Rule-based explanations provide simple reasons explaining the behavior of machine learning classifiers at given points in the feature space. We take advantage of the connection between the inherent definability of rule-based explanations and definability in topology to develop a general framework to represent explanations based on existing explanation algorithms.

Contributions
- We present a novel framework of explainability for classifiers based on existing explanation algorithms.
- We characterize explainability as a topological property relative to an explanation scheme, i.e. relative to a choice of explanation shape and a measure of explanation size. We conjecture that all classifiers “in-the-wild” satisfy this notion of explainability.
- Employing our framework, we identify two principles for explanation algorithms that apply both theoretically and in practice.

Explanation Topology
Rule-based explanations are regions of the feature space that satisfy some predicate or definable property \( \varphi \). We restrict to predicates satisfying the following properties:
- If two explanations for a given point overlap, then there exists an explanation in their intersection covering the point.
- Each point is covered by an explanation.

Sets belonging to \( \mathcal{T}_\varphi \) are called open. The main result of this paper characterizes explainability in terms of open sets and small sets.

Main Result (Explainability is a simple topological property)

**Theorem:** A classifier \( f : X \rightarrow Y \) is explainable for scheme \((X, \varphi, \mu)\) if and only if, for \( y \in Y \), there exists open set \( O_y \in \mathcal{T}_\varphi \) such that \( f^{-1}(y) = O_y \cup E_y \) and \( E_y \in \mathcal{T}_{\varphi} \)-meagre, \( \mu \)-null.

Application: Ensembles
Ensembles aggregate predictions from a collection of classifiers and are commonly used in practice e.g. Random Forests and XGBoost. Ensembles are often viewed as complex, whereas their constituent classifiers are weak or simple. We show that if the constituent classifiers are explainable for a given scheme then their ensemble is explainable.

**Theorem:** If \( f_1, \ldots, f_k \) are classifiers explainable for explanation scheme \((X, \varphi, \mu)\) and \( f \) is an ensemble of \( f_1, \ldots, f_k \), then \( f \) is explainable for \((X, \varphi, \mu)\).