1 Interpretability

> **Motivation**:

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

1.2.1 Neural Network Analogies

1.3 Attribution

1.3.1 Issues

1.3.2 Examples

2 Interpretability