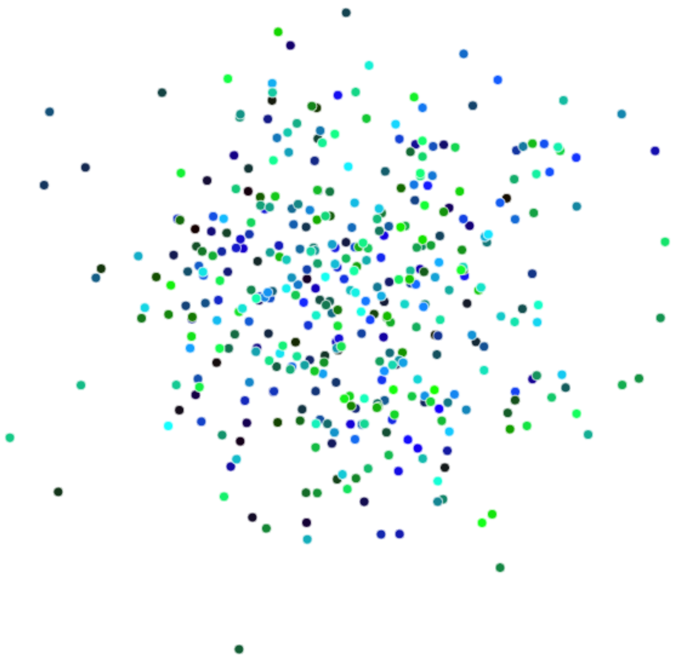
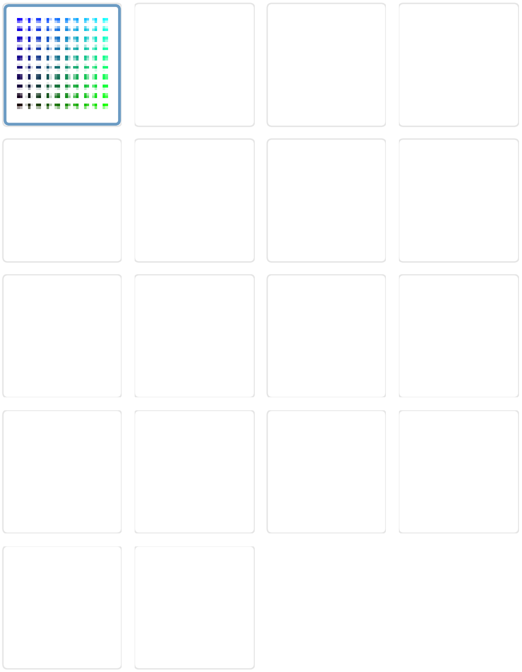
How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.





A square grid with equal spacing between points. Try convergence at different sizes.

Step
1

▶ ↻

Points Per Side 20

Perplexity 10

↻ Step:

MARTIN WATTENBER
FERNANDA VIÉGAS (
IAN JOHNSON Googl
Oct. 13 2016

Citation: Wattenberg,

A popular meth ... ten and Hinton in 2008 [1]. The ... gical ability to create compelli ... gh impressive, these images ca

We'll walk throu ... SNE technique really is useful—

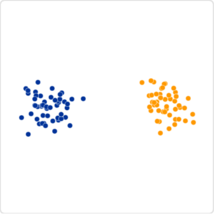
Before diving in: if you haven't encountered t-SNE before, here's what you need to know about the math behind it. The goal is to take a set of points in a high-dimensional space and find a faithful representation of those points in a lower-dimensional space, typically the 2D plane. The algorithm is non-linear and adapts to the underlying data, performing different transformations on different regions. Those differences can be a major source of confusion.

A second feature of t-SNE is a tuneable parameter, "perplexity," which says (loosely) how to balance attention between local and global aspects of your data. The parameter is, in a sense, a guess about the number of close neighbors each point has. The perplexity value has a complex effect on the resulting pictures. The original paper says, *"The performance of SNE is fairly robust to changes in the perplexity, and typical values are between 5 and 50."* But the story is more nuanced than that. Getting the most from t-SNE may mean analyzing multiple plots with different perplexities.

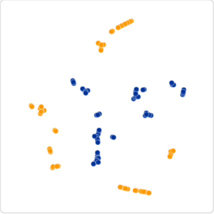
That's not the end of the complications. The t-SNE algorithm doesn't always produce similar output on successive runs, for example, and there are additional hyperparameters related to the optimization process.

1. Those hyperparameters really matter

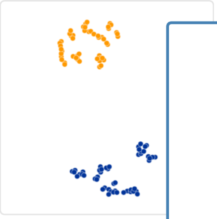
Let's start with the "hello world" of t-SNE: a data set of two widely separated clusters. To make things as simple as possible, we'll consider clusters in a 2D plane, as shown in the lefthand diagram. (For clarity, the two clusters are color coded.) The diagrams at right show t-SNE plots for five different perplexity values.



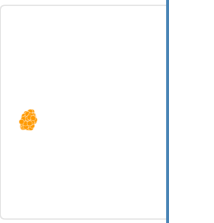
Original

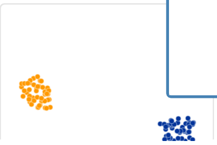


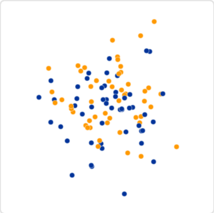
Perplexity: 2
Step: 5,000



Step:



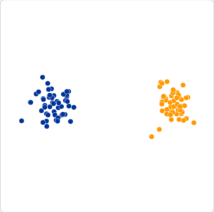




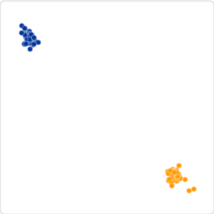
Perplexity: 100
Step: 5,000

With perplexity values in the range (5 - 50) suggested by van der Maaten & Hinton, the diagrams do show these clusters, although with very different shapes. Outside that range, things get a little weird. With perplexity 2, local variations dominate. The image for perplexity 100, with merged clusters, illustrates a pitfall: for the algorithm to operate properly, the perplexity really should be smaller than the number of points. Implementations can give unexpected behavior otherwise.

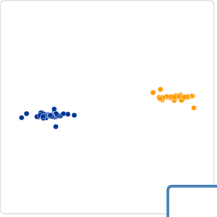
Each of the plots above was made with 5,000 iterations with a learning rate (often called "epsilon") of 10, and had reached a point of stability by step 5,000. How much of a difference do those values make? In our experience, the most important thing is to iterate until reaching a stable configuration.



Original

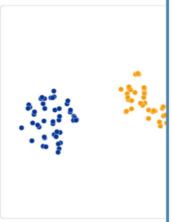


Perplexity: 30
Step: 10

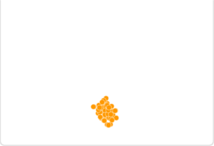


Perplexity: 30
Step: 20

Step:



Perplexity: 30
Step: 1,000



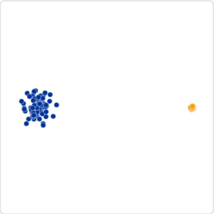
The images above show five different runs at perplexity 30. The first four were stopped before stability. After 10, 20, 60, and 120 steps you can see layouts with seeming 1-dimensional and even pointlike images of the clusters. If you see a t-SNE plot with strange "pinched" shapes, chances are the process was stopped too early. Unfortunately, there's no fixed number of steps that yields a stable result. Different data sets can require different numbers of iterations to converge.

Another natural question is whether different runs with the same hyperparameters produce the same results. In this simple two-cluster example, and most of the others we discuss, multiple runs give the same global shape. Certain data sets, however, yield markedly different diagrams on different runs; we'll give an example of one of these later.

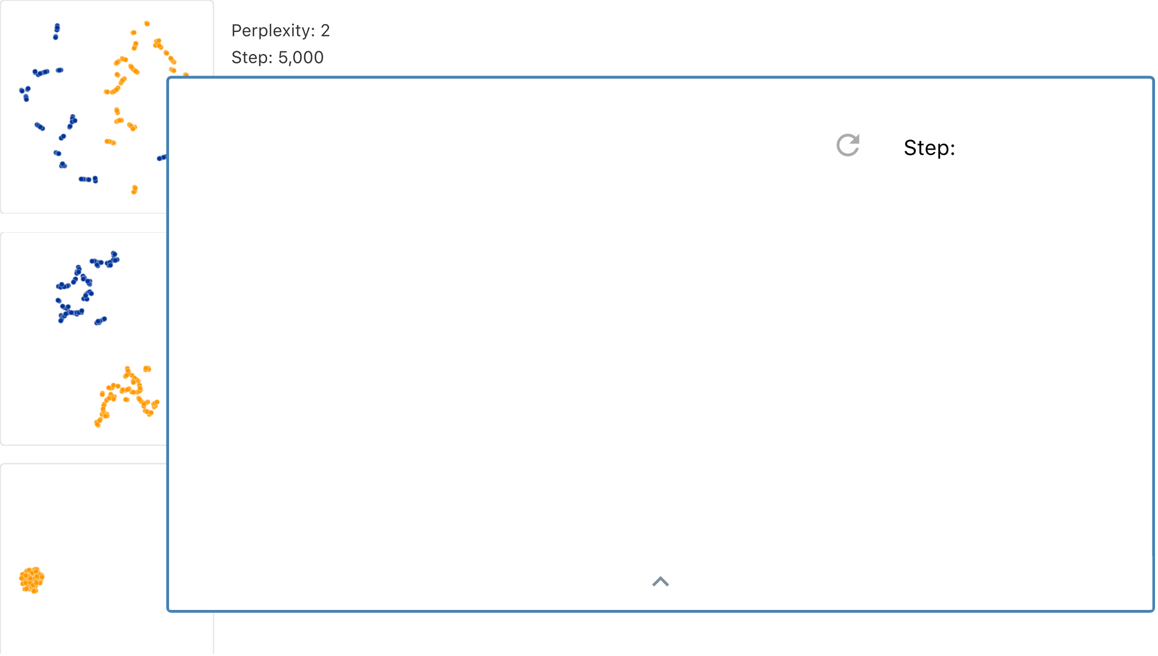
From now on, unless otherwise stated, we'll show results from 5,000 iterations. That's generally enough for convergence in the (relatively small) examples in this essay. We'll keep showing a range of perplexities, however, since that seems to make a big difference in every case.

2. Cluster sizes in a t-SNE plot mean nothing

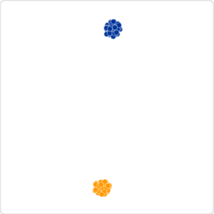
So far, so good. But what if the two clusters have different standard deviations, and so different sizes? (By size we mean bounding box measurements, not number of points.) Below are t-SNE plots for a mixture of Gaussians in plane, where one is 10 times as dispersed as the other.



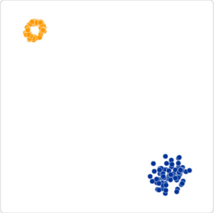
Original



Perplexity: 2
Step: 5,000



Perplexity: 50
Step: 5,000



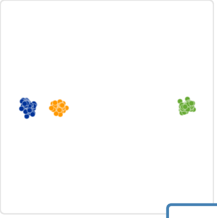
Perplexity: 100
Step: 5,000

Surprisingly, the two clusters look about same size in the t-SNE plots. What’s going on? The t-SNE algorithm adapts its notion of “distance” to regional density variations in the data set. As a result, it naturally expands dense clusters, and contracts sparse ones, evening out cluster sizes. To be clear, this is a different effect than the run-of-the-mill fact that any dimensionality reduction technique will distort distances. (After all, in this example all data was two-dimensional to begin with.) Rather, density equalization happens by design and is a predictable feature of t-SNE.

The bottom line, however, is that you cannot see relative sizes of clusters in a t-SNE plot.

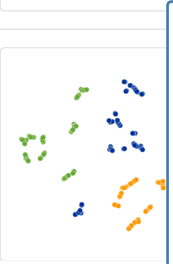
3. Distances between clusters might not mean anything

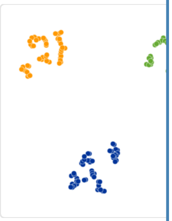
What about distances *between* clusters? The next diagrams show three Gaussians of 50 points each, one pair being 5 times as far apart as another pair.

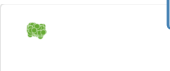


Original

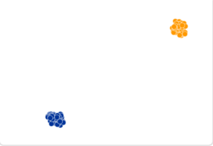
Step:

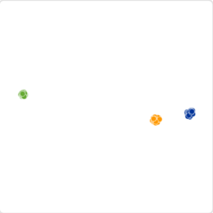




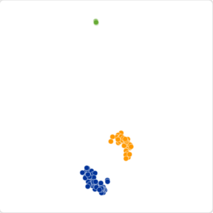


Perplexity: 30
Step: 5,000





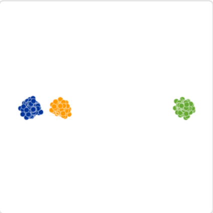
Perplexity: 50
Step: 5,000



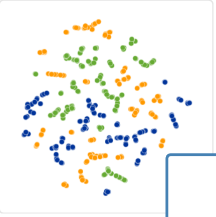
Perplexity: 100
Step: 5,000

At perplexity 50, the diagram gives a good sense of the global geometry. For lower perplexity values the clusters look equidistant. When the perplexity is 100, we see the global geometry fine, but one of the cluster appears, falsely, much smaller than the others. Since perplexity 50 gave us a good picture in this example, can we always set perplexity to 50 if we want to see global geometry?

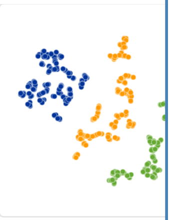
Sadly, no. If we add more points to each cluster, the perplexity has to increase to compensate. Here are the t-SNE diagrams for three Gaussian clusters with 200 points each, instead of 50. Now none of the trial perplexity values gives a good result.

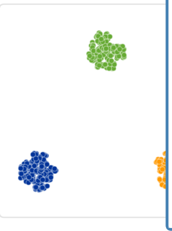


Original



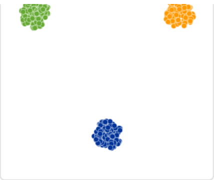
Perplexity: 2
Step: 5,000



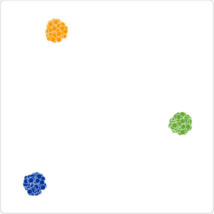


Step:

Perplexity: 50



Step: 5,000



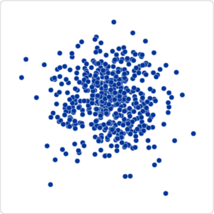
Perplexity: 100
Step: 5,000

It's bad news that seeing global geometry requires fine-tuning perplexity. Real-world data would probably have multiple clusters with different numbers of elements. There may not be one perplexity value that will capture distances across all clusters—and sadly perplexity is a global parameter. Fixing this problem might be an interesting area for future research.

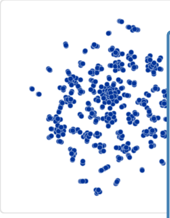
The basic message is that distances between well-separated clusters in a t-SNE plot may mean nothing.

4. Random noise doesn't always look random.

A classic pitfall is thinking you see patterns in what is really just random data. Recognizing noise when you see it is a critical skill, but it takes time to build up the right intuitions. A tricky thing about t-SNE is that it throws a lot of existing intuition out the window. The next diagrams show genuinely random data, 500 points drawn from a unit Gaussian distribution in 100 dimensions. The left image is a projection onto the first two coordinates.

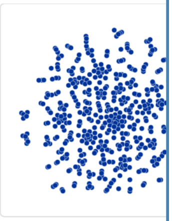


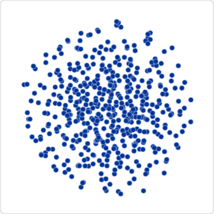
Original



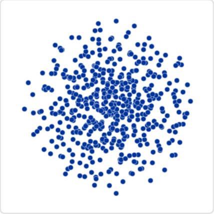
Perplexity: 2

Step:





Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

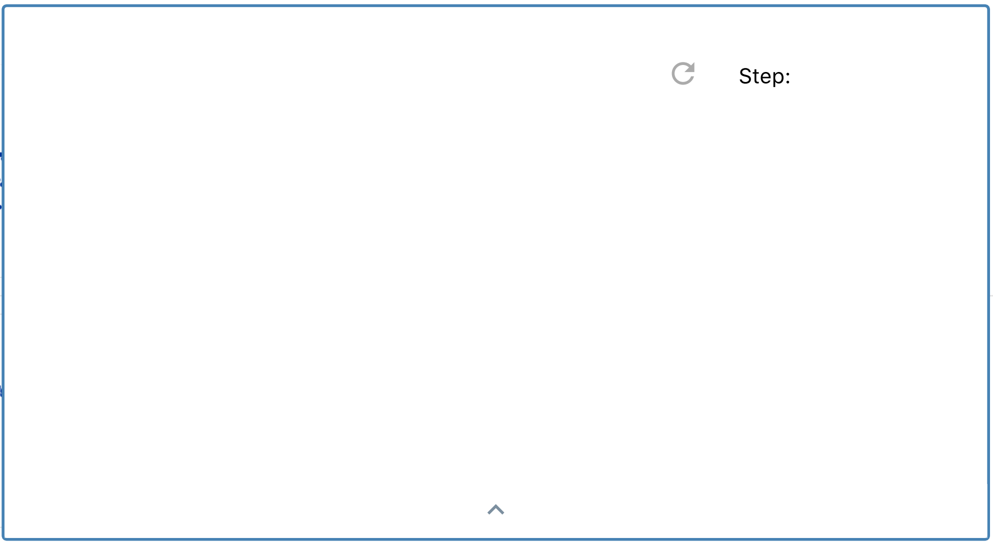
The plot with perplexity 2 seems to show dramatic clusters. If you were tuning perplexity to bring out structure in the data, you might think you'd hit the jackpot.

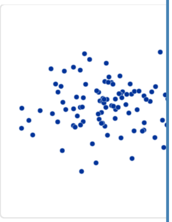
Of course, since we know the cloud of points was generated randomly, it has no statistically interesting clusters: those "clumps" aren't meaningful. If you look back at previous examples, low perplexity values often lead to this kind of distribution. Recognizing these clumps as random noise is an important part of reading t-SNE plots.

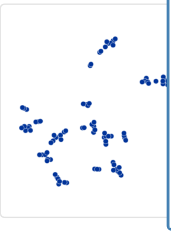
There's something else interesting, though, which may be a win for t-SNE. At first the perplexity 30 plot doesn't look like a Gaussian distribution at all: there's only a slight density difference across different regions of the cloud, and the points seem suspiciously evenly distributed. In fact, these features are saying useful things about high-dimensional normal distributions, which are very close to uniform distributions on a sphere: evenly distributed, with roughly equal spaces between points. Seen in this light, the t-SNE plot is more accurate than any linear projection could be.

5. You can see some shapes, sometimes

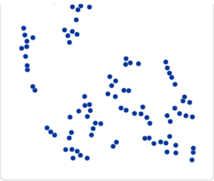
It's rare for data to be distributed in a perfectly symmetric way. Let's take a look at an axis-aligned Gaussian distribution in 50 dimensions, where the standard deviation in coordinate i is 1/i. That is, we're looking at a long-ish ellipsoidal cloud of points.



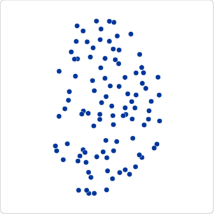




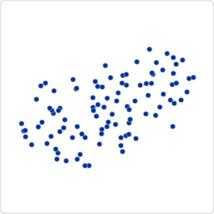
Perplexity: 5



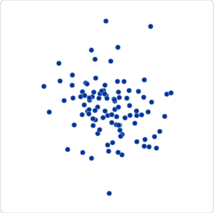
Step: 5,000



Perplexity: 30
Step: 5,000

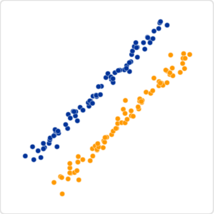


Perplexity: 50
Step: 5,000

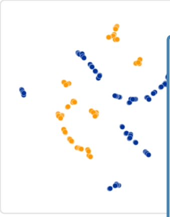


Perplexity: 100
Step: 5,000

For high enough perplexity values, the elongated shapes are easy to read. On the other hand, at low perplexity, local effects and meaningless "clumping" take center stage. More extreme shapes also come through, but again only at the right perplexity. For example, here are two clusters of 75 points each in 2D, arranged in parallel lines with a bit of noise.

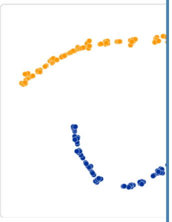


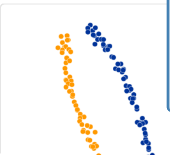
Original

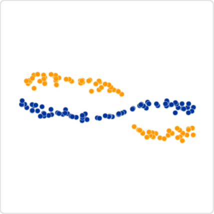


Perplexity: 2

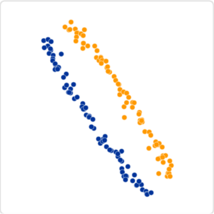
Step:







Perplexity: 50
Step: 5,000



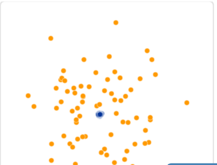
Perplexity: 100
Step: 5,000

For a certain range of perplexity the long clusters look close to correct, which is reassuring.

Even in the best cases, though, there's a subtle distortion: the lines are slightly curved outwards in the t-SNE diagram. The reason is that, as usual, t-SNE tends to expand denser regions of data. Since the middles of the clusters have less empty space around them than the ends, the algorithm magnifies them.

6. For topology, you may need more than one plot

Sometimes you can read topological information off a t-SNE plot, but that typically requires views at multiple perplexities. One of the simplest topological properties is containment. The plots below show two groups of 75 points in 50 dimensional space. Both are sampled from symmetric Gaussian distributions centered at the origin, but one is 50 times more tightly dispersed than the other. The "small" distribution is in effect contained in the large one.



Original

Step: