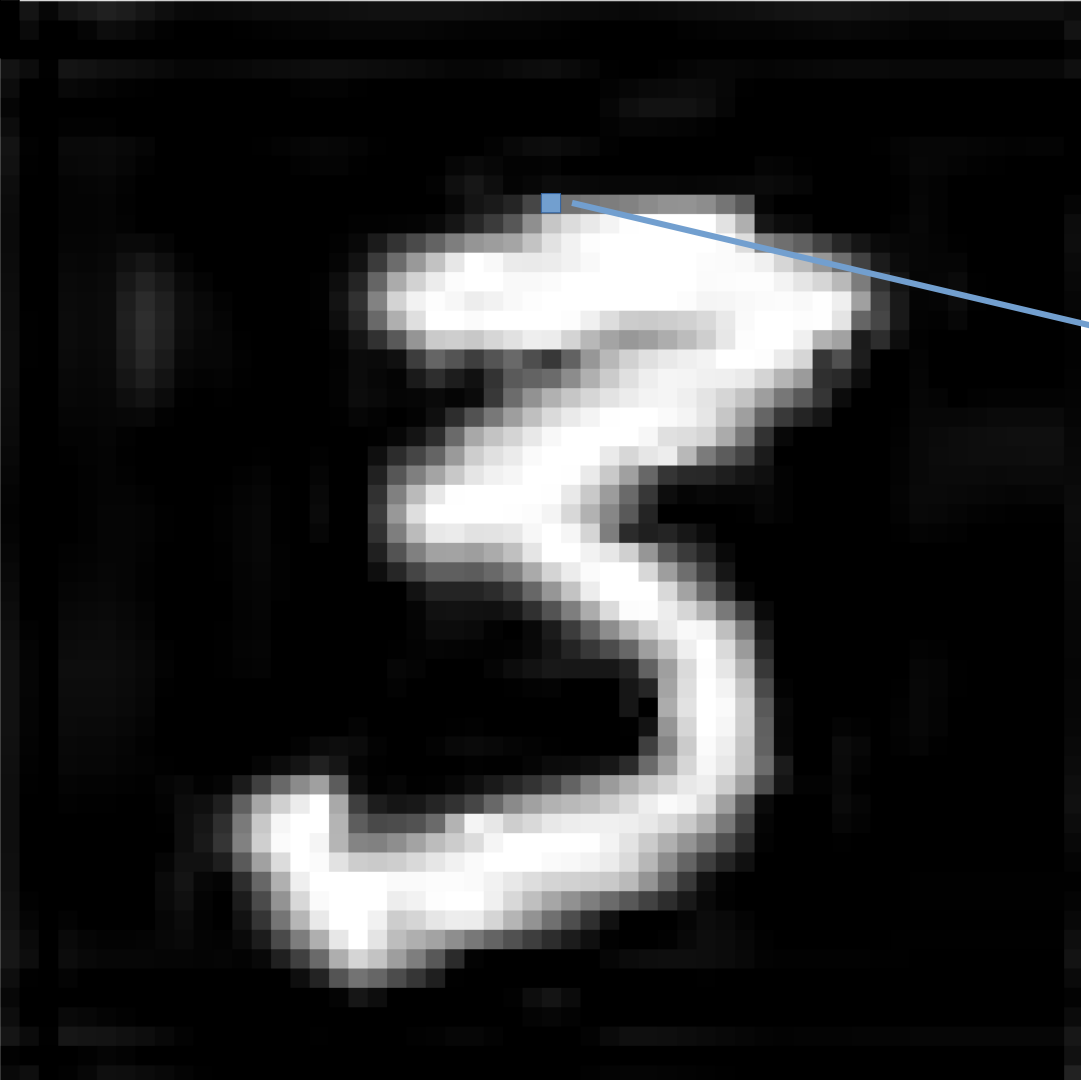

0
0
0
0

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

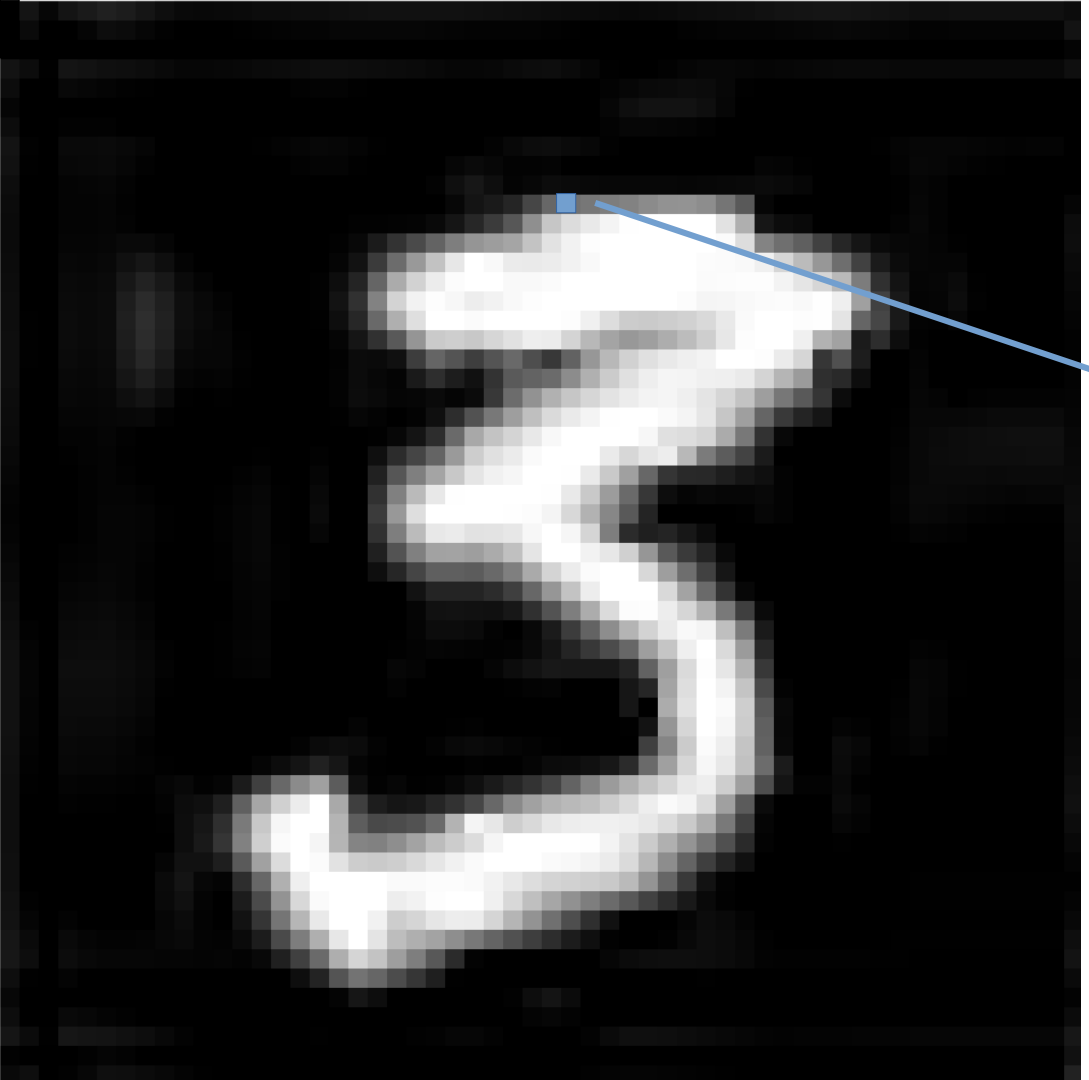
$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$



$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 120 \end{bmatrix}$$



$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 120 \\ 170 \end{bmatrix}$$



$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 120 \\ 170 \\ 190 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 120 \\ 170 \\ 190 \\ 200 \\ 200 \\ 190 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 120 \\ 170 \\ 190 \\ 200 \\ 200 \\ 190 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

0	0
0	0
0	0
0	0
⋮	⋮
120	130
170	160
190	170
200	180
200	180
190	180
⋮	⋮
0	0
0	0
0	0

0	0	0	0	0	0	0	⋯	0	0	0	0	0	0	0
0	0	0	0	0	0	0	⋯	0	0	0	0	0	0	0
0	0	0	0	0	0	0	⋯	0	0	0	0	0	0	0
0	0	0	0	0	0	0	⋯	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮
150	150	150	150	150	150	150	⋯	150	150	150	150	150	150	150
130	130	130	130	130	130	130	⋯	130	130	130	130	130	130	130
120	120	120	120	120	120	120	⋯	120	120	120	120	120	120	120
180	180	180	180	180	180	180	⋯	180	180	180	180	180	180	180
210	210	210	210	210	210	210	⋯	210	210	210	210	210	210	210
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	0	0	0	0	0	0	⋯	0	0	0	0	0	0	0
0	0	0	0	0	0	0	⋯	0	0	0	0	0	0	0
0	0	0	0	0	0	0	⋯	0	0	0	0	0	0	0

Now some actual TSNE examples

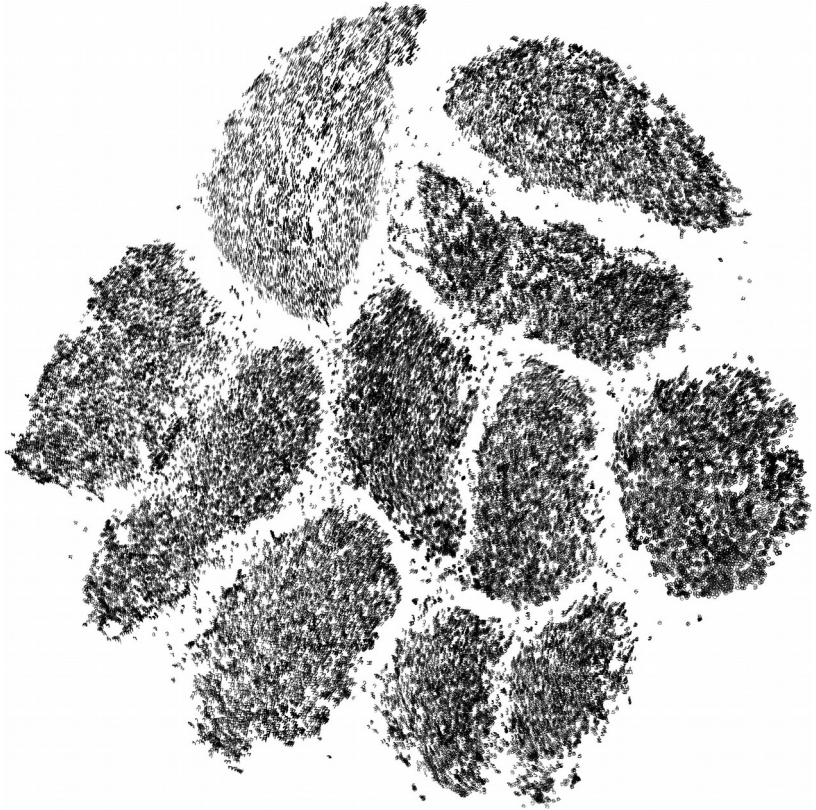
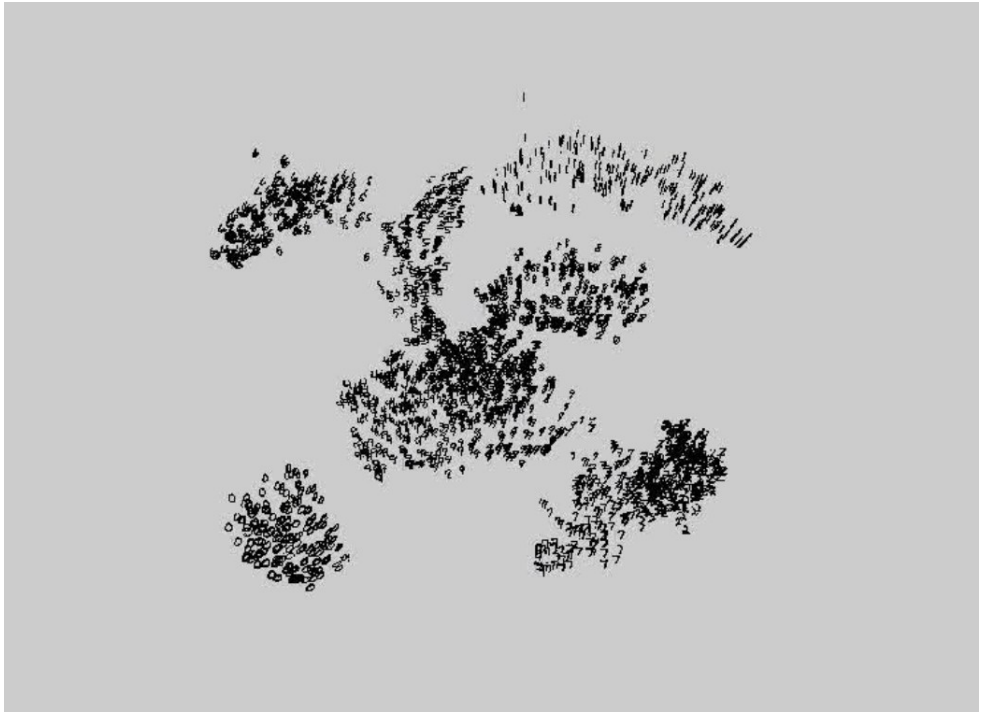


Image: Kyle McDonald

Now some actual TSNE examples



Movie: L.J.P. van der Maaten

Now some actual TSNE examples

Image: Caltech TSNE examples

Notes on parameters

How should I set the perplexity in t-SNE?
The performance of t-SNE is fairly robust under different settings of the perplexity. The most appropriate value depends on the density of your data. Loosely speaking, one could say that a larger / denser dataset requires a larger perplexity. Typical values for the perplexity range between 5 and 50.

- Laurens van der Maaten (Inventor of TSNE) - https://lvdmaaten.github.io/tsne/

perplexity : float, optional (default: 30)
The perplexity is related to the number of nearest neighbors that is used in other manifold learning algorithms. Larger datasets usually require a larger perplexity. Consider selecting a value between 5 and 50. The choice is not extremely critical since t-SNE is quite insensitive to this parameter.

early_exaggeration : float, optional (default: 12.0)
Controls how tight natural clusters in the original space are in the embedded space and how much space will be between them. For larger values, the space between natural clusters will be larger in the embedded space. Again, the choice of this parameter is not very critical. If the cost function increases during initial optimization, the early exaggeration factor or the learning rate might be too high.

learning_rate : float, optional (default: 200.0)
The learning rate for t-SNE is usually in the range [10.0, 1000.0]. If the learning rate is too high, the data may look like a ‘ball’ with any point approximately equidistant from its nearest neighbours. If the learning rate is too low, most points may look compressed in a dense cloud with few outliers. If the cost function gets stuck in a bad local minimum increasing the learning rate may help.

- sklearn.manifold.TSNE - http://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

Read this is you want to get an intuition : https://distill.pub/2016/misread-tsne/

Usefull resources for learning: TSNE

For packages, use Rtsne in R, or sklearn.manifold.TSNE / Multicore-TSNE in python

https://distill.pub/2016/misread-tsne/
https://lvdmaaten.github.io/tsne/
http://jotterbach.github.io/2016/05/23/TSNE/ <-- Curse of dim
http://colah.github.io/posts/2014-10-Visualizing-MNIST/
https://www.oreilly.com/learning/an-illustrated-introduction-to-the-t-sne-algorithm

https://github.com/oreillymedia/t-SNE-tutorial
https://www.youtube.com/watch?v=NEaUSP4YerM&t=618s

Thanks to Josh Starmer of (awesome) StatQuest for the powerpoint templete