Anomaly Detection

Dr. Ran Dubin

Lecture slides were taken from Prof. Asaf Shabtai
Malware Detection Using Machine Learning

Applications

- Network intrusion detection
- Insurance / Credit card fraud detection
- Healthcare Informatics / Medical diagnostics
- Industrial Damage Detection
- Image Processing / Video surveillance
- Novel Topic Detection in Text Mining
- Many more

Intrusion Detection

- Intrusion Detection
  - Process of monitoring the events occuring in a computer / network
  - Intrusions are defined as attempts to bypass the security mechanisms of computer network.
- Distributed Deniel Of Service (DDOS)
- Challenges
  - Traditional signature-based intrusion detection systems are based on signatures of known attacks and do not detect emerging cyber threats
- Anomaly detection is the key

Fraud Detection

- Fraud detection refers to detection of criminal activities occurring in commercial organizations.
  - Malicious users might br the actual customers of the organization or might be posing as a customer (identity theft).
- Types of fraud
  - Credit card fraud
  - Insurance claim fraud
  - Mobil/cell phone fraud.
  - Insider Trading
  - Service fraud –advertisment fraud / click jacking
- Challenges:
  - Fast and accurate real time detection

Anomaly Challenges

- Defining a normal region
- The boundries between normal and outlying behavior
- The exact notion of an outlier is different for different applications /domains.
- Considiring known changes like holidays
- Malicious adversaries
- Noisy data in the baseline

Anomaly Detection Strategies

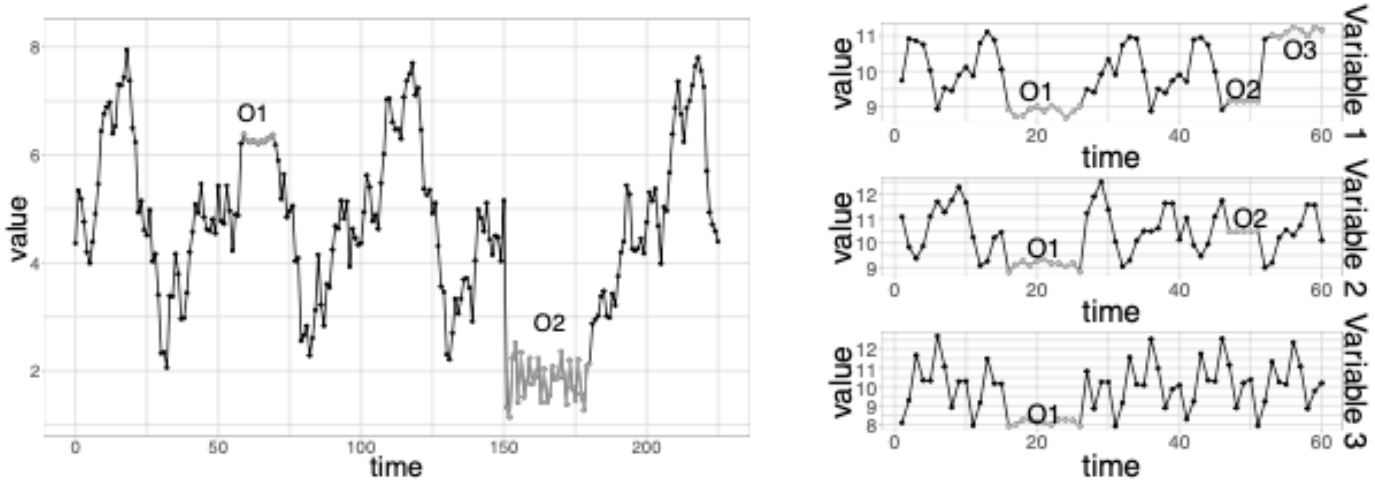
- Supervised Anomaly Detection
  - Labels avilabale for both normal data and anomalies
- Semi-supervised Anomaly Detection
  - Lables avilabale for “normal” data
- Unsupervised Anomaly Detection
  - No labels assumes (common)

Related Problems

- Rare class mining
- Chance discovery – determin significance pice of information that could have a big impact.
- Novelty Detection – aims to detect previously unobserved patterns in the data.

Univariate definition

- A univariate time series $X = \{x_t\}\, t \in T$ is an ordered set of real-valued observations, where each observation is recorded at a specific time $t \in T \subseteq Z$.
- x_t is the point or observation collected at time t and $S = x_p , x_{p+1}, \ldots, x_{p+n-1}$ the subsequence of length $n \leq |T|$
- Starting at position p of the timeseries X, for $p, t \in T$ and $p \leq |T| - n + 1$.
- It is assumed that each observation x_+ is a realized value of a certain random variable X_+ .



(a) Univariate time series. (b) Multivariate time series.

Fig. 4. Subsequence outliers in time series data.

Multivariate definition

- A multivariate time series $X = \{x_t\}\, t \in T$ is defined as an ordered set of k- dimensional vectors, each of which is recorded at a specific time $t \in T \subseteq Z+$ and consists of k real-valued observations, $x_t = (x_{1t},\ldots,x_{kt})$

- x_t is said to be appoint and $S=x_p,x_{p+1},\ldots,x_{p+n-1}$ a subsequence of length $n \leq |T|$ of the multivariate time series X, for $p, t \in T$ and $p \leq |T|-n+1$. For each dimension $j \in \{1,\ldots,k\}$, $X_j = \{x_{jt}\}\, t \in T$ is a univariate time series and each observation x_{jt} in the vector xt is a realized value of a random time-dependent variable X_{jt} in $X_t = (X_{1t},\ldots,X_{kt})$.

Challenge Outlier Detection

	Classification	Outlier Detection
Training Samples	Many from all classes	Almost all from one class
Required Quality	Enough to distinguish two classes	Perfect model of normal

- ML is better at finding similar patterns than at finding outliers.
  - Example: Recommend similar products
- ML is better for finding variants of known attacks
- But new attacks are not known.

- Outliuer assumptions:
  - Malicious activity is anomalous
  - Anomalies correspond to malicious activity

- Outlier problems:
  - User automatic code failes 5K times
  - Brute force attack false positive Vs User forgot to update script password.

Challenge Cost Of Errors

	Cost of False Negative	Cost of False Positive
Product recommendation	Low: potemntial missed sales	Low: continue shoping
Sapm detection	Low: spam finding the way to inbox But if it is phishing/malware - high	High:missed important email
Intrusion detection	High: arbitary damage	High: wasted precious analyst time

Outlier Detection

- Outlier was defined by Hawkins [1980]: “An observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism."

- Unwanted data - related to noise, erroneous, or unwanted data, which by themselves are not interesting to the analyst

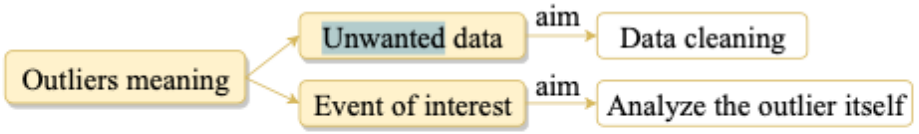
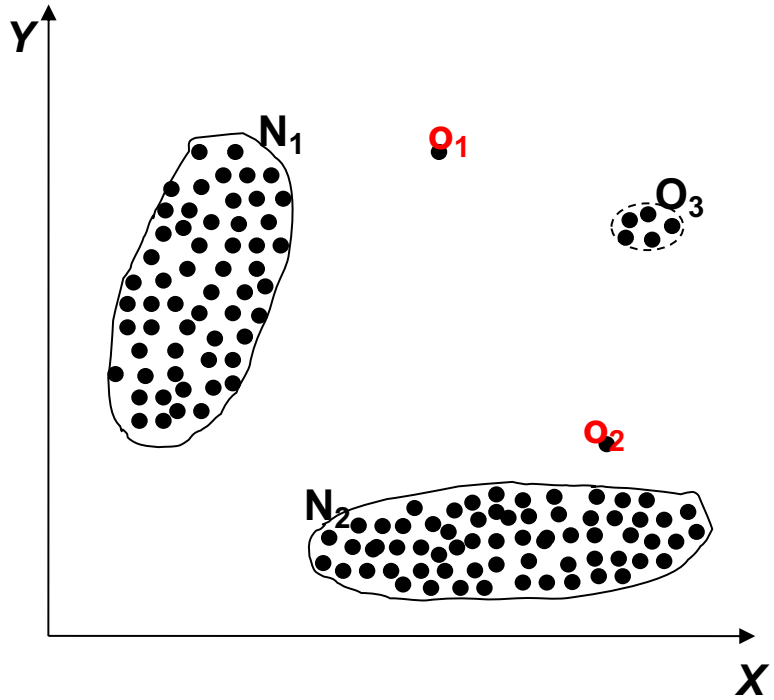


Fig. 1. Meaning of the outliers in time series data depending on the aim of the analyst.

- D. M. Hawkins. 1980. Identification of outliers. Springer Netherlands, New York.

- Blázquez-García, Ane, et al. "A review on outlier/anomaly detection in time series data." *arXiv preprint arXiv:2002.04236* (2020)



Outlier Detection Taxonomy

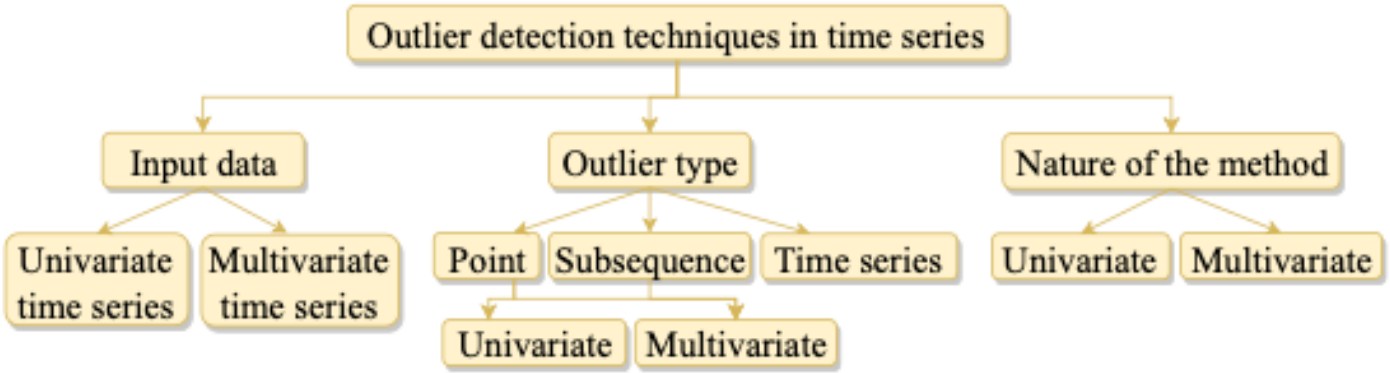


Fig. 2. Proposed taxonomy of outlier detection techniques in time series data.

DETECTING CYBER ATTACKS USING ANOMALY DETECTION WITH EXPLANATIONS AND EXPERT FEEDBACK

- Analyst sometimes has no indications about why the particular computer was identified as being “under attack”.
- analyst may have no method to provide feedback to the detector if the computer was actually identified for some benign reason.
- Use Isolation Forest for attack detection
- Generate explanations about why the detector identified certain computers as anomalous
- Use the explanations to improve the detection.

Table 1. List of Features

SuccessfulLogonRDPPortCount
UnsuccessfulLogonRDPPortCount
RDPOutboundSuccessfulCount
RDPOutboundFailedCount
RDPInboundCount
SuccessfulLogonSQLPortCount
UnsuccessfulLogonSQLPortCount
SQLOutboundSuccessfulCount
SQLOutboundFailedCount
SQLInboundCount
NtlmCount
SuccessfulLogonTypeInteractiveCount
SuccessfulLogonTypeNetworkCount
SuccessfulLogonTypeUnlockCount
SuccessfulLogonTypeRemoteInteractiveCount
SuccessfulLogonTypeOtherCount
UnsuccessfulLogonTypeInteractiveCount
UnsuccessfulLogonTypeNetworkCount
UnsuccessfulLogonTypeUnlockCount
UnsuccessfulLogonTypeRemoteInteractiveCount
UnsuccessfulLogonTypeOtherCount
DistinctSourceIPCount
DistinctDestinationIPCount

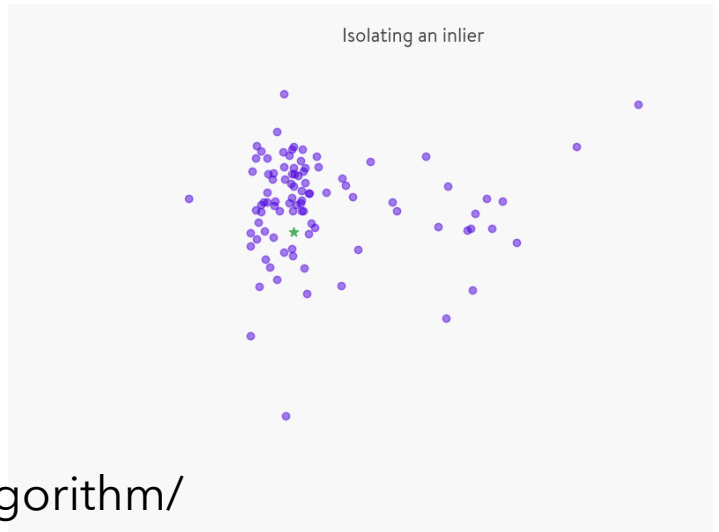
iddiqui, Md Amran, et al. "Detecting cyber attacks using anomaly detection with explanations and expert feedback." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

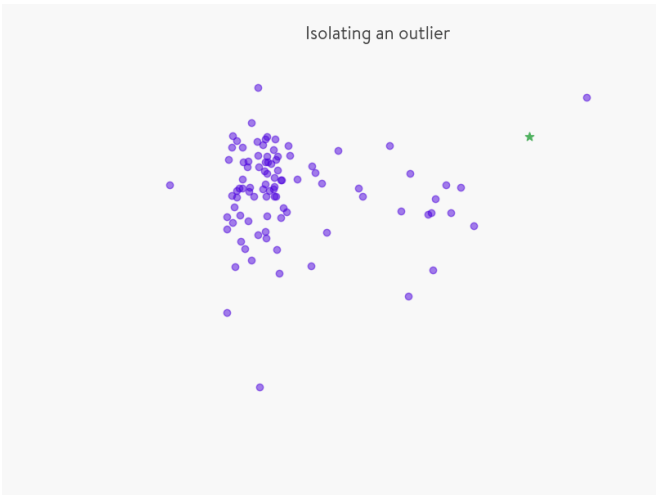
DETECTING CYBER ATTACKS USING ANOMALY DETECTION WITH EXPLANATIONS AND EXPERT FEEDBACK

- “isolate” anomalies by creating decision trees over random attributes.
- The random partitioning produces noticeable shorter paths for anomalies
- Since fewer instances (of anomalies) result in smaller partitions
- Distinguishable attribute values are more likely to be separated in early partitioning
- Hence, when a forest of random trees collectively produces shorter path lengths for some particular points, then they are highly likely to be anomalies.
- Video

Isolation of a normal point

Isolation of an anomaly





Source: https://quantdare.com/isolation-forest-algorithm/

Isolation Forest

- Choosing a record within the dataset and its variables;
- Choosing a random value within the minimum and maximum of each variable or from uniform distribution;
- Creating a node or branch: if the value of the record under consideration is greater or less than the previous random value, we repeat the exercise of evaluating our point with the minimum and maximum interval, limiting it further this time, with the cut-off point being the new maximum or minimum of the branch created.
- Executing the third step until further branching is not possible and the point to be evaluated is isolated.
- Thus, **the fewer branches needed by the tree to isolate the point, the more anomalous it will be**.

DETECTING CYBER ATTACKS USING ANOMALY DETECTION WITH EXPLANATIONS AND EXPERT FEEDBACK

- Collected data from over two millions of computers re- porting during a two week period.
  - Define activity of events window of 30 minutes.
- Divided the data into two classes: servers and clients.
  - 1. Selected the top 1000 anomalous instances and produced the explanation for each.
  - Added each anomaly explanation (over the entire data)
- In the first round the algorithm found many attacks after investigation they were RTP brute force and in the next round they focused on different attacks and found port scaning.
- They used thre outlier tree wight update based on the label data to improve.

Table 2. Explanation examples along with anomaly scores

Anomaly Score	Top 3 features Explanation
0.839	UnsuccessfulLogonTypeOtherCount = 40.0 is unusual with score 0.56 SuccessfulLogonTypeNetworkCount = 800.0 is unusual with score 0.70 SuccessfulLogonTypeInteractiveCount = 120.0 is unusual with score 0.77
0.836	SQLOutboundSuccessfulCount = 44.0 is unusual with score 0.54 UnsuccessfulLogonTypeInteractiveCount = 21.0 is unusual with score 0.70 SQLInboundCount = 1568.0 is unusual with score 0.77
0.834	UnsuccessfulLogonTypeOtherCount = 40.0 is unusual with score 0.56 SuccessfulLogonTypeInteractiveCount = 69.0 is unusual with score 0.70 SuccessfulLogonTypeNetworkCount = 547.0 is unusual with score 0.76
1.24	UnsuccessfulLogonTypeNetworkCount = 59.0 is unusual with score 0.54 RDPInboundCount = 368.0 is unusual with score 1.13 DistinctDestinationIPCount = 11.0 is unusual with score 1.17
1.24	UnsuccessfulLogonTypeNetworkCount = 79.0 is unusual with score 0.54 RDPInboundCount = 145.0 is unusual with score 1.13 DistinctDestinationIPCount = 1.0 is unusual with score 1.16

Table 3. Detection result after each feedback round

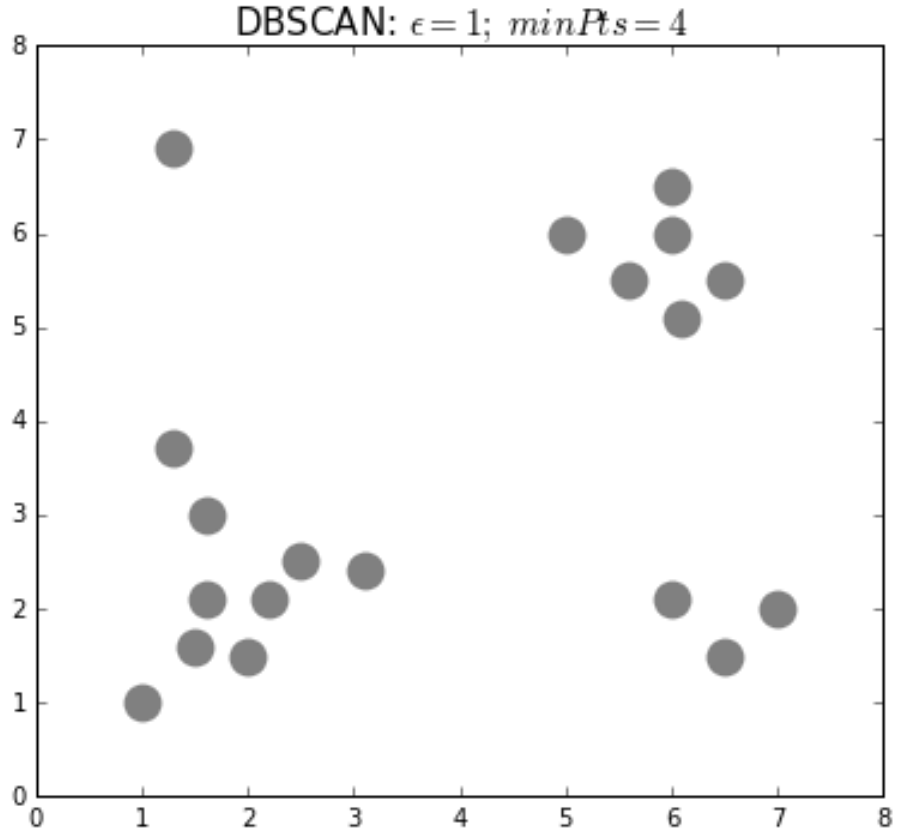
	Feedback Round 1			Feedback Round 2		
Iteration #	1	2	3	1	2	3
# TPs	1	9	20	0	1	2
# FPs	11	4	0	10	17	23

DATADOG OUTLIER DETECTION

- Provide log analysis anomaly detection
- Outlier removeal from baseline is important for better time series anomaly detection algorithm.
- Keep it simple, fast and efficent millions of time series exist.
  - Outlier have to be efficent in time and resources.
- Methods:
  - DBSCAN (density-based spatial clustering of applications with noise)
  - MAD (median absolute deviation).

DATADOG OUTLIER DETECTION- DBSCAN

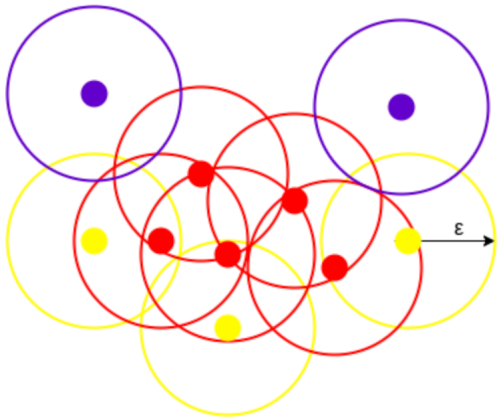
- DBSCAN-
  - DBSCAN groups together points that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. **No need to define number of clusters**
  - Marks as outliers points that are in low-density regions.
- ***eps***: *specifies how close points should be to each other to be considered a part of a cluster. It means that if the distance between two points is lower or equal to this value (eps), these points are considered neighbors.*
- ***minPoints***: *the minimum number of points to form a dense region. For example, if we set the minPoints parameter as 5, then we need at least 5 points to form a dense region.*



Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd *(Vol. 96, №34, pp. 226–231).*
Gif source: https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

DATADOG OUTLIER DETECTION- DBSCAN

- DBSCAN creates a circle of *epsilon* radius around every data point and classifies them into **Core** point, **Border** point, and **Noise**.
- A data point is a **Core** point if the circle around it contains at least '*minPoints*' number of points.
- If the number of points is less than *minPoints*, then it is classified as **Border** Point, and if there are no other data points around any data point within *epsilon* radius, then it treated as **Noise**.



Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd *(Vol. 96, №34, pp. 226-231).*

DATADOG OUTLIER DETECTION- MAD

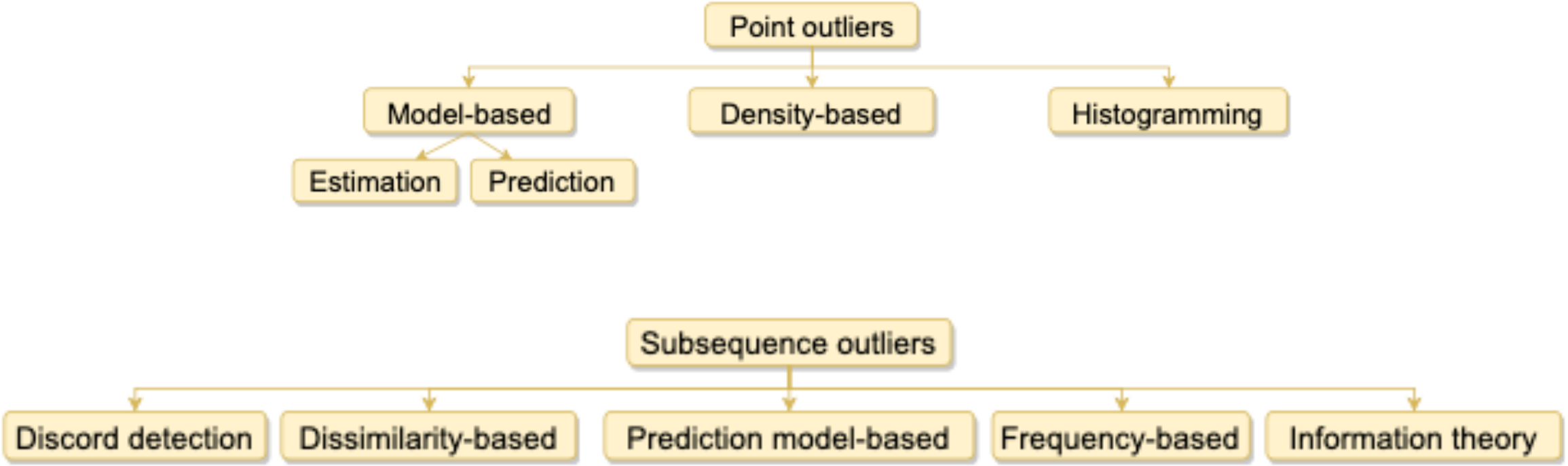
- MAD (median absolute deviation)
  - The Median Absolute Deviation is a robust measure of variability, and can be viewed as the robust analogue for standard deviation.
  - Robust statistics describe data in such a way that they are not too influenced by outliers.

- $D = \{d_1, \ldots, d_n\}$, the deviations are the difference between each d_i and median(D)

- The MAD is then the median of the absolute values of all the deviations.

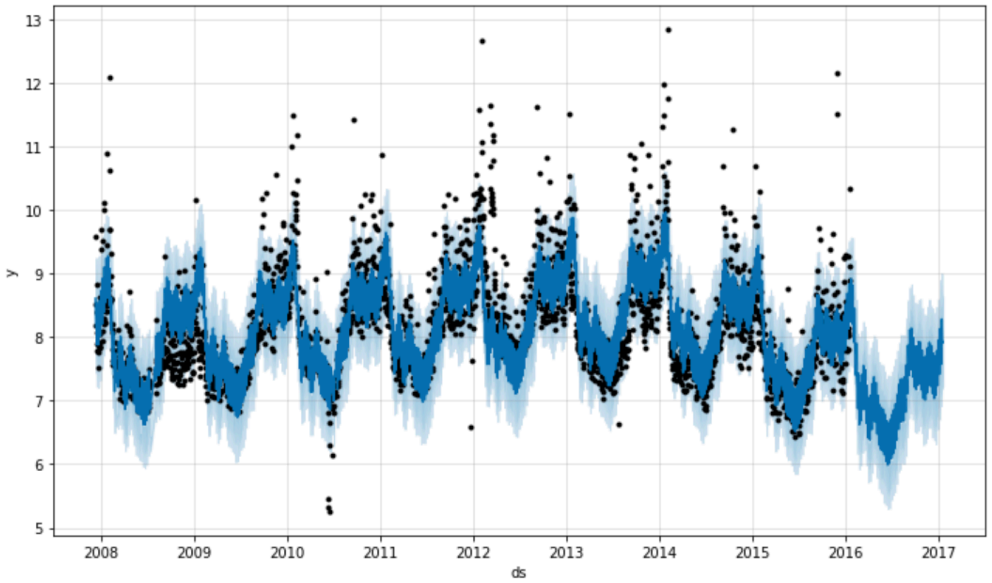
- example :
  - Data= {1, 2, 3, 4, 5, 6, 100}
  - The median is 4
  - The deviations (Median –di) : {-3, -2, -1, 0, 1, 2, 96}
  - MAD is the median of absolute values: {0, 1, 1, 2, 2, 3, 96} which is **2**
  - Note that the standard deviation by contrast is **33.8 which is high compared to the real data.**

Univariate time series



Univariate Time Series Anomaly Detection

- Prophet uses a decomposable time series model with three main model components:
  - trend, seasonality, and holidays.
  - They are combined in the following equation:
  - y(t)= g(t) + s(t) + h(t) + εt
  - g(t): piecewise linear or logistic growth curve for modeling non-periodic changes in time series
  - s(t): periodic changes (e.g. weekly/yearly seasonality) P - is days
  - h(t): effects of holidays (user provided) with irregular schedules
  - εt: error term accounts for any unusual changes not accommodated by the model
- You can try it - https://facebook.github.io/prophet/



$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right)$$

Taylor, Sean J., and Benjamin Letham. "Forecasting at scale." *The American Statistician* 72.1 (2018): 37-45.

Summary

- Anomaly detection Is important building block for various use cases.
- False positives are the adoptation key
  - Each alert force expert review.
- No single algorithm is leading for each use case.
- Please check for furthre reading:
  - Cheng, Chaoran, Fei Tan, and Zhi Wei. "DeepVar: An End-to-End Deep Learning Approach for Genomic Variant Recognition in Biomedical Literature." *Proceedings of the AAAI Conference on Artificial Intelligence*. https://ts.gluon.ai/master/api/gluonts/gluonts.model.deepvar.html