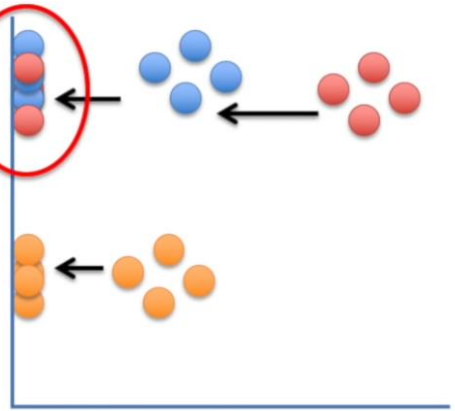
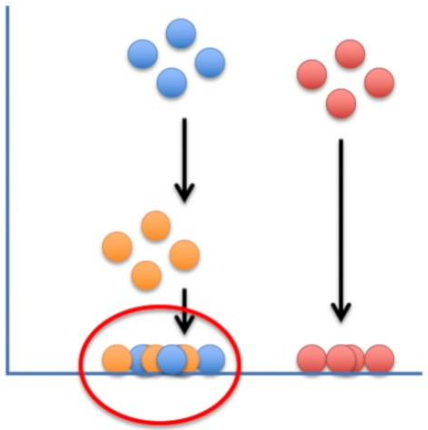
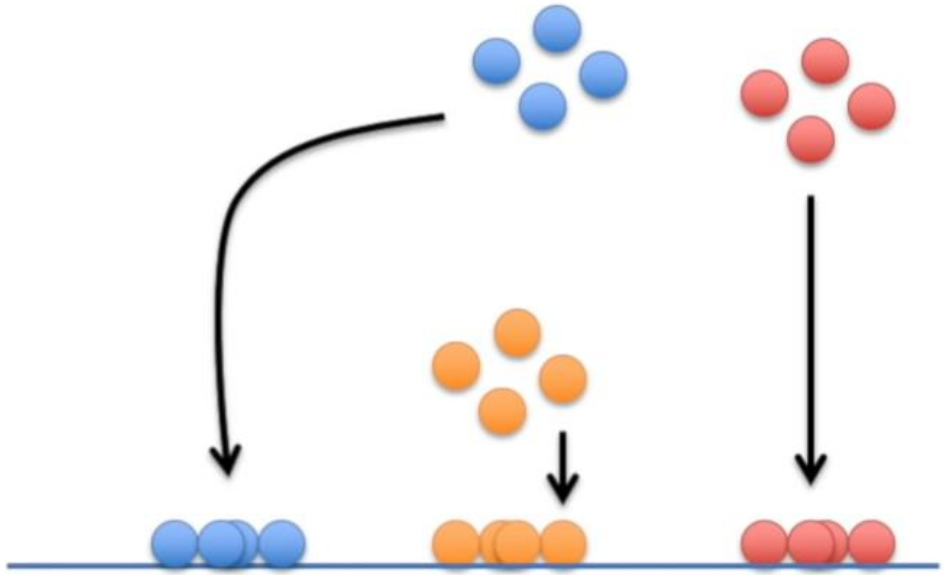
t-distributed Stochastic Neighbor Embedding (t-SNE) Explained

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





Same here...

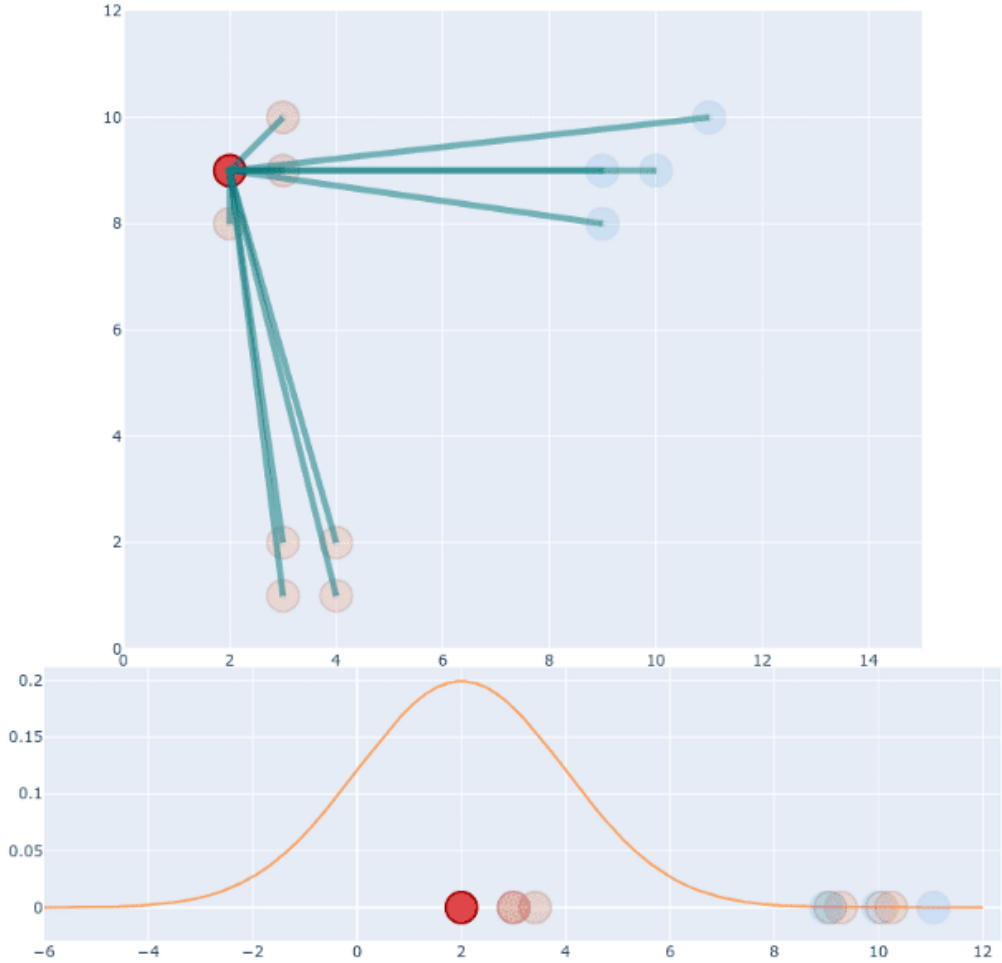


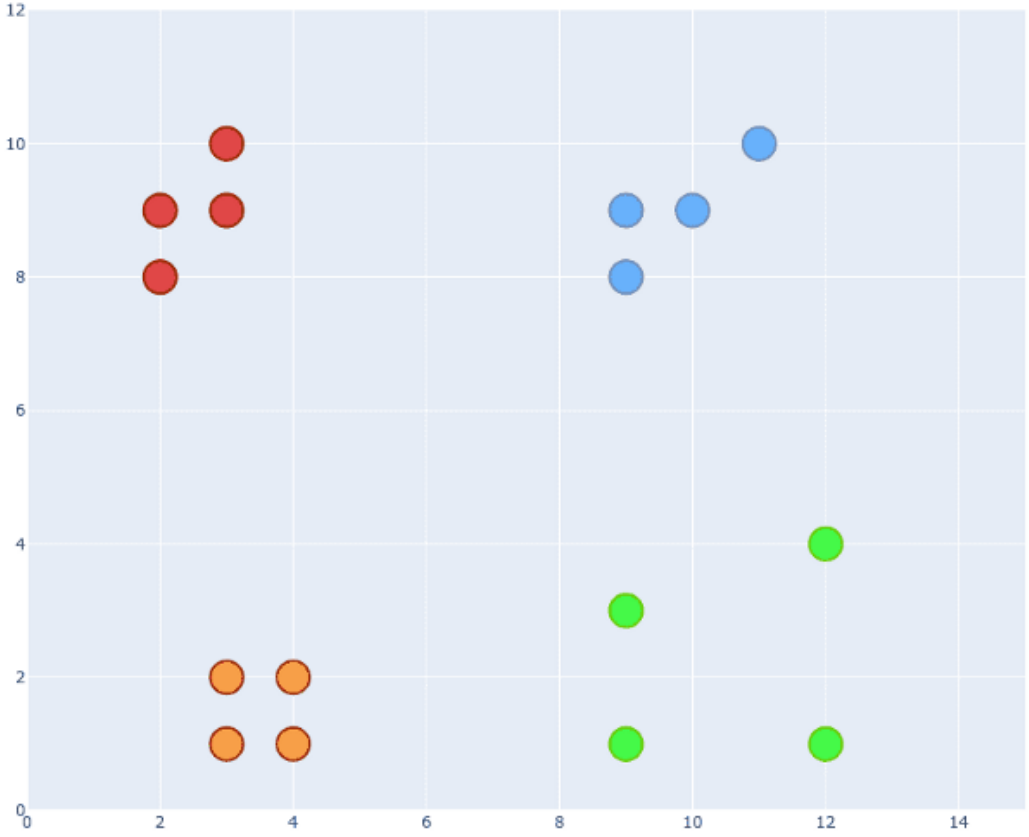
What t-SNE does is find a way to project data into a low dimensional space (in this case, the 1-D number line) so that the clustering in the high dimensional space (in this case, the 2-D scatter plot) is preserved.

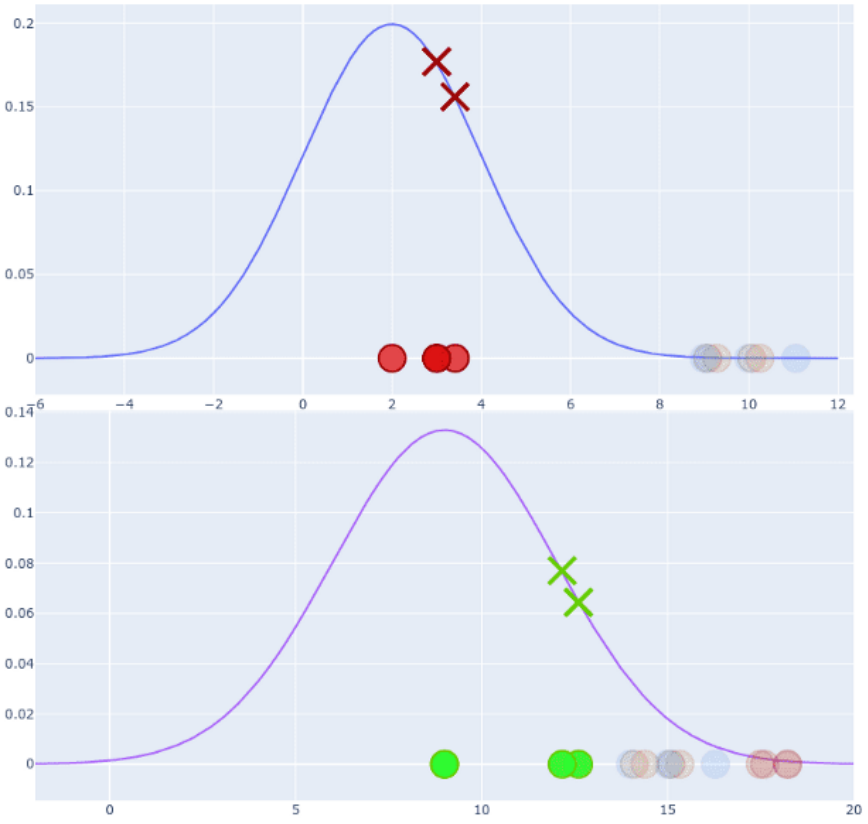
We want to create a probability distribution that represents similarities between neighbors.

Similarity of datapoint $\mathbf{x_j}$ to datapoint $\mathbf{x_i}$ is the **conditional probability** $p(j|i)$, that $\mathbf{x_i}$ would pick $\mathbf{x_j}$ as its neighbor.

Let's use Euclidean distance between them $|\mathbf{x_i}-\mathbf{x_j}|$ as a measure.







Notice that in the green example absolute values of probability are much smaller than in the red example.

We can fix that by dividing the current projection value by the sum of the projections. This scales all values to have a sum equal to 1:

$$p(j|i) = \frac{g(|x_i - x_j|)}{\sum_{k \neq i} g(|x_i - x_k|)}$$

Red example

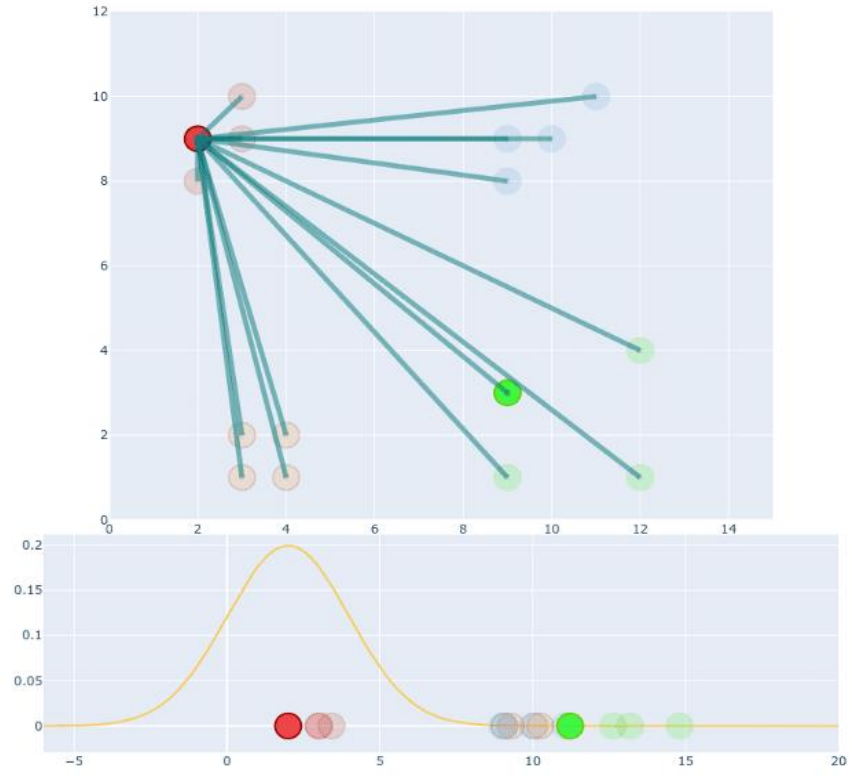
$$p(j|i) = 0.34$$

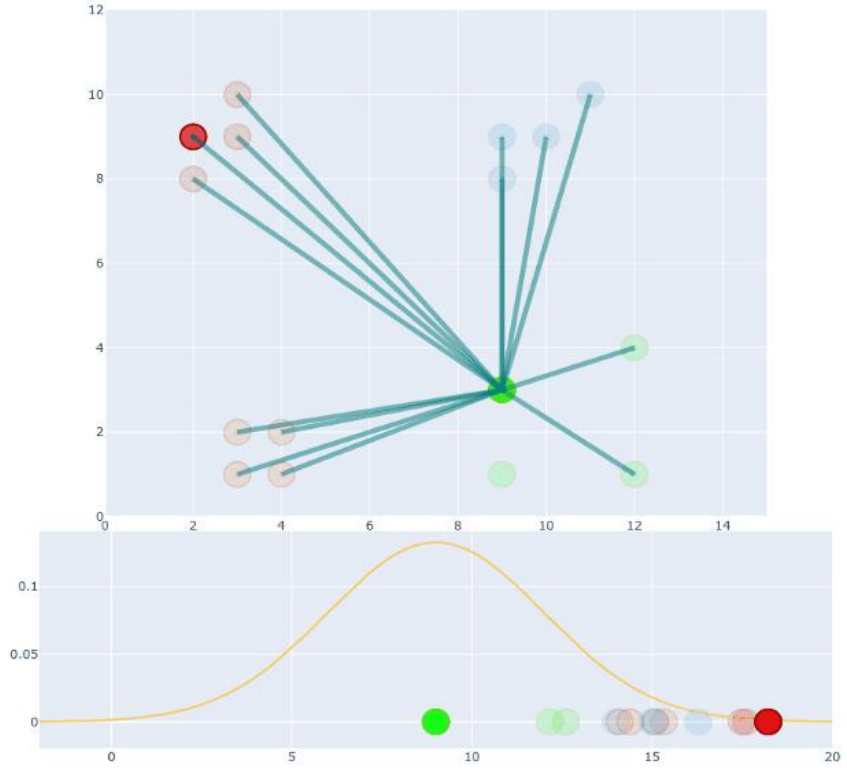
Green example

$$p(j|i) = 0.27$$

Note that $p(i|i) = 0$

If we take two points and calculate conditional probability between them then values of p(i|j) and p(j|i) will be different. Because they are coming from two different distributions. Which one to pick for the calculation then?





The reality is that t-SNE utilizes the following equation to calculate p(j|i):

$$p(j|i) = \frac{e^{\frac{-\|x_i - x_j\|^2}{2\sigma_i^2}}}{\sum_{k \neq i} e^{\frac{-\|x_i - x_k\|^2}{2\sigma_i^2}}}$$

Variance depends on Gaussian and the number of points surrounding the center of it. This is the part where perplexity value comes to play.

Think of perplexity as a target number of neighbors for our central point. The higher the perplexity is the higher value variance has, e.g., our red group is close to each other and if we set perplexity to 4, it searches the right value of to “fit” the neighbors.

t-SNE performs a binary search for the value of σ that produces probability distribution with a fixed perplexity that is specified by the user:

$$Perp(P_i) = 2^{-\sum p(j|i) \times log_2(p(j|i))}$$

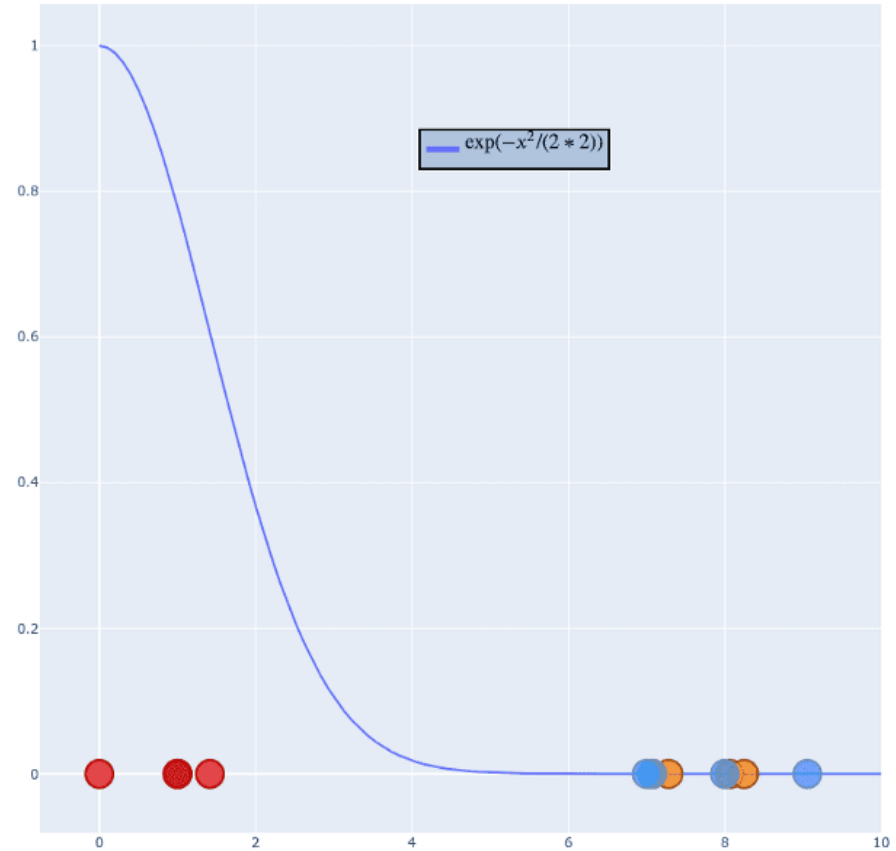
Shannon entropy

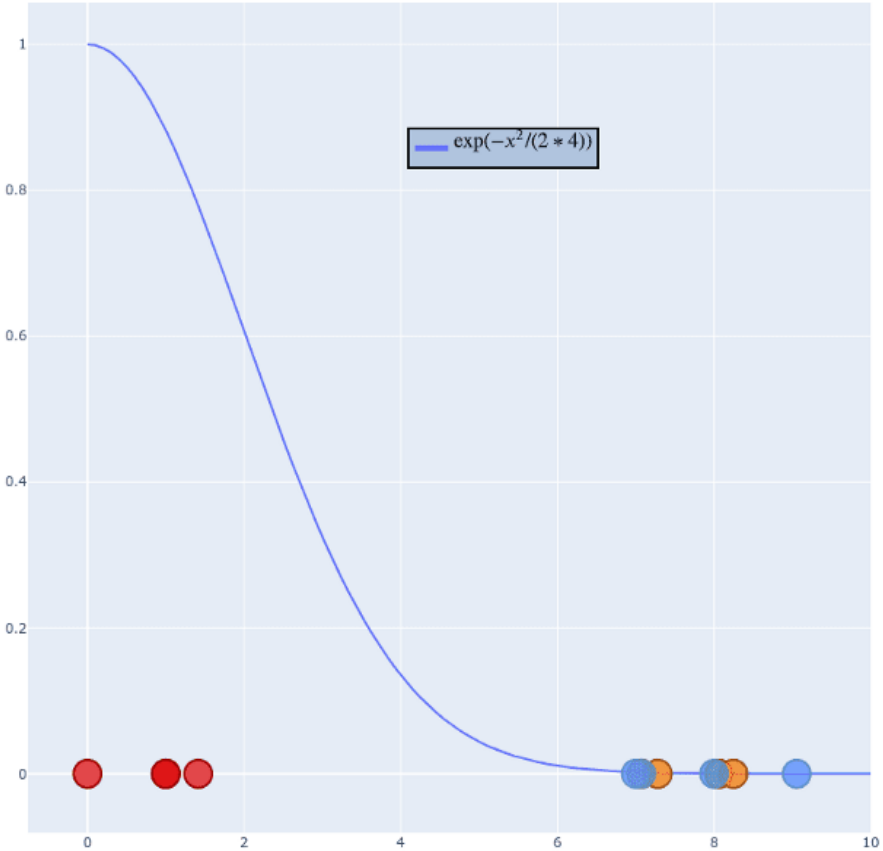
To know:

Perplexity value you choose is positively correlated with the value of μ_i and for the same perplexity you will have multiple different μ_i, based on distances. Typical perplexity value ranges between 5 and 50.

$$e^{\frac{-\|x_i - x_j\|^2}{2\sigma_i^2}}$$

If you play with σ^2 you notice that the blue curve remains fixed at x=0 it only stretches when σ^2 increases.

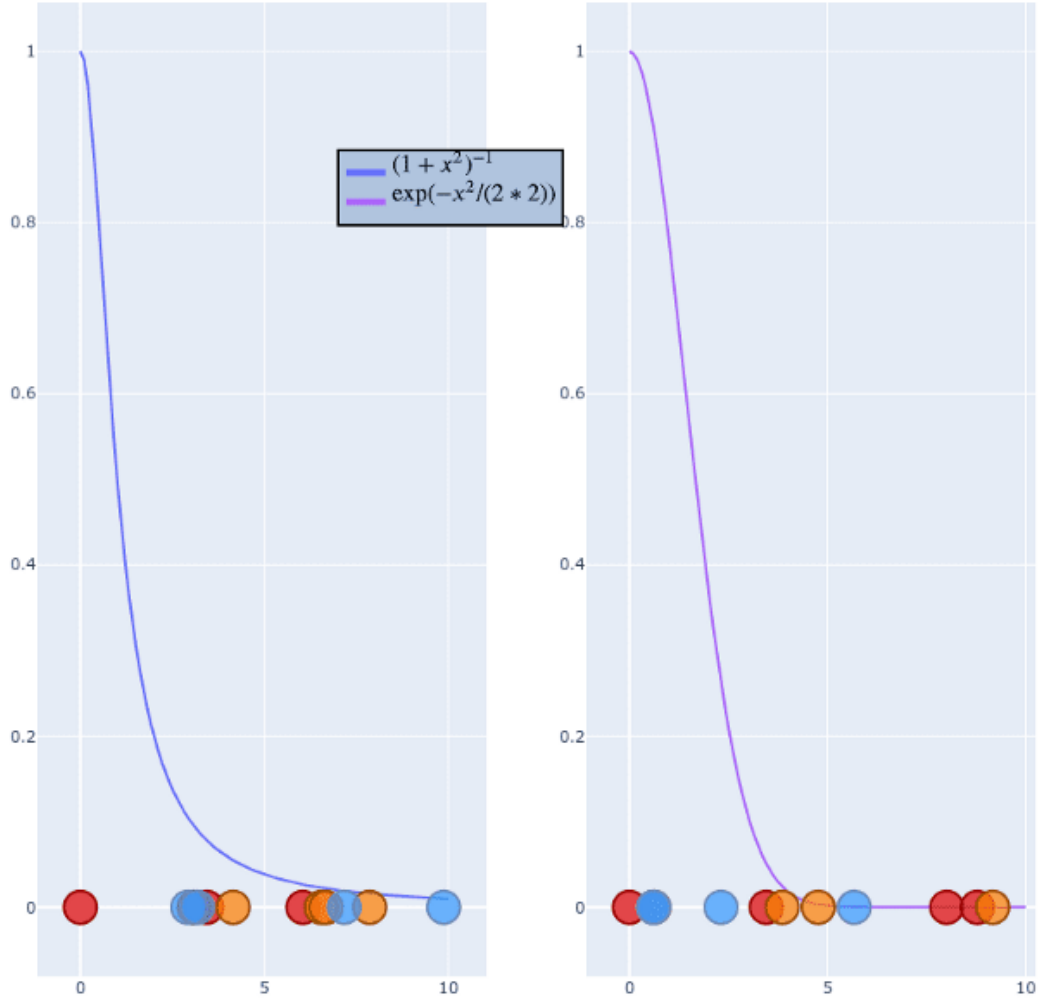




Dimension reduction:

The goal is to find similar probability distribution in low-dimensional space.

Gaussian has a “short tail” and because of that it creates a crowding problem. t-SNE uses **t-distribution**.



So, our new formula looks like:

$$q(i|j) = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

And, we had:

$$p(i|j) = \frac{e^{\frac{-\|x_i - x_j\|^2}{2\sigma_i^2}}}{\sum_{k \neq l} e^{\frac{-\|x_k - x_l\|^2}{2\sigma_i^2}}}$$

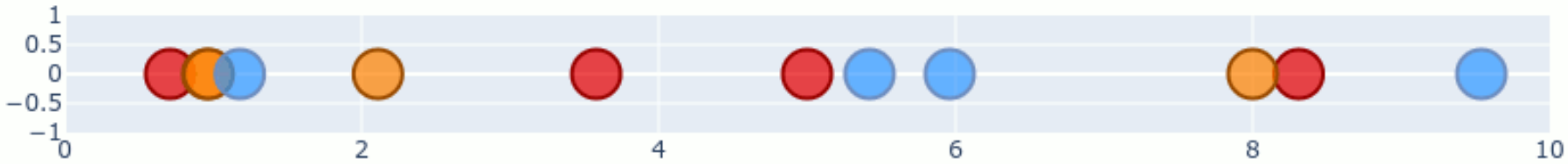
Cost function: t-SNE uses Kullback-Leibler divergence between the conditional probabilities *p* and *q:*

$$C = D_{KL}(P\|Q) = \sum_{x \in X} P(x) \times log(\frac{P(x)}{Q(x)})$$

And, it is minimized using a gradient descent method:

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p(i|j) - q(i|j))(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$$

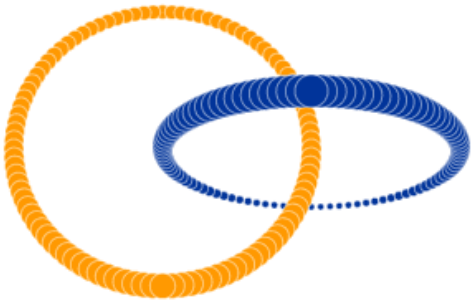
Think of the gradient as repulsion/attraction between points. A gradient is calculated for each point and describes how strong and in what direction it should be pulled.

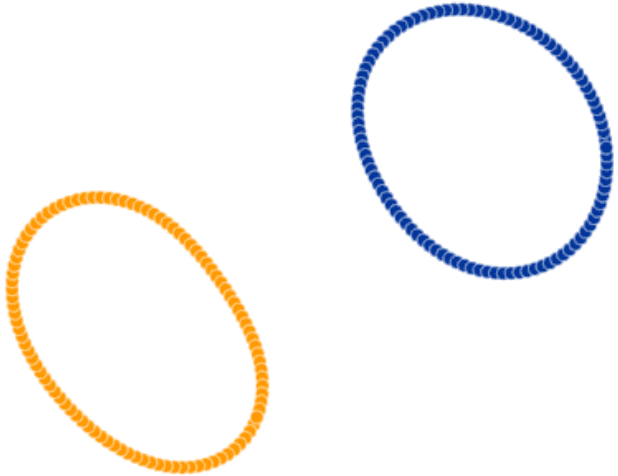


Early Compression – To prevent early clustering it adds L2 penalty to the cost function at early stages.

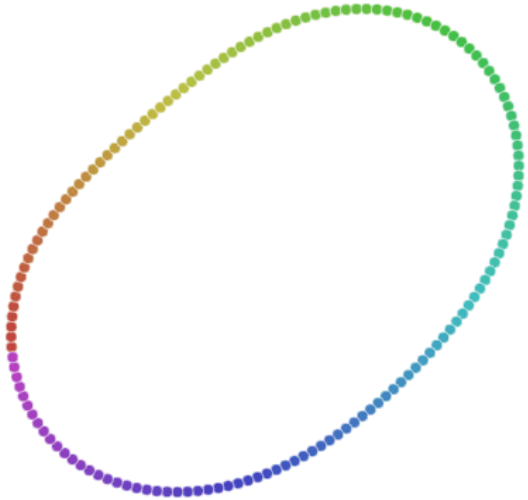
Early Exaggeration – To prevent clusters from getting in each other's ways it moves clusters of q(i|j) more. This time we're multiplying p(i|j) in early stages.

Examples





Perplexity 5



Now, go and play with this nice interactive webpage:

https://distill.pub/2016/misread-tsne/