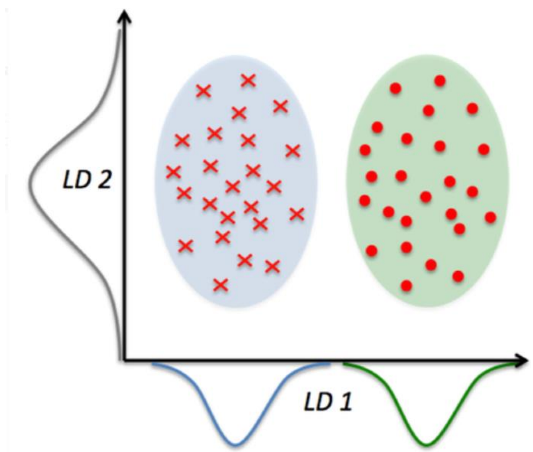
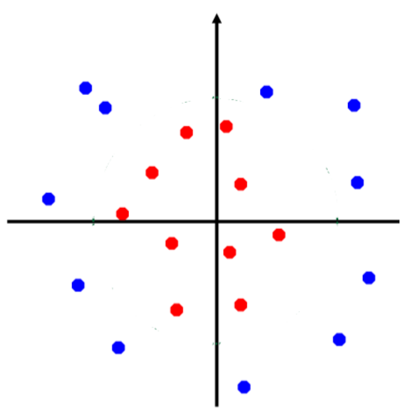
Dimentionality reduction

- Terkadang kita memiliki atribut/fitur data yang sangat banyak, sehingga:
  - ✓ Komputasi menjadi berat
  - ✓ Tidak bisa divisualisasikan → dimensi berapa yang maksimal masih bisa kita visualisasikan?
- **Ilustrasi**: diberikan data dengan 2 dimensi berikut, bagaimana cara terbaik untuk mereduksi menjadi 1 dimensi?





Dimentionality reduction – PCA

- Singkatan dari: Principal Component Analysis
- Tujuan: mereduksi data dengan dimensi m ke dimensi p, yang mana $p \leq m$
- Kategori pembelajaran: masuk dalam kategori *unsupervised learning*
- Ide: memproyeksikan data $\boldsymbol{X}$ ke data terproyeksi $\boldsymbol{Z}$ di mana $\boldsymbol{Z}$ adalah kemungkinan data terproyeksi dengan **nilai variansi tertinggi**.

$$\boxed{\boldsymbol{Z} = \boldsymbol{XW}}$$

Di mana:

Z = data terproyeksi, ukuran $= ?$

W = matriks proyeksi, ukuran $= (m, p)$

X = data yang ingin diproyeksikan, ukuran $= (n_{data}, m)$

Dimentionality reduction – PCA

$$\underset{\boldsymbol{w}}{argmax}\,(varian\;\boldsymbol{Z}) = \underset{\boldsymbol{w}}{argmax}\,((\mathbf{Z} - \bar{\mathbf{Z}}\;)^2)$$

Ada yang bisa mencoba melanjutkan? Akan mendapatkan poin plus bagi yang bisa ☺

Dimentionality reduction – PCA

$$\underset{\boldsymbol{w}}{argmax}(varian\ \boldsymbol{Z}) = \underset{\boldsymbol{w}}{argmax}((\boldsymbol{Z} - \bar{\boldsymbol{Z}}\ )^2)$$

$$\underset{\boldsymbol{w}}{argmax}(varian\ \boldsymbol{Z}) = \underset{\boldsymbol{w}}{argmax}((\boldsymbol{XW} - \bar{\boldsymbol{X}}\boldsymbol{W})^2)$$

$$\underset{\boldsymbol{w}}{argmax}(varian\ \boldsymbol{Z}) = \underset{\boldsymbol{w}}{argmax}\left(\left((\boldsymbol{X} - \bar{\boldsymbol{X}})\boldsymbol{W}\right)^2\right)$$

$$\underset{\boldsymbol{w}}{argmax}(varian\ \boldsymbol{Z}) = \underset{\boldsymbol{w}}{argmax}\left(\left((\boldsymbol{X} - \bar{\boldsymbol{X}})\boldsymbol{W}\right)^{\boldsymbol{T}}(\boldsymbol{X} - \bar{\boldsymbol{X}})\boldsymbol{W}\right)$$

$$\underset{\boldsymbol{w}}{argmax}(varian\ \boldsymbol{Z}) = \underset{\boldsymbol{w}}{argmax}(\boldsymbol{W}^T(\boldsymbol{X} - \bar{\boldsymbol{X}})^T(\boldsymbol{X} - \bar{\boldsymbol{X}})\boldsymbol{W})$$

$$\boxed{\underset{\boldsymbol{w}}{argmax}(varian\ \boldsymbol{Z}) = \underset{w}{argmax}(\boldsymbol{W}^T\boldsymbol{S}\ \boldsymbol{W})}$$

Di mana $\boldsymbol{S} = (\boldsymbol{X} - \bar{\boldsymbol{X}})^T(\boldsymbol{X} - \bar{\boldsymbol{X}})$

Dimentionality reduction – PCA

$$\boxed{\underset{w}{argmax}\,(varian\ \boldsymbol{Z}) = \underset{w}{argmax}\,(\boldsymbol{W}^T \boldsymbol{S}\ \boldsymbol{W})}$$

$$\underset{w}{argmax}\,(varian\ \boldsymbol{Z}) = \frac{d(\boldsymbol{W}^T \boldsymbol{S}\ \boldsymbol{W})}{d\boldsymbol{W}} = 0$$

Untuk mendapatkan solusi $\boldsymbol{W}$ yang *'unique'*,
tambahkan *constraint* lagrange multiplier $\boldsymbol{W}^T\boldsymbol{W} = 1$

$$\underset{w}{argmax}\,(varian\ \boldsymbol{Z}) = \frac{d(\boldsymbol{W}^T \boldsymbol{S}\ \boldsymbol{W} - \lambda(\boldsymbol{W}^T\boldsymbol{W} - 1))}{d\boldsymbol{W}} = 0$$

$$\underset{w}{argmax}\,(varian\ \boldsymbol{Z}) = \frac{d(\boldsymbol{W}^T \boldsymbol{S}\ \boldsymbol{W} - \lambda\boldsymbol{W}^T\boldsymbol{W} + \lambda)}{d\boldsymbol{W}} = 0$$

$$\underset{w}{argmax}\,(varian\ \boldsymbol{Z}) = 2\boldsymbol{S}\boldsymbol{W} - \lambda 2\boldsymbol{W} = 0$$

$$2\boldsymbol{S}\boldsymbol{W} = \lambda 2\boldsymbol{W}$$

$$\boxed{\boldsymbol{S}\boldsymbol{W} = \lambda\boldsymbol{W}} \longrightarrow$$

Solusi PCA adalah dengan dekomposisi Eigen

Dimentionality reduction – PCA

a. Diberikan data asli $\boldsymbol{X}$, hitung matriks kovarian $\boldsymbol{S}$.

$$\boldsymbol{S} = (\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}})$$

b. Dekomposisikan $\boldsymbol{S}$ ke *eigenvector* dan *eigenvalue.*

c. Jika dikehendaki direduksi ke dimensi p, maka ambil sebanyak p *eigenvector* dari hasil langkah (b) yang memiliki nilai *eigenvalue* p-terbesar. Setiap *eigenvectornya* ini kita susun sebagai matriks kolom, yang selanjutnya akan kita akan gunakan sebagai matriks proyeksi $\boldsymbol{W}$.

d. Data yang terproyeksi dapat dihitung → $\boldsymbol{Z} = \boldsymbol{XW}$

Ingat! ***Z*** ini dimensinya sudah tereduksi dan nilai variansinya paling tinggi.

Dimentionality reduction – PCA

- Hitung $S = (X - \overline{X})^T (X - \overline{X})$

```
mean:
 [185.72 151.12]
matriks kovarians:
 [[2287.04 1268.84]
 [1268.84 1304.64]]
```

- Dekomposisikan ke eigen

```
(eigenvalue, eigenvector) =
np.linalg.eig(matriks_kovarians)
```

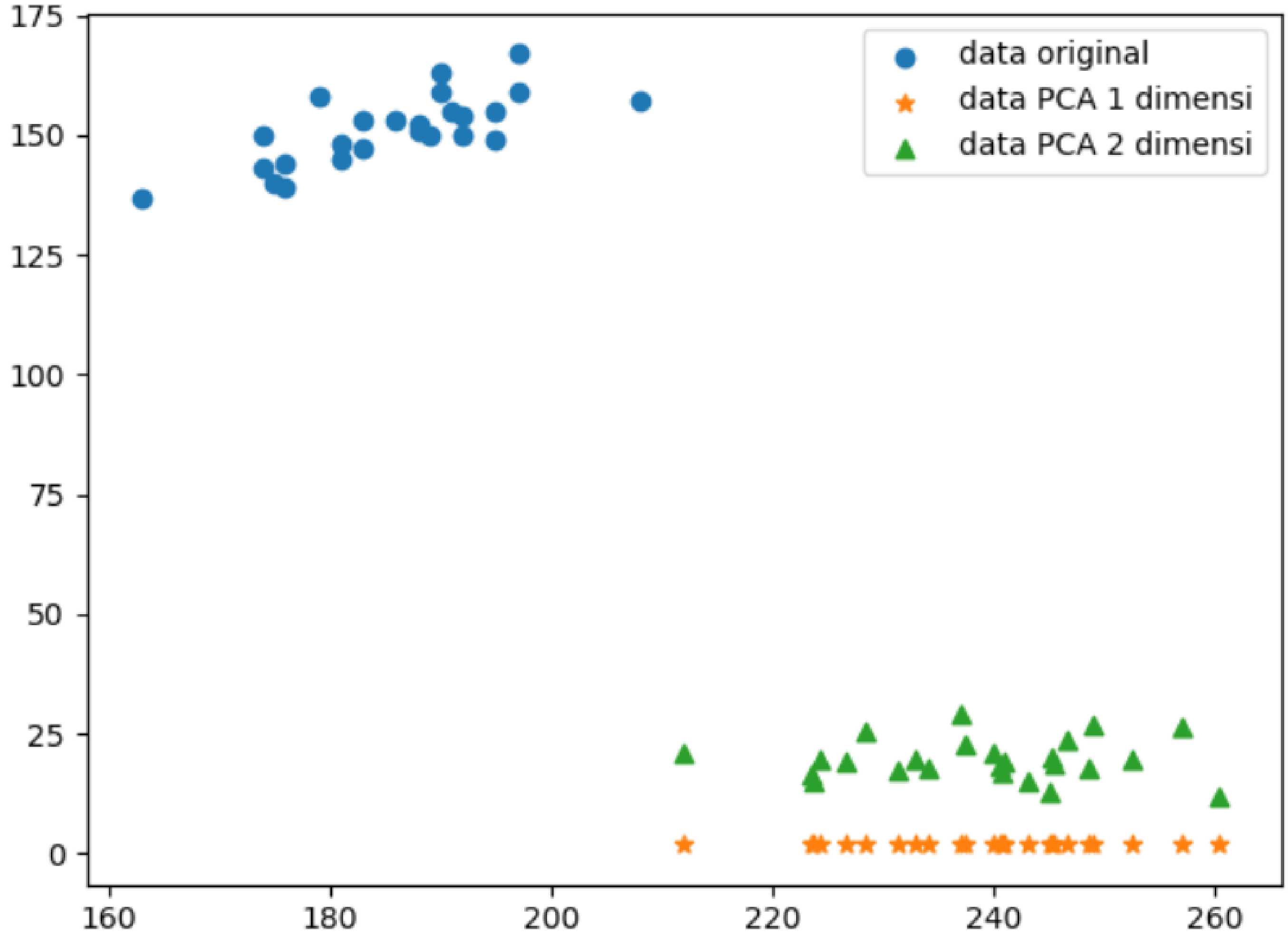
```
eigenvalue:
 [3156.44000941  435.23999059]
eigenvector:
 [[ 0.82492945 -0.5652357 ]
 [ 0.5652357   0.82492945]]
```

- Bangun matriks ***W*** dari eigenvector, dan hitung ***Z*,** contoh:

$$Z = XW = \begin{bmatrix} 191 & 155 \\ 195 & 149 \\ \vdots & \vdots \\ 190 & 163 \end{bmatrix} \begin{bmatrix} 0.825 \\ 0.565 \end{bmatrix} = \begin{bmatrix} 245.15 \\ 245.06 \\ \vdots \\ 248.8 \end{bmatrix}$$

(x_1)	(x_2)
191	155
195	149
181	148
183	153
176	144
208	157
189	150
197	159
188	152
192	150
179	158
183	147
174	150
190	159
188	151
163	137
195	155
186	153
181	145
175	140
192	154
174	143
176	139
197	167
190	163

Dimentionality reduction – PCA

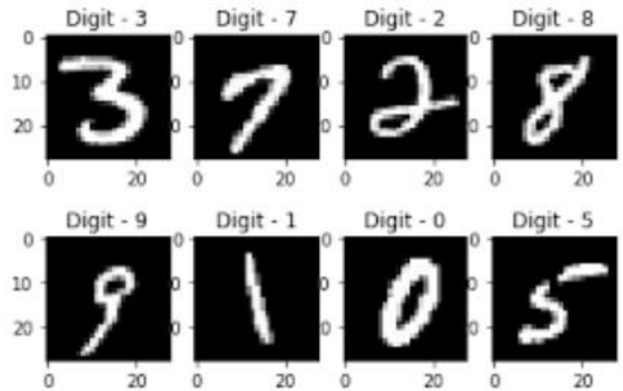


X_2

X_1

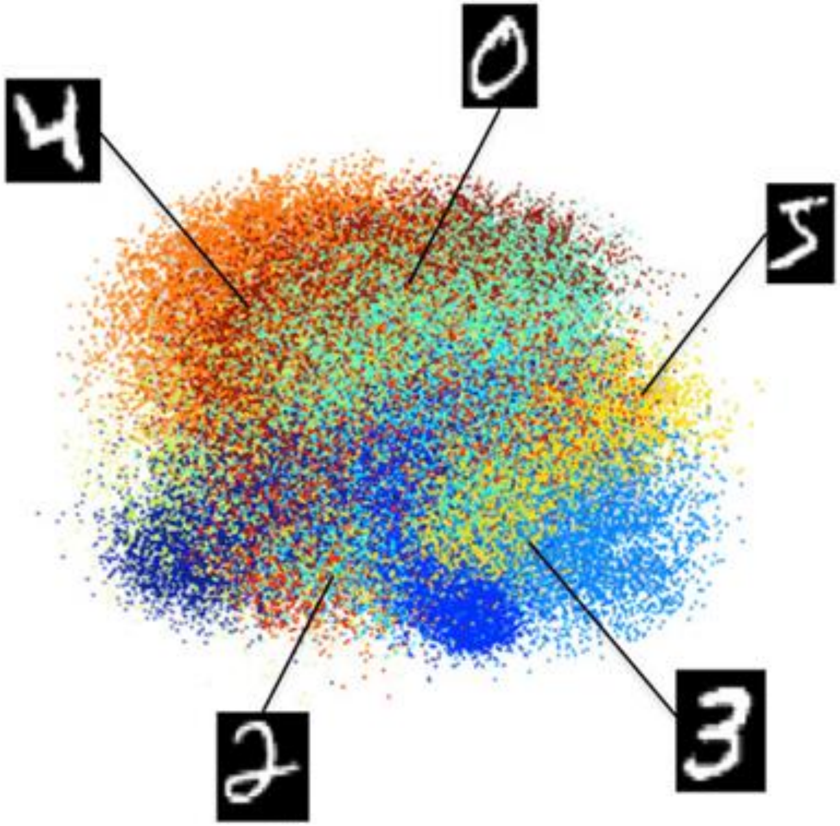
Hasil plot matriks terproyeksi ***Z*** dengan PCA dari data ***X*** (halaman sebelumnya)

Dimentionality reduction – PCA



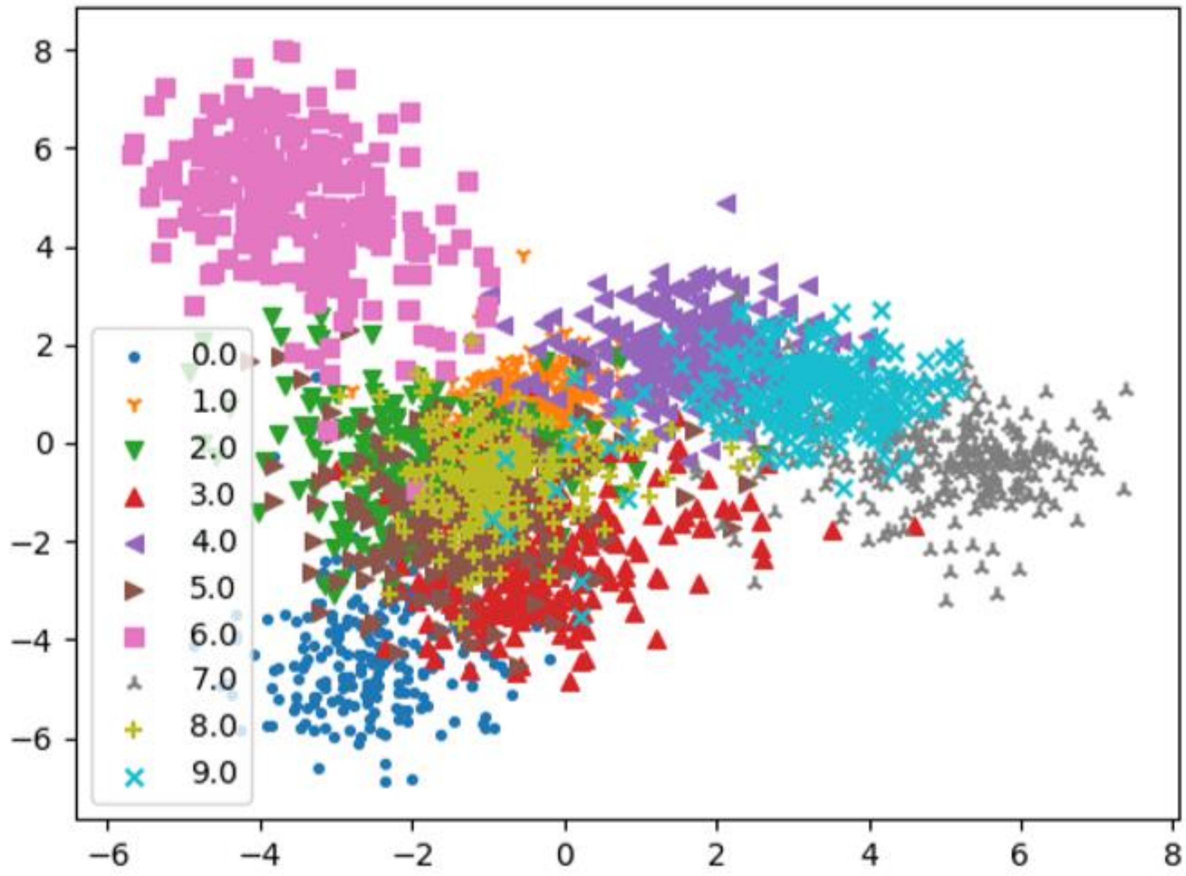
Contoh digit angka MNIST

- Angka = 0 – 9
- Ukuran = 28 x 28 = 784 fitur



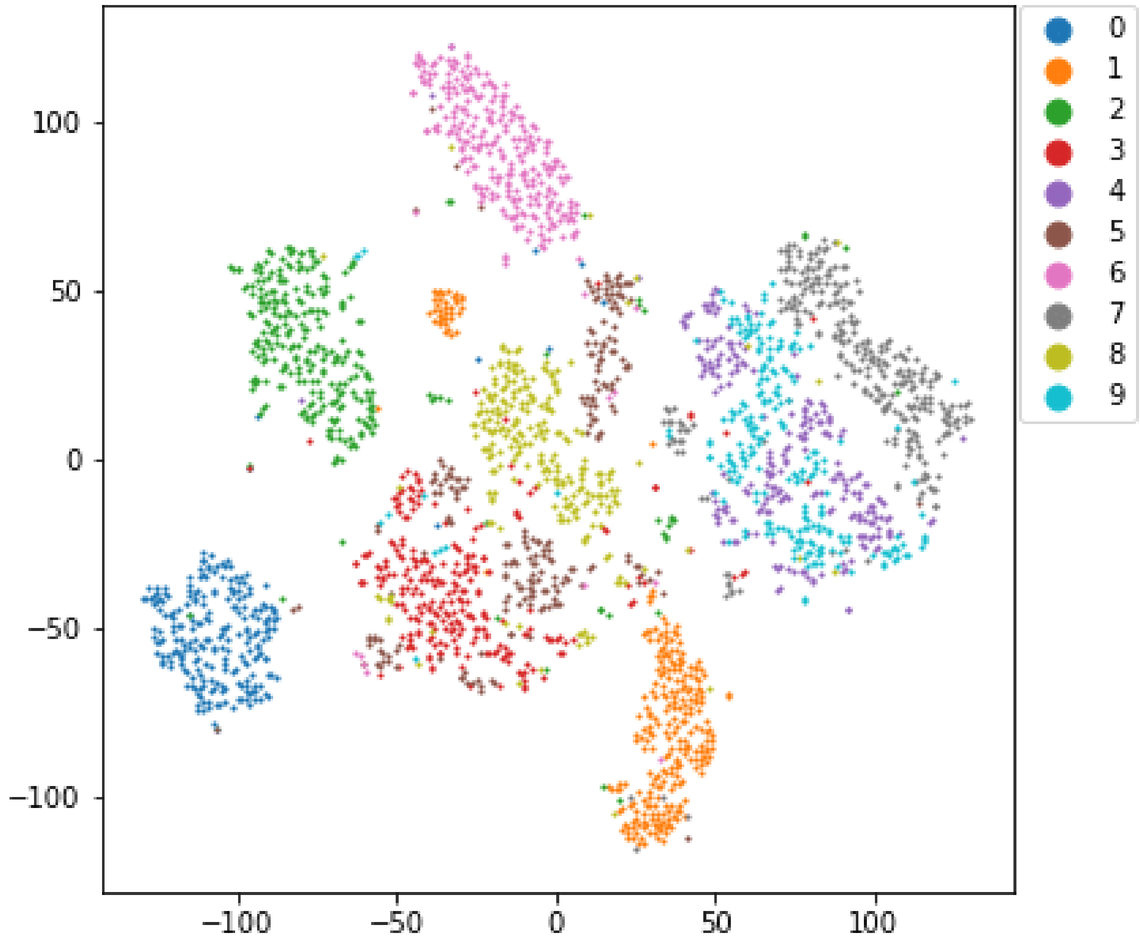
Hasil reduksi ke 2 dimensi dengan PCA

Dimentionality reduction – LDA



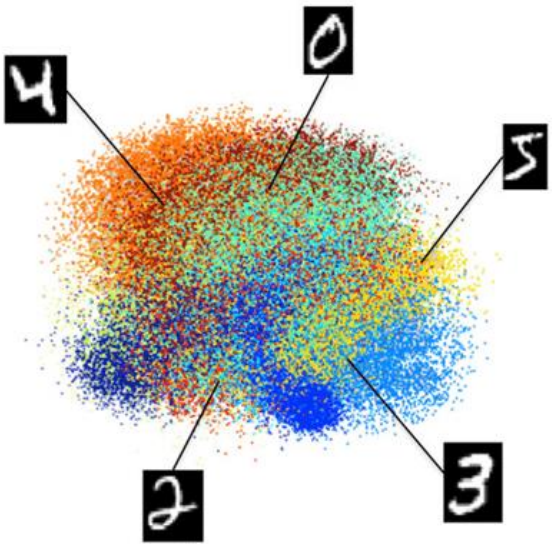
Welling, M. (2005). Fisher linear discriminant analysis. *Department of Computer Science, University of Toronto*.

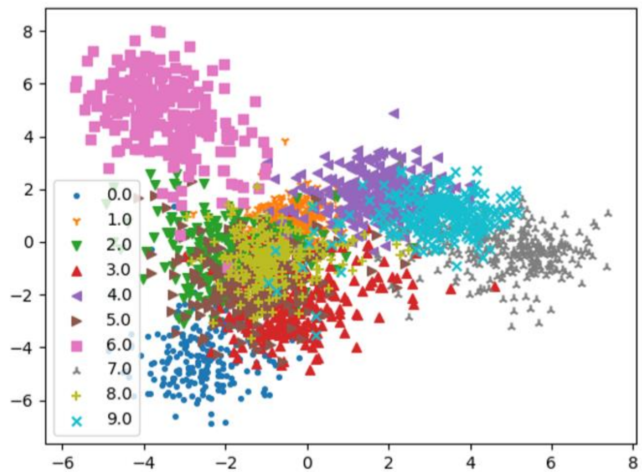
Dimentionality reduction – tSNE

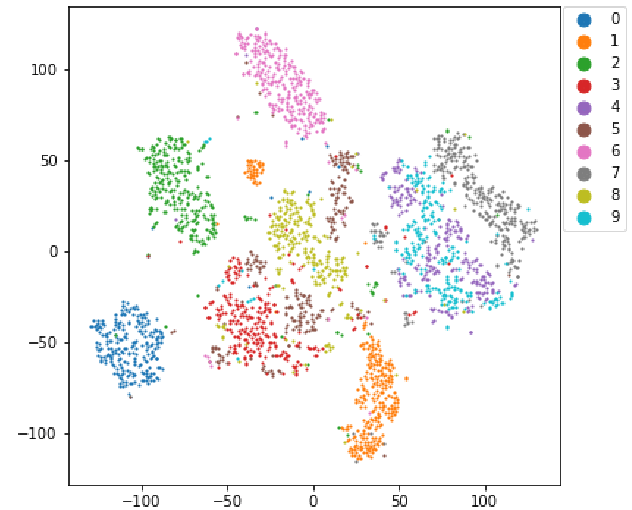


Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*

Dimentionality reduction – perbandingan







PCA
- Tidak butuh label
- Memaksimalkan variansi

LDA
- Butuh label
- Memaksimalkan *Fisher discriminant*

t-SNE
- Tidak butuh label
- Meminimalkan *Kullback-Leibler divergence / relative entropy*

Yang mana yang paling bagus?

End..