DNN SAFETY CONCERNS AND THEIR MITIGATIONS

Dr. Shervin Raafatnia
Robert Bosch GmbH

SoTA Standards and Their Applicability to DNNs

- **ISO 26262:** focus on E/E malfunctions and HW failures
  → disregards many aspects relevant to assuring safety
  DNN applicability: HW issues, tooling verification, …

- **ISO/PAS 21448 (aka. SOTIF):** addresses functional insufficiencies leading to hazardous behavior when triggered

  - "A proper understanding of the function by the user, its behavior and its limitations (including the human/machine interface) is the key to ensuring safety"

  DNN applicability: appropriate logic

BOSCH

DNNs and Safety

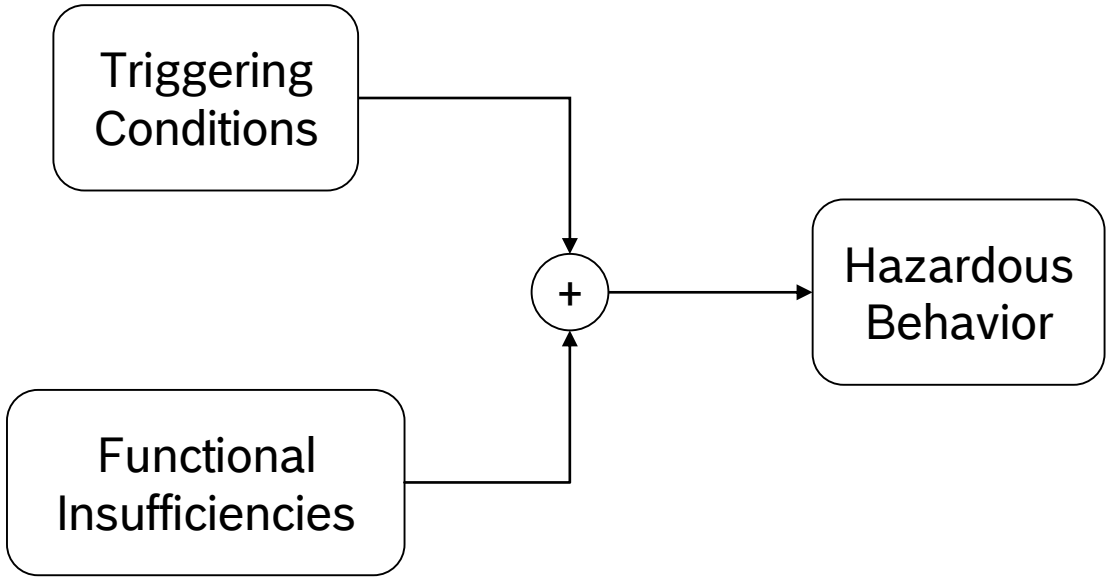
Functional Insufficiencies

DNNs and Safety

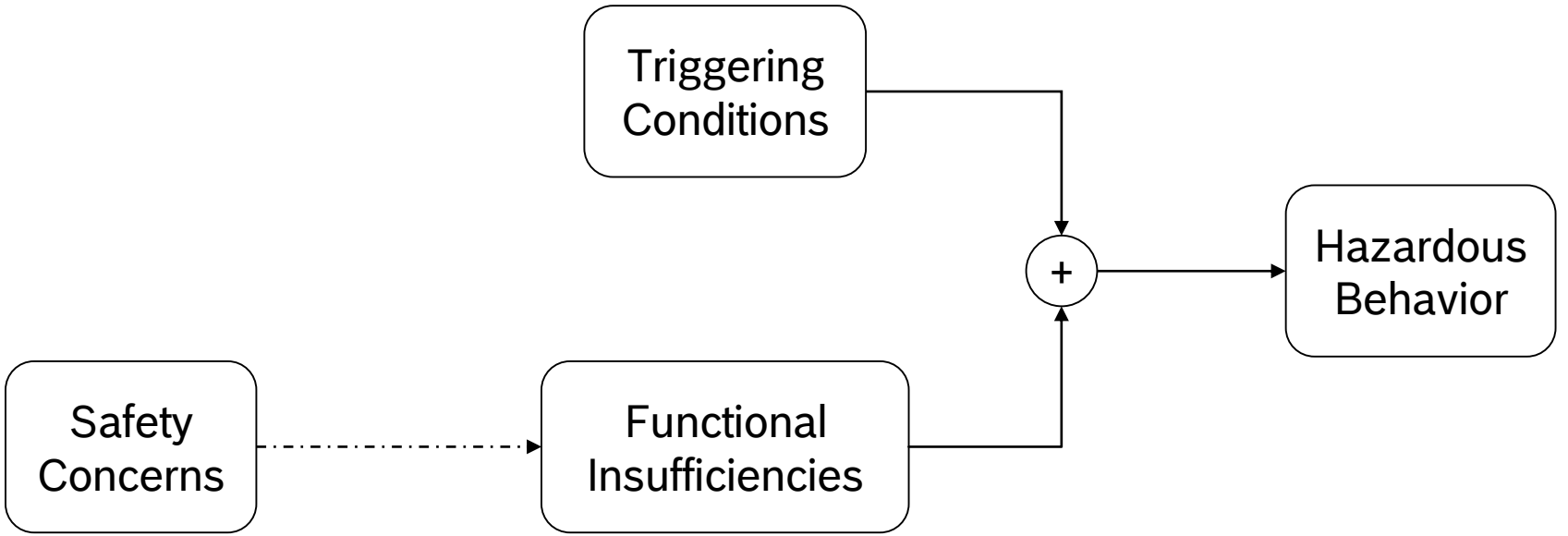
Triggering Conditions

Functional Insufficiencies

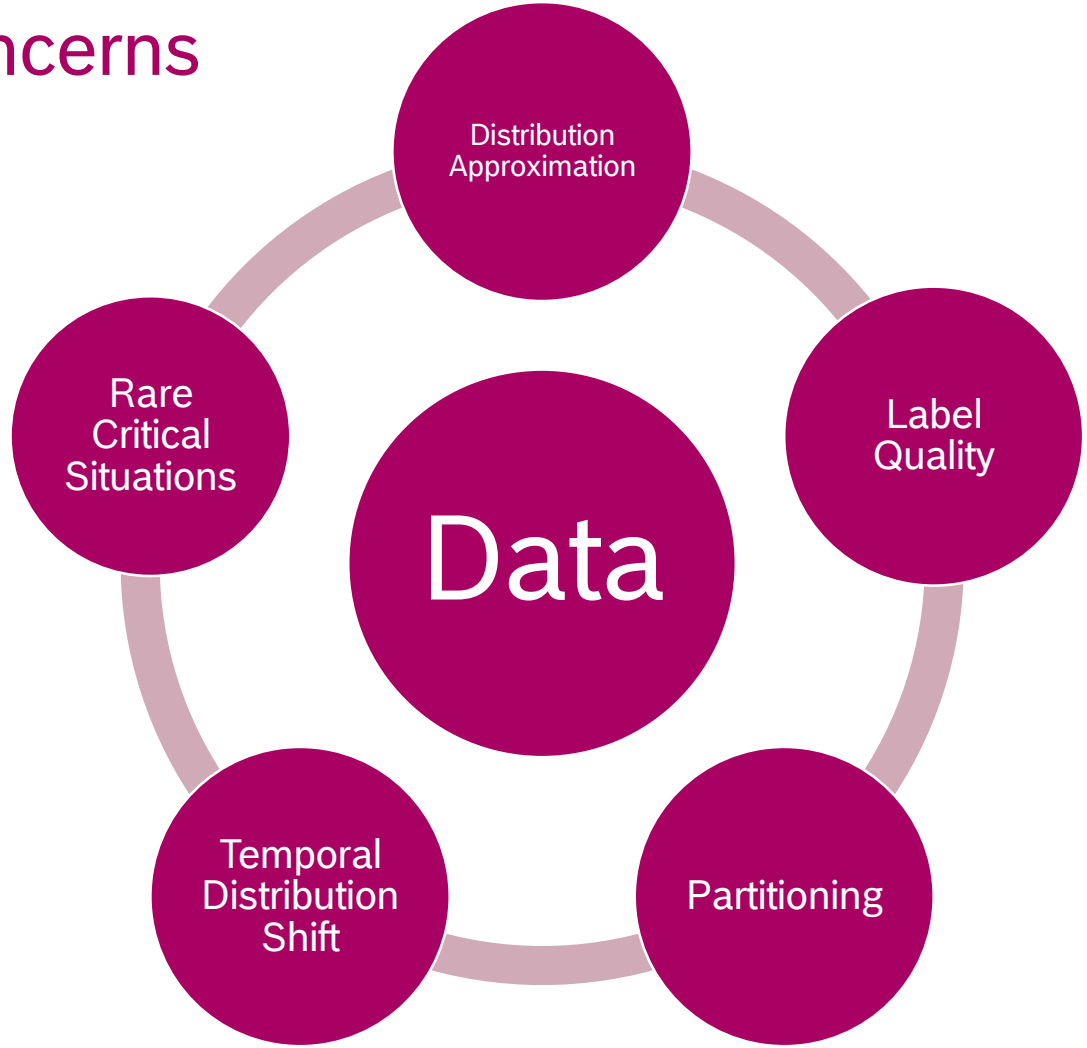
DNNs and Safety



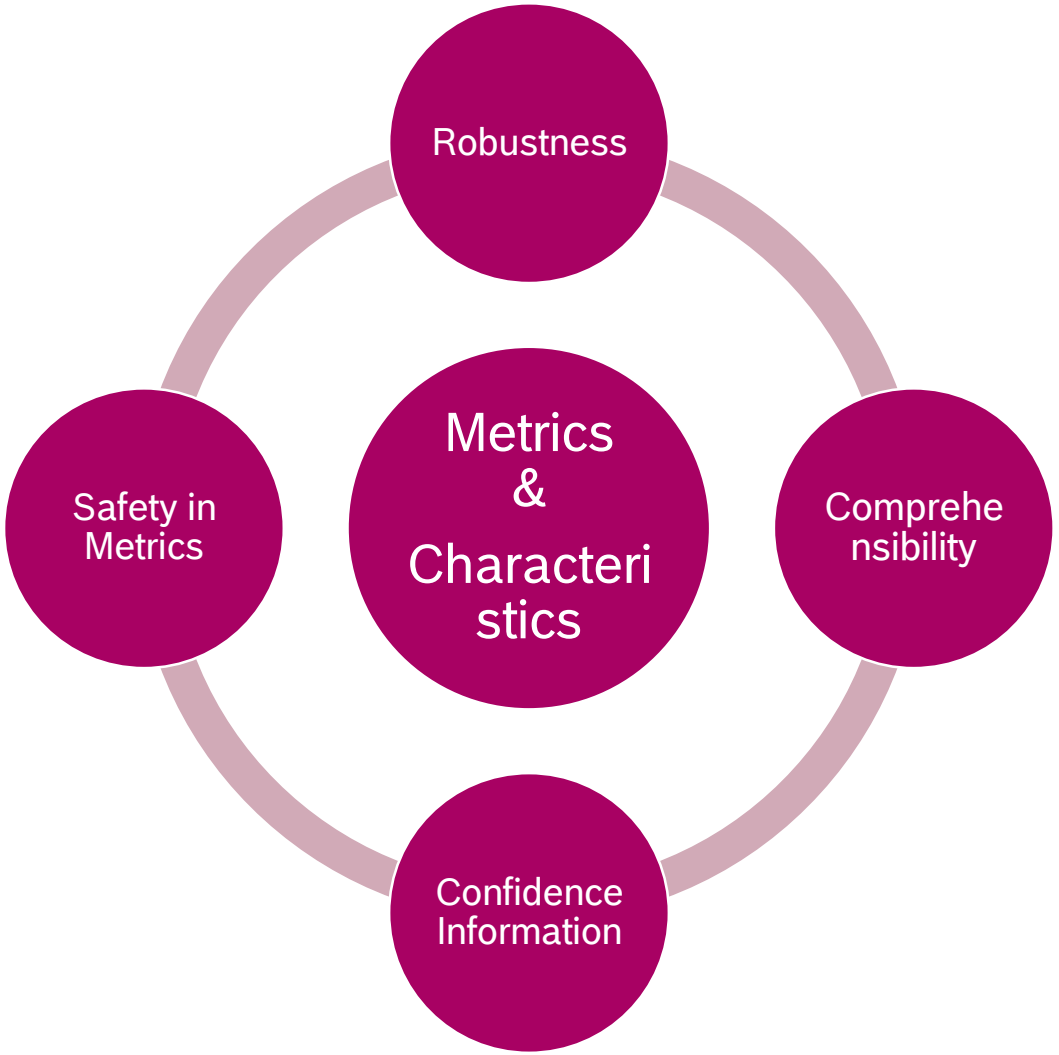
DNNs and Safety



Data-related Concerns



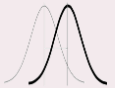
Other Concerns



Potential Mitigation Approaches

Real-World Distribution Approximation

- *Well-justified data acquisition strategy*
- *Data analysis*
- *Reliable confidence information*
- *Testing and detailed results analysis*



Distributional shift over time

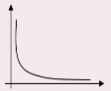
- *Reliable confidence information*
- *Continuous learning and updating*

Unreliable confidence information

- *Reliable confidence information*
- *Using Gray-box methods*
- *Testing and detailed results analysis*

Data Partitioning

- *Data partitioning guidelines*
- *Assessing the adherence to partitioning guidelines*



Unknown behavior in rare critical situations

- *Well-justified data acquisition strategy*
- *Reliable confidence information*
- *Random and dedicated testing*
- *Continuous learning and updating*

Incompre-hensible Behavior

- *Testing and detailed results analysis*
- *Using gray-box methods*

Dependence on labeling quality

- *Labeling guidelines*
- *Continuous quality checks*

Brittleness of DNNs

- *Reliable confidence information*
- *Adversarial attacks defenses*
- *Targeted testing and detailed analysis development process*
- *Continuous learning and up-dating*

Insufficient consideration of safety in metrics

- *Defining heuristics to assess criticality/relevance of different objects*
- *Performance evaluation with respect to safety*

THANKS FOR YOUR ATTENTION