1 Interpretability

Motivation: A single metric is an incomplete description of most real-world tasks.

1. Improve models
2. Justify models
   a. Creators
   b. Operators
   c. Executors
   d. Decision
   e. Auditors
   f. Data Subjects

   } Stakeholders of a AI System
3. Discover Insights

1.1 What is Interpretability?

Definition: Interpretability is:

1. Degree to which a human can understand the cause of a decision
2. Where a user can correctly and efficiently predict the method's results.
3. Science of understanding AI models from the inside out.

1.2 Types of Interpretability

Summary:

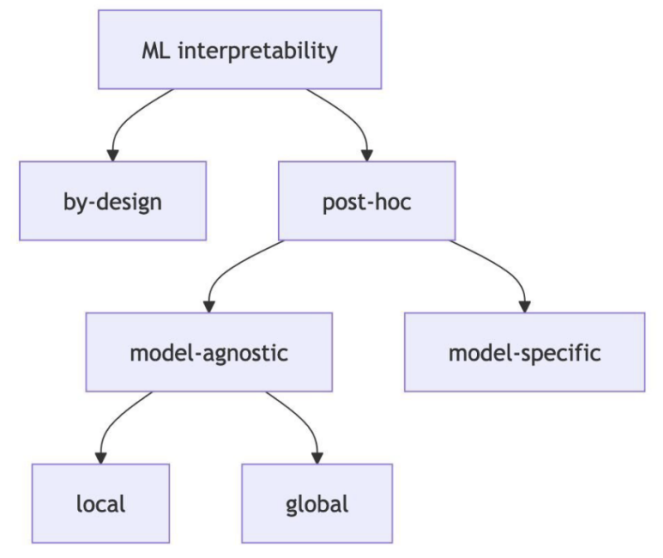

Figure 1

- **ML Interpretability:**
  - **By-design:** Interpretability built directly into the model (e.g., decision trees, linear models).
  - **Post-hoc:** Interpretability techniques are applied after the model is trained, without altering the model.
    - **Model-agnostic:** Interpretation methods that can be applied to any model.
      - **Global:** Provides an overall understanding of the model's behavior across the entire dataset.
      - **Local:** Explains the model's prediction for a specific input instance.
    - **Model-specific:** Interpretation methods that are tailored to specific models.

1.3 Attribution

1.3.1 Issues

1.3.2 Examples

2 Interpretability