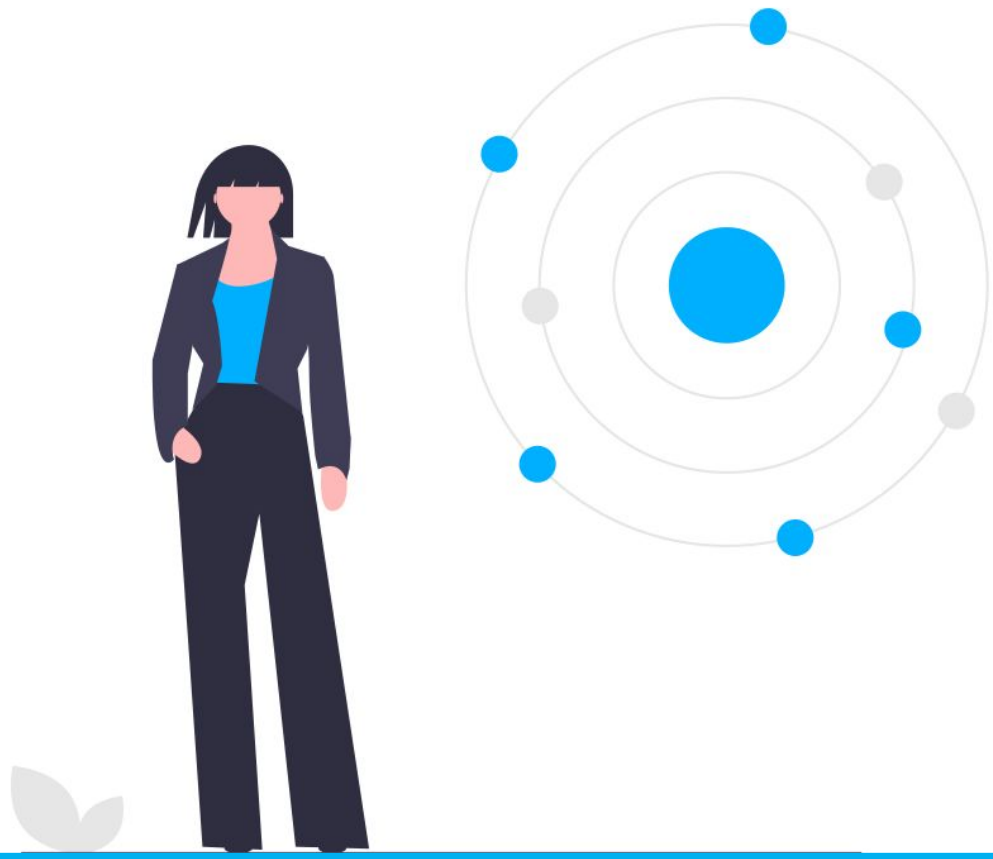
WHITE PAPER
SERIES

2025

Definitive Guide to Anomaly Detection



adsp.ai

APPLIED
DATA SCIENCE PARTNERS

APPLIED
DATA SCIENCE PARTNERS

Introduction

Everything You Need To Know About Anomaly Detection - Types, Techniques, FAQs & More!

If you have ever looked inside a dataset, chances are you have seen an anomaly.

An empty cell, a date in the place of a name, a fake account. Malicious or not, anomalies introduce inconsistencies that can later hinder extracting value from your data.

In the next chapters of our guide we will go through the fundamentals of anomalies. We will lay focus on the interplay between anomaly detection and data science, where we believe lies the true potential of this method.

At the end of your reading, you will be able to effectively reason and communicate about anomaly detection models used in business and industry.

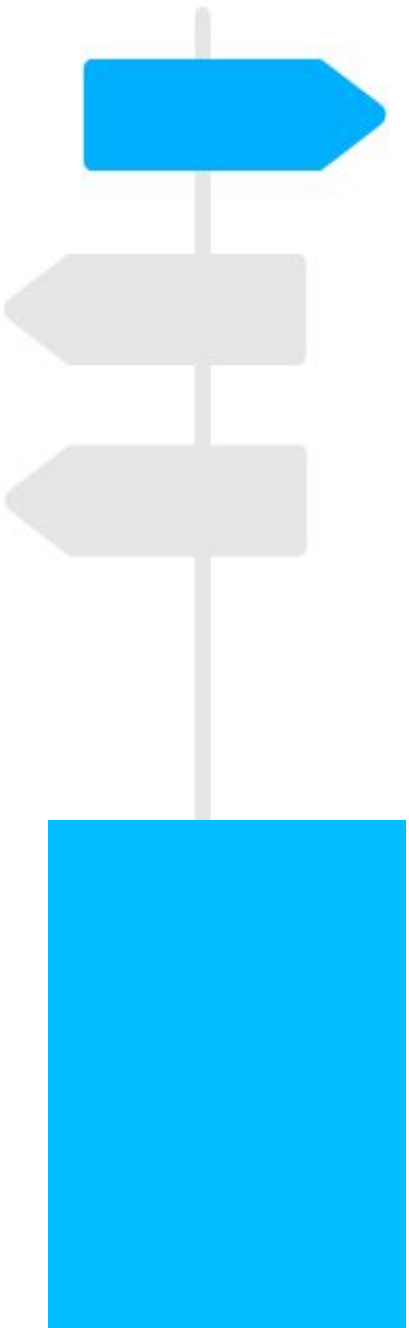
David Foster
Partner
Applied Data Science Partners (ADSP)

APPLIED
DATA SCIENCE PARTNERS

Focus Areas

This whitepaper series is a practical guide to developing your understanding of Anomaly Detection.

It is organised around 8 key topics:



Chapter One:
What Is An Anomaly?

Chapter Two:
What Is Anomaly Detection?

Chapter Three:
What Are The Three Types Of Anomalies?

Chapter Four:
Why Your Company Needs Anomaly Detection

Chapter Five:
What Are The Different Anomaly Detection Techniques?

Chapter Six:
Benefits Of Anomaly Detection And Machine Learning

Chapter Seven:
Anomaly Detection Machine Learning

Chapter Eight:
Frequently Asked Anomaly Detection Questions
Conclusion & Case Study

APPLIED
DATA SCIENCE PARTNERS

1 What Is An Anomaly?

An anomaly in your dataset is a data point that falls outside of the range of usual behaviour.

There is no single cause for anomalies. It may be that the process generating the data has changed, an error was introduced during data collection or your in-house data processing pipeline is not properly interacting with input.

While anomalies may make up a tiny part of your data, they can have a disproportionately large effect on later stages of data processing.

Based on their content we can broadly classify anomalies into two categories:

Expected Anomalies
A spike in sales due to Black Friday or a drop in the sales of print magazines is unusual, but not surprising. Anomalies like this require dedicated mechanisms that can detect seasonal patterns and long-term trends, but are easily dealt with.

Unknown Anomalies
Other anomalies may require more effort to understand. Why is there a spike in the number of complaints in a social network? What caused a drop in the number of orders on a retailer's website?

It is this type of anomaly that has made anomaly detection paramount.

Detection techniques here need to provide an understanding of the causal mechanism generating the irregularities.

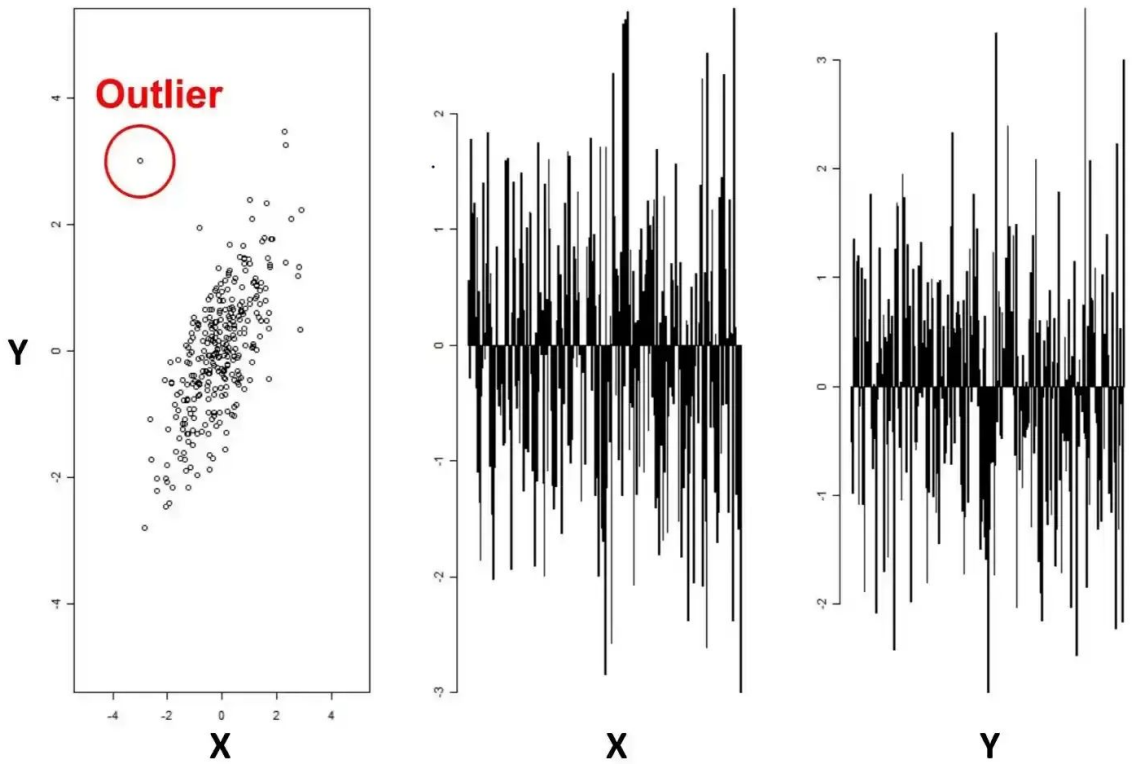
Based on their nature anomalies can be classified in three categories:

Outliers
Short/small anomalous patterns that appear in a non-systematic way in data collection.

Change in Events
Systematic or sudden change from the previous normal behaviour.

Drifts
Slow, undirected, long-term change in the data. This nature of the anomaly is important for picking the right detection method and optimally designing the anomaly detection system.

Take a look at these plots of two variables X and Y for example. If you only look at them separately you will have a hard time spotting the outlier that is so obvious when you plot them on the two axes of a single graph:



2 What Is Anomaly Detection?

How does anomaly detection work?

Anomaly detection is the art and science of detecting anomalies in your data.

Depending on the application under consideration,, you may encounter the following uses of anomaly detection:

Intrusion detection

Databases and networks are vulnerable to malicious attacks. An Intrusion Detection System is a monitoring system that detects suspicious activities and automatically generates alerts.

Anomalous data in this case often corresponds to abnormal behaviour of individual users or unexpected generalised traffic, such as DOS attack.

Fraud detection

Fraud detection is a set of activities undertaken to prevent money or property from being obtained through false pretences. Fraud detection is applied to many industries such as banking or insurance.

In banking, fraud may include forging checks or using stolen credit cards. Other forms of fraud may involve exaggerating losses or causing an accident with the sole intent for the pay out.

To detect this type of anomalies, detection needs to be real-time and user-centric.

Health monitoring

Anomalies can also arise internally due to sub-optimally designed or maintained internal systems. Structural health monitoring is the process of monitoring vital components of the software and hardware pipeline to proactively identify potential malfunctioning.

Defect detection

Defective products may appear even in a perfect pipeline. Defect defection became very widely adopted in the industry after the deep learning revolution that started around 2007 , that improved the efficiency and accuracy of computer vision. Such systems can automatically recognize faults in products or equipment.

The success of anomaly detection largely depends on how you interact with the data.

3 What Are The Three Types Of Anomalies?

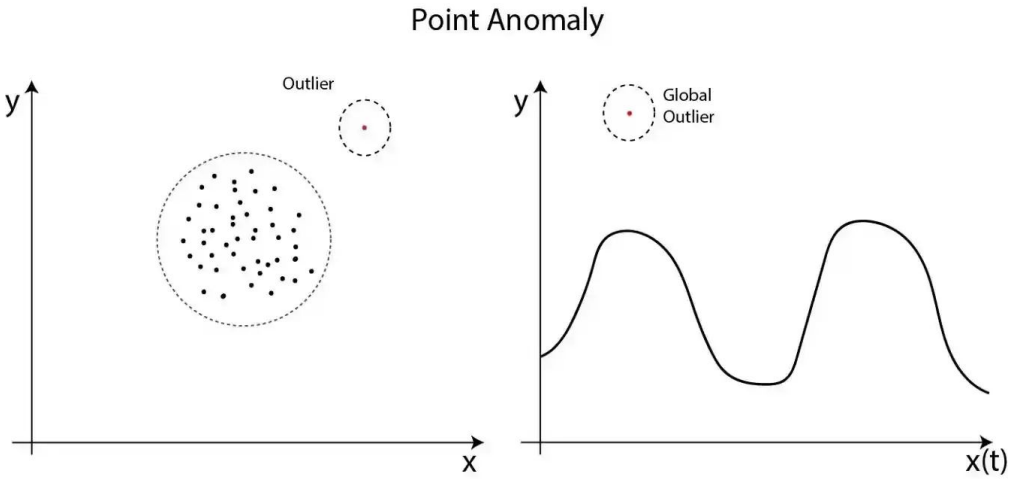
What does it mean for a data point to differ in practice?

Depending on how you compare the data, you may be dealing with one of three types of anomalies:

Global (or point) anomalies

When a single data point (or datum) or an observation in the data set is far off from the rest of the data, then it's said to be a point anomaly. They represent an extremum, irregularity, or deviation that occurs randomly with no association with the common pattern in the data.

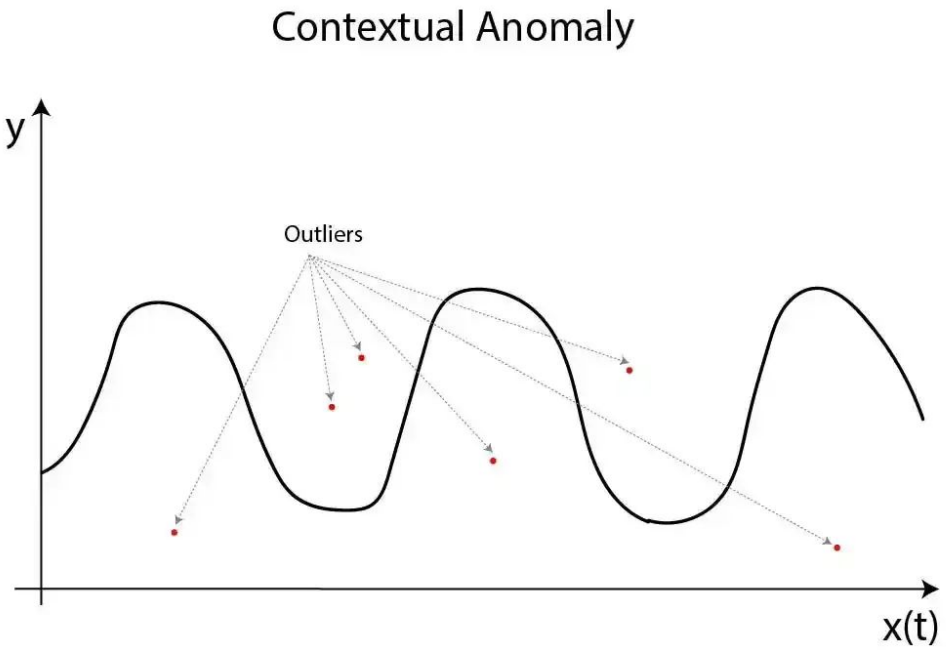
This is a strong type of anomaly that is less often encountered than the following two.



Contextual anomalies

Data points that differ greatly from other data within the same context are called contextual outliers.

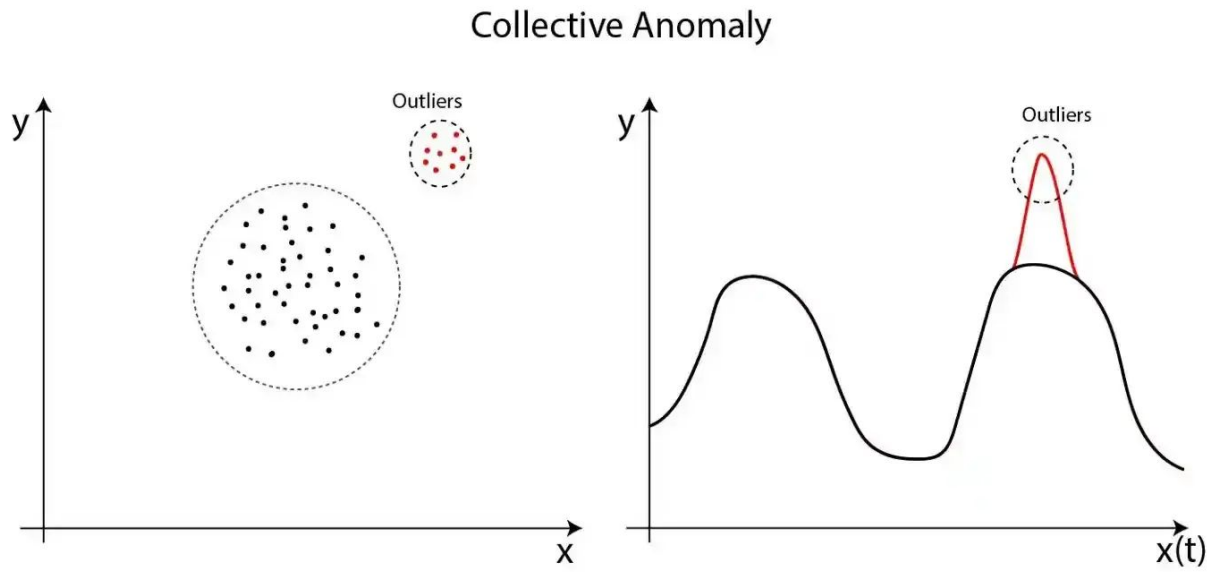
In time-series data, the observations that do not follow the pattern in the time-series data as well as significantly away from the observed pattern — are considered to be anomalous.



Collective anomalies

When a group of correlated, interconnected, or sequential instances significantly differ from the rest of the data, then those data points, collectively, are considered to be anomalous.

For the time-series data, this might appear as typical peaks and valleys happening outside of a time period when a seasonal sequence is usual, or as a set of time series that are in an outlier condition.



4 Why Your Company Needs Anomaly Detection

It is critical for humans to be able to identify changing performance and take actions on that insight.

A shift in a metric could be innocuous, or it could represent a detrimental event happening within the business, or a positive opportunity for growth.

By being alerted to these instances via anomaly detection, users can discern between insignificant changes and those that are truly unusual, driving insight and action.

Here are some specific metrics that anomaly detection can help improve:

Product quality

Health monitoring and fault detection are vital operations for making sure that the output of your business is on-par with expectations. By detecting anomalies early, you can decrease the adversarial effects and cost of malfunctioning.

User experience

By identifying and addressing user experience issues early on, you can prevent them from becoming bigger problems down the line. Anomaly detection can help you identify issues with your website, app, or other online service so that you can fix them before they cause major problems for your users.

Marketing performance

Anomaly detection can help improve marketing performance in a number of ways including; by identifying potential issues and problems early on, helping to optimise campaigns, and providing insights that can help improve future marketing efforts.

5 What Are The Different Anomaly Detection Techniques?

When it comes to choosing a technique for applying anomaly detection, data availability is the most important parameter. Anomaly detection techniques can be classified as:

Unsupervised

Unsupervised anomaly detection techniques are used when there are no labelled data points. These techniques are used to find patterns in data that do not conform to expected behaviour.

The most common unsupervised anomaly detection technique is clustering. Clustering is a method of grouping data points together so that data points in the same group are more similar to each other than data points in other groups.

Semi-supervised

Semi-supervised anomaly detection techniques are used when there are some labelled data points, but not enough to train a supervised anomaly detection model.

Semi-supervised anomaly detection techniques use both labelled and unlabelled data to find patterns in data that do not conform to expected behaviour.

The most common semi-supervised anomaly detection technique is one-class classification. One-class classification is a method of classifying data points as either inliers or outliers. Inliers are data points that are similar to the data points in the training set. Outliers are data points that are not similar to the data points in the training set.

Supervised

Supervised anomaly detection techniques are used when there are enough labelled data points to train a supervised machine learning model. Supervised anomaly detection techniques use a training set of labelled data points to find patterns in data that do not conform to expected behaviour.

The most common supervised anomaly detection technique is support vector machines. Support vector machines are a type of machine learning model that can be used for both classification and regression.

6 Benefits Of Anomaly Detection And Machine Learning

With the reduced cost of capturing data through sensors, as well as the increased connectivity between devices, being able to extract valuable information from data is becoming increasingly important.

Finding patterns in large quantities of data is the realm of machine learning, which holds the technology required to realise the potential of anomaly detection.

Detection methods of the past required domain knowledge and heavily relied on manual clearing and tuning. More importantly, they were limited to identifying simple patterns and could not draw causal relationships.

Using machine learning we can build robust pipelines that can discover complex patterns in large amounts of data with industry-leading accuracy.

In the next chapter we will review specific algorithms that make this possible.

7 Market Basket Analysis Examples

There is a simple, common idea behind all the algorithms we will review.

A machine learning algorithm for anomaly detection statistically analyses data to build a predictive model that, for a given business metric, outputs expected behaviour for a given input.

Anomalies are detected by comparing the output of the predictive model to real data.

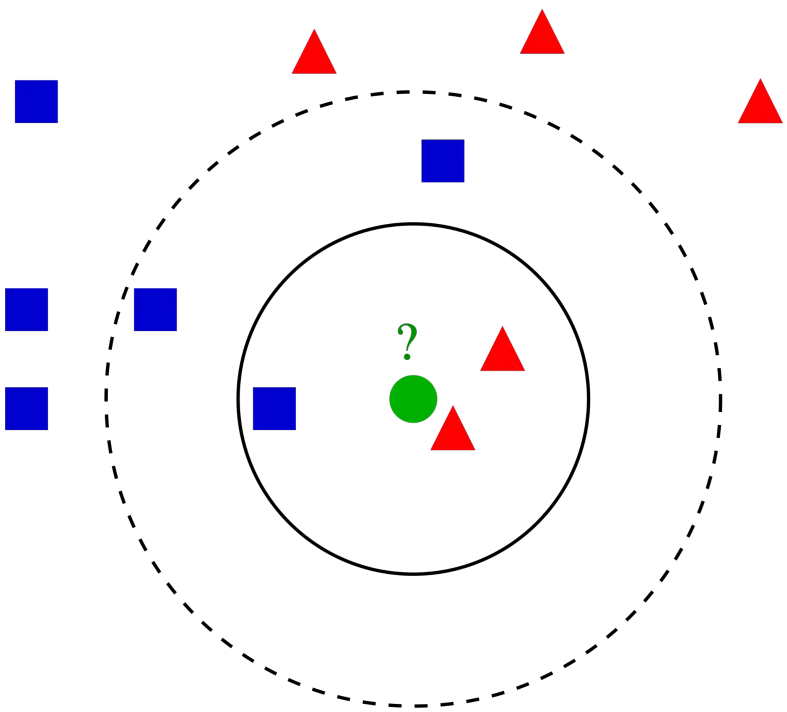
Algorithms are different in the mathematical approach they use to define the model, which can help us understand their trade-offs in performance and complexity.

K-nearest neighbours

kNN is a supervised ML algorithm frequently used for classification problems (sometimes regression problems as well) in data science. It is one of the simplest yet widely used algorithms with good use cases such as building recommender systems, face detection applications etc.

The fundamental assumption in the nearest-neighbour family is that similar observations are in proximity to each other and outliers are usually lonely observations, staying farther from the cluster of similar observations.

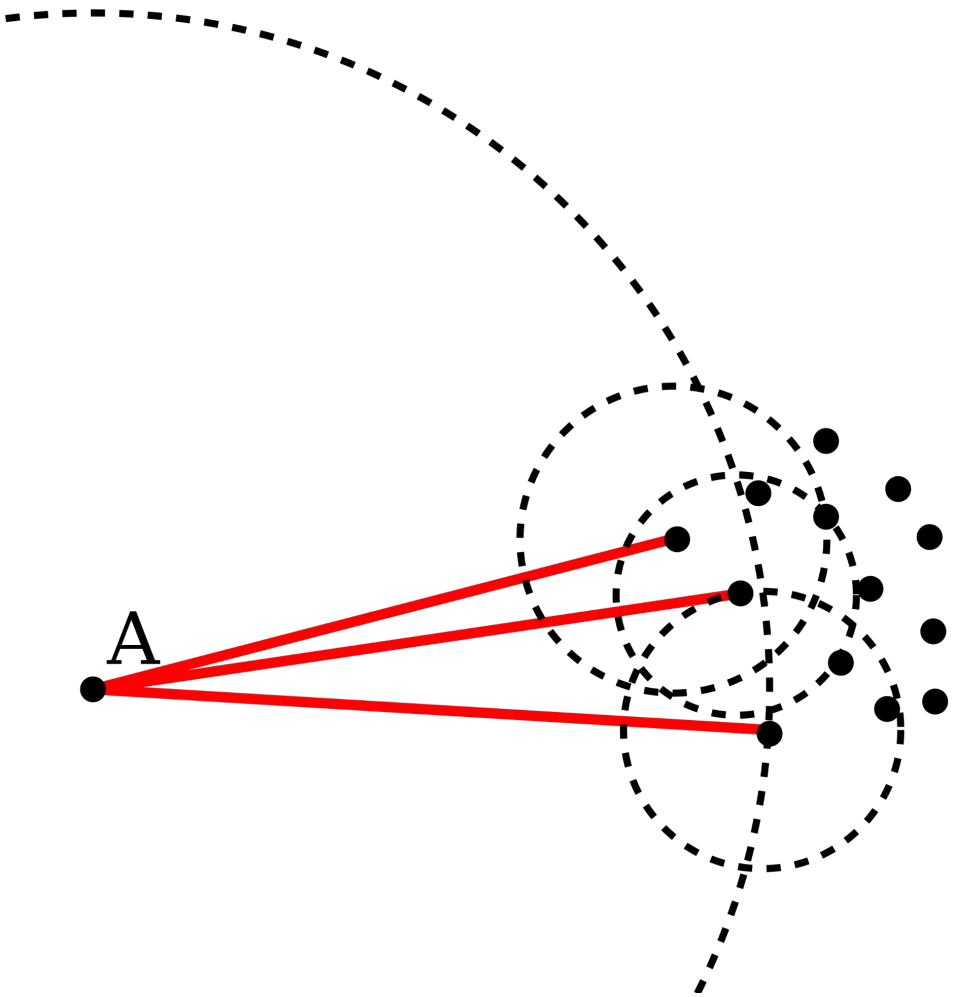
This algorithm operates best with low-dimensional data and offers interpretable results.



Local outlier factor

Local Outlier Factor is another anomaly detection technique that takes the density of data points into consideration to decide whether a point is an anomaly or not. The local outlier factor computes an anomaly score that measures how isolated the point is with respect to the surrounding neighbourhood. It takes into account the local as well as the global density to compute the anomaly score.

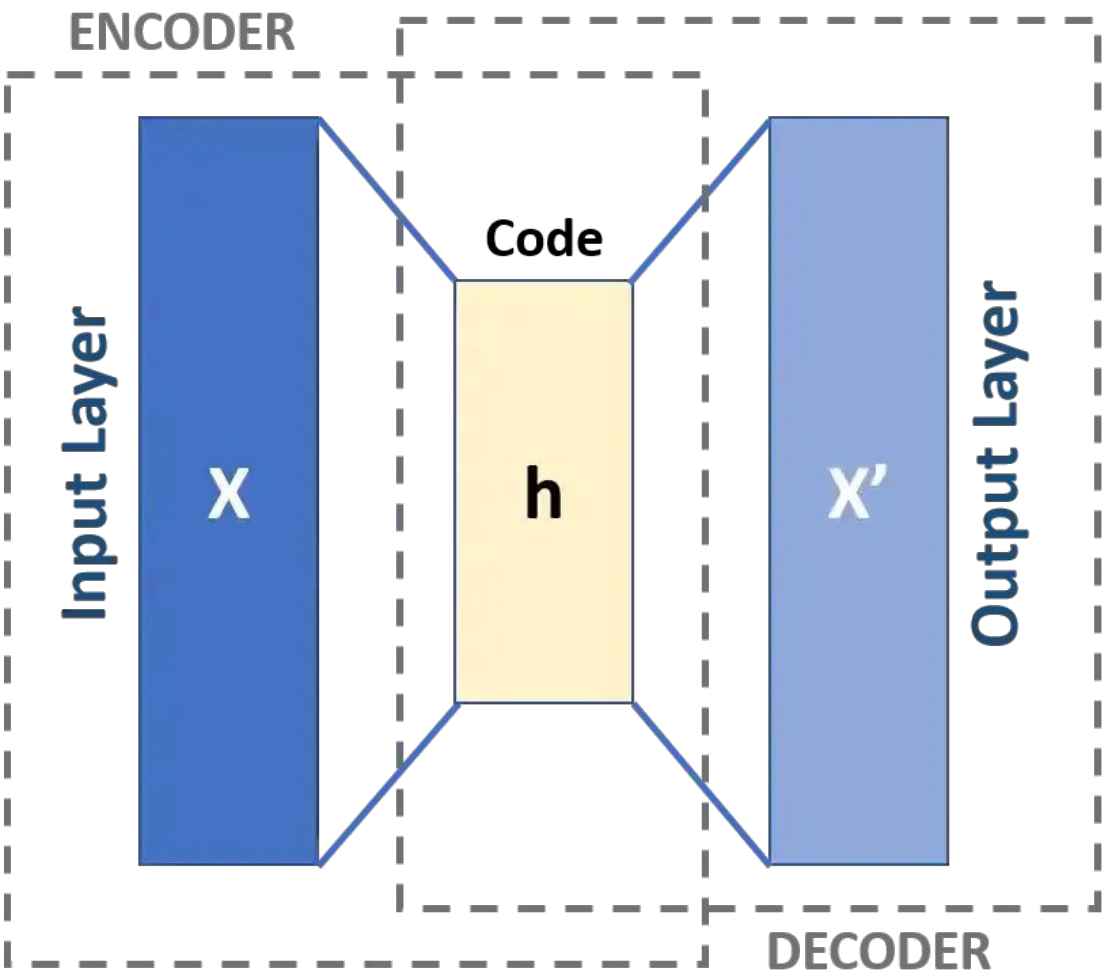
This algorithm comes with similar benefits and limitations to the K-nearest neighbour. Arguably the latter is often preferred due to its simplicity, although LOF can yield more accurate results under careful design.



Autoencoders

Autoencoders are an unsupervised Artificial Neural Network that attempts to encode the data by compressing it into the lower dimensions and then decoding the data to reconstruct the original input. The bottleneck layer (or code) holds the compressed representation of the input data. The number of hidden units in the code is called code size. The reconstruction errors are used as the anomaly scores.

Autoencoders can leverage the full power of deep learning models while not requiring large amounts of labelled data. For this reason, they are a state-of-the-art solution for anomaly detection.

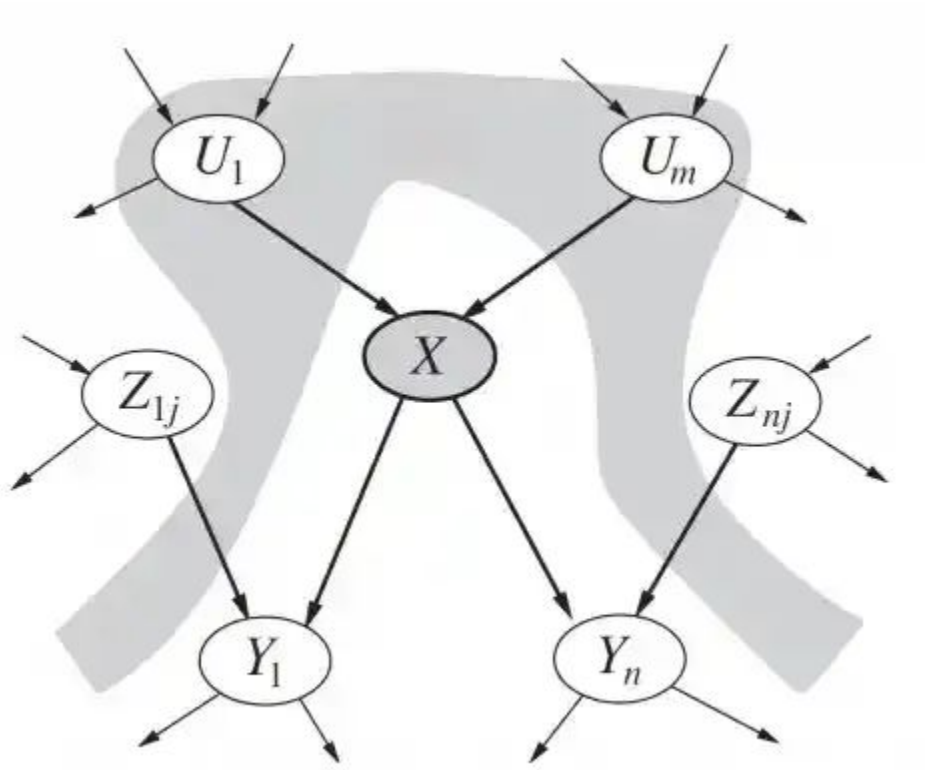


Bayesian Networks

Bayesian networks are probabilistic graphical models that can help you understand the causal relationship between problem variables.

To use them you need to identify the random variables of interest, such as user traffic and product prices. Then, the algorithm will automatically create a graph that captures how variables interact with each other under normal operation. They can also help you answer counterfactuals and anticipate the effect of interventions.

The strength of Bayesian Networks lies in their ability to provide causal relationships, while most machine learning algorithms only identify correlation.



There is no one size fits all solution for choosing the right machine learning algorithm. As we have emphasised in our description, different approaches come with different benefits and limitations.

8 Frequently Asked Anomaly Detection Questions

What is anomaly detection used for?

Anomaly detection can be useful in intrusion detection, fraud detection, health monitoring and defect detection.

What does anomaly detection mean?

Anomaly detection is the process of finding data points that fall outside of the range of usual behaviour.

Anomaly detection is the process of finding data points that fall outside of the range of usual behaviour. What is an anomaly detection example?

An intrusion detection algorithm that automatically alerts when the traffic on a server is unusual is an example of anomaly detection.

What are the types of anomaly detection?

Based on their content, anomalies are classified into expected and unknown anomalies. Based on their nature, anomalies are classified into outliers, change in events and drifts.

What are the three basic approaches to anomaly detection?

Depending on the data availability, anomaly detection may follow an unsupervised, semi-supervised or supervised approach.

9 Conclusion

While anomalies make up a small part of your data, they are the most pressing to address.

This is because abnormal behaviour can affect vital business aspects, such as product quality and user experience.

You cannot put all anomalous data in the same basket as they may be caused by very different processes and require specific treatment.