Parallel Session 1.2

Towards automated data cleaning: a process for the auto-detection of data anomalies and inconsistencies

Presented by: **Jennifer Bradford**

Director, Data Science, PHASTAR

Towards automated data cleaning: a process for the auto-detection of data anomalies and inconsistencies.

Jennifer Bradford

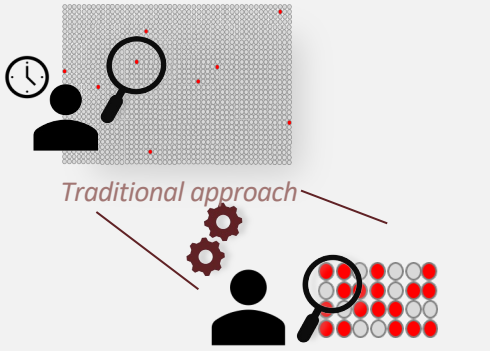
Introduction

- Data Science and Data Management experts working closely together.
- This close collaboration enables the team to Identify areas that data science techniques may provide benefit.
- Two projects to explore the application of data science and machine learning approaches within data management.

phastar

Intelligent Approach to Data Cleaning

Quality data is critical for a successful clinical trial and data managers manually inspect the data to check for errors or inconsistencies. At PHASTAR we are looking to drive efficiencies in this process through the use of different data science approaches.



Auto-Query detection

Apply rule-based approach to highlight potential data issues, supporting the data management team prioritise their activities.

Anomaly detection

Application of machine learning to identify anomalies; initial focus on centralised statistical monitoring & high-risk site identification.

Data Visualisation

Effective visualisations can help teams generate insights and monitor data, particularly if they are targeted and intuitive.

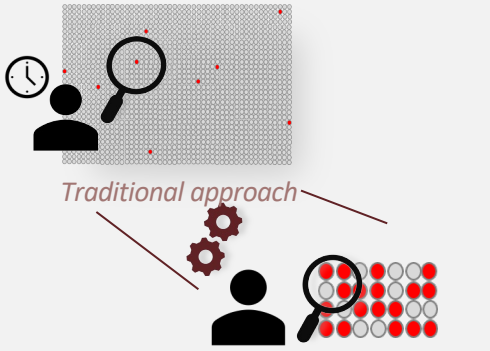
Query Miner

Looking at the free text queries, which aspects of the data are the most problematics, where can we focus effort?

phastar

Intelligent Approach to Data Cleaning

Quality data is critical for a successful clinical trial and data managers manually inspect the data to check for errors or inconsistencies. At PHASTAR we are looking to drive efficiencies in this process through the use of different data science approaches.



Auto-Query detection

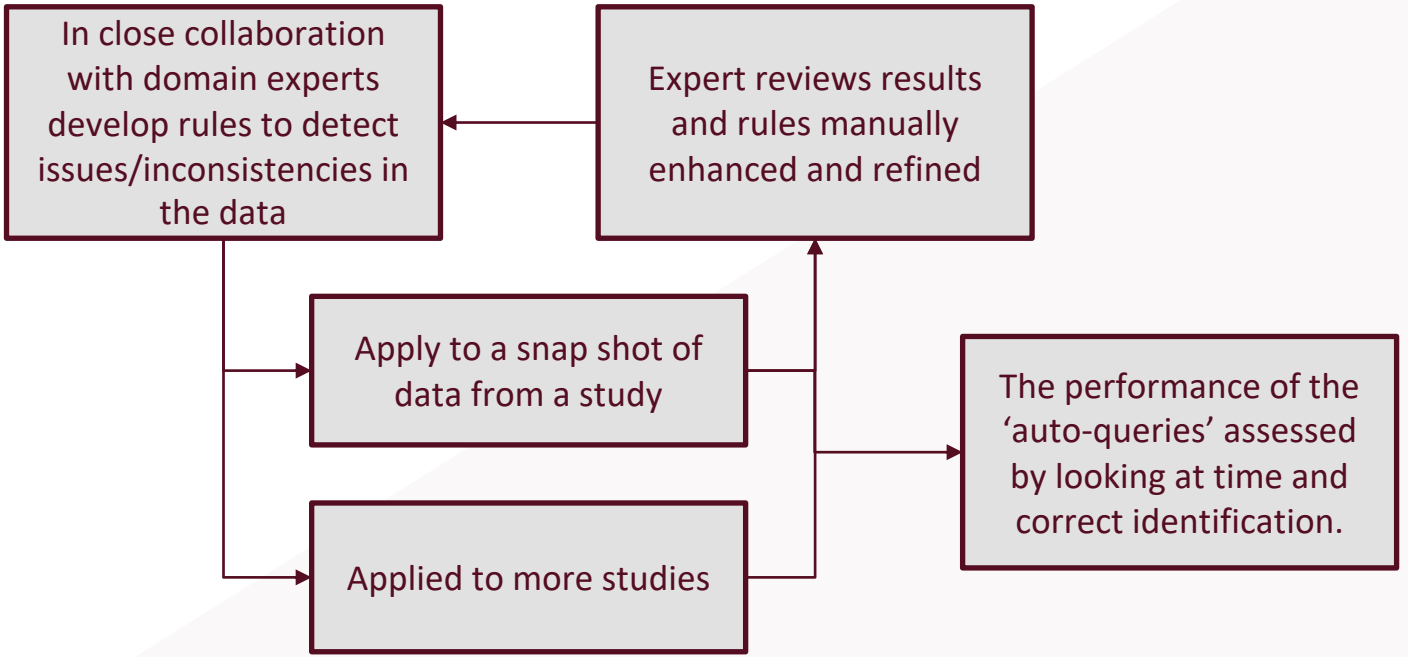
Apply rule-based approach to highlight potential data issues, supporting the data management team prioritise their activities.

Overall Aim: Identify potential data issues and inconsistencies.

- Use the previous Query analysis to focus efforts
- Develop a rule-based approach to detect any issues.
- Apply to multiple studies and assess the benefit
- Provide a feedback mechanism to learn from domain experts

Auto-Query Detection

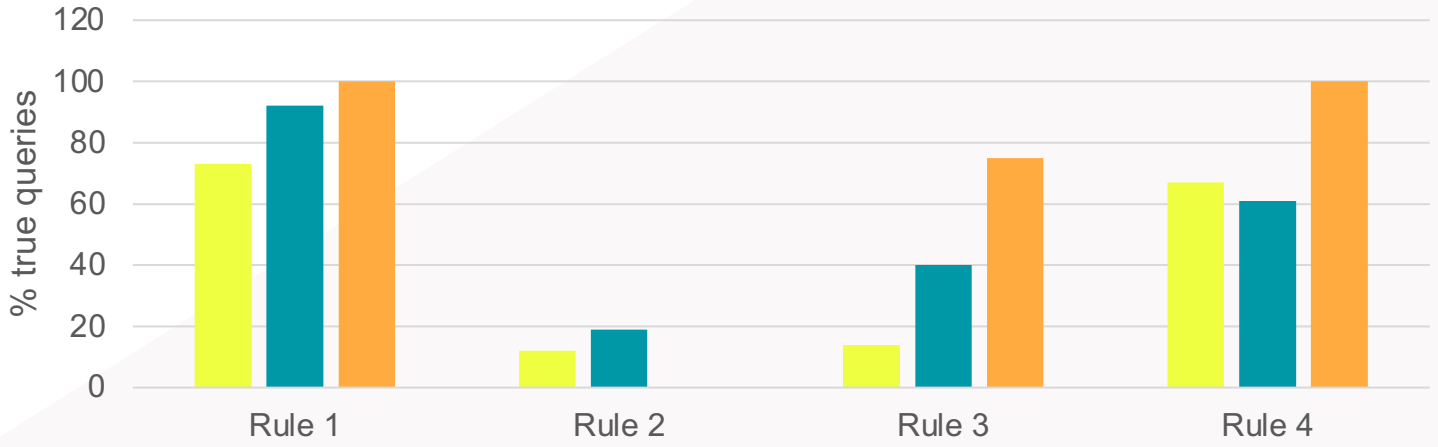
Method



Auto-Query Detection – Performance (True Queries)

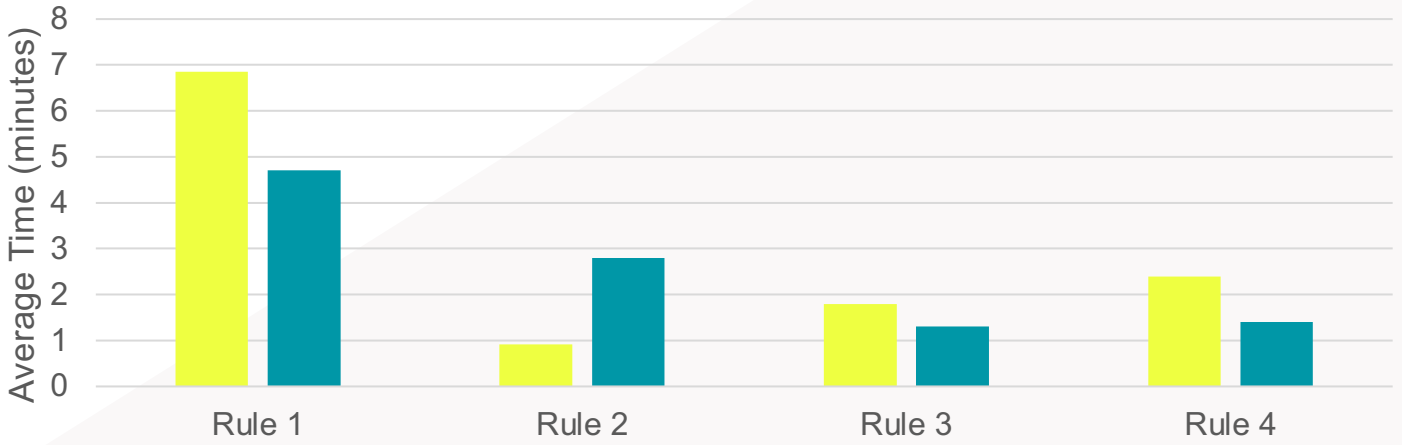
Example Adverse Events and Conmeds.

Of those auto-queries identified the following were true queries (defined as where a query was either raised or the data was subsequently changed)



Auto-Query Detection – Performance (Review Time)

The average time (minutes) to review the queries for the different rules.



Auto-Query Detection - Efficiency

The time taken to review each auto-query vs. a traditional approach

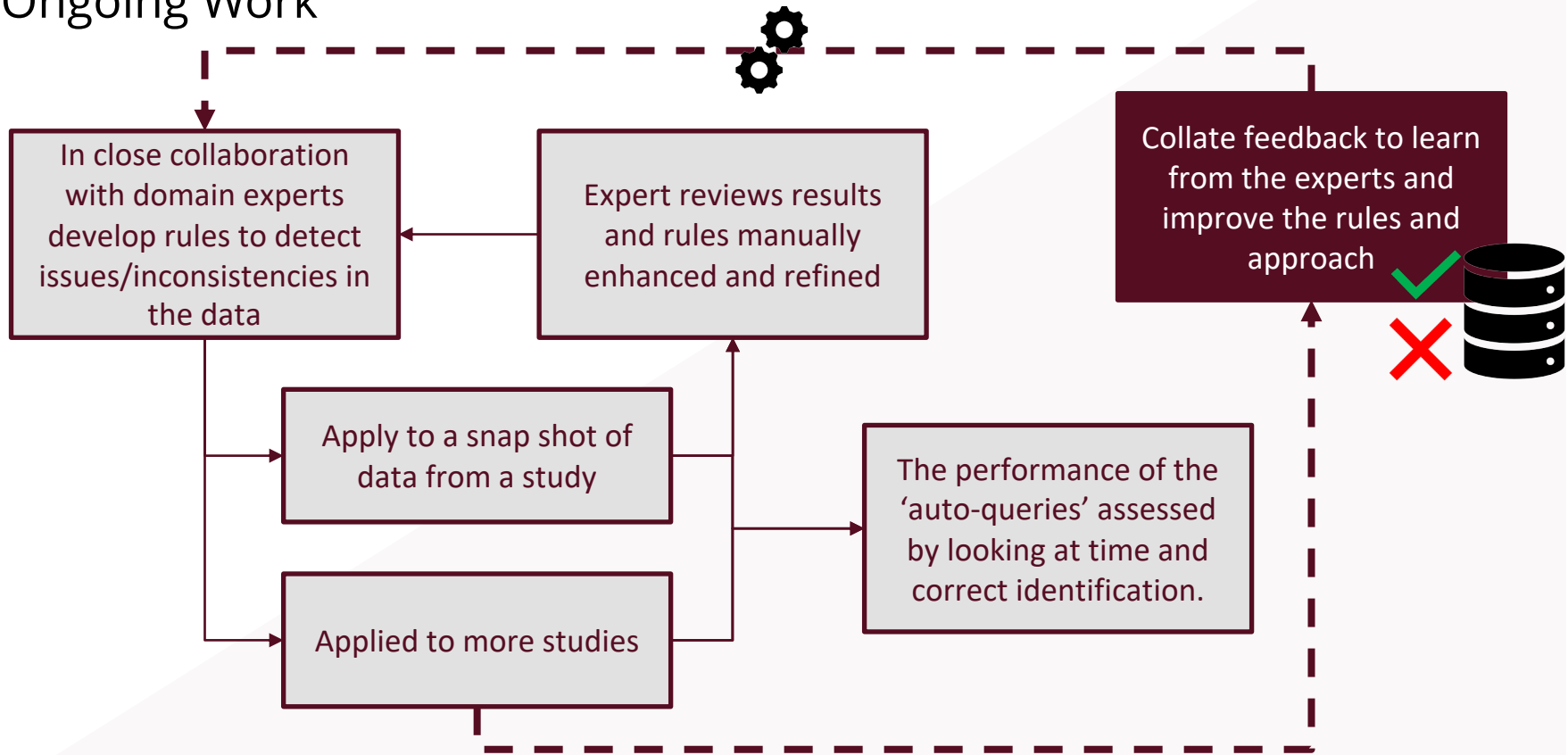
Study	Traditional Approach	
	Time to Review all AEs *	Time to review per query generated **
Study 1	3474 Minutes	60 minutes
Study 2	810 minutes	9.6 minutes
Study 3	474 minutes	8.8 minutes
Average over all studies		26.1 minutes

* Assume time to review a single AE is 3 minutes and assume all AEs are reviewed in a study
** Assume the review is only looking for the specific rules 1-4!
+ Timings were not recorded for this study so the average of study 1 and 2 were used.

phastar

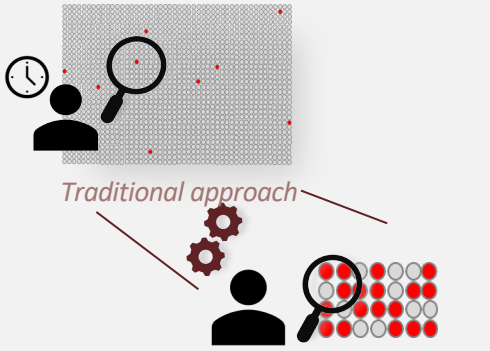
Auto-Query Detection

Ongoing Work



Intelligent Approach to Data Cleaning

Quality data is critical for a successful clinical trial and data managers manually inspect the data to check for errors or inconsistencies. At PHASTAR we are looking to drive efficiencies in this process through the use of different data science approaches.



Anomaly detection

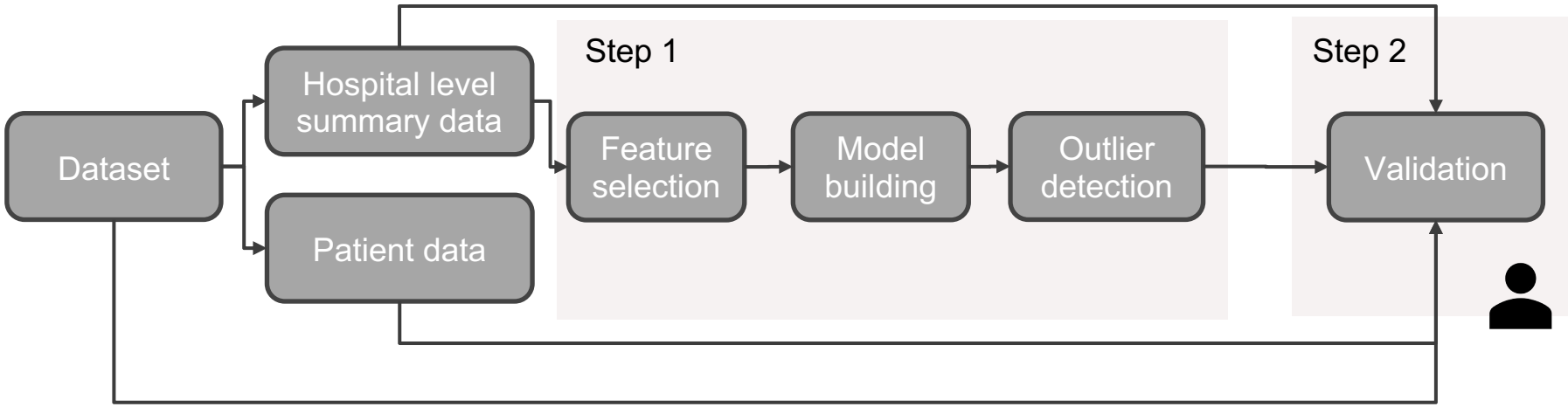
Application of machine learning to identify anomalies; initial focus on centralised statistical monitoring & high-risk site identification.

Overall Aim: to identify unusual sites or anomalous sites on a clinical trial using machine learning

- Exploring the extraction of features from the clinical data together with an unsupervised learning approach to detect anomalous sites.

An Approach to Fraud Detection

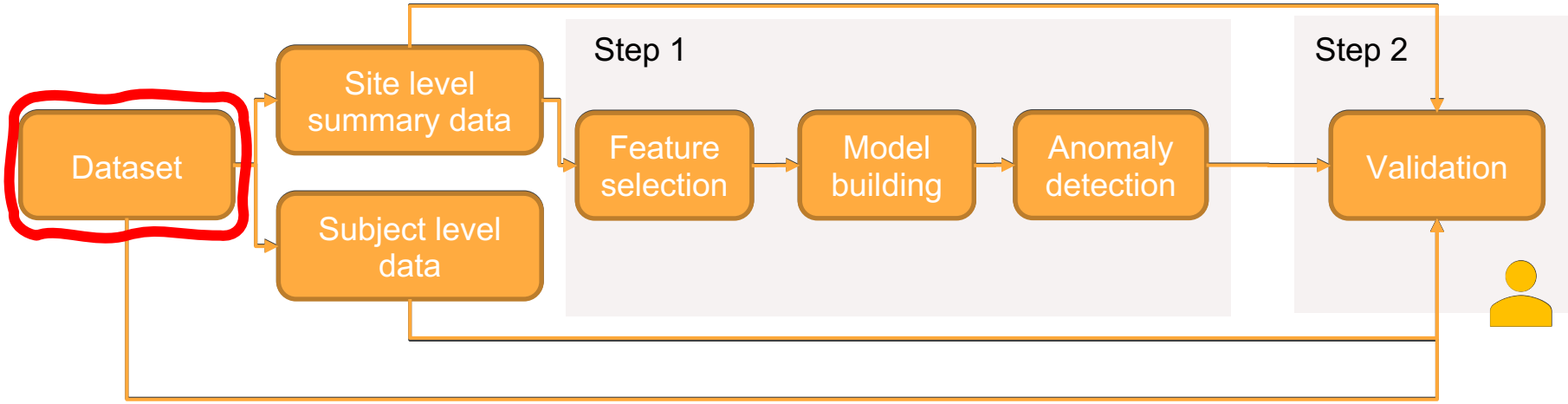
- Massi et al. described a data mining approach to detect fraud amongst hospitals



Massi, M.C., Ieva, F. and Lettieri, E. 2020 BMC Medical Informatics and Decision Making. 20 (160) 2754

An Approach to ~~Fraud~~ Anomaly Detection

The aim of this project is to explore methods to improve the detection of site level anomalies in multi-site clinical trials using machine learning to support the centralized monitoring approach.



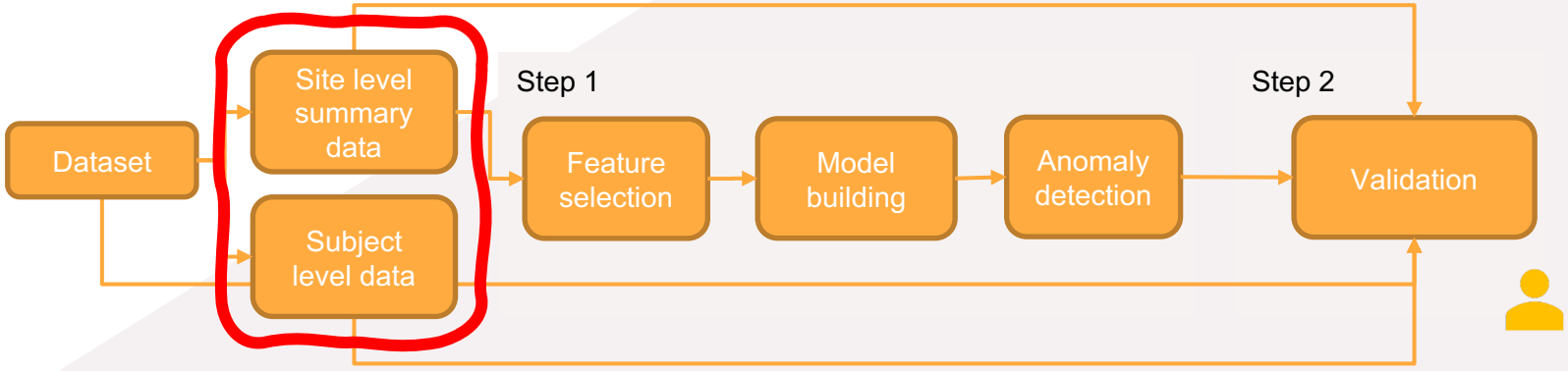
phastar

Dataset

- To explore different methods, a dataset was simulated
- The simulated dataset had the following features:
  - Multi-centre (150 sites)
  - Repeated measures
  - Skewed-normal
  - Correlated
  - Missing data (at random for some measures)
  - Different number of subjects for sites
  - Anomalies introduced (outliers and inliers)

Summarising Data

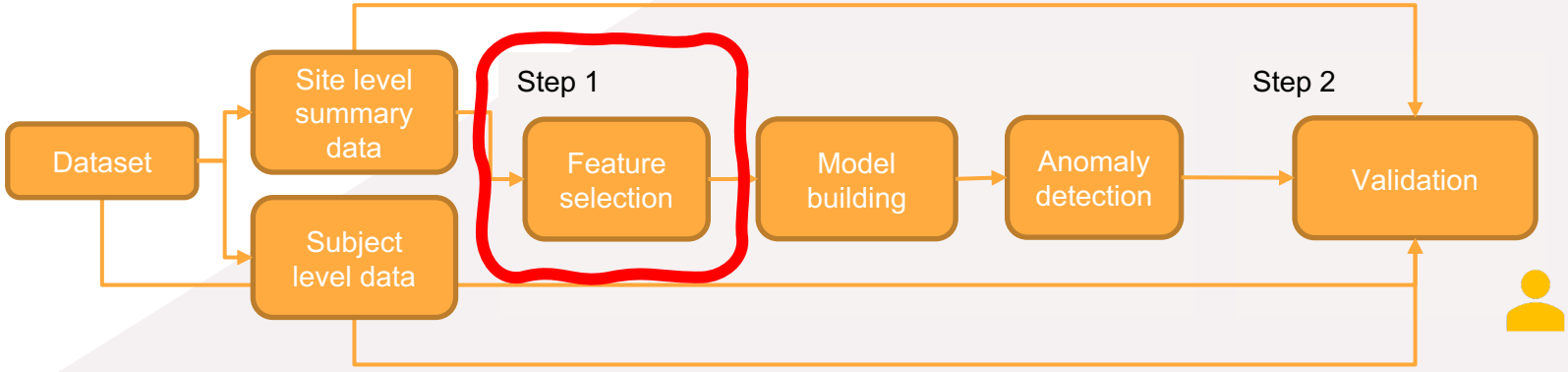
- Data was then averaged across all subjects within each site calculating a mean and standard deviation.
- Data was also summarised at the site level.



phastar

Prepare The Data for Modelling

- Feature extraction and feature engineering
- The project explored different feature engineering approaches on the different datasets i.e. the generation of meaningful features from the data.
- We explored the impact of different features on the models.



Feature Engineering - example

When scanning data for possible errors, it is usual to look for outliers. However, when looking for invented data the reverse is often more appropriate.

e.g. while it would not be surprising to find a patient with average weight, it would be unusual for all their measurements to be average

For example an inlier

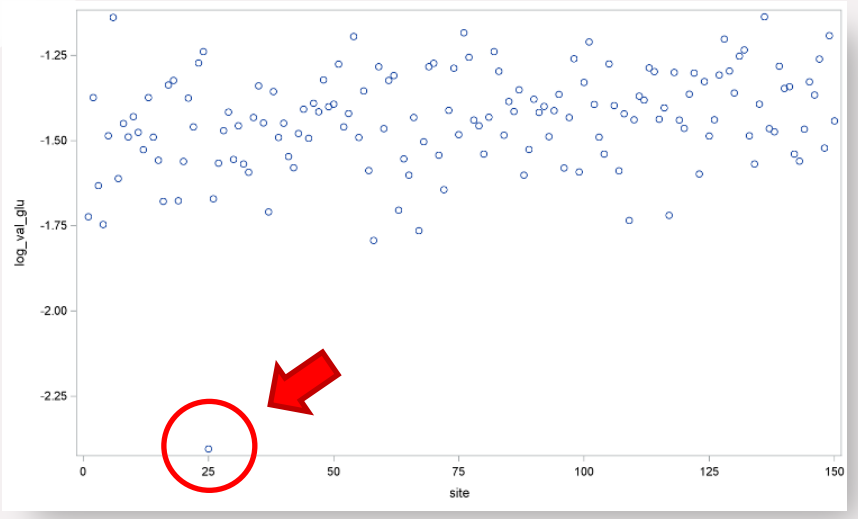
*Evans, S. Statistical aspects of the detection of fraud. Fraud and Misconduct in Medical Research. 2nd edition (1996): 226-239

Feature Engineering - example

When scanning data for possible errors, it is usual to look for outliers. However, when looking for invented data the reverse is often more appropriate.

e.g. while it would not be surprising to find a patient with average weight, it would be unusual for all their measurements to be average

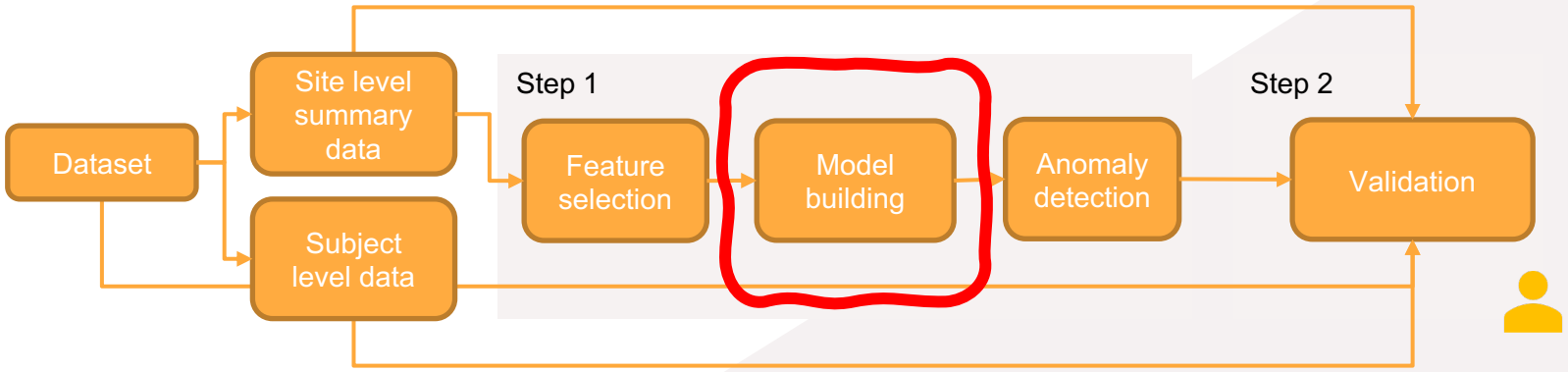
Evans* proposed a patient-based distance measure using the sum of squared z-scores of each observation from its mean, that is, compare each variable for each subject to the average/variability for that site.



*Evans, S. Statistical aspects of the detection of fraud. Fraud and Misconduct in Medical Research. 2nd edition (1996): 226-239

phastar

Model building



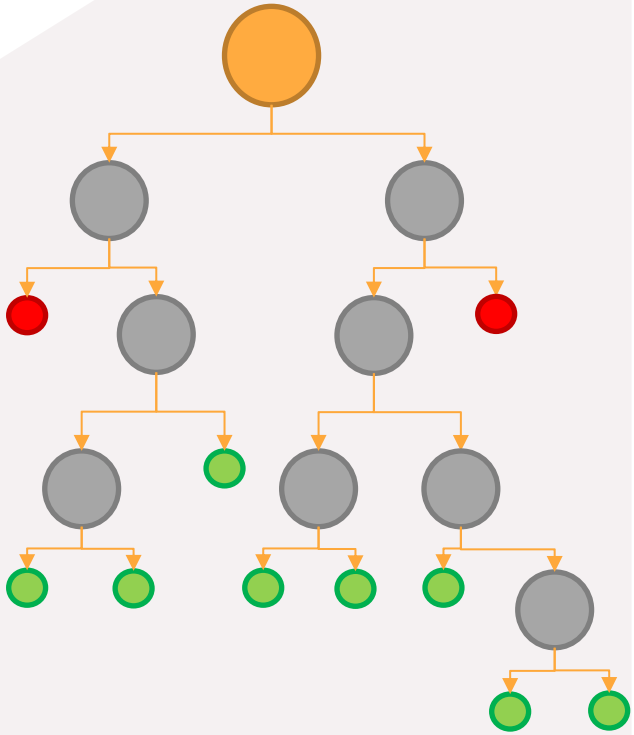
- Unsupervised Machine Learning approaches were used as in a real situation we would not know the answer i.e. which sites are potential anomalies
  - Isolation Forest
  - DBScan

Isolation Forest

- Developed by Liu *et al.* in 2008 algorithm and uses a tree-based method to detect anomalies in an unsupervised approach.
- “Outliers are few and different”

- It considers how far a point is from the rest of the data rather than modelling all of the data

- 1 parameter is required: contamination factor

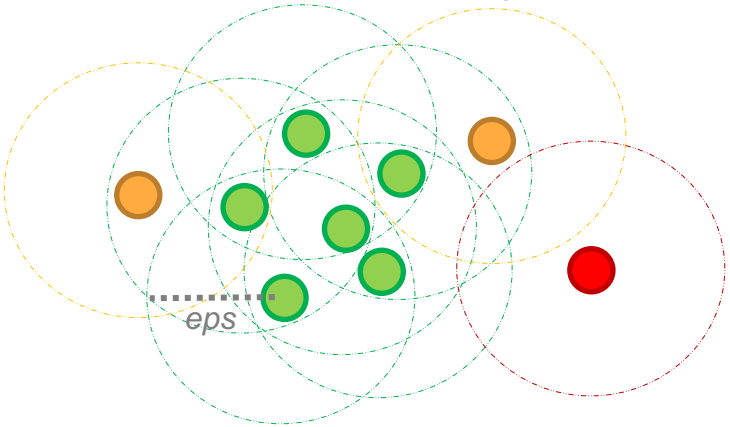


phastar

DBScan

- Density-based spatial clustering of applications with noise (DBScan) developed by Ester *et al.* in 1996
- A clustering approach aiming to separate datapoints into groups, so points within the same group have similar properties - a point belongs to a cluster if it is close to many points within that cluster.
- 2 parameters are required; *minpts* (the minimum number of points required to define a cluster as dense) and *eps* (distance measure)

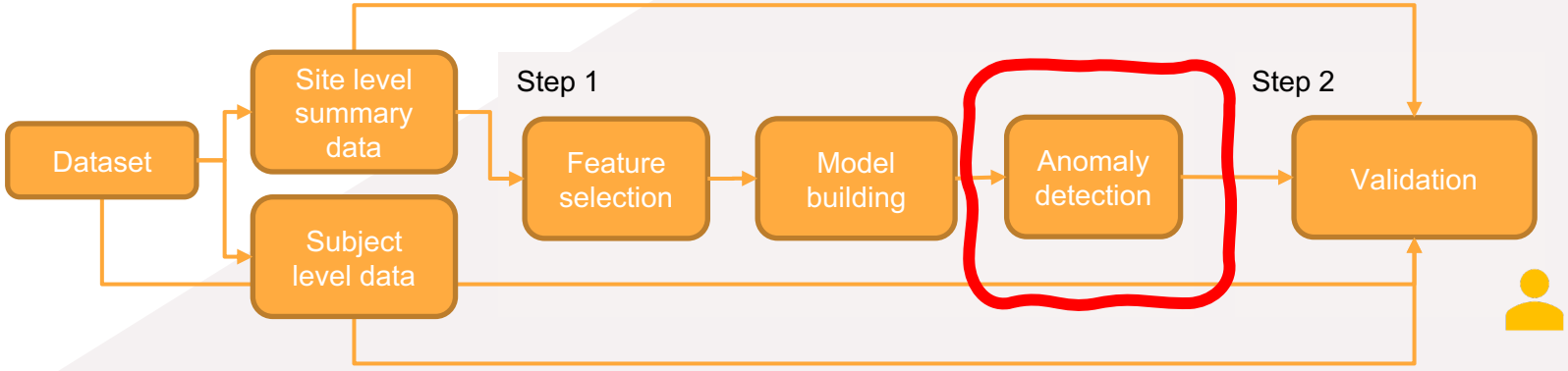
Core points
Border Points
Outlier



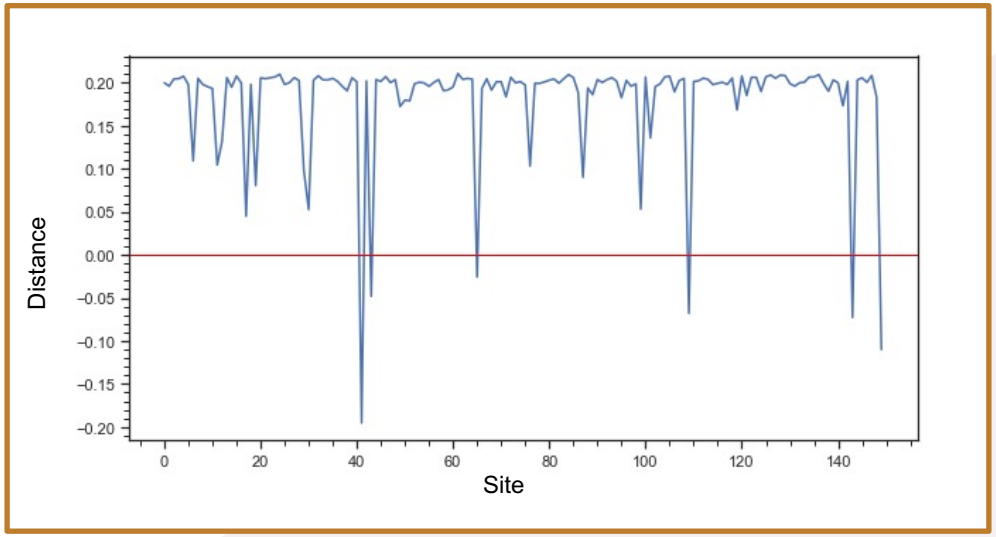
phastar

Outlier Detection

- We applied both Isolation Forest and DBScan to identify potential outliers.
- We also looked to see if we could identify those features that contributed towards that prediction.



Identification of Anomalies

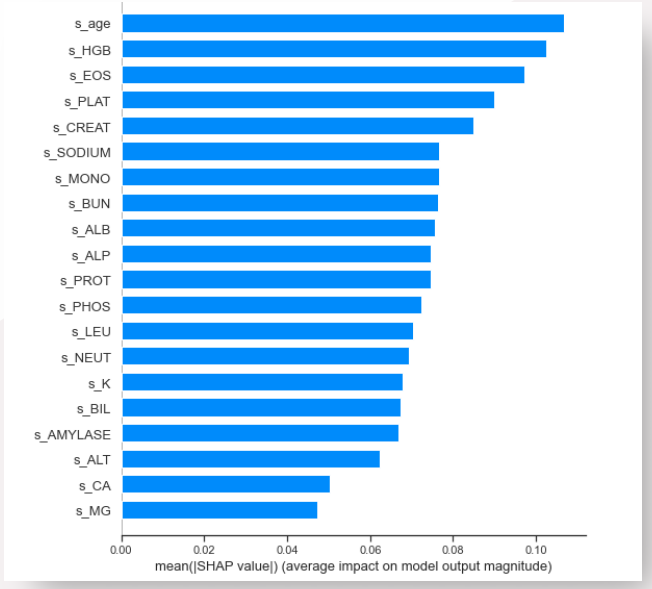


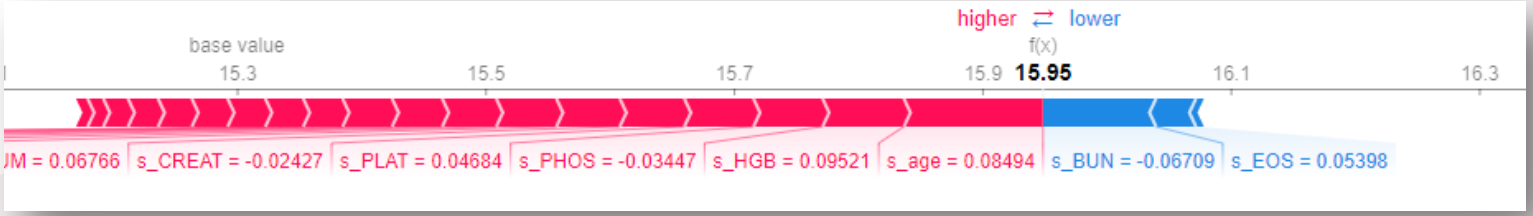
Isolation forest

- We can identify the anomalous sites
- Sites below the red line, this changes based on the parameters used (contamination)

Understanding Anomalies

- What features have led to the prediction
- Can use SHapley Additive exPlanations (SHAP)
- SHAP quantifies the contribution of each feature to the model's prediction.

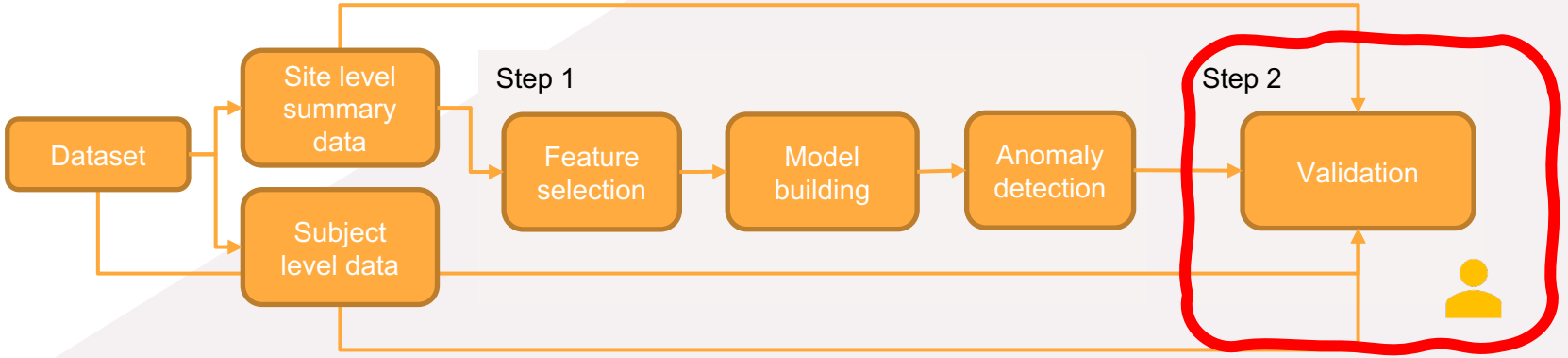




Validation

How do we know if the predicted anomalies are correct?

- In this situation we know which sites had anomalies
- With real data we wouldn't, here experts would be presented with the anomalous sites and potentially predictive features and the human experts could work towards explanation/investigation of the results.



Validation/Model Performance

- Overall, the performance was comparable across both DBScan and Isolation Forest.
- Isolation forest performance is impacted by the contamination factor.
- It is key that any anomalies identified are explainable, that is we understand why they were identified.

Model - Isolation forest
Features – site summary data with simple means for the features

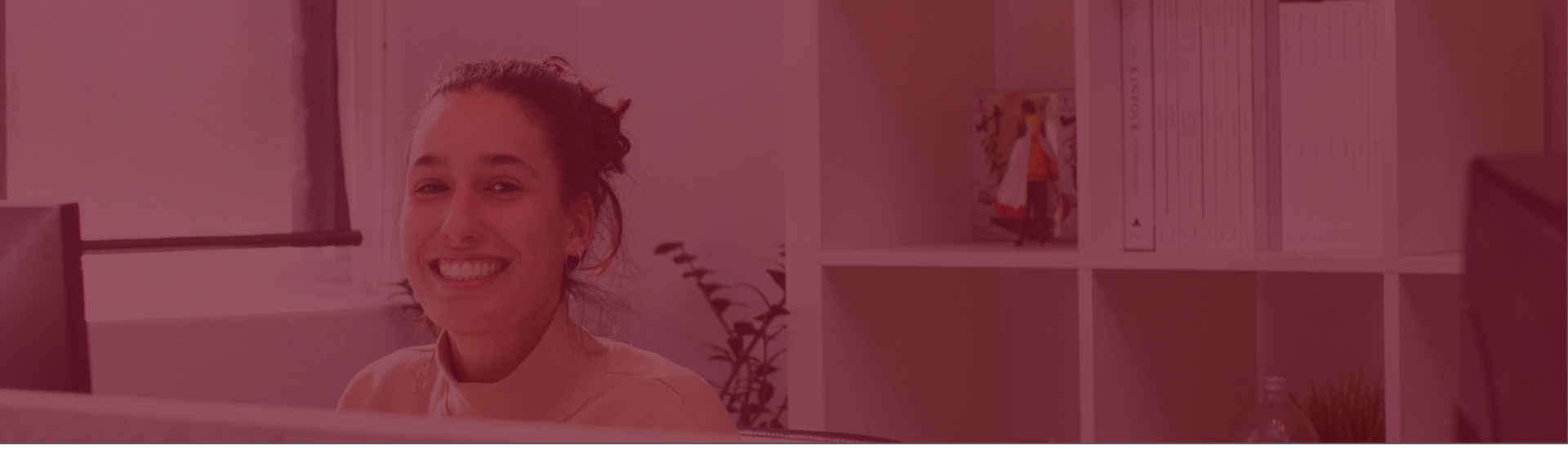
	Normal	Anomaly
Normal	125	3
Anomaly	2	20

phastar

Summary

- We have seen success with Data Science and Data Management experts working closely together.
- Provided an overview of two projects applying data science approaches within data management.
- We see efficiency gains when supporting data cleaning activities.
- Early results looking at anomaly detection are encouraging, and we plan to explore this in a real-life scenario.

phastar



Thank you for listening.

tellmemore@phastar.com

- Many thanks to the DM team at PHASTAR for support and discussion.
- Also, thanks to Charles Boachie and Jordan Bristow at PHASTAR for their work on anomaly detection.