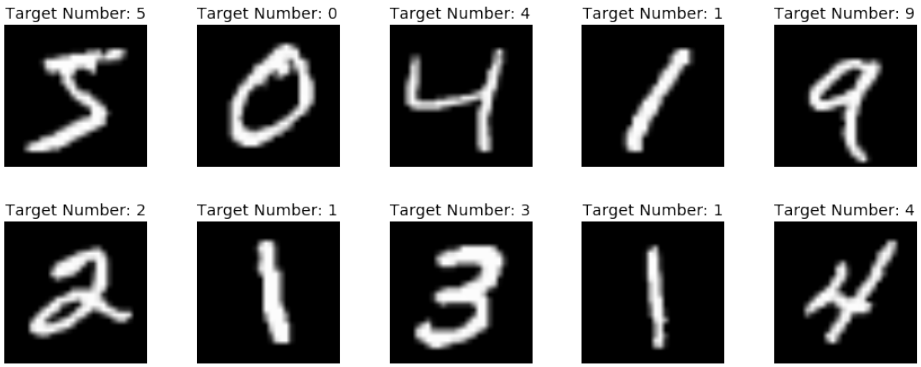
Intro Principle Component Analysis

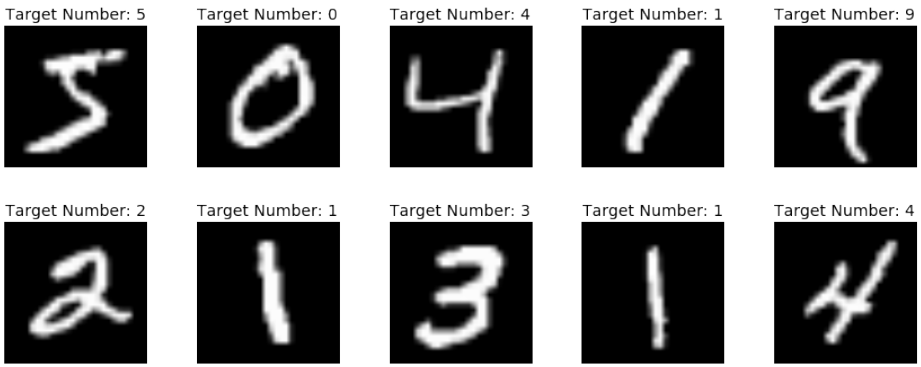
Frederik Mallmann-Trenn
6CCS3AIN

Let's say you want to recognise digits



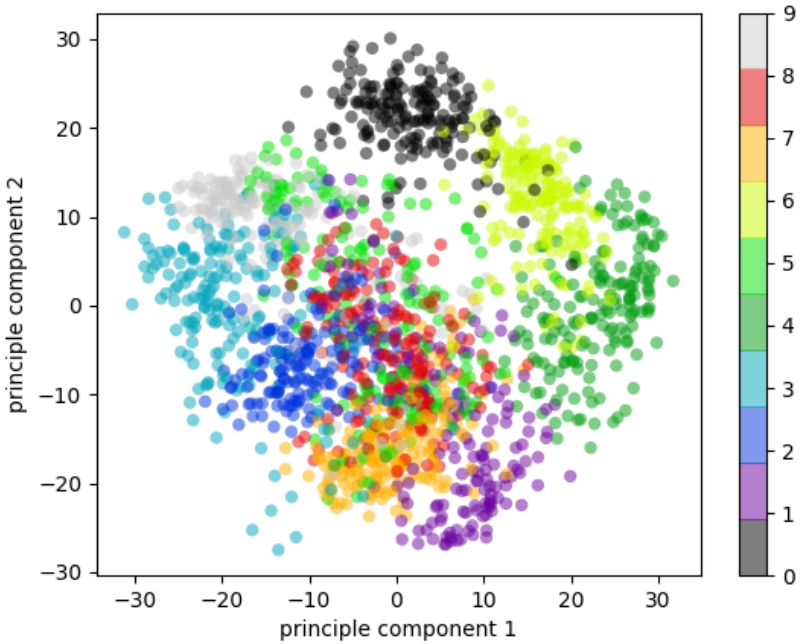
- MNIST: Very famous dataset from scikit-learn
- Let's say you want to use the large training set with examples (128x128 pixels)
- So that when I draw you a new digit, you can tell what it is! 

Let's say you want to recognise digits

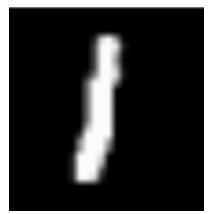


- Problem: Each digit has $128 \cdot 128 = 16,384$ features/dimensions
- Is there a nice way to reduce the number of features/dimensions?

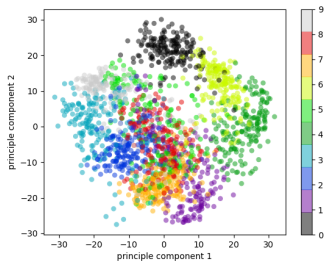
A cool way of doing this



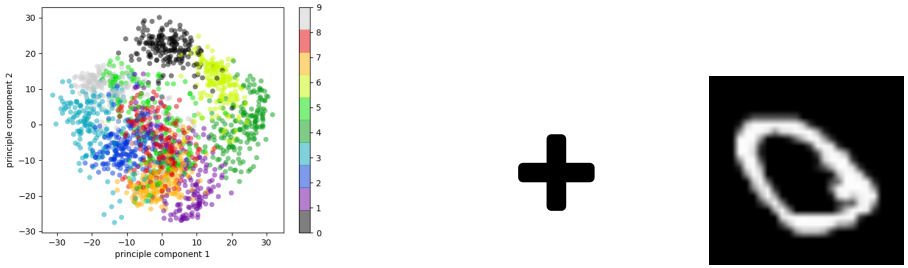
- We try to find the two components (each is a combination of the features)

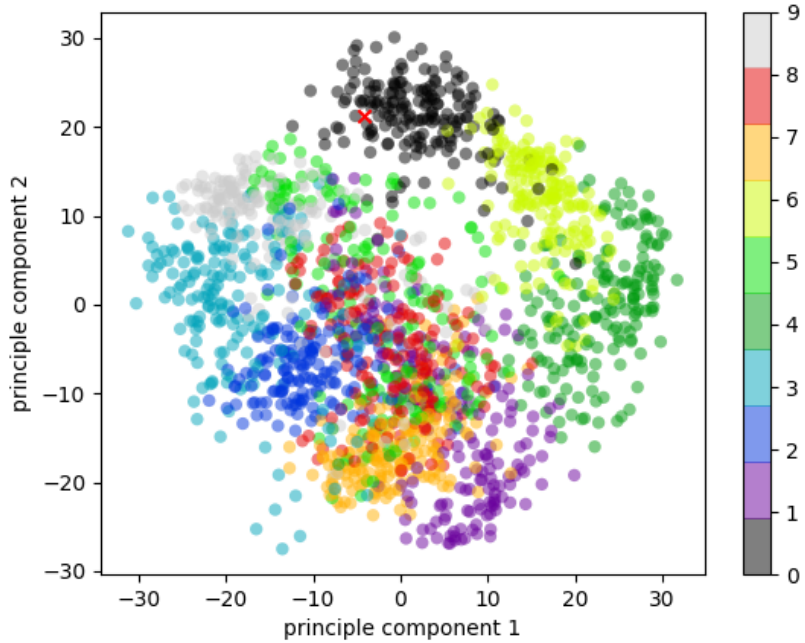


PCA



PCA





- Red cross is new input
- Easy to figure out where it belongs to..

PCA / SVD

Advantages

- State-of-the-art for many applications (supervised and unsupervised)

PCA / SVD

Advantages

- State-of-the-art for many applications (supervised and unsupervised)
- Incredibly efficient (often, almost linear time)

PCA / SVD

Advantages

- State-of-the-art for many applications (supervised and unsupervised)
- Incredibly efficient (often, almost linear time)
- Strong theoretical background

PCA / SVD

Advantages

- State-of-the-art for many applications (supervised and unsupervised)
- Incredibly efficient (often, almost linear time)
- Strong theoretical background
- Can also be used to store data in more efficient way (Image compression).

PCA / SVD

Advantages

- State-of-the-art for many applications (supervised and unsupervised)
- Incredibly efficient (often, almost linear time)
- Strong theoretical background
- Can also be used to store data in more efficient way (Image compression).
- Visual evaluation possible for a small number of components (say 2)

PCA / SVD

Advantages

- State-of-the-art for many applications (supervised and unsupervised)
- Incredibly efficient (often, almost linear time)
- Strong theoretical background
- Can also be used to store data in more efficient way (Image compression).
- Visual evaluation possible for a small number of components (say 2)

PCA / SVD

Advantages

- State-of-the-art for many applications (supervised and unsupervised)
- Incredibly efficient (often, almost linear time)
- Strong theoretical background
- Can also be used to store data in more efficient way (Image compression).
- Visual evaluation possible for a small number of components (say 2)

Small disclaimer: PCA and SVD (Singular value decomposition) are slightly different, but very very similar, we'll look at PCA (which often uses SVD)