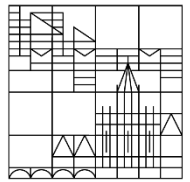
ModelSpeX

Model Specification Using Explainable Artificial Intelligence Methods

Udo Schlegel, Eren Cakmak, Daniel Keim

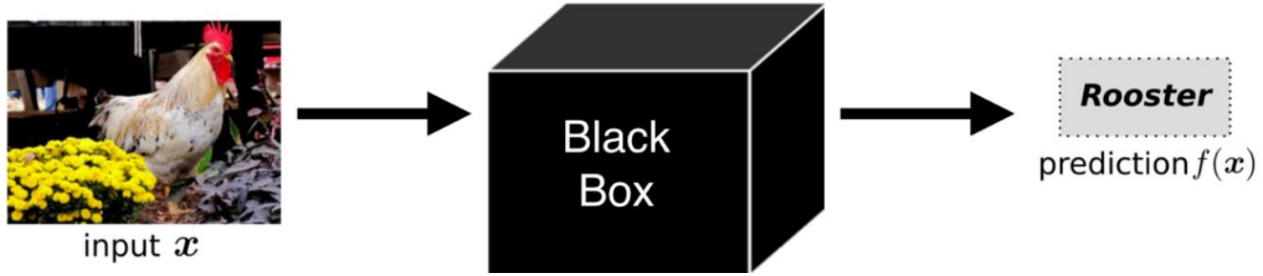
University of Konstanz, Untangle AI

Universität Konstanz



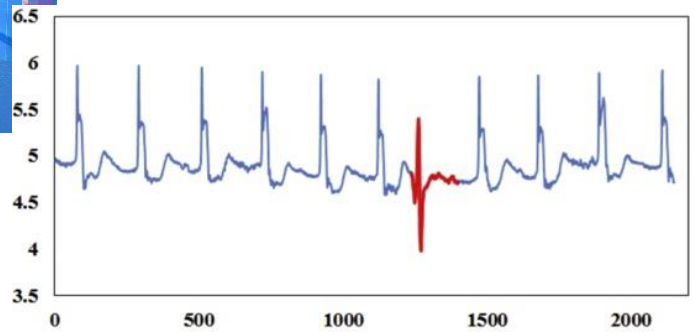
Introduction

https://github.com/HendrikStrobelt/Seq2Seq-Vis/blob/master/docs/pics/s2s_teaser.png



https://www.itu.int/en/ITU-T/AI/2018/Documents/Presentations/Wojciech%20Samek.pdf

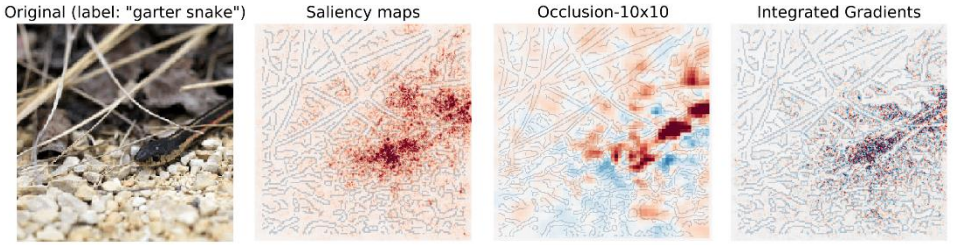
https://encrypted-tbn0.gstatic.com/images?q=tbn%3AANdgGcRrEFhzsuOpzlcqgtKuAqGWAZw8ftEsRLWc1DlWHW9Ay_kUe7jj&usqp=CAU



https://ars.els-cdn.com/content/image/1-s2.0-S0957417416301191-gr1.jpg

Current XAI

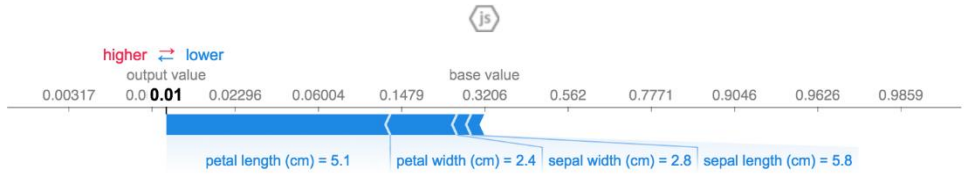
https://distill.pub/2017/feature-visualization/images/diversity/mixed4e_55_diversity.png



https://github.com/marcoancona/DeepExplain/raw/master/docs/comparison.png

IF Country = United-States **AND** Capital Loss = Low
AND Race = White **AND** Relationship = Husband
AND Married **AND** 28 $<$ Age $\leq$ 37
AND Sex = Male **AND** High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary $>$ \$50K

https://homes.cs.washington.edu/~marcotcr/aaai18.pdf

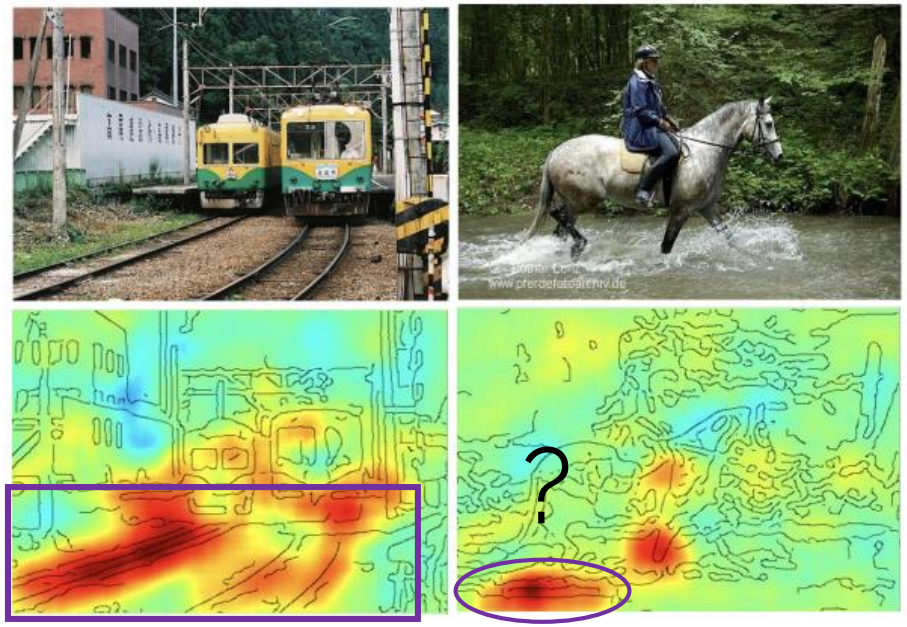


https://raw.githubusercontent.com/slundberg/shap/master/docs/artwork/iris_instance.png

Problems to Solve

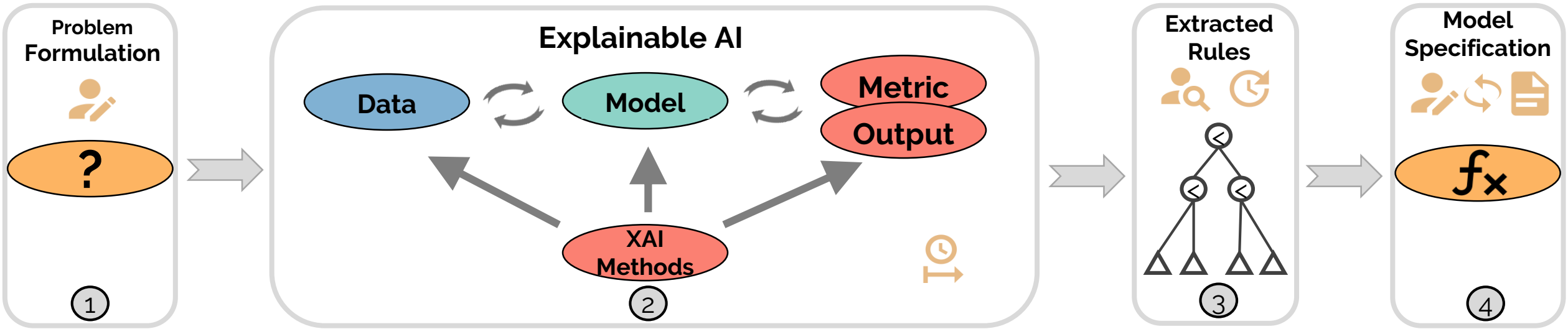
- Mismatch between problem solution and trained target
  E.g., Clever-Hans problem

- Verify learned patterns with domain knowledge

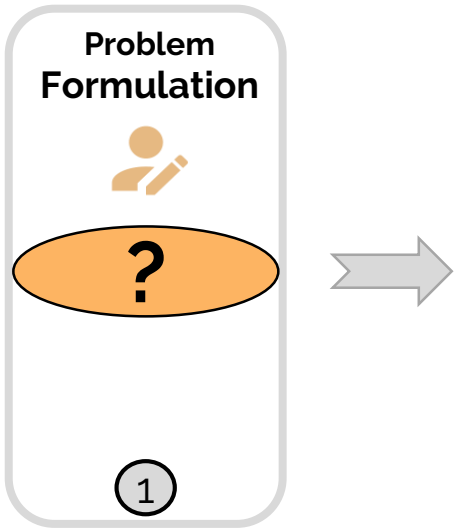


http://xai.unist.ac.kr/static/img/event/ICCV_2019_VXAI_Samek_Talk.pdf

The Workflow



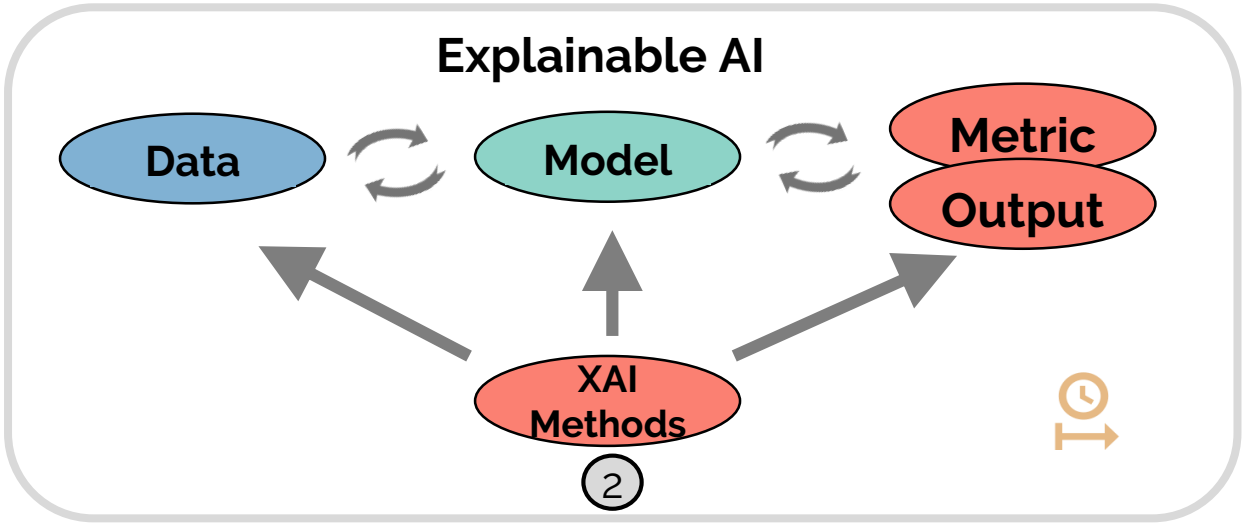
Generate Problem Formulation



- Intention to solve a task

- Faulty formulation leads to faulty models

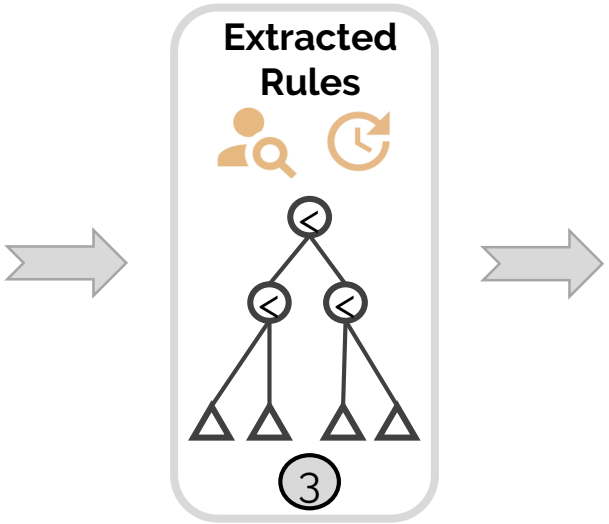
Applying Explainable AI



- XAI methods on top of data, model, output

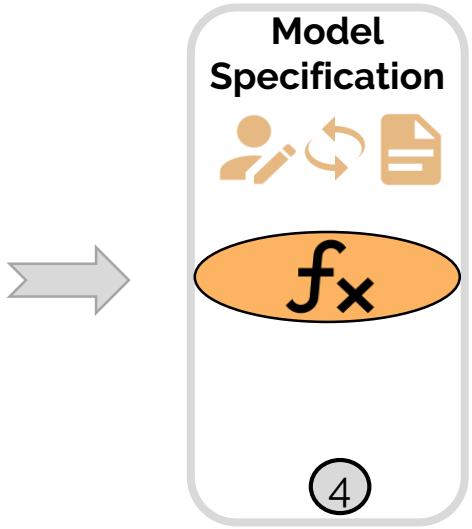
- Extract rules on these levels and during training

Adjusting Extracted Rules



- Adjust extracted rule sets
- Iterativeley and interactively pruning

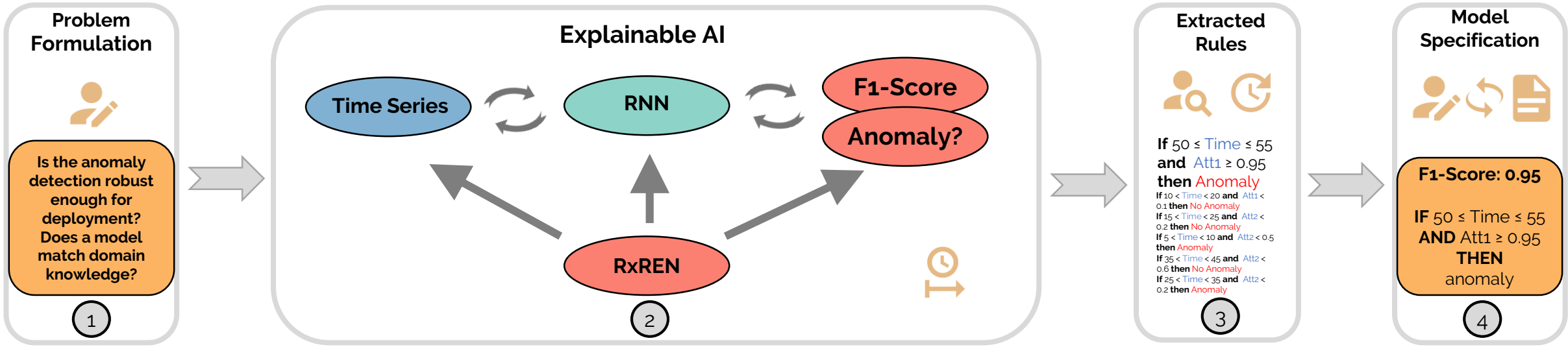
Resulting Model Specification



- Machine-readable description in collaboration of model and human

- Reduce bias of single expert by mulitple analysts refinements

Use Case Example



Predictive Maintenance:
Predict when an engine will fail
Based on sensor data

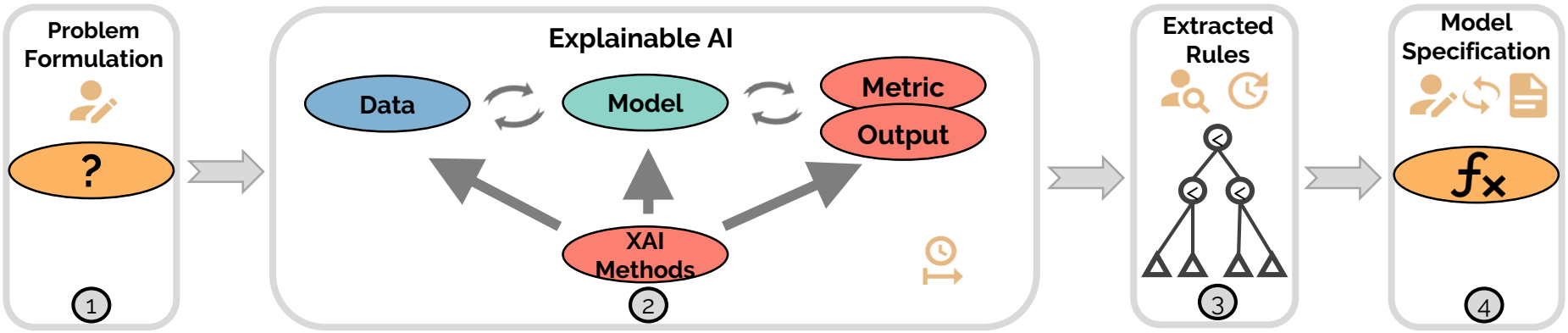
Research Opportunities

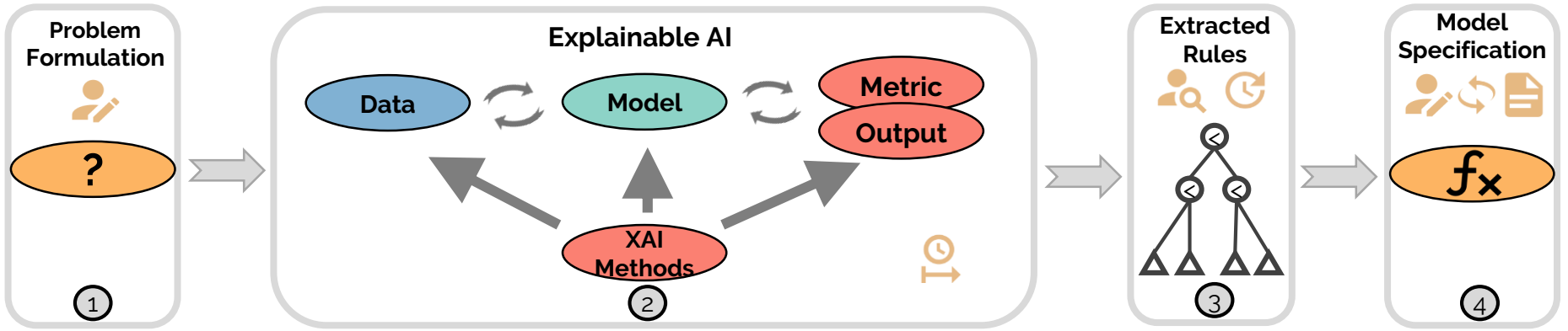
- Real-world applications of the workflow

- Scalable and interactive visualizations for decision rule lists

- Human-in-the-loop rule extraction methods

Summary


Problem Formulation
?
1
Explainable AI
Data
Model
Metric
Output
XAI Methods
2
Extracted Rules
3
Model Specification
fx
4



Thank you for your attention

Udo Schlegel
u.schlegel@uni-konstanz.de

Universität Konstanz

Data Analysis and Visualization