Outlier detection in pension asset data

DeNederlandscheBank
EUROSYSTEEM

Iris Nonneman

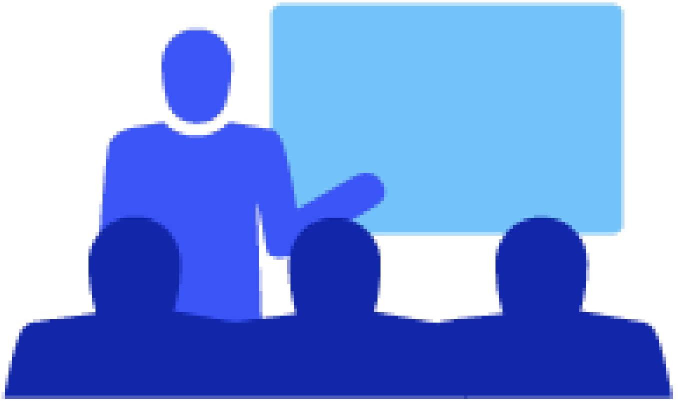
05-03-2024

Goal

Automatically detect outliers in the line-by-line data of pension funds and insurers using machine learning techniques

Agenda

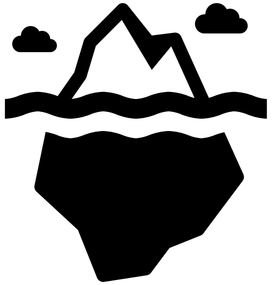
- Why detecting outliers
- Motivation
- Model approach
- From data to model
- Results and performance
- Use cases for actuaries



Detecting outliers?

Natural causes of outliers in data

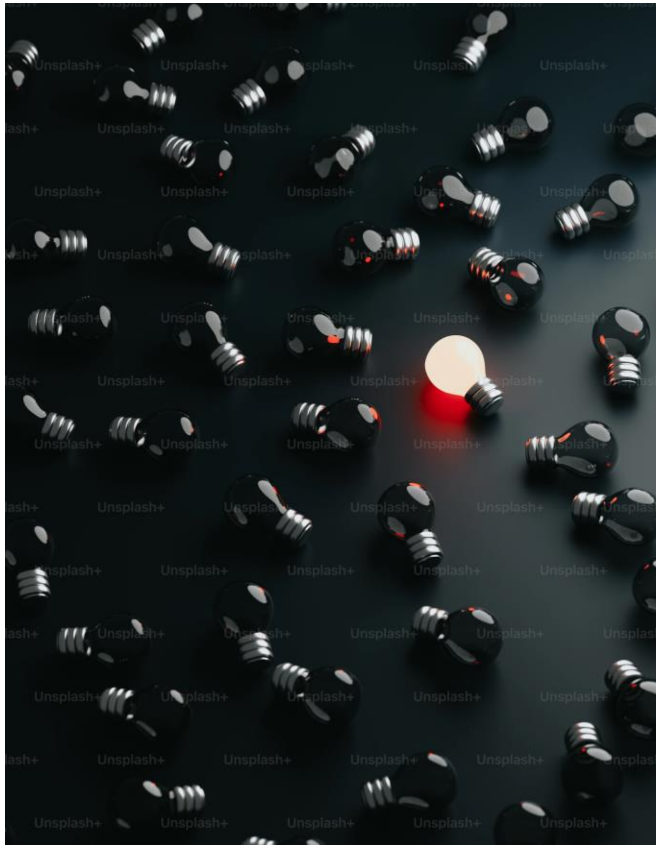
1. Data error: wrong measurement data observation
2. Natural occurrence but different than expected



Problems caused by outliers

1. Outliers in the data influence model fitting (linear models)
2. Outliers can inflate metrics which give higher weights to large errors (like RMSE).

Why detecting outliers?



[A light bulb surrounded by many black ones photo – Innovation Image on Unsplash]

Outliers can be informative

1. Outliers that are data errors influence results

2. Outliers that are not errors (anomalies), are informative: show behavior that is different from “expected” / the bulk of data

Background project

DeNederlandscheBank

EUROSYSTEEM

FTK data

Verslagstaten
- Pension funds report asset data

Dataloop
- Reports in dataloop
- Rule-based checks

Re-reporting in case of data errors

Used in supervision
- Used in supervision

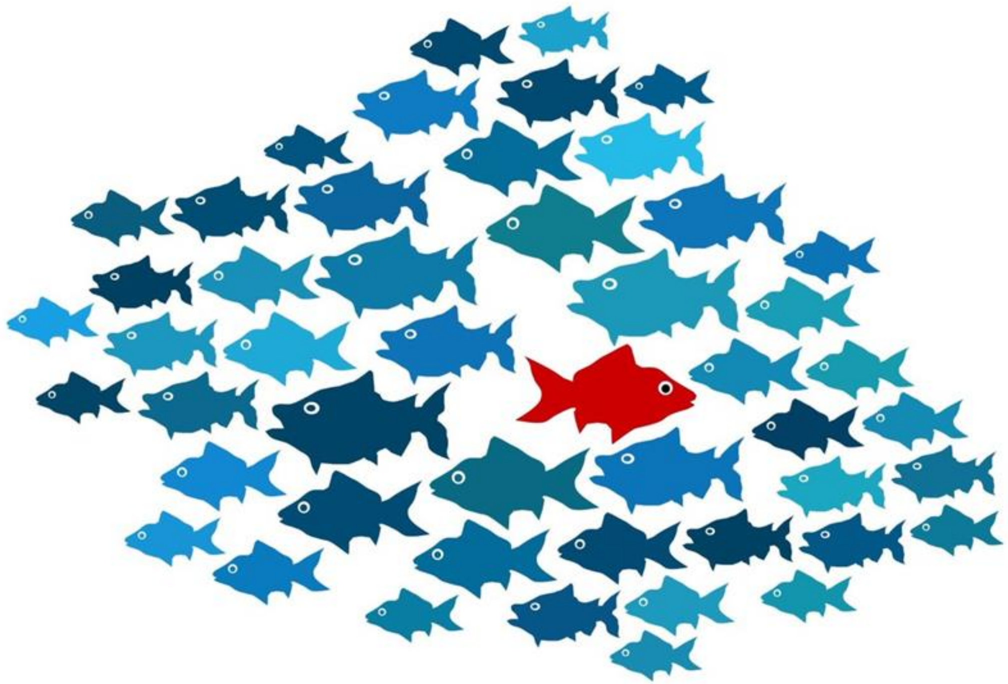
Motivation for using ML algorithm

Characteristics asset data

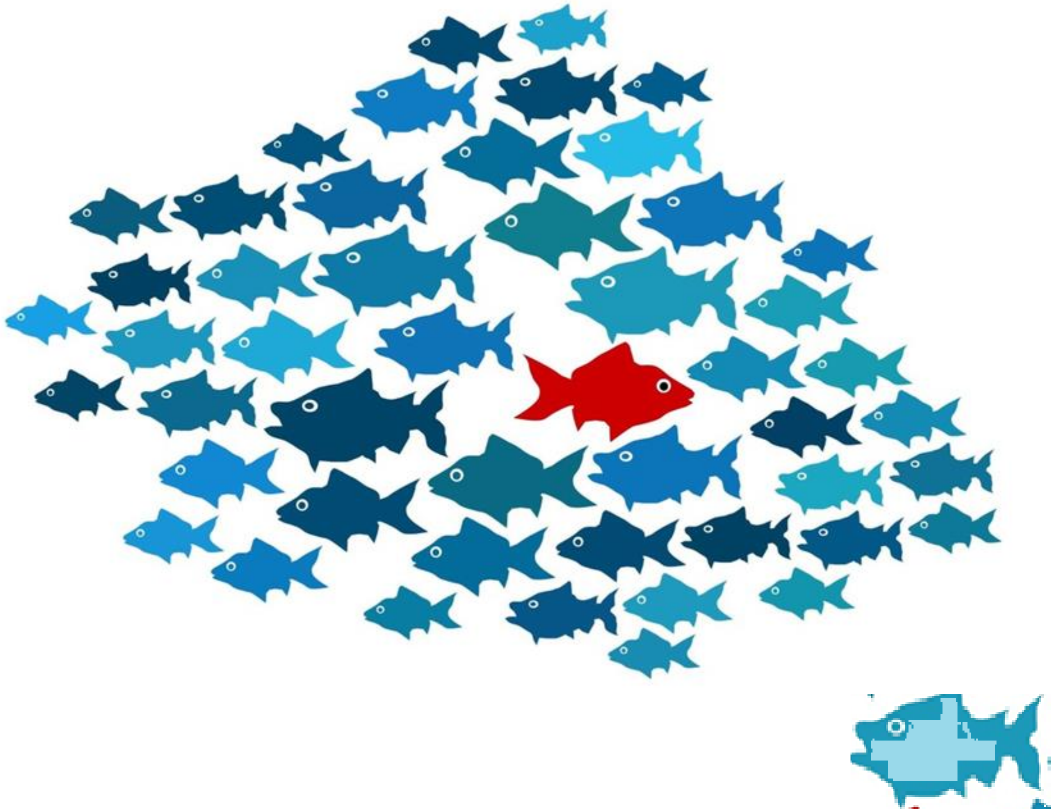
Large amounts of data

Sensitive to errors

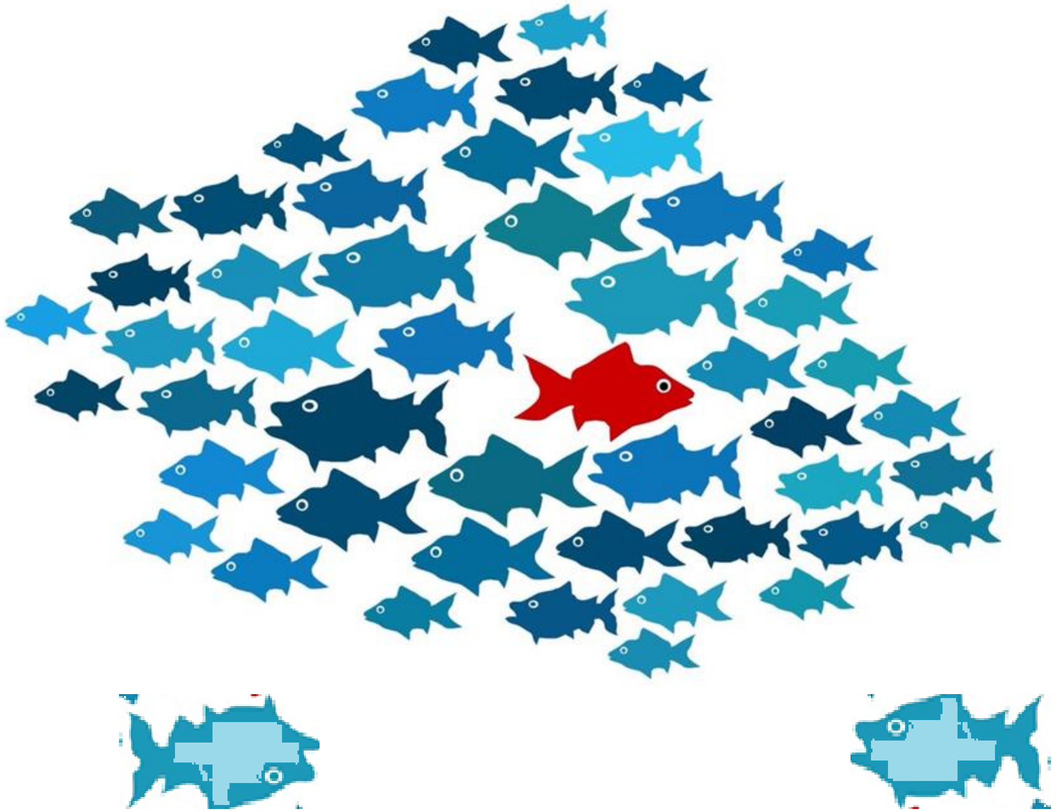
Rule based controls



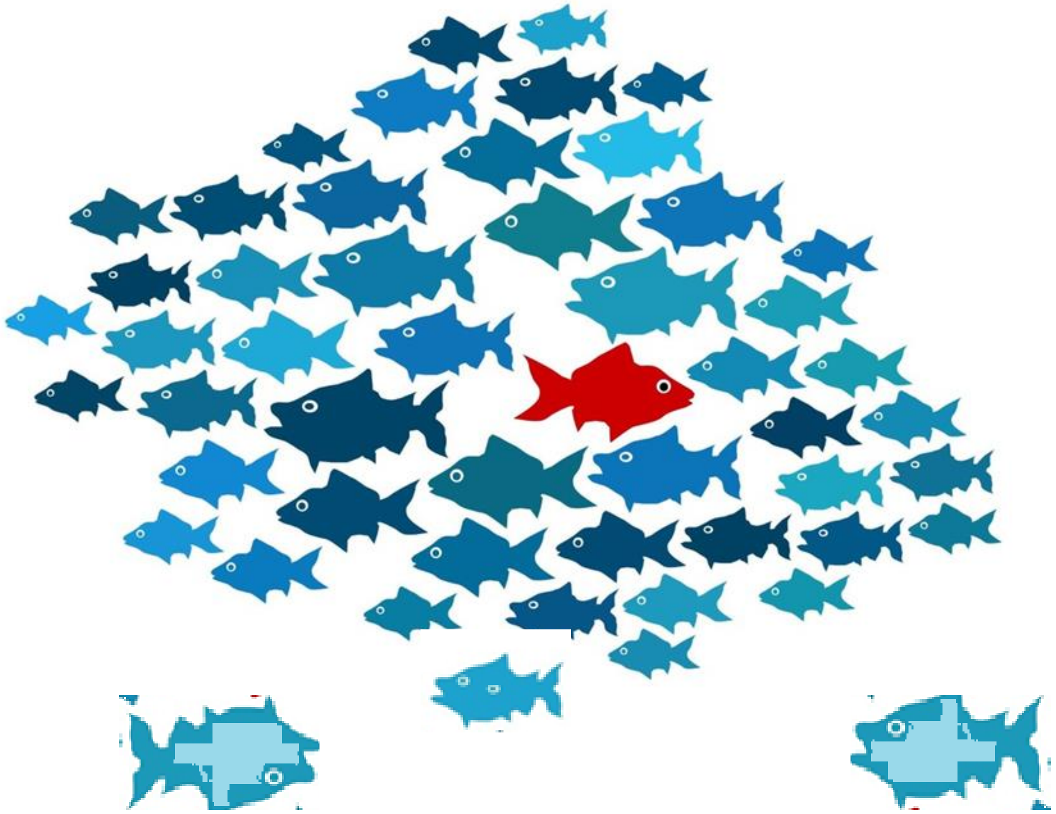
Rule based controls



Rule based controls



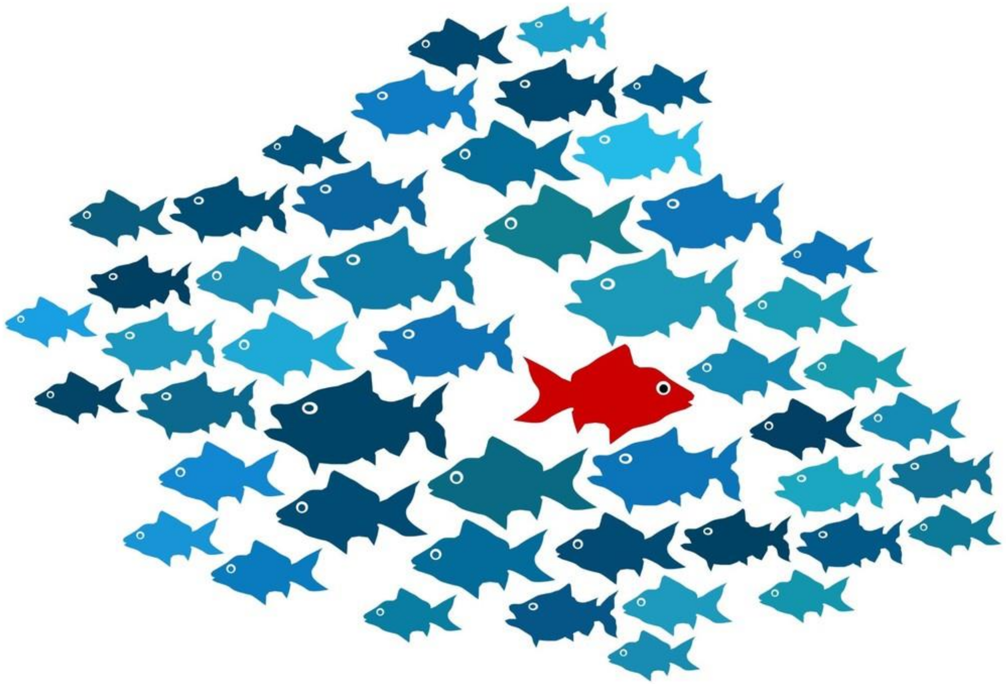
Rule based controls are not enough



Motivation for using ML algorithm

Outlier detection model

- Automatically detect errors
- Speed up data cleaning process
- Support rule-based controls
- Next step 'anomaly' detection



Models

DeNederlandscheBank

EUROSYSTEEM

Outlier detection

Ensemble learner

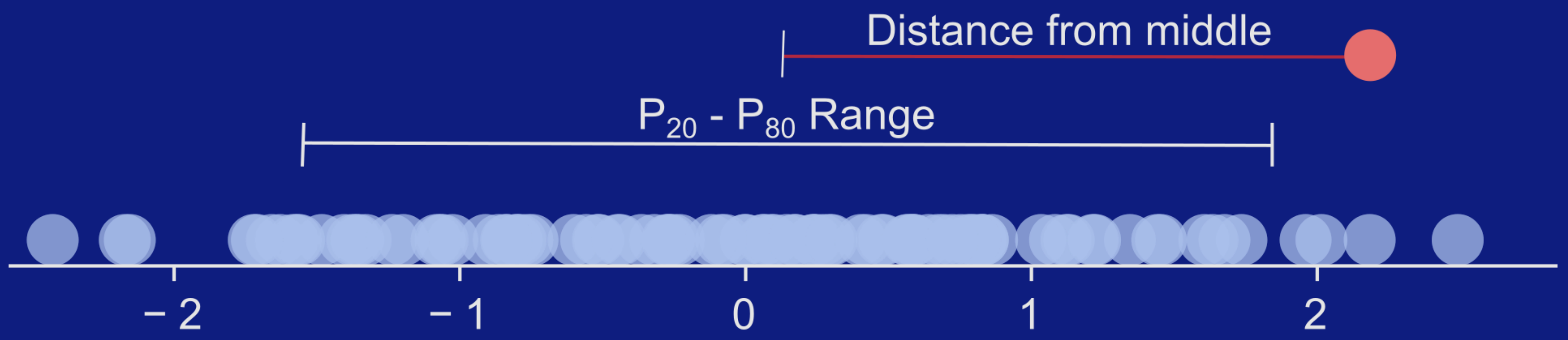
1. Make groups for the detection of outliers
2. Detect outliers with an ensemble of 3 or 3+ methods
3. Define a total score for each observation based on the different methods
4. Highest scoring observations are potential outliers

Ensemble methods

- Interquantile range
- **Nearest neighbourgh distance** (2)
- **Local outlier factor** (0)
- **Kmeans** (0)
- XGboost

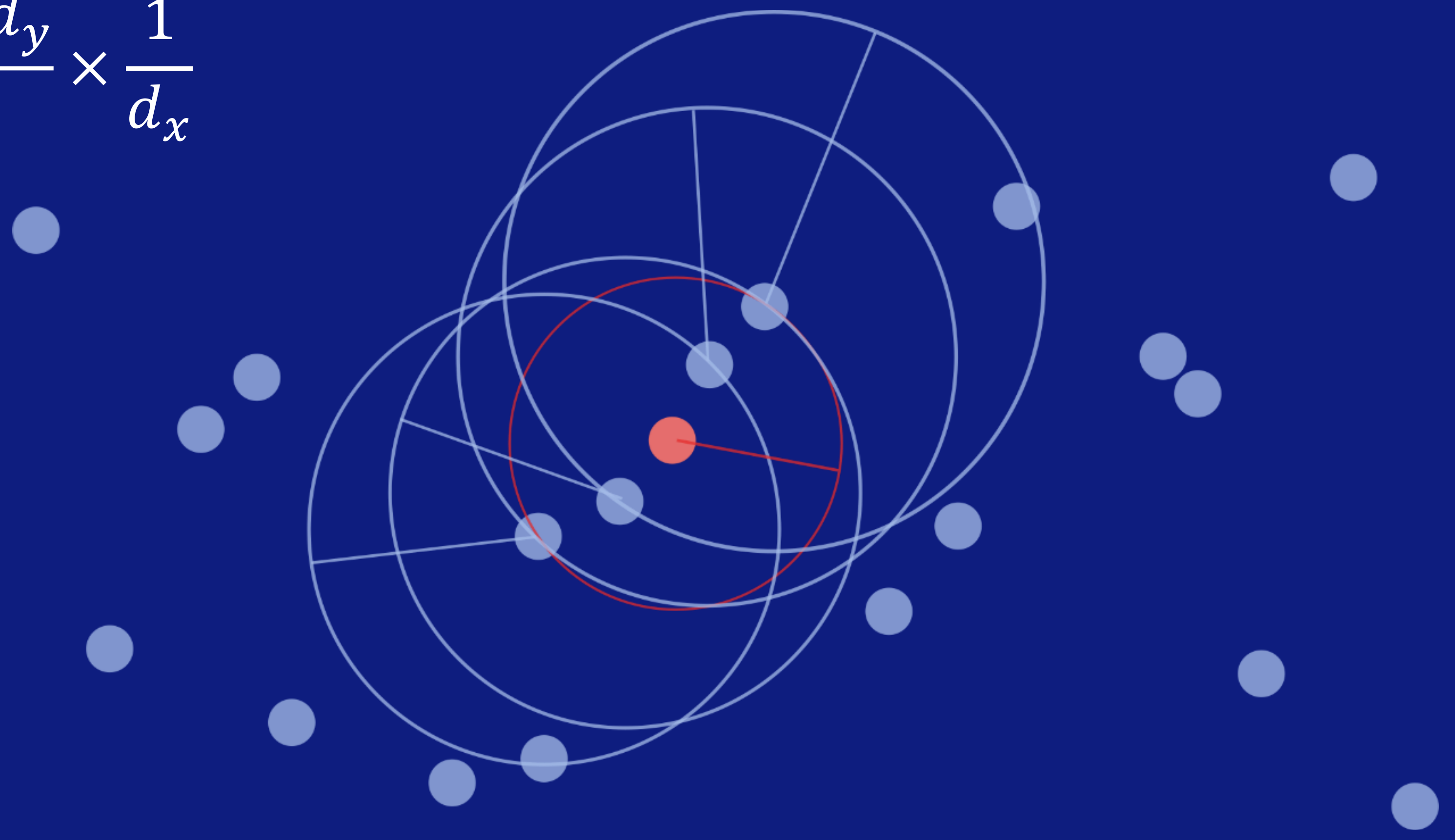
Interquantile range

$$iqr = \frac{x - (P_{80} - P_{20})/2}{P_{80} - P_{20}}$$



Local outlier factor

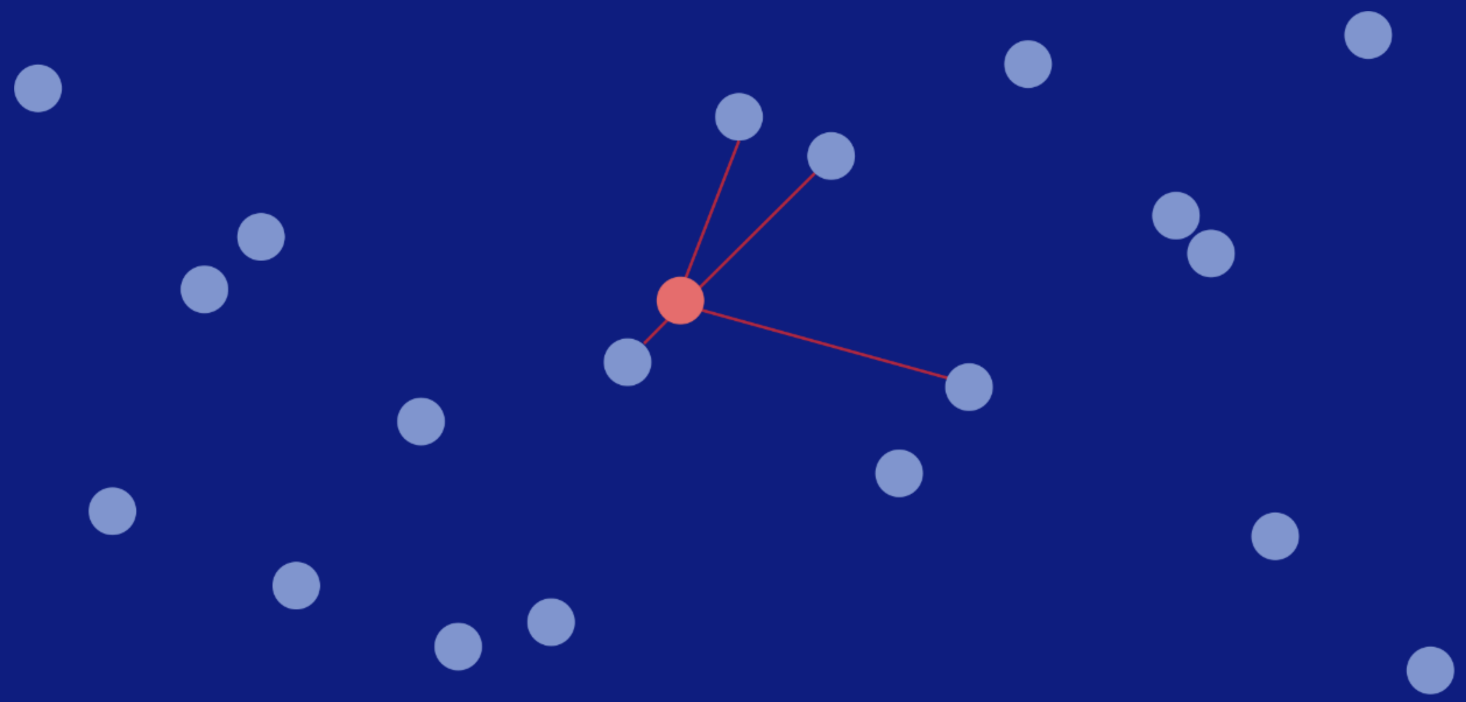
$$lof = \frac{\sum_{y \in N_x} d_y}{|N_x|} \times \frac{1}{d_x}$$



Outlier detectie

Nearest neighbourgh distance

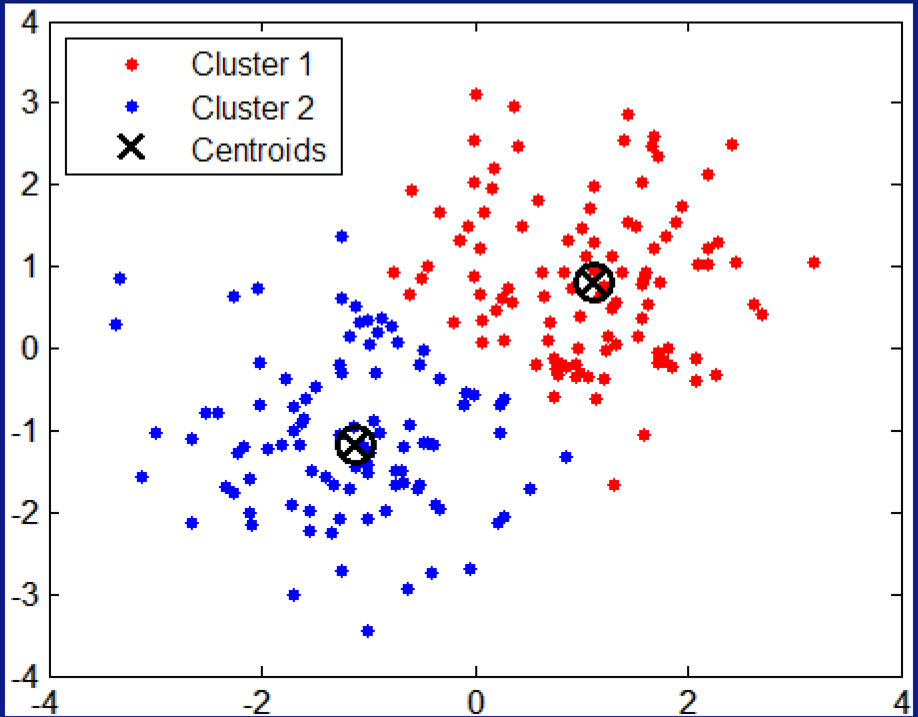
$$nn = \sum |x - x_{nn}|$$



Kmeans

K-means identifies *k* number of centroids
and allocates every data point to the nearest cluster
keeping the centroids as small as possible.

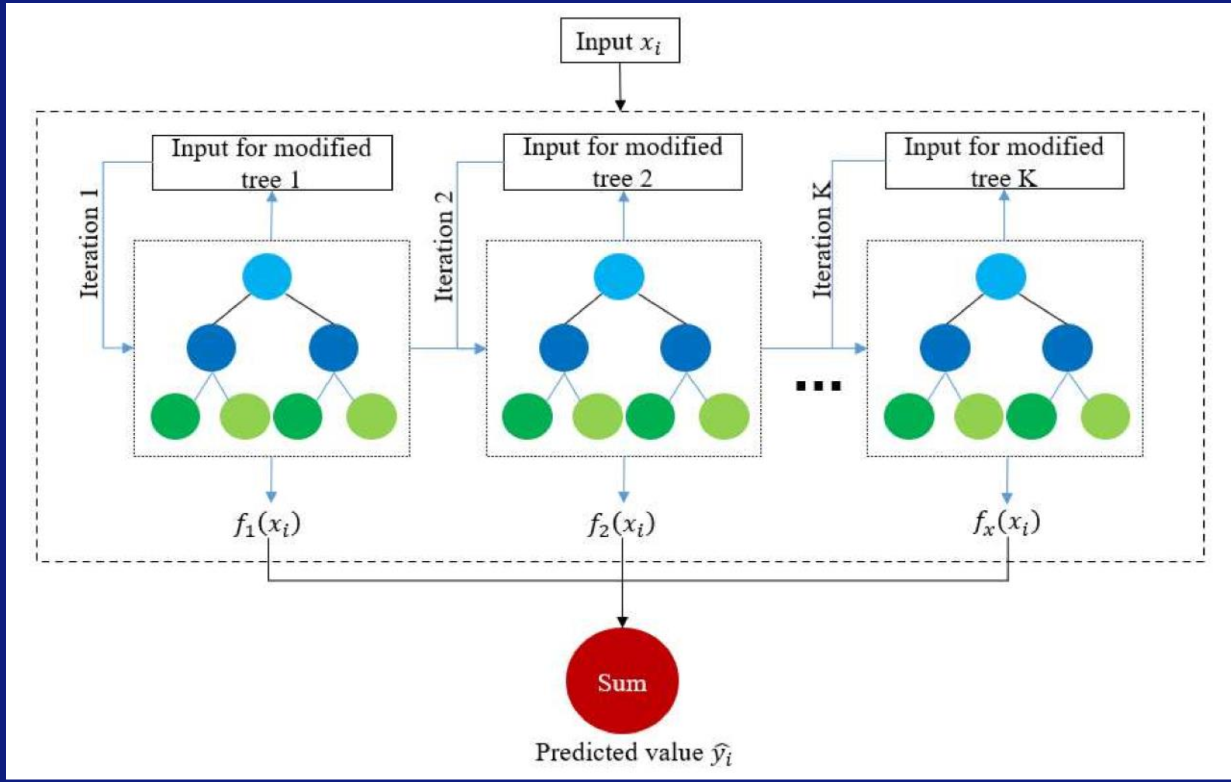
Towards outlier score:
Outliers are scored by calculating their **z-score**,
which is defined as the observation value minus
the centroid divided by the centroid's standard deviation.



XGBoost

Is a gradient-boosting decision tree algorithm.
It trains a number of trees sequentially and uses the fit of the previous tree to improve next fit.
It combines all the trees to create the ultimate predicted value.

As predictors, we (can) use several explanatory variables, like sector, credit quality and valuta.



Scoring

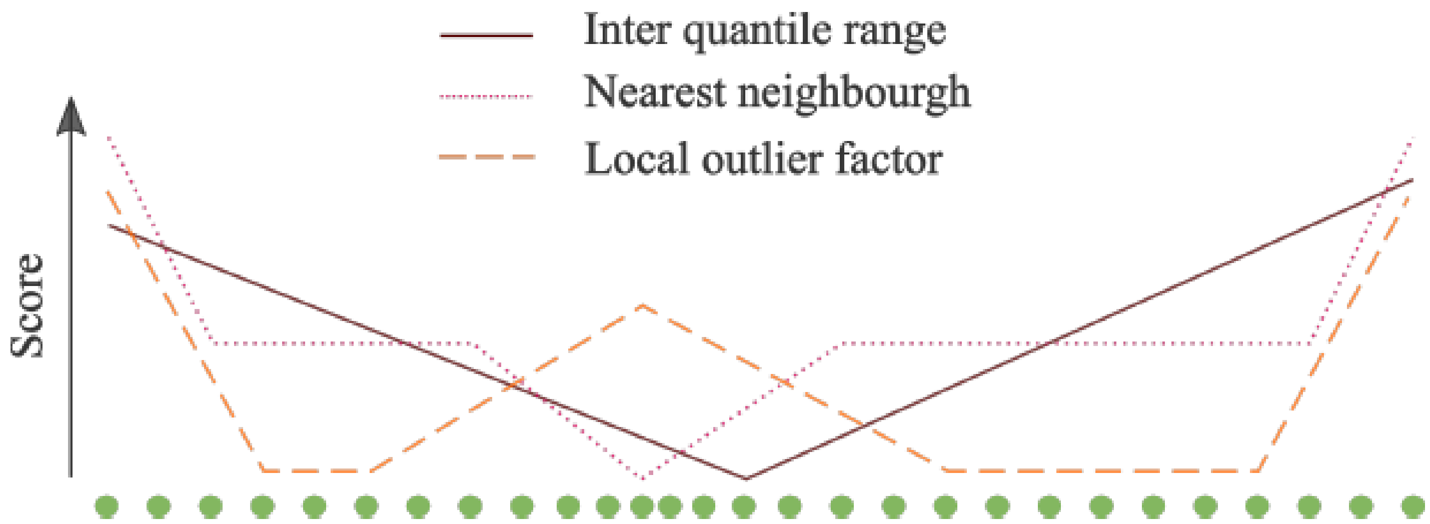
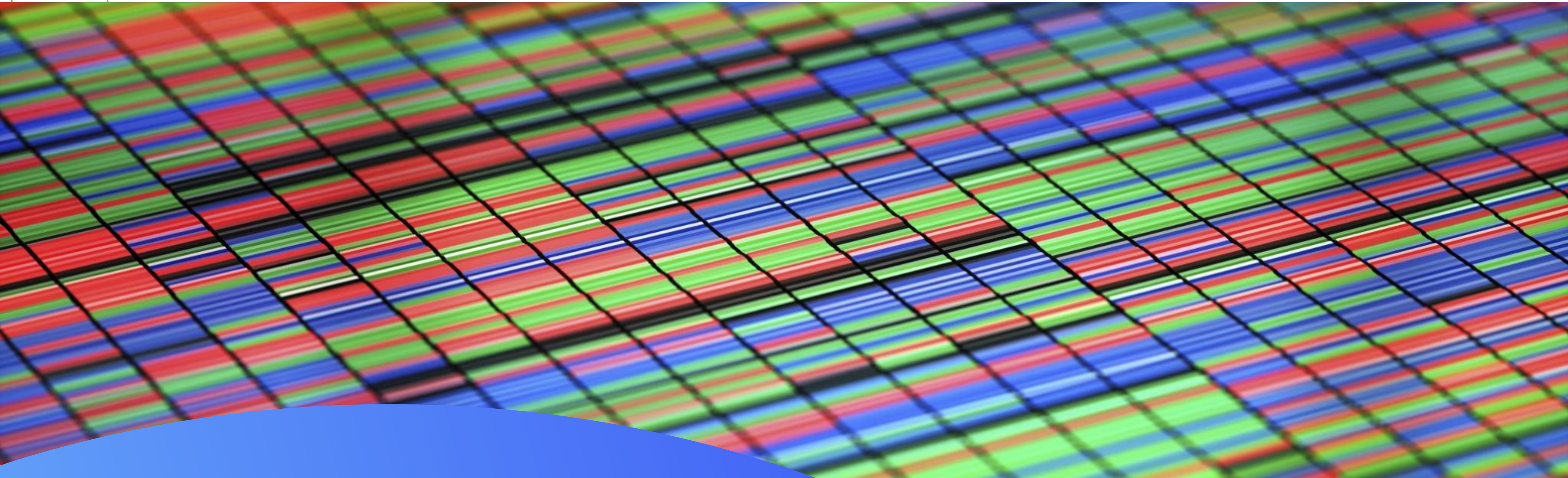


Figure 1: An overview of how the different methods would score points on a one-dimensional scale

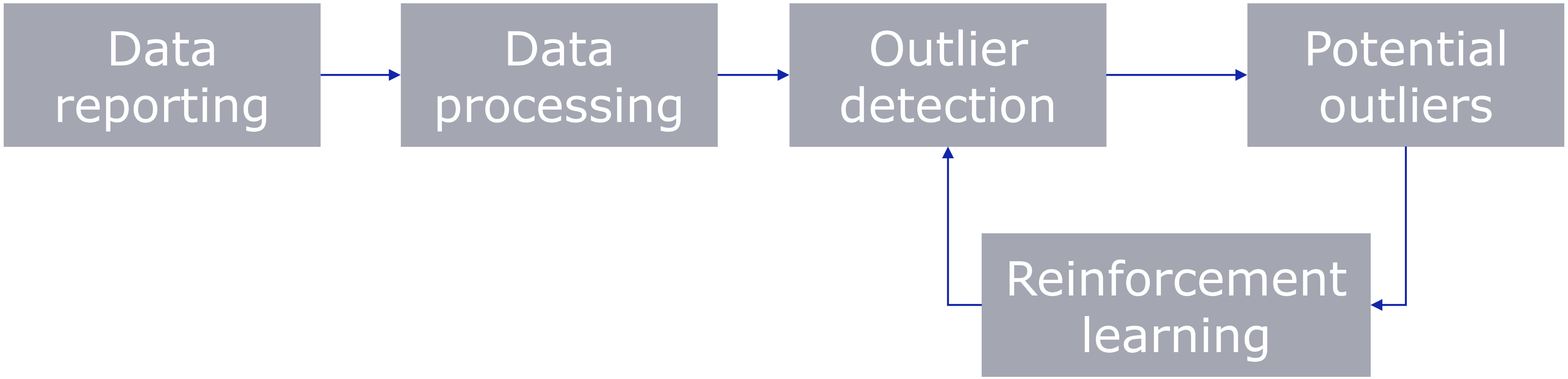


Approach: data to model

DeNederlandscheBank

EUROSYSTEEM

Approach


Data reporting
Data processing
Outlier detection
Potential outliers
Reinforcement learning

Approach

Data reporting → Data processing → Outlier detection → Potential outliers → Reinforcement learning → Outlier detection

- Combine reporting data
- Define waarde per eenheid
  - Difference w.r.t. median of previous period
- Remove stock splits

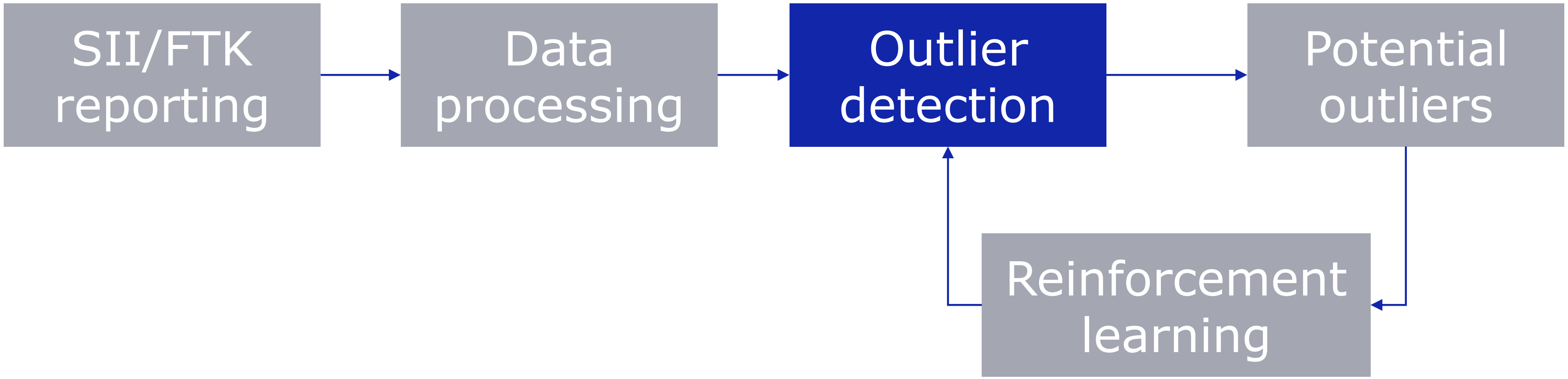
Data preprocessing

- Define 'waarde per eenheid'
  - Median of previous period
  - Percentage difference with previous period
- Create categorical variables
- Stock splits cleaning

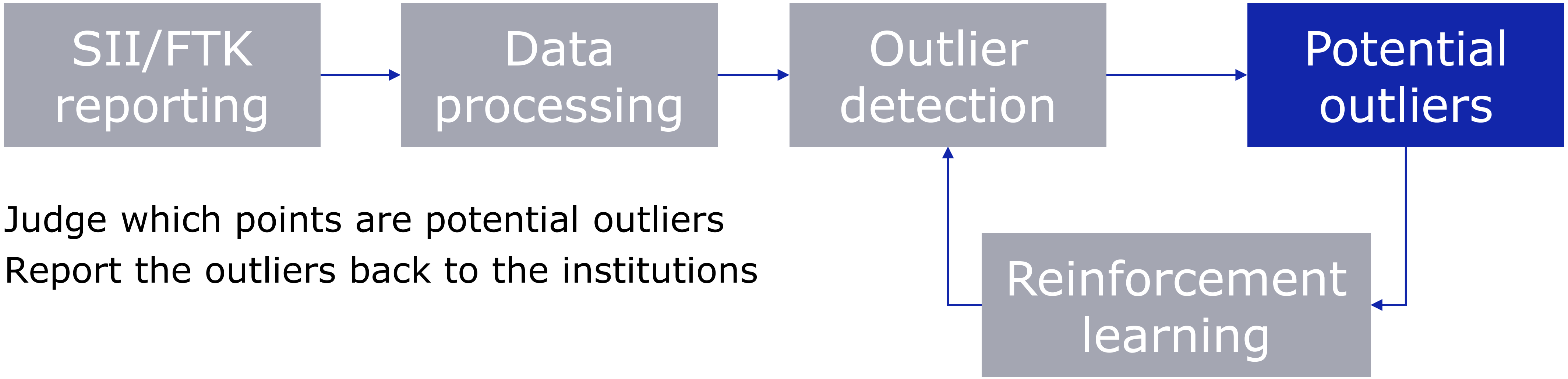
relatienummer	
URI	
aantal	
nominaal_bedrag	
aankoopwaarde	
marktwaarde	
aangegroeide_rente	
land	
valuta	
CIC	
rating	
kredietkwalitei	
interne_rating	
prijs_per_eenheid	
percentage_nominaal_bedrag	
URI_cat	
sector_main	
sector_sub	
sector_sub_sub	
waarde_per_eenheid	
waarde_per_eenheid_dt	

DeNederlandscheBank
EUROSYSTEEM

Approach

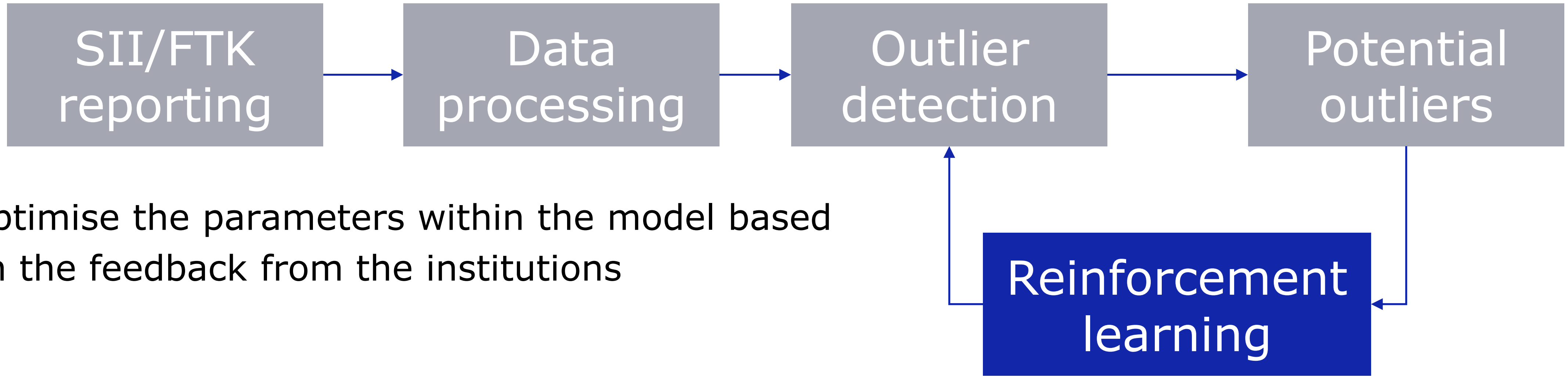

SII/FTK reporting
Data processing
Outlier detection
Potential outliers
Reinforcement learning

Approach



- Judge which points are potential outliers
- Report the outliers back to the institutions

Approach



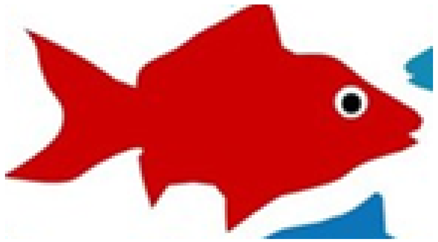
Optimise the parameters within the model based on the feedback from the institutions

Reinforcement learning

Reinforcement Learning (RL) is a ML technique that enables us to create a model that learns by trial and error through exposure with its environment.

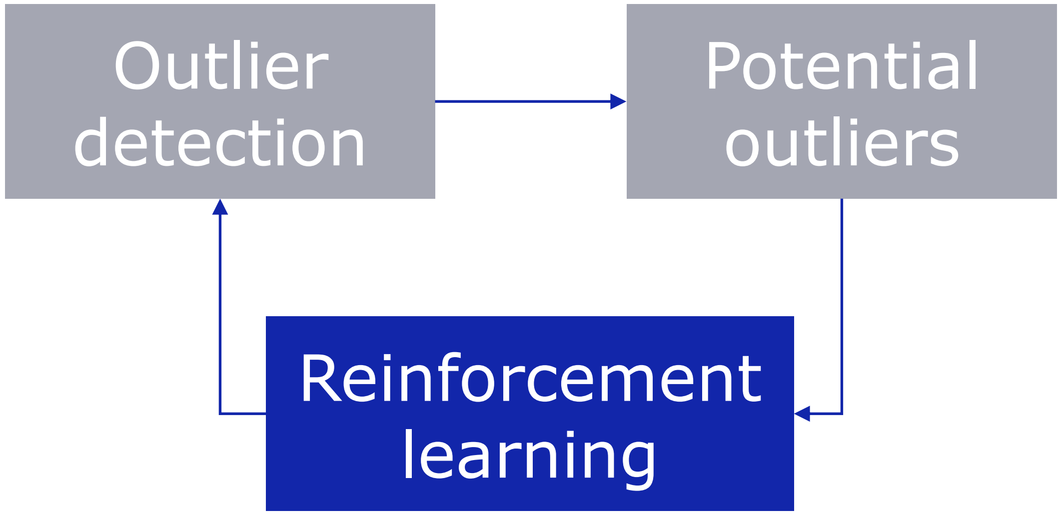
Reinforcement learning

- Our data is not on forehand **classified**: we do not know whether a point is an outlier or not:
    - *Do we deal with red or blue fish?*



Reinforcement learning

- Our data is not on forehand **classified**: we do not know whether a point is an outlier or not:
- Thus, setting optimal **model coefficients** on forehand is difficult
- Let the model learn over time, when we know which outliers identified: using **reinforcement learning**



Our reinforcement learner

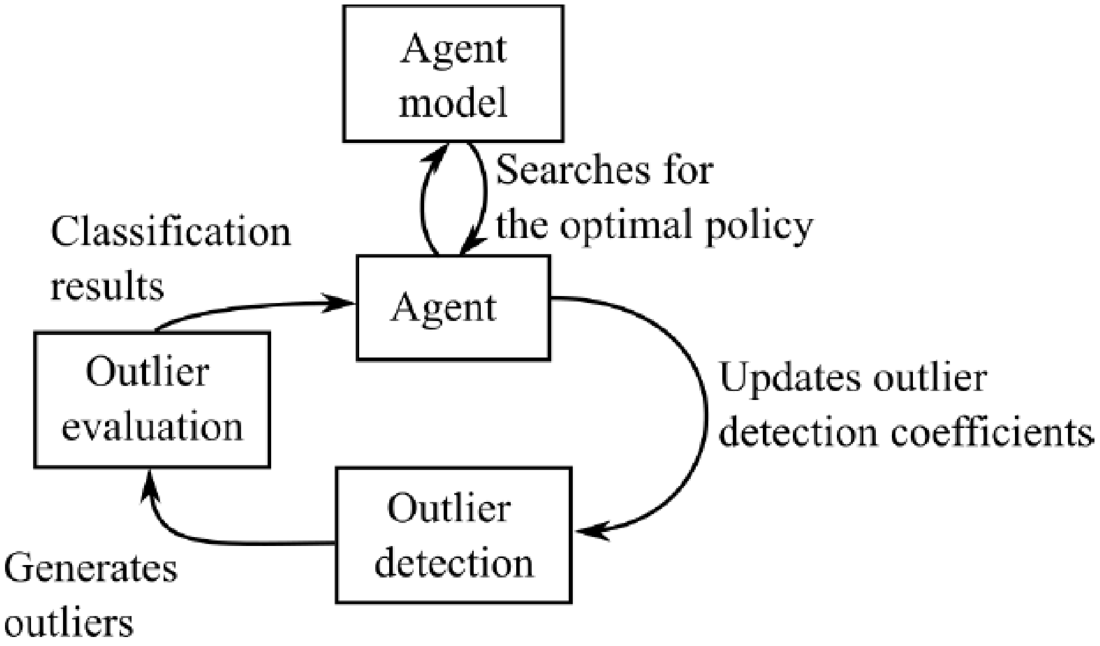


Figure 2: An overview of the reinforcement learning algorithm applied to the outlier detection

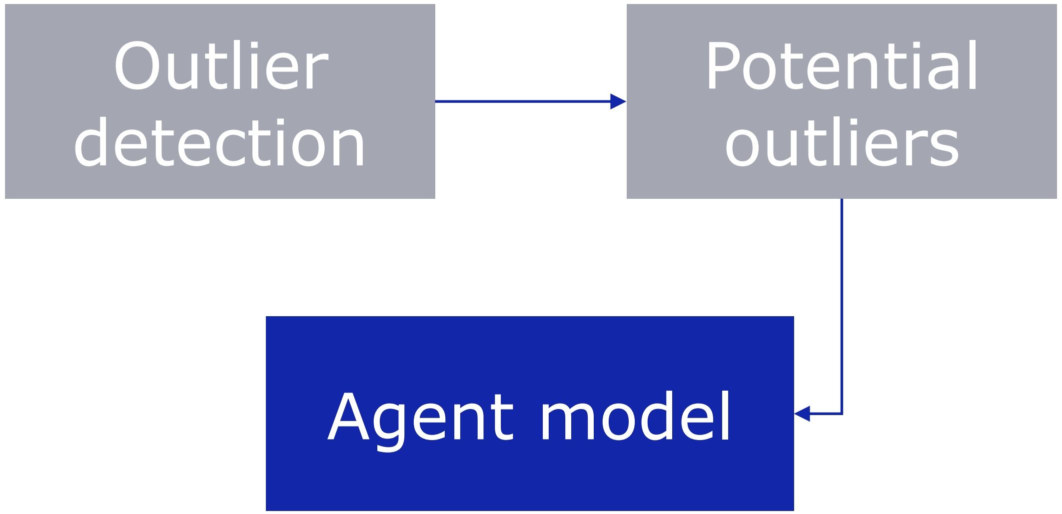
Our reinforcement learner

- The results are shared with the institutions, who file ammended reporting data

Outlier detection → Outlier evaluation

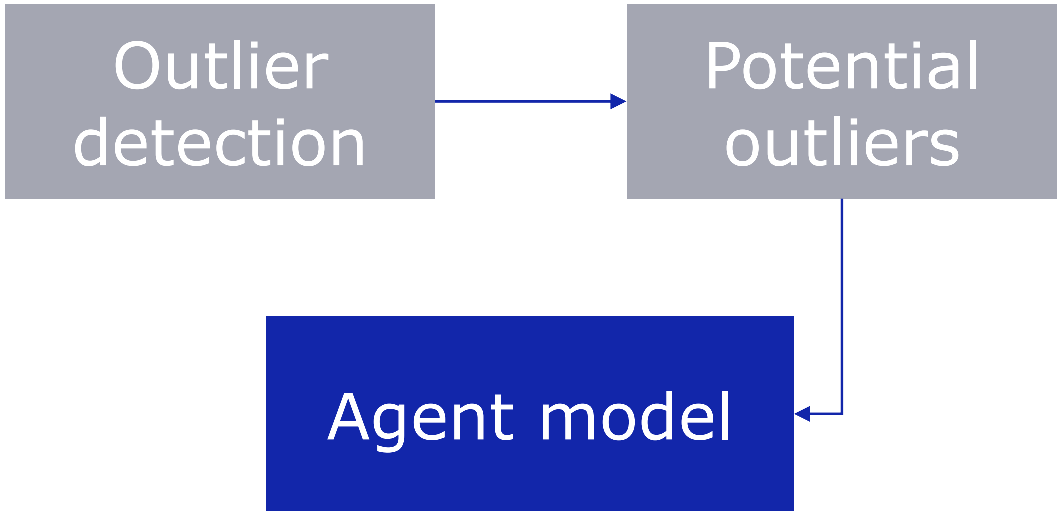
Our reinforcement learner

- The differences between the initial and amended reporting, and the outlier scores are recorded



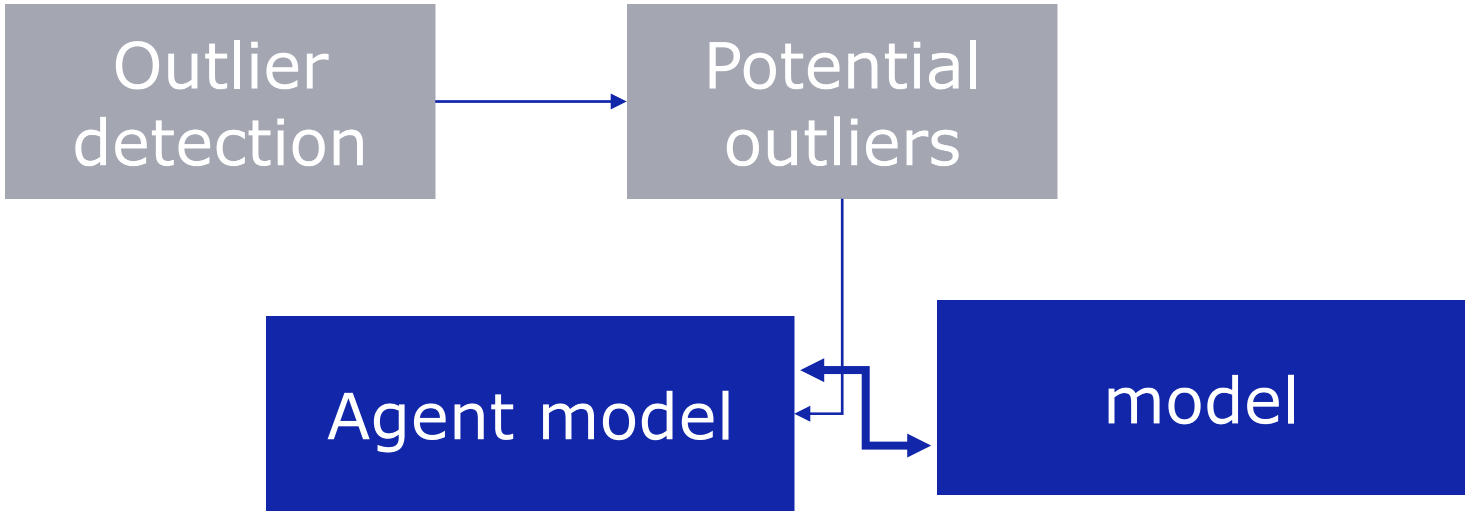
Our reinforcement learner

- The differences between the initial and amended reporting, and the outlier scores are recorded



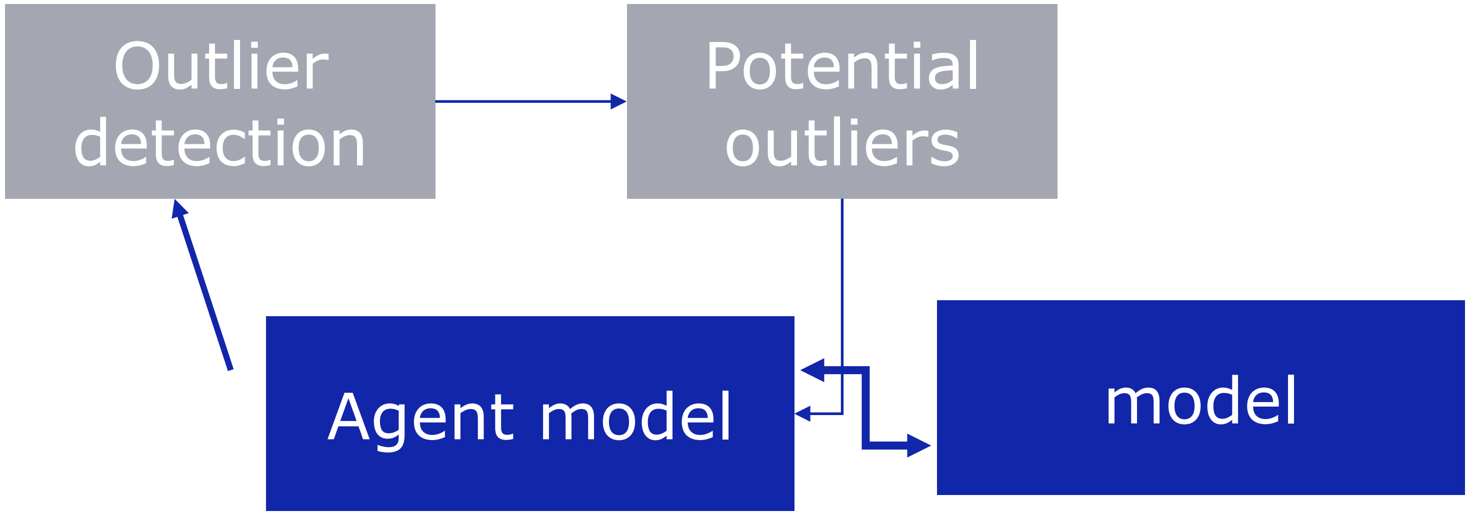
Our reinforcement learner

- The agent optimises the parameters of the outlier detection algorithm



Our reinforcement learner

- Parameters of the outlier detection are updated



Reinforcement Learning Assisted Outlier Detection for Costly to Verify Data

Michiel Nijhuis*

Data Science Hub, De Nederlandsche Bank, Amsterdam, 1000 AB, The Netherlands

Iman P.P. Van Lelyveld

Data Science Hub, De Nederlandsche Bank, Amsterdam, 1000 AB, The Netherlands

Department of Finance, VU Amsterdam, Amsterdam, 1081 HV , The Netherlands

Abstract

Within the financial reporting outliers in data are often due to data quality issues. However, some data anomalies are real and are of interest. Often extreme data points can be verified and the ground truth for a data point can be established. With the increasing granularity of data, checking all the data points is time-consuming, moreover the underlying issues leading to

DeNederlandscheBank
EUROSYSTEEM

Want to know more?

Read the paper!

Results

DeNederlandscheBank

EUROSYSTEEM

Same asset, different value?

Fonds	Periode	aantal	Aankoop-waarde	Marktwaarde	Asset	Prijs per eenheid	Waarde per eenheid
1	2018Q1	2072	€ 27,800.93	€ 41,109.01	French company	€ 18.32	€ 19.8
2	2018Q1	0	€ -	€ 2,100.76	French company	€ 18.32	€ 18.3
3	2018Q1	4598	€ 51,130.96	€ 84,200,23	French company	€ 18.32	€ 18.3

*Examples with fictional data

Did they sell the assets or not?

periode	aantal	aankoopwaarde	marktwaarde	asset	waarde per eenheid
2020Q4	20600	€ 81,701.17	€ 121,942.05	American company	€ 5.92
2021Q1	19400	€ 89,705.56	€ 130,008.09	American company	€ 6.70
2021Q2	18700	€ 70,528.19	€ 130,607.77	American company	€ 6.98
2021Q3	16003	€ 288,343.92	€ 1,310,654.20	American company	€ 81.90
2021Q4	1	€ 1.08	€ 1,200,899.20	American company	€ 1,200,899.20

*Examples with fictional data

Potential of outlier detection

DeNederlandscheBank

EUROSYSTEEM

Use cases outlier detection for actuaries

Fraud in claims

Outliers in reserving data

Data cleaning

Questions?

DeNederlandscheBank

EUROSYSTEEM