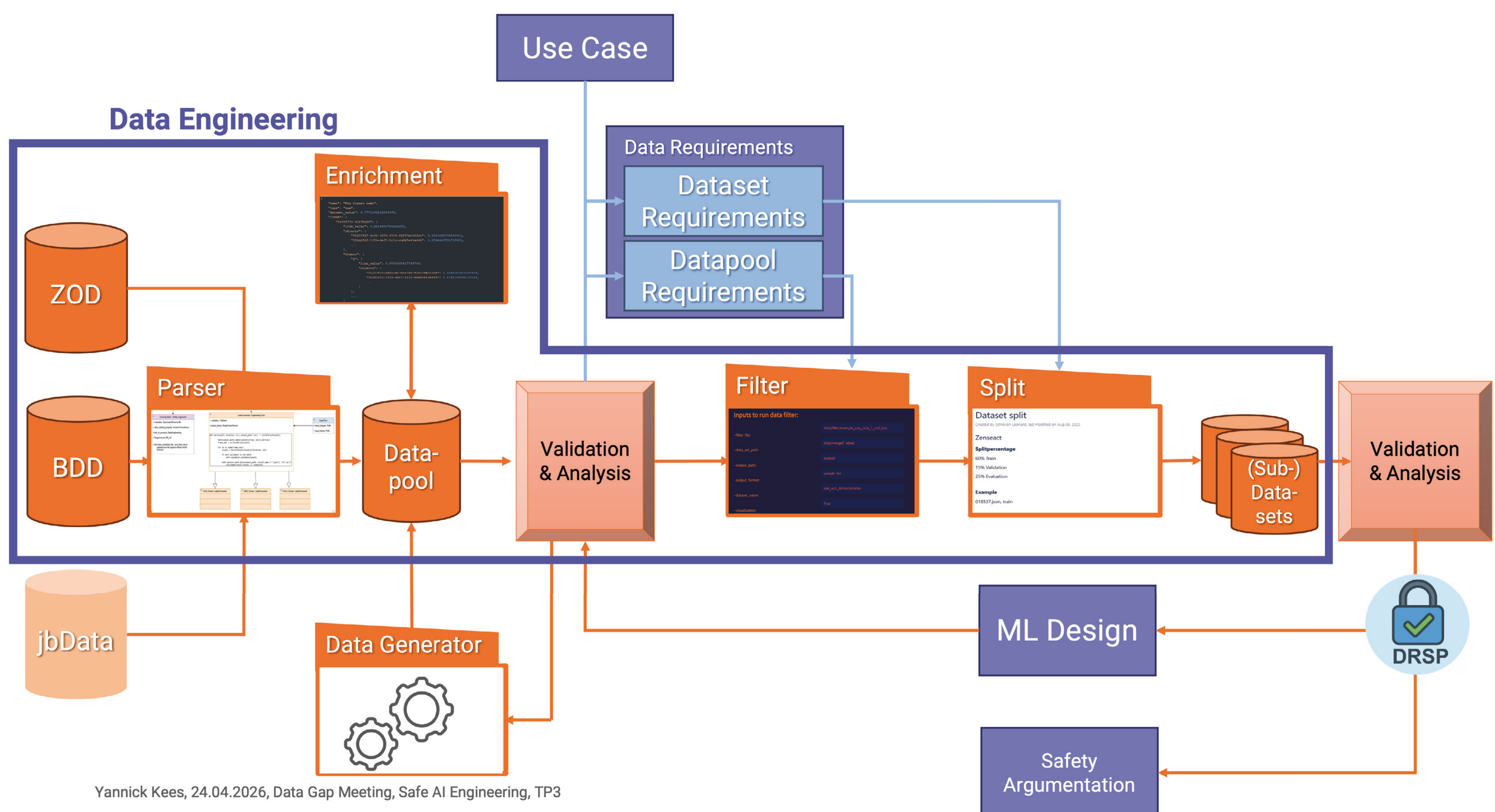


Data Specification and Generation

Leonard Schroven (Capgemini Engineering)
Philip Schoerner (FZI Forschungsinstitut Informatik)



Yannick Kees, 24.04.2026, Data Gap Meeting, Safe AI Engineering, TP3

Problem Statement

Data is the foundation of any security argument. Only when we understand what information our data contains, the quality with which it represents that information, and what gaps exist can we derive specific requirements, ensure security objectives, and develop AI systems

We collaborated with the other subprojects to define the aspects of the Safe AI methodology that must be reflected in the data

Enrichment

To better support safety-related analyses, we enriched the ground truth data directly from the annotation labels. Each project partner developed a set of metrics, collectively referred to as Scenario Descriptive Metrics (SDMs), to capture safety-relevant characteristics of the traffic environment and participant interactions. These metrics were subsequently consolidated into a unified framework.

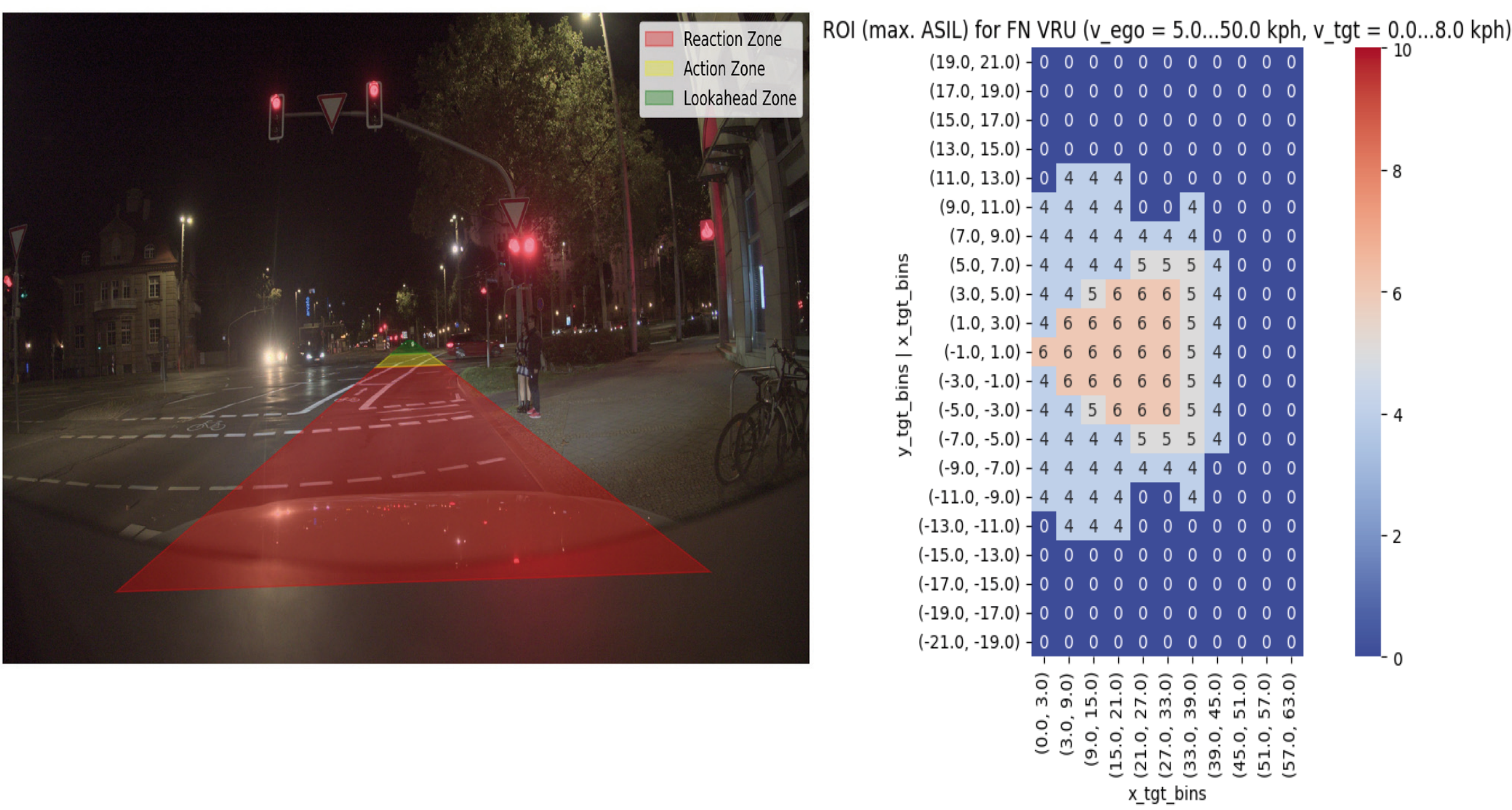


Figure 1: Example of the Safety Zone Metric by Bosch on a Frame from the ZOD Datasets*

Data Filter

The Data Filter uses enriched ground truth data and Scenario Descriptive Metrics to create targeted dataset distributions for specific objectives. By selecting and combining scenario characteristics, users can generate datasets that focus on particular operational conditions, safety-critical situations, or coverage gaps. This enables subprojects to efficiently filter data according to their needs, ensuring that relevant data is selected for training, evaluation, and verification. Using the available metrics, users can create targeted subsets aligned with specific Operational Design Domains (ODDs), safety requirements, or scenario characteristics.

Datagap Identification

In addition to the current capabilities, our future work will focus on data gap analysis to systematically identify missing or underrepresented data within the datasets. The objective is to understand where coverage is insufficient. safety requirements, and scenario characteristics, and to determine how these gaps can be addressed effectively. A key research area will be the integration of multiple data sources, including real-world data, simulation-generated data, and generative AI (GenAI) data. We aim to investigate how these different sources can be combined in a reliable and traceable manner to enhance dataset coverage while maintaining data quality and consistency.

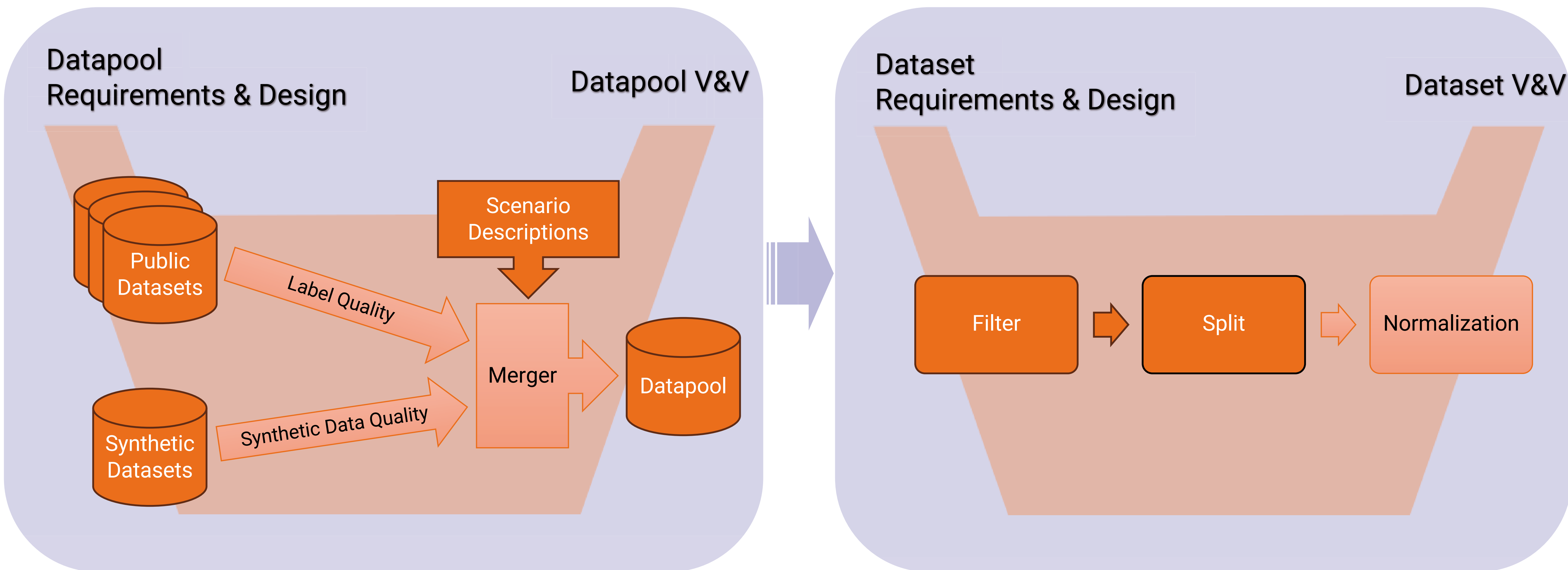


Figure 2: Data Engineering Process

* For this dataset, Zenseact AB has taken all reasonable measures to remove all personally identifiable information, including faces and license plates. To the extent that you like to request removal of specific images from the dataset, please contact privacy@zenseact.com <https://zod.zenseact.com/>

For more information regarding the poster, contact:

leonard.schroven@capgemini.com,
schoerner@fzi.de

Safe AI Engineering is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Energy.



[linkedin.com/company/ki-familie](https://www.linkedin.com/company/ki-familie)
www.safe-ai-engineering.de