Explainable Artificial Intelligence (XAI) - Concepts, Methodologies & Challenges

European Robotics Forum
21st of March 2019, Bucharest, Romania

Nadia El Bekri
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB
Karlsruhe, Germany

© Fraunhofer

Fraunhofer IOSB

Explainable Artificial Intelligence (XAI)
Introduction

- Machine learning and deep learning applications are everywhere
  - Current algorithms achieve high accuracy
  - Opacity of the algorithms complicates their use
  - Models are difficult for humans to understand

- Explainable Artificial Intelligence (AI)
  - Programmed to describe the rationale and decision-making process in a way that can be understood by a human

Explainable AI makes complex models accessible to humans

© Fraunhofer

Explainable Artificial Intelligence (XAI)

Need for Explainable AI

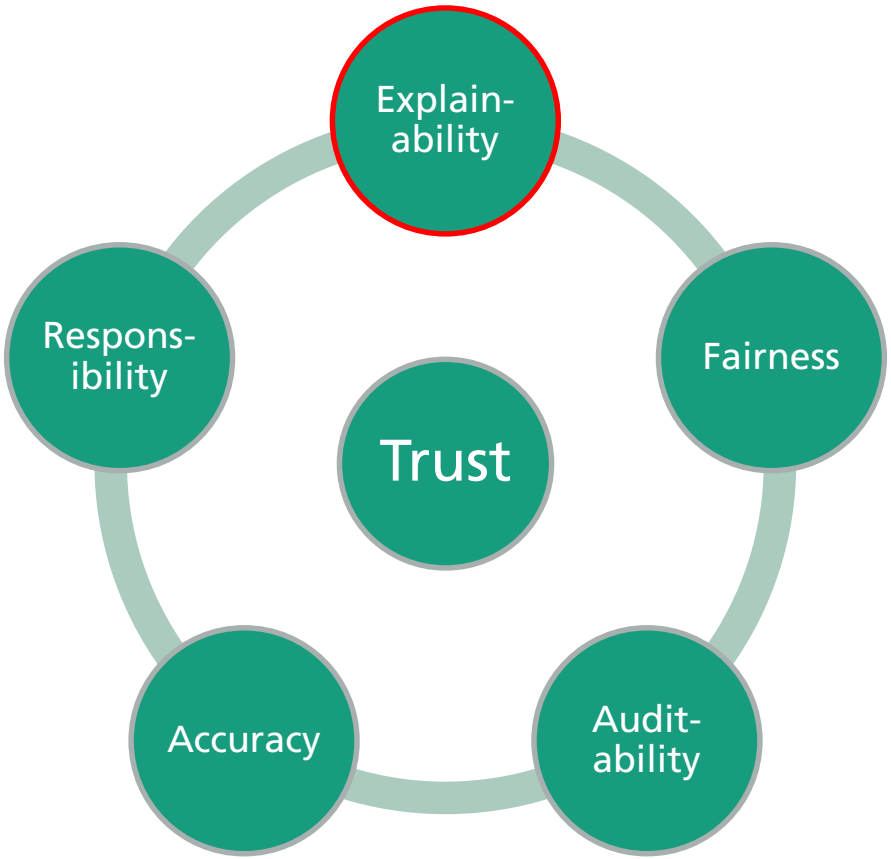
- **Verification** of the system
  - To optimize the model
  - To spot bias
- For **Legal** and **Ethical** Reasons
  - General Data Protection Regulation (GDPR)
- To **keep** the **human** still **in the loop**
  - Enhance **transparency**
  - Build and strengthen **trust**

Explainable AI is needed to access sensitive application domains

© Fraunhofer

Explainable Artificial Intelligence (XAI)

Trust in AI relies on multiple factors

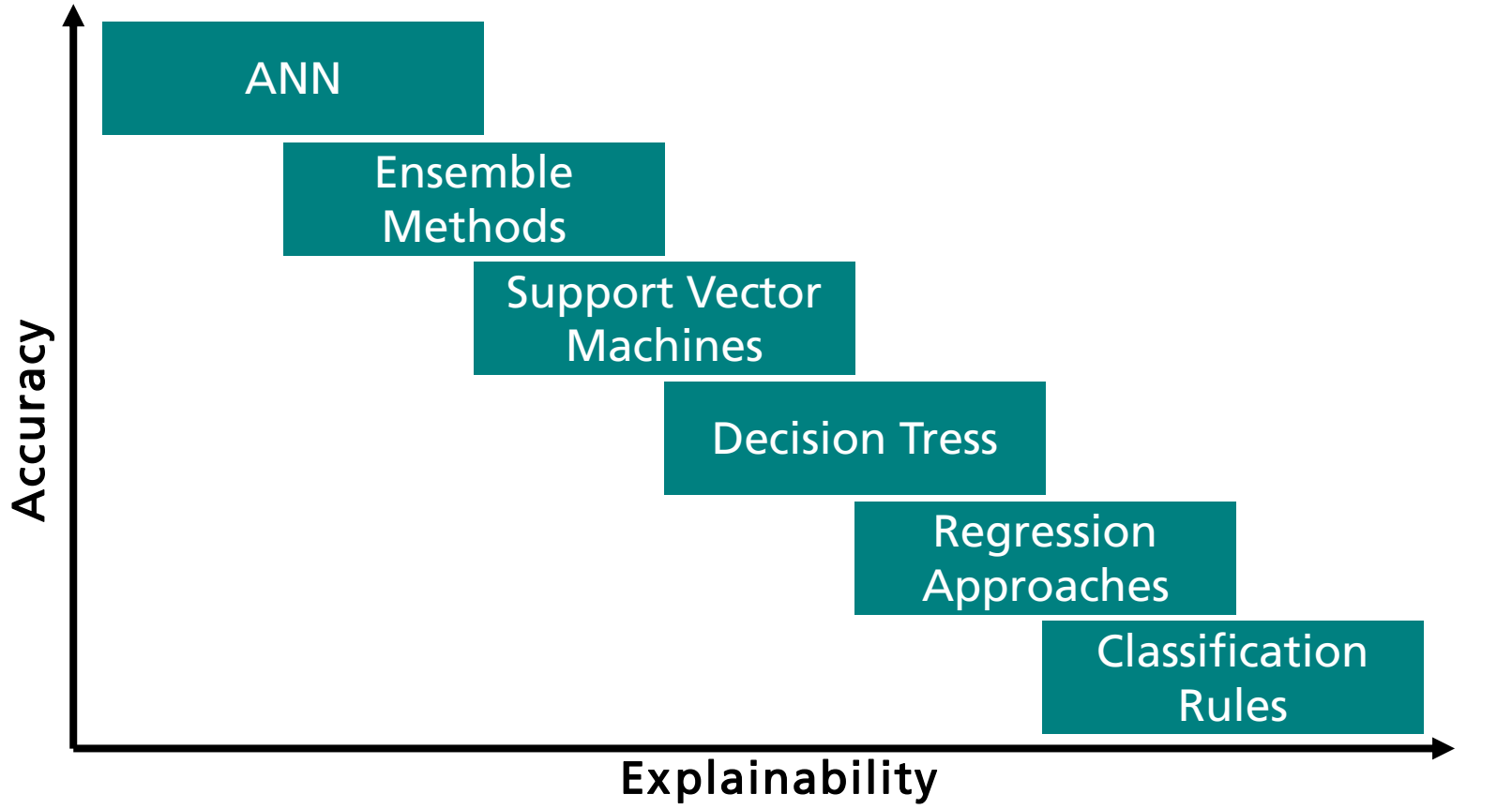


Explainable AI is one of the key enabler to trust AI

© Fraunhofer

Explainable Artificial Intelligence (XAI)

Trade-off between Accuracy and Explainability

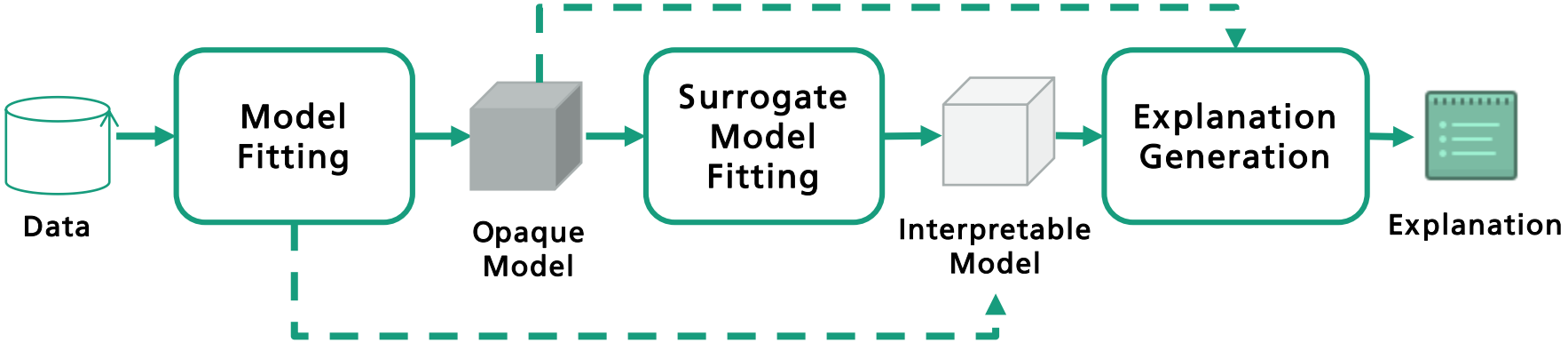


Higher accuracy is related to less explainability

© Fraunhofer

Explainable Artificial Intelligence (XAI)

Types of explainability approaches

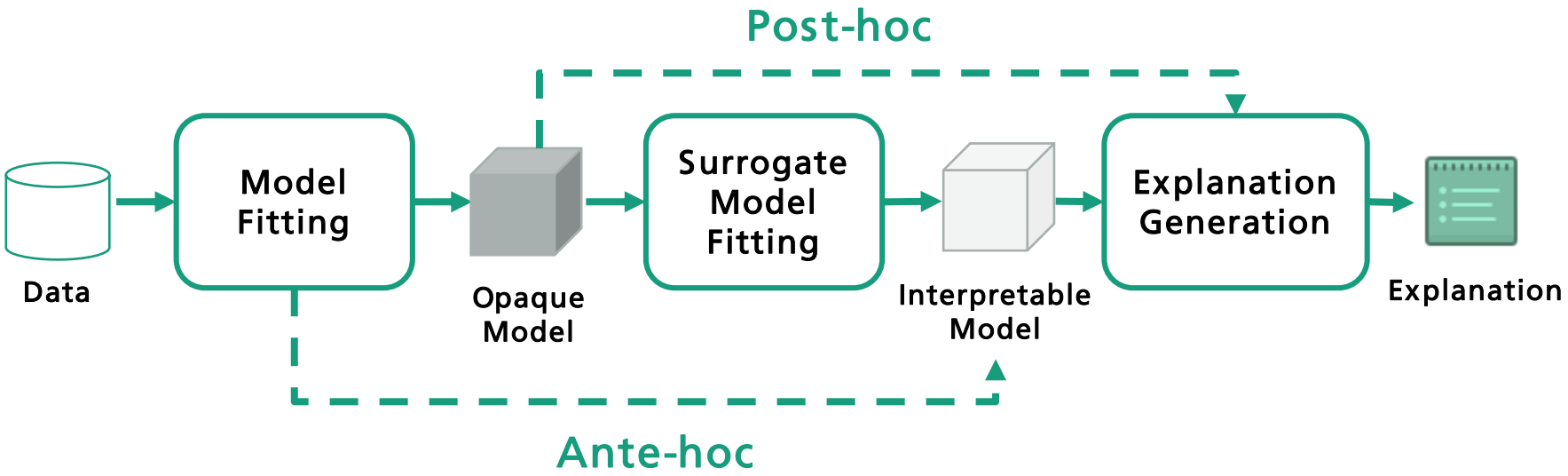


Three main approaches to create explanations

© Fraunhofer

Explainable Artificial Intelligence (XAI)

Taxonomy of Explainable AI

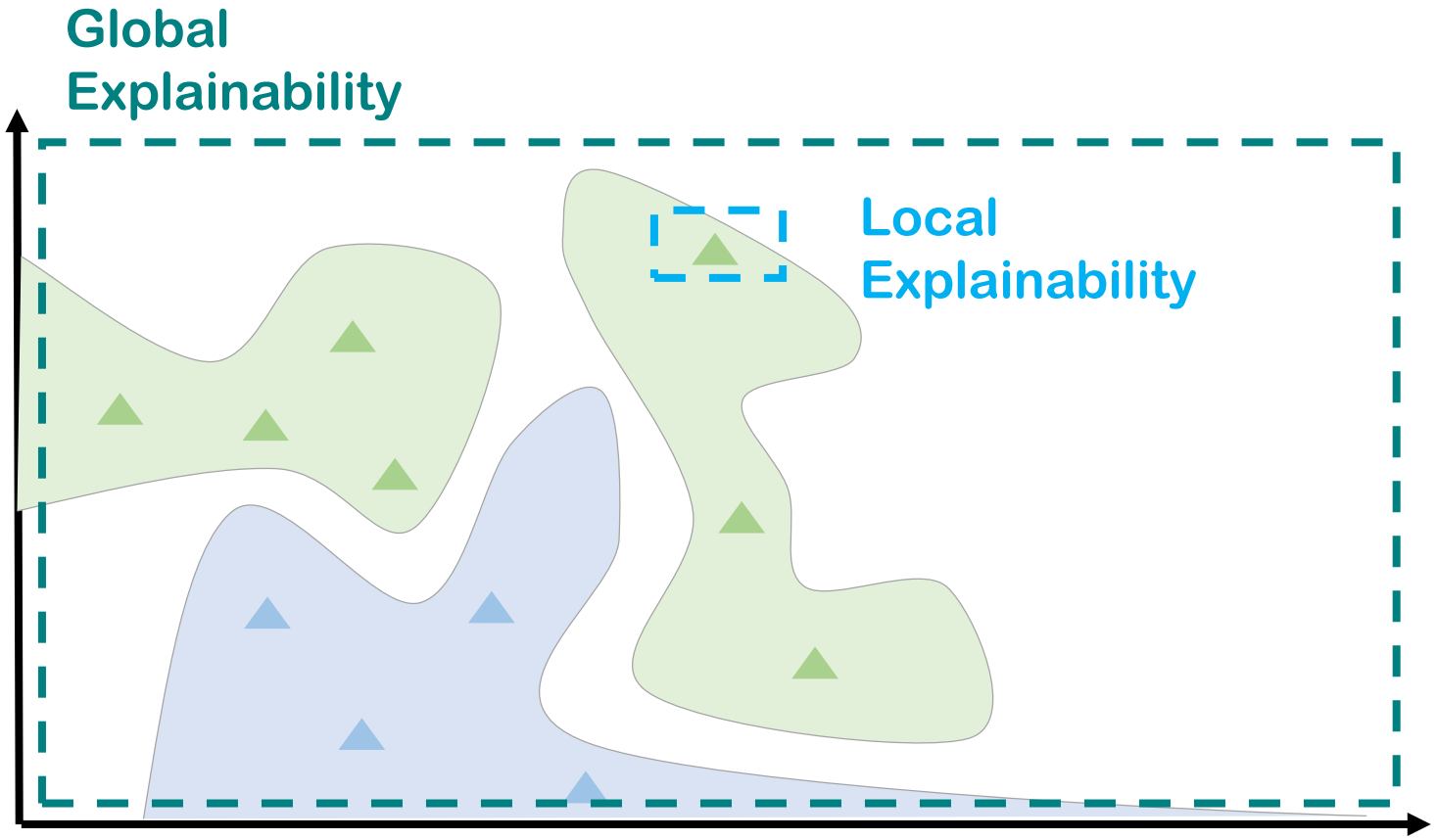


Ante-hoc approaches are trained directly on the data,
post-hoc approaches use the opaque model

© Fraunhofer

Explainable Artificial Intelligence (XAI)

Taxonomy of Explainable AI

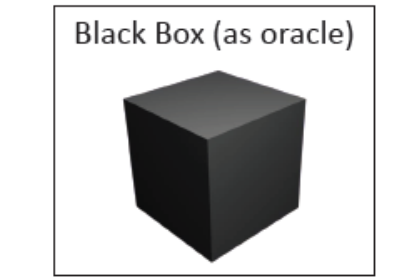


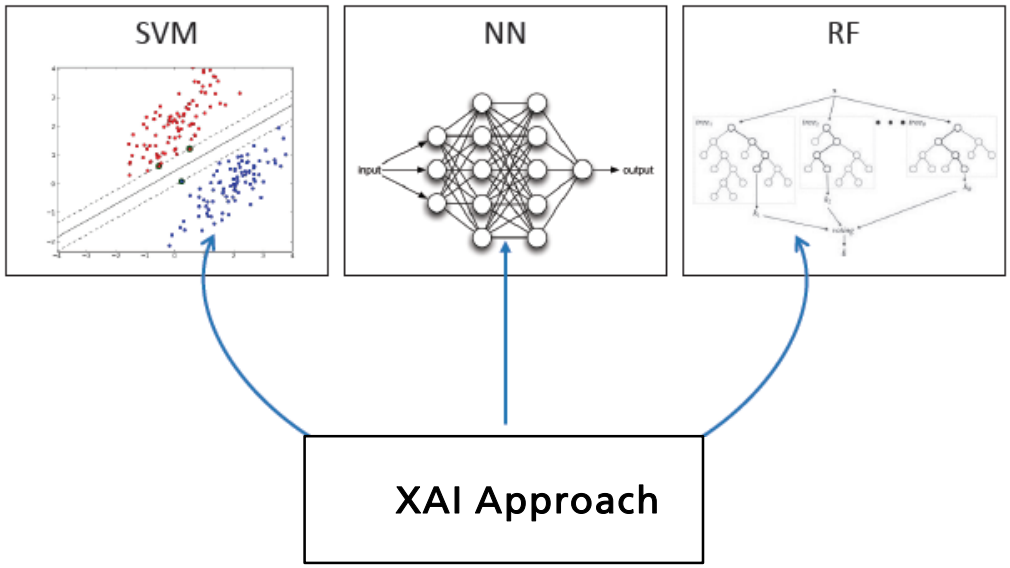
Approaches can either explain the entire model or just certain parts of it

© Fraunhofer

Explainable Artificial Intelligence (XAI)

Taxonomy: Model-agnostic vs. Model-specific



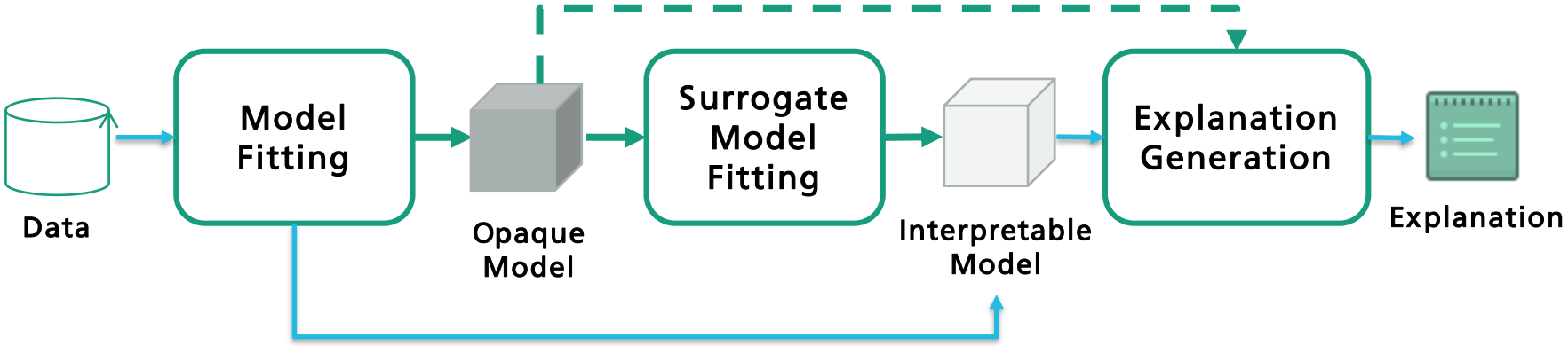


Model-agnostic approaches can be used with any black box, model-specific use the internal structure of the model

© Fraunhofer

Explainable Artificial Intelligence (XAI)

Explainability categories



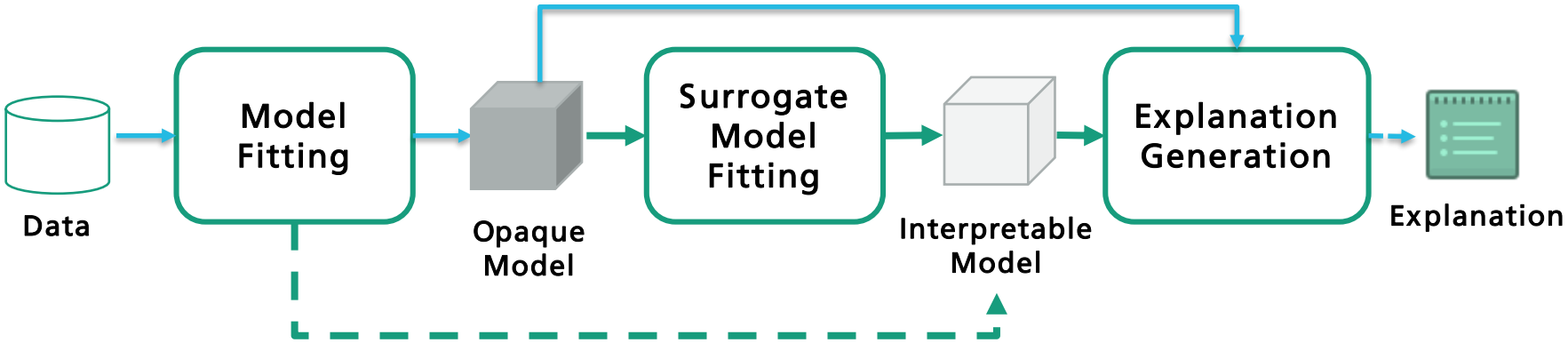
1) Interpretable model
   - Naturally (e.g. small decision trees)
   - By design (interpretability is forced into the model e.g. via sparsity constraints)
   - Ante-hoc, global, model-specific

Goal: Provide interpretable model directly trained on the data

© Fraunhofer

Explainable Artificial Intelligence (XAI)

Explainability categories



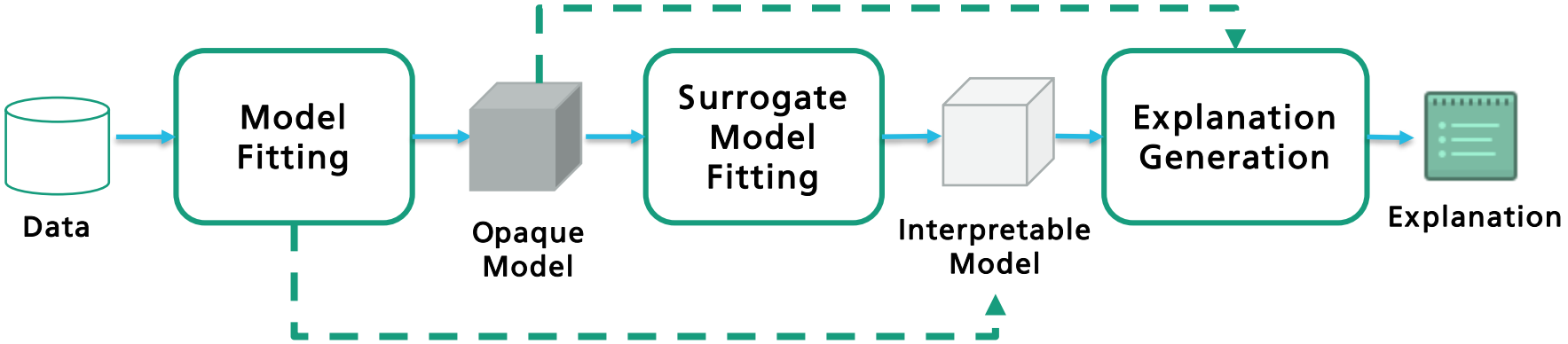
2) Data-driven model explanation
   - E. g based on sensitivity analysis, by observing the change in the predictions when varying the input
   - Post-hoc, local & global, model-specific & model-agnostic

Goal: Provide a representation (visual or textual) for understanding e.g. why the black box returns certain predictions more likely than another

© Fraunhofer

Explainable Artificial Intelligence (XAI)

Explainability categories



3) Data-independent model explanation
   - Interpretable model approximates the black box
   - post-hoc, local & global, model-specific & model-agnostic

Goal: Provide an interpretable model which is able to mimic the behavior of the black box

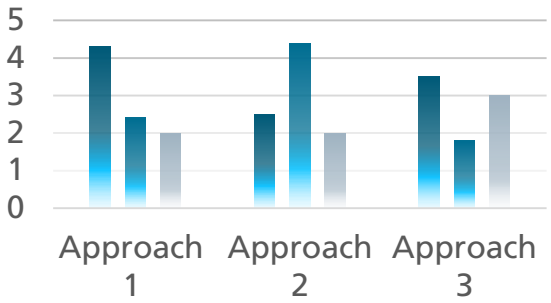
© Fraunhofer

Explainable Artificial Intelligence (XAI)

Open Challenges for XAI

- Interpretable vs. black box models

- Procedure to compare XAI approaches



- XAI approaches with focus on Deep Learning and Reinforcement Learning

Explainable AI paves the way for building trust in sensitive domains

© Fraunhofer

Thank You!

Contact Information

Nadia El Bekri, M.Sc.
Fraunhofer IOSB
Fraunhoferstraße 1
76131 Karlsruhe
Germany

nadia.elbekri@iosb.fraunhofer.de

+49 721 6091-619

© Fraunhofer

Fraunhofer IOSB