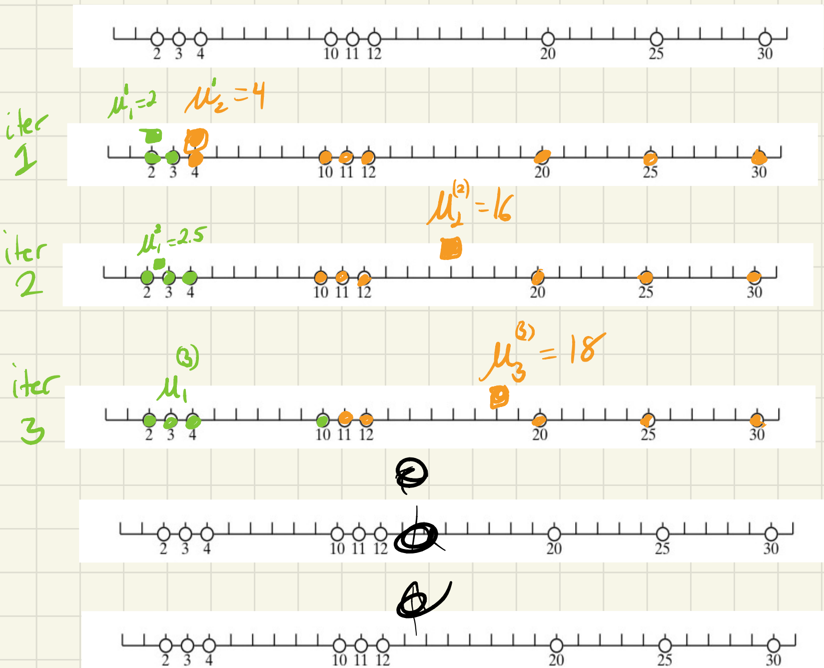
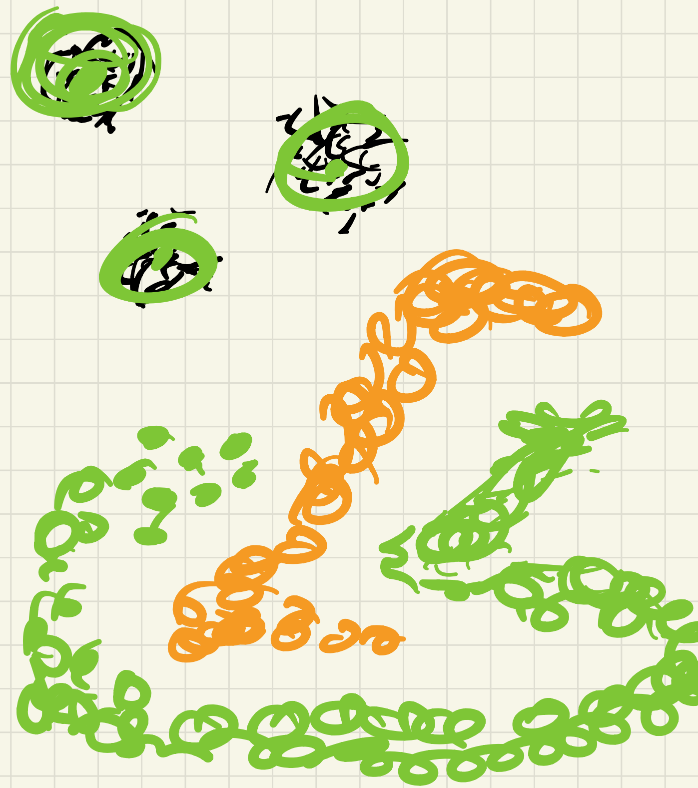
DBSCAN

Clustering So Far

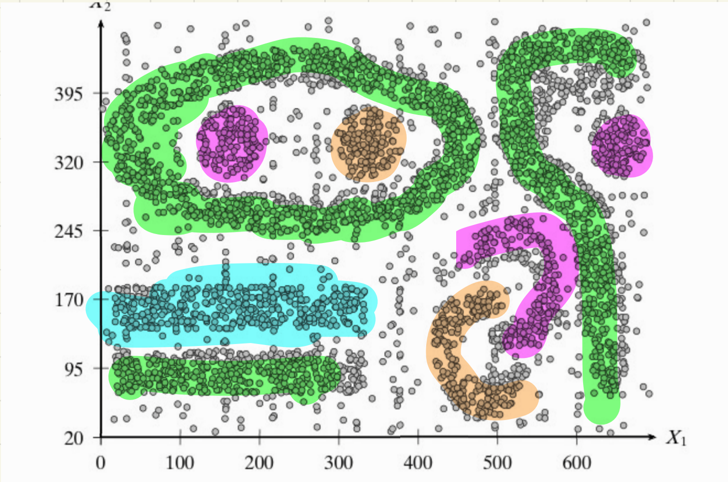
K-means



In 2D



Even More Complex



Soln: DBSCAN

Density

Based

Spatial

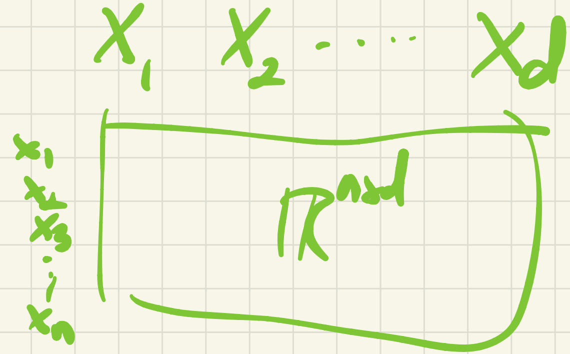
Clustering of

Applications w/

Noise

Defs

Given Dataset D

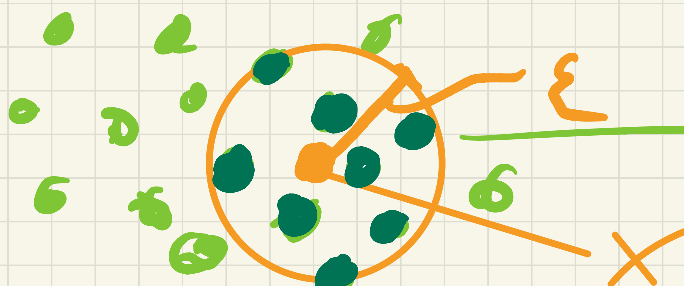


Let $\varepsilon > 0$ $\quad x \in \mathbb{R}^d$

ε-neighborhood of x

$$N_\varepsilon(x) = \{ y \in D : \|y - x\| \le \varepsilon \}$$

all pts of D w/in ε of x



all dark green are in $N_\varepsilon(x)$

Given param minpts $\in \mathbb{Z}^+$

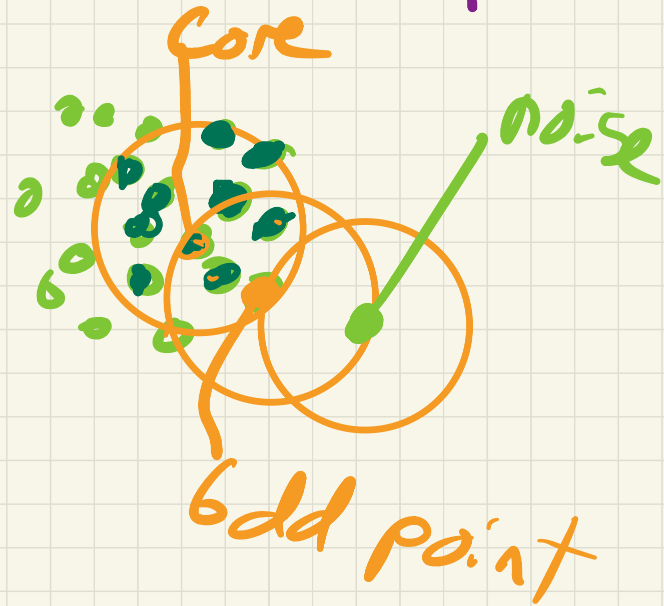
For any $x \in D$

x is a core point if there are at least minpts in its ε-neighborhood

x is a border point if not a core pt and in ε-neighborhood of core pt

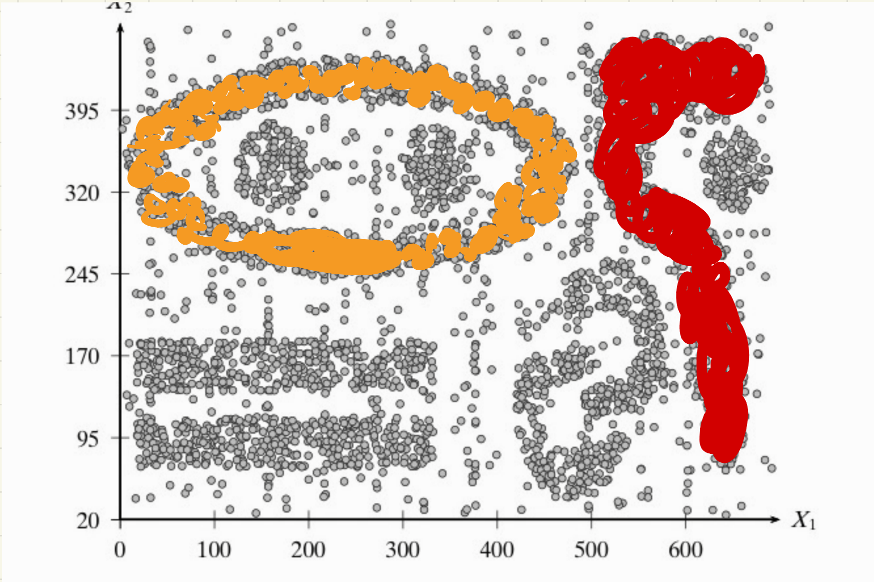
x is noise if !border & !core

Example minpts = 6



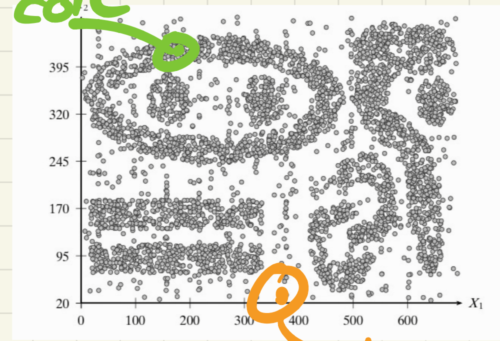
DBSCAN algo (high level)

1. $\forall x \in D$
   - compute $N_\varepsilon(x)$
   - check if core pt
2. $\forall x \in$ Core Points
   - recursively find all "density connected" points



Pseudo Code (Find Corepts)

core



not core

```
Core = ∅
for x ∈ D
    compute N_ε(x)
    id(x) = ∅
    if |N_ε(x)| ≥ minpts
        Core = Core ∪ {x}
```

Compute $N_\varepsilon(x)$

← sets cluster id to $\emptyset$

adds to core if enough points in neighborhood

Pseudocode (find clusters)

$k = 0$

for $x \in Core$ w/ $id(x) = \phi$ — For each unlabel corept

  $k = k + 1$ — inc cluster label

  $id(x) = k$ — set label

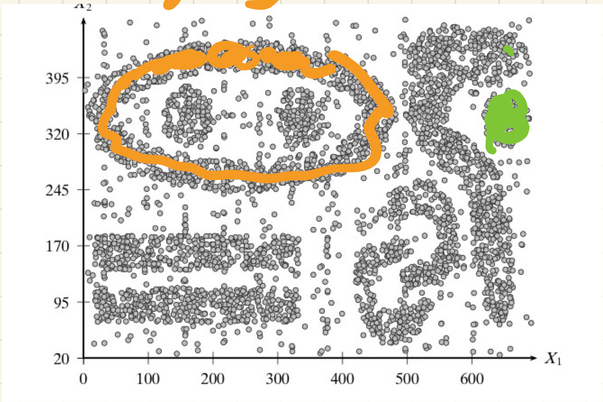
  DesityConnected(x, k) ← search for denity connected neighbors

$C = [id(x_1), id(x_2), \ldots, id(x_n)]$

$Noise = \{x \in D \mid id(x) = \phi\}$

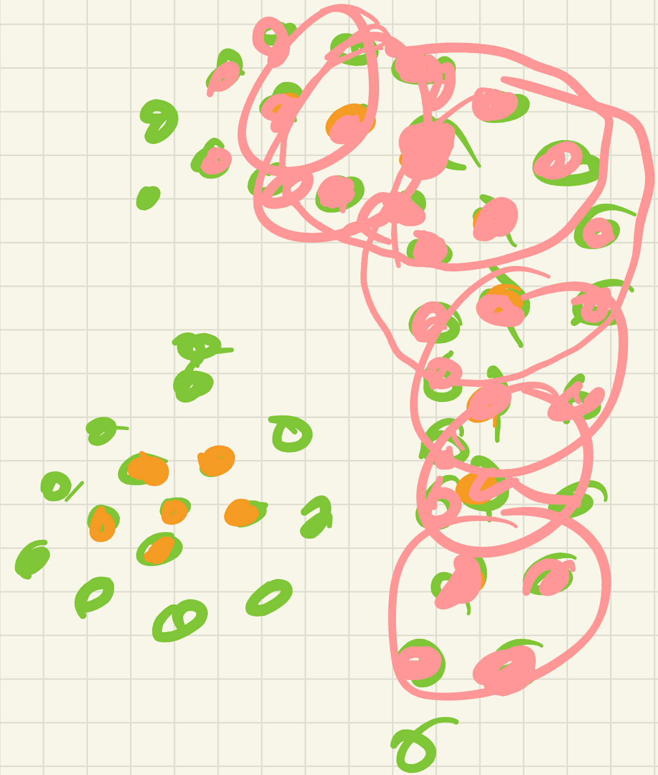
$Border = D \setminus (Core \cup Noise)$

return C, $Core$, $Border$, $Noise$



Pseudocode (Density Connected)

```
DensityConnected(x, k):
    for y ∈ N_ε(x)
        id(y) = k
        if y ∈ Core:
            DensityConnected(y, k)
```



Full Pseudocode

$|D| = n$

Range Query algos

Algorithm 15.1: Density-based Clustering Algorithm

DBSCAN ($\mathbf{D}, \epsilon, minpts$):

$Core \leftarrow \emptyset$

foreach $\mathbf{x}_i \in \mathbf{D}$ **do** `// Find the core points`

  Compute $N_\epsilon(\mathbf{x}_i)$ — $O(n)$

  $id(\mathbf{x}_i) \leftarrow \emptyset$ `// cluster id for` $\mathbf{x}_i$ — $O(1)$

  if $N_\epsilon(\mathbf{x}_i) \geq minpts$ **then** $Core \leftarrow Core \cup \{\mathbf{x}_i\}$ — $O(1)$

$O(n^2)$ if we are not fancy

$k \leftarrow 0$ `// cluster id`

foreach $\mathbf{x}_i \in Core$, *such that* $id(\mathbf{x}_i) = \emptyset$ **do**

  $k \leftarrow k + 1$

  $id(\mathbf{x}_i) \leftarrow k$ `// assign` $\mathbf{x}_i$ `to cluster id` k

  DensityConnected $(\mathbf{x}_i, k)$

$O(n)$ basically Depth first search

$\mathcal{C} \leftarrow \{C_i\}_{i=1}^k$, where $C_i \leftarrow \{\mathbf{x} \in \mathbf{D} \mid id(\mathbf{x}) = i\}$

$Noise \leftarrow \{\mathbf{x} \in \mathbf{D} \mid id(\mathbf{x}) = \emptyset\}$

$Border \leftarrow \mathbf{D} \setminus \{Core \cup Noise\}$

return $\mathcal{C}, Core, Border, Noise$

DensityConnected ($\mathbf{x}, k$):

foreach $\mathbf{y} \in N_\epsilon(\mathbf{x})$ **do**

  $id(\mathbf{y}) \leftarrow k$ `// assign` $\mathbf{y}$ `to cluster id` k

  if $\mathbf{y} \in Core$ **then** DensityConnected $(\mathbf{y}, k)$