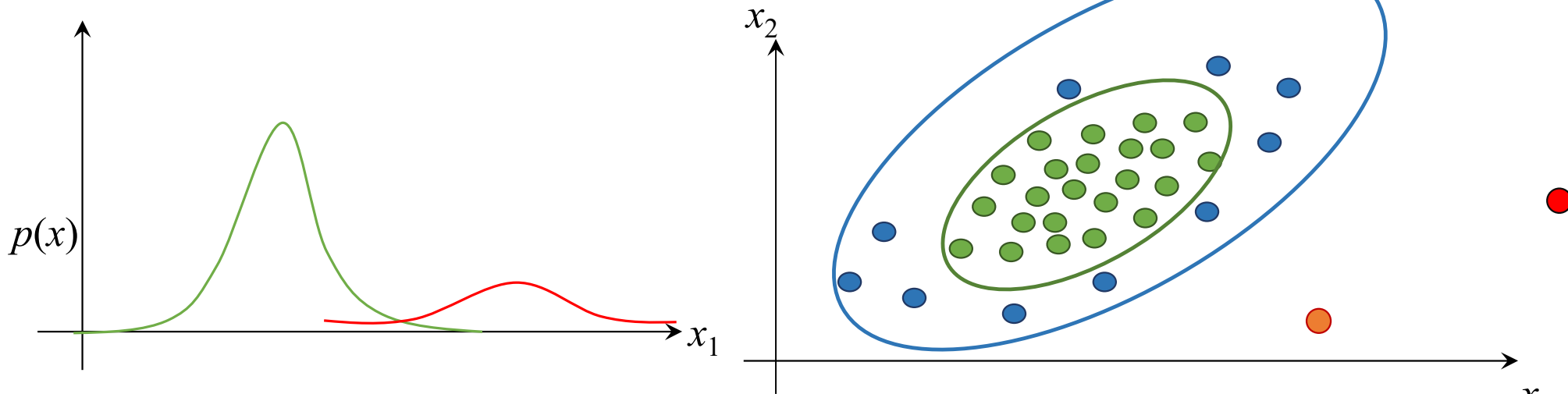


Anomaly/Outlier Detection

What are Anomalies?

- Anomalies are data points that are considerably different from the rest of the dataset, and do not confirm the normal behaviour of the data.
- Can be broadly classified into different categories:
 - Outliers: Small anomalous patterns that appears in a non-systematic way in data
 - Change in Events: Systematic or sudden change from the previous normal behaviour
 - Drifts: Slow, unidirectional, long-term change in the data



Anomalies vs. Noise

- Anomalies may represent rare events , errors, or genuine but unusual phenomena within the data. E.g.: Unusually high blood pressure
- Noise refers to random fluctuations or errors in data that do not carry meaningful information and reduces data quality. Follows a random distribution

Financial Data:

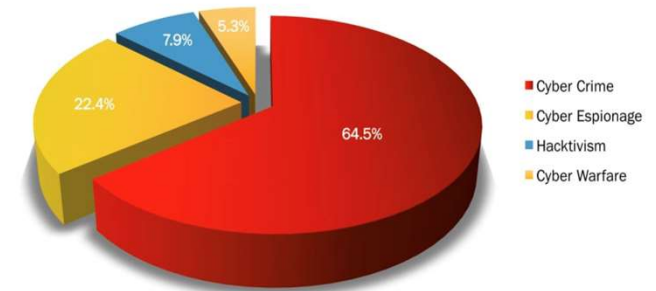
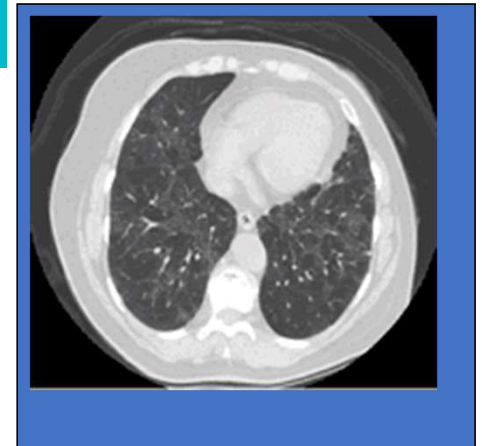
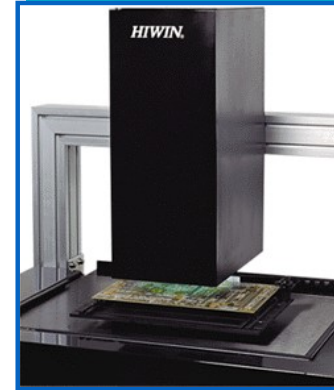
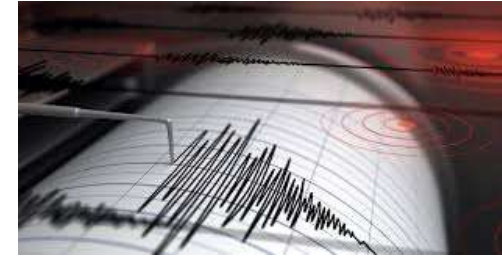
- Noise: Random fluctuations in stock price throughout the day due to short-term trading activity
- Anomaly: A sudden and significant drop in the value of a particular stock, potentially indicating financial trouble

- ✓ Understand the Data and Context
- ✓ Consider the Magnitude and Duration

- ✓ Look for Contextual Clues
- ✓ Utilize Statistical Techniques

Anomaly Detection - Applications

- **Credit Card Fraud:** An abnormally high purchase made on a credit card
- **Cyber Intrusions:** A web server involved in ftp traffic, threat detection
- **Medicine:** Unusual symptoms or test results may indicate potential health problems of a patient
- **Public Health:** The occurrence of a particular disease, e.g. tetanus, scattered across various hospitals of a city indicate problems with the corresponding vaccination program in that city
- **Earthquake Prediction**



Anomaly Detection - Categorization

Supervised Scenario:

- Training data with normal and abnormal data objects are provided
- There may be multiple normal and/or abnormal classes
- The classification scenario is often **highly imbalanced**

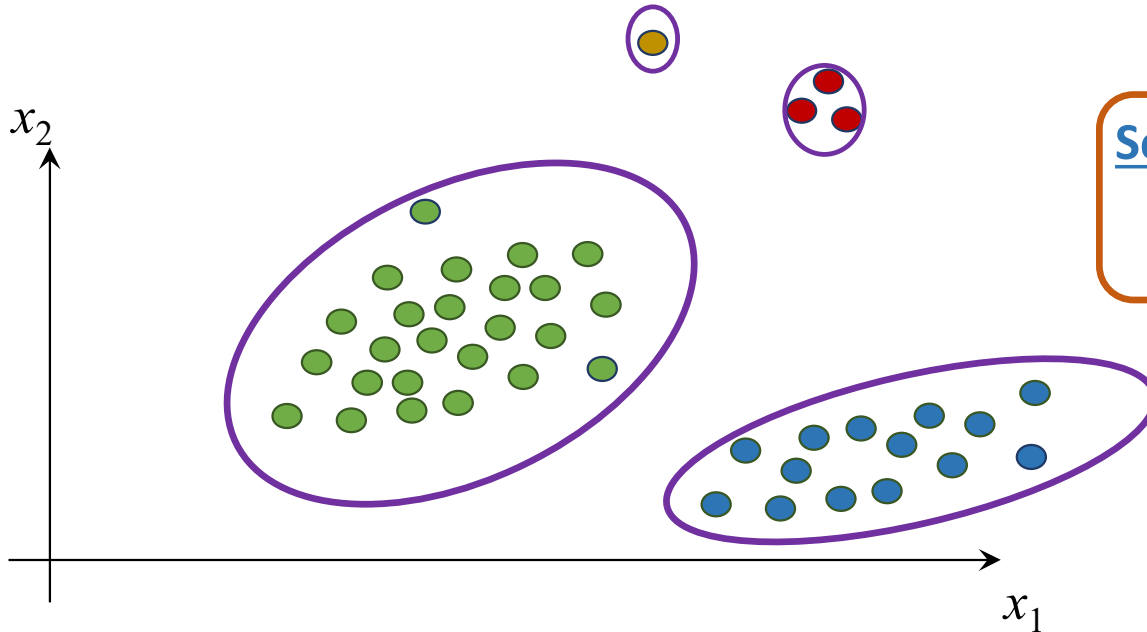
Image Source: From the slides of Tan,
Steinbach and Kumar

<i>Tid</i>	SrcIP	Start time	Dest IP	Dest Port	Number of bytes	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	199	No
5	206.163.37.95	11:14:32	160.94.179.254	139	19	Yes
6	206.163.37.95	11:14:35	160.94.179.253	139	177	No
7	206.163.37.95	11:14:36	160.94.179.252	139	172	No
8	206.163.37.95	11:14:38	160.94.179.251	139	285	Yes
9	206.163.37.95	11:14:41	160.94.179.250	139	195	No
10	206.163.37.95	11:14:44	160.94.179.249	139	163	Yes

Anomaly Detection - Categorization

Unsupervised Scenario:

- Each data input does not have such a label
- It is considered as an outlier, depending on its relation with the rest of data



Semi-Supervised

- Training data is all normal
- Test data contains anomalous points

Anomalies w.r.t. Clustering

- Many Clustering Algorithms do not assign all points to clusters but account for noise objects
- Can we consider outliers as just a side product of some clustering algorithms and look for them by applying one of the clustering algorithms and retrieve the outlier set

Problems

- ❑ A set of many abnormal data objects that are similar to each other would be recognised as a cluster rather than as noise/outlier
- ❑ Accuracy of the outlier detection depends on how good the clustering algorithm captures the structure of the clusters

Anomaly Detection Approaches

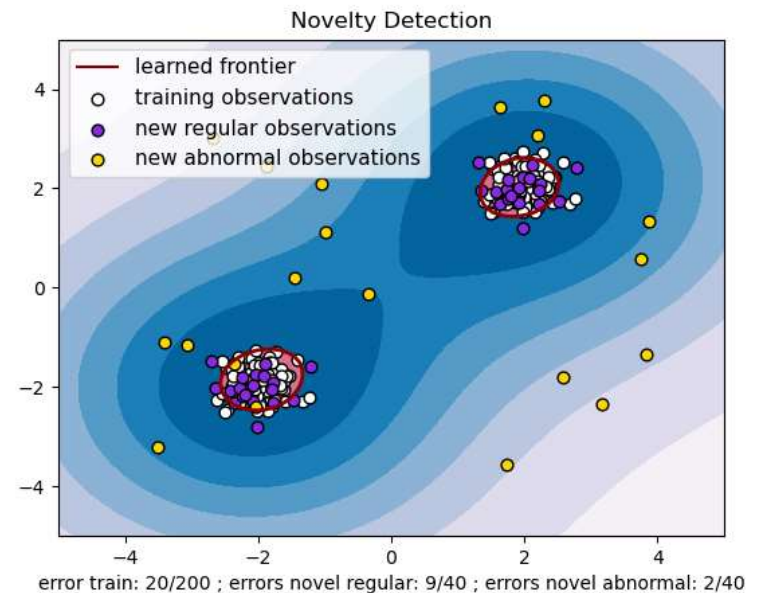
- Density Based
 - Kernel Density Estimation
 - Gaussian Mixture Models
 - Histogram-based Outlier Detection
- Quantile Based
 - One-class SVM
 - Support Vector Data Description
- Nearest Neighbor Approaches
 - Local Outlier Factor
 - Reverse Nearest Neighbor
 - kNN Angle-Based Outlier Detection
- Projection-based Models
 - Isolation Forest
 - Lightweight Online Detector of Anomalies

Anomaly Detection - Challenges

- Modelling normal objects and anomalies properly, as it is hard to enumerate all possible normal behaviour with respect to a dataset
- Choice of distance for outlier detection is often application dependent, with medical data considering a small deviation as an outlier, whereas larger fluctuations are required in marketing analysis
- Noise may distort normal object and reduce the effectiveness of outlier detection
- Difficulty in dealing with high-dimensional data: Use Dimensionality Reduction
- Extremely unbalanced data
- Needs Large Amount Of Data
 - Density Estimation: Exponential
 - Quantile Methods: Polynomial
- Explainability of Anomalies

One-Class SVM

- An unsupervised learning approach, converting into one class classification problem
- Parameter ν approximates the fraction of training errors
- OC-SVM separates the entire set of training data from the origin, i.e. finds a small region where most of the data lies and label data points in this region as one class
 - Maximizing the distance between the origin and the hypersphere (keeping the decision boundary as tight as possible to the normal data).
 - Minimizing the number of training points excluded from the hypersphere (allowing for a small percentage of training data considered as outliers).
- OC-SVM uses the "kernel trick" to map data into higher-dimensional spaces where it may be easier to find a separating boundary



Nearest Neighbour Based Techniques

- Based on the assumption that normal points have close neighbours whereas anomalies are located far from other points
- Follows a two step approach:
 - Compute the neighbourhood for each data point
 - Analyze the neighbourhood to determine whether the data record is anomaly or not

Categories:

1. **Distance based methods:** Anomalies are considered as data points most distant from other points
2. **Density based methods:** Anomalies are data points in low density regions

Nearest Neighbours Based Techniques

Pros:

- ✓ Can be used in both unsupervised as well as semi-supervised setting as they do not make any assumptions about data distribution

Cons:

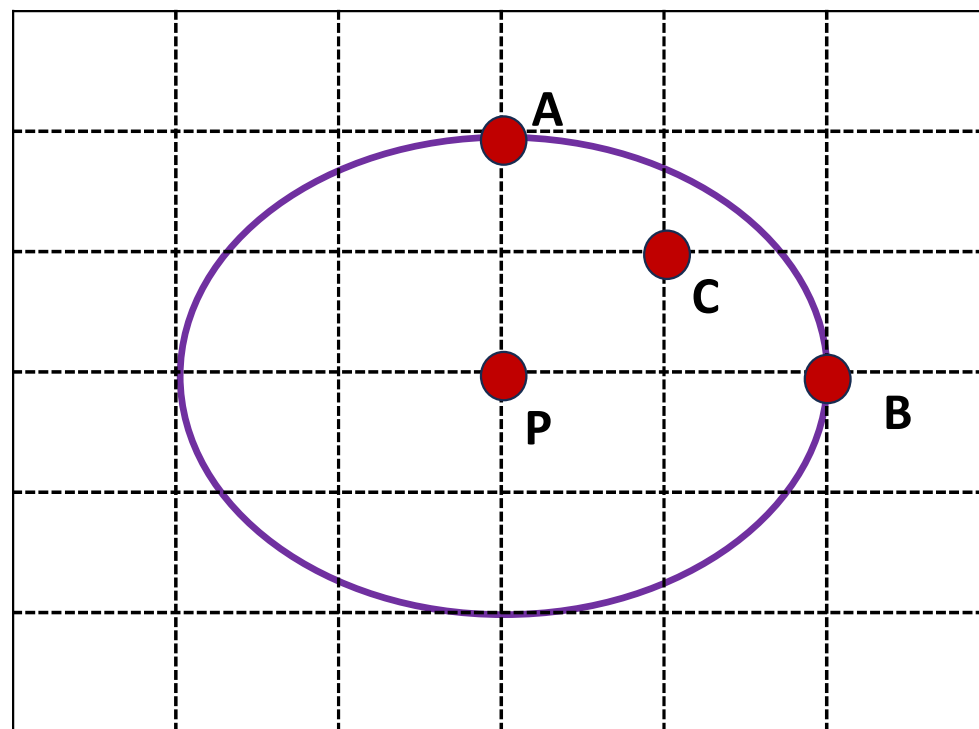
- × Computationally Expensive
- × This technique may fail, provided the normal points do not have sufficient number of neighbours
- × Data points become sparse in high dimensional space -- the concept of similarity no longer holds as distances between any two data records may become quite similar

Local Outlier Factor - K-Distance and K-Neighbours

- LOF measures how much a data point deviates from its local neighbourhood

K-Distance and K-Neighbours

- K-distance is the distance between the point and its K^{th} nearest neighbour
- K-Neighbours ($N_K(P)$) includes a set of points that lie in or on the circle of radius K-distance
- Note: K-Neighbours can be more than or equal to the value of K
- Small Value of K – Algorithm becomes sensitive to noise
- Large Value of K – May not recognise Anomalies

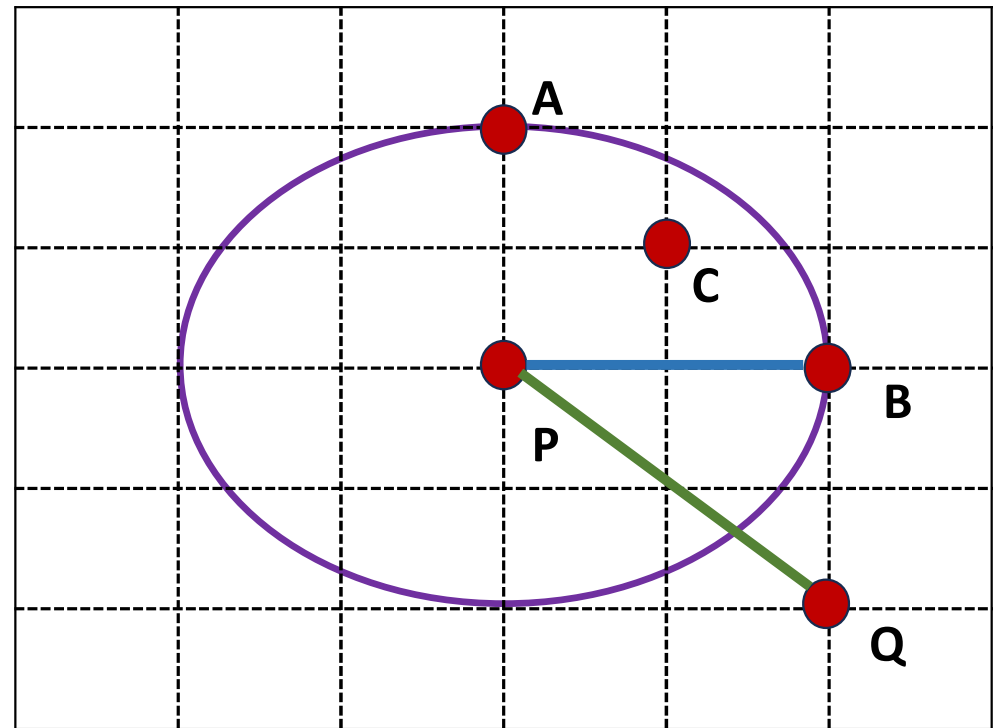


Local Outlier Factor – Reachability Distance

Reachability Distance (RD)

- Reachability Distance between points 'P' and 'Q' is defined as the maximum of the actual distance between two points and the K-distance of the point 'Q'.
- If point 'Q' lies within the K-neighbours of 'P', the RD will be the K-distance of 'P' [blue], else RD will be the distance between 'P' and 'Q' [green]

$$\text{RD}(P, Q) = \max\{K - \text{dist}(P), \text{dist}(P, Q)\}$$



Local Outlier Factor – Local Reachability Distance

Local Reachability Density (LRD)

- LRD is the inverse of average reachability distance of 'P' from its neighbours.

$$\text{LRD}_K(P) = \frac{1}{\sum_{X_j \in N_K(P)} \frac{\text{RD}(P, X_j)}{\|N_K(P)\|}}$$

- More the average RD, less density of points are present around 'P'
- Refers to how far we need to go from the point we are to reach the next point/set of points
- Low values of LRD → The closest cluster is far from the point

Local Outlier Factor - LOF

- LOF is the ratio of the average LRD of the K neighbours of 'P' to the LRD of 'P'.

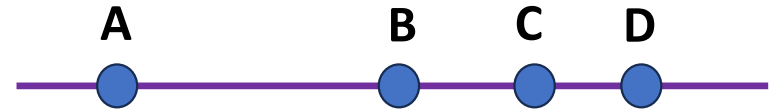
$$\mathbf{LOF_K(P)} = \frac{\sum_{X_j \in N_K(P)} \mathbf{LRD_K(X_j)}}{\|N_K(P)\|} \times \frac{1}{\mathbf{LRD_K(P)}}$$

- If the point is not an outlier, the ratio of average LRD of neighbors is approximately equal to the LRD of a point
 - **LOF = 1** → Similar density as the neighborhood
 - **LOF < 1** → Higher density than the neighborhood
 - **LOF > 1** → Lower density than the neighborhood (outlier)
- LOF can effectively identify local outliers as a point will be considered as an outlier if it is at a small distance from an extremely dense cluster

Since there is no threshold value of LOF, the selection of a point as an outlier is user-dependent.

Reverse Nearest Neighbour

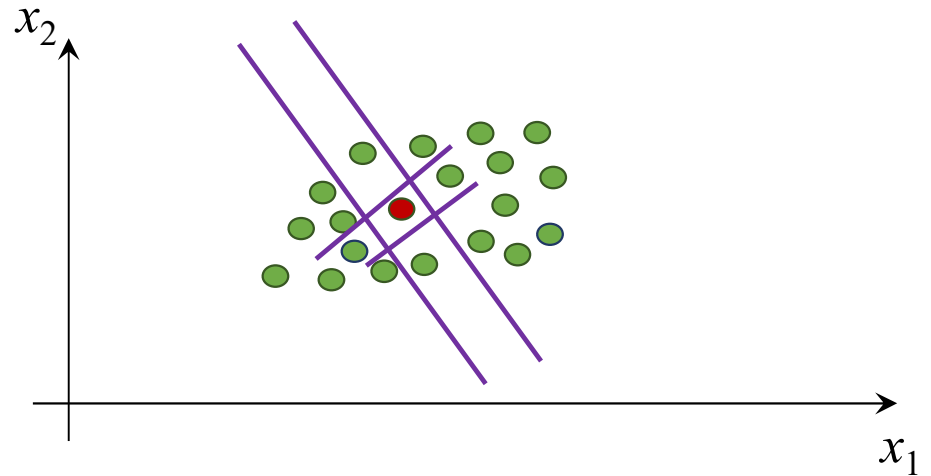
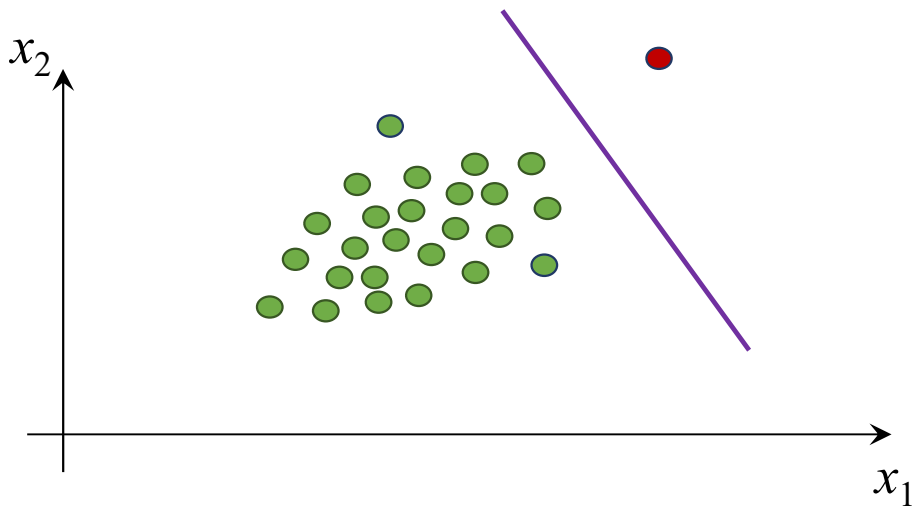
- Given a query object, reverse nearest neighbour search finds all objects in the database whose nearest neighbours are the query object
- Idea: Outliers are not so friendly
 - Find the number of samples that consider the point P as a K-Nearest Neighbor
 - This is the reverse K-NN count
 - Outliers have low reverse K-NN count
- Angle-Based Outlier Detection (ABOD)
 - Quite Robust



	NN	RNN(s)
A	B	-----
B	C	A
C	D	{B,D}
D	C	C

Isolation Forest

- Isolation Forest, an unsupervised algorithm based on the Decision Tree algorithm takes advantage of the following properties of anomalous samples
 - Fewness: Anomalous samples are a minority and there are only a few of them in any dataset
 - Different: Anomalous samples have values/attributes that are very different from those of normal samples



Isolation Forest

- The algorithm isolates anomalies by making them reach isolation points faster compared to normal data points
- Step 1: The algorithm randomly selects a feature (attribute) from the data.
- Step 2: Based on the chosen feature, a random split value is chosen between the minimum and maximum values of that feature. This creates two partitions of the data.
- Steps 1 and 2 are repeated until each data point is isolated or a maximum depth is reached.
- Build multiple isolation trees (ensemble method) and use the average **anomaly score** from each tree for the final prediction.

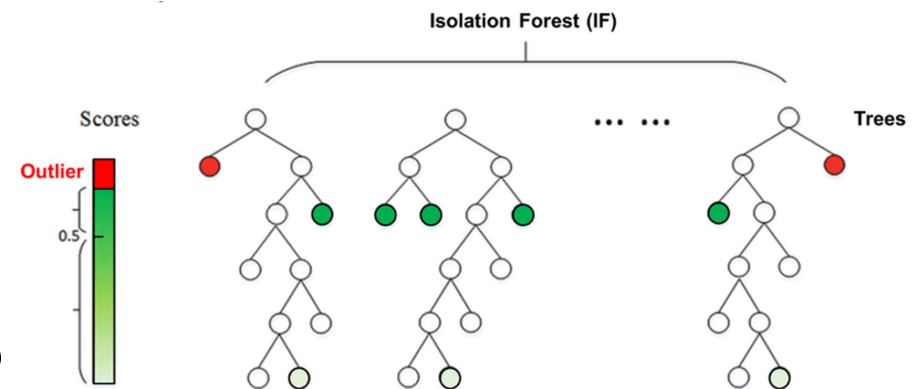


Image Source: <https://wiki.datrics.ai/isolation-forest-model>

Isolation Forest

- Each data point has a path length, which is the number of splits it went through before being isolated.
- The expected path length is calculated based on the number of data points and the number of splits performed.
- The anomaly score for a data point is calculated based on its path length and the expected path length. Data points with shorter path lengths (isolated earlier) are considered more likely to be anomalies.

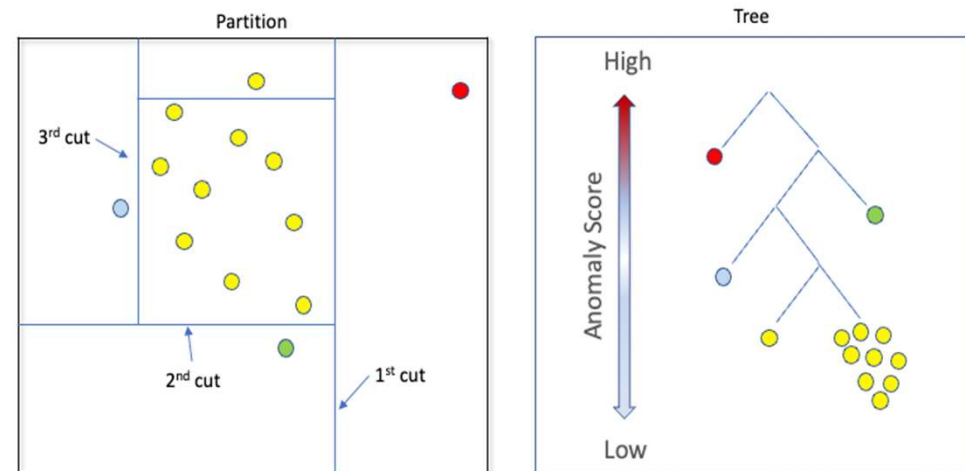


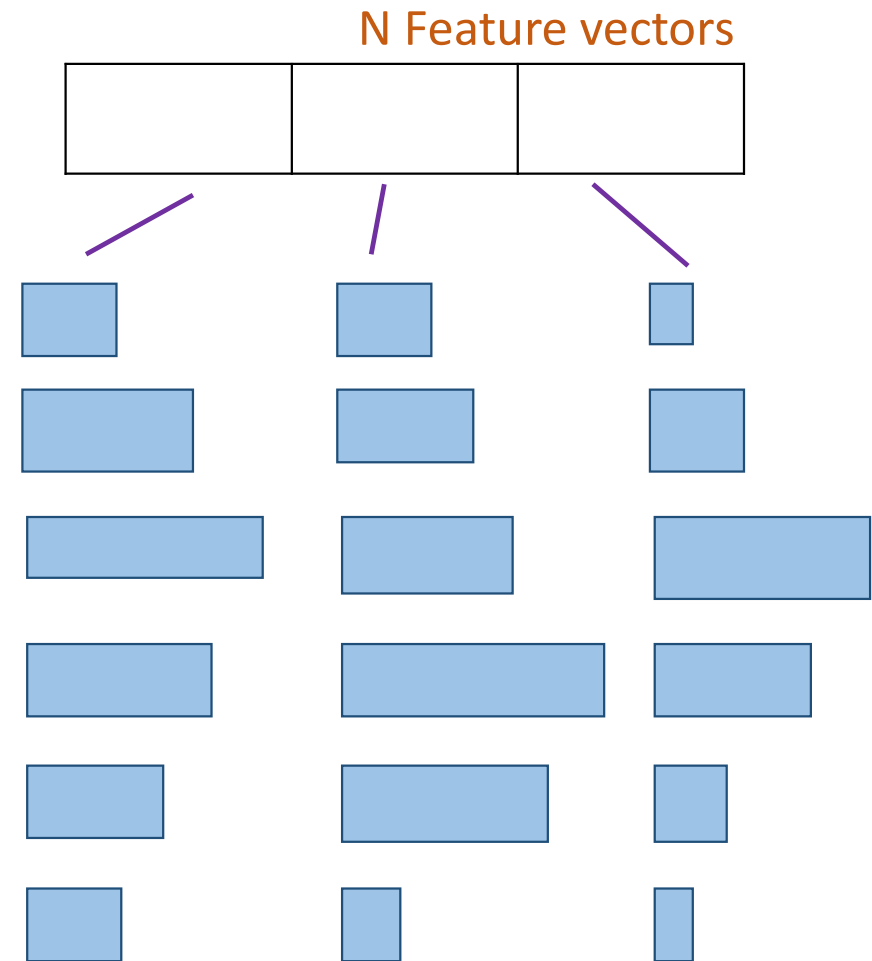
Image Source: <https://www.google.com/>

Histogram Based Outlier Score (HBOS)

- A histogram of the data for each feature is created and a score is calculated based on how likely a particular data point is to fall within the histogram bins for each dimension

Step 1: Building Histogram for each feature with max height normalized to 1

- Choose a binning strategy (e.g., fixed-width or dynamic binning).
- Construct a histogram representing the distribution of the feature values.
- Each bin in the histogram counts the number of data points falling within that specific range of values.



Histogram Based Outlier Score (HBOS)

Step 1: Anomaly score calculation for each data point

- The height of the bin is used to measure the anomaly score as most of the observations (outliers) belong to the bin of high(low) frequency.
 - Define a univariate outlier score: $1/\text{hist}_i(P)$, where $\text{hist}_i(P)$ is the height of the bin of feature i , where data point P belongs to .
- Repeat this for each feature
 - As more features label a sample as outlier, our confidence grows.
 - HBOS is a combination of the univariate outlier scores

$$\text{HBOS}(P) = \sum_{i=1}^N \log \left(\frac{1}{\text{hist}_i(P)} \right)$$

- Can deal with each variable (type) separately
 - Categorical: histogram is count by category.
 - Numeric: histogram is discretized into bins of equal width to derive the count statistic

Evaluating Anomaly Detection

- Accuracy
 - Data tends to be highly biased
- Analyze each type of error
 - False Accept Rate (FAR)
 - False Reject Rate (FRR)
- Depending on application, you might want a specific False Accept Rate (for a security application) or a specific False Reject Rate (for a customer care application)
- Popular Metrics
 - FRR @FAR=0.001
 - ROC and AUC

Summary of Anomaly Detection

- Several other approaches as well
 - Density based approaches
 - Clustering based
 - Subspace/Manifold learning
 - PCA, k-PCA, Autoencoders
 - Deviation from ***association rules*** and ***frequent itemsets***
- Python Anomaly Detection Module (PyOD)
- Open Datasets:
 - ODDS (<http://odds.cs.stonybrook.edu>)
 - OD Benchmark Datasets (<https://www.dbs.uni.lmu.de/research/outlier-evaluation/>)