

Outlier Detection

[Unsupervised Learning]

→ Types

→ Elliptical Envelope

→ Isolation Forest

→ Local Outlier factor.

Outliers

Q: Why do outliers exist?

→ Human / Sensor error

→ Read "unusual" data

→ Anomaly → Not "Normal"

exnola

↳ Novelty → Never happened before

Eg:

- Fraudulent transactions exnola
- Sales during covid → upi transactions ISRs.
- Electric cars → mileage data

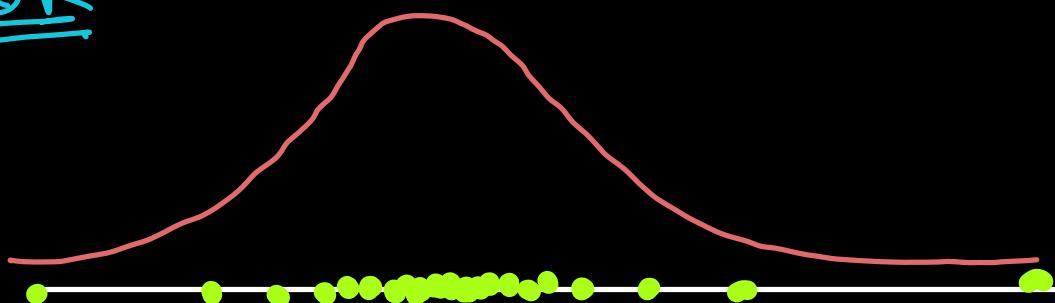
Q: What are some ways to detect outliers?

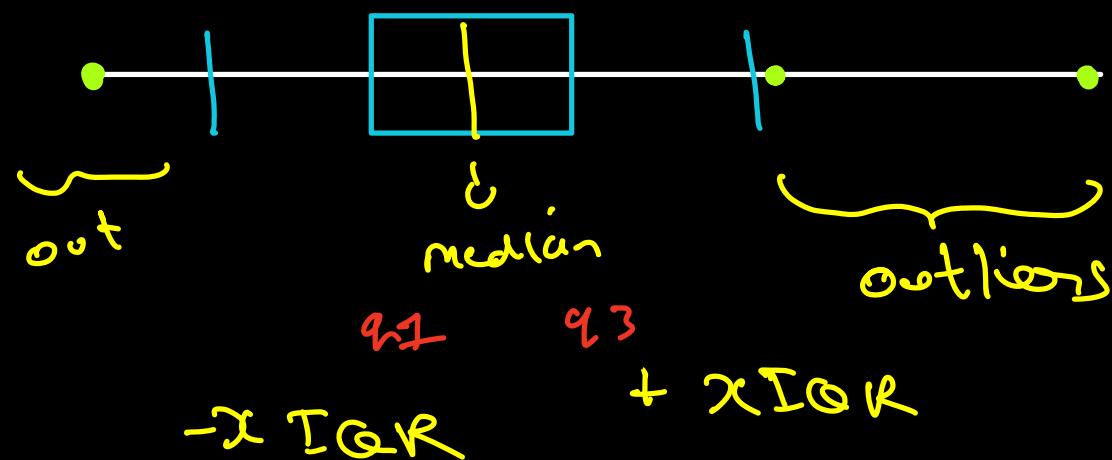
\rightarrow IQR \rightarrow KNN \rightarrow DBSCAN, etc.
Review

Elliptical Envelope

Let's build better techniques using above ideas.

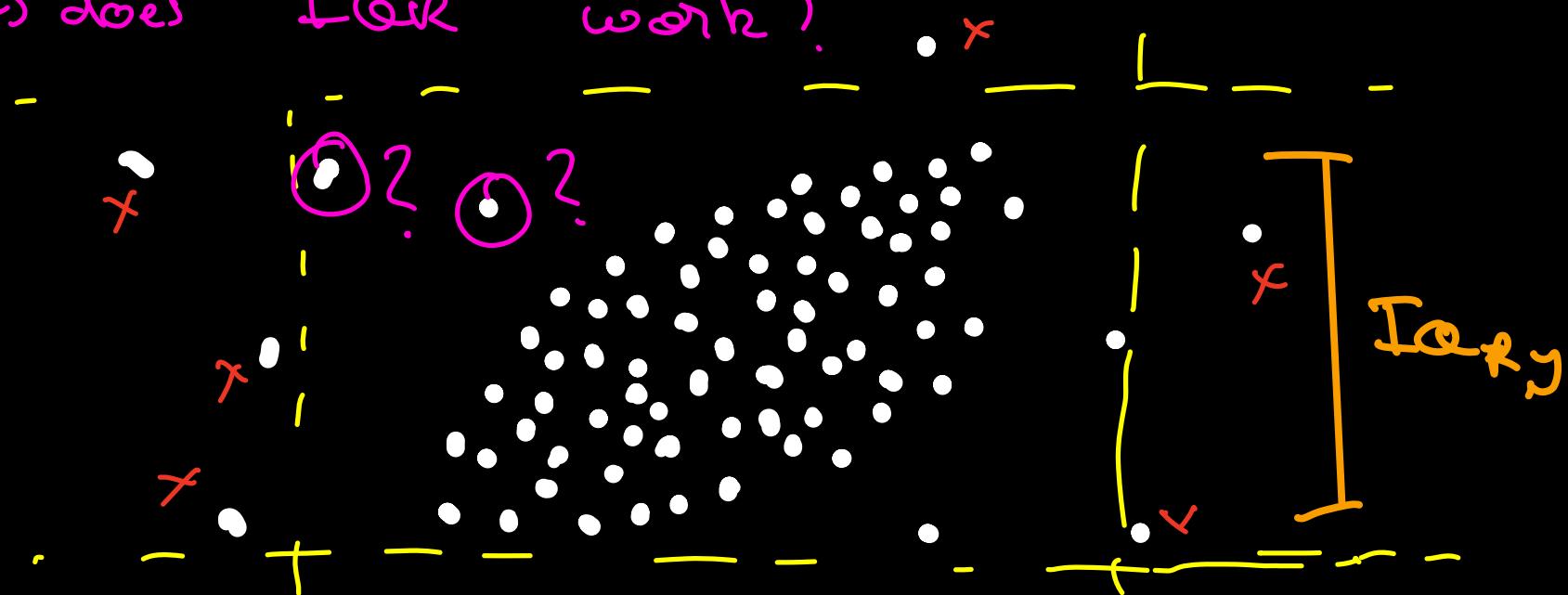
IQR





But what abt $2/n$ dimensions?

→ does IQR work?

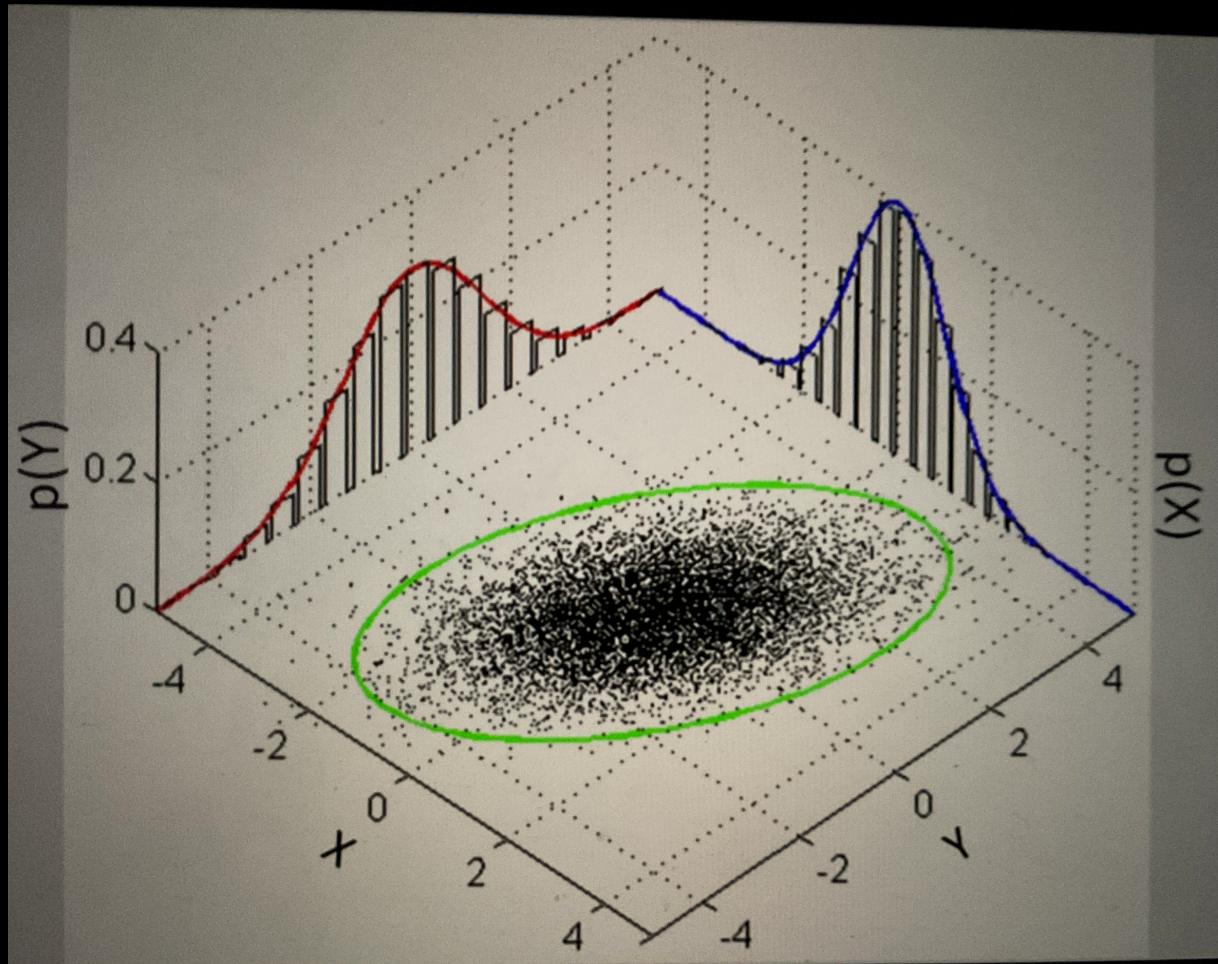




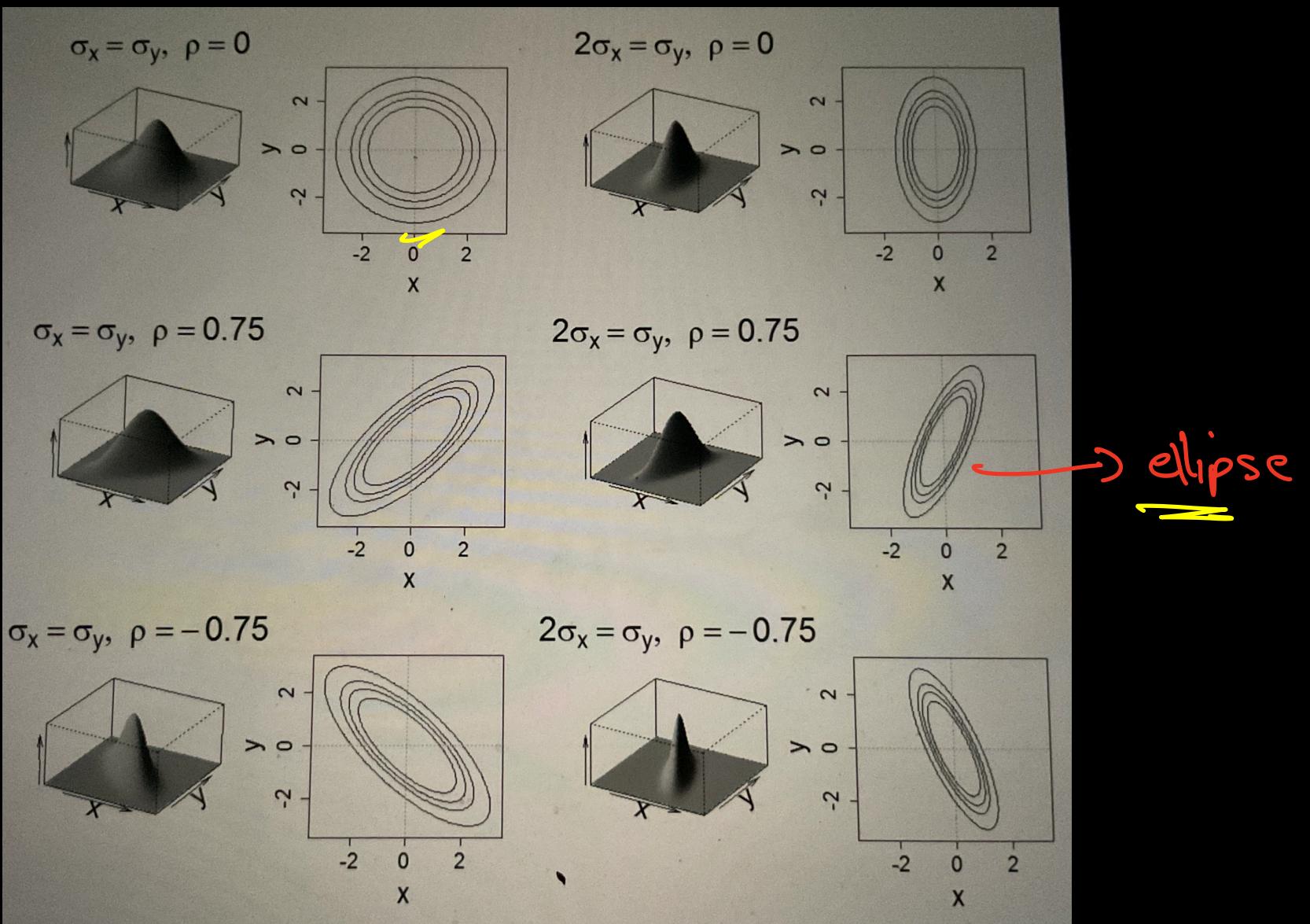
→ N -dimensional IGR will form cuboidal boundaries, whereas most high dimensional data may fit "globular" boundaries.

→ Any solutions?

↳ Multivariate gaussians!



→ wie
pdf
bij voorke



So we can fit multivariate gaussian to the data and points with low prob can be called outliers.

→ Assumptions : points follow gaussian distribution

How do we learn the parameters?

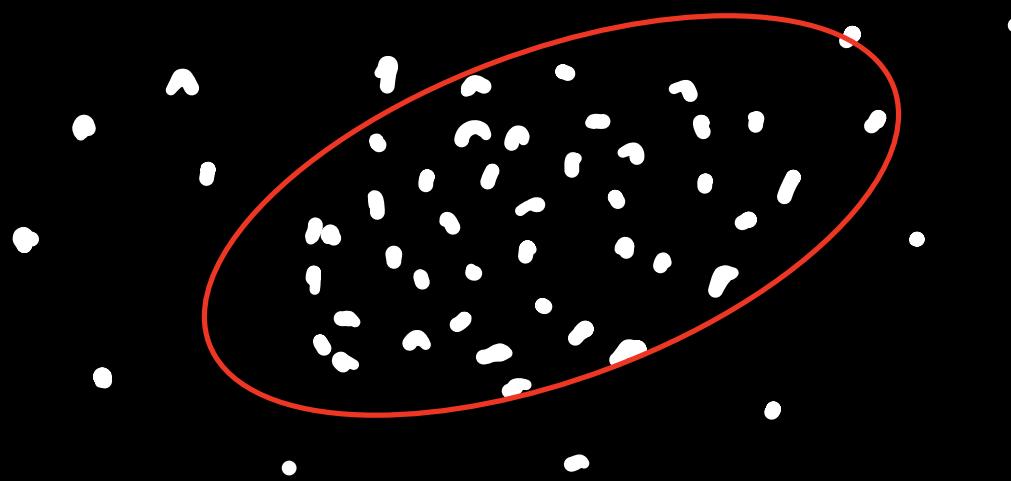
↳ μ, σ, Σ will be affected by outliers.



For this we can use robust angles ↴

RANSAC

Random sample consensus



→ Randomly sample some points.

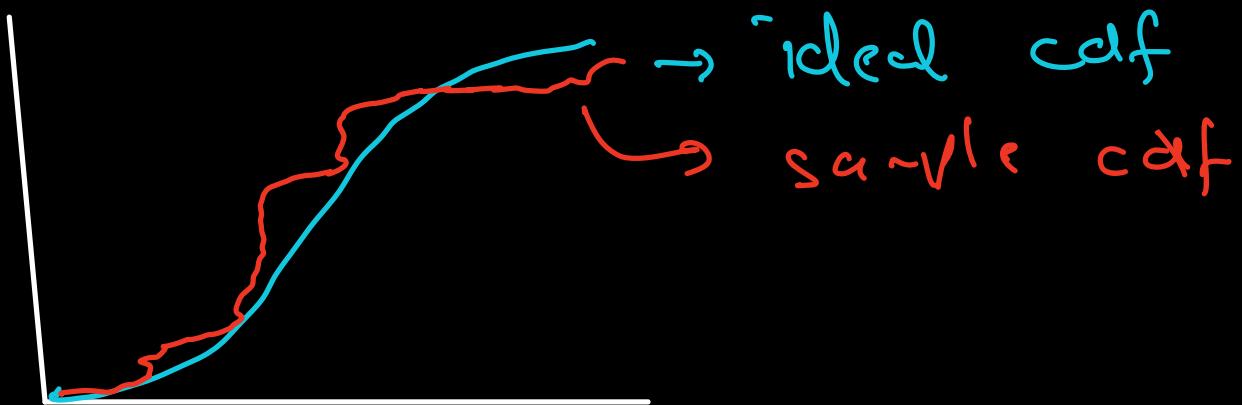
Find μ, σ, ρ of this sample.

Now compute ideal distribution

blw μ, σ, ρ

→ Compare ideal dist with actual sample.

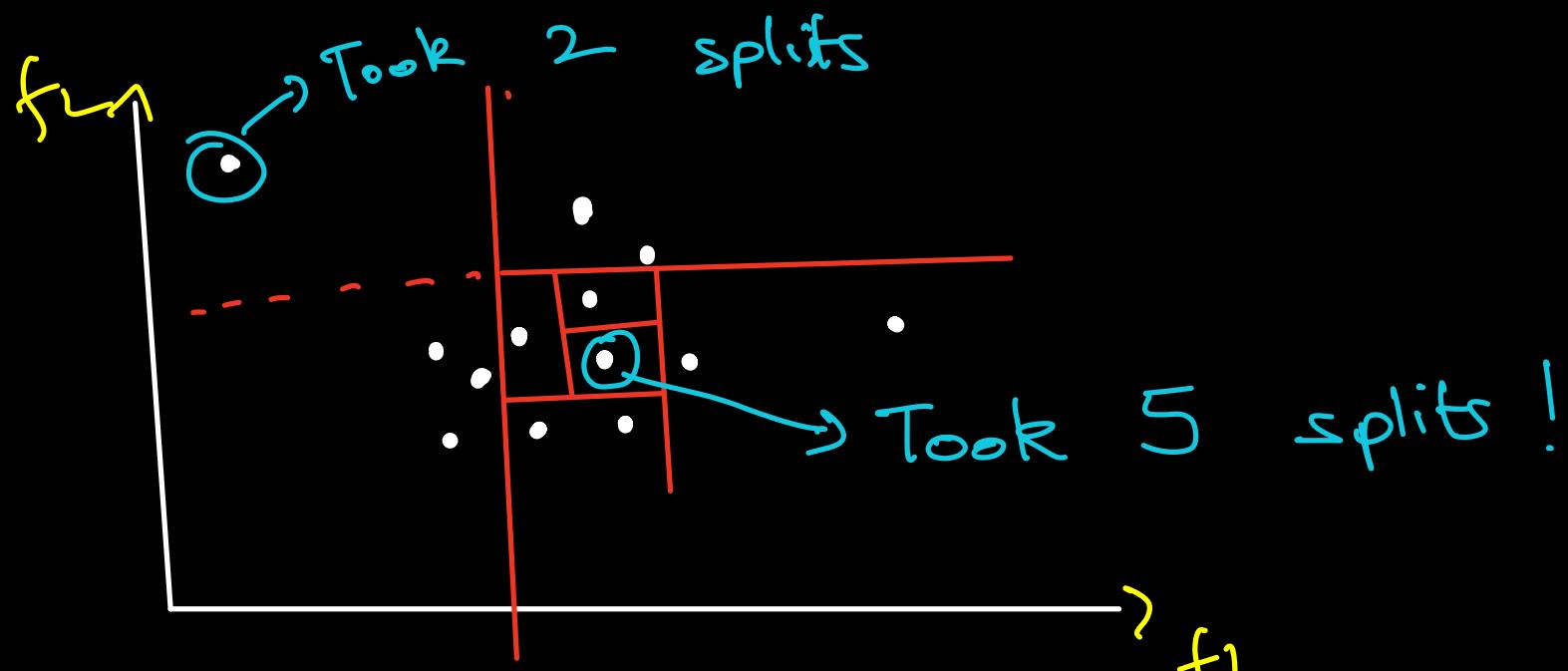
→ Choose the parameters that fit most of the samples / best fits one of the sample / avg, etc.



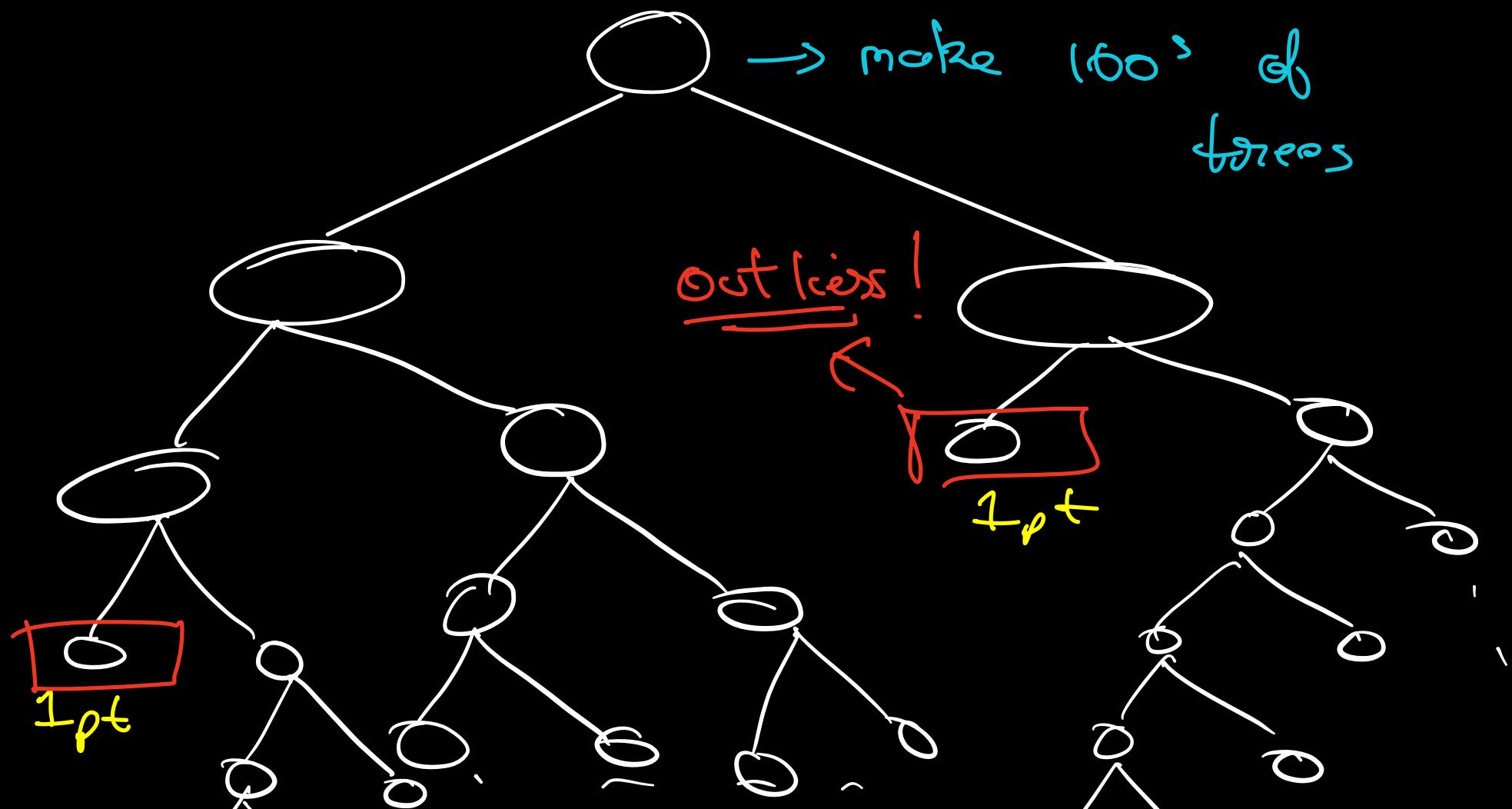
Isolation Forests

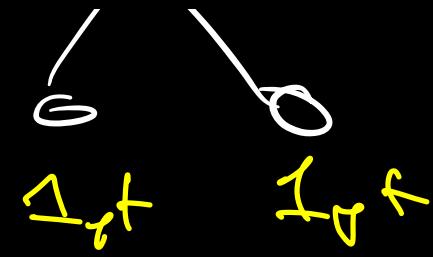
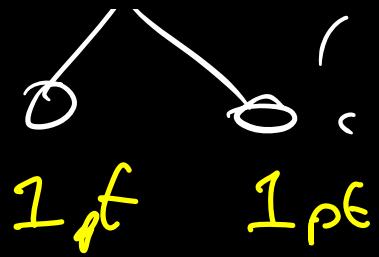
Idea !

↳ Build a lot of trees (random)
until each point is a separate
node .



⇒ On avg outliers will take less
splits to be separated.



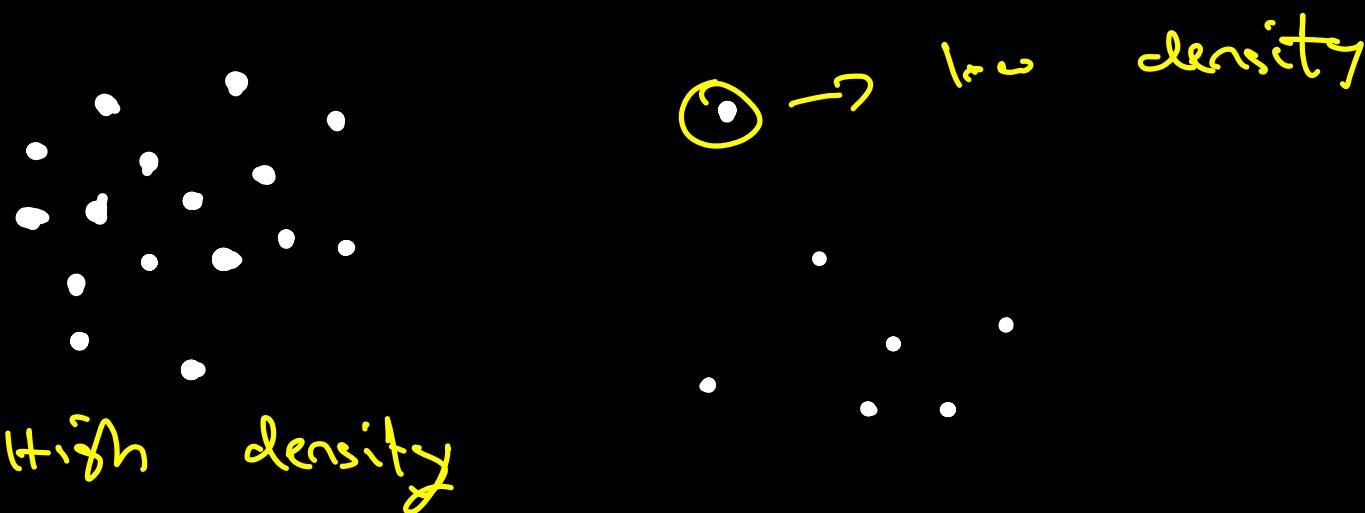


On an avg → the depth of outer nodes / pts will be lower.

Local Outlier Factor

KNN + DBSCAN

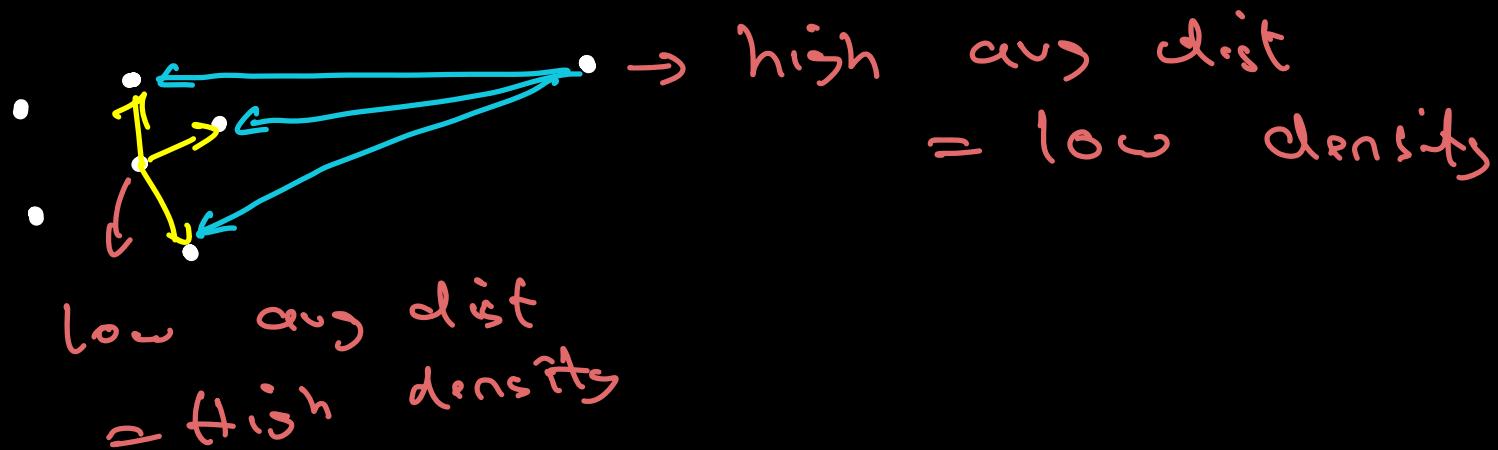
Idea: Outliers exist in low density regions



Density :

Density can be represented as avg distance between points (inverse)

K=3



Steps:

↳ Find K_{nearest} neighbours of point 'p'

→ calc avg distance of K-nn.

$$\text{density}(p) = \frac{1}{\text{avg } \underline{\text{k-dist}}}$$

Variation:

|

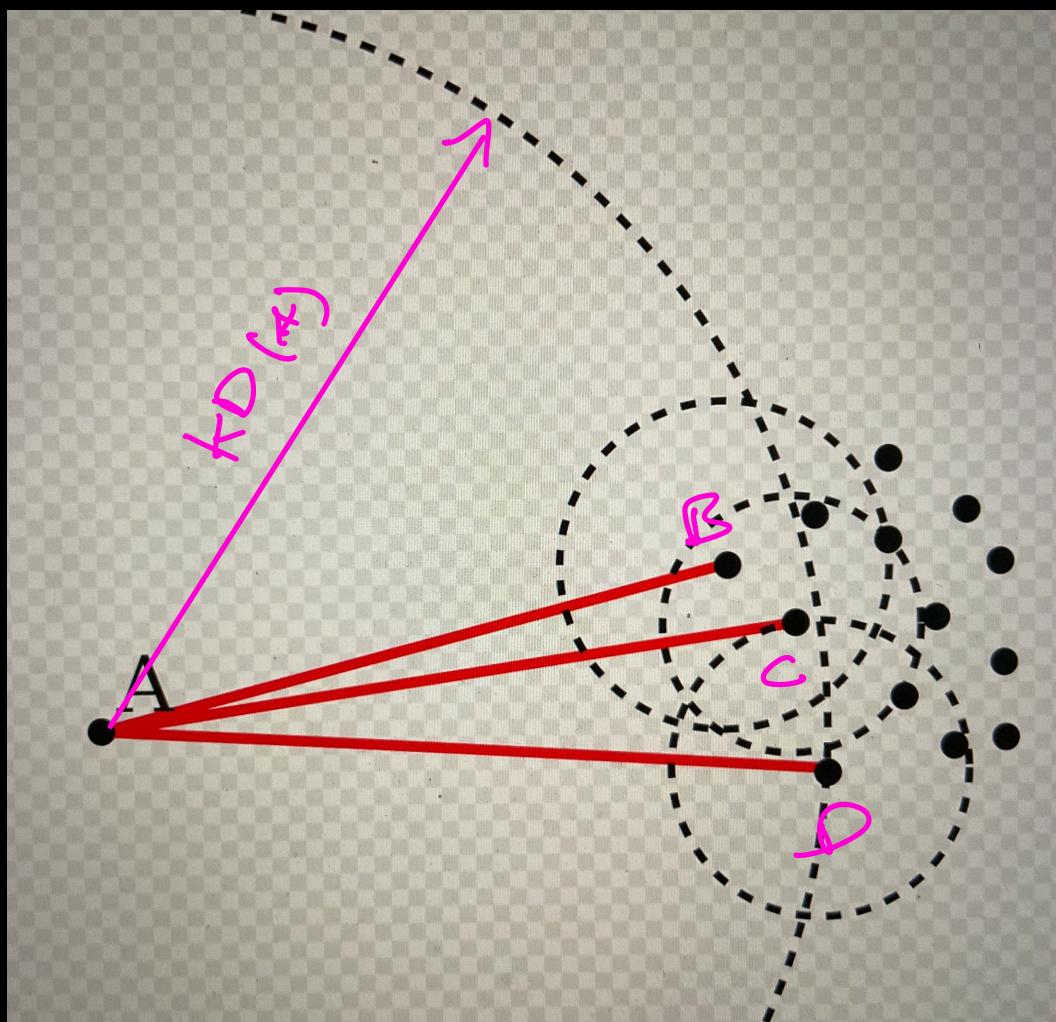
↳ it is recommended , by researchers
that max k-dist works better
than avg. (more stable boundaries)

max(K neighbours) = distance from

k-dist  k^{th} neighbour

Hence,

$$\text{density}(P) \doteq \frac{1}{KD(P)}$$



A has large
radius, i.e
low density

Now we can compute

Local outlier factor =

$$\frac{\text{Avg density of } k \text{ neighbours}}{\text{Density of } P}$$

Q: Outlier when?

A) LOF >> 1

D) LOF << 1

C) LOF ~ 0

D) LOF < 0

This algo that we discussed is called "simplified LOF" because it uses

→ K-distance

However, there is a middle ground between avg and max that avoids some edge cases:

Reachability Distance:

$$RD(A, B) = \max(KD_B, \text{dist}(A, B))$$

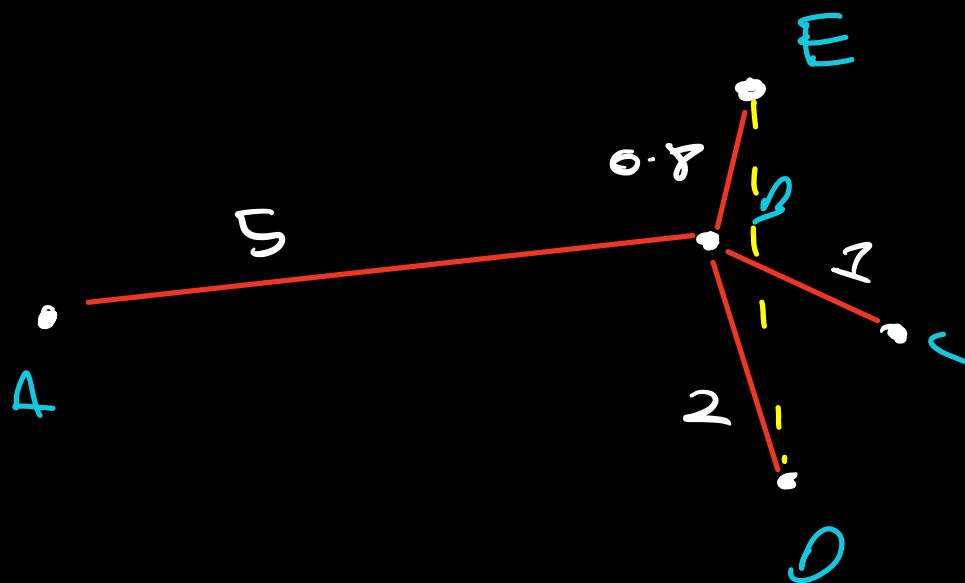
$$\text{Density}(A) = \frac{1}{K} \sum_{B} R_D(A, \text{Neighbours})$$

This means that we will use distance between A, B but it has to be min equal to K-dist(B)
in other words!

RDC me , you):
if you are in my top 'K' but I
am not in your top 'K' then
let's use actual distance but if

we both are in each other's top 'k'
then let's use your k-dist

Note: RD is not symmetric.



$$RD(A, B) = 5 \quad \# \text{ not } 1$$
$$RD(C, B) = 2$$

$$RD(E, B) = 2$$

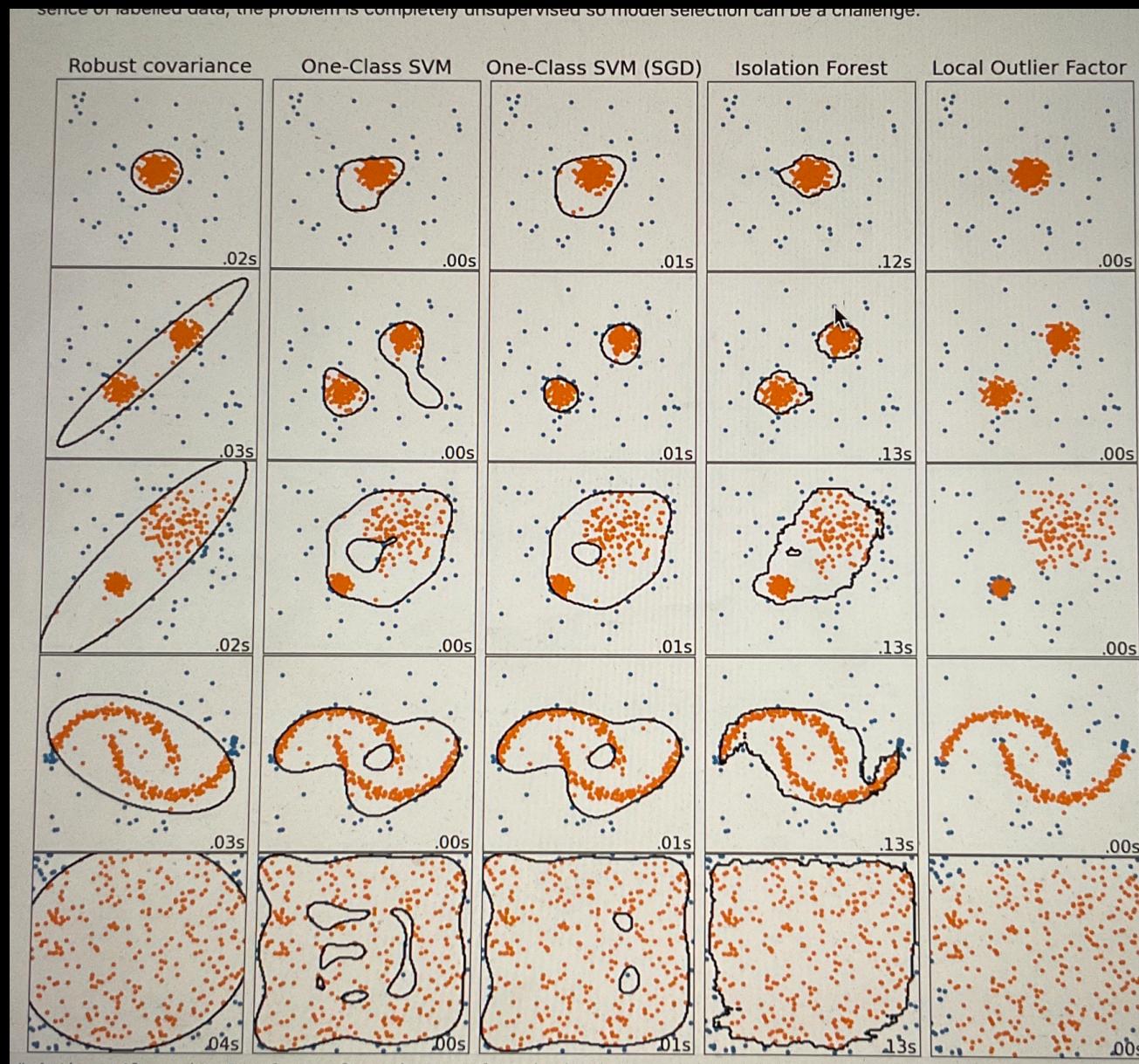
$$RD(B, E) = \underline{2.5}$$

$$RD(D, B) = 2$$

This is less intuitive but has some benefits when 1st neighbour and kth neighbour distance may be very different.

Comprehension

In the absence of labelled data, the problem is completely unsupervised so model selection can be a challenge.



Author: Alfonso Flores <alflores@mit.edu>