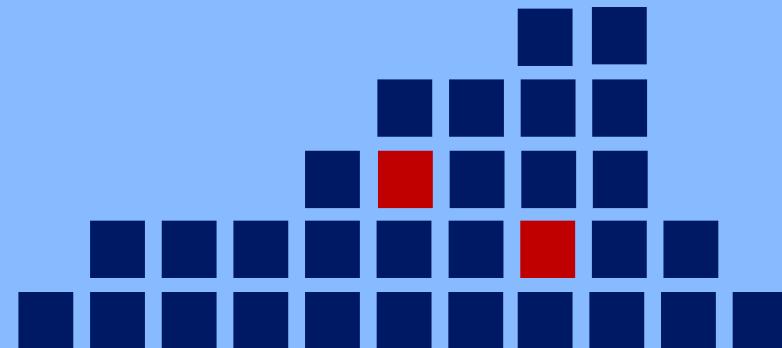


Beyond Outliers: Unsupervised Anomaly Detection in Clinical Trial Data





Kian Norouzi

DM Process and Innovation Specialist
Novo Nordisk



Helena Yaben

DM Process and Innovation Professional
Novo Nordisk

Views and opinions expressed are those of the speaker
and not Novo Nordisk

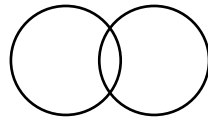
A proposed unsupervised ML pipeline to detect anomalies and provide a user-friendly output



Why anomaly detection?



Why machine learning?

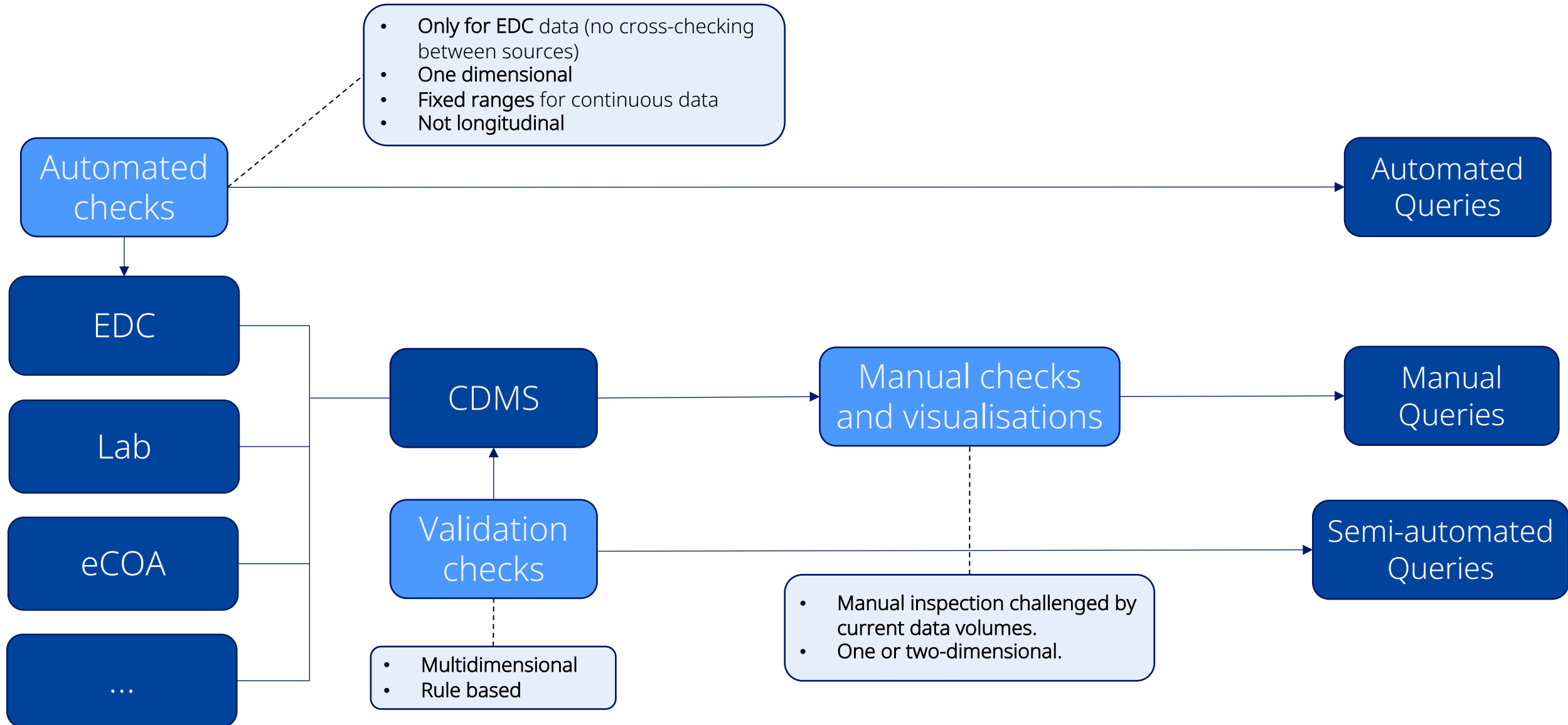


Why unsupervised?



Why user-friendly outputs?

Current Process

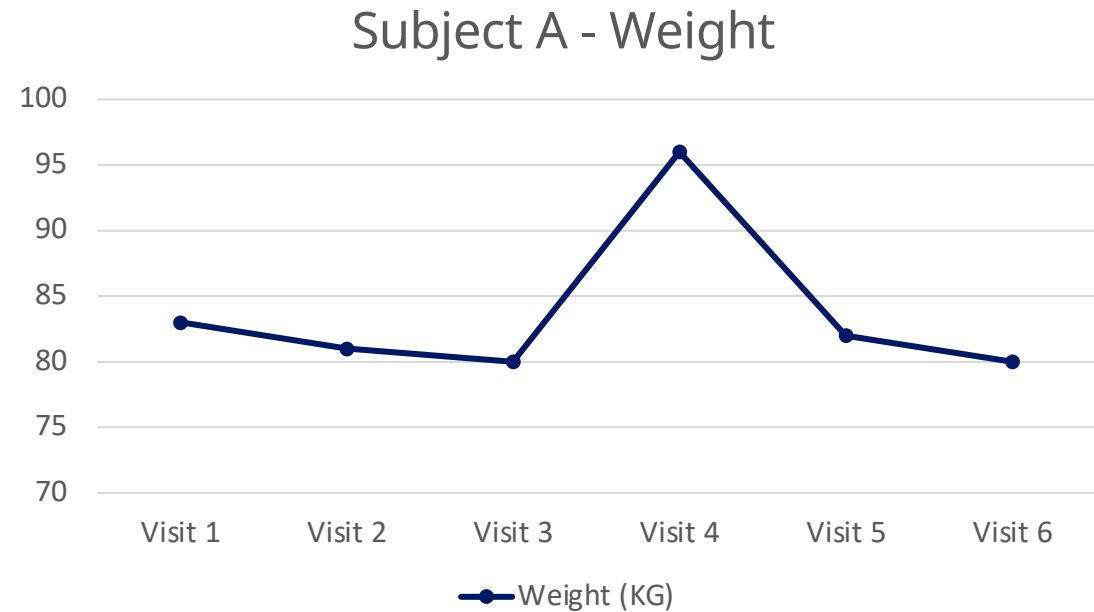


Edit Checks Limitation

- Only useful for detecting extreme values

Weight must be $\geq$ [consWeightRangeLower]
AND
Weight must be $\leq$ [consWeightRangeUpper]

Weight must be $\geq$ 40KG
AND
Weight must be $\leq$ 200KG



Data Visualisation Limitations



Data Understanding



1D, 2D, and 3D Anomalies



Temporal Anomalies



Manual process



Subjective



Time-consuming

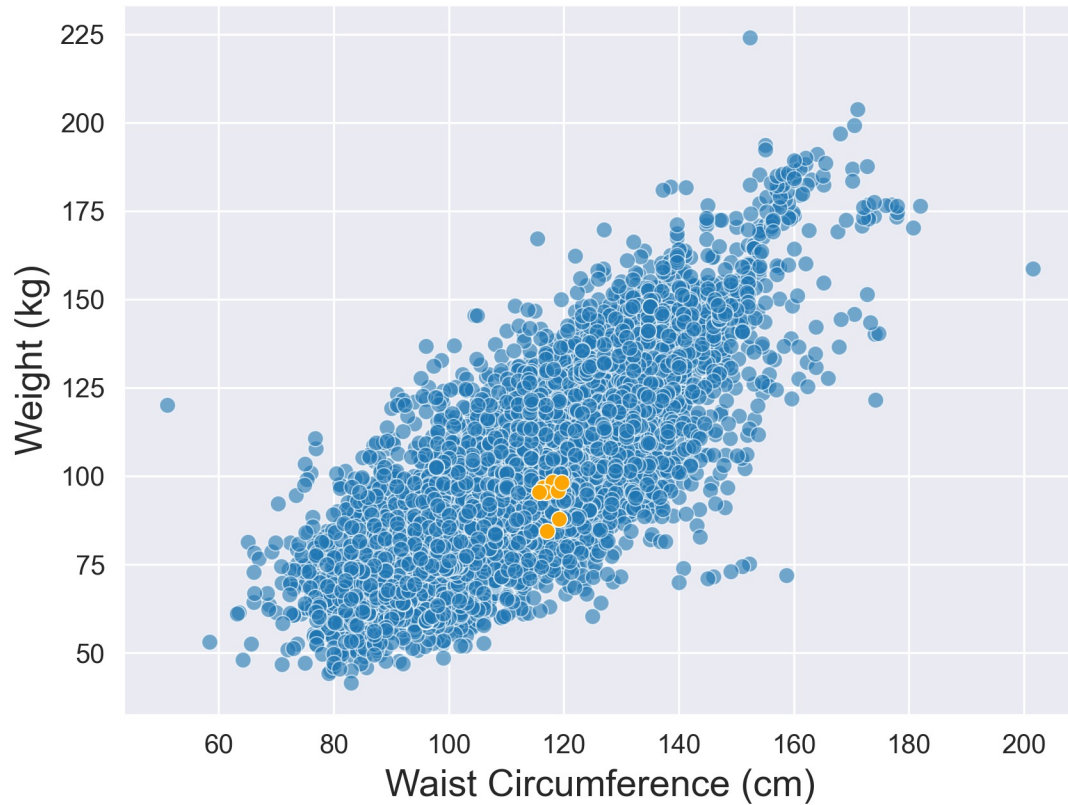
Data Visualisation: Example



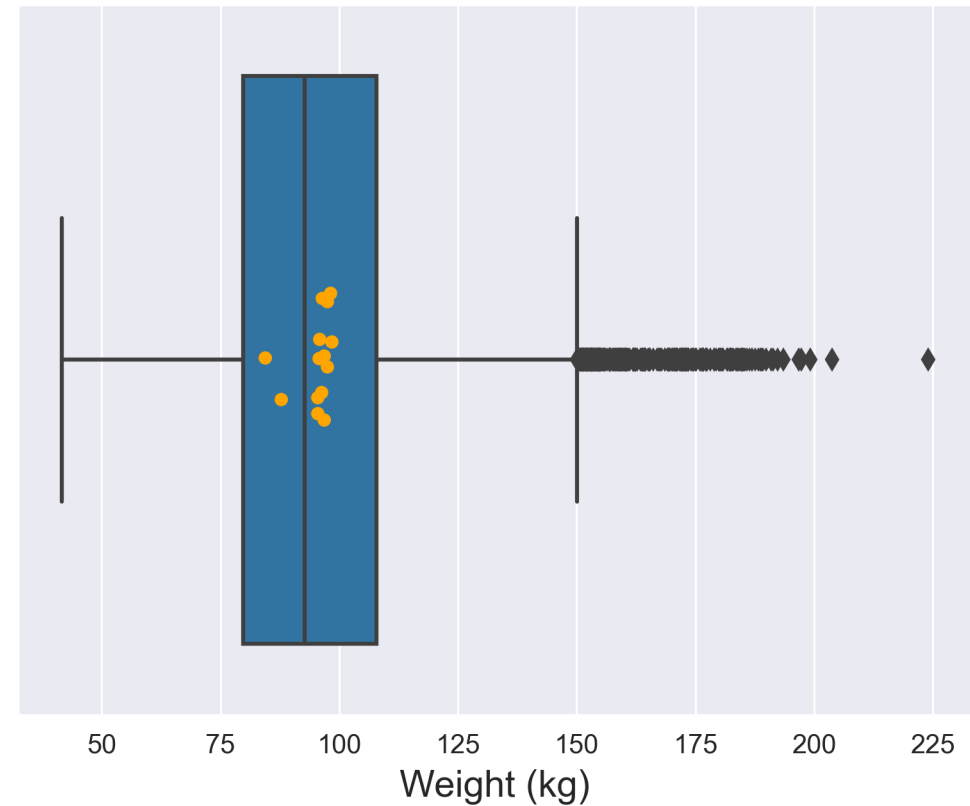
Subject A seems to have two discrepant observations for weight at visits 3 and 9

Data Visualisation: Example

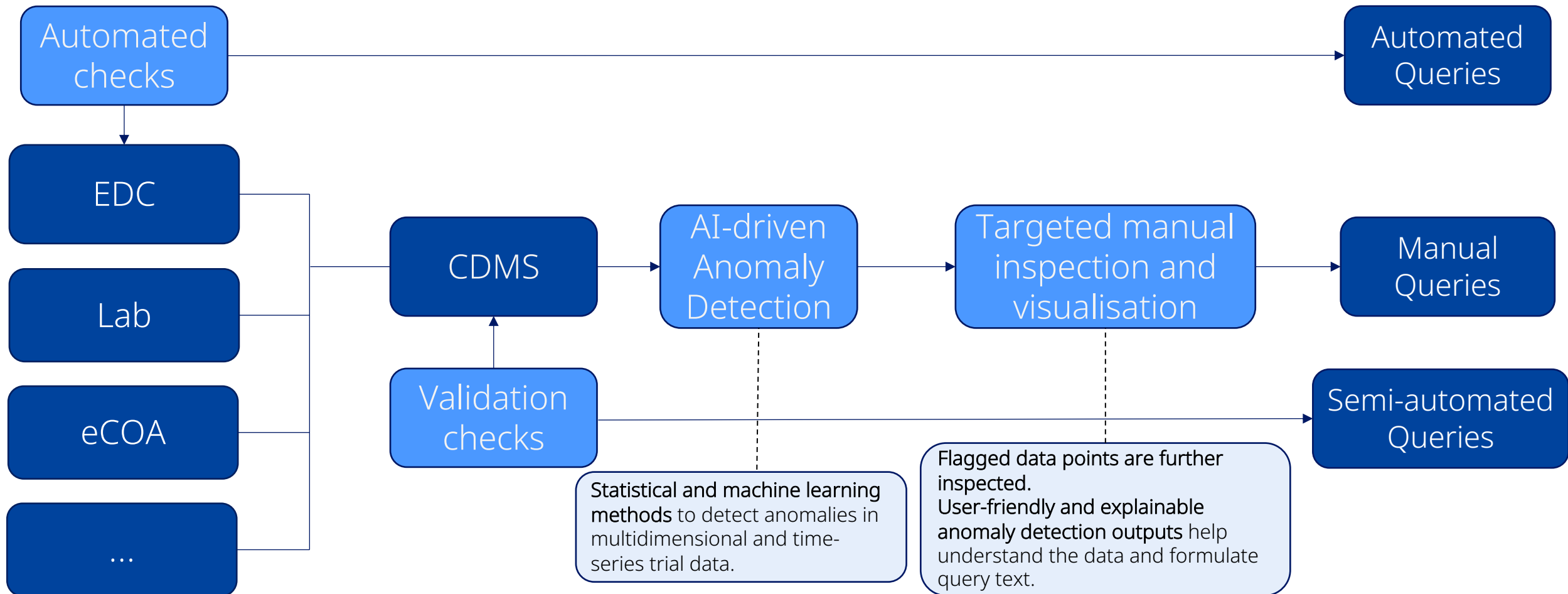
Waist circumference vs weight for all participants in the trial.
Data for subject A in orange



Box plot of weight for all participants in the trial.
Data for subject A in orange



Proposed Process



Data

Standardised datasets



Vital signs



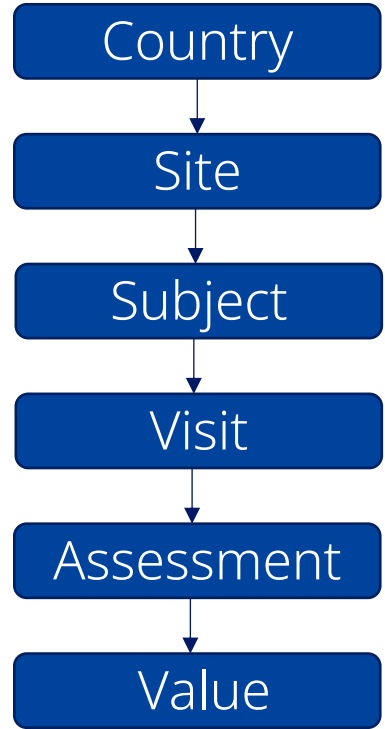
Laboratory

Numerical, continuous, and longitudinal
data

Granularity of Anomaly Detection

Reconstruction based anomaly detection methods can provide anomaly scores on the granularity needed for the data review process

Central Monitoring

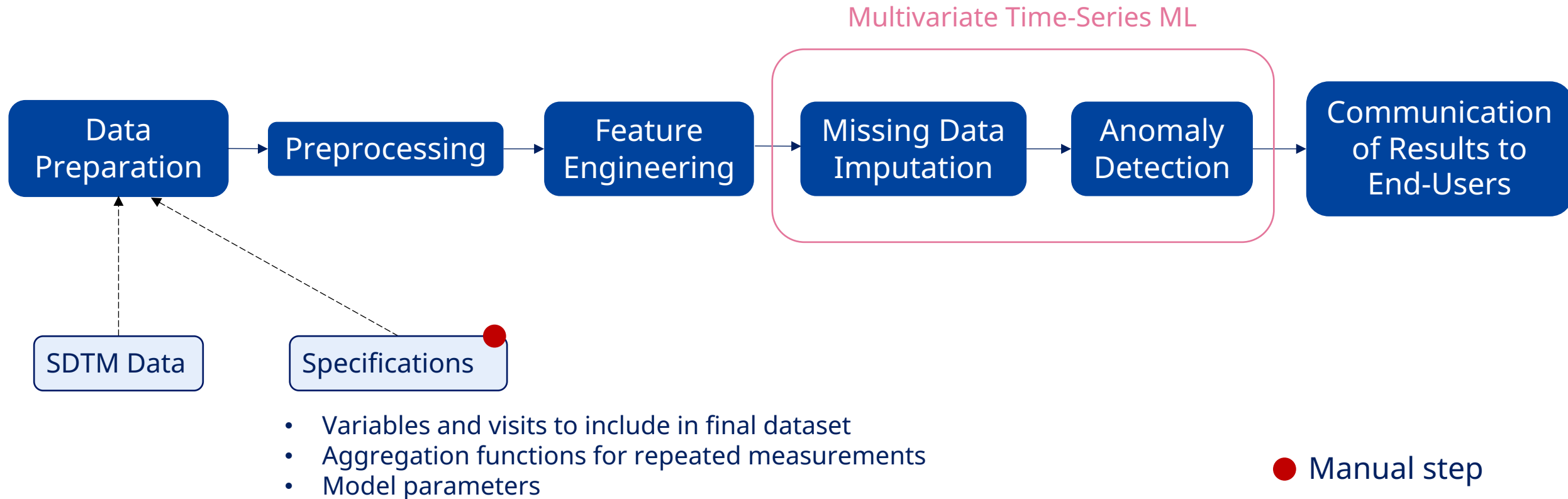


Data Cleaning

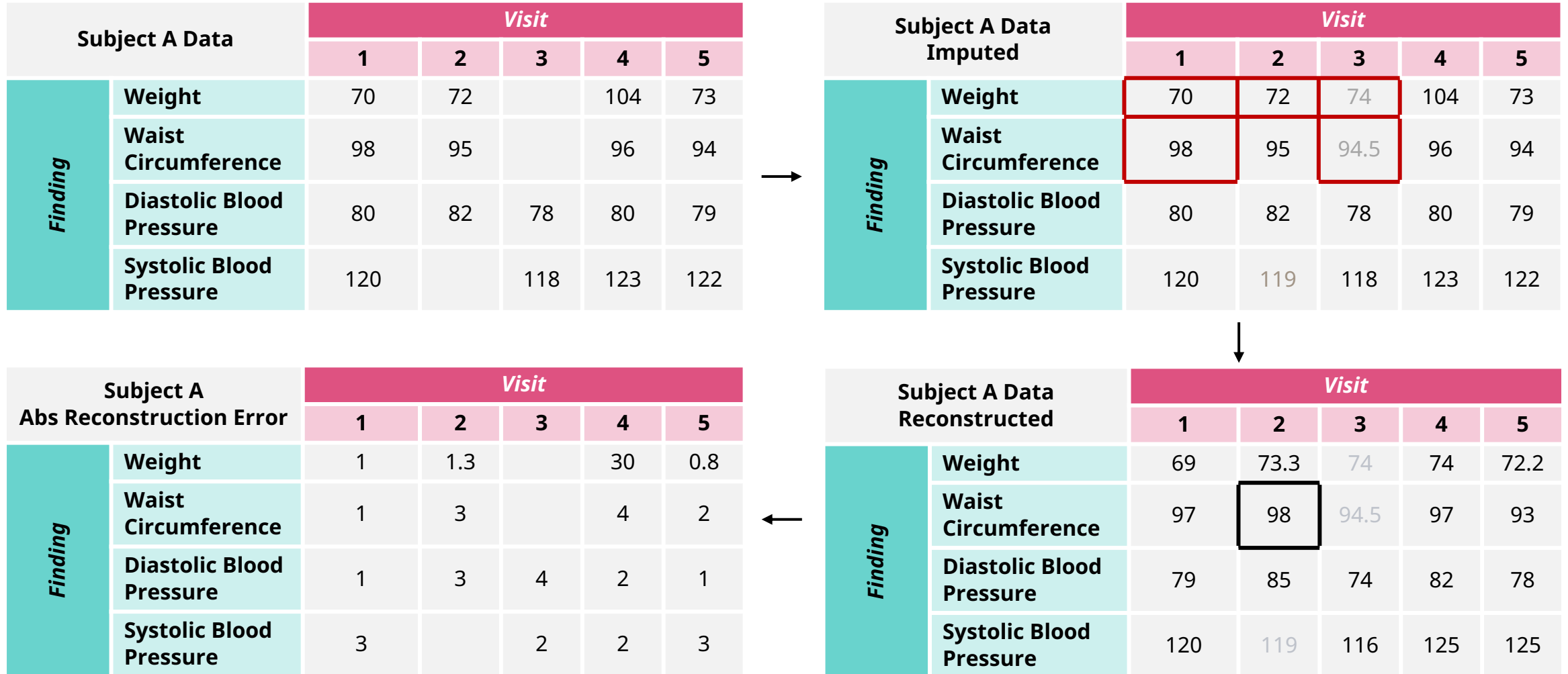
Country	Site	Subject	Visit	Assessment	Value	Anomaly score
USA	1	1	3	WEIGHT	82	0.5

Anomaly Detection Pipeline

 Developed in Python

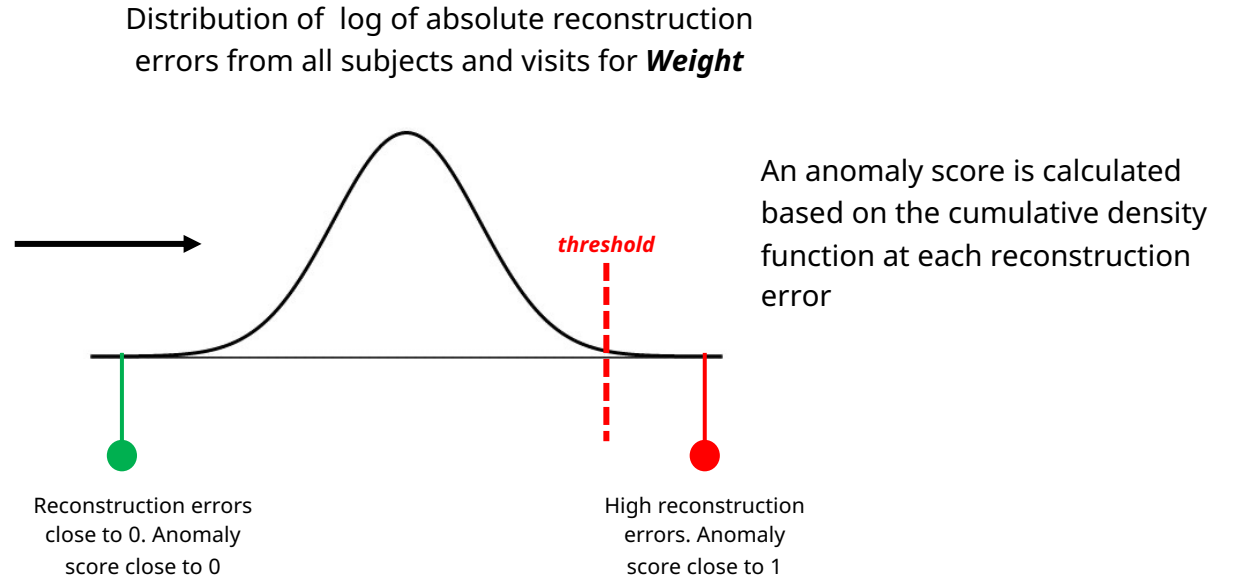


Anomaly Detection Pipeline - Example



Example

Subject A		Visit				
Abs Reconstruction Error		1	2	3	4	5
Finding	Weight	1	1.3		30	0.8
	Waist Circumference	1	3		4	2
	Diastolic Blood Pressure	1	3	4	2	1
	Systolic Blood Pressure	3		2	2	3



Subject A		Visit				
Anomaly Scores		1	2	3	4	5
Finding	Weight	0.4	0.5		0.99	0.3
	Waist Circumference	0.2	0.7		0.75	0.3
	Diastolic Blood Pressure	0.3	0.5	0.6	0.3	0.25
	Systolic Blood Pressure	0.6		0.3	0.3	0.4

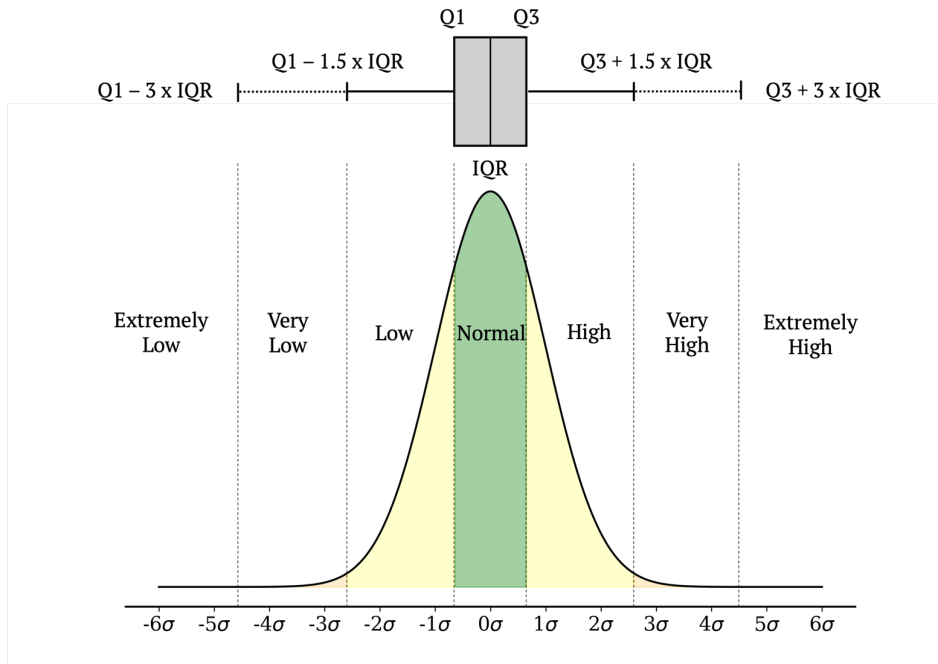


Subject A		Visit				
Anomaly Labels		1	2	3	4	5
Finding	Weight	0	0		1	0
	Waist Circumference	0	0		0	0
	Diastolic Blood Pressure	0	0	0	0	0
	Systolic Blood Pressure	0		0	0	0

Additional tests

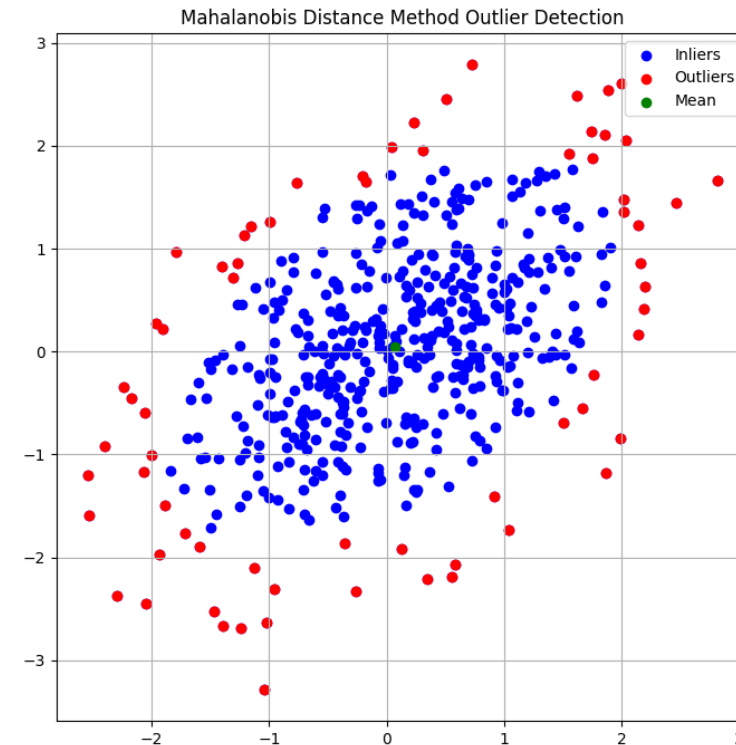
IQR:

- Univariate label
- Temporal univariate label



Mahalanobis:

- Target variable and most correlated one



Explainability

Subject A Data		Visit				
		1	2	3	4	5
Finding	Weight	70	72		104	73
	Waist Circumference	98	95		96	94
	Diastolic Blood Pressure	80	82	78	80	79
	Systolic Blood Pressure	120		118	123	122

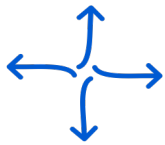
From reconstruction

- **Predicted value:** 74
- **Anomaly score:** 0.99
- **Anomaly label:** 1
- **Feature importance in prediction:**
 - Weight from previous visit
 - Weight from next visit
 - Waist circumference at current visit
 - Waist circumference at previous visit
 - Waist circumference at next visit

From univariate and bivariate methods

- **Classification based on univariate test on value:** High
- **Classification based on univariate test on absolute change since previous or next visit / days since previous or next visit:** Extremely high
- **Classification based on bivariate test for combination with most correlated feature at current visit (when both collected):** Odd ($p < 0.01$)

Conclusion



Multivariate



Time series



Missing values



Standardised



Automated



Explainable

This will serve as a recommendation pipeline, and we still need the human in the loop!

Thank You!

Kian Norouzi
vxmn@novonordisk.com
 /in/knorouzi

