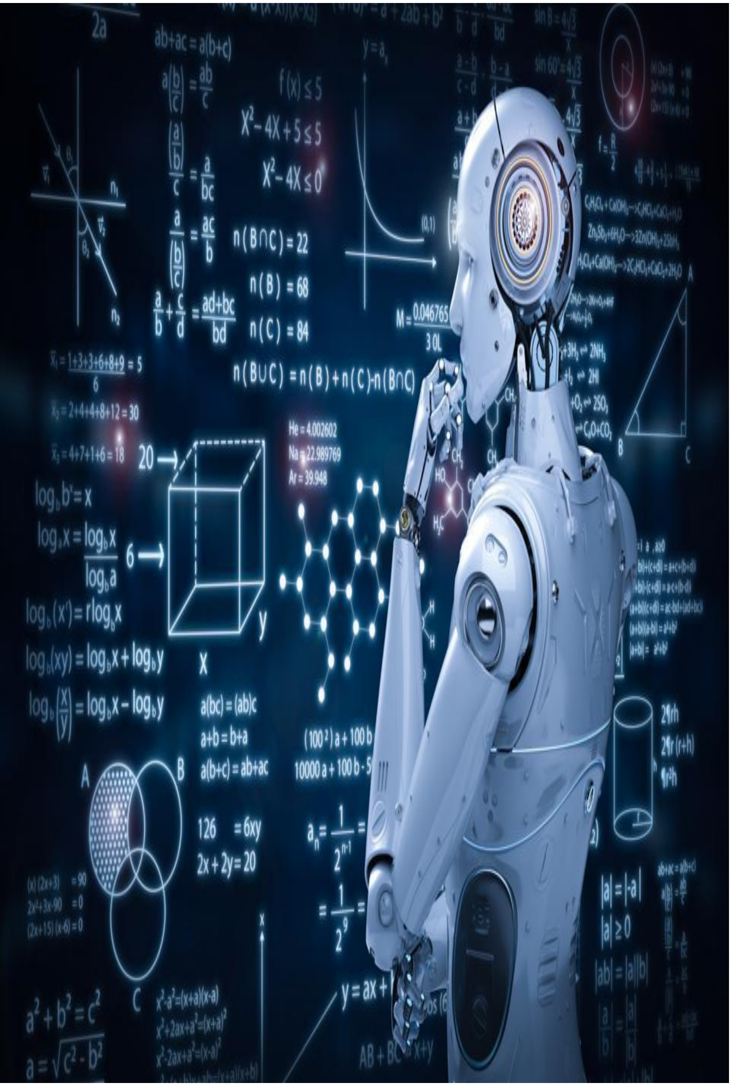


Practical Machine Learning

Day 15: Mar23 DBDA

Kiran Waghmare

Agenda

- Anomaly Detection

Anomaly detection

Anomalies and outliers
are essentially
the same thing:

objects that are different from most other objects

The techniques used for detection are the same.

Causes of anomalies

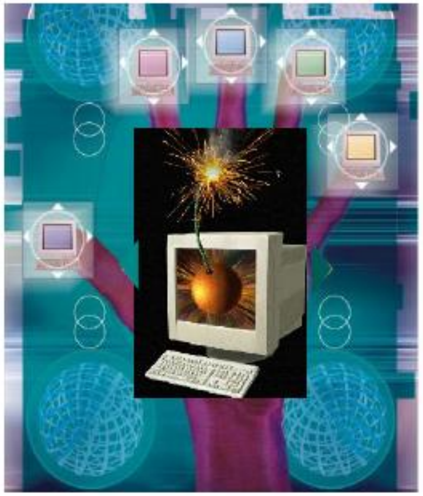
- Data from different class of object or underlying mechanism
  - disease vs. non-disease
  - fraud vs. not fraud
- Natural variation
  - tails on a Gaussian distribution
- Data measurement and collection errors

Applications of anomaly detection

- Network intrusion
- Insurance / credit card fraud
- Healthcare informatics / medical diagnostics
- Industrial damage detection
- Image processing / video surveillance
- Novel topic detection in text mining
- …

Intrusion detection

- Intrusion detection
  - Monitor events occurring in a computer system or network and analyze them for intrusions
  - Intrusions defined as attempts to bypass the security mechanisms of a computer or network
- Challenges
  - Traditional intrusion detection systems are based on signatures of known attacks and cannot detect emerging cyber threats
  - Substantial latency in deployment of newly created signatures across the computer system
- Anomaly detection can alleviate these limitations

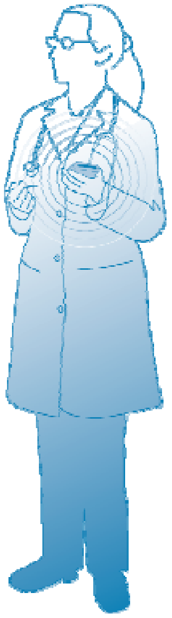


Fraud detection

- Detection of criminal activities occurring in commercial organizations.
- Malicious users might be:
  - Employees
  - Actual customers
  - Someone posing as a customer (identity theft)
- Types of fraud
  - Credit card fraud
  - Insurance claim fraud
  - Mobile / cell phone fraud
  - Insider trading
- Challenges
  - Fast and accurate real-time detection
  - Misclassification cost is very high

Healthcare informatics

- Detect anomalous patient records
  - Indicate disease outbreaks, instrumentation errors, etc.
- Key challenges
  - Only normal labels available
  - Misclassification cost is very high
  - Data can be complex: spatio-temporal



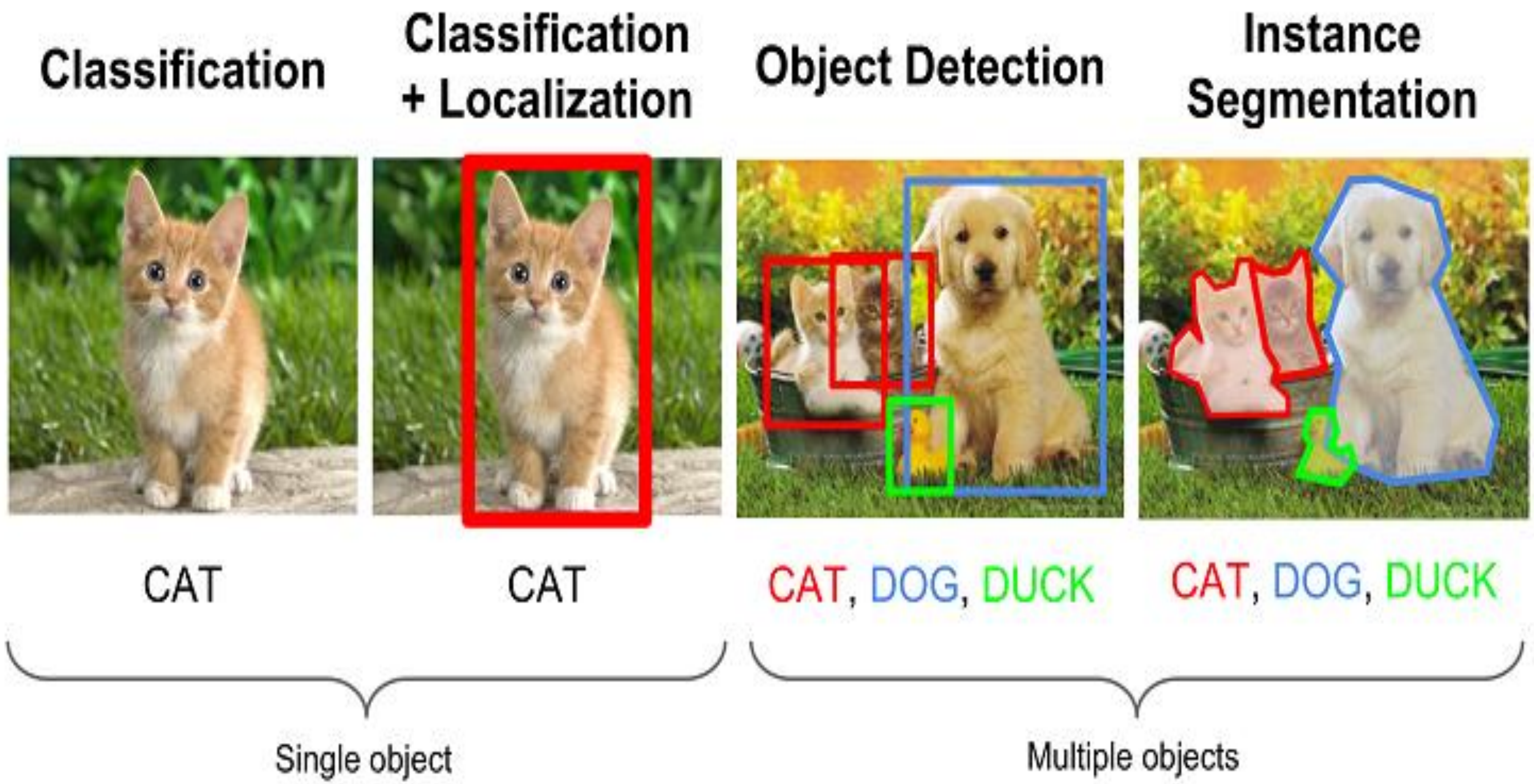
Industrial damage detection

- Detect faults and failures in complex industrial systems, structural damages, intrusions in electronic security systems, suspicious events in video surveillance, abnormal energy consumption, etc.
  - Example: aircraft safety
    - anomalous aircraft (engine) / fleet usage
    - anomalies in engine combustion data
    - total aircraft health and usage management

- Key challenges
  - Data is extremely large, noisy, and unlabelled
  - Most of applications exhibit temporal behavior
  - Detected anomalous events typically require immediate intervention

Image processing

- Detecting outliers in a image monitored over time
- Detecting anomalous regions within an image
- Used in
  - mammography image analysis
  - video surveillance
  - satellite image analysis
- Key Challenges
  - Detecting collective anomalies
  - Data sets are very large


Classification
Classification
+ Localization
Object Detection
Instance
Segmentation
CAT
CAT
CAT, DOG, DUCK
CAT, DOG, DUCK
Single object
Multiple objects

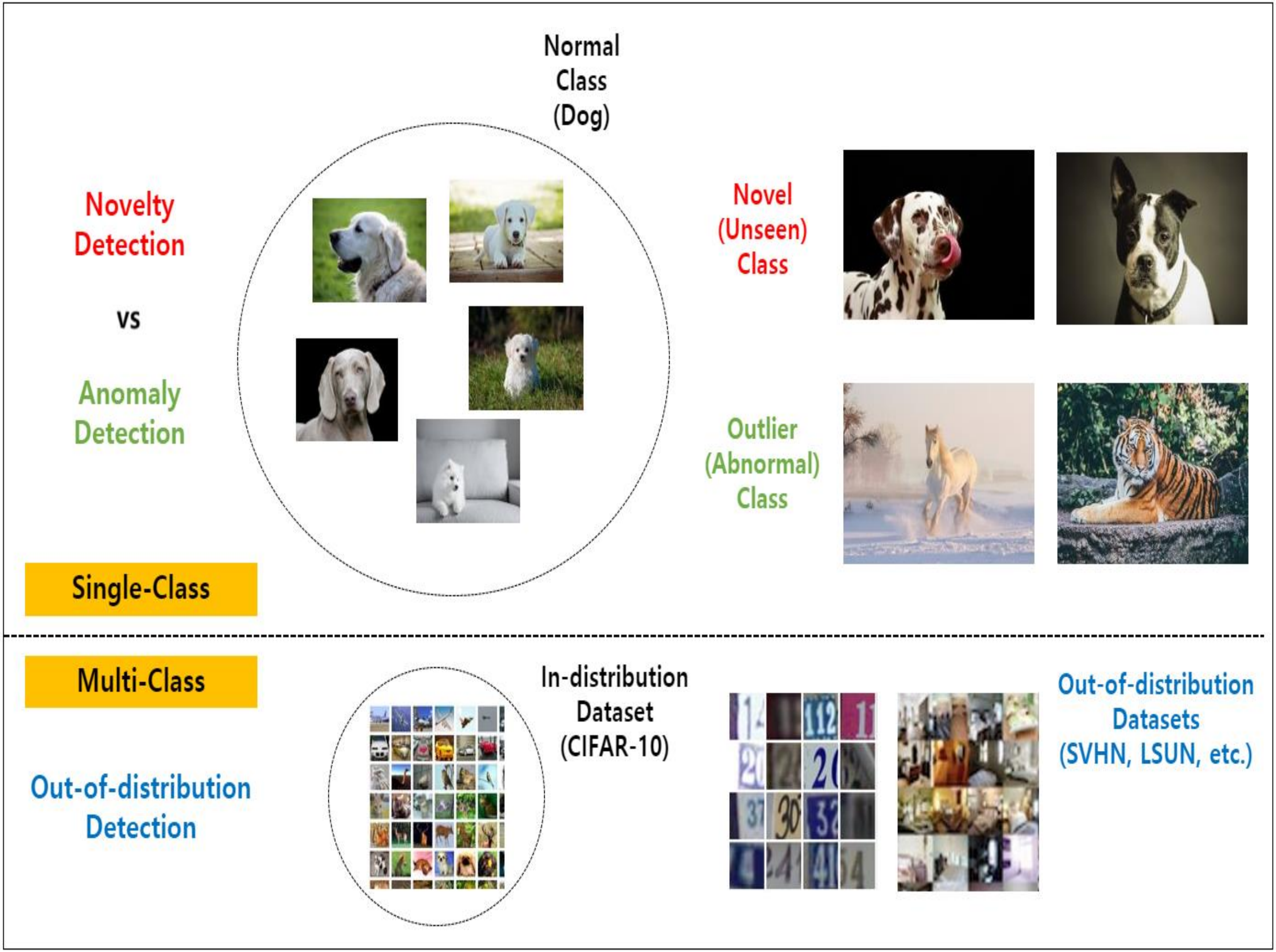
Use of data labels in anomaly detection

- Supervised anomaly detection
  - Labels available for both normal data and anomalies
  - Similar to classification with high class imbalance
- Semi-supervised anomaly detection
  - Labels available only for normal data
- Unsupervised anomaly detection
  - No labels assumed
  - Based on the assumption that anomalies are very rare compared to normal data

Output of anomaly detection

- Label
  - Each test instance is given a *normal* or *anomaly* label
  - Typical output of classification-based approaches

- Score
  - Each test instance is assigned an anomaly score
    - allows outputs to be ranked
    - requires an additional threshold parameter



Normal Class (Dog)
Novelty Detection
vs
Anomaly Detection
Novel (Unseen) Class
Outlier (Abnormal) Class
Single-Class
Multi-Class
Out-of-distribution Detection
In-distribution Dataset (CIFAR-10)
Out-of-distribution Datasets (SVHN, LSUN, etc.)

WHAT IS ANOMALY DETECTION?

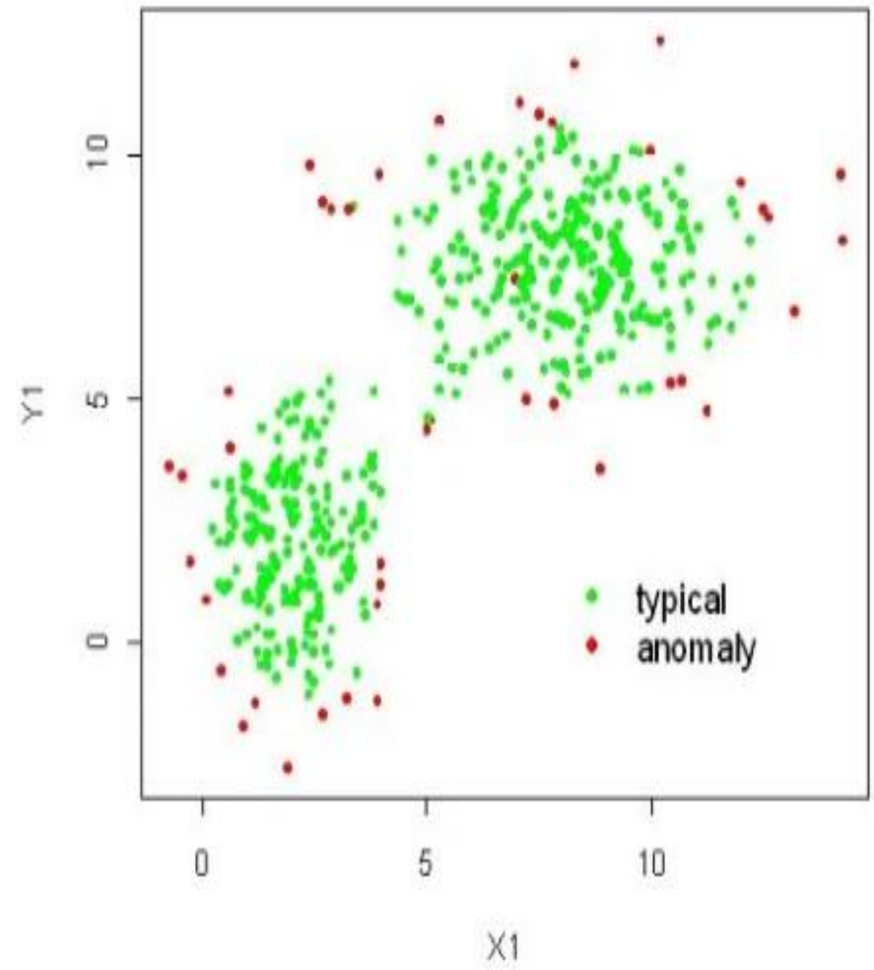
- **Anomaly Detection** (or outlier **detection**) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset.

WHAT IS ANOMALY DETECTION?

- **Anomaly detection** (also known as **outlier detection**) is the search for items or events which do not conform to an expected pattern.

- The patterns thus detected are called anomalies and often translate to critical and actionable information in several application domains.

- Anomalies are also referred to as outliers, change, deviation, surprise, aberrant, peculiarity, intrusion, etc.

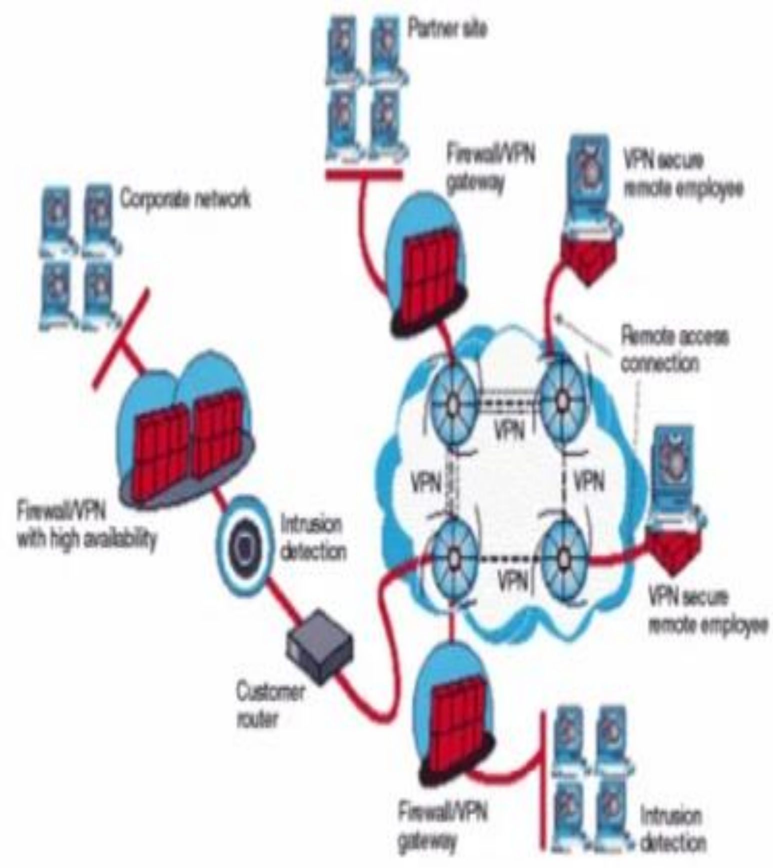


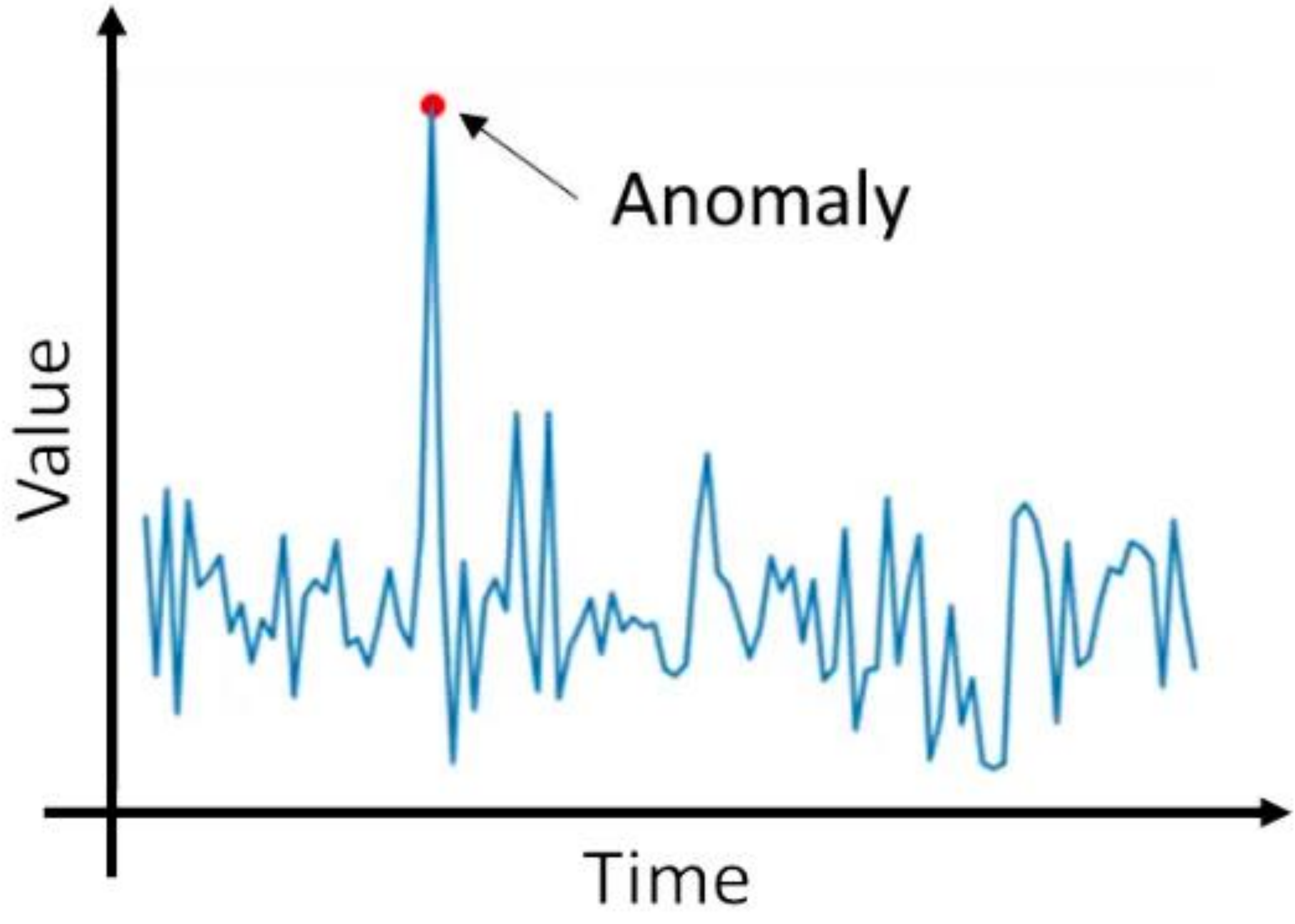
WHAT IS ANOMALY DETECTION?

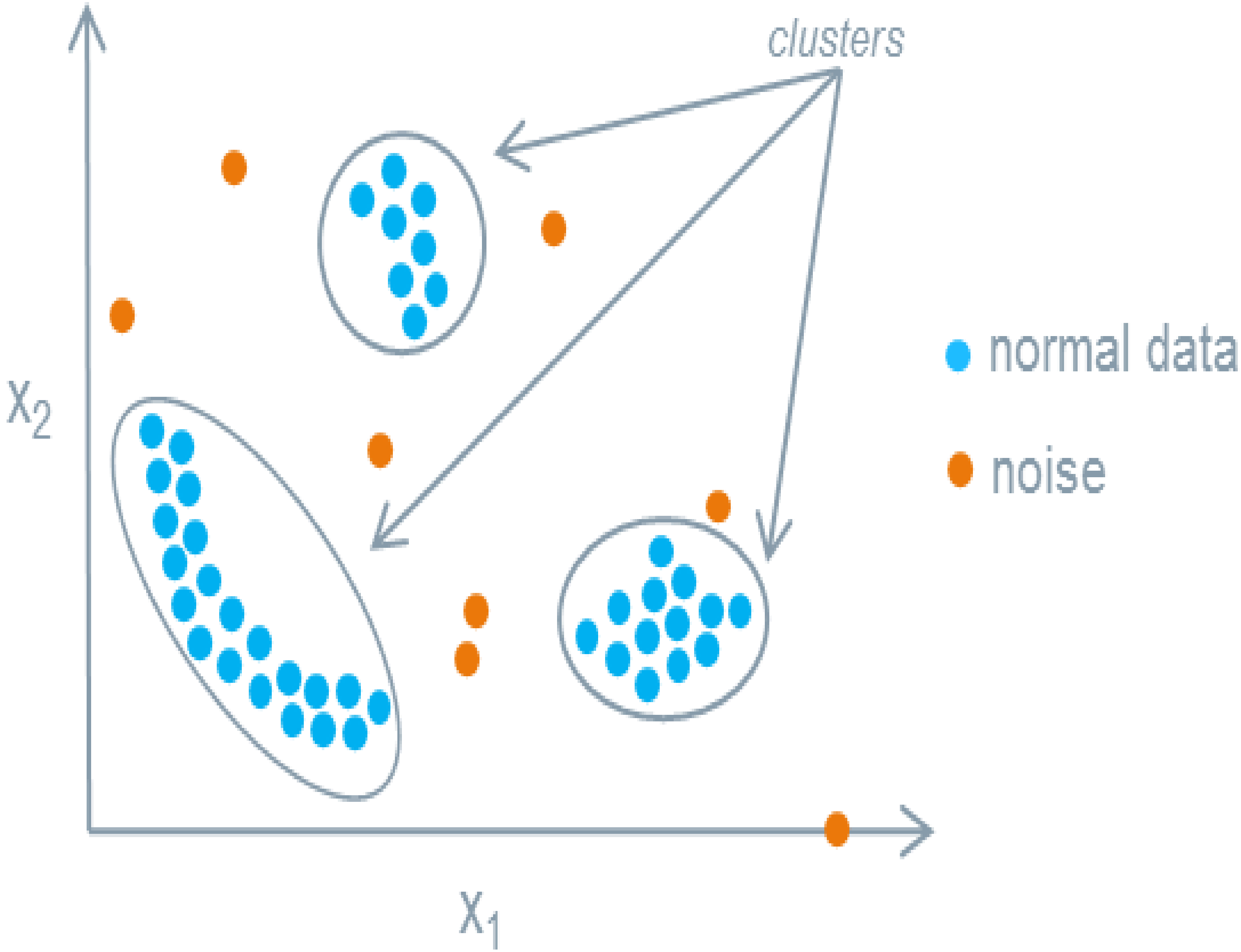
- Anomaly detection is applicable in a variety of domains, such as intrusion detection, fraud detection, fault detection, system health monitoring, event detection in sensor networks, and detecting Eco-system disturbances.

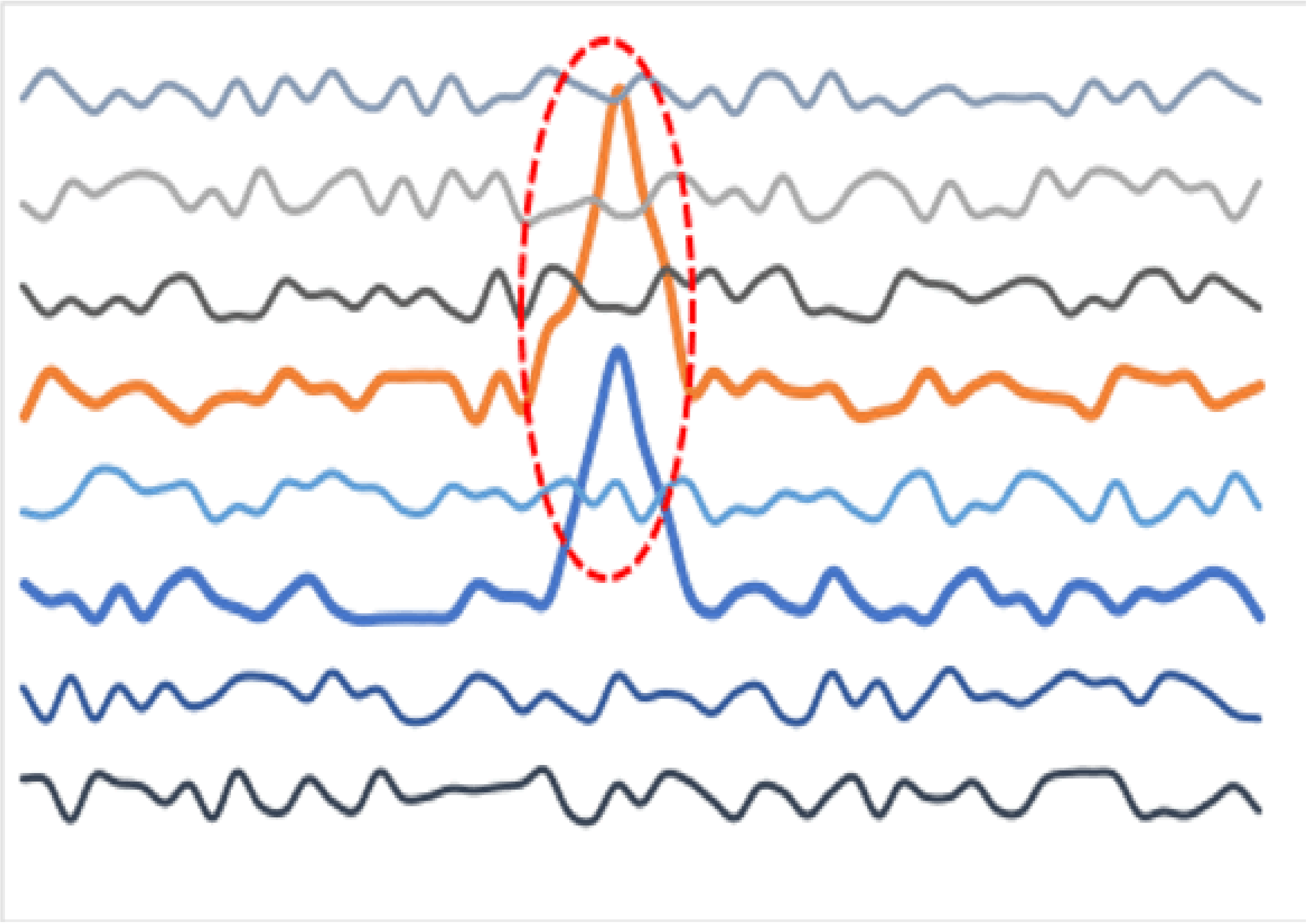
- It is often used in preprocessing to remove anomalous data from the dataset.

- In supervised learning, removing the anomalous data from the dataset often results in a statistically significant increase in accuracy.




Anomaly
Value
Time


clusters
normal data
noise
X_2
X_1



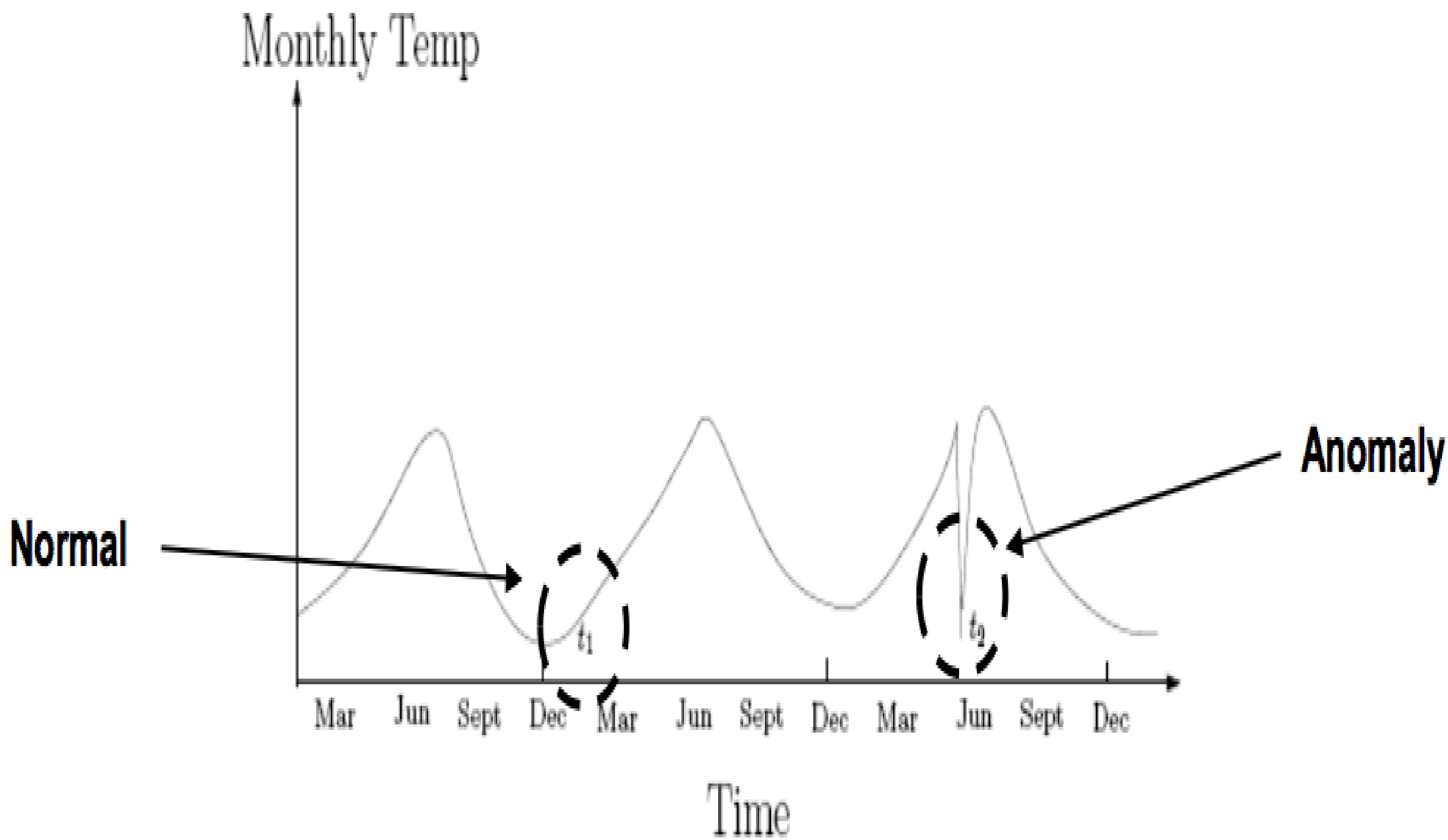
Structure of anomalies

- Point anomalies
- Contextual anomalies
- Collective anomalies

May-22	1:14 pm	FOOD	Monaco Café	$\boxed{\$1,127.80}$
May-22	2:14 pm	WINE	Wine Bistro	$28.00
...				
Jun-14	2:14 pm	MISC	Mobil Mart	$75.00
Jun-14	2:05 pm	MISC	Mobil Mart	$75.00
Jun-15	2:06 pm	MISC	Mobil Mart	$75.00
Jun-15	11:49 pm	MISC	Mobil Mart	$75.00
May-28	6:14 pm	WINE	Acton shop	$31.00
May-29	8:39 pm	FOOD	Crossroads	$128.00
Jun-16	11:14 am	MISC	Mobil Mart	$75.00
Jun-16	11:49 am	MISC	Mobil Mart	$75.00

Point Anomaly

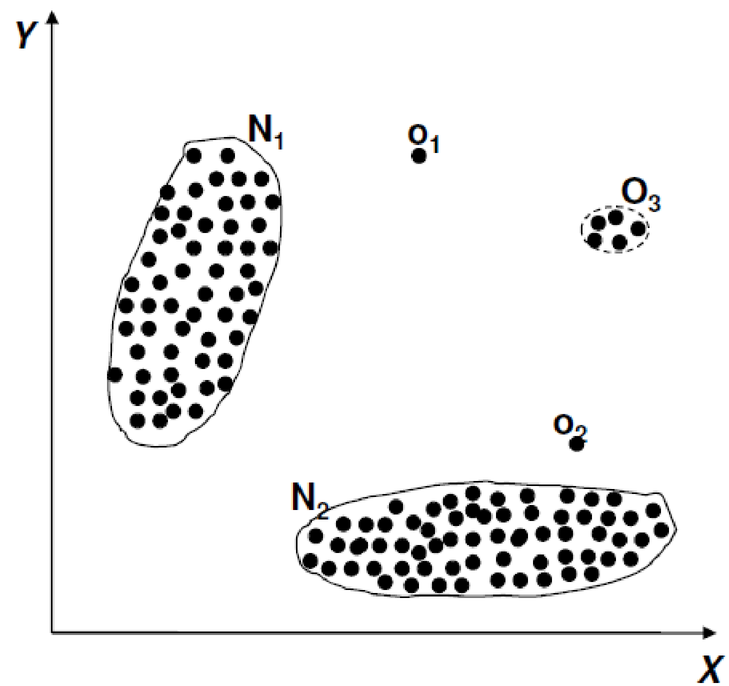
Collective Anomaly



Monthly Temp
Normal
Anomaly
t_1
t_2
Mar
Jun
Sept
Dec
Mar
Jun
Sept
Dec
Mar
Jun
Sept
Dec
Time

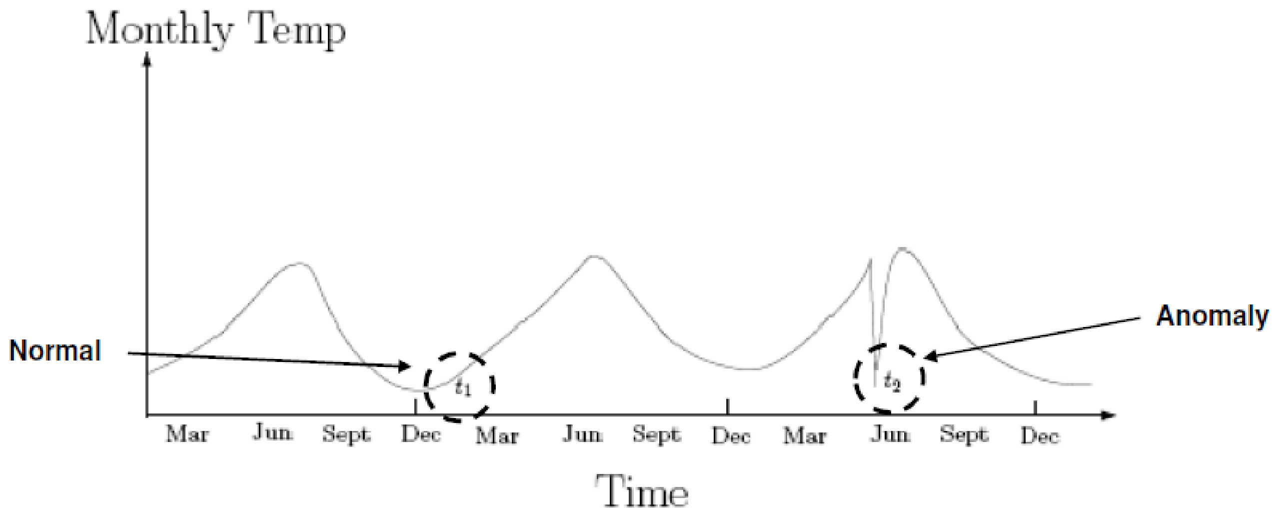
Point anomalies

- An individual data instance is anomalous with respect to the data



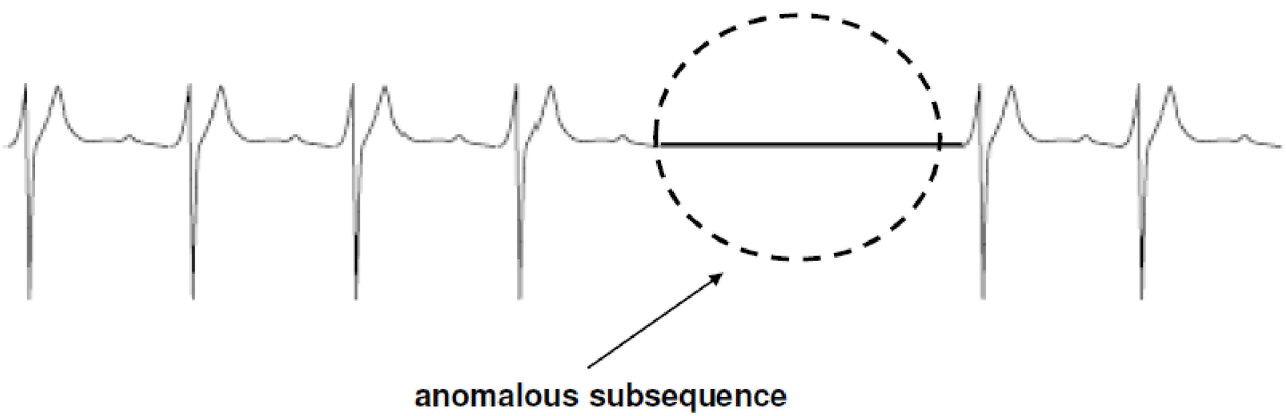
Contextual anomalies

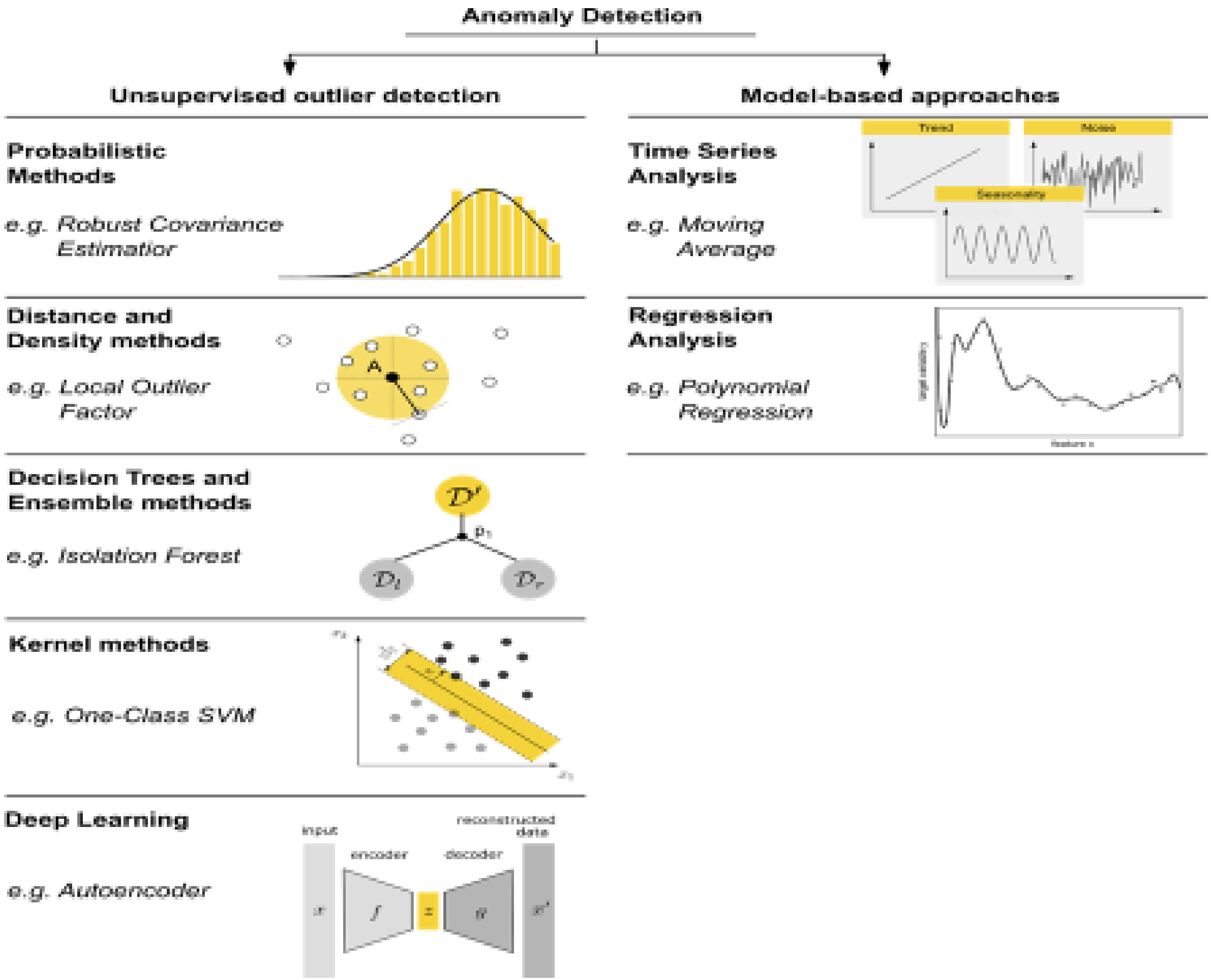
- An individual data instance is anomalous within a context
- Requires a notion of context
- Also referred to as conditional anomalies *



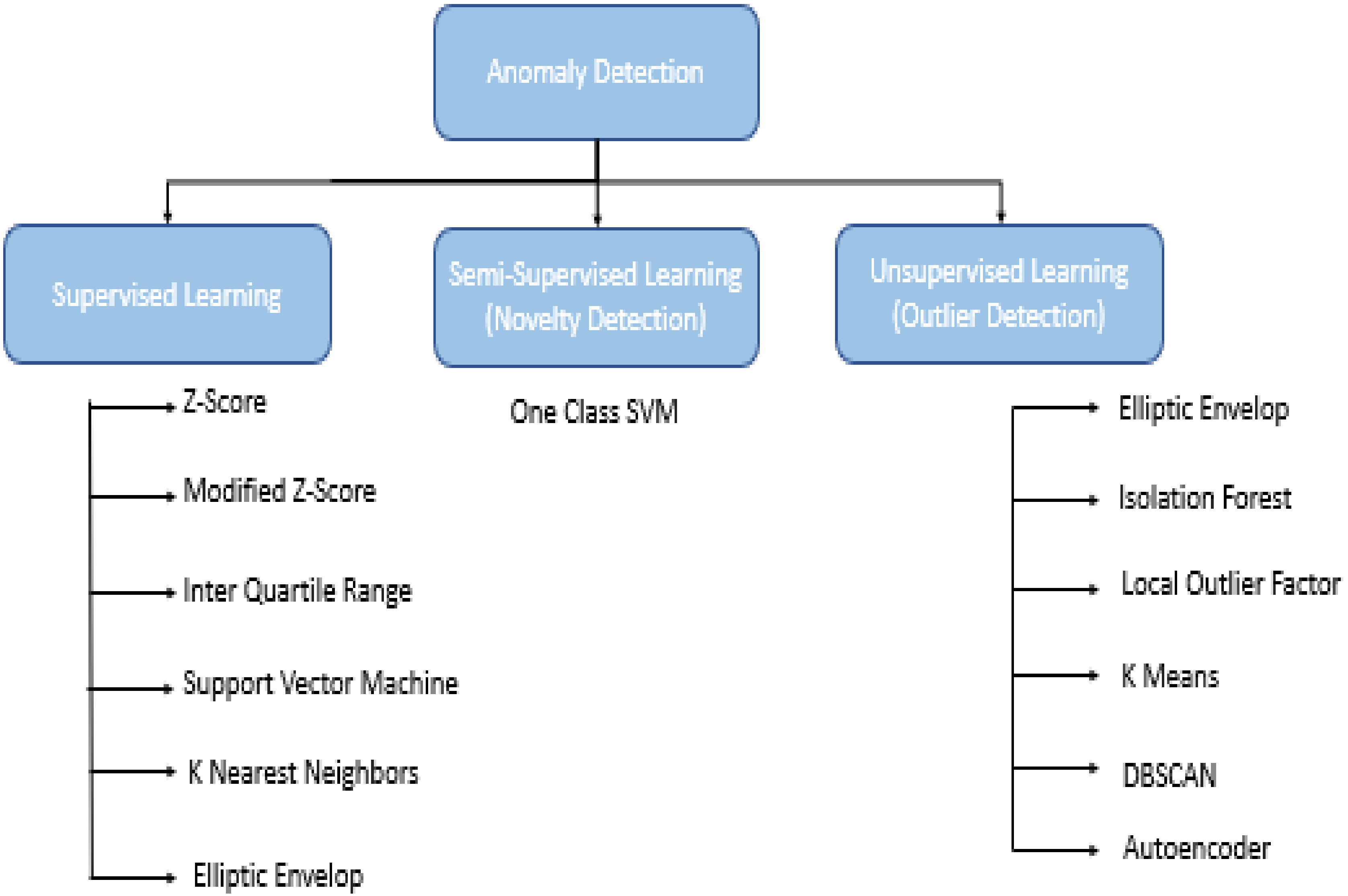
Collective anomalies

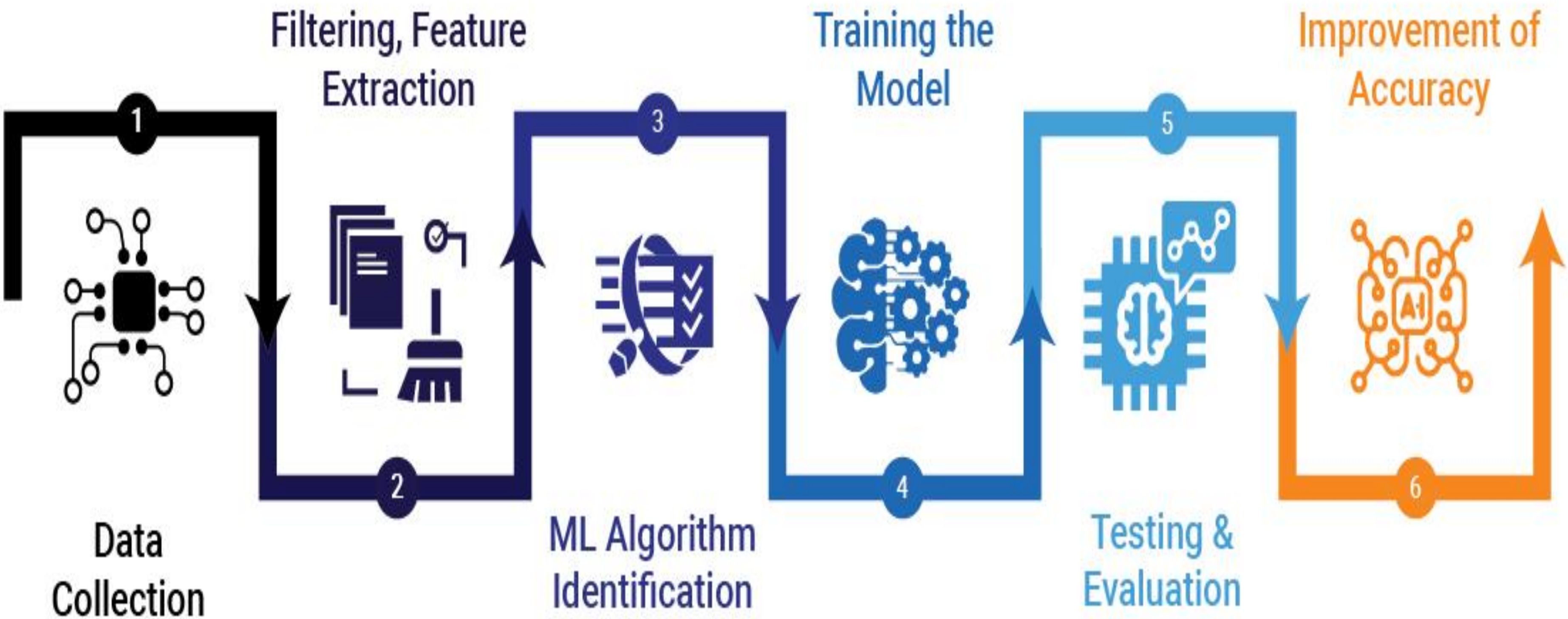
- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential data
  - Spatial data
  - Graph data
- The individual instances within a collective anomaly are not anomalous by themselves





Anomaly Detection
Unsupervised outlier detection
Model-based approaches
Probabilistic Methods
e.g. Robust Covariance Estimator
Distance and Density methods
e.g. Local Outlier Factor
A
Decision Trees and Ensemble methods
e.g. Isolation Forest
Kernel methods
e.g. One-Class SVM
Deep Learning
e.g. Autoencoder
input
encoder
decoder
reconstructed data
Time Series Analysis
e.g. Moving Average
Trend
Noise
Seasonality
Regression Analysis
e.g. Polynomial Regression


Anomaly Detection
Supervised Learning
Semi-Supervised Learning (Novelty Detection)
Unsupervised Learning (Outlier Detection)
Z-Score
Modified Z-Score
Inter Quartile Range
Support Vector Machine
K Nearest Neighbors
Elliptic Envelop
One Class SVM
Elliptic Envelop
Isolation Forest
Local Outlier Factor
K Means
DBSCAN
Autoencoder



Filtering, Feature Extraction
Training the Model
Improvement of Accuracy
1
3
5
A I
2
4
6
Data Collection
ML Algorithm Identification
Testing & Evaluation