

Keeping It Real

A Tiered Approach to Real-World Data Quality by
Leveraging Synthetic Data

Lisa Murch
William Kuan





AS YOU CAN SEE, NEW CASES OF THIS NEW DISEASE
HAVE FALLEN BELOW ZERO.
BUT I'M NOT 100% SURE ABOUT THOSE NUMBERS...

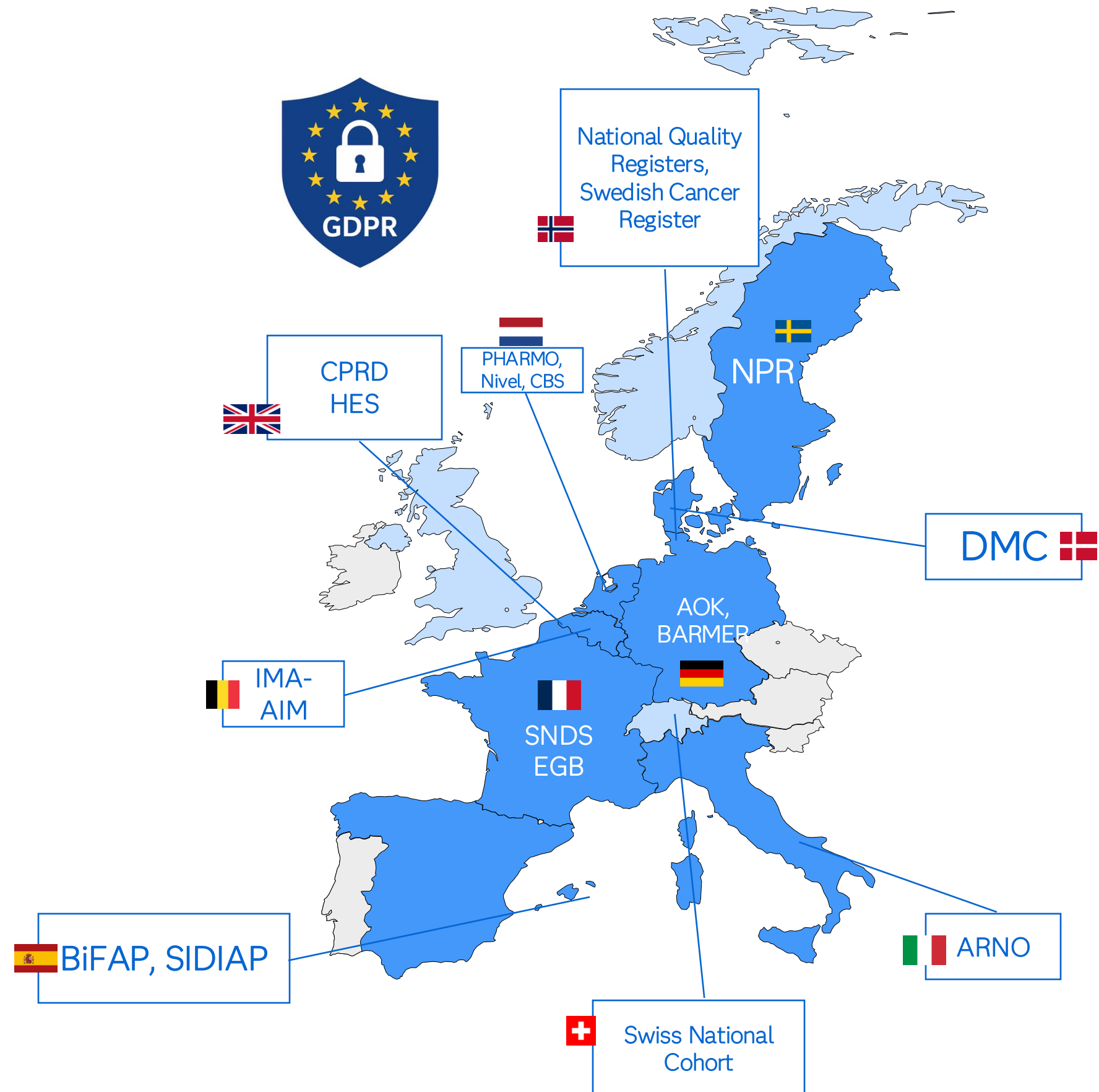
 Dataedo /cartoon

Pio M@Dataedo

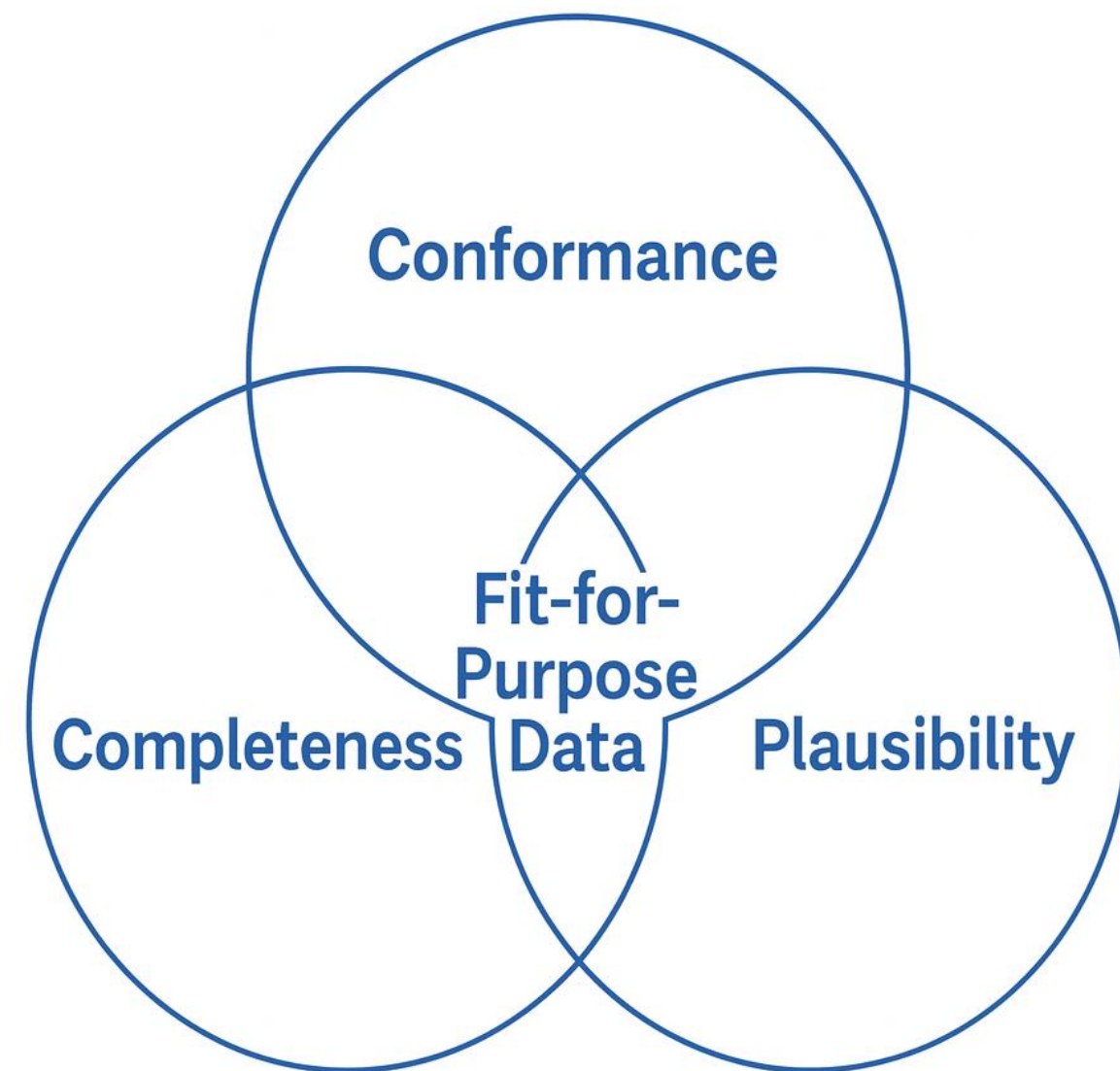
Europe RWD landscape is notoriously fragmented

Multiple systems across Europe, all built for local needs, with incompatible coding, structures & governance

DARWIN-EU and EHDEN were initiated to resolve these issues



What is Data Quality?



EMA bases its data quality framework on 5 pillars:

1. Extensiveness
2. Coherence
3. Timeliness
4. Relevance
5. Reliability

FDA emphasizes:

1. Completeness
2. Conformance
3. Plausibility
4. Integrity

We explore a 3-layered approach to data quality

1

Data Quality Layer

Profile and detect – make data quality issues visible

2

Analytics Layer

Quantify impact and look for problematic trends

3

Action Layer

Remediate – make data fit for purpose

Tiered approach to data quality

Data Quality Tier

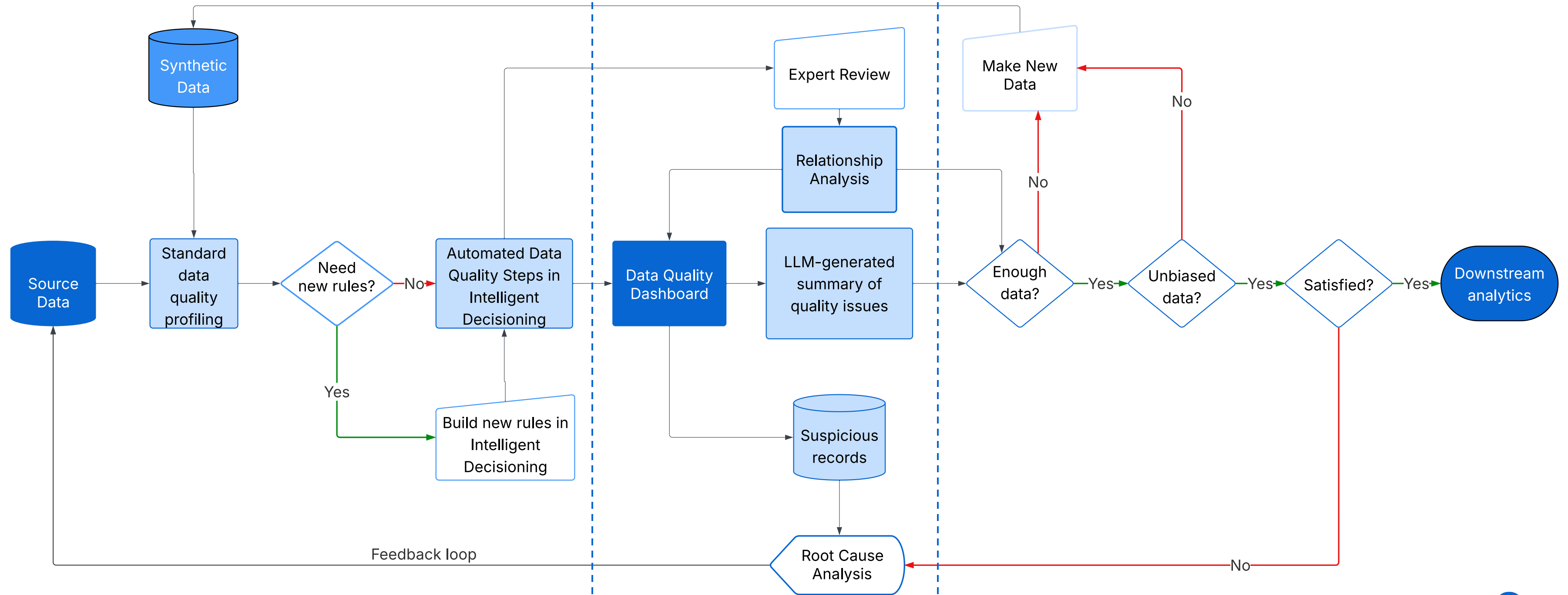
Automated profiling and custom data quality rules

Analytics Tier

Diagnose quality issues

Action Tier

Decide whether to regenerate, fix or approve for downstream use



The Tiered Data Quality Approach

1

Data Quality Layer

Profile and detect – make data quality issues visible

2

Analytics Layer

Quantify impact and look for problematic trends

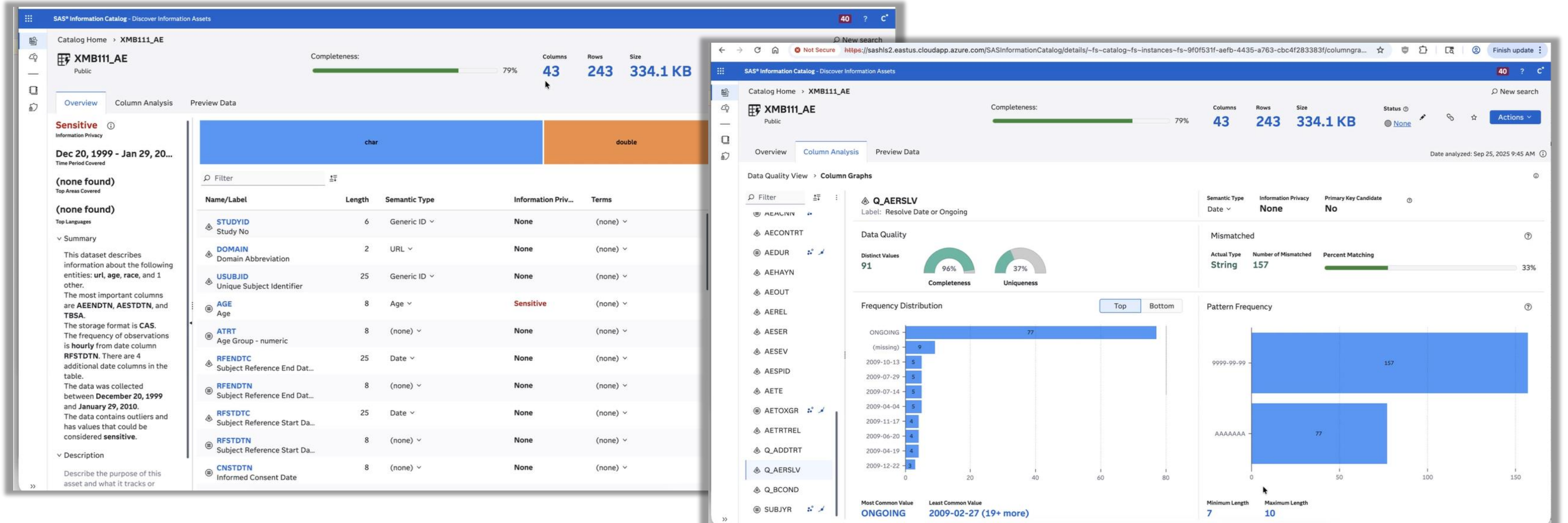
3

Action Layer

Remediate – make data fit for purpose

Data Quality Layer

Automatic Data Profiling



- Centralised data assets and metadata management
- Grab a view on data heterogeneity and completeness issues

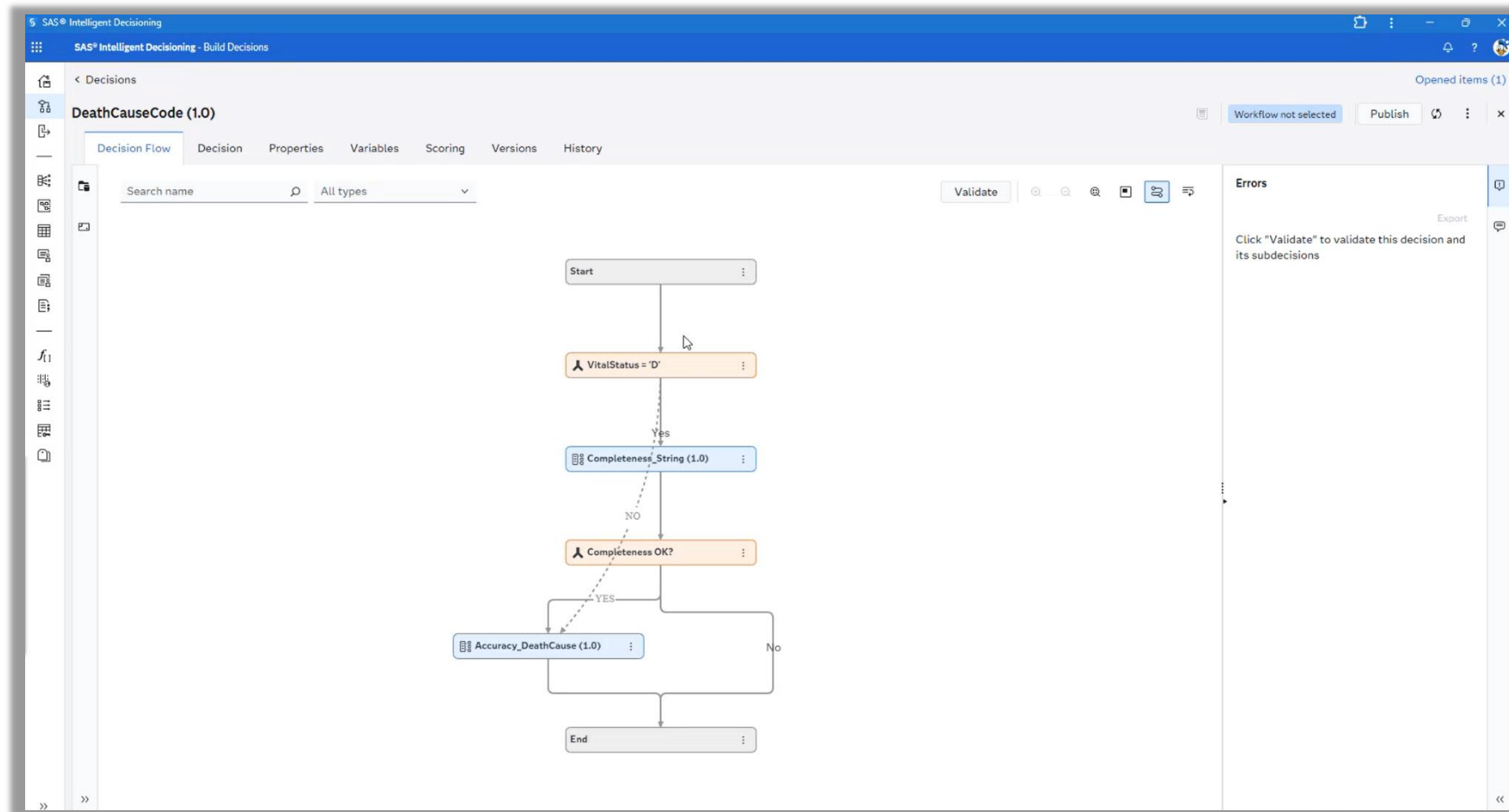
Data Quality Layer

Define Additional Data Quality Check Rules

Dimension	Description	Example
Reliability	-Data values and distributions agree with internal measurements or local knowledge. -Logical constraints between values agree with common knowledge. -Observed or derived values conform to expected temporal properties.	-Height and weight are positive values. -The patient's sex agrees with sex-specific contexts (e.g., pregnancy) -Discharge date/time happens after admission date/time
Extensiveness	-Missing required values. -Coverage of a population.	-<u>"Sex" variable</u> should not be missing
Coherence	-Unique (key) data values are not duplicated. -Data values conform to allowable values or ranges. - Data values conform to the representational constraints based on external standards. -The precision of values is fitting a target standard	-A patient ID number is only assigned to a single patient. -Sex only has values "M", "F" or "U". -Diagnosis code conforms to ICD-10 standards. -Two decimal digits are used in continuous variable.
Relevance	- The number of variables available in a given dataset matches the number of required variables	-All the necessary variables based on the research questions are available.

Data Quality Layer

Orchestrate Data Quality Decision Flow for Transparency and Reproducibility



- Orchestration layer for executing data quality rules at scale
- Combine with rule sets, lookup tables, AI/ML models, conditional logics etc.

Data Quality Layer

Build & Manage Data Quality Check Rules

The image displays three overlapping screenshots of the SAS Intelligent Decisioning interface, illustrating the process of building and managing data quality check rules.

Top Left Screenshot: Rule Set Editor
This window shows the 'Validity_Gender (1.0)' rule set. The 'Rule set' tab is active, displaying a logic statement: 'IF value_Gender NOT IN ('Male', 'Female') THEN ASSIGN isValid_Gender False'. The 'Properties' tab is also visible.

Top Middle Screenshot: Rule Sets List
This window shows a list of rule sets. The 'Rule Sets' tab is active, displaying a table with columns: Name, Latest Version, Checked Out, Description, and Location. The table lists several rule sets, including 'Validity_VitalStatus', 'Validity_Gender', 'Accuracy_DeathCause', 'Consistency_VitalStatus', and 'Completeness_String'.

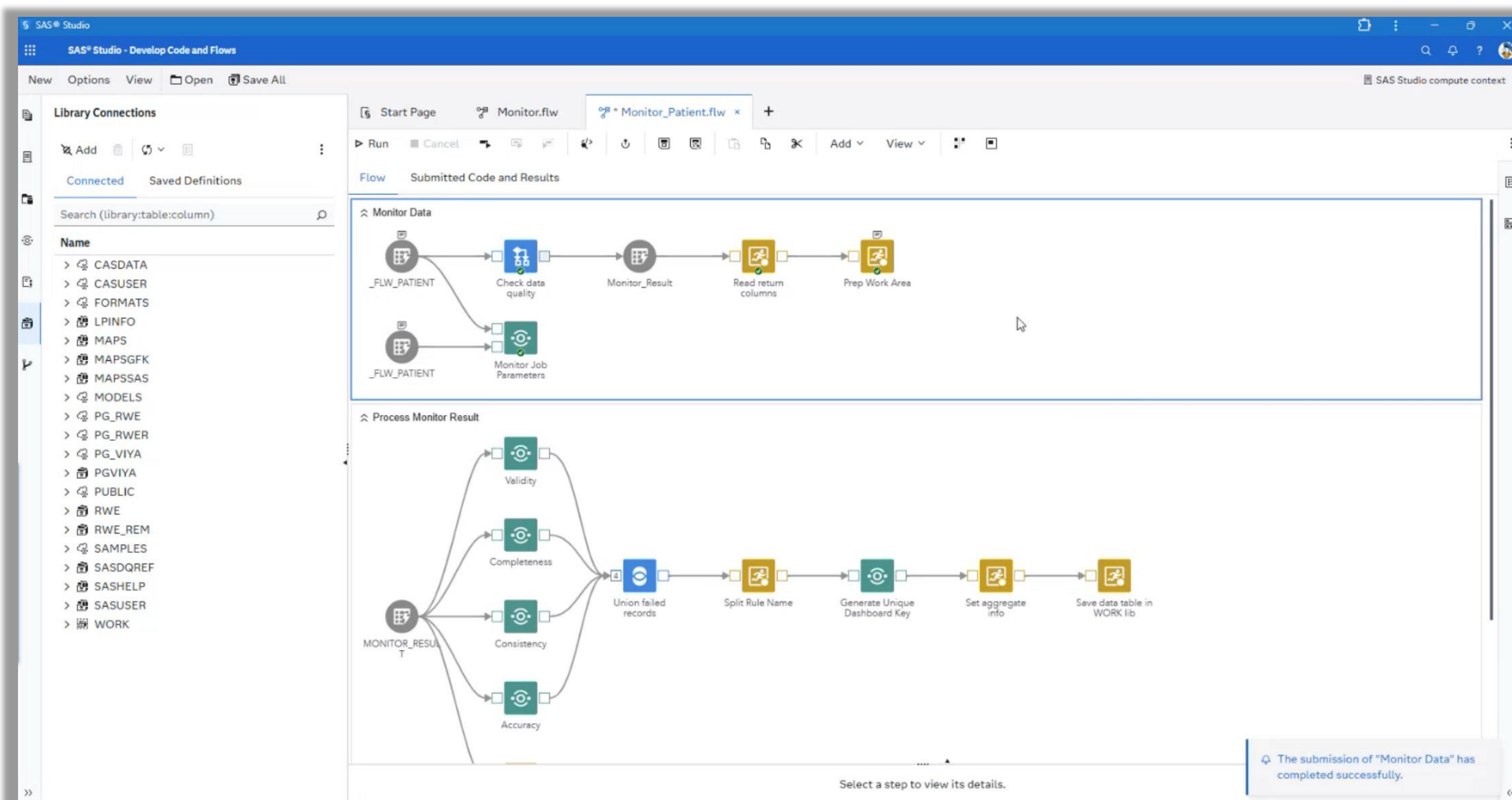
Top Right Screenshot: Lookup Table Editor
This window shows the 'ICD10 (1.1)' lookup table. The 'Lookup Table' tab is active, displaying a table with columns: Key (key) and Value (value). The table lists various ICD10 codes and their corresponding values.

- Rule Set: A collection of logic statements that define conditions to evaluate data quality

Data Quality Layer

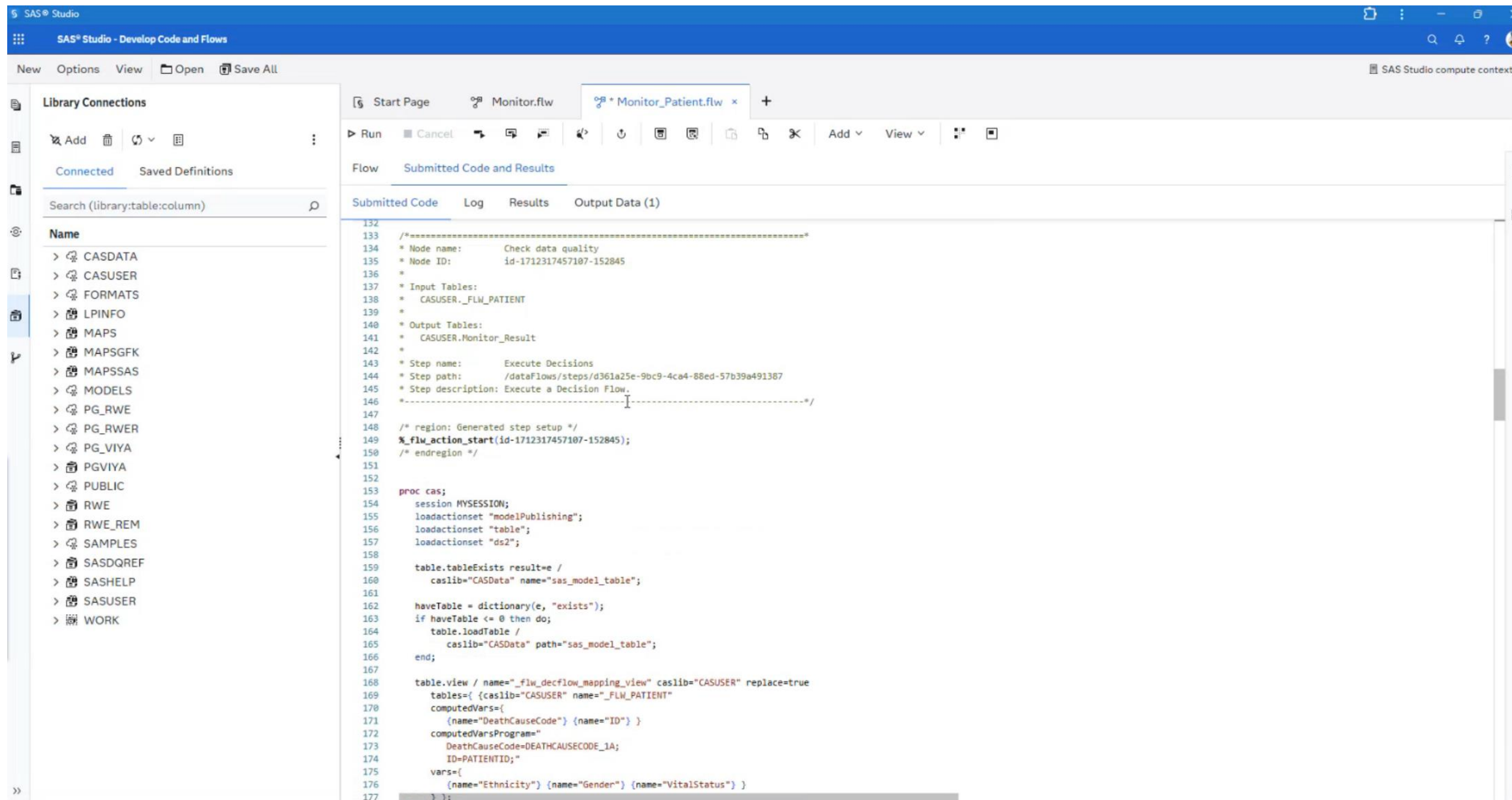
Automate Data Quality Check Pipeline/Workflow

- Deploy a decision flow into a bigger pipeline to automate end-to-end data quality check, data cleaning, and data wrangling tasks.



Data Quality Layer

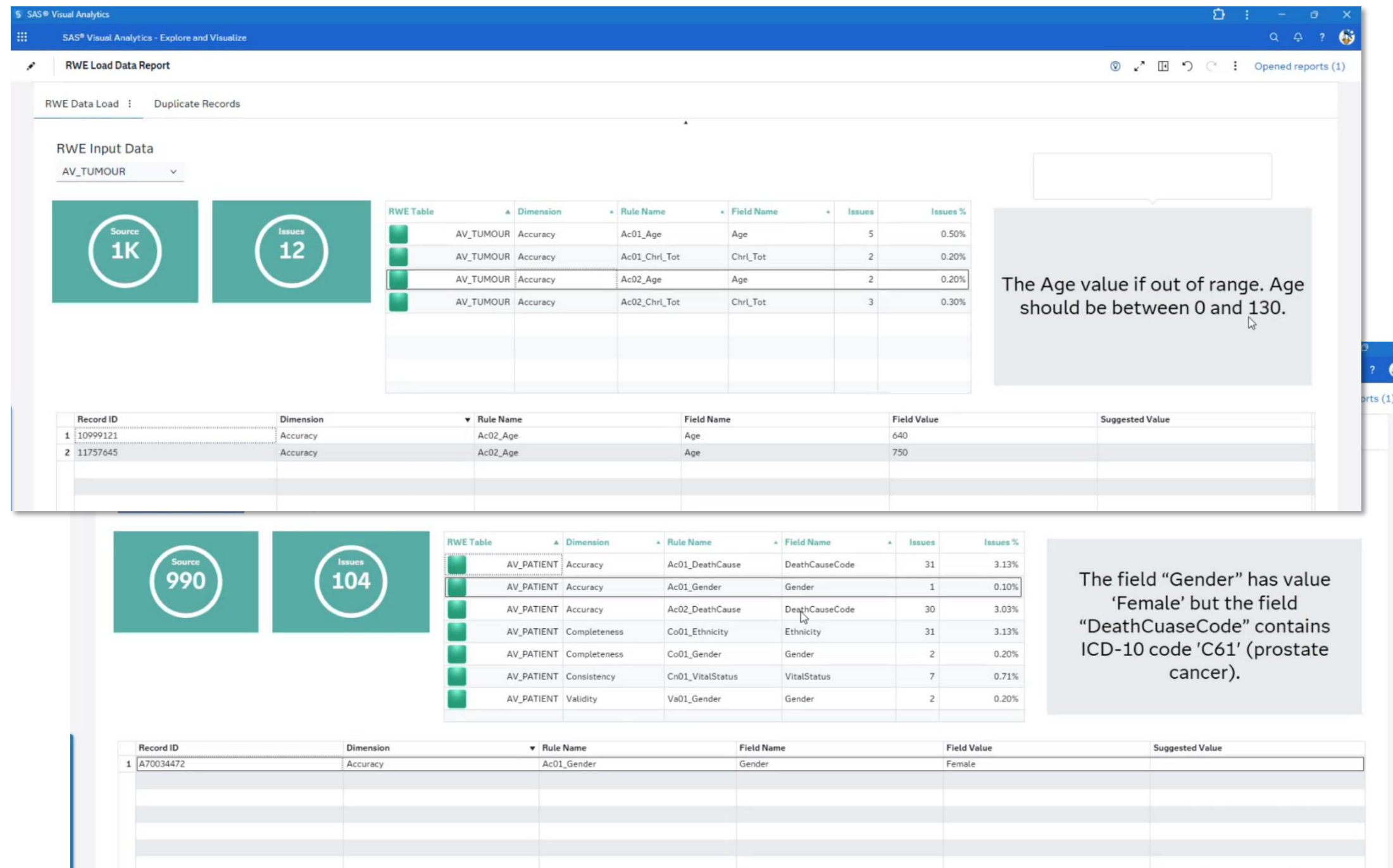
Full SAS Codes and Logs for Traceability



- The full SAS codes and logs are available behinds the pipeline and workflow built from GUI.

Data Quality Layer

Monitor Data Quality Metrics via Interactive Dashboard



- Visible and actionable insights on data quality issues.
- Update data quality dashboard automatically with a scheduled run

The Tiered Data Quality Approach

1

Data Quality Layer

Profile and detect – make data quality issues visible

2

Analytics Layer

Quantify impact and look for problematic trends

3

Action Layer

Remediate – make data fit for purpose

Adapt analytics layer to the specific use case

Automate checks which will be repeated

Quantify impact of data gaps

Measure how missing or inconsistent data changes model output or endpoint validity

Explore relationships

Use unsupervised methods to reveal anomalies or surprising patterns for further investigation

Detect systematic bias

Identify if under-recording or missingness disproportionately affects subpopulations (e.g., older patients, rare indications, or low-volume sites).

Cross reference with external sources

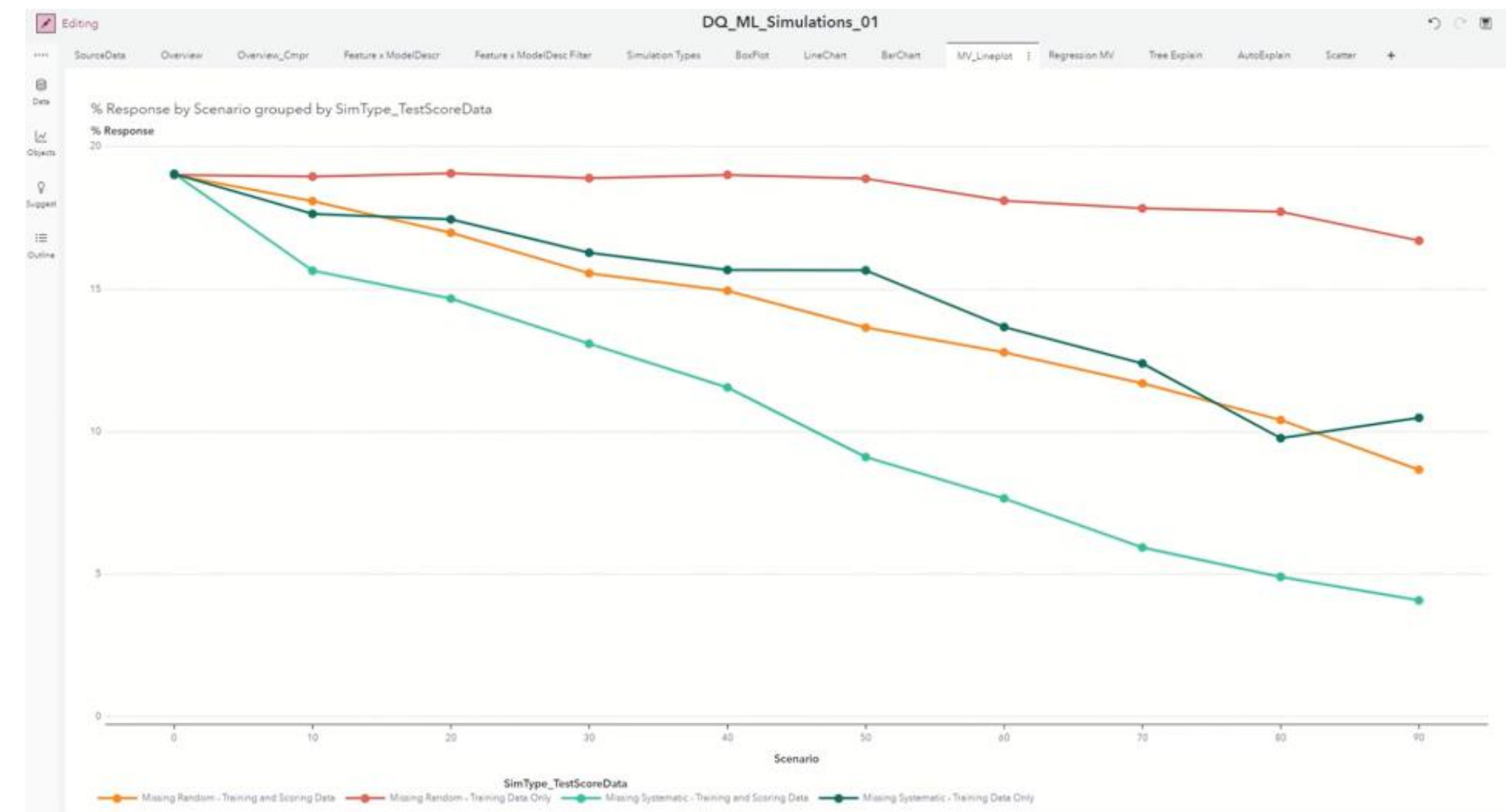
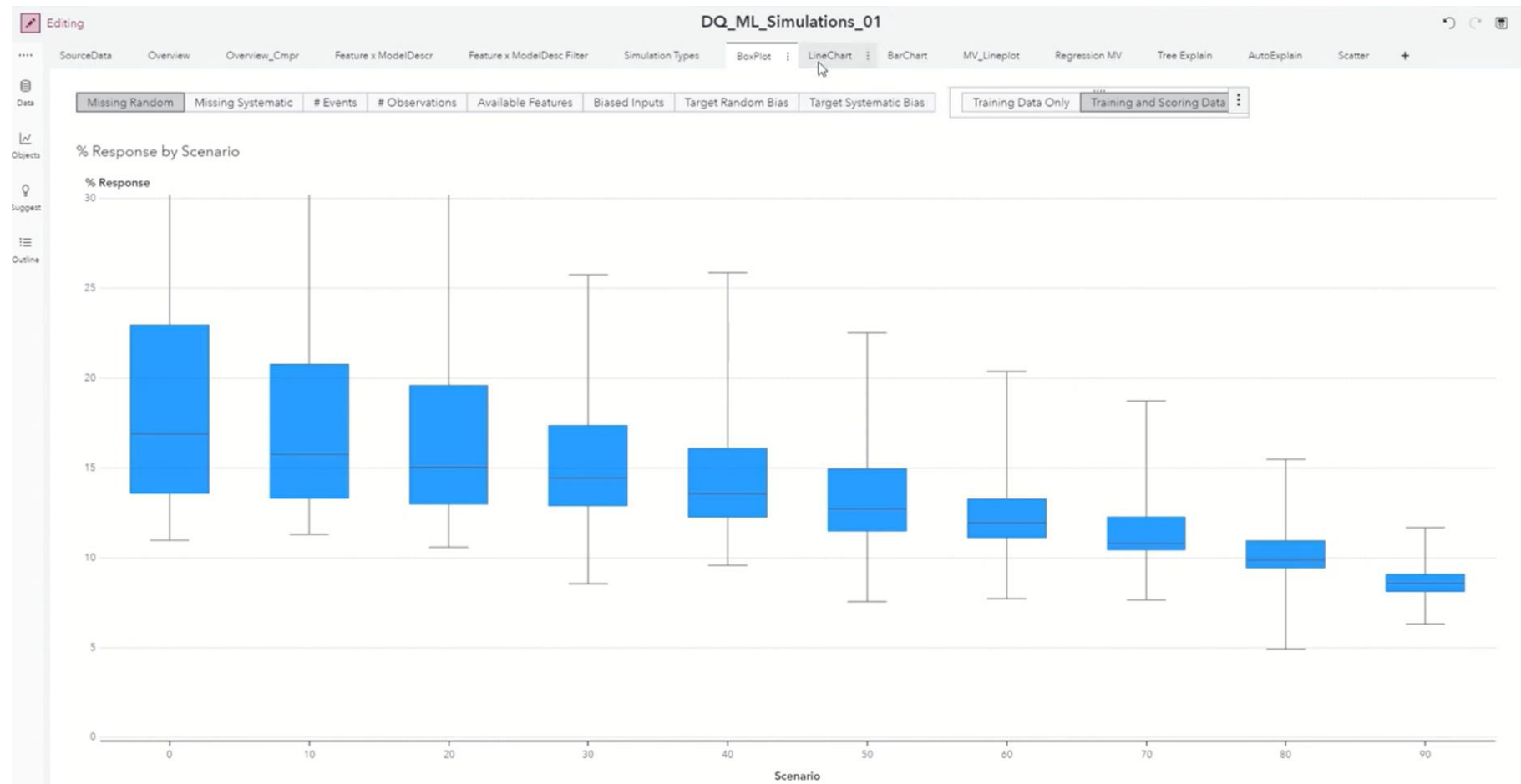
Integrate literature, other data sources to test representativeness

Look for recurring patterns in suspicious data over time

Track anomalies longitudinally, e.g., data drift, shifts in coding practice, or variable completeness over successive data refreshes.

Analytics Layer

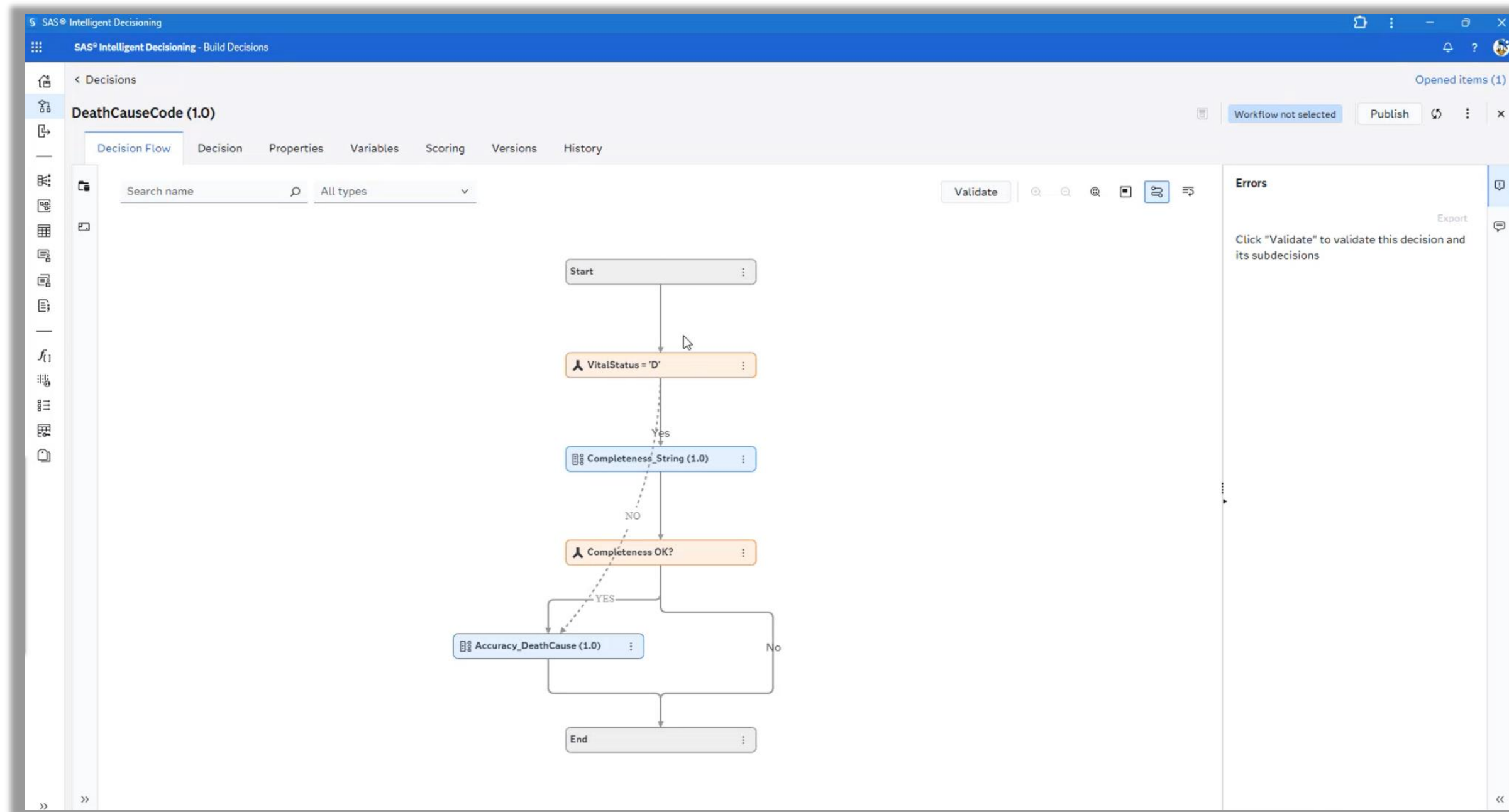
Quantifying the downstream impact of the data quality issues



- Use SAS Visual Analytics to run simulations to quantify the impact of missing values on the fidelity of a machine learning model

Data Quality Layer

Orchestrate Data Quality Decision Flow for Transparency and Reproducibility



- Build those thresholds into the Decision Flow to flag up if data quality issues are too high to use the data

The Tiered Data Quality Approach

1

Data Quality Layer

Profile and detect – make data quality issues visible

2

Analytics Layer

Quantify impact and look for problematic trends

3

Action Layer

Remediate – make data fit for purpose

Tiered approach to data quality

Data Quality Tier

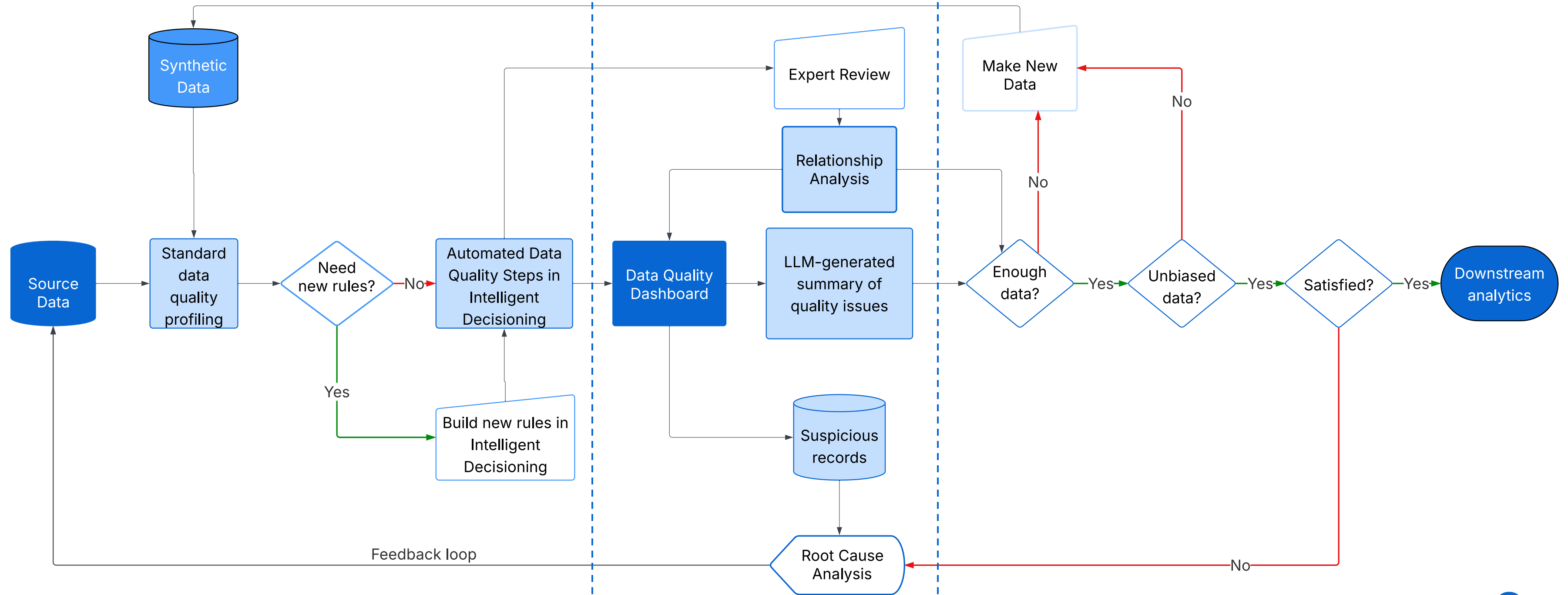
Automated profiling and custom data quality rules

Analytics Tier

Diagnose quality issues

Action Tier

Decide whether to regenerate, fix or approve for downstream use





Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving data science tasks

**The
Alan Turing
Institute**

**THE
ROYAL
SOCIETY**

Hype around Synthetic Data

“
The cumulative sum of human knowledge has been exhausted in AI training... The only way to supplement that is with synthetic data.
”
– Elon Musk

“
The rise of synthetic data will completely transform the economics, ownership, strategic dynamics, even (geo)politics of data.
”
– Forbes

[Synthetic Data Is About To Transform Artificial Intelligence \(forbes.com\)](#)

“
Synthetic data could revolutionise how we access and interact with healthcare datasets.”
”
– Michaela van der Schaar

[Synthetic data: breaking the data logjam in machine learning for healthcare // van der Schaar Lab](#)

“
Through 2030, for data used to train AI models, synthetic structured data will grow at least three times as fast as real structured data.
”
– Gartner

Emerging Tech: Top Use Cases for Tabular Synthetic Data, Gartner, September 22, 2022

“
As long as you can get over the synthetic data event horizon where the model is smart enough to make good synthetic data, it should be alright.”
”
– Sam Altman, Open AI



New Options View Open Save All

GPCI Demos

Start Page Synthetic Data Generation - HMEQ.flw x

Run Cancel Add View

Flow Generated Code Submitted Code and Results

Connect to CAS (if required)

Create CAS
Session and...

Baseline



Input data

Distribution of
Target Variable

Generate Synthetic Data using Generative Adversarial Networks (GAN)

Generate Synthetic Data using Synthetic Minority Oversampling Technique (SMOTE)

Example SMOTE application to boost a class

Assess Generated Data (GAN)

Assess Generated Data (SMOTE)

Persist table

Create Analytical Base Table (ABT)

Create Analytical Base Table (ABT) based on SMOTE

Wind Down (if required)

Steps



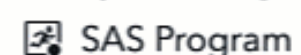
Type to filter list

SAS Steps : Shared

Data (Input and Output)



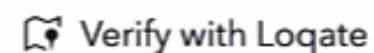
Develop



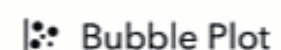
Transform Data

Integrate

Enrichment



Visualize Data



Wrap-up

Tiered approach to data quality

Data Quality Tier

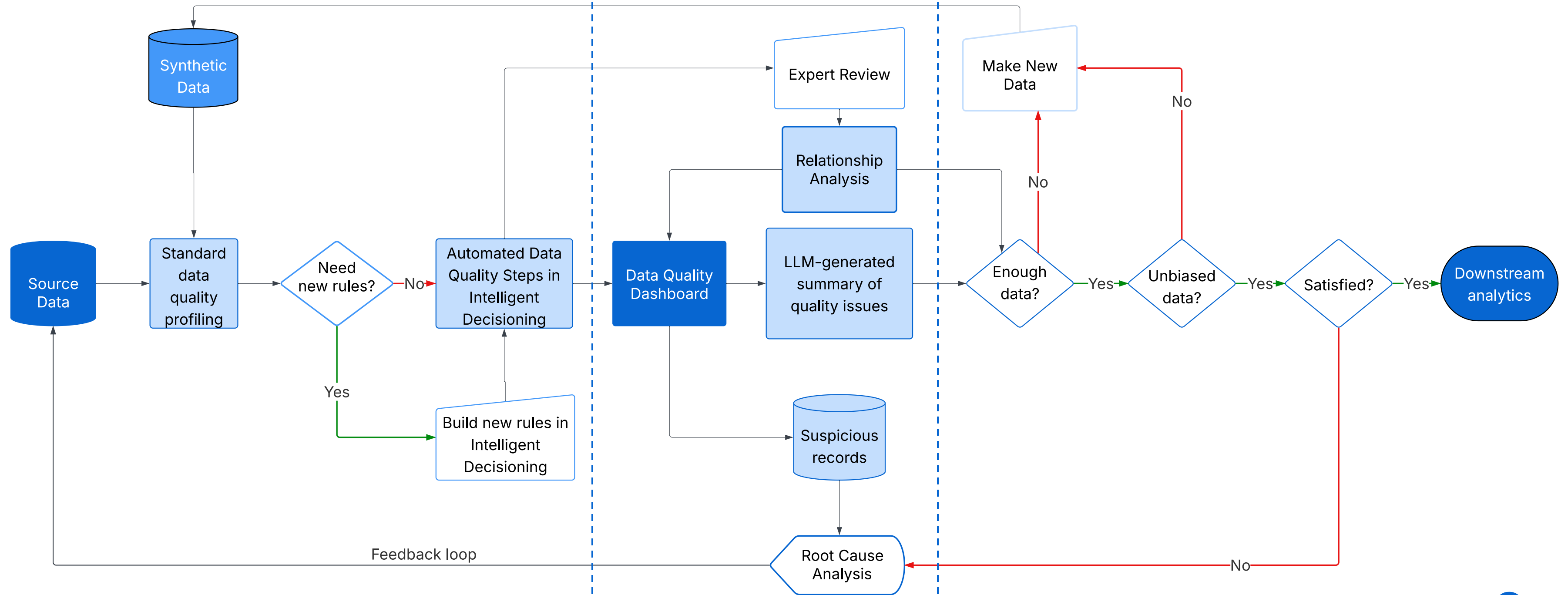
Automated profiling and custom data quality rules

Analytics Tier

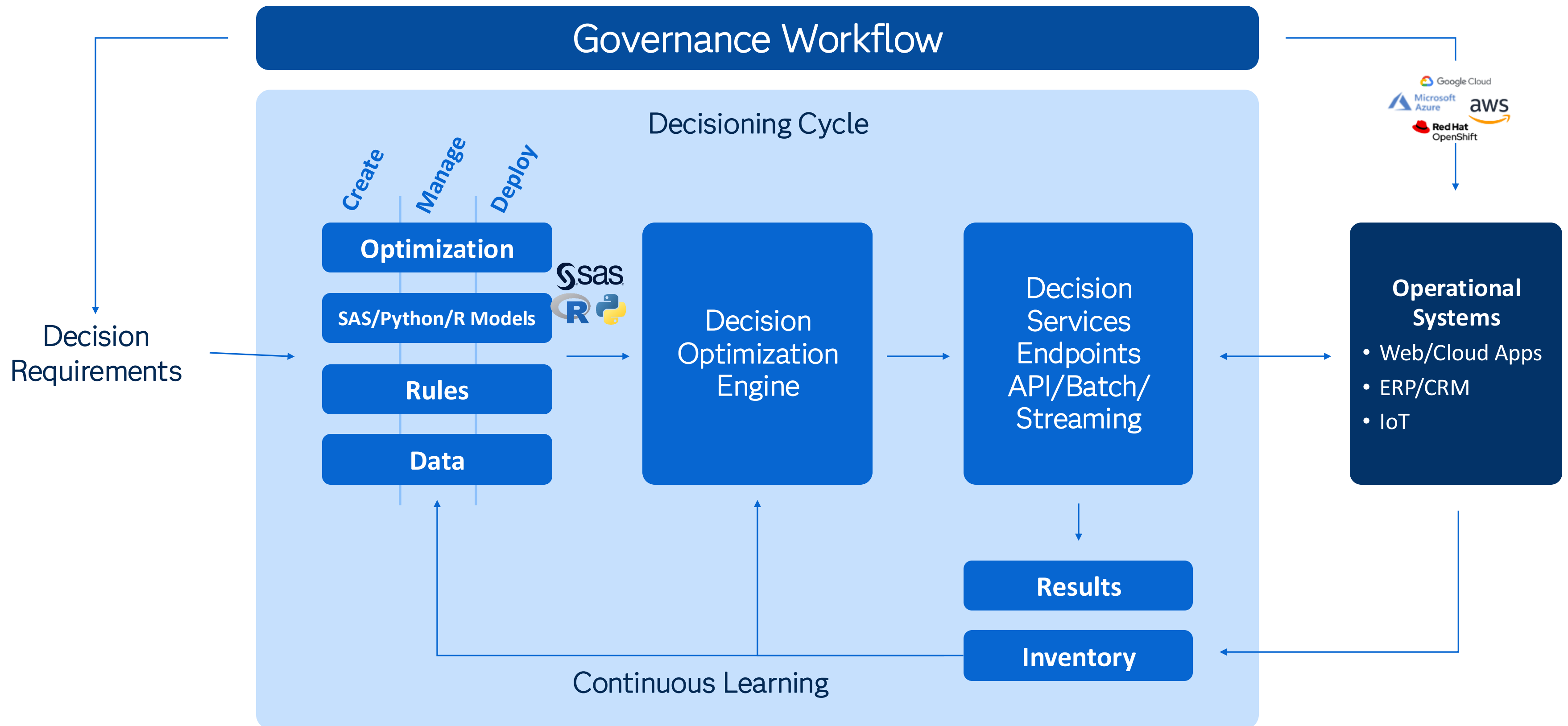
Diagnose quality issues

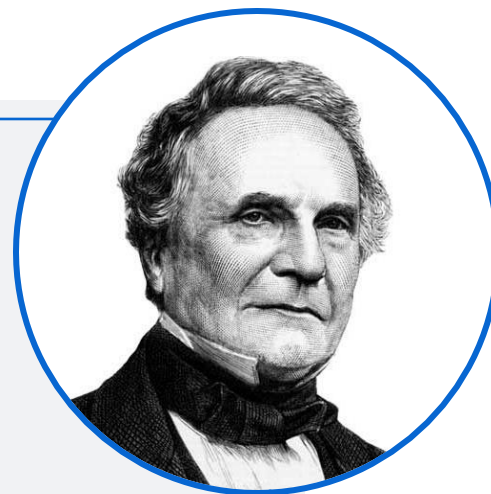
Action Tier

Decide whether to regenerate, fix or approve for downstream use



The Decisioning Cycle





Errors using
inadequate
data are much
less than
using no data
at all

- Charles Babbage



Data quality is
a critical
element for
realising the
full potential
of data-driven
regulation

- European Medicines Agency

Thank you!

Lisa Murch, lisa.murch@sas.com
William Kuan, William.kuan@sas.com



Appendx

Pre-generation tasks

Right Data begets Right Data

Data Assessment and Quality Checks

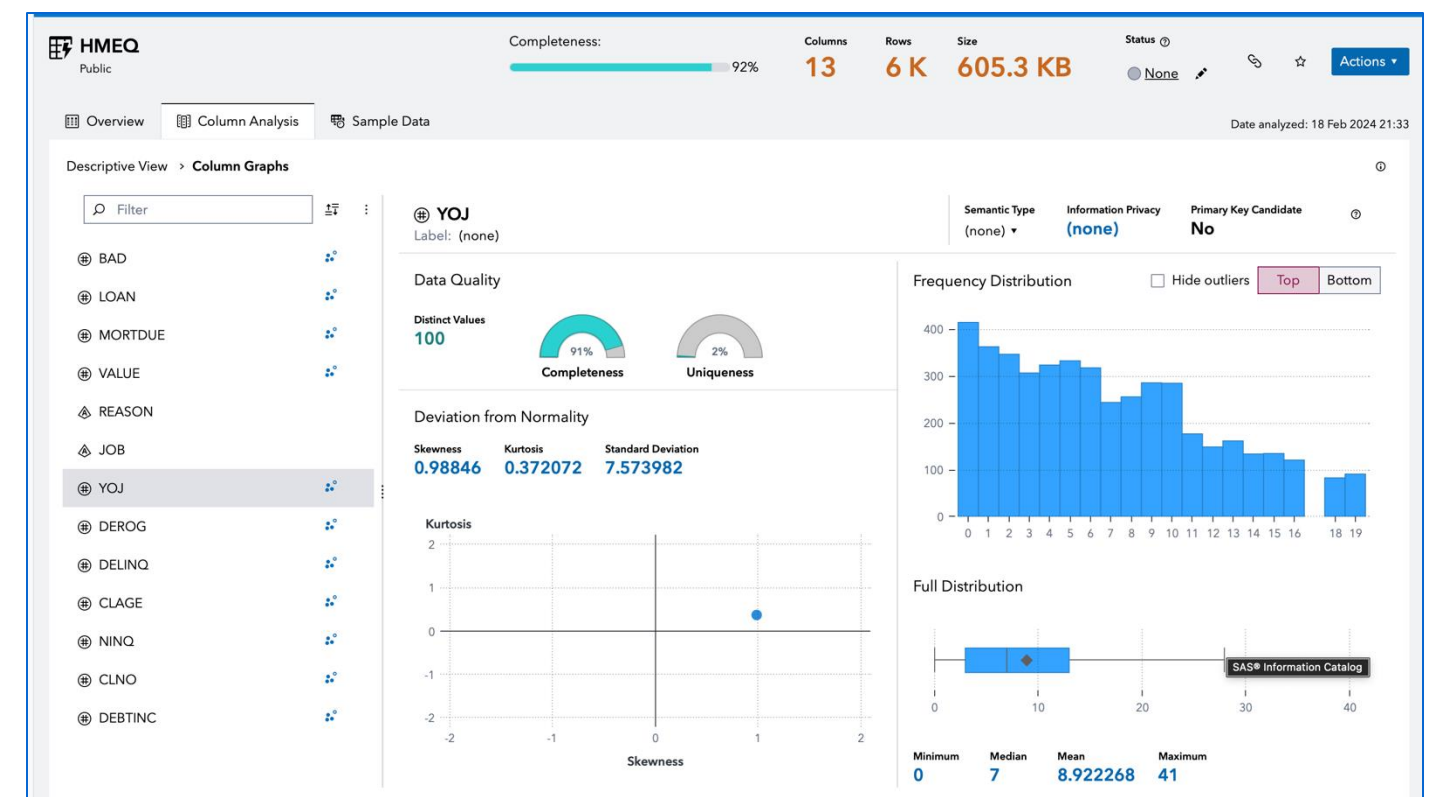
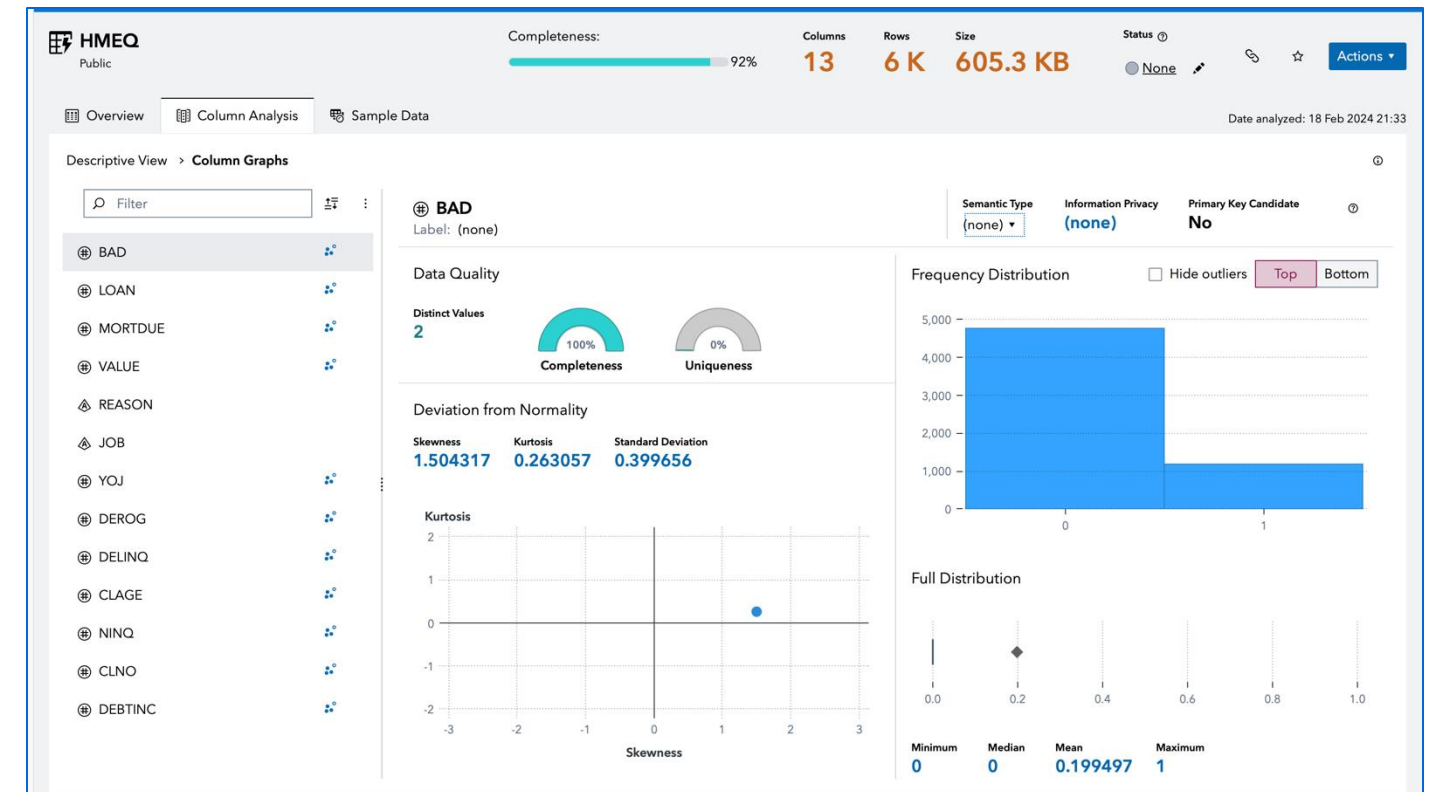
- De-identification
- Imputing missing values
- Identifying anomalies

Data Transformation

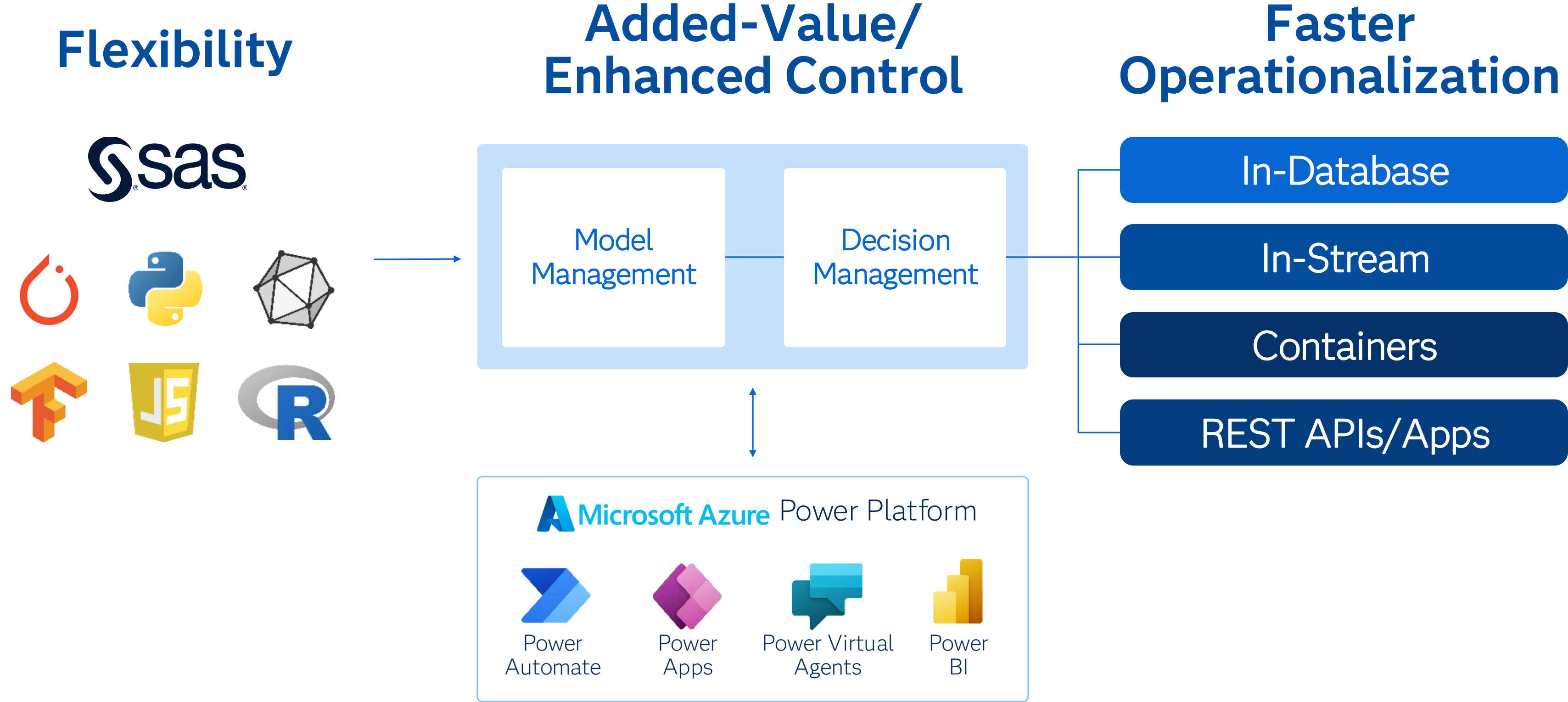
- Variable transformation and encoding
- Normalisation

'Learning' the Distribution

- Joint Probability Distributions
- Clustering & 'right' number of clusters



The SAS Approach for ModelOps



Improving Data Quality in Oncology Electronic Health Record (EHR)-Derived Databases for Research and Regulatory Decision-Making

Zhaohui Su, PhD; Lisa Herms, PhD; Aniketh Talwai, MBA; Jessica Paulus, ScD; Amy K. O'Sullivan, PhD; Sarah Spark, RN; Jennifer Frytak, PhD; Robyn Harrell, MS; Jay Hugalavalli, BS; Janet Espirito, PharmD — Ontada, Boston, MA, USA

Objectives

- Implement the FDA-led initiative on the oncology quality, characterization, and assessment of real-world data (RWD) (QCARD¹) to evaluate the relevance, reliability, and external validity of EHR data.
- Facilitate the use of robust RWD and real-world evidence (RWE) in oncology.

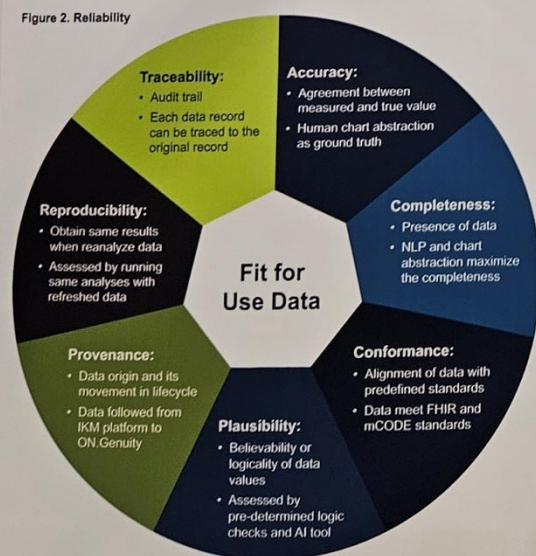
Results

- Common challenges to RWD were mitigated by integrating data from multiple sources (Table 1).
- The RWD quality dimensions defined in the FDA QCARD initiative were implemented (Figure 1-3).
- Assessment of relevance revealed consistent high availability of 250+ standardized variables across 20+ clinical domains.
- Mortality and recurrence-free survival data demonstrated high consistency with the external National Death Index (NDI) data source and clinical trial data (placebo arm) (Figure 4).

Figure 1. Relevance

Availability (data elements)
• 250+ standardized variables across 60+ cancers • Oncology-specific EHR data
Feasibility (representative patients)
• Largest network of community oncologists in the US with 2,700+ providers • Nearly 1 in 4 U.S. cancer patients treated in The Network • 15-year longitudinal near real-time linked data

Figure 2. Reliability



References

1. Rivers DR, Eckert JC, Rodriguez-Watson C, et al. The Oncology QCARD Initiative: Fostering efficient evaluation of initial real-world data proposals. *Pharmacoepidemiol Drug Saf*. 2024 Nov;33(11):e5818.
2. Osterman TJ, Terry M, Miller RS. Improving Cancer Data Interoperability: The Promise of the Minimal Common Oncology Data Elements (mCODE) Initiative. *JCO Clin Cancer Inform*. 2020 Oct;4:993-1001.

Methods

The RWD quality dimensions in the FDA QCARD initiative were implemented using Ontada's ON.Genuity RWD platform, which integrates the iKnowMed® EHR data from ~500 US community oncology clinics with claims and external mortality data.

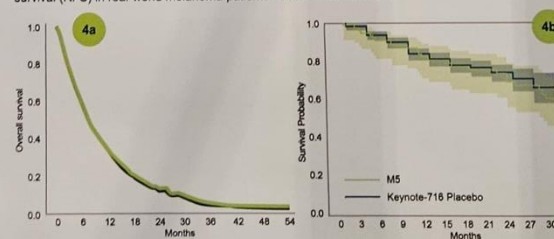
Table 1. Data Sources in Ontada's ON.Genuity RWD

Data source	Structured iKnowMed® EHR Data	Chart abstraction	Non-EHR Data
Example Variables	Gender, race, ethnicity, vital signs, labs, performance status, stage	Oral medication stop dates, surgery date, hospitalizations, disease progression status	Claims data (amount paid, payer type), supplemental mortality data
Processing Method	Ingestion, standardization, reconciliation	Technology-enabled human abstraction, natural language processing (NLP)	Linked data at patient level
Highlights	• 250+ variables were standardized following mCODE² initiative. • EHR data were linked based on unique and accurate ID within iKnowMed®.		

Figure 3. External Validity

Generalizability (inference to broader population)	Replicability (confirm findings in different populations)	Transparency (clear communication and description)
Generalizable to community oncology setting in the US	Validation against National Death Index, published clinical trial data	Detailed documentation of derivations, design, methods, analyses

Figure 4. Implementation of external validity. (a) external validity of OS for metastatic pancreatic cancer showing identical curves between EHR and NDI, and (b) Kaplan-Meier for recurrence-free survival (RFS) in real-world melanoma patients vs KN-716 control arm



Discussion

Ontada's ON.Genuity EHR database was standardized using FHIR and mCODE², with increased relevance, reliability and external validity by linking to external structured data at a patient level, human chart abstraction, and leveraging NLP technology.

The quality and fit-for-use of RWD should be carefully evaluated for each study, including a detailed assessment of relevance, reliability and external validity in relation to the research objectives.

Conclusions

- The QCARD initiative represents a significant advancement in EHR-based oncology research, enhancing RWE from RWD and improving decision-making in clinical practice and policy.
- The quality and fit-for-use of RWD should be carefully evaluated in accordance with the research objectives.