



SAFE AI
ENGINEERING

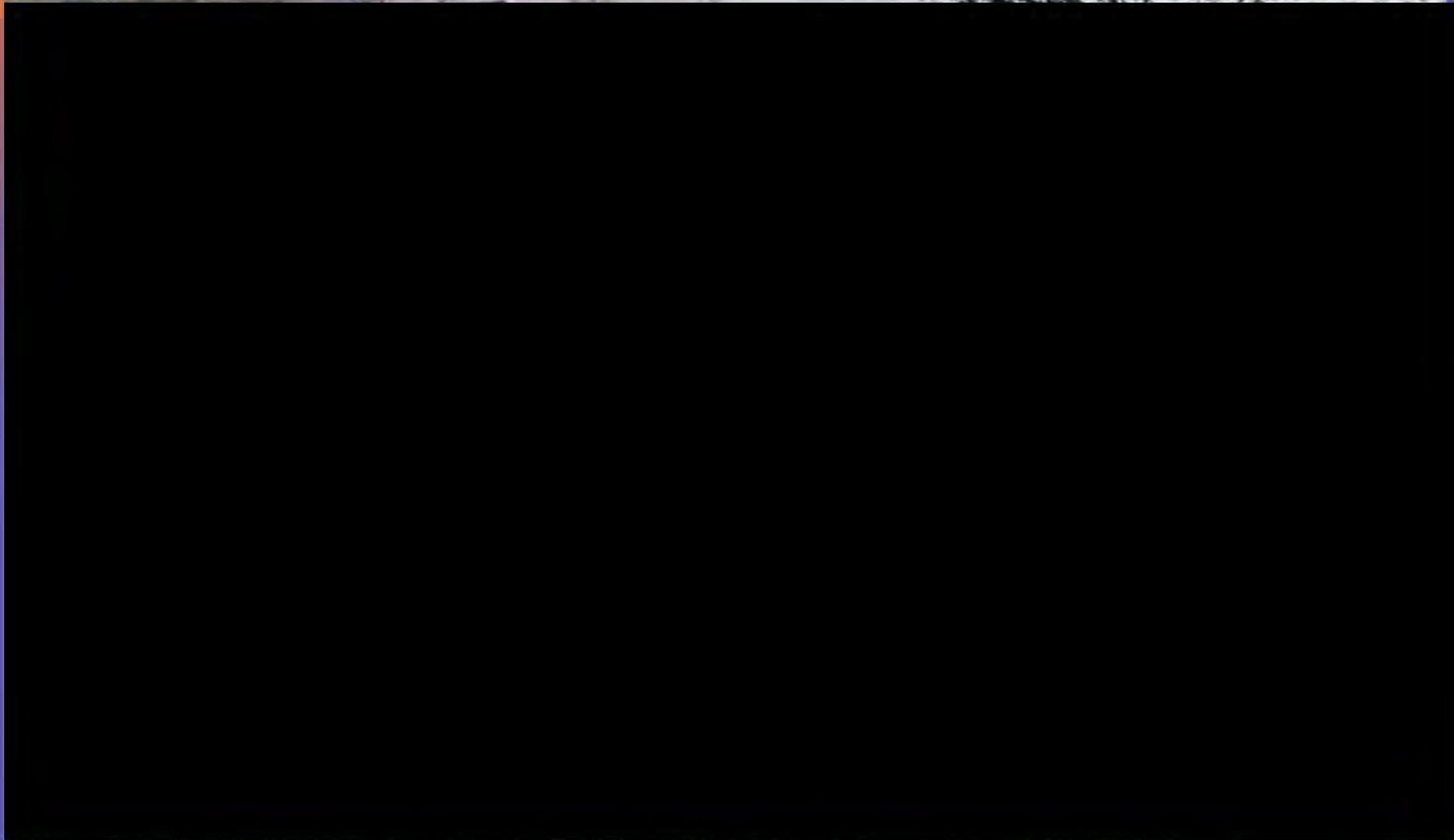
September 16, 2026 | Safe AI Engineering

Methodology Enabling AI Safety Assurance

Torsten Stiehm | DXC Luxoft



Methodology Enabling AI Safety Assurance, no red Flag



Safe AI is Becoming a Prerequisite for Deployment, Regulation and Acceptance.



The aim is to encourage the development and uptake of safe and trustworthy AI systems.

(EU AI Act, Regulation (EU) 2024/1689)
Establishing trust as the foundation for the successful deployment and adoption of AI in safety-critical domains.

Confidence, not capability, will determine adoption.

(World Economic Forum)
Trustworthy AI is emerging as a strategic enabler for innovation, adoption and sustainable growth.

As AVs become more prevalent, explaining their behavior and decisions to stakeholders and the public is crucial for building trust and confidence.

(European Transport Research Review, 2025)



EU Artificial Intelligence Act



We are Joining Forces to Enable AI Safety Assurance!



Supported by:



Federal Ministry
for Economic Affairs
and Energy

on the basis of a decision
by the German Bundestag



www.safe-ai-engineering.de/en/home

www.linkedin.com/company/ki-familie

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.

Images: Safe AI Engineering Consortium

What do we Mean by Safety?



What could possibly go wrong?

Take wrong turn



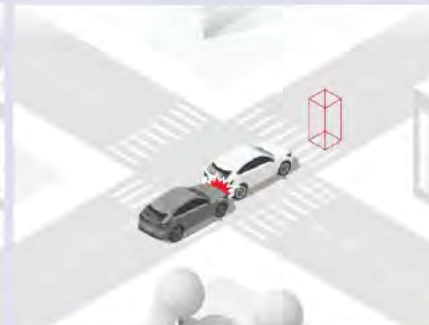
Hit an obstacle



Engage unknown object



Phantom detection

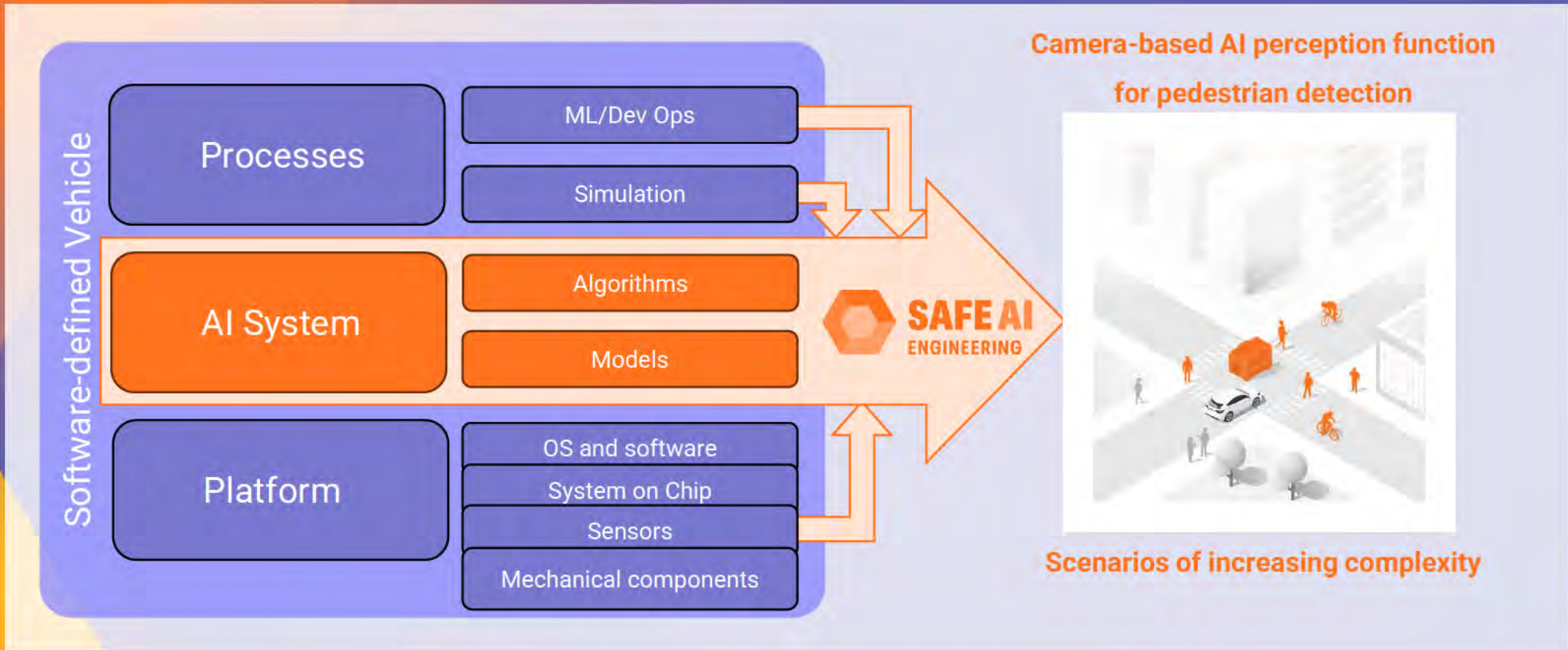


Hit a vulnerable road user

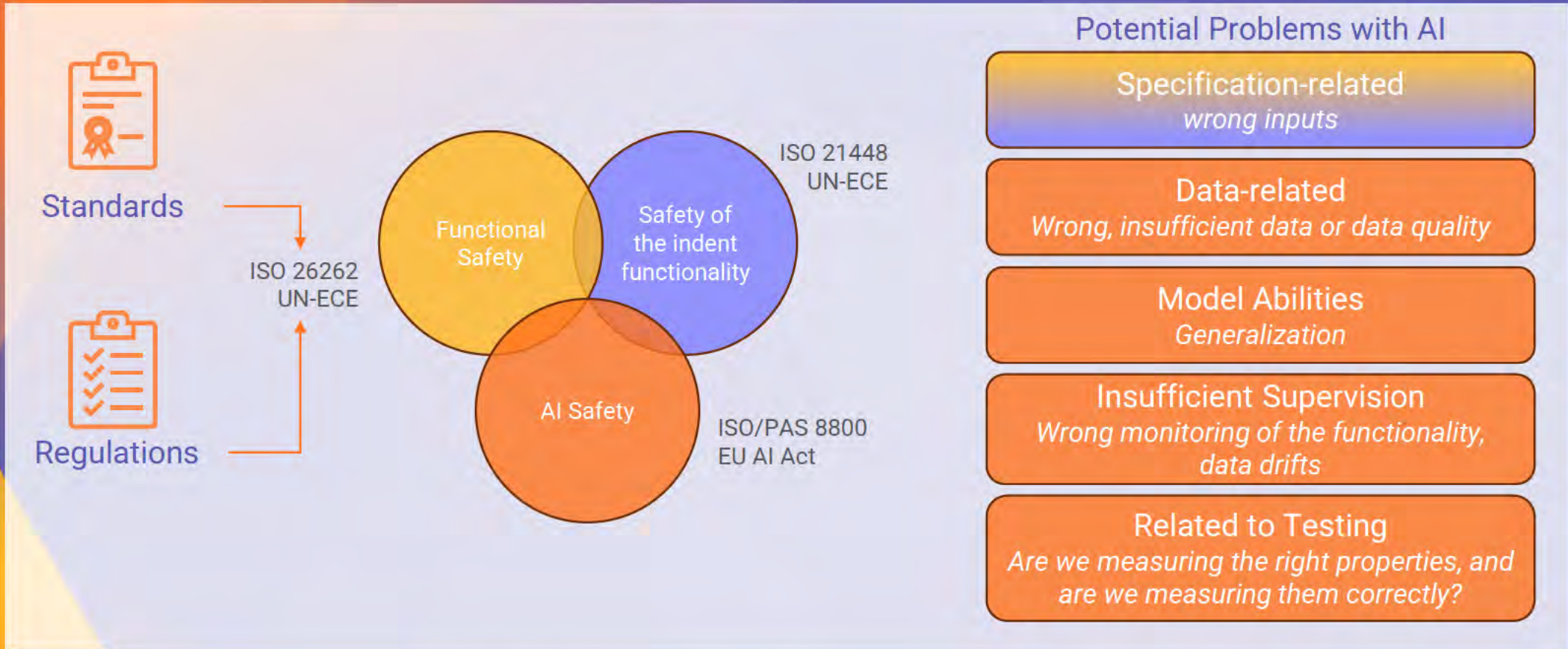


Safety is absence of unacceptable risk in a physical system

Where can Things go Wrong?



Safety of AI and why things can go wrong



Braiding a Safety Rope



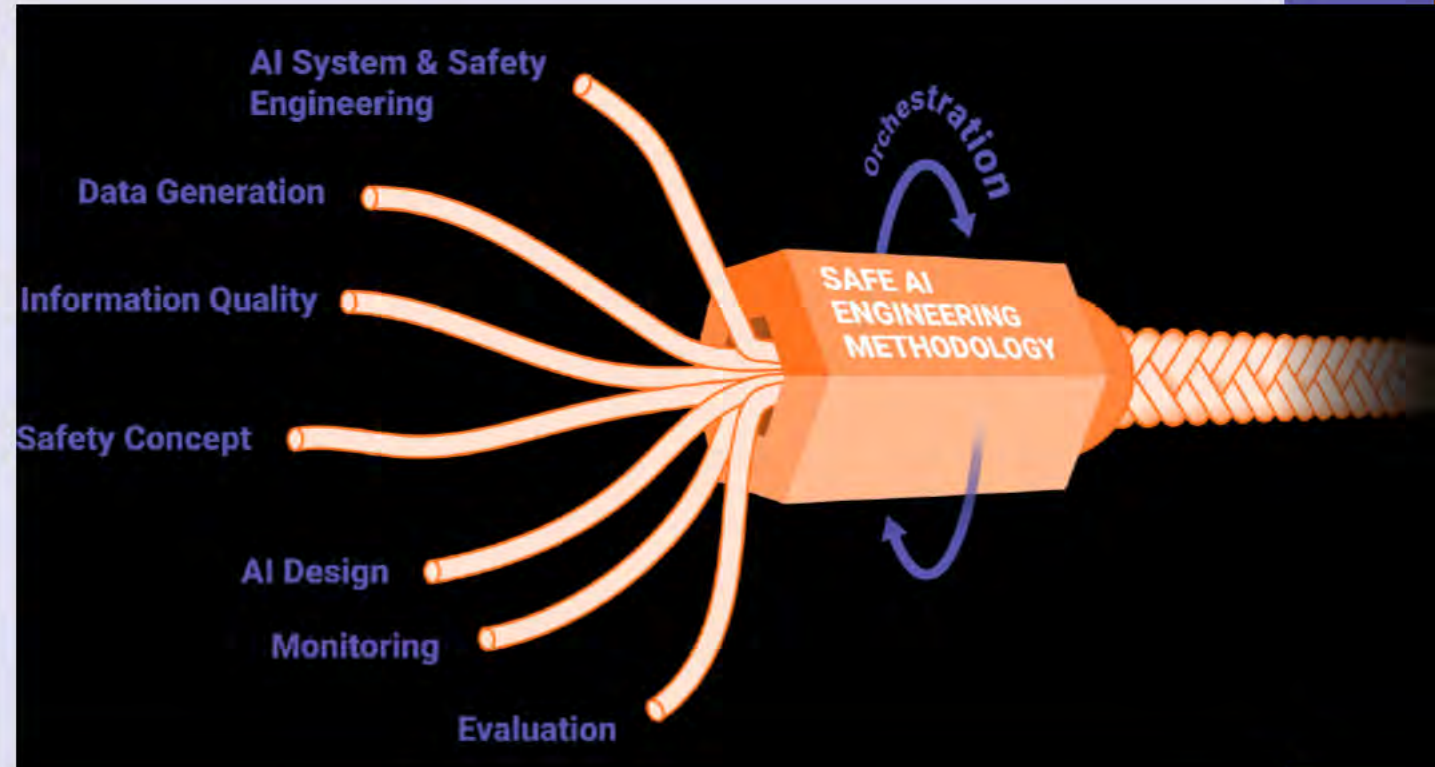
Specification-related
Wrong inputs

Data-related
Wrong, insufficient data or data quality

Model abilities
Generalization

Insufficient supervision
*Wrong monitoring of the functionality,
data drifts*

Related to Testing
*Are we measuring the right properties, and
are we measuring them correctly?*



Our Goal: Holistic Assurance of AI Functions in Automated Driving



With the AI engineering methodology, a **generally accepted, practical proof of safety for AI** is to be developed – **over the entire life cycle** of an AI function.



Camera-based AI perception function

A development cycle based on scenarios of **increasing complexity**

Development of **processes, methods and metrics** for **safe AI engineering**

Reflecting existing norms and standards



Aligning interconnections between different AI development domains to enable safe homologation argumentation



Braiding a Safety Rope



Braiding a Safety Rope





SAFE AI
ENGINEERING

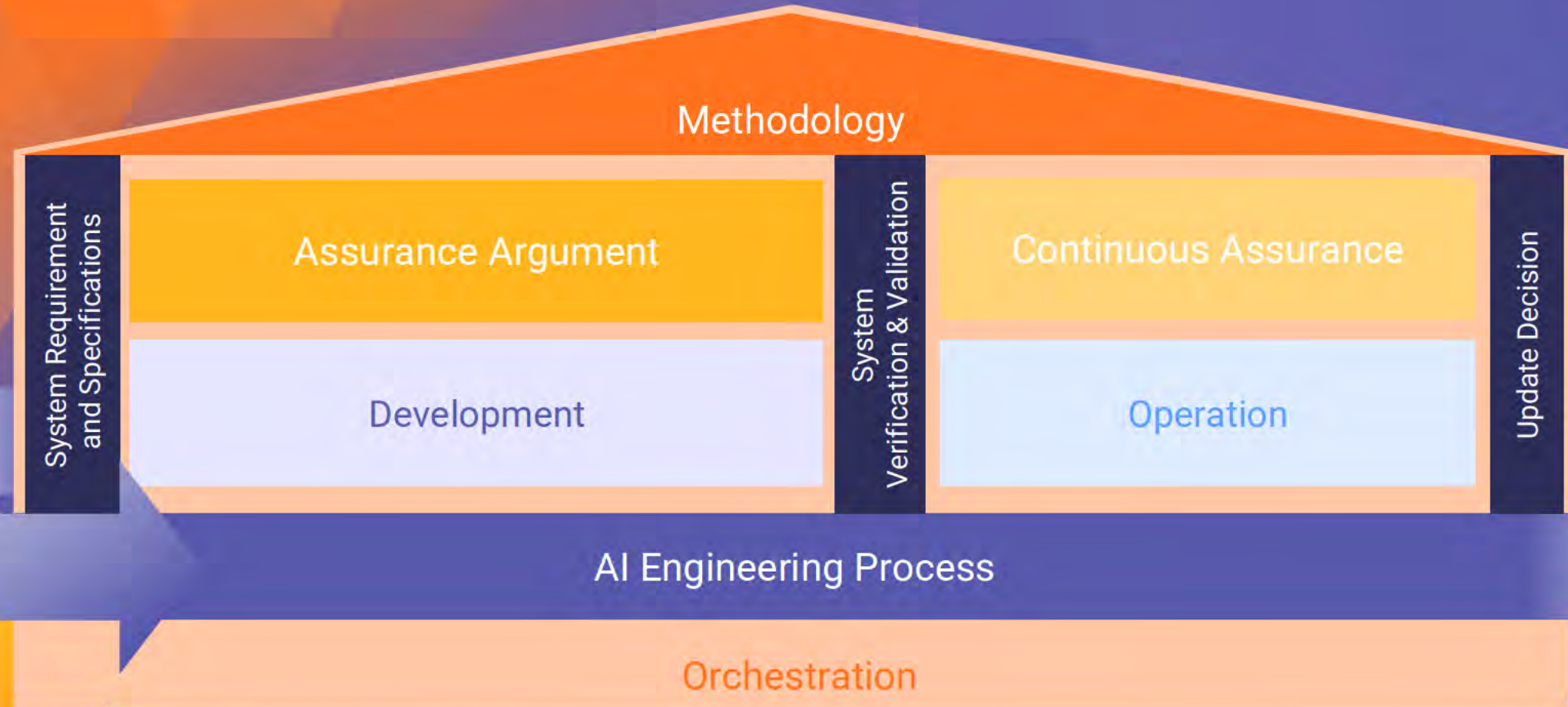
September 16, 2026 | IAA Transportation

Methodology

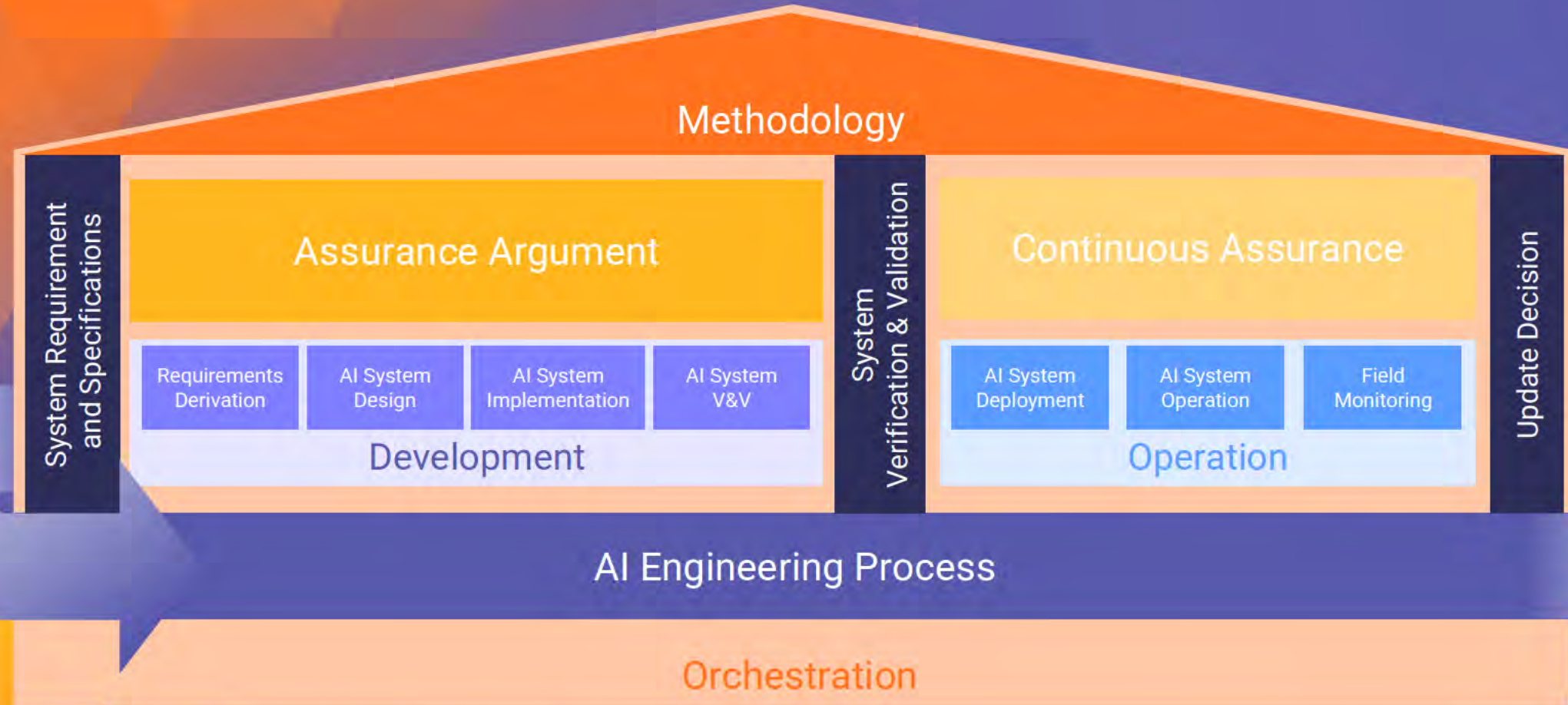
Elena Hoemann | DLR



Methodology Level 1



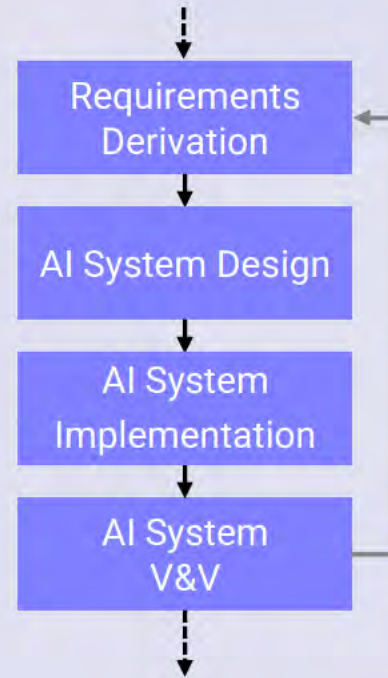
Methodology Level 2



Process

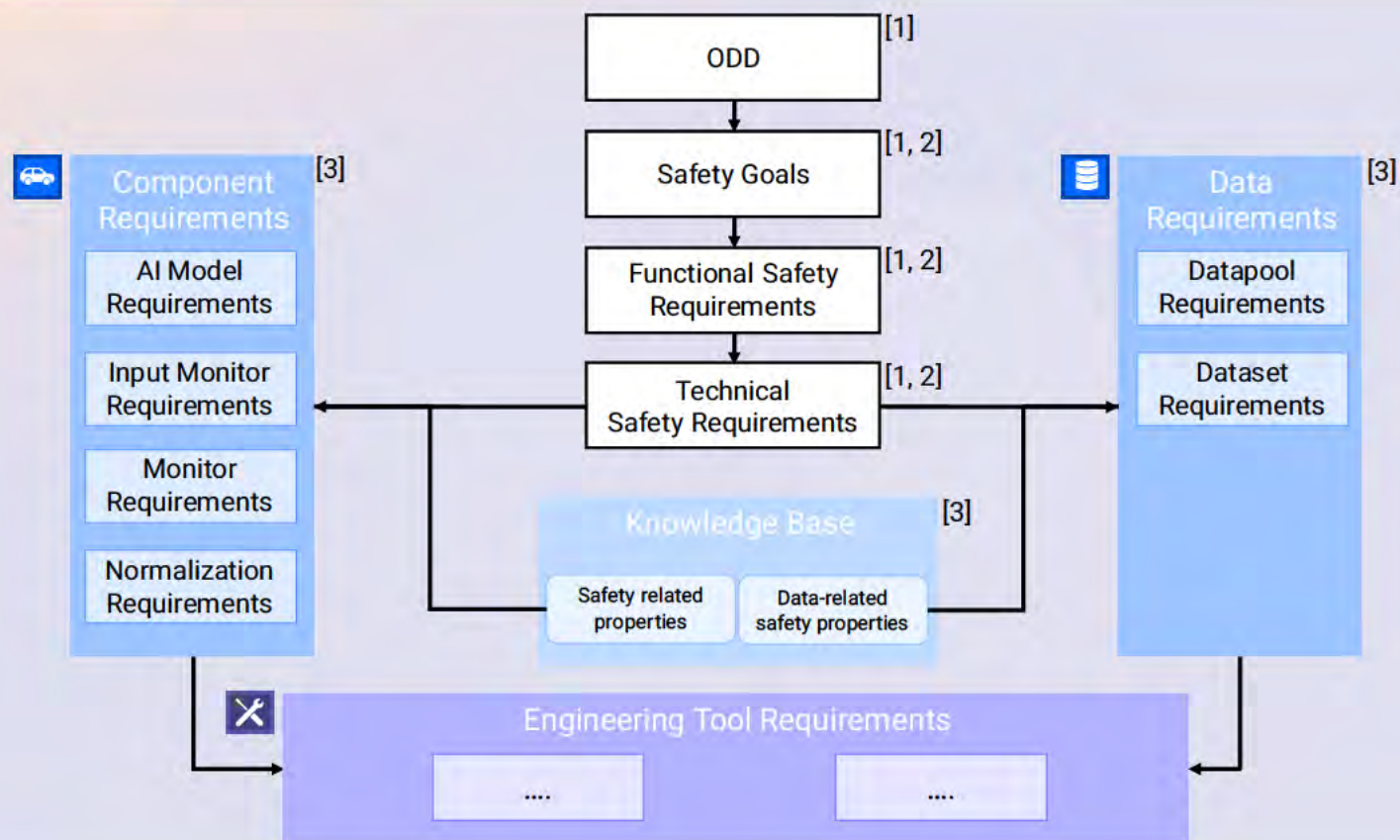


AI Development



Assurance Argument

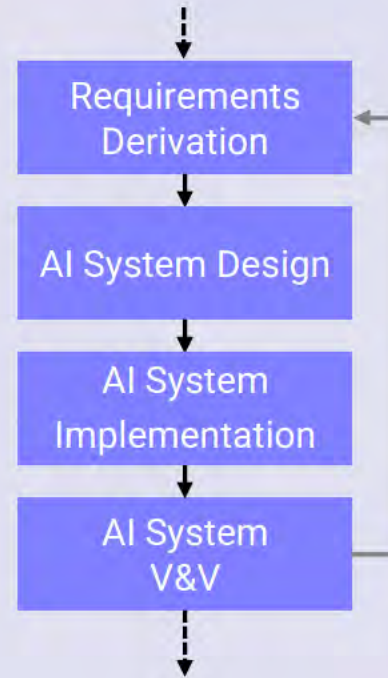
Requirements Derivation



Process

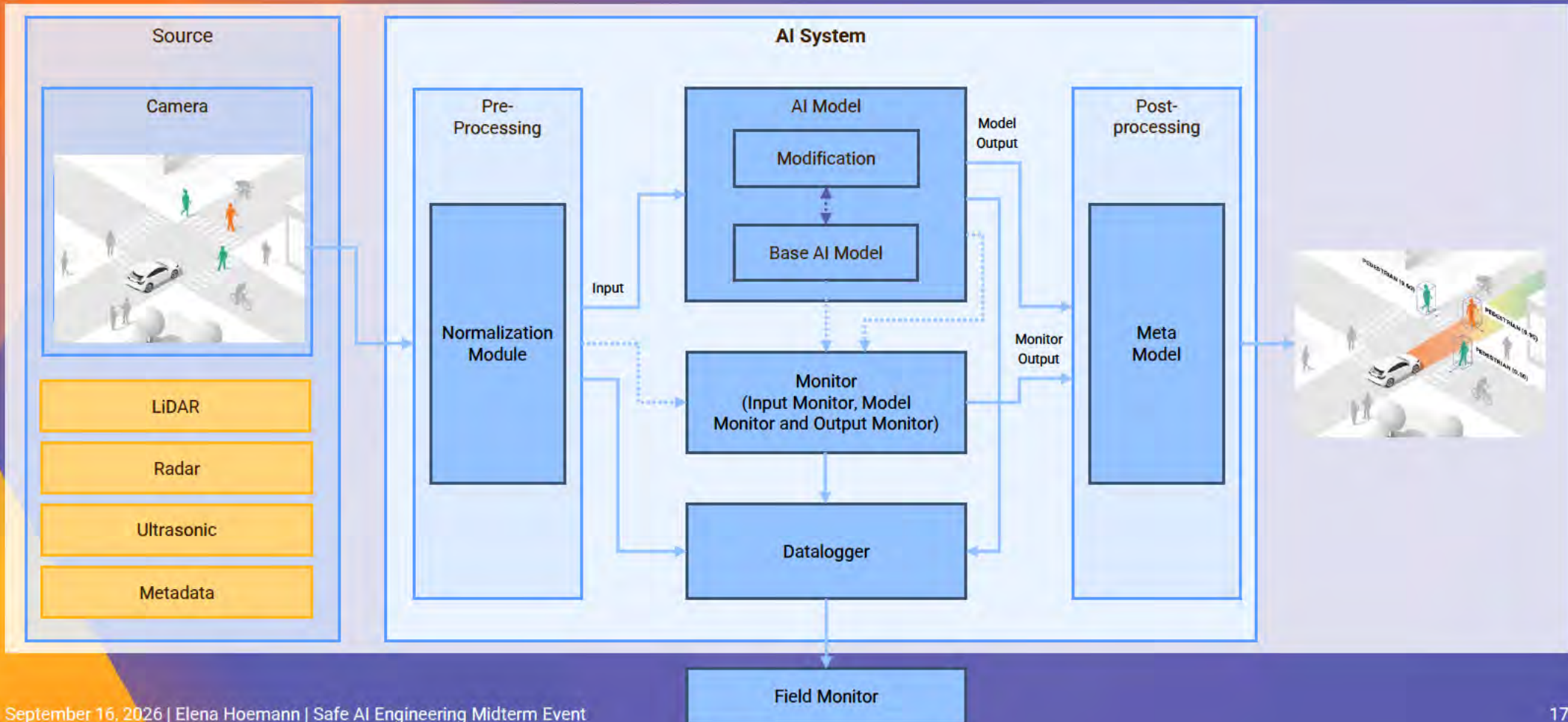


AI Development



Assurance Argument

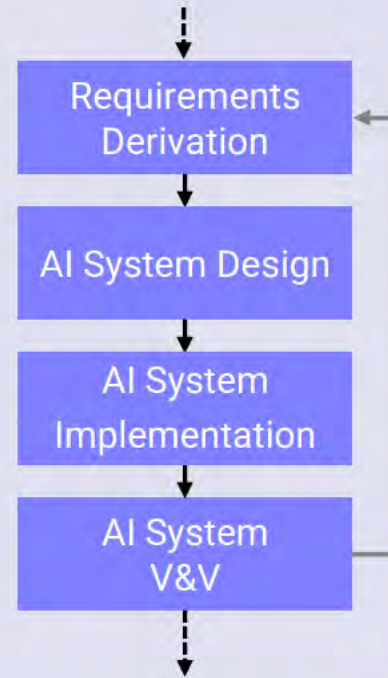
AI System Design



Process

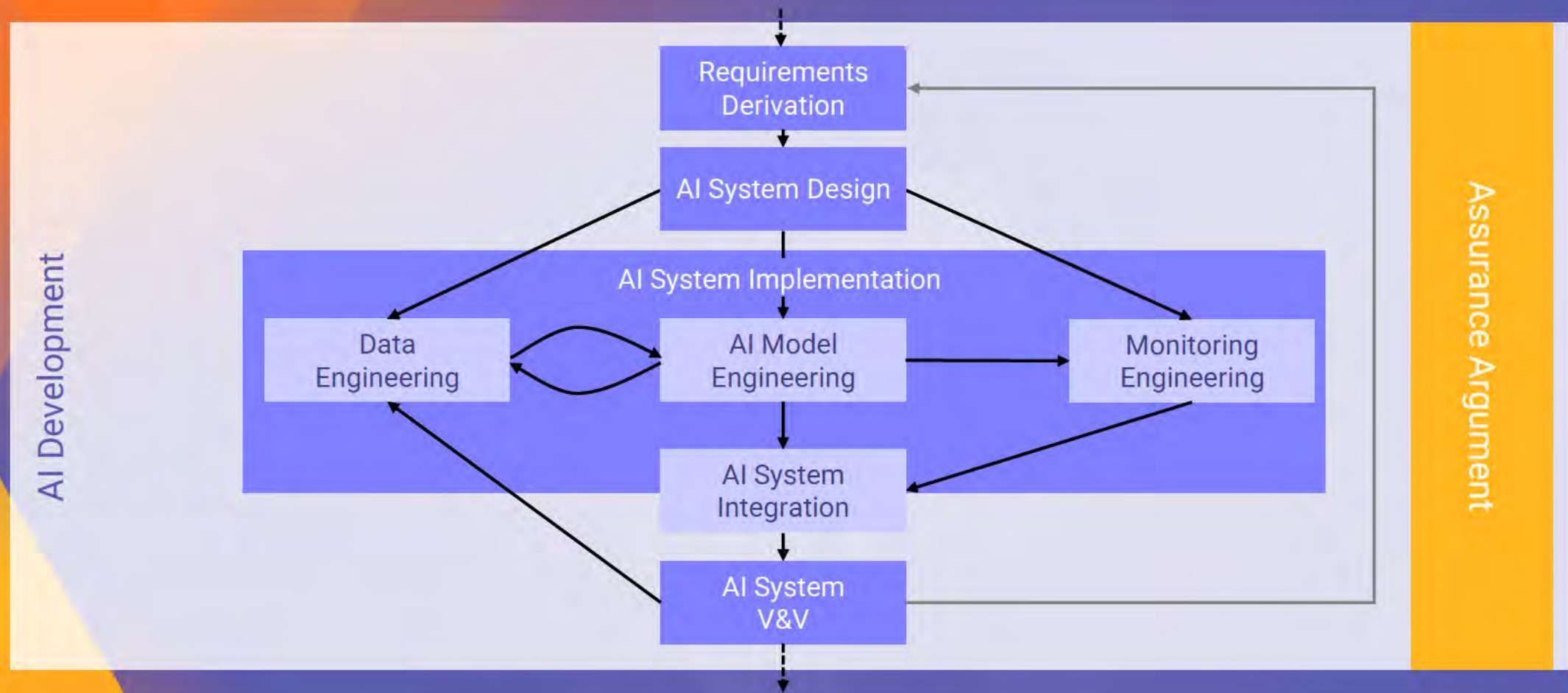


AI Development

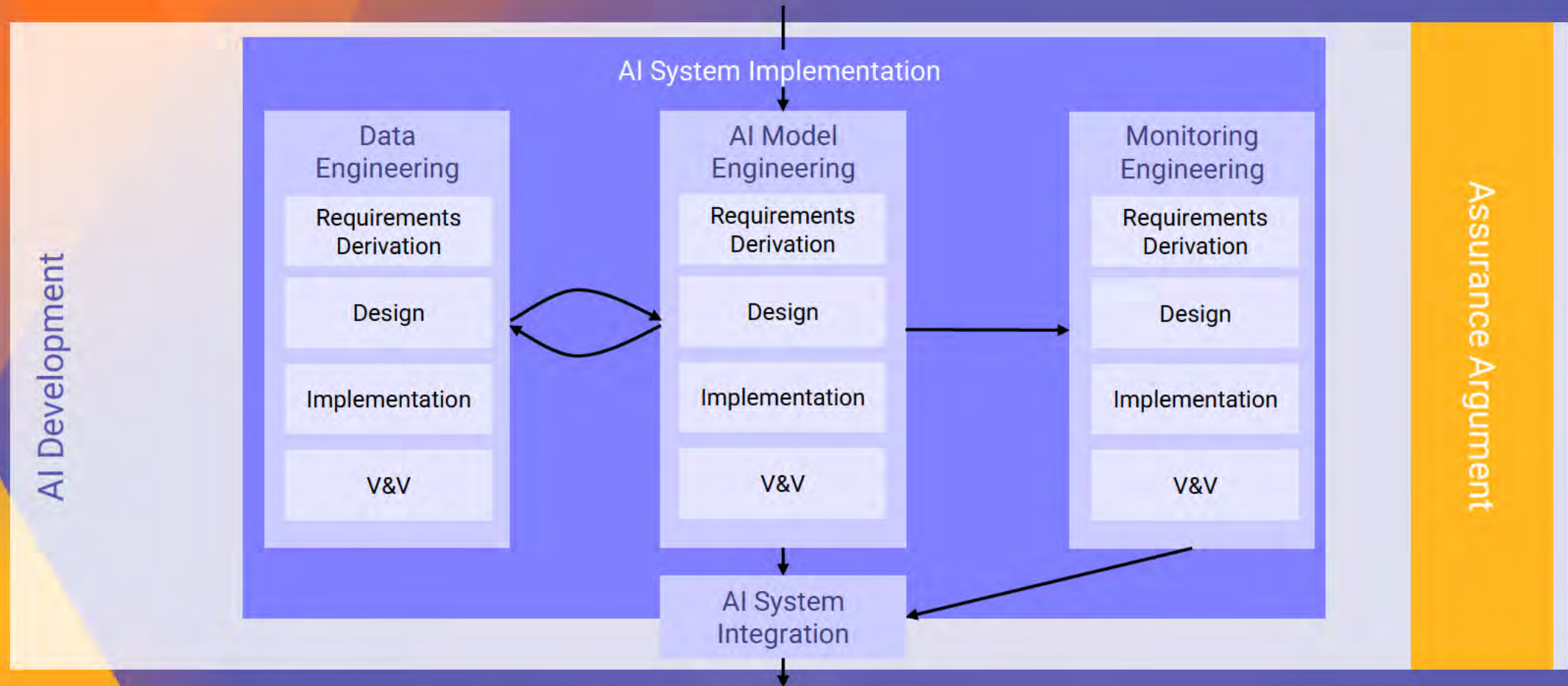


Assurance Argument

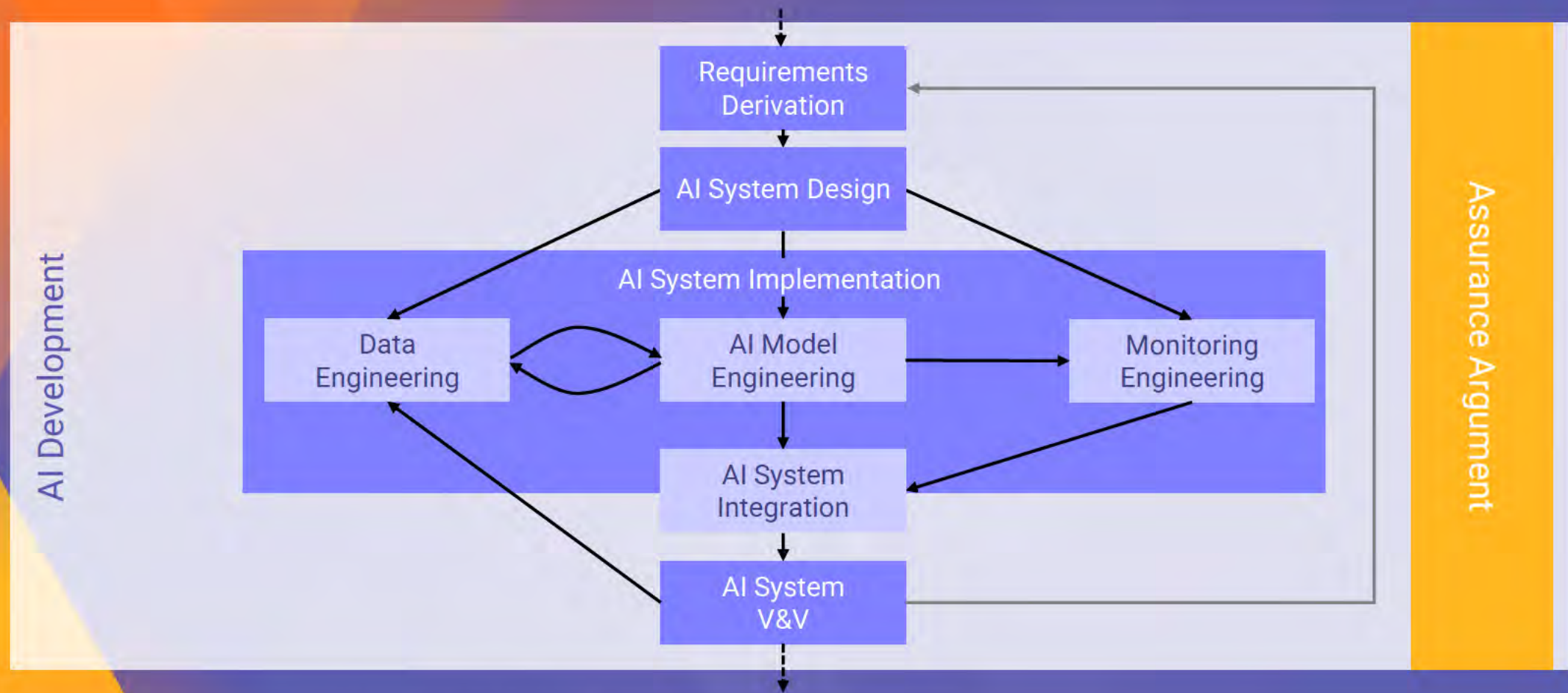
Process



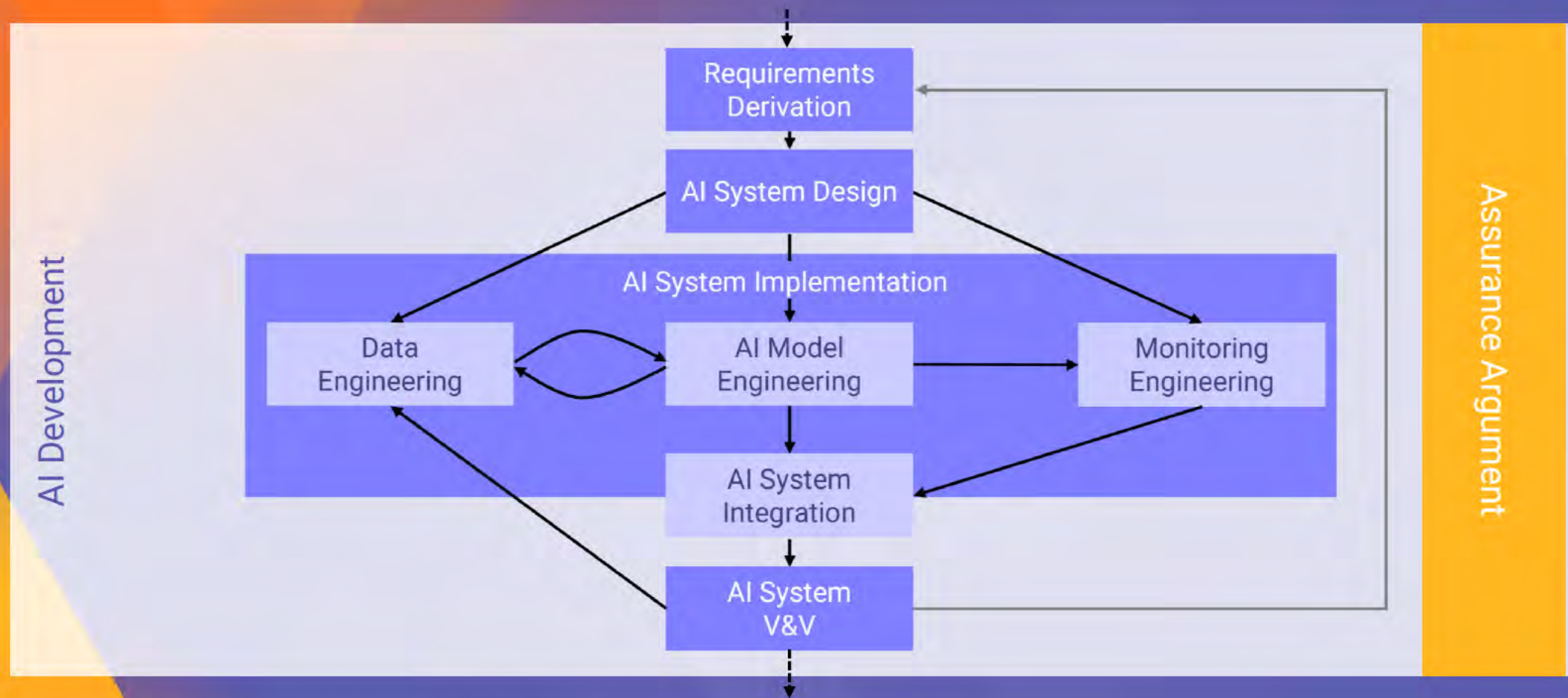
Process



Process

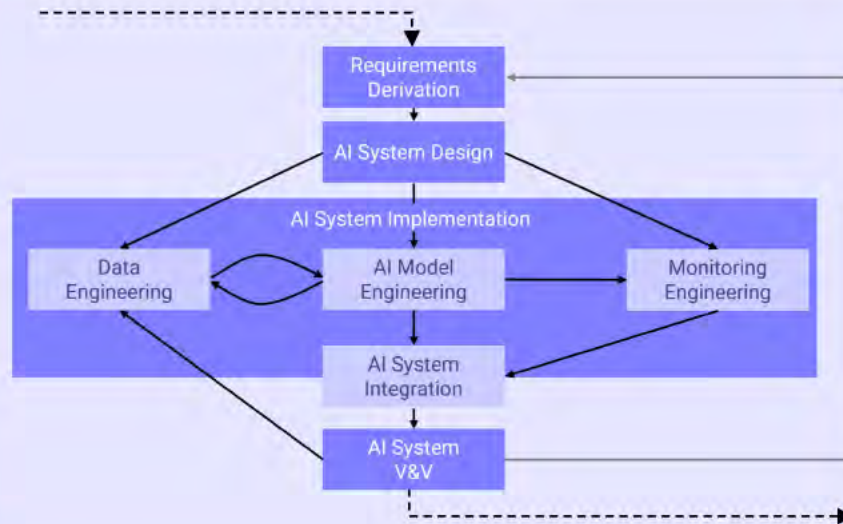


Process

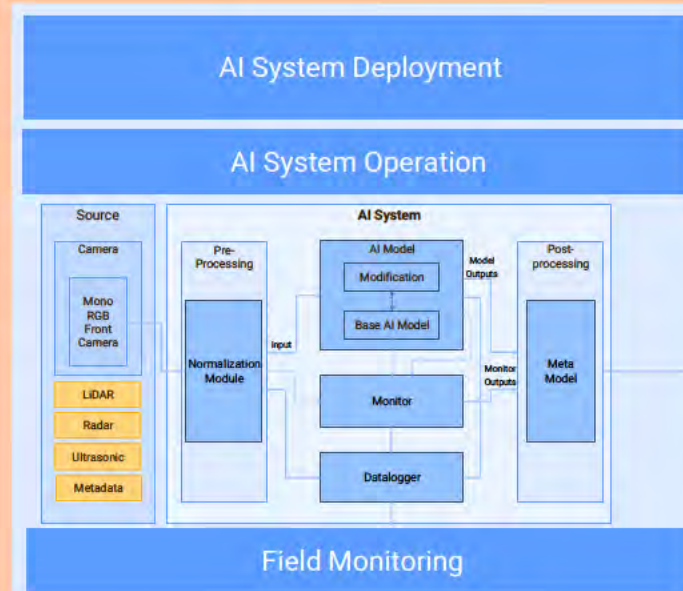




Assurance Argument



Continuous Assurance



System Requirements
and Specifications

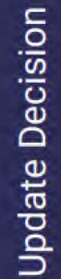
System
Verification & Validation

Update Decision

AI Engineering Process

System Requirements and Specifications

System Verification & Validation



Orchestration

System Requirements and Specifications



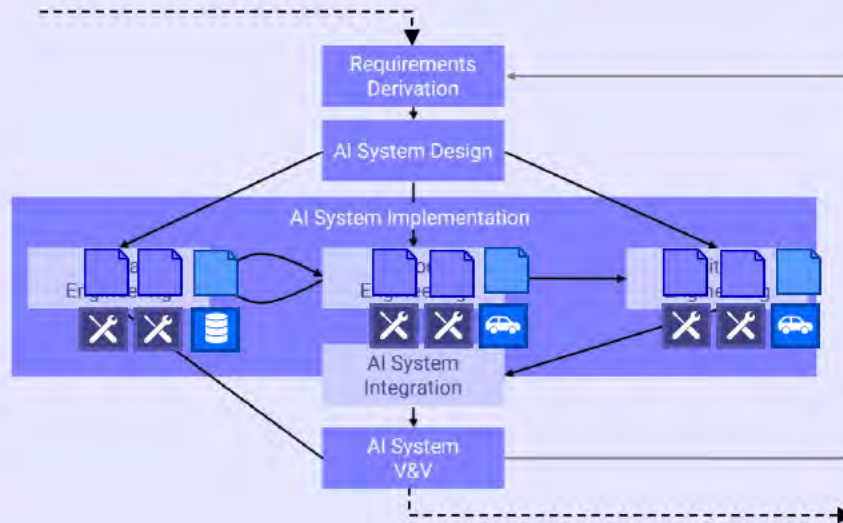
Update Decision





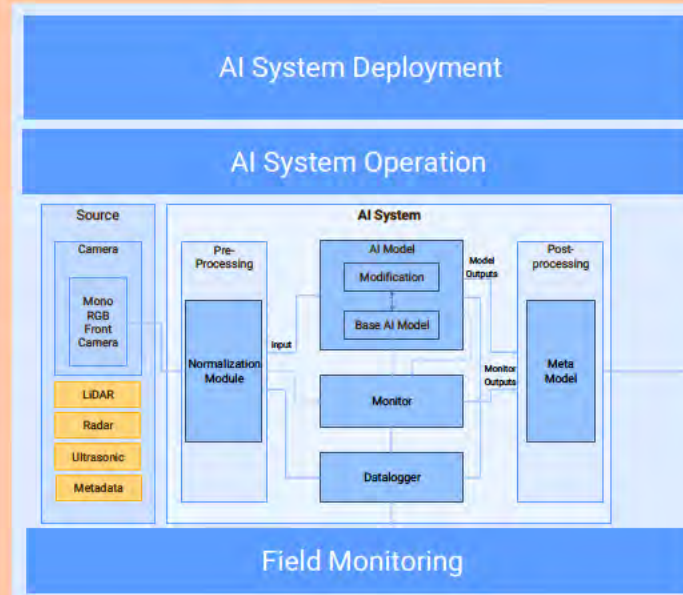
System Requirements
and Specifications

Assurance Argument



System
Verification & Validation

Continuous Assurance

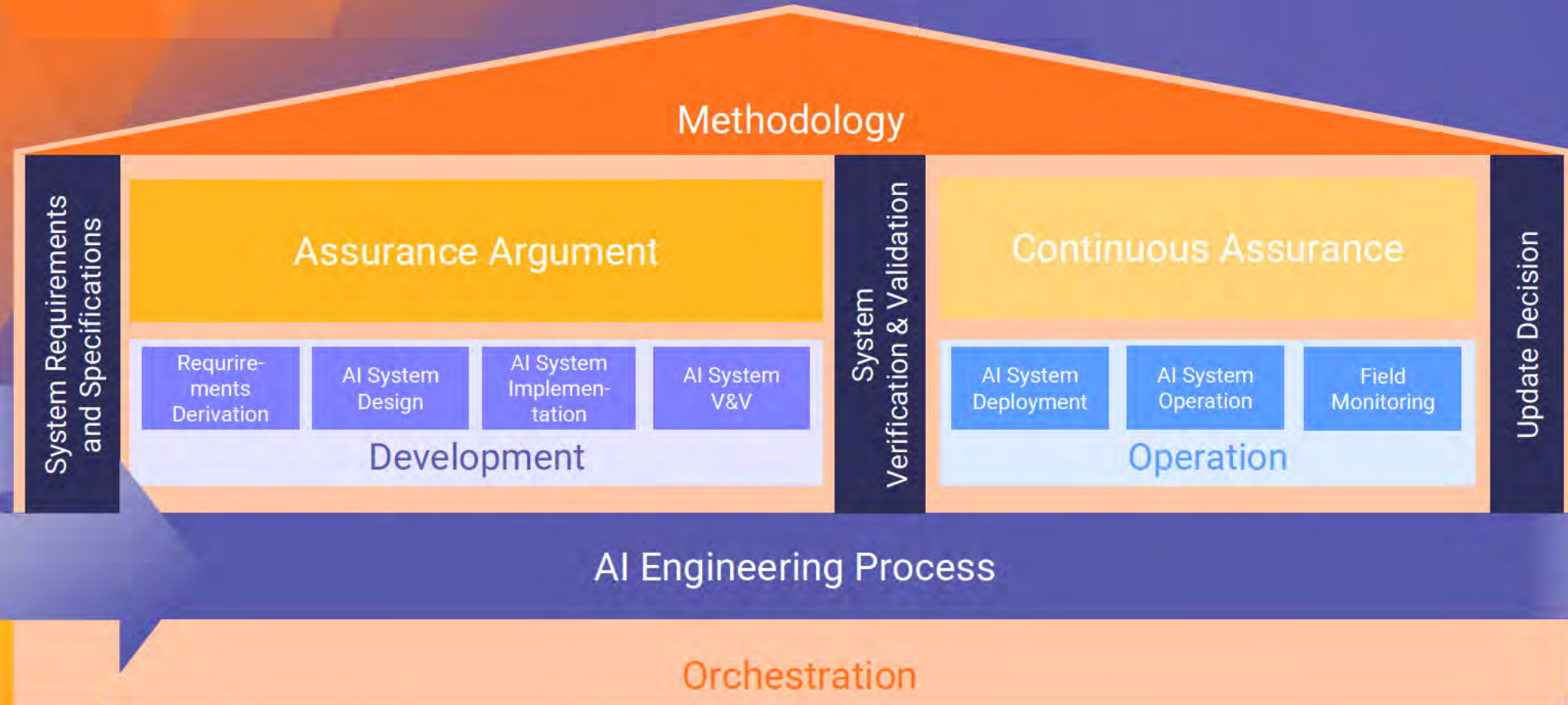


Update Decision

AI Engineering Process

Orchestration

Methodology Level 2



Braiding a Safety Rope



Braiding a Safety Rope





SAFE AI
ENGINEERING

September 16, 2026 | IAA Transportation

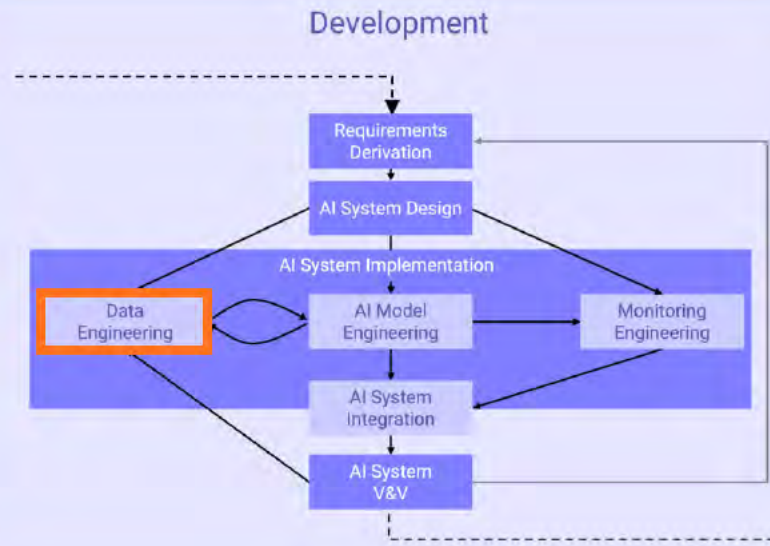
Data-Driven Engineering

Leonard Schroven | Capgemini
Yannick Kees | DLR



Safe AI Engineering Methodology

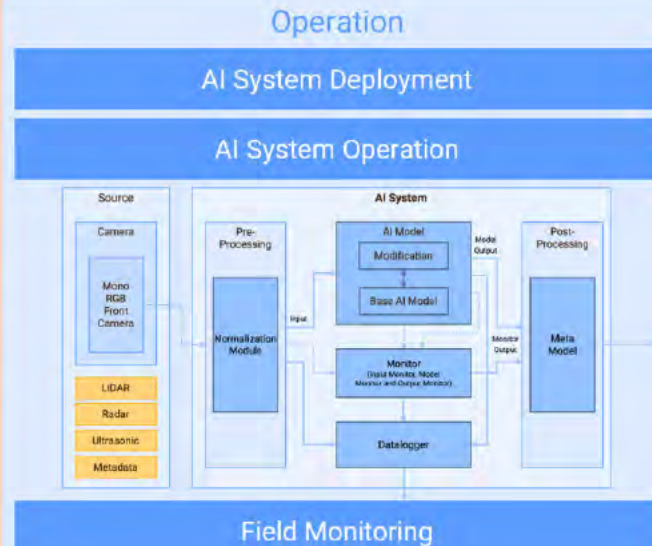
Assurance Argument



System Requirements and Specifications

System Verification & Validation

Continuous Assurance



Update Decision

AI Engineering Process

Orchestration

Data-Driven Engineering: Motivation



Data as the Foundation of the AI System

- Data determines what an AI system can learn.
- Quality of data limits system's quality
- Enables traceability and controllability

Data as the Basis for Safety Arguments

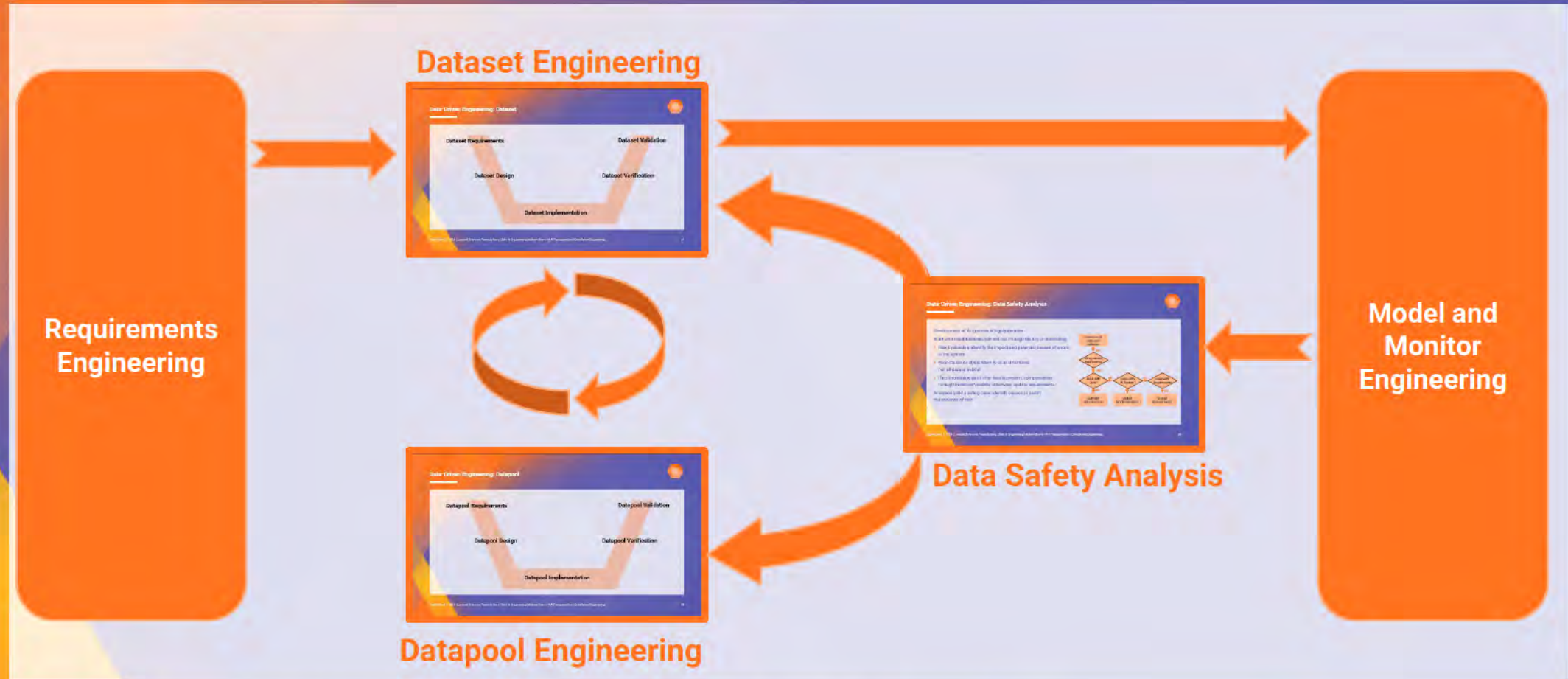
- Safety-related information must in the data.
- Uncovered scenarios lead to gaps in verification.
- Safety verification begins with data foundation.

Data Engineering as a Continuous Process

- Evaluate data coverage against requirements.
- Identify and close data gaps.
- Expand the data foundation through measurement, simulation, and synthesis.



Data-Driven Engineering: Methodology





Datapool Requirements

Datapool Validation

Datapool Design

Datapool Verification

Datapool Implementation



Datapool Requirements

Data-related Safety-Properties

- Accuracy
- Completeness
- ...

Datapool Validation

Datapool Design

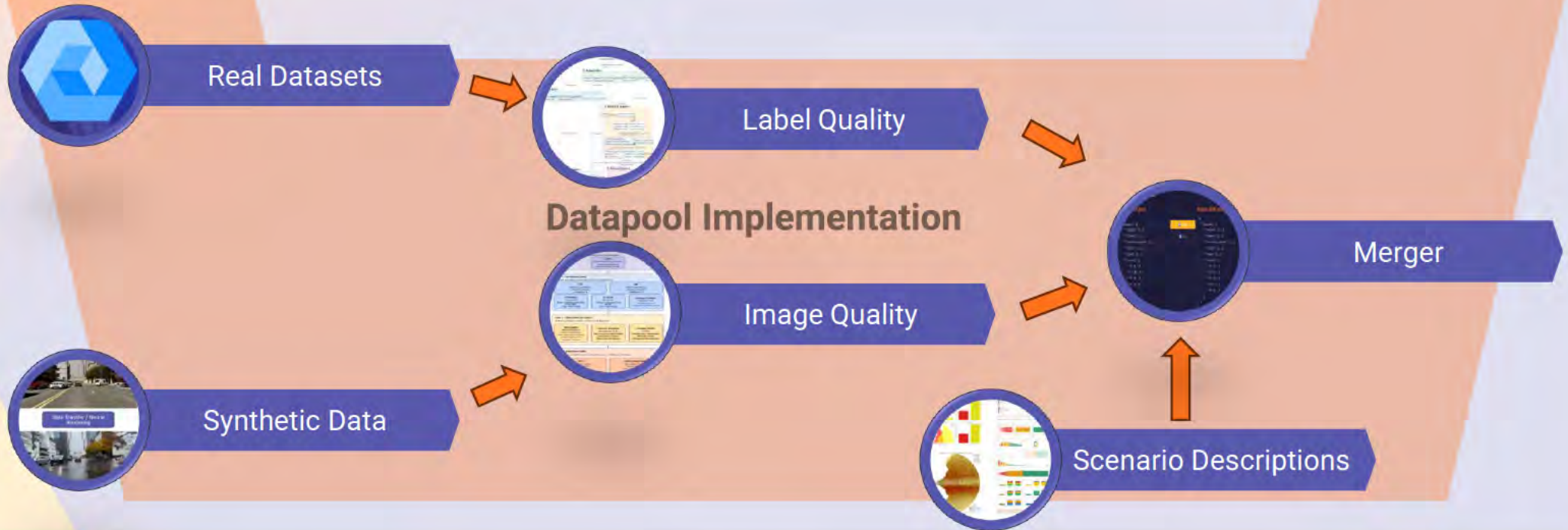
Datapool Verification

Datapool Implementation

Data-Driven Engineering: Datapool

Datapool Design

Datapool Verification





Dataset Requirements

Dataset Validation

Dataset Design

Dataset Verification

Dataset Implementation



Dataset Requirements

Data-related Safety-Properties

- Coverage
- Representativeness
- ...

Dataset Validation

Dataset Design

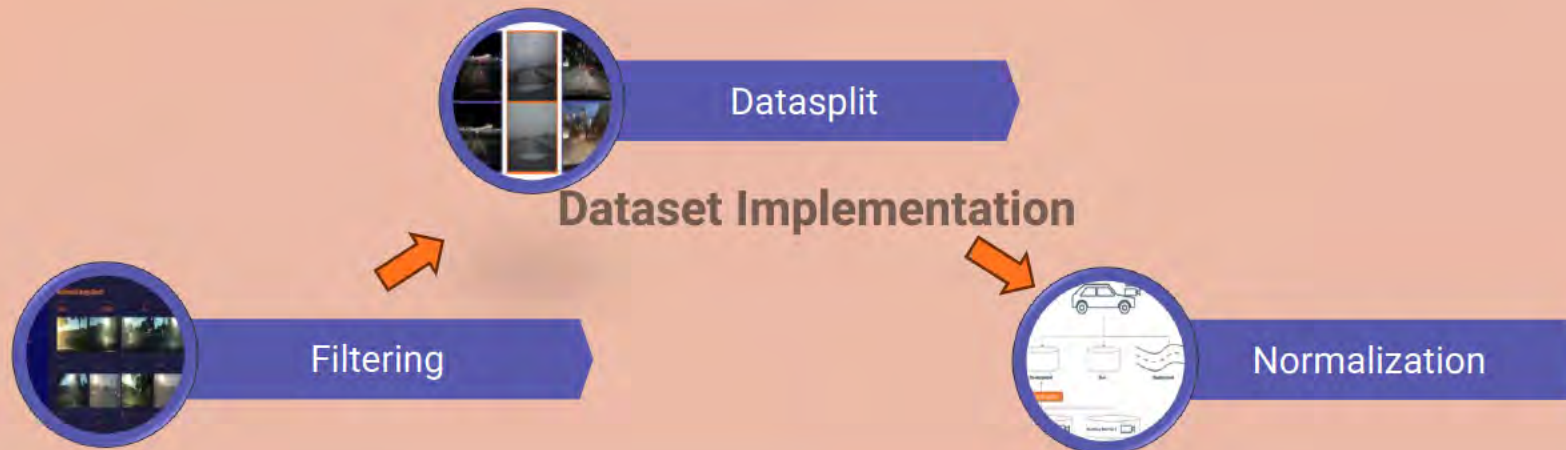
Dataset Verification

Dataset Implementation

Data-Driven Engineering: Dataset

Dataset Design

Dataset Verification





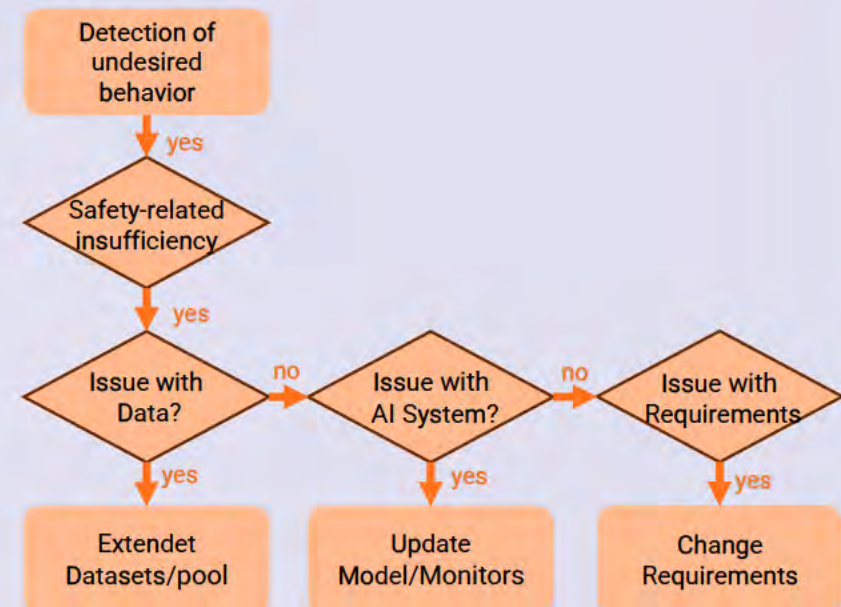
Data-Driven Engineering: Data Safety Analysis

Development of AI systems is highly iterative.

Start with insufficiencies pointed out through testing or monitoring

- Risk Evaluation: identify the impact and potential causes of errors in the system
- Root-Cause Analysis: Identify crucial features; not all data is helpful
- If no knowledge gap in the data is present, compensation through monitors/models; otherwise, update requirements.

Analyses build a safety case: identify causes to justify the absence of risk



Braiding a Safety Rope



Braiding a Safety Rope





SAFE AI
ENGINEERING

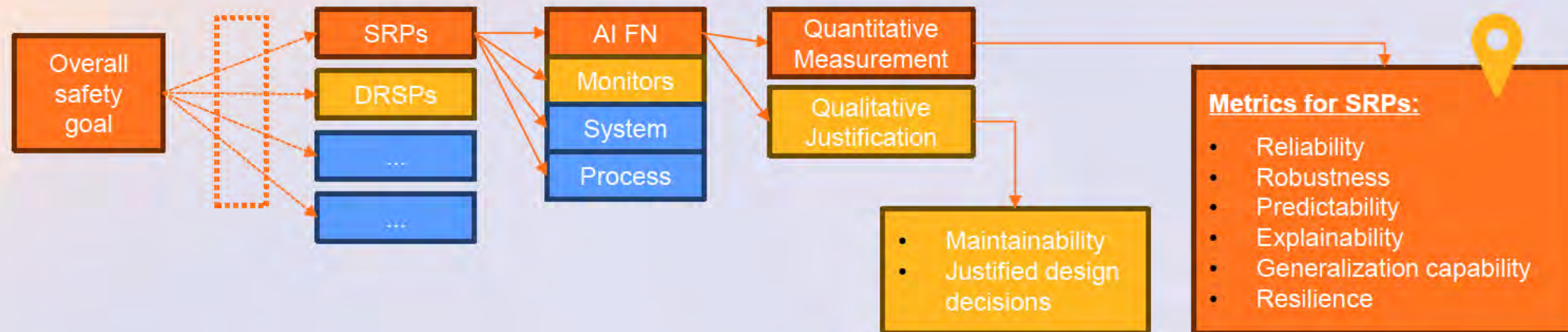
September 16, 2026 | IAA Transportation

Safety Metrics Evaluation Pipeline

Nenad Ocelić | DXC Luxoft



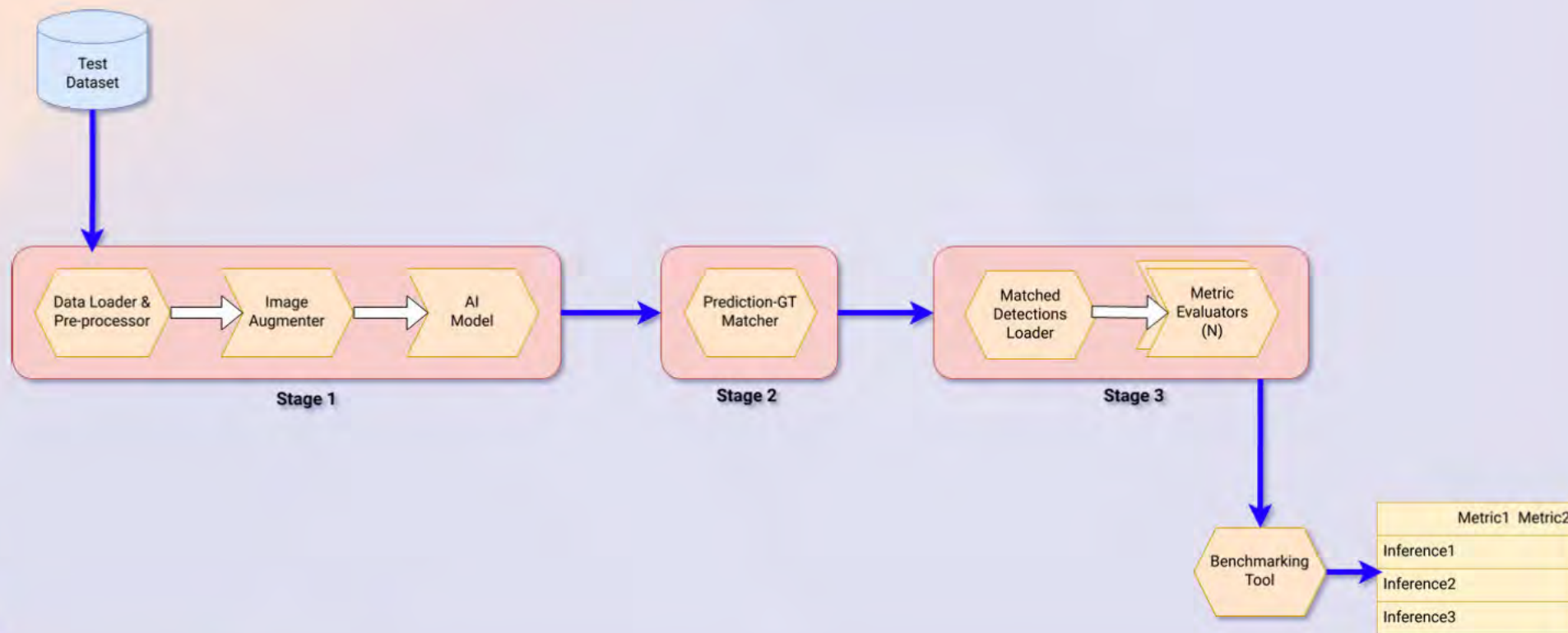
Safety-Related Properties and Metrics



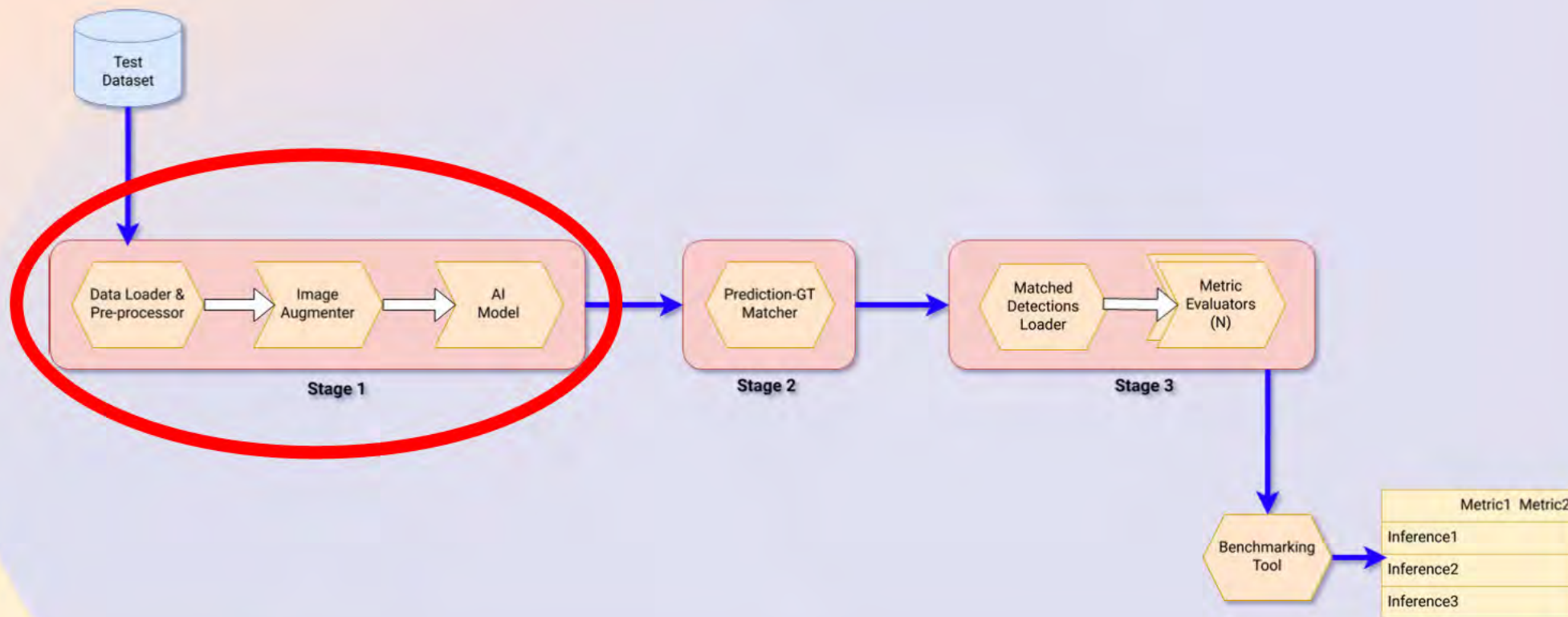
Example Metrics:

- Reliability: FP/FN Error Rate of object classification, Location RMSD vs. GT
- Robustness: The same, but after image perturbation (Noise, Blur, Weather, Digital) vs. before perturbation

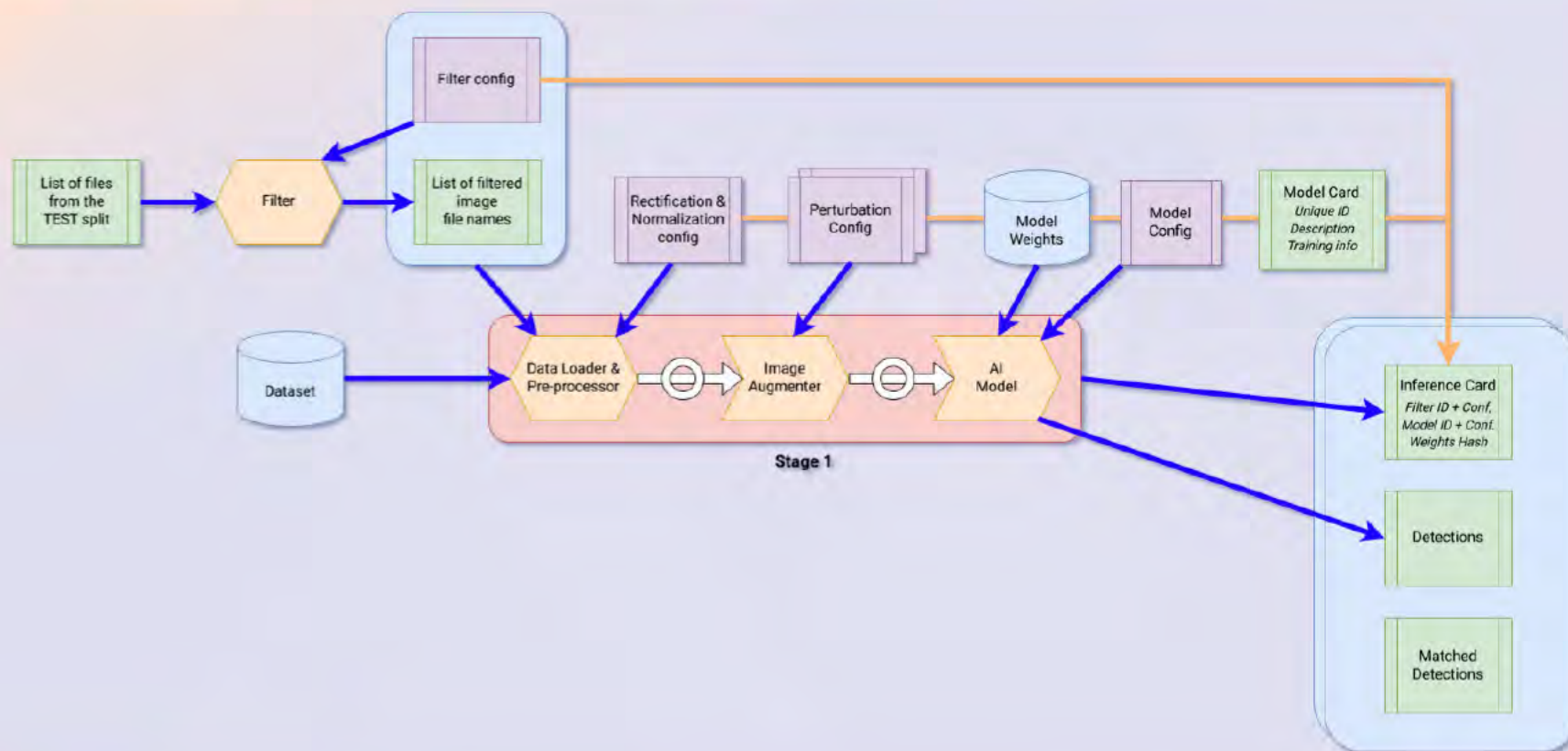
Metrics Evaluation Pipeline



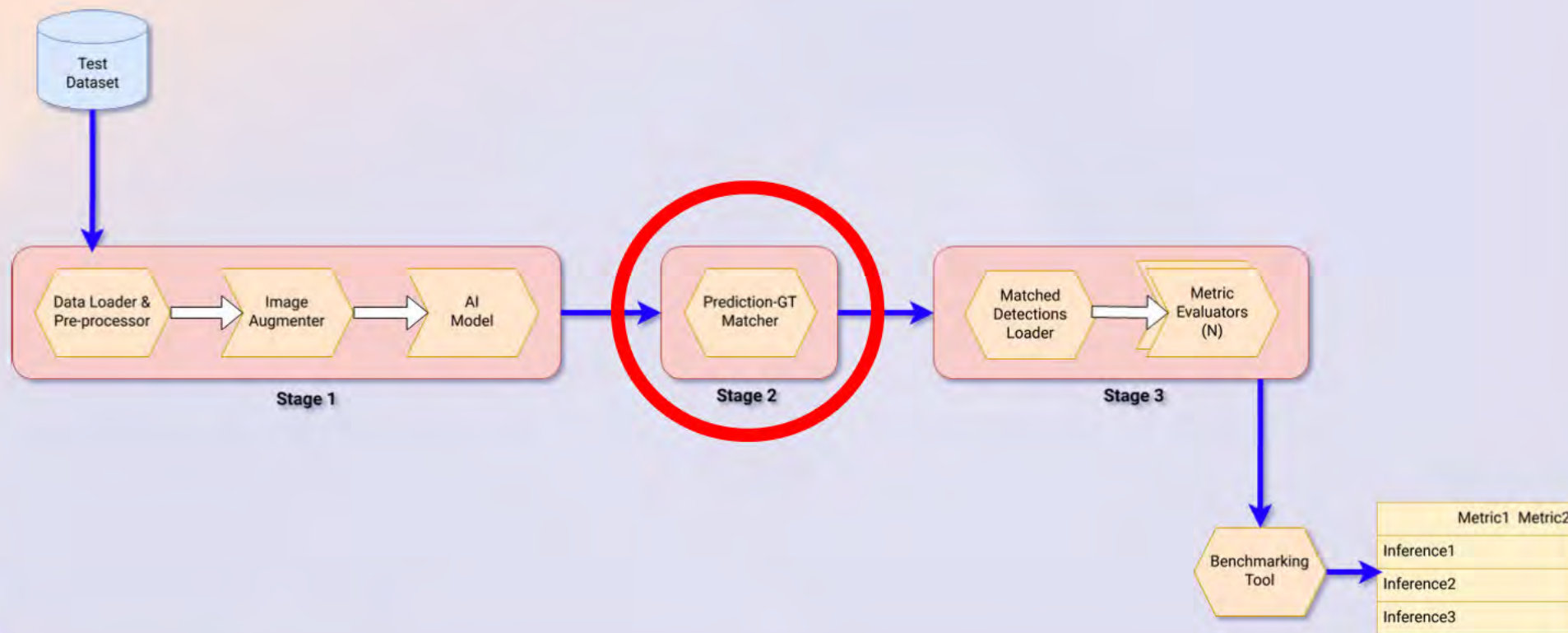
Metrics Evaluation Pipeline - Stage 1



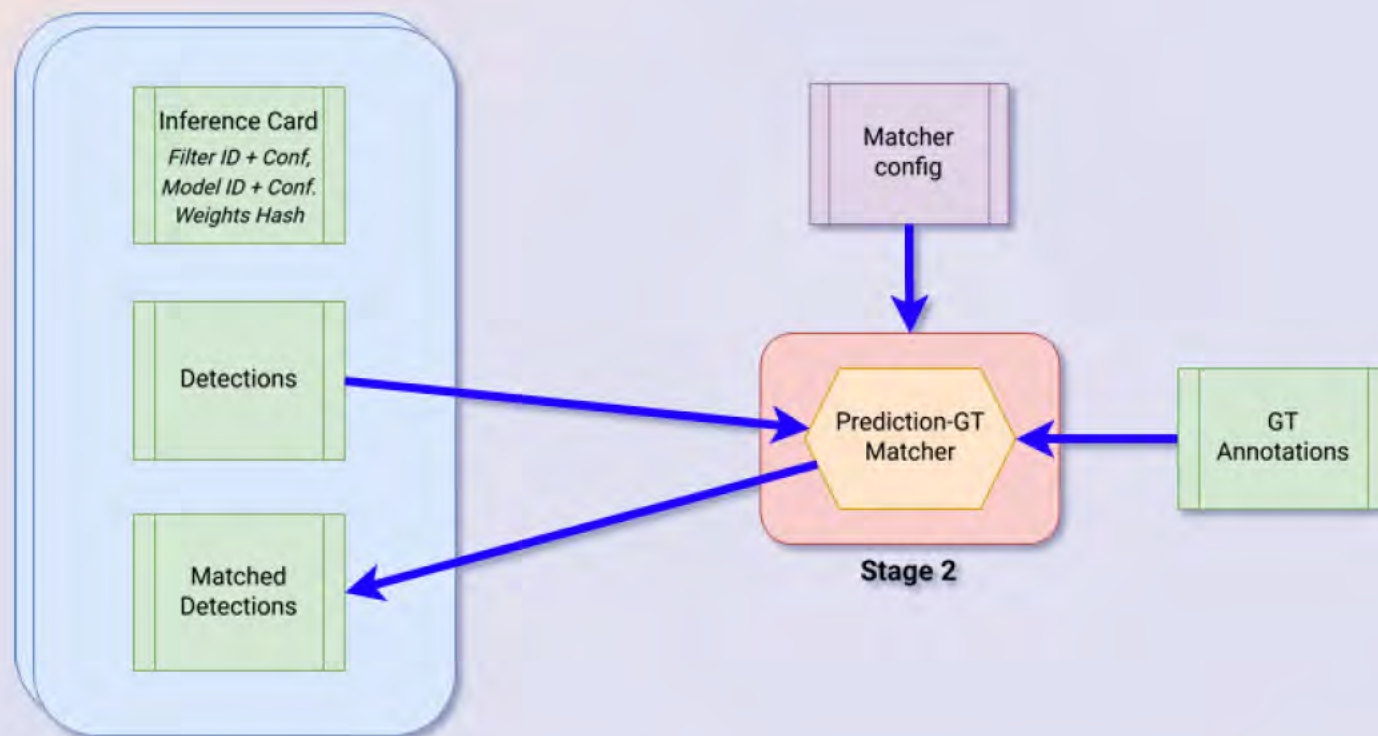
Stage 1: Data Loader, Image Augmenter, Object Detector



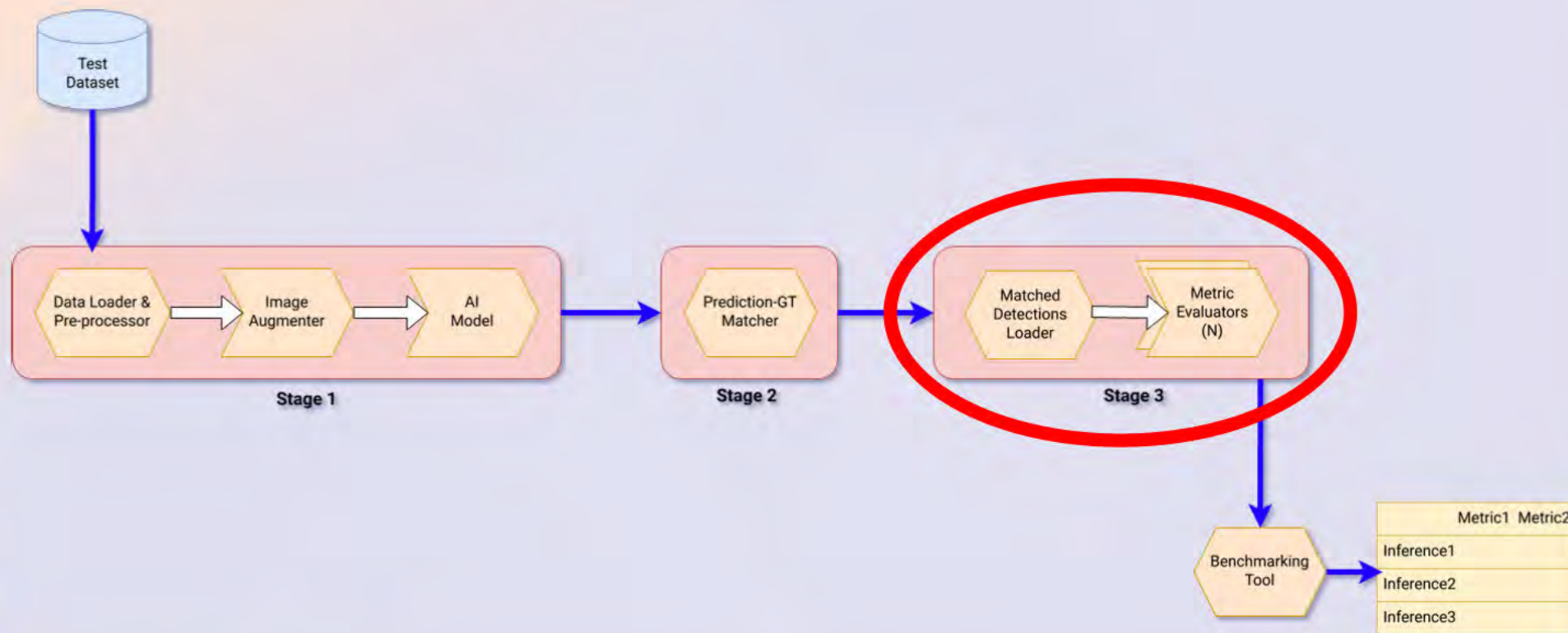
Metrics Evaluation Pipeline – Stage 2



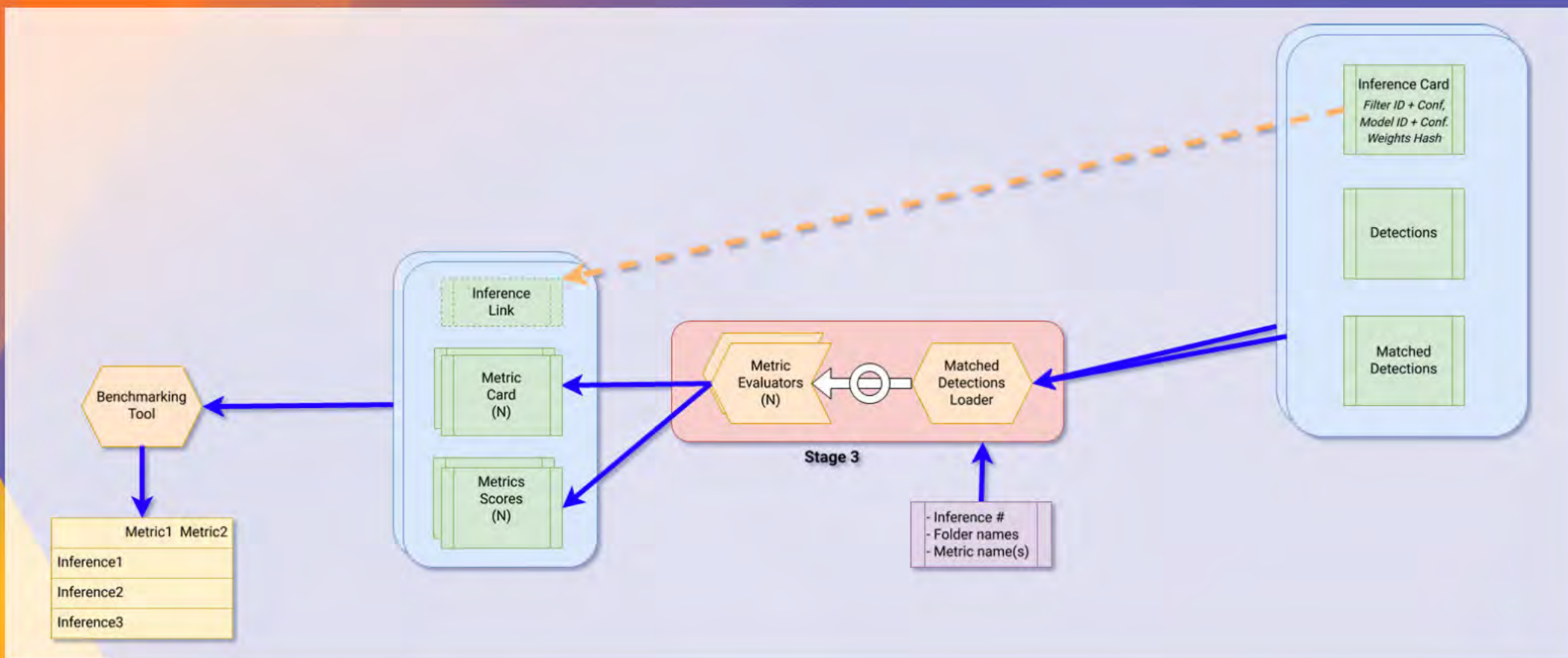
Stage 2: Predictions to Ground Truth Matcher



Metrics Evaluation Pipeline – Stage 3



Stage 3: Metrics Evaluator





SAFE AI
ENGINEERING

September 16, 2026 | IAA Transportation

Monitoring

Antje Loyal | Aumovio





When your AI says "trust me", what evidence do you have?

Monitoring



So far, we looked at:

- **Data** used for training, validating and testing
- How to evaluate the **data quality** and consistency
- How to select and **design our AI model**
- How to evaluate and **benchmark our AI model**



Monitoring – The Core Idea



So far, we looked at:

- **Data** used for training, validating and testing
- How to evaluate the **data quality** and consistency
- How to select and **design our AI model**
- How to evaluate and **benchmark** our AI model

Now:

- The component that asks, "Are you sure?"



Scenarios for Monitoring in Action



Known Scenarios

- Scenarios are well represented in training and validation data
- Model confidence is sufficient to confirm operation within the expected domain
- Monitoring verifies that the system remains in familiar conditions.

Question answered by Monitor:

"Am I operating in a scenario I already know?"



Scenarios for Monitoring in Action



Anticipated but Hard-to-Cover Scenarios

- Infinite variations exist (weather, lighting, occlusion, crowd density)
- Complete coverage during development is impossible
- Uncertainty estimates indicate whether the prediction remains reliable

Question answered by Monitor:

"Can I still trust this prediction despite increased ambiguity?"



Scenarios for Monitoring in Action

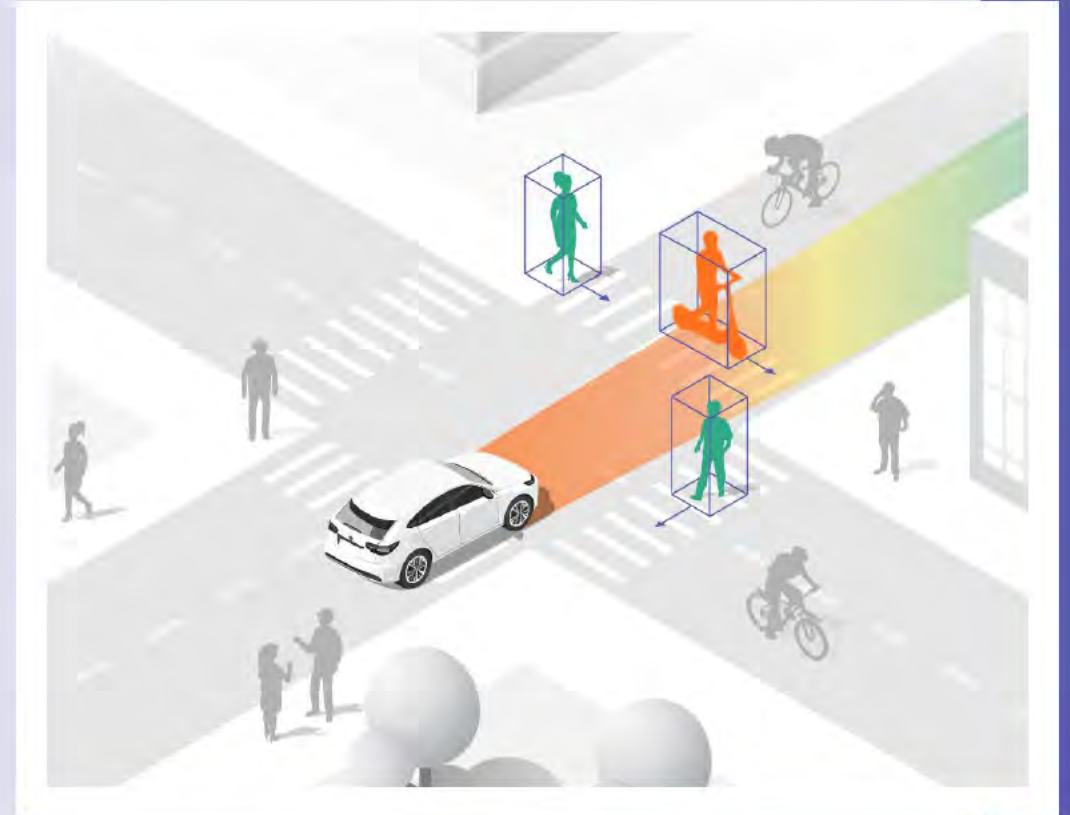


Evolving Environment Over Time

- New traffic participants and infrastructure emerge gradually
- Individual events may not be safety-critical
- Long-term monitoring identifies changes in the operational environment

Question answered by Monitor:

"Has the real world changed enough that the model needs adaptation?"



Scenarios for Monitoring in Action



Unknown & Unexpected Scenarios

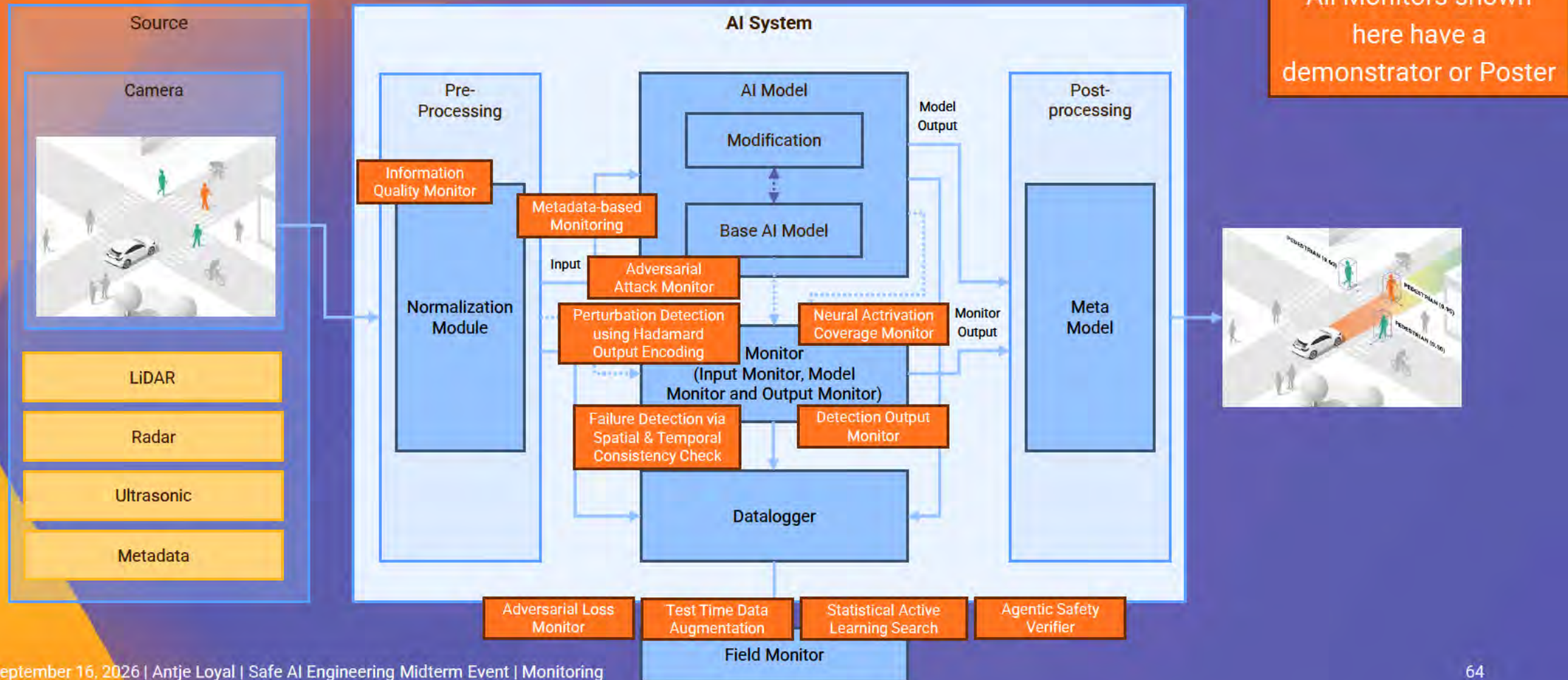
- Scenarios not considered during development
- Unexpected objects, behaviors, or environmental conditions
- The monitor detects that the model is outside its validated domain

Question answered by Monitor:

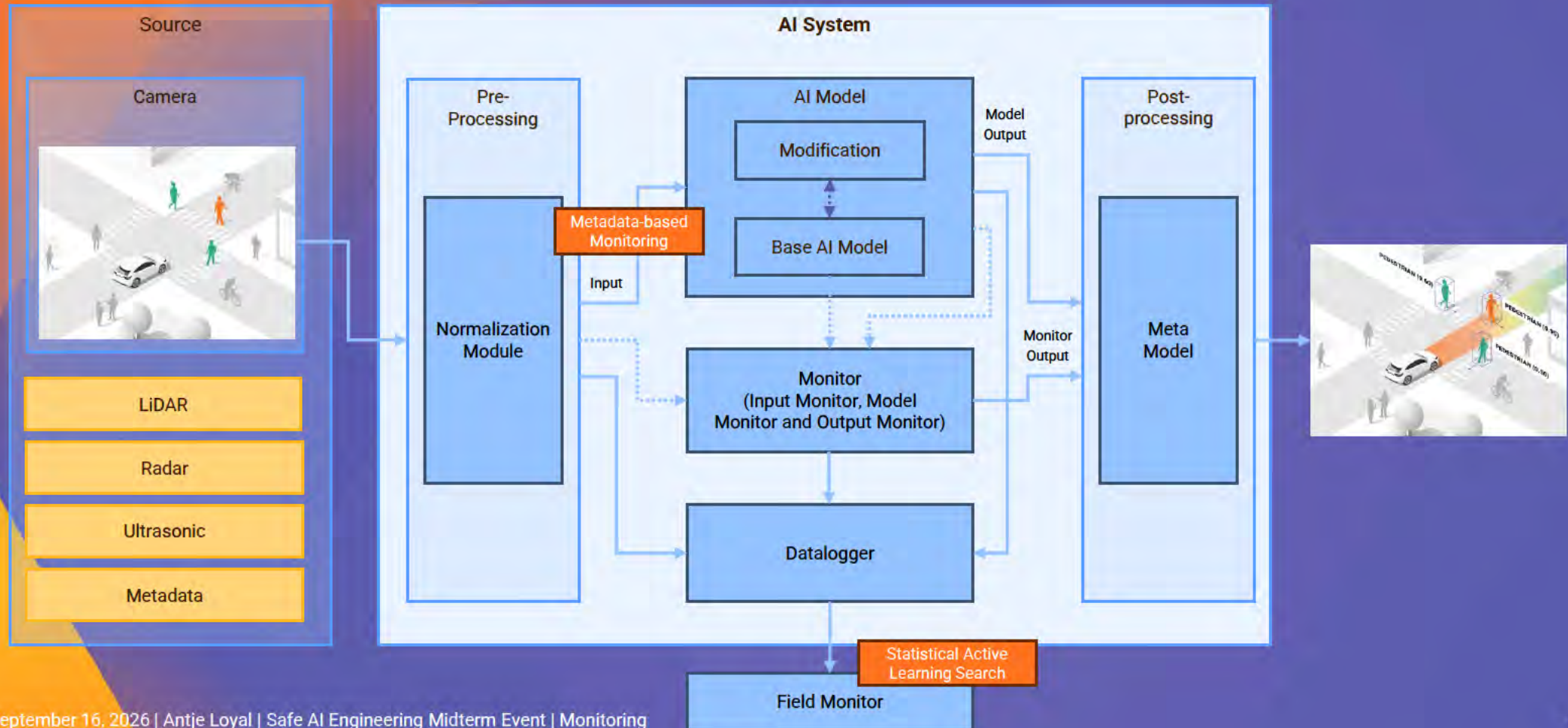
"Is this situation unknown and should I mistrust the model output?"



Architecture and Monitors



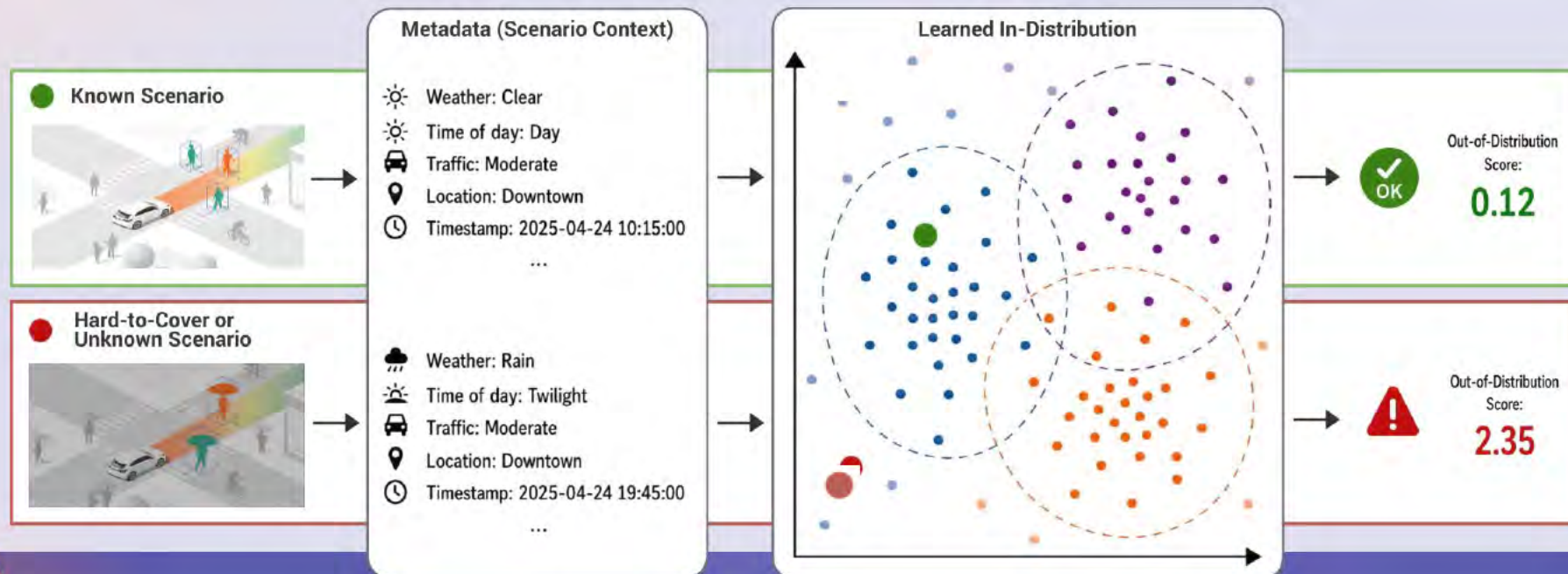
Architecture and Monitors



Monitor Example: Metadata-based Monitoring



- Rich contextual metadata provides valuable information beyond the camera image
- The monitor analyzes this data to detect unusual conditions and identify operation outside the validated domain



Field Monitor Example: Statistical Active Search Monitor



Key Challenge

- More data improves accuracy but also increases the search and analysis space.
- This leads to the curse of dimensionality, making comprehensive exploration difficult.

Our Approach

- Avoid uniform exploration of the entire search space.
- Use guided exploration based on results from previous iterations.
- Focus on identified weak spots and failure modes.
- Curate targeted test datasets to zoom in on high-risk areas.

Result

- Clear identification of failure modes.
- More accurate and defensible safety statements.

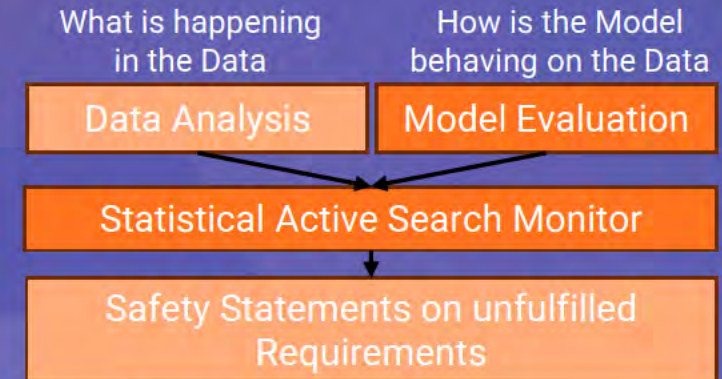


Figure 1: Process Diagram

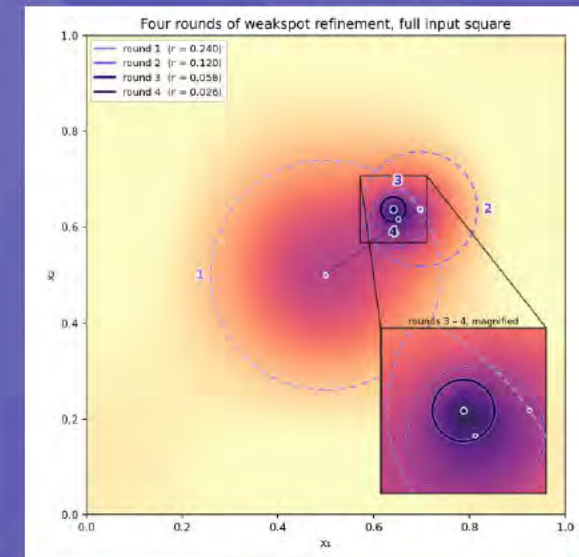


Figure 2: Example of Weakspot Drift and Reduction over Time

Braiding a Safety Rope



Braiding a Safety Rope





SAFE AI
ENGINEERING

September 16, 2026 | Safe AI Engineering

Hybrid Safety Concept

Andreas Rohatschek Striebel | Robert Bosch GmbH



The Challenge

How can we assure the safety of AI functions?



AI is **probabilistic**.



Safety is **not**.



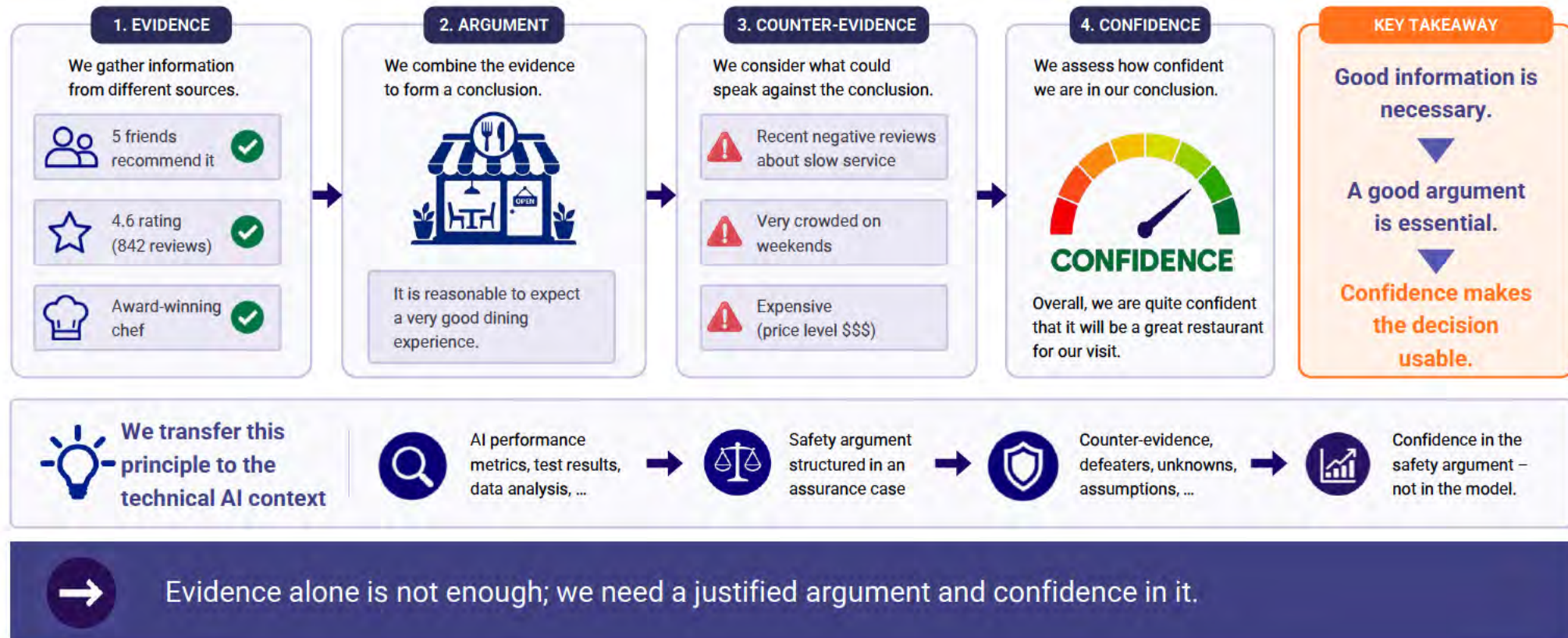
Traditional metrics alone are **not enough**.



A model can achieve very good performance –
and we can still have **insufficient confidence** that the safety claim is actually justified.

Decision Making under Uncertainty

A little Argumentation Story about finding a good Restaurant



The Conceptual Shift

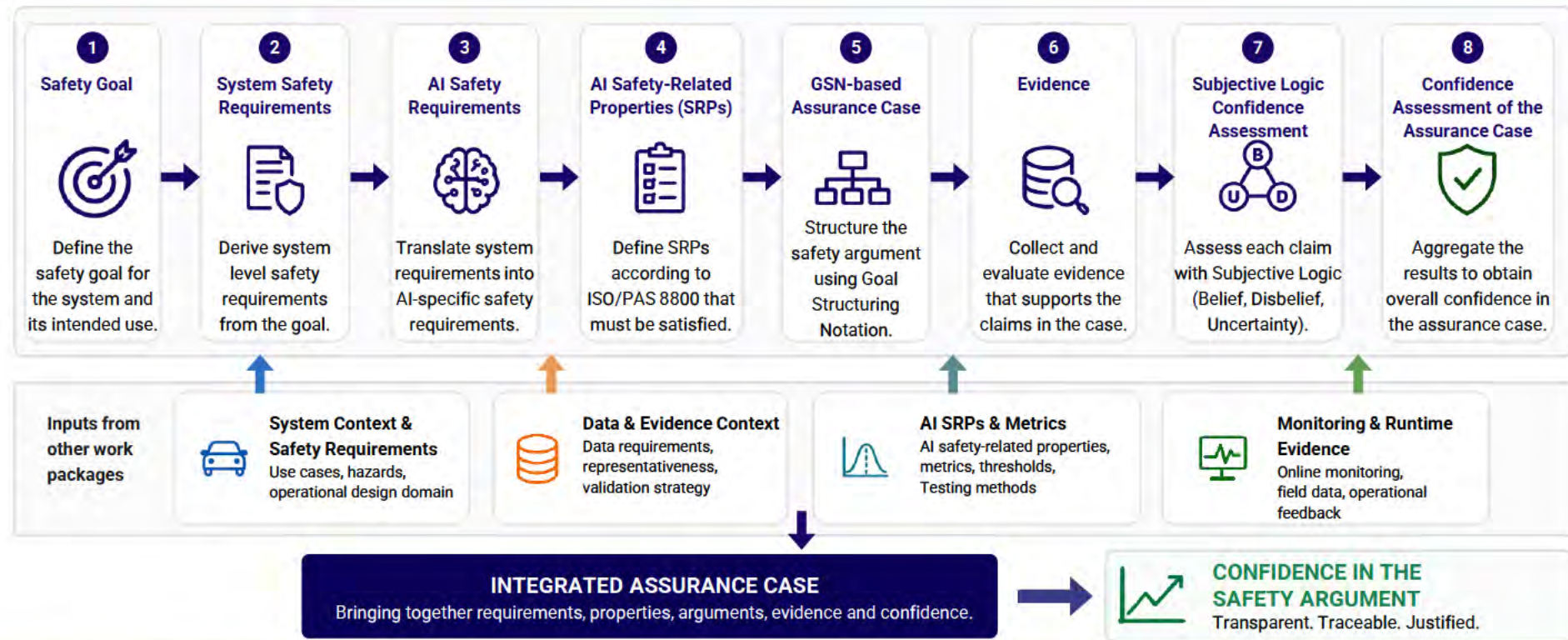
From confidence in the AI model to confidence in the safety argument



Our goal is not to prove the AI model is perfect.
Our goal is to build **confidence in the safety argument.**

The Hybrid Safety Concept

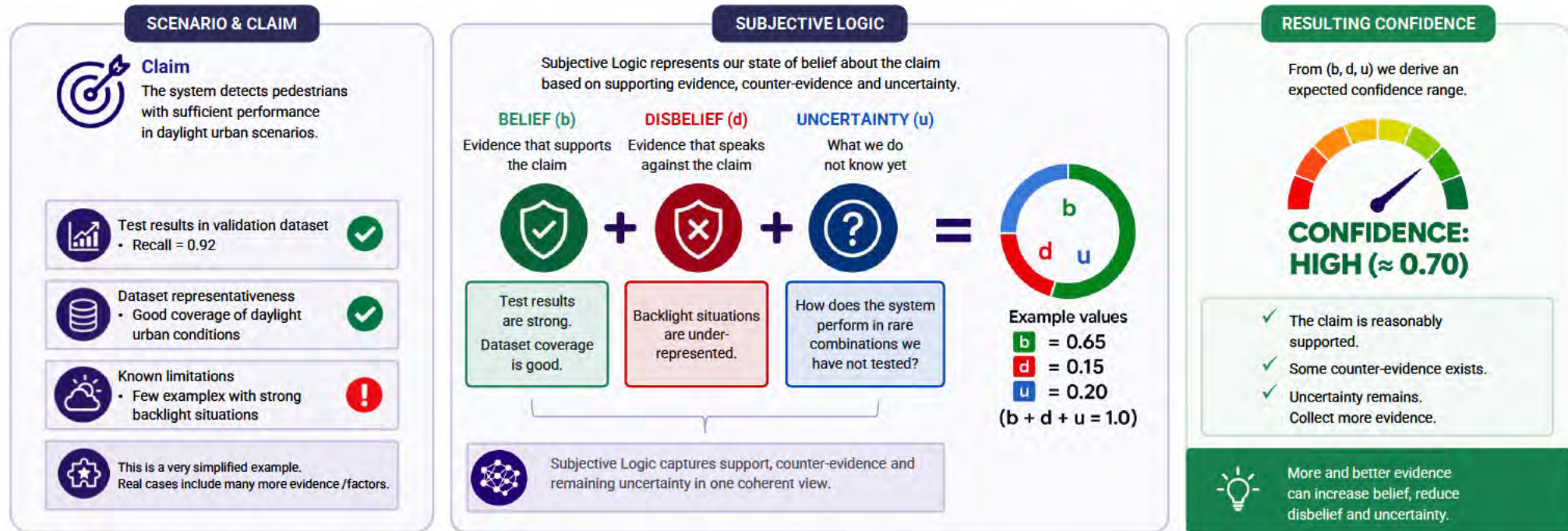
A structured path from Safety Goal to Confidence Assessment



Every step closes a different gap; together they build confidence that the system is sufficiently safe.

A Simple Worked Example

Confidence is built from evidence – and limited by what remains unknown.

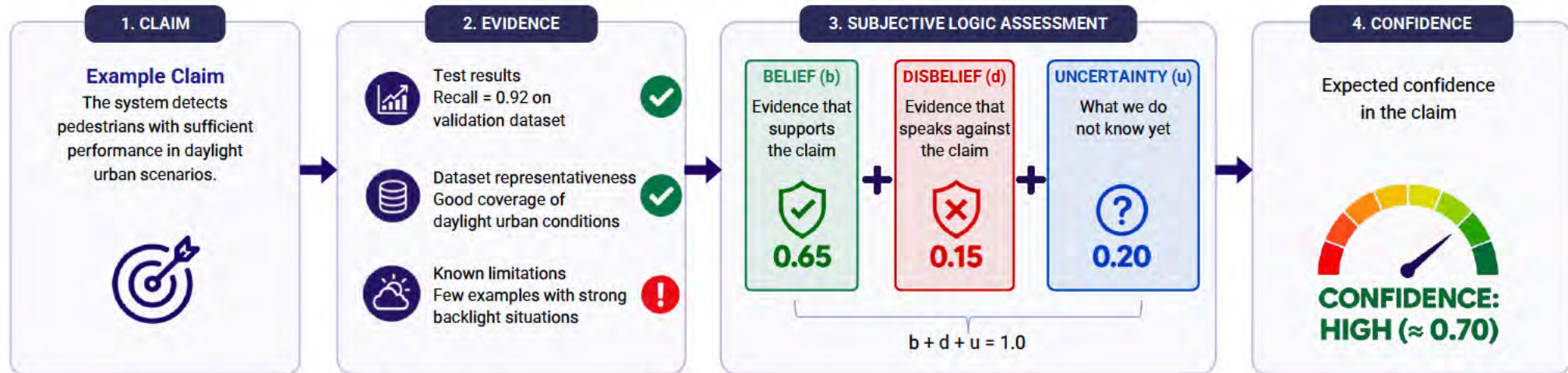


This simple example shows:

Evidence **supports** a claim, but confidence depends on what speaks **against** it, and what we still **do not know**.

From Evidence to Confidence

Subjective Logic makes support, counter-evidence and uncertainty explicit.



Key idea

Confidence is not just about how much evidence we have, but about the balance between support, counter-evidence and remaining uncertainty.



More support (b) increases confidence.



More counter-evidence (d) decreases confidence.



More uncertainty (u) limits confidence.



Subjective Logic gives us a **transparent** and **quantifiable** way to reason about **confidence** in a safety claim.



This is the foundation of our confidence assessment.

Why “Hybrid”? – Integrating Multiple Disciplines

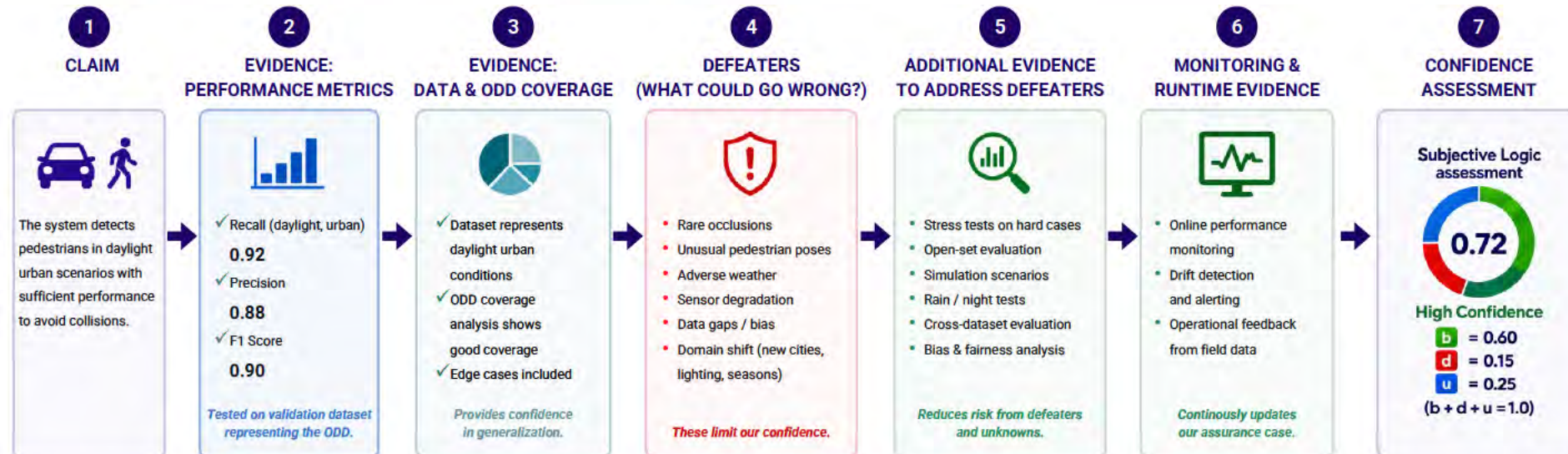
We bring together complementary contributions from across the project.



Hybrid means: combining strengths, covering different aspects, building one convincing assurance case. Different inputs. One approach. **One confidence.**

One Real Safety Claim

A good metric alone does not establish a convincing safety claim.



A metric is just one piece of evidence.

Confidence in the claim emerges from the balance of support (b), counter-evidence (d) and uncertainty (u).



A strong safety claim is well-supported, addresses its weaknesses, and acknowledges what remains unknown.

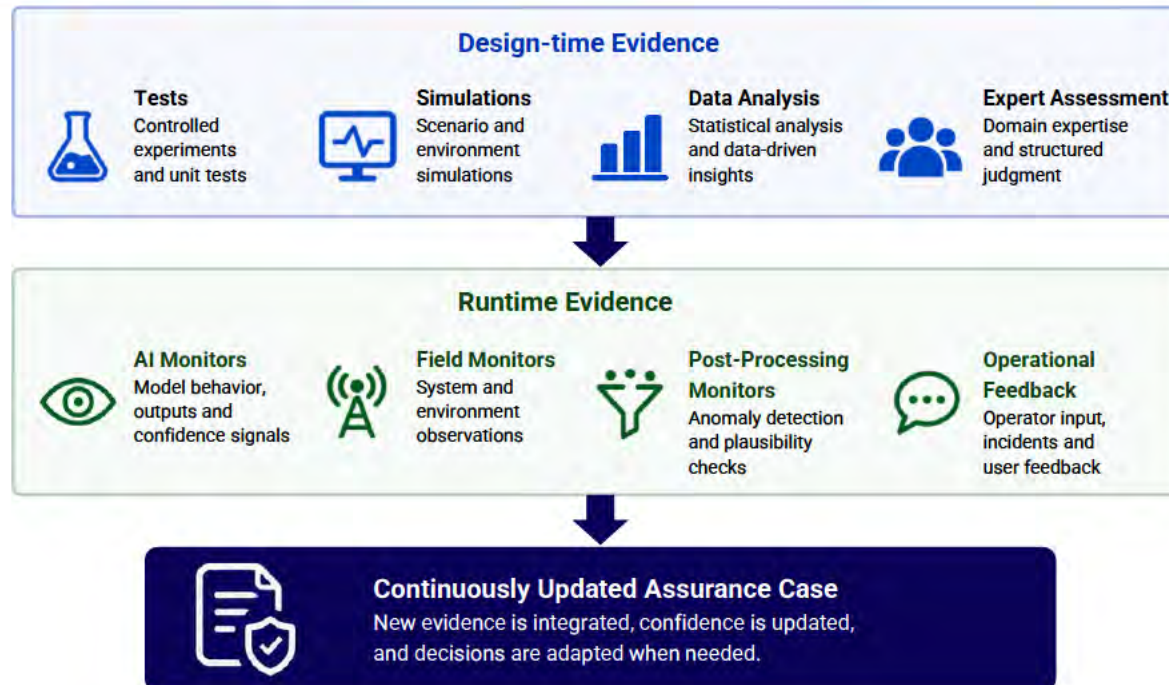


Take-away: Good performance is necessary, but not sufficient.

Convincing safety requires evidence, context, and explicit handling of what could go wrong.

Continuous Assurance

Safety assurance continues after deployment.



Why Continuous Assurance Matters

- Reality evolves**
Systems, data, and environments change over time. (Icon: Refresh)
- New evidence emerges**
Runtime data reveals what design-time evidence cannot. (Icon: Magnifying glass)
- Risks are detected early**
Monitoring and feedback enable timely intervention. (Icon: Shield)
- Confidence evolves**
More relevant evidence leads to better informed decisions. (Icon: Bar chart with upward arrow)
- Trust is sustained**
Stakeholders see that safety is actively managed – not a one-time activity. (Icon: Group of people)



Key message: Continuous assurance ensures that our safety claims remain valid and trustworthy throughout the system's entire lifecycle.

Midterm Status & Take-Away

We have made strong progress toward confidence-based safety assurance for AI systems.



✓ What we have achieved



Integrated Hybrid Safety Concept

We have developed and aligned the hybrid integration of design-time and runtime evidence to support continuous assurance.



GSN-based Assurance Case

We have defined a structured safety argument using GSN, linking claims, strategies, evidence, and assumptions.



Subjective Logic Confidence Assessment

We have applied Subjective Logic to quantify and propagate confidence throughout the assurance case.



Cross-TP Integration

We have connected contributions across TPs and established shared concepts, metrics, and interfaces.



First Concrete Evidence & Monitoring Contributions

We have identified initial evidence sources, defined monitoring concepts, and demonstrated feasibility in example use cases.

⌚ What remains



Refinement of Metrics & Evidence Qualification

We will further refine metrics, thresholds, and criteria to ensure consistent and meaningful evidence qualification.



Maturation of Confidence Aggregation

We will improve techniques for combining confidence across heterogeneous evidence while managing dependencies and uncertainties.



Broader Consortium Alignment

We will ensure broader adoption, feedback, and alignment of our concepts across all relevant stakeholders and use cases.



Validation Across Use Cases

We will validate and stress-test our approach in diverse, realistic scenarios to build robustness and generalizability.



Further Maturation Toward Implementation

We will pave the way for scalable tool support, integration into processes, and practical adoption.



From AI performance to **confidence** in the safety argument.

Our mission: Build trustworthy AI by making safety evidence visible, measurable, and continuously strengthened.

Braiding a Safety Rope



Braiding a Safety Rope





SAFE AI
ENGINEERING

September 16, 2026 | Safe AI Engineering

AI System Integration and Evaluation

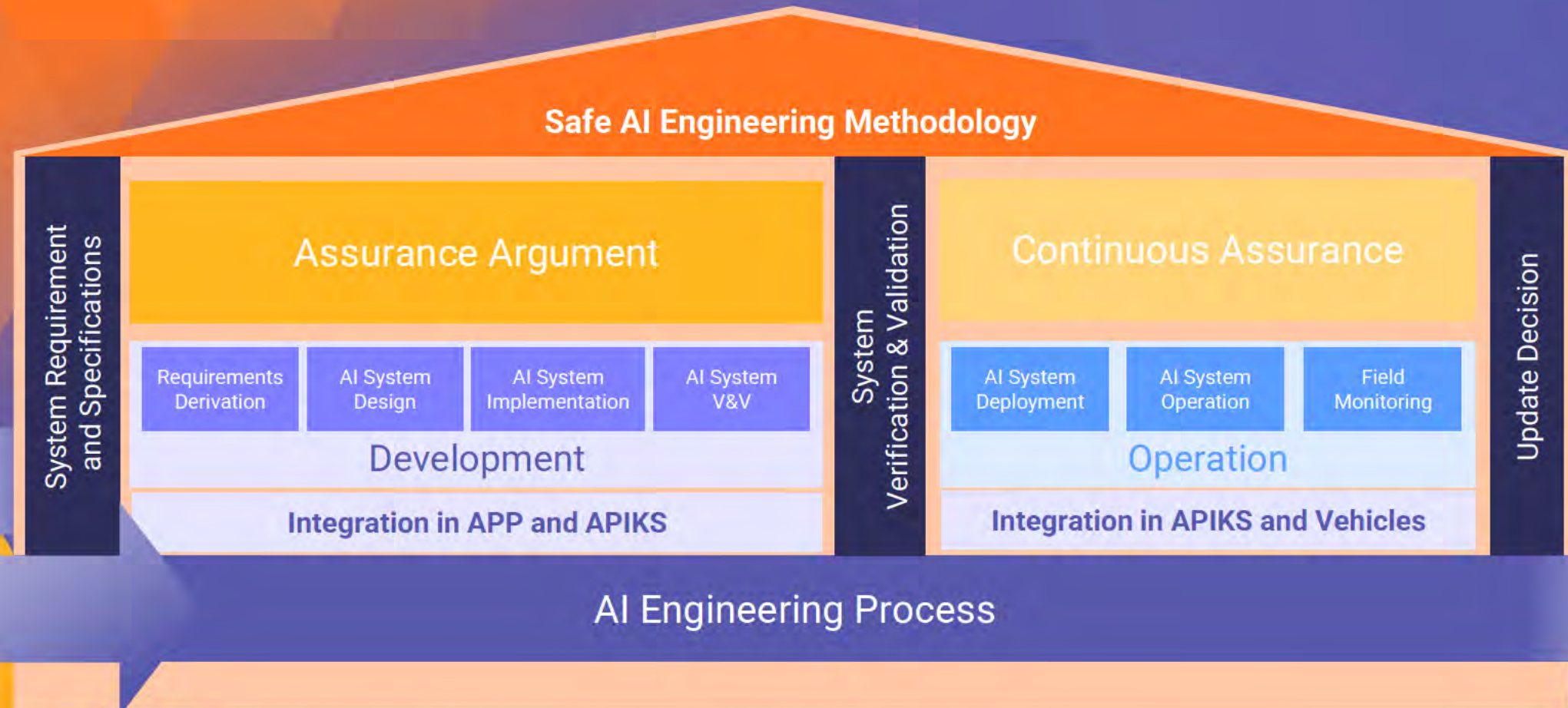
Núria Mata | Fraunhofer IKS





How to integrate the developed methods and evaluate the development process?

AI System Integration and Evaluation





AI Function with Increasing Complexity

Use Case #1:

- Detect pedestrians using 2D bounding boxes.
- Scenarios involving a stationary pedestrian on a straight road and a stationary ego vehicle.

Use Case #2:

- Detect pedestrians using 3D bounding boxes.
- Scenarios involving a dynamic pedestrian on a straight road with a stationary ego vehicle.

Use Case #3:

- Detect pedestrians using 3D bounding boxes and tracking methods.
- Scenarios involving multiple dynamic pedestrians at an intersection with a dynamic ego vehicle.



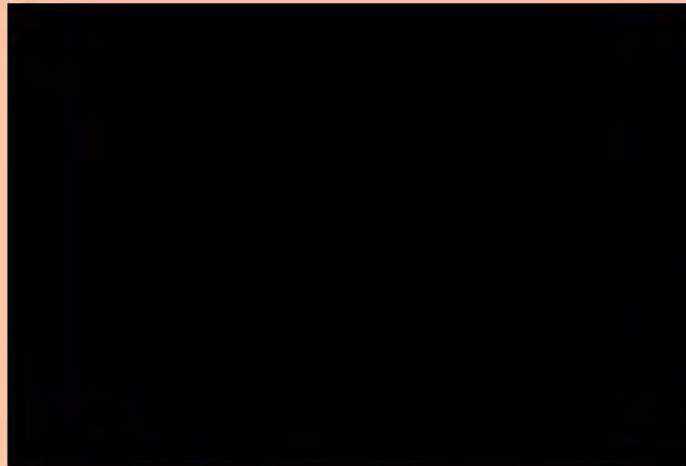
Safe AI Function in Simulation



Safe AI Engineering Methodology

Integration in APP and APIKS

Interoperability – ROS2
Interfaces



parallel walk → hesitation → approach → cross



Evaluate Assurance Arguments

Evaluate the AI Engineering Process

Safe AI Function in Test Vehicles



Safe AI Engineering Methodology

Integration in Vehicle Demonstrators



Interoperability – ROS2
Interfaces



Evaluate Assurance Arguments

Evaluate the AI Engineering Process



Ask questions to be assured on your case!

Introduction

Torsten Stiehm | Luxoft
Torsten.Stiehm@dxs.com

Methodology

Elena Hoemann | DLR
Elena.hoemann@dlr.de

Data Driven Engineering

Leonard Schroven | Capgemini
leonard.schroven@capgemini.com

Yannick Kees | DLR
yannick.kees@dlr.de

Safety Metrics Evaluation Pipeline

Nenad Ocelic | Luxoft
nenad.ocelic@dxs.com

Monitoring

Antje Loyal | Aumovio
antje.loyal@aumovio.com

Hybrid Safety Concept

Andreas Rohatschek Striebel | Robert Bosch GmbH
Andreas-Juergen.Rohatschek@de.bosch.com

AI System Integration and Evaluation

Núria Mata | Fraunhofer IKS
nuria.mata@iks.fraunhofer.de

Supported by:



Federal Ministry
for Economic Affairs
and Energy

on the basis of a decision
by the German Bundestag



www.safe-ai-engineering.de/en/home
www.linkedin.com/company/ki-familie

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.

Images: Safe AI Engineering Consortium