# umap Documentation

*Release 0.3*

**Leland McInnes**

**Jan 29, 2020**

# User Guide / Tutorial:

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualisation similarly to t-SNE, but also for general non-linear dimension reduction. The algorithm is founded on three assumptions about the data

1. The data is uniformly distributed on Riemannian manifold;
2. The Riemannian metric is locally constant (or can be approximated as such);
3. The manifold is locally connected.

From these assumptions it is possible to model the manifold with a fuzzy topological structure. The embedding is found by searching for a low dimensional projection of the data that has the closest possible equivalent fuzzy topological structure.

The details for the underlying mathematics can be found in our paper on ArXiv:

McInnes, L, Healy, J, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426, 2018

You can find the software on github.

**Installation**

Conda install, via the excellent work of the conda-forge team:

```
conda install -c conda-forge umap-learn
```

The conda-forge packages are available for linux, OS X, and Windows 64 bit.

PyPI install, presuming you have numba and sklearn and all its requirements (numpy and scipy) installed:

```
pip install umap-learn
```

CHAPTER 1

# How to Use UMAP

UMAP is a general purpose manifold learning and dimension reduction algorithm. It is designed to be compatible with scikit-learn, making use of the same API and able to be added to sklearn pipelines. If you are already familiar with sklearn you should be able to use UMAP as a drop in replacement for t-SNE and other dimension reduction classes. If you are not so familiar with sklearn this tutorial will step you through the basics of using UMAP to transform and visualise data.

First we'll need to import a bunch of useful tools. We will need numpy obviously, but we'll use some of the datasets available in sklearn, as well as the `train_test_split` function to divide up data. Finally we'll need some plotting tools (matplotlib and seaborn) to help us visualise the results of UMAP, and pandas to make that a little easier.

```
import numpy as np
from sklearn.datasets import load_iris, load_digits
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
%matplotlib inline
```

```
sns.set(style='white', context='notebook', rc={'figure.figsize':(14,10)})
```

## 1.1 Iris data

The next step is to get some data to work with. To ease us into things we'll start with the iris dataset. It isn't very representative of what real data would look like, but it is small both in number of points and number of features, and will let us get an idea of what the dimension reduction is doing. We can load the iris dataset from sklearn.

```
iris = load_iris()
print(iris.DESCR)
```

```
Iris Plants Database
====================

Notes
-----
Data Set Characteristics:
    :Number of Instances: 150 (50 in each of three classes)
    :Number of Attributes: 4 numeric, predictive attributes and the class
    :Attribute Information:
        - sepal length in cm
        - sepal width in cm
        - petal length in cm
        - petal width in cm
        - class:
                - Iris-Setosa
                - Iris-Versicolour
                - Iris-Virginica
    :Summary Statistics:

    ============== ==== ==== ======= ===== ====================
                    Min  Max   Mean    SD   Class Correlation
    ============== ==== ==== ======= ===== ====================
    sepal length:   4.3  7.9   5.84   0.83    0.7826
    sepal width:    2.0  4.4   3.05   0.43   -0.4194
    petal length:   1.0  6.9   3.76   1.76    0.9490  (high!)
    petal width:    0.1  2.5   1.20  0.76     0.9565  (high!)
    ============== ==== ==== ======= ===== ====================

    :Missing Attribute Values: None
    :Class Distribution: 33.3% for each of 3 classes.
    :Creator: R.A. Fisher
    :Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
    :Date: July, 1988

This is a copy of UCI ML iris datasets.
http://archive.ics.uci.edu/ml/datasets/Iris

The famous Iris database, first used by Sir R.A Fisher

This is perhaps the best known database to be found in the
pattern recognition literature.  Fisher's paper is a classic in the field and
is referenced frequently to this day.  (See Duda & Hart, for example.)  The
data set contains 3 classes of 50 instances each, where each class refers to a
type of iris plant.  One class is linearly separable from the other 2; the
latter are NOT linearly separable from each other.

References
----------
   - Fisher,R.A. "The use of multiple measurements in taxonomic problems"
     Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to
     Mathematical Statistics" (John Wiley, NY, 1950).
   - Duda,R.O., & Hart,P.E. (1973) Pattern Classification and Scene Analysis.
     (Q327.D83) John Wiley & Sons.  ISBN 0-471-22361-1.  See page 218.
   - Dasarathy, B.V. (1980) "Nosing Around the Neighborhood: A New System
     Structure and Classification Rule for Recognition in Partially Exposed
     Environments".  IEEE Transactions on Pattern Analysis and Machine
     Intelligence, Vol. PAMI-2, No. 1, 67-71.
```

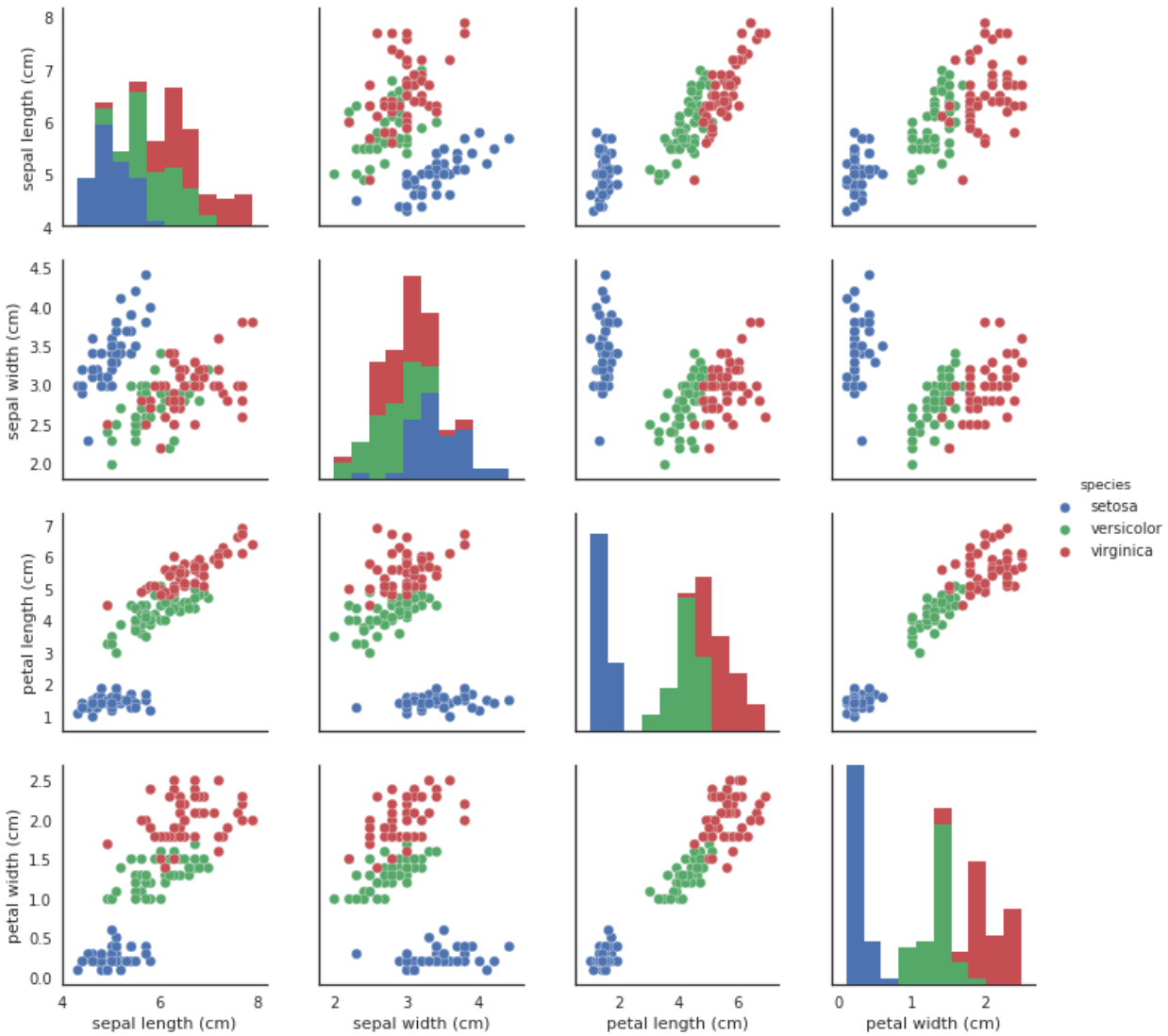
(continues on next page)

(continued from previous page)

```
  - Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule".  IEEE Transactions
    on Information Theory, May 1972, 431-433.
  - See also: 1988 MLC Proceedings, 54-64.  Cheeseman et al"s AUTOCLASS II
    conceptual clustering system finds 3 classes in the data.
  - Many, many more ...
```

The description tells us a fair amount about the dataset – it consists of measurements of petals and sepals of iris flowers. There are 3 species of flower represented, each with 50 sets of measurements. Visualizing this data is a little bit tricky since we can't plot in 4 dimensions easily. Fortunately four is not that large a number, so we can just to a pairwise feature scatterplot matrix to get an ideas of what is going on. Seaborn makes this easy (once we get the data into a pandas dataframe).

```
iris_df = pd.DataFrame(iris.data, columns=iris.feature_names)
iris_df['species'] = pd.Series(iris.target).map(dict(zip(range(3),iris.target_names)))
sns.pairplot(iris_df, hue='species');
```




This gives us some idea of what the data looks like by giving as all the 2D views of the data. Four dimensions is low enough that we can (sort of) reconstruct what the full dimensional data looks like in our heads. Now that we sort of

know what we are looking at, the question is what can a dimension reduction technique like UMAP do for us? By reducing the dimension in a way that preserves as much of the structure of the data as possible we can get a visualisable representation of the data allowing us to "see" the data and its structure and begin to get some inuitions about the data itself.

To use UMAP for this task we need to first construct a UMAP object that will do the job for us. That is as simple as instantiating the class. So let's import the umap library and do that.

```
import umap
```

```
reducer = umap.UMAP()
```

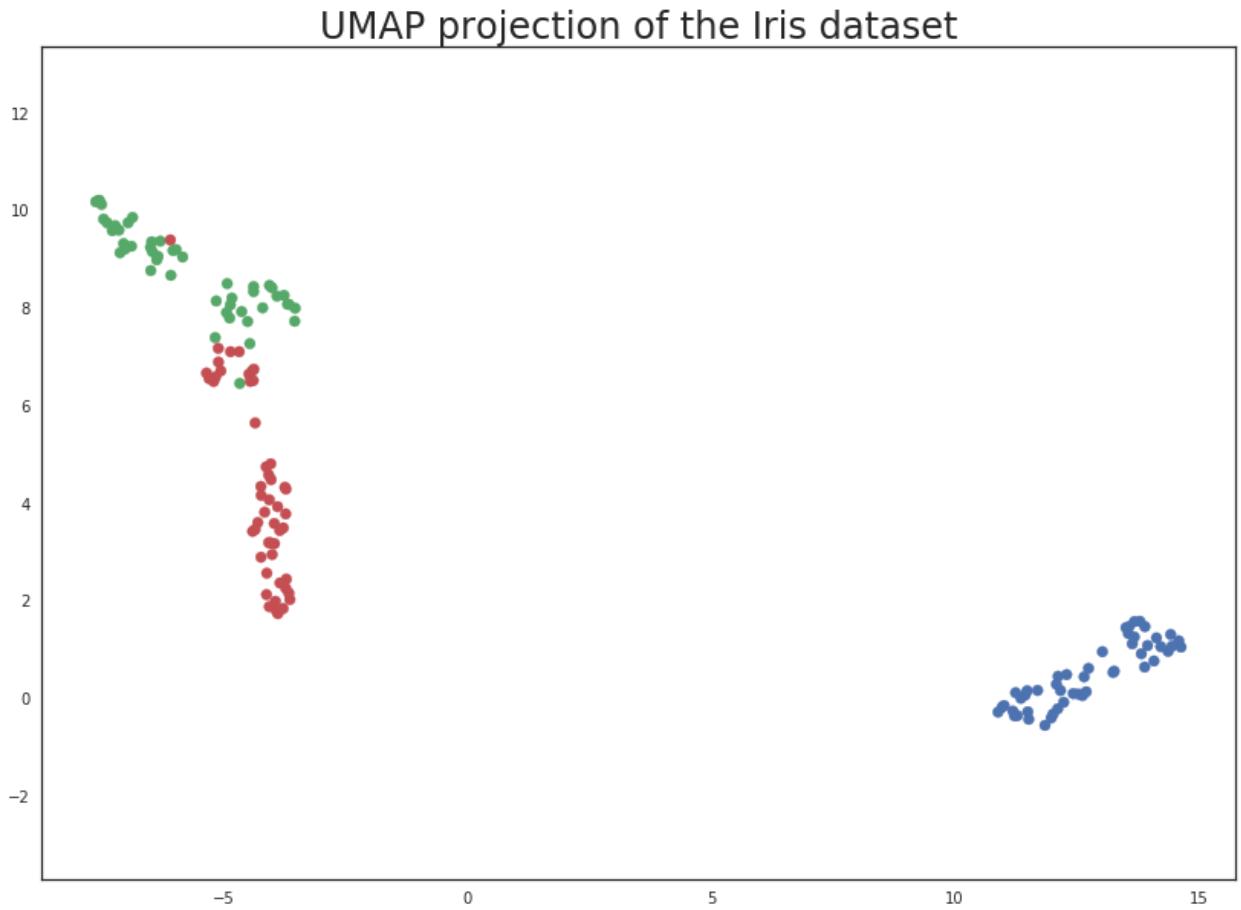
Now we need to train our reducer, letting it learn about the manifold. For this UMAP follows the sklearn API and has a method `fit` which we pass the data we want the model to learn from. Since, at the end of the day, we are going to want to reduced representation of the data we will use, instead, the `fit_transform` method which first calls `fit` and then returns the transformed data as a numpy array.

```
embedding = reducer.fit_transform(iris.data)
embedding.shape
```

```
(150, 2)
```

The result is an array with 150 samples, but only two feature columns (instead of the four we started with). This is because, by default, UMAP reduces down to 2D. Each row of the array is a 2-dimensional representation of the corresponding flower. Thus we can plot the `embedding` as a standard scatterplot and color by the target array (since it applies to the transformed data which is in the same order as the original).

```
plt.scatter(embedding[:, 0], embedding[:, 1], c=[sns.color_palette()[x] for x in iris.
↪target])
plt.gca().set_aspect('equal', 'datalim')
plt.title('UMAP projection of the Iris dataset', fontsize=24);
```




This does a useful job of capturing the structure of the data, and as can be seen from the matrix of scatterplots this is relatively accurate. Of course we learned at least this much just from that matrix of scatterplots – which we could do since we only had four differnt dimensions to analyse. If we had data with a larger number of dimensions the scatterplot matrix would quickly become unwieldy to plot, and far harder to interpret. So moving on from the Iris dataset, let's consider the digits dataset.

## 1.2 Digits data

First we will load the dataset from sklearn.

```
digits = load_digits()
print(digits.DESCR)
```

```
Optical Recognition of Handwritten Digits Data Set
==================================================

Notes
-----
Data Set Characteristics:
    :Number of Instances: 5620
    :Number of Attributes: 64
    :Attribute Information: 8x8 image of integer pixels in the range 0..16.
    :Missing Attribute Values: None
```

(continues on next page)

(continued from previous page)

```
    :Creator: E. Alpaydin (alpaydin '@' boun.edu.tr)
    :Date: July; 1998

This is a copy of the test set of the UCI ML hand-written digits datasets
http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits

The data set contains images of hand-written digits: 10 classes where
each class refers to a digit.

Preprocessing programs made available by NIST were used to extract
normalized bitmaps of handwritten digits from a preprinted form. From a
total of 43 people, 30 contributed to the training set and different 13
to the test set. 32x32 bitmaps are divided into nonoverlapping blocks of
4x4 and the number of on pixels are counted in each block. This generates
an input matrix of 8x8 where each element is an integer in the range
0..16. This reduces dimensionality and gives invariance to small
distortions.

For info on NIST preprocessing routines, see M. D. Garris, J. L. Blue, G.
T. Candela, D. L. Dimmick, J. Geist, P. J. Grother, S. A. Janet, and C.
L. Wilson, NIST Form-Based Handprint Recognition System, NISTIR 5469,
1994.

References
----------
  - C. Kaynak (1995) Methods of Combining Multiple Classifiers and Their
    Applications to Handwritten Digit Recognition, MSc Thesis, Institute of
    Graduate Studies in Science and Engineering, Bogazici University.
  - E. Alpaydin, C. Kaynak (1998) Cascading Classifiers, Kybernetika.
  - Ken Tang and Ponnuthurai N. Suganthan and Xi Yao and A. Kai Qin.
    Linear dimensionalityreduction using relevance weighted LDA. School of
    Electrical and Electronic Engineering Nanyang Technological University.
    2005.
  - Claudio Gentile. A New Approximate Maximal Margin Classification
    Algorithm. NIPS. 2000.
```

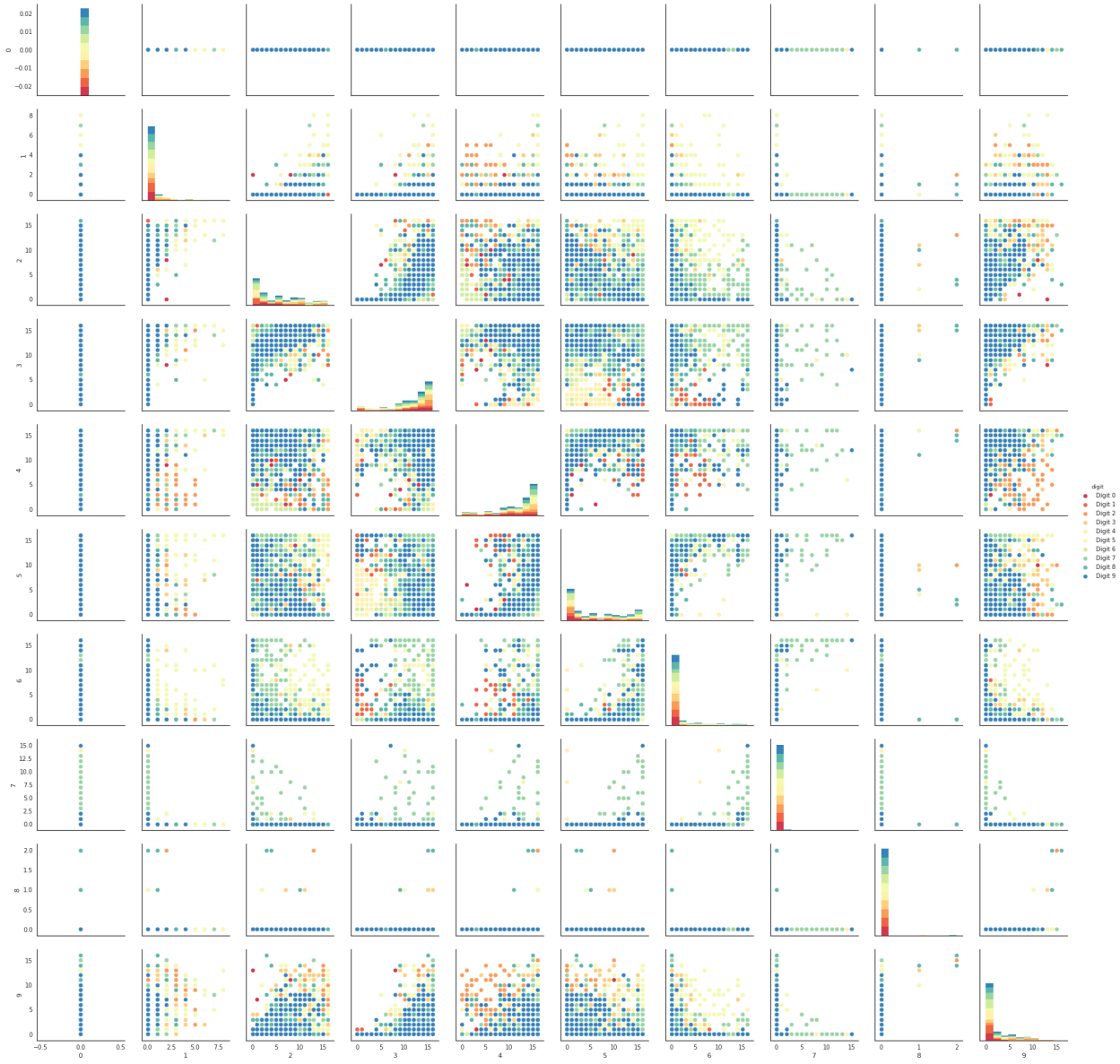
We can plot a number of the images to get an idea of what we are looking at. This just involves matplotlib building a grid of axes and then looping through them plotting an image into each one in turn.

```
fig, ax_array = plt.subplots(20, 20)
axes = ax_array.flatten()
for i, ax in enumerate(axes):
    ax.imshow(digits.images[i], cmap='gray_r')
plt.setp(axes, xticks=[], yticks=[], frame_on=False)
plt.tight_layout(h_pad=0.5, w_pad=0.01)
```

As you can see these are quite low resolution images – for the most part they are recognisable as digits, but there are a number of cases that are sufficiently blurred as to be questionable even for a human to guess at. The zeros do stand out as the easiest to pick out as notably different and clearly zeros. Beyond that things get a little harder: some of the squashed thing eights look awfully like ones, some of the threes start to look a little like crossed sevens when drawn badly, and so on.

Each image can be unfolded into a 64 element long vector of grayscale values. It is these 64 dimensional vectors that we wish to analyse: how much of the digits structure can we discern? At least in principle 64 dimensions is overkill for this task, and we would reasonably expect that there should be some smaller number of "latent" features that would be sufficient to describe the data reasonably well. We can try a scatterplot matrix – in this case just of the first 10 dimensions so that it is at least plottable, but as you can quickly see that approach is not going to be sufficient for this data.

```
digits_df = pd.DataFrame(digits.data[:,:10])
digits_df['digit'] = pd.Series(digits.target).map(lambda x: 'Digit {}'.format(x))
sns.pairplot(digits_df, hue='digit', palette='Spectral');
```




In contrast we can try using UMAP again. It works exactly as before: construct a model, train the model, and then look at the transformed data. TO demonstrate more of UMAP we'll go about it differently this time and simply use the `fit` method rather than the `fit_transform` approach we used for Iris.

```
reducer = umap.UMAP(random_state=42)
reducer.fit(digits.data)
```

```
UMAP(a=1.576943460405378, alpha=1.0, angular_rp_forest=False,
   b=0.8950608781227859, bandwidth=1.0, gamma=1.0, init='spectral',
   local_connectivity=1.0, metric='euclidean', metric_kwds={},
   min_dist=0.1, n_components=2, n_epochs=None, n_neighbors=15,
   negative_sample_rate=5, random_state=42, set_op_mix_ratio=1.0,
   spread=1.0, target_metric='categorical', target_metric_kwds={},
   transform_queue_size=4.0, transform_seed=42, verbose=False)
```

Now, instead of returning an embedding we simply get back the reducer object, now having trained on the dataset we passed it. To access the resulting transform we can either look at the `embedding_` attribute of the reducer object, or

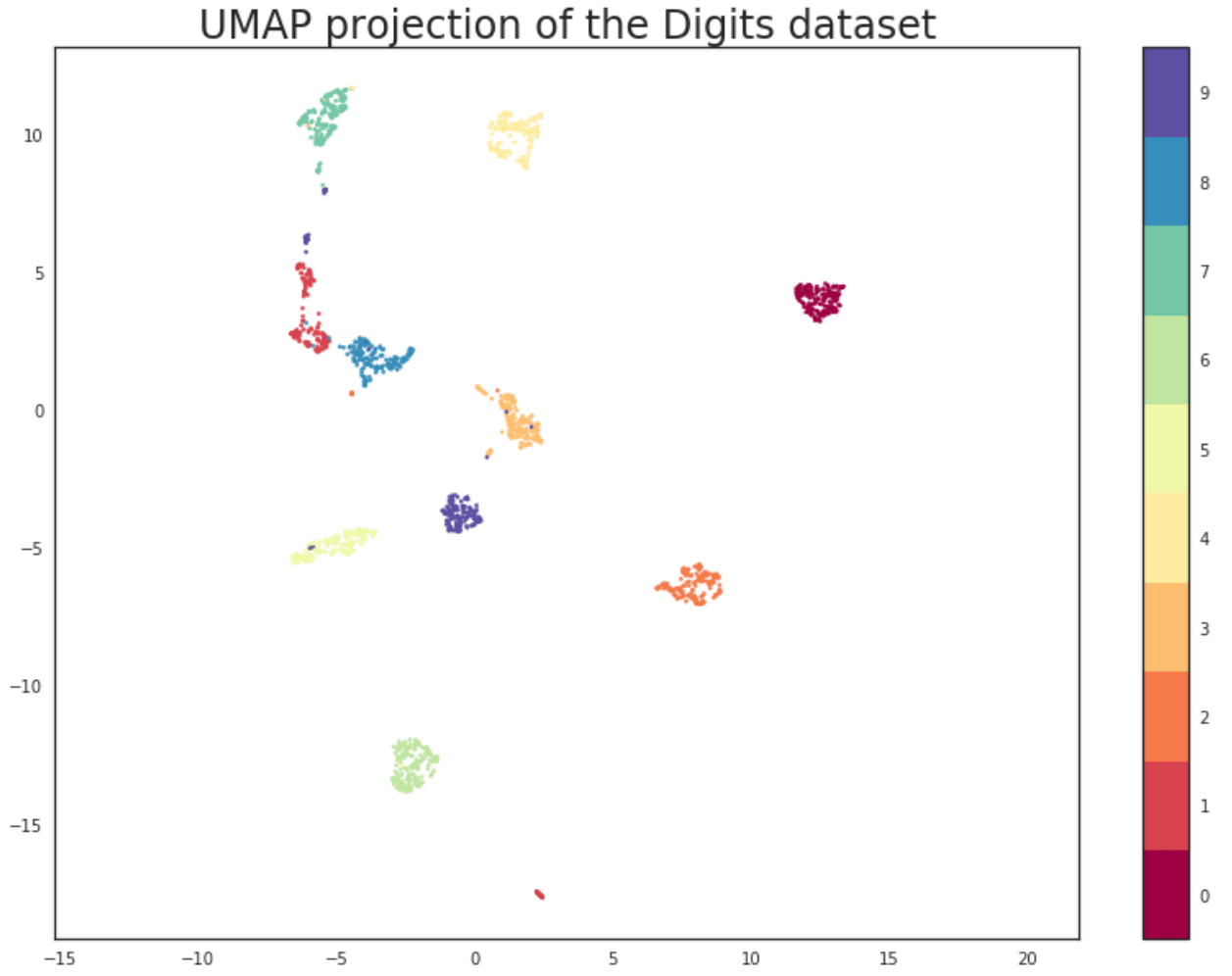
call transform on the original data.

```
embedding = reducer.transform(digits.data)
# Verify that the result of calling transform is
# idenitical to accessing the embedding_ attribute
assert(np.all(embedding == reducer.embedding_))
embedding.shape
```

```
(1797, 2)
```

We now have a dataset with 1797 rows (one for each hand-written digit sample), but only 2 columns. As with the Iris example we can now plot the resulting embedding, coloring the data points by the class that theyr belong to (i.e. the digit they represent).

```
plt.scatter(embedding[:, 0], embedding[:, 1], c=digits.target, cmap='Spectral', s=5)
plt.gca().set_aspect('equal', 'datalim')
plt.colorbar(boundaries=np.arange(11)-0.5).set_ticks(np.arange(10))
plt.title('UMAP projection of the Digits dataset', fontsize=24);
```




We see that UMAP has successfully captured the digit classes. There are also some interesting effects as some digit classes blend into one another (see the eights, ones, and sevens, with some nines in between), and also cases where digits are pushed away as clearly distinct (the zeros on the right, the fours at the top, and a small subcluster of ones at the bottom come to mind). To get a better idea of why UMAP chose to do this it is helpful to see the actual digits

involve. One can do this using bokeh and mouseover tooltips of the images.

First we'll need to encode all the images for inclusion in a dataframe.

```
from io import BytesIO
from PIL import Image
import base64
```

```
def embeddable_image(data):
    img_data = 255 - 15 * data.astype(np.uint8)
    image = Image.fromarray(img_data, mode='L').resize((64, 64), Image.BICUBIC)
    buffer = BytesIO()
    image.save(buffer, format='png')
    for_encoding = buffer.getvalue()
    return '' + base64.b64encode(for_encoding).decode()
```

Next we need to load up bokeh and the various tools from it that will be needed to generate a suitable interactive plot.

```
from bokeh.plotting import figure, show, output_notebook
from bokeh.models import HoverTool, ColumnDataSource, CategoricalColorMapper
from bokeh.palettes import Spectral10

output_notebook()
```

Finally we generate the plot itself with a custom hover tooltip that embeds the image of the digit in question in it, along with the digit class that the digit is actually from (this can be useful for digits that are hard even for humans to classify correctly).

```
digits_df = pd.DataFrame(embedding, columns=('x', 'y'))
digits_df['digit'] = [str(x) for x in digits.target]
digits_df['image'] = list(map(embeddable_image, digits.images))

datasource = ColumnDataSource(digits_df)
color_mapping = CategoricalColorMapper(factors=[str(9 - x) for x in digits.target_
↪names],
                                       palette=Spectral10)

plot_figure = figure(
    title='UMAP projection of the Digits dataset',
    plot_width=600,
    plot_height=600,
    tools=('pan, wheel_zoom, reset')
)

plot_figure.add_tools(HoverTool(tooltips="""
<div>
    <div>
        <img src='@image' style='float: left; margin: 5px 5px 5px 5px'/>
    </div>
    <div>
        <span style='font-size: 16px; color: #224499'>Digit:</span>
        <span style='font-size: 18px'>@digit</span>
    </div>
</div>
"""))

plot_figure.circle(
    'x',
```

(continues on next page)

(continued from previous page)

```
    'y',
    source=datasource,
    color=dict(field='digit', transform=color_mapping),
    line_alpha=0.6,
    fill_alpha=0.6,
    size=4
)
show(plot_figure)
```

As can be seen, the nines that blend between the ones and the sevens are odd looking nines (that aren't very rounded) and do, indeed, interpolate surprisingly well between ones with hats and crossed sevens. In contrast the small disjoint cluster of ones at the bottom of the plot is made up of ones with feet (a horizontal line at the base of the one) which are, indeed, quite distinct from the general mass of ones.

This concludes our introduction to basic UMAP usage – hopefully this has given you the tools to get started for yourself. Further tutorials, covering UMAP parameters and more advanced usage are also available when you wish to dive deeper.

CHAPTER 2

# Basic UMAP Parameters

UMAP is a fairly flexible non-linear dimension reduction algorithm. It seeks to learn the manifold structure of your data and find a low dimensional embedding that preserves the essential topological structure of that manifold. In this notebook we will generate some visualisable 4-dimensional data, demonstrate how to use UMAP to provide a 2-dimensional representation of it, and then look at how various UMAP parameters can impact the resulting embedding. This documentation is based on the work of Philippe Rivière for visionscarto.net.

To start we'll need some basic libraries. First `numpy` will be needed for basic array manipulation. Since we will be visualising the results we will need `matplotlib` and `seaborn`. Finally we will need `umap` for doing the dimension reduction itself.

```
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import seaborn as sns
import umap
%matplotlib inline
```

```
sns.set(style='white', context='poster', rc={'figure.figsize':(14,10)})
```

Next we will need some data to embed into a lower dimensional representation. To make the 4-dimensional data "visualisable" we will generate data uniformly at random from a 4-dimensional cube such that we can interpret a sample as a tuple of (R,G,B,a) values specifying a color (and translucency). Thus when we plot low dimensional representations each point can colored according to its 4-dimensional value. For this we can use `numpy`. We will fix a random seed for the sake of consistency.

```
np.random.seed(42)
data = np.random.rand(800, 4)
```

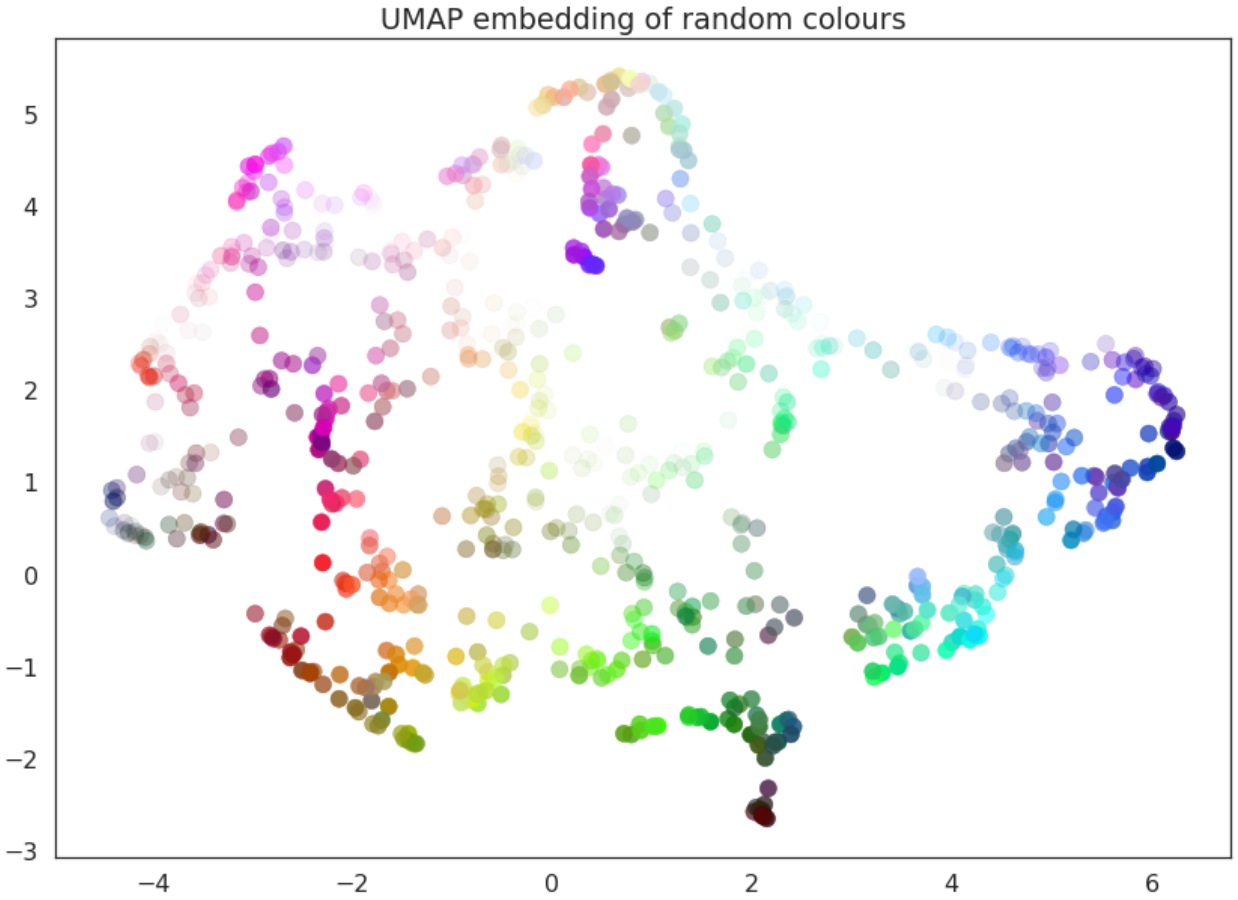
Now we need to find a low dimensional representation of the data. As in the Basic Usage documentation, we can do this by using the *`fit_transform()`* method on a *`UMAP`* object.

```
fit = umap.UMAP()
%time u = fit.fit_transform(data)
```

```
CPU times: user 7.73 s, sys: 211 ms, total: 7.94 s
Wall time: 6.8 s
```

The resulting value `u` is a 2-dimensional representation of the data. We can visualise the result by using `matplotlib` to draw a scatter plot of `u`. We can color each point of the scatter plot by the associated 4-dimensional color from the source data.

```
plt.scatter(u[:,0], u[:,1], c=data)
plt.title('UMAP embedding of random colours');
```




As you can see the result is that the data is placed in 2-dimensional space such that points that were nearby in in 4-dimensional space (i.e. are similar colors) are kept close together. Since we drew a random selection of points in the color cube there is a certain amount of induced structure from where the random points happened to clump up in color space.

UMAP has several hyperparameters that can have a significant impact on the resulting embedding. In this notebook we will be covering the four major ones:

- `n_neighbors`
- `min_dist`
- `n_components`
- `metric`

Each of these parameters has a distinct effect, and we will look at each in turn. To make exploration simpler we will first write a short utility function that can fit the data with UMAP given a set of parameter choices, and plot the result.

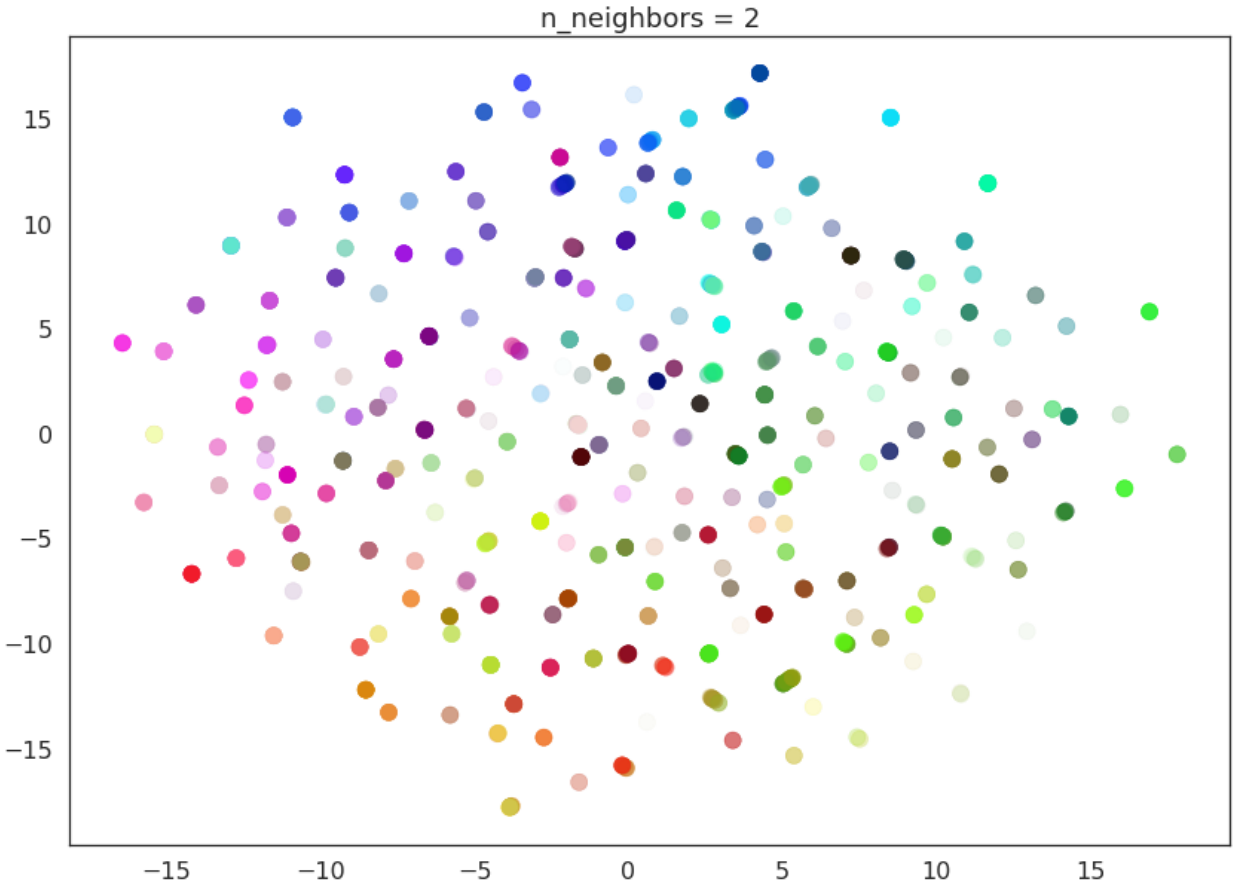
```
def draw_umap(n_neighbors=15, min_dist=0.1, n_components=2, metric='euclidean', title=
↪''):
    fit = umap.UMAP(
        n_neighbors=n_neighbors,
        min_dist=min_dist,
        n_components=n_components,
        metric=metric
    )
    u = fit.fit_transform(data);
    fig = plt.figure()
    if n_components == 1:
        ax = fig.add_subplot(111)
        ax.scatter(u[:,0], range(len(u)), c=data)
    if n_components == 2:
        ax = fig.add_subplot(111)
        ax.scatter(u[:,0], u[:,1], c=data)
    if n_components == 3:
        ax = fig.add_subplot(111, projection='3d')
        ax.scatter(u[:,0], u[:,1], u[:,2], c=data, s=100)
    plt.title(title, fontsize=18)
```

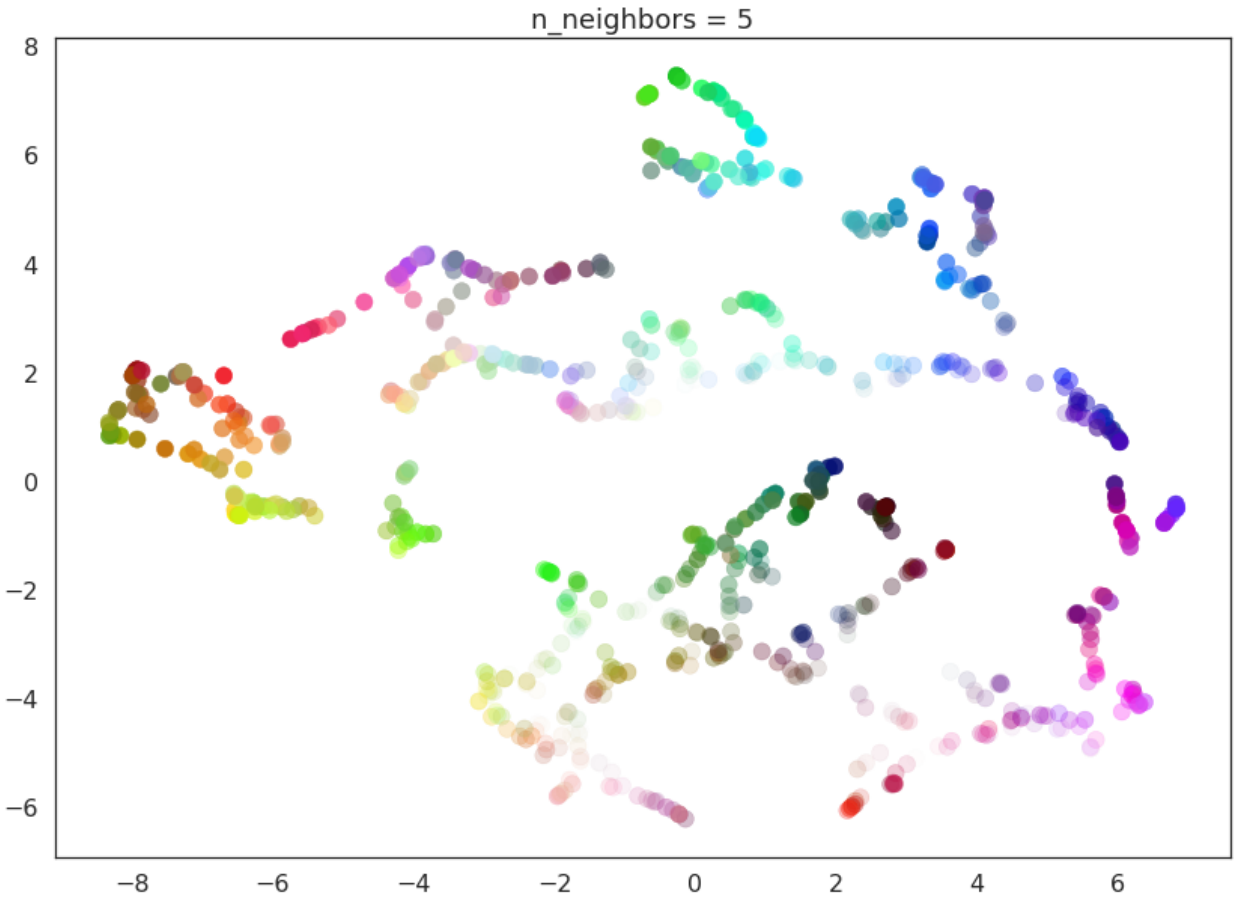
## 2.1 n_neighbors

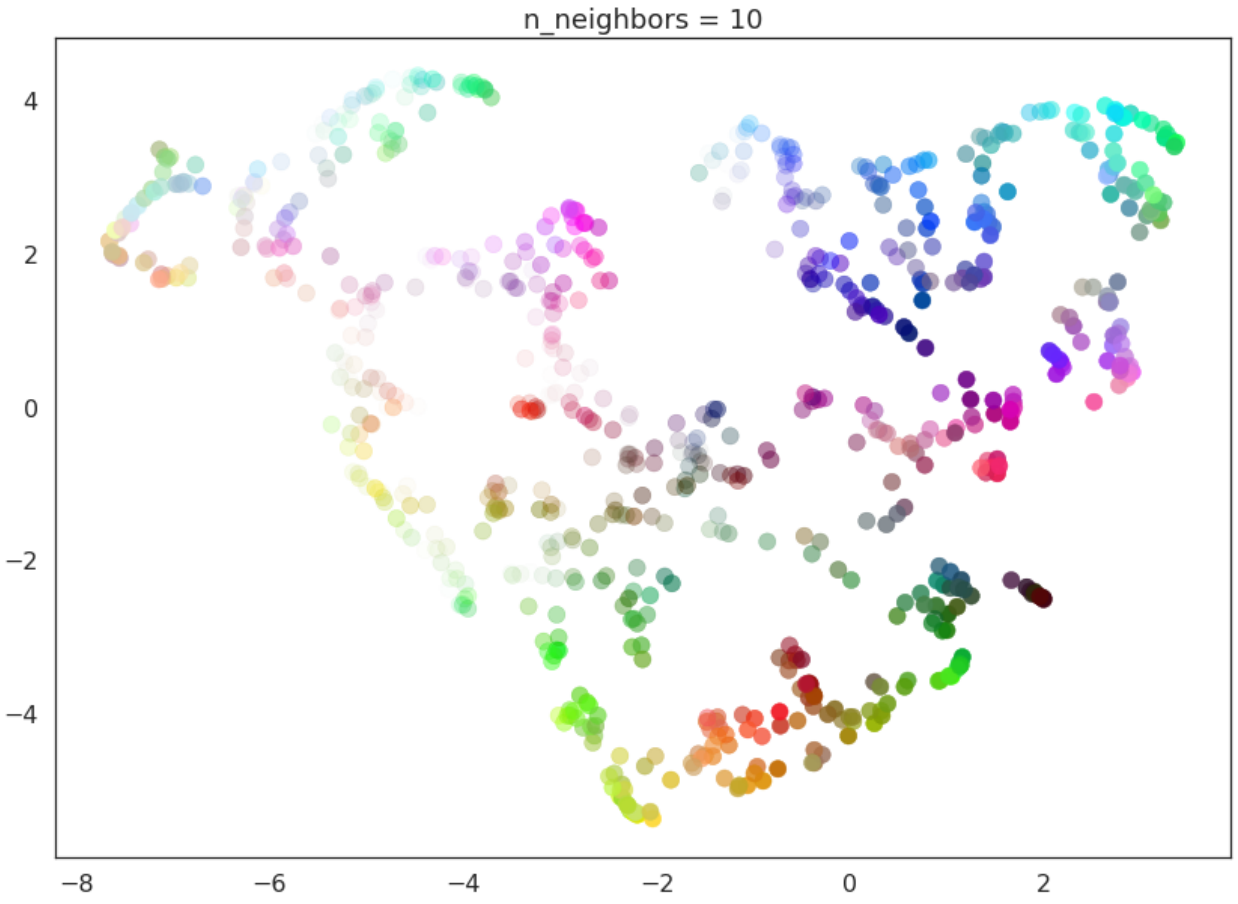
This parameter controls how UMAP balances local versus global structure in the data. It does this by constraining the size of the local neighborhood UMAP will look at when attempting to learn the manifold structure of the data. This means that low values of `n_neighbors` will force UMAP to concentrate on very local structure (potentially to the detriment of the big picture), while large values will push UMAP to look at larger neighborhoods of each point when estimating the manifold structure of the data, losing fine detail structure for the sake of getting the broader of the data.

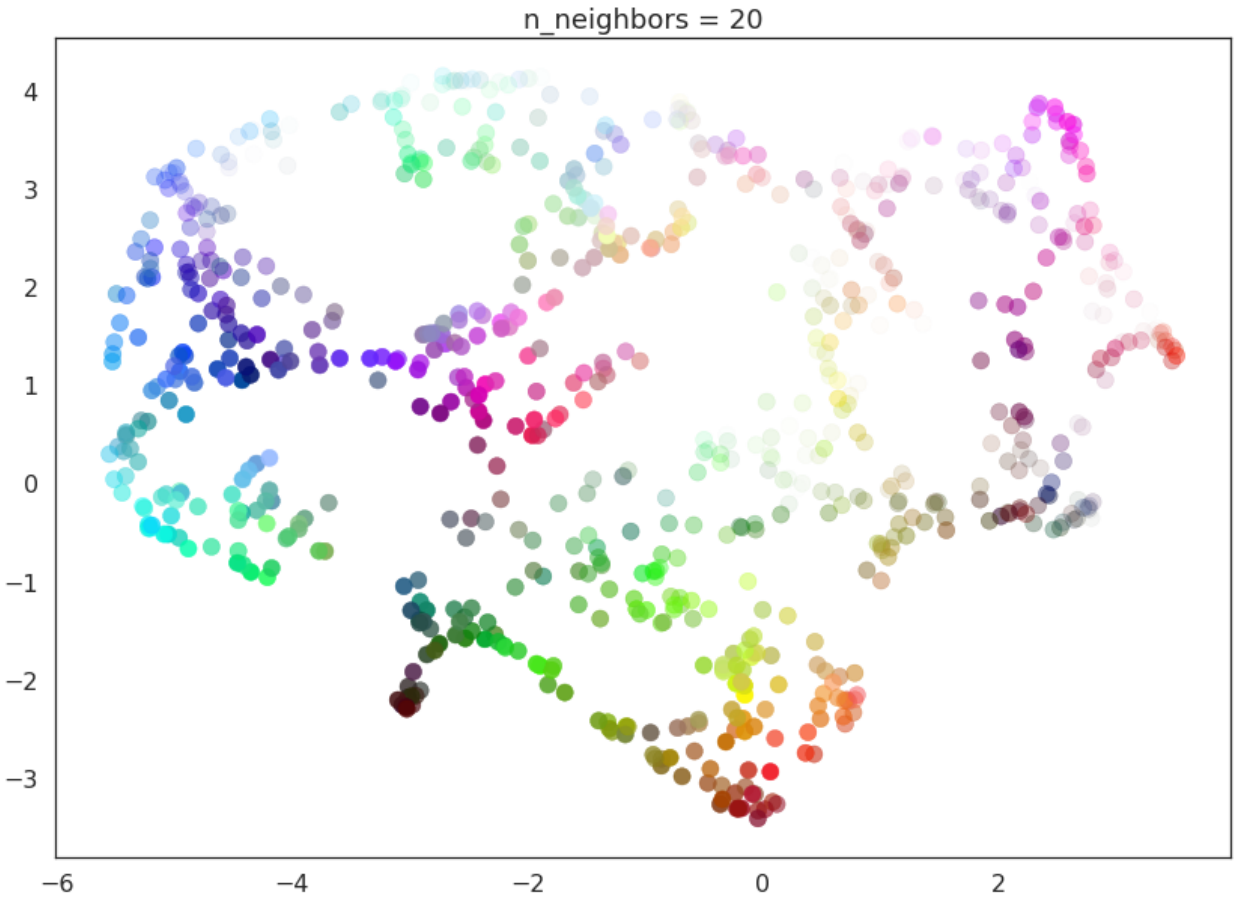
We can see that in practice by fitting our dataset with UMAP using a range of `n_neighbors` values. The default value of `n_neighbors` for UMAP (as used above) is 15, but we will look at values ranging from 2 (a very local view of the manifold) up to 200 (a quarter of the data).

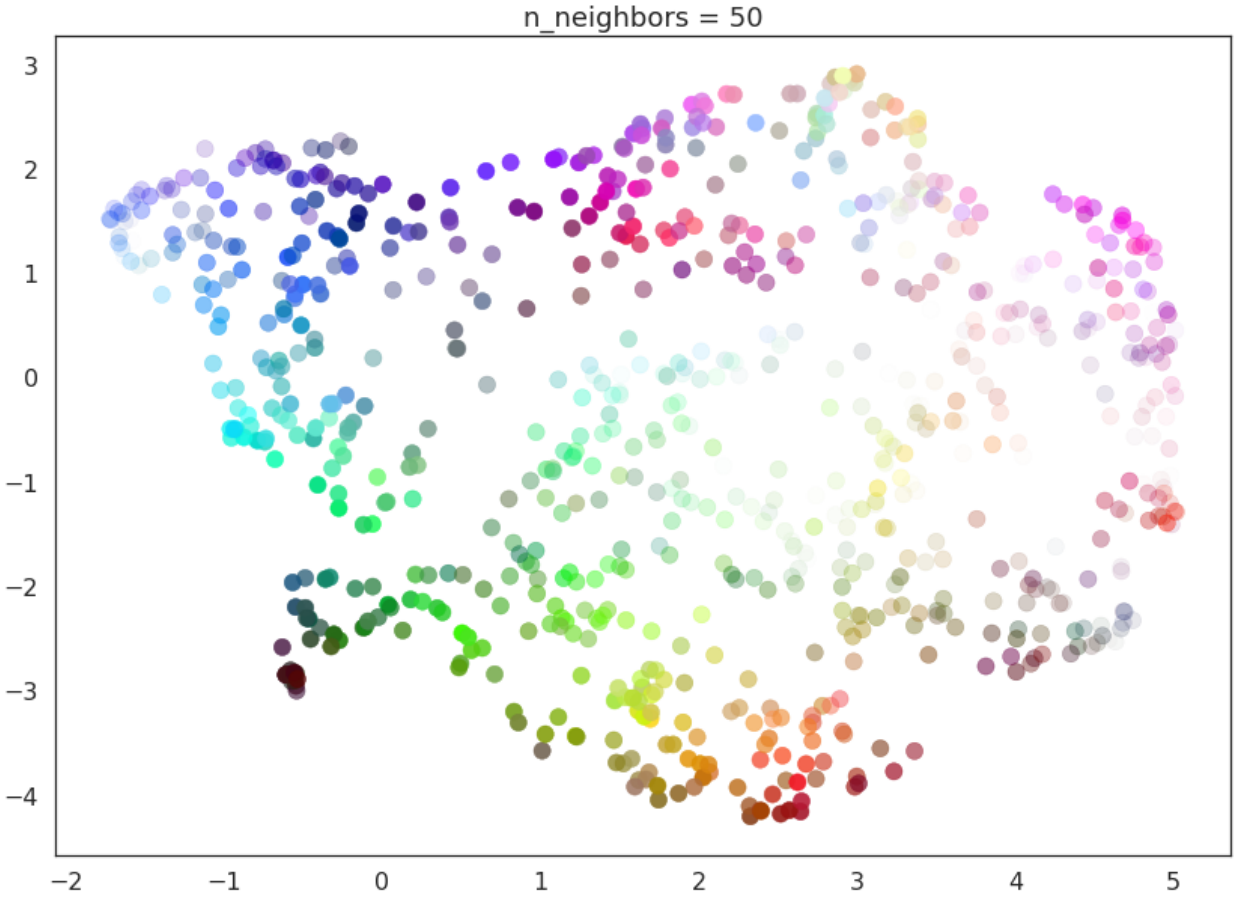
```
for n in (2, 5, 10, 20, 50, 100, 200):
    draw_umap(n_neighbors=n, title='n_neighbors = {}'.format(n))
```

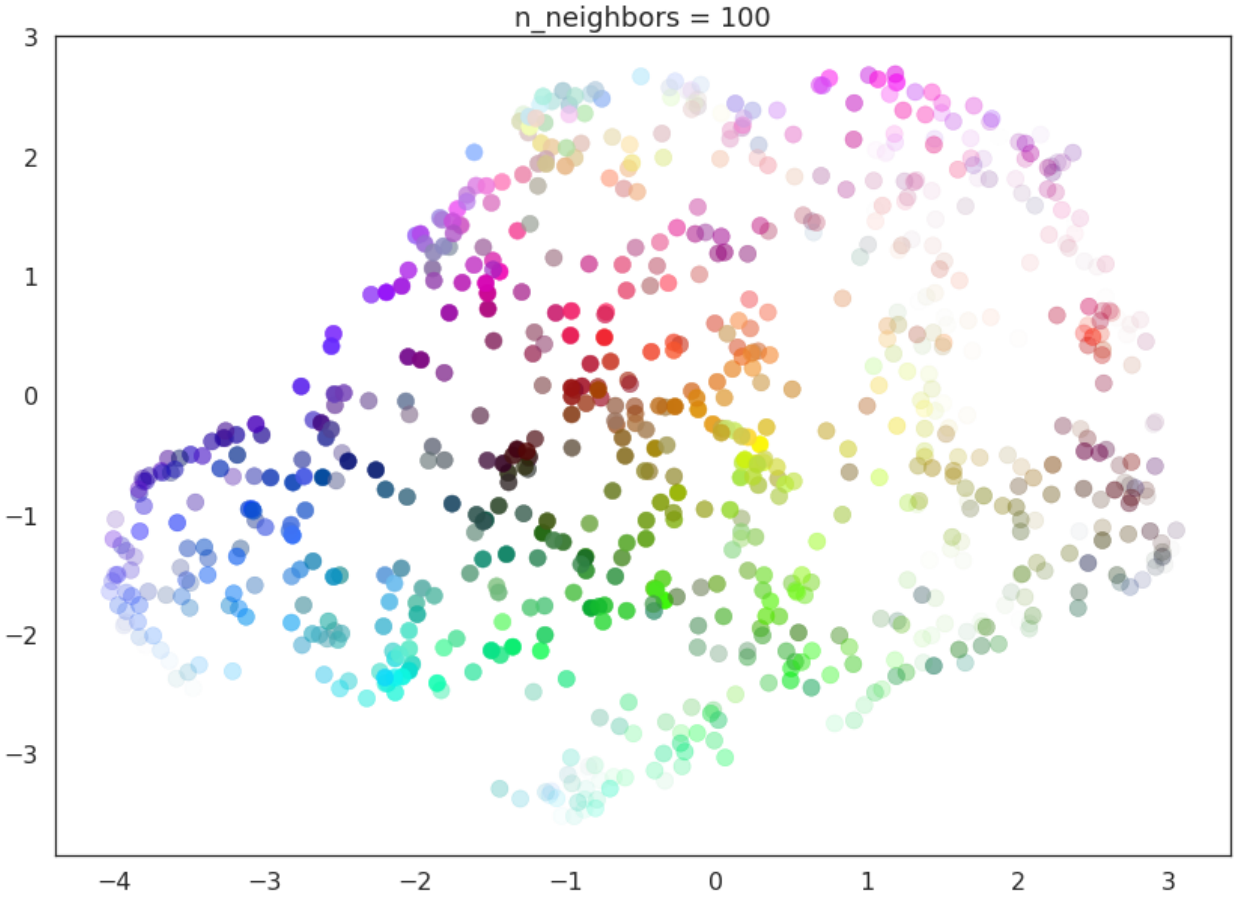

n_neighbors = 2
15
10
5
0
−5
−10
−15
−15
−10
−5
0
5
10
15

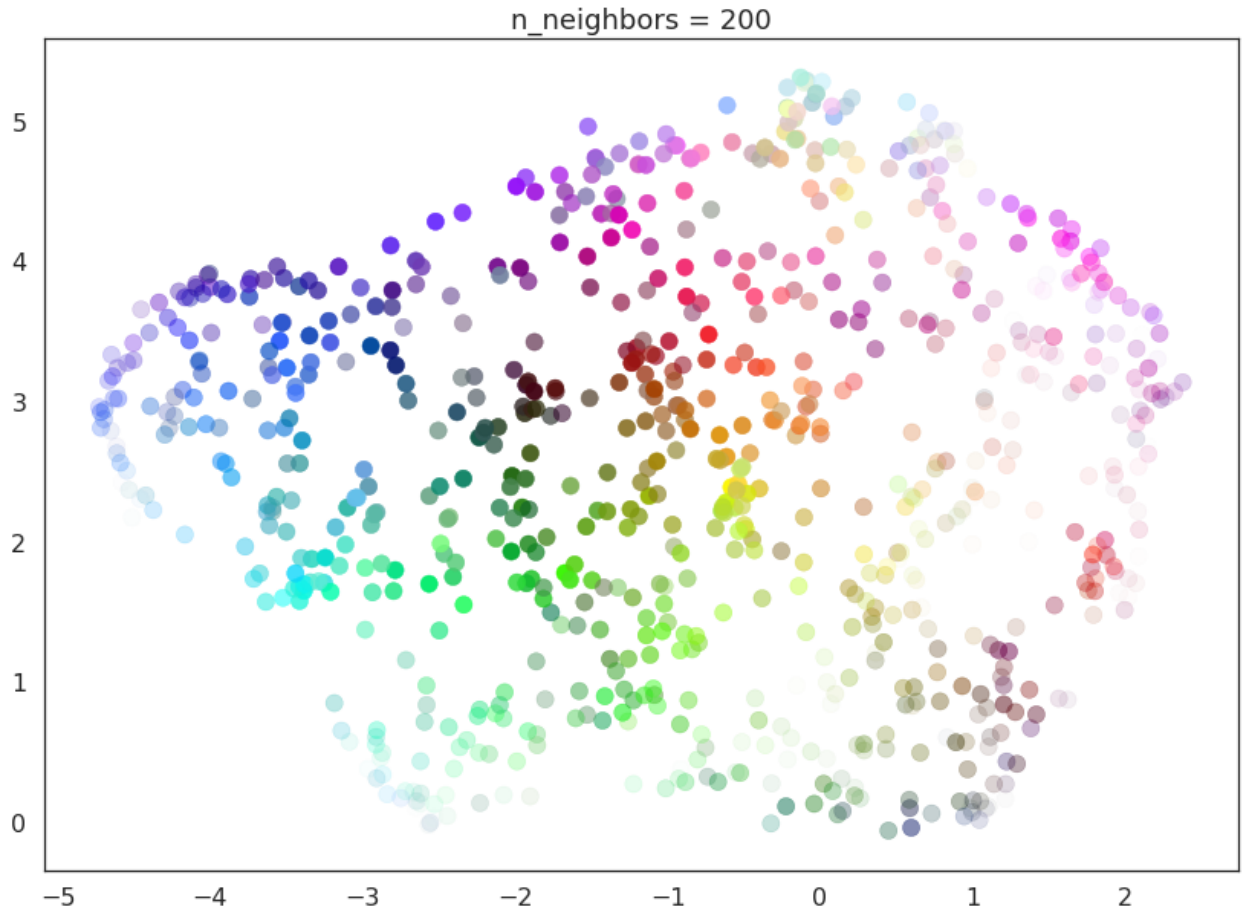

n_neighbors = 5
8
6
4
2
0
−2
−4
−6
−8
−6
−4
−2
0
2
4
6


n_neighbors = 10
4
2
0
−2
−4
−8
−6
−4
−2
0
2


n_neighbors = 20
4
3
2
1
0
−1
−2
−3
−6
−4
−2
0
2


n_neighbors = 50
3
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2
3
4
5


n_neighbors = 100
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3




With a value of `n_neighbors=2` we see that UMAP merely glues together small chains, but due to the narrow/local view, fails to see how those connect together. It also leaves many different components (and even singleton points). This represents the fact that from a fine detail point of view the data is very disconnected and scattered throughout the space.

As `n_neighbors` is increased UMAP manages to see more of the overall structure of the data, gluing more components together, and better coverying the broader structure of the data. By the stage of `n_neighbors=20` we have a fairly good overall view of the data showing how the various colors interelate to each other over the whole dataset.

As `n_neighbors` increases further more and more focus in placed on the overall structure of the data. This results in, with `n_neighbors=200` a plot where the overall structure (blues, greens, and reds; high luminance versus low) is well captured, but at the loss of some of the finer local sturcture (individual colors are no longer necessarily immediately near their closest color match).

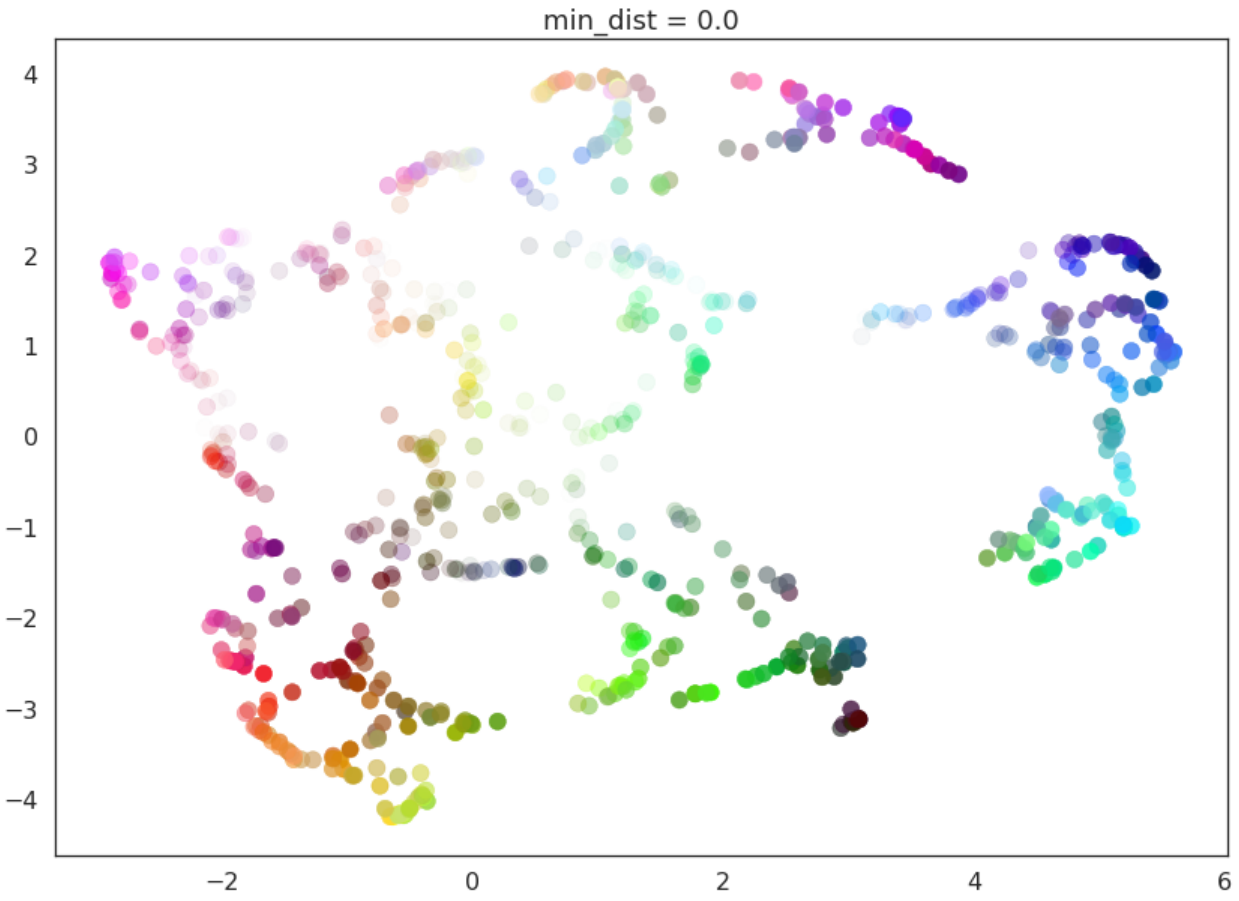
This effect well exemplifies the local/global tradeoff provided by `n_neighbors`.

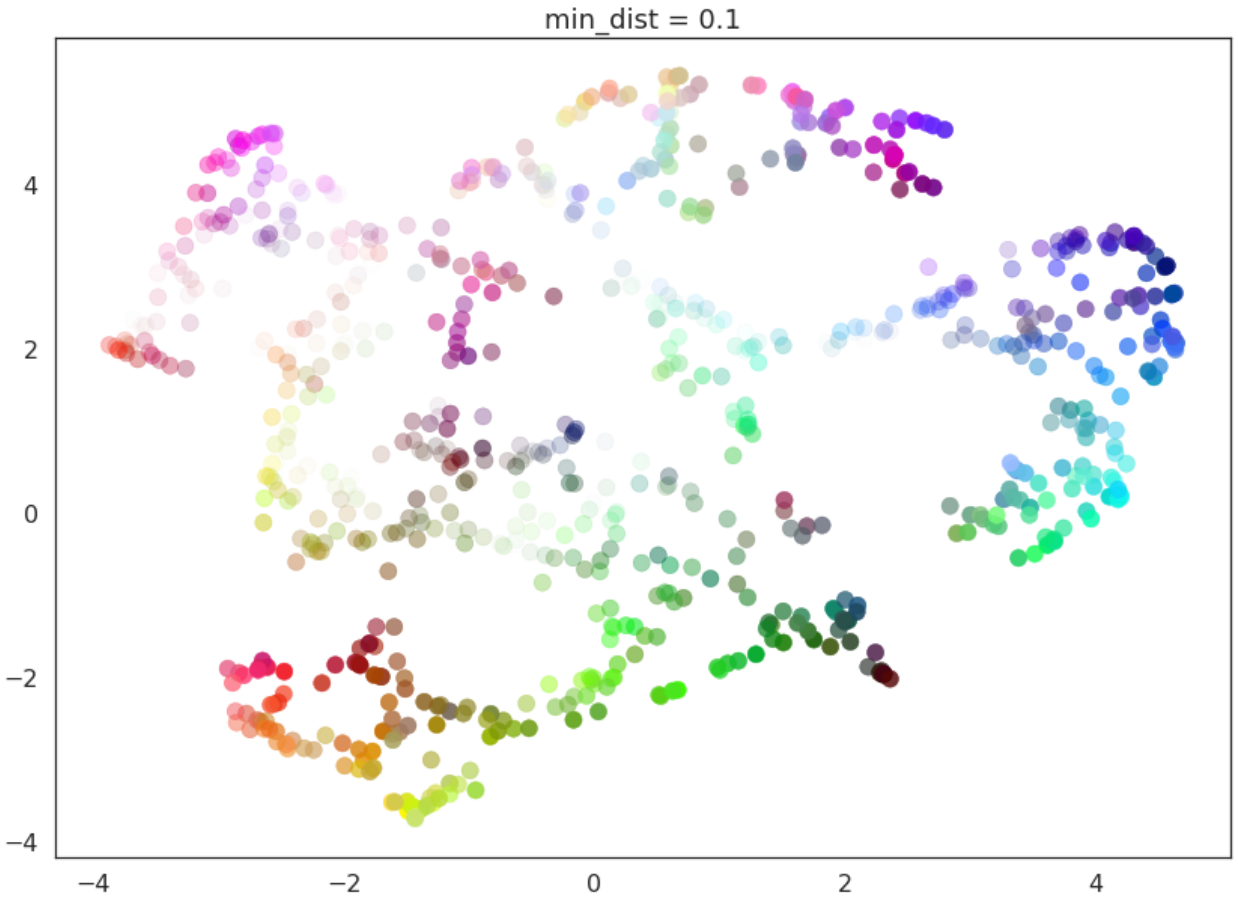
## 2.2 `min_dist`

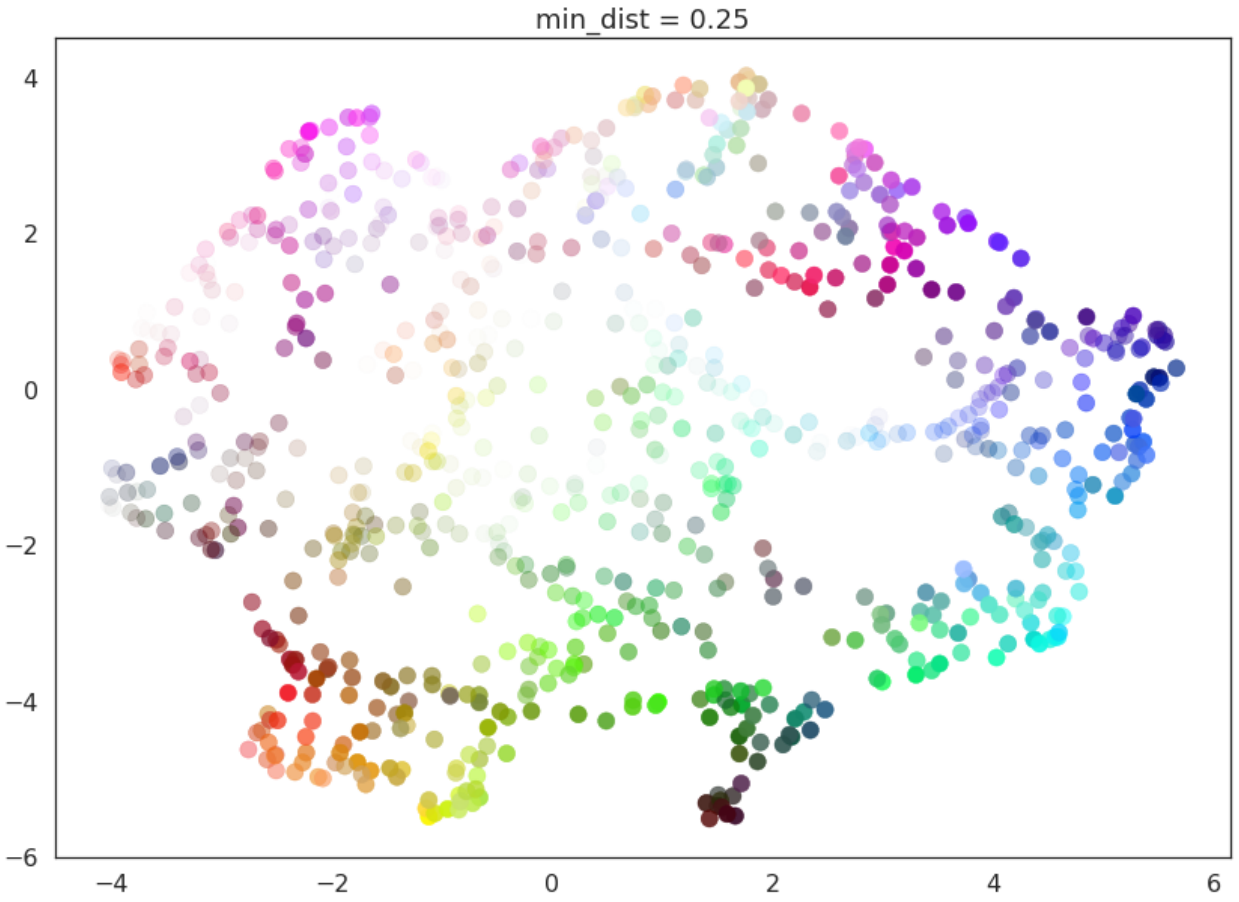
The `min_dist` parameter controls how tightly UMAP is allowed to pack points together. It, quite literally, provides the minimum distance apart that points are allowed to be in the low dimensional representation. This means that low values of `min_dist` will result in clumpier embeddings. This can be useful if you are interested in clustering, or in finer topological structure. Larger values of `min_dist` will prevent UMAP from packing point together and will focus instead on the preservation of the broad topological structure instead.

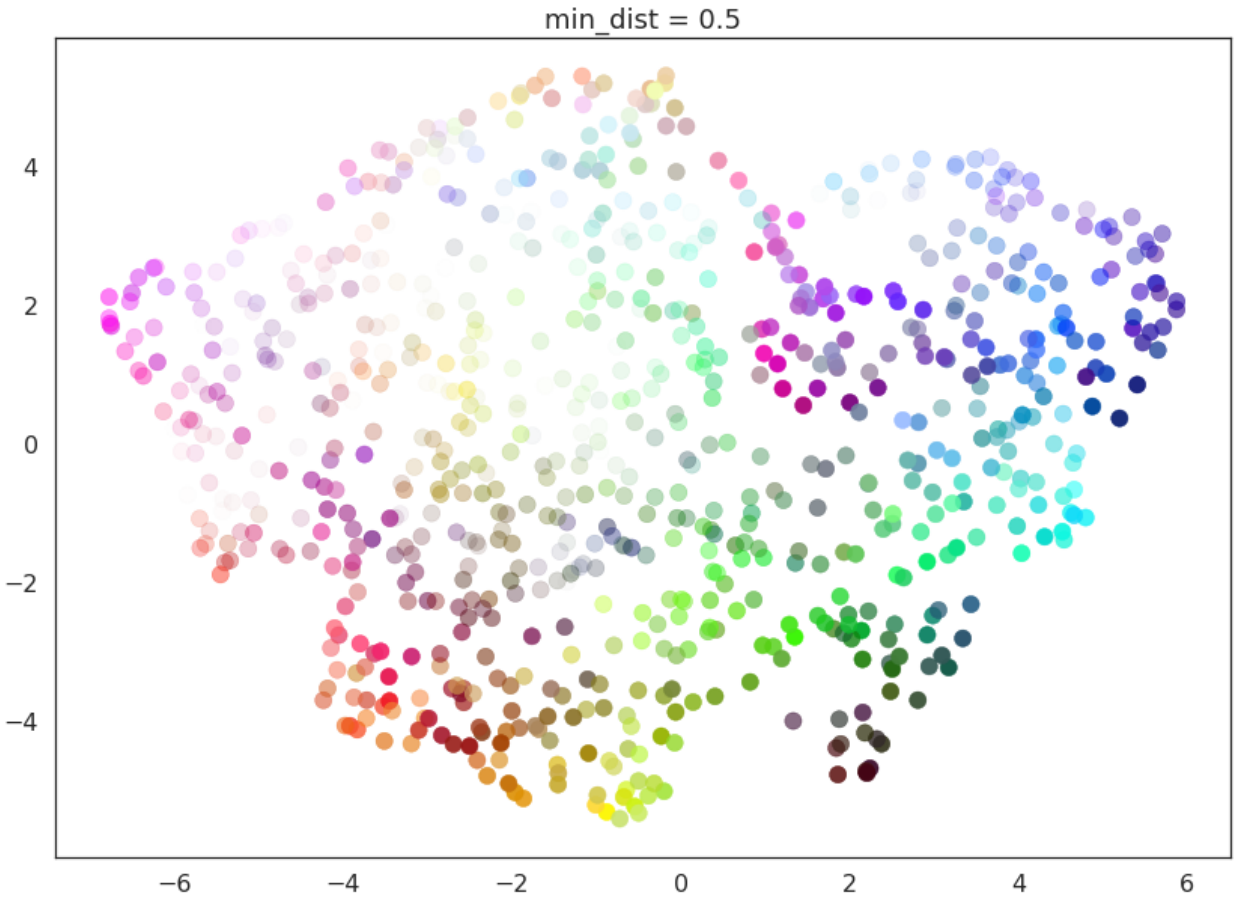
The default value for `min_dist` (as used above) is 0.1. We will look at a range of values from 0.0 through to 0.99.

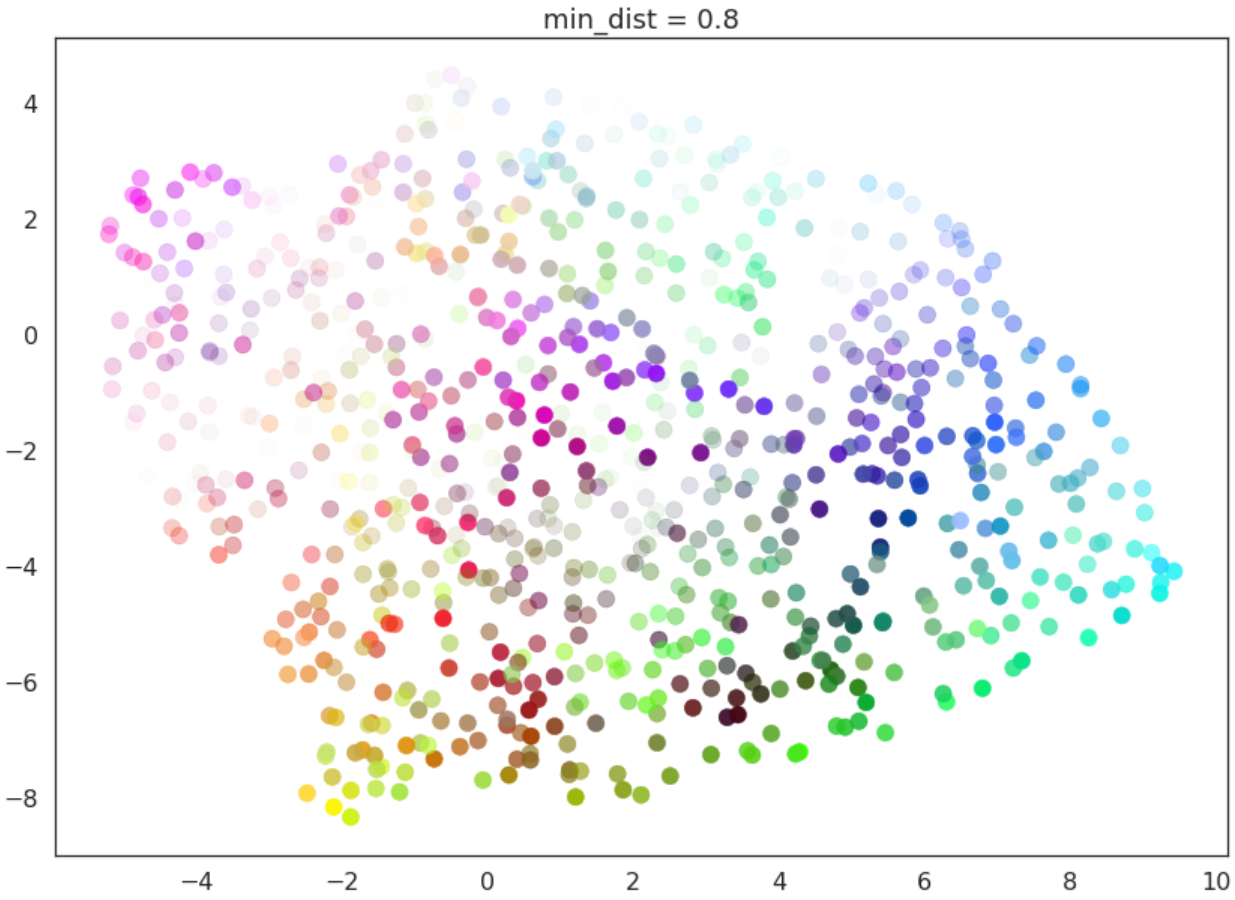
```
for d in (0.0, 0.1, 0.25, 0.5, 0.8, 0.99):
    draw_umap(min_dist=d, title='min_dist = {}'.format(d))
```

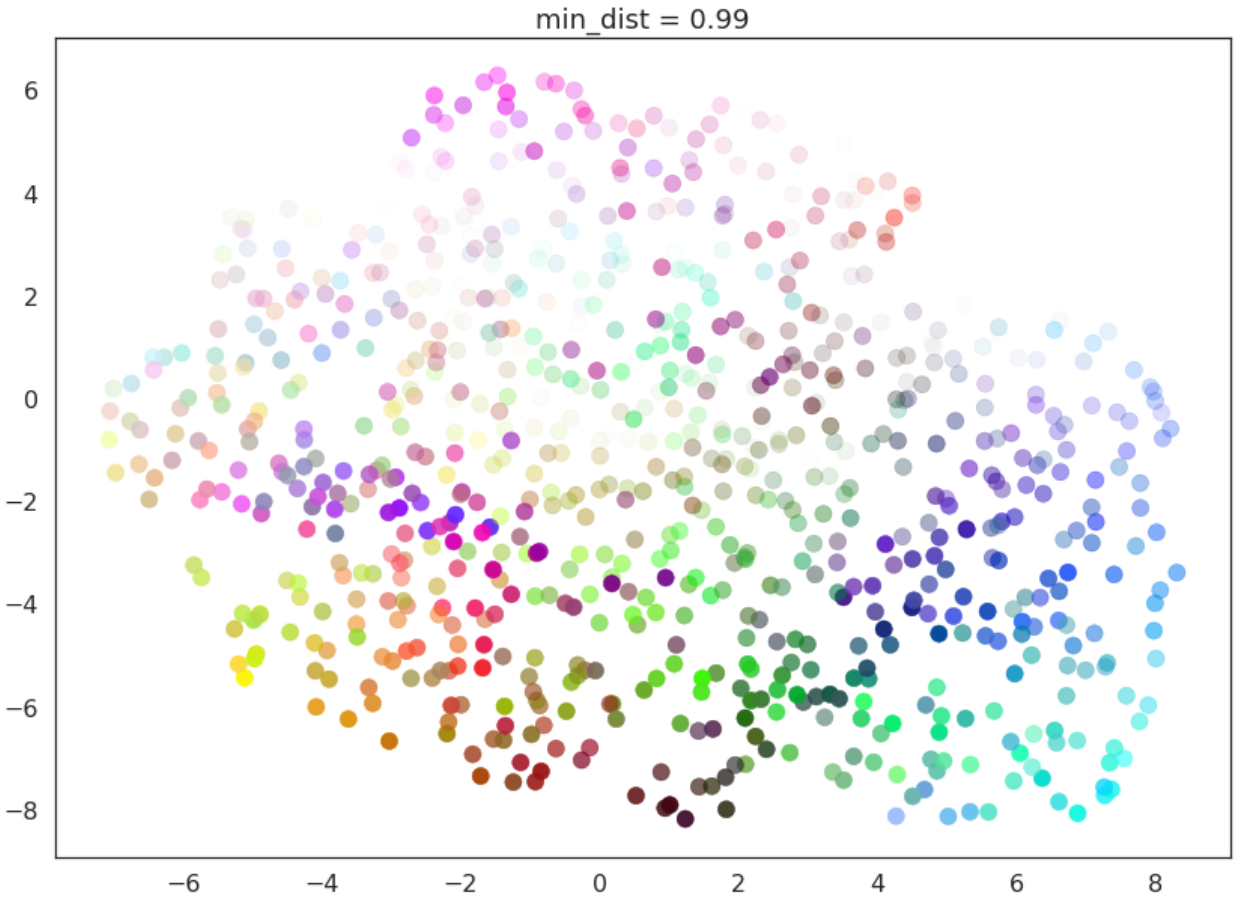



min_dist = 0.1
4
2
0
−2
−4
−4
−2
0
2
4


min_dist = 0.25
4
2
0
−2
−4
−6
−4
−2
0
2
4
6


min_dist = 0.5
4
2
0
−2
−4
−6
−4
−2
0
2
4
6


min_dist = 0.8
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10




Here we see that with `min_dist=0.0` UMAP manages to find small connected components, clumps and strings in the data, and emphasises these features in the resulting embedding. As `min_dist` is increased these structures are pushed apart into softer more general features, providing a better overarching view of the data at the loss of the more detailed topological structure.

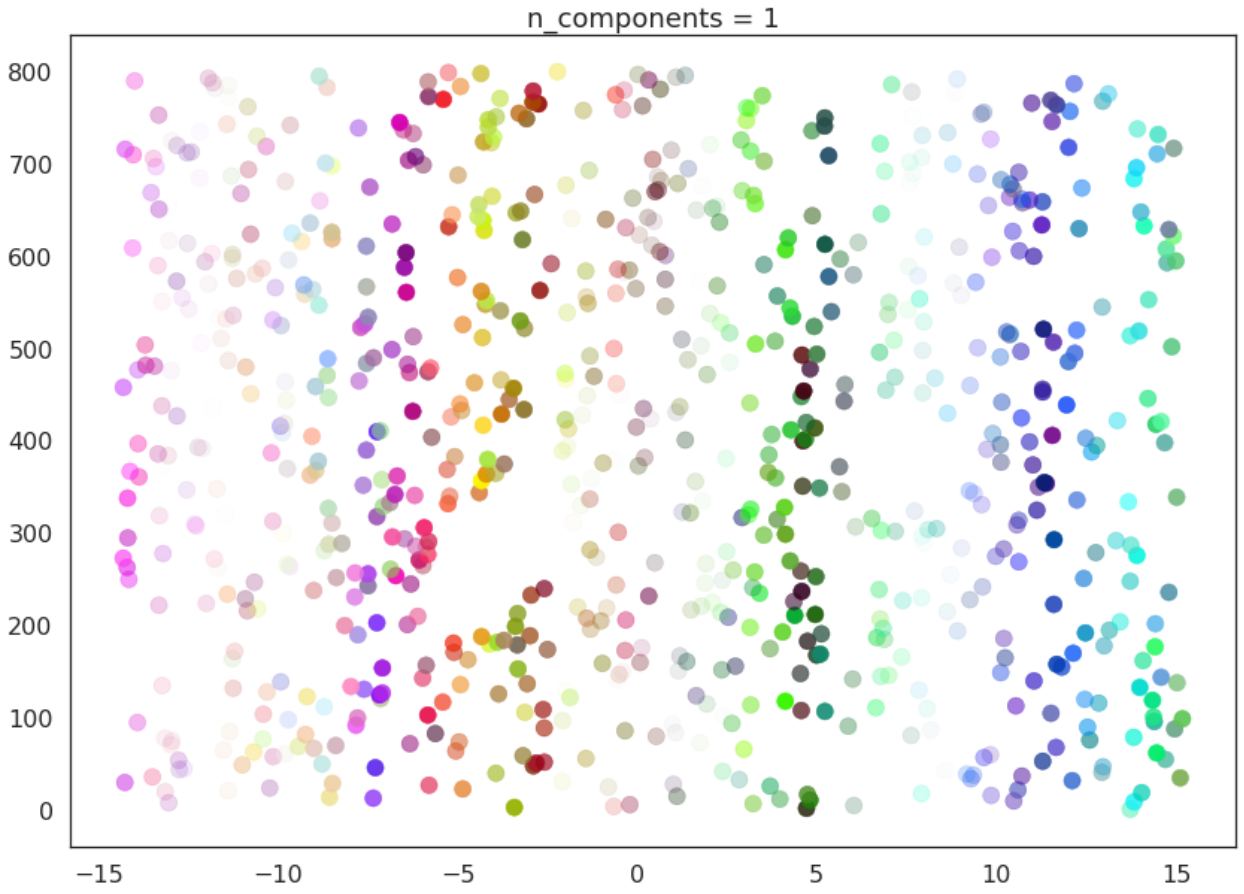
## 2.3 n_components

As is standard for many `scikit-learn` dimension reduction algorithms UMAP provides a `n_components` parameter option that allows the user to determine the dimensionality of the reduced dimension space we will be embedding the data into. Unlike some other visualisation algorithms such as t-SNE UMAP scales well in embedding dimension, so you can use it for more than just visualisation in 2- or 3-dimensions.

For the purposes of this demonstration (so that we can see the effects of the parameter) we will only be looking at 1-dimensional and 3-dimensional embeddings, which we have some hope of visualizing.

First of all we will set `n_components` to 1, forcing UMAP to embed the data in a line. For visualisation purposes we will randomly distribute the data on the y-axis to provide some separation between points.

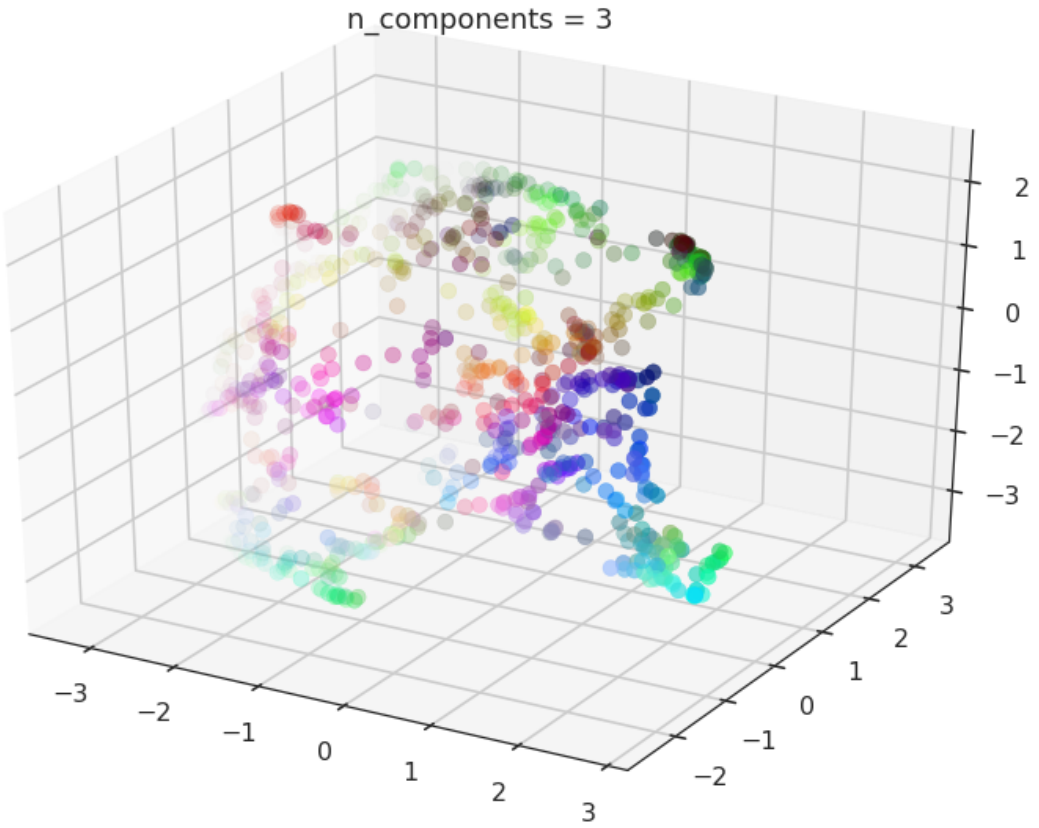
```
draw_umap(n_components=1, title='n_components = 1')
```




Now we will try `n_components=3`. For visualisation we will make use of `matplotlib`'s basic 3-dimensional plotting.

```
draw_umap(n_components=3, title='n_components = 3')
```

```
/opt/anaconda3/envs/umap_dev/lib/python3.6/site-packages/sklearn/metrics/pairwise.
↪py:257: RuntimeWarning: invalid value encountered in sqrt
  return distances if squared else np.sqrt(distances, out=distances)
```




Here we can see that with more dimensions in which to work UMAP has an easier time separating out the colors in a way that respects the topological structure of the data.

As mentioned, there is really no requirement to stop at `n_components` at 3. If you are interested in (density based) clustering, or other machine learning techniques, it can be beneficial to pick a larger embedding dimension (say 10, or 50) closer to the the dimension of the underlying manifold on which your data lies.

## 2.4 `metric`

The final UMAP parameter we will be considering in this notebook is the `metric` parameter. This controls how distance is computed in the ambient space of the input data. By default UMAP supports a wide variety of metrics, including:

**Minkowski style metrics**

- euclidean
- manhattan
- chebyshev
- minkowski

**Miscellaneous spatial metrics**

- canberra
- braycurtis

- haversine

**Normalized spatial metrics**

- mahalanobis
- wminkowski
- seuclidean

**Angular and correlation metrics**

- cosine
- correlation

**Metrics for binary data**

- hamming
- jaccard
- dice
- russellrao
- kulsinski
- rogerstanimoto
- sokalmichener
- sokalsneath
- yule

Any of which can be specified by setting `metric='<metric name>'`; for example to use cosine distance as the metric you would use `metric='cosine'`.

UMAP offers more than this however – it supports custom user defined metrics as long as those metrics can be compiled in `nopython` mode by numba. For this notebook we will be looking at such custom metrics. To define such metrics we'll need numba ...

```
import numba
```

For our first custom metric we'll define the distance to be the absolute value of difference in the red channel.

```
@numba.njit()
def red_channel_dist(a,b):
    return np.abs(a[0] - b[0])
```

To get more adventurous it will be useful to have some colorspace conversion – to keep things simple we'll just use HSL formulas to extract the hue, saturation, and lightness from an (R,G,B) tuple.

```
@numba.njit()
def hue(r, g, b):
    cmax = max(r, g, b)
    cmin = min(r, g, b)
    delta = cmax - cmin
    if cmax == r:
        return ((g - b) / delta) % 6
    elif cmax == g:
        return ((b - r) / delta) + 2
    else:
```

(continues on next page)

(continued from previous page)

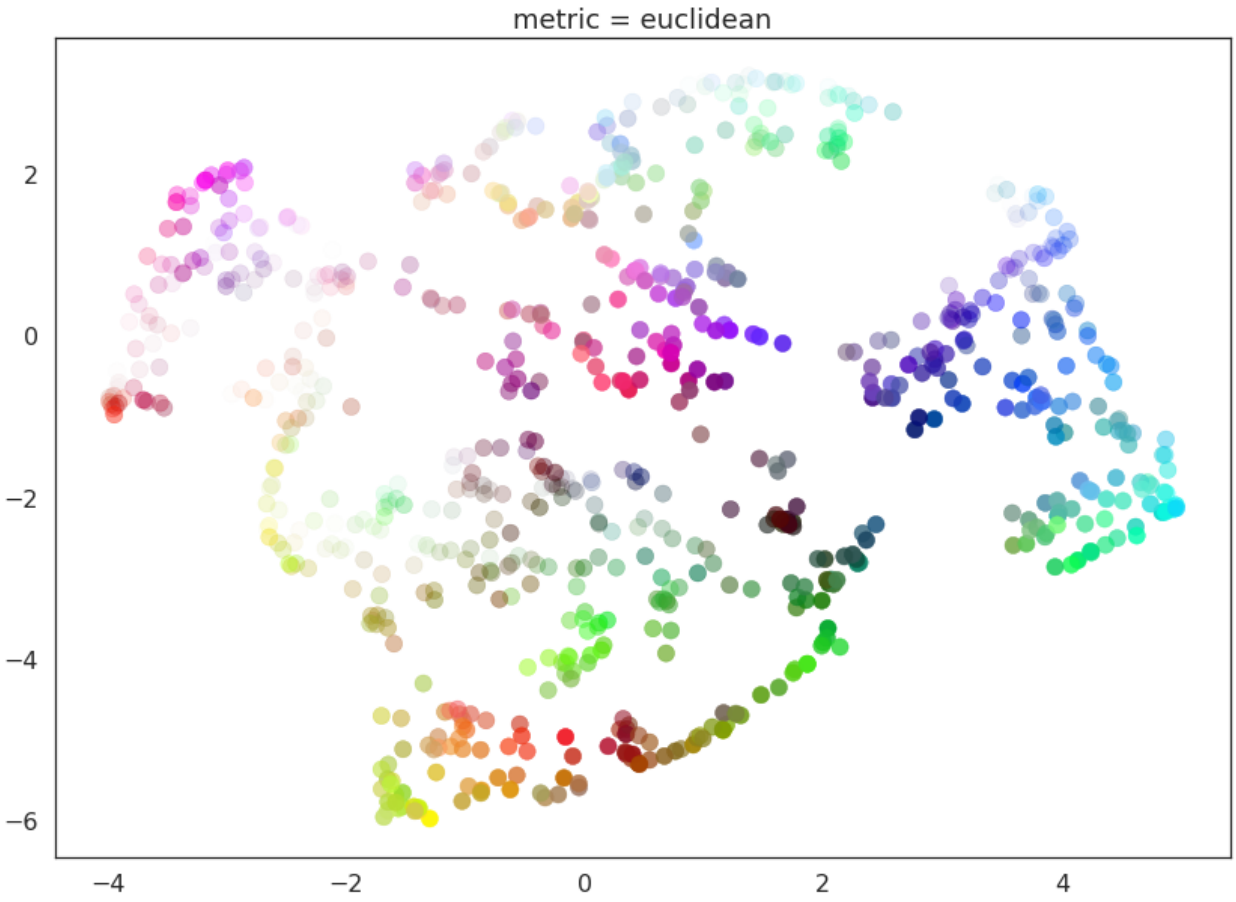
```
        return ((r - g) / delta) + 4

@numba.njit()
def lightness(r, g, b):
    cmax = max(r, g, b)
    cmin = min(r, g, b)
    return (cmax + cmin) / 2.0

@numba.njit()
def saturation(r, g, b):
    cmax = max(r, g, b)
    cmin = min(r, g, b)
    chroma = cmax - cmin
    light = lightness(r, g, b)
    if light == 1:
        return 0
    else:
        return chroma / (1 - abs(2*light - 1))
```

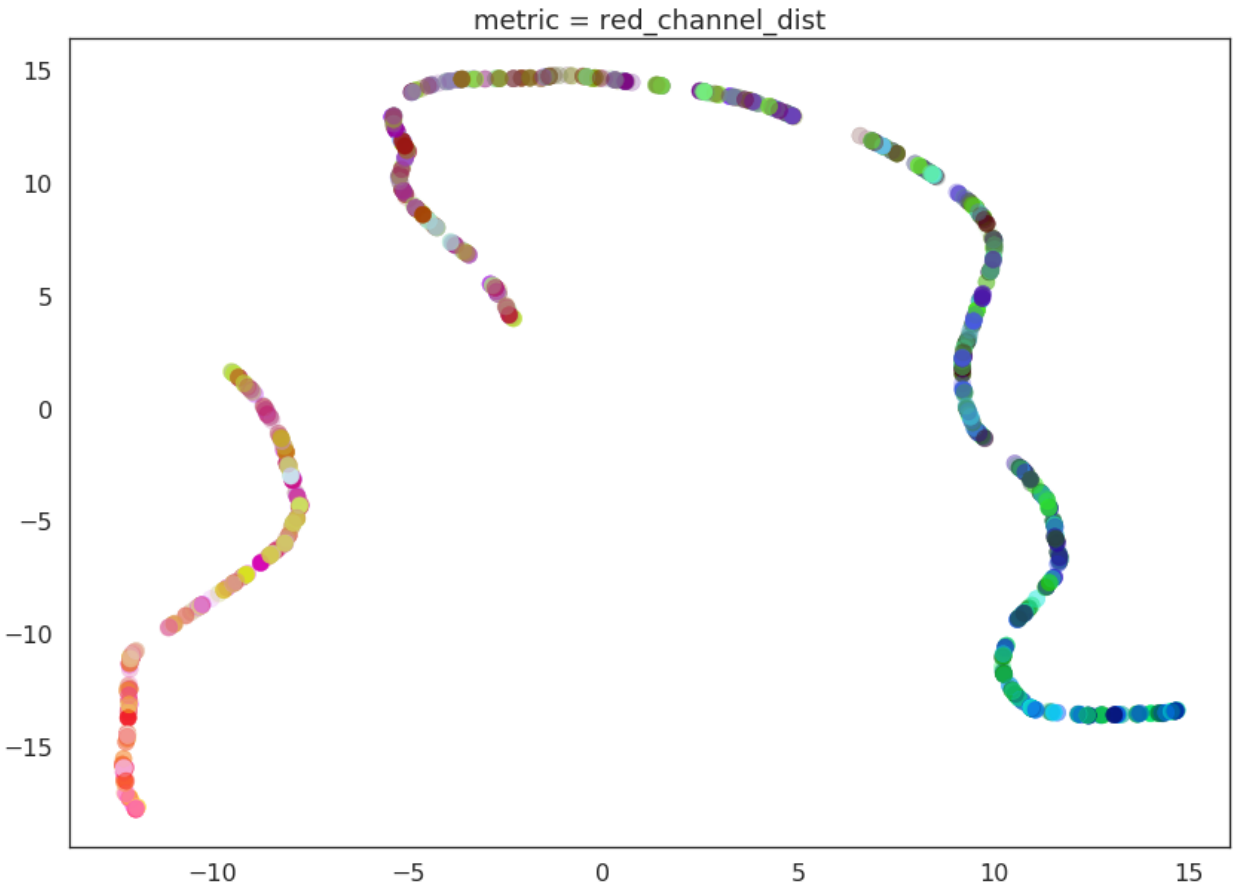
With that in hand we can define three extra distances. The first simply measures the difference in hue, the second measures the euclidean distance in a combined saturation and lightness space, while the third measures distance in the full HSL space.

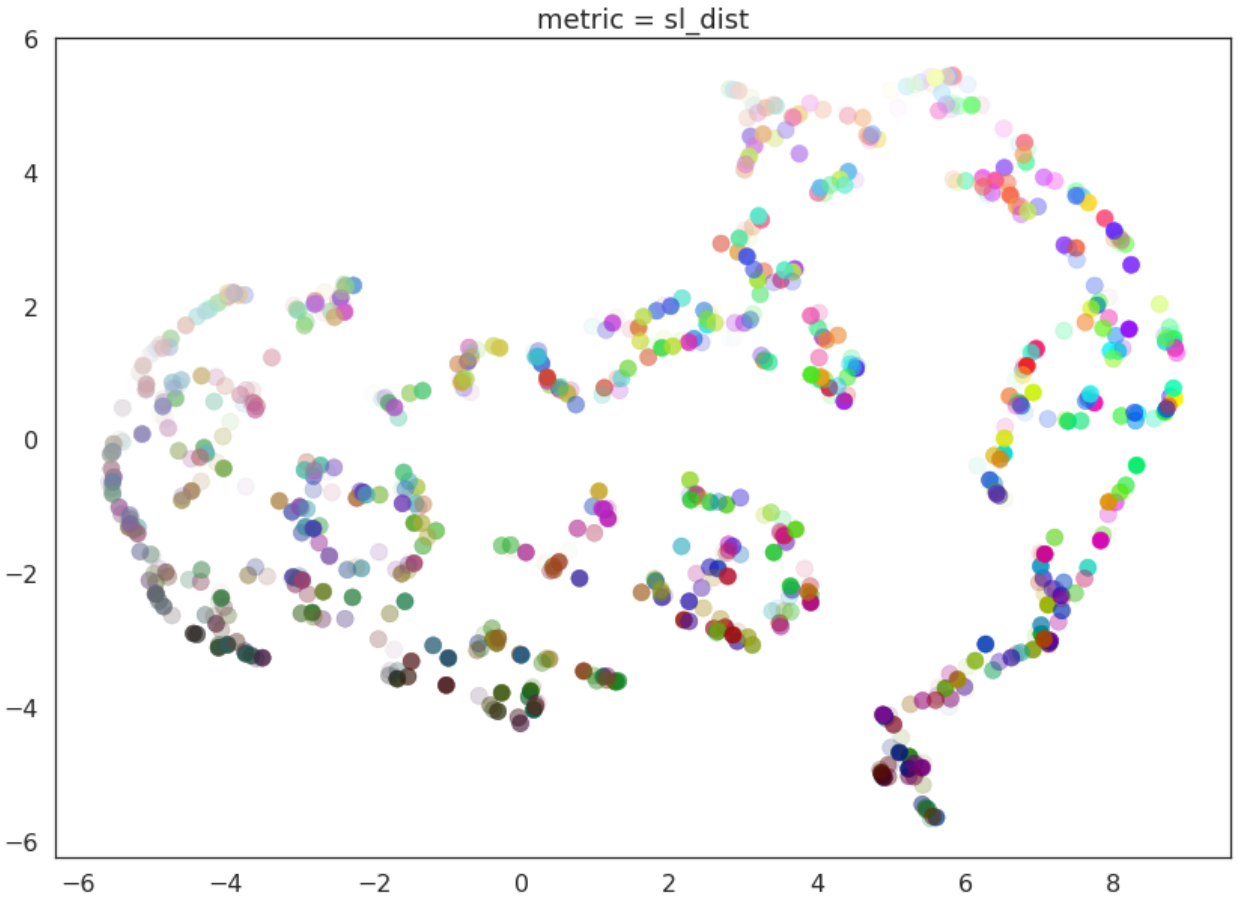
```
@numba.njit()
def hue_dist(a, b):
    diff = (hue(a[0], a[1], a[2]) - hue(b[0], b[1], b[2])) % 6
    if diff < 0:
        return diff + 6
    else:
        return diff

@numba.njit()
def sl_dist(a, b):
    a_sat = saturation(a[0], a[1], a[2])
    b_sat = saturation(b[0], b[1], b[2])
    a_light = lightness(a[0], a[1], a[2])
    b_light = lightness(b[0], b[1], b[2])
    return (a_sat - b_sat)**2 + (a_light - b_light)**2

@numba.njit()
def hsl_dist(a, b):
    a_sat = saturation(a[0], a[1], a[2])
    b_sat = saturation(b[0], b[1], b[2])
    a_light = lightness(a[0], a[1], a[2])
    b_light = lightness(b[0], b[1], b[2])
    a_hue = hue(a[0], a[1], a[2])
    b_hue = hue(b[0], b[1], b[2])
    return (a_sat - b_sat)**2 + (a_light - b_light)**2 + (((a_hue - b_hue) % 6) / 6.0)
```

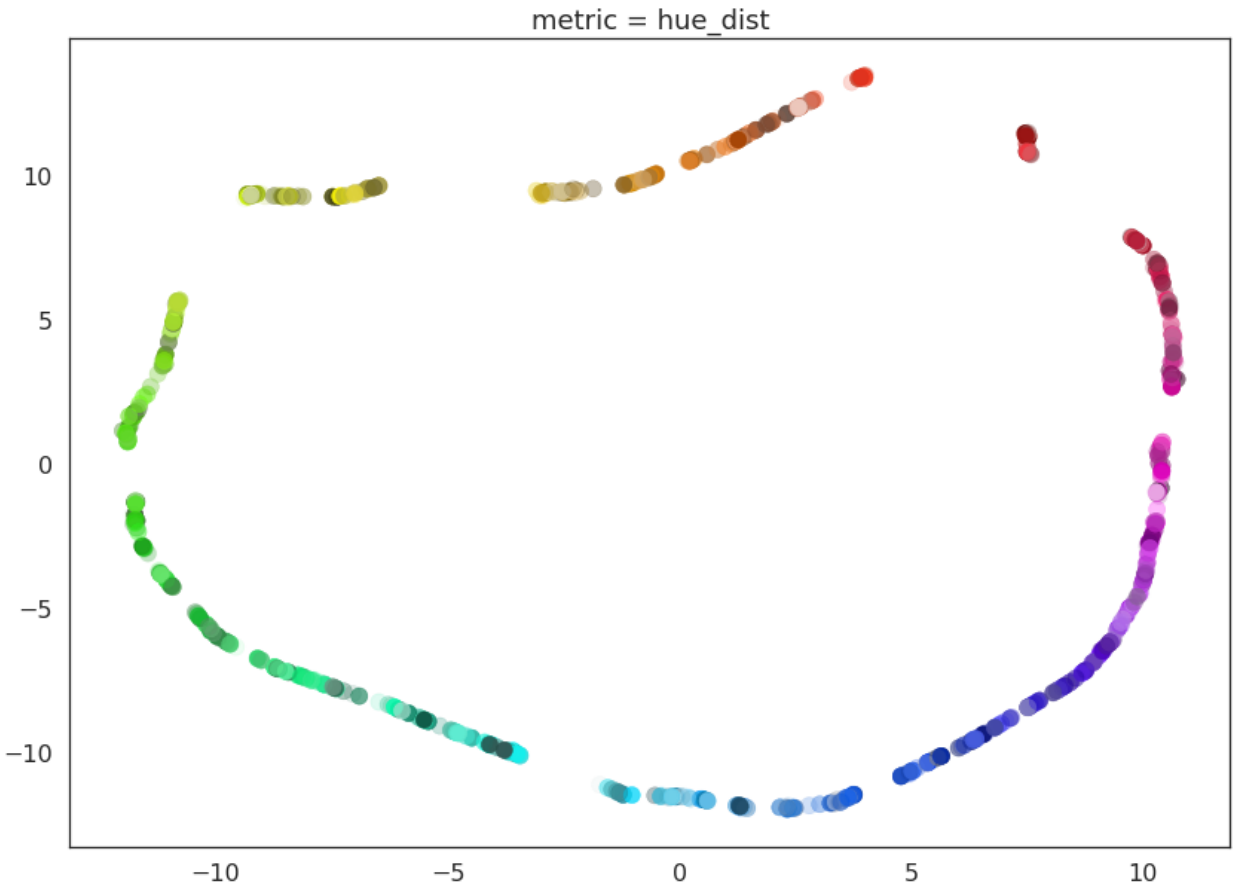
With such custom metrics in hand we can get UMAP to embed the data using those metrics to measure distance between our input data points. Note that `numba` provides significant flexibility in what we can do in defining distance functions. Despite this we retain the high performance we expect from UMAP even using such custom functions.

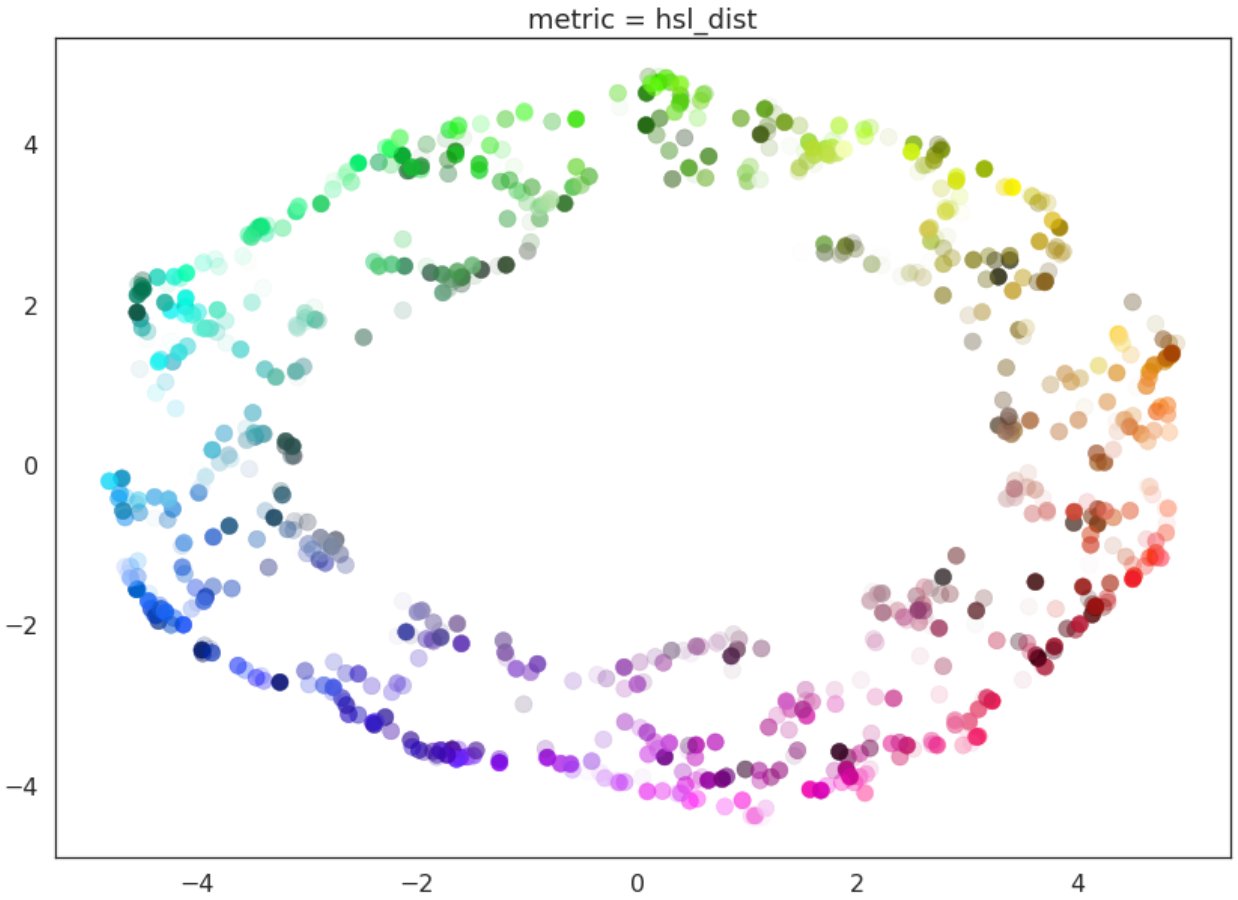
```
for m in ("euclidean", red_channel_dist, sl_dist, hue_dist, hsl_dist):
    name = m if type(m) is str else m.__name__
    draw_umap(n_components=2, metric=m, title='metric = {}'.format(name))
```


metric = euclidean
2
0
−2
−4
−6
−4
−2
0
2
4


metric = red_channel_dist
15
10
5
0
−5
−10
−15
−10
−5
0
5
10
15


metric = sl_dist
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6
8


metric = hue_dist
10
5
0
−5
−10
−10
−5
0
5
10




And here we can see the effects of the metrics quite clearly. The pure red channel correctly see the data as living on a one dimensional manifold, the hue metric interprets the data as living in a circle, and the HSL metric fattens out the circle according to the saturation and lightness. This provides a reasonable demonstration of the power and flexibility of UMAP in understanding the underlying topology of data, and finding a suitable low dimensional representation of that topology.

# CHAPTER 3

# Plotting UMAP results

UMAP is often used for visualization by reducing data to 2-dimensions. Since this is such a common use case the umap package now includes utility routines to make plotting UMAP results simple, and provide a number of ways to view and diagnose the results. Rather than seeking to provide a comprehensive solution that covers all possible plotting needs this umap extension seeks to provide a simple to use interface to make the majority of plotting needs easy, and help provide sensible plotting choices wherever possible. To get started looking at the plotting options let's load a variety of data to work with.

```
import sklearn.datasets
import pandas as pd
import numpy as np
import umap
```

```
pendigits = sklearn.datasets.load_digits()
mnist = sklearn.datasets.fetch_openml('mnist_784')
fmnist = sklearn.datasets.fetch_openml('Fashion-MNIST')
```

To start we will fit a UMAP model to the pendigits data. This is as simple as running the fit method and assigning the result to a variable.

```
mapper = umap.UMAP().fit(pendigits.data)
```

If we want to do plotting we will need the `umap.plot` package. While the umap package has a fairly small set of requirements it is worth noting that if you want to using `umap.plot` you will need a variety of extra libraries that are not in the default requirements for umap. In particular you will need:

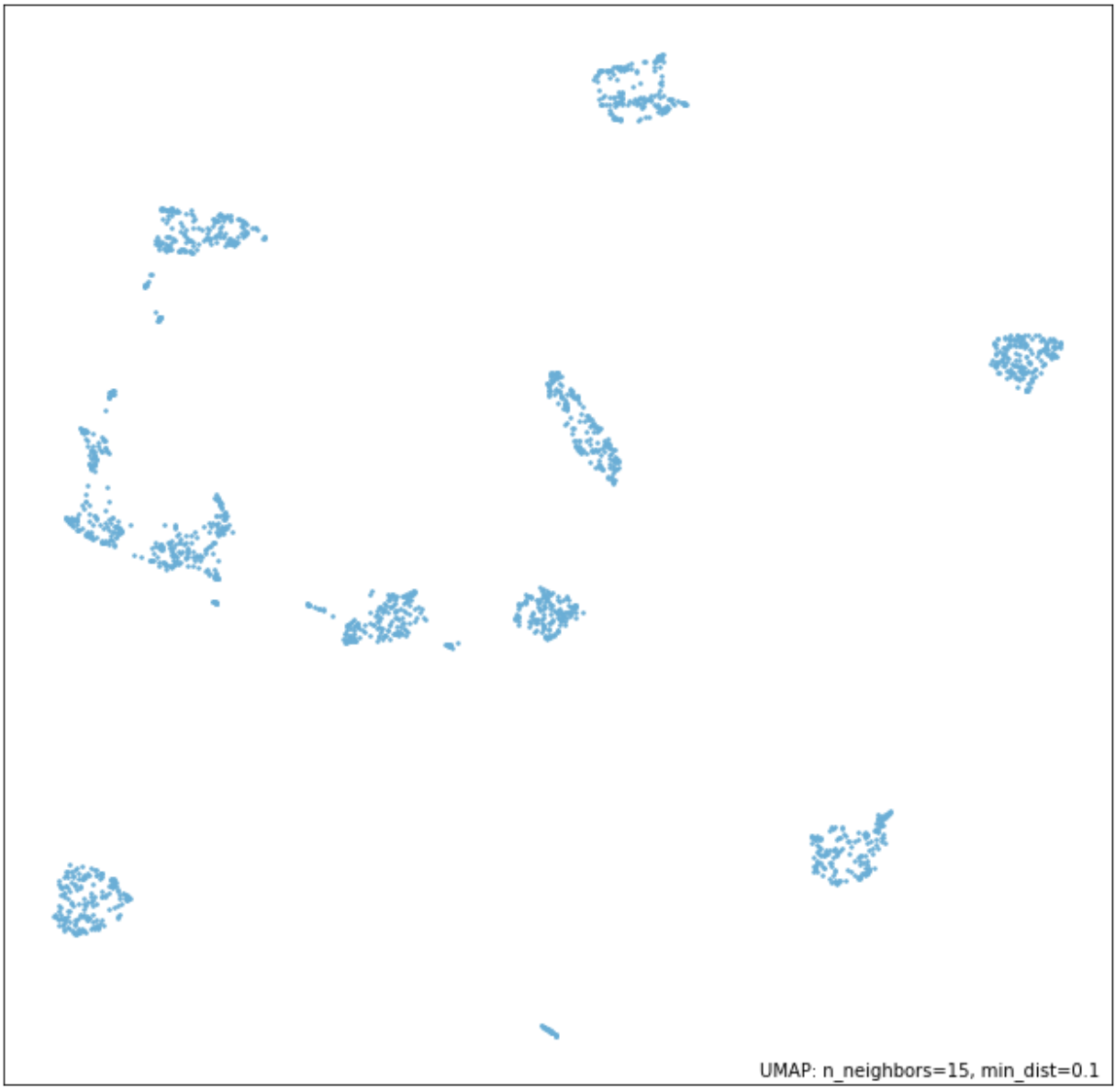
- matplotlib
- pandas
- datashader
- bokeh
- holoviews

All should be either pip or conda installable. With those in hand you can import the `umap.plot` package.

```
import umap.plot
```

Now that we have the package loaded, how do we use it? The most straightforward thing to do is plot the umap results as points. We can achieve this via the function `umap.plot.points`. In its most basic form you can simply pass the trained UMAP model to `umap.plot.points`:

```
umap.plot.points(mapper)
```

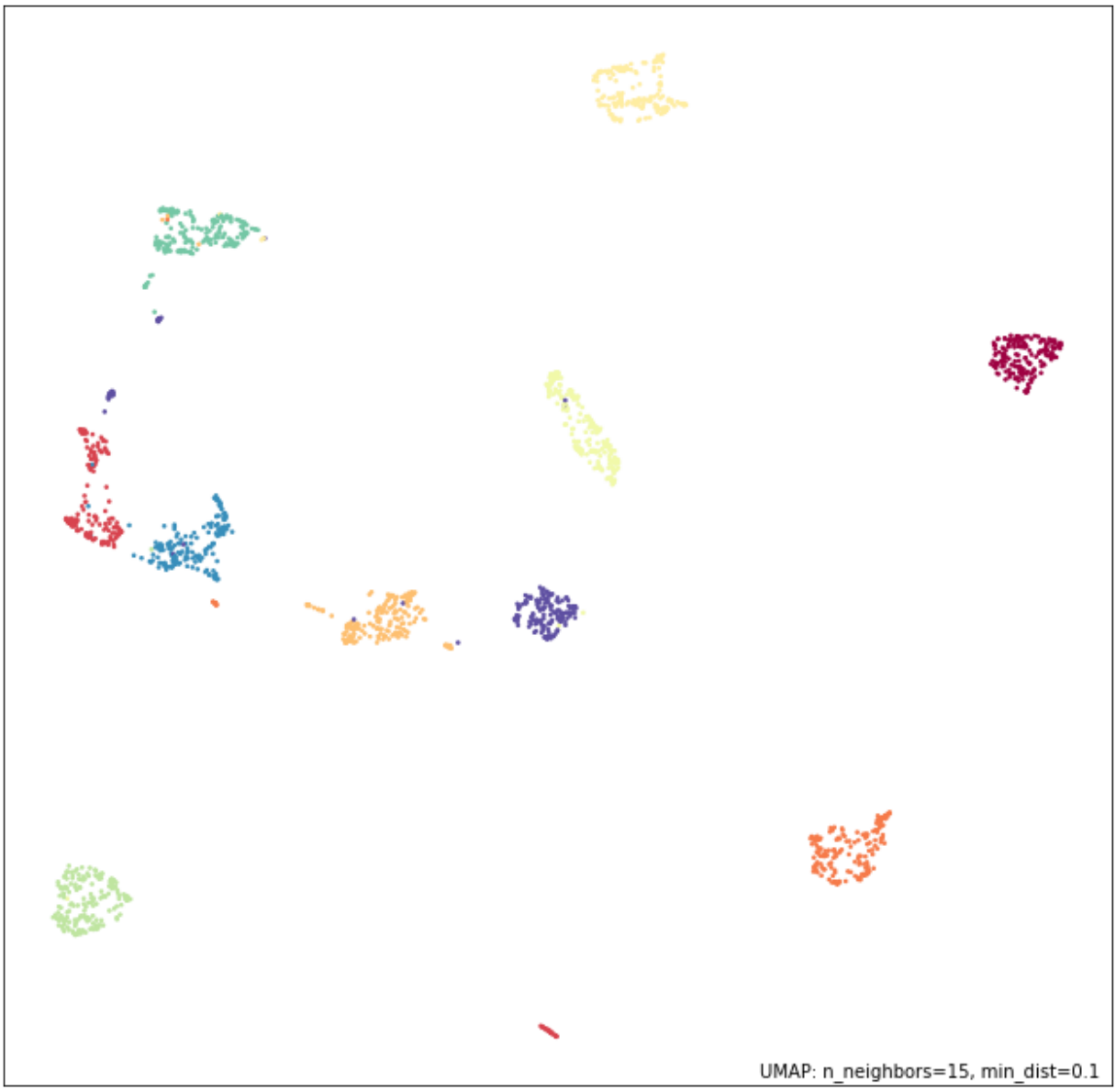



As you can see we immediately get a scatterplot of the UMAP embedding. Of note the function automatically selects a point-size based on the data density, and watermarks the image with the UMAP parameters that were used (this will include the metric if it is non-standard). The function also returns the matplotlib axes object associated to the plot, so further matplotlib functions, such as adding titles, axis labels etc. can be done by the user if required.

It is common for data passed to UMAP to have an associated set of labels, which may have been derived from ground-truth, from clustering, or via other means. In such cases it is desirable to be able to color the scatterplot according

to the labelling. We can do this by simply passing the array of label information in with the `labels` keyword. The `umap.plot.points` function will the color the data with a categorical colormap according to the labels provided.

```
umap.plot.points(mapper, labels=pendigits.target)
```




Alternatively you may have extra data that is continuous rather than categorical. In this case you will want to use a continuous colormap to shade the data. Again this is straightforward to do – pass in the continuous data with the `values` keyword and data will be colored accordingly using a continuous colormap.

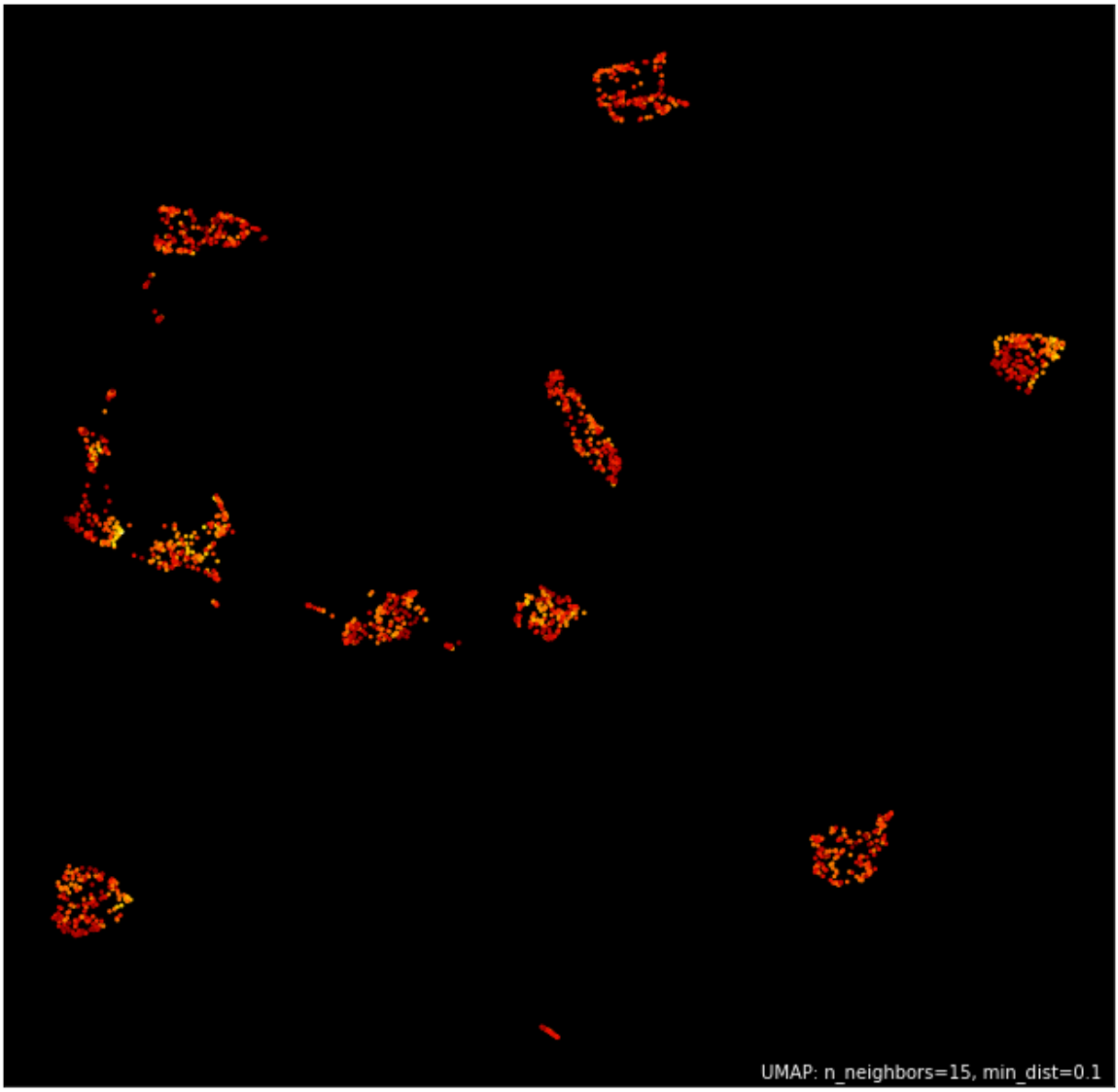
Furthermore, if you don't like the default color choices the `umap.plot.points` function offers a number of 'themes' that provide predefined color choices. Themes include:

- fire
- viridis
- inferno

- blue
- red
- green
- darkblue
- darkred
- darkgreen

Here we will make use of the 'fire' theme to demonstrate how simple it is to change the aesthetics.

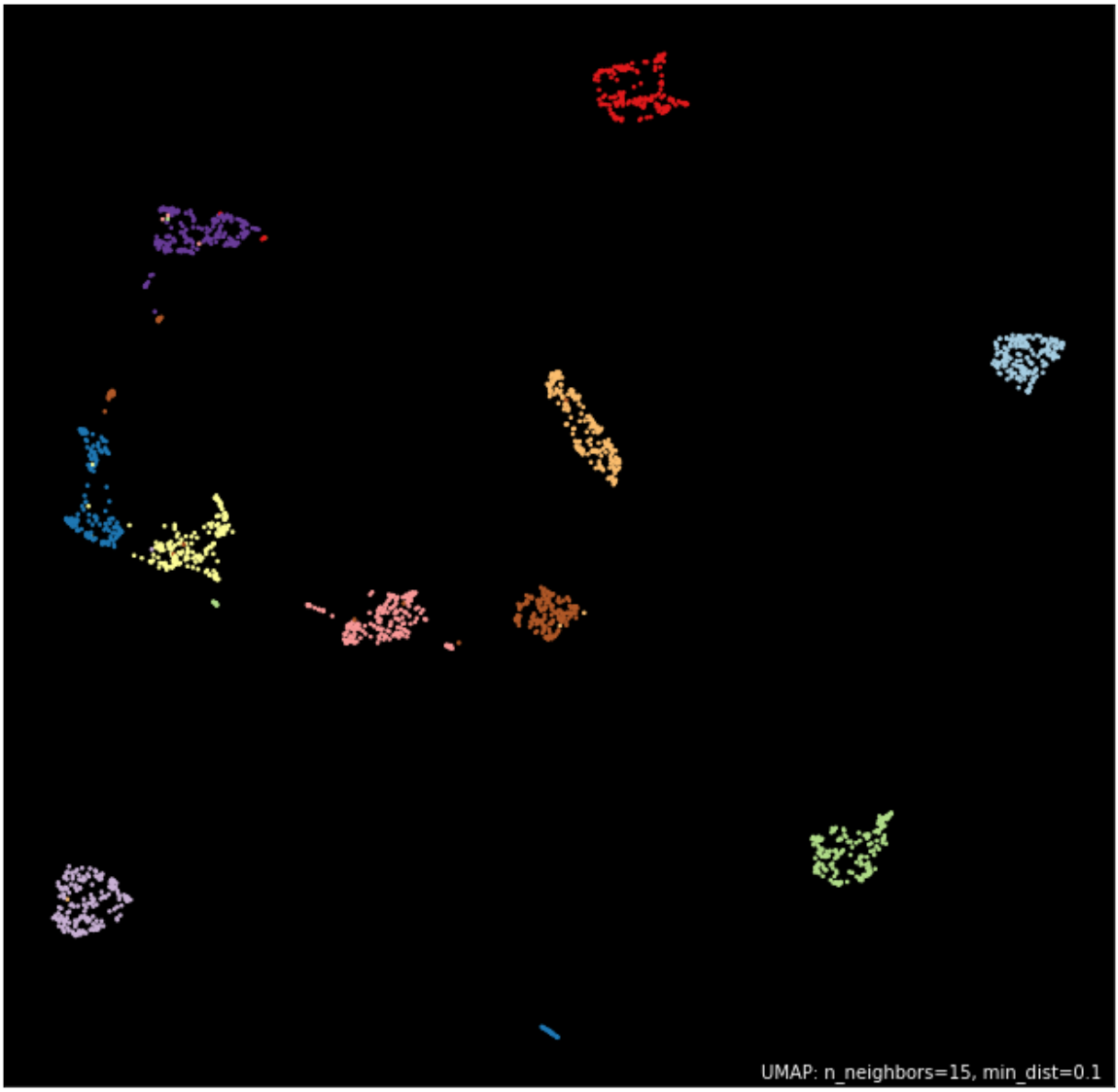
```
umap.plot.points(mapper, values=pendigits.data.mean(axis=1), theme='fire')
```




If you want greater control you can specify exact colormaps and background colors. For example here we want to color the data by label, but use a black background and use the 'Paired' colormap for the categorical coloring (passed

as `color_key_cmap`; the `cmap` keyword defines the continuous colormap).

```
umap.plot.points(mapper, labels=pendigits.target, color_key_cmap='Paired', background=
↪'black')
```




Many more options are available including a `color_key` to specify a dictionary mapping of discrete labels to colors, `cmap` to specify the continous colormap, or the width and height of the resulting plot. Again, this does not provide comprehensive control of the plot aesthetics, but the goal here is to provide a simple to use interface rather than the ability for the user to fine tune all aspects – users seeking such control are far better served making use of the individual underlying packages (matplotlib, datashader, and bokeh) by themselves.

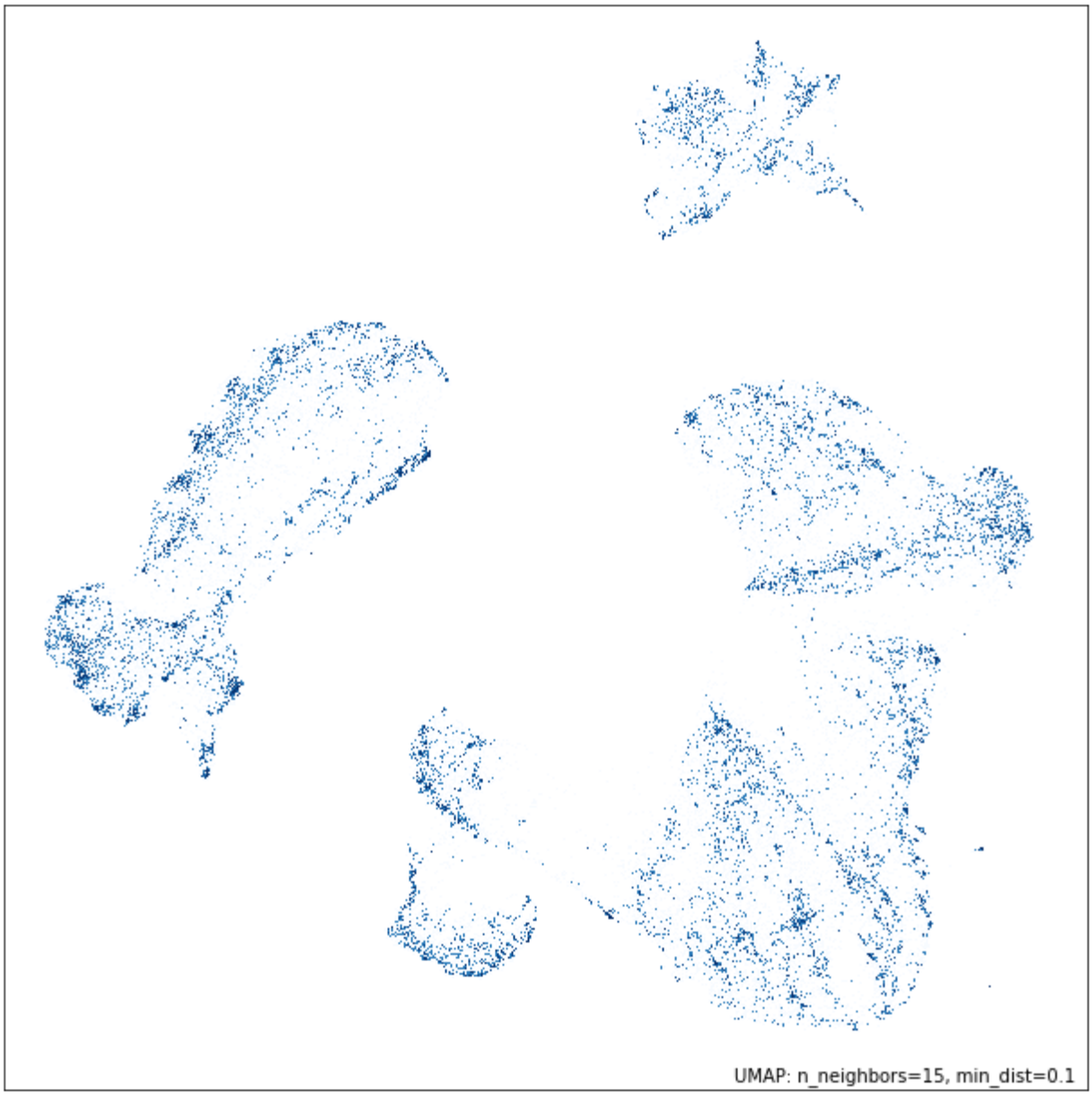
## 3.1 Plotting larger datasets

Once you have a lot of data it becomes easier for a simple scatter plot to lie to you. Most notably overplotting, where markers for points overlap and pile up on top of each other, can deceive you into thinking that extremely dense clumps may only contain a few points. While there are things that can be done to help remedy this, such as reducing the point size, or adding an alpha channel, few are sufficient to be sure the plot isn't subtly lying to you in some way. This essay in the datashader documentation does an excellent job of describing the issues with overplotting, why the obvious solutions are not quite sufficient, and how to get around the problem. To make life easier for users the `umap.plot` package will automatically switch to using datashader for rendering once your dataset gets large enough. This helps to ensure you don't get fooled by overplotting. We can see this in action by working with one of the larger datasets such as Fashion-MNIST.

```
mapper = umap.UMAP().fit(fmnist.data)
```

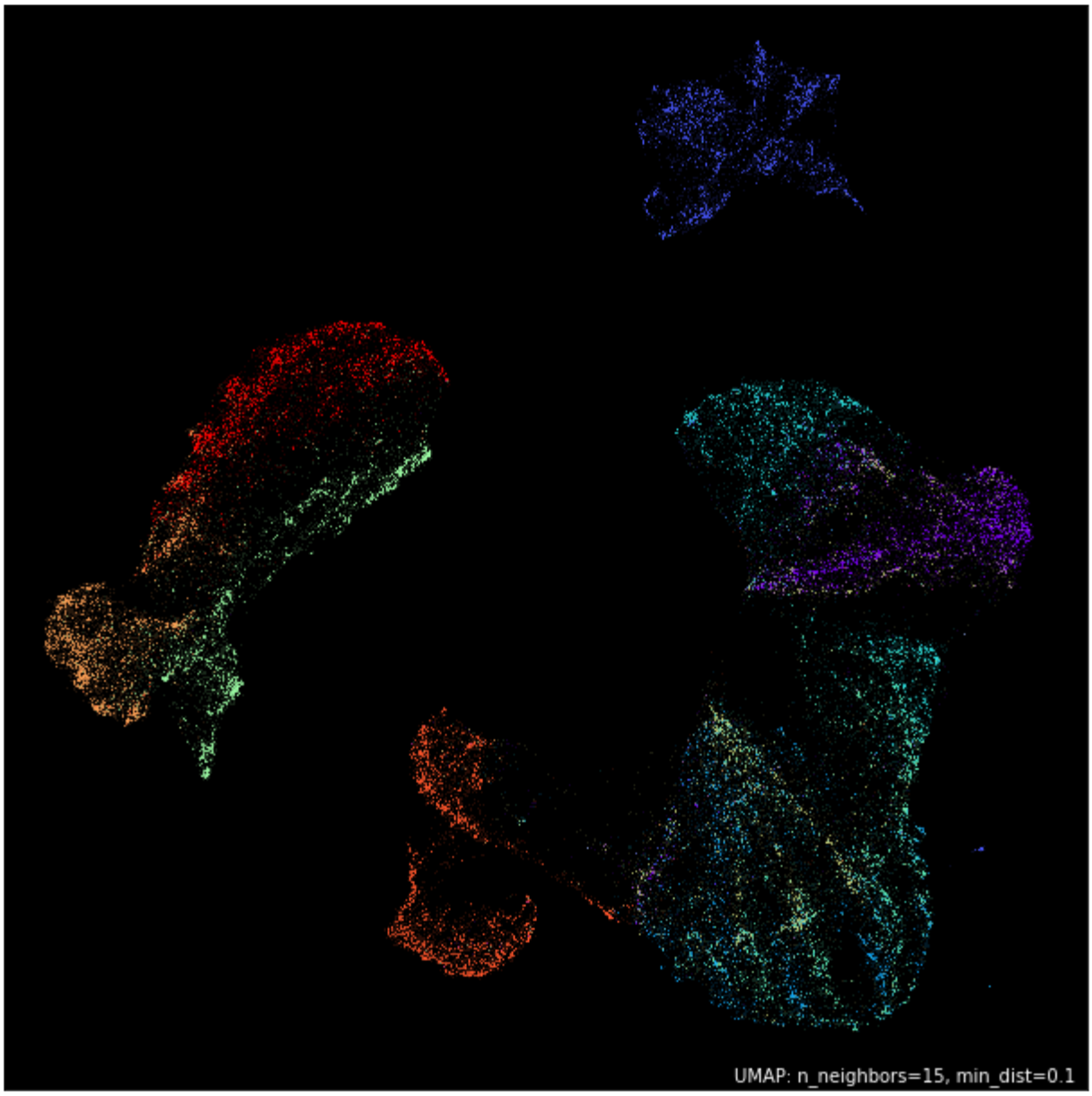
Having fit the data with UMAP we can call `umap.plot.points` exactly as before, but this time, since the data is large enough to have potential overplotting, datashader will be used in the background for rendering.

```
umap.plot.points(mapper)
```




All the same plot options as before hold, so we can color by labels, and apply the same themes, and it will all seamlessly use datashader for the actual rendering. Thus, regardless of how much data you have `umap.plot.points` will render it well with a transparent user interface. You, as a user, don't need to worry about switching to plotting with datashader, or how to convert your plotting to its slightly different API – you can just use the same API and trust the resuts you get.

```
umap.plot.points(mapper, labels=fmnist.target, theme='fire')
```




## 3.2 Interactive plotting, and hover tools

Rendering good looking static plots is important, but what if you want to be able to interact with your data – pan around, and zoom in on the clusters to see the finer structure? What if you want to annotate your data with more complex labels than merely colors? Wouldn't it be good to be able to hover over data points and get more information about the individual point? Since this is a very common use case `umap.plot` tries to make it easy to quickly generate such plots, and provide basic utilities to allow you to have annotated hover tools working quickly. Again, the goal is not to provide a comprehensive solution that can do everything, but rather a simple to use and consistent API to get users up and running fast.

To make a good example of this let's use a subset of the Fashion MNIST dataset. We can quickly train a new mapper object on that.

```
mapper = umap.UMAP().fit(fmnist.data[:30000])
```

The goal is to be able to hover over different points and see data associated with the given point (or points) under the cursor. For this simple demonstration we'll just use the target information of the point. To create hover information you need to construct a dataframe of all the data you would like to appear in the hover. Each row should correspond to a source of data points (appearing in the same order), and the columns can provide whatever extra data you would like to display in the hover tooltip. In this case we'll need a dataframe that can include the index of the point, its target number, and the actual name of the type of fashion item that target corresponds to. This is easy to quickly put together using pandas.

```
hover_data = pd.DataFrame({'index':np.arange(30000),
                           'label':fmnist.target[:30000]})
hover_data['item'] = hover_data.label.map(
    {
        '0':'T-shirt/top',
        '1':'Trouser',
        '2':'Pullover',
        '3':'Dress',
        '4':'Coat',
        '5':'Sandal',
        '6':'Shirt',
        '7':'Sneaker',
        '8':'Bag',
        '9':'Ankle Boot',
    }
)
```

For interactive use the `umap.plot` package makes use of bokeh. Bokeh has several output methods, but in the approach we'll be outputting inline in a notebook. We have to enable this using the `output_notebook` function. Alteratively we could use `output_file` or other similar options – see the bokeh documentation for more details.

```
umap.plot.output_notebook()
```

Now we can make an interactive plot using `umap.plot.interactive`. This has a very similar API to the `umap.plot.points` approach, but also supports a `hover_data` keyword which, if passed a suitable dataframe, will provide hover tooltips in the interactive plot. Since bokeh allows different outputs, to display it in the notebook we will have to take the extra stop of calling `show` on the result.

```
p = umap.plot.interactive(mapper, labels=fmnist.target[:30000], hover_data=hover_data,
↪ point_size=2)
umap.plot.show(p)
```

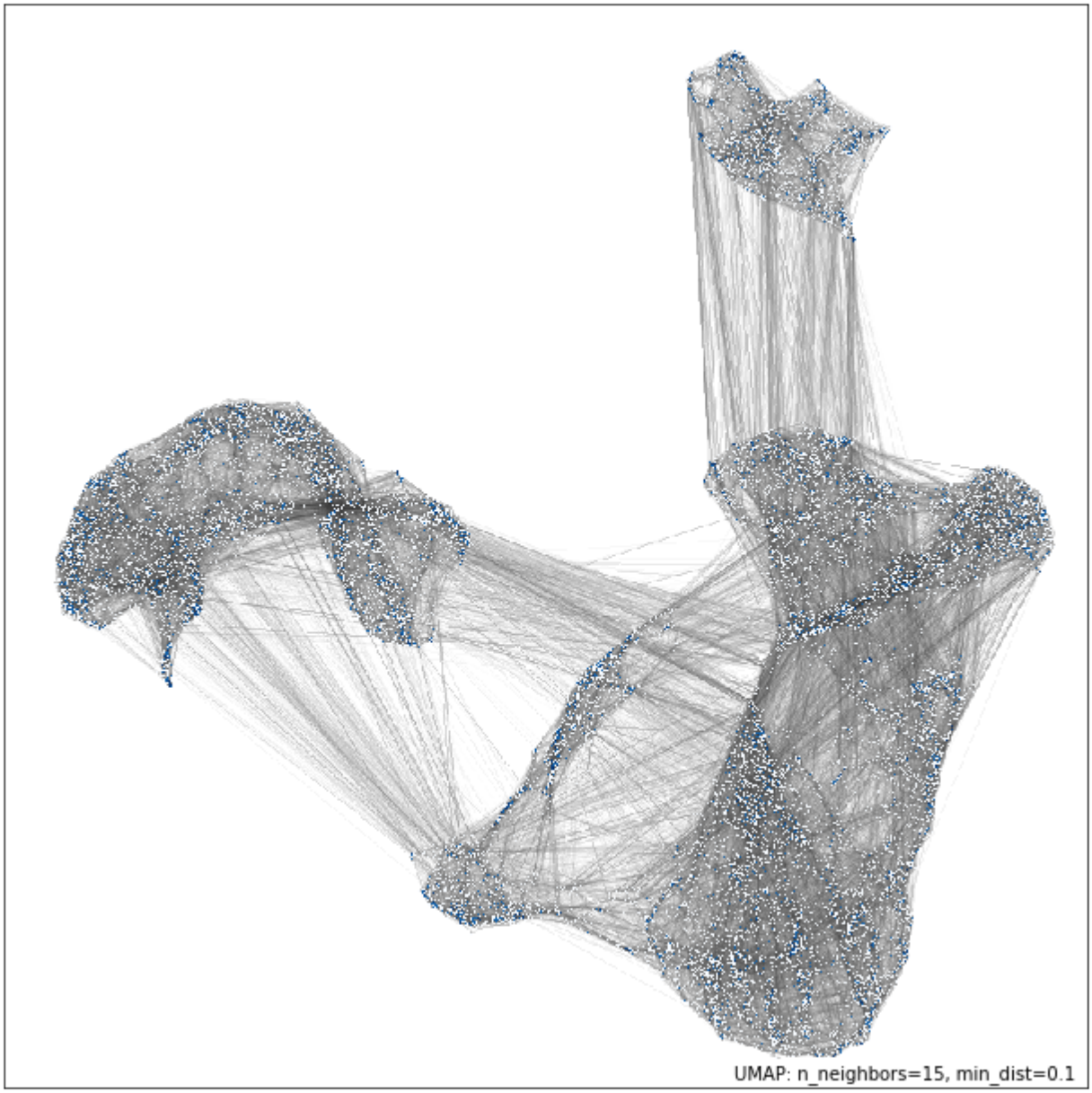
We get the sort of result one would like – a fully interactive plot that can be zoomed in on, and more, but we also now have an interactive hover tool which presents the data from the dataframe we constructed. This allows a quick and easy method to get up and running with a richer interactive exploration of your UMAP plot. `umap.plot.interactive` supports all the same aesthetic parameters as `umap.plot.points` so you can theme your plot, color by label or value, and other similar operations explained above for `umap.plot.points`.

## 3.3 Plotting connectivity

UMAP works by constructing an intermediate topological representation of the approximate manifold the data may have been sampled from. In practice this structure can be simplified down to a weighted graph. Sometimes it can be beneficial to see how that graph (representing connectivity in the manifold) looks with respect to the resulting embedding. It can be used to better understand the embedding, and for diagnostic purposes. To see the connectivity

you can use the `umap.plot.connectivity` function. It works very similarly to the `umap.plot.points` function, and has the option as to whether to display the embedding point, or just the connectivity. To start let's do a simple plot showing the points:

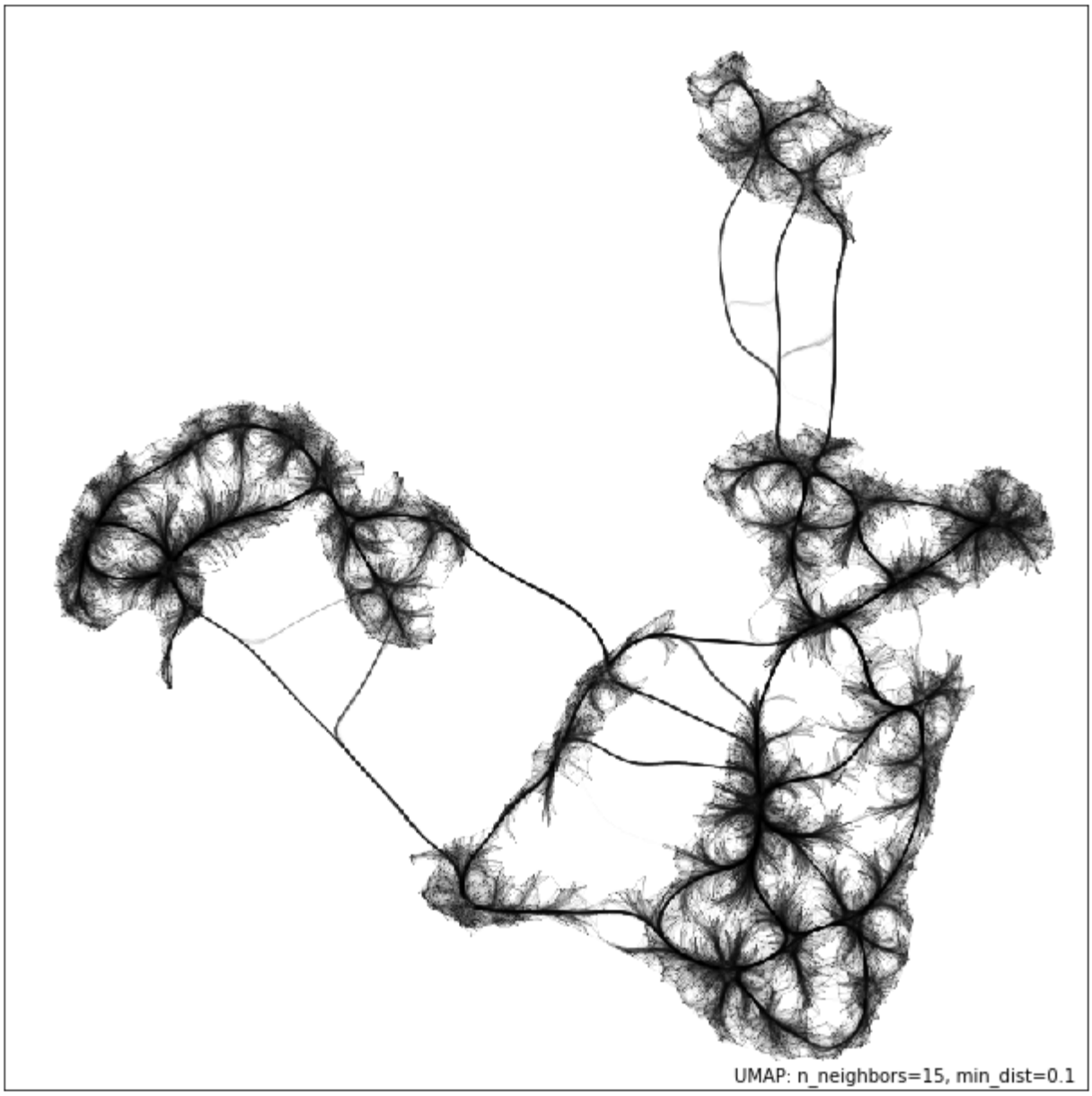
```
umap.plot.connectivity(mapper, show_points=True)
```




As with `umap.plot.points` there are options to control the basic aesthetics, including theme options and an `edge_cmap` keyword argument to specify the colormap used for displaying the edges.

Since this approach already leverages datashader for edge plotting, we can go a step further and make use of the edge-bundling options available in datashader. This can provide a less busy view of connectivity, but can be expensive to compute, particularly for larger datasets.

```
umap.plot.connectivity(mapper, edge_bundling='hammer')
```




## 3.4 Diagnostic plotting

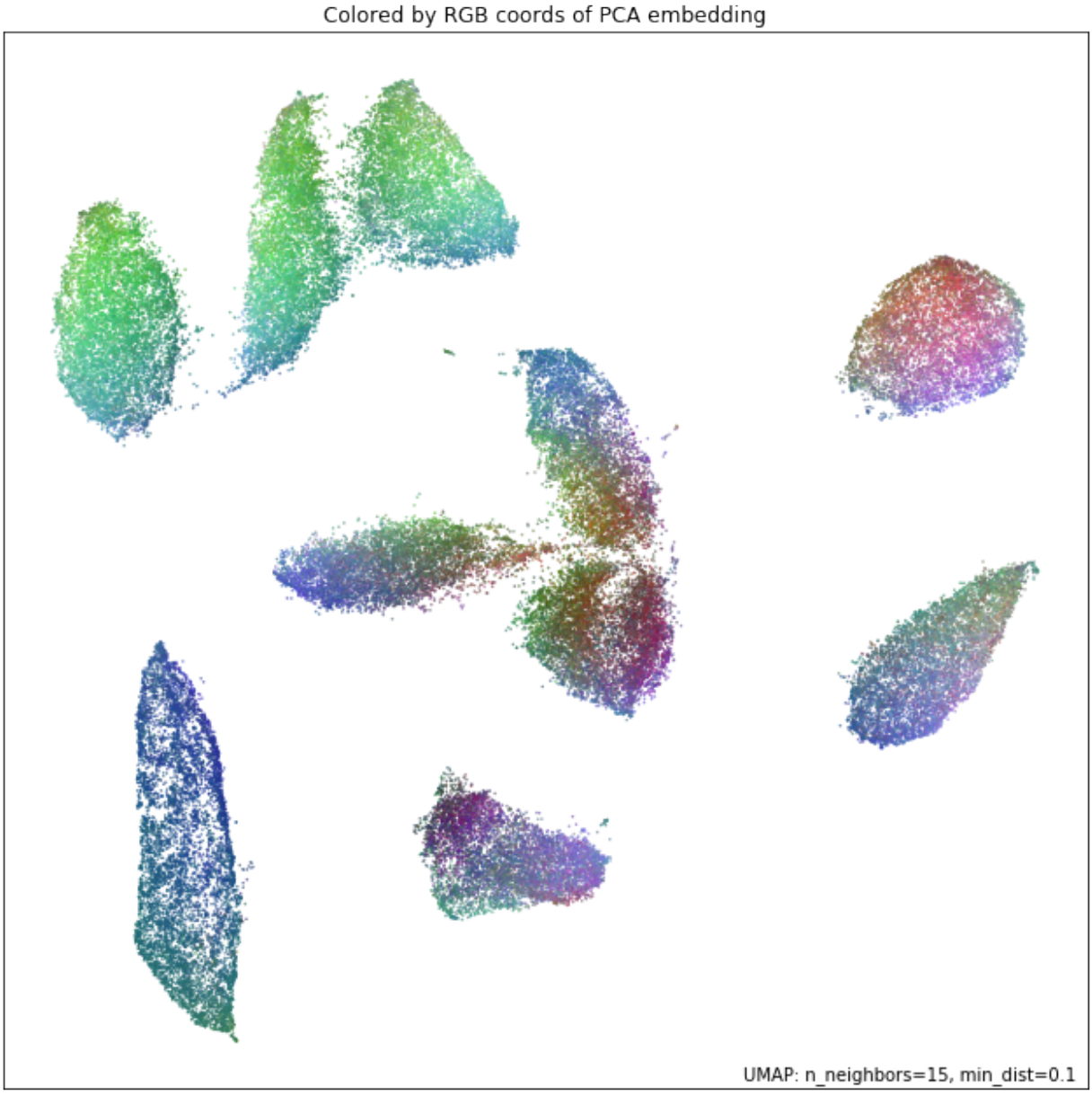
Plotting the connectivity provides at least one basic diagnostic view that helps a user understand what is going on with an embedding. More views on data are better, of course, so `umap.plot` includes a `umap.plot.diagnostic` function that can provide various diagnostic plots. We'll look at a few of them here. To do so we'll use the full MNIST digits data set.

```
mapper = umap.UMAP().fit(mnist.data)
```

The first diagnostic type is a Principal Components Analysis based diagnostic, which you can select with `diagnostic_type='pca'`. The essence of the approach is that we can use PCA, which preserves global structure, to reduce the data to three dimensions. If we scale the results to fit in a 3D cube we can convert the 3D PCA

coordinates of each point into an RGB description of a color. By then coloring the points in the UMAP embedding with the colors induced by the PCA it is possible to get a sense of how some of the more large scale global structure has been represented in the embedding.

```
umap.plot.diagnostic(mapper, diagnostic_type='pca')
```

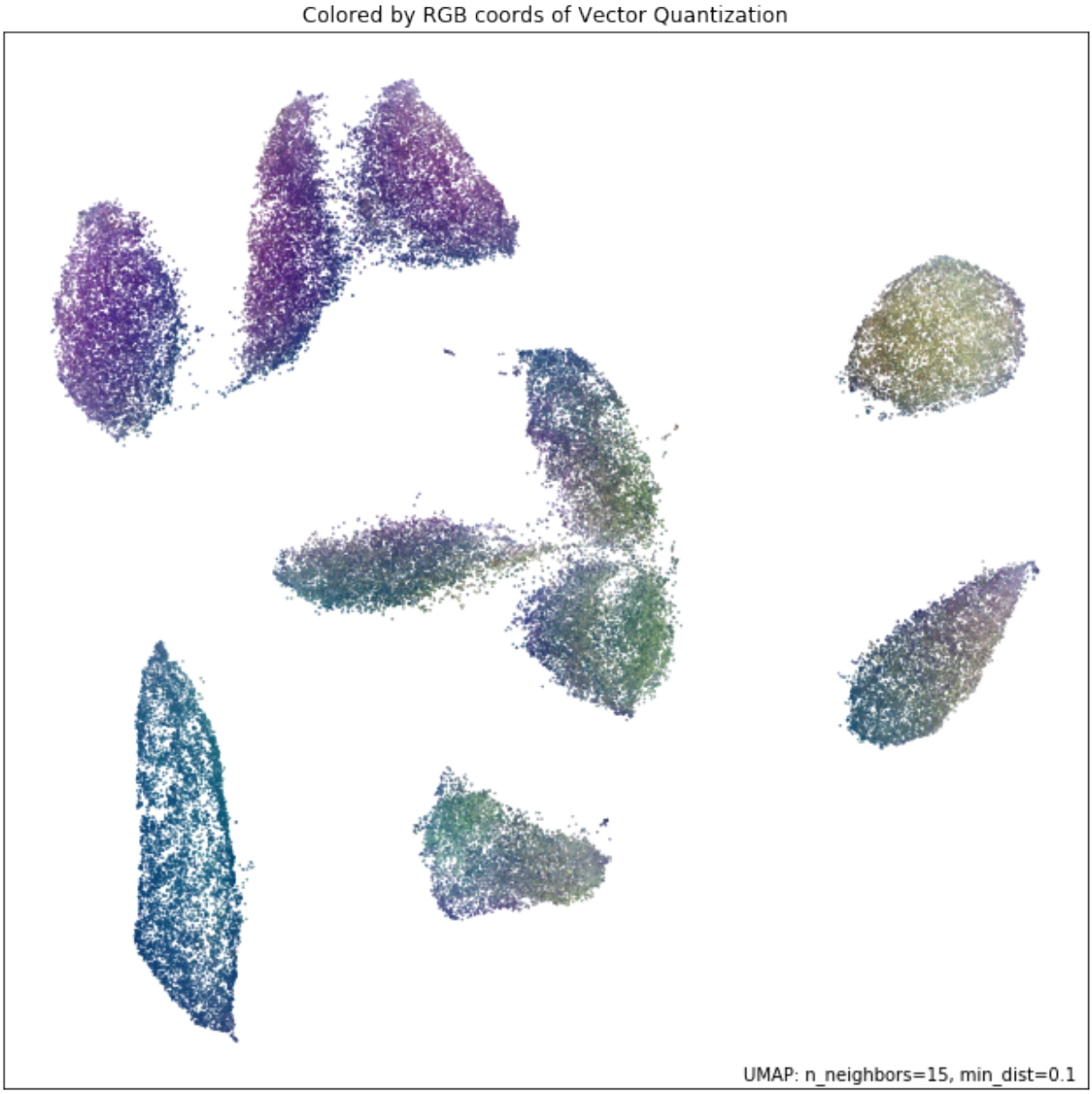



What we are looking for here is a generally smooth transition of colors, and an overall layout that broadly respects the color transitions. In this case the far left has a bottom cluster that transitions from dark green at the bottom to blue at the top, and this matches well with the cluster in the upper right which have a similar shade of blue at the bottom before transitioning to more cyan and blue. In contast in the right of the plot the lower cluster runs from purplish pink to green from top to bottom, while the cluster above it has its bottom edge more purple than green, suggesting that perhaps one or the other of these clusters has been flipped vertically during the optimization process, and this was never quite corrected.

An alternative, but similar, approach is to use vector quantization as the method to generate a 3D embedding to generate

colors. Vector quantization effectively finds 3 representative centers for the data, and then describes each data point in terms of its distance to these centers. Clearly this, again, captures a lot of the broad global structure of the data.

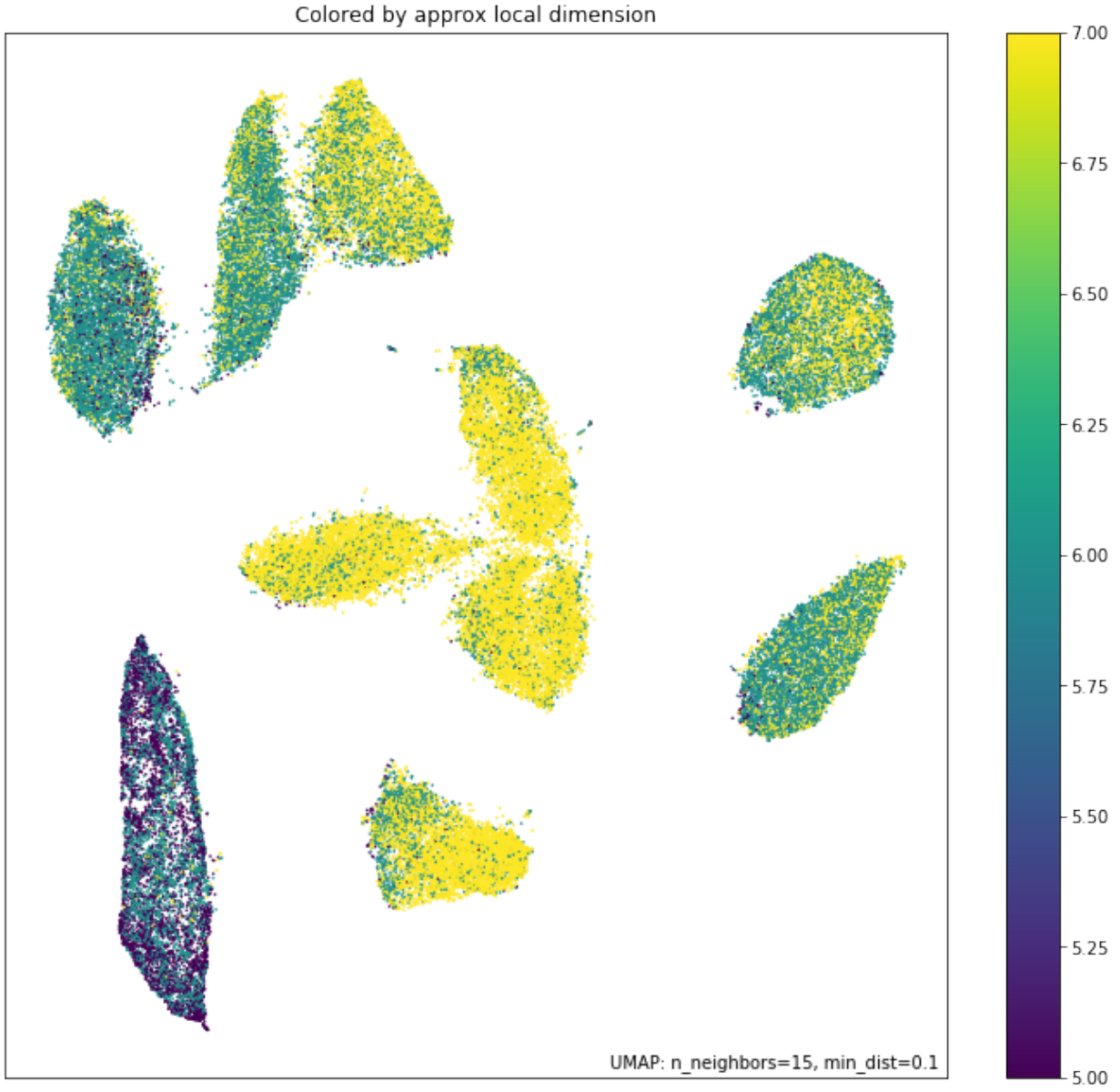
```
umap.plot.diagnostic(mapper, diagnostic_type='vq')
```




Again we are looking for largely smooth transitions, and for related colors to match up between clusters. This view supports the fact that the left hand side of the embedding has worked well, but looking at the right hand side it seems clear that it is the upper two of the clusters that has been inadvertently flipped vertically. By contrasting views like this one can get a better sense of how well the embedding is working.

For a different perspective we can look at approximations of the local dimension around each data point. Ideally the local dimension should match the embedding dimension (although this is often a lot to hope for. In practice when the local dimension is high this represents points (or areas of the space) that UMAP will have a harder time embedding as well. Thus one can trust the embedding to be more accurate in regions where the points have consistently lower local dimension.

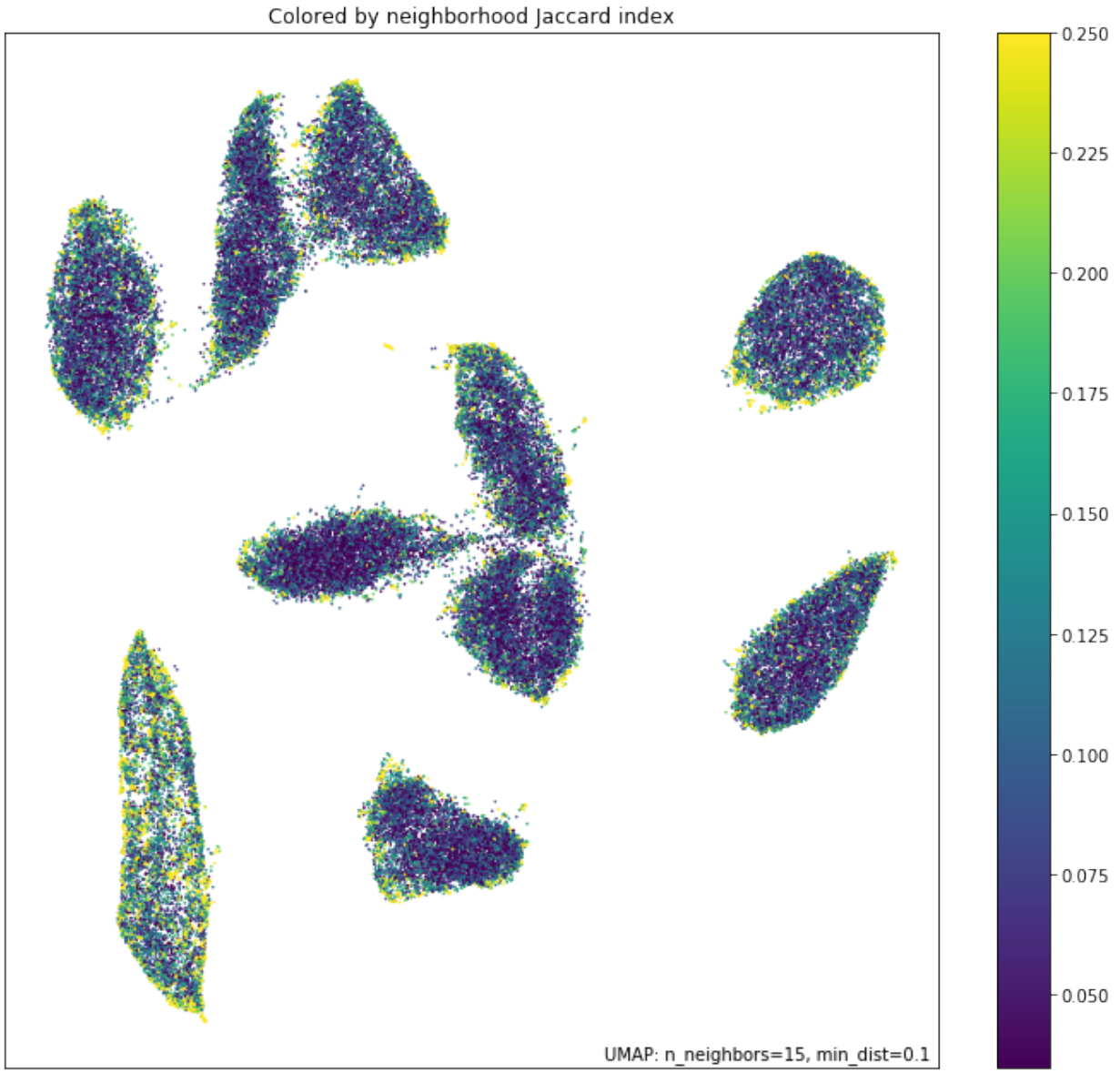
```
local_dims = umap.plot.diagnostic(mapper, diagnostic_type='local_dim')
```




As you can see, the local dimension of the data varies quite widely across the data. In particular the lower left cluster has the lowest local dimension – this is actually unsurprising as this is the cluster corresponding to the digits 1: there are relatively few degrees of freedom over how a person draws a number one, and so the resulting local dimension is lower. In contrast the clusters in the middle have a much higher local dimension. We should expect the embedidng to be a little less accurate in these regions: it is hard to represent seven dimensional data well in only two dimensions, and compromises will need to be made.

The final diagnostic we'll look at is how well local neighborhoods are preserved. We can measure this in terms of the Jaccard index of the local neighborhood in the high dimensional space compared to the equivalent neighborhood in the embedding. The Jaccard index is essentially the ratio of the number of neighbors that the two neighborhoods have in common over the total number of unique neighbors across the two neighborhoods. Higher values mean that the local neighborhood has been more accurately preserved.

```
umap.plot.diagnostic(mapper, diagnostic_type='neighborhood')
```




As one might expect the local neighborhood preservation tends to be a lot better for those points that had a lower local dimension (as seen in the last plot). There is also a tendency for the edges of clusters (where there were clear boundaries to be followed) have a better preservation of neighborhoods than the centers of the clusters that had higher local dimension. Again, this provides a view on which areas of the embedding you can have greater trust in, and which regions had to make compromises to embed into two dimensions.

CHAPTER 4

# UMAP Reproducibility

UMAP is a stochastic algorithm – it makes use of randomness both to speed up approximation steps, and to aid in solving hard optimization problems. This means that different runs of UMAP can produce different results. UMAP is relatively stable – thus the variance between runs should ideally be relatively small – but different runs may have variations none the less. To ensure that results can be reproduced exactly UMAP allows the user to set a random seed state.

Since version 0.4 UMAP also support multi-threading for faster performance; when performing optimization this exploits the fact that race conditions between the threads are acceptable within certain optimization phases. Unfortunately this means that the randomness in UMAP output for the multi-threaded case depends not only on the random seed input, but also on race conditions between thereads during optimization, over which no control can be had. This means that multi-threaded UMAP results cannot be explicitly reproduced.

In this tutorial we'll look at how UMAP can be used in multi-threaded mode for performance purposes, and alternatively how we can fix random states to ensure exact reproducibility at the cost of some performance. First let's load the relevant libraries and get some data; in this case the MNIST digits dataset.

```
import numpy as np
import sklearn.datasets
import umap
import umap.plot
```

```
data, labels = sklearn.datasets.fetch_openml(
    'mnist_784', version=1, return_X_y=True
)
```

With data in hand let's run UMAP on it, and note how long it takes to run:

```
%%time
mapper1 = umap.UMAP().fit(data)
```

```
CPU times: user 3min 18s, sys: 3.84 s, total: 3min 22s
Wall time: 1min 29s
```

The thing to note here is that the "Wall time" is significantly smaller than the CPU time – this means that multiple CPU cores were used. For this particular demonstration I am making use of the latest version of PyNNDescent for nearest neighbor search (UMAP will use it if you have it installed) which supports multi-threading as well. The result is a very fast fitting to the data that does an effective job of using several cores. If you are on a large server with many cores available and don't wish to use them *all* (which is the default situtation) you can currently control the number of cores used by setting the numba environment variable `NUMBA_NUM_THREADS`; see the numba documentation for more details.

Now let's plot our result to see what the embedding looks like:

```
umap.plot.points(mapper1, labels=labels)
```


Now, let's run UMAP again and compare the results to that of our first run.

```
%%time
mapper2 = umap.UMAP().fit(data)
```

```
CPU times: user 2min 53s, sys: 4.16 s, total: 2min 57s
Wall time: 1min 5s
```

You will note that this time we ran *even faster*. This is because during the first run numba was still JIT compiling some of the code in the background. In contrast, this time that work has already been done, so it no longer takes up any of our run-time. We see that we are still making use of mutliple cores well.

Now let's plot the results of this second run and compare to the first:

```
umap.plot.points(mapper2, labels=labels)
```


Qualitatively this looks very similar, but a little closer inspection will quickly show that the results are actually different

between the runs. Note that even in versions of UMAP prior to 0.4 this would have been the case – since we fixed no specific random seed, and were thus using the current random state of the system which will ntaurally differ between runs. This is the default behaviour, as is standard with sklearn estimators that are stochastic. Rather than having a default random seed the user is required to explicitly provide one should they want a reproducible result. As noted by Vito Zanotelli

> ... setting a random seed is like signing a waiver "I am aware that this is a stochastic algorithm and I have done sufficient tests to confirm that my main conclusions are not affected by this randomess".

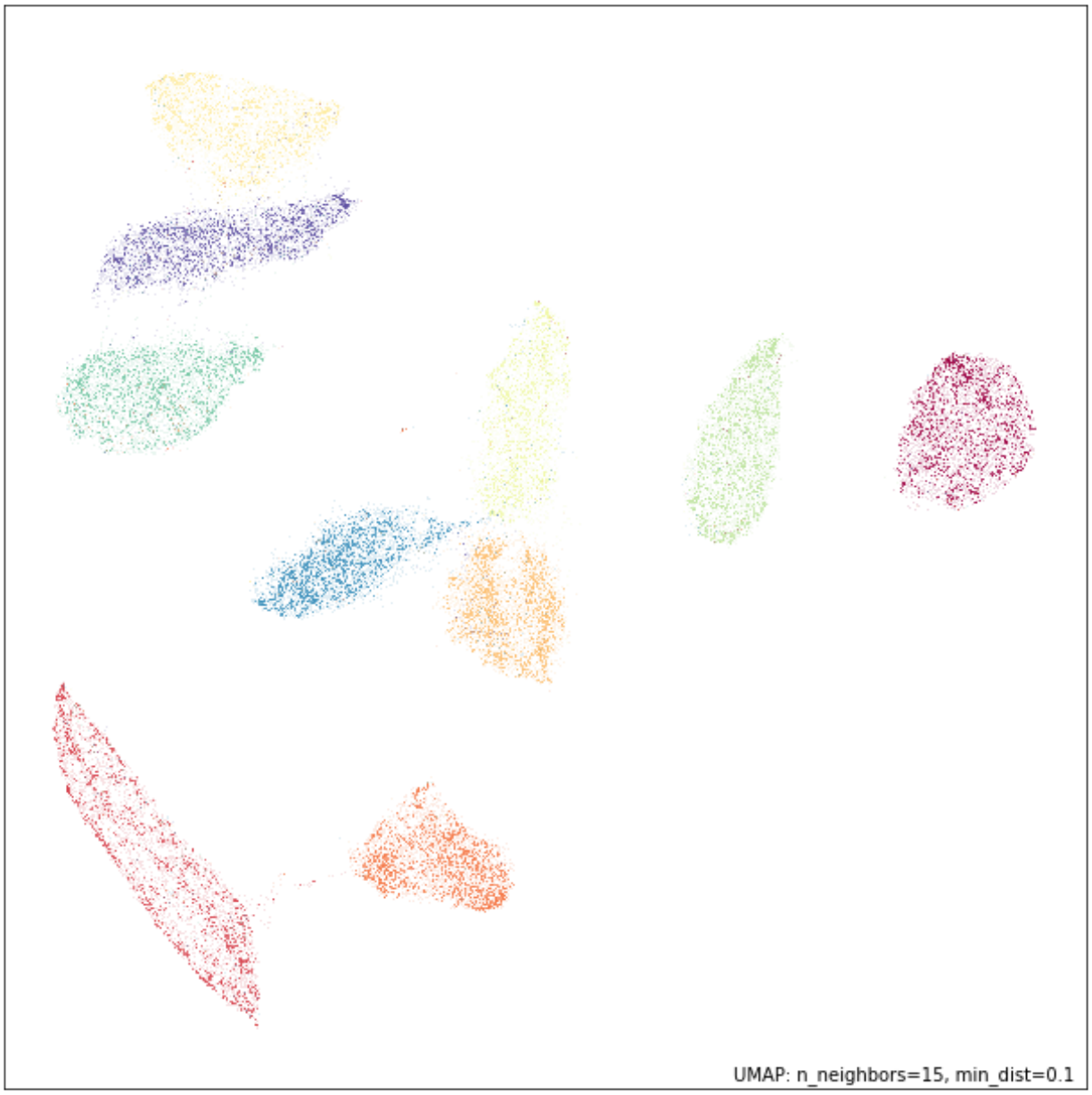
With that in mind, let's see what happens if we set an explicit `random_state` value:

```
%%time
mapper3 = umap.UMAP(random_state=42).fit(data)
```

```
CPU times: user 2min 27s, sys: 4.16 s, total: 2min 31s
Wall time: 1min 56s
```

The first thing to note that that this run took signifcantly longer (despite having all the functions JIT compiled by numba already). Then note that the Wall time and CPU times are now much closer to each other – we are not longer exploiting multiple cores to anywhere near the same degree. This is because by setting a `random_state` we are effectively turning off any of the multi-threading that does not support explicit reproducibility. Let's plot the results:

```
umap.plot.points(mapper3, labels=labels)
```



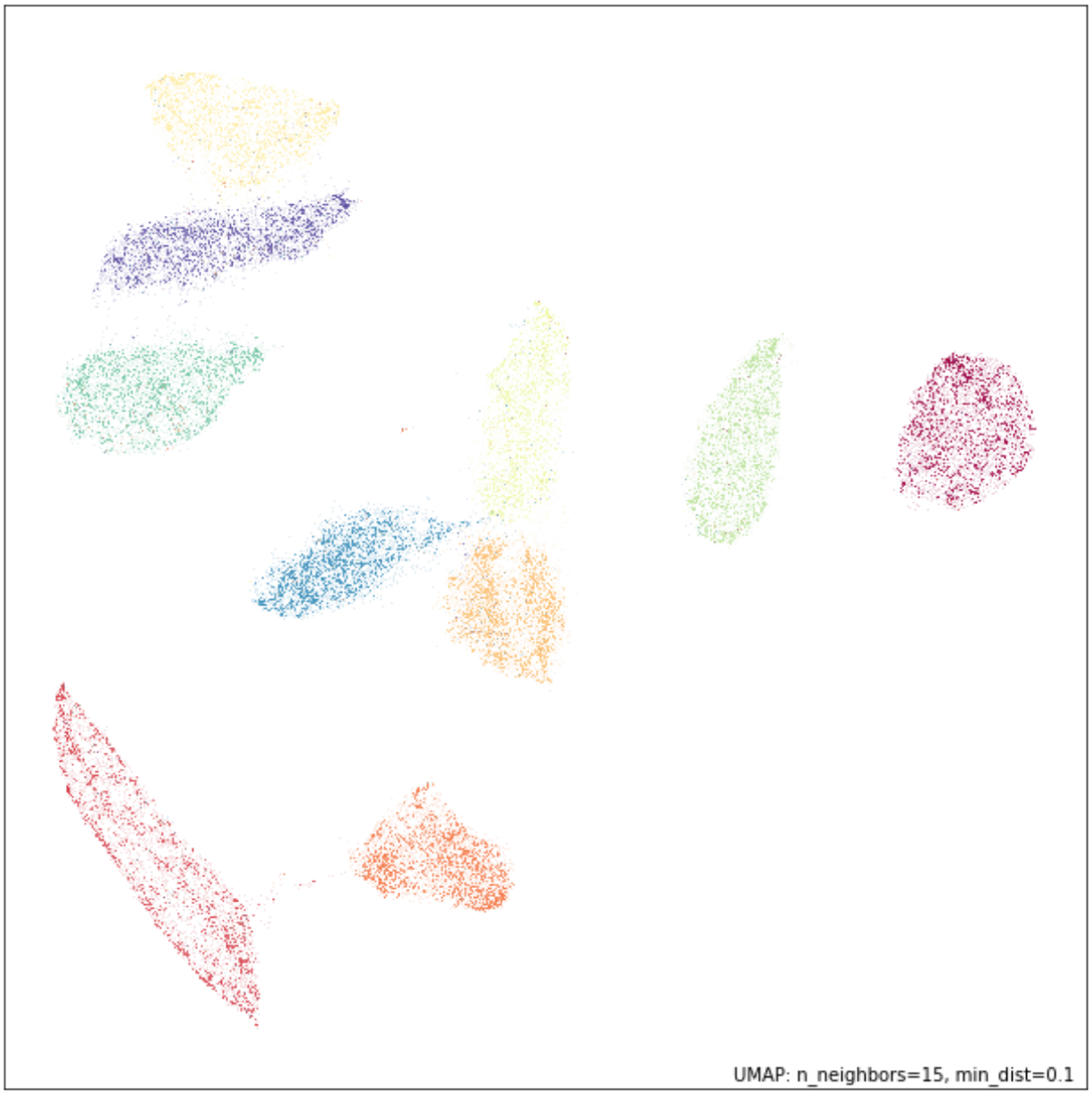

We arrive at much the same results as before from a qualitative point of view, but again inspection will show that there are some differences. More importantly this result should now be reproducible. Thus we can run UMAP again, with the same `random_state` set ...

```
%%time
mapper4 = umap.UMAP(random_state=42).fit(data)
```

```
CPU times: user 2min 26s, sys: 4.13 s, total: 2min 30s
Wall time: 1min 54s
```

Again, this takes longer than the earlier runs with no `random_state` set. However when we plot the results of the second run we see that they look not merely qualitatively similar, but instead appear to be almost identical:

```
umap.plot.points(mapper4, labels=labels)
```




We can, in fact, check that the results are identical by verifying that each and every coordinate of the resulting embeddings match perfectly:

```
np.all(mapper3.embedding_ == mapper4.embedding_)
```

```
True
```

So we have, in fact, reproduced the embedding exactly.

# CHAPTER 5

## Transforming New Data with UMAP

UMAP is useful for generating visualisations, but if you want to make use of UMAP more generally for machine learning tasks it is important to be be able to train a model and then later pass new data to the model and have it transform that data into the learned space. For example if we use UMAP to learn a latent space and then train a classifier on data transformed into the latent space then the classifier is only useful for prediction if we can transform data for which we want a prediction into the latent space the classifier uses. Fortunately UMAP makes this possible, albeit more slowly than some other transformers that allow this.

To demonstrate this functionality we'll make use of scikit-learn and the digits dataset contained therein (see *How to Use UMAP* for an example of the digits dataset). First let's load all the modules we'll need to get this done.

```
import numpy as np
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC

import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

```
sns.set(context='notebook', style='white', rc={'figure.figsize':(14,10)})
```

```
digits = load_digits()
```

To keep everything honest let's use sklearn `train_test_split` to separate out a training and test set (stratified over the different digit types). By default `train_test_split` will carve off 25% of the data for testing, which seems suitable in this case.

```
X_train, X_test, y_train, y_test = train_test_split(digits.data,
                                                    digits.target,
                                                    stratify=digits.target,
                                                    random_state=42)
```

Now to get a benchmark idea of what we are looking at let's train a couple of different classifiers and then see how well they score on the test set. For this example lets try a support vector classifier and a KNN classifier. Ideally we should be tuning hyper-parameters (perhaps a grid search using k-fold cross validation), but for the purposes of this simple demo we will simply use default parameters for both classifiers.

```
svc = SVC().fit(X_train, y_train)
knn = KNeighborsClassifier().fit(X_train, y_train)
```

The next question is how well these classifiers perform on the test set. Conveniently sklearn provides a `score` method that can output the accuracy on the test set.

```
svc.score(X_test, y_test), knn.score(X_test, y_test)
```

```
(0.62, 0.9844444444444445)
```

The result is that the support vector classifier apparently had poor hyper-parameters for this case (I expect with some tuning we could build a much more accurate mode) and the KNN classifier is doing very well.

The goal now is to make use of UMAP as a preprocessing step that one could potentially fit into a pipeline. We will therefore obviously need the `umap` module loaded.

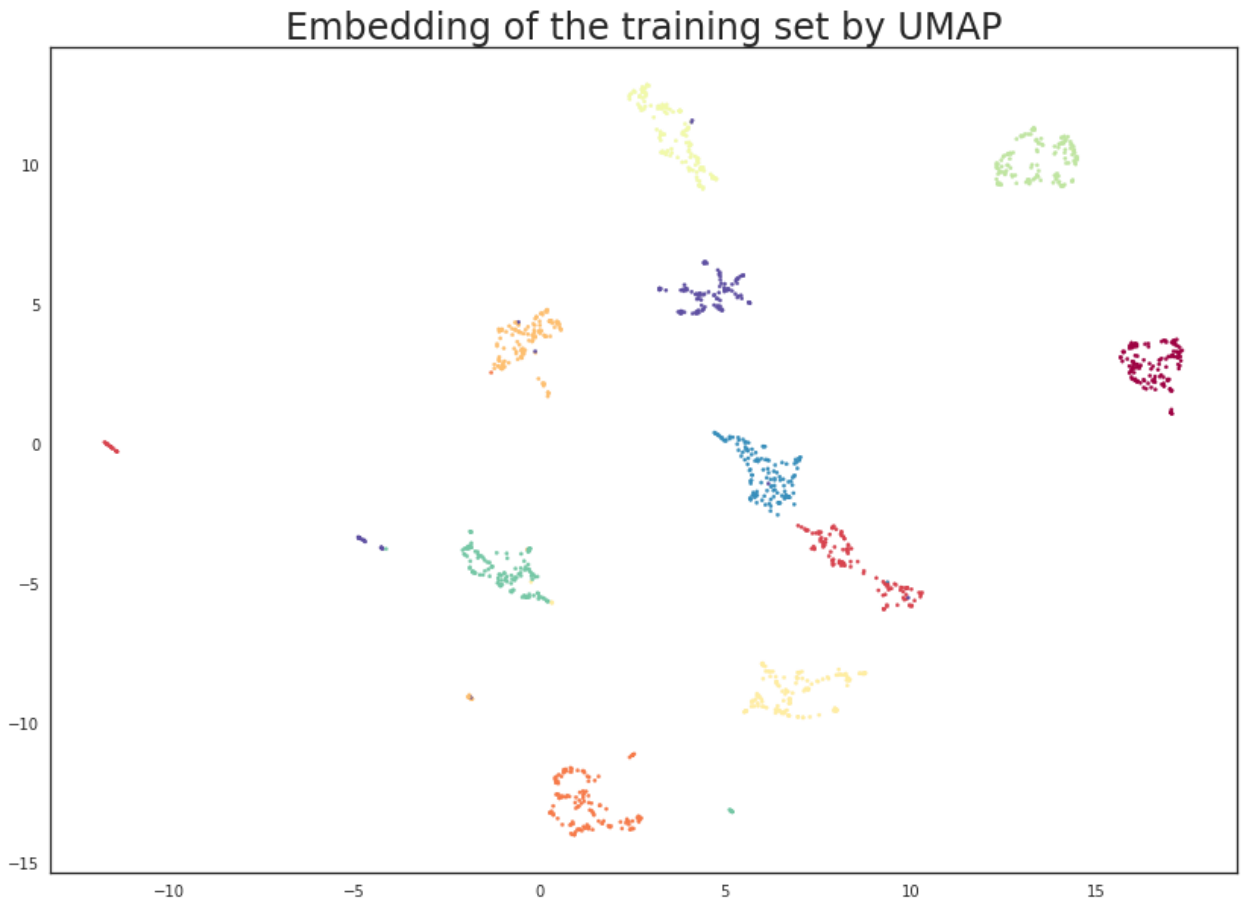
```
import umap
```

To make use of UMAP as a data transformer we first need to fit the model with the training data. This works exactly as in the *How to Use UMAP* example using the fit method. In this case we simply hand it the training data and it will learn an appropriate (two dimensional by default) embedding.

```
trans = umap.UMAP(n_neighbors=5, random_state=42).fit(X_train)
```

Since we embedded to two dimensions we can visualise the results to ensure that we are getting a potential benefit out of this approach. This is simply a matter of generating a scatterplot with data points colored by the class they come from. Note that the embedded training data can be accessed as the `.embedding_` attribute of the UMAP model once we have fit the model to some data.

```
plt.scatter(trans.embedding_[:, 0], trans.embedding_[:, 1], s= 5, c=y_train, cmap=
↪'Spectral')
plt.title('Embedding of the training set by UMAP', fontsize=24);
```




This looks very promising! Most of the classes got very cleanly separated, and that gives us some hope that it could help with classifier performance. It is worth noting that this was a completely unsupervised data transform; we could have used the training label information, but that is the subject of *a later tutorial*.

We can now train some new models (again an SVC and a KNN classifier) on the embedded training data. This looks exactly as before but now we pass it the embedded data. Note that calling `transform` on input identical to what the model was trained on will simply return the `embedding_` attribute, so sklearn pipelines will work as expected.

```
svc = SVC().fit(trans.embedding_, y_train)
knn = KNeighborsClassifier().fit(trans.embedding_, y_train)
```

Now we want to work with the test data which none of the models (UMAP or the classifiers) have seen. To do this we use the standard sklearn API and make use of the `transform` method, this time handing it the new unseen test data. We will assign this to `test_embedding` so that we can take a closer look at the result of applying an existing UMAP model to new data.

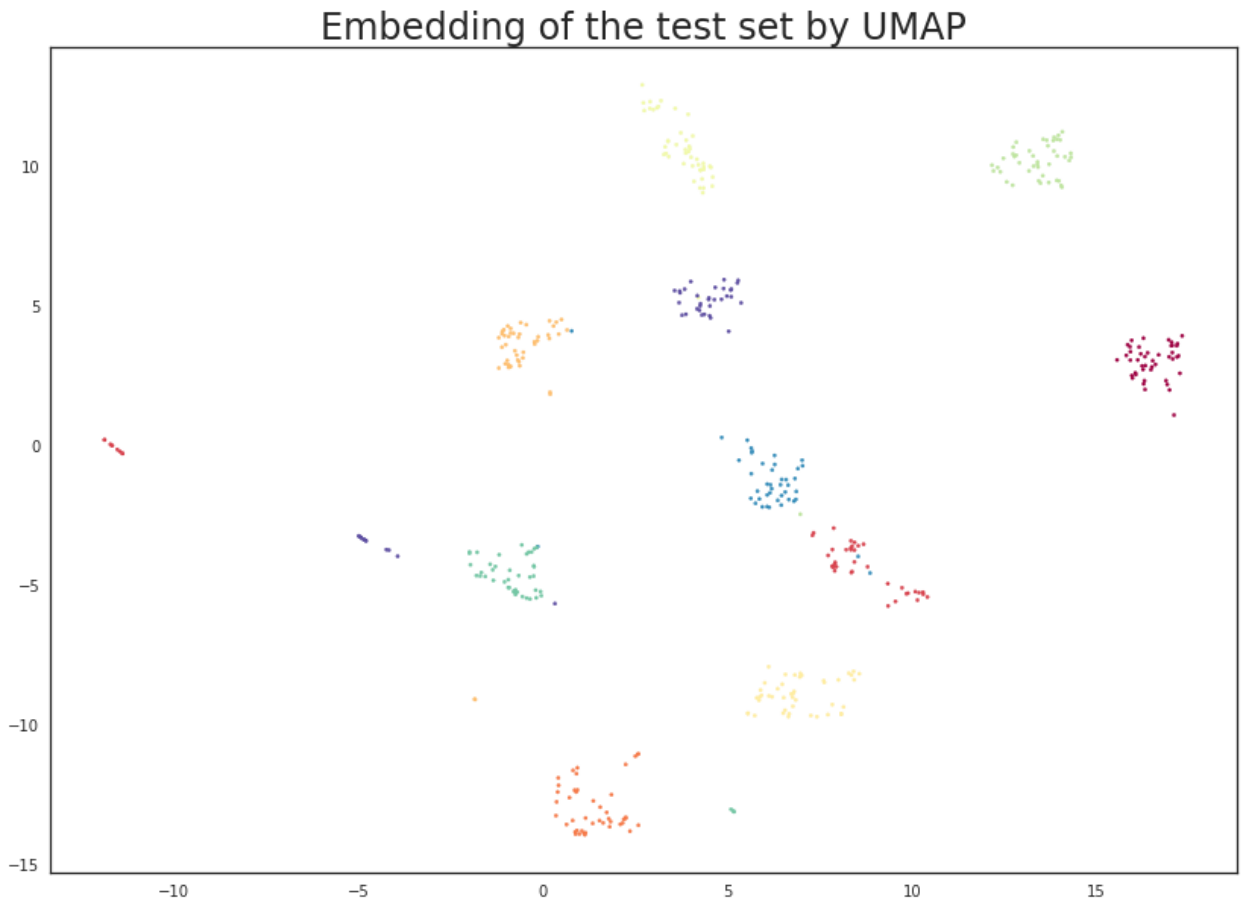
```
%time test_embedding = trans.transform(X_test)
```

```
CPU times: user 867 ms, sys: 70.7 ms, total: 938 ms
Wall time: 335 ms
```

Note that the transform operations works very efficiently – taking less than half a second. Compared to some other transformers this is a little on the slow side, but it is fast enough for many uses. Note that as the size of the training and/or test sets increase the performance will slow proportionally. It's also worth noting that the first call to transform may be slow due to Numba JIT overhead – further runs will be very fast.

The next important question is what the transform did to our test data. In principle we have a new two dimensional representation of the test-set, and ideally this should be based on the existing embedding of the training set. We can check this by visualising the data (since we are in two dimensions) to see if this is true. A simple scatterplot as before will suffice.

```
plt.scatter(test_embedding[:, 0], test_embedding[:, 1], s= 5, c=y_test, cmap='Spectral
↪')
plt.title('Embedding of the test set by UMAP', fontsize=24);
```




The results look like what we should expect; the test data has been embedded into two dimensions in exactly the locations we should expect (by class) given the embedding of the training data visualised above. This means we can now try out of models that were trained on the embedded training data by handing them the newly transformed test set.

```
svc.score(trans.transform(X_test), y_test), knn.score(trans.transform(X_test), y_test)
```

```
(0.9844444444444445, 0.9844444444444445)
```

The results are pretty good. While the accuracy of the KNN classifier did not improve there was not a lot of scope for improvement given the data. On the other hand the SVC has improved to have equal accuracy to the KNN classifier. Of course we could probably have achieved this level of accuracy by better setting SVC hyper-parameters, but the point here is that we can use UMAP as if it were a standard sklearn transformer as part of an sklearn machine learning pipeline.

Just for fun we can run the same experiments, but this time reduce to ten dimensions (where we can no longer visualise). In practice this will have little gain in this case – for the digits dataset two dimensions is plenty for UMAP

and more dimensions won't help. On the other had for more complex datasets where more dimensions may allow for a much more faithful embedding it is worth noting that we are not restricted to only two dimension.

```
trans = umap.UMAP(n_neighbors=5, n_components=10, random_state=42).fit(X_train)
```

```
svc = SVC().fit(trans.embedding_, y_train)
knn = KNeighborsClassifier().fit(trans.embedding_, y_train)
```

```
svc.score(trans.transform(X_test), y_test), knn.score(trans.transform(X_test), y_test)
```

```
(0.9822222222222222, 0.9822222222222222)
```

And we see that in this case we actually marginally lowered our accuracy scores (within the potential noise in such scoring mind you). However for more interesting datasets the larger dimensional embedding may have been a significant gain – it is certainly worth exploring as one of the parameters in a grid search across a pipeline that includes UMAP.

CHAPTER 6

# Inverse transforms

UMAP has some support for inverse transforms – generating a high dimensional data sample given a location in the low dimensional embedding space. To start let's load all the relavent libraries.

```
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.gridspec import GridSpec
import seaborn as sns
import sklearn.datasets
import umap
import umap.plot
```

We will need some data to test with. To start we'll use the MNIST digits dataset. This is a dataset of 70000 handwritten digits encoded as grayscale 28x28 pixel images. Our goal is to use UMAP to reduce the dimension of this dataset to something small, and then see if we can generate new digits by sampling points from the embedding space. To load the MNIST dataset we'll make use of sklearn's `fetch_openml` function.

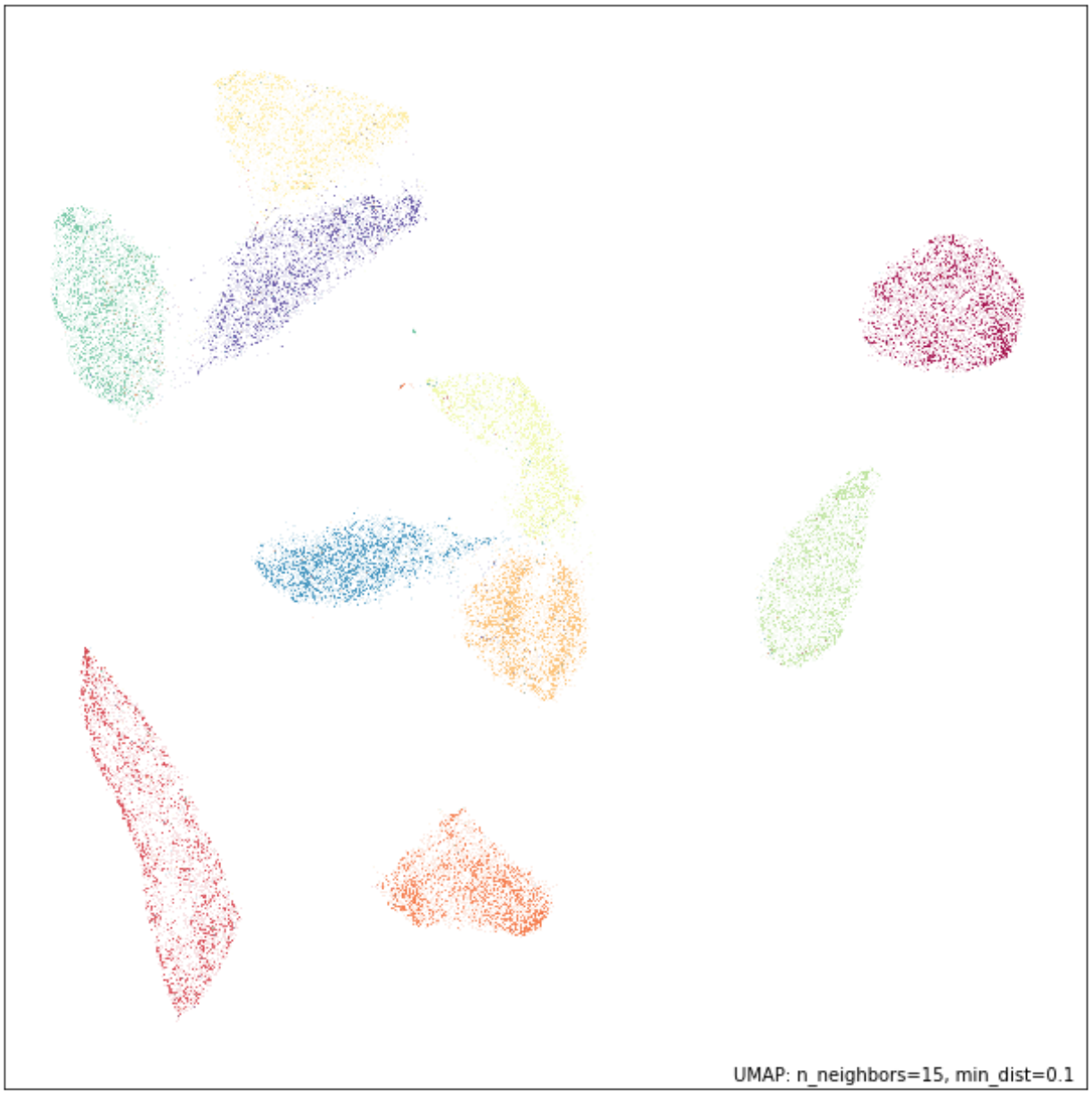
```
data, labels = sklearn.datasets.fetch_openml('mnist_784', version=1, return_X_y=True)
```

Now we need to generate a reduced dimension representation of this data. This is straightforward with umap, but in this case rather than using `fit_transform` we'll use the fit method so that we can retain the trained model for later generating new digits based on samples from the embedding space.

```
mapper = umap.UMAP(random_state=42).fit(data)
```

To ensure that things worked correctly we can plot the data (since we reduced it to two dimensions). We'll use the `umap.plot` functionality to do this.

```
umap.plot.points(mapper, labels=labels)
```




This looks much like we would expect. The different digit classes have been decently separated. Now we need to create a set of samples in the emebdding space to apply the `inverse_transform` operation to. To do this we'll generate a grid of samples linearly interpolating between four corner points. To make out selection interesting we'll carefully choose the corners to span over the dataset, and sample different digits so that we can better see the transitions.

```
corners = np.array([
    [-5, -10],  # 1
    [-7, 6],  # 7
    [2, -8],  # 2
    [12, 4],  # 0
])

test_pts = np.array([
    (corners[0]*(1-x) + corners[1]*x)*(1-y) +
```

(continues on next page)

(continued from previous page)

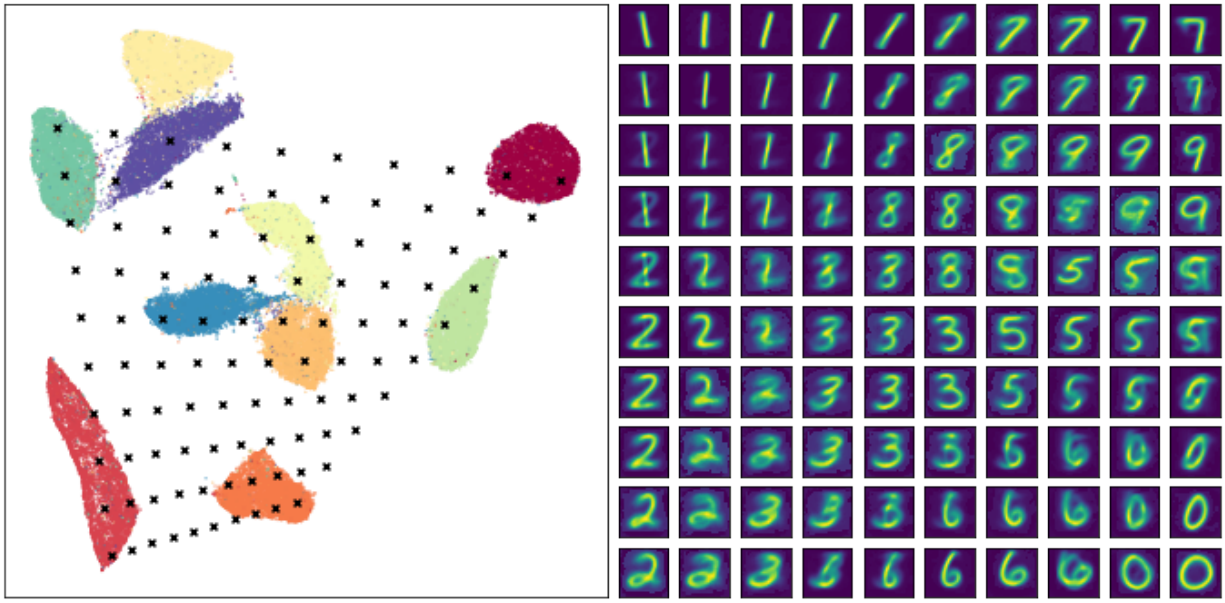
```
    (corners[2]*(1-x) + corners[3]*x)*y
    for y in np.linspace(0, 1, 10)
    for x in np.linspace(0, 1, 10)
])
```

Now we can apply the `inverse_transform` method to this set of test points. Each test point is a two dimensional point lying somewhere in the embedding space. The `inverse_transform` method will convert this in to an approximation of the high dimensional representation that would have been embedded into such a location. Following the sklearn API this is as simple to use as calling the `inverse_transform` method of the trained model and passing it the set of test points that we want to convert into high dimensional representations. Be warned that this can be quite expensive computationally.

```
inv_transformed_points = mapper.inverse_transform(test_pts)
```

Now the goal is to visualize how well we have done. Effectively what we would like to do is show the test points in the embedding space, and then show a grid of the corresponding images generated by the inverse transform. To get all of this in a single matplotlib figure takes a little setting up, but is quite manageable – mostly it is just a matter of managing `GridSpec` formatting. Once we have that setup we just need a scatterplot of the embedding, a scatterplot of the test points, and finally a grid of the images we generated (converting the inverse transformed vectors into images is just a matter of reshaping them back to 28 by 28 pixel grids and using `imshow`).

```
# Set up the grid
fig = plt.figure(figsize=(12,6))
gs = GridSpec(10, 20, fig)
scatter_ax = fig.add_subplot(gs[:, :10])
digit_axes = np.zeros((10, 10), dtype=object)
for i in range(10):
    for j in range(10):
        digit_axes[i, j] = fig.add_subplot(gs[i, 10 + j])

# Use umap.plot to plot to the major axis
# umap.plot.points(mapper, labels=labels, ax=scatter_ax)
scatter_ax.scatter(mapper.embedding_[:, 0], mapper.embedding_[:, 1],
                   c=labels.astype(np.int32), cmap='Spectral', s=0.1)
scatter_ax.set(xticks=[], yticks=[])

# Plot the locations of the text points
scatter_ax.scatter(test_pts[:, 0], test_pts[:, 1], marker='x', c='k', s=15)

# Plot each of the generated digit images
for i in range(10):
    for j in range(10):
        digit_axes[i, j].imshow(inv_transformed_points[i*10 + j].reshape(28, 28))
        digit_axes[i, j].set(xticks=[], yticks=[])
```



The end result looks pretty good – we did indeed generate plausible looking digit images, and many of the transitions (from 1 to 7 across the top row for example) seem pretty natural and make sense. This can help you to understand the structure of the cluster of 1s (it transitions on the angle, sloping toward what will eventually be 7s), and why 7s and 9s are close together in the embedding. Of course there are also some stranger transitions, especially where the test points fell into large gaps between clusters in the embedding – in some sense it is hard to interpret what should go in some of those gaps as they don't really represent anything resembling a smooth transition).

A further note: None of the test points chosen fall outside the convex hull of the embedding. This is deliberate – the inverse transform function operates poorly outside the bounds of that convex hull. Be warned that if you select points to inverse transform that are outside the bounds about the embedding you will likely get strange results (often simply snapping to a particular source high dimensional vector).

Let's continue the demonstration by looking at the Fashion MNIST dataset. As before we can load this through sklearn.

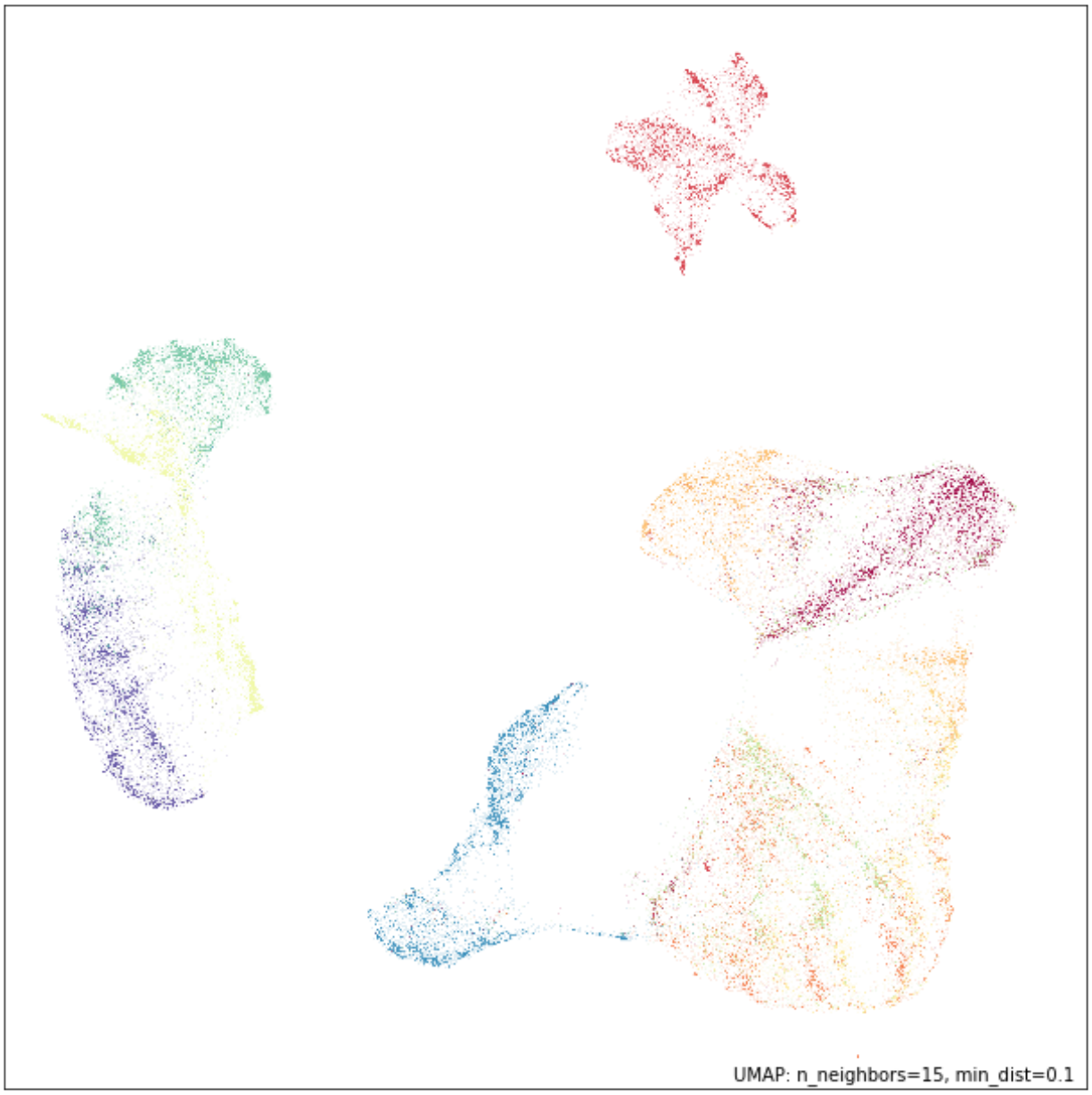
```
data, labels = sklearn.datasets.fetch_openml('Fashion-MNIST', version=1, return_X_
↪y=True)
```

Again we can fit this data with UMAP and get a mapper object.

```
mapper = umap.UMAP(random_state=42).fit(data)
```

Let's plot the embedding to see what we got as a result:

```
umap.plot.points(mapper, labels=labels)
```




Again we'll generate a set of test points by making a grid interpolating between four corners. As before we'll select the corners so that we can stay within the convex hull of the embedding points and ensure nothign to strange happens with the inverse transforms.

```
corners = np.array([
    [-2, -6],  # bags
    [-9, 3],  # boots?
    [7, -5],  # shirts/tops/dresses
    [4, 10],  # pants
])

test_pts = np.array([
    (corners[0]*(1-x) + corners[1]*x)*(1-y) +
    (corners[2]*(1-x) + corners[3]*x)*y
```

(continues on next page)

(continued from previous page)

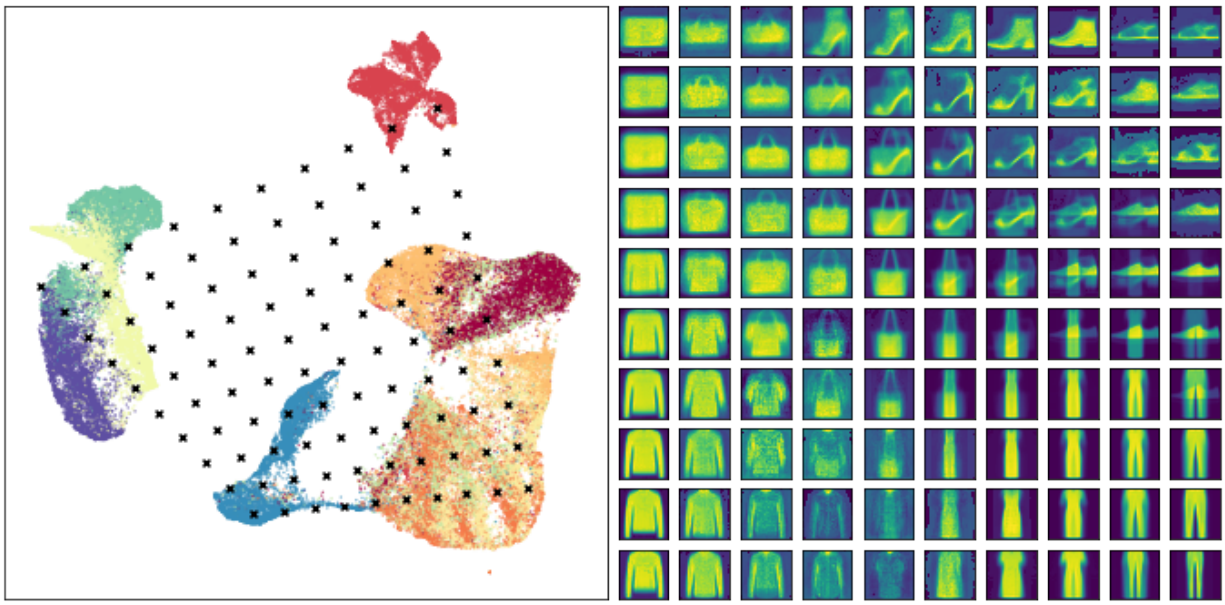
```
    for y in np.linspace(0, 1, 10)
    for x in np.linspace(0, 1, 10)
])
```

Now we simply apply the inverse transform just as before. Again, be warned, this is quite expensive computationally and may take some time to complete.

```
inv_transformed_points = mapper.inverse_transform(test_pts)
```

And now we can use similar code as above to set up out plot of the embedding with test points overlaid, and the generated images.

```
# Set up the grid
fig = plt.figure(figsize=(12,6))
gs = GridSpec(10, 20, fig)
scatter_ax = fig.add_subplot(gs[:, :10])
digit_axes = np.zeros((10, 10), dtype=object)
for i in range(10):
    for j in range(10):
        digit_axes[i, j] = fig.add_subplot(gs[i, 10 + j])

# Use umap.plot to plot to the major axis
# umap.plot.points(mapper, labels=labels, ax=scatter_ax)
scatter_ax.scatter(mapper.embedding_[:, 0], mapper.embedding_[:, 1],
                   c=labels.astype(np.int32), cmap='Spectral', s=0.1)
scatter_ax.set(xticks=[], yticks=[])

# Plot the locations of the text points
scatter_ax.scatter(test_pts[:, 0], test_pts[:, 1], marker='x', c='k', s=15)

# Plot each of the generated digit images
for i in range(10):
    for j in range(10):
        digit_axes[i, j].imshow(inv_transformed_points[i*10 + j].reshape(28, 28))
        digit_axes[i, j].set(xticks=[], yticks=[])
```



This time we see some of the interpolations between items looking rather strange – particularly the points that lie somewhere between shoes and pants – ultimately it is doing the best it can with a difficult problem. At the same time many of the other transitions seem to work pretty well, so it is, indeed, providing useful information about how the embedding is structured.

CHAPTER 7

# UMAP on sparse data

Sometimes datasets get very large, and potentially very very high dimensional. In many such cases, however, the data itself is sparse – that is, while there are many many features, any given sample has only a small number of non-zero features observed. In such cases the data can be represented much more efficiently in terms of memory usage by a sparse matrix data structure. It can be hard to find dimension reduction techniques that work directly on such sparse data – often one applies a basic linear technique such as `TruncatedSVD` from sklearn (which does accept sparse matrix input) to get the data in a format amenable to other more advanced dimension reduction techniques. In the case of UMAP this is not necessary – UMAP can run directly on sparse matrix input. This tutorial will walk through a couple of examples of doing this. First we'll need some libraries loaded. We need `numpy` obviously, but we'll also make use of `scipy.sparse` which provides sparse matrix data structures. One of our examples will be purely mathematical, and we'll make use of `sympy` for that; the other example is test based and we'll use sklearn for that (specifically `sklearn.feature_extraction.text`). Beyond that we'll need umap, and plotting tools.

```
import numpy as np
import scipy.sparse
import sympy
import sklearn.datasets
import sklearn.feature_extraction.text
import umap
import umap.plot
import matplotlib.pyplot as plt
%matplotlib inline
```

## 7.1 A mathematical example

Our first example constructs a sparse matrix of data out of pure math. This example is inspired by the work of John Williamson, and if you haven't looked at that work you are strongly encouraged to do so. The dataset under consideration will be the integers. We will represent each integer by a vector of its divisibility by distinct primes. Thus our feature space is the space of prime numbers (less than or equal to the largest integer we will be considering) – potentially very high dimensional. In practice a given integer is divisible by only a small number of distinct primes, so each sample will be mostly made up of zeros (all the primes that the number is not divisible by), and thus we will have a very sparse dataset.

To get started we'll need a list of all the primes. Fortunately we have `sympy` at our disposal and we can quickly get that information with a single call to `primerange`. We'll also need a dictionary mapping the different primes to the column number they correspond to in our data structure; effectively we'll just be enumerating the primes.

```
primes = list(sympy.primerange(2, 110000))
prime_to_column = {p:i for i, p in enumerate(primes)}
```

Now we need to construct our data in a format we can put into a sparse matrix easily. At this point a little background on sparse matrix data structures is useful. For this purpose we'll be using the so called "LIL" format. LIL is short for "List of Lists", since that is how the data is internally stored. There is a list of all the rows, and each row is stored as a list giving the column indices of the non-zero entries. To store the data values there is a parallel structure containing the value of the entry corresponding to a given row and column.

To put the data together in this sort of format we need to construct such a list of lists. We can do that by iterating over all the integers up to a fixed bound, and for each integer (i.e. each row in our dataset) generating the list of column indices which will be non-zero. The column indices will simply be the indices corresponding to the primes that divide the number. Since `sympy` has a function `primefactors` which returns a list of the unique prime factors of any integer we simply need to map those through our dictionary to covert the primes into column numbers.

Parallel to that we'll construct the corresponding structure of values to insert into a matrix. Since we are only concerned with divisibility this will simply be a one in every non-zero entry, so we can just add a list of ones of the appropriate length for each row.

```
%%time
lil_matrix_rows = []
lil_matrix_data = []
for n in range(100000):
    prime_factors = sympy.primefactors(n)
    lil_matrix_rows.append([prime_to_column[p] for p in prime_factors])
    lil_matrix_data.append([1] * len(prime_factors))
```

```
CPU times: user 2.07 s, sys: 26.4 ms, total: 2.1 s
Wall time: 2.1 s
```

Now we need to get that into a sparse matrix. Fortunately the `scipy.sparse` package makes this easy, and we've already built the data in a fairly useful structure. First we create a sparse matrix of the correct format (LIL) and the right shape (as many rows as we have generated, and as many columns as there are primes). This is essentially just an empty matrix however. We can fix that by setting the `rows` attribute to be the rows we have generated, and the `data` attribute to be the corresponding structure of values (all ones). The result is a sparse matrix data structure which can then be easily manipulated and converted into other sparse matrix formats easily.

```
factor_matrix = scipy.sparse.lil_matrix((len(lil_matrix_rows), len(primes)), dtype=np.
↪float32)
factor_matrix.rows = np.array(lil_matrix_rows)
factor_matrix.data = np.array(lil_matrix_data)
factor_matrix
```

```
<100000x10453 sparse matrix of type '<class 'numpy.float32'>'
    with 266398 stored elements in LInked List format>
```

As you can see we have a matrix with 100000 rows and over 10000 columns. If we were storing that as a numpy array it would take a great deal of memory. In practice, however, there are only 260000 or so entries that are not zero, and that's all we really need to store, making it much more compact.

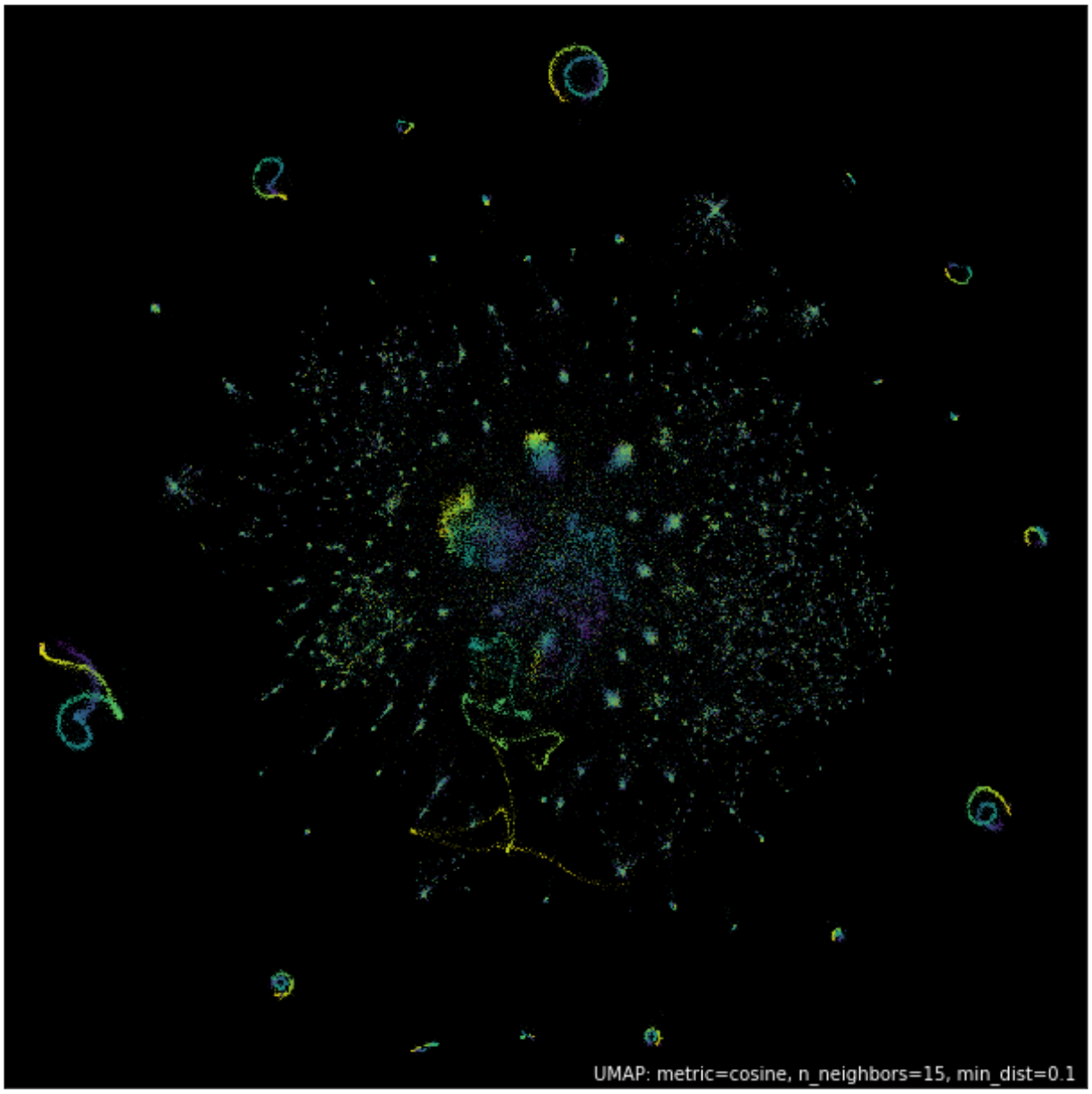
The question now is how can we feed that sparse matrix structure into UMAP to have it learn an embedding. The answer is surprisingly straightforward – we just hand it directly to the fit method. Just like other sklearn estimators that can handle sparse input UMAP will detect the sparse matrix and just do the right thing.

```
%%time
mapper = umap.UMAP(metric='cosine', random_state=42, low_memory=True).fit(factor_
↪matrix)
```

```
CPU times: user 9min 36s, sys: 6.76 s, total: 9min 43s
Wall time: 9min 7s
```

That was easy! But is it really working? We can easily plot the results:

```
umap.plot.points(mapper, values=np.arange(100000), theme='viridis')
```




And this looks very much in line with the results John Williamson got with the proviso that we only used 100,000 integers instead of 1,000,000 to ensure that most users should be able to run this example (the full million may require a large memory compute node). So it seems like this is working well. The next question is whether we can use the

`transform` functionality to map new data into this space. To test that we'll need some more data. Fortunately there are more integers. We'll grab the next 10,000 and put them in a sparse matrix, much as we did for the first 100,000.

```
%%time
lil_matrix_rows = []
lil_matrix_data = []
for n in range(100000, 110000):
    prime_factors = sympy.primefactors(n)
    lil_matrix_rows.append([prime_to_column[p] for p in prime_factors])
    lil_matrix_data.append([1] * len(prime_factors))
```

```
CPU times: user 214 ms, sys: 1.99 ms, total: 216 ms
Wall time: 222 ms
```

```
new_data = scipy.sparse.lil_matrix((len(lil_matrix_rows), len(primes)), dtype=np.
↪float32)
new_data.rows = np.array(lil_matrix_rows)
new_data.data = np.array(lil_matrix_data)
new_data
```

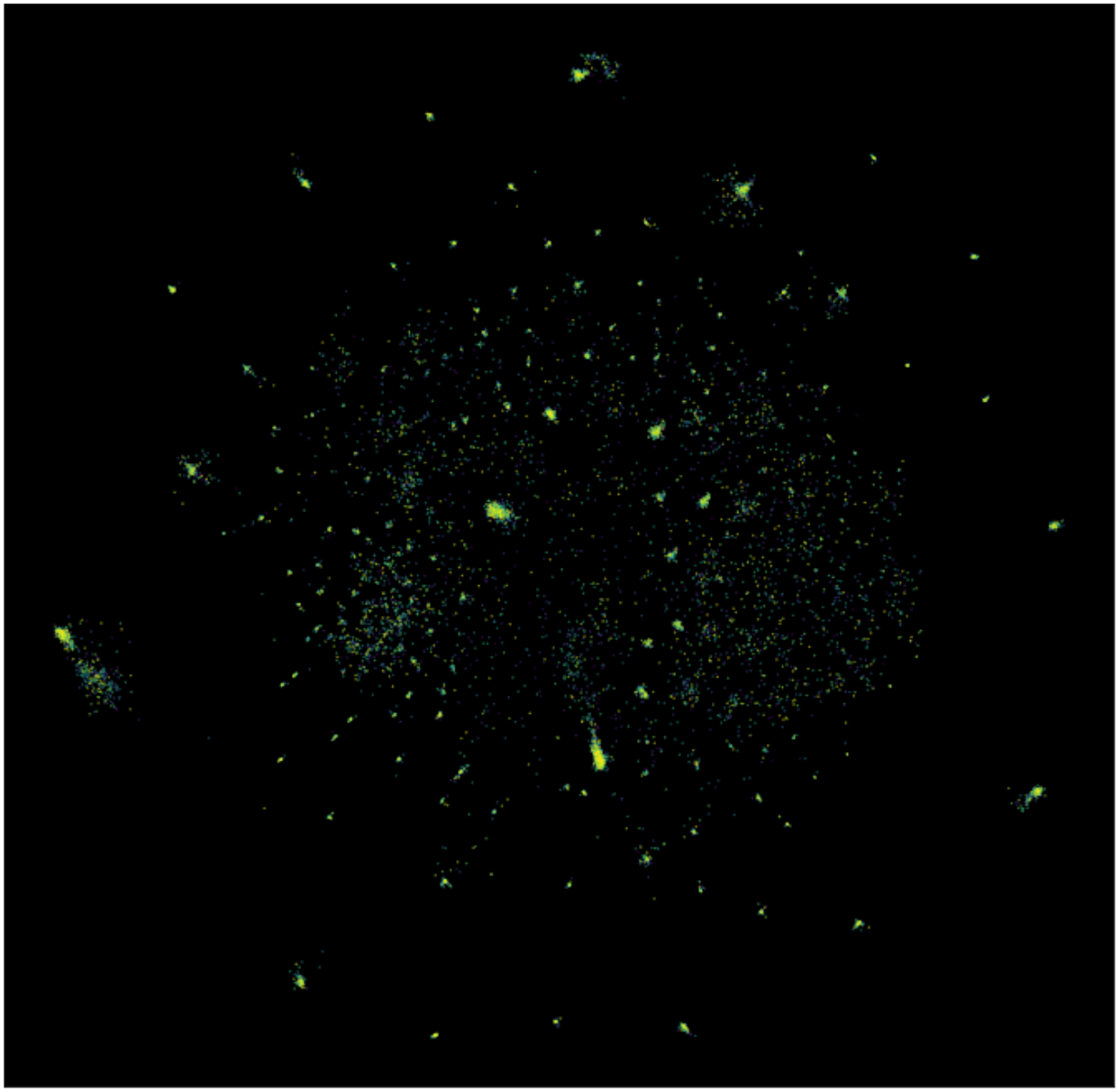
```
<10000x10453 sparse matrix of type '<class 'numpy.float32'>'
    with 27592 stored elements in LInked List format>
```

To map the new data we generated we can simply hand it to the `transform` method of our trained model. This is a little slow, but it does work.

```
new_data_embedding = mapper.transform(new_data)
```

And we can plot the results. Since we just got the locations of the points this time (rather than a model) we'll have to resort to matplotlib for plotting.

```
fig = plt.figure(figsize=(12,12))
ax = fig.add_subplot(111)
plt.scatter(new_data_embedding[:, 0], new_data_embedding[:, 1], s=0.1, c=np.
↪arange(10000), cmap='viridis')
ax.set(xticks=[], yticks=[], facecolor='black');
```



The color scale is different in this case, but you can see that the data has been mapped into locations corresponding to the various structures seen in the original embedding. Thus, even with large sparse data we can embed the data, and even add new data to the embedding.

## 7.2 A text analysis example

Let's look at a more classical machine learning example of working with sparse high dimensional data – working with text documents. Machine learning on text is hard, and there is a great deal of literature on the subject, but for now we'll just consider a basic approach. Part of the difficulty of machine learning with text is turning language into numbers, since numbers are really all most machine learning algorithms understand (at heart anyway). One of the most straightforward ways to do this for documents is what is known as the "bag-of-words" model. In this model we view a document as simply a multi-set of the words contained in it – we completely ignore word order. The result can be viewed as a matrix of data by setting the feature space to be the set of all words that appear in any document, and a

document is represented by a vector where the value of the *i*th entry is the number of times the *i*th word occurs in that document. This is a very common approach, and is what you will get if you apply sklearn's `CountVectorizer` to a text dataset for example. The catch with this approach is that the feature space is often *very* large, since we have a feature for each and every word that ever occurs in the entire corpus of documents. The data is sparse however, since most documents only use a small portion of the total possible vocabulary. Thus the default output format of `CountVectorizer` (and other similar feature extraction tools in sklearn) is a `scipy.sparse` format matrix.

For this example we'll make use of the classic 20newsgroups dataset, a sampling of newsgroup messages from the old NNTP newsgroup system covering 20 different newsgroups. The `sklearn.datasets` module can easily fetch the data, and, in fact, we can fetch a pre-vectorized version to save us the trouble of running `CountVectorizer` ourselves. We'll grab both the training set, and the test set for later use.

```
news_train = sklearn.datasets.fetch_20newsgroups_vectorized(subset='train')
news_test = sklearn.datasets.fetch_20newsgroups_vectorized(subset='test')
```

If we look at the actual data we have pulled back, we'll see that sklearn has run a `CountVectorizer` and produced the data is sparse matrix format.

```
news_train.data
```

```
<11314x130107 sparse matrix of type '<class 'numpy.float64'>'
    with 1787565 stored elements in Compressed Sparse Row format>
```

The value of the sparse matrix format is immediately obvious in this case; while there are only 11,000 samples there are 130,000 features! If the data were stored in a standard `numpy` array we would be using up 10GB of memory! And most of that memory would simply be storing the number zero, over and over again. In sparse matrix format it easily fits in memory on most machines. This sort of dimensionality of data is very common with text workloads.

The raw counts are, however, not ideal since common words the "the" and "and" will dominate the counts for most documents, while contributing very little information about the actual content of the document. We can correct for this by using a `TfidfTransformer` from sklearn, which will convert the data into TF-IDF format. There are lots of ways to think about the transformation done by TF-IDF, but I like to think of it intuitively as follows. The information content of a word can be thought of as (roughly) proportional to the negative log of the frequency of the word; the more often a word is used, the less information it tends to carry, and infrequent words carry more information. What TF-IDF is going to do can be thought of as akin to re-weighting the columns according to the information content of the word associated to that column. Thus the common words like "the" and "and" will get down-weighted, as carrying less information about the document, while infrequent words will be deemed more imporant and have their associated columns up-weighted. We can apply this transformation to both the train and test sets (using the same transformer trained on the training set).

```
tfidf = sklearn.feature_extraction.text.TfidfTransformer(norm='l1').fit(news_train.
↪data)
train_data = tfidf.transform(news_train.data)
test_data = tfidf.transform(news_test.data)
```

The result is still a sparse matrix, since TF-IDF doesn't change the zero elements at all, nor the number of features.

```
train_data
```

```
<11314x130107 sparse matrix of type '<class 'numpy.float64'>'
    with 1787565 stored elements in Compressed Sparse Row format>
```

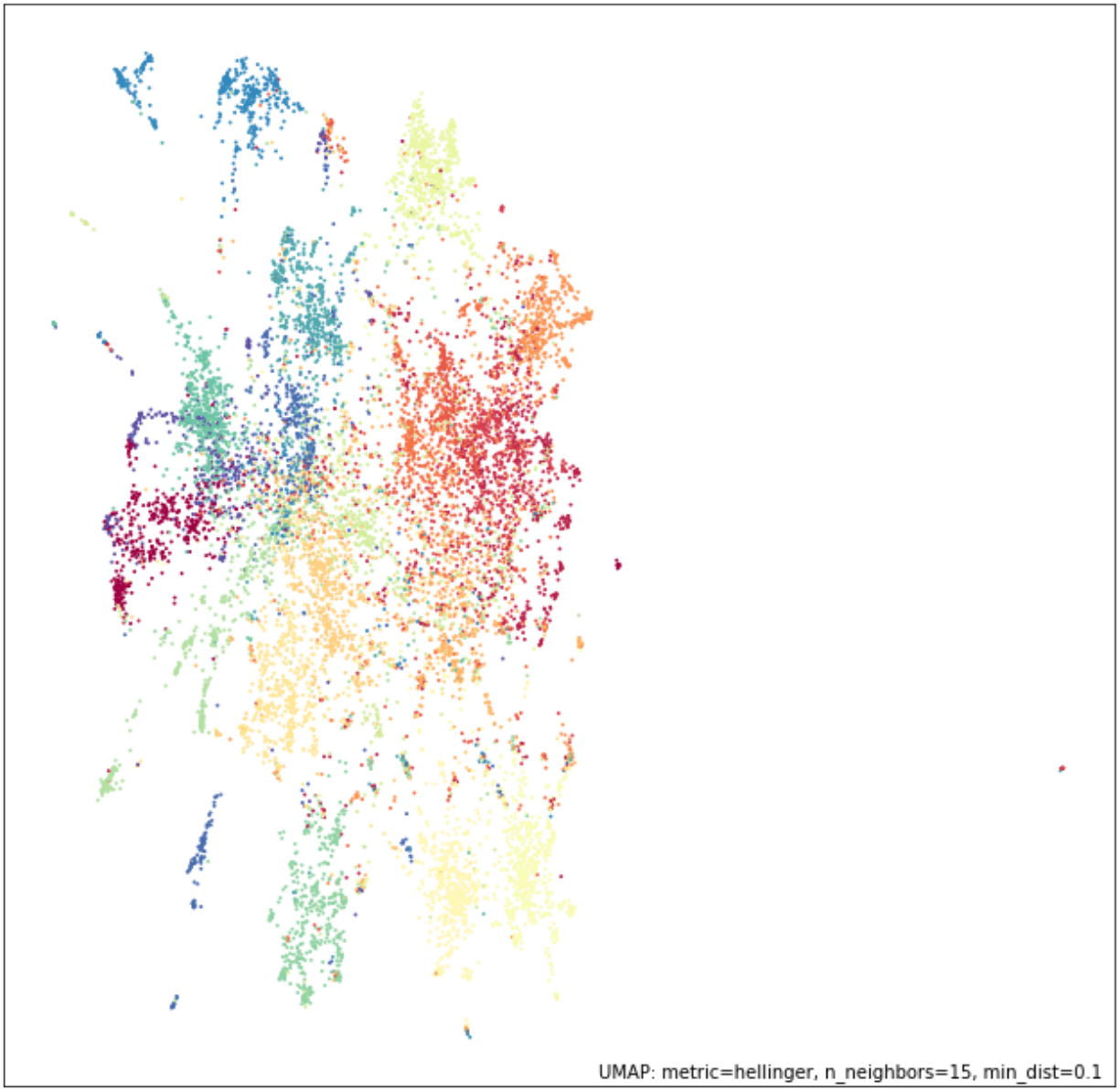
Now we need to pass this very high dimensional data to UMAP. Unlike some other non-linear dimension reduction techniques we don't need to apply PCA first to get the data down to a reasonable dimensionality; nor do we need to use other techniques to reduce the data to be able to be represented as a dense `numpy` array; we can work directly on the 130,000 dimensional sparse matrix.

```
%%time
mapper = umap.UMAP(metric='hellinger', random_state=42).fit(train_data)
```

```
CPU times: user 8min 40s, sys: 3.07 s, total: 8min 44s
Wall time: 8min 43s
```

Now we can plot the results, with labels according to the target variable of the data – which newsgroup the posting was drawn from.

```
umap.plot.points(mapper, labels=news_train.target)
```



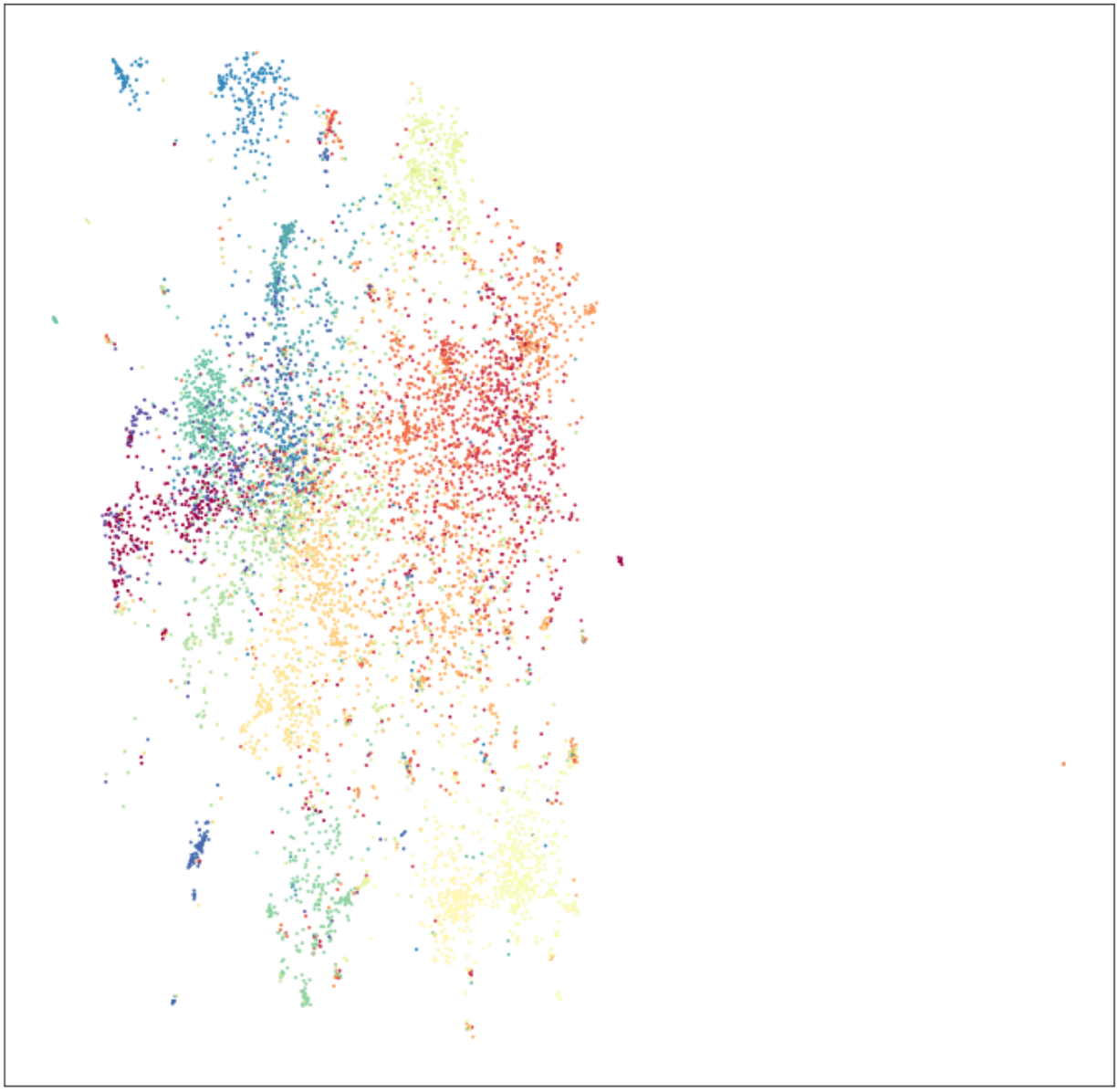

We can see that even going directly from a 130,000 dimensional space down to only 2 dimensions UMAP has done a decent job of separating out many of the different newsgroups.

We can now attempt to add the test data to the same space using the `transform` method.

```
test_embedding = mapper.transform(test_data)
```

While this is somewhat expensive computationally, it does work, and we can plot the end result:

```
fig = plt.figure(figsize=(12,12))
ax = fig.add_subplot(111)
plt.scatter(test_embedding[:, 0], test_embedding[:, 1], s=1, c=news_test.target, cmap=
↪'Spectral')
ax.set(xticks=[], yticks=[]);
```



CHAPTER 8

# UMAP for Supervised Dimension Reduction and Metric Learning

While UMAP can be used for standard unsupervised dimension reduction the algorithm offers significant flexibility allowing it to be extended to perform other tasks, including making use of categorical label information to do supervised dimension reduction, and even metric learning. We'll look at some examples of how to do that below.

First we will need to load some base libraries – `numpy`, obviously, but also `mnist` to read in the Fashion-MNIST data, and matplotlib and seaborn for plotting.

```
import numpy as np
from mnist import MNIST
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set(style='white', context='poster')
```

Our example dataset for this exploration will be the Fashion-MNIST dataset from Zalando Research. It is desgined to be a drop-in replacement for the classic MNIST digits dataset, but uses images of fashion items (dresses, coats, shoes, bags, etc.) instead of handwritten digits. Since the images are more complex it provides a greater challenge than MNIST digits. We can load it in (after downloading the dataset) using the mnist library. We can then package up the train and test sets into one large dataset, normalise the values (to be in the range [0,1]), and set up labels for the 10 classes.

```
mndata = MNIST('fashion-mnist/data/fashion')
train, train_labels = mndata.load_training()
test, test_labels = mndata.load_testing()
data = np.array(np.vstack([train, test]), dtype=np.float64) / 255.0
target = np.hstack([train_labels, test_labels])
classes = [
    'T-shirt/top',
    'Trouser',
    'Pullover',
    'Dress',
    'Coat',
    'Sandal',
    'Shirt',
```

(continues on next page)

(continued from previous page)

```
    'Sneaker',
    'Bag',
    'Ankle boot']
```

Next we'll load the `umap` library so we can do dimension reduction on this dataset.

```
import umap
```

## 8.1 UMAP on Fashion MNIST

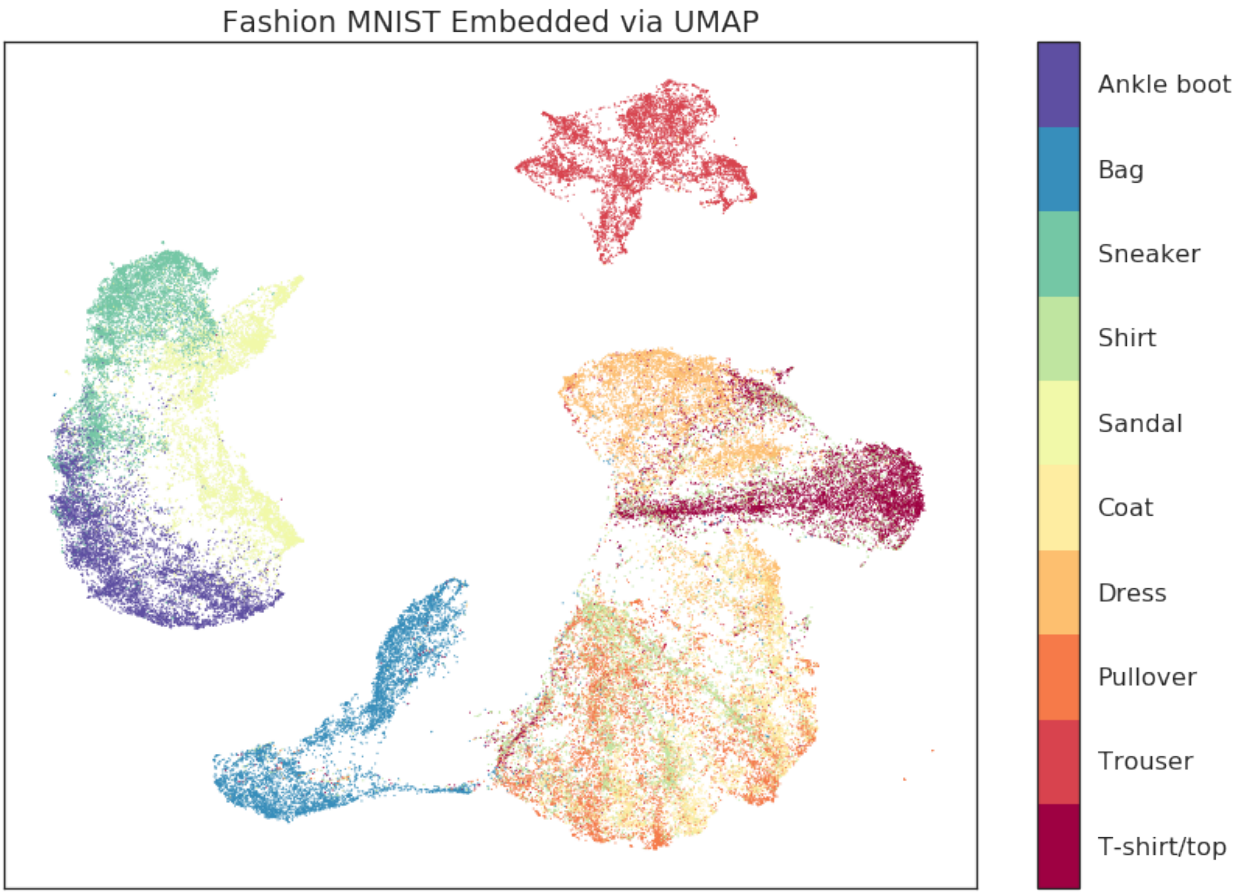
First we'll just do standard unsupervised dimension reduction using UMAP so we have a baseline of what the results look like for later comparison. This is simply a matter of instiantiating a *UMAP* object (in this case setting the `n_neighbors` parameter to be 5 – we are interested mostly in very local information), then calling the *fit_transform()* method with the data we wish to reduce. By default UMAP reduces to two dimensions, so we'll be able to view the results as a scatterplot.

```
%%time
embedding = umap.UMAP(n_neighbors=5).fit_transform(data)
```

```
CPU times: user 1min 45s, sys: 7.22 s, total: 1min 52s
Wall time: 1min 26s
```

That took a little time, but not all that long considering it is 70,000 data points in 784 dimensional space. We can simply plot the results as a scatterplot, colored by the class of the fashion item. We can use matplotlibs colorbar with suitable tick-labels to give us the color key.

```
fig, ax = plt.subplots(1, figsize=(14, 10))
plt.scatter(*embedding.T, s=0.3, c=target, cmap='Spectral', alpha=1.0)
plt.setp(ax, xticks=[], yticks=[])
cbar = plt.colorbar(boundaries=np.arange(11)-0.5)
cbar.set_ticks(np.arange(10))
cbar.set_ticklabels(classes)
plt.title('Fashion MNIST Embedded via UMAP');
```




The result is fairly good. We successfully separated a number of the classes, and the global structure (separating pants and footwear from shirts, coats and dresses) is well preserved as well. Unlike results for MNIST digits, however, there were a number of classes that did not separate quite so cleanly. In particular T-shirts, shirts, dresses, pullovers, and coats are all a little mixed. At the very least the dresses are largely separated, and the T-shirts are mostly in one large clump, but they are not well distinguished from the others. Worse still are the coats, shirts, and pullovers (somewhat unsruprisingly as these can certainly look very similar) which all have significant overlap with one another. Ideally we would like much better class separation. Since we have the label information we can actually give that to UMAP to use!

## 8.2 Using Labels to Separate Classes (Supervised UMAP)

How do we go about coercing UMAP to make use of target labels? If you are familiar with the sklearn API you'll know that the *fit()* method takes a target parameter `y` that specifies supervised target information (for example when training a supervised classification model). We can simply pass the *UMAP* model that target data when fitting and it will make use of it to perform supervised dimension reduction!

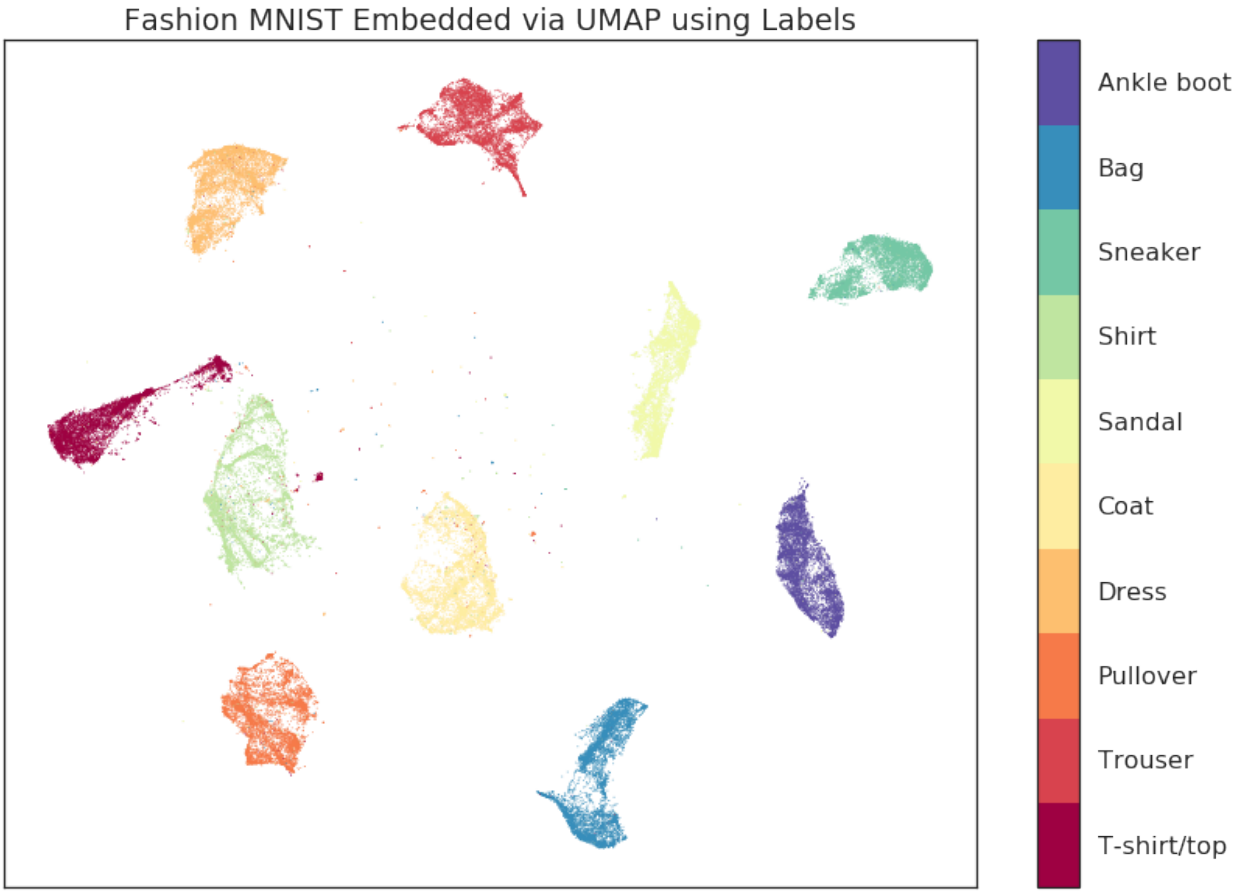
```
%%time
embedding = umap.UMAP().fit_transform(data, y=target)
```

```
CPU times: user 3min 28s, sys: 9.17 s, total: 3min 37s
Wall time: 2min 45s
```

This took a little longer – both because we are using a larger `n_neighbors` value (which is suggested when doing supervised dimension reduction; here we are using the default value of 15), and because we need to condition on the

label data. As before we have reduced the data down to two dimensions so we can again visualize the data with a scatterplot, coloring by class.

```
fig, ax = plt.subplots(1, figsize=(14, 10))
plt.scatter(*embedding.T, s=0.1, c=target, cmap='Spectral', alpha=1.0)
plt.setp(ax, xticks=[], yticks=[])
cbar = plt.colorbar(boundaries=np.arange(11)-0.5)
cbar.set_ticks(np.arange(10))
cbar.set_ticklabels(classes)
plt.title('Fashion MNIST Embedded via UMAP using Labels');
```




The result is a cleanly separated set of classes (and a little bit of stray noise – points that were sufficiently different from their class as to not be grouped with the rest). Aside from the clear class separation however (which is expected – we gave the algorithm all the class information), there are a couple of important points to note. The first point to note is that we have retained the internal structure of the individual classes. Both the shirts and pullovers still have the distinct banding pattern that was visible in the original unsupervised case; the pants, t-shirts and bags both retained their shape and internal structure; etc. The second point to note is that we have also retained the global structure. While the individual classes have been cleanly seprated from one another, the inter-relationships among the classes have been preserved: footwear classes are all near one another; trousers and bags are at opposite sides of the plot; and the arc of pullover, shirts, t-shirts and dresses is still in place.

The key point is this: the important structural properties of the data have been retained while the known classes have been cleanly pulled apart and isolated. If you have data with known classes and want to seprate them while still having a meaningful embedding of individual points then supervised UMAP can provide exactly what you need.

## 8.3 Using Partial Labelling (Semi-Supervised UMAP)

What if we only have some of our data labelled, however, and a number of items are without labels. Can we still make use of the label information we do have? This is now a semi-supervised learning problem, and yes, we can work with those cases to. To set up the example we'll mask some of the target information – we'll do this by using the sklearn standard of having unlabelled point be given the label of -1 (such as, for example, the noise points from a DBSCAN clustering).

```
masked_target = target.copy().astype(np.int8)
masked_target[np.random.choice(70000, size=10000, replace=False)] = -1
```

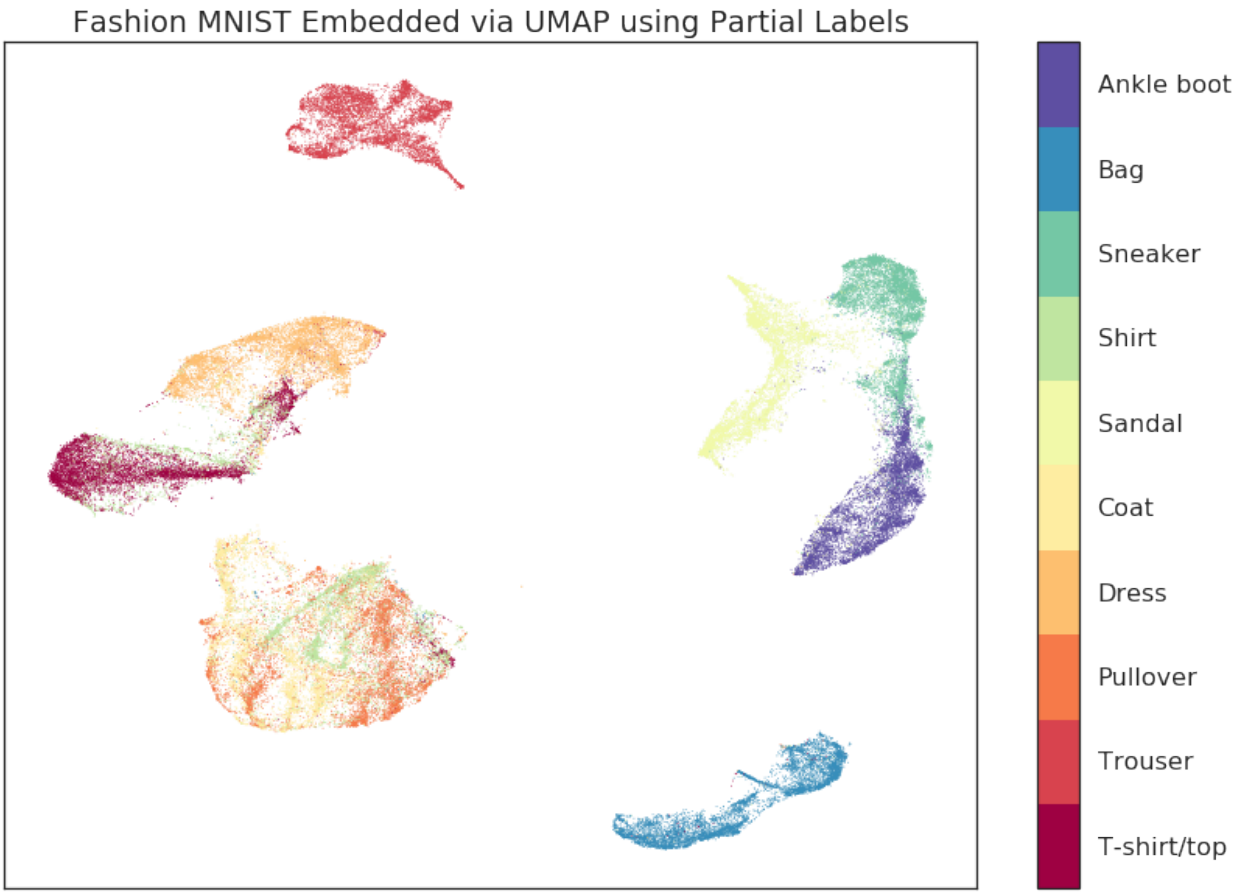
Now that we have randomly masked some of the labels we can try to perform supervised learning again. Everything works as before, but UMAP will interpret the -1 label as beingan unlabelled point and learn accordingly.

```
%%time
fitter = umap.UMAP().fit(data, y=masked_target)
embedding = fitter.embedding_
```

```
CPU times: user 3min 8s, sys: 7.85 s, total: 3min 16s
Wall time: 2min 40s
```

Again we can look at a scatterplot of the data colored by class.

```
fig, ax = plt.subplots(1, figsize=(14, 10))
plt.scatter(*embedding.T, s=0.1, c=target, cmap='Spectral', alpha=1.0)
plt.setp(ax, xticks=[], yticks=[])
cbar = plt.colorbar(boundaries=np.arange(11)-0.5)
cbar.set_ticks(np.arange(10))
cbar.set_ticklabels(classes)
plt.title('Fashion MNIST Embedded via UMAP using Partial Labels');
```




The result is much as we would expect – while we haven't cleanly separated the data as we did in the totally supervised case, the classes have been made cleaner and more distinct. This semi-supervised approach provides a powerful tool when labelling is potentially expensive, or when you have more data than labels, but want to make use of that extra data.

## 8.4 Training with Labels and Embedding Unlabelled Test Data (Metric Learning with UMAP)

If we have learned a supervised embedding, can we use that to embed new previously unseen (and now unlabelled) points into the space? This would provide an algorithm for metric learning, where we can use a labelled set of points to learn a metric on data, and then use that learned metric as a measure of distance between new unlabelled points. This can be particularly useful as part of a machine learning pipline where we learn a supervised embedding as a form of supervised feature engineering, and then build a classifier on that new space – this is viable as long as we can pass new data to the embedding model to be transformed to the new space.

To try this out with UMAP let's use the train/test split provided by Fashion MNIST:

```
train_data = np.array(train)
test_data = np.array(test)
```

Now we can fit a model to the training data, making use of the training labels to learn a supervised embedding.

```
%%time
mapper = umap.UMAP(n_neighbors=10).fit(train_data, np.array(train_labels))
```

```
CPU times: user 2min 18s, sys: 7.53 s, total: 2min 26s
Wall time: 1min 52s
```

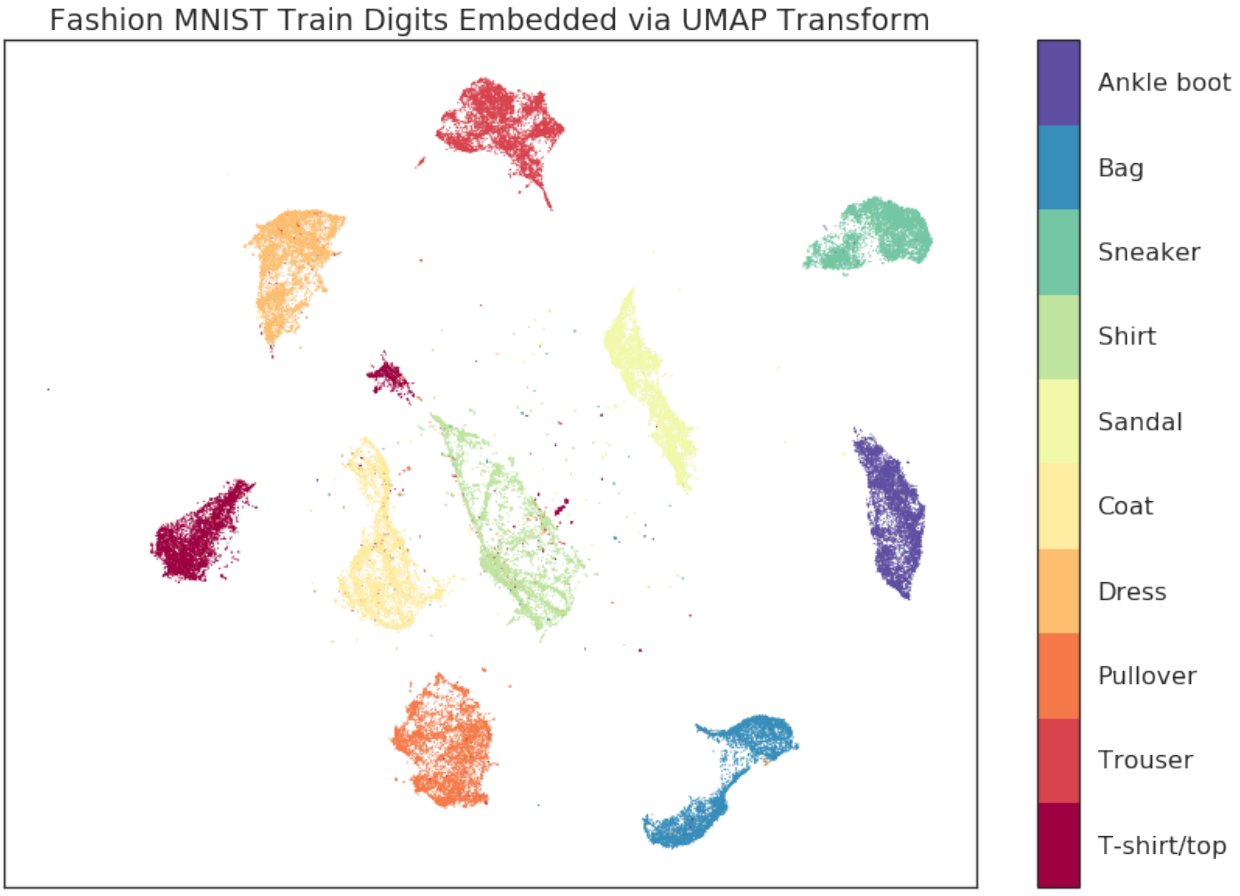
Next we can use the *transform()* method on that model to transform the test set into the learned space. This time we won't pass the label information and let the model attempt to place the data correctly.

```
%%time
test_embedding = mapper.transform(test_data)
```

```
CPU times: user 17.3 s, sys: 986 ms, total: 18.3 s
Wall time: 15.4 s
```

UMAP transforms are not as fast as some approaches, but as you can see this was still fairly efficient. The important question is how well we managed to embed the test data into the existing learned space. To start let's visualise the embedding of the training data so we can get a sense of where things *should* go.

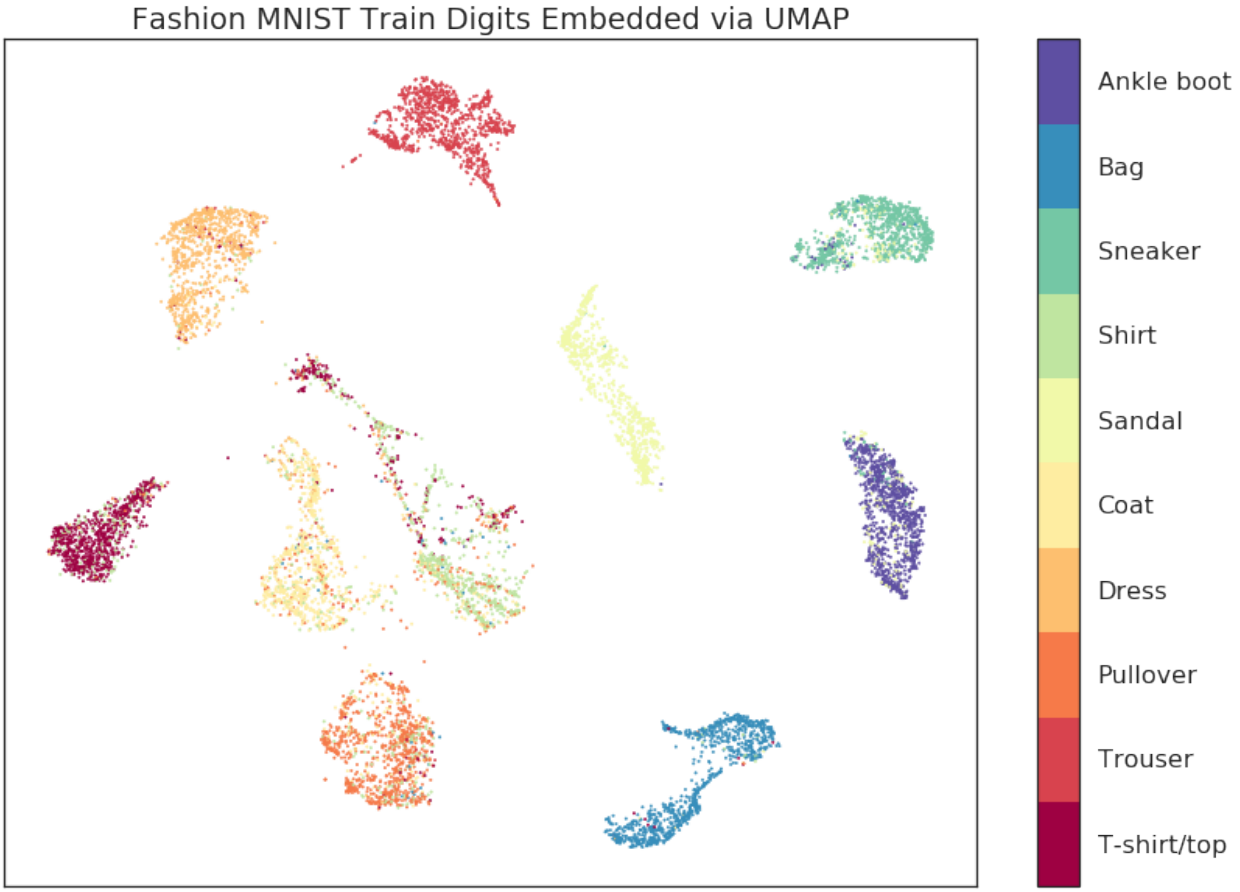
```
fig, ax = plt.subplots(1, figsize=(14, 10))
plt.scatter(*mapper.embedding_.T, s=0.3, c=np.array(train_labels), cmap='Spectral',␣
↪alpha=1.0)
plt.setp(ax, xticks=[], yticks=[])
cbar = plt.colorbar(boundaries=np.arange(11)-0.5)
cbar.set_ticks(np.arange(10))
cbar.set_ticklabels(classes)
plt.title('Fashion MNIST Train Digits Embedded via UMAP Transform');
```




As you can see this has done a similar job as before, successfully embedding the separate classes while retaining both

the internal structure and the overall global structure. We can now look at how the test set, for which we provided no label information, was embedded via the *transform()* method.

```
fig, ax = plt.subplots(1, figsize=(14, 10))
plt.scatter(*test_embedding.T, s=2, c=np.array(test_labels), cmap='Spectral', alpha=1.
↪0)
plt.setp(ax, xticks=[], yticks=[])
cbar = plt.colorbar(boundaries=np.arange(11)-0.5)
cbar.set_ticks(np.arange(10))
cbar.set_ticklabels(classes)
plt.title('Fashion MNIST Train Digits Embedded via UMAP');
```




As you can see we have replicated the layout of the training data, including much of the internal structure of the classes. For the most part assignment of new points follows the classes well. The greatest source of confusion in some t-shirts that ended up in mixed with the shirts, and some pullovers which are confused with the coats. Given the difficulty of the problemn this is a good result, particularly when compared with current state-of-the-art approaches such as siamese and triplet networks.

# CHAPTER 9

# Using UMAP for Clustering

UMAP can be used as an effective preprocessing step to boost the performance of density based clustering. This is somewhat controversial, and should be attempted with care. For a good discussion of some of the issues involved in this please see the various answers in this stackoverflow thread on clustering the results of t-SNE. Many of the points of concern raised there are salient for clustering the results of UMAP. The most notable is that UMAP, like t-SNE, does not completely preserve density. UMAP, like t-SNE, can also create tears in clusters that are not actually present, resulting in a finer clustering than is necessarily present in the data. Despite these concerns there are still valid reasons to use UMAP as a preprocessing step for clustering. As with any clustering approach one will want to do some exploration and evaluation of the clusters that come out to try to validate them if possible.

With all of that said, let's work through an example to demonstrate the difficulties that can face clustering approaches and how UMAP can provide a powerful tool to help overcome them.

First we'll need a selection of libraries loaded up. Obviously we'll need data, and we can use sklearn's `fetch_mldata` to get it. We'll also need the usual tools of numpy, and plotting. Next we'll need umap, and some clustering options. Finally, since we'll be working with labeled data, we can make use of strong cluster evaluation metrics Adjusted Rand Index and Adjusted Mutual Information.

```
from sklearn.datasets import fetch_mldata
from sklearn.decomposition import PCA
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# Dimension reduction and clustering libraries
import umap
import hdbscan
import sklearn.cluster as cluster
from sklearn.metrics import adjusted_rand_score, adjusted_mutual_info_score
```

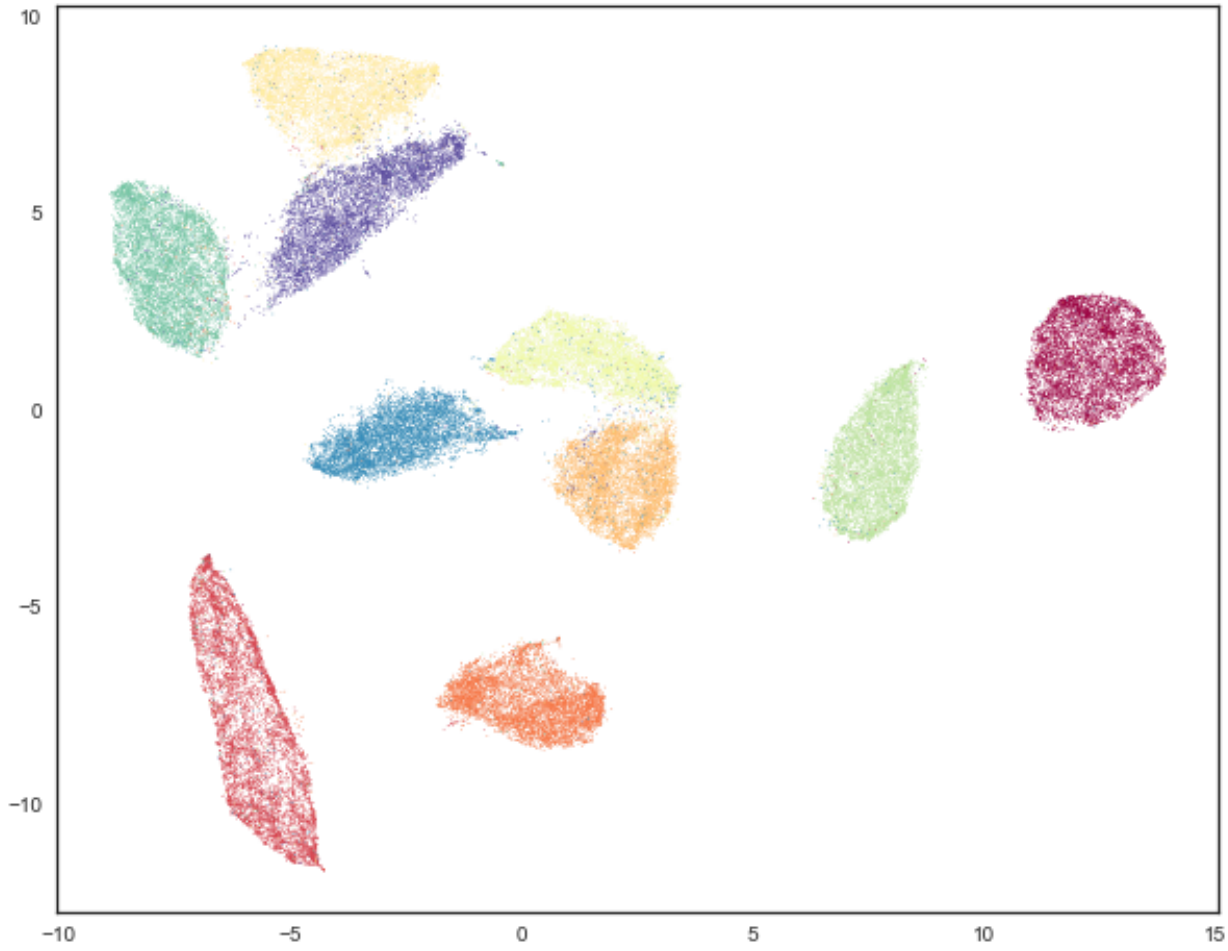
Now let's set up the plotting and grab the data we'll be using – in this case the MNIST handwritten digits dataset. MNIST consists of 28x28 pixel grayscale images of handwritten digits (0 through 9). These can be unraveled such that each digit is described by a 784 dimensional vector (the gray scale value of each pixel in the image). Ideally we would like the clustering to recover the digit structure.

```
sns.set(style='white', rc={'figure.figsize':(10,8)})
```

```
mnist = fetch_mldata('MNIST Original')
```

For visualization purposes we can reduce the data to 2-dimensions using UMAP. When we cluster the data in high dimensions we can visualize the result of that clustering. First, however, we'll view the data colored by the digit that each data point represents – we'll use a different color for each digit. This will help frame what follows.

```
standard_embedding = umap.UMAP(random_state=42).fit_transform(mnist.data)
plt.scatter(standard_embedding[:, 0], standard_embedding[:, 1], c=mnist.target, s=0.1,
↪ cmap='Spectral');
```



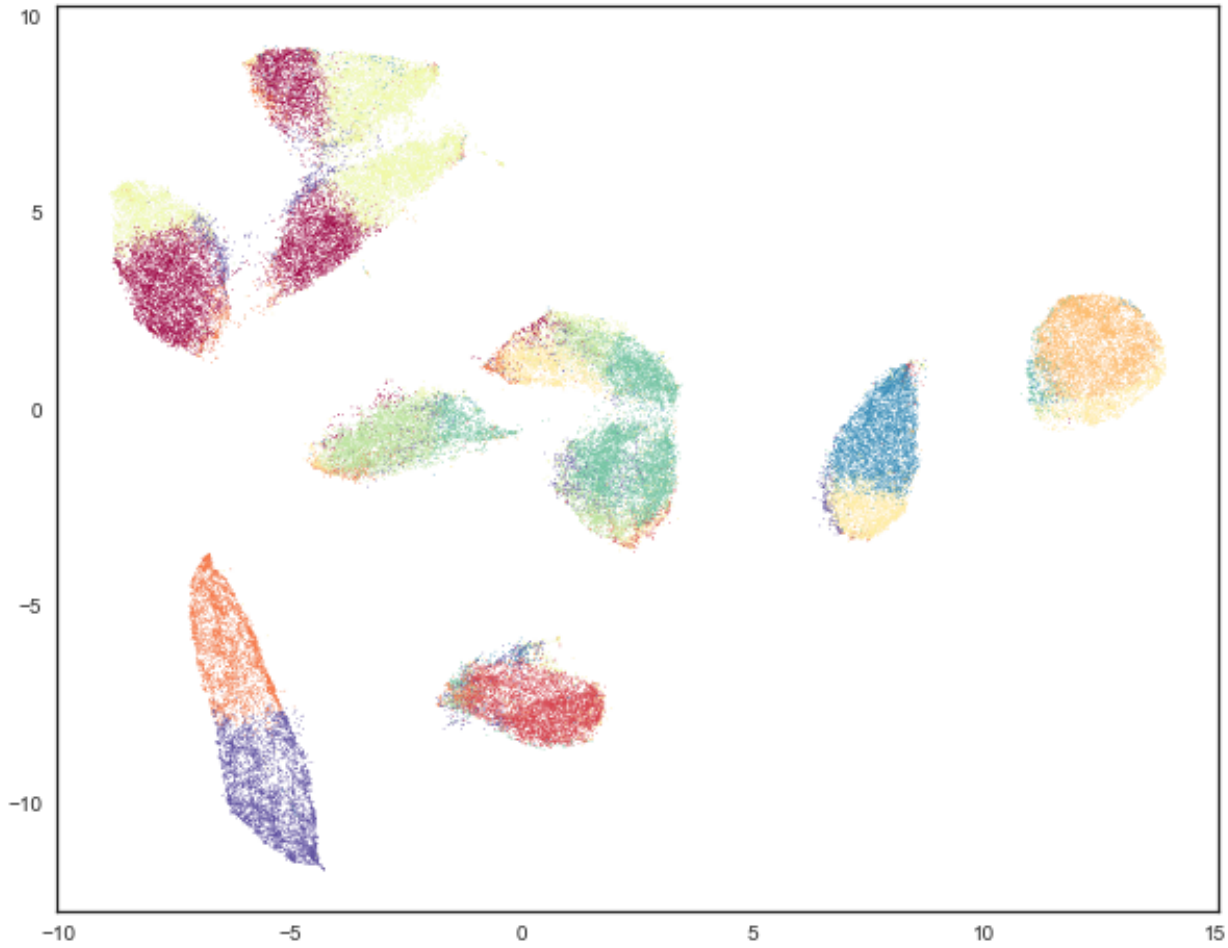

## 9.1 Traditional clustering

Now we would like to cluster the data. As a first attempt let's try the traditional approach: K-Means. In this case we can solve one of the hard problems for K-Means clustering – choosing the right k value, giving the number of clusters we are looking for. In this case we know the answer is exactly 10. We will use sklearns K-Means implementation looking for 10 clusters in the original 784 dimensional data.

```
kmeans_labels = cluster.KMeans(n_clusters=10).fit_predict(mnist.data)
```

And how did the clustering do? We can look at the results by coloring out UMAP embedded data by cluster membership.

```
plt.scatter(standard_embedding[:, 0], standard_embedding[:, 1], c=kmeans_labels, s=0.
↪1, cmap='Spectral');
```




This is not really the result we were looking for (though it does expose interesting properties of how K-Means chooses clusters in high dimensional space, and how UMAP unwraps manifolds by finding manifold boundaries). While K-Means gets some cases correct – the two clusters are the far right are mostly correct, most of the rest of the data looks somewhat arbitrarily carved up among the remaining clusters. We can put this impression to the test by evaluating the adjusted Rand score and adjusted mutual information for this clustering as compared with the true labels.

```
(
    adjusted_rand_score(mnist.target, kmeans_labels),
    adjusted_mutual_info_score(mnist.target, kmeans_labels)
)
```

```
(0.36675295135972552, 0.49614118437750965)
```

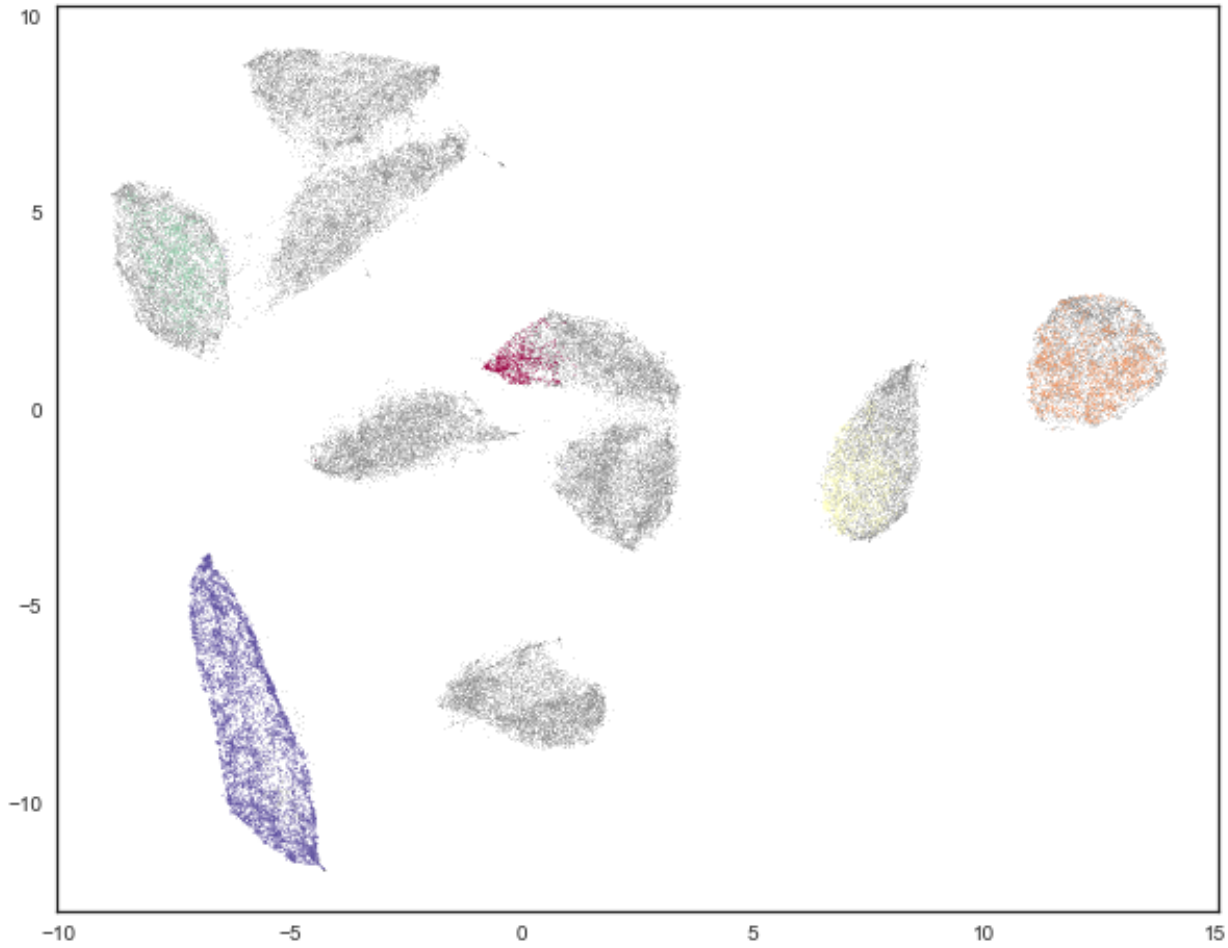
As might be expected, we have not done a particularly good job – both scores take values in the range 0 to 1, with 0 representing a bad (essentially random) clustering and 1 representing perfectly recovering the true labels. K-Means definitely was not random, but it was also quite a long way from perfectly recovering the true labels. Part of the problem is the way K-Means works, based on centroids with an assumption of largely spherical clusters – this is responsible for some of the sharp divides that K-Means puts across digit classes. We can potentially improve on this

by using a smarter density based algorithm. In this case we've chosen to try HDBSCAN, which we believe to be among the most advanced density based tehcniques. For the sake of performance we'll reduce the dimensionality of the data down to 50 dimensions via PCA (this recovers most of the variance), since HDBSCAN scales somewhat poorly with the dimensionality of the data it will work on.

```
lowd_mnist = PCA(n_components=50).fit_transform(mnist.data)
hdbscan_labels = hdbscan.HDBSCAN(min_samples=10, min_cluster_size=500).fit_
↪predict(lowd_mnist)
```

We can now inspect the results. Before we do, however, it should be noted that one of the features of HDBSCAN is that it can refuse to cluster some points and classify the as "noise". To visualize this aspect we will color points that were classified as noise gray, and then color the remaining points according to the cluster membership.

```
clustered = (hdbscan_labels >= 0)
plt.scatter(standard_embedding[~clustered, 0],
            standard_embedding[~clustered, 1],
            c=(0.5, 0.5, 0.5),
            s=0.1,
            alpha=0.5)
plt.scatter(standard_embedding[clustered, 0],
            standard_embedding[clustered, 1],
            c=hdbscan_labels[clustered],
            s=0.1,
            cmap='Spectral');
```



This looks somewhat underwhelming. It meets HDBSCAN's approach of "not being wrong" by simply refusing the classify the majority of the data. The result is a clustering that almost certainly fails to recover all the labels. We can verify this by looking at the clustering validation scores.

```
(
    adjusted_rand_score(mnist.target, hdbscan_labels),
    adjusted_mutual_info_score(mnist.target, hdbscan_labels)
)
```

```
(0.053830107882840102, 0.19756104096566332)
```

These scores are far worse than K-Means! Partially this is due to the fact that these scores assume that the noise points are simply an extra cluster. We can instead only look at the subset of the data that HDBSCAN was actually confident enough to assign to clusters – a simple sub-selection will let us recompute the scores for only that data.

```
clustered = (hdbscan_labels >= 0)
(
    adjusted_rand_score(mnist.target[clustered], hdbscan_labels[clustered]),
    adjusted_mutual_info_score(mnist.target[clustered], hdbscan_labels[clustered])
)
```

```
(0.99843407988303912, 0.99405521087764015)
```

And here we see that where HDBSCAN was willing to cluster it got things almost entirely correct. This is what it was designed to do – be right for what it can, and defer on anything that it couldn't have sufficient confidence in. Of course the catch here is that it deferred clustering a lot of the data. How much of the data did HDBSCAN actually assign to clusters? We can compute that easily enough.

```
np.sum(clustered) / mnist.data.shape[0]
```

```
0.17081428571428572
```

It seems that less than 18% of the data was clustered. While HDBSCAN did a great job on the data it could cluster it did a poor job of actually managing to cluster the data. The problem here is that, as a density based clustering algorithm, HDBSCAN tends to suffer from the curse of dimensionality: high dimensional data requires more observed samples to produce much density. If we could reduce the dimensionality of the data more we would make the density more evident and make it far easier for HDBSCAN to cluster the data. The problem is that trying to use PCA to do this is going to become problematic. While reducing the 50 dimensions still explained a lot of the variance of the data, reducing further is going to quickly do a lot worse. This is due to the linear nature of PCA. What we need is strong manifold learning, and this is where UMAP can come into play.

## 9.2 UMAP enhanced clustering

Our goal is to make use of UMAP to perform non-linear manifold aware dimension reduction so we can get the dataset down to a number of dimensions small enough for a density based clustering algorithm to make progress. One advantage of UMAP for this is that it doesn't require you to reduce to only two dimensions – you can reduce to 10 dimensions instead since the goal is to cluster, not visualize, and the performance cost with UMAP is minimal. As it happens MNIST is such a simple dataset that we really can push it all the way down to only two dimensions, but in general you should explore different embedding dimension options.

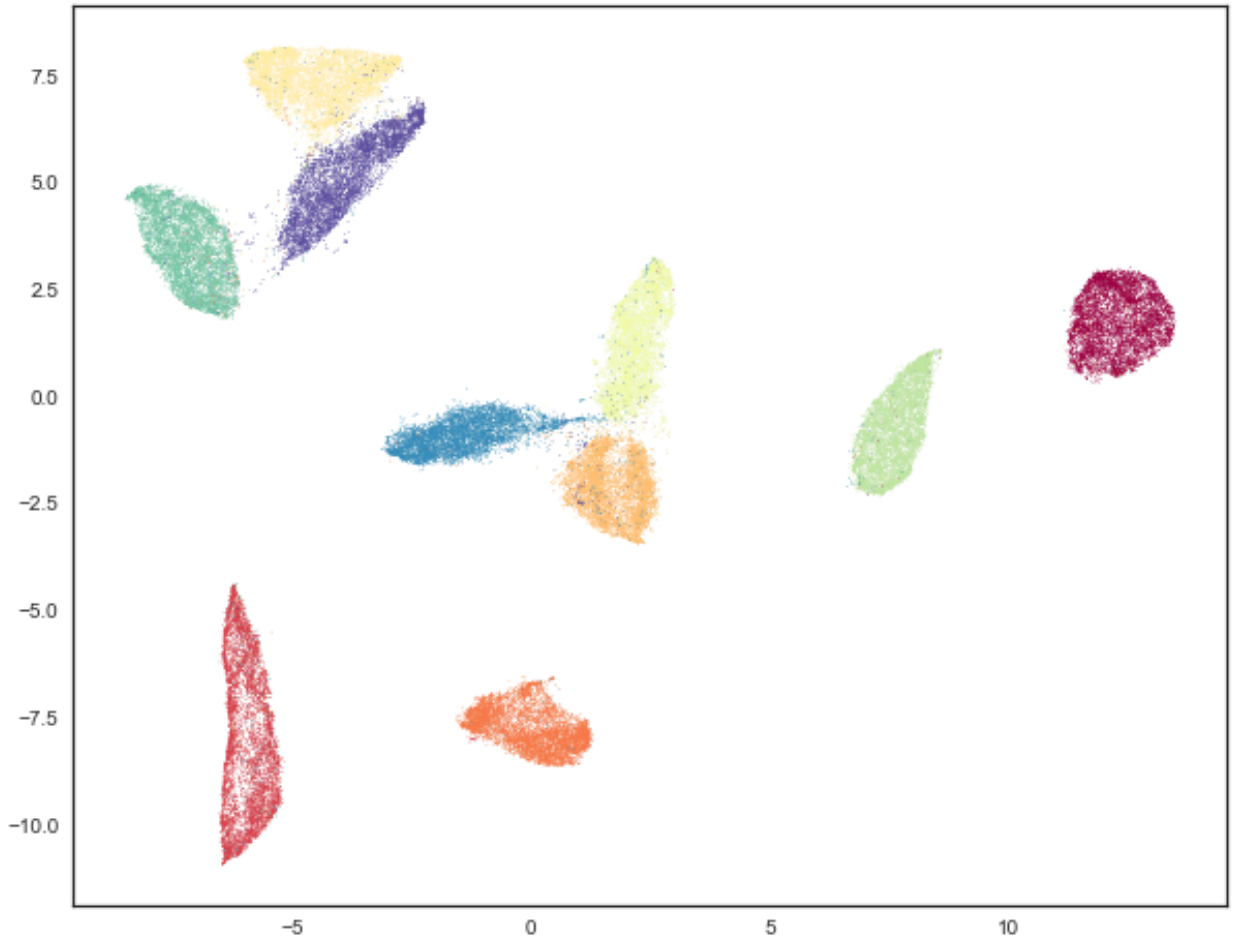
The next thing to be aware of is that when using UMAP for dimension reduction you will want to select different parameters than if you were using it for visualization. First of all we will want a larger `n_neighbors` value – small values will focus more on very local structure and are more prone to producing fine grained cluster structure that may be more a result of patterns of noise in the data than actual clusters. In this case we'll double it from the

default 15 up to 30. Second it is beneficial to set `min_dist` to a very low value. Since we actually want to pack points together densely (density is what we want after all) a low value will help, as well as making cleaner separations between clusters. In this case we will simply set `min_dist` to be 0.

```
clusterable_embedding = umap.UMAP(
    n_neighbors=30,
    min_dist=0.0,
    n_components=2,
    random_state=42,
).fit_transform(mnist.data)
```

We can visualize the results of this so see how it compares with more visualization attuned parameters:

```
plt.scatter(clusterable_embedding[:, 0], clusterable_embedding[:, 1],
            c=mnist.target, s=0.1, cmap='Spectral');
```




As you can see we still have the general global structure, but we are packing points together more tightly within clusters, and consequently we can see larger gaps between the clusters. Ultimately this embedding was for clustering purposes only, and we will go back to the original embedding for visualization purposes from here on out.

The next step is to cluster this data. We'll use HDBSCAN again, with the same parameter setting as before.

```
labels = hdbscan.HDBSCAN(
    min_samples=10,
```

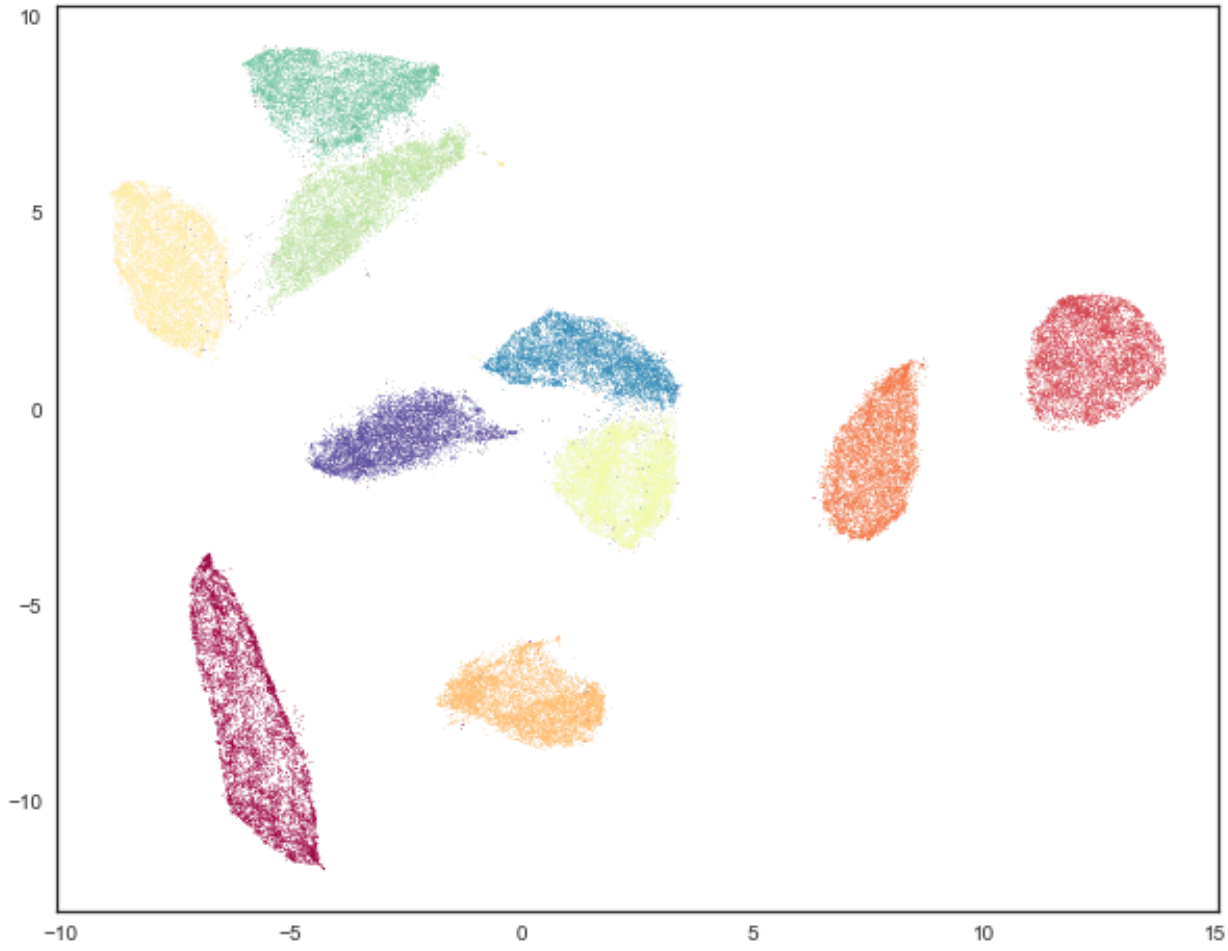
(continues on next page)

(continued from previous page)

```
    min_cluster_size=500,
).fit_predict(clusterable_embedding)
```

And now we can visualize the results, just as before.

```
clustered = (labels >= 0)
plt.scatter(standard_embedding[~clustered, 0],
            standard_embedding[~clustered, 1],
            c=(0.5, 0.5, 0.5),
            s=0.1,
            alpha=0.5)
plt.scatter(standard_embedding[clustered, 0],
            standard_embedding[clustered, 1],
            c=labels[clustered],
            s=0.1,
            cmap='Spectral');
```




We can see that we have done a much better job of finding clusters rather than merely assigning the majority of data as noise. This is because we no longer have to try to cope with the relative lack of density in 50 dimensional space and now HDBSCAN can more cleanly discern the clusters.

We can also make a quantitative assessment by using the clustering quality measures as before.

```
adjusted_rand_score(mnist.target, labels), adjusted_mutual_info_score(mnist.target,␣
↪labels)
```

```
(0.9239306564265013, 0.903026716411337 36)
```

Where before HDBSCAN performed very poorly, we now have score of 0.9 or better. This is because we actually clustered far more of the data. As before we can also look at how the clustering did on just the data that HDBSCAN was confident in clustering.

```
clustered = (labels >= 0)
(
    adjusted_rand_score(mnist.target[clustered], labels[clustered]),
    adjusted_mutual_info_score(mnist.target[clustered], labels[clustered])
)
```

```
(0.93240371696811541, 0.91912906363537572)
```

This is a little worse than the original HDBSCAN, but it is unsurprising that you are going to be wrong more often if you make more predictions. The question is how much more of the data is HDBSCAN actually clustering? Previously we were clustering only 17% of the data.

```
np.sum(clustered) / mnist.data.shape[0]
```

```
0.99164285714285716
```

Now we are clustering over 99% of the data! And our results in terms of adjusted Rand score and adjusted mutual information are in line with the current state of the art techniques using convolutional autoencoder techniques. That's not bad for an approach that is simply viewing the data as arbitrary 784 dimensional vectors.

Hopefully this has outlined how UMAP can be beneficial for clustering. As with all thing care must be taken, but clearly UMAP can provide significantly better clustering results when used judiciously.

# CHAPTER 10

## Outlier detection using UMAP

While an earlier tutorial looked at using UMAP for clustering, it can also be used for outlier detection, providing that some care is taken. This tutorial will look at how to use UMAP in this manner, and what to look out for, by finding anomalous digits in the MNIST handwritten digits dataset. To start with let's load the relevant libraries:

```
import numpy as np
import sklearn.datasets
import sklearn.neighbors
import umap
import umap.plot
import matplotlib.pyplot as plt
%matplotlib inline
```

With this in hand, let's grab the MNIST digits dataset from the internet, using the new `fetch_ml` loader in sklearn.

```
data, labels = sklearn.datasets.fetch_openml('mnist_784', version=1, return_X_y=True)
```

Before we get started we should try looking for outliers in terms of the native 784 dimensional space that MNIST digits live in. To do this we will make use of the Local Outlier Factor (LOF) method for determining outliers since sklearn has an easy to use implementation. The essential intuition of LOF is to look for points that have a (locally approximated) density that differs significantly from the average density of their neighbors. In our case the actual details are not so important – it is enough to know that the algorithm is reasonably robust and effective on vector space data. We can apply it using the `fit_predict` method of the sklearn class. The LOF class take a parameter `contamination` which specifies the percentage of data that the user expects to be noise. For this use case we will set it to 0.001428 since, given the 70,000 samples in MNIST, this will result in 100 outliers, which we can then look at in more detail.

```
%%time
outlier_scores = sklearn.neighbors.LocalOutlierFactor(contamination=0.001428).fit_
↪predict(data)
```

```
CPU times: user 1h 29min 10s, sys: 12.4 s, total: 1h 29min 22s
Wall time: 1h 29min 53s
```

It is worth noting how long that took. Over an hour and a half! Why did it take so long? Because LOF requires a notion of density, which in turn relies on a nearest neighbor type computation – which is expensive in sklearn for high dimensional data. This alone is potentially a reason to look at reducing the dimension of the data – it makes it more amenable to existing techniques like LOF.

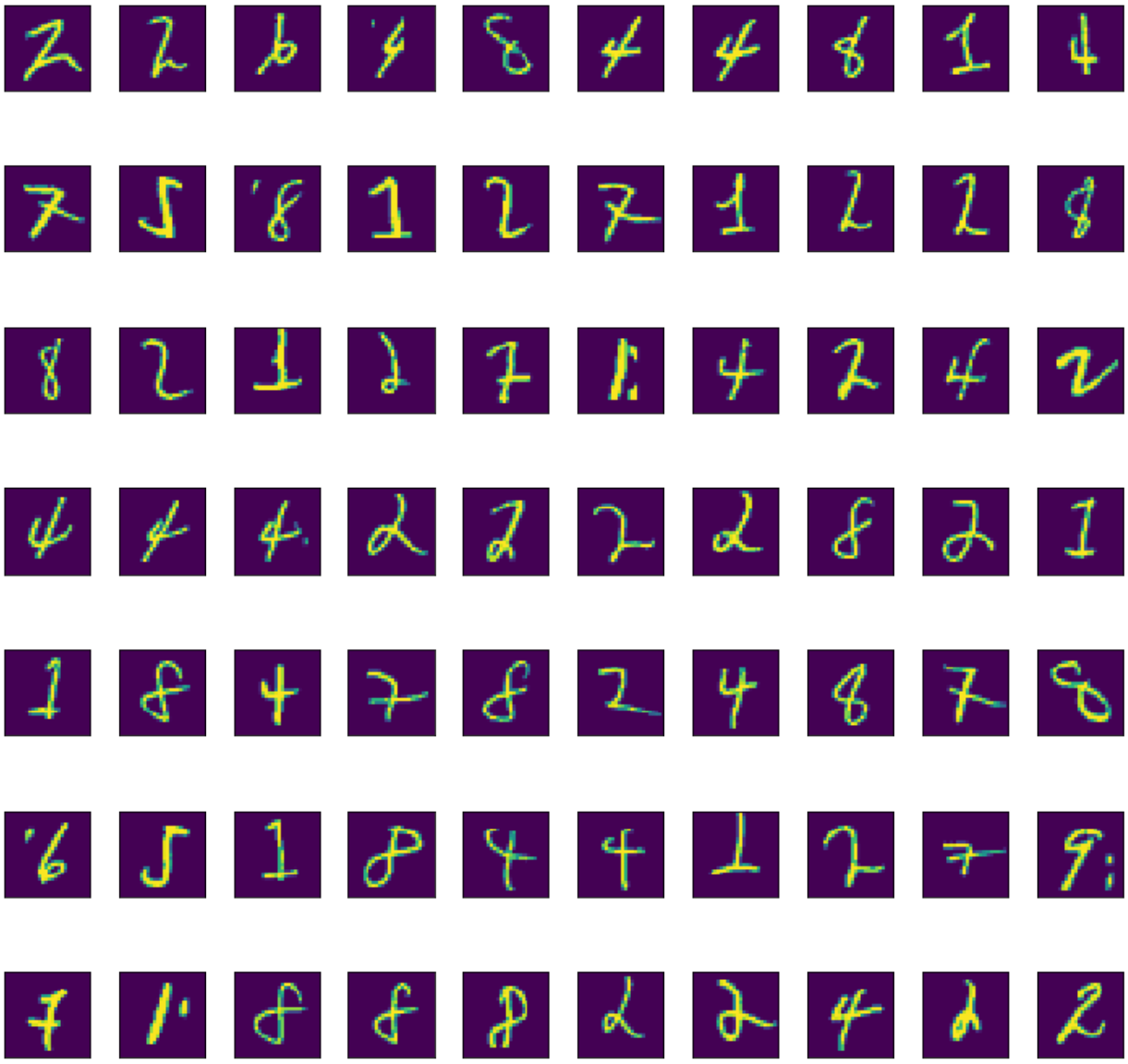
Now that we have a set of outlier scores we can find the actual outlying digit images – these are the ones with scores equal to -1. Let's extract out that data, and check that we got 100 different digit images.

```
outlying_digits = data[outlier_scores == -1]
outlying_digits.shape
```

```
(100, 784)
```

Now that we have the outlying digit images the first question we should be asking is "what do they look like?". Fortunately for us we can convert the 784 dimensional vectors back into image and plot them, making it easier to look at. Since we extracted the 100 most outlying digit images we can just display a 10x10 grid of them.

```
fig, axes = plt.subplots(7, 10, figsize=(10,10))
for i, ax in enumerate(axes.flatten()):
    ax.imshow(outlying_digits[i].reshape((28,28)))
    plt.setp(ax, xticks=[], yticks=[])
plt.tight_layout()
```



These do certainly look like somewhat strange looking handwritten digits, so our outlier detection seems to be working to some extent.

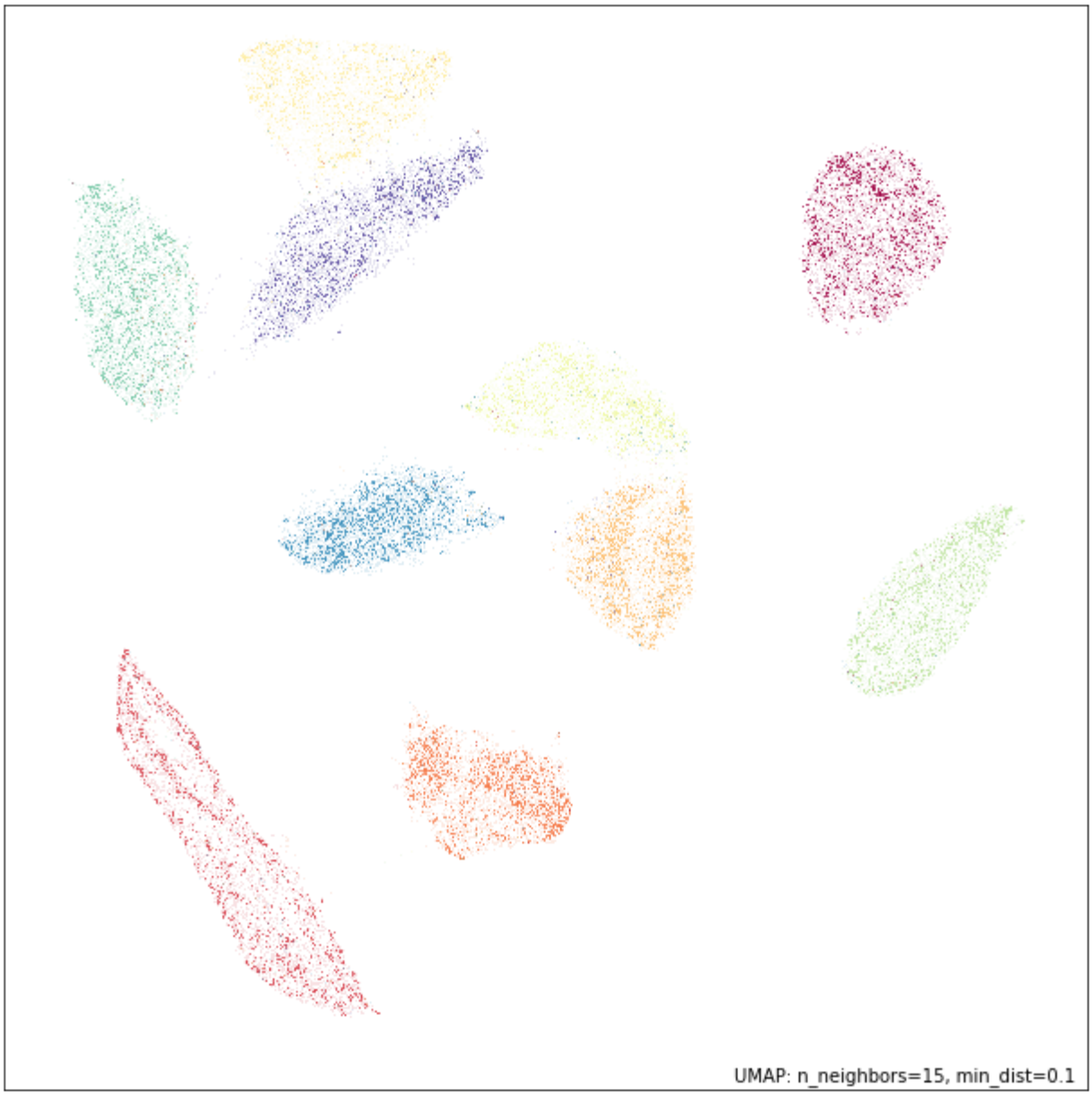
Now let's try a naive approach using UMAP and see how far that gets us. First let's just apply UMAP directly with default parameters to the MNIST data.

```
mapper = umap.UMAP().fit(data)
```

Now we can see what we got using the new plotting tools in umap.plot.

```
umap.plot.points(mapper, labels=labels)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1c3db71358>
```




That looks like what we have come to expect from a UMAP embedding of MNIST. The question is have we managed to preserve outliers well enough that LOF can still find the bizarre digit images, or has the embedding lost that information and contracted the outliers into the individual digit clusters? We can simply apply LOF to the embedding and see what that returns.

```
%%time
outlier_scores = sklearn.neighbors.LocalOutlierFactor(contamination=0.001428).fit_
↪predict(mapper.embedding_)
```

This was obviously much faster since we are operating in a much lower dimensional space that is more amenable to the spatial indexing methods that sklearn uses to find nearest neighbors. As before we extract the outlying digit images, and verify that we got 100 of them,

```
outlying_digits = data[outlier_scores == -1]
```

(continues on next page)

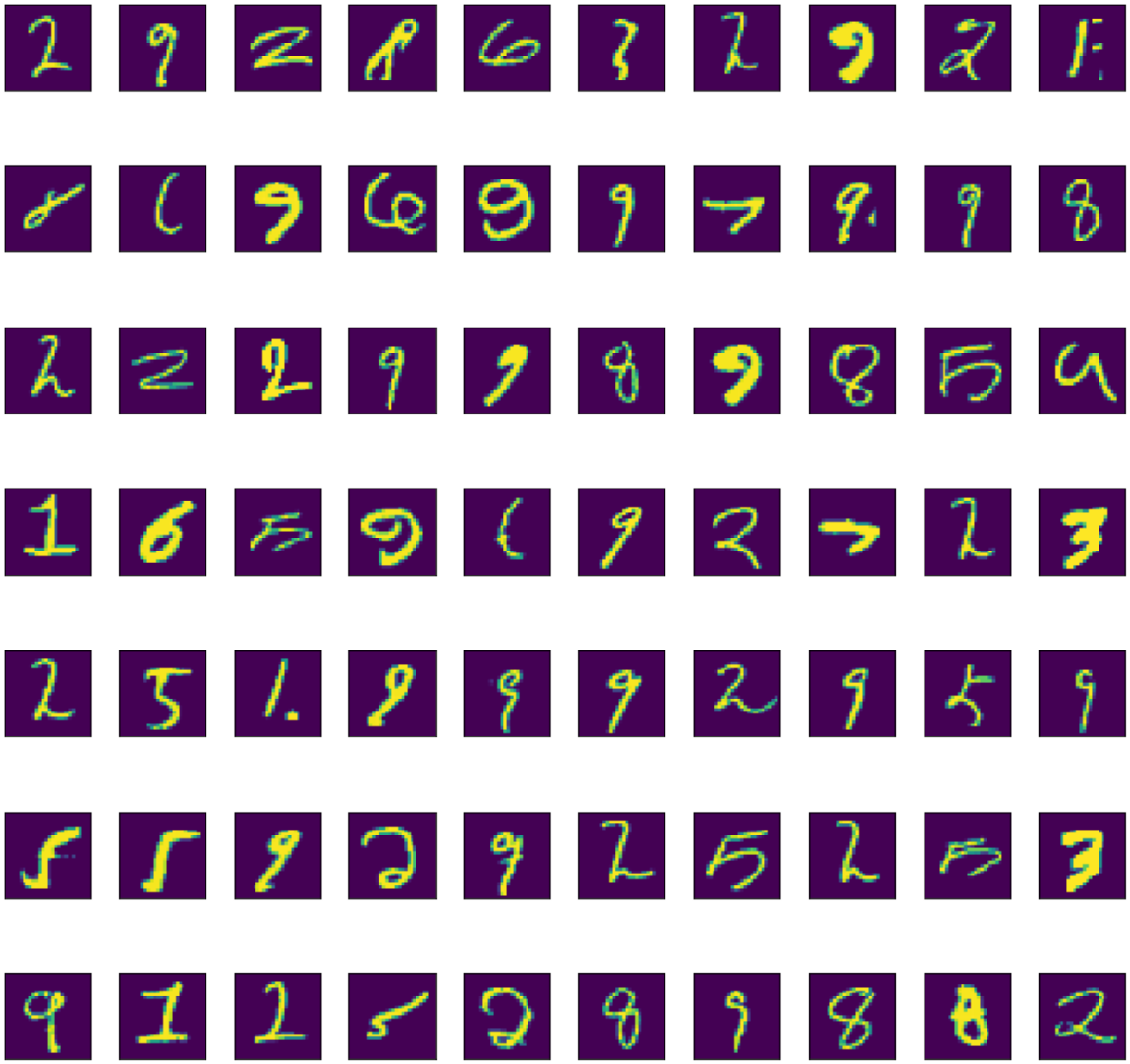
(continued from previous page)

```
outlying_digits.shape
```

```
(100, 784)
```

Now we need to plot the outlying digit images to see what kinds of digit images this approach found to be particularly strange.

```
fig, axes = plt.subplots(7, 10, figsize=(10,10))
for i, ax in enumerate(axes.flatten()):
    ax.imshow(outlying_digits[i].reshape((28,28)))
    plt.setp(ax, xticks=[], yticks=[])
plt.tight_layout()
```



In many way this looks to be a *better* result than the original LOF in the high dimensional space. While the digit images that the high dimensional LOF found to be strange were indeed somewhat odd looking, many of these digit images are considerably stranger – significantly odd line thickness, warped shapes, and images that are hard to even

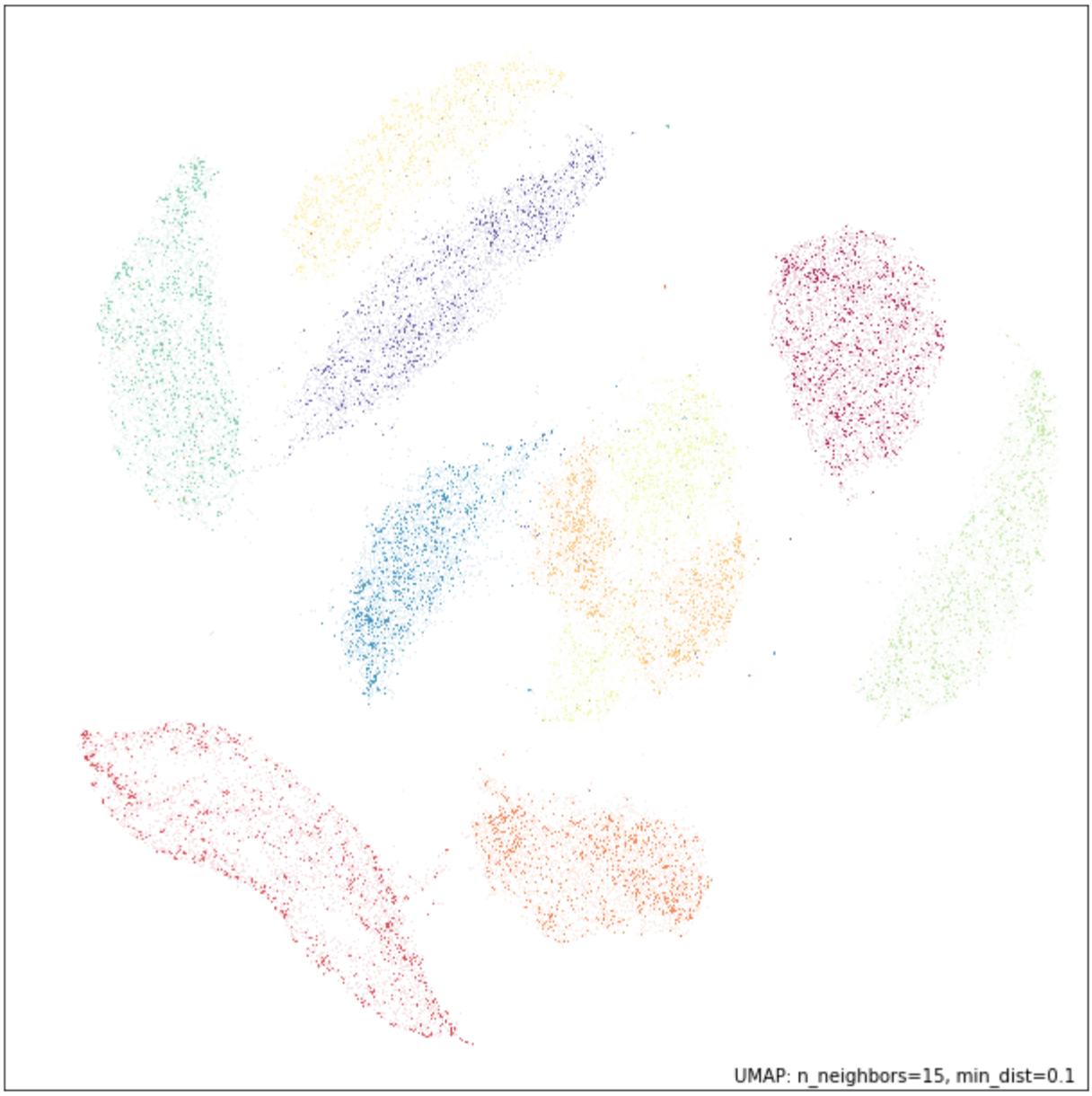
recognise as digits. This helps to demonstrate a certain amount of confirmation bias when examining outliers: since we expect things tagged as outliers to be strange we tend to find aspects of them that justify that classification, potentially unaware of how much stranger some of the data may in fact be. This should make us wary of even this outlier set: what else might lurk in the dataset?

We can, in fact, potentially improve on this result by tuning the UMAP embedding a little for the task of finding outliers. When UMAP combines together the different local simplicial sets (see how_umap_works for more details) the standard approach uses a union, but we could instead take an intersection. An intersection ensures that outliers remain disconnected, which is certainly beneficial when seeking to find outliers. A downside of the intersection is that it tends to break up the resulting simplicial set into many disconnected components and a lot of the more non-local and global structure is lost, resulting in a lot lower quality of embedding. We can, however, interpolate between the union and intersection. In UMAP this is given by the `set_op_mix_ratio`, where a value of 0.0 represents an intersection, and a value of 1.0 represents a union (the default value is 1.0). By setting this to a lower value, say 0.25, we can encourage the embedding to do a better job of preserving outliers as outlying, while still retaining the benefits of a union operation.

```
mapper = umap.UMAP(set_op_mix_ratio=0.25).fit(data)
```

```
umap.plot.points(mapper, labels=labels)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1c3f496908>
```




As you can see the embedding is not as well structured overall as when we had a `set_op_mix_ratio` of 1.0, but we have potentially done a better job of ensuring that outliers remain outlying. We can test that hypothesis by running LOF on this embedding and looking at the resulting digit images we get out. Ideally we should expect to find some potentially even stranger results.

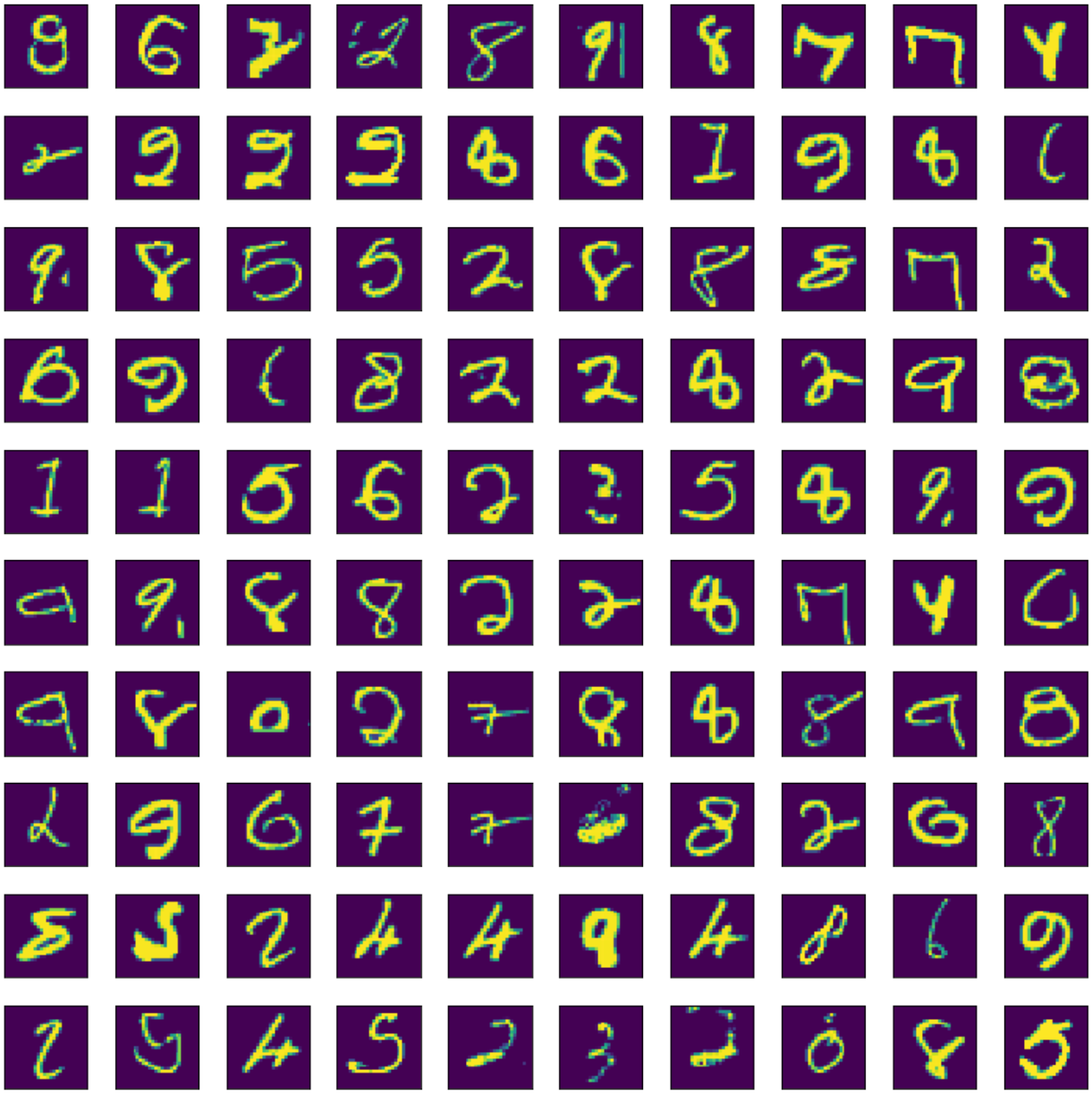
```
%%time
outlier_scores = sklearn.neighbors.LocalOutlierFactor(contamination=0.001428).fit_
↪predict(mapper.embedding_)
```

```
outlying_digits = data[outlier_scores == -1]
outlying_digits.shape
```

```
(100, 784)
```

We have the expected 100 most outlying digit images, so let's visualise the results and see if they really are particularly strange.

```
fig, axes = plt.subplots(10, 10, figsize=(10,10))
for i, ax in enumerate(axes.flatten()):
    ax.imshow(outlying_digits[i].reshape((28,28)))
    plt.setp(ax, xticks=[], yticks=[])
plt.tight_layout()
```



Here we see that the line thickness variation (particularly "fat" digits, or particularly "fine" lines) that the original embedding helped surface come through even more strongly here. We also see a number of clearly corrupted images with extra lines, dots, or strange blurring occurring.

So, in summary, using UMAP to reduce dimension prior to running classical outlier detection methods such as LOF can improve both the speed with which the algorithm runs, and the quality of results the outlier detection can find. Fur-

thermore we have introduced the `set_op_mix_ratio` parameter, and explained how it can be used to potentially improve the performance of outlier detection approaches applied to UMAP embeddings.

# CHAPTER 11

# Embedding to non-Euclidean spaces

By default UMAP embeds data into Euclidean space. For 2D visualization that means that data is embedded into a 2D plane suitable for a scatterplot. In practice, however, there aren't really any major constraints that prevent the algorithm from working with other more interesting embedding spaces. In this tutorial we'll look at how to get UMAP to embed into other spaces, how to embed into your own custom space, and why this sort of approach might be useful.

To start we'll load the usual selection of libraries. In this case we will not be using the `umap.plot` functionality, but working with matplotlib directly since we'll be generating some custom visualizations for some of the more unique embedding spaces.

```
import numpy as np
import numba
import sklearn.datasets
import matplotlib.pyplot as plt
import seaborn as sns
from mpl_toolkits.mplot3d import Axes3D
import umap
%matplotlib inline
```

```
sns.set(style='white', rc={'figure.figsize':(10,10)})
```

As a test dataset we'll use the PenDigits dataset from sklearn – embedding into exotic spaces can be considerably more computationally taxing, so a simple relatively small dataset is going to be useful.

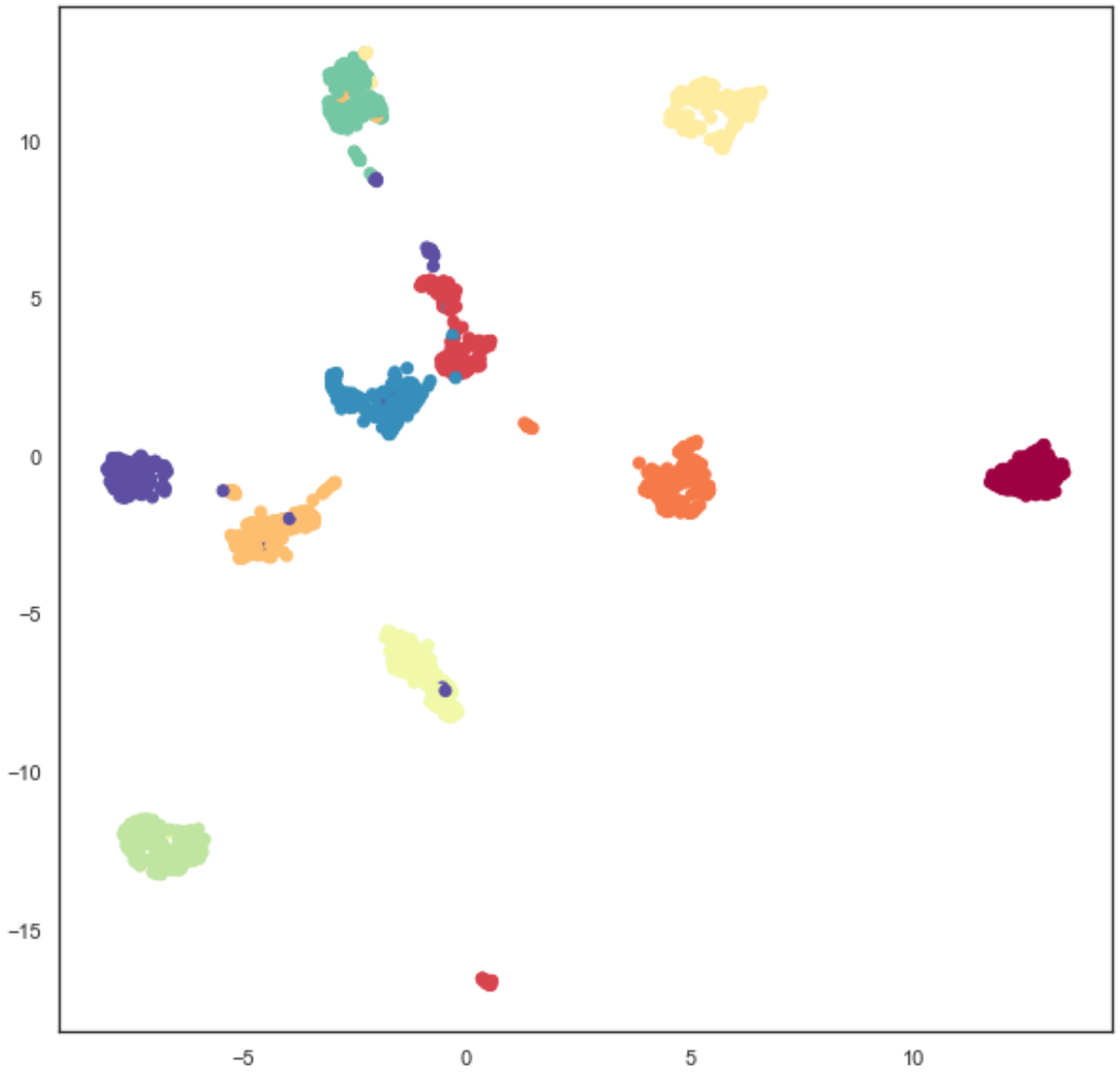
```
digits = sklearn.datasets.load_digits()
```

## 11.1 Plane embeddings

Plain old plane embeddings are simple enough – it is the default for UMAP. Here we'll run through the example again, just to ensure you are familiar with how this works, and what the result of a UMAP embedding of the PenDigits dataset looks like in the simple case of embedding in the plane.

```
plane_mapper = umap.UMAP(random_state=42).fit(digits.data)
```

```
plt.scatter(plane_mapper.embedding_.T[0], plane_mapper.embedding_.T[1], c=digits.
↪target, cmap='Spectral')
```




## 11.2 Spherical embeddings

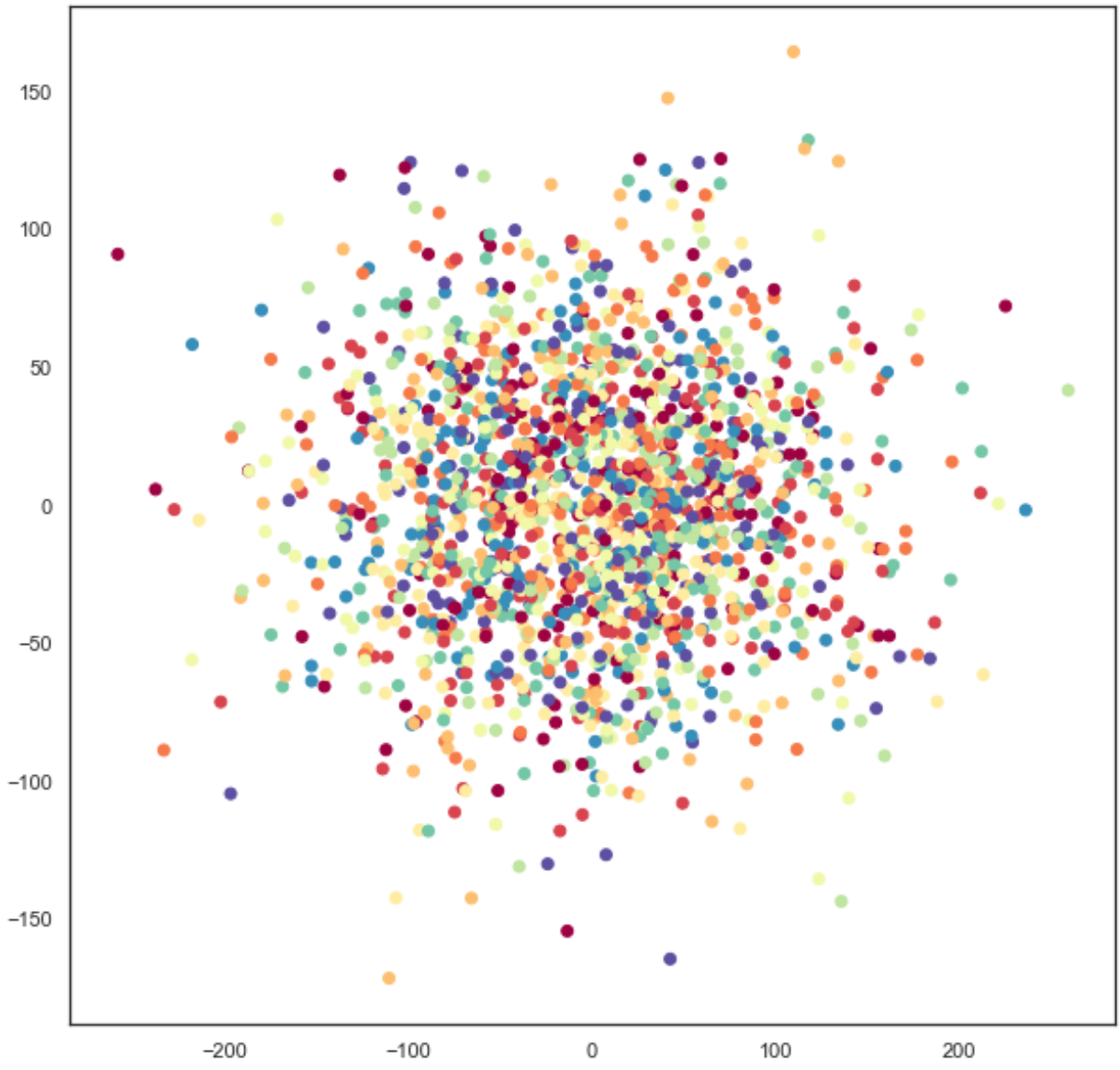
What if we wanted to embed data onto a sphere rather than a plane? This might make sense, for example, if we have reason to expect some sort of periodic behaviour or other reasons to expect that no point can be infinitely far from any other. To make UMAP embed onto a sphere we need to make use of the `output_metric` parameter, which specifies what metric to use for the **output** space. By default UMAP uses a Euclidean `output_metric` (and even has a special faster code-path for this case), but you can pass in other metrics. Among the metrics UMAP supports is the Haversine metric, used for measuring distances on a sphere, given in latitude and longitude (in radians). If we set

the `output_metric` to `"haversine"` then UMAP will use that to measure distance in the embedding space.

```
sphere_mapper = umap.UMAP(output_metric='haversine', random_state=42).fit(digits.data)
```

The result is the pendigits data embedded with respect to haversine distance on a sphere. The catch is that if we visualize this naively then we will get nonsense.

```
plt.scatter(sphere_mapper.embedding_.T[0], sphere_mapper.embedding_.T[1], c=digits.
↪target, cmap='Spectral')
```




What has gone astray is that under the embedding distance metric a point at $(0, \pi)$ is distance zero from a point at $(0, 3\pi)$ since that will wrap all the way around the equator. You'll note that the scales on the x and y axes of the above plot go well outside the ranges $(-\pi, \pi)$ and $(0, 2\pi)$, so this isn't the right representation of the data. We can, however, use straightforward formulas to map this data onto a sphere embedded in 3-space.

```
x = np.sin(sphere_mapper.embedding_[:, 0]) * np.cos(sphere_mapper.embedding_[:, 1])
```

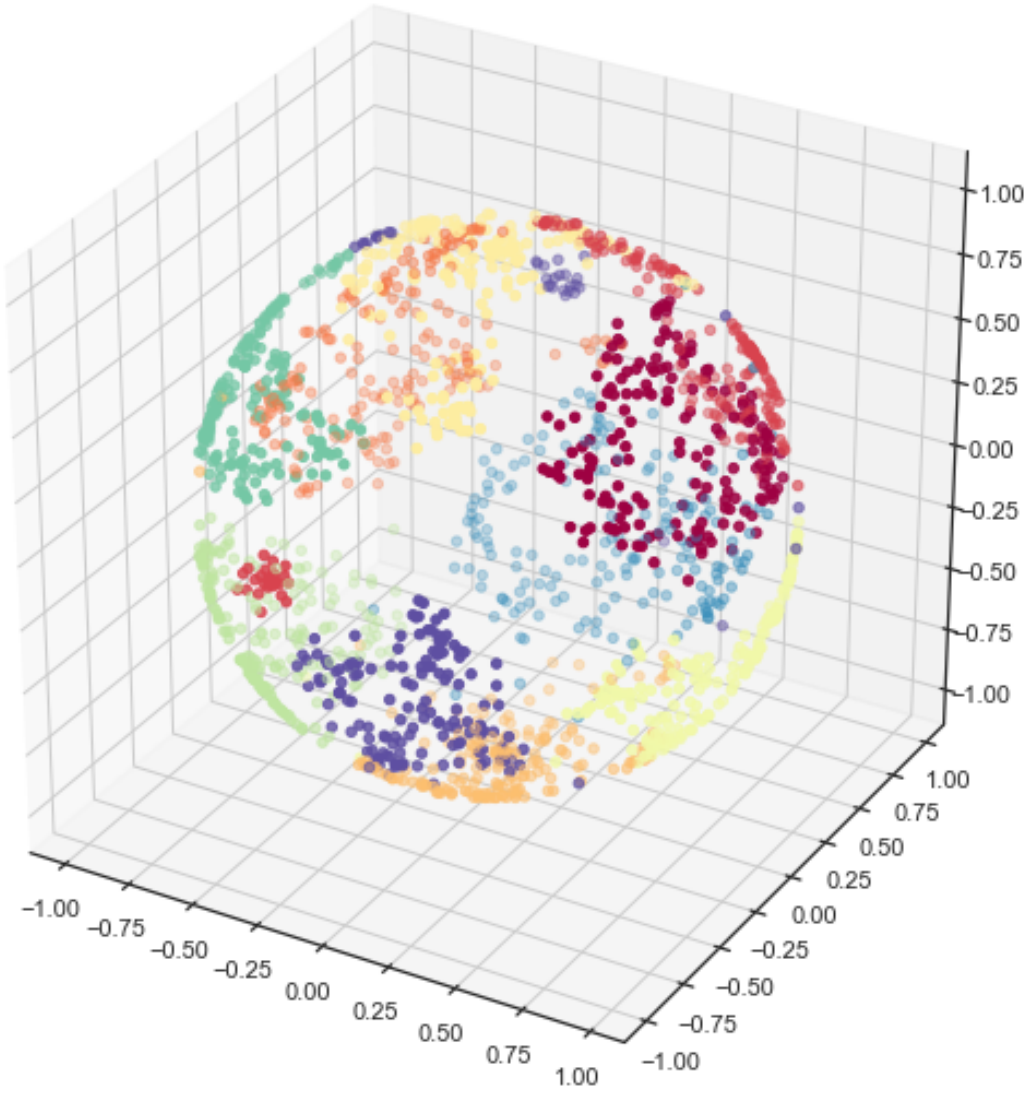
(continues on next page)

(continued from previous page)

```
y = np.sin(sphere_mapper.embedding_[:, 0]) * np.sin(sphere_mapper.embedding_[:, 1])
z = np.cos(sphere_mapper.embedding_[:, 0])
```

Now `x`, `y`, and `z` give 3d coordinates for each embedding point that lie on the surface of a sphere. We can visualize this using matplotlib's 3d plotting capabilities, and see that we have in fact induced a quite reasonable embedding of the data onto the surface of a sphere.

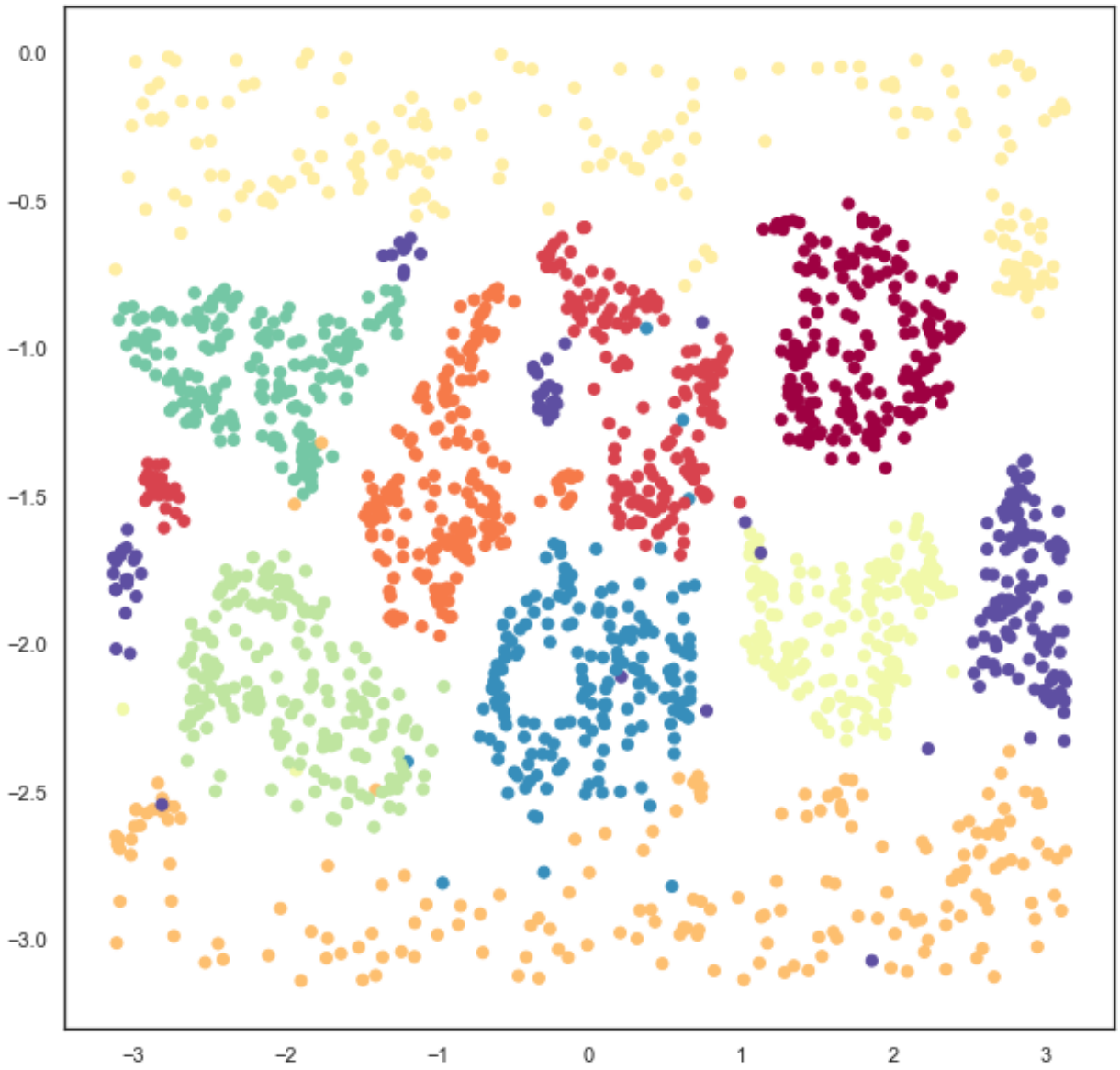
```
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.scatter(x, y, z, c=digits.target, cmap='Spectral')
```




If you prefer a 2d plot we can convert these into lat/long coordinates in the appropriate ranges and get the equivalent of a map projection of the sphere data.

```
x = np.arctan2(x, y)
y = -np.arccos(z)
```

```
plt.scatter(x, y, c=digits.target.astype(np.int32), cmap='Spectral')
```




## 11.3 Embedding on a Custom Metric Space

What if you have some other custom notion of a metric space that you would like to embed data into? In the same way that UMAP can support custom written distance metrics for the input data (as long as they can be compiled with numba), the `output_metric` parameter can accept custom distance functions. One catch is that, to support gradient descent optimization, the distance function needs to return both the distance, and a vector for the gradient of the distance. This latter point may require a little bit of calculus on the users part. A second catch is that it is highly beneficial to parameterize the embedding space in a way that has no coordinate constraints – otherwise the gradient

descent may step a point outside the embedding space, resulting in bad things happening. This is why, for example, the sphere example simply has points wrap around rather than constraining coordinates to be in the appropriate ranges.

Let's work through an example where we construct a distance metric and gradient for a different sort of space: a torus. A torus is essentially just the outer surface of a donut. We can parameterize the torus in terms of x, y coordinates with the caveat that we can "wrap around" (similar to the sphere). In such a model distances are mostly just euclidean distances, we just have to check for which is the shorter direction – across or wrapping around – and ensure we account for the equivalence of wrapping around several times. We can write a simple function to calculate that.

```
@numba.njit(fastmath=True)
def torus_euclidean_grad(x, y, torus_dimensions=(2*np.pi,2*np.pi)):
    """Standard euclidean distance.

    ..math::
        D(x, y) = \sqrt{\sum_i (x_i - y_i)^2}
    """
    distance_sqr = 0.0
    g = np.zeros_like(x)
    for i in range(x.shape[0]):
        a = abs(x[i] - y[i])
        if 2*a < torus_dimensions[i]:
            distance_sqr += a ** 2
            g[i] = (x[i] - y[i])
        else:
            distance_sqr += (torus_dimensions[i]-a) ** 2
            g[i] = (x[i] - y[i]) * (a - torus_dimensions[i]) / a
    distance = np.sqrt(distance_sqr)
    return distance, g/(1e-6 + distance)
```

Note that the gradient just derives from the standard euclidean gradient, we just have to check the direction according to the way we've wrapped around to compute the distance. We can now plug that function directly in to the `output_metric` parameter and end up embedding data on a torus.

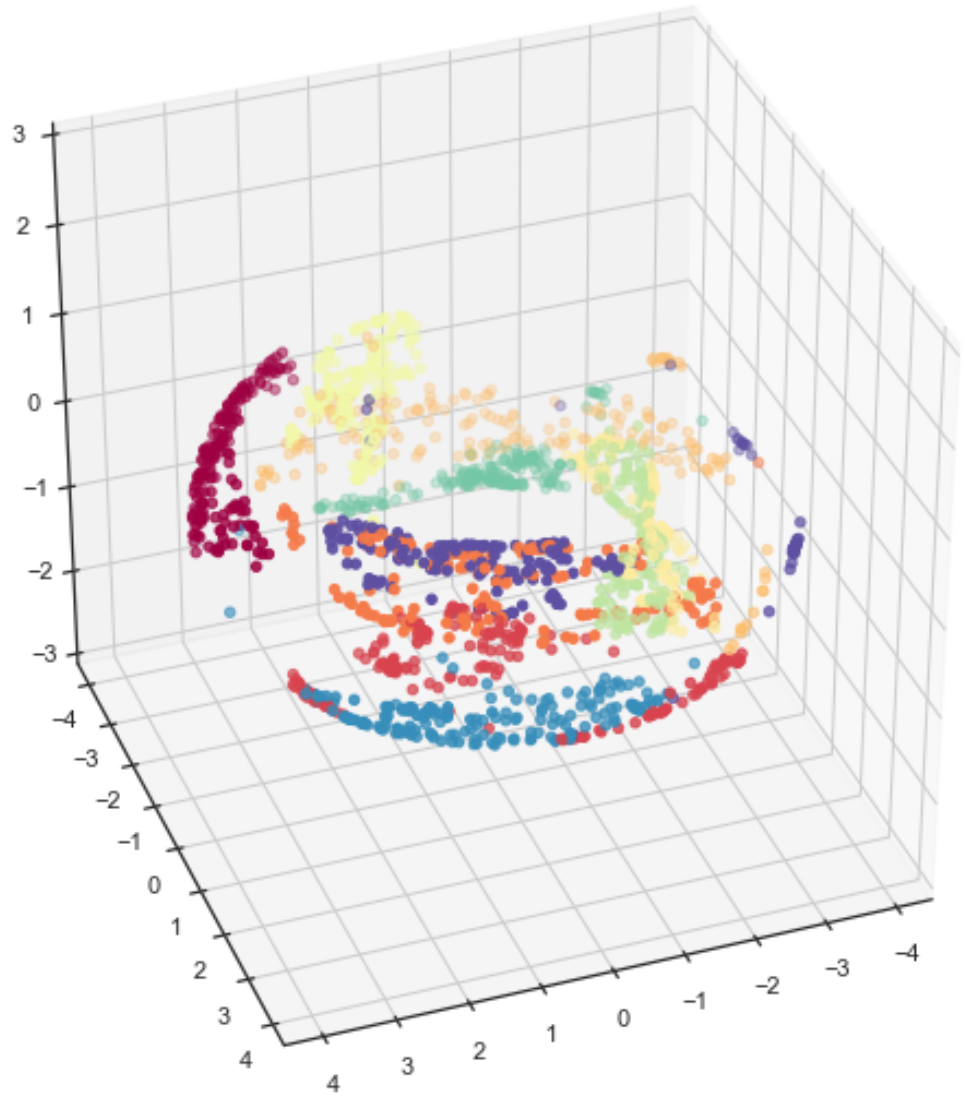
```
torus_mapper = umap.UMAP(output_metric=torus_euclidean_grad, random_state=42).
↪fit(digits.data)
```

As with the sphere case, a naive visualisation will look strange, due the the wrapping around and equivalence of looping several times. But, also just like the torus, we can construct a suitable visualization by computing the 3d coordinates for the points using a little bit of straightforward geometry (yes, I still had to look it up to check).

```
R = 3 # Size of the doughnut circle
r = 1 # Size of the doughnut cross-section

x = (R + r * np.cos(torus_mapper.embedding_[:, 0])) * np.cos(torus_mapper.embedding_
↪[:, 1])
y = (R + r * np.cos(torus_mapper.embedding_[:, 0])) * np.sin(torus_mapper.embedding_
↪[:, 1])
z = r * np.sin(torus_mapper.embedding_[:, 0])
```

Now we can visualize the result using matplotlib and see that, indeed, the data has been suitably embedded onto a torus.

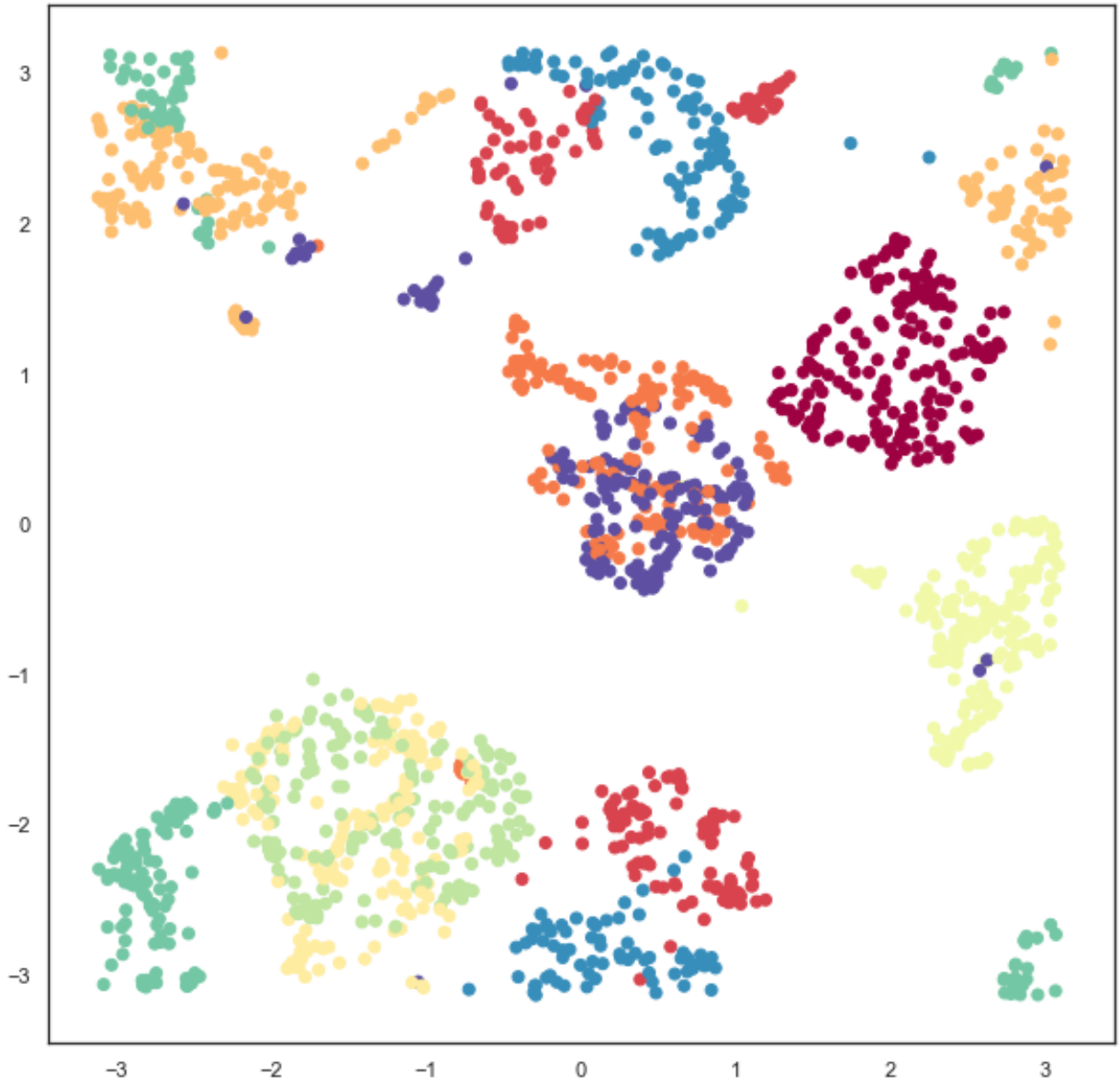
```
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.scatter(x, y, z, c=digits.target, cmap='Spectral')
ax.set_zlim3d(-3, 3)
ax.view_init(35, 70)
```




And as with the torus we can do a little geometry and unwrap the torus into a flat plane with the appropriate bounds.

```
u = np.arctan2(x,y)
v = np.arctan2(np.sqrt(x**2 + y**2) - R, z)
```

```
plt.scatter(u, v, c=digits.target, cmap='Spectral')
```




## 11.4 A Practical Example

While the examples given so far may have some use (because some data does have suitable periodic or looping structures that we expect will be better represented in a sphere or a torus), most data doesn't really fall in the realm of something that a user can, apriori, expect to lie on an exotic manifold. Are there more practical uses for the ability to embed in other spaces? It turns out that there are. One interesting example to consider is the space formed by 2d-Gaussian distributions. We can measure the distance between two Gaussians (parameterized by a 2d vector for the mean, and 2x2 matrix giving the covariance) by the negative log of the inner product between the PDFs (since this has a nice closed form solution, and is reasonably computable). That gives us a metric space to embed into where samples are represented not as points in 2d, but as Gaussian distributions in 2d, encoding some uncertainty in how each sample in the high dimensional space is to be embedded.

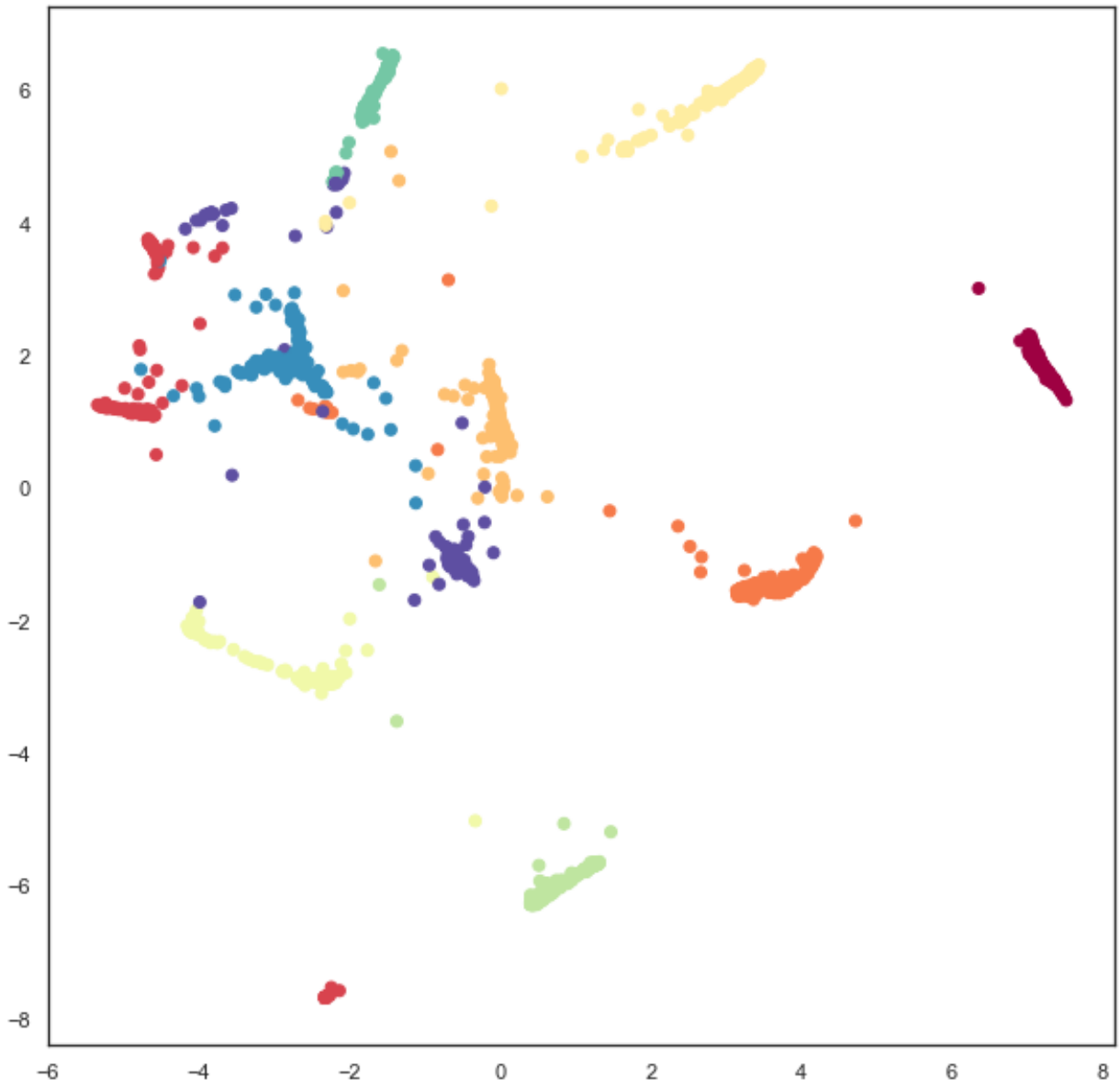
Of course we still have the issues of parameterizations that are suitable for SGD – requiring that the covariance matrix be symmetric and positive definite is challenging. Instead we can parameterize the covariance in terms of a width,

height and angle, and recover the covariance matrix from these if required. That gives us a total of 5 components to embed into (two for the mean, 3 for parameters describing the covariance). We can simply do this since the appropriate metric is defined already. Note that we have to specifically pass `n_components=5` since we need to explicitly embed into a 5 dimensional space to support all the covariance parameters associated to 2d Gaussians.

```
gaussian_mapper = umap.UMAP(output_metric='gaussian_energy',
                            n_components=5,
                            random_state=42).fit(digits.data)
```

Since we have embedded the data into a 5 dimensional space visualization is not as trivial as it was earlier. We can get a start on visualizing the results by looking at just the means, which are the 2d locations of the modes of the Gaussians. A traditional scatter plot will suffice for this.

```
plt.scatter(gaussian_mapper.embedding_.T[0], gaussian_mapper.embedding_.T[1],␣
↪c=digits.target, cmap='Spectral')
```

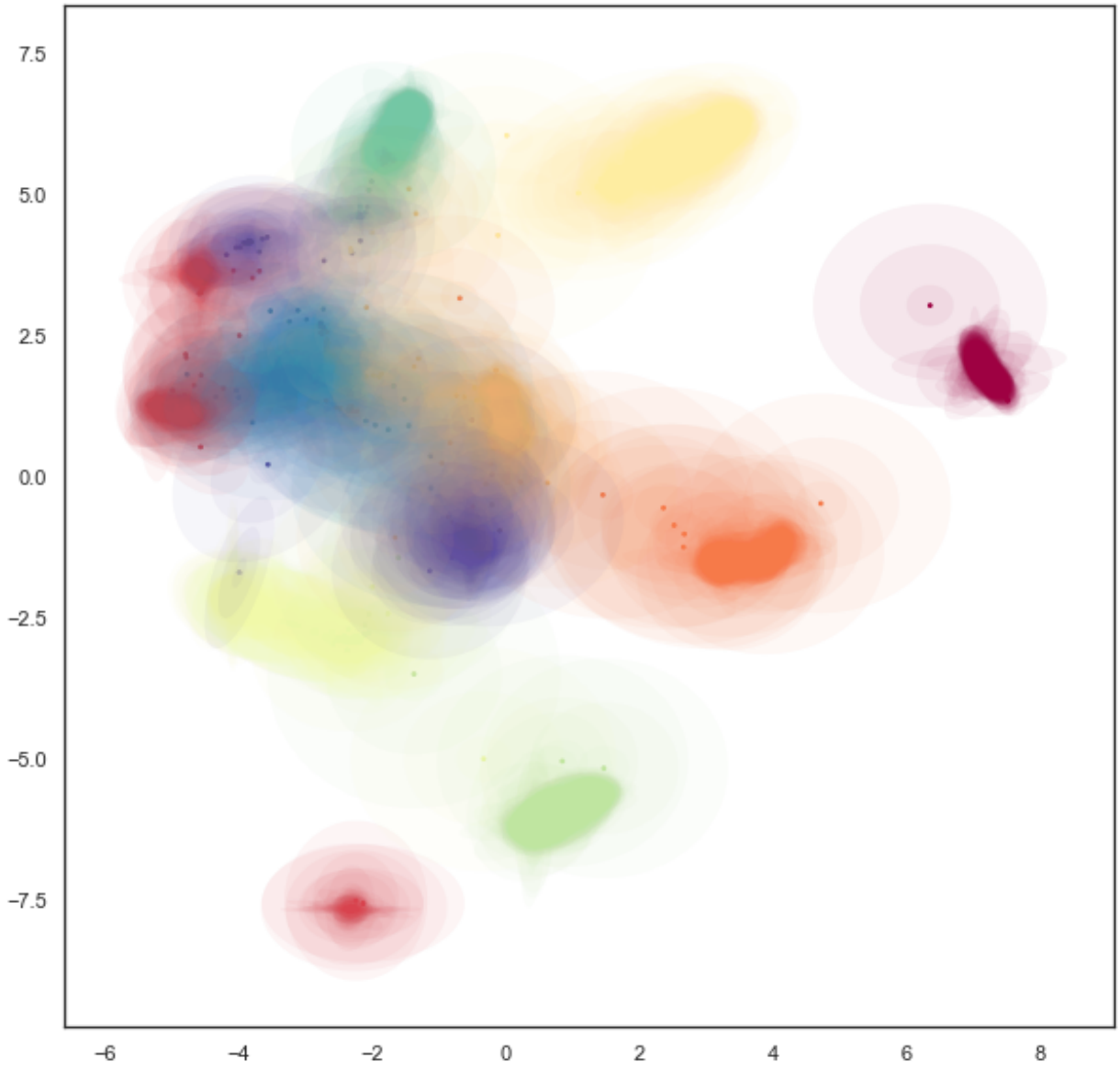


We see that we have gotten a result similar to a standard embedding into euclidean space, but with less clear clustering, and more points between clusters. To get a clearer idea of what is going on it will be necessary to devise a means to display some of the extra information contained in the extra 3 dimensions providing covariance data. To do this it will be helpful to be able to draw ellipses corresponding to super-level sets of the PDF of the 2d Gaussian. We can start on this by writing a simple function to draw ellipses on a plot accoriding to a position, a with, a height, and and angle (since this is the format the embedding computed the data).

```
from matplotlib.patches import Ellipse

def draw_simple_ellipse(position, width, height, angle,
                        ax=None, from_size=0.1, to_size=0.5, n_ellipses=3,
                        alpha=0.1, color=None,
                        **kwargs):
    ax = ax or plt.gca()
    angle = (angle / np.pi) * 180
    width, height = np.sqrt(width), np.sqrt(height)
    # Draw the Ellipse
    for nsig in np.linspace(from_size, to_size, n_ellipses):
        ax.add_patch(Ellipse(position, nsig * width, nsig * height,
                             angle, alpha=alpha, lw=0, color=color, **kwargs))
```

Now we can plot the data by providing a scatterplot of the centers (as before), but overlaying that over a super-level-set ellipses of the associated Gaussians. The obvious catch is that this will induce a lot of over-plotting, but it will at least provide a way to start understanding the embedding we have produced.

```
fig = plt.figure(figsize=(10,10))
ax = fig.add_subplot(111)
colors = plt.get_cmap('Spectral')(np.linspace(0, 1, 10))
for i in range(gaussian_mapper.embedding_.shape[0]):
    pos = gaussian_mapper.embedding_[i, :2]
    draw_simple_ellipse(pos, gaussian_mapper.embedding_[i, 2],
                        gaussian_mapper.embedding_[i, 3],
                        gaussian_mapper.embedding_[i, 4],
                        ax, color=colors[digits.target[i]],
                        from_size=0.2, to_size=1.0, alpha=0.05)
ax.scatter(gaussian_mapper.embedding_.T[0],
           gaussian_mapper.embedding_.T[1],
           c=digits.target, cmap='Spectral', s=3)
```




Now we can see that the covariance structure for the points can vary greatly, both in absolute size, and in shape. We note that many of the points falling between clusters have much larger variances, in a sense representing the greater uncertainty of the location of the embedding. It is also worth noting that the shape of the ellipses can vary significantly – there are several very stretched ellipses, quite distinct from many of the very round ellipses; in a sense this represents where the uncertainty falls more along a single line for example.

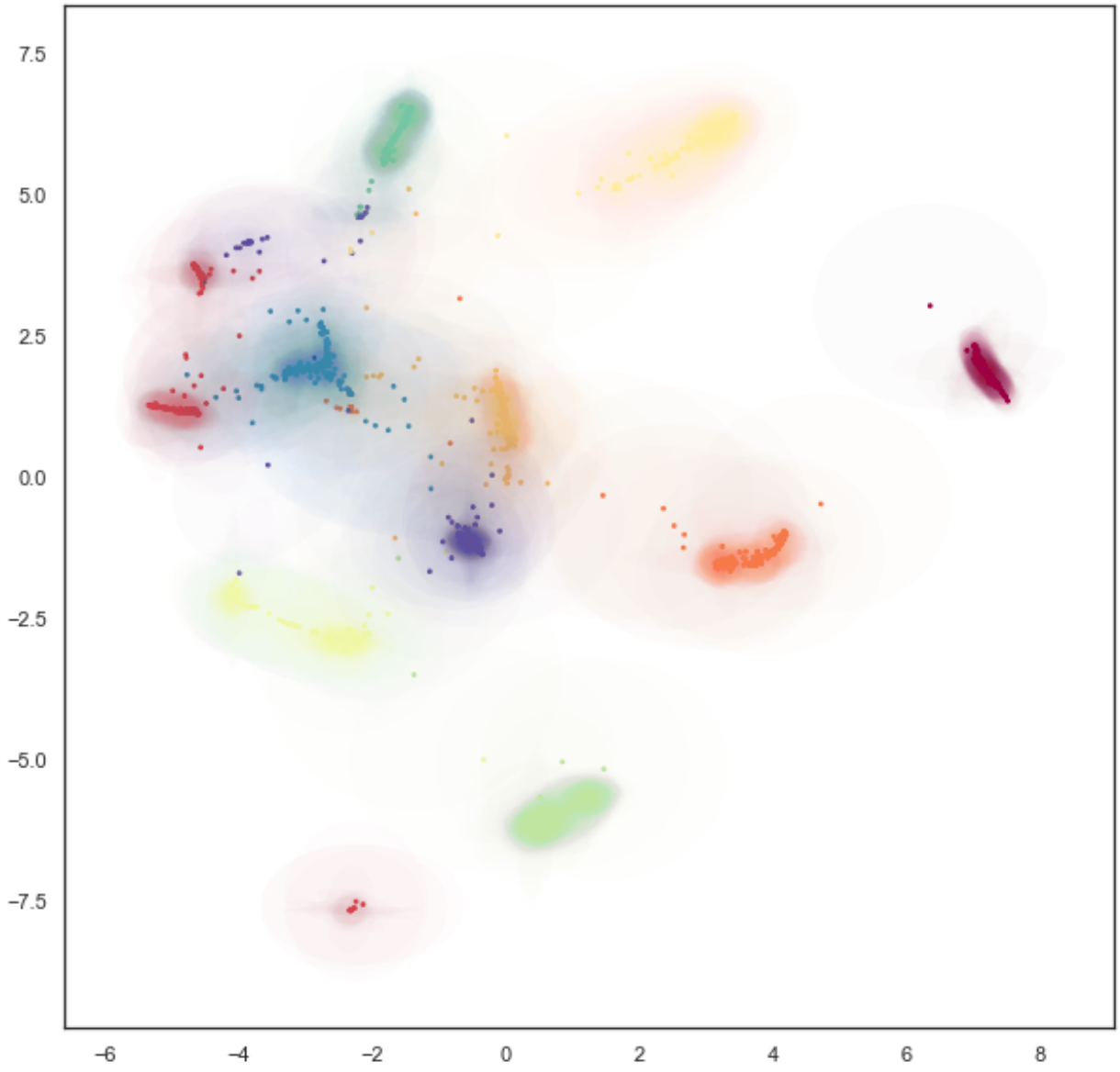
While this plot highlights some of the covariance structure in the outlying points, in practice the overplotting here obscures a lot of the more interesting structure in the clusters themselves. We can try to see this structure better by plotting only a single ellipse per point and using a lower alpha channel value for the ellipses, making them more translucent.

```
fig = plt.figure(figsize=(10,10))
ax = fig.add_subplot(111)
for i in range(gaussian_mapper.embedding_.shape[0]):
    pos = gaussian_mapper.embedding_[i, :2]
    draw_simple_ellipse(pos, gaussian_mapper.embedding_[i, 2],
```

(continues on next page)

(continued from previous page)

```
                        gaussian_mapper.embedding_[i, 3],
                        gaussian_mapper.embedding_[i, 4],
                        ax, n_ellipses=1,
                        color=colors[digits.target[i]],
                        from_size=1.0, to_size=1.0, alpha=0.01)
ax.scatter(gaussian_mapper.embedding_.T[0],
           gaussian_mapper.embedding_.T[1],
           c=digits.target, cmap='Spectral', s=3)
```




This lets us see the variation of density of clusters with respect to the covariance structure – some clusters have consistently very tight covariance,while others are more spread out (and hence have, in a sense, greater associated uncertainty. Of course we still have a degree of overplotting even here, and it will become increasingly difficult to tune alpha channels to make things visible. Instead what we would want is an actual density plot, showing the the density of the sum over all of these Gaussians.

To do this we'll need to some functions (which we'll use numba to accelerate): the evaluation of the density of a 2d

Gaussian at a given point; an evaluation of the density of a given point summing over a set of several Gaussians; and a function to generate the density for each point in some grid (summing only over nearby Gaussians to make this naive approach more computable).

```
from sklearn.neighbors import KDTree

@numba.njit(fastmath=True)
def eval_gaussian(x, pos=np.array([0, 0]), cov=np.eye(2, dtype=np.float32)):
    det = cov[0,0] * cov[1,1] - cov[0,1] * cov[1,0]
    if det > 1e-16:
        cov_inv = np.array([[cov[1,1], -cov[0,1]], [-cov[1,0], cov[0,0]]]) * 1.0 / det
        diff = x - pos
        m_dist = cov_inv[0,0] * diff[0]**2 - \
            (cov_inv[0,1] + cov_inv[1,0]) * diff[0] * diff[1] + \
            cov_inv[1,1] * diff[1]**2
        return (np.exp(-0.5 * m_dist)) / (2 * np.pi * np.sqrt(np.abs(det)))
    else:
        return 0.0

@numba.njit(fastmath=True)
def eval_density_at_point(x, embedding):
    result = 0.0
    for i in range(embedding.shape[0]):
        pos = embedding[i, :2]
        t = embedding[i, 4]
        U = np.array([[np.cos(t), np.sin(t)], [np.sin(t), -np.cos(t)]])
        cov = U @ np.diag(embedding[i, 2:4]) @ U
        result += eval_gaussian(x, pos=pos, cov=cov)
    return result

def create_density_plot(X, Y, embedding):
    Z = np.zeros_like(X)
    tree = KDTree(embedding[:, :2])
    for i in range(X.shape[0]):
        for j in range(X.shape[1]):
            nearby_points = embedding[tree.query_radius([[X[i,j],Y[i,j]]], r=2)[0]]
            Z[i, j] = eval_density_at_point(np.array([X[i,j],Y[i,j]]), nearby_points)
    return Z / Z.sum()
```

Now we simply need an appropriate grid of points. We can use the plot bounds seen above, and a grid size selected for the sake of computability. The numpy `meshgrid` function can supply the actual grid.

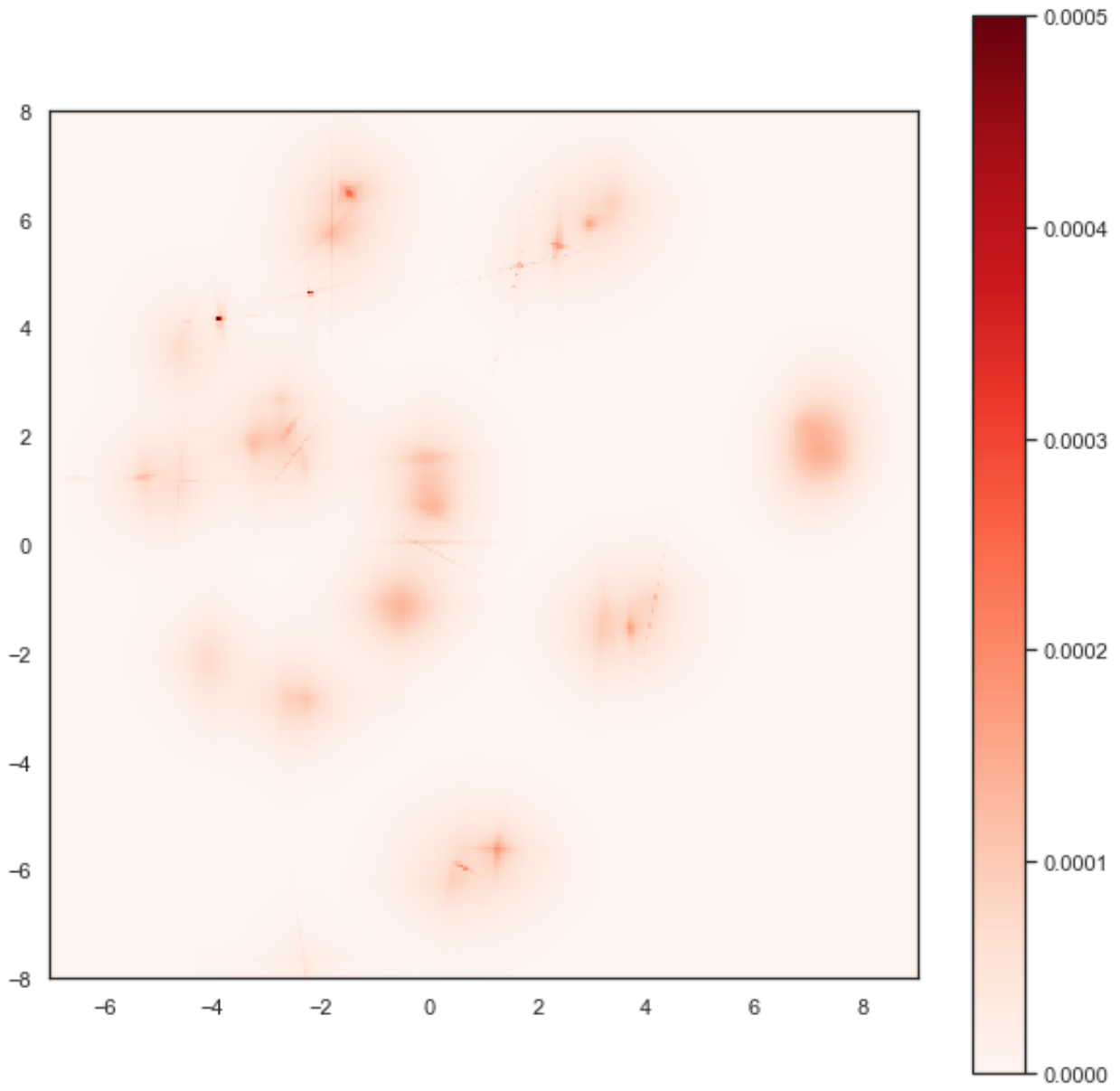
```
X, Y = np.meshgrid(np.linspace(-7, 9, 300), np.linspace(-8, 8, 300))
```

Now we can use the function defined above to compute the density at each point in the grid, given the Gaussians produced by the embedding.

```
Z = create_density_plot(X, Y, gaussian_mapper.embedding_)
```

Now we can view the result as a density plot using `imshow`.

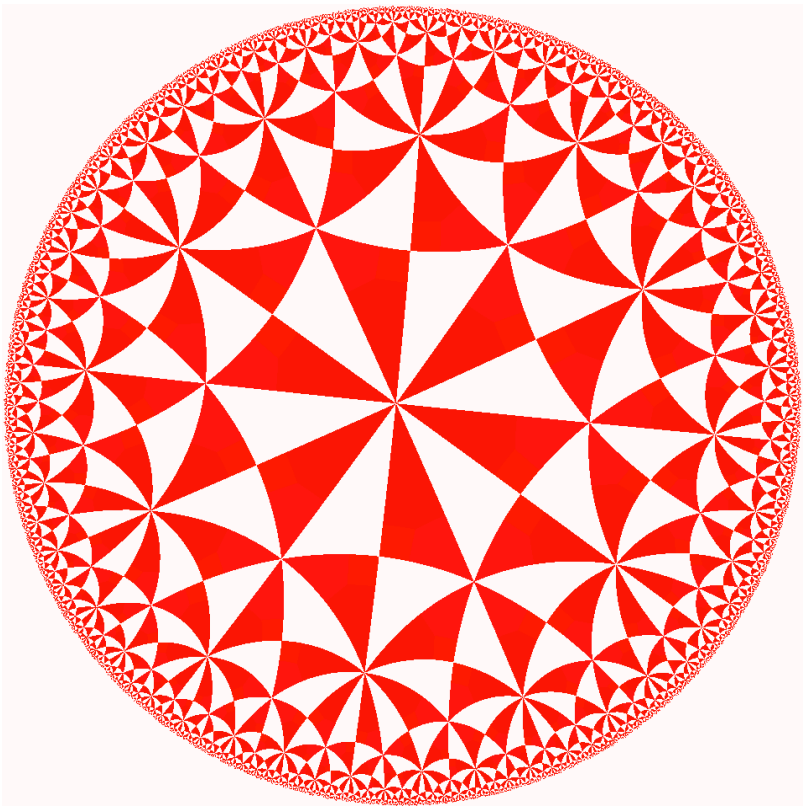
```
plt.imshow(Z, origin='lower', cmap='Reds', extent=(-7, 9, -8, 8), vmax=0.0005)
plt.colorbar()
```




Here we see the finer structure within the various clusters, including some of the interesting linear structures, demonstrating that this Gaussian uncertainty based embedding has captured quite detailed and useful information about the inter-relationships among the PenDigits dataset.

## 11.5 Bonus: Embedding in Hyperbolic space

As a bonus example let's look at embedding data into hyperbolic space. The most popular model for this for visualization is Poincare's disk model. An example of a regular tiling of hyperbolic space in Poincare's disk model is shown below; you may note it is similar to famous images by M.C. Escher.

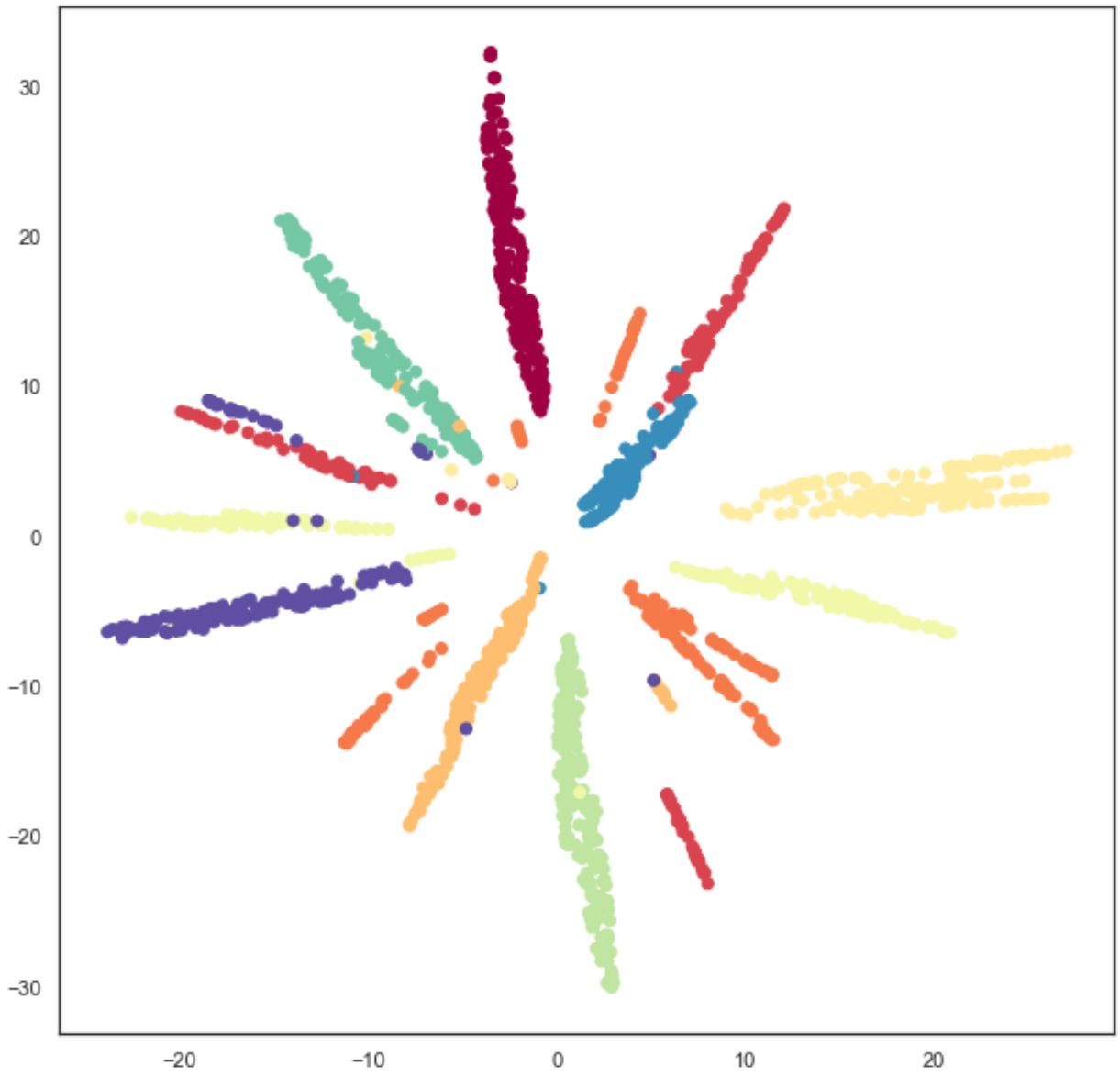


Ideally we would be able to embed directly into this Poincare disk model, but in practice this proves to be very difficult. The issue is that the disk has a "line at infinity" in a circle of radius one bounding the disk. Outside of that circle things are not well defined. As you may recall from the discussion of embedding onto spheres and toruses it is best if we can have a parameterisation of the embedding space that it is hard to move out of. The Poincare disk model is almost the opposite of this – as soon as we move outside the unit circle we have moved off the manifold and further updates will be badly defined. We therefore instead need a different parameterisation of hyperbolic space that is less constrained. One option is the Poincare half-plane model, but this, again, has a boundary that it is easy to move beyond. The simplest option is the hyperboloid model. Under this model we can simply move in x and y coordinates, and solve for the corresponding z coordinate when we need to compute distances. This model has been implemented under the distance metric `"hyperboloid"` so we can simply use it out-of-the-box.

```
hyperbolic_mapper = umap.UMAP(output_metric='hyperboloid',
                              random_state=42).fit(digits.data)
```

A straightforward visualization option is to simply view the x and y coordinates we have arrived at:

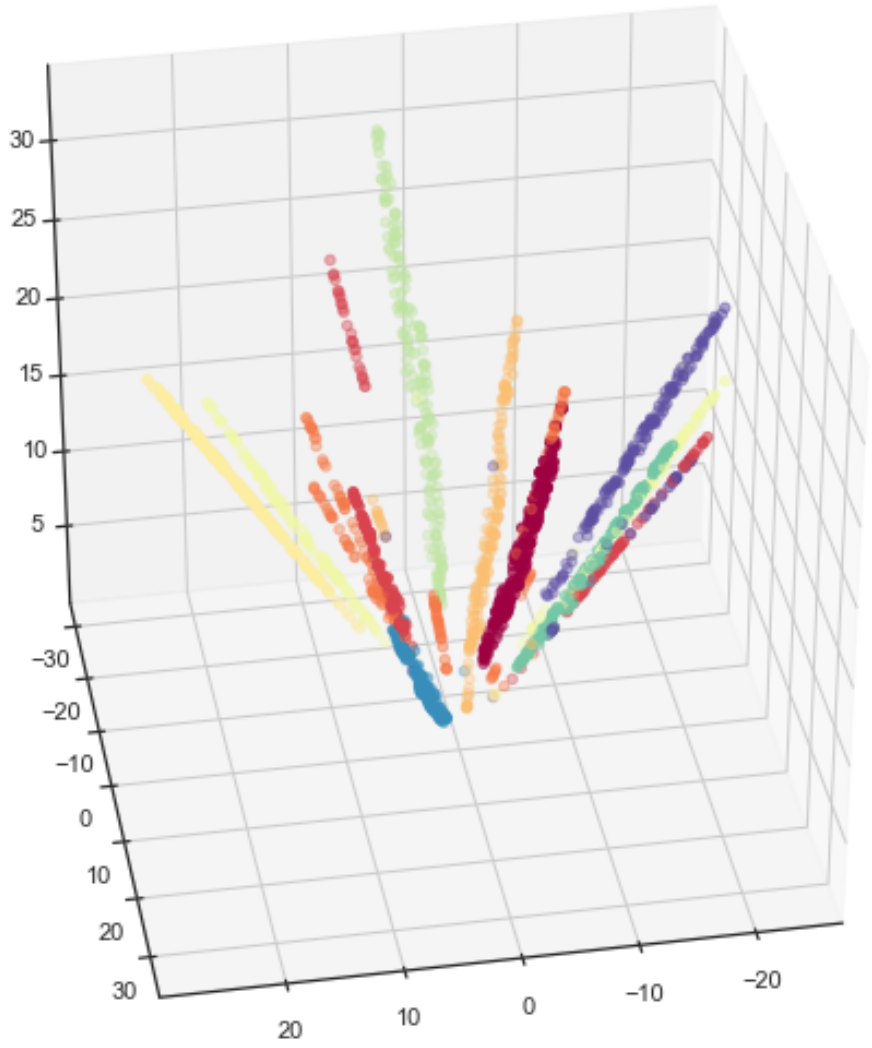
```
plt.scatter(hyperbolic_mapper.embedding_.T[0],
            hyperbolic_mapper.embedding_.T[1],
            c=digits.target, cmap='Spectral')
```




We can also solve for the z coordinate and view the data lying on a hyperboloid in 3d space.

```
x = hyperbolic_mapper.embedding_[:, 0]
y = hyperbolic_mapper.embedding_[:, 1]
z = np.sqrt(1 + np.sum(hyperbolic_mapper.embedding_**2, axis=1))
```

```
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.scatter(x, y, z, c=digits.target, cmap='Spectral')
ax.view_init(35, 80)
```

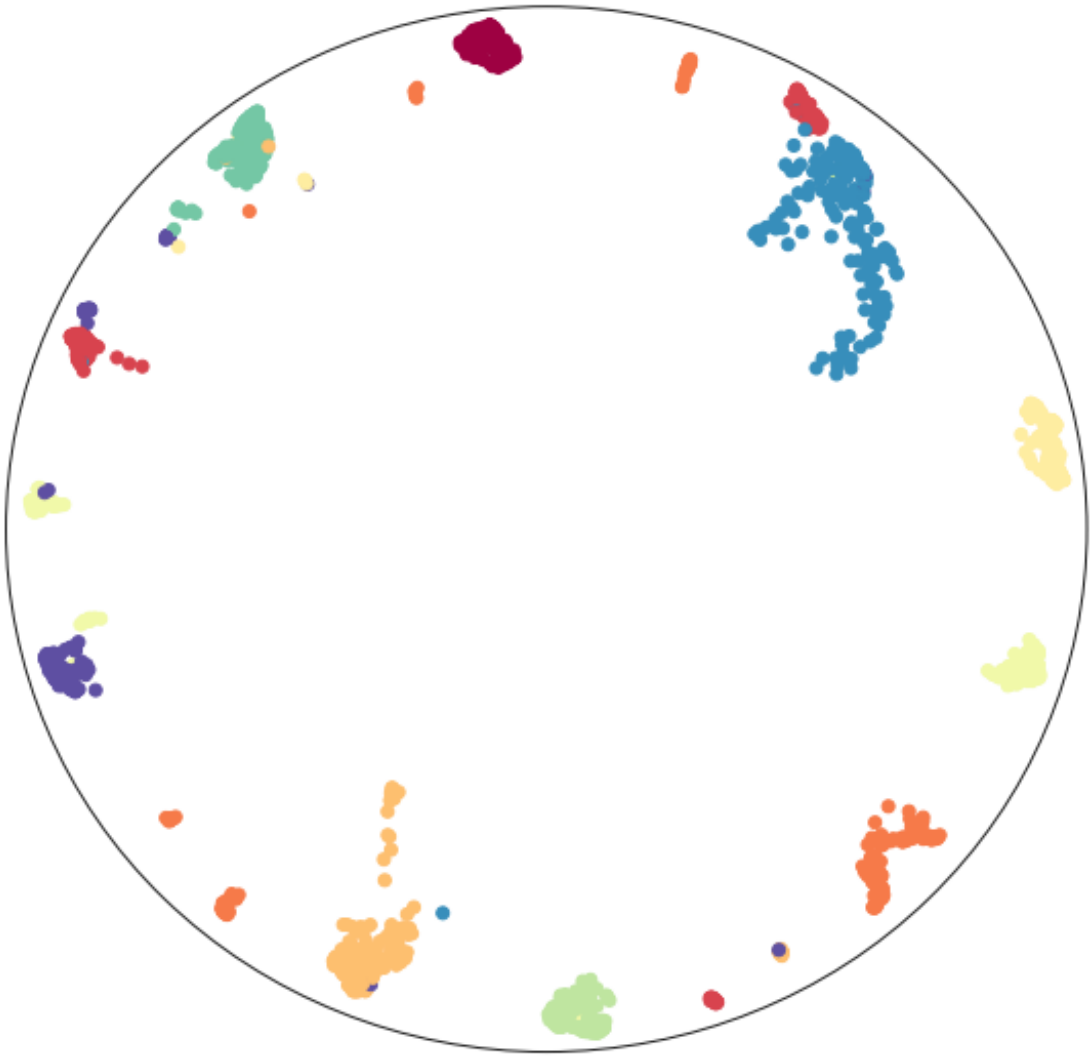



But we can do more – since we have embedded the data successfully in hyperbolic space we can map the data into the Poincare disk model. This is, in fact, a straightforward computation.

```
disk_x = x / (1 + z)
disk_y = y / (1 + z)
```

Now we can visualize the data in a Poincare disk model embedding as we first wanted. For this we simply generate a scatterplot of the data, and then draw in the bounding circle of the line at infinity.

```
fig = plt.figure()
ax = fig.add_subplot(111)
ax.scatter(disk_x, disk_y, c=digits.target, cmap='Spectral')
boundary = plt.Circle((0,0), 1, fc='none', ec='k')
ax.add_artist(boundary)
ax.axis('off');
```



Hopefully this has provided a useful example of how to go about embedding into non-euclidean spaces. This last example ideally highlights the limitations of this approach (we really need a suitable parameterisation), and some potential approaches to get around this: we can use an alternative parameterisation for the embedding, and then transform the data into the the desired representation.

# CHAPTER 12

# Gallery of Examples of UMAP usage

A small gallery collection examples of UMAP usage. Do you have an interesting UMAP plot that uses publicly available data? Submit a pull request to have it added as an example!

**Note:** Click *here* to download the full example code

## 12.1 UMAP on the MNIST Digits dataset

A simple example demonstrating how to use UMAP on a larger dataset such as MNIST. We first pull the MNIST dataset and then use UMAP to reduce it to only 2-dimensions for easy visualisation.

Note that UMAP manages to both group the individual digit classes, but also to retain the overall global structure among the different digit classes – keeping 1 far from 0, and grouping triplets of 3,5,8 and 4,7,9 which can blend into one another in some cases.

```
import umap
from sklearn.datasets import fetch_openml
import matplotlib.pyplot as plt
import seaborn as sns

sns.set(context="paper", style="white")

mnist = fetch_openml("mnist_784", version=1)

reducer = umap.UMAP(random_state=42)
embedding = reducer.fit_transform(mnist.data)

fig, ax = plt.subplots(figsize=(12, 10))
color = mnist.target.astype(int)
plt.scatter(embedding[:, 0], embedding[:, 1], c=color, cmap="Spectral", s=0.1)
plt.setp(ax, xticks=[], yticks=[])
```

(continues on next page)

(continued from previous page)

```
plt.title("MNIST data embedded into two dimensions by UMAP", fontsize=18)

plt.show()
```

**Total running time of the script:** ( 0 minutes 0.000 seconds)

**Note:** Click *here* to download the full example code

## 12.2 UMAP on the MNIST Digits dataset

A simple example demonstrating how to use UMAP on a larger dataset such as MNIST. We first pull the MNIST dataset and then use UMAP to reduce it to only 2-dimensions for easy visualisation.

Note that UMAP manages to both group the individual digit classes, but also to retain the overall global structure among the different digit classes – keeping 1 far from 0, and grouping triplets of 3,5,8 and 4,7,9 which can blend into one another in some cases.

```
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.datasets import fetch_openml
from sklearn.model_selection import train_test_split

import umap

sns.set(context="paper", style="white")

mnist = fetch_openml('mnist_784', version=1)
X_train, X_test, y_train, y_test = train_test_split(
    mnist.data,
    mnist.target,
    stratify=mnist.target,
    random_state=42
)

reducer = umap.UMAP(random_state=42)
embedding_train = reducer.fit_transform(X_train)
embedding_test = reducer.transform(X_test)

fig, ax = plt.subplots(1, 2, sharex=True, sharey=True, figsize=(12, 10))
ax[0].scatter(
    embedding_train[:, 0], embedding_train[:, 1], c=y_train, cmap="Spectral"  # , s=0.
↪1
)
ax[1].scatter(
    embedding_test[:, 0], embedding_test[:, 1], c=y_test, cmap="Spectral"  # , s=0.1
)
plt.setp(ax[0], xticks=[], yticks=[])
plt.setp(ax[1], xticks=[], yticks=[])
plt.suptitle("MNIST data embedded into two dimensions by UMAP", fontsize=18)
ax[0].set_title("Training Set", fontsize=12)
ax[1].set_title("Test Set", fontsize=12)
plt.show()
```

**Total running time of the script:** ( 0 minutes 0.000 seconds)

**Note:** Click *here* to download the full example code

## 12.3 UMAP as a Feature Extraction Technique for Classification

The following script shows how UMAP can be used as a feature extraction technique to improve the accuracy on a classification task. It also shows how UMAP can be integrated in standard scikit-learn pipelines.

The first step is to create a dataset for a classification task, which is performed with the function `sklearn.datasets.make_classification`. The dataset is then split into a training set and a test set using the `sklearn.model_selection.train_test_split` function.

Second, a linear SVM is fitted on the training set. To choose the best hyperparameters automatically, a gridsearch is performed on the training set. The performance of the model is then evaluated on the test set with the accuracy metric.

> Third, the previous step is repeated with a slight modification: UMAP is used as a feature extraction technique. This small change results in a substantial improvement compared to the model where raw data is used.

```
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.pipeline import Pipeline
from sklearn.svm import LinearSVC

from umap import UMAP

# Make a toy dataset
X, y = make_classification(
    n_samples=1000,
    n_features=300,
    n_informative=250,
    n_redundant=0,
    n_repeated=0,
    n_classes=2,
    random_state=1212,
)

# Split the dataset into a training set and a test set
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

# Classification with a linear SVM
svc = LinearSVC(dual=False, random_state=123)
params_grid = {"C": [10 ** k for k in range(-3, 4)]}
clf = GridSearchCV(svc, params_grid)
clf.fit(X_train, y_train)
print(
    "Accuracy on the test set with raw data: {:.3f}".format(clf.score(X_test, y_test))
)

# Transformation with UMAP followed by classification with a linear SVM
umap = UMAP(random_state=456)
pipeline = Pipeline([("umap", umap), ("svc", svc)])
```

(continues on next page)

(continued from previous page)

```
params_grid_pipeline = {
    "umap__n_neighbors": [5, 20],
    "umap__n_components": [15, 25, 50],
    "svc__C": [10 ** k for k in range(-3, 4)],
}


clf_pipeline = GridSearchCV(pipeline, params_grid_pipeline)
clf_pipeline.fit(X_train, y_train)
print(
    "Accuracy on the test set with UMAP transformation: {:.3f}".format(
        clf_pipeline.score(X_test, y_test)
    )
)
```

**Total running time of the script:** ( 0 minutes 0.000 seconds)

**Note:** Click *here* to download the full example code

## 12.4 UMAP on the Fashion MNIST Digits dataset using Datashader

This is a simple example of using UMAP on the Fashion-MNIST dataset. The goal of this example is largely to demonstrate the use of datashader as an effective tool for visualising UMAP results. In particular datashader allows visualisation of very large datasets where overplotting can be a serious problem. It supports coloring by categorical variables (as shown in this example), or by continuous variables, or by density (as is common in datashader examples).

```
import umap
import numpy as np
import pandas as pd
import requests
import os
import datashader as ds
import datashader.utils as utils
import datashader.transfer_functions as tf
import matplotlib.pyplot as plt
import seaborn as sns

sns.set(context="paper", style="white")

if not os.path.isfile("fashion-mnist.csv"):
    csv_data = requests.get("https://www.openml.org/data/get_csv/18238735/phpnBqZGZ")
    with open("fashion-mnist.csv", "w") as f:
        f.write(csv_data.text)
source_df = pd.read_csv("fashion-mnist.csv")

data = source_df.iloc[:, :784].values.astype(np.float32)
target = source_df["class"].values

pal = [
    "#9e0142",
    "#d8434e",
    "#f67a49",
```

(continues on next page)

(continued from previous page)

```
    "#fdbf6f",
    "#feeda1",
    "#f1f9a9",
    "#bfe5a0",
    "#74c7a5",
    "#378ebb",
    "#5e4fa2",
]
color_key = {str(d): c for d, c in enumerate(pal)}

reducer = umap.UMAP(random_state=42)
embedding = reducer.fit_transform(data)

df = pd.DataFrame(embedding, columns=("x", "y"))
df["class"] = pd.Series([str(x) for x in target], dtype="category")

cvs = ds.Canvas(plot_width=400, plot_height=400)
agg = cvs.points(df, "x", "y", ds.count_cat("class"))
img = tf.shade(agg, color_key=color_key, how="eq_hist")

utils.export_image(img, filename="fashion-mnist", background="black")

image = plt.imread("fashion-mnist.png")
fig, ax = plt.subplots(figsize=(6, 6))
plt.imshow(image)
plt.setp(ax, xticks=[], yticks=[])
plt.title(
    "Fashion MNIST data embedded\n"
    "into two dimensions by UMAP\n"
    "visualised with Datashader",
    fontsize=12,
)

plt.show()
```

**Total running time of the script:** ( 0 minutes 0.000 seconds)

**Note:** Click *here* to download the full example code

## 12.5 Comparison of Dimension Reduction Techniques

A comparison of several different dimension reduction techniques on a variety of toy datasets. The datasets are all toy datasets, but should provide a representative range of the strengths and weaknesses of the different algorithms.

The time to perform the dimension reduction with each algorithm and each dataset is recorded in the lower right of each plot.

Things to note about the datasets:

- **Blobs: A set of five gaussian blobs in 10 dimensional** space. This should be a prototypical example of something that should clearly separate even in a reduced dimension space.
- **Iris: a classic small dataset with one distinct class** and two classes that are not clearly separated.

- **Digits: handwritten digits – ideally different digit** classes should form distinct groups. Due to the nature of handwriting digits may have several forms (crossed or uncrossed sevens, capped or straight line oes, etc.)
- **Wine: wine characteristics ideally used for a toy** regression. Ultimately the data is essentially one dimensional in nature.
- **Swiss Roll: data is essentially a rectangle, but** has been "rolled up" like a swiss roll in three dimensional space. Ideally a dimension reduction technique should be able to "unroll" it. The data has been coloured according to one dimension of the rectangle, so should form a rectangle of smooth color variation.
- **Sphere: the two dimensional surface of a three** dimensional sphere. This cannot be represented accurately in two dimensions without tearing. The sphere has been coloured with hue around the equator and black to white from the south to north pole.

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import time

from sklearn import datasets, decomposition, manifold, preprocessing
from colorsys import hsv_to_rgb

import umap

sns.set(context="paper", style="white")

blobs, blob_labels = datasets.make_blobs(
    n_samples=500, n_features=10, centers=5, random_state=42
)
iris = datasets.load_iris()
digits = datasets.load_digits(n_class=10)
wine = datasets.load_wine()
swissroll, swissroll_labels = datasets.make_swiss_roll(
    n_samples=1000, noise=0.1, random_state=42
)
sphere = np.random.normal(size=(600, 3))
sphere = preprocessing.normalize(sphere)
sphere_hsv = np.array(
    [
        (
            (np.arctan2(c[1], c[0]) + np.pi) / (2 * np.pi),
            np.abs(c[2]),
            min((c[2] + 1.1), 1.0),
        )
        for c in sphere
    ]
)
sphere_colors = np.array([hsv_to_rgb(*c) for c in sphere_hsv])

reducers = [
    (manifold.TSNE, {"perplexity": 50}),
    # (manifold.LocallyLinearEmbedding, {'n_neighbors':10, 'method':'hessian'}),
    (manifold.Isomap, {"n_neighbors": 30}),
    (manifold.MDS, {}),
    (decomposition.PCA, {}),
    (umap.UMAP, {"n_neighbors": 30, "min_dist": 0.3}),
]

test_data = [
```

(continues on next page)

(continued from previous page)

```
    (blobs, blob_labels),
    (iris.data, iris.target),
    (digits.data, digits.target),
    (wine.data, wine.target),
    (swissroll, swissroll_labels),
    (sphere, sphere_colors),
]
dataset_names = ["Blobs", "Iris", "Digits", "Wine", "Swiss Roll", "Sphere"]

n_rows = len(test_data)
n_cols = len(reducers)
ax_index = 1
ax_list = []

# plt.figure(figsize=(9 * 2 + 3, 12.5))
plt.figure(figsize=(10, 8))
plt.subplots_adjust(
    left=0.02, right=0.98, bottom=0.001, top=0.96, wspace=0.05, hspace=0.01
)
for data, labels in test_data:
    for reducer, args in reducers:
        start_time = time.time()
        embedding = reducer(n_components=2, **args).fit_transform(data)
        elapsed_time = time.time() - start_time
        ax = plt.subplot(n_rows, n_cols, ax_index)
        if isinstance(labels[0], tuple):
            ax.scatter(*embedding.T, s=10, c=labels, alpha=0.5)
        else:
            ax.scatter(*embedding.T, s=10, c=labels, cmap="Spectral", alpha=0.5)
        ax.text(
            0.99,
            0.01,
            "{:.2f} s".format(elapsed_time),
            transform=ax.transAxes,
            size=14,
            horizontalalignment="right",
        )
        ax_list.append(ax)
        ax_index += 1
plt.setp(ax_list, xticks=[], yticks=[])

for i in np.arange(n_rows) * n_cols:
    ax_list[i].set_ylabel(dataset_names[i // n_cols], size=16)
for i in range(n_cols):
    ax_list[i].set_xlabel(repr(reducers[i][0]()).split("(")[0], size=16)
    ax_list[i].xaxis.set_label_position("top")

plt.tight_layout()
plt.show()
```

**Total running time of the script:** ( 0 minutes 0.000 seconds)

# CHAPTER 13

# Frequently Asked Questions

Compiled here are a set of frequently asked questions, along with answers. If you don't find your question listed here then please feel free to add an issue on github. More questions are always welcome, and the authors will do their best to answer. If you feel you have a common question that isn't answered here then please suggest that the question (and answer) be added to the FAQ when you file the issue.

## 13.1 Should I normalise my features?

The default answer is yes, but, of course, the real answer is "it depends". If your features have meaningful relationships with one another (say, latitude and longitude vales) then normalising per feature is not a good idea. For features that are essentially independent it does make sense to get all the features on (relatively) the same scale. The best way to do this is to use pre-processing tools from scikit-learn. All the advice given there applies as sensible preprocessing for UMAP, and since UMAP is scikit-learn compatible you can put all of this together into a scikit-learn pipeline

## 13.2 Can I cluster the results of UMAP?

This is hard to answer well, but essentially the answer is "yes, with care". To start with it matters what clustering algorithm you are going to use. Since UMAP does not necessarily produce clean spherical clusters something like K-Means is a poor choice. I would recommend HDBSCAN or similar. The catch here is that UMAP, with its uniform density assumption, does not preserve density well. What UMAP will do, however, is contract connected components of the manifold together. Providing you have enough data for UMAP to distinguish that information then you can get *useful* clustering results out since algorithms like HDBSCAN will easily pick out the components after applying UMAP.

UMAP does offer significant improvements over algorithms like t-SNE for clustering. First, by preserving more global structure and creating meaningful separation between connected components of the manifold on which the data lies, UMAP offers more meaningful clusters. Second, because it supports arbitrary embedding dimensions, UMAP allows embedding to larger dimensional spaces that make it more amenable to clustering.

## 13.3 The clusters are all squashed together and I can't see internal structure

One of UMAPs goals is to have distance between clusters of points be meaningful. This means that clusters can end up spread out with a fair amount of space between them. As a result the clusters themselves can end up more visually packed together than in, say, t-SNE. This is intended. A catch, however, is that many plots (for example matplotlib's scatter plot with default parameters) tend to show the clusters only as indistinct blobs with no internal structure. The solution for this is really a matter of tuning the plot more than anything else.

If you are using matplotlib consider using the `s` parameter that specifies the glyph size in scatter plots. Depending on how much data you have reducing this to anything from 5 to 0.001 can have a notable effect. The `size` parameter in bokeh is similarly useful (but does not need to be quite so small).

More generally the real solution, particular with large datasets, is to use datashader for plotting. Datashader is a plotting llibrary that handles aggregation of large scale data in scatter plots in a way that can better show the underlying detail that can otherwise be lost. We highly recommend investing the time to learn datashader for UMAP plot particularly for larger datasets.

## 13.4 I ran out of memory. Help!

For some datasets the default options for approximate nearest neighbor search can result in excessive memory use. If your dataset is not especially large but you have found that UMAP runs out of memory when operating on it consider using the `low_memory=True` option, which will switch to a slower but less memory intensive approach to computing the approximate nearest neighbors. This may alleviate your issues.

## 13.5 UMAP is eating all my cores. Help!

If run without a random seed UMAP will use numba's parallel implementation to do multithreaded work and use many cores. By default this will make use of as many cores as are available. If you are on a shared machine or otherwise don't wish to use *all* the cores at once you can restrict the number of threads that numba uses by making use of the environment variable `NUMBA_NUM_THREADS`; see the numba documentation for more details.

## 13.6 Is there GPU or multicore-CPU support?

There is basic multicore support as of version 0.4. In the future it is possible that GPU support may be added.

There is a UMAP implementation for GPU available in the NVIDIA RAPIDS cuML library, so if you need GPU support that is currently the best palce to go.

## 13.7 Can I add a custom loss function?

To allow for fast performance the SGD phase of UMAP has been hand-coded for the specific needs of UMAP. This makes custom loss functions a little difficult to handle. Now that Numba (as of version 0.38) supports passing functions it is posisble that future versions of UMAP may support such functionality. In the meantime you should definitely look into smallvis, a library for t-SNE, LargeVis, UMAP, and related algorithms. Smallvis only works for small datasets, but provides much greater flexibility and control.

## 13.8 Is there support for the R language?

Yes! A number of people have worked hard to make UMAP available to R users.

If you want to use the reference implementation under the hood but want a nice R interface then we recommend umap, which wraps the python code with reticulate. Another reticulate interface is umapr, but it may not be under active development.

If you want a pure R version then we recommend uwot at this time. umap also provides a pure R implementation in addition to its reticulate wrapper.

Both umap and uwot are available on CRAN.

## 13.9 Is there a C/C++ implementation?

Not that we are aware of. For now Numba has done a very admirable job of providing high performance and the developers of UMAP have not felt the need to move to lower level languages. At some point a multithreaded C++ implementation may be made available, but there are no time-frames for when that would happen.

## 13.10 I can't get UMAP to run properly!

There are, inevitably, a number of issues and corner cases that can cause issues for UMAP. Some know issues that can cause problems are:

- UMAP doesn't currently support 32-bit Windows. This is due to issues with Numba of that platform and will not likely be resolved soon. Sorry :-(
- If you have pip installed the package `umap` at any time (instead of `umap-learn`) this can cause serious issues. You will want to purge/remove everything umap related in your `site-packages` directory and re-install `umap-learn`.
- Having any files called `umap.py` in the current directory you will have issues as that will be loaded instead of the `umap` module.

It is worth checking the issues page on github for potential solutions. If all else fails please add an issue on github.

## 13.11 What is the difference between PCA / UMAP / VAEs?

This is an example of an embedding for a popular Fashion MNIST dataset.

Note that FMNIST is mostly a toy dataset (MNIST on steroids). On such a simplistic case UMAP shows distillation results (i.e. if we use its embedding in a downsteam task like classification) comparable to VAEs, which are more computationally expensive.

By definition:

- PCA is linear transformation, you can apply it to mostly any kind of data in an unsupervised fashion. Also it works really fast. For most real world tasks its embeddings are mostly too simplistic / useless.
- VAE is a kind of encoder-decoder neural network, trained with KLD loss and BCE (or MSE) loss to enforce the resulting embedding to be continuous. VAE is and extension of auto-encoder network, which by design should produce embeddings that are not only relevant to actually encoding the data, but are also smooth.

From a more practical standpoint:

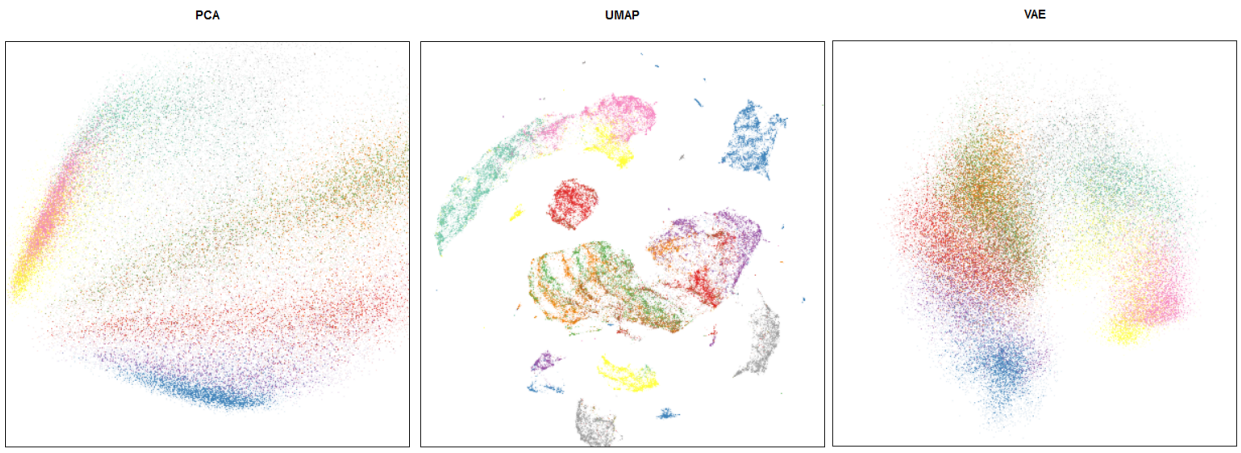



Fig. 1: Comparison of PCA / UMAP / VAE embeddings

- PCA mostly works for any reasonable dataset on a modern machine. (up to tens or hundreds of millions of rows);
- VAEs have been shown to work only for toy datasets and to our knowledge there was no real life useful application to a real world sized dataset (i.e. ImageNet);
- Applying UMAP to real world tasks usually provides a good starting point for downstream tasks (data visualization, clustering, classification) and works reasonably fast;
- Consider a typical pipeline: high-dimensional embedding (300+) => PCA to reduce to 50 dimensions => UMAP to reduce to 10-20 dimensions => HDBSCAN for clustering / some plain algorithm for classification;

Which tool should I use?

- PCA for very large or high dimensional datasets (or maybe consider finding a domain specific matrix factorization technique, e.g. topic modelling for texts);
- UMAP for smaller datasets;
- VAEs are mostly experimental;

Where can I learn more?

- While PCA is ubiqutous, you may look at this example comparing PCA / UMAP / VAEs;

## 13.12 Successful use-cases

UMAP can be / has been Successfully applied to the following domains:

- Single cell data visualization in biology;
- Mapping malware based on behavioural data;
- Pre-processing phrase vectors for clustering;
- Pre-processing image embeddings (Inception) for clustering;

and many more – if you have a successful use-case please submit a pull request adding it to this list!

# CHAPTER 14

# How UMAP Works

UMAP is an algorithm for dimension reduction based on manifold learning techniques and ideas from topological data analysis. It provides a very general framework for approaching manifold learning and dimension reduction, but can also provide specific concrete realizations. This article will discuss how the algorithm works in practice. There exist deeper mathematical underpinnings, but for the sake of readability by a general audience these will merely be referenced and linked. If you are looking for the mathematical description please see the UMAP paper.

To begin making sense of UMAP we will need a little bit of mathematical background from algebraic topology and topological data analysis. This will provide a basic algorithm that works well in theory, but unfortunately not so well in practice. The next step will be to make use of some basic Riemannian geometry to bring real world data a little closer to the underlying assumptions of the topological data analysis algorithm. Unfortunately this will introduce new complications, which will be resolved through a combination of deep math (details of which will be elided) and fuzzy logic. We can then put the pieces back together again, and combine them with a new approach to finding a low dimensional representation more fitting to the new data structures at hand. Putting this all together we arrive at the basic UMAP algorithm.

## 14.1 Topological Data Analysis and Simplicial Complexes

Simplicial complexes are a means to construct topological spaces out of simple combinatorial components. This allows one to reduce the complexities of dealing with the continuous geometry of topological spaces to the task of relatively simple combinatorics and counting. This method of taming geometry and topology will be fundamental to our approach to topological data analysis in general, and dimension reduction in particular.

The first step is to provide some simple combinatorial building blocks called *simplices*. Geometrically a simplex is a very simple way to build an $k$-dimensional object. A $k$ dimensional simplex is called a $k$-simplex, and it is formed by taking the convex hull of $k+1$ independent points. Thus a 0-simplex is a point, a 1-simplex is a line segment (between two zero simplices), a 2-simplex is a triangle (with three 1-simplices as "faces"), and a 3-simplex is a tetrahedron (with four 2-simplices as "faces"). Such a simple construction allows for easy generalization to arbitrary dimensions.

This has a very simple combinatorial underlying structure, and ultimately one can regard a $k$-simplex as an arbitrary set of $k+1$ objects with faces (and faces of faces etc.) given by appropriately sized subsets – one can always provide a "geometric realization" of this abstract set description by constructing the corresponding geometric simplex.

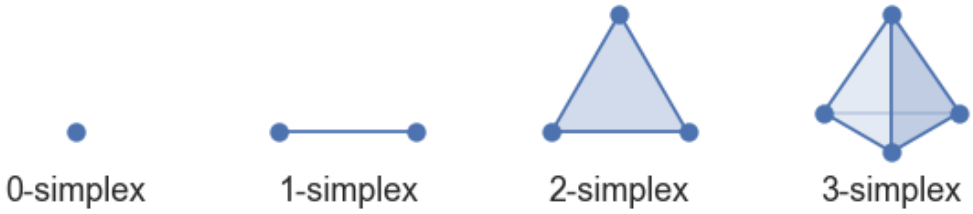



Fig. 1: Low dimensional simplices

Simplices can provide building blocks, but to construct interesting topological spaces we need to be able to glue together such building blocks. This can be done by constructing a *simplicial complex*. Ostensibly a simplicial complex is a set of simplices glued together along faces. More explicitly a simplicial complex $\mathcal{K}$ is a set of simplices such that any face of any simplex in $\mathcal{K}$ is also in $\mathcal{K}$ (ensuring all faces exist), and the intersection of any two simplices in $\mathcal{K}$ is a face of both simplices. A large class of topological spaces can be constructed in this way – just gluing together simplices of various dimensions along their faces. A little further abstraction will get to *simplicial sets* which are purely combinatorial, have a nice category theoretic presentation, and can generate a much broader class of topological spaces, but that will take us too far afield for this article. The intuition of simplicial complexes will be enough to illustrate the relevant ideas and motivation.

How does one apply these theoretical tools from topology to finite sets of data points? To start we'll look at how one might construct a simplicial complex from a topological space. The tool we will consider is the construction of a Čech complex given an open cover of a topological space. That's a lot of verbiage if you haven't done much topology, but we can break it down fairly easily for our use case. An open cover is essentially just a family of sets whose union is the whole space, and a Čech complex is a combinatorial way to covert that into a simplicial complex. It works fairly simply: let each set in the cover be a 0-simplex; create a 1-simplex between two such sets if they have a non-empty intersection; create a 2-simplex between three such sets if the triple intersection of all three is non-empty; and so on. Now, that doesn't sound very advanced – just looking at intersections of sets. The key is that the background topological theory actually provides guarantees about how well this simple process can produce something that represents the topological space itself in a meaningful way (the Nerve theorem is the relevant result for those interested). Obviously the quality of the cover is important, and finer covers provide more accuracy, but the reality is that despite its simplicity the process captures much of the topology.

Next we need to understand how to apply that process to a finite set of data samples. If we assume that the data samples are drawn from some underlying topological space then to learn about the topology of that space we need to generate an open cover of it. If our data actually lie in a metric space (i.e. we can measure distance between points) then one way to approximate an open cover is to simply create balls of some fixed radius about each data point. Since we only have finite samples, and not the topological space itself, we cannot be sure it is truly an open cover, but it might be as good an approximation as we could reasonably expect. This approach also has the advantage that the Čech complex associated to the cover will have a 0-simplex for each data point.

To demonstrate the process let's consider a test dataset like this

If we fix a radius we can then picture the open sets of our cover as circles (since we are in a nice visualizable two dimensional case). The result is something like this

We can then depict the the simplicial complex of 0-, 1-, and 2-simplices as points, lines, and triangles

It is harder to easily depict the higher dimensional simplices, but you can imagine how they would fit in. There are two things to note here: first, the simplicial complex does a reasonable job of starting to capture the fundamental topology of the dataset; second, most of the work is really done by the 0- and 1-simplices, which are easier to deal with computationally (it is just a graph, in the nodes and edges sense). The second observation motivates the Vietoris-Rips

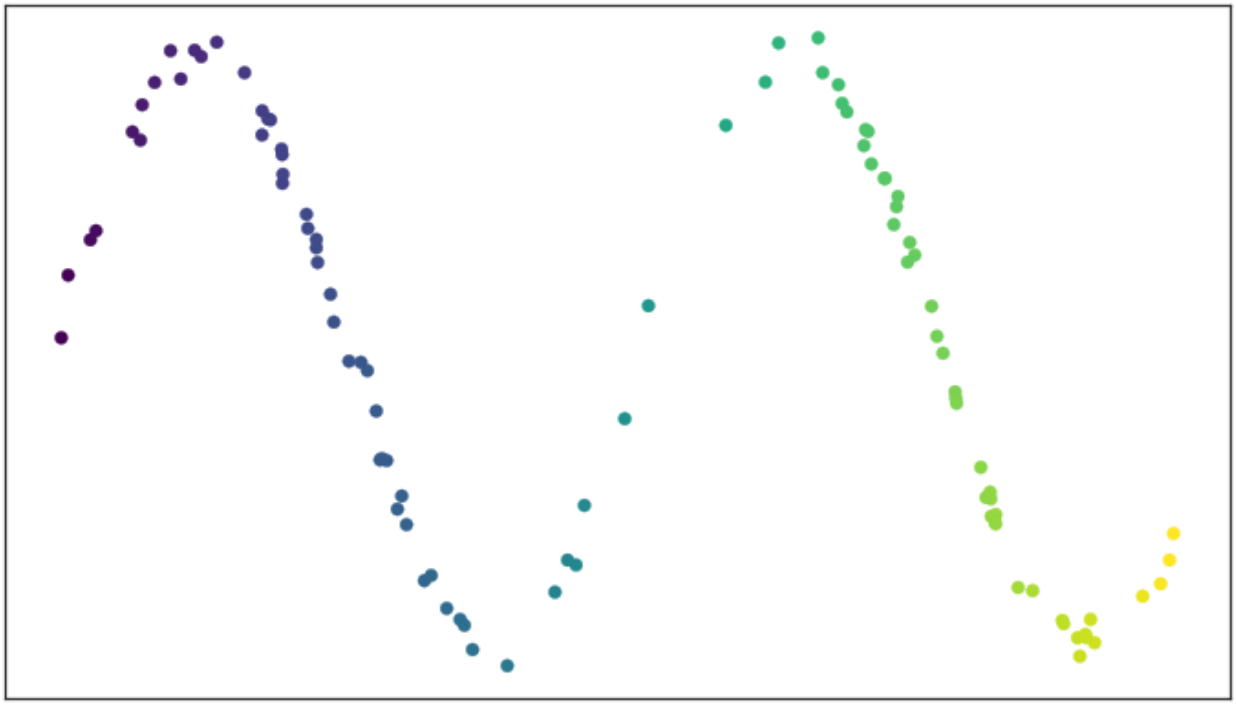


Fig. 2: Test data set of a noisy sine wave

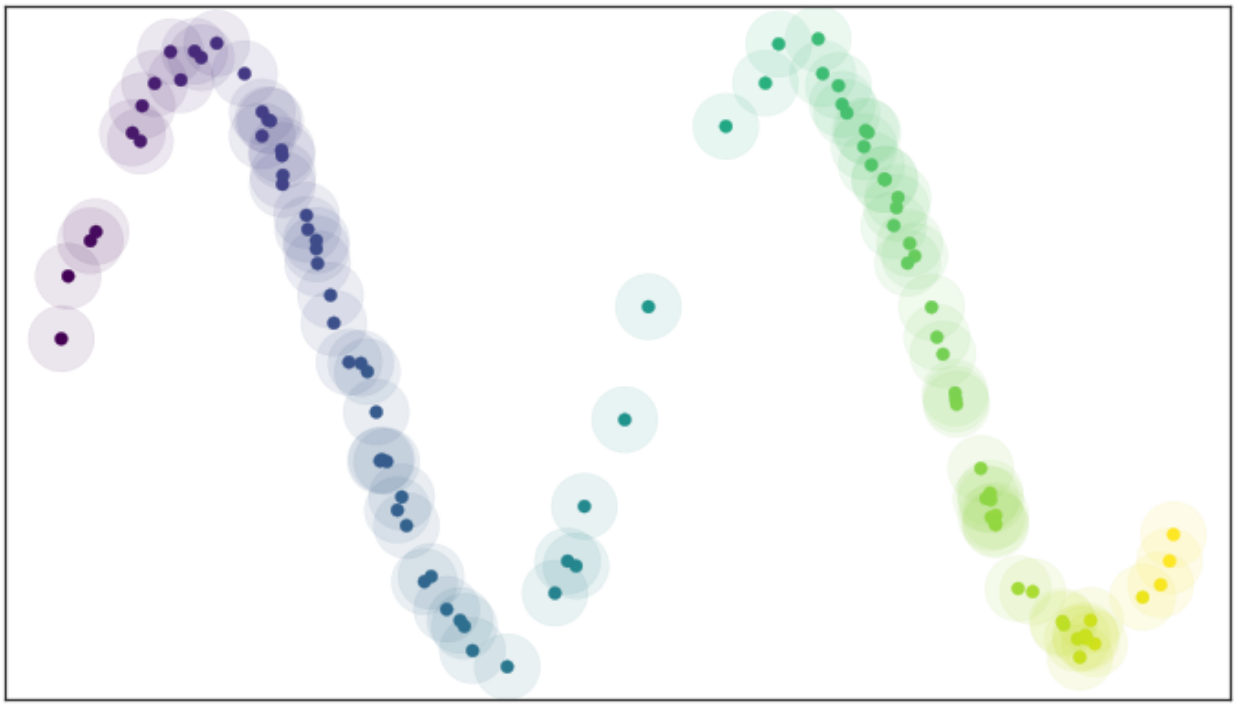


Fig. 3: A basic open cover of the test data

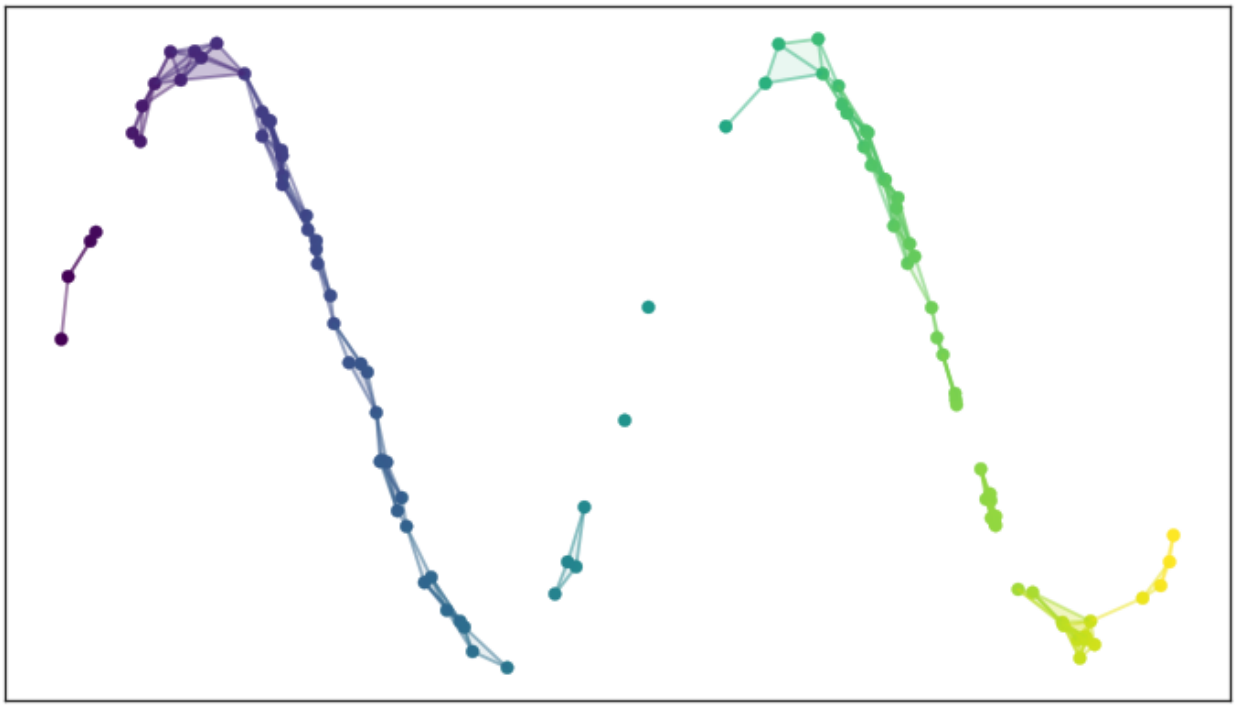


Fig. 4: A simplicial complex built from the test data

complex, which is similar to the Čech complex but is entirely determined by the 0- and 1-simplices. Vietoris-Rips complexes are much easier to work with computationally, especially for large datasets, and are one of the major tools of topological data analysis.

If we take this approach to get a topological representation then we can build a dimension reduction algorithm by finding a low dimensional representation of the data that has a similar topological representation. If we only care about the 0- and 1-simplices then the topological representation is just a graph, and finding a low dimensional representation can be described as a **'graph layout problem <>'__**. If one wants to use, for example, spectral methods for graph layout then we arrive at algorithms like **'Laplacian eigenmaps <>'__** and **'Diffusion maps <>'__**. Force directed layouts are also an option, and provide algorithms closer to **'MDS <>'__** or **'Sammon mapping <>'__** in flavour.

I would not blame those who have read this far to wonder why we took such an abstract roundabout road to simply building a neighborhood-graph on the data and then laying out that graph. There are a couple of reasons. The first reason is that the topological approach, while abstract, provides sound theoretical justification for what we are doing. While building a neighborhood-graph and laying it out in lower dimensional space makes heuristic sense and is computationally tractable, it doesn't provide the same underlying motivation of capturing the underlying topological structure of the data faithfully – for that we need to appeal to the powerful topological machinery I've hinted lies in the background. The second reason is that it is this more abstract topological approach that will allow us to generalize the approach and get around some of the difficulties of the sorts of algorithms described above. While ultimately we will end up with a process that is fairly simple computationally, understanding *why* various manipulations matter is important to truly understanding the algorithm (as opposed to merely computing with it).

## 14.2 Adapting to Real World Data

The approach described above provides a nice theory for why a neighborhood graph based approach should capture manifold structure when doing dimension reduction. The problem tends to come when one tries to put the theory into practice. The first obvious difficulty (and we can see it even our example above) is that choosing the right radius for

the balls that make up the open cover is hard. If you choose something too small the resulting simplicial complex splits into many connected components. If you choose something too large the simplicial complex turns into just a few very high dimensional simplices (and their faces etc.) and fails to capture the manifold structure anymore. How should one solve this?

The dilemma is in part due to the theorem (called the Nerve theorem) that provides our justification that this process captures the topology. Specifically, the theorem says that the simplicial complex will be (homtopically) equivalent to the union of the cover. In our case, working with finite data, the cover, for certain radii, doesn't cover the whole of the manifold that we imagine underlies the data – it is that lack of coverage that results in the disconnected components. Similarly, where the points are too bunched up, our cover does cover "too much" and we end up with higher dimensional simplices than we might ideally like. If the data were *uniformly distributed* across the manifold then selecting a suitable radius would be easy – the average distance between points would work well. Moreover with a uniform distribution we would be guaranteed that our cover would actually cover the whole manifold with no "gaps" and no unnecessarily disconnected components. Similarly, we would not suffer from those unfortunate bunching effects resulting in unnecessarily high dimensional simplices.

If we consider data that is uniformly distributed along the same manifold it is not hard to pick a good radius (a little above half the average distance between points) and the resulting open cover looks pretty good:

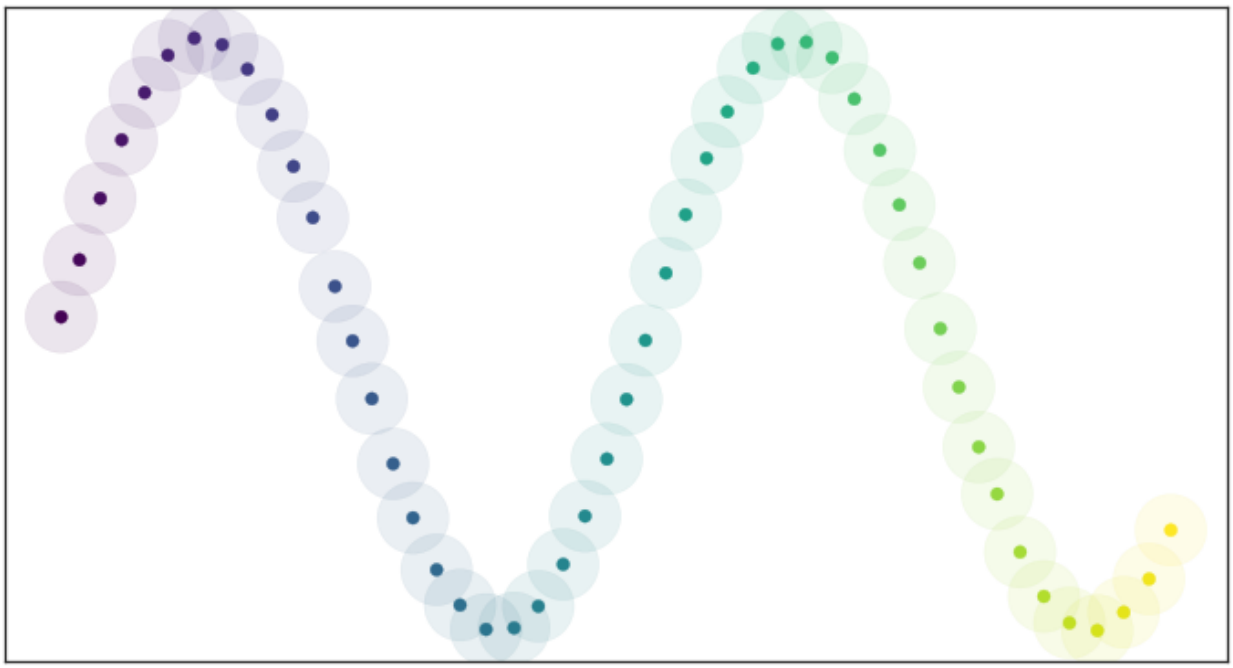


Fig. 5: Open balls over uniformly_distributed_data

Because the data is evenly spread we actually cover the underlying manifold and don't end up with clumping. In other words, all this theory works well assuming that the data is uniformly distributed over the manifold.

Unsurprisingly this uniform distribution assumption crops up elsewhere in manifold learning. The proofs that Laplacian eigenmaps work well require the assumption that the data is uniformly distributed on the manifold. Clearly if we had a uniform distribution of points on the manifold this would all work a lot better – but we don't! Real world data simply isn't that nicely behaved. How can we resolve this? By turning the problem on its head: assume that the data is uniformly distributed on the manifold, and ask what that tells us about the manifold itself. If the data *looks* like it isn't uniformly distributed that must simply be because the notion of distance is varying across the manifold – space itself is warping: stretching or shrinking according to where the data appear sparser or denser.

By assuming that the data is uniformly distributed we can actually compute (an approximation of) a local notion of distance for each point by making use of a little standard Riemannian geometry. In practical terms, once you push the

math through, this turns out to mean that a unit ball about a point stretches to the $k$-th nearest neighbor of the point, where $k$ is the sample size we are using to approximate the local sense of distance. Each point is given its own unique distance function, and we can simply select balls of radius one with respect to that local distance function!

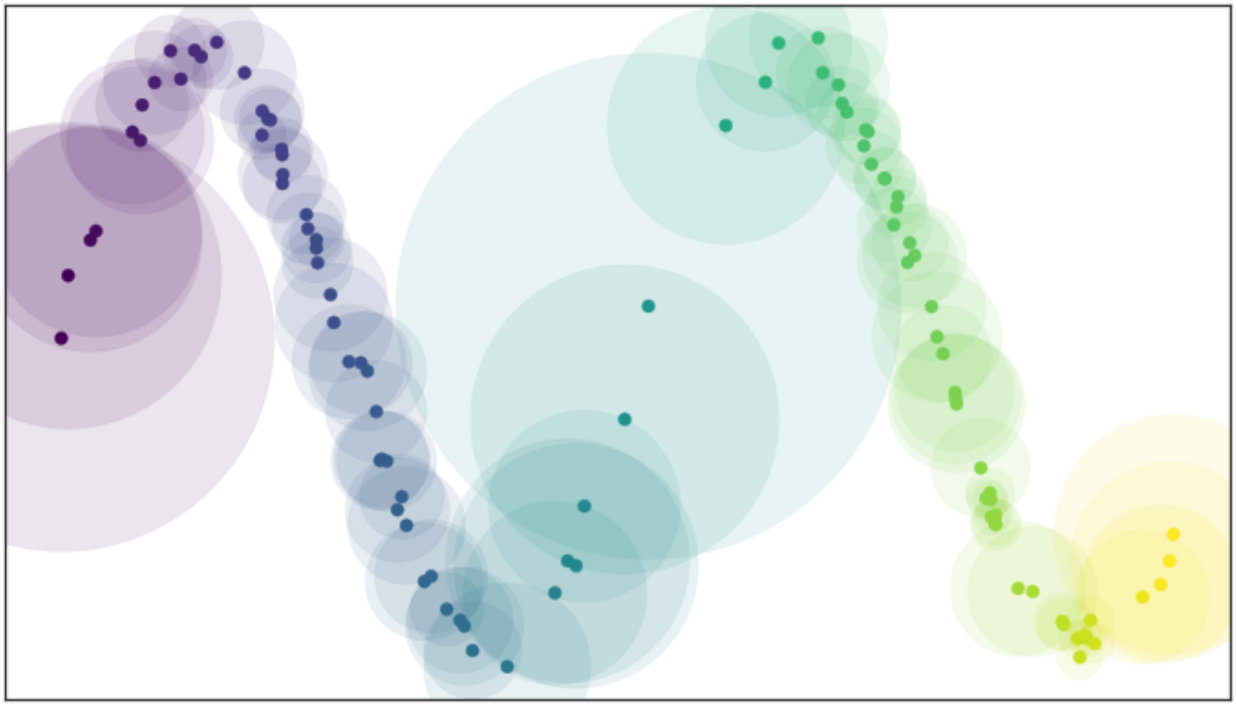


Fig. 6: Open balls of radius one with a locally varying metric

This theoretically derived result matches well with many traditional graph based algorithms: a standard approach for such algorithms is to use a $k$-neighbor graph instead of using balls of some fixed radius to define connectivity. What this means is that each point in the dataset is given an edge to each of its $k$ nearest neighbors – the effective result of our locally varying metric with balls of radius one. Now, however, we can explain why this works in terms of simplicial complexes and the Nerve theorem.

Of course we have traded choosing the radius of the balls for choosing a value for $k$. However it is often easier to pick a resolution scale in terms of number of neighbors than it is to correctly choose a distance. This is because choosing a distance is very dataset dependent: one needs to look at the distribution of distances in the dataset to even begin to select a good value. In contrast, while a $k$ value is still dataset dependent to some degree, there are reasonable default choices, such as the 10 nearest neighbors, that should work acceptably for most datasets.

At the same time the topological interpretation of all of this gives us a more meaningful interpretation of $k$. The choice of $k$ determines how locally we wish to estimate the Riemannian metric. A small choice of $k$ means we want a very local interpretation which will more accurately capture fine detail structure and variation of the Riemannian metric. Choosing a large $k$ means our estimates will be based on larger regions, and thus, while missing some of the fine detail structure, they will be more broadly accurate across the manifold as a whole, having more data to make the estimate with.

We also get a further benefit from this Riemannian metric based approach: we actually have a local metric space associated to each point, and can meaningfully measure distance, and thus we could weight the edges of the graph we might generate by how far apart (in terms of the local metric) the points on the edges are. In slightly more mathematical terms we can think of this as working in a fuzzy topology where being in an open set in a cover is no longer a binary yes or no, but instead a fuzzy value between zero and one. Obviously the certainty that points are in a ball of a given radius will decay as we move away from the center of the ball. We could visualize such a fuzzy cover as looking something like this

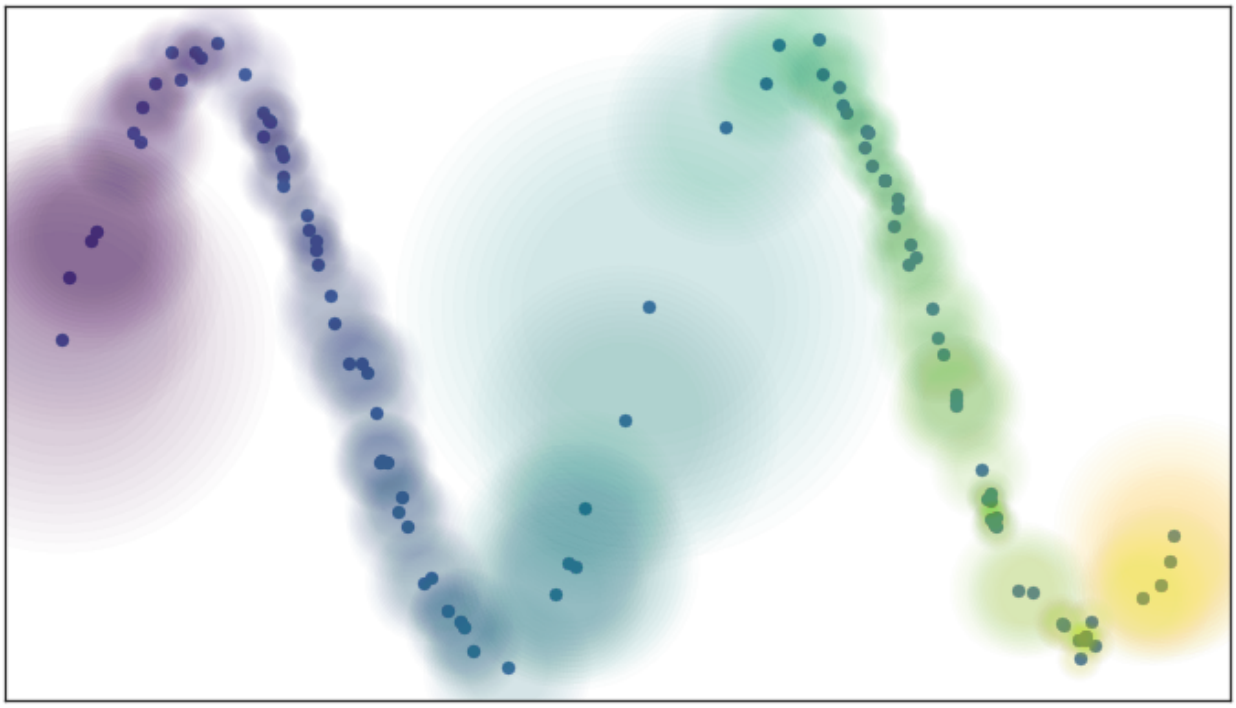


Fig. 7: Fuzzy open balls of radius one with a locally varying metric

None of that is very concrete or formal – it is merely an intuitive picture of what we would like to have happen. It turns out that we can actually formalize all of this by stealing the singular set and geometric realization functors from algebraic topology and then adapting them to apply to metric spaces and fuzzy simplicial sets. The mathematics involved in this is outside the scope of this exposition, but for those interested you can look at the original work on this by David Spivak and our paper. It will have to suffice to say that there is some mathematical machinery that lets us realize this intuition in a well defined way.

This resolves a number of issues, but a new problem presents itself when we apply this sort of process to real data, especially in higher dimensions: a lot of points become essentially totally isolated. One would imagine that this shouldn't happen if the manifold the data was sampled from isn't pathological. So what property are we expecting that manifold to have that we are somehow missing with the current approach? What we need to add is the idea of local connectivity.

Note that this is not a requirement that the manifold as a whole be connected – it can be made up of many connected components. Instead it is a requirement that at any point on the manifold there is some sufficiently small neighborhood of the point that *is* connected (this "in a sufficiently small neighborhood" is what the "local" part means). For the practical problem we are working with, where we only have a finite approximation of the manifold, this means that no point should be *completely* isolated – it should connect to at least one other point. In terms of fuzzy open sets what this amounts to is that we should have complete confidence that the open set extends as far as the closest neighbor of each point. We can implement this by simply having the fuzzy confidence decay in terms of distance *beyond* the first nearest neighbor. We can visualize the result in terms of our example dataset again.

Again this can be formalized in terms of the aforementioned mathematical machinery from algebraic topology. From a practical standpoint this plays an important role for high dimensional data – in high dimensions distances tend to be larger, but also more similar to one another (see the curse of dimensionality). This means that the distance to the first nearest neighbor can be quite large, but the distance to the tenth nearest neighbor can often be only slightly larger (in relative terms). The local connectivity constraint ensures that we focus on the difference in distances among nearest neighbors rather than the absolute distance (which shows little differentiation among neighbors).

Just when we think we are almost there, having worked around some of the issues of real world data, we run aground

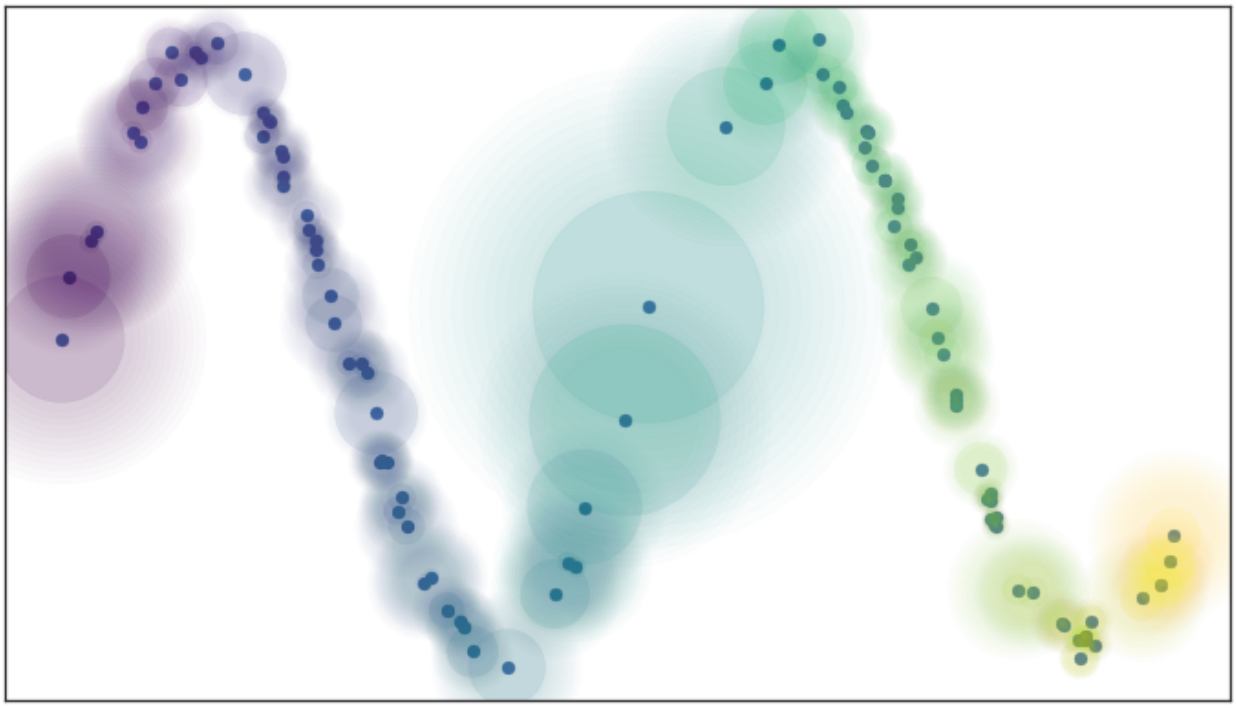


Fig. 8: Local connectivity and fuzzy open sets

on a new obstruction: our local metrics are not compatible! Each point has its own local metric associated to it, and from point *a*'s perspective the distance from point *a* to point *b* might be 1.5, but from the perspective of point *b* the distance from point *b* to point *a* might only be 0.6. Which point is right? How do we decide? Going back to our graph based intuition we can think of this as having directed edges with varying weights something like this.

Between any two points we might have up to two edges and the weights on those edges disagree with one another. There are a number of options for what to do given two disagreeing weights – we could take the maximum, the minimum, the arithmetic mean, the geometric mean, or something else entirely. What we would really like is some principled way to make the decision. It is at this point that the mathematical machinery we built comes into play. Mathematically we actually have a family of fuzzy simplicial sets, and the obvious choice is to take their union – a well defined operation. There are a a few ways to define fuzzy unions, depending on the nature of the logic involved, but here we have relatively clear probabilistic semantics that make the choice straightforward. In graph terms what we get is the following: if we want to merge together two disagreeing edges with weight *a* and *b* then we should have a single edge with combined weight $a + b - a \cdot b$. The way to think of this is that the weights are effectively the probabilities that an edge (1-simplex) exists. The combined weight is then the probability that at least one of the edges exists.

If we apply this process to union together all the fuzzy simplicial sets we end up with a single fuzzy simplicial complex, which we can again think of as a weighted graph. In computational terms we are simply applying the edge weight combination formula across the whole graph (with non-edges having a weight of 0). In the end we have something that looks like this.

So in some sense in the end we have simply constructed a weighted graph (although we could make use of higher dimensional simplices if we wished, just at significant extra computational cost). What the mathematical theory lurking in the background did for us is determine *why* we should construct *this* graph. It also helped make the decisions about exactly *how* to compute things, and gives a concrete interpretation of *what* this graph means. So while in the end we just constructed a graph, the math answered the important questions to get us here, and can help us determine what to do next.

So given that we now have a fuzzy topological representation of the data (which the math says will capture the topology

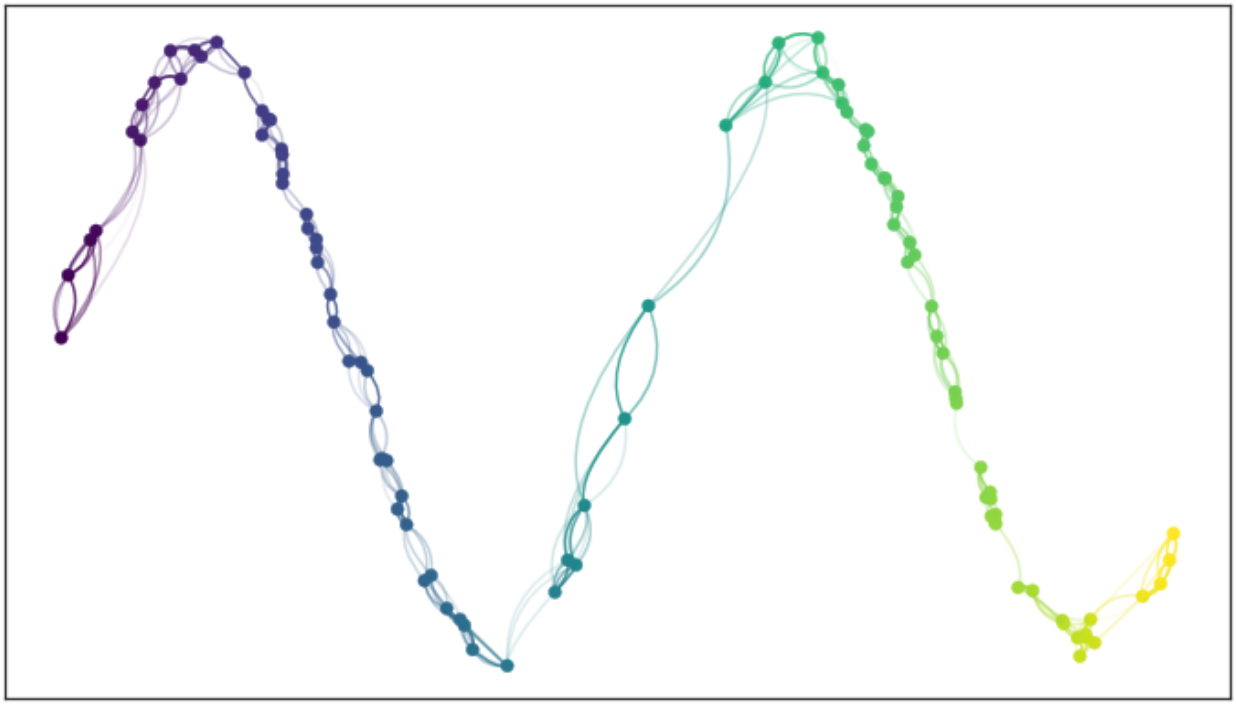


Fig. 9: Edges with incompatible weights

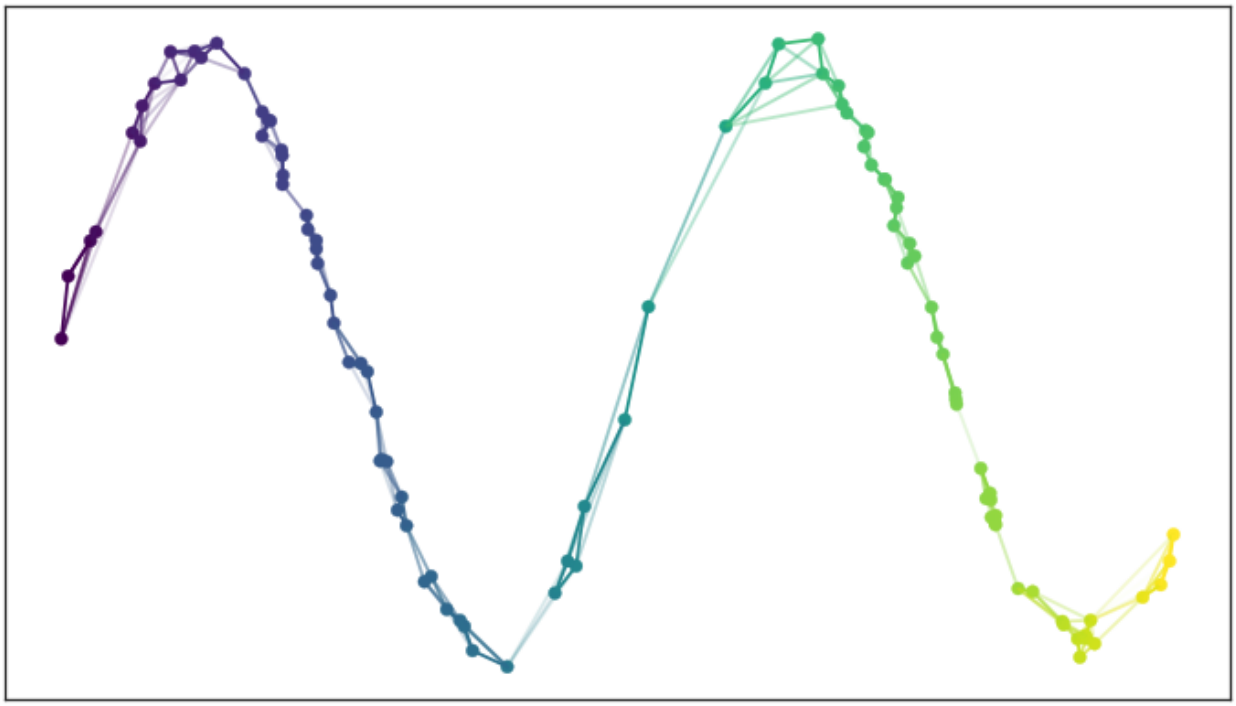


Fig. 10: Graph with combined edge weights

of the manifold underlying the data), how do we go about converting that into a low dimensional representation?

## 14.3 Finding a Low Dimensional Representation

Ideally we want the low dimensional representation to have as similar a fuzzy topological structure as possible. The first question is how do we determine the fuzzy topological structure of a low dimensional representation, and the second question is how do we find a good one.

The first question is largely already answered – we should presumably follow the same procedure we just used to find the fuzzy topological structure of our data. There is a quirk, however: this time around the data won't be lying on some manifold, we'll have a low dimensional representation that is lying on a very particular manifold. That manifold is, of course, just the low dimensional euclidean space we are trying to embed into. This means that all the effort we went to previously to make vary the notion of distance across the manifold is going to be misplaced when working with the low dimensional representation. We explicitly *want* the distance on the manifold to be standard euclidean distance with respect to the global coordinate system, not a varying metric. That saves some trouble. The other quirk is that we made use of the distance to the nearest neighbor, again something we computed given the data. This is also a property we would like to be globally true across the manifold as we optimize toward a good low dimensional representation, so we will have to accept it as a hyper-parameter `min_dist` to the algorithm.

The second question, 'how do we find a good low dimensional representation', hinges on our ability to measure how "close" a match we have found in terms of fuzzy topological structures. Given such a measure we can turn this into an optimization problem of finding the low dimensional representation with the closest fuzzy topological structure. Obviously if our measure of closeness turns out to have various properties the nature of the optimization techniques we can apply will differ.

Going back to when we were merging together the conflicting weights associated to simplices, we interpreted the weights as the probability of the simplex existing. Thus, since both topological structures we are comparing share the same 0-simplices, we can imagine that we are comparing the two vectors of probabilities indexed by the 1-simplices. Given that these are Bernoulli variables (ultimately the simplex either exists or it doesn't, and the probability is the parameter of a Bernoulli distribution), the right choice here is the cross entropy.

Explicitly, if the set of all possible 1-simplices is $E$, and we have weight functions such that $w_h(e)$ is the weight of the 1-simplex $e$ in the high dimensional case and $w_l(e)$ is the weight of $e$ in the low dimensional case, then the cross entropy will be

$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right)$$

This might look complicated, but if we go back to thinking in terms of a graph we can view minimizing the cross entropy as a kind of force directed graph layout algorithm.

The first term, $w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right)$, provides an attractive force between the points $e$ spans whenever there is a large weight associated to the high dimensional case. This is because this term will be minimized when $w_l(e)$ is as large as possible, which will occur when the distance between the points is as small as possible.

In contrast the second term, $(1 - w_h(e)) \log\left(\frac{1-w_h(e)}{1-w_l(e)}\right)$, provides a repulsive force between the ends of $e$ whenever $w_h(e)$ is small. This is because the term will be minimized by making $w_l(e)$ as small as possible.

On balance this process of pull and push, mediated by the weights on edges of the topological representation of the high dimensional data, will let the low dimensional representation settle into a state that relatively accurately represents the overall topology of the source data.

## 14.4 The UMAP Algorithm

Putting all these pieces together we can construct the UMAP algorithm. The first phase consists of constructing a fuzzy topological representation, essentially as described above. The second phase is simply optimizing the low dimensional representation to have as close a fuzzy topological representation as possible as measured by cross entropy.

When constructing the initial fuzzy topological representation we can take a few shortcuts. In practice, since fuzzy set membership strengths decay away to be vanishingly small, we only need to compute them for the nearest neighbors of each point. Ultimately that means we need a way to quickly compute (approximate) nearest neighbors efficiently, even in high dimensional spaces. We can do this by taking advantage of the Nearest-Neighbor-Descent algorithm of Dong et al. The remaining computations are now only dealing with local neighbors of each point and are thus very efficient.

In optimizing the low dimensional embedding we can again take some shortcuts. We can use stochastic gradient descent for the optimization process. To make the gradient descent problem easier it is beneficial if the final objective function is differentiable. We can arrange for that by using a smooth approximation of the actual membership strength function for the low dimensional representation, selecting from a suitably versatile family. In practice UMAP uses the family of curves of the form $\frac{1}{1+ax^{2b}}$. Equally we don't want to have to deal with all possible edges, so we can use the negative sampling trick (as used by word2vec and LargeVis), to simply sample negative examples as needed. Finally since the Laplacian of the topological representation is an approximation of the Laplace-Beltrami operator of the manifold we can use spectral embedding techniques to initialize the low dimensional representation into a good state.

Putting all these pieces together we arrive at an algorithm that is fast and scalable, yet still built out of sound mathematical theory. Hopefully this introduction has helped provide some intuition for that underlying theory, and for how the UMAP algorithm works in practice.

# CHAPTER 15

# Performance Comparison of Dimension Reduction Implementations

Different dimension reduction techniques can have quite different computational complexity. Beyond the algorithm itself there is also the question of how exactly it is implemented. These two factors can have a significant role in how long it actually takes to run a given dimension reduction. Furthermore the nature of the data you are trying to reduce can also matter – mostly the involves the dimensionality of the original data. Here we will take a brief look at the performance characterstics of a number of dimension reduction implementations.

To start let's get the basic tools we'll need loaded up – numpy and pandas obviously, but also tools to get and resample the data, and the time module so we can perform some basic benchmarking.

Next we'll need the actual dimension reduction implementations. For the purposes of this explanation we'll mostly stick with scikit-learn, but for the sake of comparison we'll also include the MulticoreTSNE implementation of t-SNE, which has significantly better performance than the current scikit-learn t-SNE.

Next we'll need out plotting tools, and, of course, some data to work with. For this performance comparison we'll default to the now standard benchmark of manifold learning: the MNIST digits dataset. We can use scikit-learn's `fetch_mldata` to grab it for us.

Now it is time to start looking at performance. To start with let's look at how performance scales with increasing dataset size.

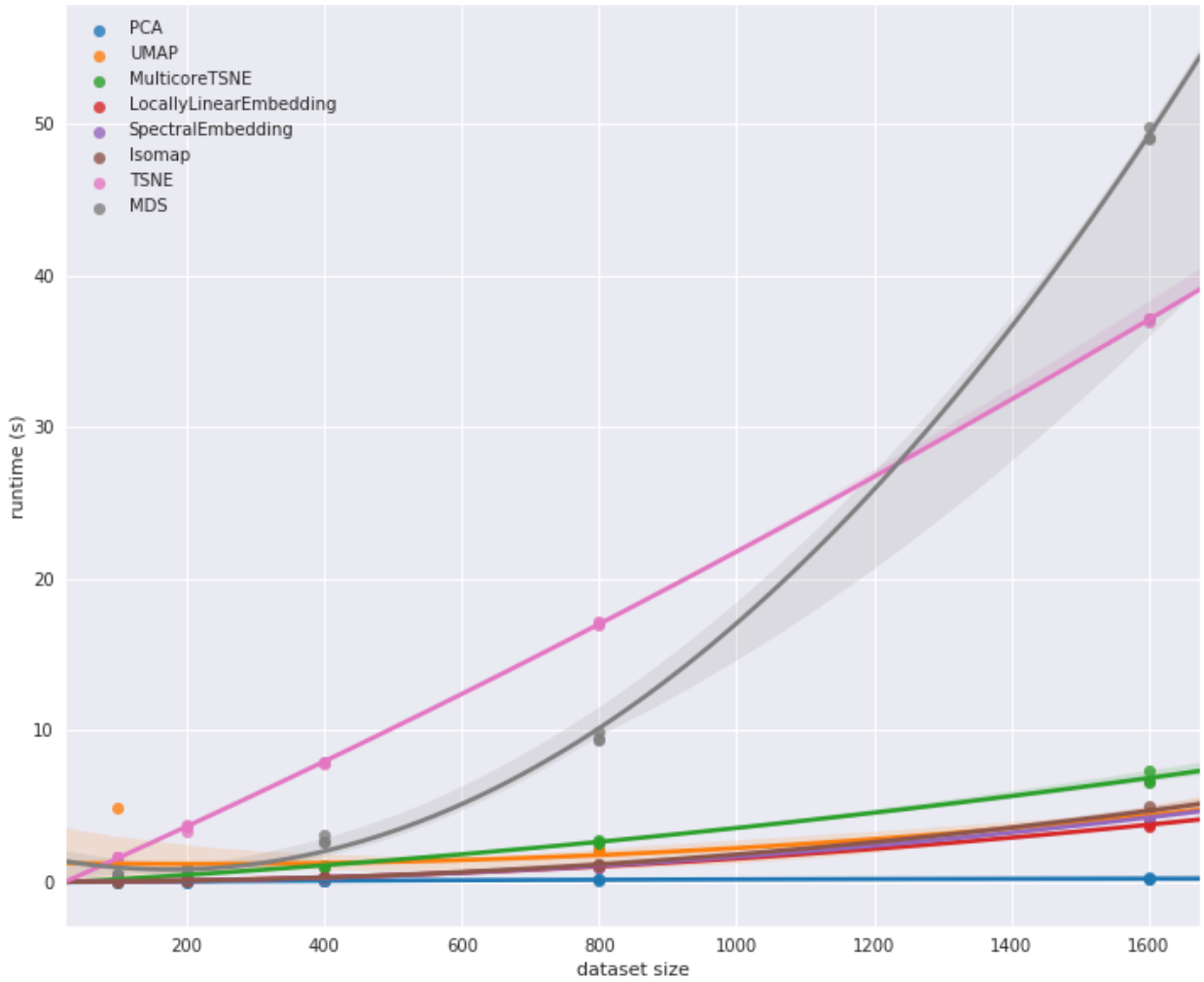
## 15.1 Performance scaling by dataset size

As the size of a dataset increases the runtime of a given dimension reduction algorithm will increase at varying rates. If you ever want to run your algorithm on larger datasets you will care not just about the comparative runtime on a single small dataset, but how the performance scales out as you move to larger datasets. We can similate this by subsampling from MNIST digits (via scikit-learn's convenient `resample` utility) and looking at the runtime for varying sized subsamples. Since there is some randomness involved here (both in the subsample selection, and in some of the algorithms which have stochastic aspects) we will want to run a few examples for each dataset size. We can easily package all of this up in a simple function that will return a convenient pandas dataframe of dataset sizes and runtimes given an algorithm.

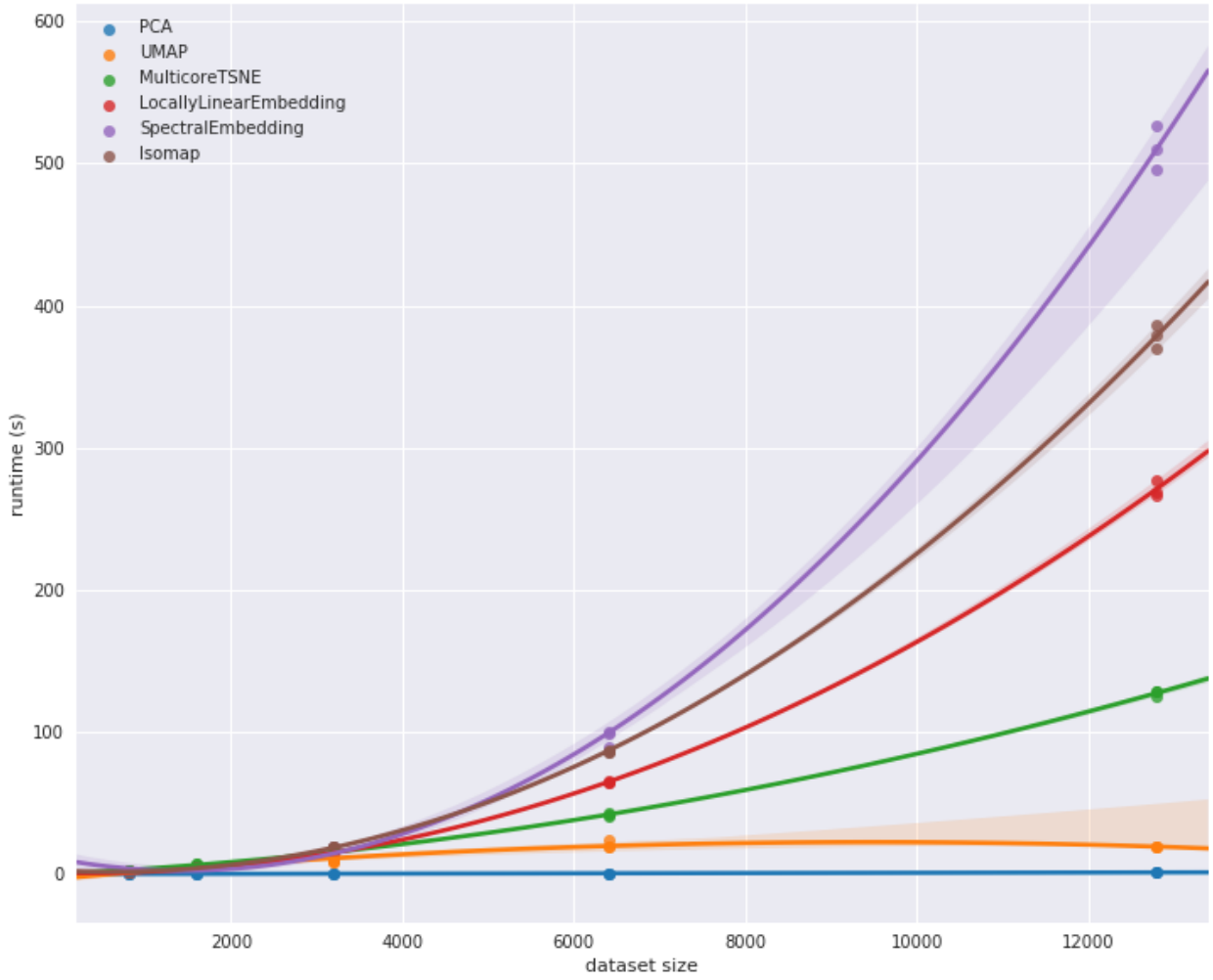
Now we just want to run this for each of the various dimension reduction implementations so we can look at the results. Since we don't know how long these runs might take we'll start off with a very small set of samples, scaling up to only

1600 samples.

Now let's plot the results so we can see what is going on. We'll use seaborn's regression plot to interpolate the effective scaling.

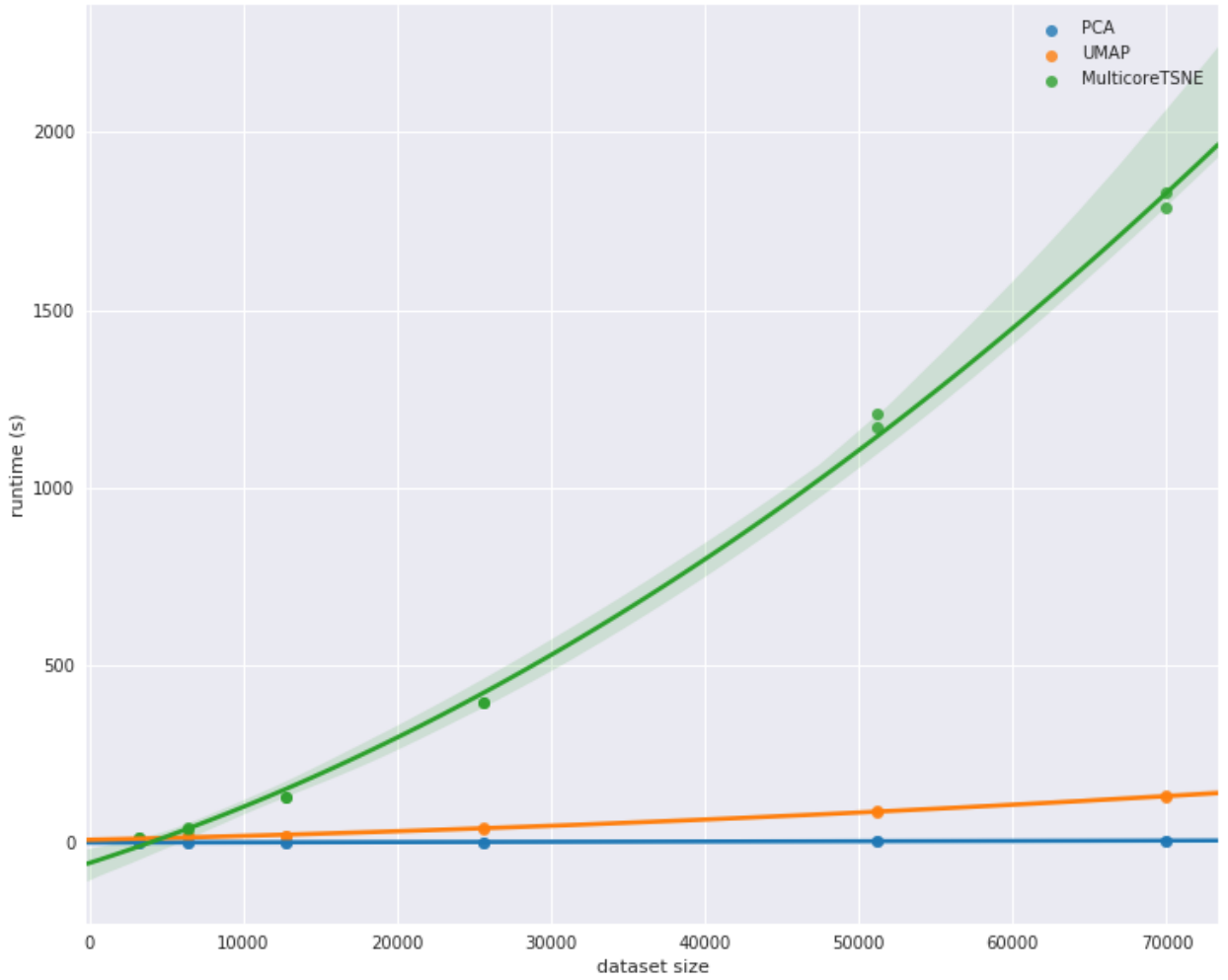



We can see straight away that there are some outliers here. The scikit-learn t-SNE is clearly much slower than most of the other algorithms. It does not have the scaling properties of MDS however; for larger dataset sizes MDS is going to quickly become completely unmanageable. At the same time MulticoreTSNE demonstrates that t-SNE can run fairly efficiently. It is hard to tell much about the other implementations other than the fact that PCA is far and away the fastest option. To see more we'll have to look at runtimes on larger dataset sizes. Both MDS and scikit-learn's t-SNE are going to take too long to run so let's restrict ourselves to the fastest performing implementations and see what happens as we extend out to larger dataset sizes.




At this point we begin to see some significant differentiation among the different implementations. In the earlier plot MulticoreTSNE looked to be slower than some of the other algorithms, but as we scale out to larger datasets we see that its relative scaling performance is far superior to the scikit-learn implementations of Isomap, spectral embedding, and locally linear embedding.

It is probably worth extending out further – up to the full MNIST digits dataset. To manage to do that in any reasonable amount of time we'll have to restrict out attention to an even smaller subset of implementations. We will pare things down to just MulticoreTSNE, PCA and UMAP.




Here we see UMAP's advantages over t-SNE really coming to the forefront. While UMAP is clearly slower than PCA, its scaling performance is dramatically better than MulticoreTSNE, and for even larger datasets the difference is only going to grow.

This concludes our look at scaling by dataset size. The short summary is that PCA is far and away the fastest option, but you are potentially giving up a lot for that speed. UMAP, while not competitive with PCA, is clearly the next best option in terms of performance among the implementations explored here. Given the quality of results that UMAP can provide we feel it is clearly a good option for dimension reduction.

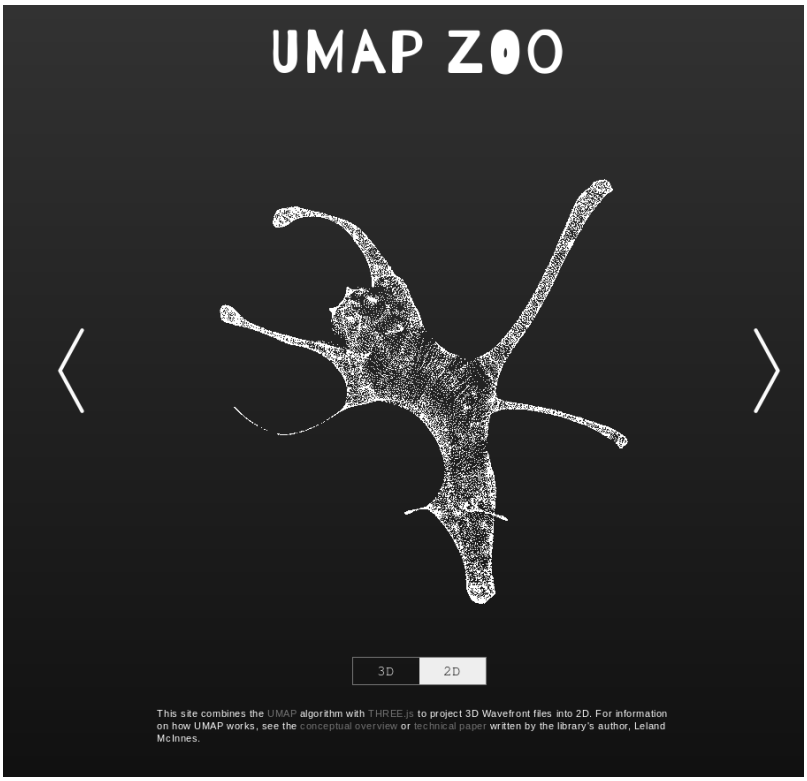
# CHAPTER 16

# Interactive Visualizations

UMAP has found use in a number of interesting interactive visualization projects, analyzing everything from images from photo archives, to word embedding, animal point clouds, and even sound. Sometimes it has also been used in interesting interactive tools that simply help a user to get an intuition for what the algorithm is doing (by applying it to intuitive 3D data). Below are some amazing projects that make use of UMAP.

## 16.1 UMAP Zoo

An exploration of how UMAP behaves when dimension reducing point clouds of animals. It is interactive, letting you switch between 2D and 3D representations and has a wide selection of different animals. Attempting to guess the animal from the 2D UMAP representation is a fun game. In practice this tool can go a long way to helping to build at least some intuitions for what UMAP tends to do with data.

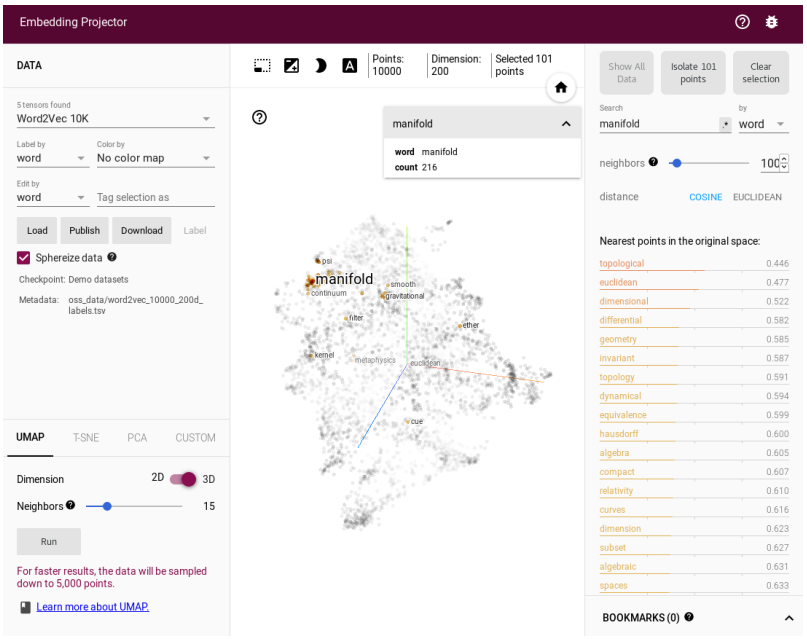



UMAP Zoo

Thanks to Douglas Duhaime.

## 16.2 Tensorflow Embedding Projector

If you just want to explore UMAP embeddings of datasets then the Embedding Projector from Tensorflow is a great way to do that. As well as having a good interactive 3D view it also has facilities for inspecting and searching labels and tags on the data. By default it loads up word2vec vectors, but you can uplod any data you wish. You can then select the UMAP option among the tabs for embeddings choices (alongside PCA and t-SNE).

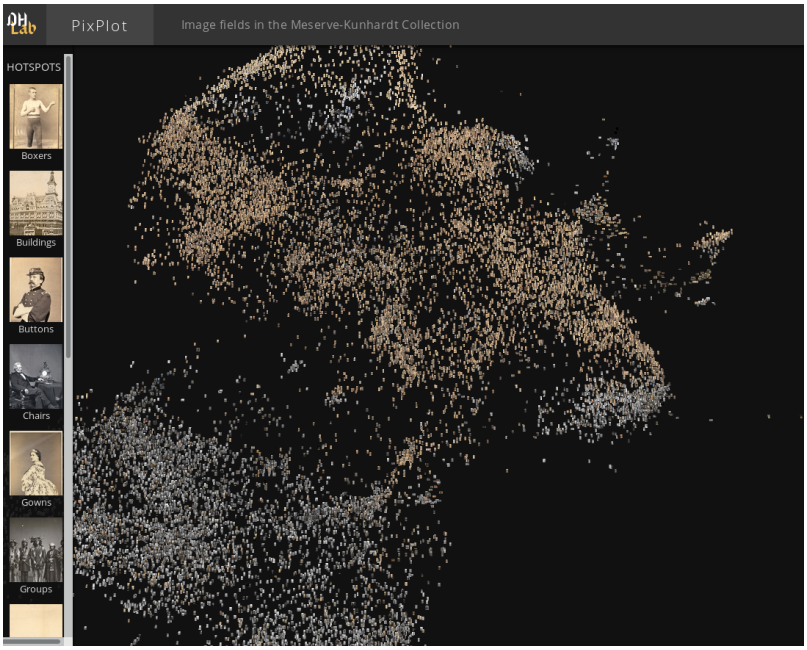



Embedding Projector

Thanks to Andy Coenen and the Embedding Projector team.

## 16.3 PixPlot

PixPlot provides an overview of large photo-collections. In the demonstration app from Yale's Digital Humanities lab it provides a window on the Meserve-Kunhardt Collection of historical photographs. The approach uses convolutional neural nets to reduce the images to 2048 dimensions, and then uses UMAP to present them in a 2-dimensional map which the user can interactive pan and zoom around in. This process results in similar photos ending up in similar regions of the map allowing for easy perusal of large photo collections. The PixPlot project is also available on github in case you wish to train it on your own photo collection.




PixPlot

Thanks to Douglas Duhaime and the Digital Humanities lab at Yale.

## 16.4 UMAP Explorer

A great demonstration of building a web based app for interactively exploring a UMAP embedding. In this case it provides an exploration of UMAP run on the MNIST digits dataset. Each point in the embeddijng is rendered as the digit image, and coloured according to the digit class. Mousing over the images will make them larger and provide a view of the digit in the upper left. You can also pan and zoom around the emebdding to get a better understanding of how UMAP has mapped the different styles of handwritten digits down to 2 dimensions.

UMAP Explorer

Thanks for Grant Custer.

## 16.5 Audio Explorer

The Audio Explorer uses UMAP to embed sound samples into a 2 dimensional space for easy exploration. The goal here is to take a large library of sounds samples and put similar sounds in similar regions of the map, allowing a user to quickly mouse over and listen to various variations of a given sample to quickly find exactly the right sound sample to use. Audio explorer uses MFCCs and/or WaveNet to provide an initial useful vector representation of the sound samples, before applying UMAP to generate the 2D embedding.


Audio Explorer

Thanks to Leon Fedden.

# CHAPTER 17

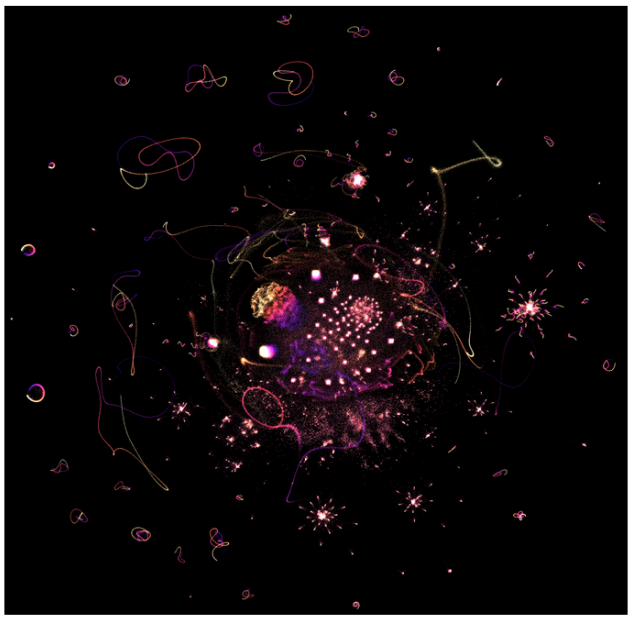
# Exploratory Analysis of Interesting Datasets

UMAP is a useful tool for general exploratory analysis of data – it can provide a unique lens through which to view data that can highlight structures and properties hiding in data that are not as apparent when analysed with other techniques. Below is a selection of uses cases of UMAP being used for interesting explorations of intriguing datasets – everything from pure math and outputs of neural networks, to philosophy articles, and scientific texts.

## 17.1 Prime factorizations of numbers

What would happen if we applied UMAP to the integers? First we would need a way to express an integer in a high dimensional space. That can be done by looking at the prime factorization of each number. Next you have to take enough numbers to actually generate an interesting visualization. John Williamson set about doing exactly this, and the results are fascinating. While they may not actually tell us anything new about number theory they do highlight interesting structures in prime factorizations, and demonstrate how UMAP can aid in interesting explorations of datasets that we might think we know well. It's worth visiting the linked article below as Dr. Williamson provides a rich and detiled exploration of UMAP as applied to prime factorizations of integers.

1 What do numbers look like?



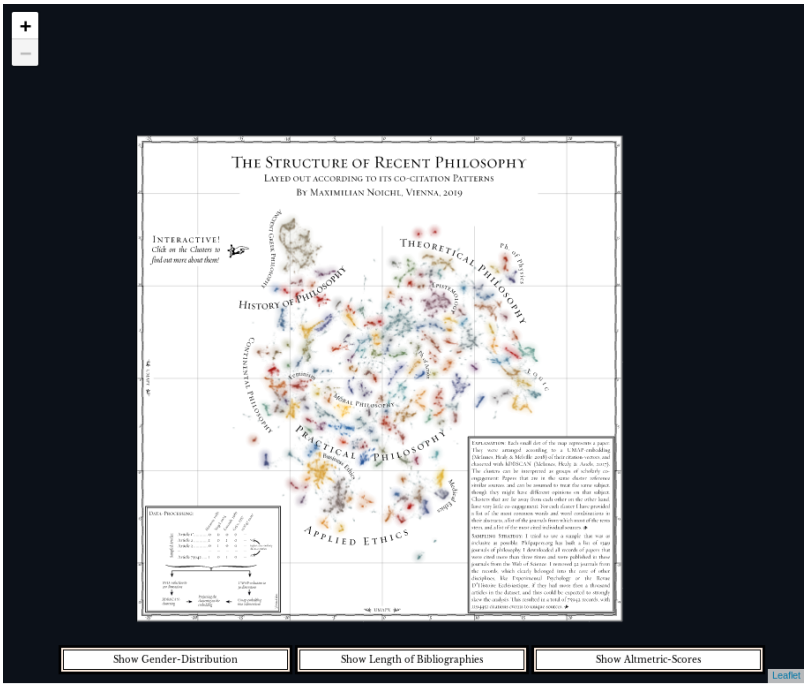
*One million integers embedded into 2D space with UMAP*

UMAP on prime factorizations

Thanks to John Williamson.

## 17.2 Structure of Recent Philosophy

Philosophy is an incredibly diverse subject, ranging from social and moral philosophy to logic and philosophy of math; from analysis of ancient Greek philosophy to modern business ethics. If we could get an overview of all the philosophy papers published in the last century what might it look like? Maximilian Noichl provides just such an exploration, looking at a large sampling of philosophy papers and comparing them according to their citations. The results are intriguing, and can be explored interactively in the viewer Maximilian built for it.

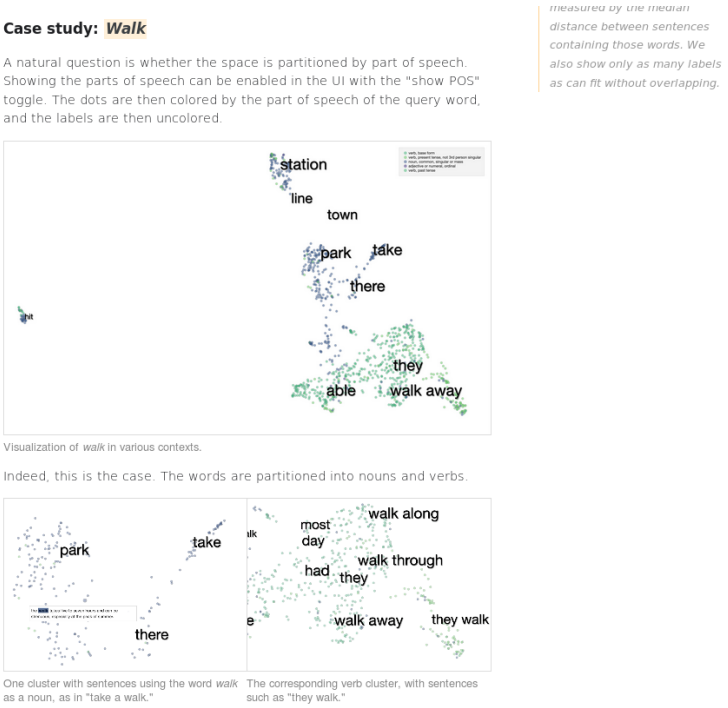



Structure of Recent Philosophy

Thanks to Maximilian Noichl

## 17.3 Language, Context, and Geometry in Neural Networks

Among recent developments in natural language processing is the BERT neural network based technique for analysis of language. Among many things that BERT can do one is context sensitive embeddings of words – providing numeric vector representations of words that are sentive to the context of how the word is used. Exactly what goes on inside the neural network to do this is a little mysterious (since the network is very complex with many many parameters). A tram of researchers from Google set out to explore the word embedding space generated by BERT, and among the tools used was UMAP. The linked blog post provides a detailed and inspirign analysis of what BERT's word embeddings look like, and how the different layers of BERT represent different aspects of language.




measured by the median distance between sentences containing those words. We also show only as many labels as can fit without overlapping.

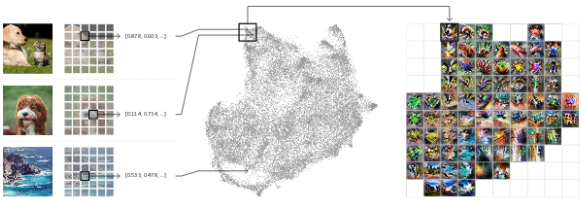
Language, Context, and Geometry in Neural Networks

Thanks to Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg.

## 17.4 Activation Atlas

Understanding the image processing capabilities (and deficits!) of modern convolutional neural networks is a challenge. Certainly these models are capable of amazing feats in, for example, image classification. They can also be brittle in unexpected ways, with carefully designed images able to induce otherwise baffling mis-classifications. To better understand this researchers from Google and OpenAI built the activation atlas – analysing the space of activations of a neural network. Here UMAP provides a means to compress the activation landscape down to 2 dimensions for visualization. The result was an impressive interactive paper in the Distill journal, providing rich visualizations and new insights into thw working of convolutional neural networks.

activation vectors, but we also need to aggregate into a more manageable number of elements — one million dots would be hard to interpret. We'll do this by drawing a grid over the 2D layout we created with dimensionality reduction. For each cell in our grid, we average all the activations that lie within the boundaries of that cell, and use feature visualization to create an iconic representation.



A randomized set of one million images is fed through the network, collecting one random spatial activation per image.

The activations are fed through UMAP to reduce them to two dimensions. They are then plotted, with similar activations placed near each other.

We then draw a grid and average the activations that fall within a cell and run feature inversion on the averaged activation. We also optionally size the grid cells according to the density of the number of activations that are averaged within.

TRY IN A NOTEBOOK

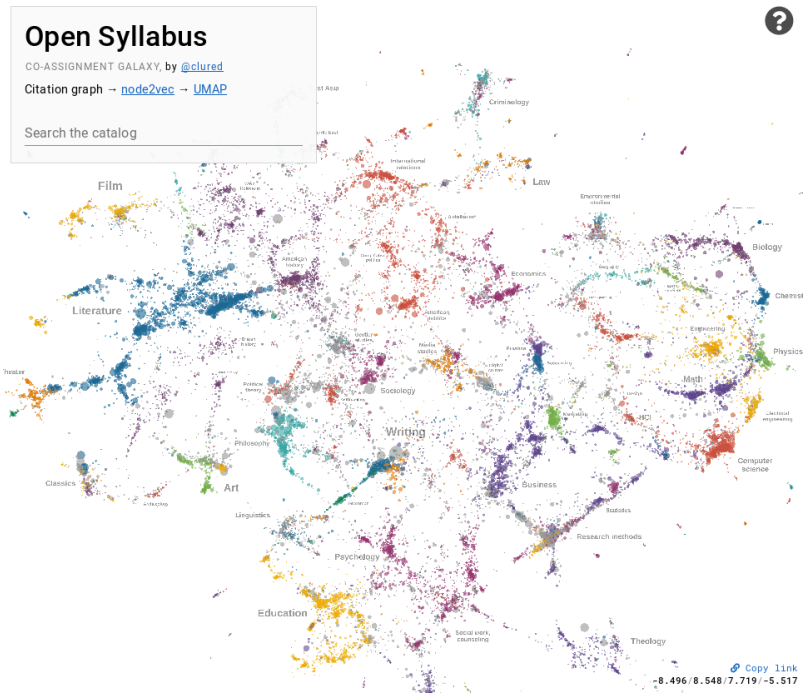
We perform feature visualization with the regularizations described in Feature Visualization [2] (in particular, transformation robustness). However, we use a slightly non-standard objective. Normally, to visualize a direction in activation space, $v$, one

### The Activation Atlas

Thanks to Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah

## 17.5 Open Syllabus Galaxy

Suppose you wanted to explore the space of commonly assigned texts from Open Syllabus? That gives you over 150,000 texts to consider. Since the texts are open you can actually analyse the text content involved. With some NLP and neural network wizardry David McClure build a network of such texts and then used node2vec and UMAP to generate a map of them. The result is a galaxy of textbooks showing inter-relationships between subjects, similar and related texts, and genrally just a an interesting ladscape of science to be explored. As with some of the other projects here David made a great interactive viewer allowing for rich exploration of the results.




Open Syllabus Galaxy

Thanks to David McClure.

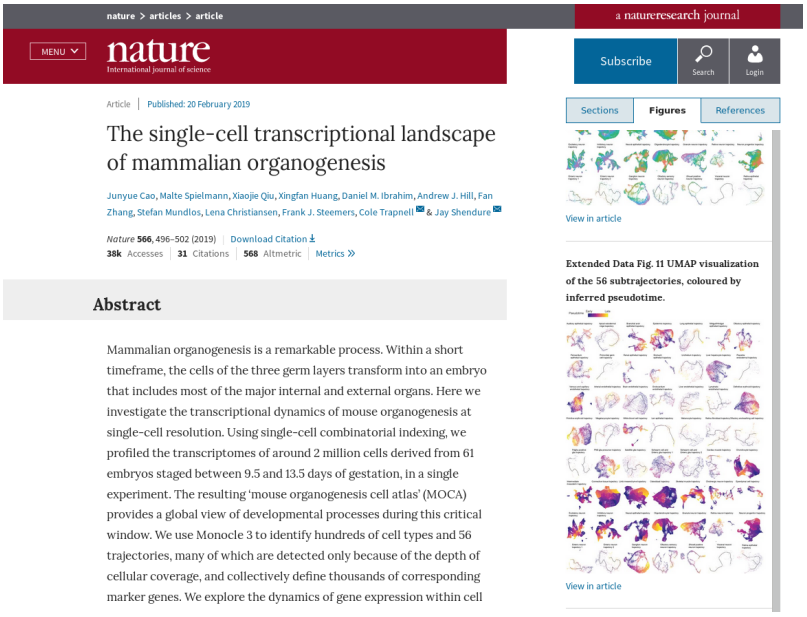
# CHAPTER 18

# Scientific Papers

UMAP has been used in a wide variety of scientific publications from a diverse range of fields. Here we will highlight a small selection of papers that demonstrate both the depth of analysis, and breadth of subjects, UMAP can be used for. These range from biology, to machine learning, and even social science.

## 18.1 The single-cell transcriptional landscape of mammalian organogenesis

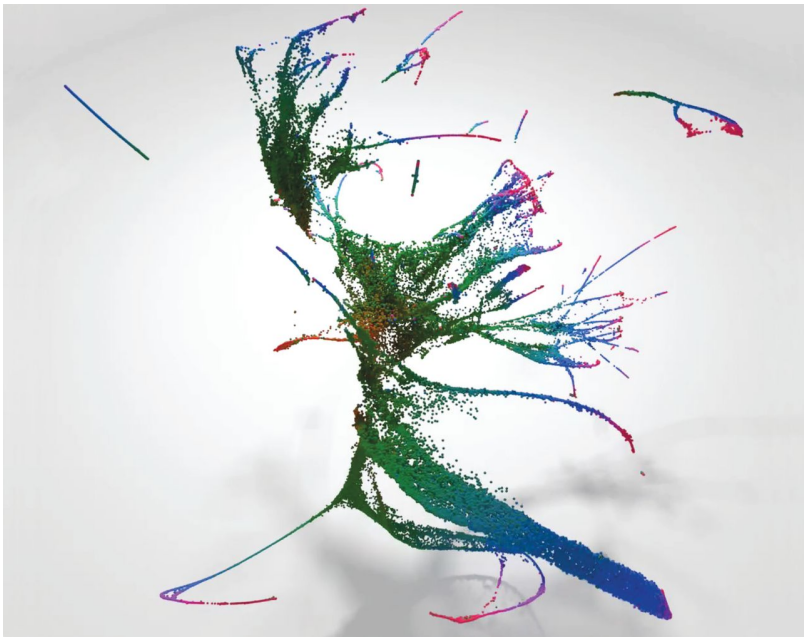
A detailed look at the development of mouse embryos from a single-cell view. UMAP is used as a core piece of The Monocle3 software suite for identifying cell types and trajectories. This was a major paper in Nature, demonstrating the power of UMAP for large scale scientific endeavours.



Link to the paper

## 18.2 A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution

Still in the realm of single cell biology this paper looks at the developmental landscape of the round-word C. elegans. UMAP is used for detailed analysis of the developmental trajectories of cells, looking at global scales, and then digging down to look at individual organs. The result is an impressive array of UMAP visualisations that tease out ever finer structures in cellular development.

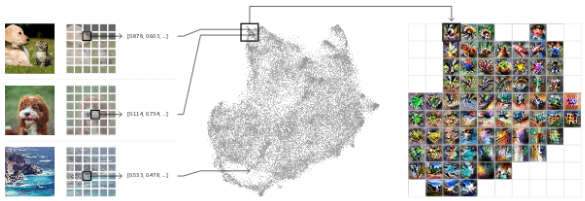


Link to the paper

## 18.3 Exploring Neural Networks with Activation Atlases

Understanding the image processing capabilities (and deficits!) of modern convolutional neural networks is a challenge. This interactive paper from Distill seeks to provide a way to "peek inside the black box" by looking at the activations throughout the network. By mapping this high dimensional data down to 2D with UMAP the authors can construct an "atlas" of how different images are perceived by the network.

activation vectors, but we also need to aggregate into a more manageable number of elements — one million dots would be hard to interpret. We'll do this by drawing a grid over the 2D layout we created with dimensionality reduction. For each cell in our grid, we average all the activations that lie within the boundaries of that cell, and use feature visualization to create an iconic representation.



A randomized set of one million images is fed through the network, collecting one random spatial activation per image.

The activations are fed through UMAP to reduce them to two dimensions. They are then plotted, with similar activations placed near each other.

We then draw a grid and average the activations that fall within a cell and run feature inversion on the averaged activation. We also optionally size the grid cells according to the density of the number of activations that are averaged within.

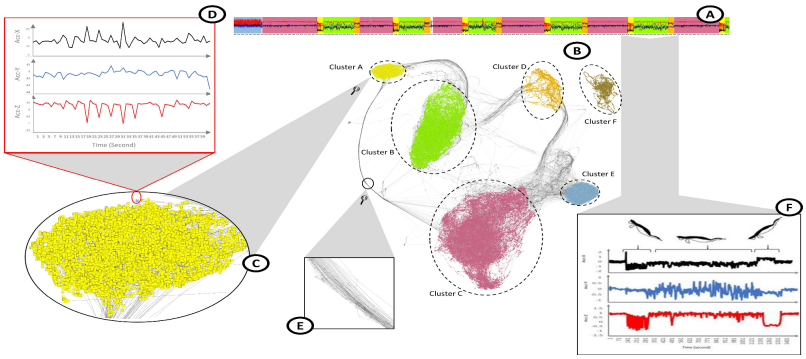
TRY IN A NOTEBOOK

We perform feature visualization with the regularizations described in Feature Visualization [2] (in particular, transformation robustness). However, we use a slightly non-standard objective. Normally, to visualize a direction in activation space, $v$, one

Link to the paper

## 18.4 TimeCluster: dimension reduction applied to temporal data for visual analytics

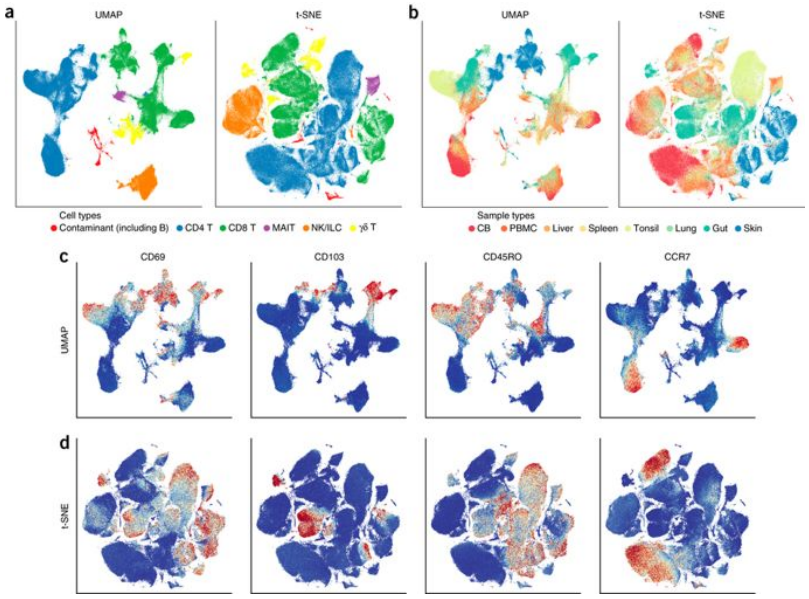
An interesting approach to time-series analysis, targeted toward cases where the time series has repeating patterns – though no necessarily of a consistently periodic nature. The approach involves dimension reduction and clustering of sliding window blocks of the time-series. The result is a map where repeating behaviour is exposed as loop structures. This can be useful for both clustering similar blocks within a time-series, or finding outliers.




Link to the paper

## 18.5 Dimensionality reduction for visualizing single-cell data using UMAP

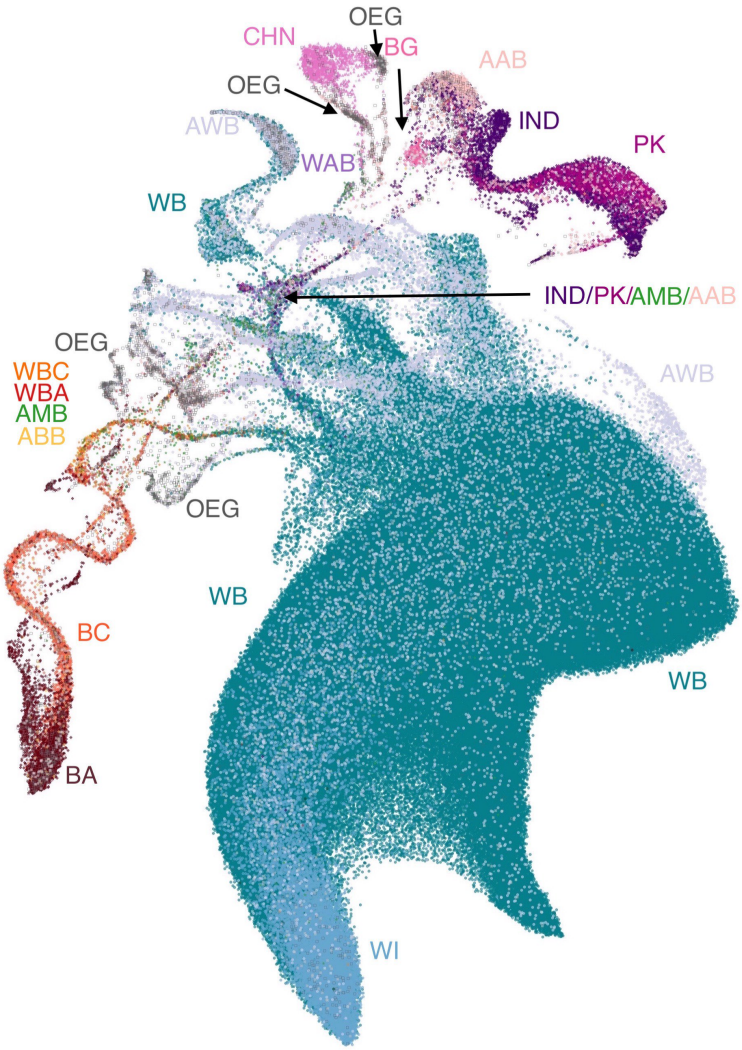
An early paper on applying UMAP to single-cell biology data. It looks at both gene-expression data and flow-cytometry data, and compares UMAP to t-SNE both in terms of performance and quality of results. This is a good introduction to using UMAP for single-cell biology data.




Link to the paper

## 18.6 Revealing multi-scale population structure in large cohorts

A paper looking at population genetics which uses UMAP as a means to visualise population structures. This produced some intriguing visualizations, and was one of the first of several papers taking this visualization approach. It also includes some novel visualizations using UMAP projections to 3D as RGB color specifications for data points, allowing the UMAP structure ot be visualized in geographic maps based on where the samples where drawn from.

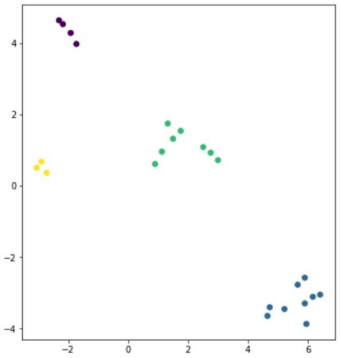



Link to the paper

## 18.7 Understanding Vulnerability of Children in Surrey

An example of the use of UMAP in sociological studies – in this case looking at children in Surrey, British Columbia. Here UMAP is used as a tool to aid in general data analysis, and proves effective for the tasks to which it was put.

## Validating Clustering results with UMAP




UMAP Clustering (Right) shows four distinct clusters on all-waves.

Hopkins Statistic (Below) to reject the null hypothesis that these clusters reasonably random.

| | t-SNE A-clusters | | | | | |
|---|---|---|---|---|---|---|
| Cluster | 0 | 1 | **2** | 3 | 4 | **5** |
| $H$ | 0.4563 | 0.5478 | **0.5706** | 0.4166 | 0.6080 | **0.4311** |

Table 2: Hopkin's statistic over the t-SNE all-wave clusters.

| | UMAP UA-clusters | | | |
|---|---|---|---|---|
| Cluster | **0** | 1 | 2 | **3** |
| $H$ | **0.5706** | 0.5023 | 0.5308 | **0.4311** |

Table 3: Hopkin's statistic over the UMAP all-wave clusters.

Link to the paper

# CHAPTER 19

# UMAP API Guide

UMAP has only a single class `UMAP`.

## 19.1 UMAP

**class** umap.umap_.**UMAP**(*n_neighbors=15*, *n_components=2*, *metric='euclidean'*, *metric_kwds=None*, *output_metric='euclidean'*, *output_metric_kwds=None*, *n_epochs=None*, *learning_rate=1.0*, *init='spectral'*, *min_dist=0.1*, *spread=1.0*, *low_memory=False*, *set_op_mix_ratio=1.0*, *local_connectivity=1.0*, *repulsion_strength=1.0*, *negative_sample_rate=5*, *transform_queue_size=4.0*, *a=None*, *b=None*, *random_state=None*, *angular_rp_forest=False*, *target_n_neighbors=-1*, *target_metric='categorical'*, *target_metric_kwds=None*, *target_weight=0.5*, *transform_seed=42*, *force_approximation_algorithm=False*, *verbose=False*, *unique=False*)

Uniform Manifold Approximation and Projection

Finds a low dimensional embedding of the data that approximates an underlying manifold.

**n_neighbors: float (optional, default 15)** The size of local neighborhood (in terms of number of neighboring sample points) used for manifold approximation. Larger values result in more global views of the manifold, while smaller values result in more local data being preserved. In general values should be in the range 2 to 100.

**n_components: int (optional, default 2)** The dimension of the space to embed into. This defaults to 2 to provide easy visualization, but can reasonably be set to any integer value in the range 2 to 100.

**metric: string or function (optional, default 'euclidean')** The metric to use to compute distances in high dimensional space. If a string is passed it must match a valid predefined metric. If a general metric is required a function that takes two 1d arrays and returns a float can be provided. For performance purposes it is required that this be a numba jit'd function. Valid string metrics include:

- euclidean
- manhattan

- chebyshev
- minkowski
- canberra
- braycurtis
- mahalanobis
- wminkowski
- seuclidean
- cosine
- correlation
- haversine
- hamming
- jaccard
- dice
- russelrao
- kulsinski
- ll_dirichlet
- hellinger
- rogerstanimoto
- sokalmichener
- sokalsneath
- yule

Metrics that take arguments (such as minkowski, mahalanobis etc.) can have arguments passed via the metric_kwds dictionary. At this time care must be taken and dictionary elements must be ordered appropriately; this will hopefully be fixed in the future.

**n_epochs: int (optional, default None)** The number of training epochs to be used in optimizing the low dimensional embedding. Larger values result in more accurate embeddings. If None is specified a value will be selected based on the size of the input dataset (200 for large datasets, 500 for small).

**learning_rate: float (optional, default 1.0)** The initial learning rate for the embedding optimization.

**init: string (optional, default 'spectral')**

**How to initialize the low dimensional embedding. Options are:**

- 'spectral': use a spectral embedding of the fuzzy 1-skeleton
- 'random': assign initial embedding positions at random.
- A numpy array of initial embedding positions.

**min_dist: float (optional, default 0.1)** The effective minimum distance between embedded points. Smaller values will result in a more clustered/clumped embedding where nearby points on the manifold are drawn closer together, while larger values will result on a more even dispersal of points. The value should be set relative to the `spread` value, which determines the scale at which embedded points will be spread out.

**spread: float (optional, default 1.0)** The effective scale of embedded points. In combination with `min_dist` this determines how clustered/clumped the embedded points are.

**low_memory: bool (optional, default False)** For some datasets the nearest neighbor computation can consume a lot of memory. If you find that UMAP is failing due to memory constraints consider setting this option to True. This approach is more computationally expensive, but avoids excessive memory use.

**set_op_mix_ratio: float (optional, default 1.0)** Interpolate between (fuzzy) union and intersection as the set operation used to combine local fuzzy simplicial sets to obtain a global fuzzy simplicial sets. Both fuzzy set operations use the product t-norm. The value of this parameter should be between 0.0 and 1.0; a value of 1.0 will use a pure fuzzy union, while 0.0 will use a pure fuzzy intersection.

**local_connectivity: int (optional, default 1)** The local connectivity required – i.e. the number of nearest neighbors that should be assumed to be connected at a local level. The higher this value the more connected the manifold becomes locally. In practice this should be not more than the local intrinsic dimension of the manifold.

**repulsion_strength: float (optional, default 1.0)** Weighting applied to negative samples in low dimensional embedding optimization. Values higher than one will result in greater weight being given to negative samples.

**negative_sample_rate: int (optional, default 5)** The number of negative samples to select per positive sample in the optimization process. Increasing this value will result in greater repulsive force being applied, greater optimization cost, but slightly more accuracy.

**transform_queue_size: float (optional, default 4.0)** For transform operations (embedding new points using a trained **model_** this will control how aggressively to search for nearest neighbors. Larger values will result in slower performance but more accurate nearest neighbor evaluation.

**a: float (optional, default None)** More specific parameters controlling the embedding. If None these values are set automatically as determined by `min_dist` and `spread`.

**b: float (optional, default None)** More specific parameters controlling the embedding. If None these values are set automatically as determined by `min_dist` and `spread`.

**random_state: int, RandomState instance or None, optional (default: None)** If int, random_state is the seed used by the random number generator; If RandomState instance, random_state is the random number generator; If None, the random number generator is the RandomState instance used by *np.random*.

**metric_kwds: dict (optional, default None)** Arguments to pass on to the metric, such as the `p` value for Minkowski distance. If None then no arguments are passed on.

**angular_rp_forest: bool (optional, default False)** Whether to use an angular random projection forest to initialise the approximate nearest neighbor search. This can be faster, but is mostly on useful for metric that use an angular style distance such as cosine, correlation etc. In the case of those metrics angular forests will be chosen automatically.

**target_n_neighbors: int (optional, default -1)** The number of nearest neighbors to use to construct the target simplcial set. If set to -1 use the `n_neighbors` value.

**target_metric: string or callable (optional, default 'categorical')** The metric used to measure distance for a target array is using supervised dimension reduction. By default this is 'categorical' which will measure distance in terms of whether categories match or are different. Furthermore, if semi-supervised is required target values of -1 will be trated as unlabelled under the 'categorical' metric. If the target array takes continuous values (e.g. for a regression problem) then metric of 'l1' or 'l2' is probably more appropriate.

**target_metric_kwds: dict (optional, default None)** Keyword argument to pass to the target metric when performing supervised dimension reduction. If None then no arguments are passed on.

**target_weight: float (optional, default 0.5)** weighting factor between data topology and target topology. A value of 0.0 weights entirely on data, a value of 1.0 weights entirely on target. The default of 0.5 balances the weighting equally between data and target.

**transform_seed: int (optional, default 42)** Random seed used for the stochastic aspects of the transform operation. This ensures consistency in transform operations.

**verbose: bool (optional, default False)** Controls verbosity of logging.

**unique: bool (optional, default False)** Controls if the rows of your data should be uniqued before being embedded. If you have more duplicates than you have n_neighbour you can have the identical data points lying in different regions of your space. It also violates the definition of a metric.

**fit** (*X*, *y=None*)
Fit X into an embedded space.

Optionally use y for supervised dimension reduction.

**X** [array, shape (n_samples, n_features) or (n_samples, n_samples)] If the metric is 'precomputed' X must be a square distance matrix. Otherwise it contains a sample per row. If the method is 'exact', X may be a sparse matrix of type 'csr', 'csc' or 'coo'.

**y** [array, shape (n_samples)] A target array for supervised dimension reduction. How this is handled is determined by parameters UMAP was instantiated with. The relevant attributes are `target_metric` and `target_metric_kwds`.

**fit_transform** (*X*, *y=None*)
Fit X into an embedded space and return that transformed output.

**X** [array, shape (n_samples, n_features) or (n_samples, n_samples)] If the metric is 'precomputed' X must be a square distance matrix. Otherwise it contains a sample per row.

**y** [array, shape (n_samples)] A target array for supervised dimension reduction. How this is handled is determined by parameters UMAP was instantiated with. The relevant attributes are `target_metric` and `target_metric_kwds`.

**X_new** [array, shape (n_samples, n_components)] Embedding of the training data in low-dimensional space.

**inverse_transform** (*X*)
Transform X in the existing embedded space back into the input data space and return that transformed output.

**X** [array, shape (n_samples, n_components)] New points to be inverse transformed.

**X_new** [array, shape (n_samples, n_features)] Generated data points new data in data space.

**transform** (*X*)
Transform X into the existing embedded space and return that transformed output.

**X** [array, shape (n_samples, n_features)] New data to be transformed.

**X_new** [array, shape (n_samples, n_components)] Embedding of the new data in low-dimensional space.

A number of internal functions can also be accessed separately for more fine tuned work.

## 19.2 Useful Functions

**class** umap.umap_.**DataFrameUMAP**(*metrics*, *n_neighbors=15*, *n_components=2*, *output_metric='euclidean'*, *output_metric_kwds=None*, *n_epochs=None*, *learning_rate=1.0*, *init='spectral'*, *min_dist=0.1*, *spread=1.0*, *set_op_mix_ratio=1.0*, *local_connectivity=1.0*, *repulsion_strength=1.0*, *negative_sample_rate=5*, *transform_queue_size=4.0*, *a=None*, *b=None*, *random_state=None*, *angular_rp_forest=False*, *target_n_neighbors=-1*, *target_metric='categorical'*, *target_metric_kwds=None*, *target_weight=0.5*, *transform_seed=42*, *verbose=False*)

**class** umap.umap_.**UMAP**(*n_neighbors=15*, *n_components=2*, *metric='euclidean'*, *metric_kwds=None*, *output_metric='euclidean'*, *output_metric_kwds=None*, *n_epochs=None*, *learning_rate=1.0*, *init='spectral'*, *min_dist=0.1*, *spread=1.0*, *low_memory=False*, *set_op_mix_ratio=1.0*, *local_connectivity=1.0*, *repulsion_strength=1.0*, *negative_sample_rate=5*, *transform_queue_size=4.0*, *a=None*, *b=None*, *random_state=None*, *angular_rp_forest=False*, *target_n_neighbors=-1*, *target_metric='categorical'*, *target_metric_kwds=None*, *target_weight=0.5*, *transform_seed=42*, *force_approximation_algorithm=False*, *verbose=False*, *unique=False*)

Uniform Manifold Approximation and Projection

Finds a low dimensional embedding of the data that approximates an underlying manifold.

**n_neighbors: float (optional, default 15)** The size of local neighborhood (in terms of number of neighboring sample points) used for manifold approximation. Larger values result in more global views of the manifold, while smaller values result in more local data being preserved. In general values should be in the range 2 to 100.

**n_components: int (optional, default 2)** The dimension of the space to embed into. This defaults to 2 to provide easy visualization, but can reasonably be set to any integer value in the range 2 to 100.

**metric: string or function (optional, default 'euclidean')** The metric to use to compute distances in high dimensional space. If a string is passed it must match a valid predefined metric. If a general metric is required a function that takes two 1d arrays and returns a float can be provided. For performance purposes it is required that this be a numba jit'd function. Valid string metrics include:

- euclidean
- manhattan
- chebyshev
- minkowski
- canberra
- braycurtis
- mahalanobis
- wminkowski
- seuclidean
- cosine
- correlation
- haversine

- hamming
- jaccard
- dice
- russelrao
- kulsinski
- ll_dirichlet
- hellinger
- rogerstanimoto
- sokalmichener
- sokalsneath
- yule

Metrics that take arguments (such as minkowski, mahalanobis etc.) can have arguments passed via the metric_kwds dictionary. At this time care must be taken and dictionary elements must be ordered appropriately; this will hopefully be fixed in the future.

**n_epochs: int (optional, default None)** The number of training epochs to be used in optimizing the low dimensional embedding. Larger values result in more accurate embeddings. If None is specified a value will be selected based on the size of the input dataset (200 for large datasets, 500 for small).

**learning_rate: float (optional, default 1.0)** The initial learning rate for the embedding optimization.

**init: string (optional, default 'spectral')**

**How to initialize the low dimensional embedding. Options are:**

- 'spectral': use a spectral embedding of the fuzzy 1-skeleton
- 'random': assign initial embedding positions at random.
- A numpy array of initial embedding positions.

**min_dist: float (optional, default 0.1)** The effective minimum distance between embedded points. Smaller values will result in a more clustered/clumped embedding where nearby points on the manifold are drawn closer together, while larger values will result on a more even dispersal of points. The value should be set relative to the `spread` value, which determines the scale at which embedded points will be spread out.

**spread: float (optional, default 1.0)** The effective scale of embedded points. In combination with `min_dist` this determines how clustered/clumped the embedded points are.

**low_memory: bool (optional, default False)** For some datasets the nearest neighbor computation can consume a lot of memory. If you find that UMAP is failing due to memory constraints consider setting this option to True. This approach is more computationally expensive, but avoids excessive memory use.

**set_op_mix_ratio: float (optional, default 1.0)** Interpolate between (fuzzy) union and intersection as the set operation used to combine local fuzzy simplicial sets to obtain a global fuzzy simplicial sets. Both fuzzy set operations use the product t-norm. The value of this parameter should be between 0.0 and 1.0; a value of 1.0 will use a pure fuzzy union, while 0.0 will use a pure fuzzy intersection.

**local_connectivity: int (optional, default 1)** The local connectivity required – i.e. the number of nearest neighbors that should be assumed to be connected at a local level. The higher this value the more connected the manifold becomes locally. In practice this should be not more than the local intrinsic dimension of the manifold.

**repulsion_strength: float (optional, default 1.0)** Weighting applied to negative samples in low dimensional embedding optimization. Values higher than one will result in greater weight being given to negative samples.

**negative_sample_rate: int (optional, default 5)** The number of negative samples to select per positive sample in the optimization process. Increasing this value will result in greater repulsive force being applied, greater optimization cost, but slightly more accuracy.

**transform_queue_size: float (optional, default 4.0)** For transform operations (embedding new points using a trained **model_** this will control how aggressively to search for nearest neighbors. Larger values will result in slower performance but more accurate nearest neighbor evaluation.

**a: float (optional, default None)** More specific parameters controlling the embedding. If None these values are set automatically as determined by `min_dist` and `spread`.

**b: float (optional, default None)** More specific parameters controlling the embedding. If None these values are set automatically as determined by `min_dist` and `spread`.

**random_state: int, RandomState instance or None, optional (default: None)** If int, random_state is the seed used by the random number generator; If RandomState instance, random_state is the random number generator; If None, the random number generator is the RandomState instance used by *np.random*.

**metric_kwds: dict (optional, default None)** Arguments to pass on to the metric, such as the `p` value for Minkowski distance. If None then no arguments are passed on.

**angular_rp_forest: bool (optional, default False)** Whether to use an angular random projection forest to initialise the approximate nearest neighbor search. This can be faster, but is mostly on useful for metric that use an angular style distance such as cosine, correlation etc. In the case of those metrics angular forests will be chosen automatically.

**target_n_neighbors: int (optional, default -1)** The number of nearest neighbors to use to construct the target simplcial set. If set to -1 use the `n_neighbors` value.

**target_metric: string or callable (optional, default 'categorical')** The metric used to measure distance for a target array is using supervised dimension reduction. By default this is 'categorical' which will measure distance in terms of whether categories match or are different. Furthermore, if semi-supervised is required target values of -1 will be trated as unlabelled under the 'categorical' metric. If the target array takes continuous values (e.g. for a regression problem) then metric of 'l1' or 'l2' is probably more appropriate.

**target_metric_kwds: dict (optional, default None)** Keyword argument to pass to the target metric when performing supervised dimension reduction. If None then no arguments are passed on.

**target_weight: float (optional, default 0.5)** weighting factor between data topology and target topology. A value of 0.0 weights entirely on data, a value of 1.0 weights entirely on target. The default of 0.5 balances the weighting equally between data and target.

**transform_seed: int (optional, default 42)** Random seed used for the stochastic aspects of the transform operation. This ensures consistency in transform operations.

**verbose: bool (optional, default False)** Controls verbosity of logging.

**unique: bool (optional, default False)** Controls if the rows of your data should be uniqued before being embedded. If you have more duplicates than you have n_neighbour you can have the identical data points lying in different regions of your space. It also violates the definition of a metric.

**`fit`** (*X*, *y=None*)
Fit X into an embedded space.

Optionally use y for supervised dimension reduction.

**X** [array, shape (n_samples, n_features) or (n_samples, n_samples)] If the metric is 'precomputed' X must be a square distance matrix. Otherwise it contains a sample per row. If the method is 'exact', X may be a sparse matrix of type 'csr', 'csc' or 'coo'.

**y** [array, shape (n_samples)] A target array for supervised dimension reduction. How this is handled is determined by parameters UMAP was instantiated with. The relevant attributes are `target_metric` and `target_metric_kwds`.

**fit_transform**(*X*, *y=None*)

Fit X into an embedded space and return that transformed output.

**X** [array, shape (n_samples, n_features) or (n_samples, n_samples)] If the metric is 'precomputed' X must be a square distance matrix. Otherwise it contains a sample per row.

**y** [array, shape (n_samples)] A target array for supervised dimension reduction. How this is handled is determined by parameters UMAP was instantiated with. The relevant attributes are `target_metric` and `target_metric_kwds`.

**X_new** [array, shape (n_samples, n_components)] Embedding of the training data in low-dimensional space.

**inverse_transform**(*X*)

Transform X in the existing embedded space back into the input data space and return that transformed output.

**X** [array, shape (n_samples, n_components)] New points to be inverse transformed.

**X_new** [array, shape (n_samples, n_features)] Generated data points new data in data space.

**transform**(*X*)

Transform X into the existing embedded space and return that transformed output.

**X** [array, shape (n_samples, n_features)] New data to be transformed.

**X_new** [array, shape (n_samples, n_components)] Embedding of the new data in low-dimensional space.

umap.umap_.**compute_membership_strengths**

Construct the membership strength data for the 1-skeleton of each local fuzzy simplicial set – this is formed as a sparse matrix where each row is a local fuzzy simplicial set, with a membership strength for the 1-simplex to each other data point.

**knn_indices: array of shape (n_samples, n_neighbors)** The indices on the `n_neighbors` closest points in the dataset.

**knn_dists: array of shape (n_samples, n_neighbors)** The distances to the `n_neighbors` closest points in the dataset.

**sigmas: array of shape(n_samples)** The normalization factor derived from the metric tensor approximation.

**rhos: array of shape(n_samples)** The local connectivity adjustment.

**rows: array of shape (n_samples * n_neighbors)** Row data for the resulting sparse matrix (coo format)

**cols: array of shape (n_samples * n_neighbors)** Column data for the resulting sparse matrix (coo format)

**vals: array of shape (n_samples * n_neighbors)** Entries for the resulting sparse matrix (coo format)

umap.umap_.**discrete_metric_simplicial_set_intersection**(*simplicial_set*, *discrete_space*, *unknown_dist=1.0*, *far_dist=5.0*, *metric=None*, *metric_kws={}*, *metric_scale=1.0*)

Combine a fuzzy simplicial set with another fuzzy simplicial set generated from discrete metric data using discrete distances. The target data is assumed to be categorical label data (a vector of labels), and this will update the fuzzy simplicial set to respect that label data.

TODO: optional category cardinality based weighting of distance

**simplicial_set: sparse matrix** The input fuzzy simplicial set.

**discrete_space: array of shape (n_samples)** The categorical labels to use in the intersection.

**unknown_dist: float (optional, default 1.0)** The distance an unknown label (-1) is assumed to be from any point.

**far_dist: float (optional, default 5.0)** The distance between unmatched labels.

**metric: str (optional, default None)** If not None, then use this metric to determine the distance between values.

**metric_scale: float (optional, default 1.0)** If using a custom metric scale the distance values by this value – this controls the weighting of the intersection. Larger values weight more toward target.

**simplicial_set: sparse matrix** The resulting intersected fuzzy simplicial set.

umap.umap_.**fast_intersection**

Under the assumption of categorical distance for the intersecting simplicial set perform a fast intersection.

**rows: array** An array of the row of each non-zero in the sparse matrix representation.

**cols: array** An array of the column of each non-zero in the sparse matrix representation.

**values: array** An array of the value of each non-zero in the sparse matrix representation.

**target: array of shape (n_samples)** The categorical labels to use in the intersection.

**unknown_dist: float (optional, default 1.0)** The distance an unknown label (-1) is assumed to be from any point.

**far_dist float (optional, default 5.0)** The distance between unmatched labels.

None

umap.umap_.**fast_metric_intersection**

Under the assumption of categorical distance for the intersecting simplicial set perform a fast intersection.

**rows: array** An array of the row of each non-zero in the sparse matrix representation.

**cols: array** An array of the column of each non-zero in the sparse matrix representation.

**values: array of shape** An array of the values of each non-zero in the sparse matrix representation.

**discrete_space: array of shape (n_samples, n_features)** The vectors of categorical labels to use in the intersection.

**metric: numba function** The function used to calculate distance over the target array.

**scale: float** A scaling to apply to the metric.

None

umap.umap_.**find_ab_params**(*spread*, *min_dist*)

Fit a, b params for the differentiable curve used in lower dimensional fuzzy simplicial complex construction. We want the smooth curve (from a pre-defined family with simple gradient) that best matches an offset exponential decay.

umap.umap_.**fuzzy_simplicial_set**(*X*, *n_neighbors*, *random_state*, *metric*, *metric_kwds={}*, *knn_indices=None*, *knn_dists=None*, *angular=False*, *set_op_mix_ratio=1.0*, *local_connectivity=1.0*, *apply_set_operations=True*, *verbose=False*)

Given a set of data X, a neighborhood size, and a measure of distance compute the fuzzy simplicial set (here represented as a fuzzy graph in the form of a sparse matrix) associated to the data. This is done by locally approximating geodesic distance at each point, creating a fuzzy simplicial set for each such point, and then combining all the local fuzzy simplicial sets into a global one via a fuzzy union.

**X: array of shape (n_samples, n_features)** The data to be modelled as a fuzzy simplicial set.

**n_neighbors: int** The number of neighbors to use to approximate geodesic distance. Larger numbers induce more global estimates of the manifold that can miss finer detail, while smaller values will focus on fine manifold structure to the detriment of the larger picture.

**random_state: numpy RandomState or equivalent** A state capable being used as a numpy random state.

**metric: string or function (optional, default 'euclidean')** The metric to use to compute distances in high dimensional space. If a string is passed it must match a valid predefined metric. If a general metric is required a function that takes two 1d arrays and returns a float can be provided. For performance purposes it is required that this be a numba jit'd function. Valid string metrics include:

- euclidean (or l2)
- manhattan (or l1)
- cityblock
- braycurtis
- canberra
- chebyshev
- correlation
- cosine
- dice
- hamming
- jaccard
- kulsinski
- ll_dirichlet
- mahalanobis
- matching
- minkowski
- rogerstanimoto
- russellrao
- seuclidean
- sokalmichener

- sokalsneath
- sqeuclidean
- yule
- wminkowski

Metrics that take arguments (such as minkowski, mahalanobis etc.) can have arguments passed via the metric_kwds dictionary. At this time care must be taken and dictionary elements must be ordered appropriately; this will hopefully be fixed in the future.

**metric_kwds: dict (optional, default {})** Arguments to pass on to the metric, such as the `p` value for Minkowski distance.

**knn_indices: array of shape (n_samples, n_neighbors) (optional)** If the k-nearest neighbors of each point has already been calculated you can pass them in here to save computation time. This should be an array with the indices of the k-nearest neighbors as a row for each data point.

**knn_dists: array of shape (n_samples, n_neighbors) (optional)** If the k-nearest neighbors of each point has already been calculated you can pass them in here to save computation time. This should be an array with the distances of the k-nearest neighbors as a row for each data point.

**angular: bool (optional, default False)** Whether to use angular/cosine distance for the random projection forest for seeding NN-descent to determine approximate nearest neighbors.

**set_op_mix_ratio: float (optional, default 1.0)** Interpolate between (fuzzy) union and intersection as the set operation used to combine local fuzzy simplicial sets to obtain a global fuzzy simplicial sets. Both fuzzy set operations use the product t-norm. The value of this parameter should be between 0.0 and 1.0; a value of 1.0 will use a pure fuzzy union, while 0.0 will use a pure fuzzy intersection.

**local_connectivity: int (optional, default 1)** The local connectivity required – i.e. the number of nearest neighbors that should be assumed to be connected at a local level. The higher this value the more connected the manifold becomes locally. In practice this should be not more than the local intrinsic dimension of the manifold.

**verbose: bool (optional, default False)** Whether to report information on the current progress of the algorithm.

**fuzzy_simplicial_set: coo_matrix** A fuzzy simplicial set represented as a sparse matrix. The (i, j) entry of the matrix represents the membership strength of the 1-simplex between the ith and jth sample points.

umap.umap_.**init_transform**

Given indices and weights and an original embeddings initialize the positions of new points relative to the indices and weights (of their neighbors in the source data).

**indices: array of shape (n_new_samples, n_neighbors)** The indices of the neighbors of each new sample

**weights: array of shape (n_new_samples, n_neighbors)** The membership strengths of associated 1-simplices for each of the new samples.

**embedding: array of shape (n_samples, dim)** The original embedding of the source data.

**new_embedding: array of shape (n_new_samples, dim)** An initial embedding of the new sample points.

umap.umap_.**make_epochs_per_sample**(*weights*, *n_epochs*)

Given a set of weights and number of epochs generate the number of epochs per sample for each weight.

**weights: array of shape (n_1_simplices)** The weights ofhow much we wish to sample each 1-simplex.

**n_epochs: int** The total number of epochs we want to train for.

An array of number of epochs per sample, one for each 1-simplex.

umap.umap_.**nearest_neighbors**(*X*, *n_neighbors*, *metric*, *metric_kwds*, *angular*, *random_state*, *low_memory=False*, *use_pynndescent=True*, *verbose=False*)

Compute the n_neighbors nearest points for each data point in X under metric. This may be exact, but more likely is approximated via nearest neighbor descent.

**X: array of shape (n_samples, n_features)** The input data to compute the k-neighbor graph of.

**n_neighbors: int** The number of nearest neighbors to compute for each sample in X.

**metric: string or callable** The metric to use for the computation.

**metric_kwds: dict** Any arguments to pass to the metric computation function.

**angular: bool** Whether to use angular rp trees in NN approximation.

**random_state: np.random state** The random state to use for approximate NN computations.

**low_memory: bool (optional, default False)** Whether to pursue lower memory NNdescent.

**verbose: bool (optional, default False)** Whether to print status data during the computation.

**knn_indices: array of shape (n_samples, n_neighbors)** The indices on the n_neighbors closest points in the dataset.

**knn_dists: array of shape (n_samples, n_neighbors)** The distances to the n_neighbors closest points in the dataset.

**rp_forest: list of trees** The random projection forest used for searching (if used, None otherwise)

umap.umap_.**reset_local_connectivity**(*simplicial_set*, *reset_local_metric=False*)

Reset the local connectivity requirement – each data sample should have complete confidence in at least one 1-simplex in the simplicial set. We can enforce this by locally rescaling confidences, and then remerging the different local simplicial sets together.

**simplicial_set: sparse matrix** The simplicial set for which to recalculate with respect to local connectivity.

**simplicial_set: sparse_matrix** The recalculated simplicial set, now with the local connectivity assumption restored.

umap.umap_.**simplicial_set_embedding**(*data*, *graph*, *n_components*, *initial_alpha*, *a*, *b*, *gamma*, *negative_sample_rate*, *n_epochs*, *init*, *random_state*, *metric*, *metric_kwds*, *output_metric=CPUDispatcher(<function euclidean_grad>)*, *output_metric_kwds={}*, *euclidean_output=True*, *parallel=False*, *verbose=False*)

Perform a fuzzy simplicial set embedding, using a specified initialisation method and then minimizing the fuzzy set cross entropy between the 1-skeletons of the high and low dimensional fuzzy simplicial sets.

**data: array of shape (n_samples, n_features)** The source data to be embedded by UMAP.

**graph: sparse matrix** The 1-skeleton of the high dimensional fuzzy simplicial set as represented by a graph for which we require a sparse matrix for the (weighted) adjacency matrix.

**n_components: int** The dimensionality of the euclidean space into which to embed the data.

**initial_alpha: float** Initial learning rate for the SGD.

**a: float** Parameter of differentiable approximation of right adjoint functor

**b: float** Parameter of differentiable approximation of right adjoint functor

**gamma: float** Weight to apply to negative samples.

**negative_sample_rate: int (optional, default 5)** The number of negative samples to select per positive sample in the optimization process. Increasing this value will result in greater repulsive force being applied, greater optimization cost, but slightly more accuracy.

**n_epochs: int (optional, default 0)** The number of training epochs to be used in optimizing the low dimensional embedding. Larger values result in more accurate embeddings. If 0 is specified a value will be selected based on the size of the input dataset (200 for large datasets, 500 for small).

**init: string**

**How to initialize the low dimensional embedding. Options are:**

- 'spectral': use a spectral embedding of the fuzzy 1-skeleton
- 'random': assign initial embedding positions at random.
- A numpy array of initial embedding positions.

**random_state: numpy RandomState or equivalent** A state capable being used as a numpy random state.

**metric: string or callable** The metric used to measure distance in high dimensional space; used if multiple connected components need to be layed out.

**metric_kwds: dict** Key word arguments to be passed to the metric function; used if multiple connected components need to be layed out.

**output_metric: function** Function returning the distance between two points in embedding space and the gradient of the distance wrt the first argument.

**output_metric_kwds: dict** Key word arguments to be passed to the output_metric function.

**euclidean_output: bool** Whether to use the faster code specialised for euclidean output metrics

**parallel: bool (optional, default False)** Whether to run the computation using numba parallel. Running in parallel is non-deterministic, and is not used if a random seed has been set, to ensure reproducibility.

**verbose: bool (optional, default False)** Whether to report information on the current progress of the algorithm.

**embedding: array of shape (n_samples, n_components)** The optimized of `graph` into an `n_components` dimensional euclidean space.

`umap.umap_.`**`smooth_knn_dist`**

Compute a continuous version of the distance to the kth nearest neighbor. That is, this is similar to knn-distance but allows continuous k values rather than requiring an integral k. In essence we are simply computing the distance such that the cardinality of fuzzy set we generate is k.

**distances: array of shape (n_samples, n_neighbors)** Distances to nearest neighbors for each samples. Each row should be a sorted list of distances to a given samples nearest neighbors.

**k: float** The number of nearest neighbors to approximate for.

**n_iter: int (optional, default 64)** We need to binary search for the correct distance value. This is the max number of iterations to use in such a search.

**local_connectivity: int (optional, default 1)** The local connectivity required – i.e. the number of nearest neighbors that should be assumed to be connected at a local level. The higher this value the more connected the manifold becomes locally. In practice this should be not more than the local intrinsic dimension of the manifold.

**bandwidth: float (optional, default 1)** The target bandwidth of the kernel, larger values will produce larger return values.

**knn_dist: array of shape (n_samples,)** The distance to kth nearest neighbor, as suitably approximated.

**nn_dist: array of shape (n_samples,)** The distance to the 1st nearest neighbor for each point.

CHAPTER 20

# Indices and tables

- genindex
- modindex
- search

# Python Module Index

## u

# Index