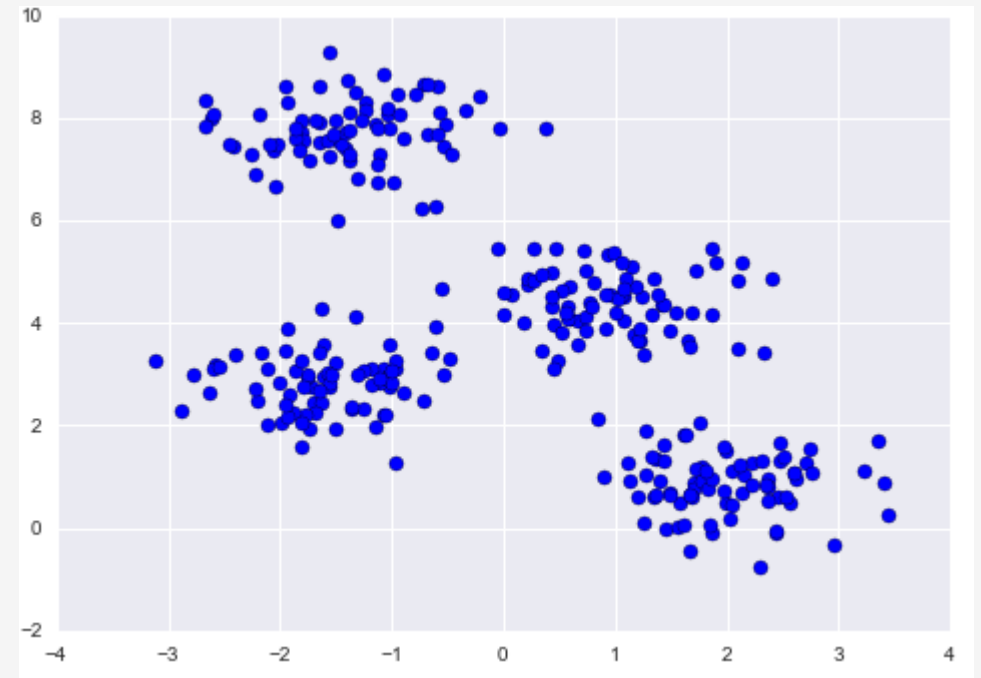


Clustering

- Unsupervised Learning
- Goal of clustering is to group set of objects based on similar characteristics
- Help find meaningful structure among your data, group similar data together and discover underlying patterns



Iris dataset

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa



75 × 477 Iris Versicolor



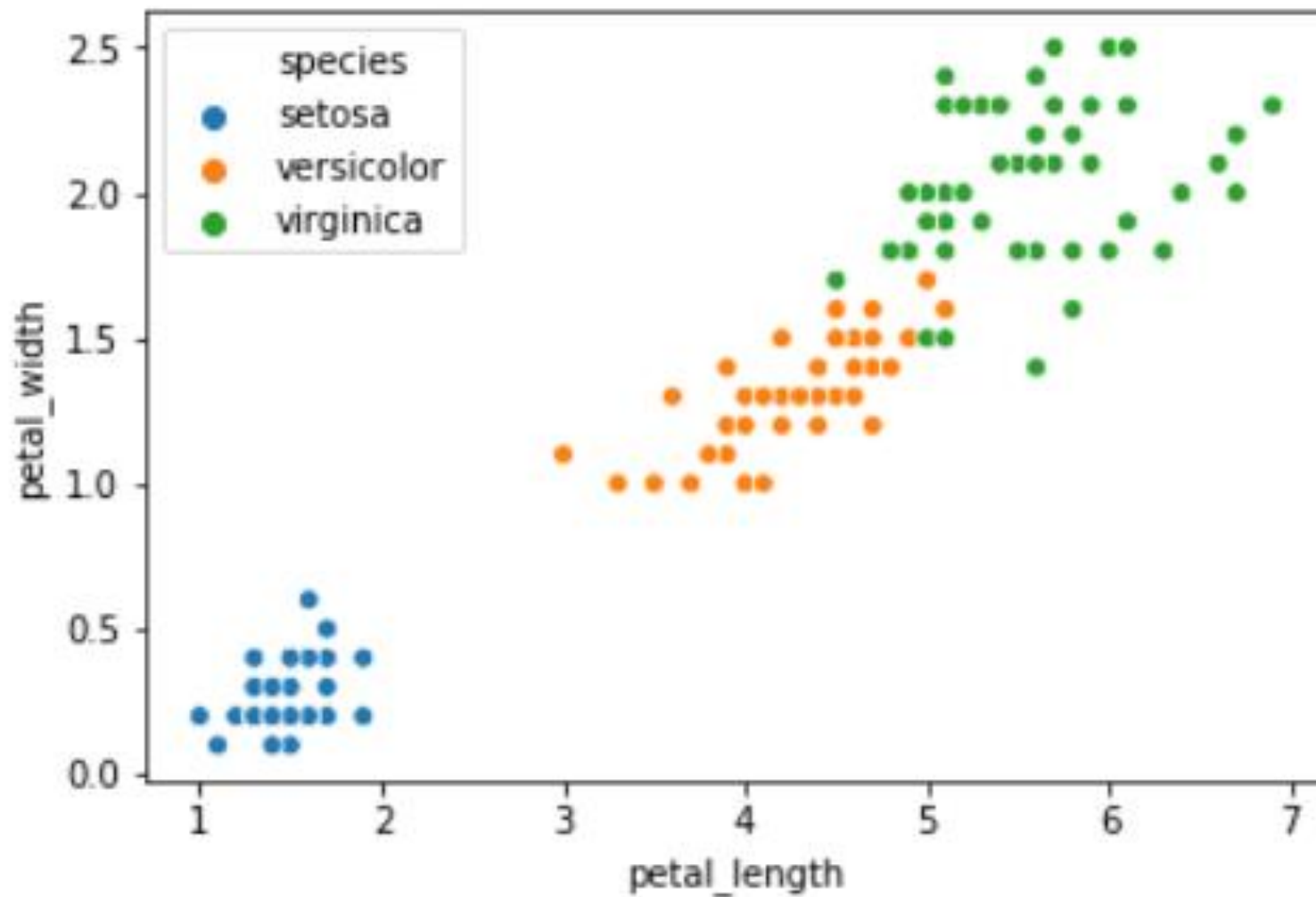
Iris Setosa



Iris Virginica



Iris dataset



K-Means Algorithm

K-means Pseudo Code — TEXTBOOK VERSION OF THIS ON
PG 143

1. Pick a value for k (the number of clusters to create)
2. Initialize k 'centroids' (starting points) in your data
3. Create your clusters. Assign each point to the nearest centroid.
4. Make your clusters better. Move each centroid to the center of its cluster.
5. Repeat steps 3–4 until your centroids converge.

Video shows how K-Cluster works

<https://www.youtube.com/watch?v=aWzGGNrcio>

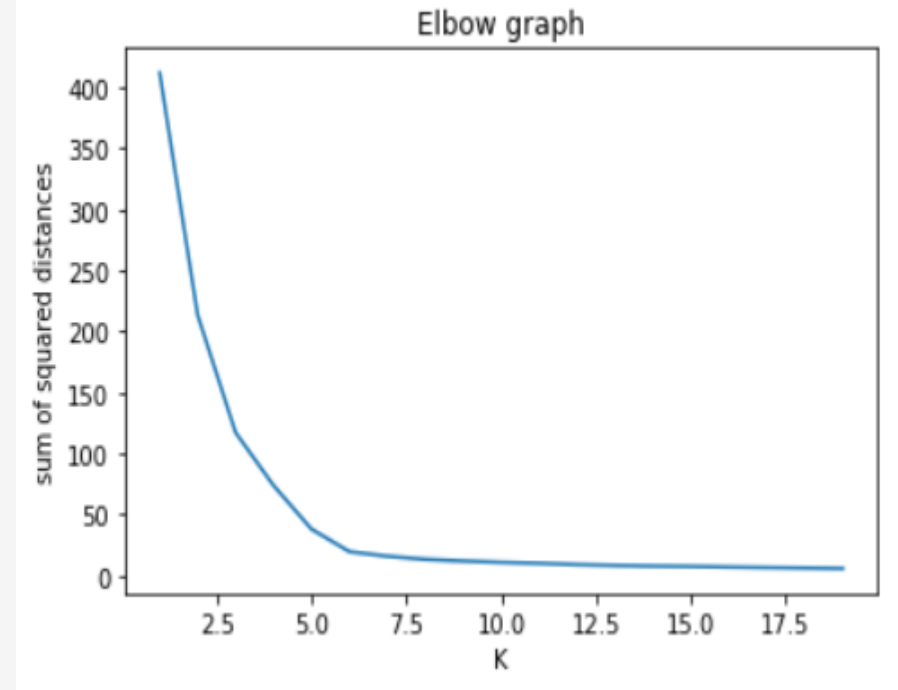
Starting from 4:20

How to choose K?

- Sum of squared distance of each data point to its closest centroid should be small if our clusters make sense
- So if try different value of K, this sum of squared should decreases
- After certain value of K, the marginal benefit of adding more cluster would not help
- The resulting graph looks like an elbow and one can pick K by looking at the point of inflection. The graph is called an elbow graph

This is the ELBOW METHOD for determining the amount of K clusters to have when we CANT determine a good amount of clusters

In the beginning, the centroid is picked randomly. The centroid is moved until no data points move to another cluster. At K clusters there is very little movement of clusters so there is no point in trying to try more clusters because the tradeoff isn't worth it. In this example, K = 5 would be best because after the tradeoff isn't worth it



The more clusters the smaller the sum of squared distance

K-Means Clustering

Learning by doing