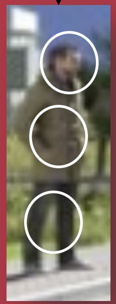
KI WISSEN
Automotive AI Powered by Knowledge

KI Wissen Final Event | 21-22 March 2024

Advancements of Local and Global Explanation Methods for Failure Case Detection

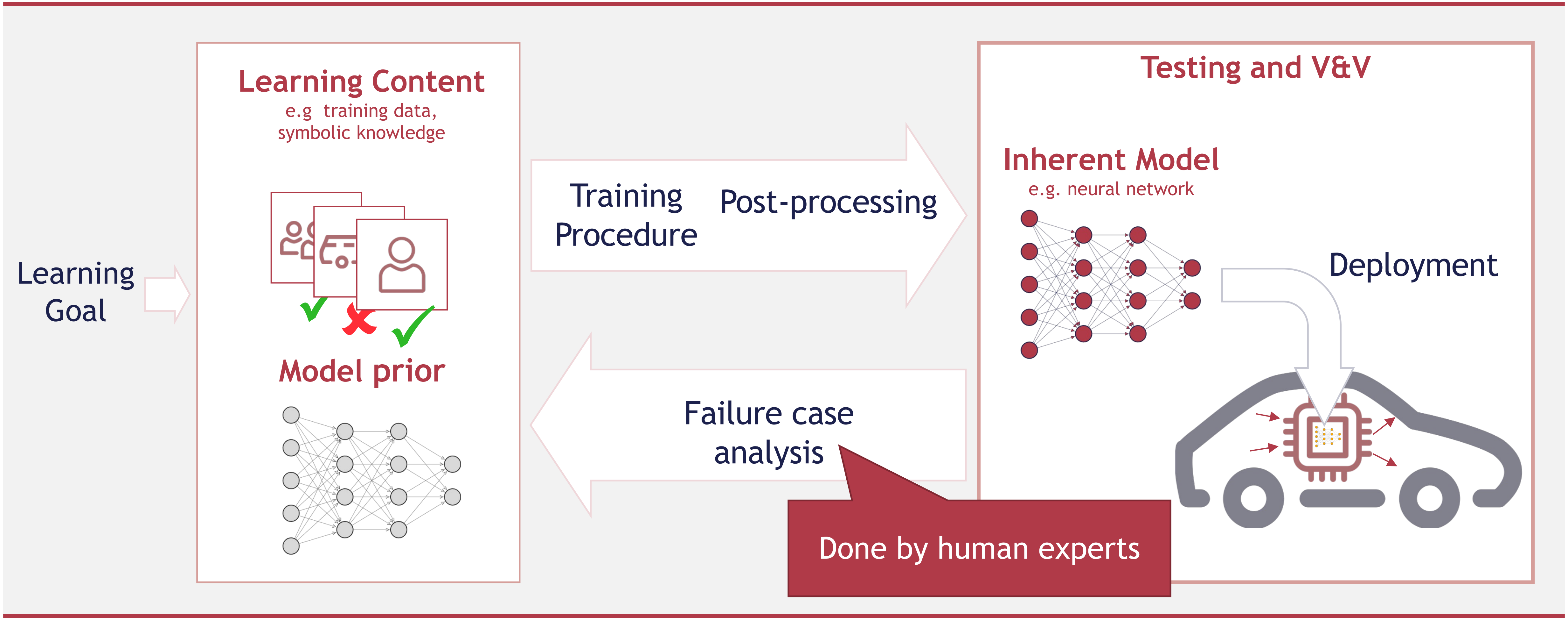
Christian Hellert | Continental

KI
WISSEN
Automotive AI Powered by Knowledge



Continental

Need for Explainability



Agenda

1. Introduction
   - Hypothesis
   - Concept and XAI
2. Concept Stability and Similarity
   - Concept Embedding Analysis
   - Results on Stability and Similarity
3. Concept Attribution and Analysis
   - Backpropagation-based Feature Attribution
   - Results on Concept Attribution and Analysis
4. Conclusion and Outlook

1

Introduction

Introduction

Hypothesis and Questions

Hypothesis

DNNs learn (semantic) concepts to detect objects, which are embedded in the latent space in different layers and thus learn a relationship between concepts and classes (objects).

- Question 1: Can we extract the concepts from a DNN model robustly?
- Question 2: Can we use the concept to detect/show failures?

Introduction
What are Concepts?

Domain of XAI

Semantic concept, e.g. „wheel“

ML Model

✓ Person

✗ Other

Association with latent space representation

Introduction

Explainable Artificial Intelligence (XAI)

Explainable decision system = There exists a

- **mechanism** providing an **explanation** (= explanator)
- to a **human** (= explainee)
- that allows them to **understand**
- one of (= explanandum)
  - the **model** resp. parts thereof,
  - evidence for a **model output**, or
  - the **context** of the system's reasoning.

Understanding = successful update of mental model; can be

- mechanistical = how it works, or
- functional = what is its purpose

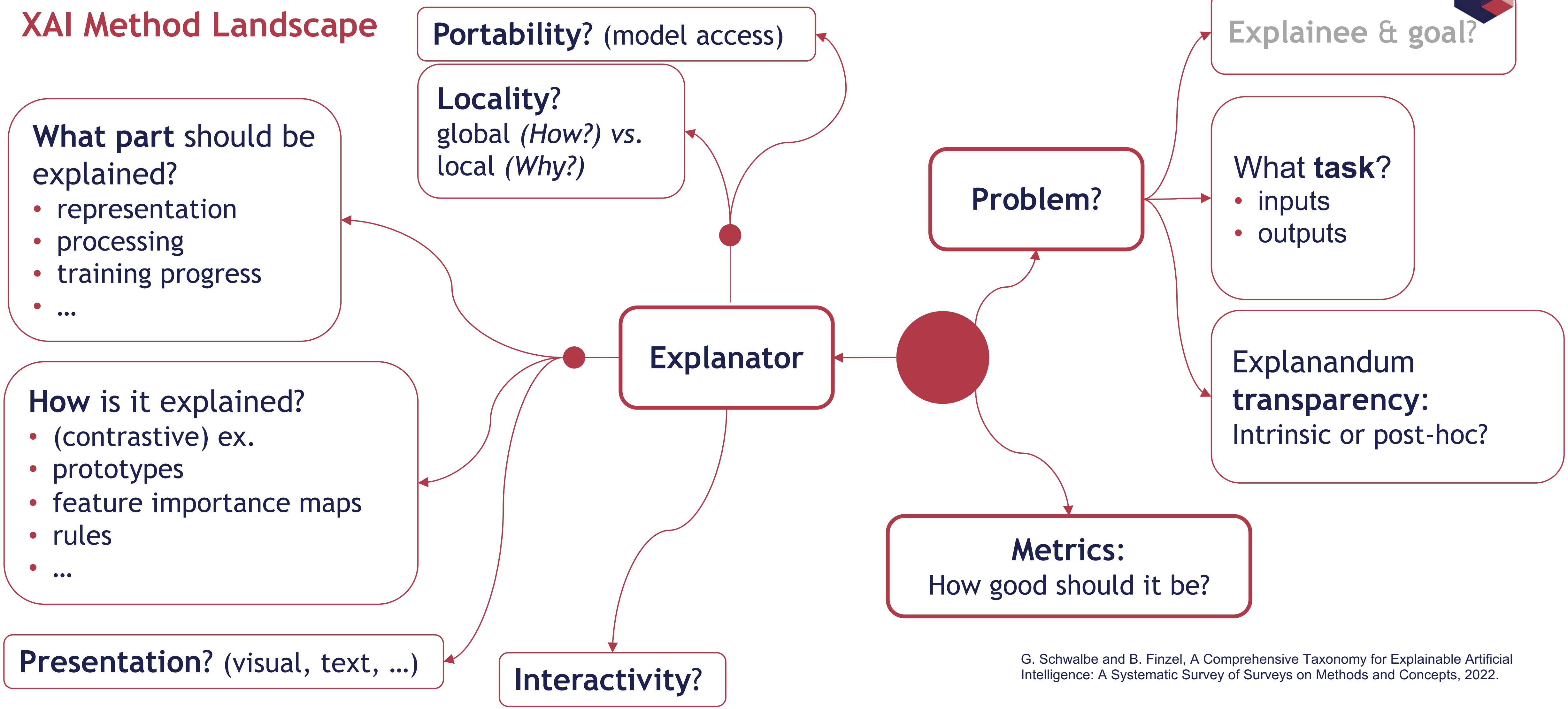
XAI = lots of cognitive science!

Levels of **transparency** of a model (= mechanistic understanding):

- simulatable (= understandable as a whole)
- decomposable (into simulatable parts)
- algorithmically transparent (= mathematical understanding)

Introduction

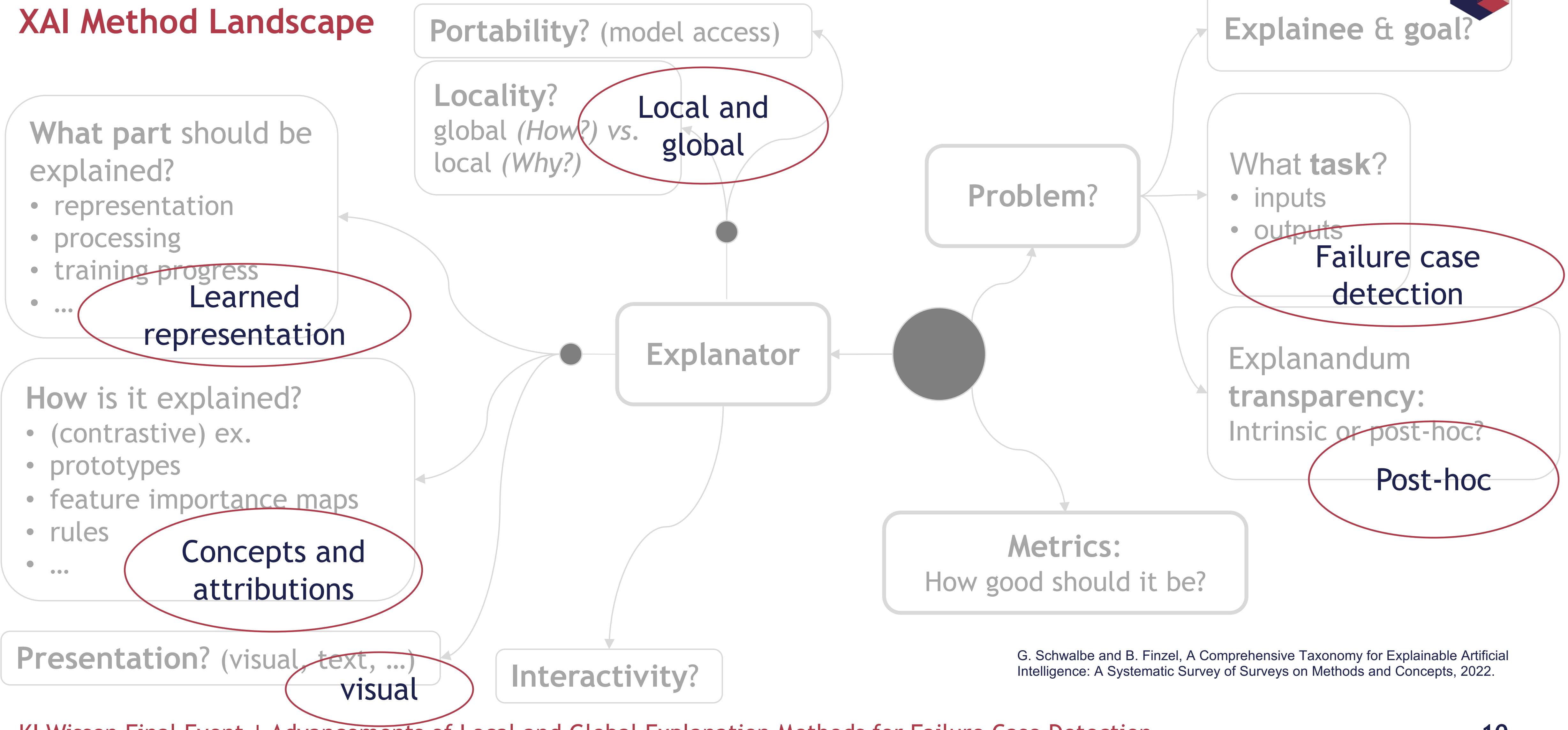
XAI Method Landscape



G. Schwalbe and B. Finzel, A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts, 2022.

Introduction

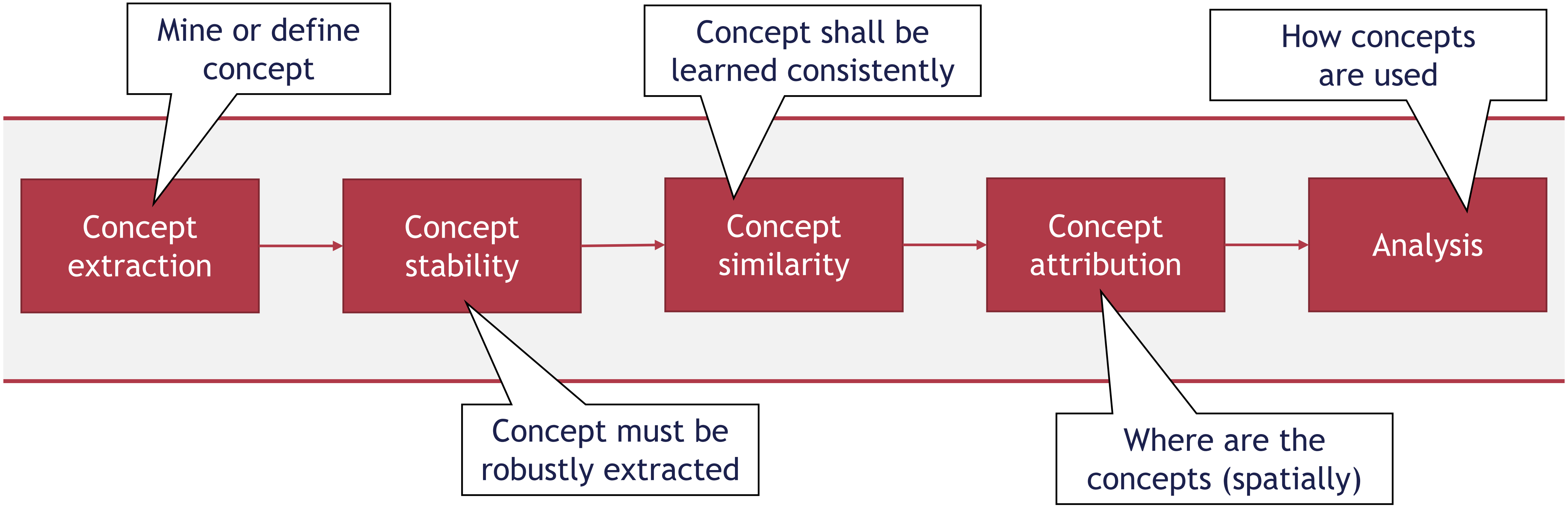
XAI Method Landscape



G. Schwalbe and B. Finzel, A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts, 2022.

Introduction

Pipeline Towards Failure Case Detection



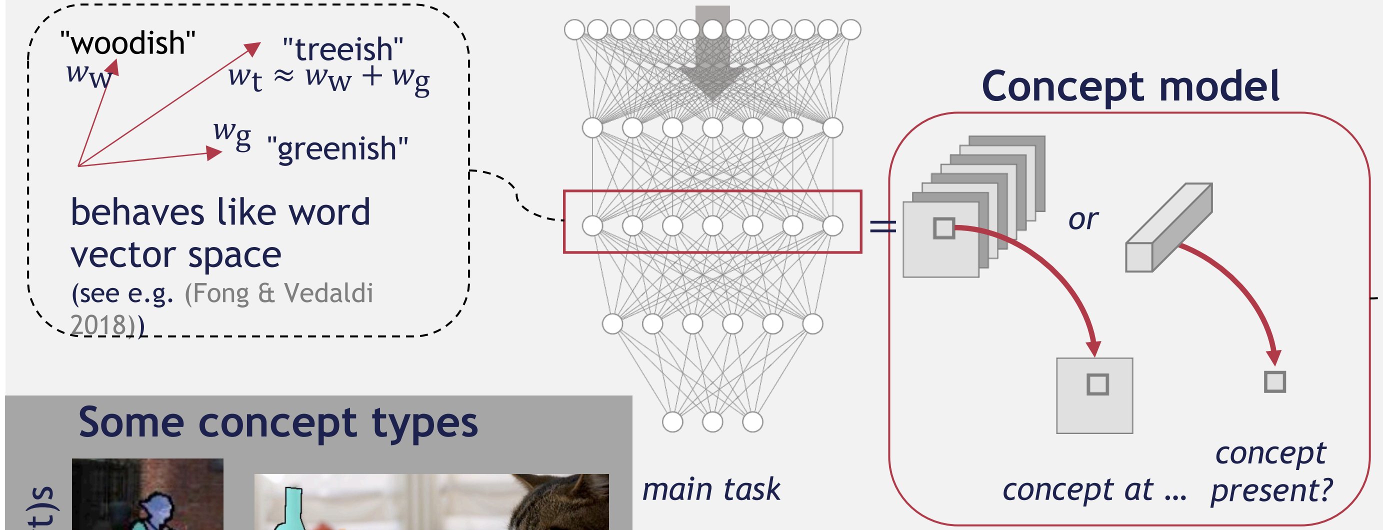
2

Concept Stability and Similarity

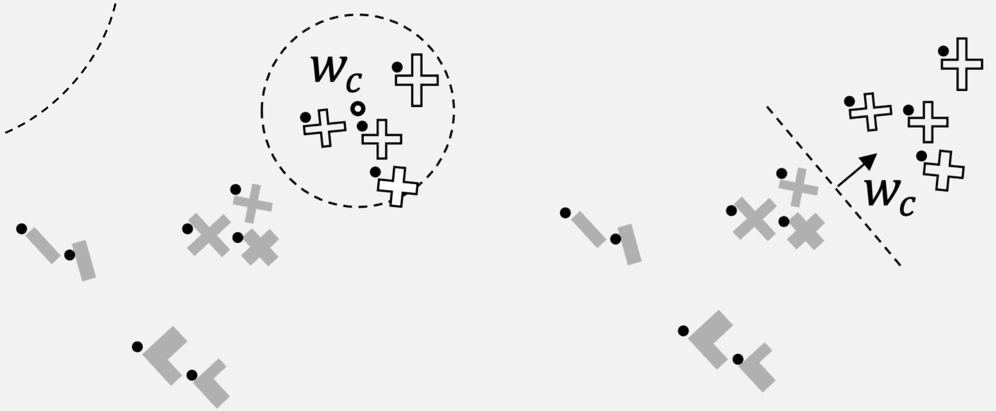
Concept Stability and Similarity

Latent Space Analysis: Concept Embedding Analysis

- *Goal: Associate semantic concept w/ latent space vector / subspace*
- *Idea*: Vector as parameters of simple predictor for concept (*concept model*)



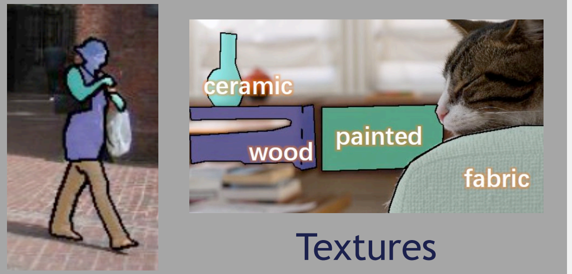
Main types of concept models:



Clustering & matrix factorization
e.g. ACE (Ghorbani et al. 2019), NCAV (Zhang et al. 2021)

Linear
$f_c = \cdot \circ w_c + t_c$
e.g. TCAV (Kim et al. 2018), Net2Vec (Fong & Vedaldi 2018)

Some concept types

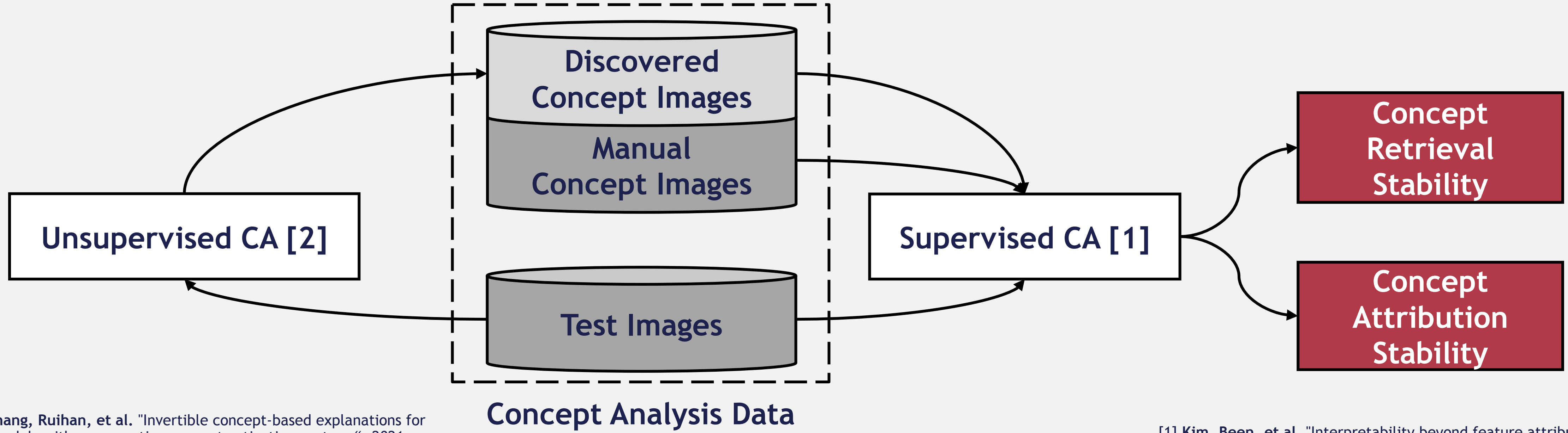


Object(part)s

Textures

(Xiao et al. 2018), p. 6, Fig. 3

Concept Stability and Similarity



[2] **Zhang, Ruihan, et al.** "Invertible concept-based explanations for cnn models with non-negative concept activation vectors.", 2021.

[1] **Kim, Been, et al.** "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).", 2018.

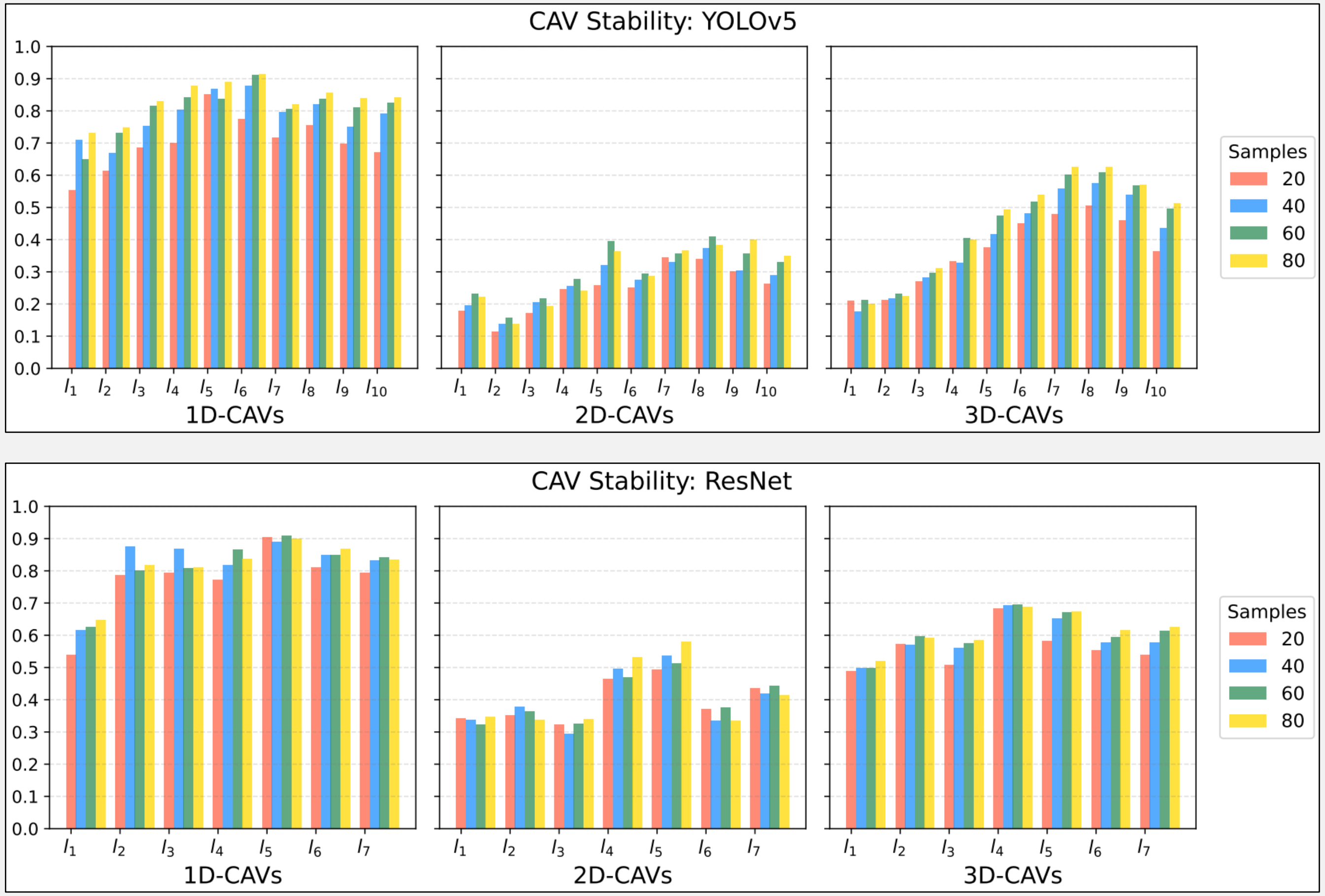
- **Unsupervised CA** – concept discovery
- **Supervised CA** – extraction of user-defined concepts

Benefits: **Combine strengths** of concept extraction and discovery with **minimal manual effort.**

Concept Stability and Similarity

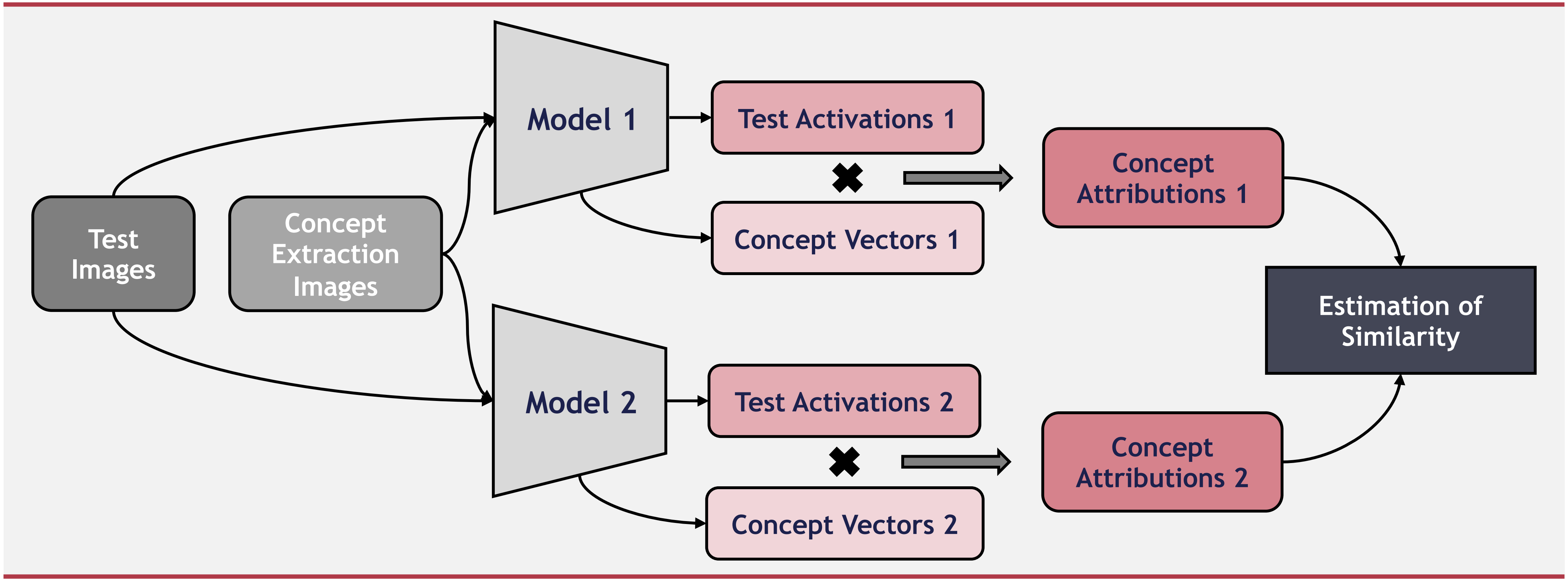
Results: Concept Stability

- **1D-CAVs are the most stable**
  - **Faster** to evaluate, need **less memory.**
- **40 (ideally more than 60) concept samples** for high stability
- Network **architecture has impact** on behavior of stability
  - E.g., top-stability is achieved in different relative backbone depth



Concept Stability and Similarity

Concept-based Semantics Comparison

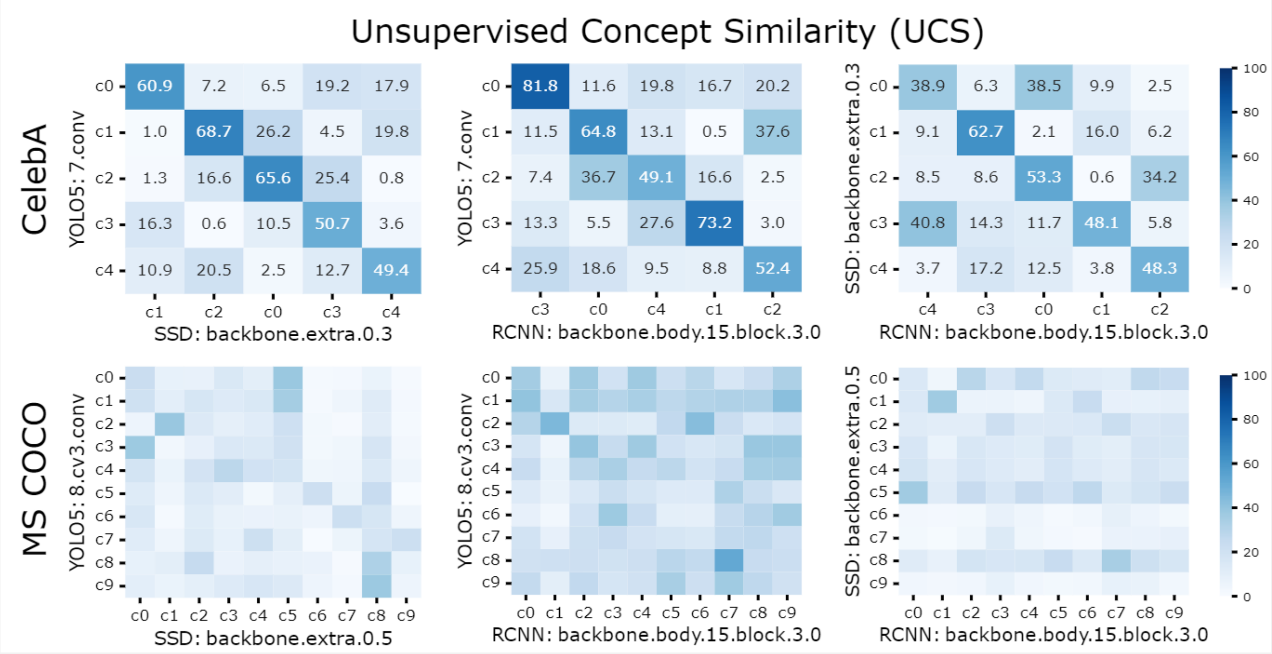


Indirect feature space comparison via semantic concepts and sample attributions

Concept Stability and Similarity

Results: Unsupervised Saliency-based Similarity

- Test **data diversity impacts the complexity** of further **inspection.**
- Different (architecture-wise) **networks learn similar concepts:**
  - Trained on MS COCO, discovered similar concepts in CelebA



Similar concepts in CelebA

Similar concepts in MS COCO

3

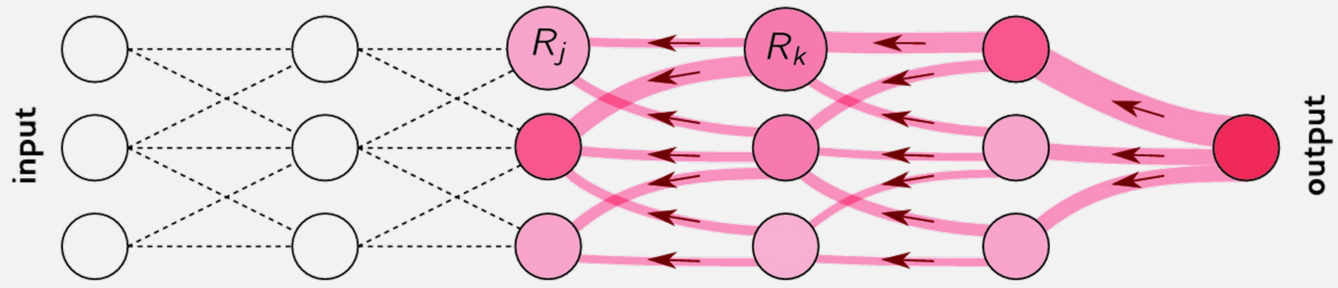
Concept Attribution and Analysis

Concept Attribution and Analysis

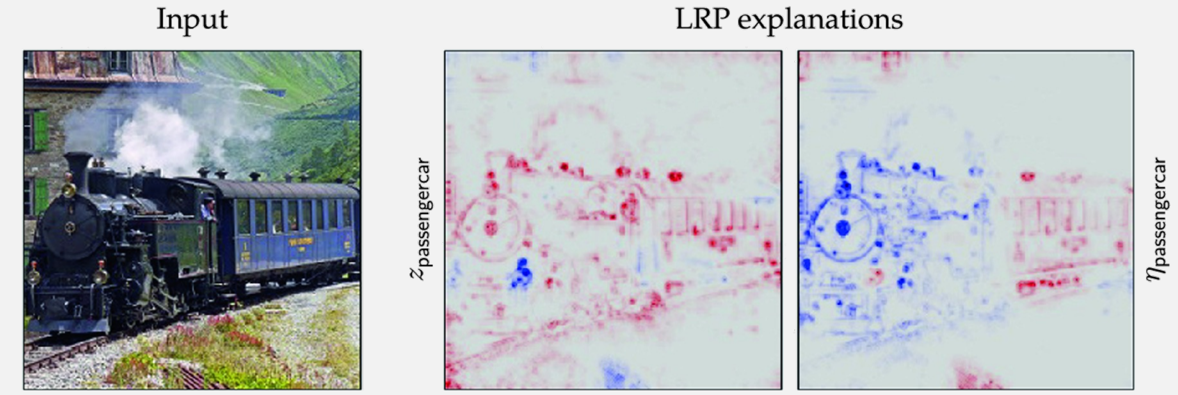
Feature Saliency: Backpropagation-based

- *Idea:*
  - **Trace back influence** (=*Relevance*) of activations from output to input
  - Total relevance within a layer l stays constant:

    $$f(x) = \cdots = \sum_i R_i^{(l-1)} = \sum_i R_i^l$$
- One additional **backwards-pass**
- Requires **access to model** internals
- Backpropagation functions **must be chosen carefully** wrt. layer type and question



(Montavon et al. 2019), Fig. 10.2

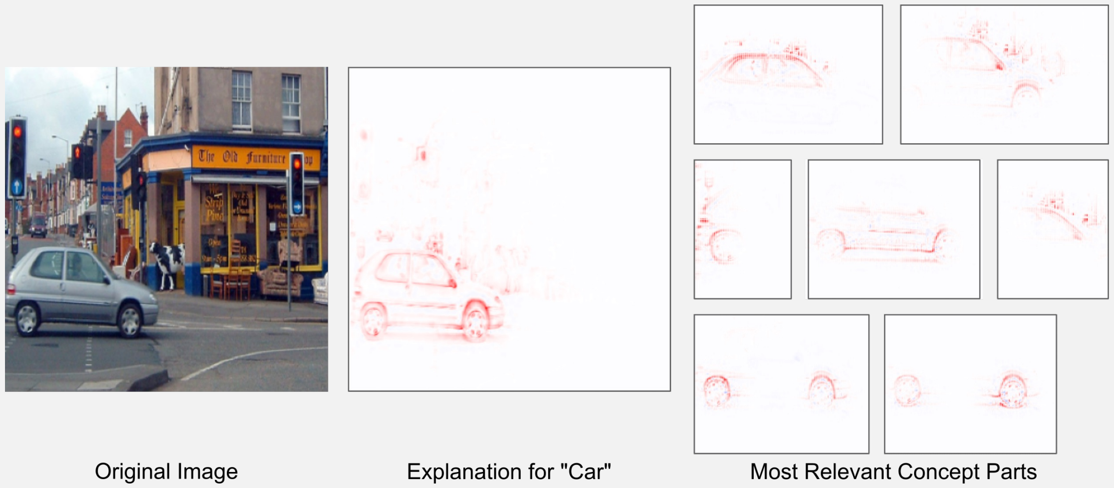


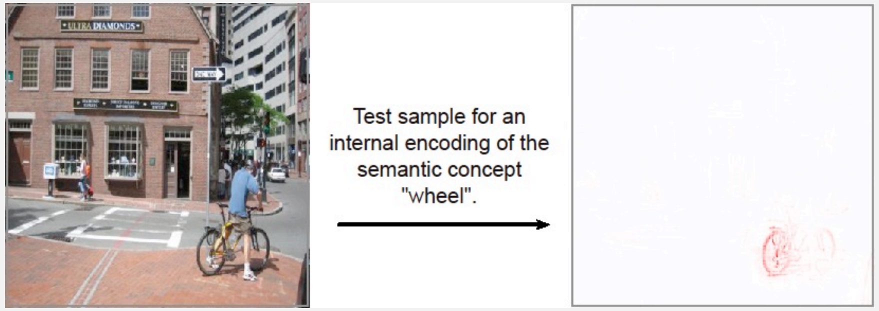
(Montavon et al. 2019), Fig. 10.5

Concept Attribution and Analysis

Concept Decomposition and Testing

- Concept Decomposition
  - Conv-filter-conditioned local attribution
  - Assigning filters to concepts
- Testing for specific concepts
  - arbitrary choice of predefined concept
  - Attribution for global concept encoding

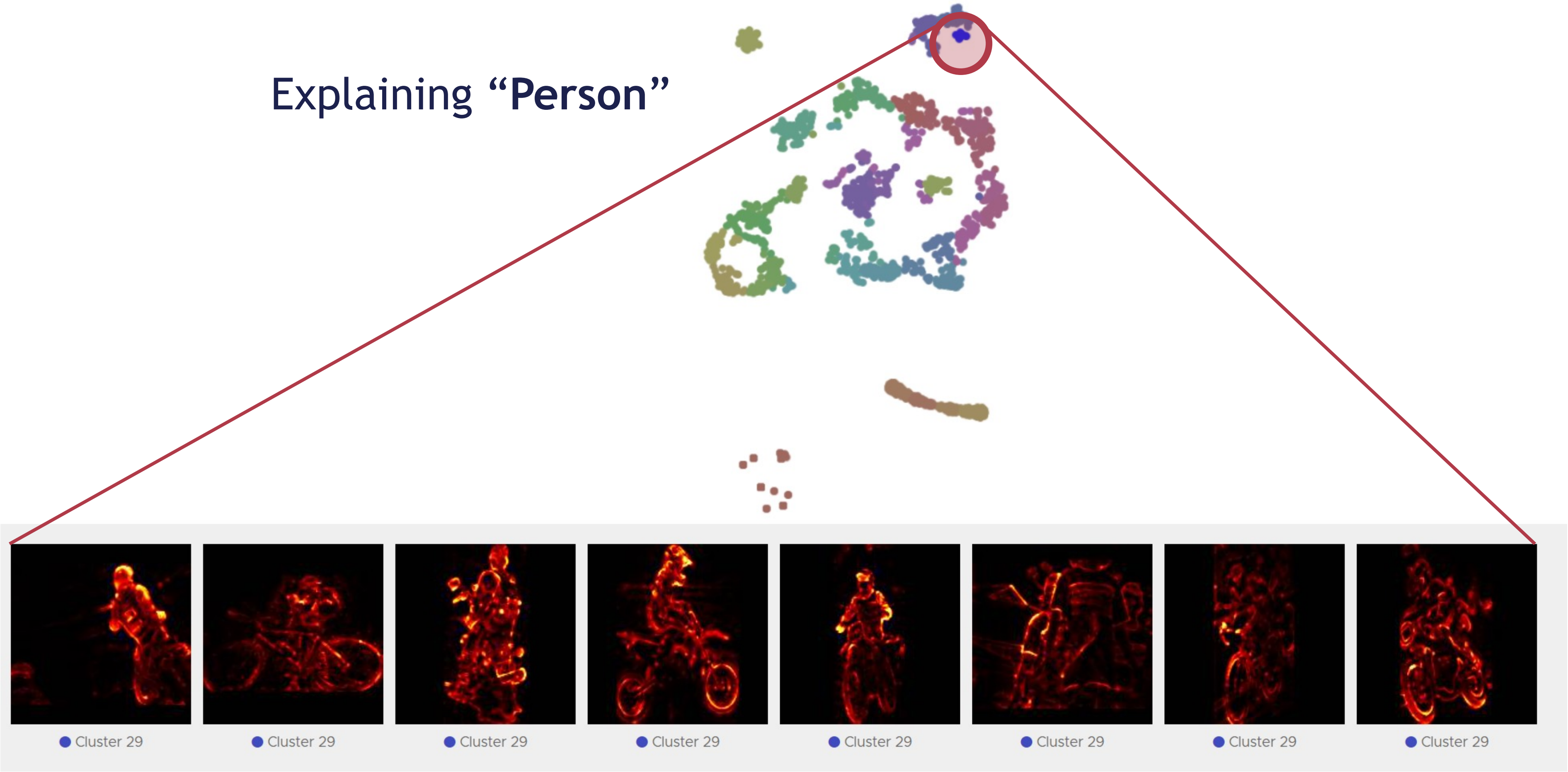




Concept Attribution and Analysis

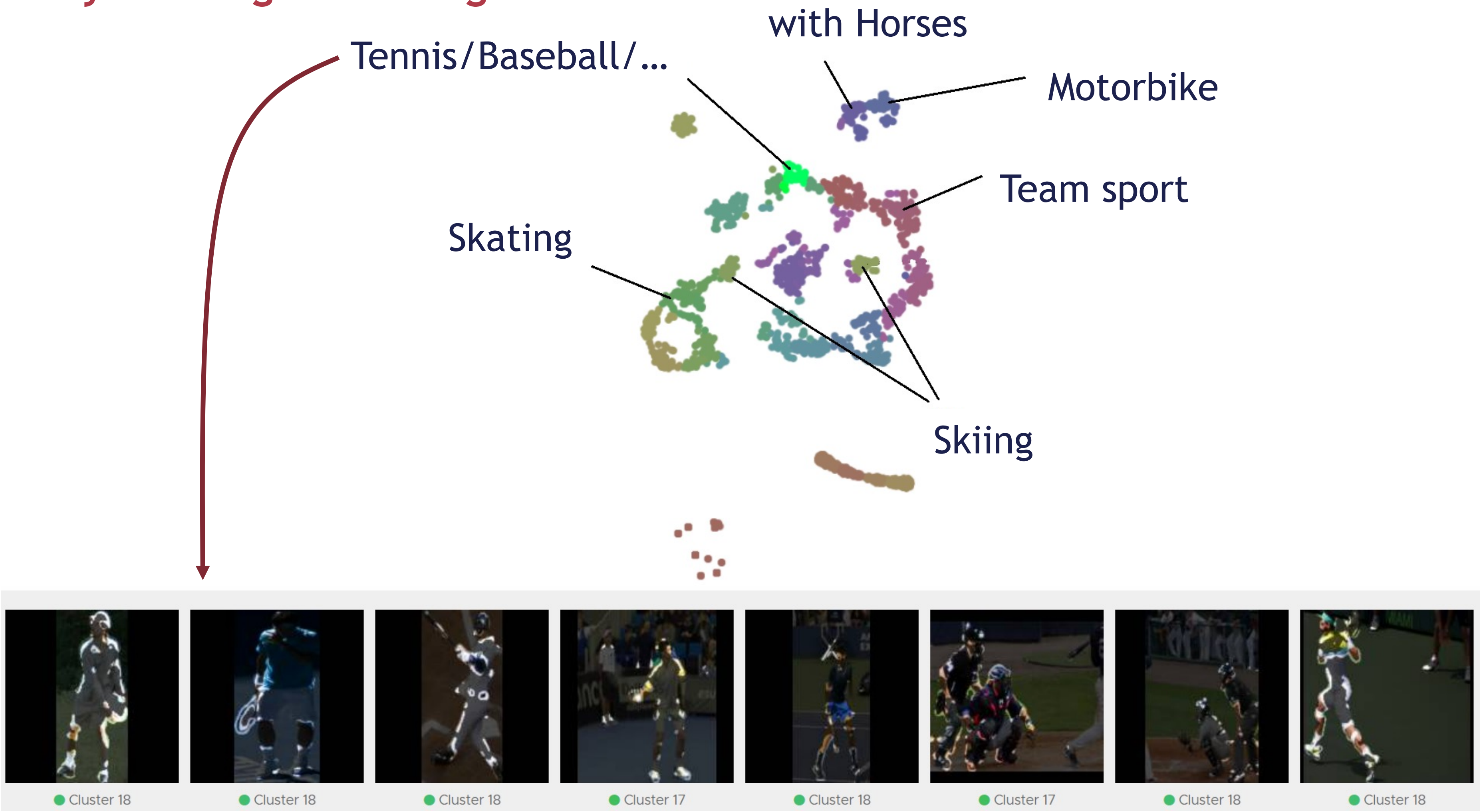
Concept Analysis Using Clustering

Explaining “**Person**”



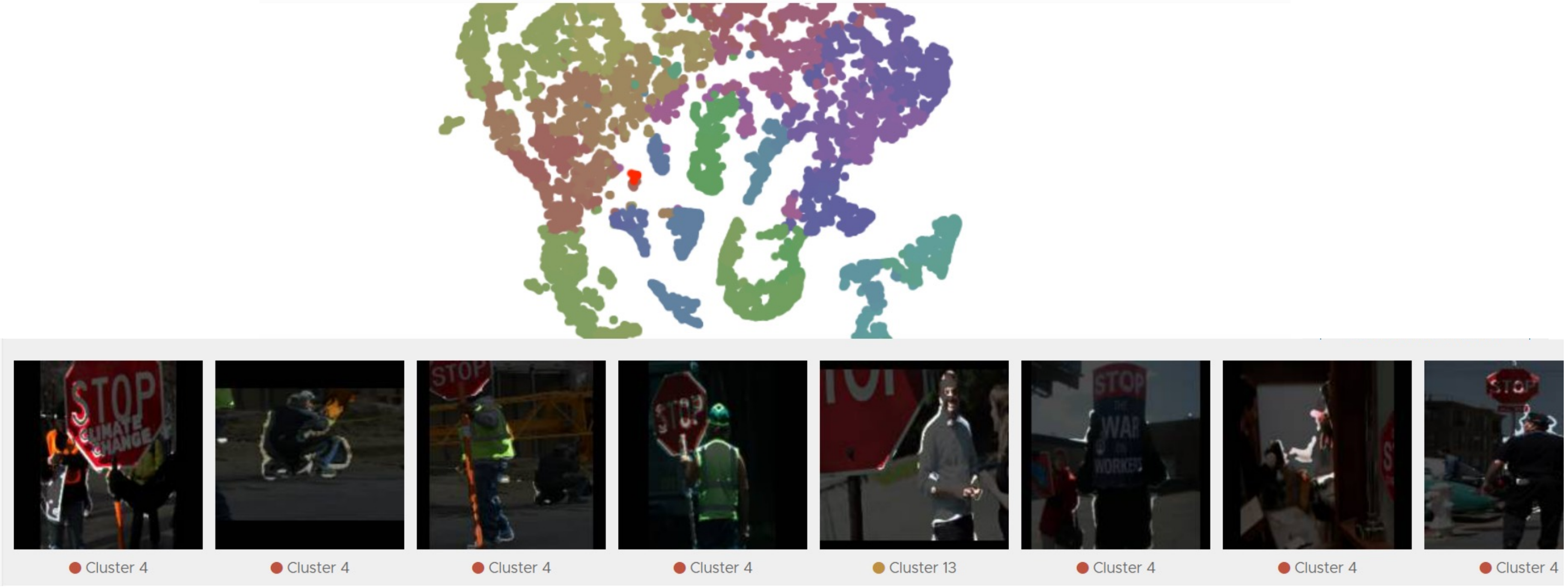
Concept Attribution and Analysis

Concept Analysis Using Clustering



Concept Attribution and Analysis

Concept Analysis Using Clustering



4

Conclusion and Outlook

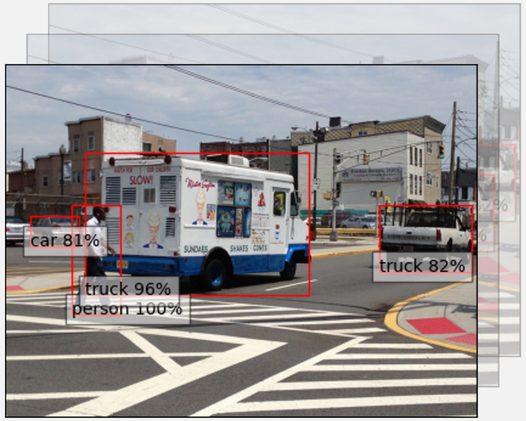
Conclusion

- **Hypothesis**: DNNs learn (semantic) concepts to detect objects, which are embedded in the latent space in different layers and thus learn a relationship between concepts and classes (objects).
  - Results in the experiments support the hypothesis
  - Most important extracted concepts are interpretable, but there are also non-interpretable concepts

- **Question 1**: Can we extract the concepts from a DNN model robustly?
  - Stability tests show that concepts can be learned robustly

- **Question 2**: Can we use the concept to detect/show failures?
  - Partially concept analysis can reveal spurious learned representations

Outlook

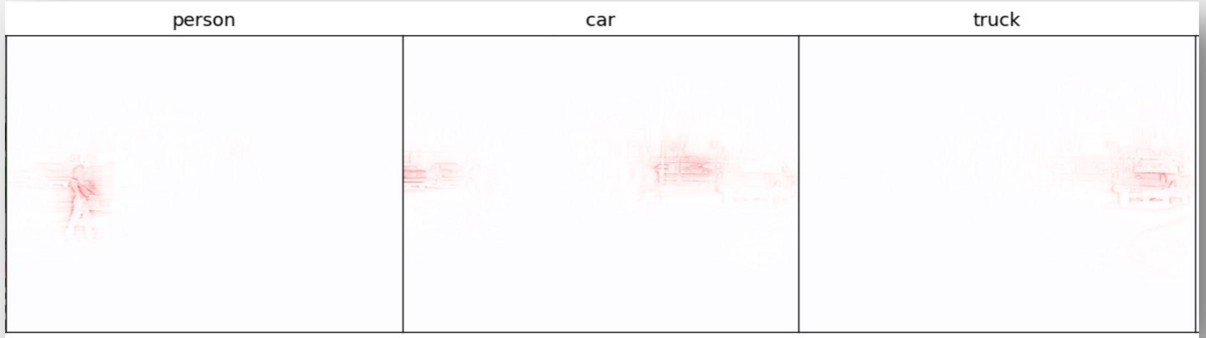


Samples with Detections

ML Model

Guided Adaptation of
Data/Model Parameters
(fixing failures)

Why is there a **person**?



Sample-specific explanations (local)
+ Concept-based explanations (global)

Validation

Domain Knowledge

Tooling for V&V activities

KI WISSEN

Automotive AI Powered by Knowledge

Christian Hellert | Continental | christian.hellert@continental.com

Georgii Mikriukov | Continental

Franz Motzkus | Continental

KI Wissen 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 Climate Action.

KI FAMILIE

VDA LEITINITIATIVE

Funded by the European Union

NextGenerationEU

Supported by:

Federal Ministry for Economic Affairs and Climate Action

on the basis of a decision by the German Bundestag

www.kiwissen.de @KI_Familie KI Familie