eXplego: An interactive tool that helps you select appropriate XAI-methods for your explainability needs

Martin Jullum¹*, Jacob Sjødin², Robindra Prabhu² and Anders Løland¹

¹Norwegian Computing Center, P.O. Box 114, Blindern, NO-0314 Oslo, Norway
²NAV IT Utvikling og Data, Arbeids- og velferdsdirektoratet, Fyrstikkalléen 1, 0661 Oslo, Norway

Abstract
The growing demand for transparency, interpretability, and explainability of machine learning models and AI systems has fueled the development of methods aimed at understanding the properties and behavior of such models (XAI). Since different methods answer different explainability questions, it is crucial to understand the kind of explanation the different XAI-methods provide, and in what situations they should be used. We introduce eXplego, an interactive tree-structured tool designed to assist users in selecting the most suitable XAI method for their use case. eXplego prompts users to answer questions regarding the type of explanation they seek, guiding them along the branches of the decision tree for further inquiries. After 2-5 questions, the tree reaches one of its leaves to suggest an XAI method aligned with the user’s explainability need. The tool also provides helpful practical examples, simplified descriptions of the suggested method’s functionality and interpretability, points to consider when using the method, and links to the paper introducing the method, additional resources, and software implementations. The tool is developed from an in-depth study to discern the characteristics of the most prominent methods and the nature of the explanations they provide. We believe eXplego will help streamline the process of XAI method selection and contribute to the practical implementation of XAI in various domains. The tool is available at explego.nr.no.

Keywords
XAI, Tool, Interactive, Methodology selection, Features, Model, Prediction, Data distribution

1. Motivation and scope
A plethora of XAI methods and variations thereof have been proposed in recent years, typically grounded in formal and narrowly defined mathematical notions of interpretability [1, 2, 3, 4, 5]. Different XAI methods address different aspects of model behavior, and may therefore provide very different results without being "wrong." Further, an increasing number of XAI methods are now available as low-entry software implementations [6, 7]. Such software packages facilitate XAI adoption, but due to the wide variety of methods available, they also pose a conundrum:
Which XAI method is best suited for my specific case or application? The lack of guidance on which questions an XAI method can address risks prioritizing ease of use and familiarity in user choices.

In recent years, various XAI taxonomies have been proposed [8, 9, 10]. While such categorization schemes organize the XAI method landscape, the principle objective of such taxonomies is rarely to assist the developer in selecting a suitable XAI method for their specific use case and explanation requirement – i.e. they often lack the practical dimension [11, 12, 13, 14, 15].

The recent taxonomy review paper [10] identifies challenges with the current state of the XAI field, and provides three concrete suggestions for overcoming them. One of these is to create a decision tree to guide method selection. In response, we have developed eXplego, an interactive tree-structured tool to help guide developers and practitioners in their assessment and choice of appropriate XAI methods, directly accessible in the web browser at explego.nr.no. The name eXplego is (combined with 'explain') derived from the Greek word 'eklégo', meaning the deliberate act of choosing or making a thoughtful selection. The tool draws inspiration from the "Fairness tree" [16], a tool designed to assist in the selection of metrics to assess bias and fairness in ML-models. Similarly, eXplego provides navigation to various XAI methods through a series of practical desiderata the users must consider in their selection of XAI methods.

We have restricted ourselves to post-hoc, model-agnostic explanation methods for tabular data models in eXplego. While explaining text and image-based models is important, their data formats require other types of questions to identify an appropriate explanation method. Moreover, we believe the need for a navigation tool is most pressing for post-hoc, model-agnostic methods, precisely because they can be widely applied – hence this additional restriction.

2. The eXplego tool

The eXplego tool prompts the users to answer questions regarding the type of explanation they seek, guiding them along the decision tree for further inquiries. Each question also comes with practical in-place examples. A section of the eXplego tool is shown in Figure 1. After 2-5 questions, the tool suggests an XAI method aligned with the user’s explainability needs. The leaves also contain a short summary of its use and interpretability, a list of method usage considerations, and links to methodological papers, additional resources, and software implementations.

eXplego is developed based on methodological and practical XAI experience, and an extensive study of the most prominent XAI methods and the kind of explanations they provide. The methods included in eXplego are listed below, along with brief justifications for their placement in the tree.

**Permutation feature importance** [17]: Whole model → Features → Observing the features → One
feature at a time
Measures the value of observing the features in the whole model as it permutes features and measures the change in model performance. Since it permutes one feature at a time it does not consider joint efforts/dependence among features.

SAGE [18]: Whole model → Features → Observing the features → Joint efforts
Decompose model loss onto features to explain the whole model in terms of the features. As Shapley values fix some subsets of features while imputing the others, it explains the value of observing the features considering their joint efforts.

ALEPlots [19]: Whole model → Features → Changing the feature values
Per-feature plots show changes in the ‘average’ prediction as one feature is altered. Thus, they explain how the whole model reacts to changes in the feature values.

Data Banzhaf [20]: Whole model → Training observations
Decomposes a performance score for the whole model on training observations. Similarly to Shapley values, subsets of observations are interchangeably fixed, while others imputed. Hence, it explains the value of observing the training observations.

Conditional Shapley values [21]: Specific predictions → Features → Observing the features → Joint efforts among features → Obey data distribution
Explains specific predictions in terms of features, by decomposing them onto the features. As subsets of observations are interchangeably fixed/imputed, it explains the value of observing the features where joint efforts are considered. Properly estimated conditional expectations ensure feature dependence is accounted for.

Marginal Shapley values [22]: Specific predictions → Features → Observing the features → Joint efforts among features → Ignore feature dependence
Exactly like Conditional Shapley values, but estimates the conditional expectations with a simpler method ignoring the feature dependence.

PredDiff (first order) [23]: Specific predictions → Features → Observing the features → One feature at a time
Explains how specific predictions are affected by observing single features (assuming others known) by measuring how the prediction changes as they are replaced by conditional expectations.

Anchors [24]: Specific predictions → Features → Changing the observed feature values → Categorical decision → Same decision
By providing feature space regions where a decision based on a prediction is unchanged, it explains specific predictions in terms of changes in features values for the same decision that was reached by the specific prediction.

Counterfactual explanation [25]: Specific predictions → Features → Changing the observed feature values → Categorical decision → Different decision
Explains how specific predictions can reach a different categorical decision (based on the prediction score) by providing examples of (minimal) changes to the feature values that would give the desired decision.

LIME [26]: Specific predictions → Features → Changing the observed feature values → Continuous prediction → Joint efforts among features
By fitting a local surrogate model to a joint feature set sampled around a prediction, the method explains specific continuous predictions directly in terms of changes in the feature values, while accounting for joint efforts among features.
ICE [27]: Specific predictions → Features → Changing the observed feature values → Continuous prediction → One feature at a time
Explain specific predictions in terms of changes to one feature value at a time, by plotting individual prediction scores against single altered features.

Shapley values for cluster importance [28]: Specific predictions → Training observations → Including the observations
Uses Shapley values to explain the value of including (clusters of) training observations by decomposing specific predictions onto the different clusters.

Influence functions for perturbing training data [29]: Specific predictions → Training observations → Changing the observed values
Explains changes in observed values in the training data for specific predictions by measuring loss change when perturbing features in the training observations.

To the best of our knowledge, the eXplego tool is unique in its form. That said, the structuring proposed in IBM’s Explainability 360 Toolkit [30] bears conceptual resemblance. eXplego differs in the following key points: eXplego is more comprehensive, covers a wider range of XAI methods, and is geared towards developers in that question prompts are more informed by the technicalities of the XAI methods. As explained above, eXplego is also interactive, and provides both in-place examples to help the user answer the questions, and detailed information beyond the method’s name in the leaves.

Since our tool is restricted to models for tabular data, we encourage other researchers to apply our format to other scenarios and model types, such as text and images. Further, our tool is limited to identifying quantitative XAI methods most befitting different use cases. Privacy, contextual and normative dimensions [31], also need to be considered when providing adequate and trustworthy explanations [5]. Questions related to compliance with any legal framework, like GDPR, are neither addressed.

Finally, our tool has been built with the open-source diagramming application draw.io. The source code for our tool is available at github.com/NorskRegnesentral/explego. Feedback and suggestions for new methods are all welcome and can be submitted by opening an issue in the GitHub repository.

3. Expected contribution to the XAI community

It is our impression that the practical difficulty of matching explainability needs with existing XAI methods is underestimated, and eXplego is a practical tool that can guide its users in selecting an appropriate explanation method.

The tool can inspire future research: As [32] puts it: “Despite the recent resurgence of explanation and interpretability in AI, most of the research and practice in this area seems to use the researchers’ intuitions of what constitutes a ‘good’ explanation.” The tool can also be used to highlight explainability questions that no XAI method addresses. For instance, that our tree lacks a question addressing feature dependence for global explanations of feature observation with joint efforts, identifies that the SAGE method lacks a counterpart using conditional Shapley values. Finally, we believe eXplego will streamline XAI method selection and contribute to practical implementation of XAI in various domains.
References


