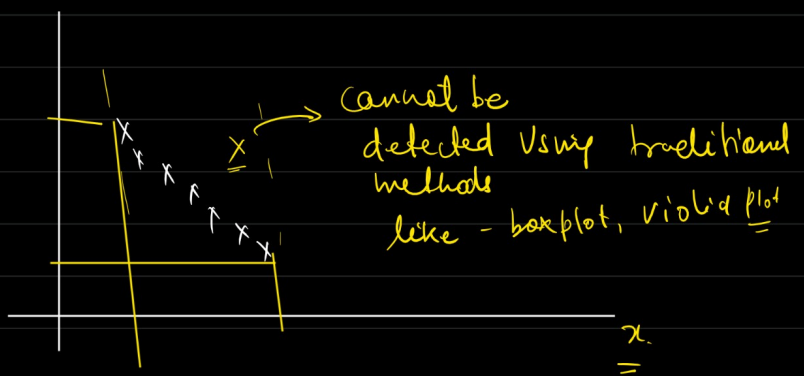
# Anomaly detection (USL)

⇓

## Abnormality




Ex1

| transaction Amt | Acc no |
|---|---|
| 1000 | — |
| 2000 | — |
| 3000 | — |
| | — |

→ 2 Lakhs → this customer is a outlier.

Ex2

Fraud transaction 3 trans → 15000

$t_1$ – 11 am
$t_2$ – 1 PM
$t_3$ 2 PM

$t_4$, $t_5$, $t_6$ → night

D
D
D

USA

Ex-3 gmail or any portal ?

Ex-4 0, 1, 0, 9, 10, 35, 76

Ex5: Fraud IP:

① Isolation forest Anomaly detection
② DBSCAN
③ Local outlier factor Anomaly detection

* Use case → Specifically solving an outlier | Anomaly problems.

* Use Case

ABC — Sbi

→ 1 lakh transactions

→ 3-5%

SL → fraud / not fraud

→ 2004 → 20 years

Utkarsh

DBSCAN

data: 1 lakh ⇒ only the transaction which are abnormal | outliers

2018, 2017, 2016

10K

6K 4K

Classification problem

① Isolation Forest (USL)

⇓

→ multiple Isolation trees

→ Isolation is like a decision tree.

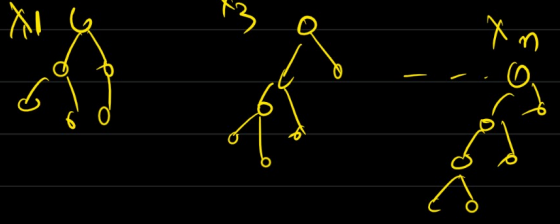
$X_1$ | $X_2$ | $X_3$ | $X_4$

Construction of Isolation tree:

→ Select a feature randomly X

→ Randomly choose a split value with range of $X_1$

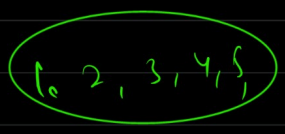
→ Repeat the process recursively to build the tree.



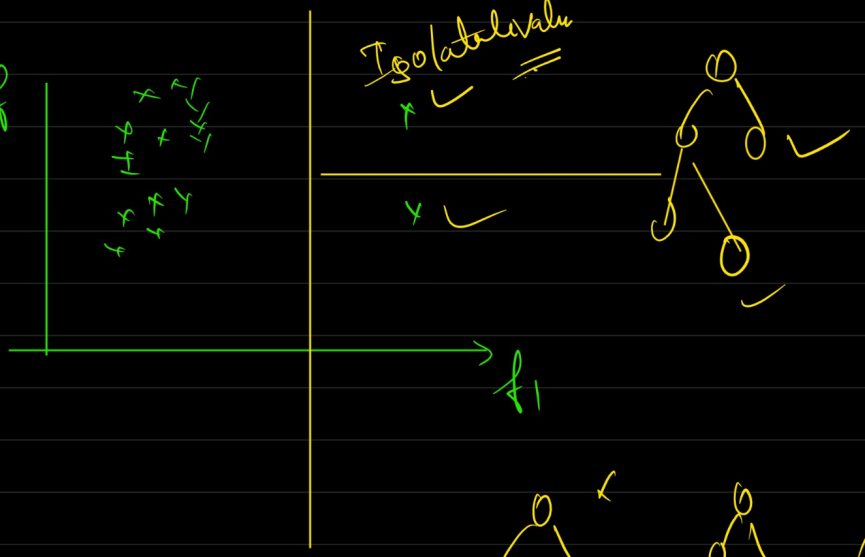

(Each of these isolated tree will be built to leaf node / pure node)

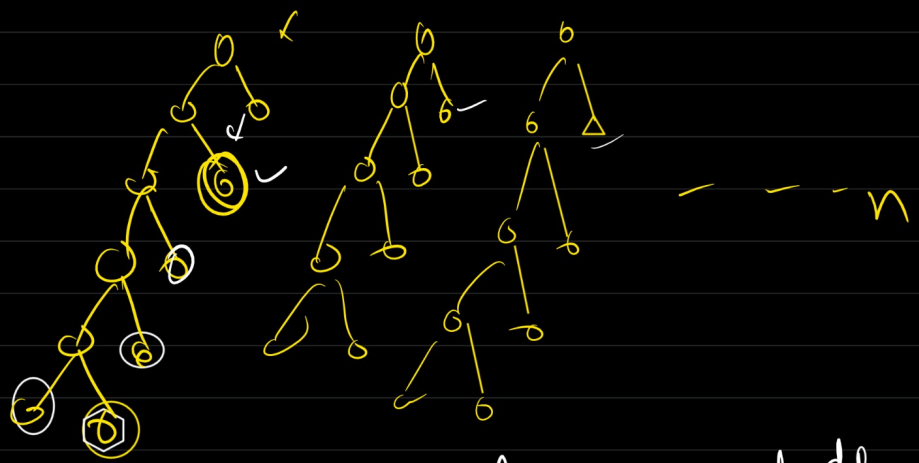
* How you will get outlier

Anomaly/outlier due to their distinctiveness, tend to end up in leaf node at shortest path.


$\leq 100$, $> 100$




Since outliers are diffent from normal dp, during construction of Isolated trees it will be isolated as a seperate leaf node.

How much confidence !?

* Anomaly score

$$S(x,m) = 2^{-\frac{E(h(x))}{C(m)}}$$

m is total no of dp's

x → datapoint for which you want to check anomal score

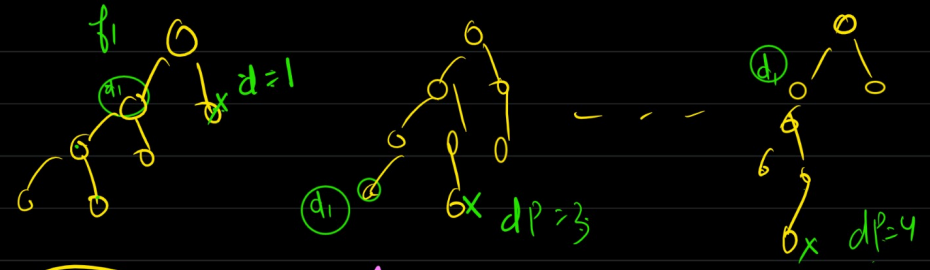
(one dp)

$E(h(x))$ → Average search depth of x in all isolated trees.

$C(m)$ = Average depth of all the dp's in all isolation trees

$E(h(x)) << c(m)$

Since $c(m)$ is avg depth of all dp's so $c(m)$ be far greater than $E(h(x))$




$$s(x,m) = 2^{-\frac{E(h(x))}{c(m)}}$$

$\frac{E(h(x))}{c(m)}$ → very small / very large

$s(x,m) = 2^{-\text{very small no}} = \frac{1}{\text{very small no}} \approx 1$ → very very small.

Generally if $s(x,m) \geq 0.5$ ⟹ outlier

threshold

$< 0.5$ ⟹ Normal dp

$E(h(x)) >>> c(m)$

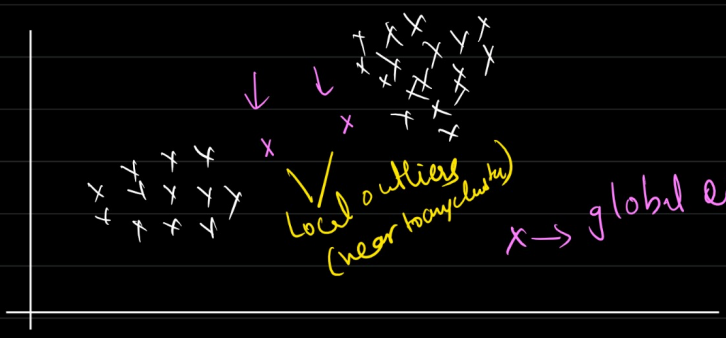
$2^{-\frac{E(x)}{c_m}} = \frac{1}{2^{\text{big}}} = 0.5$

## DBSCAN

→ Studied VSL

→ Also detects the outlier ⟹ If a dp is not core / boundary dp ⟹ That's an outlier.

# * Local Outlier factor




x → global outlier (far from all the cluster)

⇓

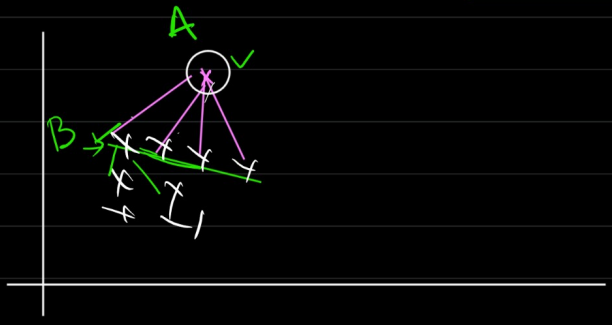
easy to detect

## How to detect local outlier.

⇓

Local outlier Factor.

* Calculate distance of a dp with its k-nearest neighbour
* dp's which are far will have more distance from its neighbour & vice-versa.
  * Higher distance ⇒ density less ⇒ outlier.




$$LOF_{score} = \frac{\text{Interested dp distance}}{\text{All dp's distance of neighbourhood dp}}$$

> threshold → outlier

LOF. (0-∞)

~1 → Normal

>1 → outlier

<1 → Normal