DBscan talk – Unsupervised learning- you do not know the number of clusters or cluster membership

There are two parameters to the algorithm,
min_samples: min number of samples to be considered a cluster (or core point)
eps which define formally what we mean when we say dense. Higher min_samples or lower eps indicate higher density necessary to form a cluster.

Its a density based algorithm, clusters are in high density regions

Problems
K-means is vulnerable to outliers
K-means works best with spherical clusters
K-means requires you to pick the number of clusters you have before you run the algorithm (how to know?)
K-means is not guaranteed to produce the same result every run (it depends on initial cluster centers)

Algorithm
Choose eps (radius) and min_samples

1. Randomly pick a point
2. while(other points to process)
   If there are at least min_samples within eps distance of that point, it is a core point
   choose another point

Now all core points are classified

3. Randomly pick a non clustered core point
4. Assign to a cluster
5. for (every core point in cluster)
6. assign all core points withen eps to that cluster
7. Add all non core points withen eps of a core point to the cluster
8. If more core points go to step 3

Any remaining points are called outliers