# Constructive Empiricism for Explainable AI

Sebastian Müller[1,2] (✉), Vanessa Toborek[1,2], Eike Stadtländer[1,2], Tamás Horváth[1,2,3], Brendan Balcerak Jackson[2,1], and Christian Bauckhage[1,2,3]

[1] University of Bonn, Bonn, Germany
[2] Lamarr Institute for Machine Learning and Artificial Intelligence, Bonn, Germany
[3] Fraunhofer IAIS, Sankt Augustin, Germany
semueller@uni-bonn.de

**Abstract** We explore what it means to build a scientific "theory" of a black-box model, drawing on van Fraassen's Constructive Empiricism (CE), and demonstrate how such a theory can be used for explainable AI (XAI). A *scientific theory* is more than just an explanation: it not only has value in its own right, but also serves as a robust framework for answering different questions. According to CE, a theory must be both empirically adequate (i.e., accurate with respect to observed data) and shaped by pragmatic virtues, such as user preferences. These criteria align closely with the needs of XAI, which require fidelity and comprehensibility. We turn CE's core notion of *empirical adequacy* into three concrete criteria: consistency, sufficient predictive performance, and algorithmic adaptability. We develop the Constructive Box Theorizer (CoBoT) algorithm within this framework. As a proof of concept, we present a qualitative discussion showing that CoBoT can produce empirically adequate theories and illustrate the utility of a *theory* in XAI.

**Keywords:** Constructive Empiricism · Explainable AI · Rule Based Explanations · Philosophy of Science · Machine Learning

## 1 Introduction

In this paper we propose to view explanations in Explainable AI (XAI) as a form of *scientific explanation* that draws from an underlying *theory*. This perspective allows us to step outside the established conceptual frameworks of XAI. XAI methods are typically classified by their scope (global explanations aim to capture the entire behavior of the model, while local explanations focus on individual predictions) or the specific questions they address [8,10,21]. However, a scientific theory is more than a single explanation: It is developed independently of particular questions and can be evaluated on its own merits. This leads us to our central question for this work: *If we treat a black-box classifier as the subject of scientific inquiry, what does it look like to build a scientific theory of its behavior and how does it support explanation?*

To answer this question, we build on the conception of science and scientific explanation known as *Constructive Empiricism* (CE), as formulated by the philosopher of science Bas van Fraassen in [4]. We find that CE naturally aligns with the core concerns of fidelity and comprehensibility in XAI. Fidelity describes the degree of "truthfulness" of the explanatory information with respect to the underlying black-box. Comprehensibility ensures that this information is accessible and meaningful to users [2,14]. We find that this aligns directly with the notions of *empirical adequacy* and *pragmatic virtues* that lie at the heart of CE.

According to CE, a scientific theory explains directly observable phenomena by positing underlying processes and structures that are not directly observable, and by describing their possible states and interactions. The aim of a scientific theory is to be *empirically adequate*, which requires a theory to be consistent with all available observations. The scientist *accepts* the theory, which entails the *belief* in its empirical adequacy, and a *commitment* to use the conceptual resources of the theory to explain *future* phenomena when they are encountered. In CE, the central aim of science is to produce useful insights. The scientist is thus free to shape the theory the way that best serves their needs, with those guiding considerations captured by the concept of *pragmatic virtues*. For XAI, we argue that empirical adequacy builds a foundation for justified trust, while pragmatic virtues allow the user's needs to inform the shape of the theory and the explanations. Our contributions are threefold:

1) We operationalize empirical adequacy for XAI by establishing three criteria: (i) to address the agreement with observations we require *consistency* of the theory with its training data, (ii) to subsequently justify acceptance, we require sufficient *predictive performance* on unseen data and (iii) computational *adaptability* that guarantees that the theory can be revised algorithmically to incorporate new observations. This mimics the scientific process of continuously adjusting the theory to account for new observations. Explanations *derived* from the theory are thus *always* based on the most complete and up-to-date representation of the model's behavior.
2) We demonstrate the utility of the CE conceptual framework to XAI by presenting a novel algorithm, called Constructive Box Theorizer (CoBoT), that incrementally builds a system of axis-aligned hyperrectangles (i.e., boxes) as a "theory of the black-box". We show how CoBoT can be integrated within the CE framework and assess its performance against a range of criteria for XAI algorithms that naturally arise within this conceptual framework. In short: Empirical adequacy is demonstrated by assessing predictive performance on seen and unseen samples. The algorithm is built to support incremental updates, addressing the need for adaptability. To justify the claims of fidelity, i.e., that the boxes reflect the actual black-box behavior, the boxsystems and subsequent explanations are constrained by an *attribution method*. The *compactness*, i.e., length of all rules, is directly controllable by the user.
3) We demonstrate the *concept* of a "theory of a black-box" and its *usefulness* by illustrating how the resulting data structure naturally supports different inquiries at various levels of granularity.

The remainder of the paper is structured as follows: Section 2 presents how CE has been discussed so far in the context of XAI and highlights two rule based explanation methods to facilitate the discussion later. Section 3 gives a practical summary of CE and explains the conceptual applicability to XAI. Finally, in Section 4 we develop our algorithm CoBoT, before illustrating its use as a theory in Section 5. Code is available on github.com/semueller/CoBoT.

## 2 Related Work

In the XAI literature, to the best of our knowledge, Constructive Empiricism is mostly discussed for the particular *theory of explanation* it provides, not for its larger scope as a *philosophy of science*. Naturally, as a philosophy of science that views science ultimately as an epistemic endeavor and promotes a pragmatic approach to scientific inquiry, CE has to provide a framework for explanation.

CE is mentioned for its theory of explanation in a line of works dealing with the assessment of explainability and how adopting the view on the explanatory process provided by a particular theory of explanation in turn impacts the evaluation of explanations [18,20]. These works deal with important questions in XAI that are very distinct from the scope of our work.

Another work [12], that we see complementary to ours, employs CE's view on explanation to describe a conceptual framework where explainability methods, given access to input data and the black-box, produce outputs in accordance with *specifications* provided by the user. In line with van Fraassen, these specifications are descriptions of "the kind of explanation required", formalized in a *relevance relation*. The relevance relation establishes an explicit connection between *observations* and *mutually exclusive propositions* (explanatory statements). The authors discuss how their framework captures explanations for Bayesian networks and for graph neural networks. For the former the hypothetical user poses a question in the form of "request for arguments" and explanations take the form of probabilistic arguments. For graph neural networks the paper shows how an existing XAI method can be expressed as a specification. While the framework directly adopts the notion of explanation from CE, it leaves out a crucial aspect: Within CE, explanations are derived from a *theory*. [12] demonstrates the conceptual utility of CE's interpretation of explanation as a suitable framework for XAI research. Our work continues with the conceptualization of CE for XAI and provides a concrete proof of concept implementation of a theory for a black-box.

*Rule Based Explanations* To ground our discussion in a concrete setting, we will build on *rule based* explanations. Framed as a *specification*, rule based explanations posit *labeled interval constraints* as propositions and establish the connection that if the values of a sample (observation) lie within the interval constraints, the proposition is fulfilled. We summarize two rule based explanation algorithms for later reference:

Anchors [16] was developed as a *local* explanation method that explores the decision surface of a black-box around an input sample. Given a user-defined

precision threshold, the algorithm performs a greedy search around the sample that sequentially adds more interval constraints, to find a box in which the black-box prediction remains the same (up to the specified threshold). The authors state that their precision preserving search criterion is naturally biased towards shorter, more compact, rules. However, the total number of constraints used by the final "Anchor" is out of control of the user. Also, the empirically obtained rules may be *unbounded* in some direction.

CFIRE [11] computes a *global* rule model with strictly bounded interval constraints. Given the black-box, a set of labeled input samples and a local explainer, CFIRE computes a set of rules for each class separately. To identify rule candidates, the algorithm performs closed frequent itemset mining on the local explanations. Samples whose local explanations are in the support of the same itemset are used to fit a decision tree. The decision tree is tasked to discriminate said samples from *all* other samples, i.e., also from samples from other classes and irrespective of their local explanation. After processing all itemsets, the algorithm greedily selects a subset of candidate rules by computing the minimal set cover. This step is still performed for each class separately. The resulting rule models are compact and accurate. However, examples provided in the paper also show that rules for different classes may *overlap*. For samples falling into the overlapping region, CFIRE applies a *post-hoc* heuristic to resolve the *ambiguity*.

## 3 Operationalizing Constructive Empiricism

In this section we give a practical introduction to Constructive Empiricism (CE) as discussed in [4] and how it relates to XAI. We briefly set the scene before discussing the notions of empirical adequacy and pragmatic virtues. We shall find that these notions address the two core aims of XAI: provide both *true* and *useful* insights into a black-box.

*Setting the scene* We consider the setting where we are given a classifier function $f : \mathbb{R}^d \to \mathrm{C}$. We assume that the classifier does not change over time. The goal is to develop an algorithm that produces an empirically adequate data structure that is useful for rule based explanations.

**Empirical Adequacy** To describe the relation between a scientific theory and the domain it aims to describe, van Fraassen uses the term *empirical adequacy*. It has three components: 1) A theory must be able to *represent all possible phenomena* (observations made through measurements). For science in general, the former can pose a challenge (e.g., think of the step from Newtonian Mechanics to the Theory of Relativity). In our case however, the hypothesis class of the black-box is known and observations are members of its domain and co-domain. 2) The theory must be internally *consistent* with all actually encountered observations. In machine learning terms this translates to *perfect training accuracy*. 3) For a scientist to *accept* a theory means a) to *believe* that it is empirically

adequate and b) to *commit* to use it. Committing means that they must be willing to work with it in the hopes that it will be successfully *applicable* to future observations or can be *adapted* to account for them. In van Fraassen's view of the scientific process, theory building is *iterative*, where observations inform theory building which leads to new experiments and so on. To justify the acceptance of a "theory of a black-box", we argue that the theory must be *algorithmically adaptable* and proof its applicability quantitatively *up to the standards of the user*, e.g., by exceeding a predefined accuracy score on a given test set.

**Pragmatic Virtues** The evaluation of XAI methods separates the evaluation of the information content of an explanation from the evaluation of its comprehensibility. For example, in [14] the authors distinguish "Content", "Presentation" and "User". The work in [2] goes one step further and argues that a content-first evaluation scheme builds the foundation for trustworthy explanation. CE's emphasis on empirical adequacy fits this content-first scheme and we argue that the theory of a black-box thus provides a foundation for justifying trust in the explanations derived from it. How do user centric criteria fit in this picture? For van Fraassen, to accept a theory is not only to believe that it is empirically adequate, but also to be committed to using the theory to make new predictions, discover new observable phenomena, and design new experiments. Thus, he argues, if a scientist is being presented with two empirically adequate theories, it is rational to prefer one over the other, solely on the basis of *usefulness* – whatever this may entail for them. The reasons the scientist may provide for preferring one theory over another are subsumed under the notion of *pragmatic virtues*. Analogously, the user is of central importance in XAI: The request for explanation implies a need (or at least desire) for insight, hence it is justified to construct the output of XAI algorithms according to user needs. Many efforts in XAI are concerned with the definition and evaluation of user needs [2,3,14,17,24]. In the particular case of our setting, rule based explanations, *compactness* and *fidelity*, are, among others, desired properties of explanations that go beyond the mere predictive agreement of rule and black-box [11,14,16].

**Summary** XAI requires both truthful and comprehensible explanations. We saw that both concerns are captured directly by the notions of empirical adequacy and pragmatic virtues, respectively. Empirical adequacy entails consistency of the theory with the available observations on the one hand and the commitment of the scientist to work with the theory on the other. From this we motivated the requirements of (i) *consistency*, (ii) *algorithmic adaptability* and (iii) *sufficient predictive performance*. Drawing a direct connection to content-first evaluation schemes in XAI, we argue that the empirical adequacy of a theory provides the necessary basis to justify trust into explanations derived from it. By emphasizing the epistemic goal of the scientist, adapting the theory to maximize its usefulness to the practitioner is a rational requirement. Thus, CE provides the grounds on which to embed user needs, that go beyond empirical adequacy, directly into the theory.

## 4 Example: The Constructive Box Theorizer Algorithm

We now describe the CoBoT algorithm (see Algorithm 1). Recall that, in order to achieve empirical adequacy and justify acceptance, the algorithm needs to achieve *consistency* with all observations, *adaptability*, and *sufficient predictive performance* on unseen samples. The first two properties are addressed algorithmically within CoBoT, while the latter will be demonstrated in Section 5. Regarding pragmatic virtues, CoBoT addresses *(internal) fidelity* and *compactness*, since these criteria are central to rule-based explanations [11,14,16], which we want the theory to support.

We first give a brief overview of the algorithm and introduce the notation. Then we go through the algorithm step-by-step and key design choices in light of empirical adequacy and pragmatic virtues. Finally, we discuss the shortcomings of Anchors and CFIRE in providing *theories*.

**In brief,** CoBoT incrementally constructs a *set of boxsystems* $\mathcal{B}$, each covering a distinct *subspace* of the input space and consisting of a *set of axis-aligned bounded boxes* $b$, each associated with a label $c$ indicating the corresponding target class. To facilitate fidelity, the subspaces are derived from a local explainer method $\Phi$ using an indicator function $\mathcal{I}$. For a given local explanation, this function selects the feature set $I \subseteq [1..d]$ , that represents the most important dimensions (e.g., all dimensions for which attribution scores exceed a positive threshold). For each unique feature set $I$ encountered, a corresponding boxystem is computed, based on all observed samples $x$ with $\mathcal{I}(\Phi(x)) = I$. These boxsystems are constructed ensuring that 1) consistency with all observations is guaranteed and 2) each observation is unambiguously associated with a single box $b$. In case CoBoT encounters a new, conflicting observation (i.e., one that falls into a box with a wrong label), only the inconsistent box is adapted. To facilitate comprehensibility, the maximal size of $I$ can be controlled by the user.

**Step-by-Step Description** *1) Initialization* As input, CoBoT requires a black-box model $f$, an input sample $x$, a local explainer function $\Phi$, an indicator function $\mathcal{I}$, a list of observations $\mathcal{O}$, and a (possibly empty) set of boxsystems $\mathcal{B}$. In lines 1 to 3, CoBoT computes the class label $c$ for the input sample $x$ and the local explanation $a$, from which it derives the feature set $I_x$ of most important dimensions. The observation triple $(x, c, I_x)$ is then added to the set of observations $\mathcal{O}$ in line 4. The local explainer is included to ensure internal fidelity with the boxes. Internal fidelity [24] (also called *reasoning-completeness* [14] or *faithfulness* [5]), requires an explanation to not only reproduce the input-output mapping but to do so in a way that reflects the actual reasoning process of the black-box. While fitting a decision tree to match the black-box predictions may satisfy external fidelity, we have no grounds to believe in its internal fidelity because of predictive multiplicity [9] (i.e., $f$ may allow for functionally different but equally well performing approximations of itself). Therefore, we need to constrain the solution space of viable boxystems. Since the black-box is (by

**Algorithm 1** CoBoT

**Input:** Black-box model $f : \mathbb{R}^d \to \mathrm{C}$ for some positive integer $d$ and finite set C, local explainer $\Phi$ for $f$, a binarization function $\mathcal{I}$ for local explanation, input sample $x$, set of boxsystems $\mathcal{B}$, set of encountered observations $\mathcal{O}$
**Output**: Set of boxsystems $\mathcal{B}$

1: $c \leftarrow f(x)$
2: $a \leftarrow \Phi(f, x, c)$
3: $I_x \leftarrow \mathcal{I}(a)$ ▷ obtain set of important dimensions
4: $\mathcal{O} \leftarrow \mathcal{O} \cup \{(x, c, I_x)\}$
5: $B_{I_x} \leftarrow B$ **if** $\exists (I, B) \in \mathcal{B}$ with $I = I_x$, **else** None
6: **if** $B_{I_x}$ = None **then** ▷ None on first encounter of itemset
7: $b_x \leftarrow \textsc{CreateSingleton}(x, c, I_x)$
8: $\mathcal{B} \leftarrow \mathcal{B} \cup \{(I_x, \{b_x\})\}$
9: **else**
10: **if** $\exists b \in B_{I_x}$ with $x \in b$ **then**
11: **if** $\textsc{Label}(b) \neq c$ **then**
12: $B_{\mathcal{O}_b} \leftarrow \{\textsc{CreateSingleton}(o, c_o, I_o) \mid (o, c_o, I_o) \in \mathcal{O} \wedge I_o = I_x \wedge o \in b\}$
13: $B'_{I_x} \leftarrow \textsc{Merge}((B_{I_x} \setminus \{b\}) \cup B_{\mathcal{O}_b})$
14: $\mathcal{B} \leftarrow (\mathcal{B} \setminus \{(I_x, B_{I_x})\}) \cup \{(I_x, B'_{I_x})\}$
15: **else** ▷ no existing box contains sample
16: $b_x \leftarrow \textsc{CreateSingleton}(x, c, I_x)$
17: $B'_{I_x} \leftarrow \textsc{Merge}(B_{I_x} \cup \{b_x\})$
18: $\mathcal{B} \leftarrow (\mathcal{B} \setminus \{(I_x, B_{I_x})\}) \cup \{(I_x, B'_{I_x})\}$
19: **return** $\mathcal{B}$, $\mathcal{O}$

definition) opaque, i.e., its "true reasoning" is *unobservable*, we use local explanation methods to obtain constraints for each observation. Only features indicated as important for a given sample are used to construct the boxes.

*2) Creating a new boxsystem* Line 6 begins with the consistency check and the updating logic. It checks whether the feature set $I_x$ is present in the set of boxsystems. If not, line 7 creates a singleton-box $b_x$ around $x$ in the subspace determined by $I_x$, associates it with the corresponding class label, and adds the tuple $(I_x, \{b_x\})$ to $\mathcal{B}$. CoBoT then returns the updated set of boxsystems $\mathcal{B}$ and observations $\mathcal{O}$.

*3) Resolving inconsistency* In case $I_x$ had been encountered before, CoBoT proceeds in line 10 and checks if there exists a box $b$ in the corresponding boxsystem $B_{I_x}$ that covers $x$. If so, line 11 checks if the box's label matches the black-box output $c$. If the labels match, the algorithm returns. Otherwise, the boxsystem is updated: Line 12 selects all samples from the current list of observations that lie within the affected box and places each into its own singleton box. In the next step (line 13), the inconsistent box $b$ is removed from $B_{I_x}$ and all singleton-boxes in $B_{\mathcal{O}_b}$ are attempted to be *merged with* the remaining boxes in $B_{I_x}$. To this end, we use an adaptation of the algorithm presented in [22]. Given a set of boxes, the algorithm greedily selects the closest two boxes having the same class, and attempts to join them. If a join leads to an overlap with any

other box , the join operation is reverted. This process continues until no further consistent mergers are possible. While the original algorithm [22] was developed for binary classifications, we adapt it mutatis mutandis to multi-class problems. The algorithm has several desirable properties for our scenario:

i It naturally guarantees that *all indicated* dimensions in $I_x$ are used. In contrast to Anchors, via $\mathcal{I}$, CoBoT provides direct control over rule complexity.
ii The algorithm guarantees to return a boxsystem that is *fully* consistent with the training data, thereby fulfilling the *consistency* requirement of empirical adequacy.
iii None of the computed boxes overlap, meaning that 1) each sample induces the construction of at most one box and 2) no sample can fall into two boxes at the same time. This avoids the *ambiguity* problem the rule systems of CFIRE struggle with.
iv All boxes are bounded on all sides (i.e.,, unlike Anchors, no box extends to $\pm\infty$ on any side) and do not extent beyond values observed in the input data. Thus, each box is strictly tied to observations.

If boxes from two different boxsystems, say $B_{I_1}$ and $B_{I_2}$, overlap, this is of no concern, because $\Phi$ indicated that $f$ used different sets of features for its decision. Continuing on line 15, the set of boxsystems is updated with the adapted boxsystem and CoBoT returns.

*4) Expanding coverage* The final *else* case is entered if a boxsystem does exist for $I_x$, but none of the boxes in $B_{I_x}$ cover the sample. In that case line 16 creates a new singleton box around $x$ and line 17 performs the merge operation described above in an attempt to extend any existing consistent box by the singleton-box. The set of boxsystems is updated and CoBoT returns.

*Classification using* $\mathcal{B}$ To assess the predictive performance of $\mathcal{B}$ on an unseen sample $x$, we first compute the feature set $I_x$ for $x$ and check if the respective boxsystem $B_{I_x}$ contains a box containing $x$ that has the correct label. If such a box with the correct label exists, the prediction is successful, else not. Constraining the prediction to the local explainer, the assessed predictive performance also reflects the internal fidelity of the rules.

**Pragmatic virtues in practice** In Section 3, we established that, given two empirically adequate theories, CE gives no objective criterion to choose one over the other, but that it is rational to prefer a theory solely based on its "usefulness". We have already hinted at several aspects of what this entails.

Since we develop CoBoT with the goal to *support* explanation, it is reasonable to incorporate XAI quality criteria into it. The first one was internal fidelity. Constraining CoBoT to "important" subspaces is not done to fulfill empirical adequacy, but rather to ensure the boxsystem reproduces the black-box's behavior in specific ways that align with its reasoning process, as far as that is observable. Internal fidelity is thus, in the sense of CE, a pragmatic virtue.

The choice of local explainer is equally pragmatic. The disagreement problem in XAI states that different local explainers may give different explanations

for the same model output, but that there is no principled way to resolve the disagreement and decide "which method is correct" [7]. [11] demonstrated that the quality of rules produced by CFIRE can greatly vary depending on the local explanation method used. Another practical consideration is that some methods have significantly lower computational costs than others (e.g., Integrated Gradients vs. SHAP). CoBoT is agnostic to the local explainer method. By pragmatic virtue, the "best" local explainer to use is the one that is most practical.

Lastly, we mentioned that *compactness* is a desirable property for rule based explanations. It quantifies the length or total number of rules in an explanation, a lower number being generally deemed better [11,14,16]. Since the intention is that CoBoT makes it easy to extract useful explanations from it, we want to have an ad-hoc way to constrain the solution space to find empirically adequate boxsystems in low dimensional subspaces. To this end we modify the indicator function $\mathcal{I}$ to extract the *up-to-top-k* most important features as per $\Phi$. This gives control over the maximal complexity of rules in each boxsystem. We denote the adapted indicator function as $\mathcal{I}^{k\uparrow}$.

**Theories from Anchors or CFIRE?** In contrast to CoBoT, neither Anchors nor CFIRE were build to provide theories. Still, it is insightful to discuss how exactly they fall short. Both algorithms can be used to obtain a set of rules, and differences to CoBoT in this regard have been made clear already (e.g., open intervals, ambiguity). To compute rules, CFIRE also relies on local explanations and performs an itemset mining step to find frequent patterns. To perform an update when a feature set is missing, CFIRE needs to run the itemset mining step again on all observations. In case the new sample does not belong to a frequent pattern, it will still not be accounted for even *after* the update; thus CFIRE cannot guarantee *consistency*. Performing an update with Anchors seems straightforward: for each sample a new Anchor is computed, which guarantees consistency. However, applying Anchor rules to unseen *test* samples does not take internal fidelity into account, because no local explainer is involved and the rules are applied ad-hoc. Artificially comparing Anchors against any local explainer will yield no clear conclusion because of the disagreement problem mentioned above. Likewise, computing an "Anchor" for a test sample and compare it against an existing one will be equally unhelpful because of the heuristic nature of Anchors.

## 5 Extracting Information from CoBoT

We now apply CoBoT to three real-world datasets. Our qualitative analysis shows that empirical adequacy can be achieved and illustrates how the choice of local explainer $\Phi$ and the maximum size $k$ of the subspace impacts the results. Further, we discuss how the theory computed by CoBoT can be used to produce different standard explanations.

**Table 1:** Overview of datasets. Name, number of classes ($|C|$), dimensionality ($d$), black-box accuracy, and number of samples for CoBoT.

| dataset | $\|C\|$ | $d$ | $f$ accuracy | #samples for CoBoT |
|---|---|---|---|---|
| Abalone [13] | 3 | 7 | 0.64 | 1671 |
| Breast-w [6] | 2 | 9 | 0.98 | 10.000 |
| Dry Bean [1] | 7 | 16 | 0.92 | 5445 |

*Setup and Hyperparameters* We apply CoBoT to three neural networks (4 hidden layers, 32 neurons each, ReLU activation) trained on three tabular tasks. An overview of the datasets is provided in Table 1; the number of classes ranges from 3-7, and the number of features from 7-16. Each dataset is split into 3 sets: one for training the black-box $f$, a second for assessing the test performance of $f$ and computing $\mathcal{B}$ with CoBoT, and a third serving as a validation set for $\mathcal{B}$. As local explainers $\Phi$ we use Lime (LI) [15] and Integrated Gradients (IG) [23].

We will discuss one particular theory produced for $f$ trained on the Abalone dataset, showcasing its evolution and its final state from different perspectives. We then vary the local explainer $\Phi \in \{\text{IG, LI}\}$ and the subspace size $k \in \{2, 4, 6\}$ for all three datasets.

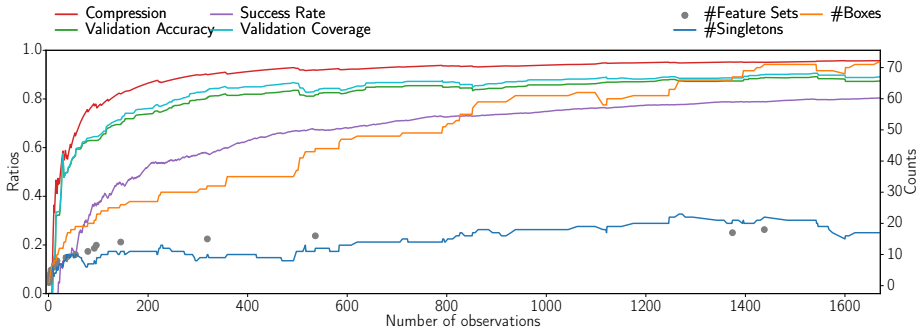


**Figure 1:** Abalone, $\Phi$=IG, k=2. Left axis ("Ratios"): Compression is defined as $1 - \frac{\#boxes}{\#samples}$ over time, Validation Accuracy tracks accuracy on test set, success rate is the ratio of samples that did *not* trigger an update. Right axis ("Counts"): #Feature Sets denotes the number of unique feature sets, #singletons and #boxes count the number of singleton and the total boxes, respectively. After processing 1671 samples, the set of boxes contain less than 20 singletons. Across all observations, 328 updates were performed.

**Evolution of the Theory** We analyze in more detail the set of boxsystems $\mathcal{B}$ computed for $f$ that was fit on the Abalone dataset. CoBoT uses $\Phi$=IG and k=2. Figure 1 tracks the performance on the test set and various other statistics of the theory during the observation phase. CoBoT started with $\mathcal{B} = \varnothing$ and processed 1671 samples (x-axis), leading to 328 updates. *Compression* (red) tracks the space saving ratio of the total number of boxes and the number of samples observed. *Success Rate* (purple) tracks the relative number of samples that were correctly covered, i.e., that did not trigger an update. *Validation Accur-*

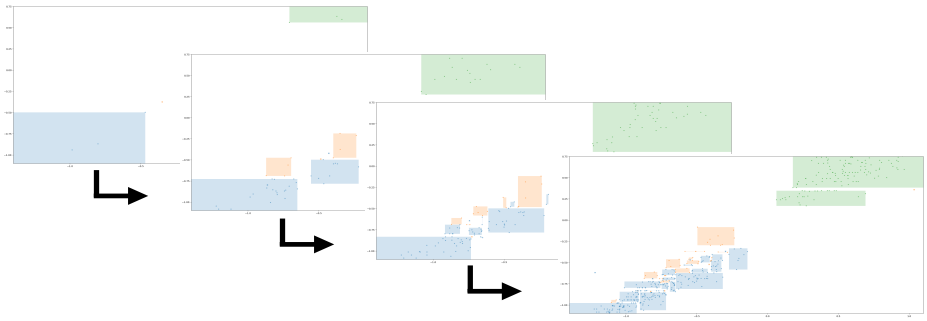

**Figure 2:** Evolution of a boxsystem. Each individual image shows the same boxsystem after different updates. Updates number [9, 40, 61, 144 (final)] (Abalone, $\Phi$=IG, $\mathcal{I}^{2\uparrow}$, $I = \{3, 6\}$).

*acy* (green) and *Coverage* (cyan) show the predictive performance of the theory on the held-out test set. We also track various counting statistics, namely the *total number of features sets* (gray), the *total number of boxes* (orange) across all features sets and the *total number of singletons* (blue) contained within the box count. In the end, $\mathcal{B}$ contained 18 unique feature sets comprised of 72 boxes. The final accuracy was 0.87 with a coverage of 0.89, meaning the false positive rate is only 0.02. As expected, the number of feature sets and boxes rise rapidly in the beginning. Around 1100 and then again near 1600 processed samples, the number of boxes and singletons both decline, indicating that an update was triggered that led to merges of several singletons into boxes.

Because we chose $k = 2$, we can directly visualize a boxysystem. Figure 2 shows a section of subspace of the boxsystem for feature set $I = \{3, 6\}$ at different points in time. Three classes are present, indicated by the three different box colors. Observations are visualized as dots. The scatterplot reveals that the samples are linearly correlated within the subspace. Blue and green samples are well separated, whereas the orange samples lead to an increase in complexity of the boxsystem over time. Of the 328 total updates, 144 were performed on this boxsystem alone.

To obtain a more global view on $\mathcal{B}$, Figure 3 shows a 2d UMAP embedding of all datapoints, color-coded by the feature set they were mapped to and the marker shape indicating the class label predicted by $f$. The color imbalance highlights that the feature sets have vastly different support. For example, there is a single one dimensional feature set $\{6\}$ (dark brown) with very low support (n=2). Feature 6 appears most often across all feature sets. Feature sets $\{2, 4\}$ (orange) and $\{4, 6\}$ (light purple) are mostly confined to the upper right region, whereas several other sets, e.g., $\{1, 4\}$ (dark blue), $\{3, 6\}$ (dark green), and $\{2, 6\}$ (olive), span across large parts of the manifold. In particular, the datapoints colored in dark green belong to the boxsystem visualized in Figure 2. We now see that this is the most populated one, explaining the comparatively large number of updates. Datapoints colored in orange all share the same marker, meaning that this feature set is particularly indicative for a single class. $f$ seems to use similar features for samples that are far apart, and different features for samples

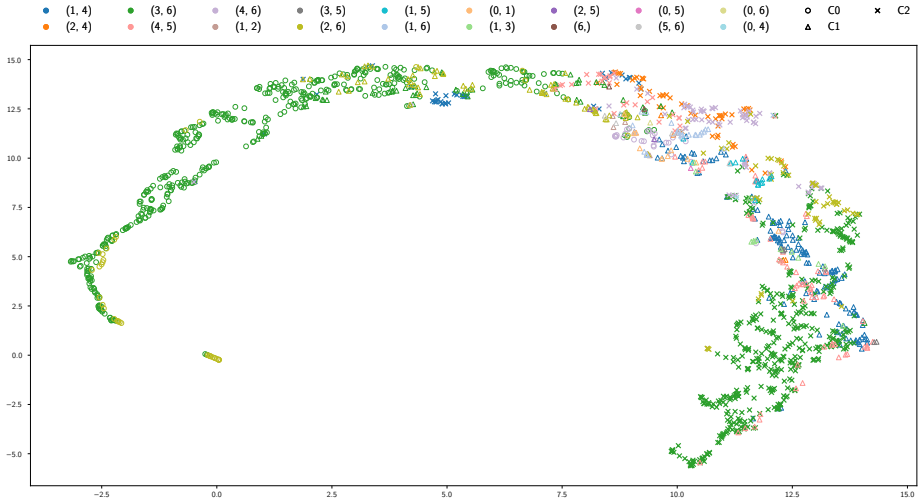


**Figure 3:** 2d UMAP embedding of samples. Colors indicate feature sets, marker style indicates predicted classes (Abalone, IG, $\mathcal{I}^{2\uparrow}$).

that are close together. We see how the theory produced by CoBoT naturally gives insights into global and regional properties of $f$ (cf. related questions in [8]).

Table 2 summarizes the results for different $\Phi$ and $k$ values. For each black-box, some configuration of $\Phi$ and $k$ reaches a final accuracy of at least 0.97, noting that both hyperparameters can have a notable impact on that result. Overall we can observe that lower values of $k$ lead to better accuracy, fewer unique feature sets and, most of the time, fewer boxes compared to larger $k$. This might be an effect of the increased amount of data needed for larger $k$. Accuracy and Coverage coincide in nearly all cases, meaning no false positives. While LI leads to the better-performing theory on both Abalone and Dry Bean, the method struggled on Breast-w, where it failed to produce usable explanations in 56% of the cases. In those cases, no attribution value was positive and the binarization $\mathcal{I}^{k\uparrow}$ mapped the sample to an empty set. Conceptually, a sample with "no supporting features" makes no sense. However, by pragmatic virtue, we are free to use IG, which did not encounter this issue.

**Retrieving Explanations** The data structure produced by CoBoT naturally supports different explanations. Algorithm 2 and Algorithm 3 extract a local and a contrastive explanation, respectively. Both algorithms operate directly on the data structure, conditioning their explanation on the local explainer to increase internal fidelity of their output. Algorithm 2 answers the question "Given $x$, why did $f$ predict $c$?" and provides explanations in the same output format as CFIRE and Anchors. Algorithm 3 answers the question "Given $x$, why did $f$ predict $c$ and what other class is the sample most similar to?", supplying in its output also a neighboring box of a different class. It could be easily adapted to give a targeted answer "Why $c$ and not $c'$ ?" by restricting the search in line 9 to $b$

**Table 2:** Results for varying $\Phi \in \{\text{IG}, \text{LI}\}$ and $k \in \{2, 4, 6\}$ for all three datasets. Testset sizes were 626, 1050, 1050 samples for Abalone, Breast-w and Dry Bean, respectively. On Breast-w, for all but 4431/10000 samples the LI explanations contained negative scores only, meaning $\mathcal{I}^{k\uparrow}$ returned an empty set which is conceptually uninterpretable (denoted by ϟ).

| dataset | $\Phi$ | $k$ | Acc. | Cover. | #updates | $\#I$ | $\#b$ | #singletons |
|---|---|---|---|---|---|---|---|---|
| Abalone | IG | 2 | 0.87 | 0.89 | 328 | 18 | 72 | 17 |
| | | 4 | 0.84 | 0.85 | 445 | 62 | 99 | 28 |
| | | 6 | 0.83 | 0.83 | 498 | 86 | 136 | 57 |
| | LI | 2 | 0.91 | 0.93 | 211 | 6 | 74 | 28 |
| | | 4 | 0.98 | 0.98 | 145 | 12 | 13 | 1 |
| | | 6 | 0.96 | 0.96 | 184 | 16 | 17 | 2 |
| Breast-w | IG | 2 | 0.98 | 0.98 | 404 | 35 | 62 | 9 |
| | | 4 | 0.92 | 0.92 | 1552 | 158 | 252 | 52 |
| | | 6 | 0.87 | 0.87 | 2325 | 304 | 440 | 131 |
| | LI ϟ | 2 | 0.47 | 0.47 | 58 | 6 | 6 | 0 |
| | ϟ | 4 | 0.47 | 0.47 | 165 | 17 | 17 | 5 |
| | ϟ | 6 | 0.46 | 0.46 | 272 | 19 | 19 | 2 |
| Dry Bean | IG | 2 | 0.94 | 0.94 | 800 | 74 | 130 | 29 |
| | | 4 | 0.81 | 0.81 | 1658 | 337 | 359 | 124 |
| | | 6 | 0.74 | 0.74 | 2080 | 567 | 588 | 283 |
| | LI | 2 | 0.97 | 0.97 | 528 | 52 | 71 | 14 |
| | | 4 | 0.87 | 0.87 | 1330 | 177 | 194 | 49 |
| | | 6 | 0.81 | 0.81 | 1813 | 291 | 304 | 102 |

associated with $c'$. A different method could answer the question "What kind of instance gets the same prediction" by returning the supporting samples of the box computed as the local explanation. All these are standard questions arising in XAI [8].

These questions can be easily answered within the theory, but of course, the theory could also be probed with additional queries to verify the information it provides. For example, instead of falling back to global statistics, local rule-based explanations could retrieve individual performance statistics for the particular box that was relevant to them. Further, explanation algorithms can not only be used to *extract* information from the theory, but their inquiries can also *inform* useful refinements. For example, they might highlight areas of the input space that contain relevant information but where the theory lacks observations. Each query thus offers an opportunity to update the data structure, incrementally increasing its empirical adequacy and explanatory power over time. This way, each explanation, regardless of the question it answers, improves the quality for *all* subsequent explanations.

## 6 Conclusion

We explored the utility of constructing a "scientific theory" of a black-box through the lens of CE. We found that CE's core concepts of empirical adequacy and

**Algorithm 2** Local Explanation

**Input:** $I_x$, $c$, $x$, $\mathcal{B}$

$B_{I_x} \leftarrow B : ((I, B) \in \mathcal{B} \wedge I = I_x)$ **else** None
**if** $B_{I_x}$ = None **then**
    **return** "Itemset unkown"
**else**
    $b_x \leftarrow b : b \in B_{I_x} : x \in b$ **else** None
    **if** $b_x$ = None **then**:
        **return** "Not Covered"
    **else if** $\textsc{Label}(b_x) = c$ **then**
        **return** $b_x$
    **else**
        **return** "Wrong Label"

**Algorithm 3** Contrastive Explanation

**Input:** $I_x$, $c$, $x$, $\mathcal{B}$, a distance function $d(.,.)$ between two boxes

$B_{I_x} \leftarrow B : ((I, B) \in \mathcal{B} \wedge I = I_x)$ **else** None
**if** $B_{I_x}$ = None **then**
    "Itemset unkown"
**else**
    $b_x \leftarrow b : b \in B_{I_x} : x \in b$ **else** None
    **if** $b_x$ = None **then**:
        **return** "Not Covered"
    **else if** $\textsc{Label}(b_x) = c$ **then**
        $b_{\neg c} \leftarrow \underset{b' \in B_{I_x} \wedge \textsc{Label}(b) \neq c}{\textbf{argmin}} d(b_x, b')$
        **return** $(b_x, b_{\neg c})$
    **else return** "Wrong Label"

pragmatic virtues naturally align with the key concerns of XAI, particularly fidelity and comprehensibility. We operationalized empirical adequacy by introducing three criteria: *consistency*, *algorithmic adaptability*, and *sufficient predictive performance*. As a proof of concept, we developed CoBoT, that produce an empirically adequate data structure as a theory. We showed how such a theory can support inquiries at various levels of granularity.

Extending our exemplary discussion of CFIRE and Anchors from the perspective of CE, future work should continue to map out existing approaches in XAI and assess their ability to produce theories.

Further, the specification framework proposed in [12] (see Section 2) offers a promising direction for formalizing black-box theories and defining their interfaces for use in different scenarios. A theory with a set of specifications seems like a natural extension to the tool developed in [19] that represents an interactive, illocutionary explanation process. The tool integrates a single explanation method and draws on a body of background information to deliver contextual, but black-box unrelated, information to the user. The theory and specifications could be used to enrich this pipeline.

## References

1. Dry Bean [Dataset]. UCI ML Repository (2020), 10.24432/C50S4B
2. Beckh, K., Müller, S., Rüping, S.: A quantitative human-grounded evaluation process for explainable machine learning. In: LWDA. pp. 13–20 (2022)
3. Donoso-Guzmán, I., Ooge, J., Parra, D., Verbert, K.: Towards a comprehensive human-centred evaluation framework for explainable ai. In: World Conference on Explainable Artificial Intelligence. pp. 183–204. Springer (2023)
4. van Fraassen, B.C.: The Scientific Image. Oxford University Press (12 1980), 10.1093/0198244274.001.0001
5. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? arxiv:2004.03685 (2020)

6. Jan van Rijn: BNG (breast-w) [Dataset] (2014), openml.org/d/251
7. Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., Lakkaraju, H.: The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv:2202.01602 (2022)
8. Liao, Q.V., Gruen, D., Miller, S.: Questioning the ai: informing design practices for explainable ai user experiences. In: Proceedings of the 2020 CHI conference on human factors in computing systems. pp. 1–15 (2020)
9. Marx, C., Calmon, F., Ustun, B.: Predictive multiplicity in classification. In: International conference on machine learning. pp. 6765–6774. PMLR (2020)
10. Molnar, C.: Interpretable machine learning. Lulu.com (2020)
11. Müller, S., Toborek, V., Horváth, T., Bauckhage, C.: Cfire: A general method for combining local explanations. arxiv:2504.00930 (2025)
12. Naik, H., Turán, G.: Explanation from specification. arxiv:2012.07179 (2020)
13. Nash, W., Sellers, T., Talbot, S., Cawthorn, A., Ford, W.: Abalone [Dataset]. UCI ML Repository (1994), 10.24432/C55C7W
14. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Computing Surveys **55**(13s), 1–42 (2023)
15. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
16. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
17. Rong, Y., Leemann, T., Nguyen, T.T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., Kasneci, E.: Towards human-centered explainable ai: A survey of user studies for model explanations. IEEE transactions on pattern analysis and machine intelligence **46**(4), 2104–2122 (2023)
18. Sovrano, F., Sapienza, S., Palmirani, M., Vitali, F.: A survey on methods and metrics for the assessment of explainability under the proposed ai act. In: Legal Knowledge and Information Systems, pp. 235–242. IOS Press (2021)
19. Sovrano, F., Vitali, F.: From philosophy to interfaces: an explanatory method and a tool inspired by achinstein's theory of explanation. In: Proceedings of the 26th International Conference on Intelligent User Interfaces. pp. 81–91 (2021)
20. Sovrano, F., Vitali, F.: Perlocution vs illocution: How different interpretations of the act of explaining impact on the evaluation of explanations and xai. In: World Conference on Explainable Artificial Intelligence. pp. 25–47. Springer (2023)
21. Speith, T.: A review of taxonomies of explainable artificial intelligence (xai) methods. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. pp. 2239–2250 (2022)
22. Stadtländer, E., Horváth, T., Wrobel, S.: Learning weakly convex sets in metric spaces. arXiv:2105.06251 (2024)
23. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
24. van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M.: Evaluating xai: A comparison of rule-based and example-based explanations. Artificial intelligence **291**, 103404 (2021)