



Bridging the Data Quality Gap

Integrating traditional and modern techniques
for a seamless pension transition.

VSAE Actuarialcongress

5 March 2024

Frans Kuys, Bence Zaupper, Marino San Lorenzo

A FRESH TAKE ON RISK AND VALUATION



Introduction





Introduction

Presenters



Frans Kuys

Principal Consultant

frans.kuys@finalyse.com

- Head of Actuarial Practice in NL.
- Actuary in life insurance & pensions.
- 15+ years of experience: investment consulting, ALM, actuarial valuations, reporting & climate risk management.
- Qualified actuary: Institute of Actuaries (UK) & Actuariële Genootschap (AAG).



Bence Zaupper

Managing Consultant

bence.zaupper@finalyse.com

- Leads Insurance Risk & Actuarial Modelling working group.
- Actuary in life insurance & pensions.
- 20+ years of experience in pension valuations, life actuarial reporting (IFRS, GAAP, SII and BMA).
- Qualified actuary: Society of Actuaries (Ireland), Deputy Chair of Data Science Committee.



Marino San Lorenzo

Senior Consultant

marino.sanlorenzo@finalyse.com

- Actuary in non-life insurance & banking.
- 7 years of experience in reserving, pricing, actuarial modelling, data science, AI and Python programming.
- Qualified actuary: Institute of Actuaries in Belgium.
- Holds MSc in Actuarial Sciences from the Université libre de Bruxelles (ULB).



100+ consultants



7 office locations:

- Amsterdam
- Brussels
- Budapest
- Dublin
- Luxembourg
- Warsaw
- Paris



75 key accounts including major European financial institutions



Delivering projects in the entire EMEA region

A Europe-wide leading consultancy founded in 1988

We are here for you when it comes to incorporating changes and innovations in valuation, risk management and regulatory compliance.

Empowering you to make good decisions, and ensuring you can focus on your core business.

By bringing a unique mix of financial and technological skills, we offer a fresh perspective, unbiased analyses and modern answers to all your questions.

Our distinctive blend of expertise, pragmatism, team spirit and fairness has already contributed to more than 35 years of successful projects and trustful relationships.



Bridging the Data Quality Gap

Table of Contents

1. Data Quality Framework
2. Rules-based engines
3. Anomaly Detection
4. Data visualisations
5. Data Imputation techniques
6. Natural Language Processing
7. Q&A



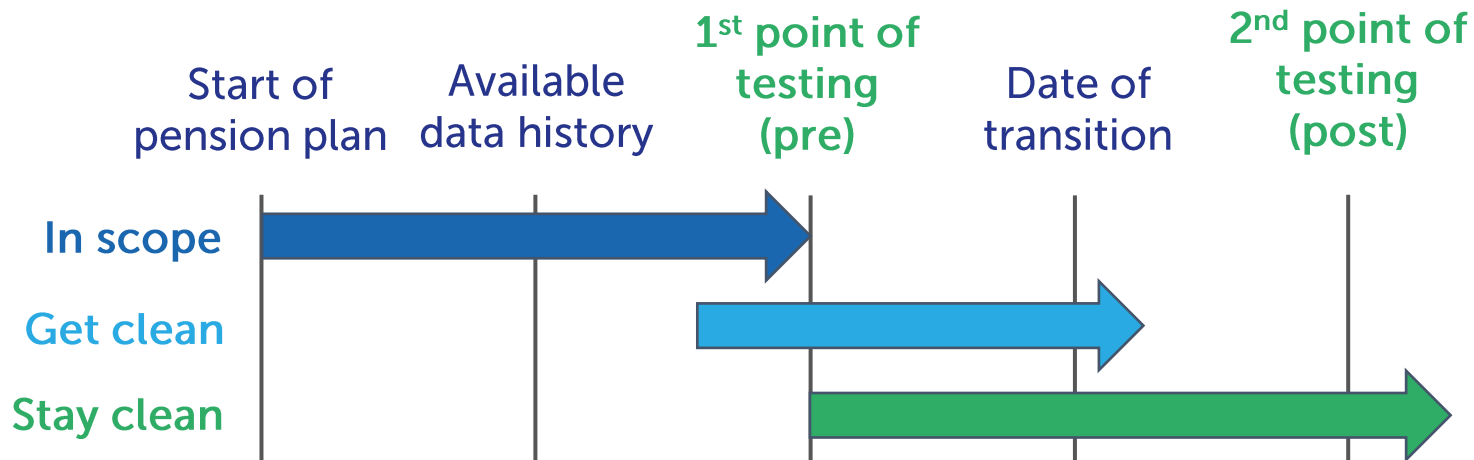
1 | Data Quality Framework





Pension Transition Deadline:

- The new pension law implemented in the Netherlands on 1 July 2023.
- All pension funds must make the transition to comply by 1 January 2028.
- Pension funds are facing a critical milestone in ensuring data quality leading up to transition date.



Source: Kader Datakwaliteit (Pensioen Federatie)



1. Design data quality policy

2. Risk Inventory & Assessment

3. Data Analysis & Partial Observations

4. Reporting & Assessment

5. Performing external audit (Accountant / IT)

6. Decision on data quality for transfer



Bridging the Data Quality Gap

Critical Data Elements (KDEs)

Personal Information

- date of birth
- gender
- marital status
-

Employment Details

- salary
- part-time %
- date of employment
-

Pensions Data

- type of scheme
- accrual %
- franchise
- contribution %
- accrued entitlement
-

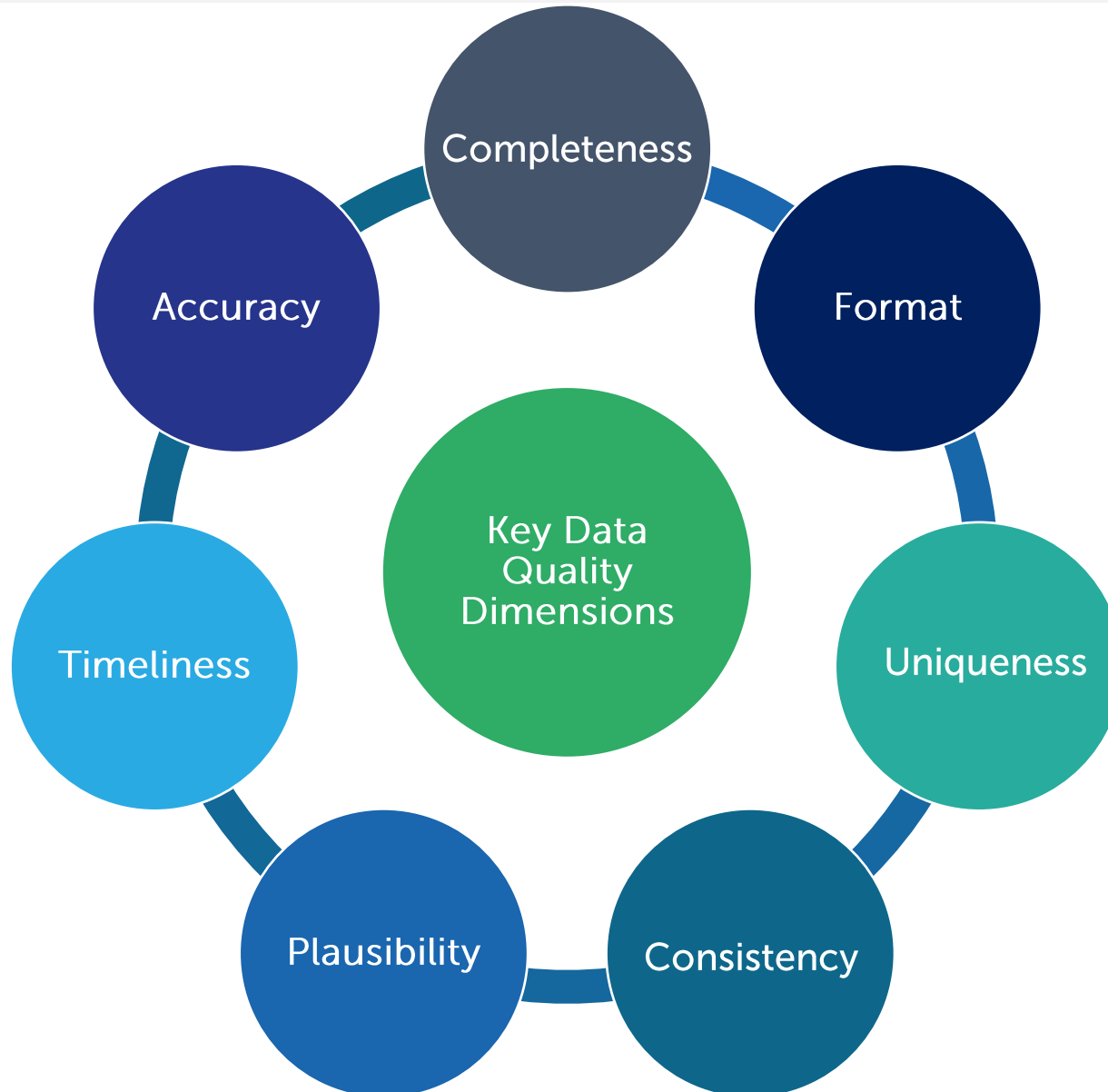
Payment Details

- bank account number
- pension benefit
- tax tables
-



Bridging the Data Quality Gap

Key Dimensions for Data Quality





2 | Rules-based engines





Examples for Rules-based Data Quality Checks

Completeness Check

- Member data compared to last reporting period
- Mandatory field (e.g. Salary) missing

Consistency Check

- Member's Gender not the same as last reporting period
- Historic salary increase of a member is above the specified threshold.

Uniqueness Check

- Check for Duplicates Records


Referential Integrity Check

- Member's Gender takes unexpected values (not 'M' or 'F')

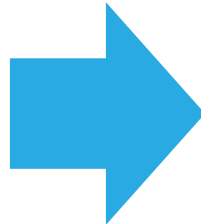



Data quality checks

Operational efficiency – rules based engine vs spreadsheets

 **Spreadsheets**

- Simple tool
- Transparent
- Difficulties in handling large datasets
- Risk of manual errors
- Lack of automation
- Manual process



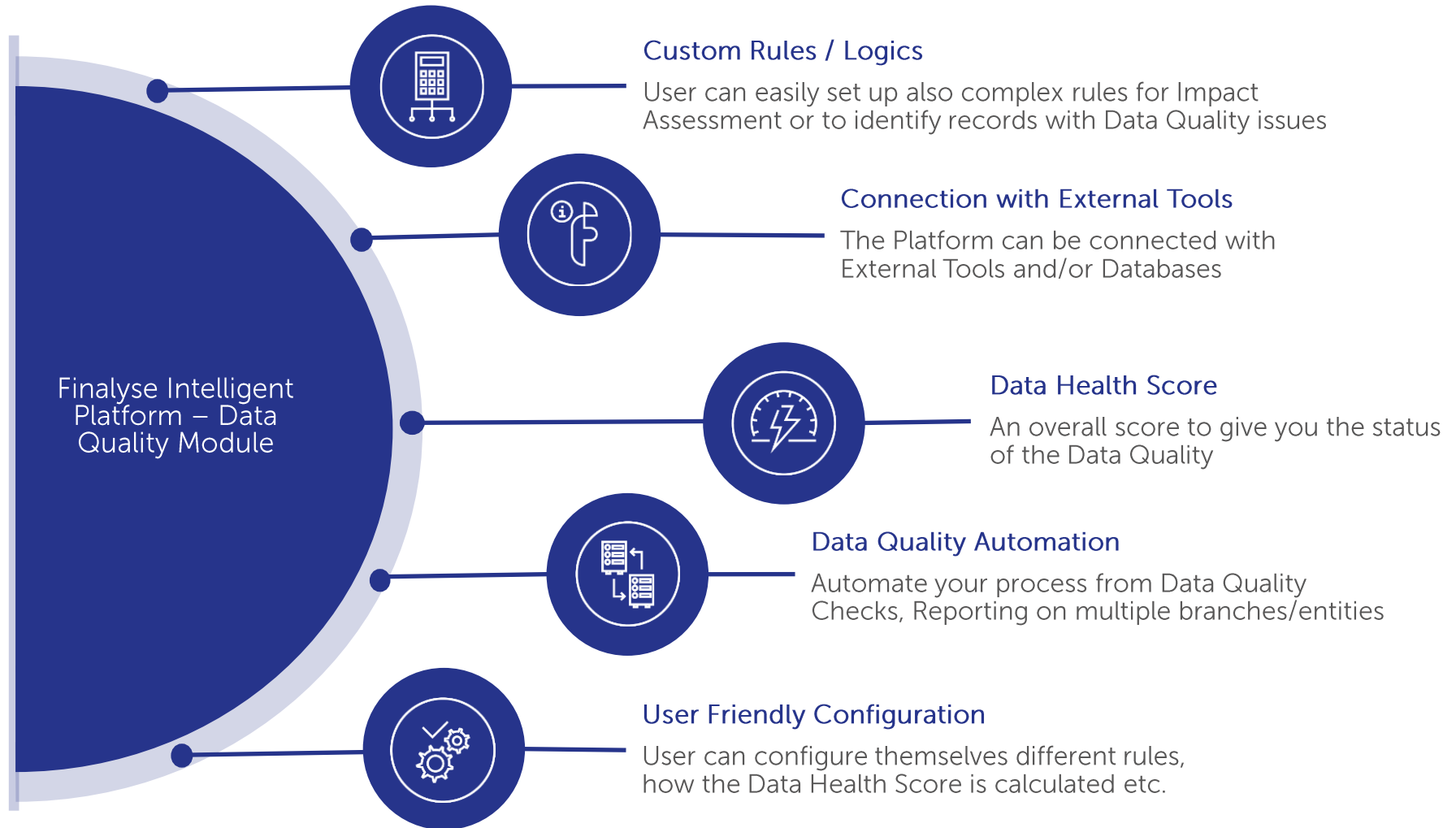
 **Rules based engine**

- Initial effort to set up
- Expertise in database technology and programming required
- Ability to handle large datasets
- Automated checks
- Customisable rules
- Automated reports and dashboards (health score)



Finalyse Intelligent Platform

Data Quality Solution





How can Management and Data Quality Specialists use the platform?

Management	Specialist
Data Health Score	Build Data Quality Rules & Fallback Rules
Data Quality Dashboards	Check Data Validation Logs sorted by Impact
Specific Reports for Management	Use built-in SQL Editor to further investigate issues
	Get Automatic Data Quality Reports based on defined Criteria



Technologies Used

- SQL Server
- Python
- HTML, CSS, JavaScript etc.

Performance

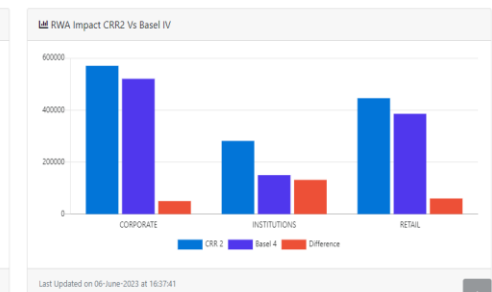
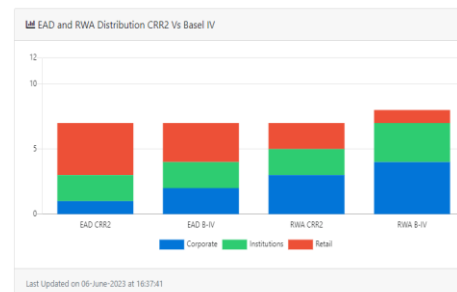
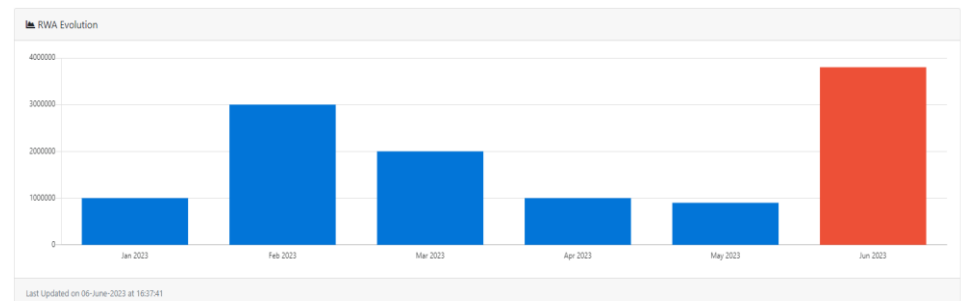
- Can easily process up to 5 Million records in each calculation

Audit

- Dedicated Audit log to track what each user has done when.
- 4 Eyes Principle (One user inputs another authorises)

Roles

- User Role Management on:
 - Functionalities
 - Processes
 - Data





3 | Anomaly Detection





- An alternative to rules-based checks
 - Data quality rules not specified by user in advance
 - Instead, the aim of the unsupervised machine learning algorithm is to learn patterns from the data independently
 - Identifies anomalies – data records that are unusual based on the rest of the dataset and patterns learnt during training
 - Human intervention is required to review any anomalies
- Anomaly detection is a natural application of deep learning models due to similarities with pattern recognition in image processing.
- The performance of the recently introduced Anomaly Transformers are promising.



Anomaly Detection Problem

- Unsupervised anomaly detection is a challenging task.
- We expect the model to learn complex dynamics of temporal data and derive a distinguishable criterion for detecting rare anomalies from plenty of normal time points.

Transformer model

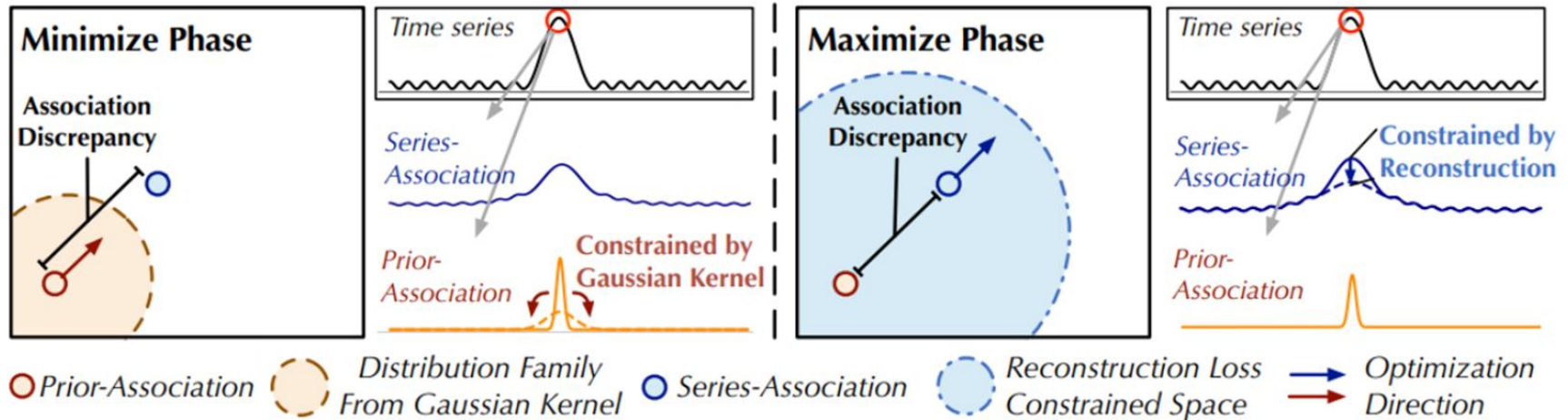
An adapted version of the Transformer model is used to solve this task with the following key features:

- Anomaly Attention mechanism
- Define Anomaly Discrepancy
- Distinguish abnormal from normal points (using Abnormal Discrepancy)



Anomaly Transformer

Deep learning model



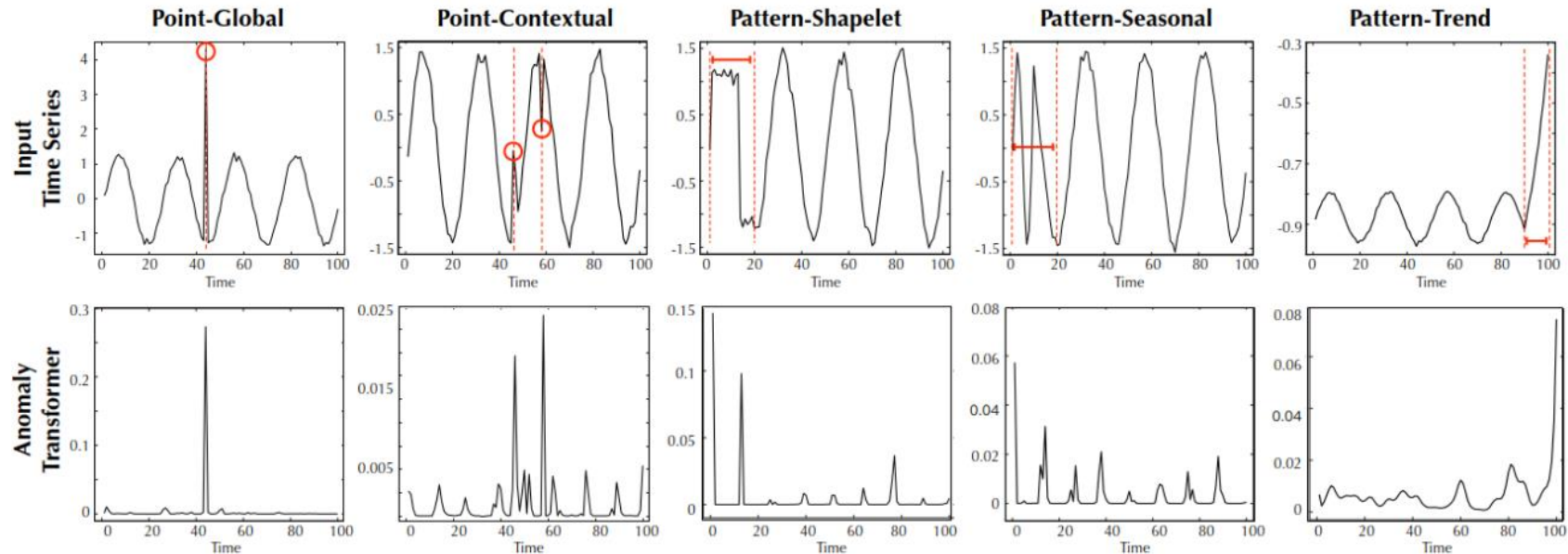
Source: [Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy \(Article\)](#)
(Article)

- Minimise phase - Prior Association determined by learning the temporal patterns from the raw series using a Gaussian kernel.
- Maximise phase - Anomaly Discrepancy maximised to detect abnormal points using reconstruction loss.



Anomaly Transformer

Example – wide range of anomalies detected on seasonal data



Source: Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy (Article) (Article)

- The anomaly transformer detects a wide range of anomalies using association discrepancy in an unsupervised setting.
- Some of these would not be spotted by traditional time series models.



Anomaly detection for historic salaries

- Historic salary data is a time series hence a similar anomaly detection algorithm could be implemented
- Aim of the algorithm is to detect members with unusual salary history without specifying what "unusual" means
- Validation of detected anomalies to be performed by human:
 - There may be a few false alarms where the data is actually correct, but still, quick identification of suspicious cases may reduce time spent on data cleaning significantly

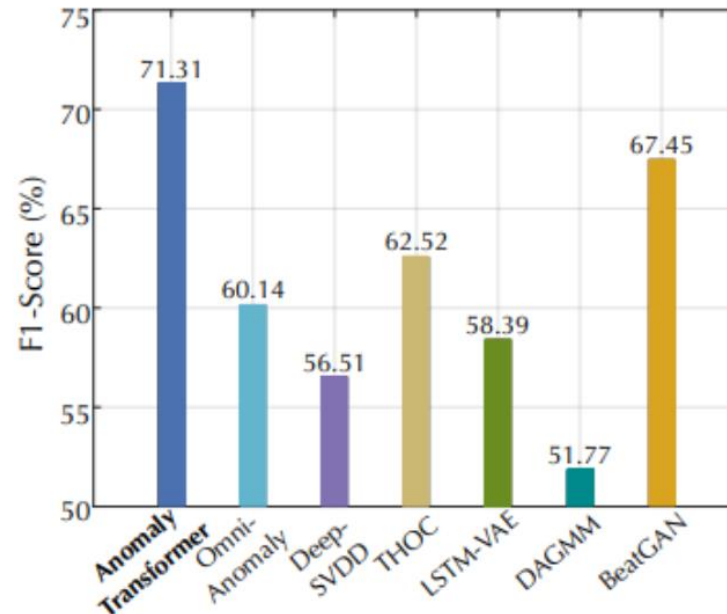
Anomaly examples

-
- In general, salaries are expected to increase over time, so members with significant salary decreases may be detected as unusual
 - Salary increases usually follow similar dynamics due to the link to inflation (with some known deviations such as promotions), therefore members with completely different patterns may be detected as anomaly
 - Other unusual patterns could trigger attention, e.g. salaries going up and down or frequent significant drops or increases
 - Although promotional increases are possible, too frequent significant salary increases may be detected as anomaly



Anomaly Transformer

Model performance



Source: Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy (Article)

- As shown on the chart, the Anomaly Transformer outperforms other models on a wide range of anomaly detection tasks. It also outperforms traditional ARIMA methods.
- The accuracy of the algorithms was measured using the F1-score (a balanced accuracy metric widely used in machine learning).



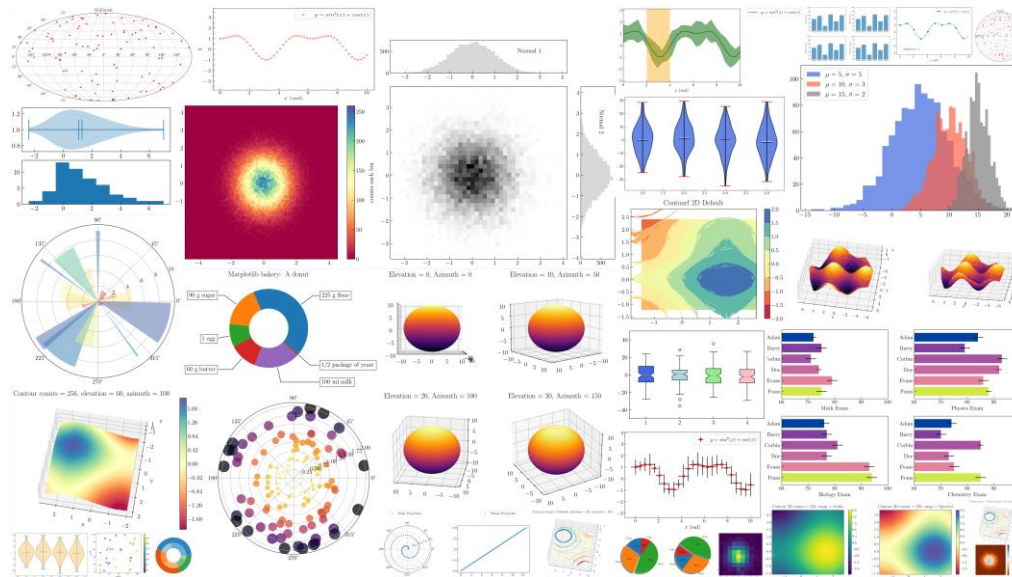
4 | Data visualisations





- What are the most common data quality issues? Missing values, outliers, inconsistent formatting, duplicates or incorrect data.
- However, it is possible that the data is correct but for its distribution has changed.
- Why can this be a problem? Because any data drift can have an impact on model performance and thus an impact on risk assessment.
- Python tools enable us to monitor and visualise such drifts. Most popular libraries are: Matplotlib, Seaborn, Plotly and Pandas.

Data visualisation with Matplotlib - examples





- Open-source dashboarding libraries (e.g. Dash or Streamlit for Python or Shiny for R) enable building interactive, transparent monitoring dashboards that are easy to create and maintain.

Dash dashboard example





Monitoring data drifts

Features above Threshold

1 (out of 7 selected) features

▼ List of features

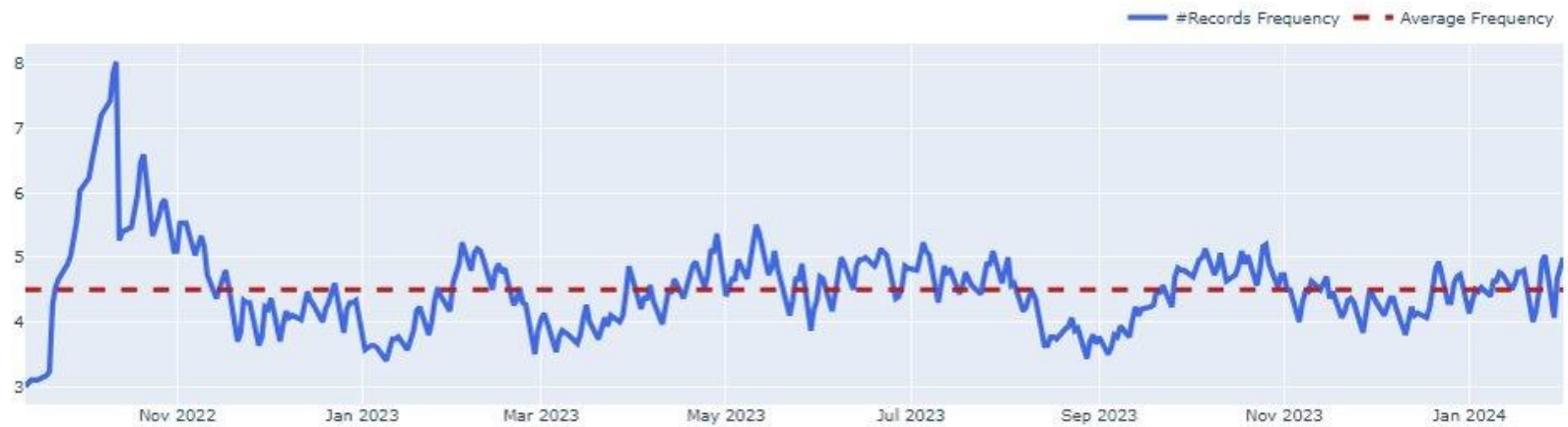
fe_thismonth_in=tc_intt_c_n

Display

Information

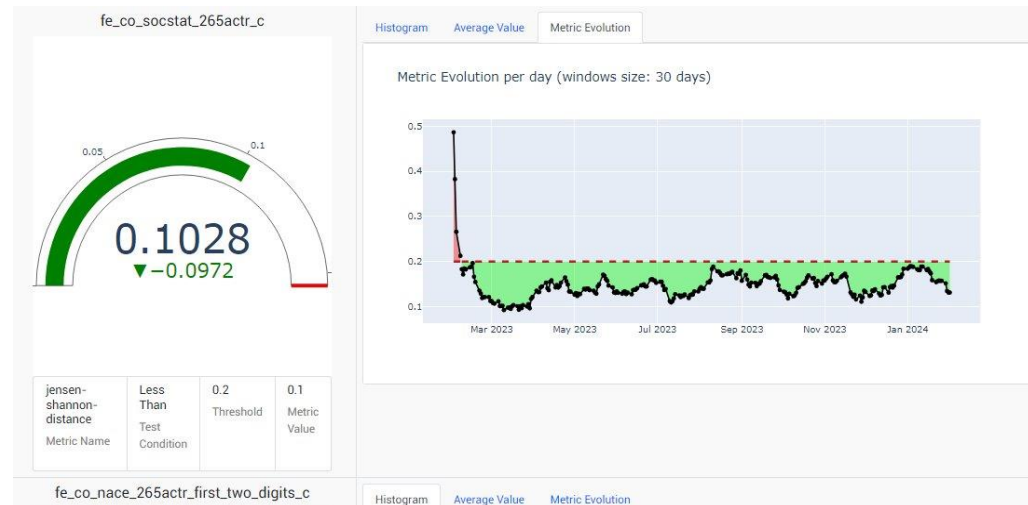
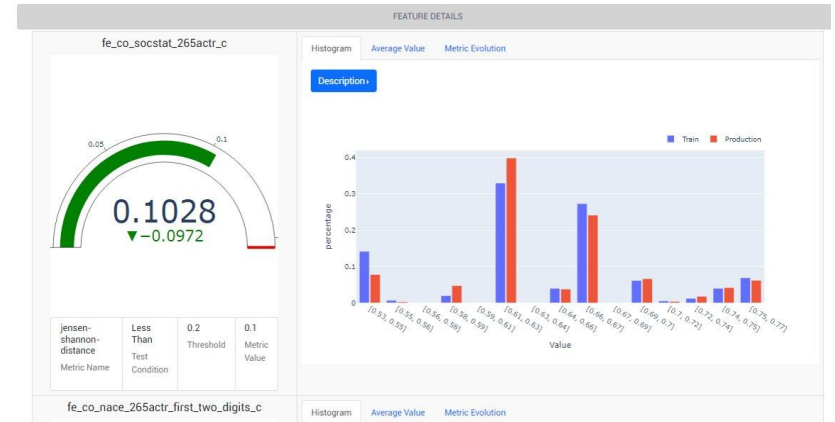
Production - #Records Frequency

#Records per day (Window size = 30 days)





Monitoring data drifts – cont'd





5 | Data Imputation techniques





Data Imputation techniques

- Ways to handle missing values in an automated manner (3 options):
 - Discard feature with proportion of missing values (above threshold)
 - Keep the feature as it is since some estimators are able to deal with missing values
 - Use data imputation techniques
- Data imputation consists of replacing a missing value using simple or more advanced statistical methods
- Generally, there are two types of techniques:

Univariate	Multivariate
<ul style="list-style-type: none">+ Easy to implement- Ignores relationships between features (issue if data is not randomly missing)	<ul style="list-style-type: none">- Can be computationally expensive+ Captures relationships between features
<i>Example:</i> Replace missing values with mean or mode of the observed values from the same feature/column	<i>Example:</i> k-nearest neighbors (KNN) or Multiple Imputation by Chained Equations (MICE)

- These techniques are implemented in Scikit-learn (one of the most popular ML libraries in Python)



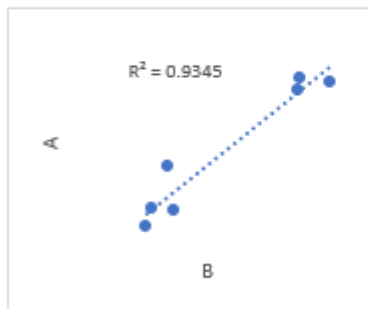


Data Imputation techniques

MICE algorithm

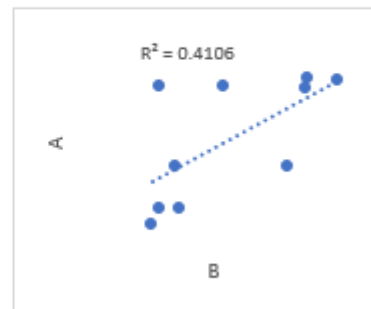
Missing data is in red. There is a strong correlation between A and B, so let's try to impute A using B and C.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45



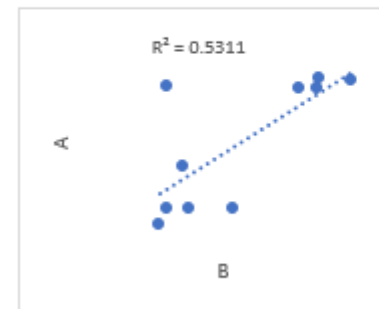
Missing data is filled in randomly. This dilutes the correlations, but allows us to impute using all available data.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45



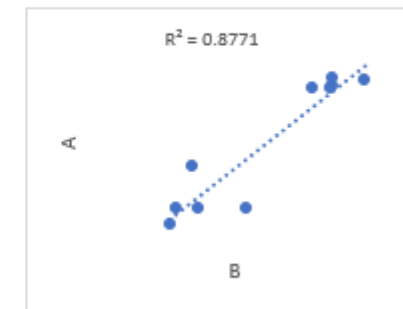
A random forest is used to predict A with B and C. Notice the correlation between A and B improved.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45



After Imputing B using A and C, we have achieved a correlation between A and B much closer to the original data.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	1.24	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45





6 | Natural Language Processing (NLP)

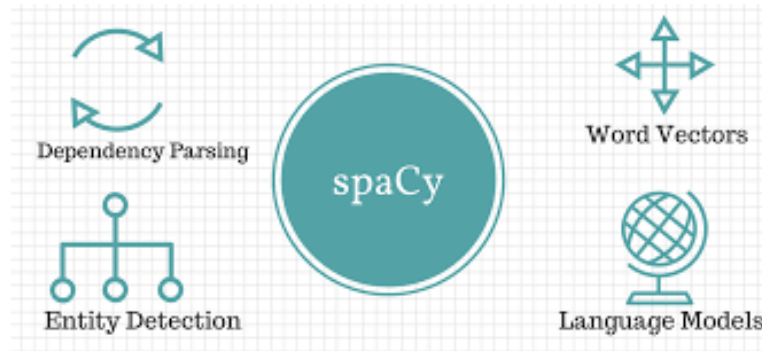




Natural Language Processing (NLP)

Data cleaning and standardisation

- NLP can be used for improving data quality and especially for standardising input data.
- For example, you might have a feature, that is a text feature which contains information or inputs written by users. You might want to summarise this feature into either a sentiment score or to categorise this text information.



- Solution: load pre-trained LLM model (OpenAI, Spacy, Lab21, ...) to answer a question with a specific prompt and returned format.
- Example of task: "extract country from those texts under this format:{"country_name":country}"

"I was in Holland and had a great time there."

"We were in the Netherlands and had a great time."

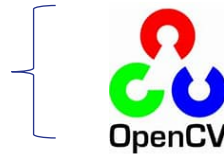



Natural Language Processing (NLP)

Extracting information from pension documents




Optical character recognition (OCR) 



Document segmentation 



Document classification 



Semantic search 



Large Language Models 





Thank you for your attention! Questions?

Amsterdam

Parktoeren
Van Heuven Goedhartlaan
13D-1181LE Amstelveen
+31 20 808 36 28

Brussels

"The Artist"
Avenue des Arts 9
B-1210 Brussels
+32 2 537 43 73

Luxembourg

12 rue Jean Engling Bte 9B
L-1466 Luxembourg

+352 260 927 (FVS)
+352 27 40 1757 (Consultancy)

Budapest

Széchenyi István tér 7-8
H-1051 Budapest

+36 1 354 18 90

Dublin

Upper Pembroke Street 28-32
D02 EK84 Dublin

+353 1 608 7705

Warsaw

Chłodna 51,
00-867 Warsaw

+48 22 22 30 559

Paris

Rue Taitbout 13-15
75009 Paris

+32 2 537 43 73

