Building a Model to Automate Anomaly Detection

Pratheerth Padman
Freelance Data Scientist

Module Overview

Anomaly detection techniques:

- STL decomposition
- Classification and regression trees
- Clustering-based anomaly detection
- Anomaly detection using autoencoders

Demo: Introduction to the problem and dataset

Demo: Exploratory data analysis and data cleaning

Demo: Building a model for anomaly detection

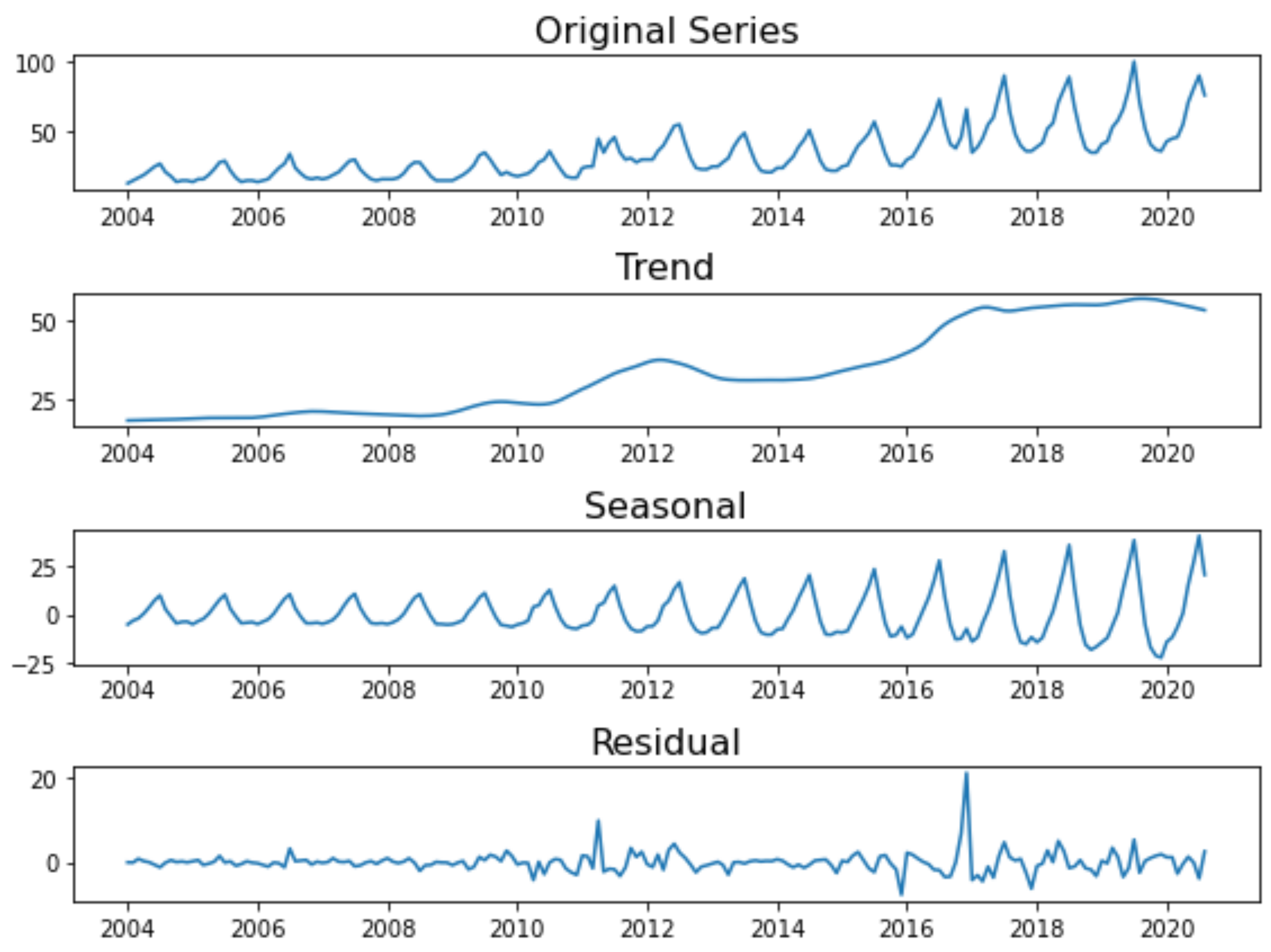
STL Decomposition

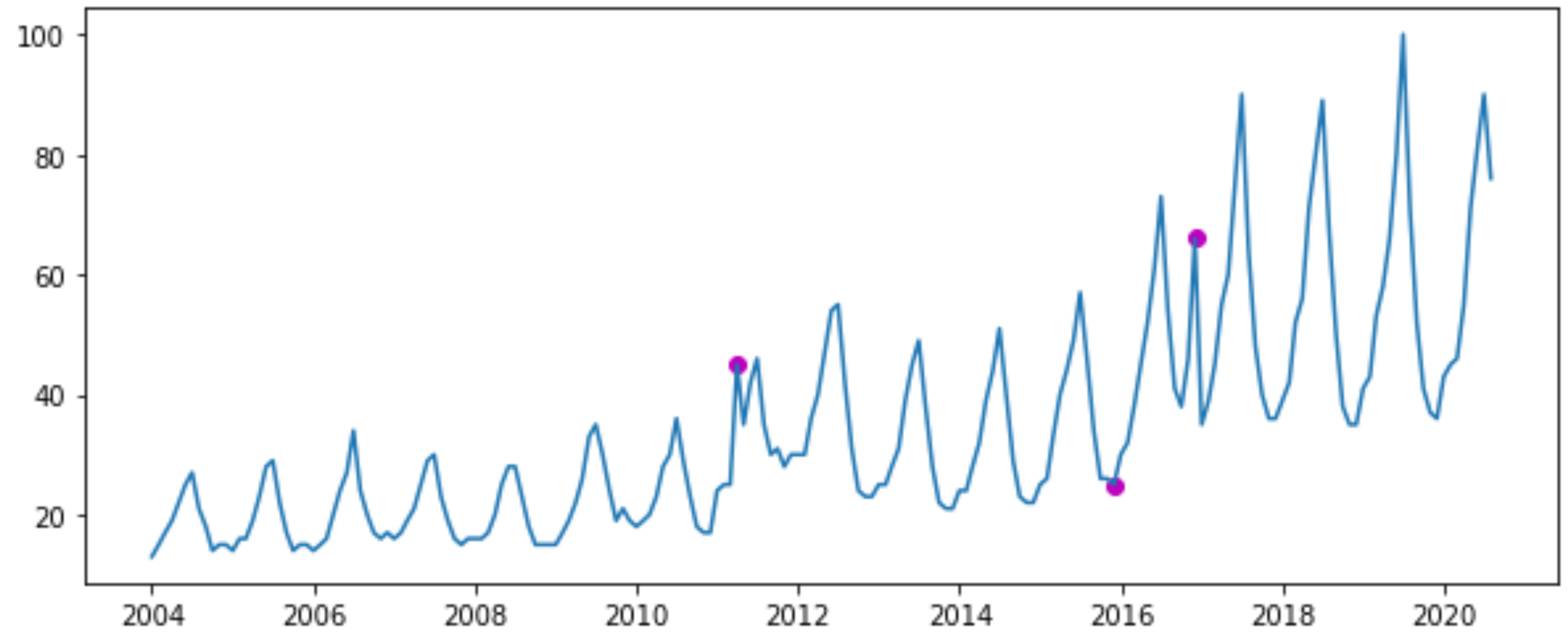
What Is STL?

STL → Seasonal and Trend decomposition using LOESS

Time Series →
- Seasonal
- Trend
- Residue

STL Decomposition





STL Decomposition – Pros and Cons

Simple, robust, and can handle lots of different situations

Rigid tweaking options

Threshold and confidence interval are the only things you can control

Classification and Regression Trees (CART)

Classification and Regression Trees

Utilizing the power of decision trees to identify anomalies

Supervised Learning

Unsupervised Learning

Classification and Regression Trees

Utilizing the power of decision trees to identify anomalies

Supervised Learning

Unsupervised Learning

Isolation forest

Isolation Forest

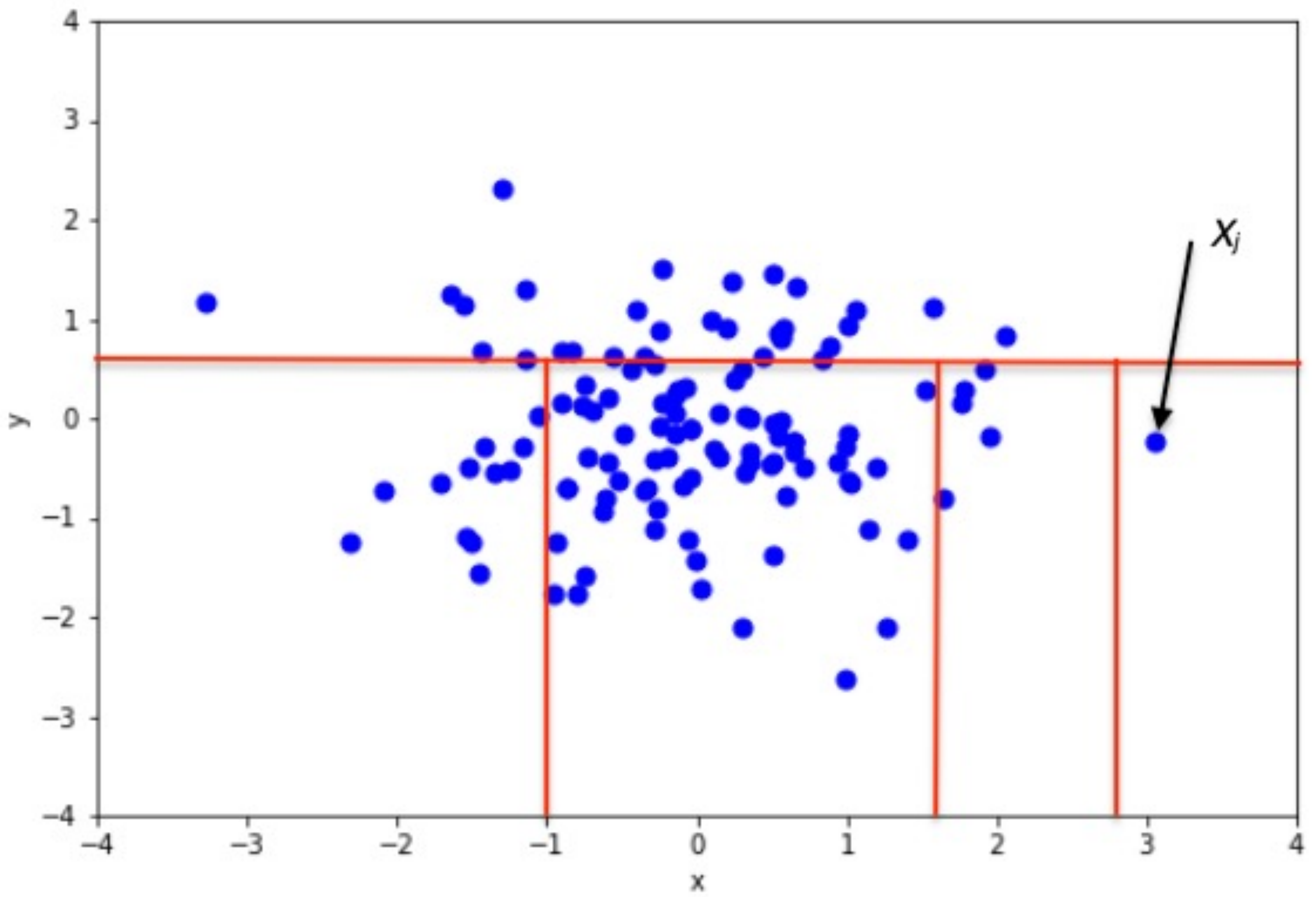
Build based on decision trees

Based on the fact that anomalies are “few and different”

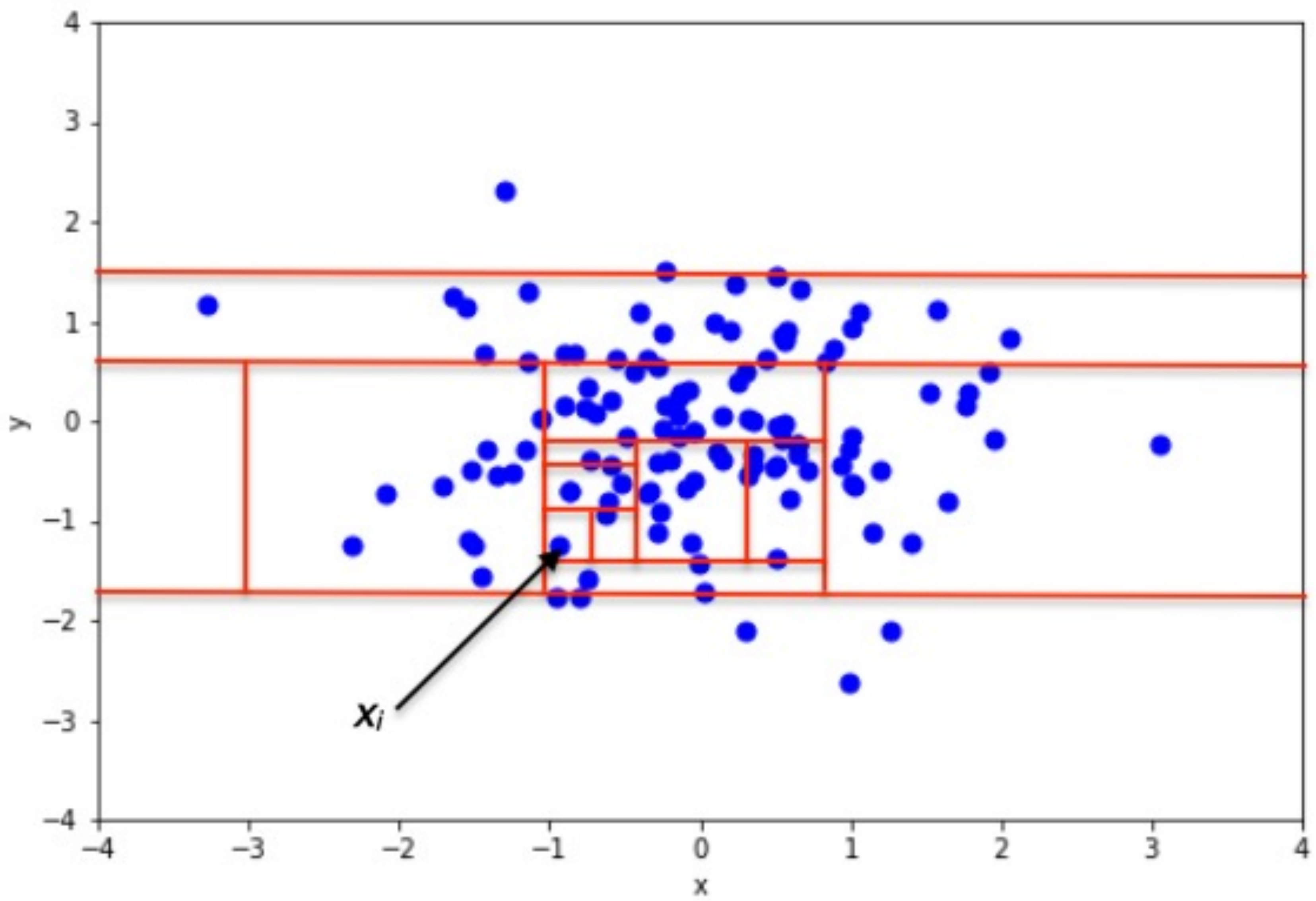
Ensemble of binary trees and each tree is called an isolation tree

Make partitions so that each data point is isolated

Isolation Forest



Isolating an anomalous point



Isolating a normal point

Isolation Forest Algorithm

Training: Building a forest of isolation trees (iTree)

- Take a sample of the dataset and build iTrees until each point is isolated
  - Randomly select a feature
  - Randomly partition along the range

During prediction, an “anomaly score” is assigned to each of the data points based on depth of tree required to arrive to that point

-1 to anomalies and 1 to normal

Isolation Forest – Pros and Cons

Introduce as many random variables or features as you like

Growing number of features can start to impact computational performance

Clustering-based Anomaly Detection

Clustering

Clustering is an unsupervised machine learning technique of dividing a dataset into a number of groups or clusters such that data points in the same cluster are similar to each other

Clustering-based Anomaly Detection

Normal data points belong to large and dense clusters while outliers belong to small and sparse clusters, or do not belong to any clusters

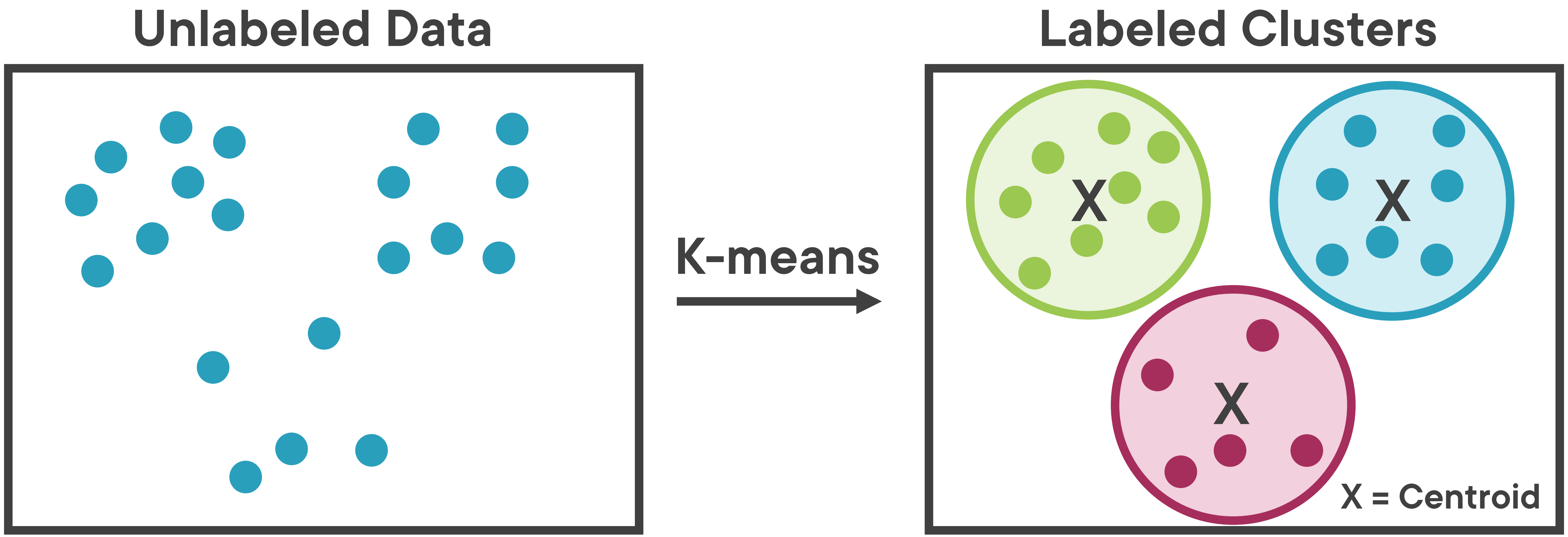
Does data point belong to any cluster? If no - outlier

Large distance between data point and cluster? If yes - outlier

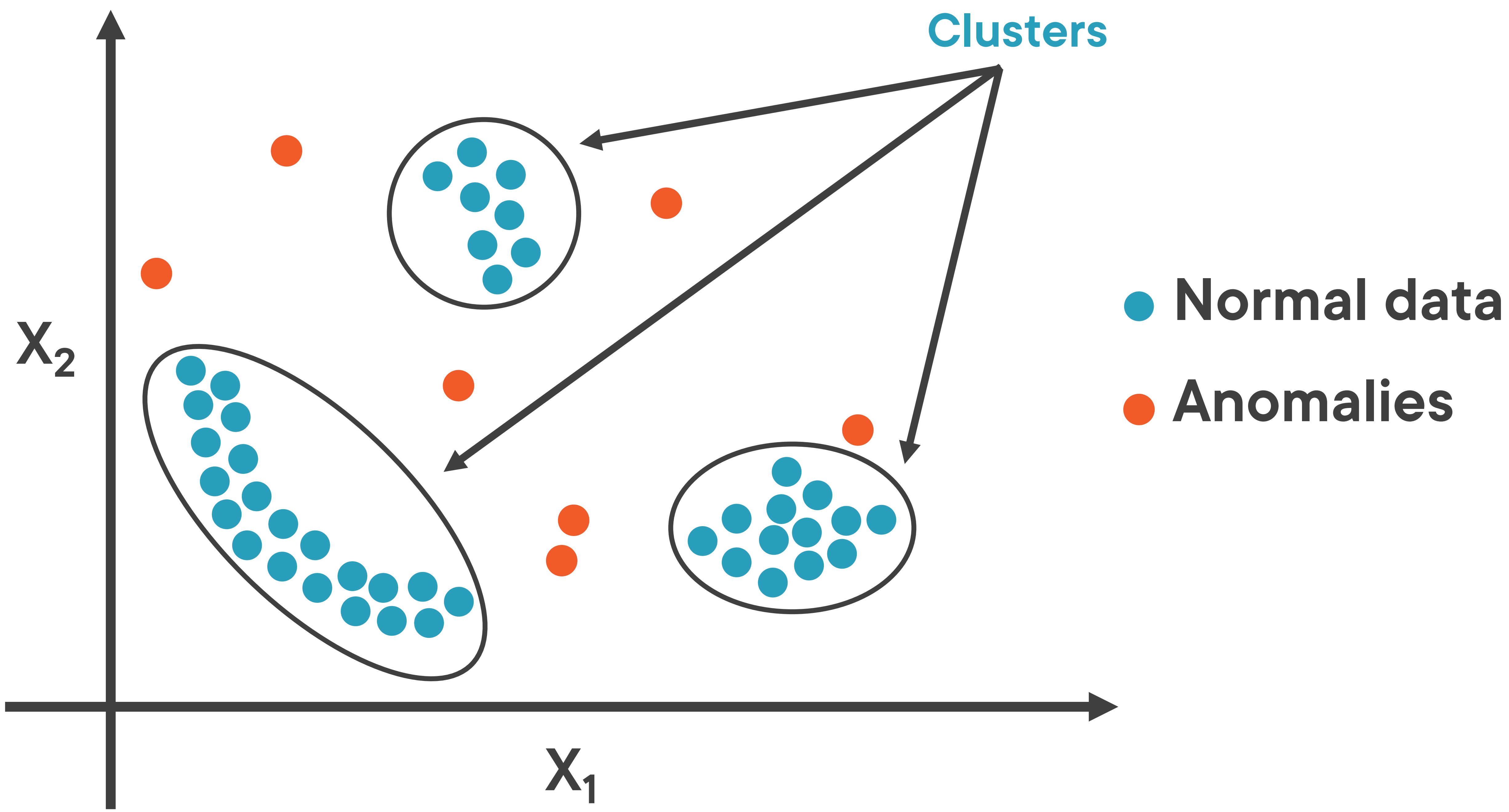
Does data point belong to a small or sparse cluster? If yes - outlier

K-means Clustering

K-means is an unsupervised clustering algorithm designed to partition unlabeled data into a certain number – denoted by "K" - of distinct groupings.



Anomaly Detection with K-means Clustering



Clustering-based Anomaly Detection – Pros and Cons

You can introduce as many variables or features as you like to make it a more sophisticated model

Growing number of features can affect computational performance

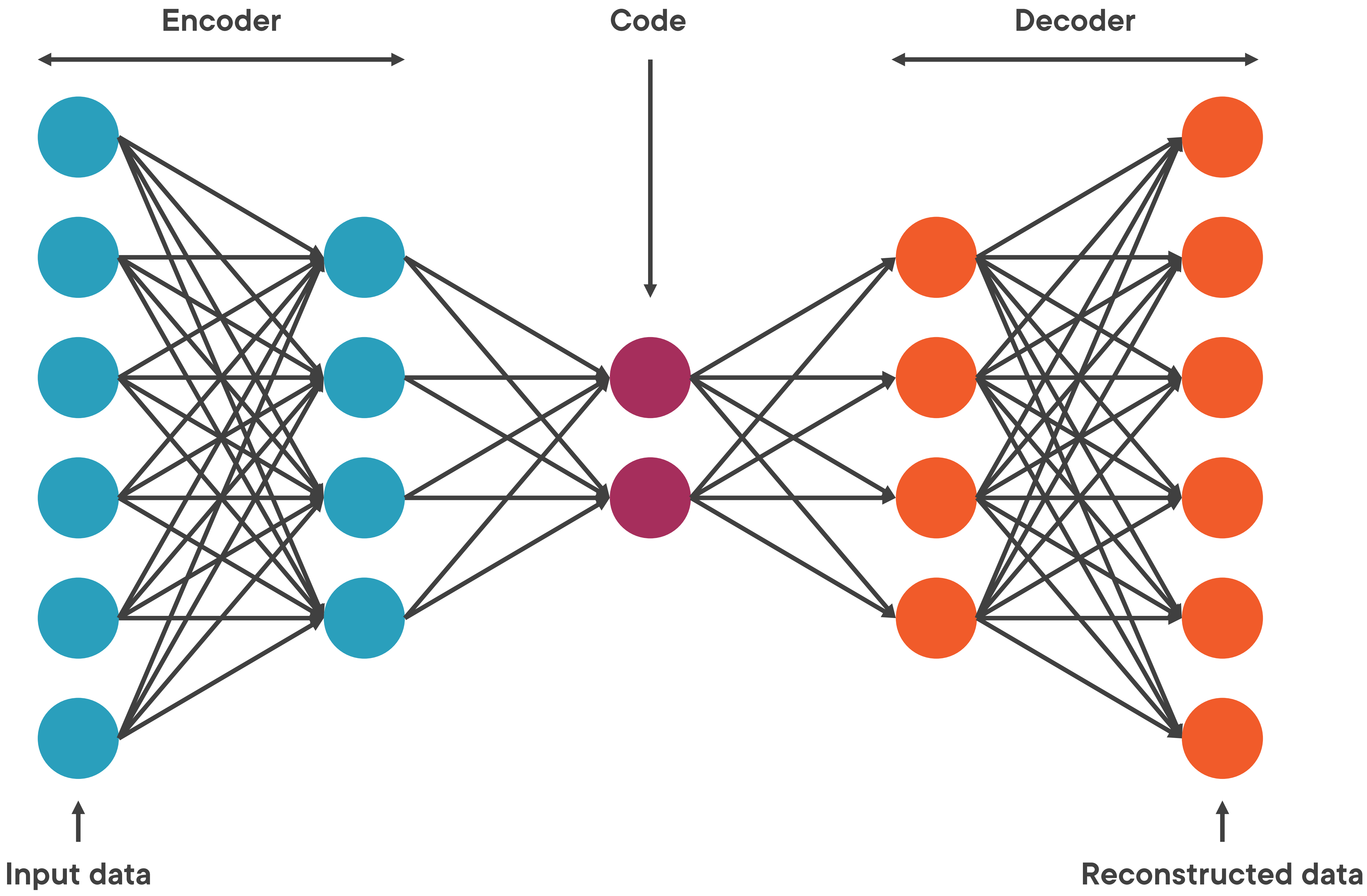
More hyper-parameters to tune, so there is a chance of high model variance in performance

Anomaly Detection Using Autoencoders

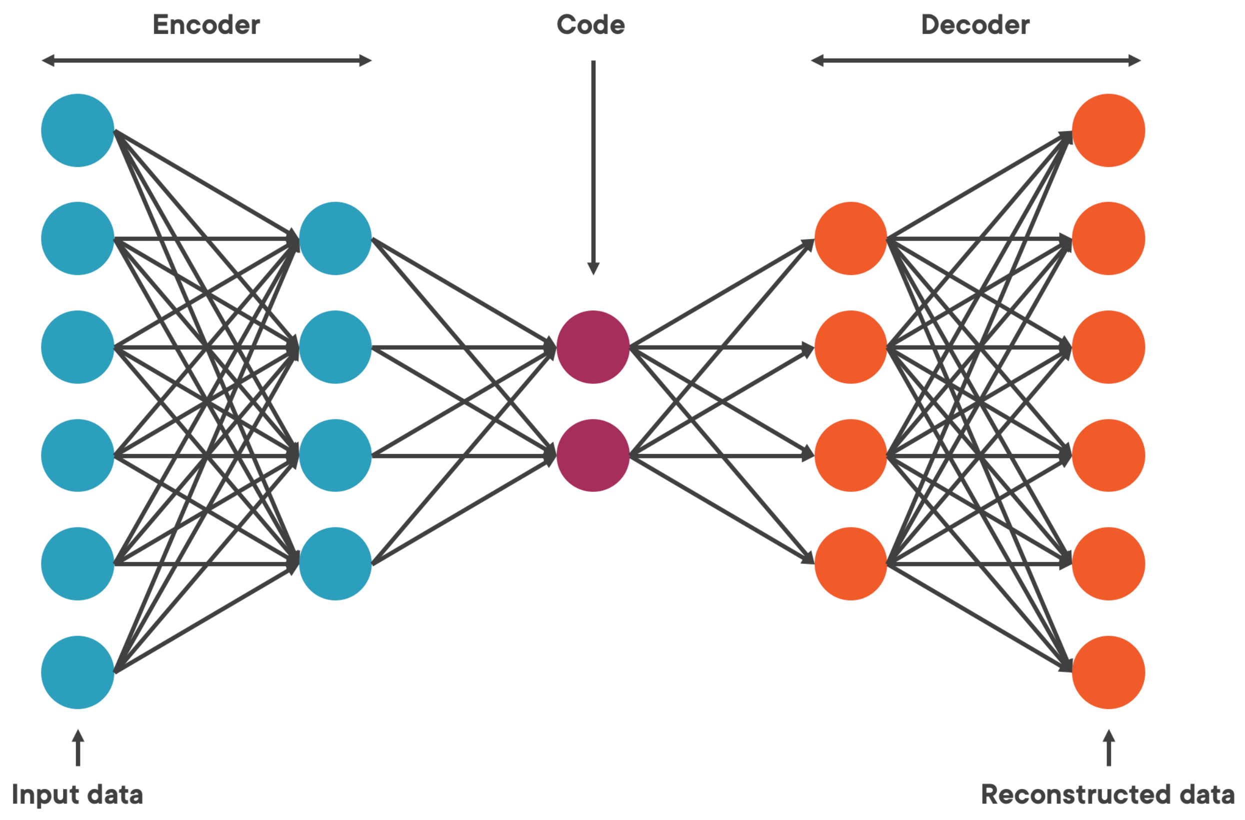
Autoencoders

Autoencoders are a type of neural network in which the input and output are identical. They're typically used for image denoising and dimensionality reduction.

Autoencoders



Autoencoders for Anomaly Detection



To check if observation is anomalous:

- Input the observation into the network
- Measure the error between original observation and the reconstructed one
- Large error between original and reconstructed observation means it is an anomaly

Autoencoders for Anomaly Detection – Pros and Cons

Autoencoders can handle high-dimensional data with ease

Since it is a deep-learning-based strategy, it will struggle if data is less

High computational cost if depth of network increases

Demo

Introduction to problem and dataset

The Problem

Manufacturing industry utilizes various types of heavy machinery such as motors, pumps, pipes, furnaces, etc.

Critical assets for the industry's operations, so their reliability and integrity is often part of the organization's core focus

Failures of these assets often result in production losses that could lead to losses of hundreds of thousands, or even millions of dollars

The ability to detect anomalies in advance and mitigate risk is therefore a very valuable capability

Demo

Exploratory data analysis and data cleaning

Demo

Data preprocessing and dimensionality reduction

Demo

Building a model for anomaly detection

Summary

STL decomposition splits time series into 3 – seasonality, trend, and residue

Isolation forest detects anomalies because they are few and different

K-means says that data points that fall outside certain thresholds are outliers

Autoencoders use the error between original observation and the reconstructed one to mark data points as outliers

We explored, cleaned, and preprocessed a time series dataset and built two anomaly detection models on it

Up Next: Model Evaluation and Dealing with Anomalies