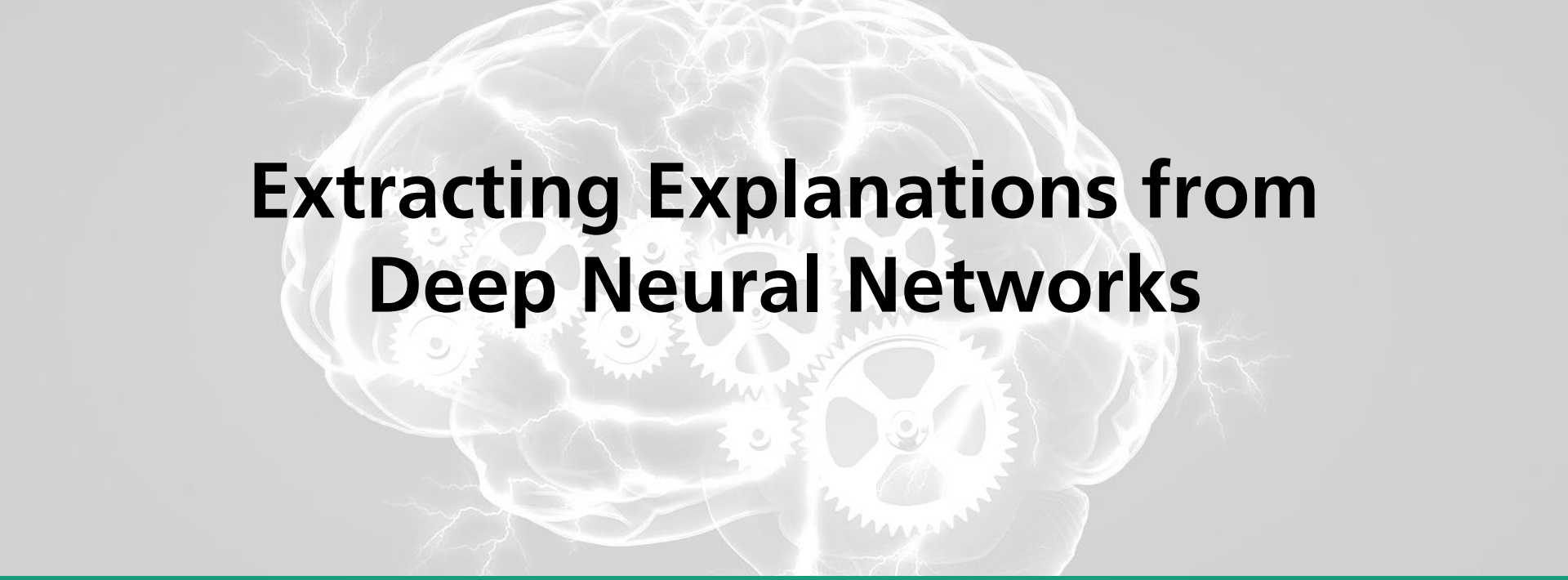


Extracting Explanations from Deep Neural Networks

Nina Schaaf, Marco Huber
marco.huber@ipa.fraunhofer.de

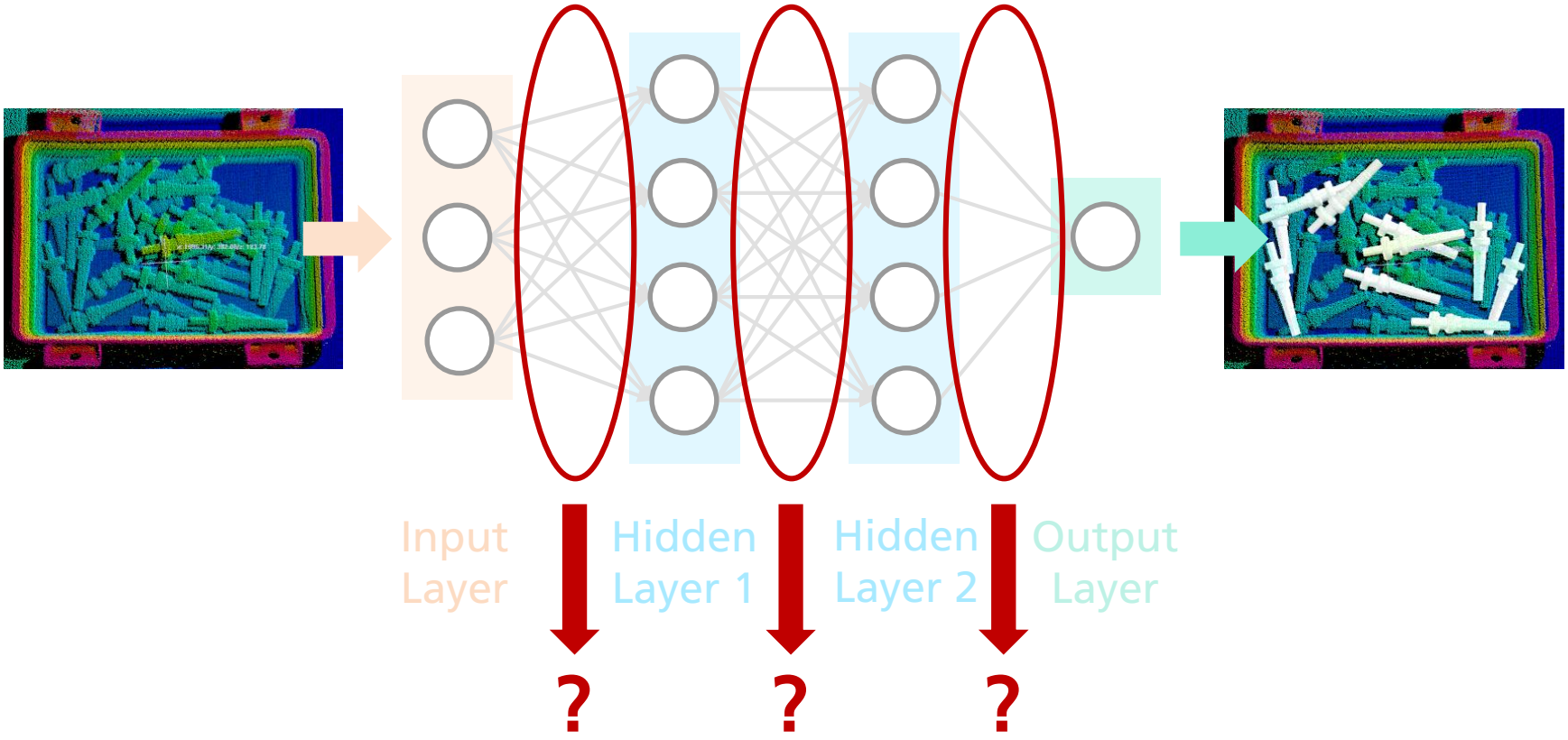
Fraunhofer IPA

Center for Cyber Cognitive Intelligence
Fraunhofer IPA
Stuttgart
https://www.ipa.fraunhofer.de

University of Stuttgart
Germany

Cognitive Production Systems
Institute of Industrial Manufacturing and Management
University of Stuttgart
https://www.iff.uni-stuttgart.de

Problem

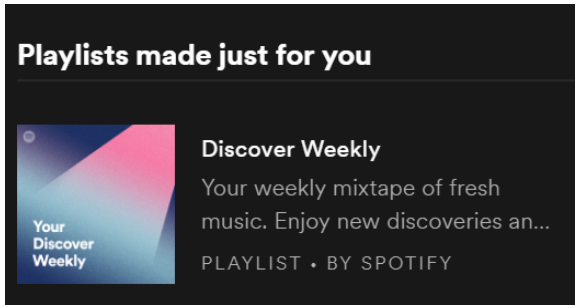


(Deep) Neural Networks are a **black box**.

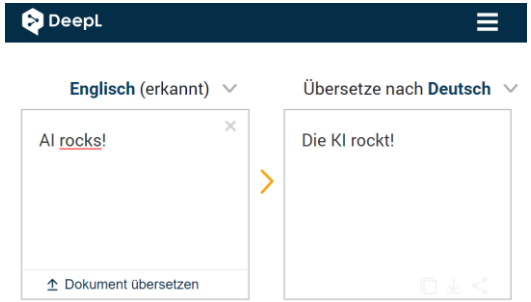
Fraunhofer IPA

Need for Explanability?

Non-critical applications



Music recommendation



Machine translation

Critical domains

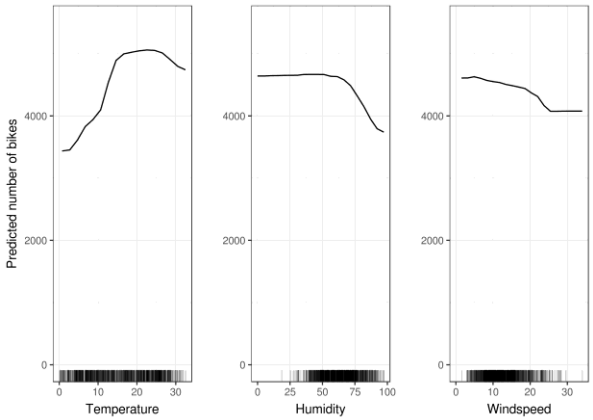
Discrimination/Bias

Law

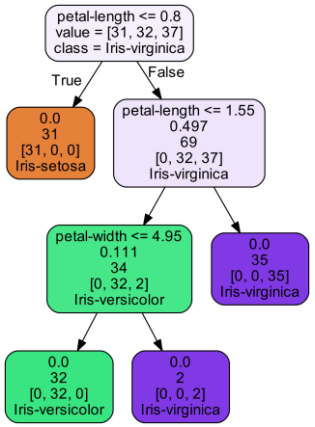
Fraunhofer IPA

Representation Formats

Heatmap (Layer-wise relevance propagation)



Partial-Dependence Plots



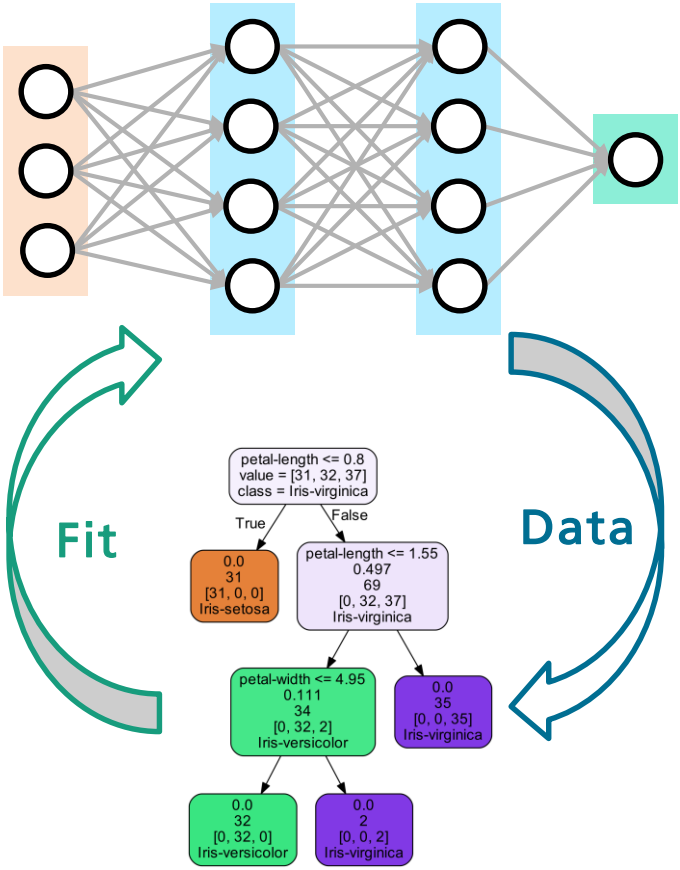
Decision Tree

IF temperature > 15 °C **AND** humidity < 60% **THEN** play tennis

Fraunhofer IPA

Objectives

- Representation format: **decision trees (rules)**
- Tree extraction from **deep** Neural Networks
- **Global** explainability
- **Classification** tasks



Naïve:

1. Considered trained neural network as ground truth
2. Fit decision tree to trained network
   - → Tree has limited representational power
   - → Low fidelity (fit to network)
   - → Low accuracy (fit to data)

Key Idea

- Regularization of network training in such a way that tree fitting is improved

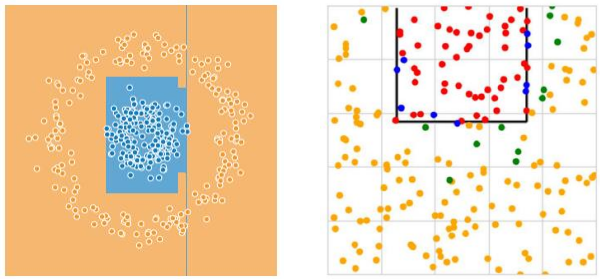
$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{n} E\left(y_i, f\left(X_i, W\right)\right) + \boldsymbol{\lambda \Omega(W)}$$

Classification loss

Regularization

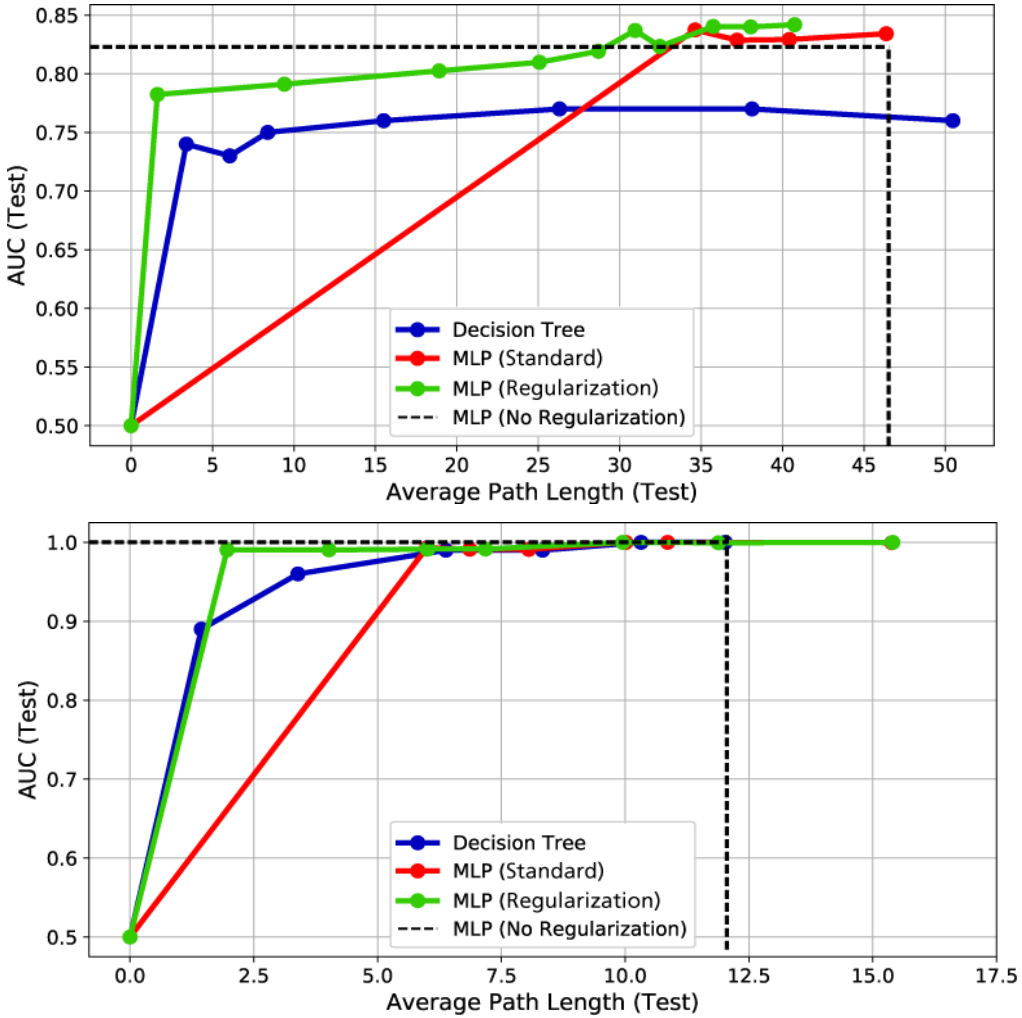
→ Easy-to-match decision boundaries for decision trees

→ Simple representation with broad feature coverage



Fraunhofer IPA

Results: Benchmark Datasets



Fraunhofer IPA

Thank You!

Prof. Marco Huber

marco.huber@ipa.fraunhofer.de
+49 711 970 1960

www.ipa.fraunhofer.de

Fraunhofer IPA