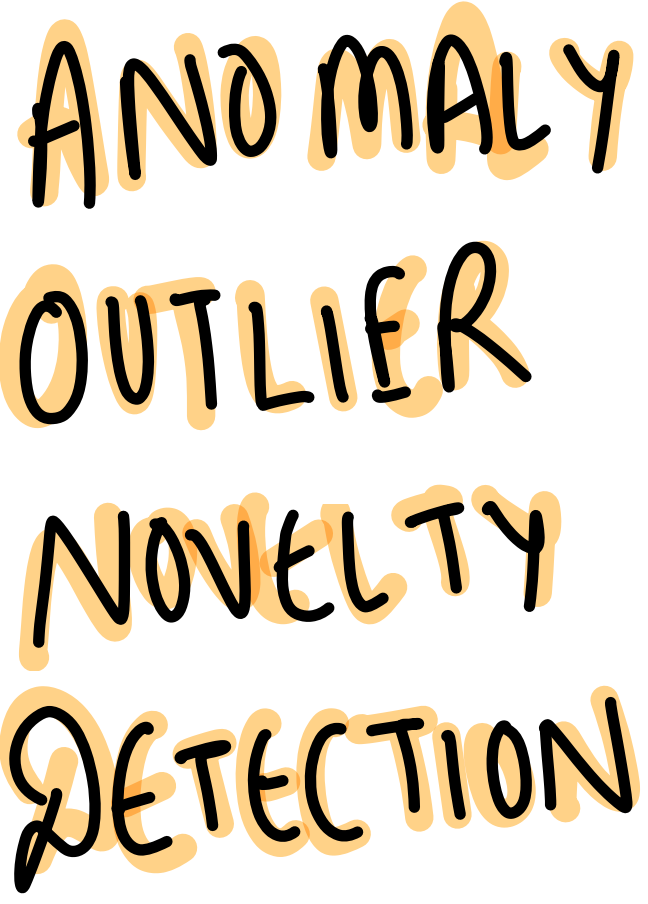

ANOMALY
OUTLIER
NOVELTY
DETECTION

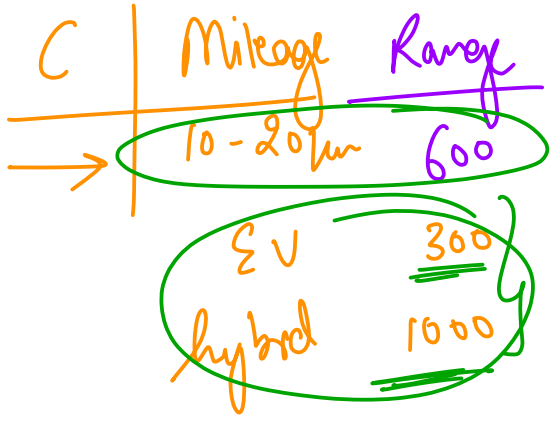
Outliers

Q Why do outliers exists?

* Human error
* Sensor error / faulty machine
* "Unusual" data
  - → Anomaly ⇒ Not Normal
  - → Novelty ⇒ New, Never happened before.

eg. * Fraud detection -

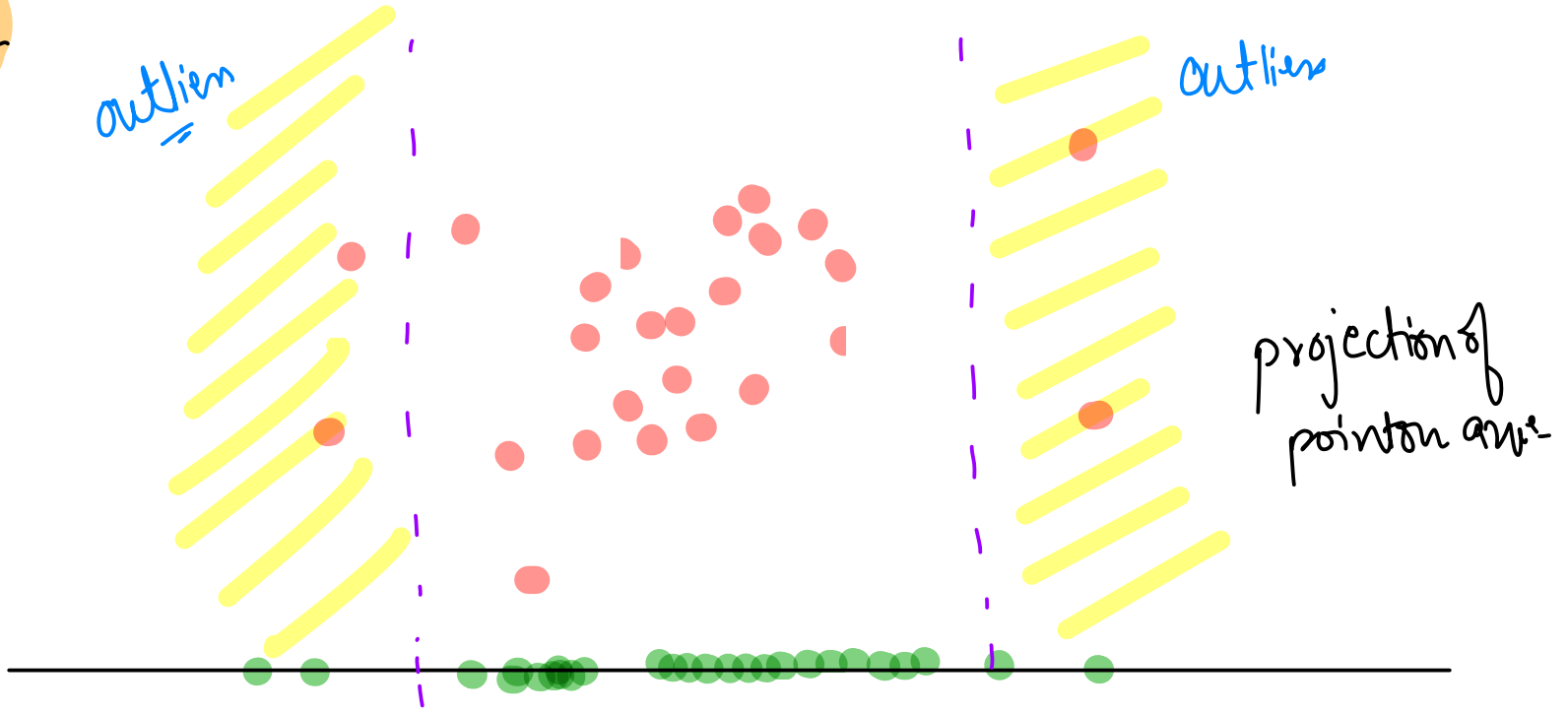
* Car Mileage EV, hybrid.



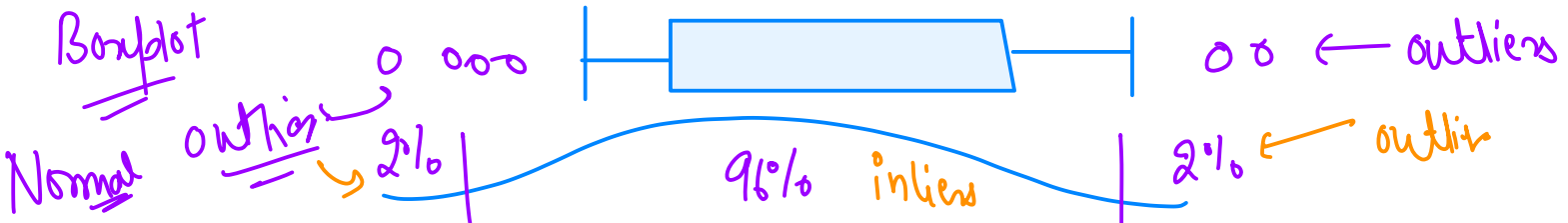
Q What all some ways to detect outlier?

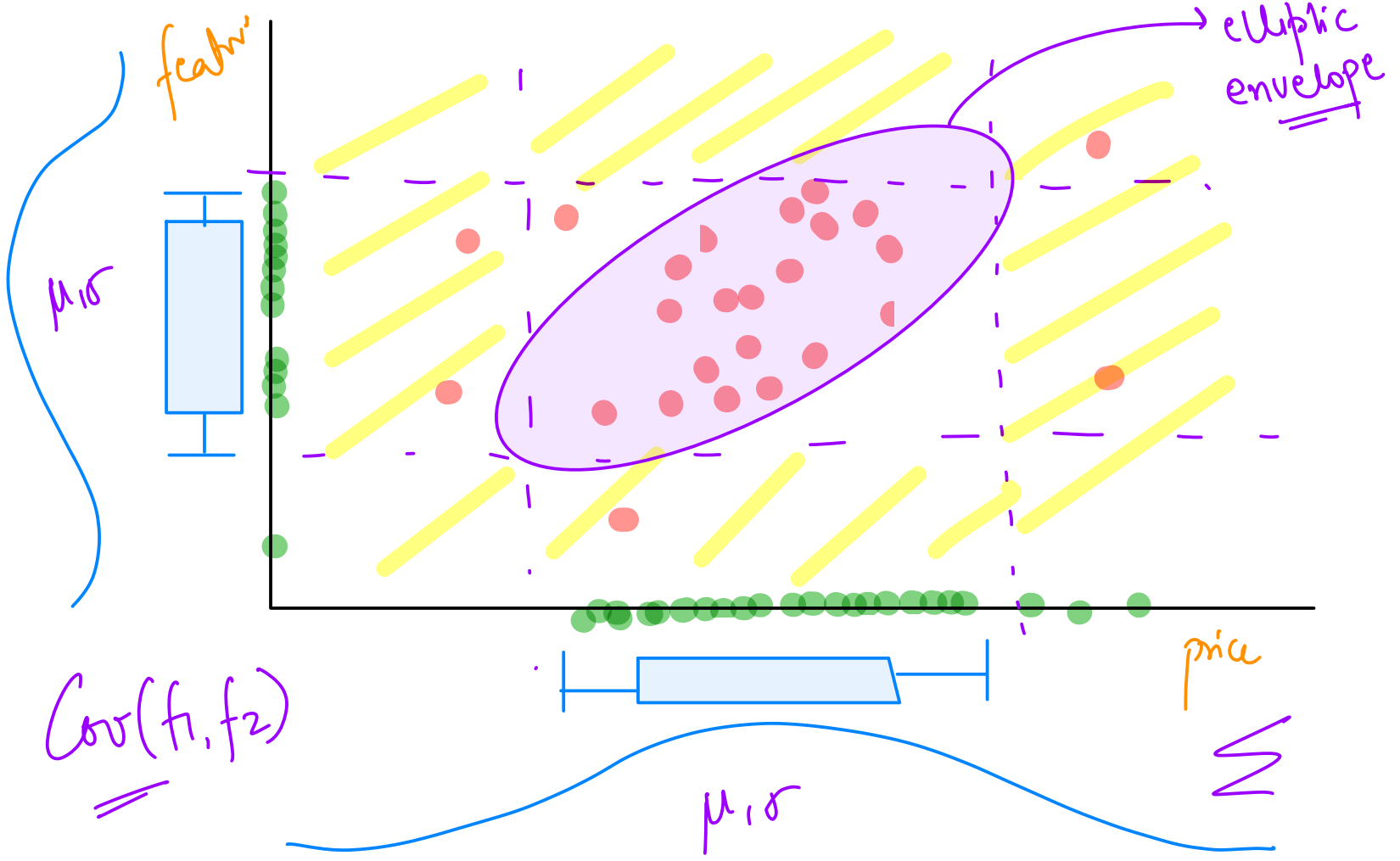
A IQR, KNN → DBSCAN

11


outliers
outliers
projection of
points on x-

Boxplot


outliers
outliers
Normal
2%
95% inliers
2%
outlier


feature
elliptic envelope
price
Cov(f1, f2)

1. IQR (BOXPLOT)

2. Multi Variate Gaussians
   - Elliptical Envelope..

$$\Sigma = \begin{array}{c} \\ f_1 \\ f_2 \\ f_3 \end{array} \begin{array}{c} \begin{array}{ccc} f_1 & f_2 & f_3 \end{array} \\ \begin{bmatrix} - & - & - \\ - & - & - \\ - & - & - \end{bmatrix} \end{array}$$

learn $\{\mu, \sigma, \Sigma\}$ $\Rightarrow$

estimate

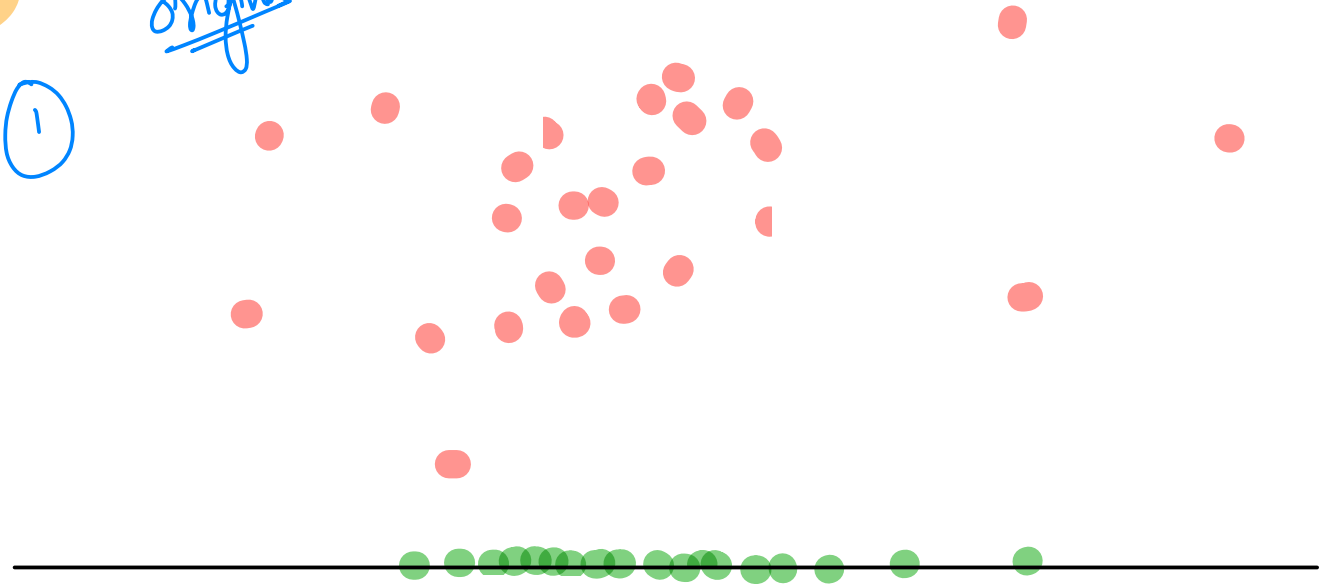
RANSAC

Random Sample Consensus

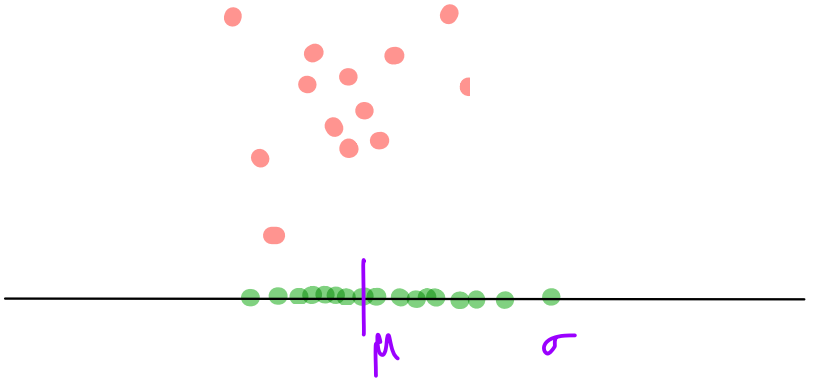
original

1

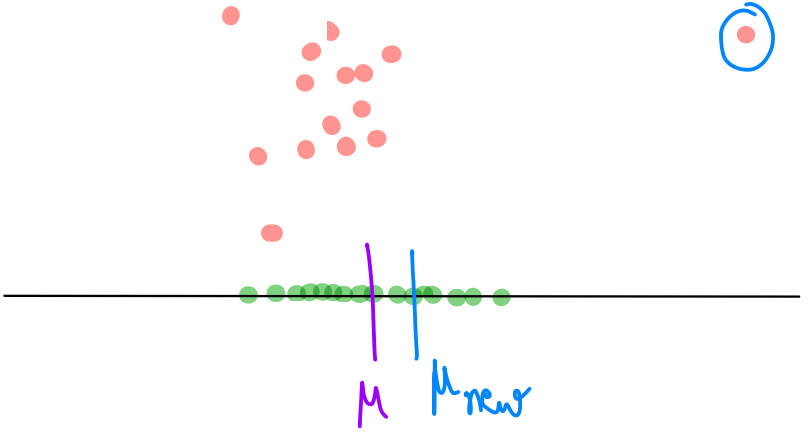
Robust



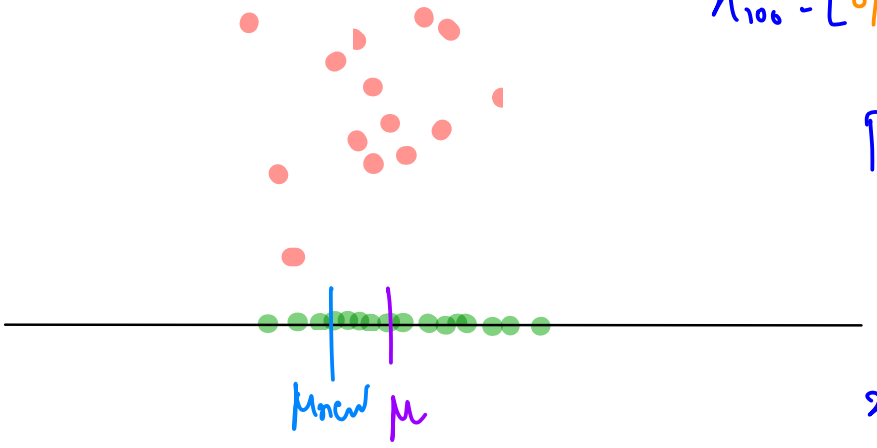
(2)



(3)



3



$x_{100} = [0/1/2 \; 0/1/2 \; 0/1/2 \; 0/1/2 \; 0/1/2]$

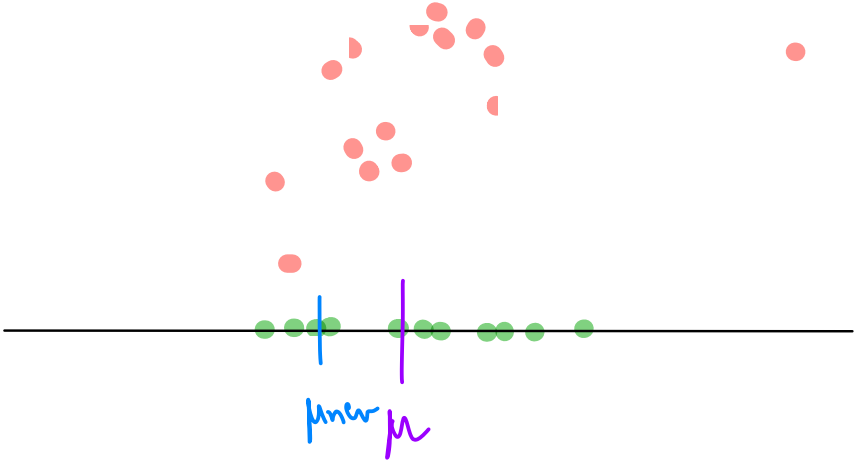
present not outlier : 1

present outlier : 2

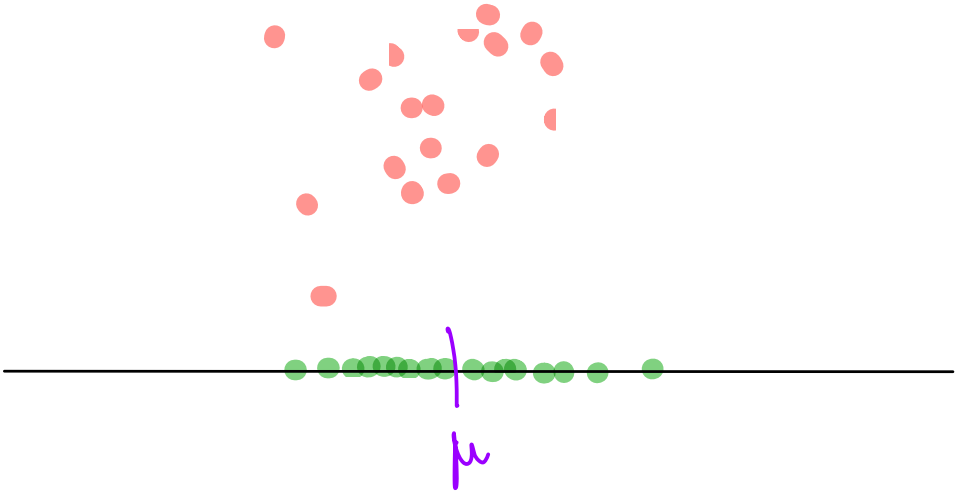
absent : 0

x_{100} [1 2 2 0 0]

4



(5)



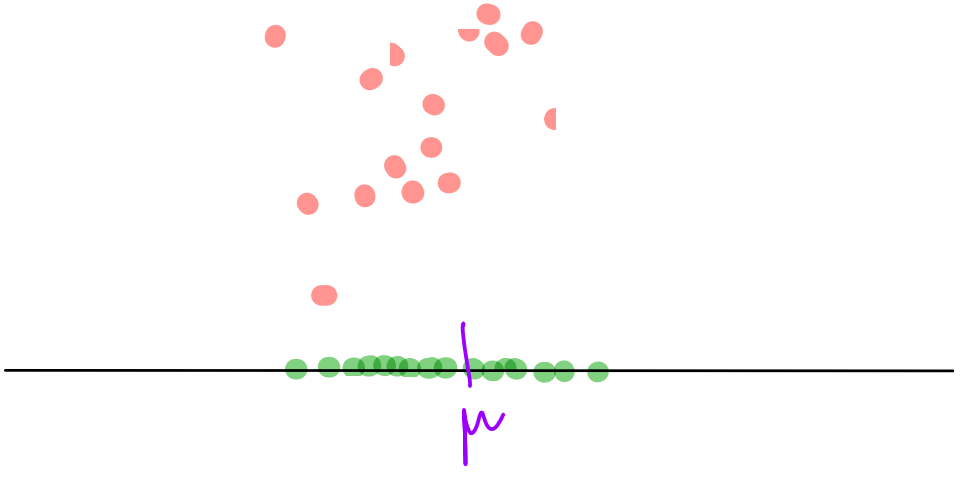
100 times

90% μ

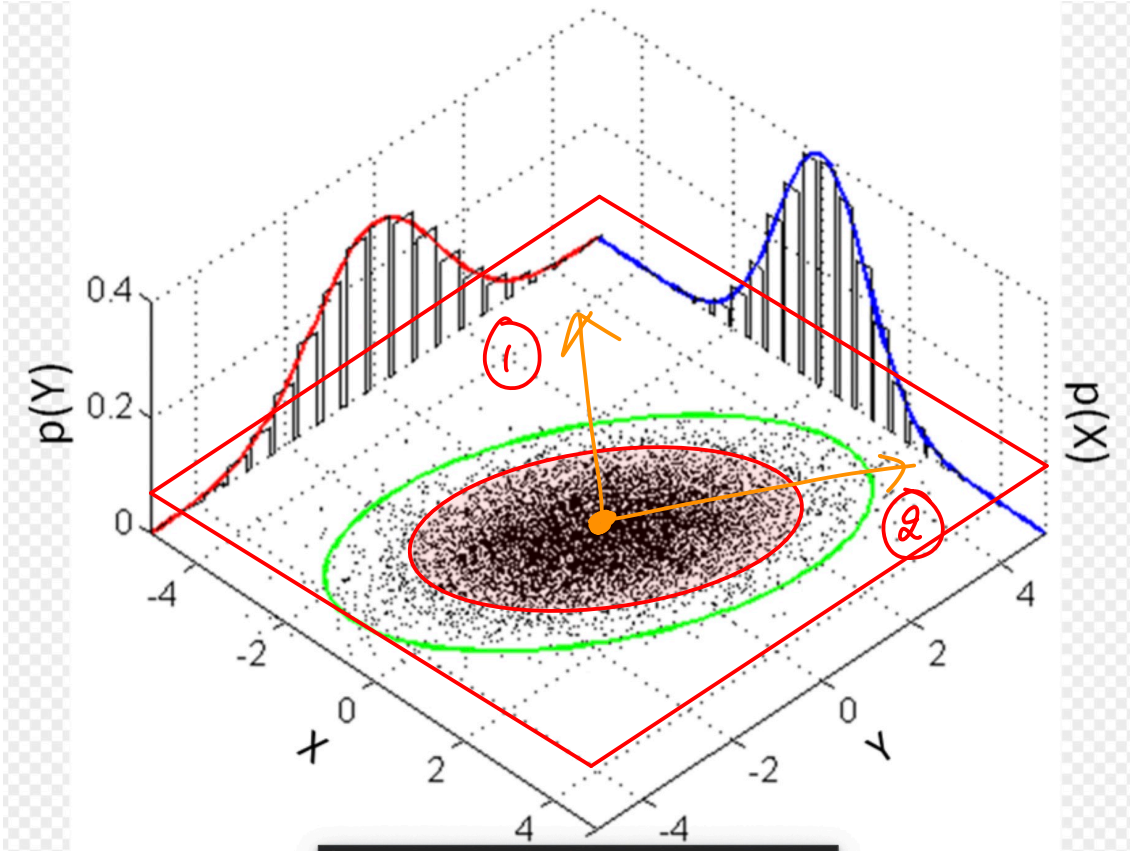
10% μ + δ

μ − δ

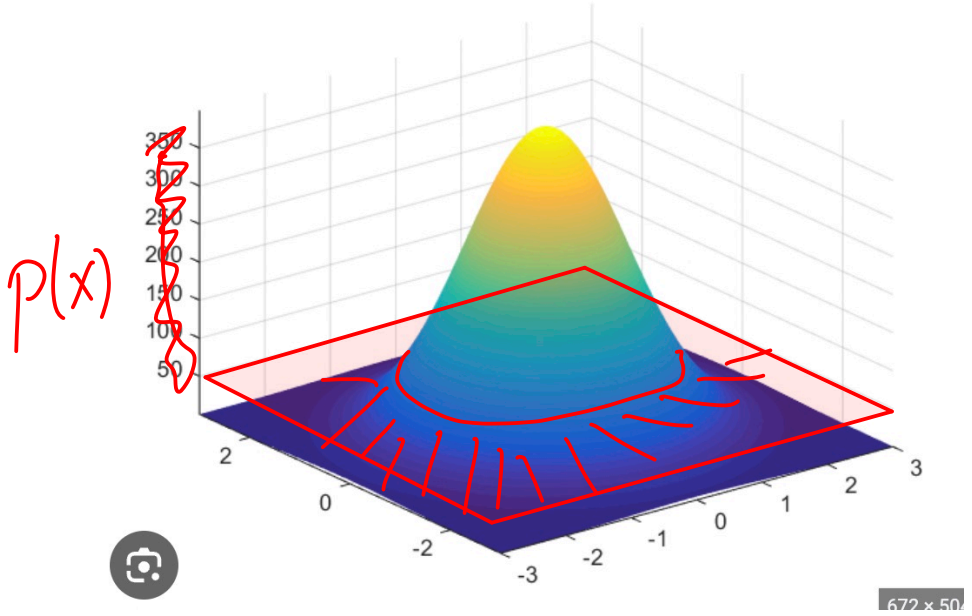
(6)

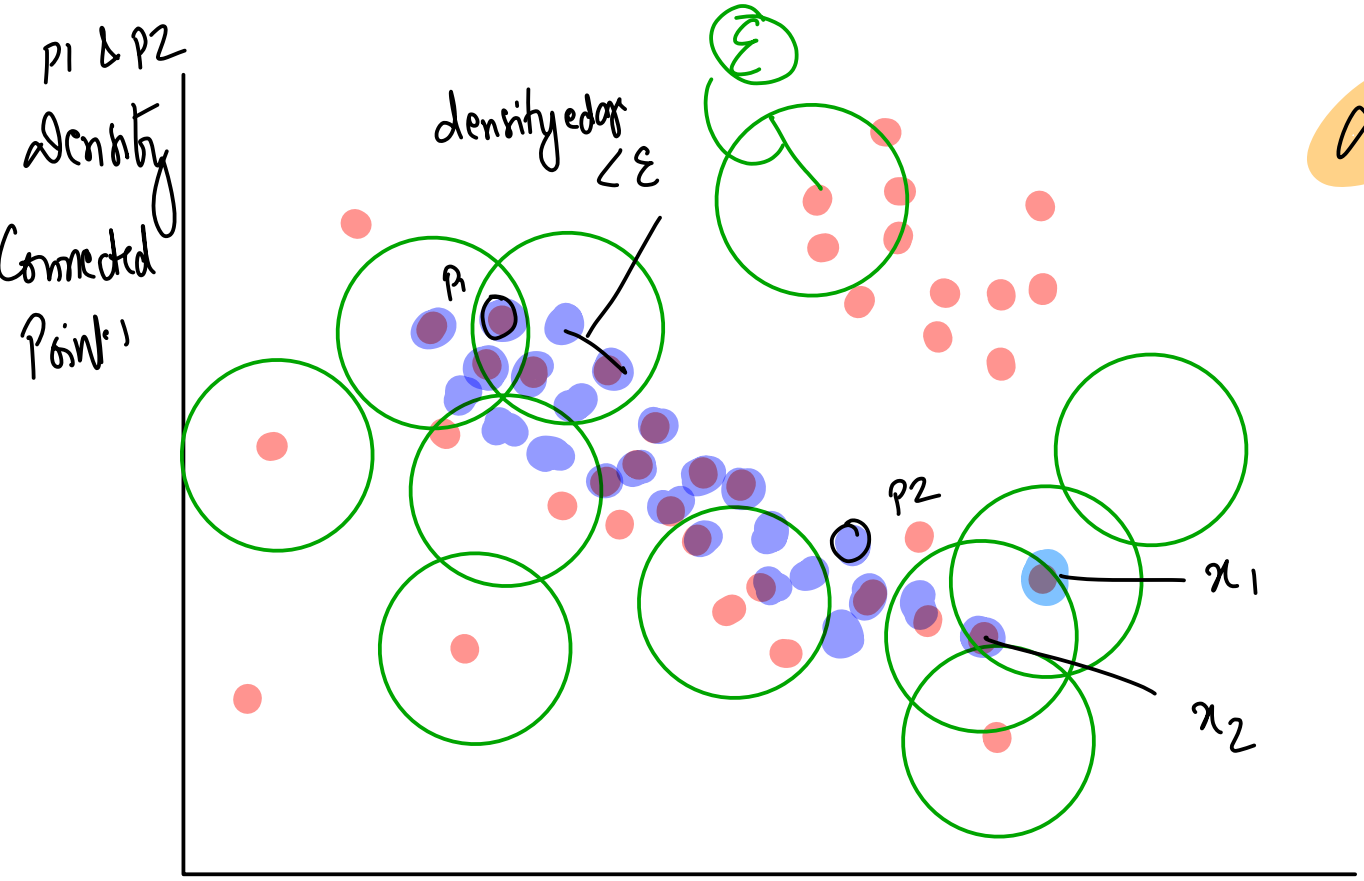


Avg ⇒ μ



$p(x) < 0.2$


p(x)
350
300
250
200
150
100
50
2
0
-2
-3
-2
-1
0
1
2
3



DBSCAN

Density Based Spatial Clustering Application with noise

1. Core point
2. Border point
3. Noise point

ε = Radius : Hyper parameter

min # of points : n
↳ 3

For ε Radius

⇒ We will check

if # of points inside circle > n

↳ Core point

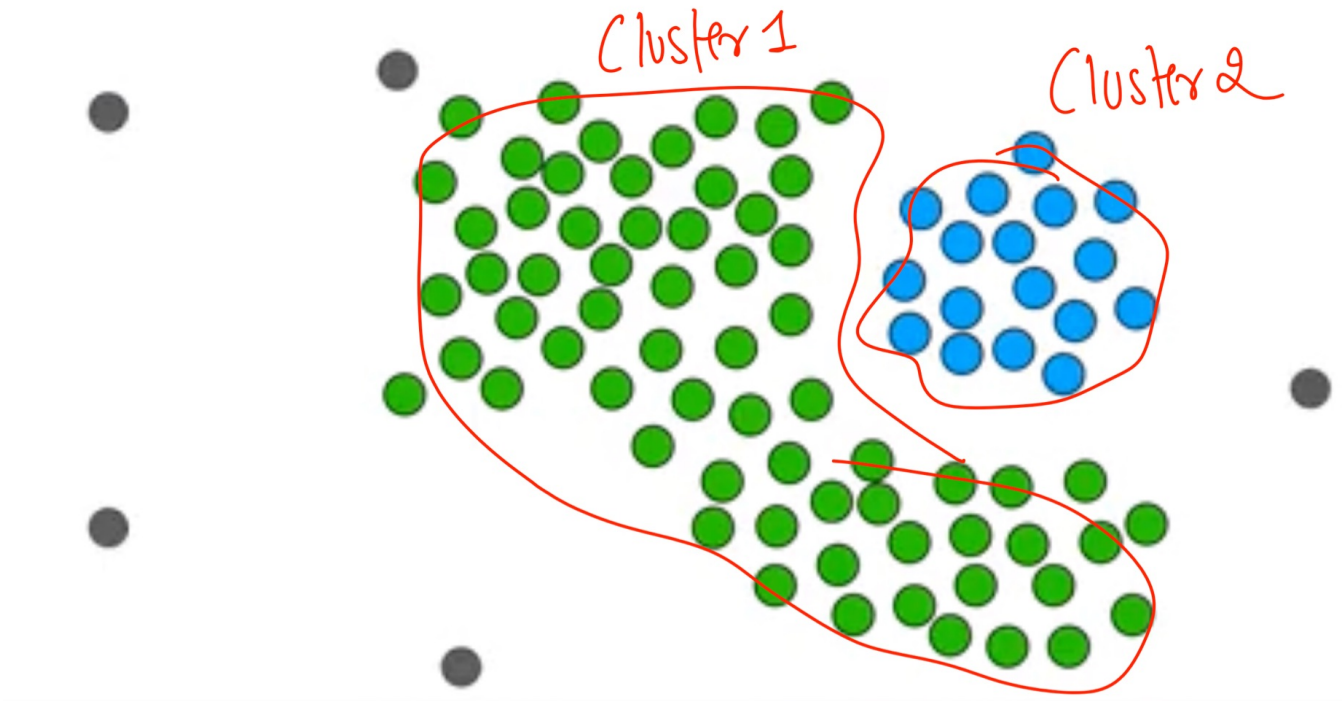
else.

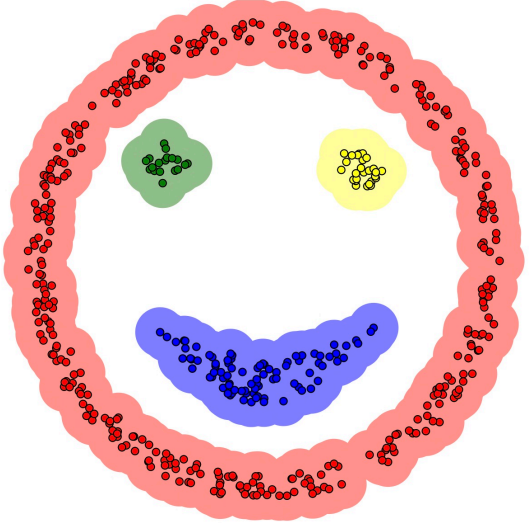
Non Core point → Border point

→ Noise point.

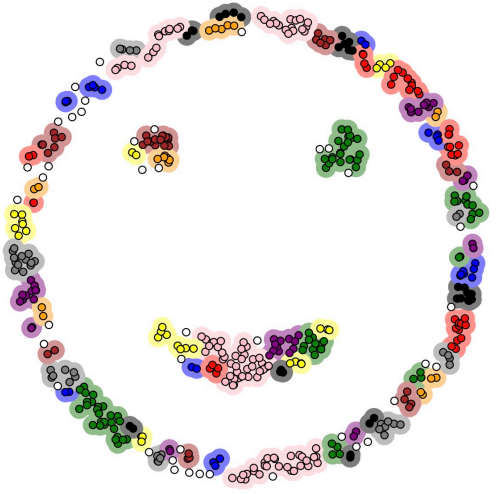
Border point : if a Non Core point lies in the Circle of any Core point, it is a Border point

else: Noise point.

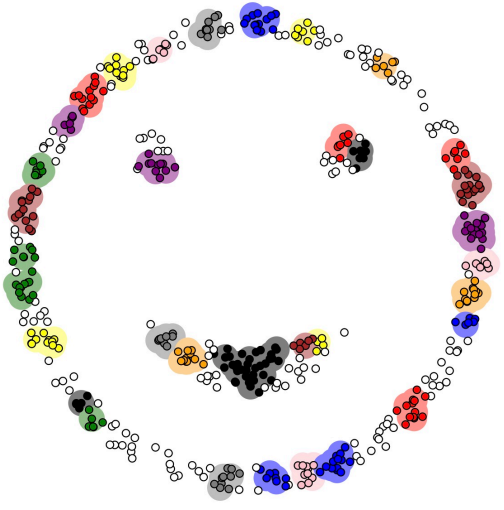

Cluster 1
Cluster 2



epsilon = 1.00
minPoints = 2



epsilon = 0.38
minPoints = 2



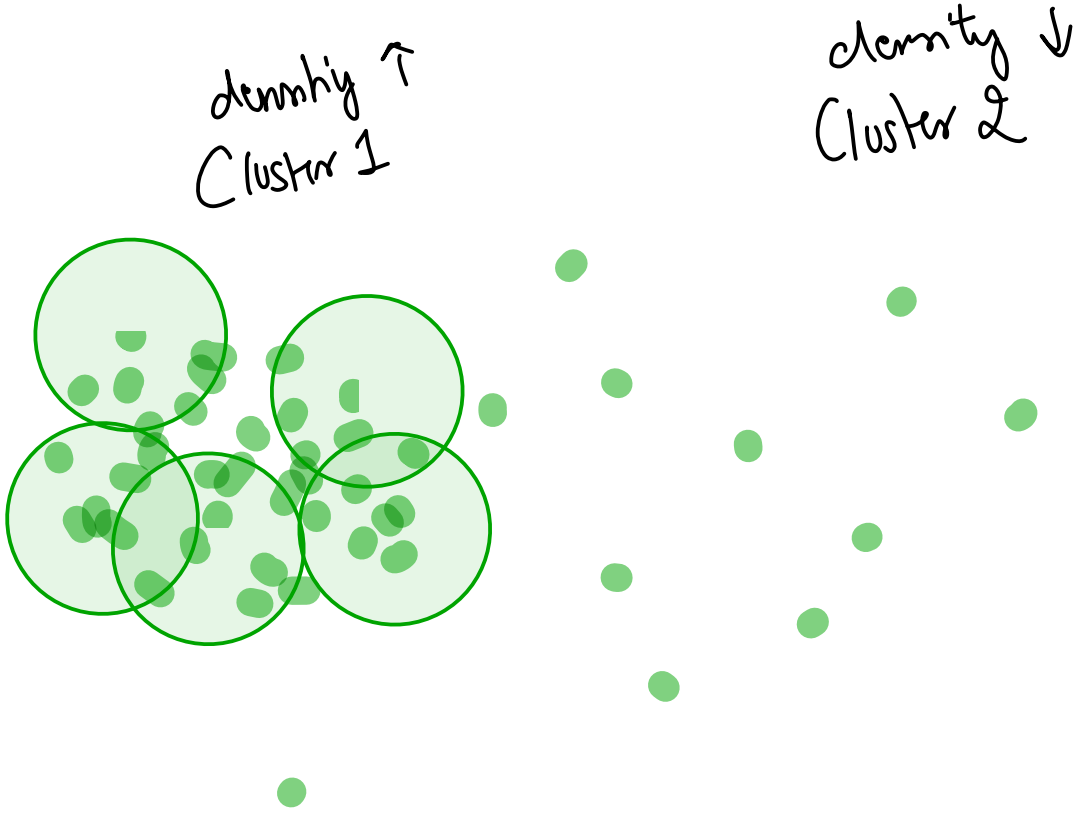
epsilon = 0.50
minPoints = 6

Pro:
1. finds out Outliers & clusters as well
2. No need to decide 'K'

Cons.
1. Very Sensitive to hyperparameter.
2. Doesn't work well with sparse data / data with diff densities.

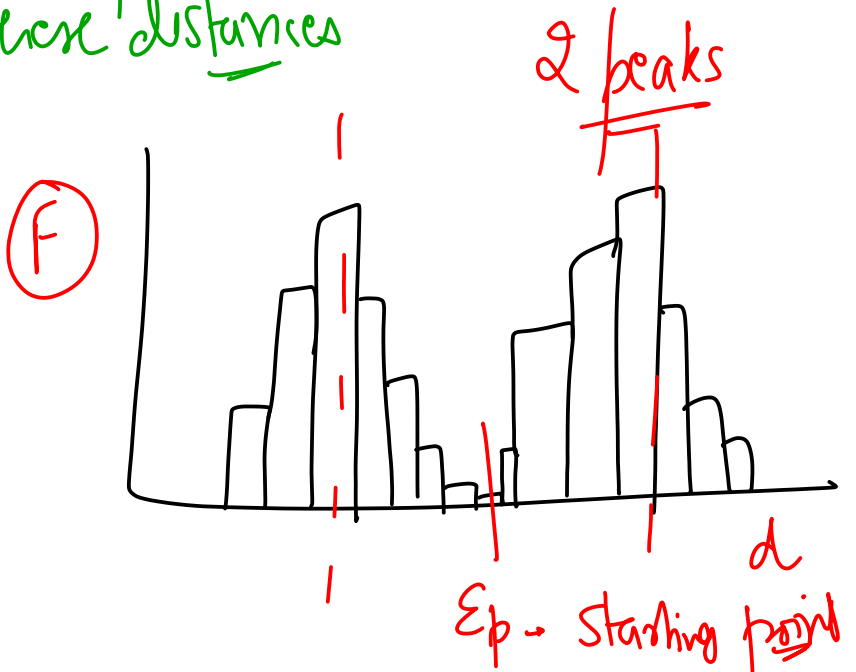
RF
↳ 100
↳ 120

DBSCAN
0.1 → 0.11


density ↑
Cluster 1
density ↓
Cluster 2

How to estimate good value of ϵ ?

1. Calc. distance for each point.
2. Plot histogram of these distances



How to estimate min # of points ??

* d dimensional
  ↳ $n \geq d+1$

* d-dimensional
  ↳ $n \geq d \times 2$

Quiz time!

Quiz Ended!

What is a core point in DBSCAN?

5 users have participated

A	A point located at the center of a cluster	0%
B	A point with the highest density in the dataset	0%
✓ C	A point that has more than Min Points within the Epsilon radius	100%
D	A point that is not part of any cluster	0%

Quiz Ended!

What is a "density edge" in DBSCAN?

6 users have participated

A	A line connecting two random points in a dataset	0%
✓ B	An edge that connects two core points with a distance less than or equal to ε (Epsilon)	100%
C	The edge of a dense cluster in the dataset	0%
D	An edge between two noise points	0%

Quiz time!

TIME LEFT: 13 Secs

How is a border point defined in DBSCAN?

6 users have participated

A	A point located on the border of the dataset	0%
B	A point with the highest density in a cluster	0%
C	A point that is not part of any cluster	0%
✓ D	A point that is not a core point but is within the Epsilon radius of a core point	100%

When are two points considered "density connected" in DBSCAN?

6 users have participated

A	When they are connected by an edge	0%
B	When they are both noise points in the dataset	0%
✓ C	When they are core points and there exists a sequence of density edges connecting them	100%
D	When they are part of the same cluster	0%

End Quiz Now