## Learning Outcomes and Prerequisites

**In this module, you will:**

* Identify different methods of achieving Explainable AI
* Recognize the role of open-source AI Explainability 360 toolkit in supporting explainability
* Describe various features and capabilities of the open-source AI Explainability 360 toolkit

**To be successful in this module, prior knowledge is recommended in:**

* Module 1: The Big Picture of Trustworthy and Explainable AI
* Data Science / Machine Learning workflows
* Evaluation metrics

## Interactive Demo

Walk through the process of explaining models to consumers with different personas in an interactive web-demo that shows a sample of capabilities available in the toolkit

Click <https://aix360.mybluemix.net/> link to open resource.

* [**Module 1: The Big Picture of Trustworthy AI and AI Explainability (Start Here)**](https://learn.ibm.com/mod/url/view.php?id=181788&forceview=1#collapse1)
* [What Will You Learn In This Course?](https://learn.ibm.com/mod/page/view.php?id=175920)
* [Welcome to the Course!](https://learn.ibm.com/mod/video/view.php?id=182260)
* [Get To Know Your Instructors!](https://learn.ibm.com/mod/page/view.php?id=171805)
* [Learning Outcomes and Prerequisites](https://learn.ibm.com/mod/page/view.php?id=171806)
* [The Need for Trusted AI](https://learn.ibm.com/mod/video/view.php?id=176026)
* [Explainable AI](https://learn.ibm.com/mod/video/view.php?id=182261)
* [Michael Hind of IBM Research on AI Explainability (Optional)](https://learn.ibm.com/mod/video/view.php?id=181677)
* [Module 1 Summary](https://learn.ibm.com/mod/page/view.php?id=181112)
* [Check for Understanding: The Big Picture of Trustworthy AI and Explainable AI](https://learn.ibm.com/mod/quiz/view.php?id=181111)
* [**Module 2: Methods for Explainable AI & Overview of AI Explainability 360 Toolkit**](https://learn.ibm.com/mod/url/view.php?id=181788&forceview=1#collapse2)
* [Learning Outcomes and Prerequisites](https://learn.ibm.com/mod/page/view.php?id=181115)
* [Methods for Explainable AI](https://learn.ibm.com/mod/video/view.php?id=182262)
* [IBM AI Explainability 360 for Explainable AI](https://learn.ibm.com/mod/video/view.php?id=181678)
* [Overview of AI Explainability 360](https://learn.ibm.com/mod/video/view.php?id=182266)
* [Interactive Demo](https://learn.ibm.com/mod/url/view.php?id=181787)
* [Explore AI Explainability 360 Open Source Toolkit](https://learn.ibm.com/mod/url/view.php?id=181788)
* [Terms Used in this Module](https://learn.ibm.com/mod/page/view.php?id=181792)
* [Module 2 Summary](https://learn.ibm.com/mod/page/view.php?id=181790)
* [Check for Understanding: Methods for Explainable AI and IBM AI Explainability 360](https://learn.ibm.com/mod/quiz/view.php?id=181791)
* [**Module 3: Hands-on with AI Explainability**](https://learn.ibm.com/mod/url/view.php?id=181788&forceview=1#collapse3)

## Explore AI Explainability 360 Open Source Toolkit

Demos, Tutorials, APIs and Additional Guidance

This extensible open source toolkit will help you to support explainability and interpretability of your machine learning models.The AI Explainability 360 [Python package](https://pypi.org/project/aix360/) includes algorithms that span the different dimensions of ways of explaining along with proxy explainability metrics.The AI Explainability 360 [interactive demo](https://aix360.mybluemix.net/data) provides a gentle introduction to the concepts and capabilities by walking through an example use case from the perspective of different consumer personas. The [tutorials](https://aix360.mybluemix.net/resources#) and other [notebooks](https://github.com/IBM/AIX360/blob/master/examples) offer a deeper, data scientist-oriented introduction. The [complete API](https://aix360.readthedocs.io/) is also available.Being a comprehensive set of capabilities, it may be confusing to figure out which class of algorithm is most appropriate for a given use case. To help, we have created some [guidance material](https://aix360.mybluemix.net/resources#) that can be consulted.We have developed the package with extensibility in mind. We encourage the contribution of your explainability metrics and algorithms. Please join the [community](https://aix360.mybluemix.net/community) to get started as a contributor.

Click <https://aix360.mybluemix.net/> link to open resource.

## Terms Used in this Module

**Black box model**   
**A complicated model that consumers are not easily able to understand, such as a deep neural network.**

**Classifier**   
**A model that predicts categorical labels from features.**

**Consumer**   
**A human receiving an explanation.**

**Directly interpretable model**   
**A model that consumers can usually understand, such as a simple decision tree or Boolean rule set.**

**Explanation**   
**A reason or justification for the predicted label. Some experts differentiate explanations from interpretations. Explanations come from surrogate models and interpretations come from the models themselves.**

**Feature**   
**An attribute containing information for predicting the label.**

**Global explanation**   
**An explanation for an entire model.**

**Label**   
**A value indicating the outcome or category for a sample.**

**Local explanation**   
An explanation for a sample.

**Machine learning**   
**A general approach for determining models from data.**

**Model**   
**A function that takes features as input and predicts labels as output.**

**Persona**   
**The role of the consumer, such as a decision maker, regulator, data scientist, or patient.**

**Post hoc explanation**   
**An explanation coming from a model that approximates a black box model. The experts that differentiate the terms explanation and interpretation limit the term explanation only to post hoc explanation.**

**Representation**   
**A mathematical transformation of data into features suitable for models.**

**Sample**   
**A single data point, instance, or example.**

**Score**   
**A continuous valued output from a classifier. Applying a threshold to a score results in a predicted label.**

**Supervised learning**   
**Determining models from data having features and labels.**

**Training data**   
**A dataset from which a model is learned.**

**Unsupervised learning**   
**Determining models or representations from data having only features, no labels.**

## Module 2 Summary

**1.**There are many ways to explain how machine learning makes predictions, including:

*Data vs. Model*: Based on if the explanation is needed for the data or the model

*Direct interpretability vs. Post hoc explanation:*

* Directly interpretable models are model formats such as decision trees, Boolean rule sets, and generalized additive models, that are fairly easily understood by people and learned straight from the training data.
* Post hoc explanation methods first train a black box model and then build another explanation model on top of the black box model. The second distinction is global vs. local explanations.

*Global vs Local:* Global explanations are for entire models whereas local explanations are for single sample points

**2.** IBM researchers have created the **Open Source AI Explainability 360** **toolkit** , which is a comprehensive Toolkit for making models more explainable

***AIX360 contains model explanation methods for all of these categories of explanation***

F,{41550040-d3d1-4aa9-8706-f32ed0646fb8}{91},13,8.208333

**3.**Lesson on AI Explainability 360 overview provides a gentle introduction to the concepts and capabilities through a demo and the tutorial