Dimensionality Reduction

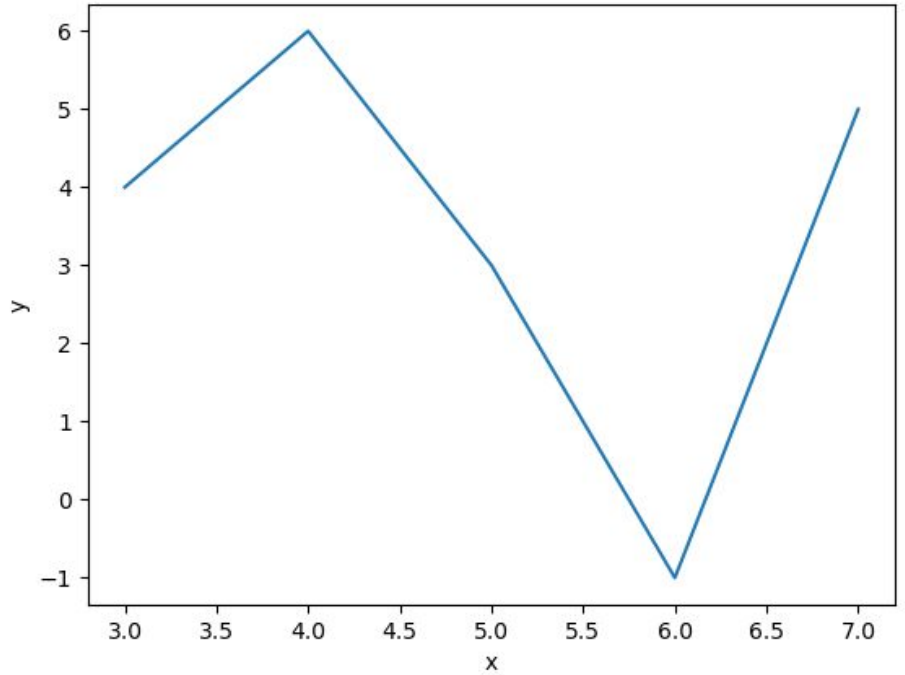
PCA & t-SNE

How would you plot this?

x	y
3	4
4	6
5	3
6	-1
7	5

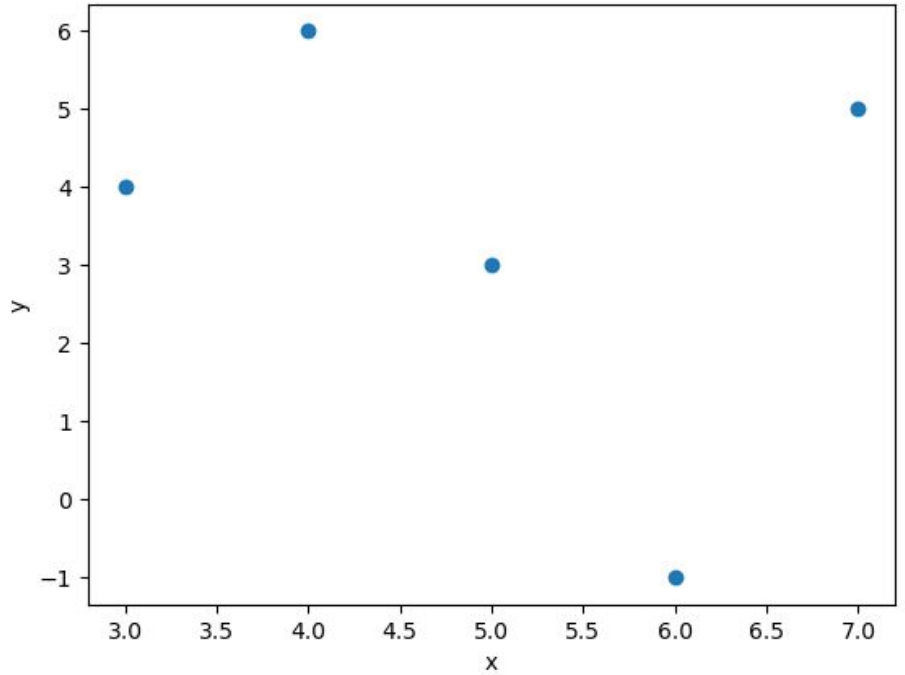
How would you plot this?

x	y
3	4
4	6
5	3
6	-1
7	5



How would you plot this?

x	y
3	4
4	6
5	3
6	-1
7	5

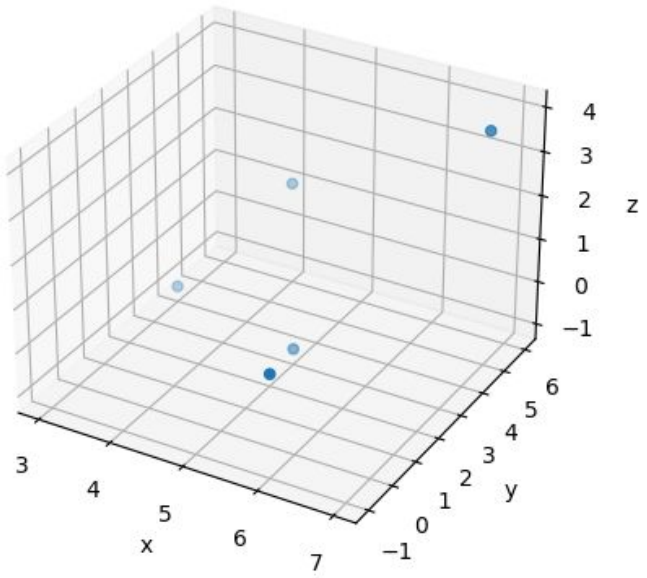


What about 3 dimensions?

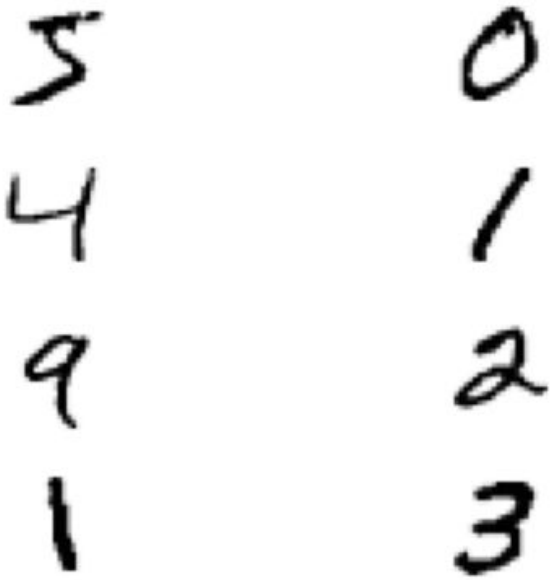
x	y	z
3	4	-1
4	6	1
5	3	-1
6	-1	1
7	5	4

What about 3 dimensions?

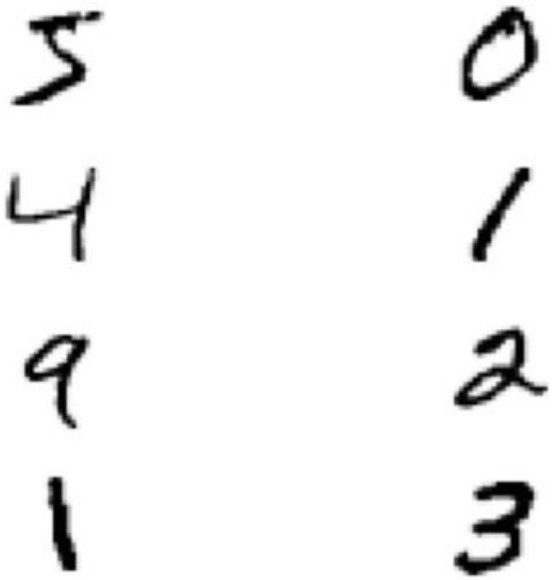
x	y	z
3	4	-1
4	6	1
5	3	-1
6	-1	1
7	5	4



What about higher-dimensional data?

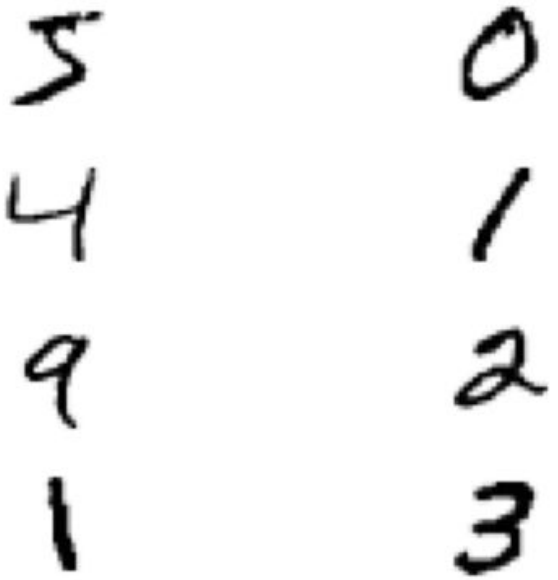


What about higher-dimensional data?



28 * 28 = 784 dimensions (pixels) per image!

Dimensionality Reduction



Goal: reduce each 784-dimensional data point down to a couple dimensions

Dimensionality Reduction Algorithms

Principal Component Analysis (PCA)

- Widely-used traditional method
- Classical; based on linear algebra
- Very efficient
- Keeps global structure (at cost of local information)

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Relatively new (2008)
- Nonlinear manifold learning algorithm
- Can capture complex structure
- Focuses on local structure (at cost of global information)