Anomaly Detection based on Operational Data

Aditya Firman Ihsan

Building Resilient Water Systems Workshop through Adaptive Dam Safety, Flood Protection and Management, River Restoration, and Groundwater Management
4 – 5 December 2024

Table of contents

01 Anomaly Detection

Basic about the task of detecting anomaly

02 Machine Learning 101

Learn a bit how machine learns

03 Data-Driven Anomaly Detection

Utilize data and machine learning for detection system

04 Real Case in Oil and Gas Industry

A case study how detection system is developed

01

Anomaly Detection

> “An outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.”
>
> – Hawkins, Identification of Outliers (1980)

Anomaly

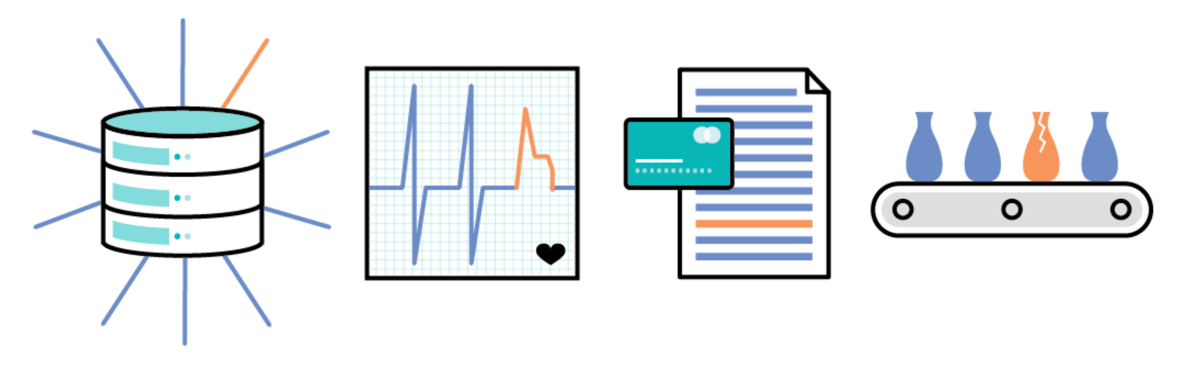
- **Anomalies (**often referred to as **outliers, abnormalities, rare events**, or **deviants**), are data points or patterns in data that do not conform to a notion of normal behavior.

- Anomaly detection is the task of finding those patterns in data that do not adhere to expected norms, given previous observations.

Anomaly Detection

Anomaly detection has applications in a variety of domains, including

- IT analytics,
- network intrusion analytics,
- medical diagnostics,
- financial fraud protection,
- manufacturing quality control,
- marketing and social media analytics,
- and more.



Different Types of Anomalies

1. **Global outliers,** or point anomalies, occur far outside the range of the rest of a data set.

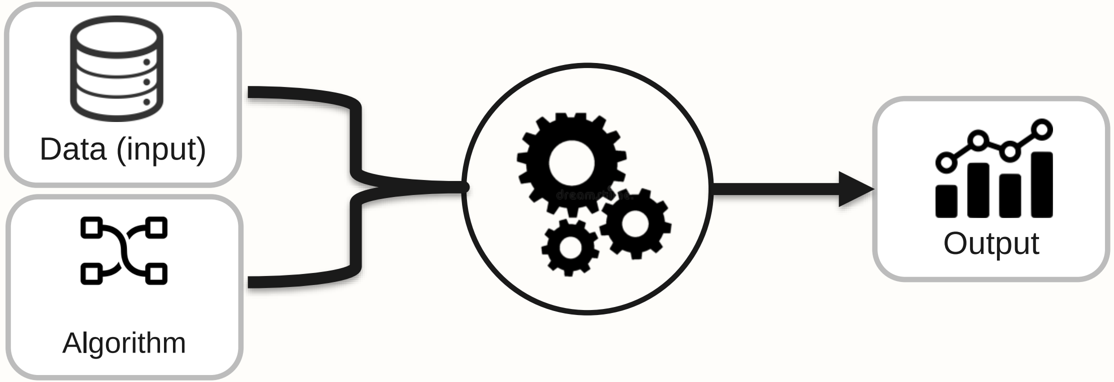
2. **Contextual outliers** deviate from other points in the same context, e.g., holiday or weekend sales.

3. **Collective outliers** occur when a range of different types of data vary when considered together, for example, ice cream sales and temperature spikes.

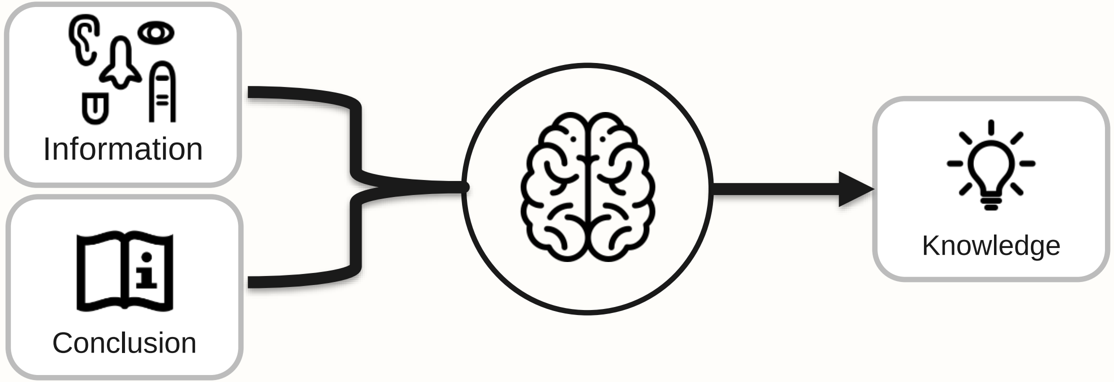
02

Machine Learning 101

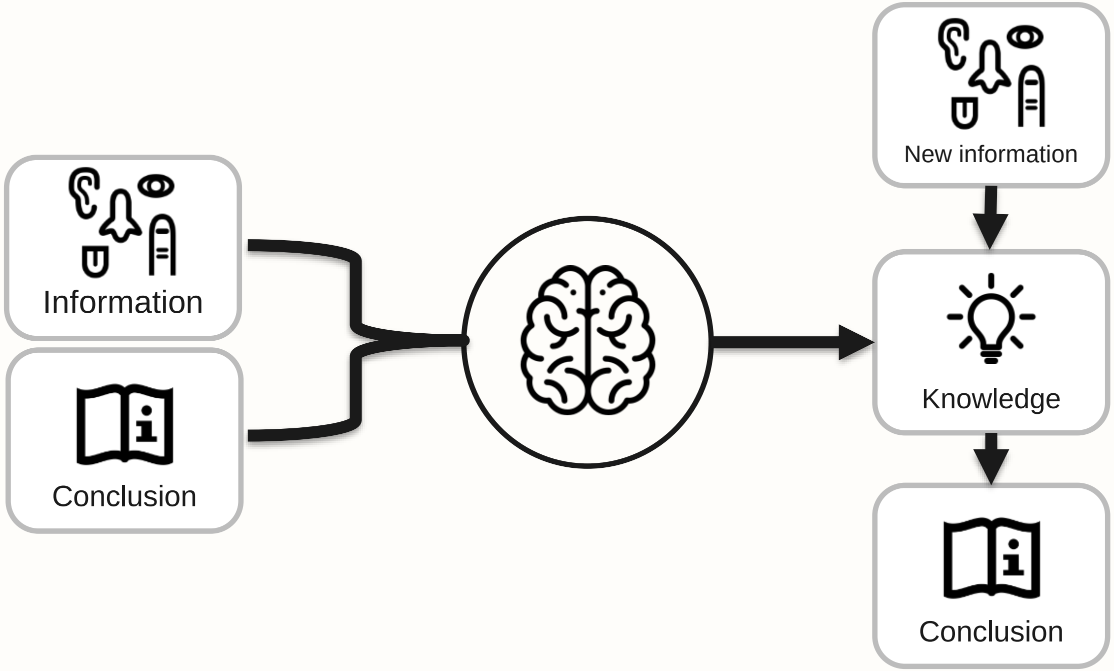
How machine works?



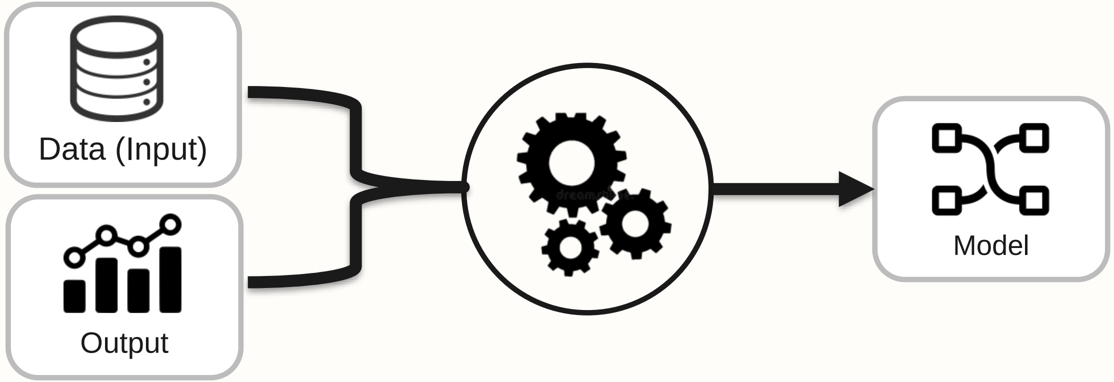
But how does exactly human learn?



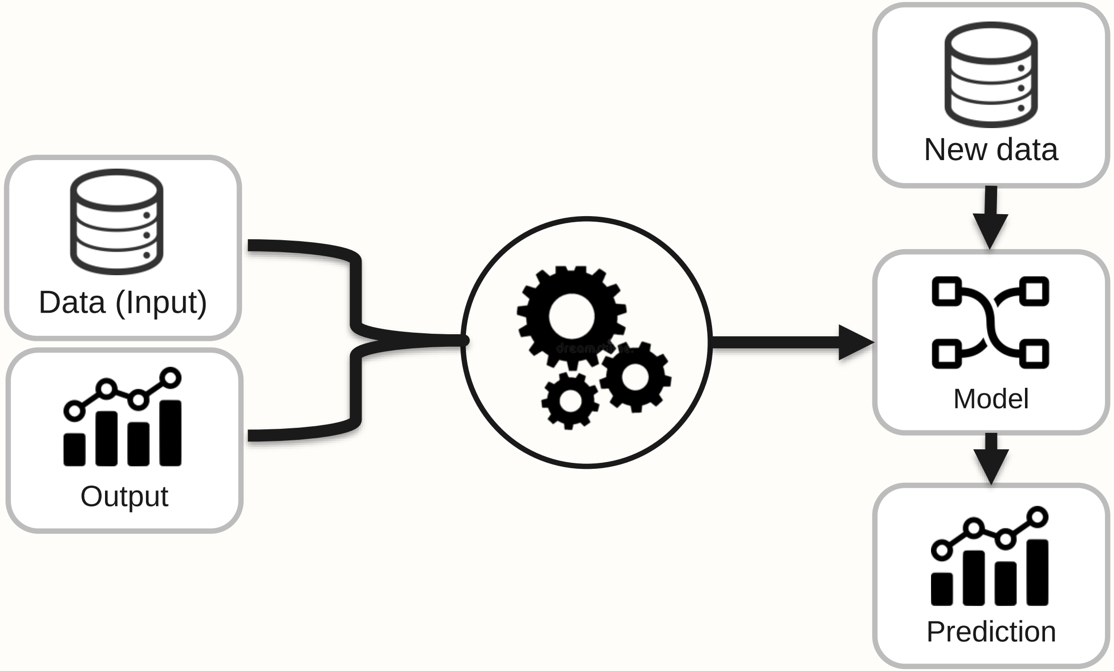
But how does exactly human learn?



We can make machine also “learns”.



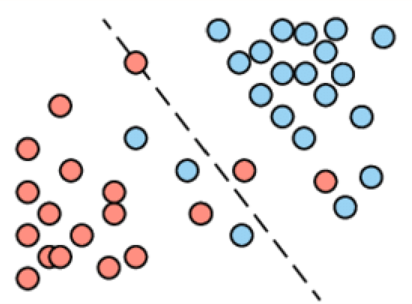
We can make machine also “learns”.

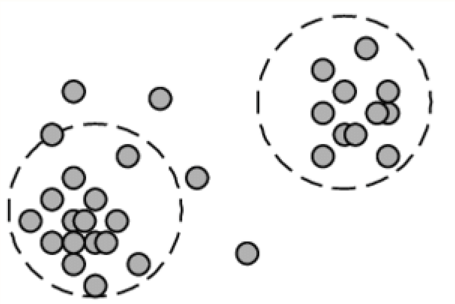


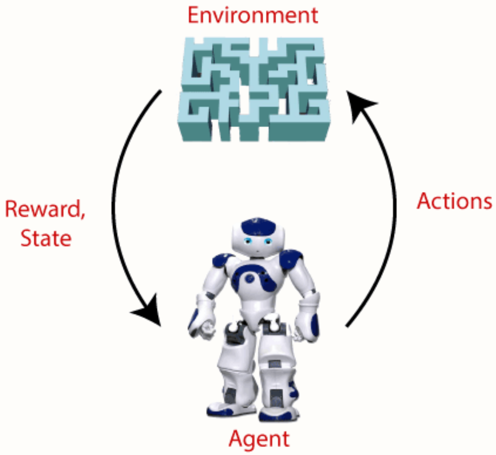
Supervised

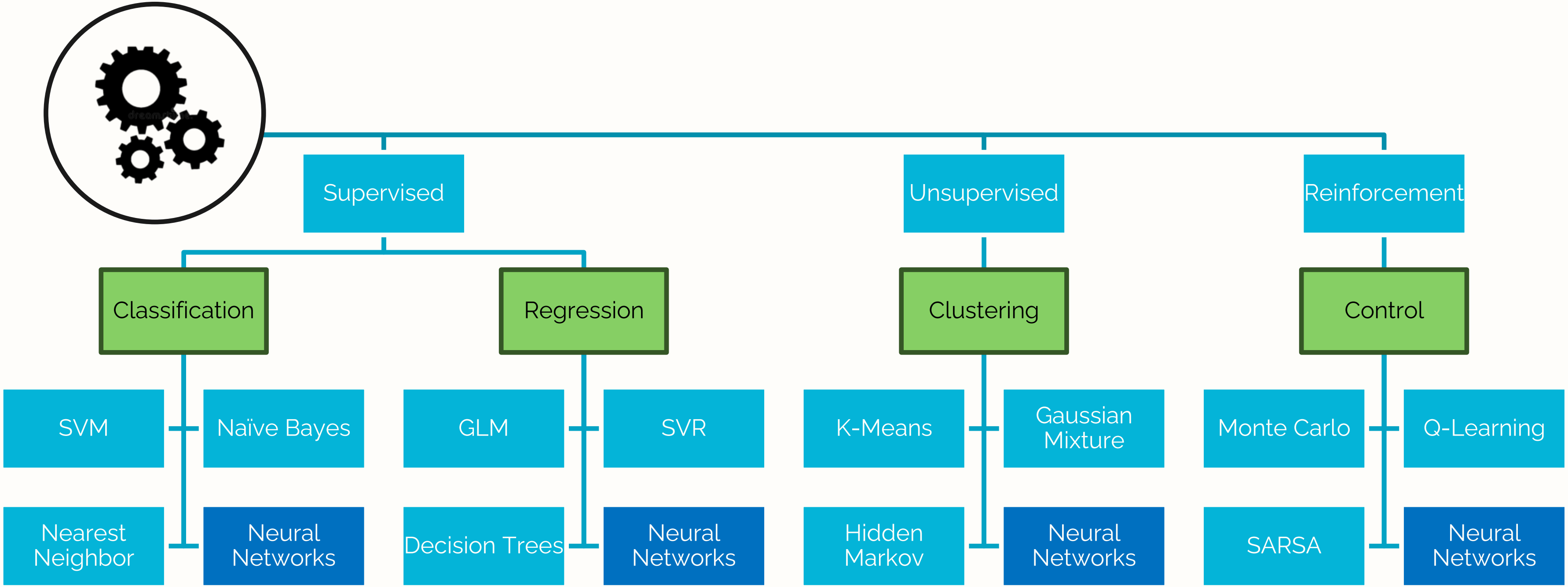
Unsupervised

Reinforcement





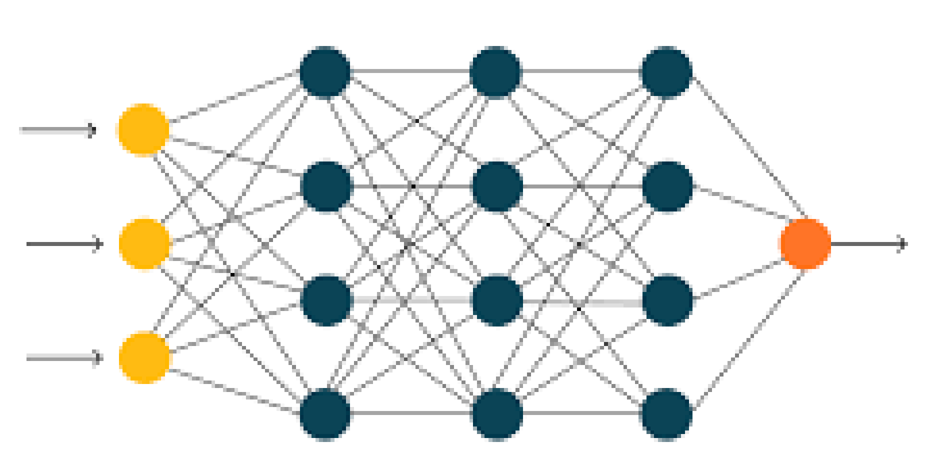


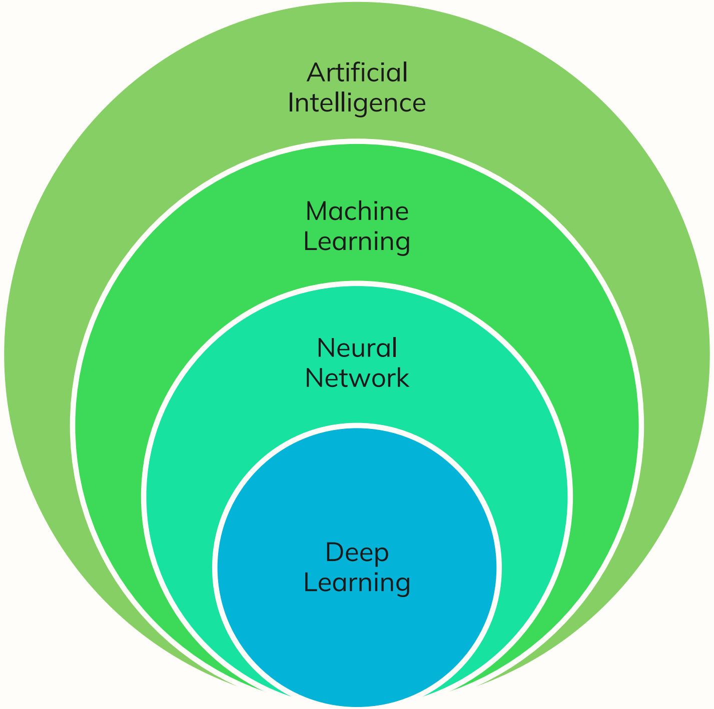


- Supervised
  - Classification
    - SVM
    - Naïve Bayes
    - Nearest Neighbor
    - Neural Networks
  - Regression
    - GLM
    - SVR
    - Decision Trees
    - Neural Networks
- Unsupervised
  - Clustering
    - K-Means
    - Gaussian Mixture
    - Hidden Markov
    - Neural Networks
- Reinforcement
  - Control
    - Monte Carlo
    - Q-Learning
    - SARSA
    - Neural Networks

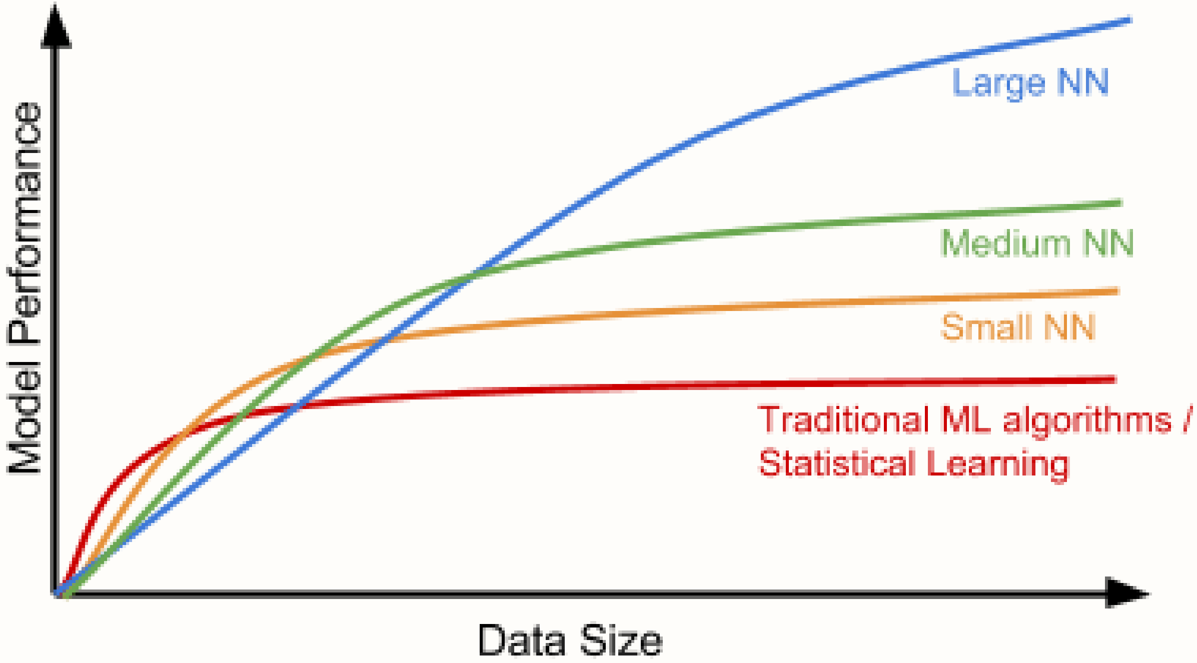
Neural Network as Modern ML

Neural Network becomes the bridge to new paradigm of ML, called deep learning





Neural Network as Modern ML



03

Data-Driven Anomaly Detection

Why Data–driven Detection

- **Processing large datasets in real-time**

AI is capable of efficient processing, labeling and categorization of large datasets in real-time

- **Automated, real-time detection**

AI enables automated, real-time detection of anomalies by consistently monitoring and learning patterns so that AI can quickly detect anomalies as they occur.

- **Effective pattern recognition**

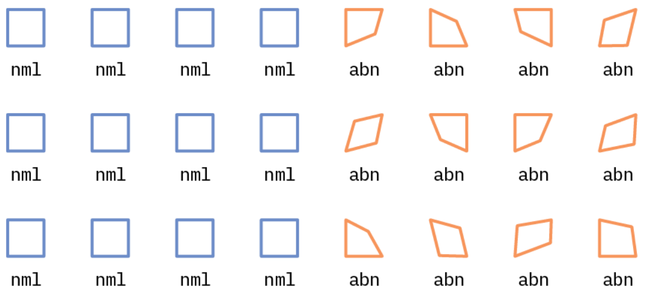
Large datasets exhibit complex behavior that traditional systems may struggle to identify. AI-powered solutions excels in recognizing patterns, learning from them, and accurately identifying any deviations or anomalies.

- **Proactive anomaly management**

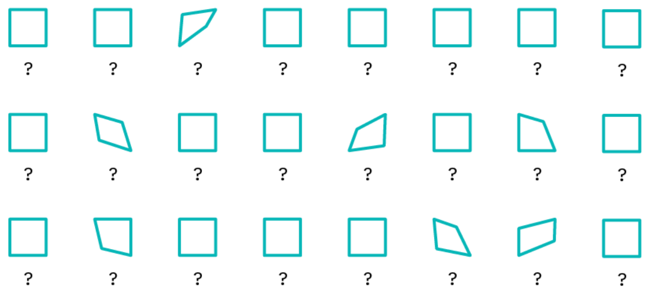
Traditionally, anomalies were handled reactively. AI-powered systems enable a proactive methodology to detect anomalies via closed-loop automation.

ML for Anomaly Detection

Supervised Learning Training

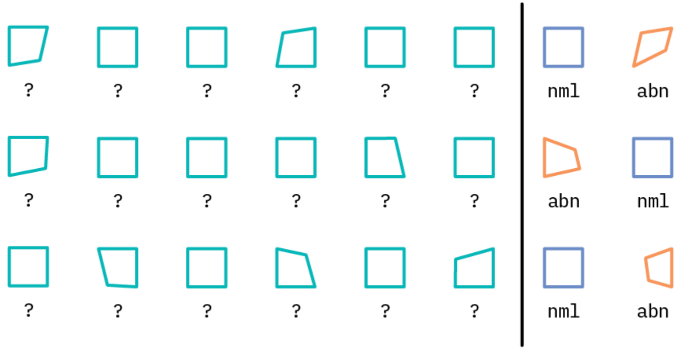

nml nml nml nml abn abn abn abn
nml nml nml nml abn abn abn abn
nml nml nml nml abn abn abn abn

Unsupervised Learning Training



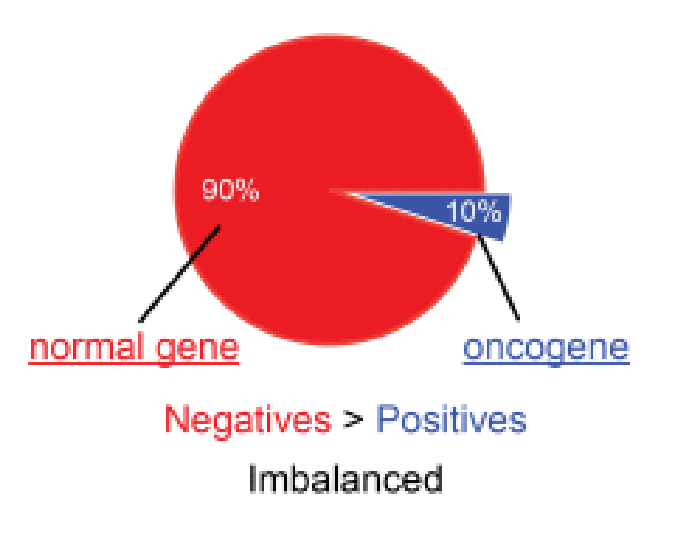
Semi-supervised Training

Testing



Supervised Anomaly Detection

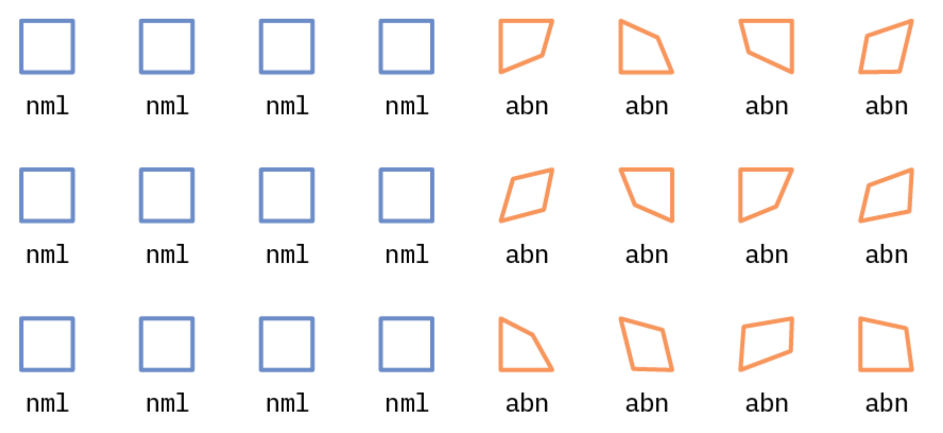
- Problem of imbalanced class -> abnormal class always a minority
- Anomaly classification tends to give bias.
- For example, if the model always give output “normal”, its accuracy is 90%.



Supervised Anomaly Detection

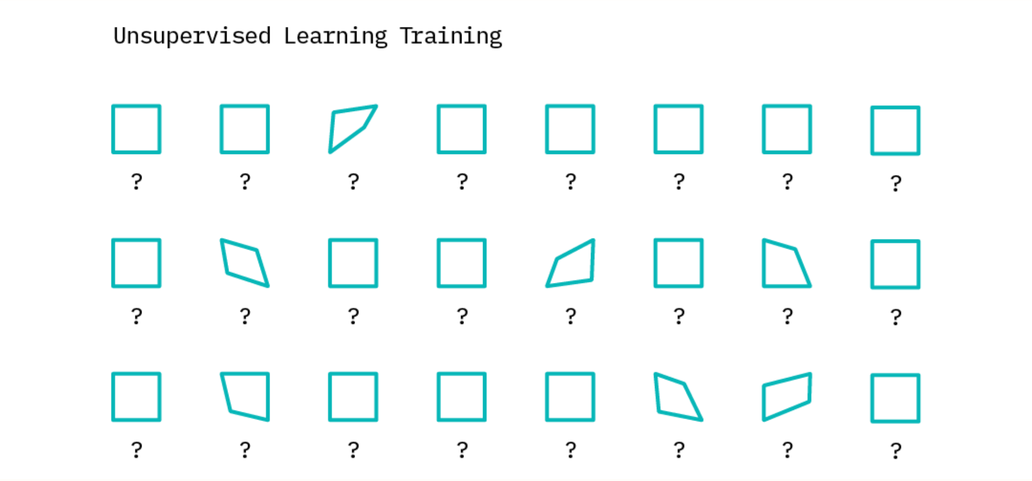
- In supervised paradigm, we need to know in advance which data is anomaly to label it.

- All supervised methods of ML can be applied



Unsupervised Anomaly Detection

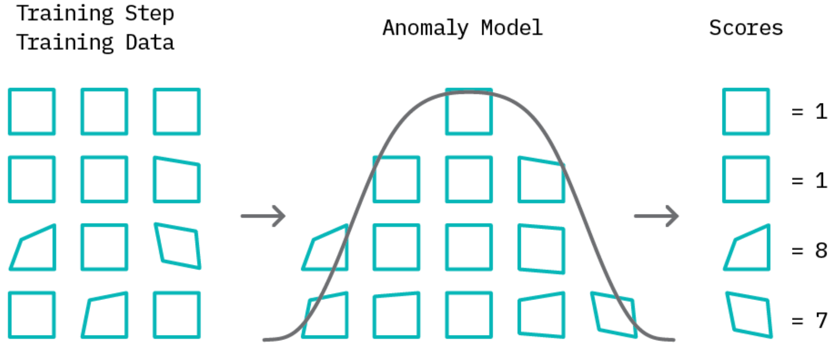
In unsupervised paradigm, anomaly is detected purely on the internal characteristic of the data.



Unsupervised Anomaly Detection

The underlying strategy for anomaly detection is to first **model normal behavior** and then exploit this knowledge to identify deviations.

How to model “normal behavior”?



(1) Statistical approach

1. **Z-score (standard score):** it measures how many standard deviations a data point is away from the mean. Generally, instances with a z-score over 3 are chosen as outliers.

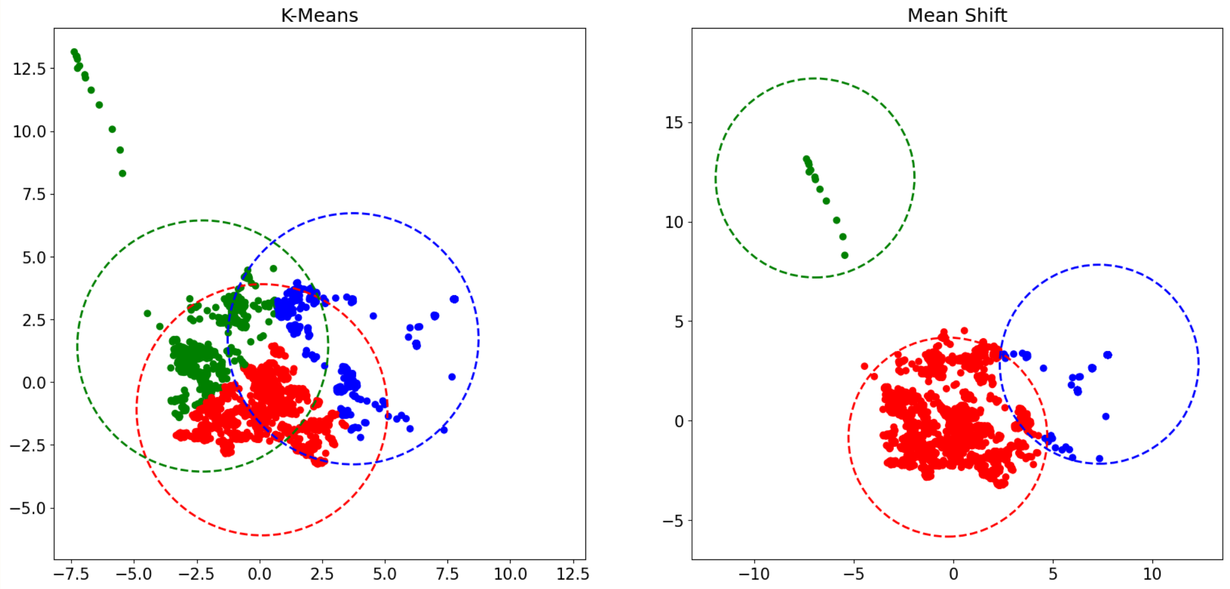
2. **Interquartile range (IQR):** When an instance is beyond Q1 or Q3 for some multiplier of IQR, they are considered outliers. The most common multiplier is 1.5, making the outlier range.

3. **Time Series decomposition:** Data can be decomposed to multiple components with noise/residual. A threshold for the residual can be taken to determine anomalies

(2) “Geometry” Based

We can determine the anomalies as data points that is far from majority of data.

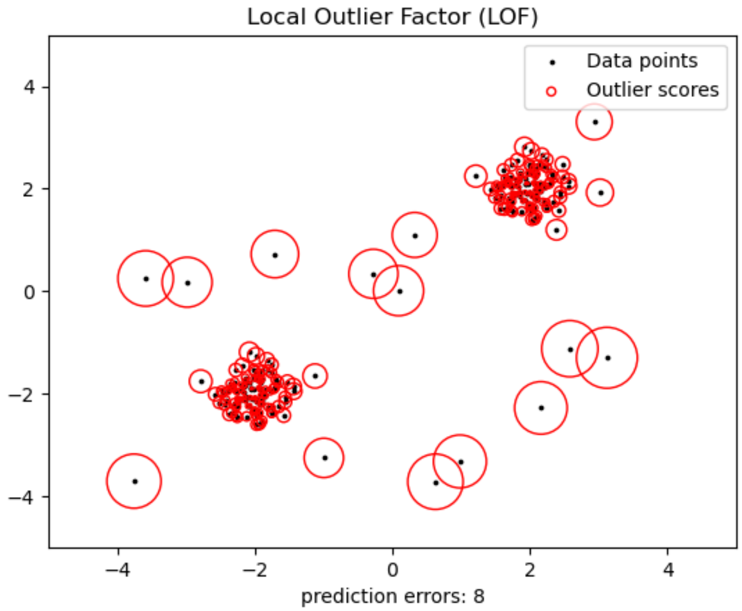
(a) Centroid Clustering



(2) “Geometry” Based

We can determine the anomalies as data points that is far from majority of data.

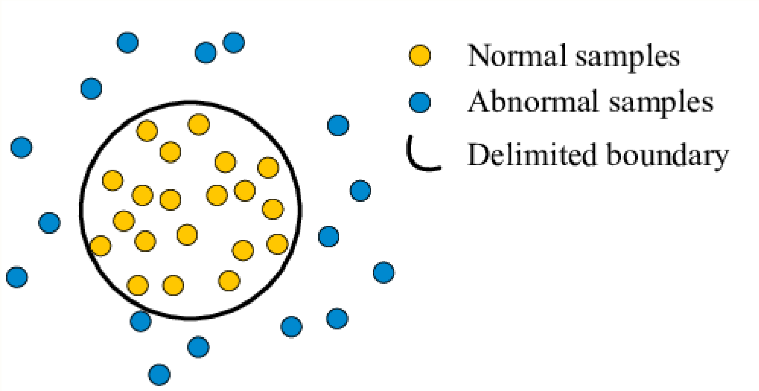
(b) Local Outlier Factor

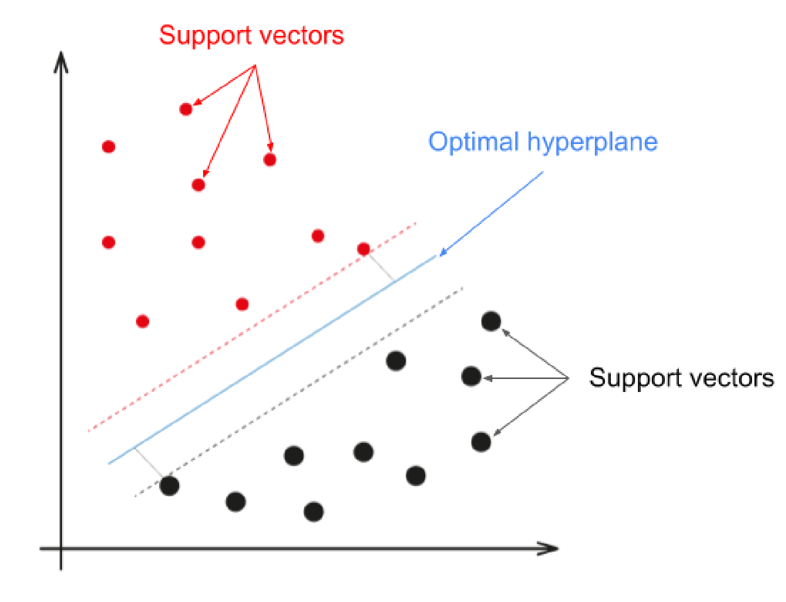


(2) “Geometry” Based

We can determine the anomalies as data points that is far from majority of data.

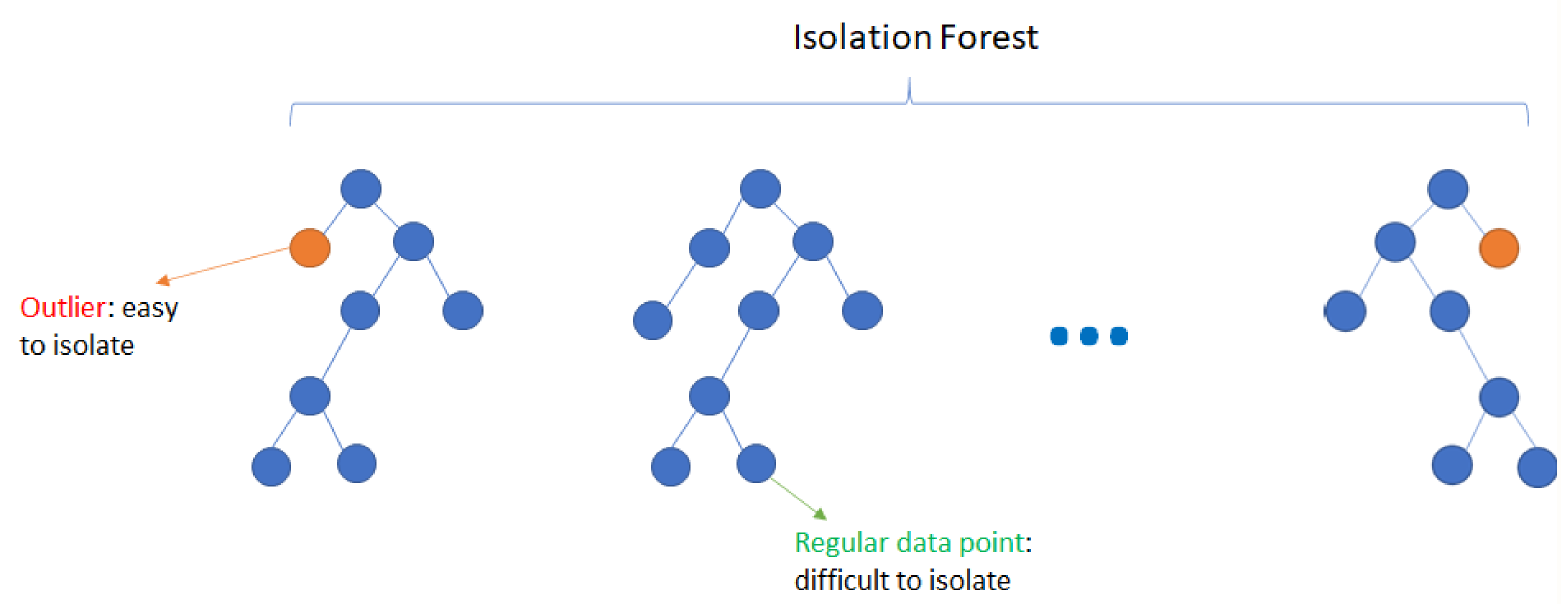
(c) One-Class SVM (OCSVM)





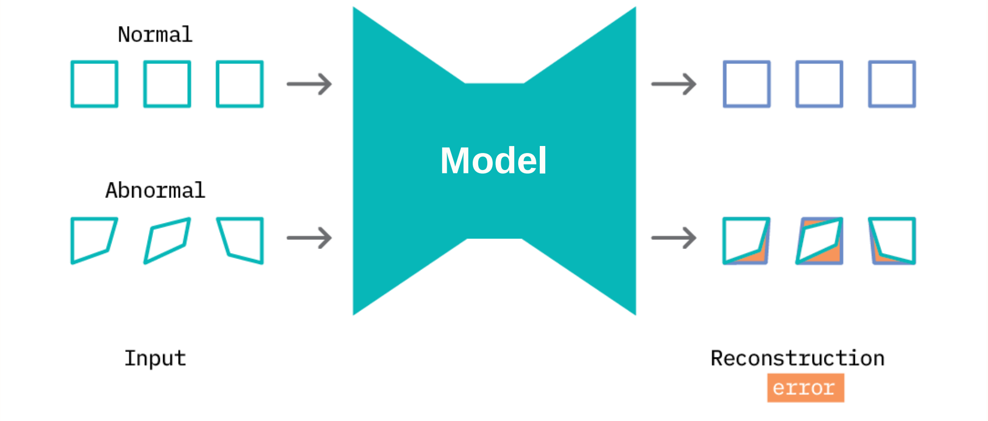
(3) Tree Based

Isolation Forest uses a collection of **decision trees** that recursively divide complex datasets until each instance is isolated. The instances that get isolated the quickest are considered outliers.



(4) Reconstruction Based (Deep Learning)

Reconstruction-based detection works by any **deep learning** model, from the simplest Neural Network to the complex one such as GAN (Generative Adversarial Network)

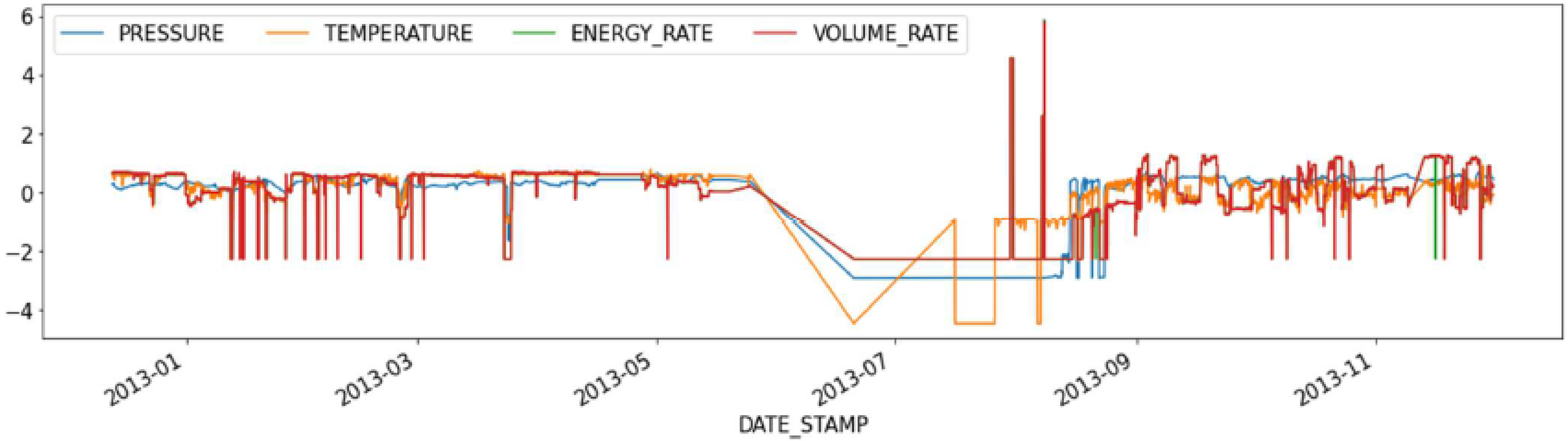


04

Real Case in Oil and Gas Industry

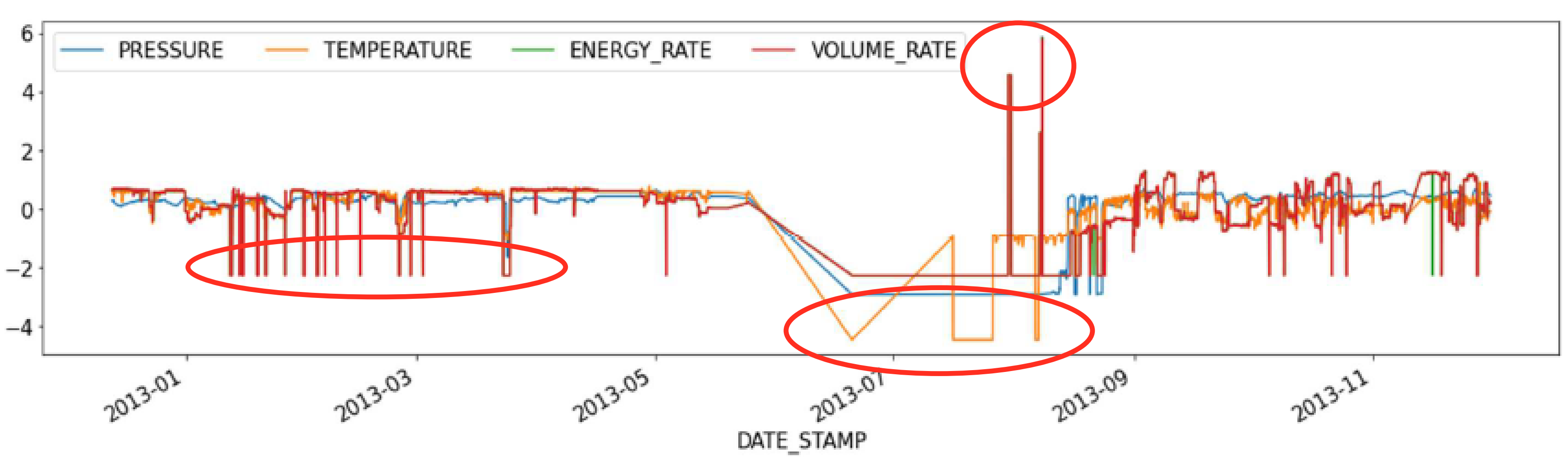
Gas pipeline system monitoring

Pipeline system of natural gas needs to be monitored carefully to make sure gas transmission and delivery is in order. Some operational variables data at some node points are collected real-time.

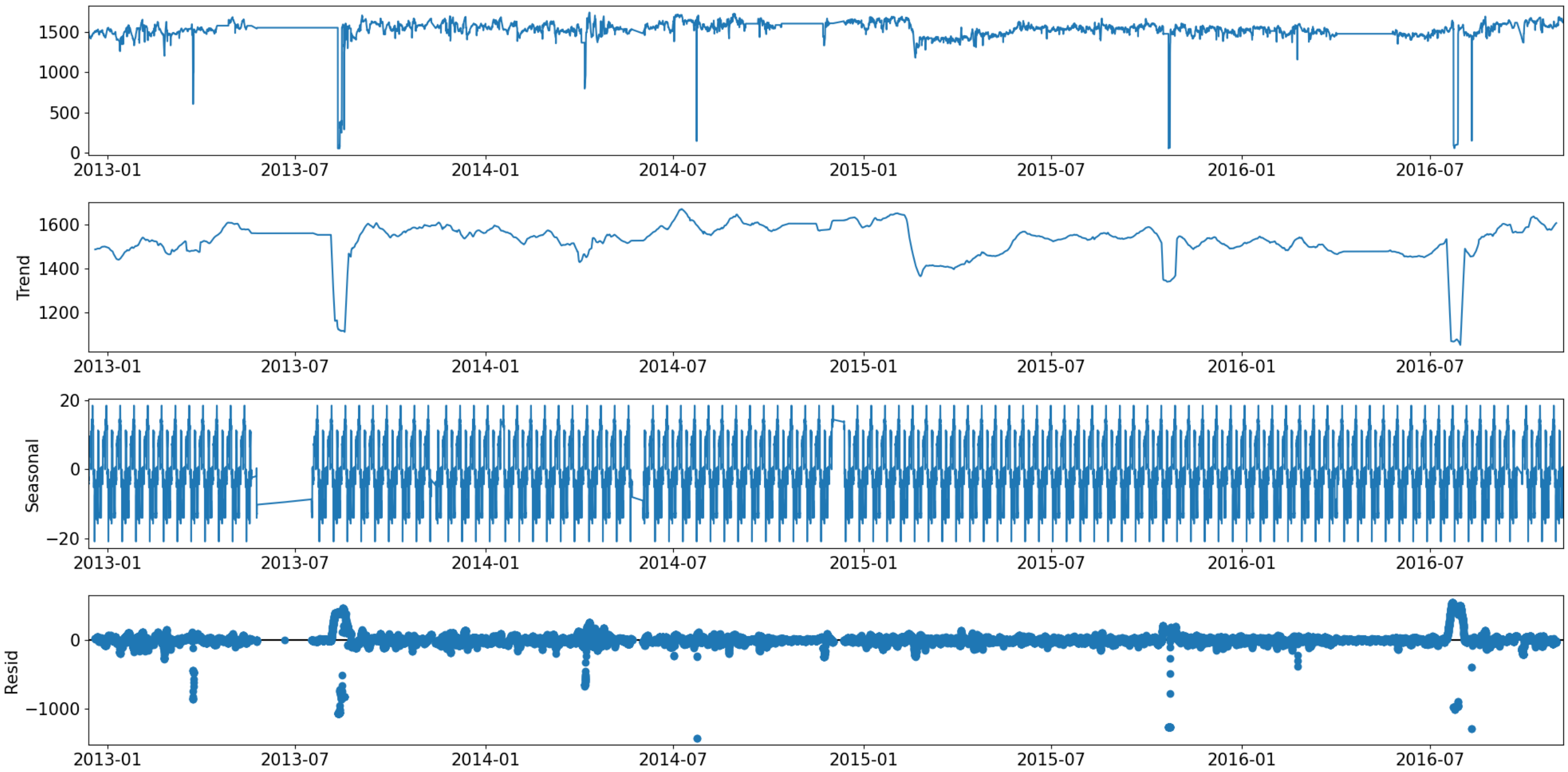


Gas pipeline system monitoring

There are many **potential anomalies** seen in the graph. But make sure anomalies can be distinguished from **“dirty data”**. Dirty data may affect “normal behavior”, so it needs to be cleaned first.

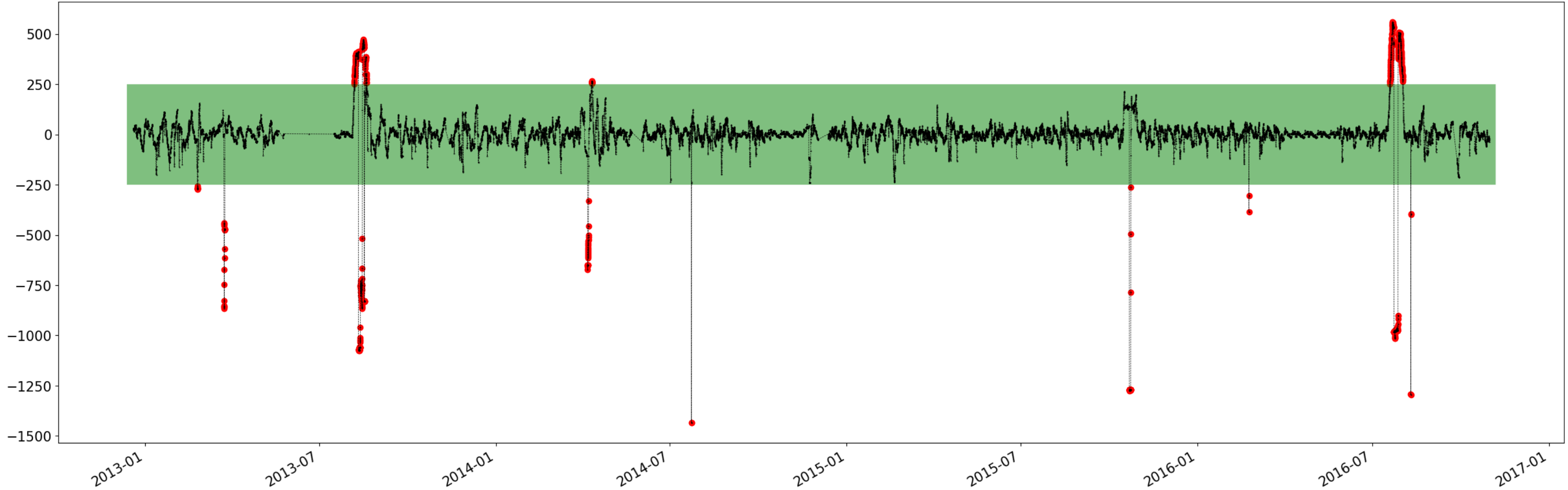


Time Series Decomposition



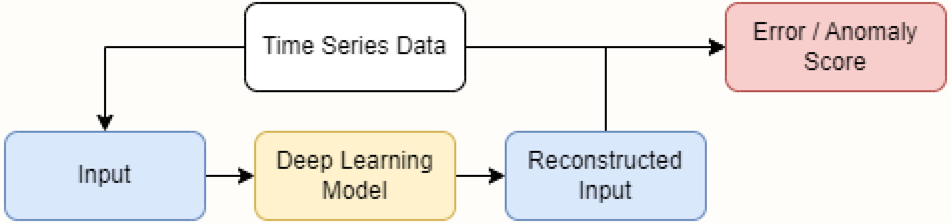
Anomalies from Residual

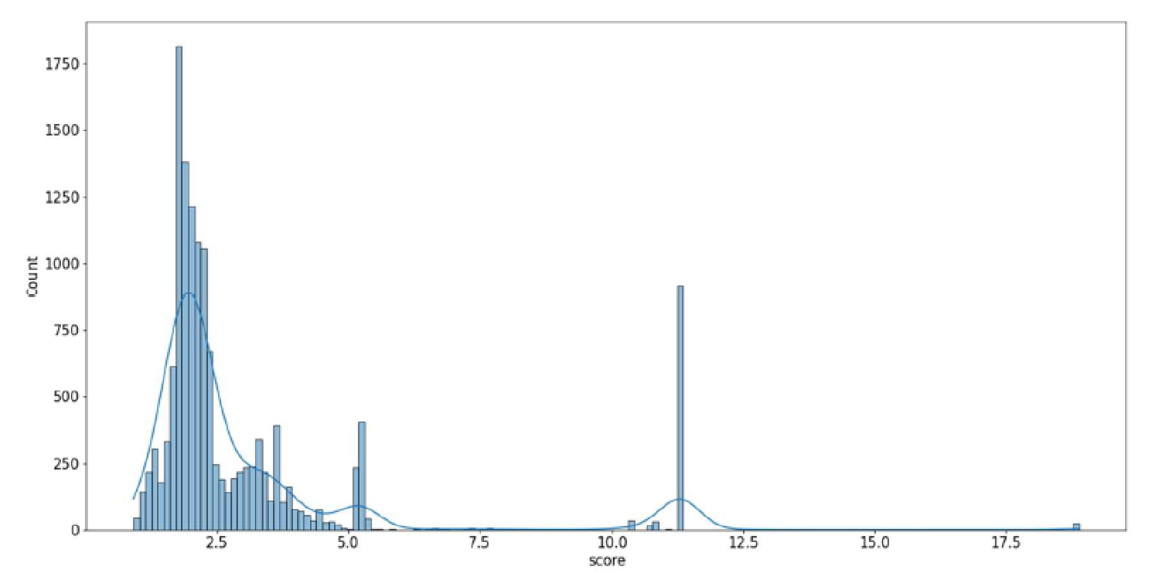
By time series decomposition, we can set a threshold to determine anomalies outside the threshold.



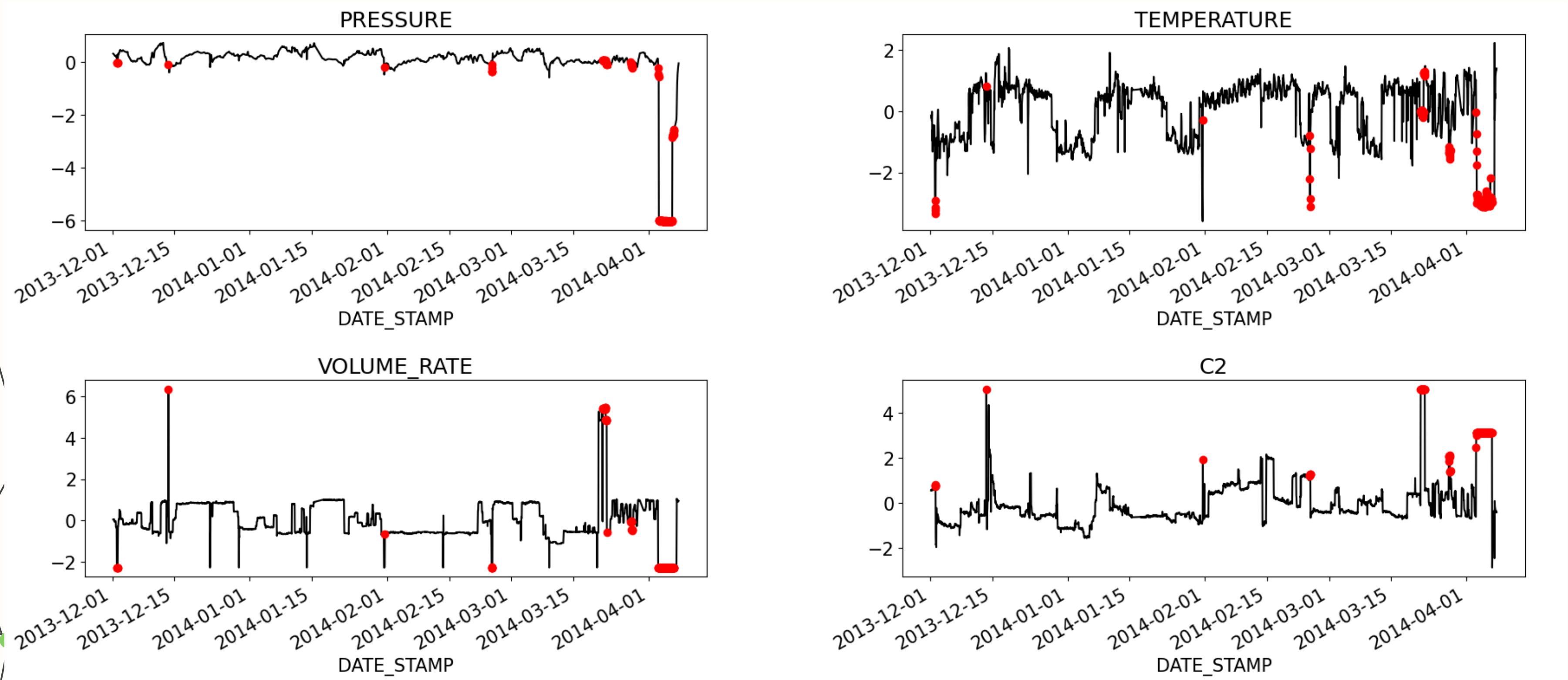
Anomalies from Reconstruction Score

If we train a deep learning model to reconstruct the data, we can compute its





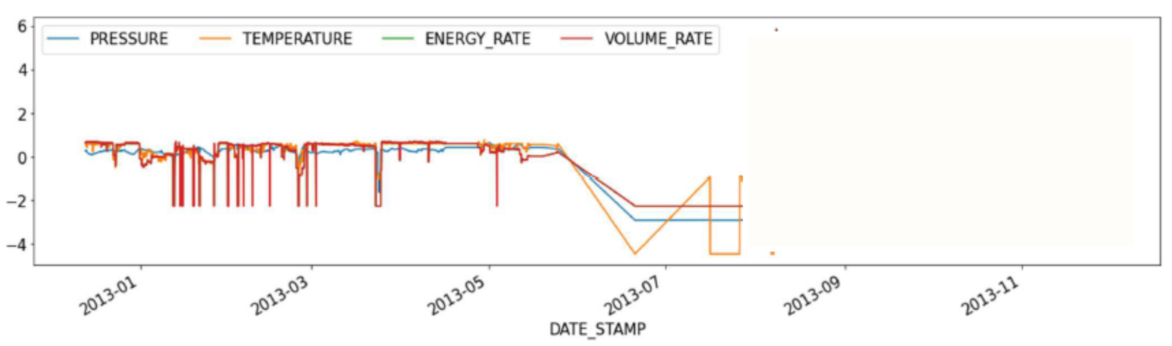
Intertwine of Anomalies



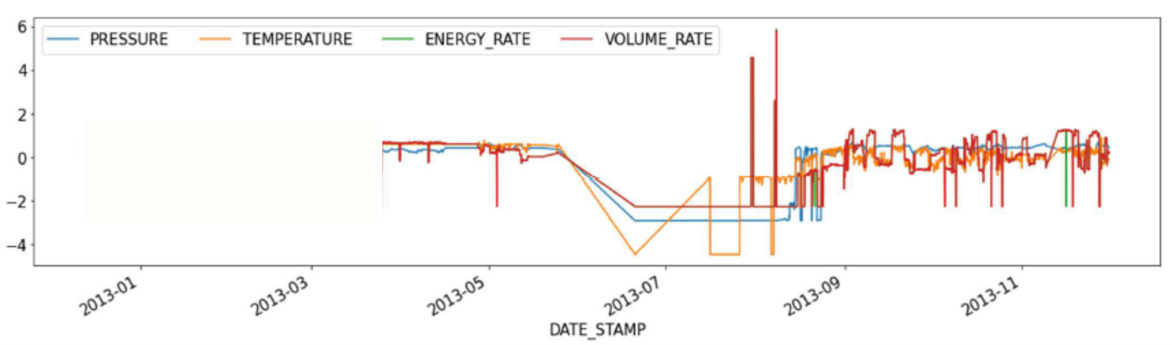
Continuous Development

Model should be updated continuously, to capture new trends in the data

First batch



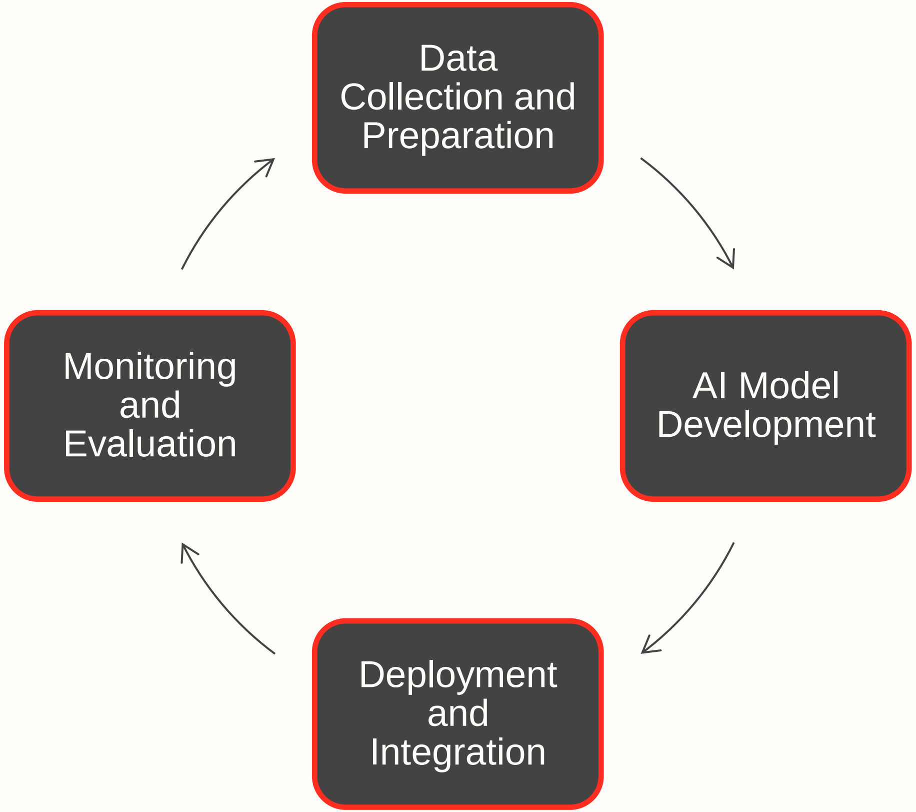
Next batch



Continuous Development

Maintenance of Anomaly Detection System should be done continuously.

An anomaly in a period of time can be seen as normal in other periods.



Data Collection and Preparation

AI Model Development

Deployment and Integration

Monitoring and Evaluation

Thanks!

Do you have any questions?

adityaihsan@telkomuniversity.ac.id
+62 857 4185 2615
phoenixfin.github.io

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**