Applied Machine Learning for Anomaly Detection on Equipment

Kristin Hornæs
Lukasz Mentel
Trung Doan

Kristin

Lukasz

Trung

ARUNDO

09:00 - 09:30	Welcome & Introduction to Anomaly Detection
09:30 - 10:00	Set-up environment
10:00 - 10:30	Walk through examples of AD methods in Jupyter notebooks
10:30 - 10:45	Coffee break
10:45 - 11:45	Improve models
11:45 - 12:00	Presentation of selected solutions

-------- LUNCH BREAK --------

ARUNDO

BUZZWORD BINGO

Digital Transformation	Operational Intelligence	Streaming Analytics
IT Operations Analytics	**Industry 4.0**	Blended Analytics

ARUNDO

Increase efficiency and productivity

Data is the jetfuel

ARUNDO

Decrease downtime

Unexpected downtime on a single asset can cost upwards of a million dollars per day

INTERCONNECTED AMBITIONS

Increase efficiency

Increase profits despite a decreasing price per barrel

Scalable, actionable insight

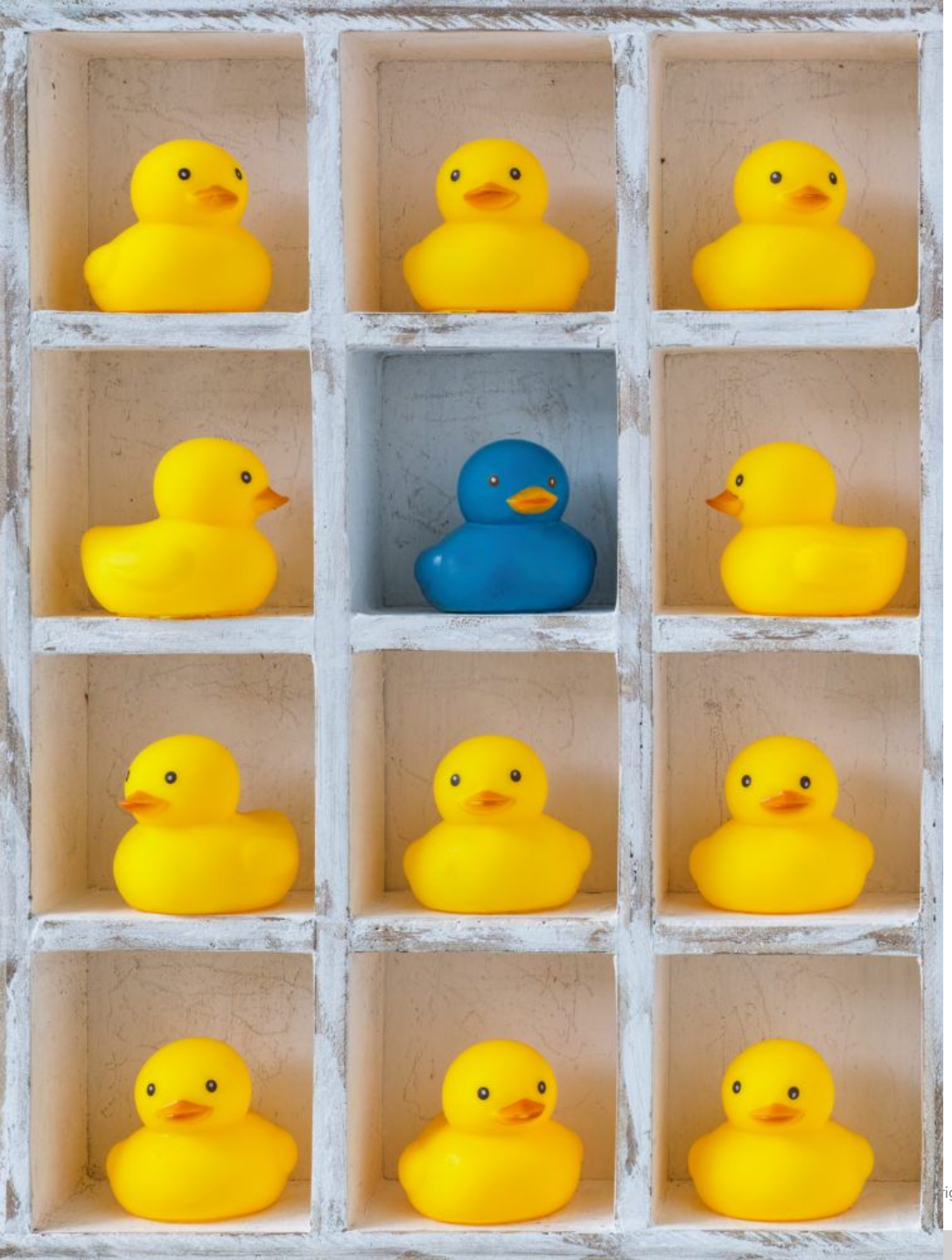
Make models which can be easily applied to the company's entire portfolio

ARUNDO

ARUNDO

provides software products to **enable** enterprise-scale machine learning and advanced analytics applications for **industrial companies**

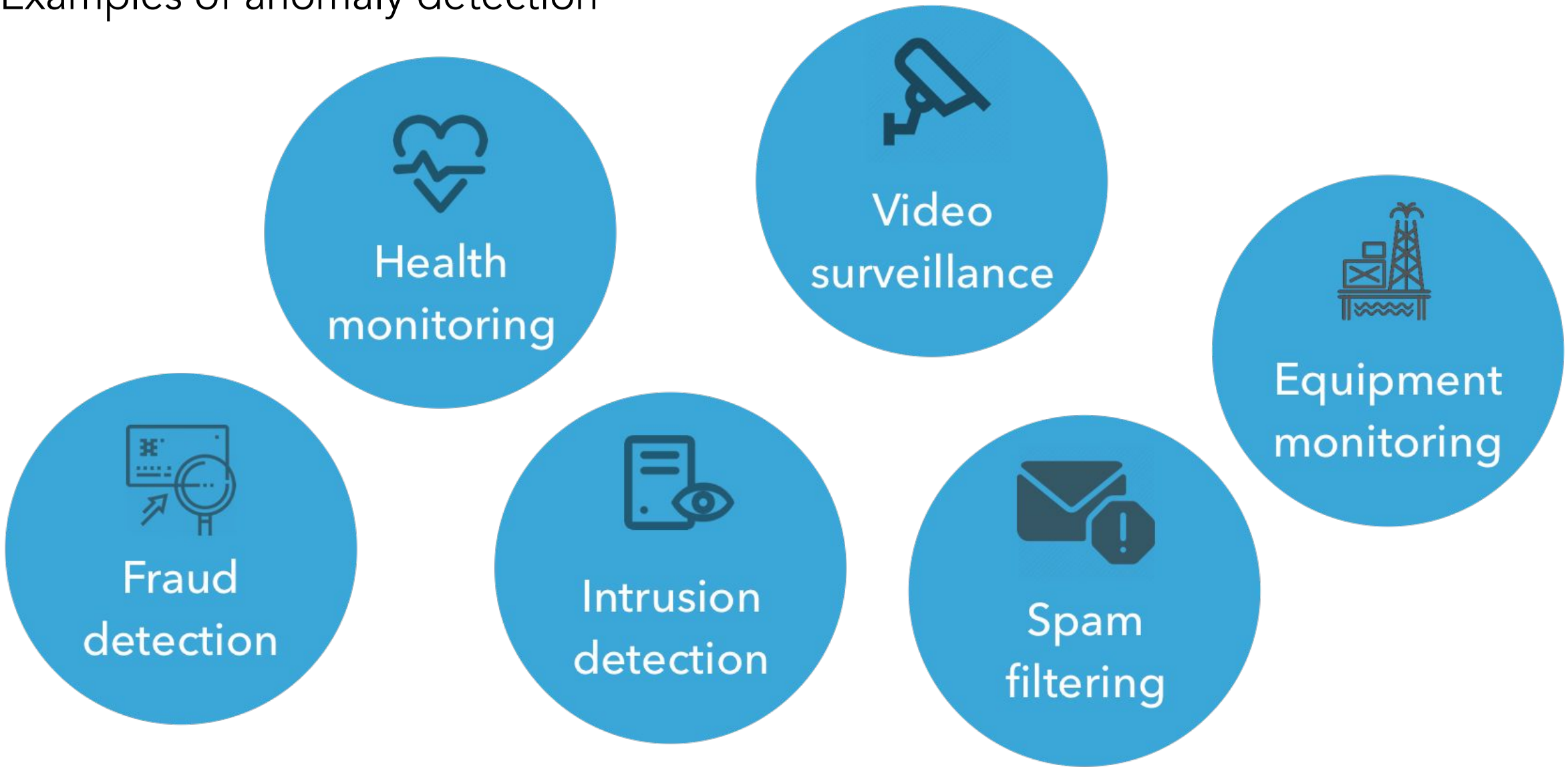

What is an anomaly?



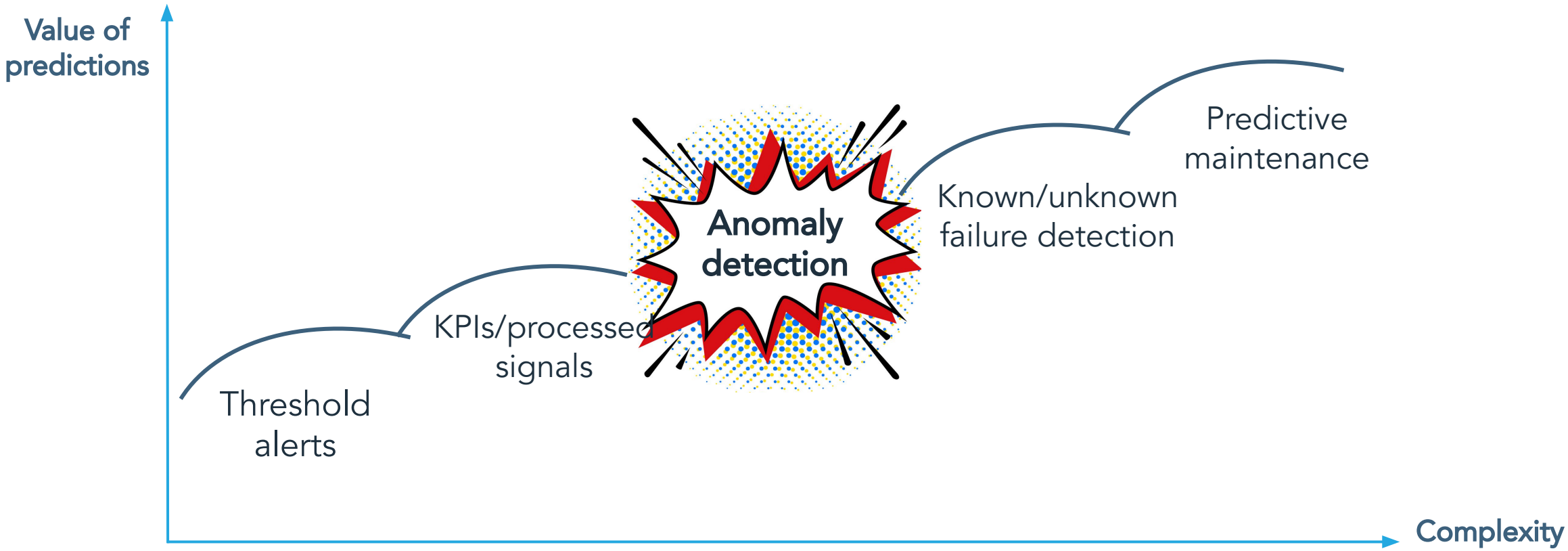
"DATAPOINTS, ITEMS, OBSERVATIONS OR EVENTS THAT DO NOT CONFORM TO THE EXPECTED PATTERN"

ARUNDO

Examples of anomaly detection



The stages of data-driven equipment monitoring

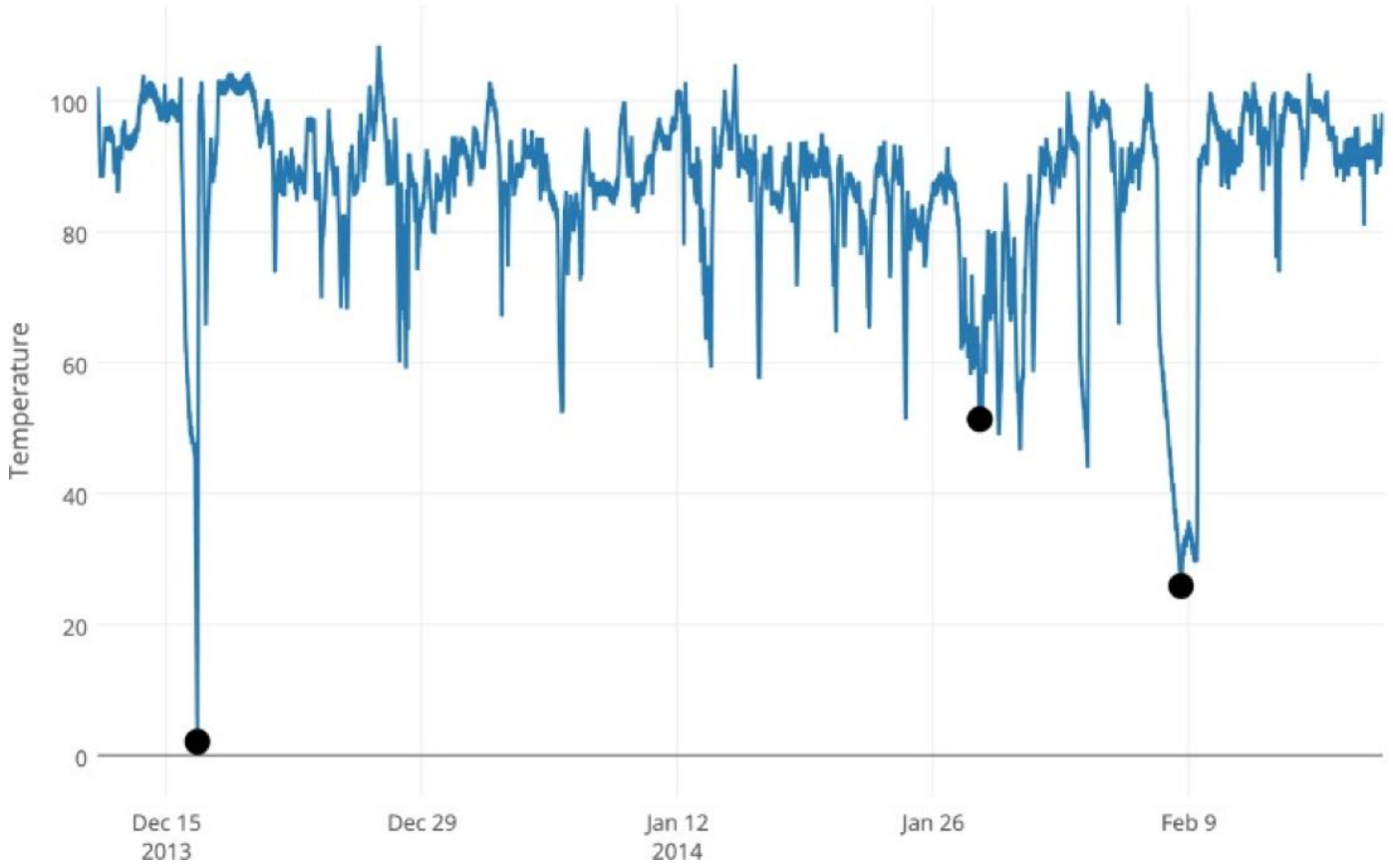


ARUNDO

Anomaly detection in **equipment monitoring**

Previously unseen patterns can be a sign of:

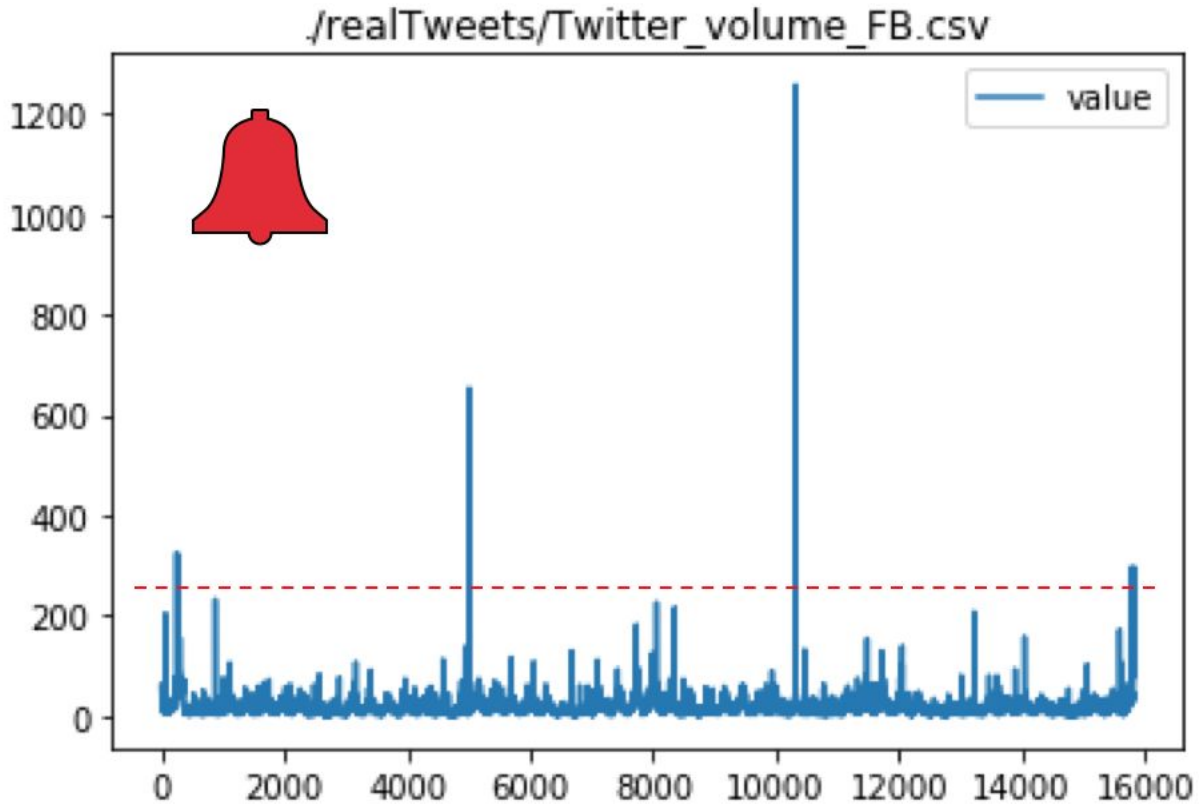
- **misconfiguration**
- increasing mechanical **wear-out**
- **unforeseen** situations



ARUNDO

How can you detect anomalies?

- **Define a threshold for each sensor channel**
- **Raise a notification once a specified threshold is violated**

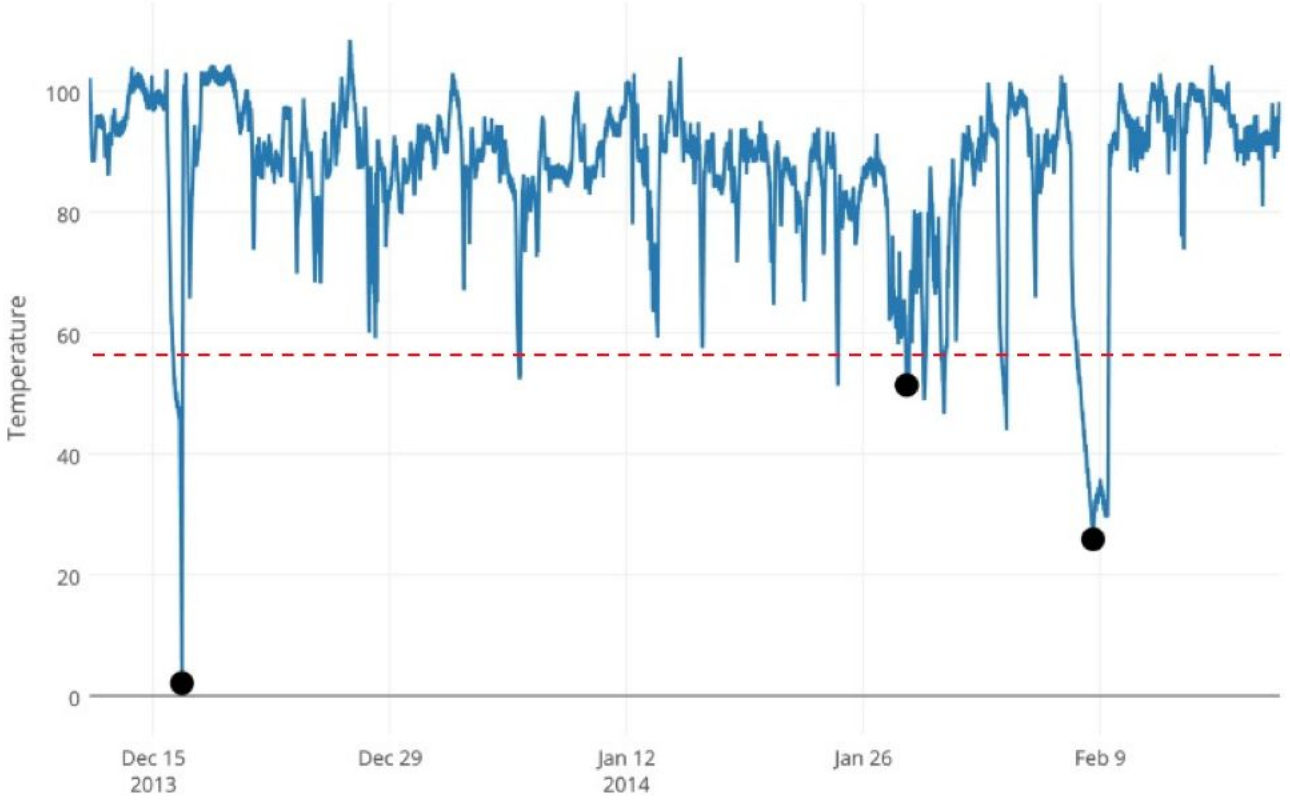


ARUNDO

An oil rig can have upwards of 15.000 sensors
(the newest have more than 50.000 sensors!!!)

How can you detect anomalies?

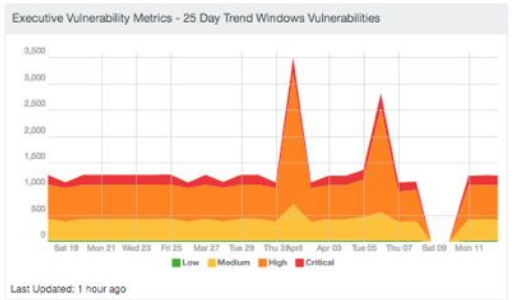
- **Causes many false alerts**
- **Does not take into account the joint characteristics of multiple channels**

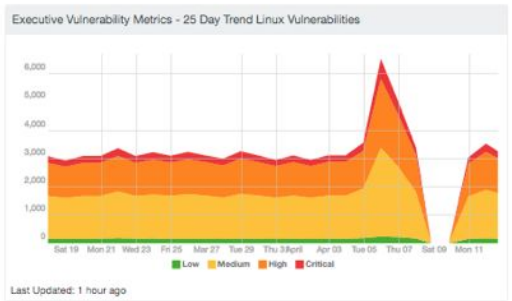


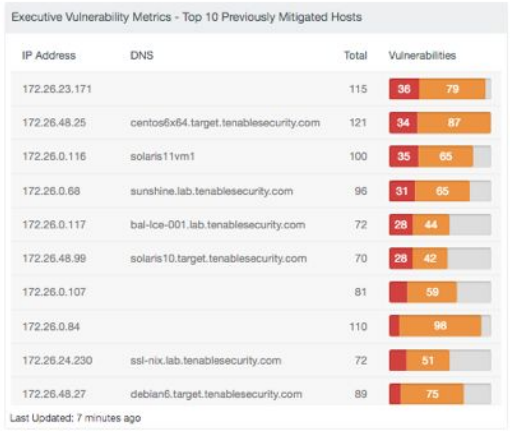
ARUNDO

Manual analysis is immensely time-consuming and unreliable

- **Massive** amount of multi-sensor data
- **Complex** systems
- **Rare** faults



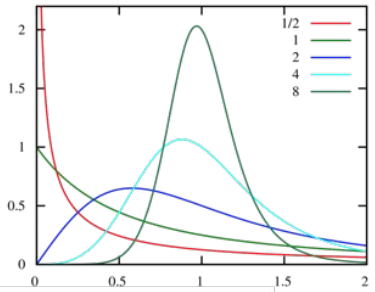


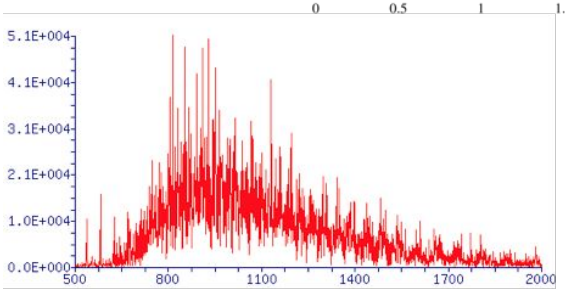


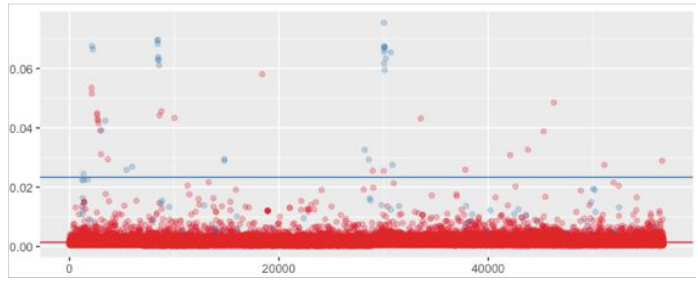
ARUNDO

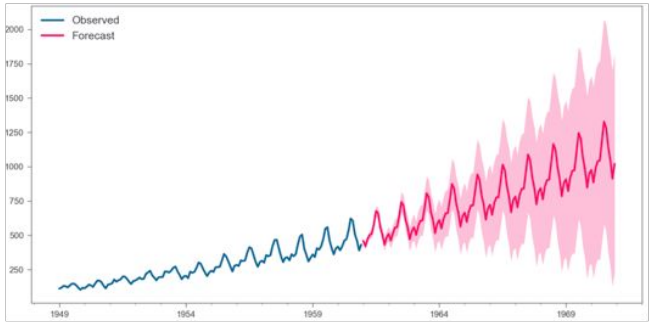
Multivariate anomaly detection

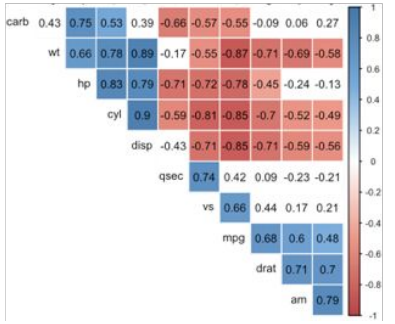
- No prior knowledge about anomalies
- No precise boundary
- Data often contain noise
- Normal behaviour keeps evolving
- Temporal dependencies
- Highly unbalanced classes
- High dimensionality and multimodal dependencies











ARUNDO

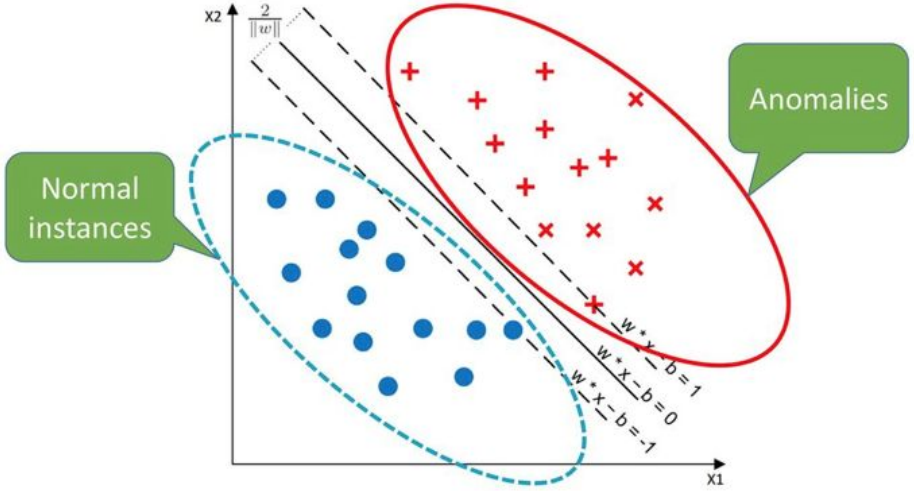
Approaches

Data

Labeled data?
Predict Y from X?

ARUNDO

Approaches



Data

Labeled data?
Predict Y from X?

yes

Supervised Learning

Develop predictive model based on both input and output data

Partition data based on labels

Classification

Regression

ARUNDO

Approaches

Data

Labeled data?
Predict Y from X?

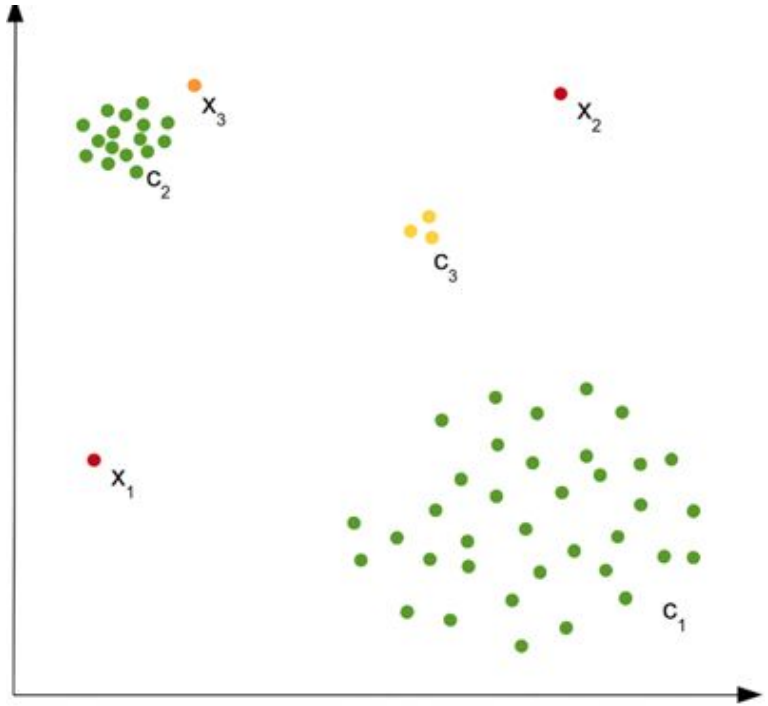
no

Unsupervised Learning

Group and interpret data based only on input data

Outliers in the distribution

Clustering



ARUNDO

Real life example - Leakage on heat exchangers

Not enough failures to make a good model

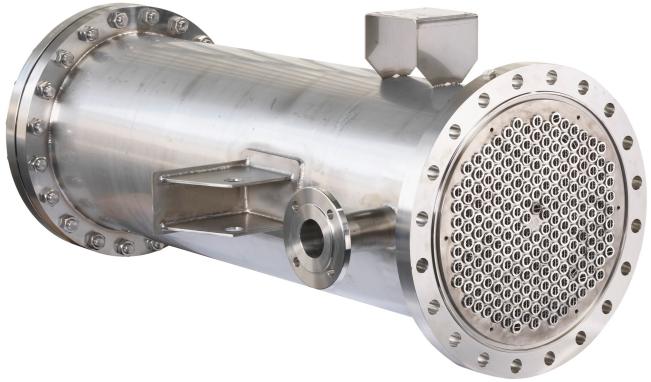
Unsupervised Learning

Group and interpret data based only on input data

What is the pattern of normal behaviour of the heat exchanger?

Sensor data+ maintenance logs

Training (simulation or historical) data available?

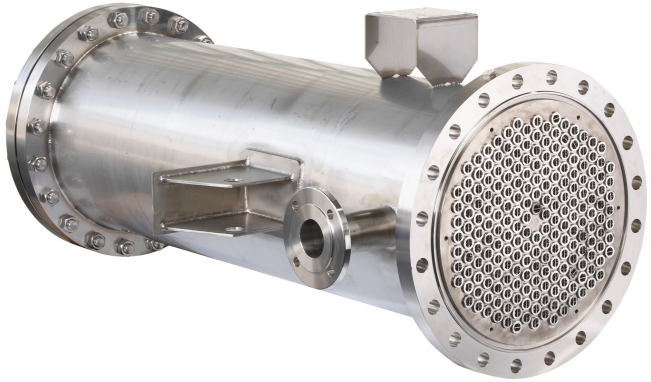


ARUNDO

Real life example - Leakage on heat exchangers

Sensor data+ maintenance logs

Training (simulation or historical) data available?



Yes but not a lot of failures

Supervised Learning

Develop predictive model based on both input and output data

discrete

Has my heat exchanger sprung a leak?

continuous

What is the predicted performance of my heat exchanger?

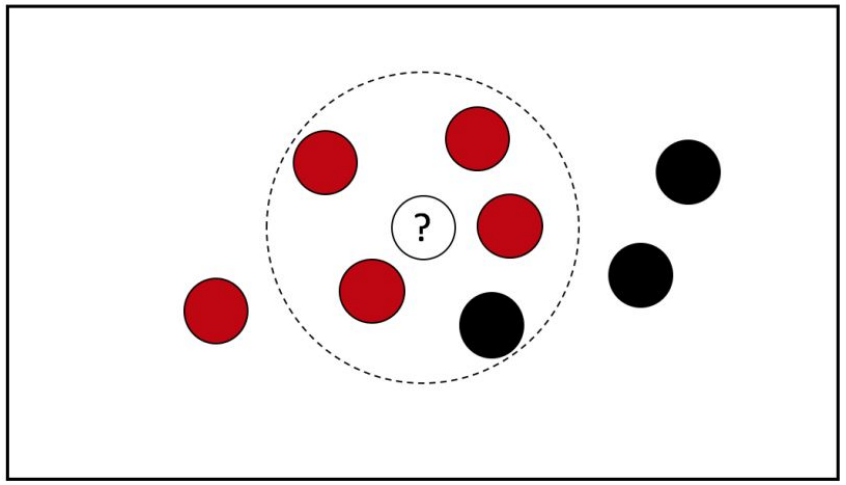
ARUNDO

Approaches

Supervised Methods	Unsupervised Methods
Random Forest	One Class SVM
Support Vector Classification	Isolation Forest
KNN Classifier	Elliptic Envelope
Logistic Regression	Local Outlier Factor
	Autoencoder

ARUNDO

k-NN Classifier



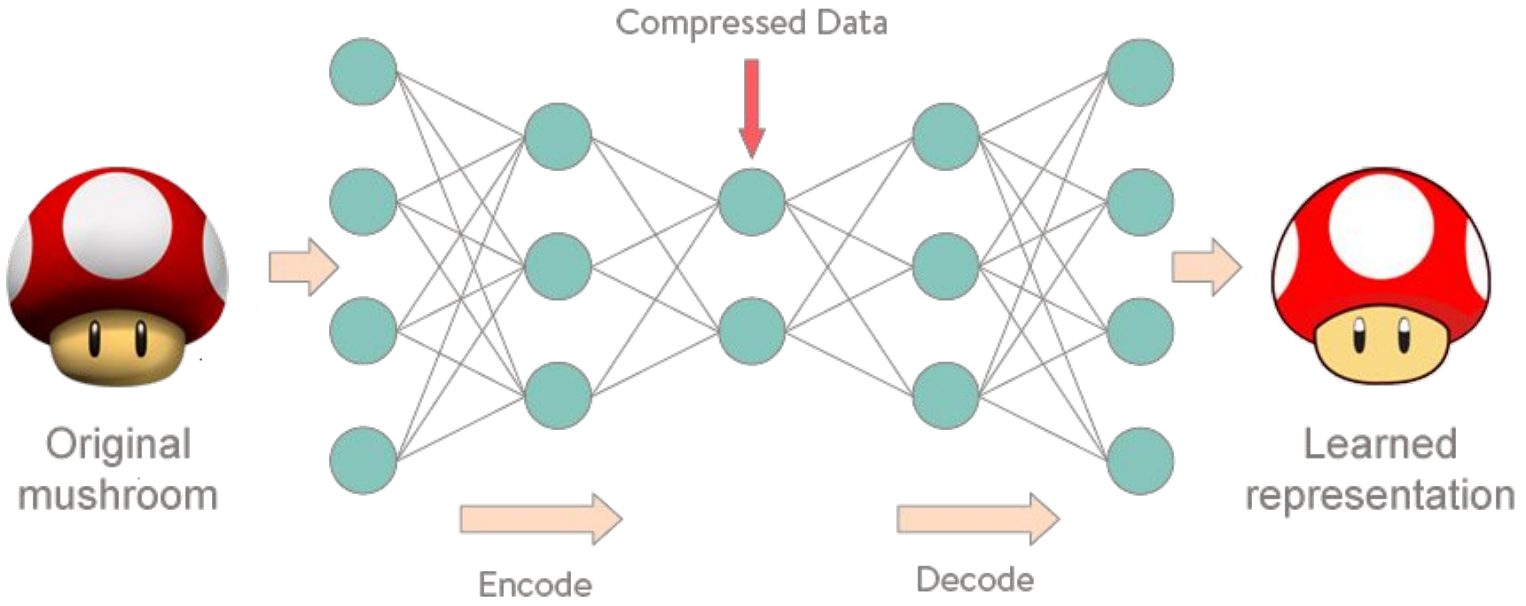
_k_-NN classifies a data point based on how its neighbors are classified.

If a data point is surrounded by 4 red points and 1 black point, that data point is likely a red point by majority vote.

Tune for *k* - the number of nearest neighbors to include in the majority voting process

ARUNDO

Autoencoders



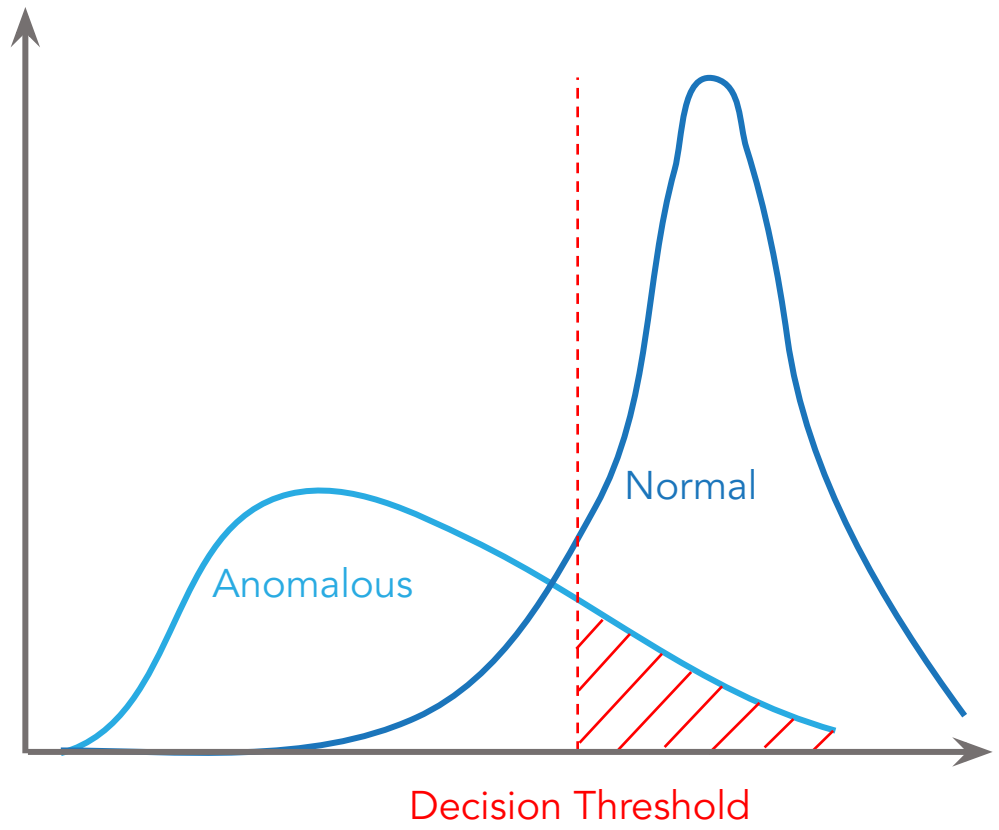
The job of those models is to predict the input, given that same input.

Learns a representation of the training data and recreates the input at the output layer.

Used for data compression and learning generative models from data.

We optimize the parameters of our Autoencoder model in such way that the reconstruction error is minimized.

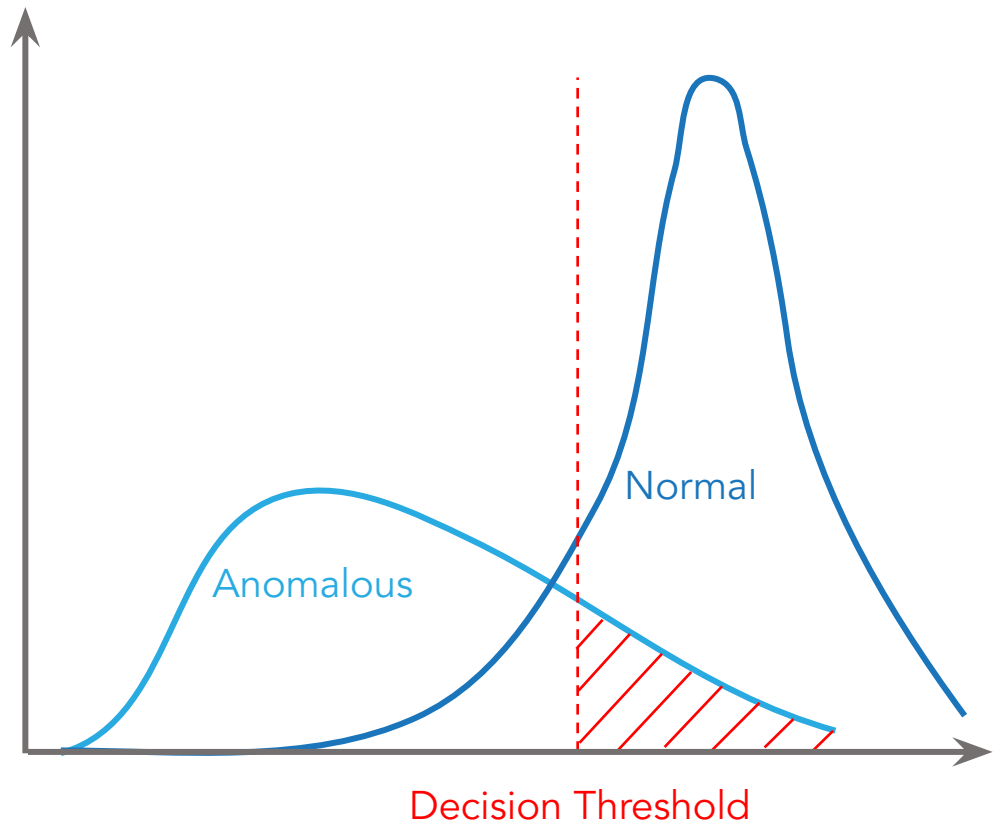
Machine Learning Predictions are not 100% Accurate



- False positives (ML predicts anomalous but system is normal) can and will occur
- False negatives (ML predicts normal but system is anomalous) can and will occur

ARUNDO

Machine Learning Predictions are not 100% Accurate



- Choose trained algorithms which minimize these effects as much as possible
- Identify any additional sources of data which may further help minimize these effects
- Evaluate the expected probabilities of each scenario using independent historical data
- Retrain the system if there is evidence that probabilities change significantly with time

ARUNDO

Workshop details

Task

- Make a model to identify anomalies based on the dataset provided
- The model will be tested on a separate test dataset
- The best model will be chosen based on f1 score

Practicalities

- Log on to jupyter hub using your github account
  - https://wids.arundo.com

Notebooks

- Opening and looking at the data
  - ➢ 01-Model Development
- Improve baseline model results
  - ➢ Please add comments to explain your code where possible

ARUNDO