DBSCAN

Chelsea Parlett-Pelleriti

DBSCAN

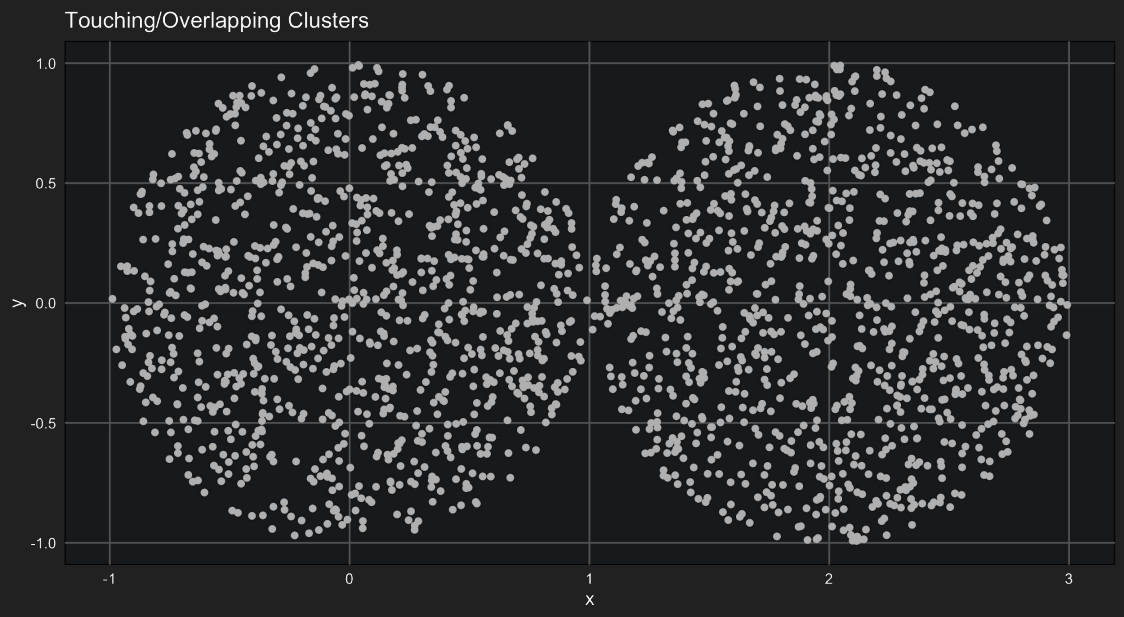
Density **B**ased **S**patial **C**lustering of **A**pplications with **N**oise

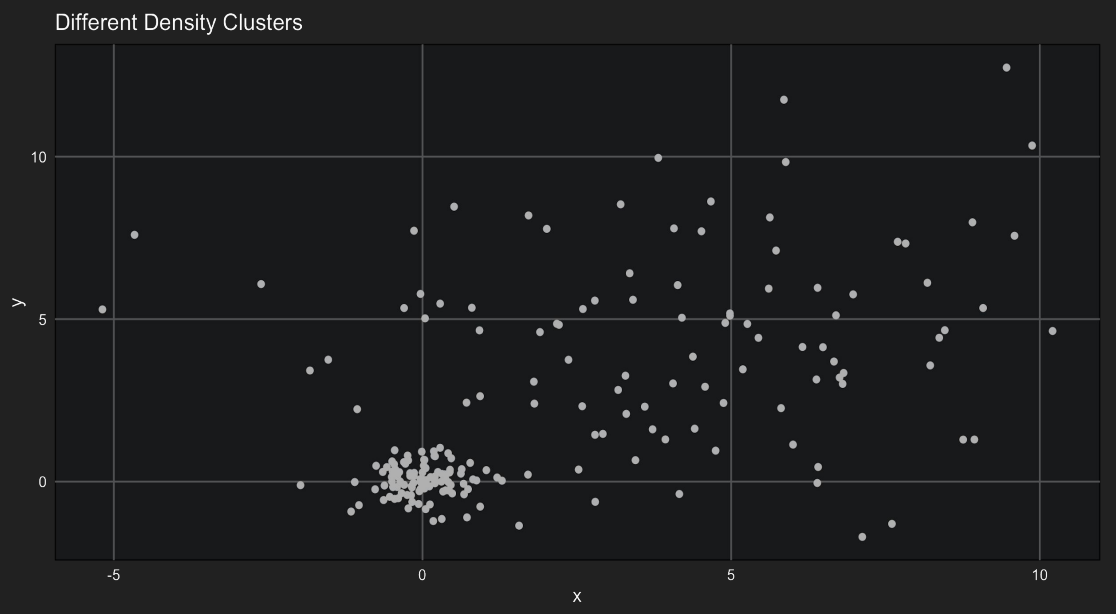
- Distance Metric
- Epsilon (*eps*)
- Minimum Points (*minpts*)

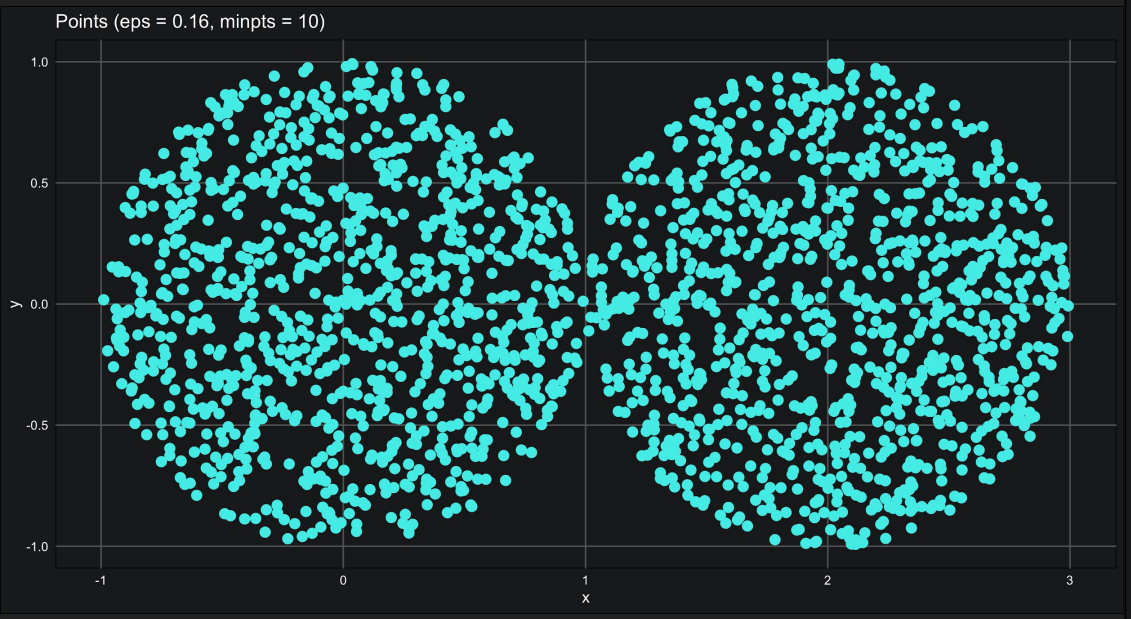
Benefits of DBSCAN

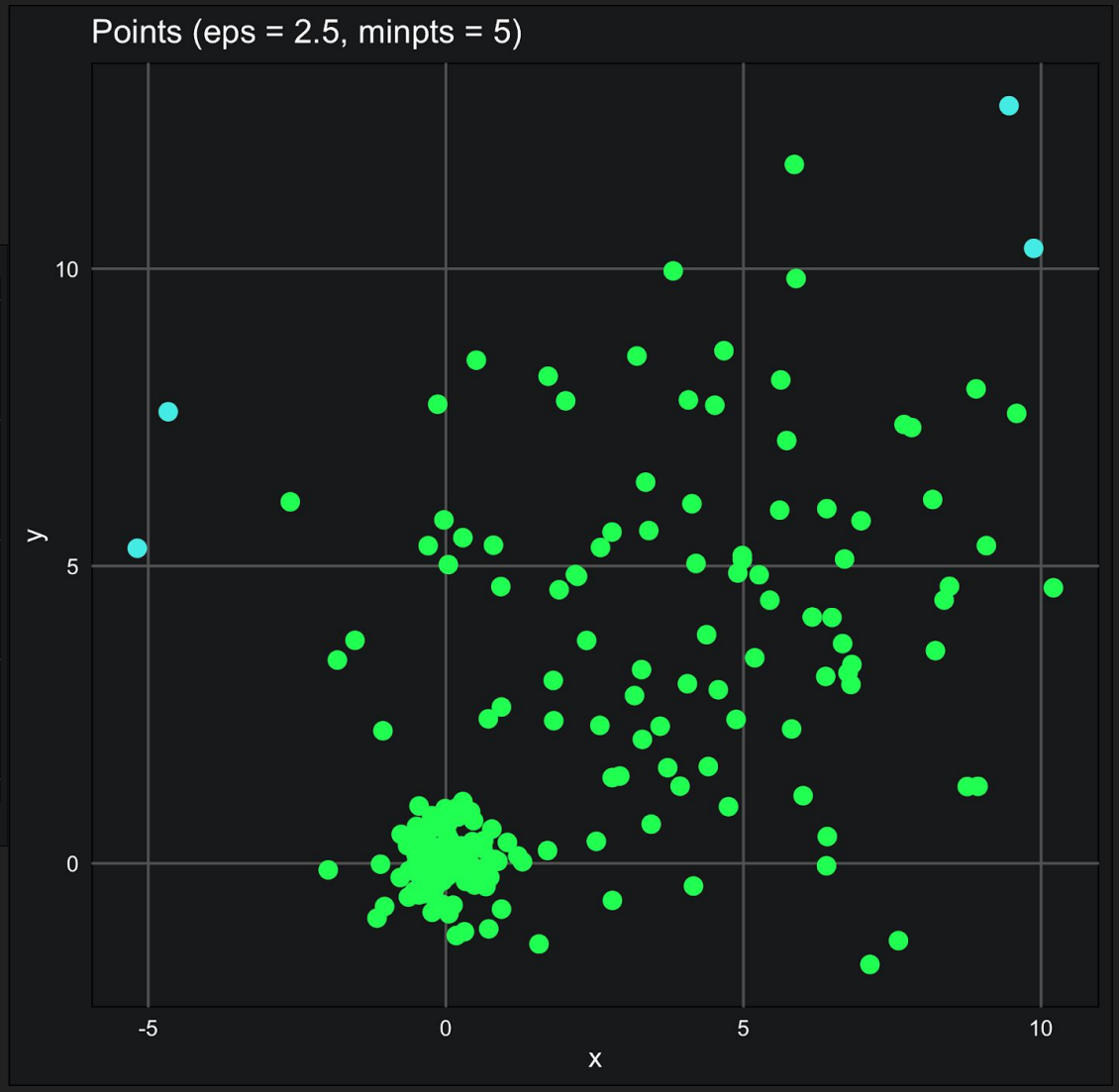
Disadvantages of DBSCAN

- Can be less effective in High Dimensional Data
- Not great with overlapping/touching clusters
- Suboptimal when clusters have different densities






Points (eps = 0.16, minpts = 10)
1.0
0.5
0.0
-0.5
-1.0
y
-1
0
1
2
3
x

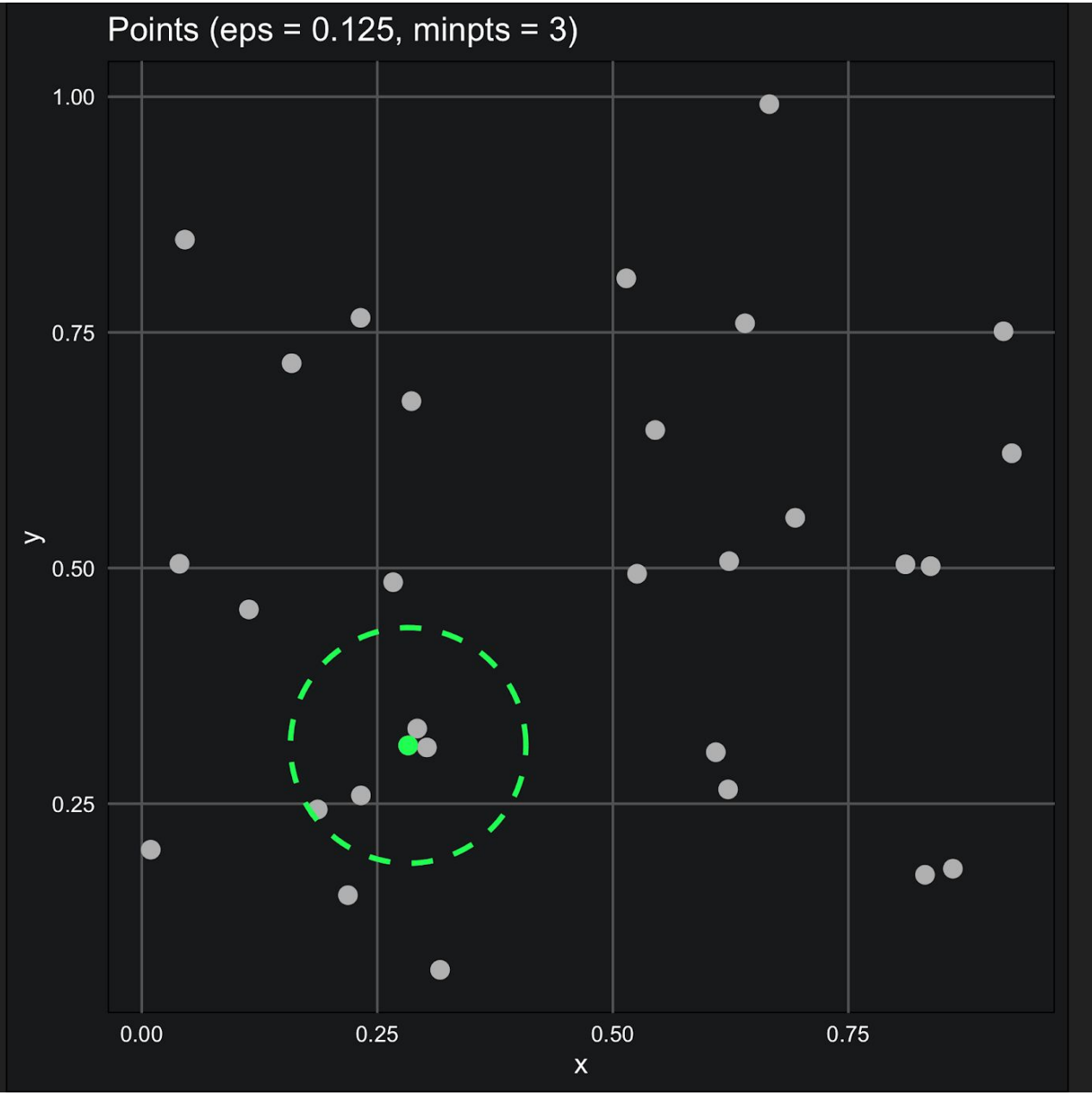

Points (eps = 2.5, minpts = 5)
10
5
0
y
-5
0
5
10
x

DBSCAN

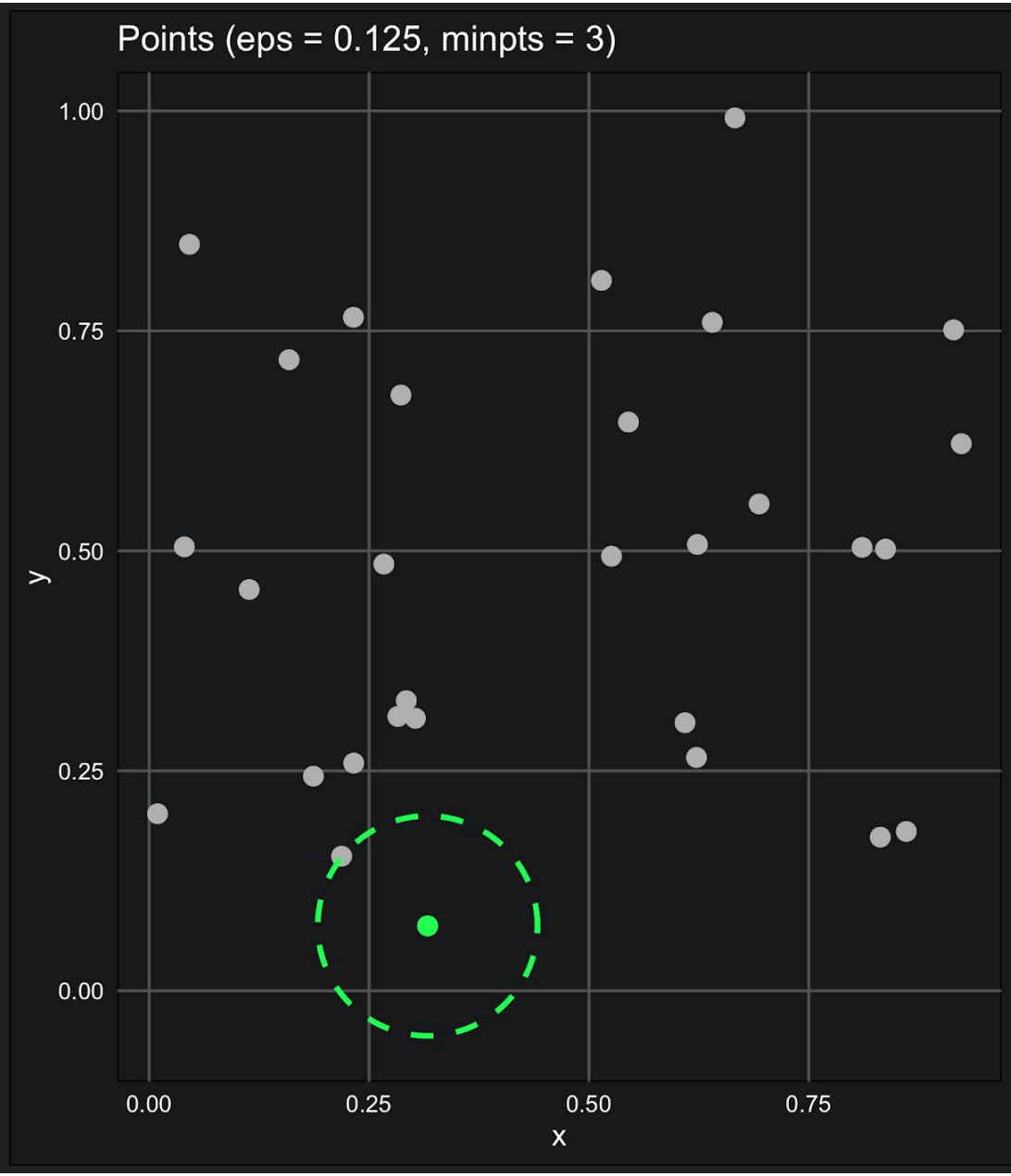
Density **B**ased **S**patial **C**lustering of **A**pplications with **N**oise

- Distance Metric
- Epsilon (*eps*)
- Minimum Points (*minpts*)

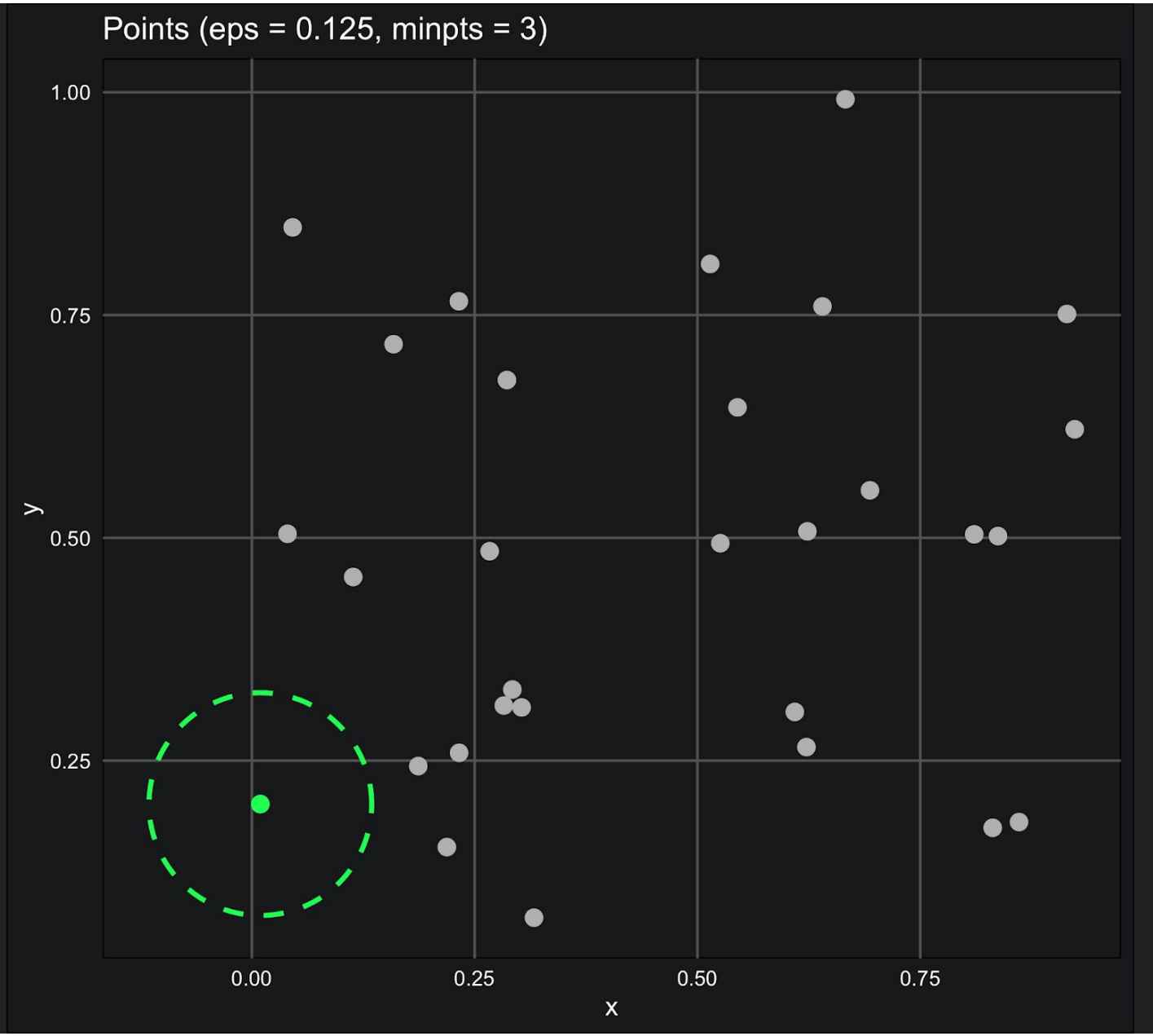
Core Point*:* **p** is a core point if it has at least *minpts* neighbors within *eps* distance of itself



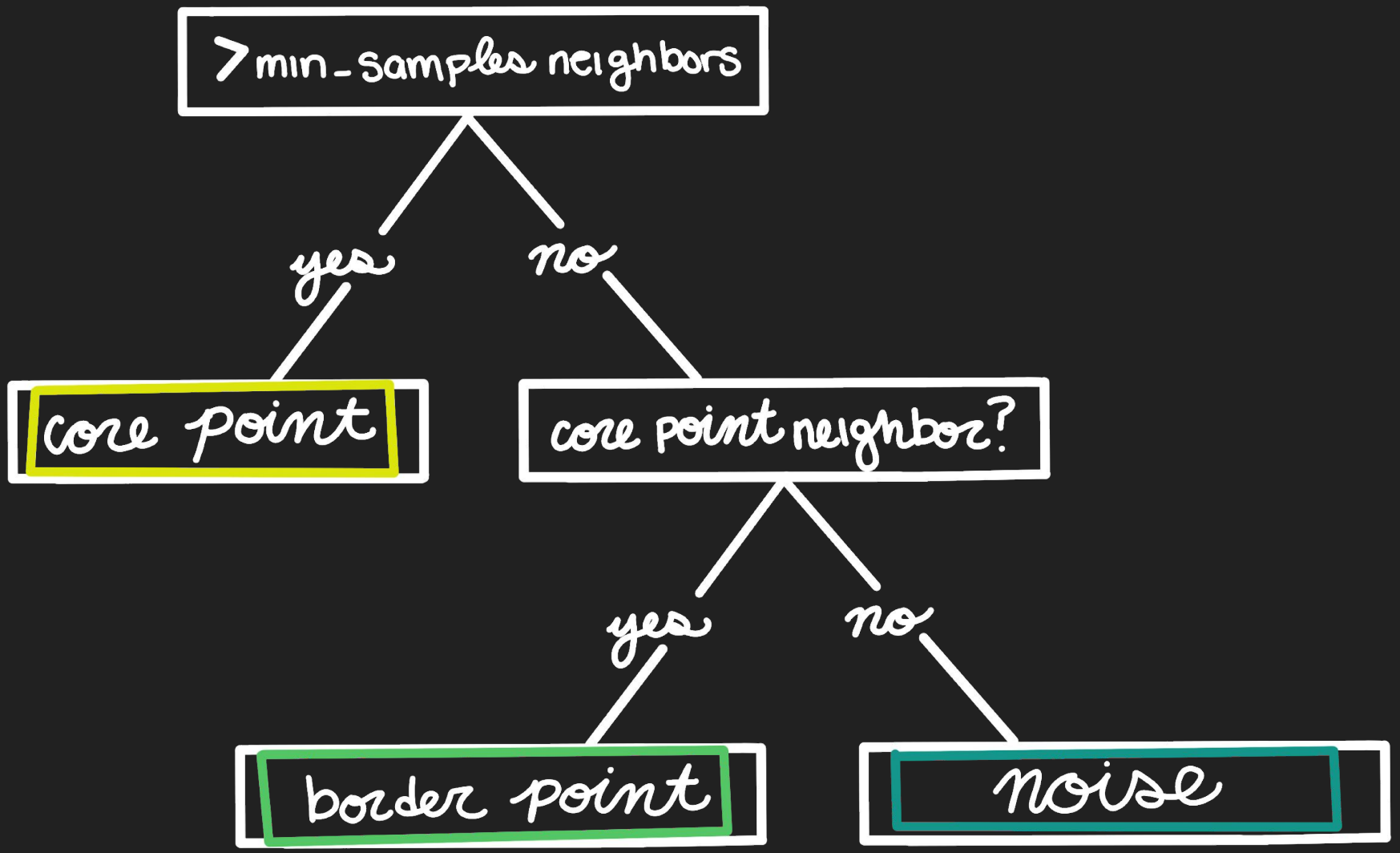
Border Point*: **p** is a border point if it DOES NOT have at least *minpts* neighbors within *eps* distance of itself, but is a neighbor of a core point



Noise*:* **p** is noise if it DOES NOT have at least *minpts* neighbors within *eps* distance of itself, and IS NOT a neighbor of a core point



DBSCAN

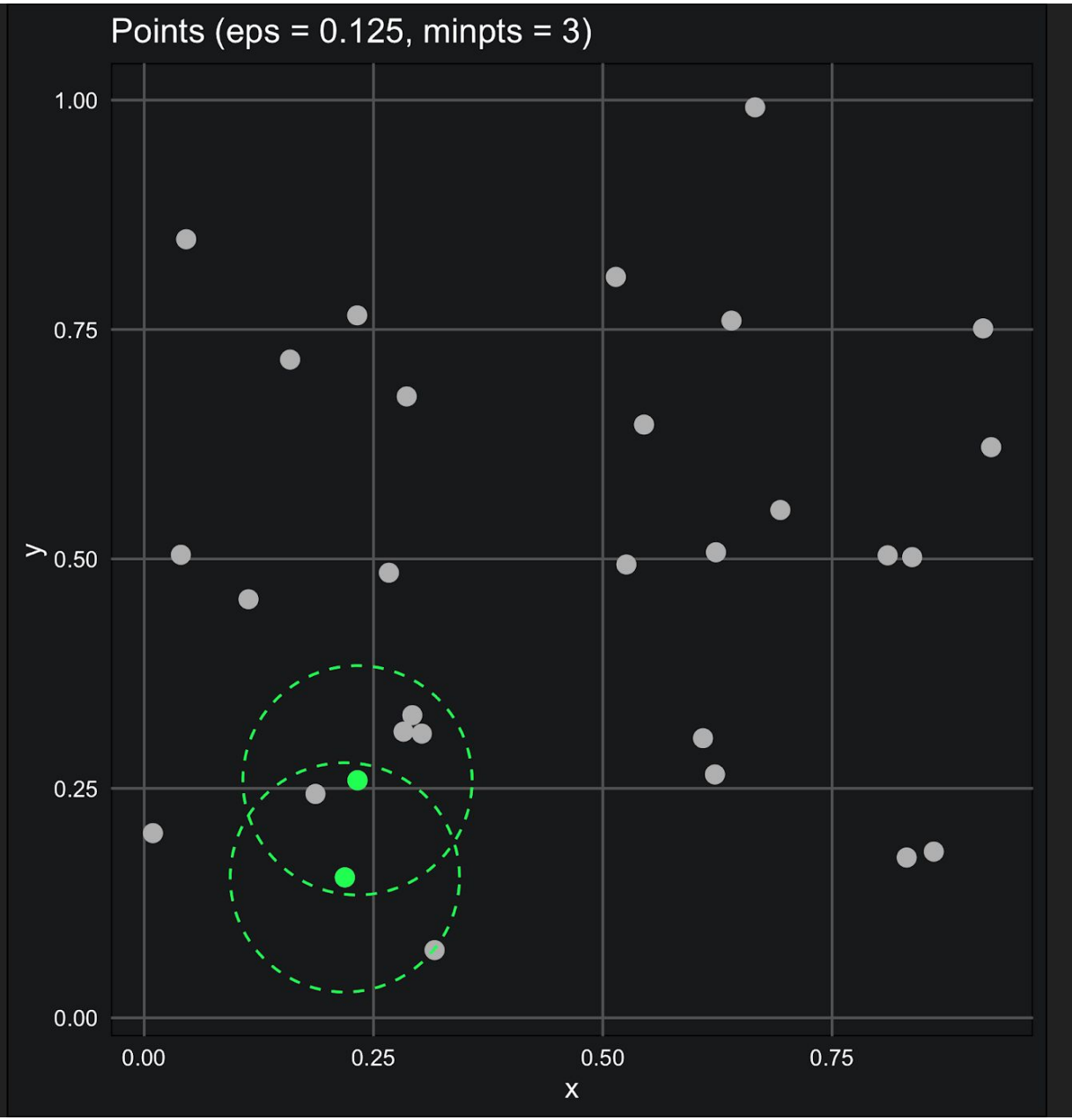

> min_samples neighbors
yes
no
core point
core point neighbor?
yes
no
border point
noise

Definitions

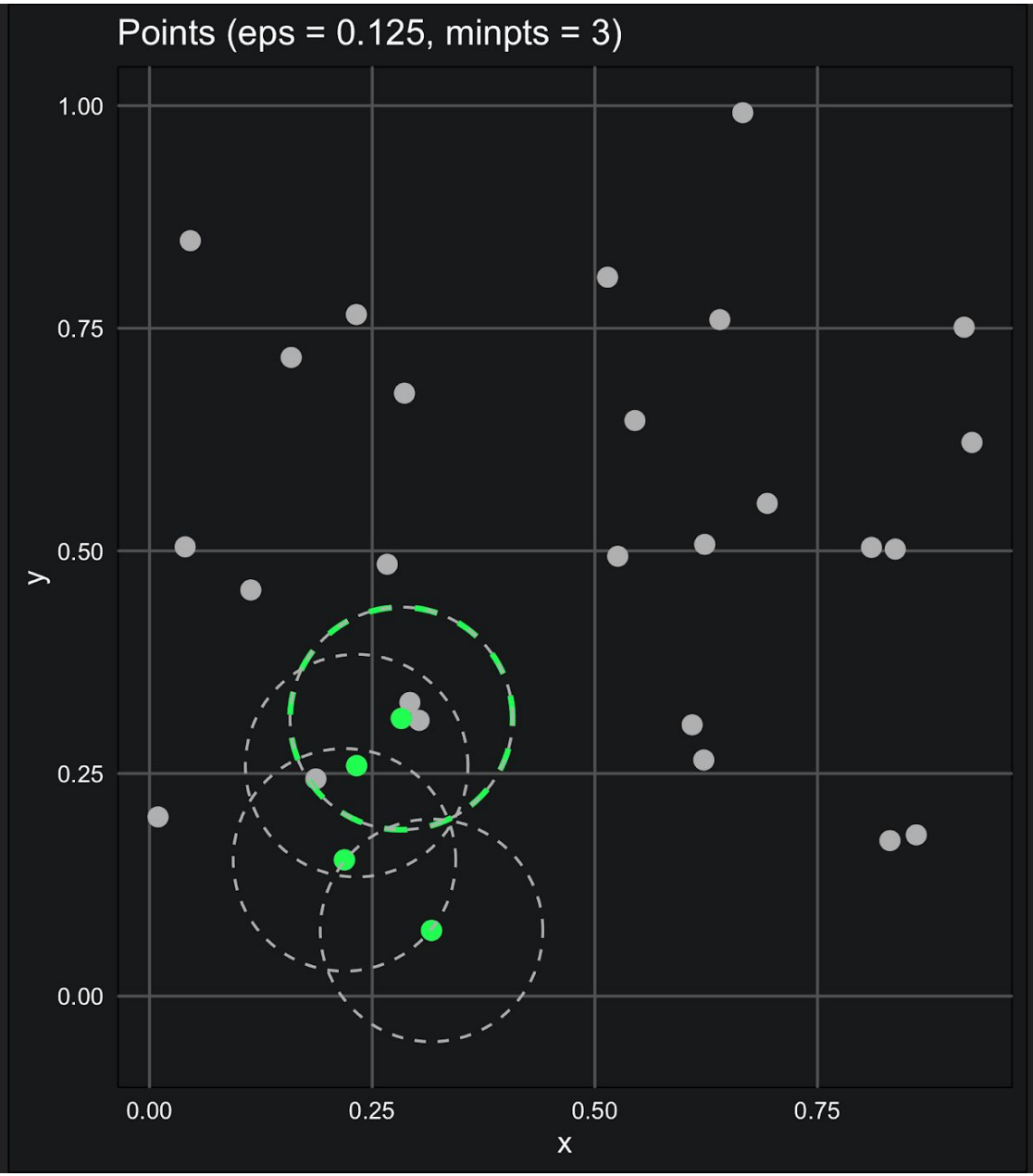
- Directly density reachable: **p** is directly density reachable from core-point **q** if it is in the neighborhood of **q**
- Density reachable*:* **p** is directly reachable from **q** if there are a chain of points that are directly density reachable from **q** to **p**
- Density connected: **p** and **q** are density connected if they are both density reachable from a third point, **o**

- Cluster: choose core point **q**, a cluster **C** contains all points density reachable by **q**
- Noise: any point not in a cluster

Directly density reachable: **p** is directly density reachable from core-point **q** if it is in the neighborhood of **q**

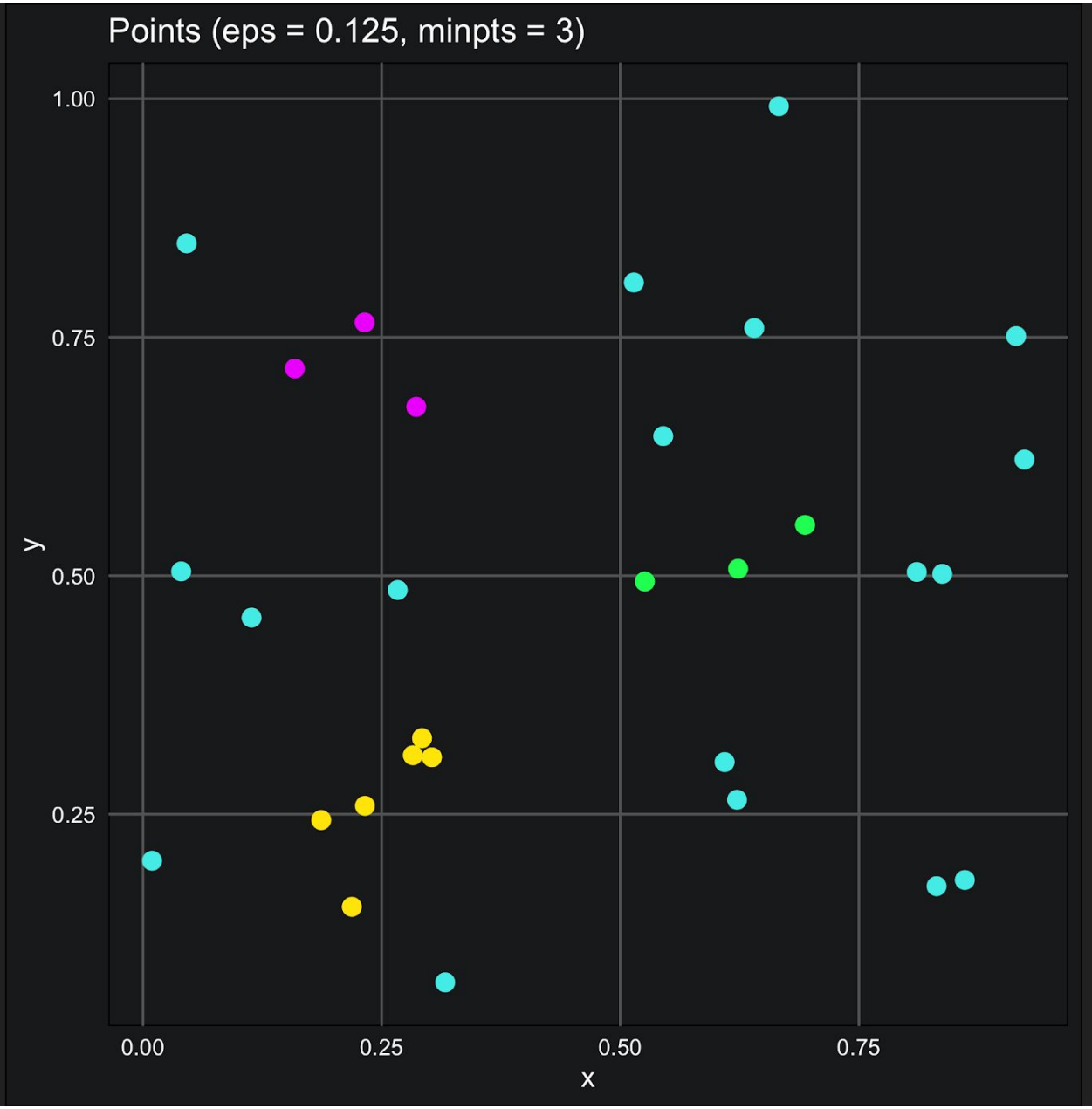


Density reachable*:* **p** is density reachable from **q** if there are a chain of points that are directly density reachable from **q** to **p**

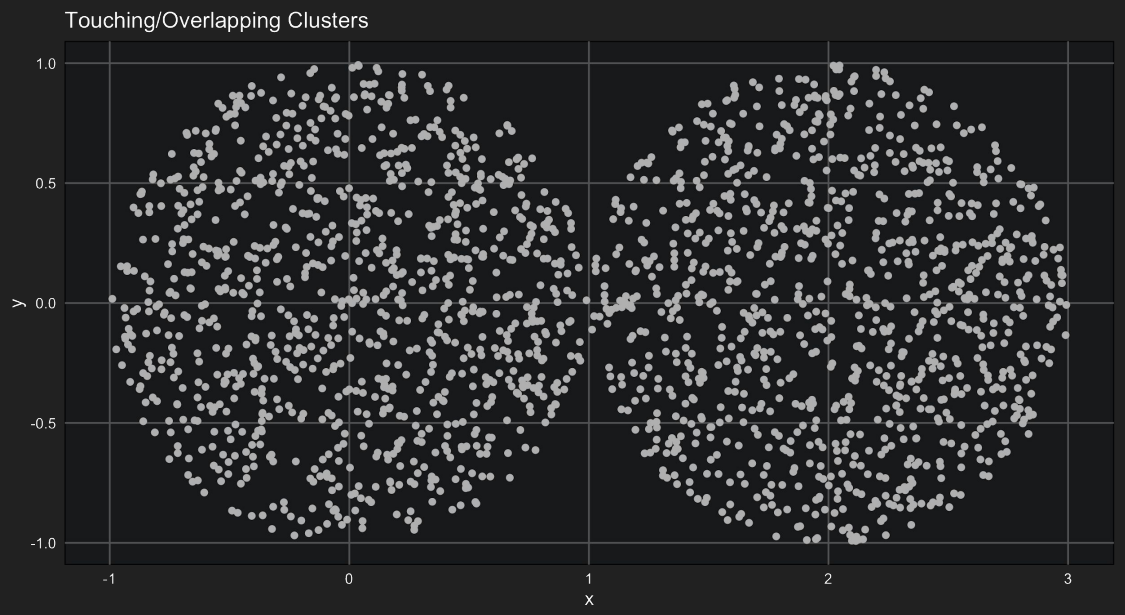


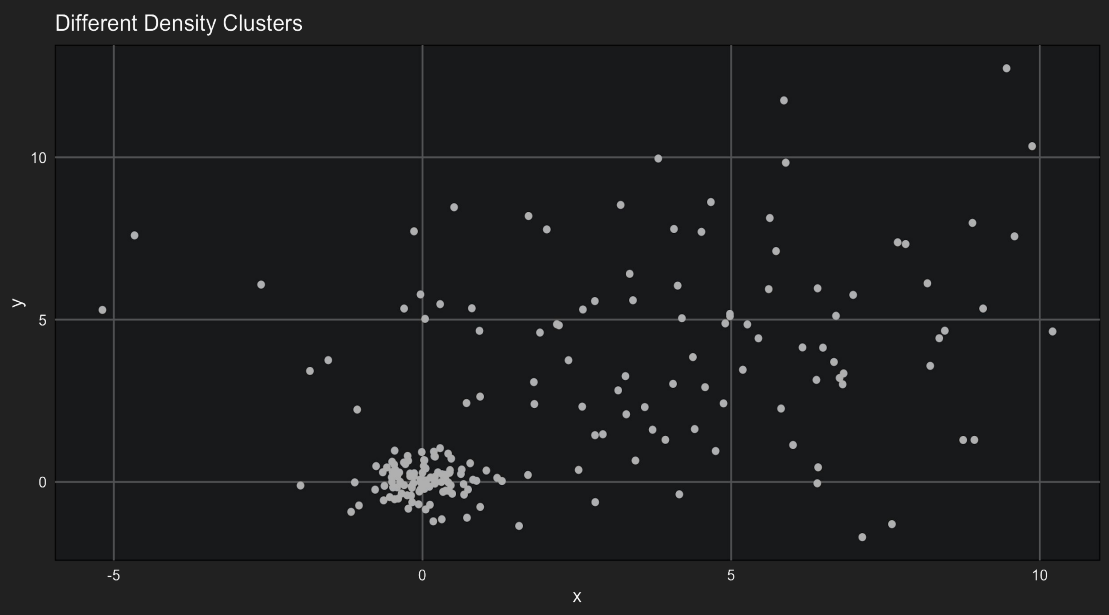
Cluster: choose core point **q**, a cluster **C** contains all points density reachable by **q**

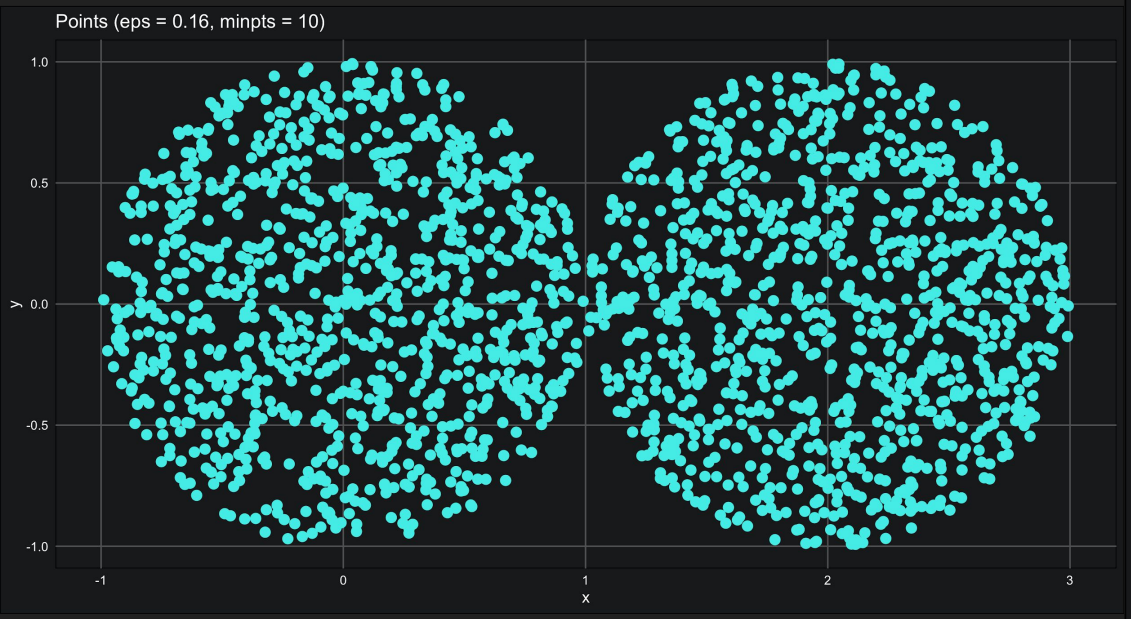
Noise: any point not in a cluster

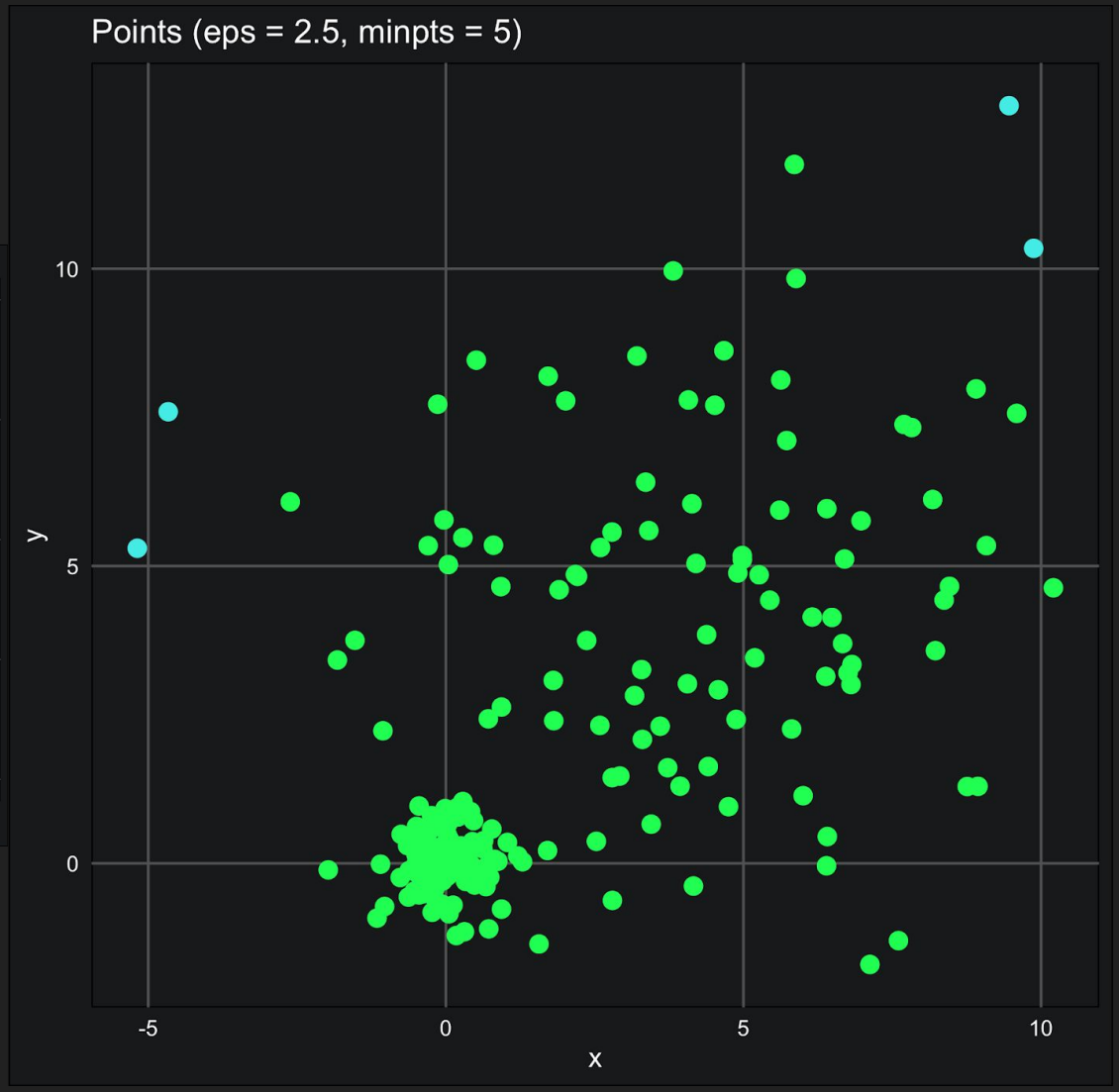


Benefits of DBSCAN


Touching/Overlapping Clusters
1.0
0.5
0.0
-0.5
-1.0
y
-1
0
1
2
3
x

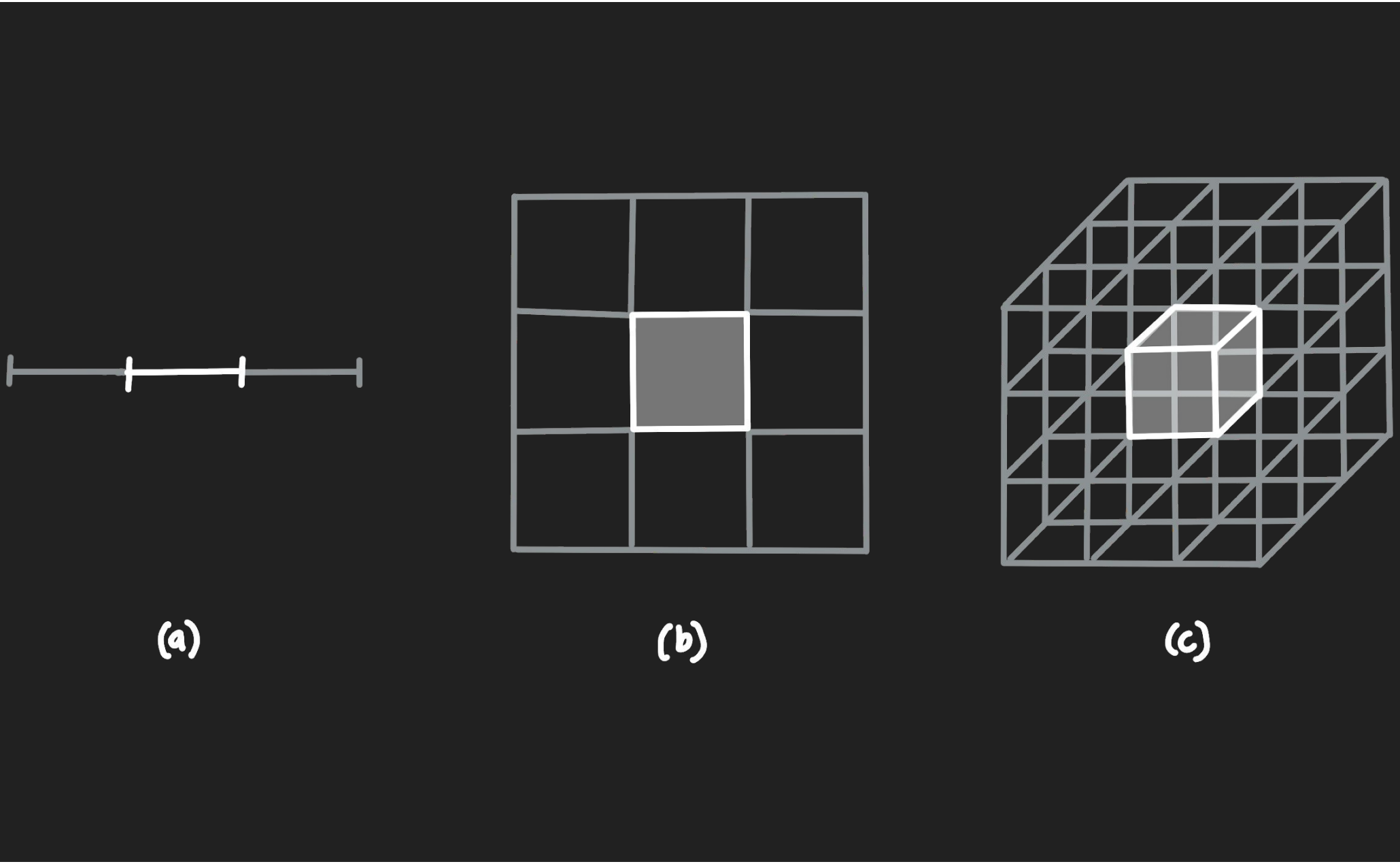

Different Density Clusters
10
5
0
y
-5
0
5
10
x


Points (eps = 0.16, minpts = 10)
1.0
0.5
0.0
-0.5
-1.0
y
-1
0
1
2
3
x


Points (eps = 2.5, minpts = 5)
10
5
0
y
-5
0
5
10
x

Choosing Minimum Points

- Domain Knowledge + Distance Metrics
- More rows = larger min_pts
- More noise = larger min_pts
- More features = larger min_pts


(a)
(b)
(c)

Choosing Epsilon

- Elbow method (k-dist)
- Domain Knowledge