Pontifícia Universidade Católica
do Rio de Janeiro

**Jorge Luiz Cataldo Falbo Santo**

# A Critical View on the Interpretability of Machine Learning Models



**Dissertação de Mestrado**

Dissertation presented to the Programa de Pós Graduação em Informática, PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor: Profª. Simone Diniz Junqueira Barbosa

Rio de Janeiro
October 2018

Pontifícia Universidade Católica do Rio de Janeiro

**Jorge Luiz Cataldo Falbo Santo**

# A Critical View on the Interpretability of Machine Learning Models

Dissertation presented to the Programa de Pós Graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática. Approved by the undersigned examining committee.

**Profª. Simone Diniz Junqueira Barbosa**
Advisor
Departamento de Informática - PUC-Rio

**Prof. Bruno Feijó**
Departamento de Informática - PUC-Rio

**Prof. Marcus Vinicius Soledade Poggi de Aragao**
Departamento de Informática - PUC-Rio

**Prof. Márcio da Silveira Carvalho**
Vice Dean of Graduate Studies
Centro Técnico Científico - PUC-Rio

Rio de Janeiro, October 1st, 2018

**Jorge Luiz Cataldo Falbo Santo**

Jorge Cataldo holds a B.Sc. in Mechanics of Aeronautical Engineering from the Brazilian Technological Institute of Aeronautics - ITA in 1987, a graduate degree in Economics from EPGE/FGV in 1990 and a M.Sc. in Business Administration from the IBMEC Business School in 2003.

Bibliographic data

Santo, Jorge Luiz Cataldo Falbo

A Critical View on the Interpretability of Machine Learning Models / Jorge Luiz Cataldo Falbo Santo; advisor: Simone Diniz Junqueira Barbosa. – 2018.

153 f. : il. color ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui bibliografia

1. Informática – Dissertações 2. Inteligência artificial. 3. Aprendizado de máquina. 4. Interpretabilidade. 5. Explanabilidade. 6. Algoritmos interpretáveis. I. Barbosa, Simone Diniz Junqueira. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

## Acknowledgments



I thank my family for their patience.
Thanks to my advisor and academic writing expert, Simone, for the untiring help.
Thanks to Alexandre Nakamura for the inspiring talk in mid-2016.
Finally, thank the researchers at New River Pharmaceuticals for the development of such a wonderful molecule.
This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance code 001.

## Abstract

Santo, Jorge Luiz Cataldo Falbo; Barbosa, Simone Diniz Junqueira (Advisor). **A Critical View on the Interpretability of Machine Learning Models.** Rio de Janeiro, 2018. 153p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

As machine learning models penetrate critical areas like medicine, the criminal justice system, and financial markets, their opacity, which hampers humans' ability to interpret most of them, has become a problem to be solved. In this work, we present a new taxonomy to classify any method, approach or strategy to deal with the problem of interpretability of machine learning models. The proposed taxonomy fills a gap in the current taxonomy frameworks regarding the subjective perception of different interpreters about the same model. To evaluate the proposed taxonomy, we have classified the contributions of some relevant scientific articles in the area.



## Keywords

Explainable AI; XAI; Machine learning; Interpretability of models; Explainability of models; Interpretable algorithms; Algorithmic transparency; Artificial intelligence; AI.

# Resumo

Santo, Jorge Luiz Cataldo Falbo; Barbosa, Simone Diniz Junqueira. **Uma Visão Crítica sobre a Interpretabilidade de Modelos de Aprendizado de Máquina.** Rio de Janeiro, 2018. 153p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

À medida que os modelos de aprendizado de máquina penetram áreas críticas como medicina, sistema de justiça criminal e mercados financeiros, sua opacidade, que impede que as pessoas interpretem a maioria deles, se tornou um problema a ser resolvido. Neste trabalho, apresentamos uma nova taxonomia para classificar qualquer método, abordagem ou estratégia para lidar com o problema da interpretabilidade de modelos de aprendizado de máquina. A taxonomia proposta que preenche uma lacuna existente nas estruturas de taxonomia atuais em relação à percepção subjetiva de diferentes intérpretes sobre um mesmo modelo. Para avaliar a taxonomia proposta, classificamos as contribuições de artigos científicos relevantes da área.



## Palavras-chave

Inteligência artificial explicável; Aprendizado de máquina; Interpretabilidade de modelos; Explanabilidade de modelos; Algoritmos interpretáveis; Transparência de algoritimos; Inteligência artificial; IA.

# Table of contents

## List of figures

## List of tables

# List of equations

*"Opaque and invisible models are the rule, and clear ones very much the exception. We are modeled as shoppers and couch potatoes, as patients and loan applicants, and very little of this do we see—even in applications we happily sign up for. Even when such models behave themselves, opacity can lead to a feeling of unfairness".*

***Cathy O'Neil***, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (2017).*

# 1
# Introduction

There is a lot of evidence that humankind is facing an embarrassing trade-off between accuracy of their most ubiquity and useful models ever and our ability to **understand** and **trust** them. Predictive models generated by techniques that enable them to learn from data sets have become very popular, both for the simplicity with which they can be generated and for their increasing accuracy. However, learning models, especially machine learning models[1], have two worrying features. First, as they are **data driven**, they are subjected to data bias. According to O'Neil (2017), we cannot expect equity and justice from **data-driven models**, as "these models are opinions embedded in mathematics", so they are not free from biases (O'Neil, 2017). Second, they are increasingly opaque. Lipton (2017) claims that, because of their nested non-linear structure, these highly successful machine learning and artificial intelligence models are usually applied in a black-box manner.

The concerns addressed by Lipton and O'Neil are typical of such models and tend to increase as the deployment of machine learning models becomes widespread and ever more complex. As these models penetrate critical areas like medicine, the criminal justice system, and financial markets, people's inability to understand them seems problematic (Lipton, 2016). Unfortunately, although understanding machine learning models has become increasingly relevant, it has also become more difficult and complex to achieve (Samek, Wiegand, & Müller, 2017).

Finding tools to deal with the current trade-off between accuracy of machine learning models and our ability to interpret them is a legitimate human demand that defines the contours of what we call in this research the "problem of human interpretability of machine learning models" or simply "the problem of interpretability."

This dissertation proposes a **new way** of approaching the strategies to deal with this problem.

[1] To simplify the text, we use the term "machine learning models" to refer to "predictive models generated by machine learning techniques".

## 1.1.Problem statement

We can note a kind of global "anxiety" because of the current advances of artificial intelligence (AI). It seems that humanity shares the feelings of **hope** and **fear** at the same time. If, on one hand, we have the **advantages** of being able to use **high performance models**, on the other hand, we experience the **discomfort** by the threat of losing control of the situation.

The second Workshop on Human Interpretation in Machine Learning (WHI/ICML 2017) makes clear on its call message the **discomfort** with the current nature of machine learning models**:**

> "*The latest trend in machine learning is to use* ***highly nonlinear complex*** *systems such as deep neural networks, kernel methods, and large ensembles of diverse classifiers. While such approaches often produce impressive, state-of-the art prediction accuracies, their* ***black-box nature*** *offers* ***little comfort*** *to decision makers. Therefore, in order for predictions to be adopted, trusted, and safely used by decision makers in mission-critical applications, it is imperative to develop machine learning methods that produce interpretable models with excellent predictive accuracy. It is in this way that machine learning methods can have impact on consequential real-world applications*" (WHI/ICML, 2017).



As another evidence of how we are trying to protect ourselves from the current growth of artificial intelligence, the European Parliament adopted in 2016 a set of comprehensive regulations for the collection, storage and use of personal information, The General Data Protection Regulation (GDPR). Slated to take effects as law across the EU in May 2018, the GDPR creates a "*right to explanation*" whereby users can ask for an explanation of an algorithmic decision that was made about them (Goodman & Flaxman, 2016).

The GDPR is an example of the current social demands to ensure that machine learning algorithms are not merely **efficient** but also **transparent** and **fair**. As consequence, by addressing some humankind concerns, such as (but not only) **fairness**, **privacy** and **trust** (Lipton, 2106; O'Neil, 2017), and some human preferences like **causality**, (Narayanan et al., 2017; Lombrozo, 2006; Keil, 2006) the increasing social demands to interpret the outputs of **opaque models** has opened a **new promising area** of research in AI and machine learning.

Initially named as **human interpretability**, but also known as **model explainability**, **interpretable algorithms**, **algorithmic transparency** or **explainable artificial intelligence**[2], the research area rapidly gained prominence with the increasing number scientific conferences hosting specialized discussion forums. Despite of being a recent field of study, since 2016 the theme "**interpretability**" has won its own workshops at the two largest international conferences on machine learning, ICML and NIPS[3].

From July 2018, in ICML, the term "explainable IA" appears as a proposal of nomenclature to bring together the various themes of study on human interpretability of machine learning models. On this occasion, the "Second XAI Workshop" grouped the coordination of the four main discussion forums on interpretability until then. As its own description states:

> "***Explainable AI (XAI) systems*** *embody explanation processes that allow users to gain insight into the system's models and decisions, with the intent of improving the user's performance on a related task. (…) Addressing this challenge has become* ***more urgent*** *with the increasing reliance on learned models in deployed applications. This raises several questions, such as:* ***How should explainable models be designed****?* ***How should user interfaces communicate decision-making? What types of user interactions should be supported? How should explanation quality be measured?*** *These questions are of interest to researchers, practitioners, and end-users, independent of what AI techniques are used. Solutions can draw from several disciplines, including* ***cognitive science****,* ***human factors****, and* ***psycholinguistics***" (Second XAI Workshop / ICML 2018).

The research on "interpretability" seems to keep the same increasing pace of the research on new techniques to develop machine learning models. Figure 1 shows that the interest of the scientific community in the theme "interpretability" has somewhat accompanied the growing interest of the scientific community in the theme "machine learning".

[2] A term popularized since 2016 by the DARPA Explainable Artificial Intelligence Program (DARPA XAI), an AI research program which aims to create a suite of machine learning techniques to generate a portfolio of methods that will provide future developers with a range of design options covering the performance-versus-explainability trade space (Gunning, 2017a).

[3] At NIPS 2017 only, there were three workshops discussing research on human interpretability of models.

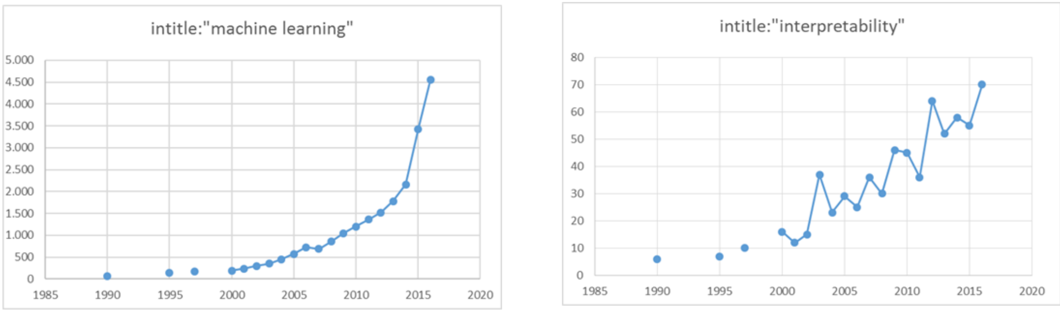


**Figure 1 - Hits on Google Scholar search engine for the words "machine learning" and "interpretability" in the title.**

Most of the academic papers have proposed techniques for interpreting the outputs of machine learning models, as well as analyzing the impact of using those techniques on the models' accuracy. According to DARPA, these **applied works** provide a **range of design options** covering the performance-versus-explainability trade space (Gunning, 2017a). However, despite the large number of **applied works** published so far, some practical issues still need to be addressed in order to have a "fully explainable" AI as part of our daily lives.

Moreover, it appears that the formal contours of the area are still diffuse and do not have a broadly accepted definition. In one of the first conceptual works of the area, Burrell (2016) made an interesting case on the types of opacity of the machine learning models. According to him, computer scientists term this **opacity** as a "**problem of interpretability**" but, though seemingly intuitive, the term "interpretability" in this context does not have a consolidated definition yet. Moreover, according to Lipton (2016), some suggest "model interpretability" as a remedy, but few articulate precisely what "interpretability" means. Despite the absence of a definition, papers frequently make claims about the interpretability of various models. In the same way, Doshi-Velez & Kim (2017) claim that, despite the challenges and the growing interest in interpretability, there is very little consensus on what interpretable machine learning is and how it should be measured (Doshi-Velez & Kim, 2017).

As a new area of research, **Explainable AI** shares with emerging research areas the need for clear definitions. So far, a formal definition of the "problem of interpretability" has proved to be a difficult task, which may be far from complete. In turn, the lack of a more formal definition may be contributing to slow the development and widespread use of **XAI Systems**[4] in our daily lives.

[4] Explanation systems or XAI Systems refer to any software that interacts with the potential interpreters to provide explanations of the outputs of a target model.

## 1.2.Motivation

The great relevance that issues related to **Explainable AI** have reached today motivates the present academic research. In line with the current challenges of the area discussed in the previous section, this academic research contributes to address the following goals:

1. To provide software developers, lawmakers and governmental agencies with a **range of design options** covering the performance-versus-explainability trade space[5].
2. To provide consulting companies with **theoretical frameworks** so that they can evaluate projects and recommend strategies to address different variants of the problem of interpretability.
3. To guide future research on issues related to a **broader and useful definition** of the problem of interpretability.

In line with these goals, this work focuses on the problem of considering **the subjective perception of different interpreters on the outputs of a same model**.

## 1.3.Research goals and questions

The main goal of this research can be phrased as:

> *"To **present** a new approach to the problem of interpreting the outputs of machine learning models that supports the development of **Explainable AI systems** which consider the **subjective perception of different interpreters on the outputs of a same model**.*

To achieve the research goal we set out to answer the following **research questions** and its respective sub-questions:

**RQ1**: *What are, and what principles underlie, the techniques[6] proposed so far to improve the interpretability of machine learning models?*

- **RQ1.1**: *How many techniques have been proposed by scientific research to improve the interpretability of machine learning models?*

[5] Also known as the DARPA XAI goal.
[6] By "technique" we mean in this research any method, approach or strategy.

- **RQ1.2:** *What are the taxonomies proposed so far to classify the techniques that improve the interpretability of machine learning models?*

**RQ2:** *How to propose a technique that considers the subjective perception of different interpreters on the outputs of a same model?*

## 1.4.Dissertation structure

The next chapter, “**Related work**”, summarizes some definitions of the problem of interpretability and presents the state of the art of research on Explainable AI.

The chapter “**Research method**” addresses the strategy chosen to answer research questions and shows the plan for putting the strategy into action.

The chapter “**A semiotic view on interpretability**” discusses the conceptual foundations of a **new proposal to address** the problem of interpretation of machine learning models and presents a procedure to apply the new view to solve problems of interpretability.

The chapter "**Evaluating the semiotic view of interpretability**" provides a framework to classify the techniques that improve the interpretability of machine learning models based on semiotic view proposed, and shows the results of the classification of technical proposals in some selected scientific articles.

The chapter “**Conclusions and future work**” presents the main results of the research and some suggestions for future work.

Finally, **“Appendix I”** brings to the public of professionals who are not familiar with the terms of **machine learning** and **semiotics**, definitions of the main concepts that are important to understand the Explainable AI concepts discussed in this dissertation.

# 2
# Related work

In this chapter, we present an overview of the current research on **Explainable AI**[7]. First, we present conceptual works, which formally define a typical problem of interpretability. We then provide an overview of the conceptual taxonomy frameworks proposed so far and, finally, we discuss the current gaps of these frameworks.

## 2.1.The "problem of interpretability"

Research on interpretability of complex models is not a new topic, but the scientific contributions dealing with machine learning models (especially deep models) are recent. **Conceptual works** on this theme are even more recent.

**To justify the study of model interpretability**, Lipton lists some objectives of model interpretations we believe important but struggle to model formally. Likewise, Doshi-Velez & Kim (2017) list some desiderata that machine learning models often do not achieve when interacting with humans.

**Defining model interpretability**, Burrell (2016) focuses on the types of opacity. Escalante et al. (2017) distinguish explainability from interpretability. Weller (2017) discusses which types of transparency are helpful to whom in which contexts and addresses the concept of machine interpretability. Lipton (2016) lists some properties of interpretable models and post-hoc techniques to interpret them and Lipton (2017) addresses the hard questions involved with the formulation of the problem of interpretability.

**Going further**, Dhurandhar et al. (2017) propose an approach for interpretability relative not only to humans, but also to a target model (Dhurandhar, Iyengar, Luss, & Shanmugam, 2017). Weller (2017) addresses a fruitful line of work which helps machines understand each other (Weller, 2017), and Offert (2017) suggests that a better understanding of the deficiencies of the intuitive notion of interpretability is needed as well.

[7] To better understand the discussions in this chapter, it is necessary for the reader to know the fundamentals of **machine learning** presented in Appendix I.

In a conceptual way, the current definitions of the problem of interpretability consider mainly “what” and “how” must be explained.

To answer “**what must be explained**”, Lipton (2016) explores the reason for interpretability, claiming that interpretations serve those objectives that we deem important but struggle to model formally: (1) trust; (2) causality; (3) transferability; (4) informativeness; and (5) fair and ethical decision-making (Lipton, 2016). Likewise, Dhurandhar et al. (2017) and Doshi-Velez & Kim (2017) proposed formal and rigorous frameworks for subjects related to model interpretability.

To answer “**how it must be explained**”, Montavon et al. (2017), considered the difference between “to interpret” and “to explain” (Montavon, Samek, & Müller, 2017). However, despite Montavon et al.’s contribution, a **semiotic view** of the potential strategies and technical approaches is still an open field for research[8]. In this sense, Dhurandhar et al. (2016) have begun an interesting discussion that addresses the model interpretability beyond the explanations for humans.



## 2.2.Taxonomy frameworks

Some researchers propose **formal taxonomy frameworks** to classify techniques whose improve the interpretability of machine learning models while others indirectly address the classification by presenting **focused classifications** to support the rationale of novel approaches. This section details the results of a search for **conceptual works** whose propose those taxonomy structures, and, additionally, for articles which do not explicitly propose a taxonomy but present **focused mappings** to explain the arguments of novel techniques.

### 2.2.1.Search results

Table 1 shows the results of the search for taxonomy frameworks and focused surveys on digital libraries, journals and conferences proceedings.

**Table 1 – Search for taxonomy frameworks and focused mappings**

| Type | Author | Title | Year |
|---|---|---|---|
| Conceptual works | Gunning | Explainable artificial intelligence (XAI) | 2018 |

[8] For this research, a “semiotic view” means the study of **meaning making**, as the semiotic field explores the study of signs and symbols as a significant part of perception and communication.

| | | | |
|---|---|---|---|
| | Narayanan et al. | How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation | 2018 |
| | Lipton | The Doctor Just Won't Accept That! | 2017 |
| | Offert | I know it when I see it. Visualization and Intuitive Interpretability | 2017 |
| | Dhurandhar et al. | A Formal Framework to Characterize Interpretability of Procedures | 2017 |
| | Weller | Challenges for Transparency | 2017 |
| | Doshi-Velez & Kim | Towards A Rigorous Science of Interpretable Machine Learning | 2017 |
| | Gunning | Explainable artificial intelligence (XAI) | 2016 |
| | Ribeiro et al. | Why should I trust you? Explaining the Predictions of Any Classifier (v.3)<br><br>Model-Agnostic Interpretability of Machine Learning | 2016a<br><br>2016b |
| | Lipton | The Mythos of Model Interpretability (v.1) | 2016 |
| | Ribeiro et al. | Why should I trust you? Explaining the Predictions of Any Classifier (v.1) | 2016a |
| | Burrell | How the machine 'thinks': Understanding opacity in machine learning algorithms | 2016 |
| Focused mapping | Lundberg & Lee | A unified approach to interpreting model predictions | 2017 |
| | Olah et al. | Feature Visualization: How neural networks build up their understanding of images | 2017 |
| | Chakraborty et al. | Interpretability of Deep Learning Models: A Survey of Results. | 2017 |
| | Samek et al. | Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models | 2017 |
| | Montavon et al. | Methods for interpreting and understanding deep neural networks | 2017 |
| | Escalante et al. | Design of an explainable machine learning challenge for video interviews | 2017 |
| | Guo at al. | Towards Interrogating Discriminative Machine Learning Models | 2017 |
| | Shrikumar et al. | Learning key features through propagating activation differences | 2017 |

| | | | |
|---|---|---|---|
| | Montavon et al. | Explaining nonlinear classification decisions with deep Taylor decomposition | 2016 |

We reviewed the academic works shown in Table 5 both by the (1) **relevance** of the taxonomy to the scientific community and by the (2) mapping **coverage** of the research. Following are some highlights of the review:

Taxonomy frameworks

The following researches presented new conceptual **views**, **definitions**, or **formal taxonomy frameworks** to deal with the problem of interpretability:

- Lipton addresses the techniques and the **model properties** that are proposed either to enable or to create model interpretations (Lipton, 2016).
- Ribeiro et al. address and name the **model-agnostic approach** (Ribeiro, Singh, & Guestrin, 2016).
- Gunning borrows the concept of **deep explanation** from the literature of **expert systems** (Gunning, 2017b).
- Gunning addresses the role of human-computer interaction (**HCI**) and **psychology** on the strategies for improving model interpretability (Gunning, 2017b).
- Doshi-Velez and Kim propose an **evaluation taxonomy** of model interpretability (Doshi-Velez & Kim, 2017).

Figure 2 shows an overview of Gunning's classification framework.

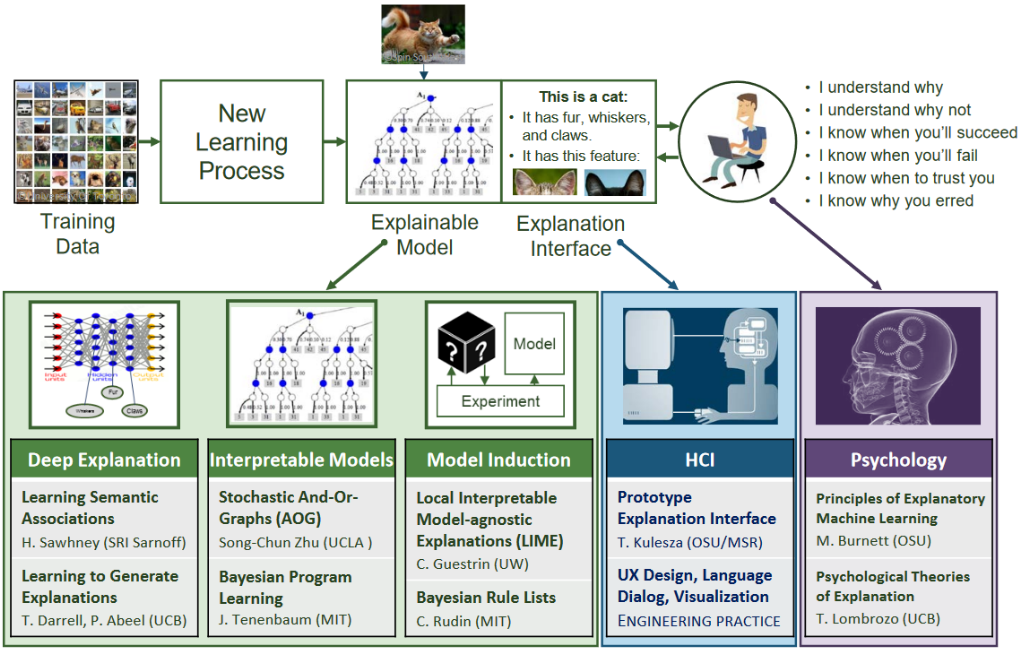


**Figure 2 - Gunning's 2017 framework**



<u>Focused mappings</u>

The following researches propose focused taxonomy frameworks as they present short and focused mappings to support the rationale of novel techniques:

- Samek et al. address some methods for visualizing, explaining and interpreting **deep learning models** (Samek et al., 2017).
- Montavon et al. provides an entry point to the problem of interpreting a deep **neural network model** by introducing some tricks and recommendations (Montavon et al., 2017).
- Chakraborty et al. outline some of the dimensions that are useful for model interpretability in terms of low-level **network parameters**, or in terms of input features used by the model (Chakraborty et al., 2017).
- Shrikumar et al. address approaches to assign an **importance score** to a given task and input example (Shrikumar, Greenside, & Shcherbina, 2017).

### 2.2.2.Summary

In this section, we group the most relevant taxonomy frameworks proposed to date using the criteria of how the techniques:

- Change the target model components;
- Induce surrogate models to interpret the target model;

- Promote the interaction of XAI Systems with potential interpreters.

Figure 3 shows the most relevant proposed taxonomies grouped by the three criteria above.

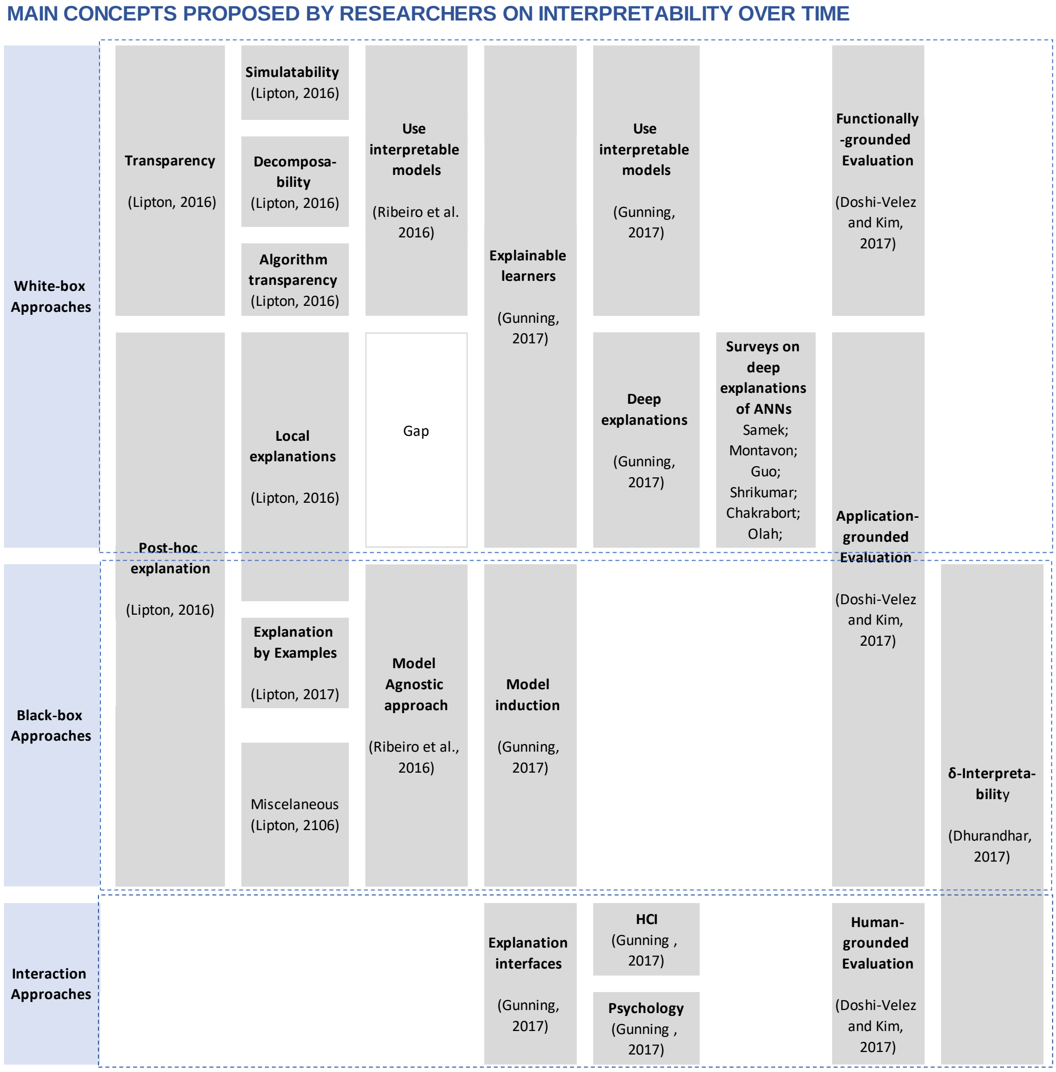


**Figure 3 – Summary of the taxonomies proposed so far**

The logic used to group the proposed taxonomies can also be used to define the categories of an agglutinating taxonomy. The following section details these categories and Table 2 summarizes them.

WHITE BOX APPROACH

Techniques that modify any component of the target model are commonly classified as "WHITE BOX" techniques.

The “WHITE BOX APPROACHES” class is equivalent to Gunning’s “EXPLAINABLE LEARNERS” class. It also includes Lipton’s “TRANSPARENCY” class and part of his “POST-HOC EXPLANATION” class. This class includes: (1) the techniques which aim to develop more interpretable models with mathematical and computational tools like, among others, multi-objective goals, dimensionality reduction, and additive functions; (2) the techniques that aim to explain the results by analogy to examples of the training dataset or previous predictions of the model; and (3) the techniques which aim to explain the output of the models with mathematical and computational tools like, among others, importance score, dimensionality reduction, and information analysis.

BLACK BOX APPROACH

The techniques to infer an auxiliary explainable model from the behavior of the target model are commonly referred as “BLACK-BOX” techniques.

The BLACK-BOX class is equivalent to Gunning’s “MODEL INDUCTION” class and Ribeiro et al.’s “MODEL AGNOSTIC” class. This class includes the techniques ‘which aim to explain the model output of both **just around a single point** of the input domain with instance level explanation tools and explanation by example, and the techniques which interpret the model outputs throughout the full range of the input domain.

INTERACTION APPROACH

Finally, the techniques whose main goal is to improve the meaning making of the interaction with the potential interpreters we named “INTERACTION” techniques.

The INTERACTION approach is equivalent to Gunning’s “EXPLANATION INTERFACES” class and Doshi-Velez and Kim’s “HUMAN-GROUNDED EVALUATION” class. It includes the techniques supported by human-computer interaction theories, which improve the interpretability of a model by changing the interaction between humans and devices, such as interfaces for text explanation, interfaces for visualization, and others.[9] It also includes the techniques, which summarize, extend, and apply current psychological theories of explanation.

[9] Note that “interaction techniques” manipulate the interpretable domains but don’t change them as the “white box techniques” do.

**Table 2– Categories of the agglutinating taxonomy**

| Class Number | Class Description |
|---|---|
| **1.** | **WHITE BOX APPROACHES** |
| 1.1 | INTERPRETABLE MODELS |
| 1.2 | EXPLANATION BY EXAMPLE |
| | |
| 1.3 | DEEP EXPLANATION |
| | |
| | <u>EXPLAIN INDIVIDUAL PREDICTIONS</u> |
| 1.3.1 | FORWARD PROPAGATION - LOCAL EXPLANATION |
| | |
| | <u>UNDERSTAND WHAT THE MODEL HAS LEARNED</u> |
| 1.3.2 | DECOMPOSITION APPROACHES |
| 1.3.3 | BACKPROPAGATION-BASED APPROACHES |
| 1.3.3.1 | *GRADIENTS / DECONVOLUTION / GUIDED BACKPROP* |
| 1.3.3.2 | *RELEVANCE PROPAGATION* |
| 1.3.3.3 | *INTEGRATED GRADIENTS* |
| | |
| 1.3.3 | <u>OTHER DEEP EXPLANATION APPROACHES</u> |
| | |
| **2.** | **BLACK-BOX APPROACHES** |
| 2.1 | MODEL INDUCTION - LOCAL EXPLANATIONS |
| 2.1 | MODEL INDUCTION - GLOBAL EXPLANATIONS |
| | |
| **3.** | **INTERACTION APPROACHES** |
| 3.1 | HCI |
| 3.2 | PSYCHOLOGY |



### 2.2.3.Gaps to fill

After having reviewed the academic works shown in Table 1, we have not found evidence of a **sufficiently comprehensive framework** to classify techniques that improve the interpretability of machine learning models.

Some taxonomy frameworks are quite broad, such as those proposed by Lipton (2016) and Gunning (2017b), but are very superficial in their unfolding, while others are deeply unfolded but much more focused on a specific model class, such as those proposed by Samek et al. (2017) and Montavon et al. (2017) to interpret neural networks. Moreover, there is a lack of a **deeper discussion** on the **interpreters' behavior** faced with the outputs of the models.

By deeper discussion, we mean a discussion that:

<u>Not only considers:</u>

- "Why" and "what" must be explained; and
- "How" it could be explained.

However, also considers:

- "To whom" the model interpretability is useful;
- The relation(s) between "what" and “to whom" to explain; and
- The relation(s) between "how" and “to whom" to explain, considering the perception of who needs the explanation.

## 2.3.The “hard” problem of interpretability

The discussion on the **interpreters’ behavior** brings to light the subjective and recursive aspects of the problem of interpreting machine-learning models. All these aspects could leverage the **problem of interpretability** to the level of some problems, which are hard to deal with, such as the “problem of consciousness”.

According to Chalmers, the “easy” problems of consciousness[10] can be explained in terms of computational or neural mechanism (also known as cognitive abilities and functions), but the broader problem of consciousness goes beyond problems about the performance of cognitive functions. The “hard” problem of consciousness is the **problem of experience**, as neural functions cannot explain the subjective aspect of perception (Chalmers, 1995).

Similarly, the subjective and recursive aspects of the problem of interpreting machine learning models could also leverage the problem of interpretability to the same level of the Chalmers’ “hard” problem of consciousness.

### 2.3.1.Problems addressed

In this academic research, we address two problems of the XAI research area that are highlighted by the gaps of the taxonomy frameworks presented in this chapter. The frameworks do not address important aspects of the problem of interpretability, as they fail to consider:

- The different perceptions of different interpreters about a same model’s output;
- Non-human interpreters (*e.g.*, other systems) in the process of interpreting model outputs.

[10] The Chalmers’ “**easy**” problems of consciousness: (1) the ability to discriminate, categorize, and react to environmental stimuli; (2) the integration of information by a cognitive system; the reportability of mental states; (3) the focus of attention; the deliberate control of behavior; and (4) the difference between wakefulness and sleep.

# 3
# Research method

In this chapter, we first describe the actions proposed to answer each of the research questions listed in Section 3.1. We then propose a plan to systematically search for the available techniques to improve the interpretability of machine learning models. Finally, we use the proposed plan to estimate the order of magnitude of the number of techniques proposed so far.

## 3.1.Research steps

The research was divided into eight stages, each one with a set of actions that seek to achieve the following objectives:

1. To search for available techniques to improve the interpretability of machine learning models, we elaborate a systematic search plan.
2. To validate the inclusion and exclusion criteria of the proposed systematic search plan and answer RQ 1.1, we extract using the plan's search strings and analyze a sample of the scientific articles obtained.
3. To answer what taxonomies are proposed so far to classify techniques to improve the interpretability of machine learning models (RQ1.2), we perform a review of the taxonomies and tools proposed so far.
4. To summarize the review, we build an agglutinating taxonomy framework that summarizes the proposed taxonomies.
5. To answer the question of "how to propose an approach that considers the subjective perception of different interpreters on the outputs of a same model" (RQ2), we study the correlation between some paradigms of machine learning, and widely accepted semiotic theories.
6. To answer RQ2, we also propose a new semiotic-based approach that considers the subjective perception of different interpreters.
7. To evaluate the semiotic-based approach, we propose a new taxonomy framework and classify some selected articles.
8. To systematize the semiotic-based approach, we propose a procedure to address typical problems of interpreting machine learning models considering the subjective perception of different interpreters

## 3.2. Systematic search plan

To answer *"how to search for available techniques to improve the interpretability of machine learning models"* (RQ1.1), in this section we present a **systematic search plan** that can guide procedures for extracting **scientific databases** of **systematic mapping studies** of varying scope.

### 3.2.1. Sources of research contributions

The **systematic search plan** uses as **search databases** the set formed by the databases of all **digital libraries**, **journals,** and **conferences proceedings** indexed by the Google Scholar search engine.

### 3.2.2. Keywords and search terms

The **search terms** of the **systematic search plan** are composed by adding the keyword "interpretability" and its variations to some related keywords[11].

<u>The broader search domain</u>

To set up the borders of the mapping we compose a **broader domain** with all scientific articles whose **text** contains any combination of the following search terms:

Keyword: Machine Learning

- Search terms: machine learning;

Keyword: Model transparency

- Search terms: transparency; black-box; blackbox, black box; opacity; deep models;

<u>The focused search domain</u>

To focus on interpretability affairs, the results of the **broader domain** are restricted by considering only the academic articles matching **in their title** any combination of the following search terms:

[11] According to Kitchenham (2007), **search terms for mapping studies** are likely to return a very large number of studies. For a **mapping study**, this is less of a problem than for **systematic reviews**, as the aim here is for broad coverage rather than narrow focus.

Keyword: Interpretability

- Search terms: interpretability; interpretable; interpreting; interpretation; interpretations; interpret; understanding;

Keyword: Explainability

- Search terms: explainability; explainable; explaining; explanation; explanations; explain;

Table 3 summarizes the search terms proposed to the broader and the focused search domain.

**Table 3 - Keywords and related search terms**

| Domain | Keyword | Search terms | Where |
|---|---|---|---|
| Broader domain | Machine Learning | machine learning; | In the text |
| | Model transparency | transparency; black-box; blackbox, black box; opacity; deep models; | |
| Focused domain | Interpretability | interpretability; interpretable; interpreting; interpretation; interpretations; interpret; understanding; rationalizing | In the title |
| | Explainability | explainability; explainable; explaining; explanation; explanations; explain; visualizing; visualization | |

## 3.2.3.Search strings

The **search strings** of the **systematic search** are composed by merging the search terms of the **broader domain** in the text and the search terms of the **focused domain** in the title of the articles. Table 4 shows the six proposed search strings.

**Table 4 - Summary of the search strings**

| Search terms | Number | Search string |
|---|---|---|
| interpretability;<br>interpretable;<br>interpreting;<br>interpretation;<br>interpret;<br>explainability;<br>explainable;<br>explaining;<br>explanation;<br>explain; | 1. | (intitle:interpretability OR intitle:explainability) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") |
| | 2. | (intitle:interpreting OR intitle:explaining) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") |

| | | |
|---|---|---|
| | 3. | (intitle:interpretable OR intitle:explainable) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") |
| | 4. | (intitle:interpretation OR intitle:explanation) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") |
| | 5. | (intitle:interpretations OR intitle:explanations) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") |
| | 6. | (intitle:interpret OR intitle:explain) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") |

We define as the **data collection** of the **systematic search plan**, the sample of **scientific works** obtained by extracting from the **search database** the results of a search using the **search strings** of Table 4**, without** applying any additional criterion and after eliminating the duplicated **scientific works**.

### 3.2.4.Inclusion criteria

Depending on the goals of the research that uses the **systematic search plan**, we could work with complete **data collection** or only with a sample of it.

To obtain samples of the **data collection,** we filter their elements by using the following **inclusion criteria**:

As criteria of <u>inclusion</u>, depending on the target of the research:

- Published from **the "lower bound year"** to the "**upper bound year**"**;**

Table 5 summarizes the motivation for the inclusion criteria.

**Table 5 - Summary of the inclusion criteria**

| Type | Number | Criteria | Motivation |
|---|---|---|---|
| Inclusion criteria | IC1 | Indexed works published from "since ever" to "the research target year"; | Depending on the research, due to time constraints to completing the work. |

To apply the **inclusion criteria**, the elements of the **data collection** are filtered by their publication date to match with the date range of the inclusion criterion 1 (IC1).

### 3.2.5.Exclusion criteria

After applying the **inclusion criteria** to the data collection, the remaining elements are filtered **again** by using the following **exclusion criteria**:

As criteria of exclusion:

- Nonscientific papers;
- Scientific works which investigate the interpretation of models not obtained by techniques of machine learning, such as **fuzzy systems**.
- Scientific works which cannot be accessed by PUC-Rio domain;

Table 6 summarizes the motivation for the exclusion criteria.



**Table 6 - Summary of the exclusion criteria**

| Type | Number | Criteria | Motivation |
|---|---|---|---|
| Exclusion criteria | EC1 | Indexed works, which cannot be accessed by PUC-Rio domain; | The research was conducted in the PUC-Rio laboratory. |
| | EC2 | Indexed works, which are not scientific papers. | To bring scientific relevance to the research. |
| | EC3 | Indexed papers, which investigate models **not obtained** by machine learning techniques. | Machine learning models are on the focus of the research (as proposed by RQ1) |
| | EC4 | Indexed papers, which do not investigate interpretability of learning models. | Interpretability is the field of the research. |
| | EC5 | Indexed papers, which do not propose a new technique to improve the interpretability of ML models. | Finding new techniques to improve the interpretability is the main goal of the systematic search plan. |

To apply the **exclusion criteria**, we suggest the following **tagging actions**:

1. The **scientific works** of the **data collection** are tagged with one of the following tags:
   - "**EC1: IS NOT ACCESSIBLE",** if the work cannot be accessed using PUC-Rio proxy domain (exclusion criterion 1).

- "**EC2: IS NOT A SCIENTIFIC PAPER**", if the work is not a **scientific paper**, despite being indexed in the Google Scholar search engine (exclusion criterion 2);
- "**EC3: DOES NOT ADDRESS ML MODELS**", if the work is a scientific paper, but does not addresses machine learning models (exclusion criterion 3);
- "**EC4: DOES NOT ADDRESS MODEL INTERPRETABILITY**", if the work is a scientific paper that addresses machine learning models, but does not address the interpretability of models (exclusion criterion 4);

At the end of this step, the set of **papers** <u>not tagged</u> by any exclusion criteria compose what we name the "**INTERPRETABILITY DOMAIN" (ID).**

2. The papers of the **INTERPRETABILITY DOMAIN** are classified based on their research goal[12].

   If the work is a scientific paper that addresses the interpretability of machine learning models, but does not directly propose new **methods**, **strategies** or **approaches** to improve the interpretability, it is tagged as:
   - **"ID/EC5: DOES NOT PROPOSE A NEW TECHNIQUE"**

   If the work is a scientific paper that addresses the interpretability of machine learning models AND proposes any new **method**, **strategy**, or **approach** to improve interpretability, it is tagged as:
   - **"ID: RESEARCH DOMAIN"**

3. Additionally, the **scientific papers** that do not propose a technique, are also tagged:
   - **"CONCEPTUAL PAPER",** including taxonomy proposals; position papers; tutorial papers; etc.
   - **"MAPPING OR REVIEW",** including systematic mappings; systematic reviews; focused mappings; etc.

[12] A paper could be classified in more than one research type.

- **“METHOD APPLICATION”,** including application reports of previously proposed techniques; papers which compares methods, etc.
- **“OTHER TYPE OF RESEARCH”,** if none of the above conditions is true.

### 3.2.6.Data collection

In order to make a sensitivity analysis of the extractions in relation to the range of the publication date, we set up the **data collection** with **upper bound year** of 2017 and **lower bound year** of “since ever”.

We first extracted from the **search database** the results of a search using the **search strings of** Table 4**,** and then we eliminated its duplicated elements. Table 7 details the extraction[13] of each **search string**.

**Table 7 - Data extraction summary**

| Search terms | Search string | Number of hits |
|---|---|---|
| interpretability; interpretable; interpreting; interpretation; interpret; explainability; explainable; explaining; explanation; explain; | (intitle:interpretability OR intitle:explainability) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") | 135 |
| | (intitle:interpreting OR intitle:explaining) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") | 168 |
| | (intitle:interpretable OR intitle:explainable) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") | 267 |
| | (intitle:interpretations OR intitle:explanations) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") | 89 |
| | (intitle:interpret OR intitle:explain) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") | 41 |

[13] We use the software Harzing’s Publish or Perish 6 (Harzing, 2007) to manage and report the search analytics. The query reports (one for each research string) of Publish or Perish 6 are shown in the Appendix II – Data extraction - Public or Perish reports.

| | | |
|---|---|---|
| | (intitle:interpret OR intitle:explain) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") | 41 |
| Total returned | | **1,060** |

We found 12 duplicated works among the sample of the extracted 1,060 works. After eliminating them, the **data collection** in 31-Dec-2017 counted 1,048 works.

Number of works per year

To analyze the impact of the publication year to the coverage of the **systematic mapping plan**, we plotted the number of the works of the **data collection** for each year of publication. Figure 4 and Table 8 show the count distribution of the **data collection** per year.



**Table 8 – Distribution of the data collection by the year of publication**

| Year | Hits | Year | Hits | Year | Hits |
|---|---|---|---|---|---|
| 2017 | 260 | 2005 | 21 | 1993 | 10 |
| 2016 | 129 | 2004 | 17 | 1992 | 8 |
| 2015 | 78 | 2003 | 20 | 1991 | 3 |
| 2014 | 79 | 2002 | 12 | 1990 | 4 |
| 2013 | 61 | 2001 | 8 | 1989 | 4 |
| 2012 | 59 | 2000 | 7 | 1988 | 2 |
| 2011 | 60 | 1999 | 11 | 1987 | 4 |
| 2010 | 30 | 1998 | 9 | 1985 | 2 |
| 2009 | 38 | 1997 | 6 | 1972 | 2 |
| 2008 | 19 | 1996 | 7 | - | 4 |
| 2007 | 37 | 1995 | 2 | | |
| 2006 | 29 | 1994 | 6 | | |

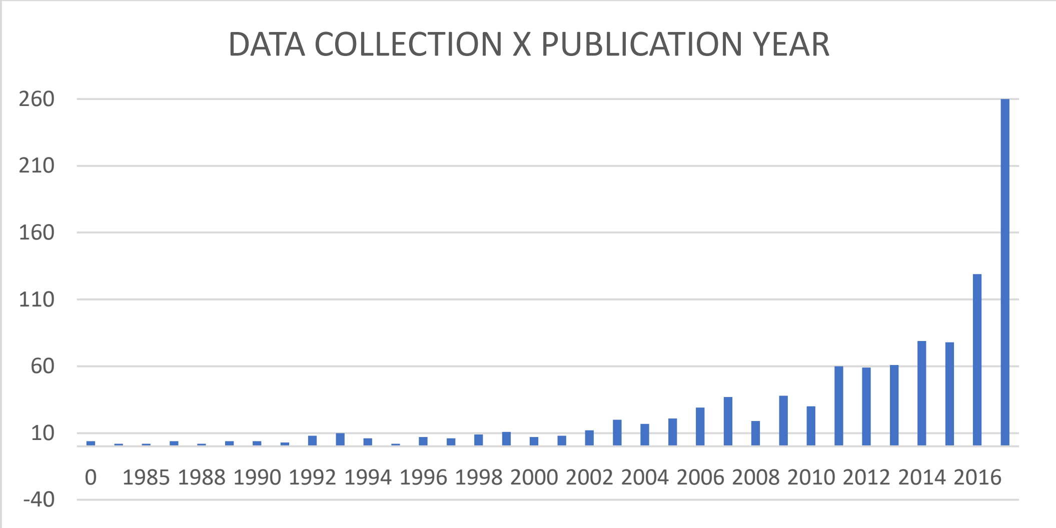


**Figure 4 -** Distribution of the data collection by the year of publication

<u>Coverage rate</u>

To estimate the impact of the **inclusion criterion** 1 (IC1) for the number of works of **data collection,** we formulate the **coverage rate** $\boldsymbol{CR}(\boldsymbol{y})$ for **inclusion criteria** that considers papers from the **lower bound year** $(LBY)$ to the **upper bound year** $(UBY)$, as being:

$$\boldsymbol{CR}(\boldsymbol{y}) = \frac{\sum_{i=y}^{UBY} \boldsymbol{N}(\boldsymbol{i})}{\sum_{i=LBY}^{UBY} \boldsymbol{N}(\boldsymbol{i})}$$

Where:

$\boldsymbol{N}(\boldsymbol{i})$ *is the total number of works of the data collection published in year i*

**Equation 1 – Coverage rate x inclusion criterion 1**

For example, CR=24.5% for y=2107 in the **data collection** shown in Table 8 means that applying an **inclusion criterion** which selects papers from 2017, the mapping covers about 24.9% of the scientific works indexed by the Google Scholar search engine.

With the numbers of Table 8 we calculated the **coverage rate** for each previous **lower bound year.** Table 9 and Figure 5 show how the **coverage rate** increases as the **lower bound year** of the **inclusion criteria increases**.

**Table 9 – Publication date's distribution and the coverage rate**

| Year | Coverage rate (%) | Year | Coverage rate (%) | Year | Coverage rate (%) |
|---|---|---|---|---|---|
| 2017 | 24.9% | 2005 | 86.2% | 1993 | 97.2% |
| 2016 | 37.3% | 2004 | 87.8% | 1992 | 98.0% |
| 2015 | 44.7% | 2003 | 89.8% | 1991 | 98.3% |
| 2014 | 52.3% | 2002 | 90.9% | 1990 | 98.7% |
| 2013 | 58.1% | 2001 | 91.7% | 1989 | 99.0% |
| 2012 | 63.8% | 2000 | 92.3% | 1988 | 99.2% |
| 2011 | 69.5% | 1999 | 93.4% | 1987 | 99.6% |
| 2010 | 72.4% | 1998 | 94.3% | 1985 | 99.8% |
| 2009 | 76.1% | 1997 | 94.8% | 1972 | 100.0% |
| 2008 | 77.9% | 1996 | 95.5% | | |
| 2007 | 81.4% | 1995 | 95.7% | | |
| 2006 | 84.2% | 1994 | 96.3% | | |



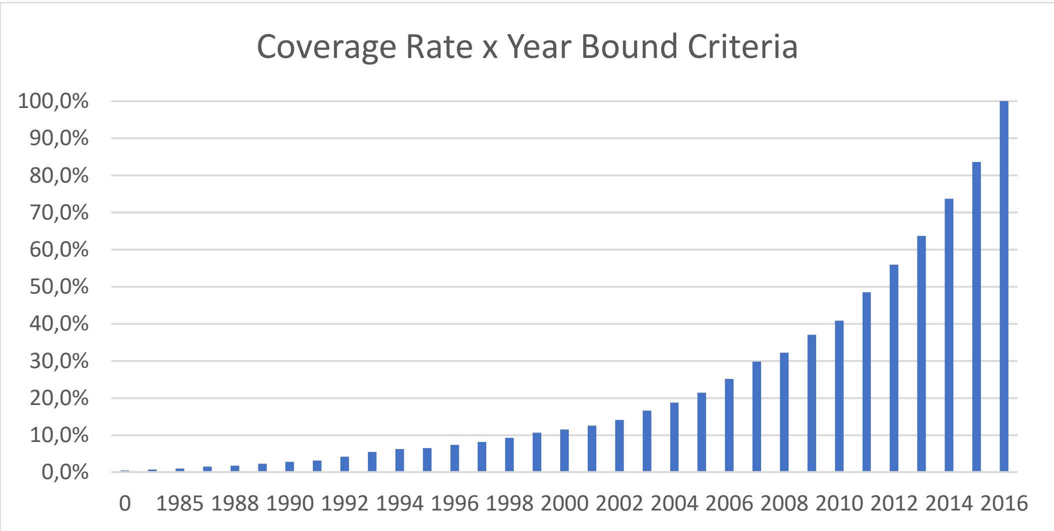


**Figure 5 –** Coverage rate per year bound criteria

## 3.3.Evaluating the plan

In this section, we evaluate the **systematic search plan** proposed in the previous section. First, we extracted the **data collection** from the research database using the plan's **search strings**. We then chose the lower and the **upper bound years** to compose the **research domain.** Finally, we estimated the order of magnitude of the number of techniques to improve the interpretability of machine learning proposed by scientific researches until 31-Dec-2017.

### 3.3.1.Research domain

We performed the **systematic mapping plan** with an **upper bound year** of 2017 and a **lower bound year** of 2017 to set up the associated r**esearch domain**, representing a statistically significant sample of 24.9% of the **data collection**.

After applying the inclusion and exclusion criteria, as proposed by the systematic mapping plan described in Section 3.2, we have selected 109 scientific papers. Table 10 summarizes the process of applying the inclusion and exclusion criteria to obtain the research domain.

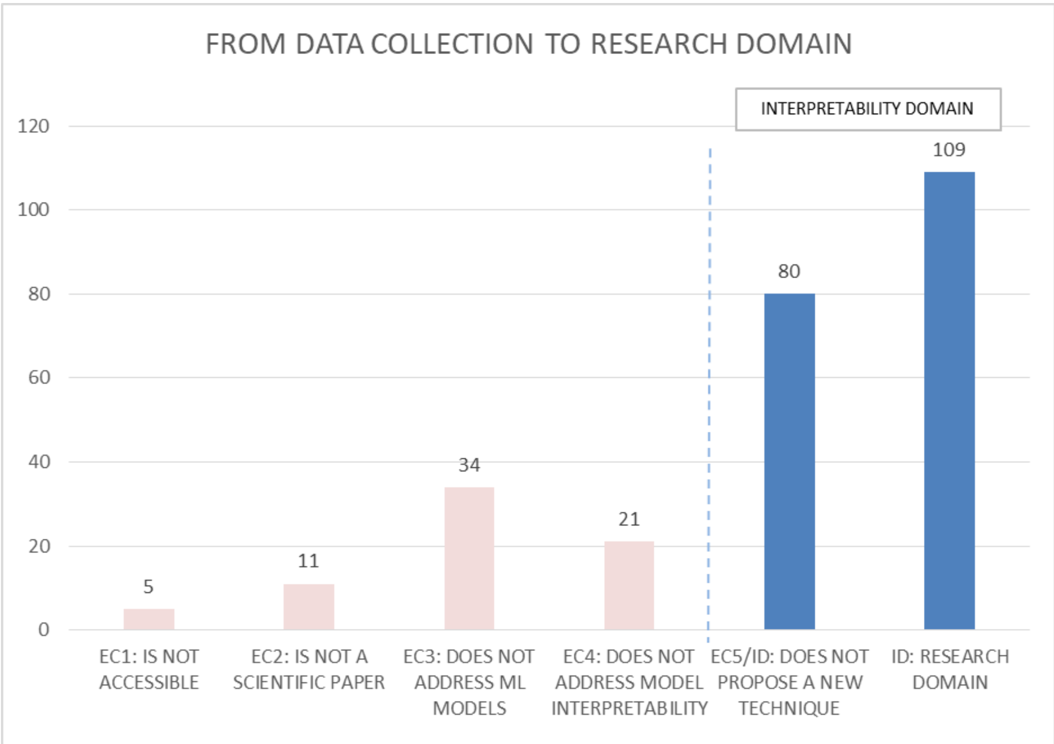


**Figure 6 shows the results of the tagging actions.**

**Table 10 – From data extraction to research domain**

| **Data set** | **Action** | **Number of articles** | **Number of articles remained** |
|---|---|---|---|
| **Data extraction** / IC1 | PoP query reports (Appendix II) / applying upper bound year = 2017 | - | **1,060** |

| | | | |
|---|---|---|---|
| Data collection 1 | Eliminating duplicates and undefined publication year. | 12 | 1,048 |
| IC1/ Data collection 2 | Applying lower bound year = 2017 | 260 | 260 |
| EC1 | Cutting works that cannot be accessed by PUC-Rio domain. | 5 | 255 |
| EC2 | Eliminating works that are not scientific papers. | 11 | 244 |
| EC3 | Eliminating works that do not address machine learning models. | 34 | 210 |
| EC4 / Interpretability Domain | Eliminating works that do not address interpretability of machine learning models. | 21 | 189 |
| ID/EC5 / **Research Domain** | Eliminating papers that do not propose new techniques to improve interpretability. | 80 | **109** |

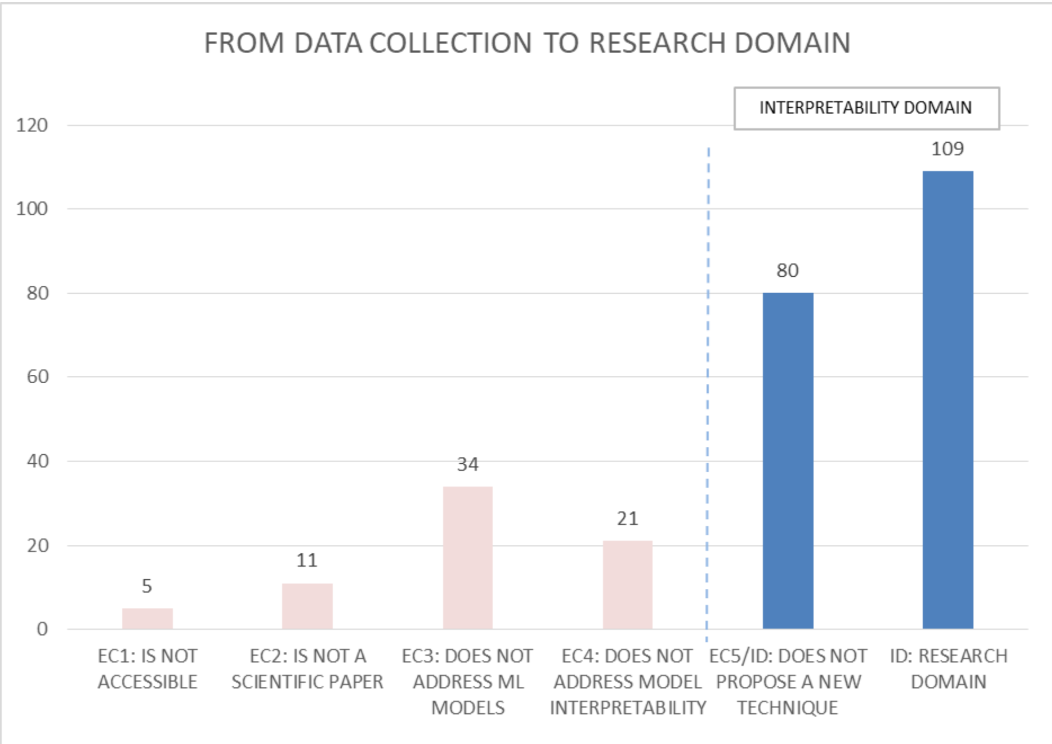


**Figure 6 – Tag distribution of setting up the research domain**

### 3.3.2.Total number of techniques

To estimate the order of magnitude of the number of techniques to improve the interpretability of the machine learning models proposed by scientific research up to December 31, 2017, we assume that the calculated contribution rate of the research domain sample is a good estimator for the contribution rate of the entire population.

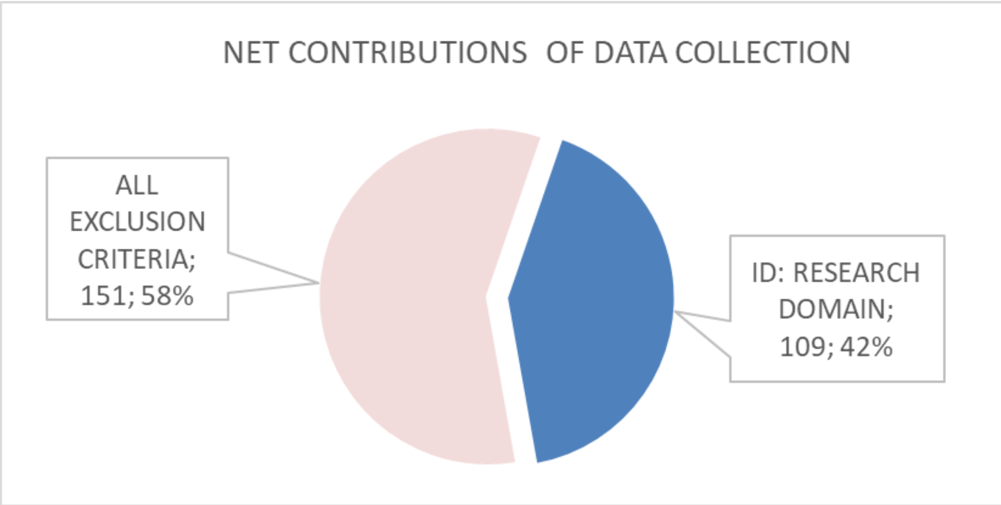


**Figure 7 – Research domain from the data collection**

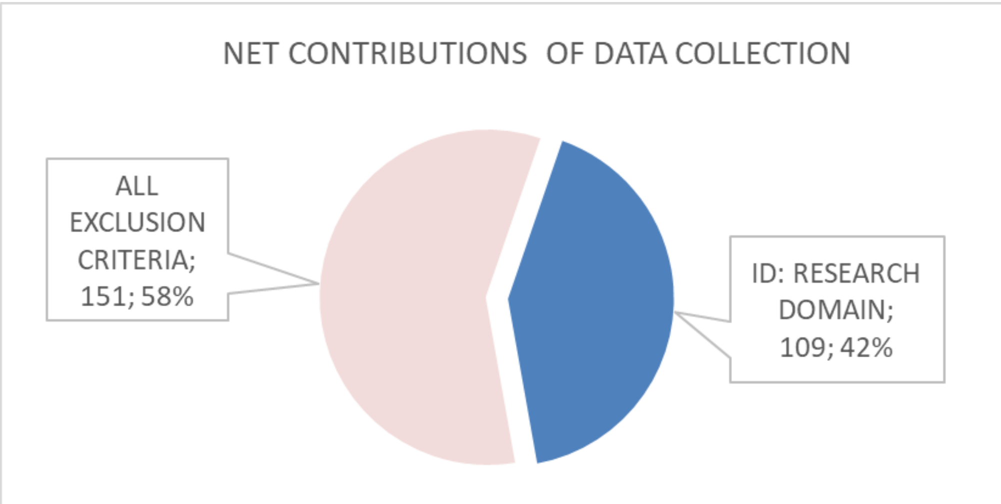




Figure 7 shows that of all the published works of this sample, only 42% of them effectively contribute with techniques to improve the interpretability of machine learning models. Thus, given that the population - aka data collection - has 1,044 scientific works, we estimate the number of proposed techniques of the order of $4x10^2$.

# 4
# A semiotic view on interpretability

Section 2.2 addresses the two gaps of the current conceptual frameworks, which are contributing to slow the development and widespread use of **XAI Systems** in our daily lives. They fail to consider: (1) the different perceptions of different interpreters about a same model's output; (2) Non-human interpreters in the process of interpreting model outputs. In this chapter, we propose a new approach to deal with the problem of interpretability, which aims to fill those gaps.

First, we outline the conceptual foundations of the new view[14], and then we present a proposal of how to approach, characterize and solve typical problems of interpretability by applying these concepts. Finally, we propose a procedure to improve the interpretability of machine learning models considering the semiotic approach for solving typical problems of interpretability.



## 4.1.Conceptual foundations

One of the foundations for understanding how to improve the interpretability of any model is the definition of "interpretation". Unfortunately, the literature on proposing solutions for increasing the interpretability of models does not address this definition in detail. Much of the works assume that the definition of "interpretation" is of common knowledge and does not semantically explore what "interpret" means.

This section presents some conceptual foundations capable of supporting a new approach to the problem of interpretability based on some theories, which formalize the generation of meaning by the process of interpreting. In the next sections, we show how these concepts will help to build a procedure to deal with a typical problem of interpretability.

[14] To better understand the discussions in this chapter, it is necessary for the reader to know the fundamentals of **semiotics** presented in Appendix I.

### 4.1.1.Interpretation as “mappings”

A rather pragmatic way of deal with interpretations that can be used for practical purposes is to view an interpretation as a mapping. Montavon et al. (2017) present a useful description of what differentiates "interpretation" from "explanation" in respect to the outputs of a target model. According to them:

- An **interpretation** is the mapping of an **abstract concept of the model** onto a domain that the human can make sense of.
- An **explanation** is the collection of features of the interpretable domain that have contributed for a given example to produce an output.

Although Montavon et al. (2017) refer in their work only to abstract concepts of **artificial neural networks**; to develop the semiotic view on interpretability we extend this range to include (1) other types of machine learning models, and (2) other possible components of the models. In this research, we consider as **abstract concepts** of machine learning models the following:

- All the input features or any subset of them, such as convolutions of an image, etc.;
- All the components of the model data structure, such as ANN layers, Bayesian nodes, etc.;
- All kind of output features, such as output classes, regression output functions, etc. and their correspondent inputs;

Moreover, we include any “training component” used to learn the model among the potential **abstract concepts** that can be mapped onto human interpretable domains. Among others, we include:

- The training datasets (or training sets);
- The training algorithms.

Similarly, we also extend the concept of collections of interpretable domains to include, in addition to traditional ones, any tacit or explicit information transmitted by symbols or entities that have meaning for humans. As examples of **human interpretable domain,** we consider, among others:

- Basic signs, such as text, images;
- Composed signs, such as heat maps;

- Language elements, such as mathematical elements, logical languages elements;
- Semantically structured sentences of languages, such as written sentences, audiovisual content;

Table 11 summarizes the possible mappings that generate human interpretations, according to the relaxation of Montavon et al.'s (2017) definition.

**Table 11 – Interpretations as "mappings" by relaxing the Montavon et al. (2017) definition.**

| | **HUMAN INTERPRETABLE DOMAIN** | | | |
|---|---|---|---|---|
| **ABSTRACT CONCEPTS OF ANY ML MODEL** | Basic Signs | Composed signs | Language elements | Sentences of structured languages |
| Training components | **INTERPRETATIONS** (the mapping of an abstract concept onto an interpretable domain) | | | |
| Input features | | | | |
| Components of the model's data structure | | | | |
| Previous outputs/inputs | | | | |

One of the immediate advantages of adopting the view of "interpretation as mapping" is to highlight the **three** main elements of a human interpretation, which must be considered when developing **XAI Systems**. To interpret any model output, the system must clearly define:

- The abstract concept of the model;
- The human interpretable domain; and
- The "mapping rule" that associates them.

The latter element is detailed in the next section.

### 4.1.2.The mapping rule as "reasonable explanatory principle"

According to De Souza et al. (2016), **semiotics** is a multifaceted discipline where **signs** and **signification** constitute the common object of study in all cases. Moreover, according to semiotics, **signs** are the result of associations between <u>expressions</u> (*aka* <u>representations</u>) and <u>content</u> (*aka* <u>information</u>)"; and

**signification** is broadly defined as the process by which signs come into existence. Based on this definition, we adopt these foundations of semiotics to formally define the "mapping rule", which is the third element of the process of interpreting model outputs.

About the associations (or mappings) between **signs** and **meaning**, De Souza et al. state that some semiotic theories will postulate that such expression-content associations <u>are carried out by some mind</u> (individual or collective, human or nonhuman). Others will postulate that they have an <u>abstract, systemic, or logic nature</u>. Yet others will consider that these associations are <u>the result of evolutionary sociocultural processes</u>.

Regardless of the nature of the association rule that drives the interpretation of model outputs, it seems clear that this association is not "processed" in a unique way for all the potential interpreters involved in the interpretation process. People have diverse levels of knowledge and what is interpretable by one person could not be interpretable by another person. Thus, to consider the perception of each particular **interpreter** on the available interpretable domain and its **effects** on the interpretability of the model, we need some definitions that formalize this perception. To do so, we use some formal elements of the Peircean semiotics.

For the Peircean semiotics, **sign** is anything that, for somebody, under some circumstance(s) and in some respect(s), stands for something else. Moreover, the three constituent parts of a sign are **representamen** (a representation), **object** (what the representation stands for), and **interpretant** (the mediating interpretation that creates a meaningful association between the other two components).

According to Santaella (2002), the **Peircean speculative grammar** addresses formally the **interpreter** as part of the study of all kinds of signs and forms of thinking that they enable, and the formal elements involved in the **meaning making** of the explanations. De Souza et al. (2016) summarize very well the relationship between these three elements of Peircean semiotics stating that: “signs only come into existence if some mind mediates (and thus creates) the association between a representation and what this representation stands for. The mediation is an **interpretation**.”

<u>Abduction, explanatory hypothesis, and semiosis</u>

Although the **principle** that "processes" this mediation may often be **previously established or incorporated** into the “physical structure” of the interpreter by, for example, humans’ previous experiences, culture, etc., many times this pre-conceptualization has not yet been established. In this case, De

Souza et al. emphasize the importance of what Peirce call **abduction** for the meaning generation process. In Peirce's words, a**bduction** is an inference (what-if) process that produces a **reasonable explanatory principle** capable of turning some surprising fact into a logical consequence of this principle. De Souza et al. state that the concept of **abduction** is important for the study of meanings in general because it describes the logic of human sense making, from practical mundane situations to elaborate philosophic argumentation. Moreover**,** they claim, "the aim of **abduction** is to create a (new) mental habit that will be used in the **interpretation** of future occurrences of the previously surprising **sign**".

Beyond **abduction**, two other concepts of the Peircean semiotics are used to design a semiotic view on the process of interpreting outputs of models. They are, according De Souza et al. (2016):

(1) **Circumstantially verifiable hypothesis** (or **explanatory hypothesis**) **is the** hypothesis that is signified and confirmed in the collection of signs that are contextually associated with the surprising fact that triggered the abductive process in the reasoner's mind.;
(2) **Semiosis** is the unlimited sense-making **abductive** process where all conclusions are provisional as they hold until they are contradicted by new facts. **.**

Finally, to develop our semiotic view on interpretability we appropriated the tools available in the semiotic theories by considering the **mapping rule** as the proposed **reasonable explanatory principle** of the **abduction** process that takes place before or during the process of interpreting model outputs. In short, according to this semiotic view, to give to potential interpreters an explanation of any model output, the **XAI system** must clearly define:

- The abstract concept of the model;
- The human interpretable domain; and
- The **reasonable explanatory principle** of the **abduction** that is conducted by the potential interpreters.

Although the **XAI system** must "know" the **abduction** process conducted by the potential interpreters, in the case of a well-defined human interpreter, the **XAI system** must "know" the personal **abduction** process conducted by the interpreter. By "knowing" the abduction process we mean knowing:

- The personal **mental habit** that is trigged while the process of interpreting the signs selected by the system, or;

- The **abduction** process, in case of a surprising **fact.**

This means that, considering interpreting a well-defined interpreter, an XAI system would need to "learn" the "reasonable explanatory principle" (or the set of "reasonable explanatory principles") that drives the abduction process of that interpreter. In our research, we will name this personal set of "reasonable explanatory principles" "**personal semiotic patterns**".

The following section proposes a way to learn some personal semiotic patterns considered in the interpretation of outputs of machine learning models, so that these patterns can be compared with the XAI system's own target model.

### 4.1.3.Interpretation as a “learning process”

Domingos (2015) presents a comprehensive classification of the principles (or paradigms) used by researchers to construct machine learning models (*aka* learners). According to Domingos (2015), there are five computational paradigms for building learners. They are:

- The **Symbolist paradigm**, in which learning is achieved through processes of manipulation of symbols (and, consequently, of languages);
- The **Bayesian paradigm**, in which learning is achieved through processes that promote the systematic reduction of uncertainties;
- The **Analogizer paradigm,** in which learning is achieved through processes of searching for similarities;
- The **Connectionist paradigm**, with their models based on neural networks, which try to simulate the learning process of biological neocortex; and
- The **Evolutionist paradigm**, which try to simulate the learning process of the biological evolution with genetic algorithms;

Curiously, the learning processes presented by Domingos are quite similar with the process, which results in a human interpretation, as described in the previous section. Moreover, three of the five learning paradigms highlighted by Domingos have a direct adherence to the way the semiotics theories explain the meaning-making of a typical interpretation process. In this sense, they can be grouped as follows:

- The <u>Connectionist</u> and <u>Evolutionist</u> paradigms are inspired by the learning observed in biological processes. In this research, we call them paradigms inspired by biology, or **biology-inspired paradigms**.
- The <u>Symbolist</u>, <u>Bayesian</u>, and <u>Analogizer</u> paradigms seem to maintain a logical adherence to the foundations of classical semiotic theories, especially with the semiotic theories of Peirce and Eco. In our research, we call them paradigms based on semiotics, or **semiotics-based paradigms**.

Because of this adherence, in the semiotic view on interpretability we use the **semiotics-based paradigms** to classify the preferences of an interpreter facing the options of "meaning-making" capable of solving typical problems of interpreting the outputs of a model.

However, a typical problem of interpretability may involve other obstacles beyond the generation of meaning that must be separated from the analysis, so that the semiotic theories can be applied. To deal with this need, the next section proposes a way to split the problem of interpretability in some subproblems.



## 4.2.Approaching a typical problem of interpretability

Within the **semiotic view on interpretability,** we propose to look to any problem of interpreting the outputs of a target model by an individual interpreter as the <u>combination</u> of three more tractable subproblems. They are:

1. A problem of accessing the components of the target model;
2. A problem of generating meaning for potential interpreters;
3. A problem of communicating a collection of the interpretable domain to the target interpreter.

The following sections detail each of these subproblems, as well as the elements, which characterize each one.

### 4.2.1.The "access" subproblem

The problem of interpretability is directly related to the level of access that the XAI system has to the components and to the input domain of the target model.

This level of access is a determining factor for the solution of what we call the **access subproblem**.

The access subproblem is the problem that most of the current classical taxonomies exclusively address. As shown in the section 2.2, the white-box approach and the black-box approach (or agnostic approach) are the two macro strategies addressed by Lipton (2016) and Ribeiro et al. (2016) with respect to the restrictions of accessing the components of a model.

The elements that characterize the **access subproblems** are:

- The level of access to the model components
- The level of access to the model input domain

### 4.2.2.The “meaning-making” subproblem

Interpreting the outputs of a model is a process that presupposes a previous generation of meaning by an XAI system, regardless of whether these outputs are outputs of the target model or of an auxiliary model.

Compared to the current classical view on interpretability, the meaning-making subproblem is equivalent to the class of problems for which Gunning (2017) propose that the solutions be classified in the class "Psychology".

The view of interpretation as mapping is very useful to address meaning-making sub–problems, as it helps to highlight the elements of the interpretation that could be considered when generating meaning. Using the Peircean semiotics to characterize the meaning-making subproblem, it is the problem of choosing suitable **representamen** to present to the potential interpreters, where the **interpretant** is an interpretable domain for the potential interpreters and the **object** is an abstract concept of the model. A meaning-making subproblem must be tackled in two steps. The first one involves the definition of the first two elements of the "interpretation as mapping" (the model's concepts and the interpretable domain) and the second step involves the definition of the “reasonable explanatory principle” (mapping rule).

The elements that characterize the **meaning-making subproblems** are:

- The mental habits of the potential interpreters;
- The learning preferences of the potential interpreters.

### 4.2.3.The "interaction" subproblem

It is reasonable to assume that the way an **XAI system** interacts with a potential interpreter can also affect its interpretability. Even in cases where interpretations are properly mapped by the two subproblems above, if the elements are not properly communicated, the model's outputs may not be understood.

Although the present approach considers the interaction between **XAI Systems** and the target interpreters as an integral part of the interpretation process, the rank of possible solutions to solve the interaction subproblem can be treated in isolation through HCI theories. Thus, the elements that characterize the interaction subproblem in question are all elements considered by these theories.

### 4.2.4.Featuring a typical problem of interpretability

Consider a typical problem of interpretability the problem of providing an explanation, in the form of a set of human-interpretable signs, about the outputs of a **target model**[15] to a particular interpreter. This section proposes some multiple-choice questions to help raising the <u>variables</u> and the <u>constraints</u>, which characterize the subproblems that make up this kind of problems. The questions are:

***Question 1:*** *Who are the potential interpreters of the target model outputs?*

The answer to this question defines the range of the best possible **interpretable domains** for the explanation system. Based on this constraint, the potential interpreters of the model's outputs can be grouped in:

1. Non-expert humans;
2. Expert humans;
3. Non-human interpreters;

***Question 2:*** *What level of access does the explanation system have to the components of the target model?*

The answer to this question defines the range of the possible **abstract concepts** of the target model, which are accessible by the explanation system to associate these concepts with interpretable domains for the potential interpreters. Based on this constraint, the levels of access by the explanation system to the model's components can be grouped in:

[15] We use "target model" to differentiate the model that is the target of the interpretation process from to the other auxiliary models used in the same task.

1. Full access to the model's components;
2. Partial access to the model's components;
3. No access to the model's components.

***Question 3:*** *What level of access does the explanation system have to the input domain of the target model (TM)?*

The answer to this question defines the range of the **model induction** possibilities for input/output simulations. Based on this constraint, the possible degrees of access of the range of input domain can be grouped in:

1. Input domain is finite, and the TM is accessible to input simulations;
2. Input domain is infinite (or very large) and the TM is accessible to input simulations;
3. Input domain is finite, but the TM is not accessible to input simulations;
4. Input domain is infinite (or very large) and the TM is not accessible to input simulations;

***Question 4:*** *What is the role of humans in the outputs of the target model?*

The answer to this question defines the possible strategies supported by the **explanation system** to interact with the interpreters. Based on this constraint, role of humans in the outputs of the target model can be grouped in:

1. Passive interpreter;
2. Interpreter in the loop;
3. Other humans in the loop;

Choosing variables and constraints driven by the above list may provide to the designers of XAI Systems an important advisory to start solving typical problems of interpretability.

## 4.3. Filling the gaps

This section suggests two approaches to deal with typical problems of interpretability, which aim to consider both the different perceptions of different interpreters about a same model's output; and non-human interpreters in the process of interpreting model outputs.

### 4.3.1.Personal semiotic patterns

The advantage of classifying the interpretability solutions based on their similarity to one of the three semiotics-based paradigms is to allow an immediate association of the "semiosis" of the technique with the "semiosis" characteristic of the potential interpreters. This association can be used to design XAI Systems that can propose different techniques of interpretability for different interpreters. By "semiosis" of the potential interpreters, we mean a set of learning preferences of the interpreters. In other words, a set of personal patterns that can be learned by similarity with one or more semiotics-based learning paradigms.

By the semiotic view on interpretability, we propose to solve the problem of interpreting the outputs of a target model considering different perceptions of different interpreters about a same model's output:

1. To identify and characterize the subproblems of the main problem of interpretability.
2. To learn some semiotic patterns of personal interpretation from the individual interpreter;
3. To suggest the technique to interpret the model outputs that most resemble those learned personal patterns.



### 4.3.2.Extending interpretations with chains of mappings

Dhurandhar (2017) takes inspiration from the theory of computation to claim that a <u>language</u> is classified as regular, context free, or something else based on the strength of the machine (i.e. program) required to recognize it. Inspired by Dhurandhar's statement, we propose an extension in the range of the definition of "collections of interpretable domains" to help fill the gap of not considering the non-humans in the loop of interpretation process. We suggest adding to the list of the collections of interpretable domains some reports of "human and/or non-human entities" which are **trustworthy** for the target interpreter. These "**trustworthy interpreters**" can be:

- <u>Other human interpreters</u>, such as other human experts or non-experts;
- <u>Non-human "interpreters"</u>, such as other explanation systems based on trustworthy[16] models, such as statistical models.

[16] Trustworthy because they are interpretable, for example.

We include trustworthy entities in the list as they really can act as collections onto which some abstract concepts of the model can be “mapped” to generate an interpretation. Moreover, assuming that "trustworthy" interpreters can also have other "trustworthy" interpreters, it is possible to design explanation systems based on "trust chains" of interpreters, which are able to interpret even more outputs that are complex to non-experts.

## 4.4.A procedure to deal with the problem of interpretability

This section proposes a systematic procedure to improve the interpretability of machine learning models considering the semiotic approach for dealing with a typical problem of interpretability presented in the previous sections. The procedure roughly consists of sequentially solving each one of the subproblems that characterize a problem of interpretability: the **access**, the **meaning-making**, and the **interaction** subproblems.

To solve the **access subproblem**, two strategies are commonly employed, depending on the level of access to the components of the target model:

4. Directly solve the **meaning-making subproblem** in cases where the target model is fully accessible by the XAI system;
5. Directly solve the **meaning-making subproblem** of a "**relaxed**" problem of interpretability that aims to set some boundaries for the interpretations of the original problem.

To solve the **meaning-making subproblem**, both of the main problem and the relaxed problem, the suggested sequence is:

1. Define the interpreter and his/her/its **personal semiotic pattern**.
2. Define which semiosis process best fits to the interpreter's semiotic pattern.
3. Define the abstract concept of the model from a list of concepts available to be mapped by the defined semiosis process.
4. Define the interpretable domain that best fits the interpreter.

Finally, use HCI tools and theories to define the strategy to communicate the target interpreter the collection of the chosen interpretable domain which best explains the output of the target model.

# 5
# Evaluating the semiotic view of interpretability

In this chapter, we present the actions that we performed to evaluate the use of the **semiotic view of interpretability** to classify techniques, which interpret machine learning outputs.

First, we propose a new **taxonomy framework** to classify these techniques based on the fundamentals of the **semiotic view on interpretability,** then we use the proposed taxonomy to classify the papers of a sample extracted from the **research domain** of Section 3.3, and, finally, we analyze the usefulness of the **semiotic view on interpretability** for this kind of classification task.

## 5.1.A taxonomy for the semiotic view

According to Bruno and Richmond (2003), the six steps to follow in developing a taxonomy are 1. plan and gather data; 2. build a draft taxonomy; 3. pilot; 4. refine and finalize; 5. user training, and 6. ensure continued development. However, this research seeks to perform only steps 1 and 2, which already contribute to achieving the goals of Explainable AI addressed in section 1.2.

In this sense, the categories of a taxonomy framework to classify techniques that improve the interpretability of machine learning models should guide planners and consultants in the most common choices when they plan to apply any method, strategy or approach to interpret these kinds of models.

### 5.1.1.Categories of the taxonomy framework

In order to define the categories of the framework, let us consider the typical problem of interpreting the outputs of a **target model** by **potential interpreters** as a <u>combination</u> of the following problems:

1. The problem of accessing the **target model's components** (*aka* the **access subproblem)**;
2. The problems of generating the meaning for each potential interpreters (*aka* the **meaning-making subproblems)**;

3. The problem of communicating a collection of human interpretable domain to each potential interpreter (*aka* the **interaction subproblem)**.

Let us also consider that the techniques proposed to interpret the outputs of the target model can be characterized by the elements that characterize the **access**, **meaning-making** and **interaction** subproblems (according to section 4.2) as the following:

- The level of access to the model components;
- The level of access to the model input domain;
- The mental habits of the potential interpreters;
- The learning preferences of the potential interpreters;
- The available interface to interact with the potential interpreters.

Based on the above assumptions, we propose that the techniques are sequentially classified according to their:

**1. Access to the components of the target model**

**Techniques** that suppose a full access to any component of the target model are classified in the "DIRECT INTERPRETATIONS" class, and the techniques that seek to infer the behavior of the target model by using other techniques to direct interpret surrogate models are classified in the "RELAXED INTERPRETATIONS" class.

**2. The nature of the relaxation**

The techniques classified in the “RELAXED INTERPRETATIONS" class do not directly solve the problem of interpreting the target model, but an "**associated relaxed problem**" of directly interpreting a substitute model whose access to the components is unrestricted.

The techniques that seek to solve an "associated relaxed problem" are also classified according with the classification of two components.

2.1 The **NATURE of the surrogate models** are classified in the following classes:

“A REGULARIZATION OF TM”

“AN EXPLAINABLE MODEL”

“A NON EXPLAINABLE MODEL”

2.2 The **ACCESS LEVEL to the surrogate models** is classified in the following classes:

"FULL ACCESS"

"ACCESS ONLY FOR SIMULATIONS"

"NO ACCESS"

2.3 The **SCOPE of the relaxation** is classified in the following classes:

"LOCAL INTERPRETATIONS"

"GLOBAL INTERPRETATIONS"

Figure 8 shows a schematic representation of relaxed interpretations.

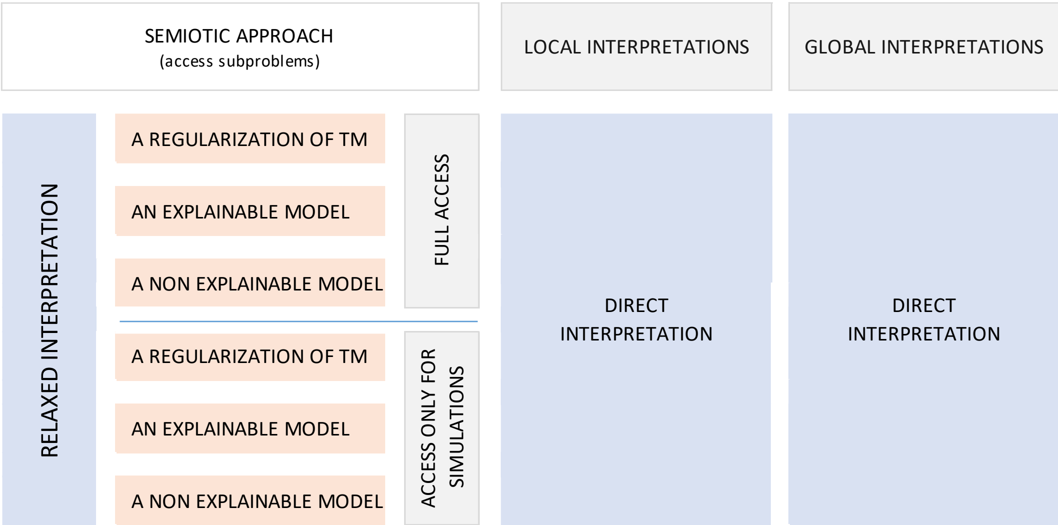


**Figure 8 – Schematic representation of relaxed interpretations**



**<u>3. The elements of direct interpretation</u>**

As **direct interpretations** suppose the <u>association</u> of a target model's <u>concept</u> with a <u>human interpretable domain</u>, the **techniques** classified in the "DIRECT INTERPRETATIONS" class are also classified according with the classification of these <u>three</u> components.

3.1 The **target model's concepts** are classified in the following classes:

"INPUT FEATURES"

"COMPONENTS OF THE DATA STRUCTURE"

"PREVIOUS EXAMPLES"

3.2 The **human interpretable domain** is classified in the following classes:
“ELEMENTARY SIGNALS FOR SENSORY PERCEPTION”
“COMPOSITE SIGNALS”
“ELEMENTS OF LANGUAGES”
“SEMANTICALLY STRUCTURED SENTENCES OF LANGUAGES”

3.3 The **association rules** are classified in the following classes:
“SYMBOLIC-BASED ASSOCIATIONS”
“SYMILARITY-BASED ASSOCIATONS”
“ASSOCIATION BASED IN REDUCING UNCERTAINTIES”

Figure 9 shows a schematic representation of direct interpretations.

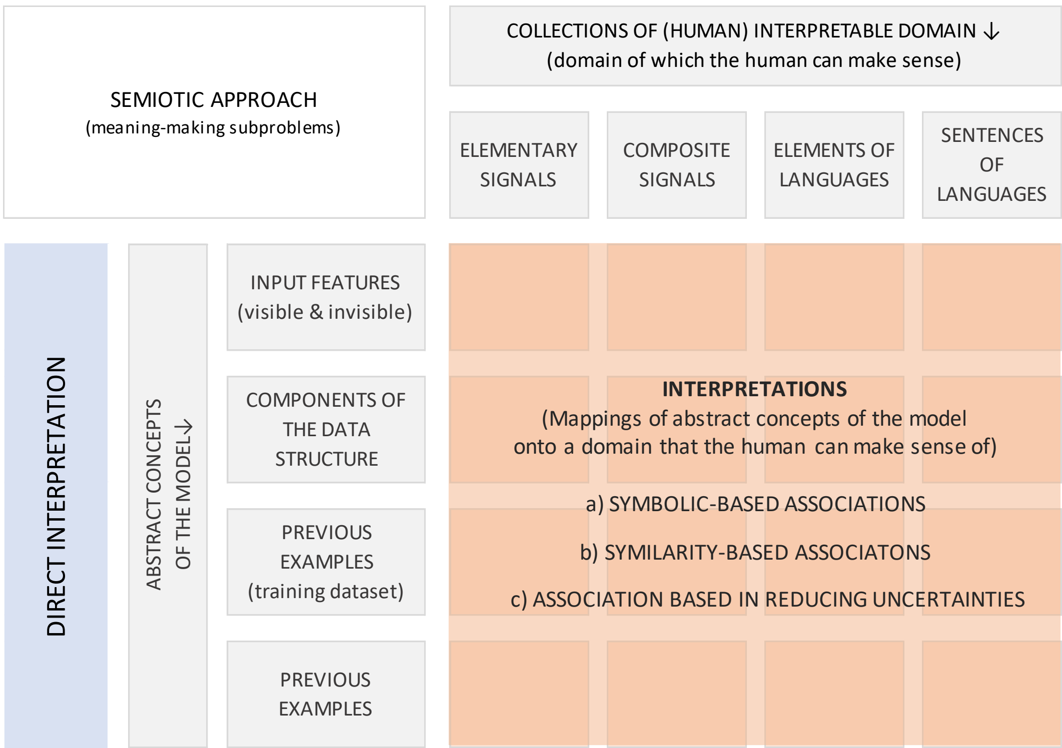


**Figure 9 – Schematic representation of direct interpretations**

<u>**4. The interaction with the potential interpreters**</u>

Although the interaction between XAI Systems and potential model interpreters is fundamental to the process of dealing with the problem of interpretability, this research will not deepen the ways of solving the **interaction subproblems** by understanding that these are problems that are already well formulated and adequately addressed by the IHC area through their theories.

### 5.1.2.Auxiliary tables

In this section, we suggest lists to assist in the classification of the techniques.

**Table 12 - Auxiliary table – model concepts**

| **CONCEPTS OF THE MODEL** |
|---|
| |
| **++ INPUT FEATURES** |
| **INPUT FEATURES** |
| **INPUT FEATURES OF TRAINNING DATASET** |
| **HIDDEN INPUT FEATURES** |
| **OTHER INPUT FEATURES** |
| |
| **++ COMPONENTS OF THE DATA STRUCTURE** |
| **+ANN STRUCTURES** |
| ANN's HIDDEN LAYERS |
| ANN's WEIGHTS |
| **OTHER COMPONENTS OF DATA STRUCTURE** |
| |
| **++ PREVIOUS EXAMPLES** |
| **EXAMPLES OF PREVIOUS OUTPUTS** |
| **EXAMPLES OF THE TRAINING DATASET** |
| |

**Table 13 - Auxiliary table – interpretable domain**

| **INTERPRETABLE DOMAIN** |
|---|
| **++ ELEMENTARY SIGNALS FOR SENSORY PERCEPTION** |
| **+ VISUAL SIGNALS** |
| SALIENCY MAPS |
| HEAPMAPS |
| **AUDITORY SIGNALS** |
| **TACTICLE SIGNALS** |
| **SMELL SIGNALS** |
| **TASTE SIGNALS** |
| |
| **++ COMPOSITE SIGNALS** |
| **+ COMPOSITE VISUAL SIGNALS** |
| IMAGES |
| CHARTS |
| GRAPH DIAGRAMS |
| 2-DIMENTIONAL GRID (t-SNE) |
| OTHER VISUAL ELEMENTS |
| **+ COMPOSITE AUDITORY SIGNALS** |
| SOUNDS |
| OTHER SONOROUS ELEMENTS |
| **OTHER COMPOSITE SIGNALS** |
| |
| **++ ELEMENTS OF LANGUAGES** |
| **+ LOGICAL ELEMENTS** |
| TRUE, FALSE, AND, OR OPERATORS |
| DNF - DISJUNCTIVE NORMAL FORM (OR-OF-ANDS) |
| OTHER LOGIC OPERATORS |

| |
|---|
| **+ ELEMENTS OF NATURAL LANGUAGE** |
| LANGUAGE APHABETS AND CHUNKS |
| SET OF NUMBERS AND METRIC SYSTEMS |
| OTHER NATURAL LANGUAGE ELEMENTS |
| **+ ELEMENTS OF MATHEMATICS** |
| ALGEBRIC OPERATORS |
| OTHER MATH OPERATORS |
| **+ ELEMENTS OF DATA STRUCTURES** |
| OTHER ELEMENTS OF DATA STRUCTURE |
| **OTHER ELEMENTS OF LANGUAGES** |
| |
| **++ SEMANTICALLY STRUCTURED SENTENCES OF LANGUAGES** |
| **+ SENTENCES OF PROPOSITIONAL LOGIC** |
| FORMAL SENTENCES |
| DECISION SETS / RULE SETS IN DNF |
| OTHER SENTENCES OF PROPOSITIONAL LOGIC |
| **+ NATURAL LANGUAGE WRITEN SENTENCES** |
| WORDS, TEXTS |
| RULE LISTS IN NATURAL LANGUAGE |
| OTHER S WRITEN ENTENCES |
| **+ NATURAL LANGUAGE SPOKEN SENTENCES** |
| VERBAL SPEECHES |
| MUSIC |
| OTHER SPOKEN SENTENCES |
| **+ AUDIOVISUAL CONTENT** |
| AUDIOS |
| VIDEOS |
| OTHER AUDIOVISUAL CONTENT |
| **+ DATA STRUCTURE** |
| SETS AND COLLECTIONS |
| ARRAYS |
| GRAPH STRUCTURES G(V,E) |
| BAYESIAN NETWORKS |
| FUZZY COGNITIVE MAPS |
| OTHER DATA STRUCTURES |
| **+ MATH SENTENCES** |
| FORMULAS |
| GRADIENTES |
| DECISION TREES / DECISION PATHS |
| OTHER MATH SENTENCES |
| OTHER SENTENCES OF LANGUAGES |
| INTERPRETABLE DOMAIN FOR HUMANS |

**Table 14 - Auxiliary table – mapping rule**

| **MAPPING RULE** |
|---|
| **++ SYMBOLIC-BASED ASSOCIATION** |
| **+ RULES** |
| DEDUCTION |
| INVERSE DEDUCTION |
| MONOTONITICY TREND |
| **+ SYMBOL ASSOCIATIONS** |
| SIMILARITY (SYMBOL) |
| **+ ASSOCIATION BASED ON THE IMPORTANCE** |
| ATTENTION |
| IMPORTANCE SCORE |
| OTHER RULE-BASED ASSOCIATION |

| |
|---|
| **++ SYMILARITY- BASED ASSOCIATONS** |
| **+ ICON ASSOCIATIONS** |
| SIMILARITY BY APARENCE (ICON - IMAGE) |
| SIMILARITY IN RELATIONS (ICON - DIAGRAM) |
| SIMILARITY IN MEANING (ICON - METAPHOR) |
| **+ INDEX ASSOCIATIONS** |
| SIMILARITY BY REFERENCE (INDEX) |
| OTHER SYMILARITY-BASED ASSOCIATION |
| |
| **++ ASSOCIATION BASED IN REDUCING UNCERTAINTIES** |
| REJECTION OF ALTERNATIVE CHOICES |
| OTHER UNCERTANTY-BASED ASSOCIATION |

## 5.2.Classification

In this section, we present the actions performed to classify some techniques proposed by XAI research area.

### 5.2.1.Validation domain

We extracted a sample of the **research domain** - presented in Section 3.3 - , choosing 79 scientific articles that were cited in the research works selected in Table 1. Table 15 summarizes the results of obtaining the **validation domain** of the **research domain**.

**Table 15 – From research domain to validation domain**

| Data set | Action | Number of articles | Number of articles remained |
|---|---|---|---|
| **Research domain** | The domain used to validate the search plan's inclusion and exclusion criteria described in Section 3.2 | - | 102 |
| **Validation Domain** | Choosing the works that were cited in the scientific works of Table 1. | 79 | 79 |

In Appendix 1, we detail the authors and titles of the scientific articles chosen.

### 5.2.2.Classifying and counting the results

Based on the abstract, the selected articles of **validation domain** were classified (1) in the categories of the taxonomy framework with the traditional view -presented in section 2.2-, and (2) in the categories of the taxonomy framework based on the semiotic vision -presented in section 5.1. Table 13 and Table 17 show the results of the classification[17].

**Table 16 – Results of the classification from the traditional point of view**

| Class Number | Class Description | Number of proposed techniques |
|---|---|---|
| **1.** | **WHITE BOX APPROACHES** | |
| 1.1 | INTERPRETABLE MODELS | 23 |
| 1.2 | EXPLANATION BY EXAMPLE | 4 |
| 1.3 | +DEEP EXPLANATION | |
| | ++EXPLAIN INDIVIDUAL PREDICTIONS | |
| 1.3.1 | FORWARD PROPAGATION - LOCAL EXPLANATION | 4 |
| | ++UNDERSTAND WHAT THE MODEL HAS LEARNED | |
| 1.3.2 | DECOMPOSITION APPROACHES | 0 |
| 1.3.3 | +++BACKPROPAGATION-BASED APPROACHES | |
| 1.3.3.1 | *GRADIENTS / DECONVOLUTION / GUIDED BACKPROP* | 23 |
| 1.3.3.2 | *RELEVANCE PROPAGATION* | 2 |
| 1.3.3.3 | *INTEGRATED GRADIENTS* | 0 |
| 1.3.3 | OTHER DEEP EXPLANATION APPROACHES | 6 |
| | | |
| **2.** | **BLACK-BOX APPROACHES** | |
| 2.1 | MODEL INDUCTION - LOCAL EXPLANATIONS | 8 |
| 2.1 | MODEL INDUCTION - GLOBAL EXPLANATIONS | 7 |
| | | |
| **3.** | **INTERACTION APPROACHES** | |
| 3.1 | HCI | 1 |
| 3.2 | PSYCHOLOGY | 1 |

**Table 17 – Results of the classification from the semiotic point of view**

| Class Number | Class Description | Number of proposed techniques |
|---|---|---|
| **1.** | **SOLVING THE ACCESS SUBPROBLEM** | |
| | | |
| **1.1** | **STRATEGIES TO SOLVE THE PROBLEM** | |
| 1.1.1 | FULL ACCESS TO TM | 64 |
| 1.1.2 | SURROGATE MODEL | 15 |
| 1.1.3 | OTHER STRATEGIES TO SOLVE THE ACCESS SUBPROBLEM | 0 |
| | | |
| **1.2.1** | **SURROGATE MODEL RELAXATION STRATEGIES** | |
| 1.2.1.1 | NATURE OF THE SURROGATE MODEL CHOICE | |
| | A REGULARIZATION OF TM | 2 |
| | AN EXPLAINABLE MODEL | 3 |
| | A NON EXPLAINABLE MODEL | 10 |
| | | |

[17] We used the auxiliary tables of Section 5.1.2.

| | | |
|---|---|---|
| **1.2.1.1** | **ACCESS LEVEL TO SURROGATE MODELS** | |
| | FULL ACCESS | 64 |
| | ACCESS ONLY FOR SIMULATIONS | 0 |
| | NO ACCESS | 15 |
| | | |
| **1.2.1.1** | **SCOPE OF INTERPRETABILITY** | |
| | LOCAL INTERPRETATIONS | 8 |
| | GLOBAL INTERPRETATIONS | 7 |
| | | |
| **2.** | **SOLVING THE MEANING-MAKING SUBPROBLEM** | |
| | | |
| **2.1** | **MODEL COMPONENTS** (Peircean dynamic object) | |
| 2.1.1 | INPUT FEATURES | 41 |
| 2.1.2 | ELEMENTS OF THE DATA STRUCTURE | 7 |
| 2.1.3 | EXAMPLES OF PREVIOUS OUTPUTS | 0 |
| 2.1.4 | EXAMPLES OF THE TRAINING DATASET | 0 |
| | | |
| **2.2** | **INTERPRETATION DOMAIN** (Peircean sign representation) | |
| 2.2.1 | ELEMENTARY SIGNALS FOR SENSORY PERCEPTION | 4 |
| 2.2.2 | COMPOUND SIGNALS | 4 |
| 2.2.3 | ELEMENTS OF LANGUAGES | 16 |
| 2.2.4 | SEMANTICALLY STRUCTURED LANGUAGE SENTENCES | 16 |
| | | |
| **2.3** | **ASSOCIATION RULES** (Peircean reasonable explanatory principle) | |
| 2.3.1 | RULE-BASED ASSOCIATION | 27 |
| 2.3.2 | SYMILARITY- BASED ASSOCIATONS | 6 |
| 2.3.3 | ASSOCIATION BASED IN REDUCING UNCERTAINTIES | 2 |



## 5.3. Analysis

In this section, we compare the results obtained by the classification of the **validation domain** techniques from the **semiotic point of view** with the one obtained from the **traditional point of view**.

### 5.3.1.Traditional point of view

Figure 10 presents an overview of the classification from the traditional point of view.

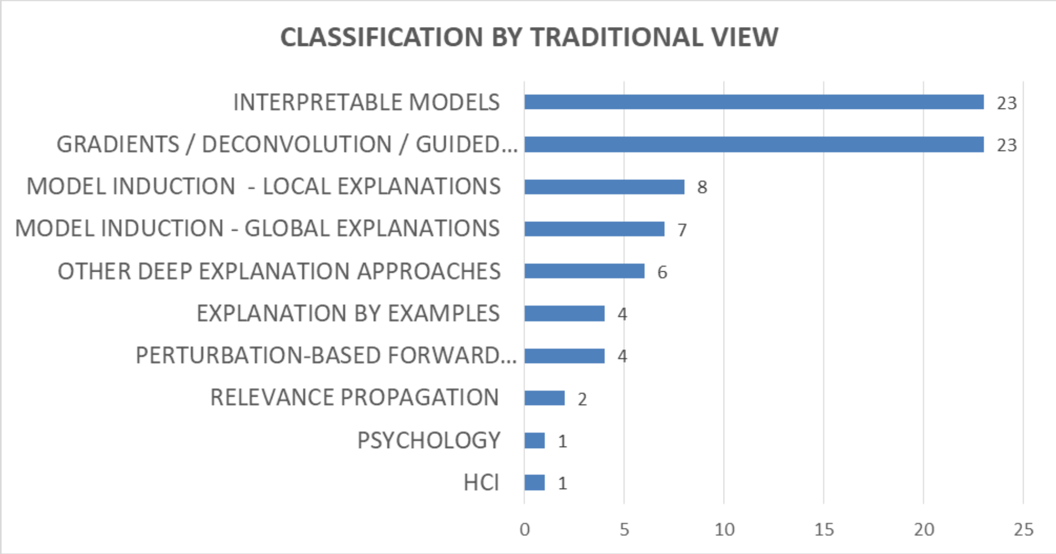


**Figure 10 - Classification from the traditional point of view**

### 5.3.2.Semiotic point of view

Figure 11 to Figure 14 present some overviews of the classification from the semiotic point of view.

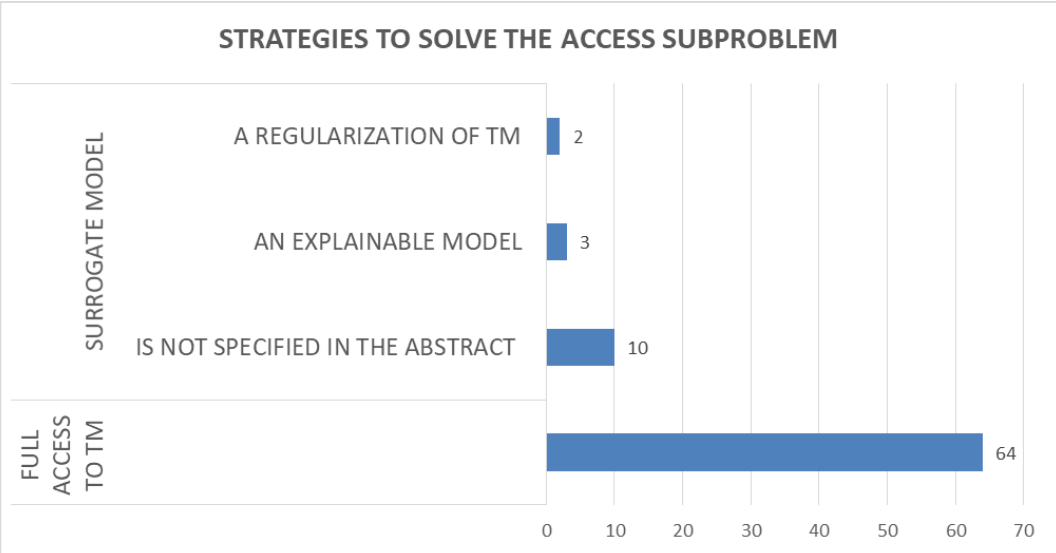


**Figure 11 – Strategies to solve access subproblems**

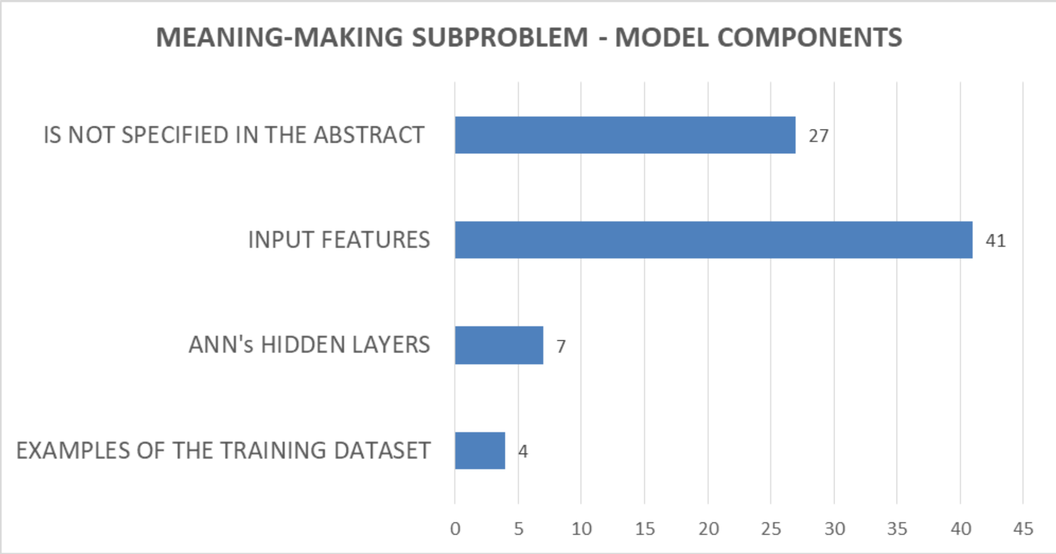


**Figure 12 - Model components for solving meaning-making subproblems**

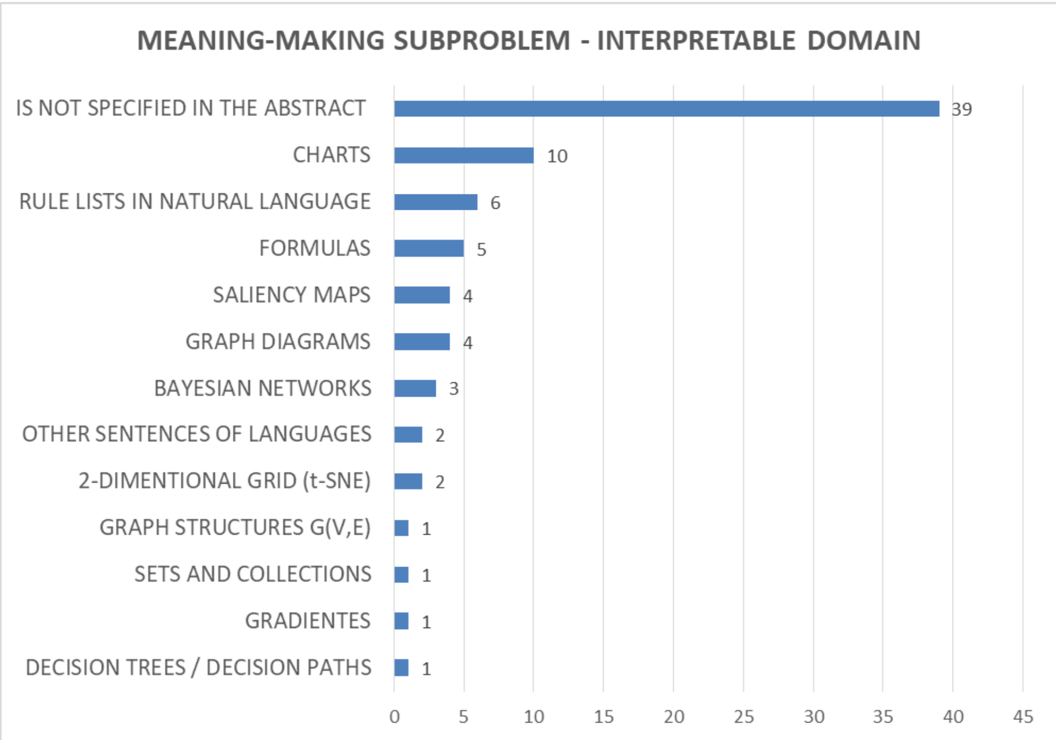


**Figure 13 - Interpretable domains for solving meaning-making subproblems**

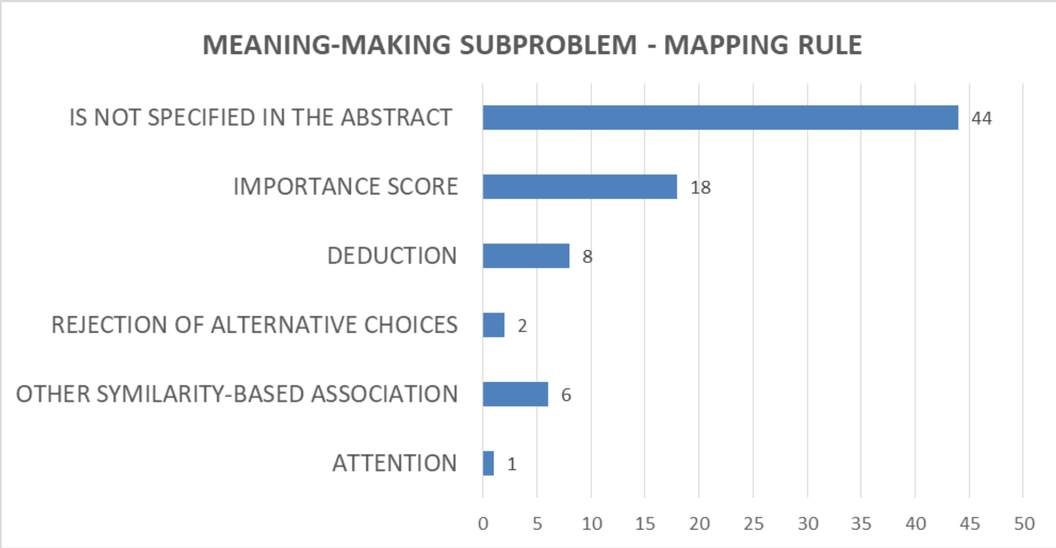


**Figure 14 - Mapping rules for solving meaning-making subproblems**

### 5.3.3. Summary of the research results

In this section, we present a brief summary of the general numbers of the research and make a comparison between the classifications carried out from the traditional point of view and from the semiotic point of view.

**Overall numbers of research**

First, we extracted 1,060 scientific articles indexed by the Google Scholar search engine according to the **systematic search plan** described in **Section 3.2**. After cutting the duplicate records and the records without publication year, we generated a **data collection** with 1,044 scientific papers. We then applied the inclusion and exclusion criteria for filtering 109 papers for the **research domain**, which stand for the coverage of about 24.9% of the papers indexed by Google Scholar. Finally, we selected a sample of 79 papers from the **research domain** to compose the **validation domain,** whose elements were classified using the categories of the taxonomy framework proposed in Section 5.1.

**Analysis of the abstract as a single source of information**

The impact of the abstract analysis on the classification of academic articles had a varied effect. In the classification by the traditional view, we did not find cases in which the abstract did not contain all the information for the classification. However, in the classification by the semiotic vision, the impact is relevant and compromises the quality of the classification due to the representative quantity of articles that could not be classified. This is the case of the classes that identify: (1) the components of the model, with 27 of 79 unidentified articles, (2) the interpretable domain with 39 in 79 unidentified articles, and (3) the mapping rule with 44 in 79 unidentified articles.

By the semiotic view, the identification of the strategies to solve the access subproblems was not impacted, but in the case of the option of the strategy to relax the original problem with surrogate models, in 10 of the 15 articles it was not possible to recognize the nature of the model only by the analysis of the abstract.

Classification using only abstract information is important because it can enable a future systematic mapping study at a lower cost than having to analyze the entire text of about four hundred articles, which is the order of magnitude estimated in section 3.3 .2, of the number of techniques proposed. However, it has not been shown to be effective for classification from a semiotic point of view.



**Comparison between views**

In general, if we consider the comparison of the classification by the two points of view, we did not observe a class where there were significant divergences. The highest similarity in the classifications occurs in the identification of the scope of the strategy to solve the access subproblem, where all 15 selected articles are classified in a similar way by the two views. On the other hand, it was not possible to correlate the classifications of the strategies for deep explanation according to the two views. Regarding this criterion, 35 articles were classified in classes 1.3.x according to the traditional view, while, according to the semiotic view that is less detailed in this point, 64 articles were classified as "total access to TM".

**Usefulness of classifications**

From the point of view of the objective discussed in Section x.x of "to provide software developers, lawmakers and government agencies with a range of design options covering performance-versus-explainability trade space", both classifications present strengths and weaknesses. From the traditional viewpoint, the strong point is the detailed classification of the strategies for deep explanation, in part due to the large number of studies published to interpret the outputs of ANNs. In these articles, the focus is the presentation of the mathematical and computational tools used by the proposed approaches. By the semiotic vision, by the very motivation for its conception, the strong point is the detailing of the form as the meaning is generated for the potential interpreters. Thus, both views, each with a specific focus, contribute to design options for XAI Systems.

# 6
# Conclusion and future work

Research on Explainable AI —a new growing research topic within AI and machine learning— proposes strategies to deal with the trade-off between the accuracy of the state-of-the-art machine learning models and our ability to understand and trust them. These strategies, in turn, are usually implemented by using XAI Systems to interpret the outputs of machine learning models. However, despite the current high growth rate of Explainable IA contributions, the widespread use of XAI Systems in our daily lives is not yet a reality.

This research addressed some open Explainable AI problems, and sheds light on this distortion between the theory behind the researches and the practice of systems used daily. In the theoretical field, we have shown that a clear definition of the "problem of interpretability" is a difficult task, and which is still far from complete. In the practical field, this research focused on finding solutions to the problem of considering the subjective perception of different interpreters in the outputs of the same model.

This final chapter shows how these challenging problems were faced by this research, and, finally, presents some proposals for future work that can contribute to follow up the advances obtained with this work.



## 6.1.Research goals

When research questions are fully answered, the goals of an academic research are achieved. This section presents the analysis of what answers were provided (or are still lacking) to each research question proposed in Section 1.3.

### 6.1.1.State-of-the-art techniques

To answer what are, and what principles underlie, the techniques proposed so far to improve the interpretability of machine learning models (RQ1), we worked on two research sub-questions.

To answer how to search for available techniques to improve the interpretability of machine learning models (RQ1.1), we present in Section 3.2 a

**systematic search plan** that can serve as a basis for a future **systematic mapping study** on these techniques. Section 3.3 describes the validation of the systematic search plan's inclusion and exclusion criteria applying them in a set of scientific articles extracted with the Google Scholar search engine.

To answer what taxonomies are proposed so far to classify techniques to improve the interpretability of machine learning models (RQ1.2), we presented in Section 2.2 a summary of the currently more accepted taxonomy frameworks. As we have not found a sufficiently comprehensive framework to classify the results of a future **systematic mapping study**, we proposed in Section 2.2.2 a synthetic framework by considering the agglutination of some elements of these main taxonomy frameworks. Finally, in Chapter 5 we validate the **semiotic view of interpretability** by classifying some papers extracted with the search terms proposed by the **systematic search plan** of Section 3.2.

### 6.1.2.Subjective perception of interpreters

To answer the question whether it is possible to propose an approach that considers the subjective perception of different interpreters on the outputs of the same model (RQ2), we presented in Section 4.4 a procedure that allows us to address a typical problem of interpreting machine learning models by different interpreters. The procedure is based on **the semiotic view of interpretability** proposed in Sections 4.1, 4.2 and 4.3, which divides a problem of interpretability into three typical subproblems that can be solved separately: the access, the meaning making, and the interaction subproblems.



## 6.2.Challenges of Explainable AI

The research area of the Explainable IA is currently facing numerous challenges, among them, the three presented in Section 1.2, which this academic research tackled. This section suggests some actions to advance the development of theoretical and practical tools to address these and other challenges of XAI.

### 6.2.1.Increasing the range of design options

The challenge of "*providing software developers, legislators, and government agencies with a range of design options covering the performance versus explainability trade space*" was addressed by this research as it **proposes**

**a systematic mapping study of the techniques proposed so far to design XAI Systems**.

Although the **systematic search plan** presented in Section 3.2 proposes a comprehensive interval, it is possible to increase this interval by:

- Including more subjective interpretability-related and more comprehensive keywords words to the search terms.
- Using multiple search engines instead of using only the Google Scholar engine.
- Relaxing the exclusion criterion that cuts across articles that approach fuzzy models because, although these articles deal with another type of model, it may be possible to find in them some elements to inspire new methods to increase models of interpretability machine learning.

In addition, design options mapped by a future **systematic mapping study** can be qualitatively enriched if the study also highlights the mathematical and computational tools used by each mapped technique.

#### 6.2.2.Testing new theoretical frameworks

The challenge of "*providing consulting firms with theoretical frameworks so that they can <u>evaluate projects</u> and <u>recommend strategies</u> to address the problem of interpretability*" was partially addressed by this academic research as it **proposes, by the Semiotic View on Interpretability, a new more comprehensive approach to classify the current available techniques to improve interpretability.** However, to evaluate projects and recommend strategies we also need other **qualitative** elements.

To advance the development of a qualitative approach, it is necessary to evaluate the scientific contributions selected by a future **systematic mapping study** as to their **applicability**, **efficiency** and **cost**. In short, a **guiding framework** for technical recommendation to improve the interpretability of machine learning models should also address the implications of using these techniques from the point of view of some areas of computer science such as:

- The construction and implementation of more efficient **algorithms** for advanced applications.
- The development and application of algorithmic methods for **handling and analyzing large volumes of data**.

- The application and analysis of highly complex techniques and **solutions in software engineering**.
- The development of more efficient **user interfaces** which provide better **human-computer interaction.**

### 6.2.3.Towards a broader definition

This research addresses the challenge of "*guiding future research on issues related to a <u>broader and useful</u> definition of the problem of interpretability*" as it presents in Section 2.2.3 a gap analysis of the taxonomy frameworks proposed so far. In this way, the **semiotic view on interpretability** presented in Section 4.1, 4.2 e 4.3 and the **procedure to solve typical problems of interpretability** presented in Section 4.4, address some elements of what a broader taxonomy must considers. In particular they consider: (1) the subjective relation between "what", "how", and "who" needs to interpret the model; and (2) the role of non-humans in the process of interpreting models.

However, despite the scientific community's effort to develop new techniques to interpret machine learning models, it is fair to expect that the available solutions today are not sufficiently effective to interpret the increasing more complex models in the future. As the complexity of machine learning algorithms and the ubiquity of applications increases; as actions need to be explained to more and more people with different **perceptions**; and as the dependency between "what to explain" and "to whom explain" becomes increasingly more personal, perhaps the "hard" problem of interpretability cannot be directly solved but only recursively bypassed with the help of other trustworthy entities, *i.e.*, entities that, in turn, could need to trust on another entities, who could trust on other entities and so on.

## 6.3.Future Work: Designing Interpretation Support Systems

The in-depth discussions on strategies to interpret machine learning models carried out by this academic research have brought to light some challenges of using XAI Systems in day-to-day applications. Two of these challenges were addressed in this work: (1) the subjective perception of different interpreters on the same model, and (2) the non-human interpreters in the process of interpreting model's outputs.

The **semiotic view on interpretability** proposed in chapter 4 opens a new front of opportunities to develop XAI Systems that face both challenges above.

However, unlike the current approach of XAI Systems to interpret directly the models, we need to work on how to support human being in the task of interpreting models under their personal point of view by considering their preferences and uncertainties, as usually do Decision Support Systems and Recommender Systems.

In short, developing systems that are more adaptable to different users' preferences and uncertainties, and consider non-human in the loop has the potential to extend the scope of XAI Systems to become sufficiently comprehensive and pragmatic to be used by us in daily tasks. This section proposes a set of future work as actions of a strategy to develop comprehensive and pragmatic systems to support human interpretation, which we call here **Interpretation Support Systems**.

### 6.3.1.Core procedure

Section 4.4 proposes a **procedure to deal with the problem of interpretability** that roughly consists of sequentially solving each of the subproblems that characterize this kind of problem: the access, the meaning making, and the interaction subproblems.

If we want to build **Interpretation Support Systems** based on this procedure, the sequential operation of these systems would be:

1. To solve the **access subproblem** by choosing, and linking the original problem with a relaxed problem.
2. To solve the **meaning-making subproblem** of the chosen relaxed problem.
3. To solve the interaction subproblem.

### 6.3.2.Choosing the relaxed problem

The idea of working on solving a relaxed problem instead of working on solving the original problem is usually employed in optimization problems. Therefore, it is fair to expect that future work that proposes solutions to solve the access subproblem may be inspired by classic optimization strategies. For example, a procedure that systematically finds lower and upper bounds could be used to propose **meaning-making subproblems** to be solved until it reaches a sufficiently narrow range of certainty.

In short, future work that seeks to solve **access subproblems** should propose formalizations for the problem, as well as algorithms that solve them efficiently.

### 6.3.3.Learning personal semiotic patterns

When solving **meaning-making subproblems** by using the procedure suggested in Section 4.4, the next task after defining the target interpreter is to **learn his, her, or its personal semiotic pattern**.

A possible strategy to learn these patterns is to use **Markov Logical Networks** (MLN) for the task. According to Domingos (2015), these networks have the advantage of simultaneously capturing the logical essence of the three **semiotic-based learning algorithm** —presented in Section 4.1.3— with a single data structure. In this sense, MLN-based learners are a super generalization of the three semiotic-based learners, since the data structure of MLNs can converge to the data structure of each of the semiotic-based learners depending on their parameterization. According to Domingos (2015), the algorithm able to learn with an “MLN data structure” would be a kind of "master algorithm" for learners.

MLNs seem to be appropriate to learn **personal semiotic patterns,** as they are capable of simultaneously capturing from the dataset: (1) Bayesian causalities; (2) cluster analogies and; (3) rule-based knowledge. Thus, we propose a future research work, which the main goal is to develop models that learn **personal semiotic patterns** using MLNs.

In short, the research should mainly propose and execute a set of assessing tests for human interpreters capable of generating a sufficiently large dataset so that a MLN could learn some **personal semiotic patterns** of these interpreters.

### 6.3.4.Solving the meaning-making subproblem

The **meaning-making subproblem** is also a typical **optimization problem**, as it seeks for the **reasonable explanatory principle** that **best fits** the interpreter's semiotic pattern. Again, future work, which proposes solutions to solve meaning-making subproblems, could be inspired by the strategies to solve optimization problems proposed so far.

Future work, which seeks to solve meaning-making subproblems, should formulate and propose an algorithm that solves efficiently the optimization problem.

### 6.3.5.Expanding XAI Systems with chains of trustworthy entities

We proposed in Section 4.3.2 that some reports of human or **non-human trustworthy entities** could be added to the list of possible human interpretable domains. According to the proposal, these trustworthy entities could be human experts, such as professionals, or non-human “interpreters”, such as other trusted XAI Systems. This strategy has the potential to expand the reach of Interpretation Support Systems to limits far beyond the current XAI Systems, since it would be possible to develop them based on "**chains of trustworthy interpreters**", which would leverages and be able to explain very complex and focused outputs even to non-expert interpreters.

A possible path to expand **Interpretation Support Systems** with chains of trustworthy interpreters is the use of “smart contracts”. According to Christidis and Devetsikiotis (2016), blockchain technology enables applications that could previously run only through a trusted intermediary to operate in a decentralized fashion (…) with the same amount of certainty. Smart contracts —self-executing scripts that reside on the blockchain— integrate all the blockchain’s fundamentals that enable trustless networks and allow for proper, distributed, heavily automated workflows. The idea is that **personal semiotic patterns** can be incorporated into blockchains in the form of smart contracts so that XAI Systems can access them selectively.

In short, future work in this field should mainly propose and apply strategies to build chains of trustworthy interpreters, which could efficiently be integrated, to Interpretation Support Systems.

## References

Bruno, D.; Richmond, H. (2013). **The truth about taxonomies**. Information Management Journal, Lenexa, v. 37, n. 2, p. 44-53.

Burrell, J. (2016). **How the machine 'thinks**': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 2053951715622512.

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., ... & Kelley, T. D. (2017). **Interpretability of deep learning models**: a survey of results. In IEEE Smart World Congress 2017 Workshop: DAIS.

Chalmers, D. (1995). **Facing up to the problem of consciousness**. Journal of Consciousness Studies. Retrieved from http://www.ingentaconnect.com/content/imp/jcs/1995/00000002/00000003/653

Christidis, K., & Devetsikiotis, M. (2016). **Blockchains and smart contracts for the internet of things**. Ieee Access, 4, 2292-2303.

De Souza, C. S., de Gusmão Cerqueira, R. F., Afonso, L. M., de Mello Brandão, R. R., & Ferreira, J. S. J. (2016). **Software developers as users**: semiotic investigations in human-centered software development. Springer.

Domingos, P. (2015). **The master algorithm:** How the quest for the ultimate learning machine will remake our world. Basic Books.

Doshi-Velez, F., & Kim, B. (2017). **Towards A Rigorous Science of Interpretable Machine Learning**. Retrieved from https://arxiv.org/pdf/1702.08608.pdf

Dhurandhar, A., Iyengar, V., Luss, R., & Shanmugam, K. (2017). **A Formal Framework to Characterize Interpretability of Procedures**. arXiv preprint arXiv:1707.03886.

Escalante, H. J., Guyon, I., Escalera, S., Jacques, J., Madadi, M., Baró, X., ... & Van Gerven, M. A. (2017, May). **Design of an explainable machine learning challenge for video interviews.** In IJCNN 2017-30th International Joint Conference on Neural Networks (pp. 1-8).

Goodman, B., & Flaxman, S. (2016). **European Union regulations on algorithmic decision-making and a right to explanation.** In ICML Workshop on Human Interpretability in Machine Learning (WHI 2016). New York, NY. Retrieved from https://arxiv.org/pdf/1606.08813.pdf

Gunning, D. (2017a). Explainable Artificial Intelligence (Page). Retrieved October 29, 2017, from https://www.darpa.mil/program/explainable-artificial-intelligence

Gunning, D. (2017b). Explainable Artificial Intelligence (XAI). In Defense Advanced Research Projects Agency (DARPA). Retrieved from https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning) IJCAI-16 DLAI WS.pdf

Guo, W., Zhang, K., Lin, L., Huang, S., & Xing, X. (2017). **Towards**

**interrogating discriminative machine learning models**. arXiv preprint arXiv:1705.08564.

Harzing, A. W. (2007). Publish or Perish. Retrieved from http://www.harzing.com/pop.htm

Kitchenham, B., & Charters, S. (2007). **Guidelines for performing Systematic Literature Reviews in Software Engineering. Engineering** (Vol. 2).

Lipton, Z. (2016). **The Mythos of Model Interpretability**. In ICML Workshop on Human Interpretability in Machine Learning (WHI 2016). New York, NY.

Lundberg, S. M., & Lee, S. I. (2017). **A unified approach to interpreting model predictions.** In Advances in Neural Information Processing Systems (pp. 4765-4774).

Montavon, G., Samek, W., & Müller, K.-R. (2017). **Methods for Interpreting and Understanding Deep Neural Networks**. Retrieved from https://arxiv.org/pdf/1706.07979.pdf

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). **Explaining nonlinear classification decisions with deep taylor decomposition**. Pattern Recognition, 65, 211-222.

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). **How do Humans Understand Explanations from Machine Learning Systems?** An Evaluation of the Human-Interpretability of Explanation. arXiv preprint arXiv:1802.00682.

O’Neil, C. (2017). **Weapons of Math Destruction:** How Big Data Increases Inequality and Threatens Democracy. Broadway Books.

Offert, F. (2017). **"I know it when I see it".** Visualization and Intuitive Interpretability. arXiv preprint arXiv:1711.08042.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). **Feature visualization.** Distill, 2(11), e7.

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). **Guidelines for conducting systematic mapping studies in software engineering**: An update. Information and Software Technology, 64, 1–18. http://doi.org/10.1016/J.INFSOF.2015.03.007

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). **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). ACM.

Samek, W., Wiegand, T., & Müller, K. R. (2017). **Explainable artificial intelligence:** Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

Santanella, L. (2002). Semiótica Alicada. São Paulo: Pioneira Thomsom Learning.

Shrikumar, A., Greenside, P., & Shcherbina, A. Y. (2017). **Not Just A Black Box:** Learning Important Features Through Propagating Activation Differences (OLD VERSION). In ICML 2017. Retrieved from https://arxiv.org/pdf/1605.01713.pdf

Weller, A. (2017). **Challenges for transparency**. arXiv preprint arXiv:1708.01870.

WHI / ICML. (2017). Workshop on Human Interpretability in Machine Learning (WHI). Retrieved October 31, 2017, from

https://2017.icml.cc/Conferences/2017/Schedule?showEvent=20

# Appendix I – Basic Concepts of Machine Learning and Semiotics

This appendix presents the basic notions of **machine learning** and **semiotics** needed to understand the discussions of this dissertation. The purpose here is not to present the basic definitions with formal rigor, but leave no doubt about them when they are mentioned in the text.

## Basic notions of machine learning

In this research, we consider **machine learning** as the field of science that studies the development and application of a specific class of model, which is usually named by the same name of the field of studies, that is, machine learning models.

### Statistical models vs**. machine learning models**

**Models** are representations used to help people to know, understand, or simulate aspects of the real world that the model represents, whether these are **empirical objects**[18] or **factual relationships**.

However, in order to model objects and relationships whose complexity or intangibility precludes the representation of all its characteristics, it is necessary to formulate simplifying hypotheses, or as they are usually called, reductionist hypotheses. A reduction consists of the identification and choice of subsets of the total of all the characteristics of the real aspect to be modeled, whose representation is both operationally feasible and able to achieve the modeling objectives.

**Conceptual models** are composed of reductionist representations, known as "concepts," which, while not representing all the characteristics of the real-world aspects to be modeled, are capable of helping people to know, to understand or to simulate these aspects.

[18] We define e**mpirical objects** as those obtained from observations of the world.

The process of **cognitive inference** to explain the concepts of a conceptual model is called conceptualization or **generalization** and is often guided by **induction** or **deduction** logics.

**Mathematical models** are conceptual models, where concepts are represented by mathematical structures. So-called **statistical models** can refer to two types of models: They refer to mathematical models that consider **random variables** - and their probabilistic distributions - as part of their structure, but also refer to models, whose real-world aspects to be modeled are empirical objects. In this second case, statistical models are used as synonymous with the class of models known as **data-driven models**.

**Statistical models** vs**. machine learning models**

Like statistical models, **machine learning models** are mathematical models and data-driven models, but their generalization processes are driven only by induction inferences, while the generalization processes of statistical models are generally composed of deduction-type inferences. In general, statistical models (in the sense of data-driven) are composed of formal representations of mathematical language supported by theorems, while machine learning models do not necessarily observe this requirement, although some of them are generalized by computational algorithms with guarantees of convergence and optimality.

In addition, although statistical models and machine learning models are both data-driven models, statistical models often seek to identify and quantify the **correlations** between empirical object variables, while machine learning models can go beyond identification and quantification to infer more complex patterns, such as **cause and effect** relationships between them.

**Elements of machine learning models**

In this dissertation, we use the nomenclature below to refer to the main elements of **machine learning models**:

Suppose the problem of developing a **computational model** capable of explaining or predicting the **behavior** of a system based on **patterns** observed in samples of variables collected from the previously observable states of that system. Thus, we identify the following elements:

**Output features** - A particular subset, among all possible subsets from the universe of observable variables of the system behavior, chosen to characterize the conceptual behavior of the system in the applications of the model.

**Input features** - A particular subset, among all possible subsets from the universe of observable variables, chosen because of their capability of affecting the conceptual behavior of the system.

**Example** - A sample, measured from the previous system behavior, where the elements are the set of values of each **input feature**, and the set of values of its correspondent output feature, if this latter is available.

**Transfer function** - A parameterized **data structure** that governs how the model outputs, that represent the model behavior, relate to the model inputs.

**Developing machine learning models**

The process of developing the machine learning models encompasses the following actions and components.

**Model training** (or **model learning**) - It is the conceptualization process of determining the values of the data structure's parameters, so that the **outputs** of the model represent, in the best possible way according to the criteria chosen for performance evaluation, the behavior of the system.

**Training data set** - It is the set of inputs available for the development of the model, which may or may not have the corresponding output information of each input. The training dataset is often spliced in two partitions: the **training data** and the **validation data**.

**Training algorithm** - It is the **optimization algorithm** used to find the best values for the data structure's parameters.

**Trained model** - It is the transfer function represented by the data structure configured with the optimal parameters obtained in the model-training step.

**Accuracy** - Is a set of drivers (quantitative or not) used as criteria to evaluate the performance of the trained model.

**Training data** - It is the **training dataset partition** whose elements are used in the model-training step.

**Validation data** - It is the **training dataset partition** whose elements are used to evaluate the final performance of the trained model.

**Overfitting** - Occurs when the performance of the trained model, often measured by accuracy indicators, is representatively higher when the model is subjected to the **training data** inputs than when subjected to the **validation data** inputs.

**Training strategy** (or **learning strategy**) – It is the set of actions that encompass the choice of the **training algorithm** and the model **data structure**.

**Model task –** is the purpose of the trained model for the users. Generally, the task performing by the model is critical for the choosing of the learning strategy.

**Tasks and learning strategies**

The typical classes of **tasks** assigned to the **machine learning models**, and their associated **learning strategies** are following described:

**Supervised learning** - Supervised learning algorithms help to develop the model from a set of data that contains both the inputs and the observed outputs. The examples of tasks that machine learning models execute when supervised learning algorithms train them are:

- **Classification** (of discrete variables) is used when the outputs are restricted to a limited set of values, such as binary or multiclass classifiers, and discriminative or generative classifiers.
- **Regression** (of continuous variables) is used when the outputs may have any numerical value within a range, such as general data regressors and time-series regressors.

**Unsupervised learning** - Unsupervised learning algorithms take a set of data that contains only inputs and find structure in the data. The examples of tasks that machine learning models execute when trained by unsupervised learning algorithms there are:

- **Clustering** (of discrete variables), i.e., the assignment of a set of observations into subsets (so-called clusters) are considered within the same cluster according to one or more predesignated criteria, such as non-hierarchical clustering and hierarchical clustering.
- **Dimensionality reduction**, i.e., the task of reducing the number of features to simplify inputs by mapping them into lower-dimensional space (such as factor analysis, feature learning -aka feature extraction-, and data transformation).

**Semi-supervised learning** – Semi-supervised learning algorithms build a model from a set of data of which one bi-partition contains both the inputs and the observed outputs, and another bi-partition contains only the inputs. The example of tasks that machine learning models execute when trained by semi-supervised learning algorithms there are:

- **Density estimation** finds the distribution of inputs in some space;

- **Low-density separation**;
- **Graph-based models**;

**Reinforcement learning** - Reinforcement learning is an area of machine learning concerned with how **software agents** ought to take actions in an environment to maximize some notion of **cumulative reward**. An example of task that machine learning models execute when trained by reinforcement learning algorithms there is:

- **Action Support,** such as Monte Carlos Methods and temporal difference methods.

**Taxonomy of learning algorithms**

In this research, we used the taxonomy of Domingos (2015) to classify the five main paradigms or computational strategies for training machine learning models. They are:

The **Symbolist paradigm**, in which learning is achieved through processes of manipulation of symbols (and consequently, languages) for inverse deduction;

- Among these processes, we can mention the **Association Rule Learning**" **algorithms**, such as "apriori" algorithm; eclat algorithm; FP-growth algorithm (frequent pattern); the D**ecision Tree algorithms**, such as "classification and regression tree" (cart); "Iterative dichotomiser" (id3); c4.5 and c5.0; chi-squared automatic interaction detection (chaid); decision stump; m5 and conditional decision trees.

The **Bayesian paradigm**, in which learning is achieved through processes of probabilistic inference that promote the systematic reduction of uncertainties;

- Among them, we can mention the **Bayesian algorithms**, such as naive Bayes; Gaussian naive Bayes; multinomial naive Bayes; averaged one-dependence estimators (AODE) and **Bayesian networks**, such as Bayesian network (BN); Bayesian belief network (BBN).

The **Analogizer paradigm**, in which learning is achieved by analogy reasoning, based on the processes of searching for similarities;

- Among them, we can mention the **Clustering algorithms**, such as k-means; k-medians; expectation maximization (EM); hierarchical clustering; the **Instance-based algorithms**, such as k-nearest neighbor (KNN); learning vector quantization (lvq); self-organizing map (SOM); locally weighted learning (LWL); and the **Kernel algorithms**, such as support vector machine (SVM) and restricted Boltzmann machine (RBM).

The **Connectionist paradigm**, with their models based on neural networks, which try to simulate the learning process of biological neocortex; and);

- Among them we can mention: The artificial neural network (ANN) algorithms, such as "perceptron"; multiclass perceptron; back-propagation; hopfield network; radial basis function network (RBFN); deep learning algorithms; deep Boltzmann machine (DBM); deep belief networks (DBN); convolutional neural network (CNN); recurrent neural network (RNN); long-short-term memory networks (LSTM); generative adversarial networks (GAN); stacked auto-encoders.

The **Evolutionist paradigm**, which tries to simulate the learning process of the biological evolution;

- Among them we can mention the class of Genetic algorithms;



## Basic notions of semiotics

The study of interpretability of machine learning models encompasses the understanding of two key concepts: **interpretation** and **explanation**. On the other hand, semiotics is a multifaceted discipline where **signs** and **signification** are common objects of study in all cases. Moreover, the semiotic theories deal with the process of human perception, in which **signs** have the key role of carrying meaning. Therefore, any study on **model interpretability** that aims to be comprehensive must consider the semiotic approach as one of its visions.

To present the basics of **semiotics,** we extracted the concepts from the discussions of the introductory chapter of De Souza et al.'s book "*Software Developers as Users. Semiotic Investigations in Human-Centric Software Development*" (De Souza et al., 2016).

**Association between signs and meaning**

According to semiotics, **signs** are the result of associations between **expressions** (aka representations) and **content** (aka information); and **signification** is broadly defined as the process by which signs come into existence. Depending on the semiotic theory, it postulates that such **expression-content associations** are carried out by:

(1) A mind (individual or collective, human or nonhuman);

(2) An abstract, systemic, or logic nature;

(3) The result of evolutionary sociocultural processes;

**Essential elements of semiotics**

According to Santaella (2002), the **Peircean speculative grammar** addresses the **interpreter** as part of the study of all kinds of **signs** and **forms of thinking** that they enable, and the formal elements involved in the meaning making of the explanations.

**Sign** - Anything that, for somebody, under some circumstance (s) and in some respect (s) stands for something else. The three constituent parts of a sign are:

- **Representamen** (a representation itself),
- **Object** (what the representation stands for), and
- **Interpretant** (the mediating That Creates a meaningful interpretation association between the other two components).

**Interpretation** - Signs only comes into existence if some mind mediates (and thus creates) the association between representation and what representation stands for. The mediation is an interpretation.

**Abduction** - An inference (what-if) process that produces a reasonable explanatory principle capable of turning some surprising fact into the logical consequence of this principle.

**Circumstantially Verifiable Hypothesis** (aka **explanatory hypothesis**) - It is the hypothesis that is signified and confirmed in the collection of signs that are contextually associated with the surprising fact that triggered the abductive process in the reasoner's mind.

**Semiosis** is the unlimited sense-making abductive process where all conclusions are provisional as they hold until new facts contradict them.

## Appendix II – Public or Perish query reports

### Query report 01 – “interpretability” AND “explainability”

# *(intitle:interpretability OR intitle:explainability) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") to 2017*

*Publish or Perish 6.21.6145.6594*

## *Search terms*

***All of the words:*** *(intitle:interpretability OR intitle:explainability) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning")*
***Years:*** *earliest to 2017*

## *Data retrieval*

***Data source:*** *Google Scholar*
***Query date:*** *21/01/2018 12:08:11*
***Cache date:*** *21/01/2018 12:08:30*
***Query result:*** *[0] The operation completed successfully.*

## *Metrics*

***Publication years:*** *1997-2017*
***Citation years:*** *21 (1997-2018)*
***Papers:*** *135*
***Citations:*** *3473*
***Citations/year:*** *165.38*
***Citations/paper:*** *25.73 (*count=12)*
***Citations/author:*** *2297.02*
***Papers/author:*** *73.21*
***Authors/paper:*** *2.44/2.0/2 (mean/median/mode)*
***Age-weighed citation rate:*** *470.29 (sqrt=21.69), 283.65/author*
***Hirsch h-index:*** *24 (a=6.03, m=1.14, 2949 cites=84.9% coverage)*
***Egghe g-index:*** *58 (g/h=2.42, 3385 cites=97.5% coverage)*
***PoP hI,norm:*** *19*

***PoP hI,annual:*** *0.90*

## *Results*

*ZC Lipton (2016)* ***The mythos of model interpretability****. arXiv preprint arXiv:1606.03490, arxiv.org, cited by 100 (50.00* per year)*

*RP Paiva, A Dourado (2004)* ***Interpretability and learning in neuro-fuzzy systems****. Fuzzy sets and systems, Elsevier, cited by 136 (9.71 per year)*

*Y Jin (2000)* ***Fuzzy modeling of high-dimensional systems: complexity reduction and interpretability improvement****. IEEE Transactions on Fuzzy Systems, ieeexplore.ieee.org, cited by 487 (27.06* per year)*

*H Ishibuchi, T Yamamoto (2003)* ***Interpretability issues in fuzzy genetics-based machine learning for linguistic modelling****. Modelling with Words, Springer, cited by 30 (2.00 per year)*

*MT Ribeiro, S Singh, C Guestrin (2016)* ***Model-agnostic interpretability of machine learning****. arXiv preprint arXiv:1606.05386, arxiv.org, cited by 14 (7.00 per year)*

*AC Haury, P Gestraud, JP Vert (2011)* ***The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures****. PloS one, journals.plos.org, cited by 193 (27.57* per year)*

*U Bodenhofer, P Bauer (2003)* ***A formal model of interpretability of linguistic variables****. Interpretability issues in fuzzy modeling, Springer, cited by 45 (3.00 per year)*

*U Bodenhofer, P Bauer (2000)* ***Towards an axiomatic treatment of "interpretability"****. Proc. IIZUKA2000, academia.edu, cited by 28 (1.56 per year)*

*I Bratko (1997)* ***Machine learning: Between accuracy and interpretability****. Learning, networks and statistics, Springer, cited by 20 (0.95 per year)*

*C Mencar, G Castellano, ... (2005)* ***Some Fundamental Interpretability Issues in Fuzzy Modeling.****. EUSFLAT Conf …, ai2-s2-pdfs.s3.amazonaws.com, cited by 28 (2.15 per year)*

*SM Zhou, JQ Gan (2008)* ***Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling****. Fuzzy Sets and Systems, Elsevier, cited by 204 (20.40* per year)*

*JM Alonso, L Magdalena (2011)* ***HILK++: an interpretability-guided fuzzy modeling methodology for learning readable and comprehensible fuzzy rule-based classifiers****. Soft Computing, Springer, cited by 55 (7.86 per year)*

*MB Gorzałczany, F Rudziński (2016)* ***A multi-objective genetic optimization for fast, fuzzy rule-based credit classification with balanced accuracy and interpretability****. Applied Soft Computing, Elsevier, cited by 24 (12.00* per year)*

*J González, I Rojas, H Pomares, LJ Herrera, ... (2007)* ***Improving the accuracy while preserving the interpretability of fuzzy function approximators by means of multi-objective evolutionary algorithms****. International Journal of …, Elsevier, cited by 43 (3.91 per year)*

*MB Gorzałczany, F Rudziński (2012)* ***Accuracy vs. interpretability of fuzzy rule-***

*based classifiers: an evolutionary approach*. *Swarm and Evolutionary Computation, Springer, cited by 25 (4.17 per year)*

*PK Shukla, SP Tripathi (2011)* ***A survey on interpretability-accuracy (IA) trade-off in evolutionary fuzzy systems****. Genetic and Evolutionary Computing ..., ieeexplore.ieee.org, cited by 16 (2.29 per year)*

*F Maes, R Fonteneau, L Wehenkel, D Ernst (2012)* ***Policy search in a space of simple closed-form formulas: Towards interpretability of reinforcement learning****. International Conference on ..., Springer, cited by 11 (1.83 per year)*

*E Lughofer (2013)* ***On-line assurance of interpretability criteria in evolving fuzzy systems–achievements, new concepts and open issues****. Information Sciences, Elsevier, cited by 68 (13.60* per year)*

*MJ Gacto, R Alcalá, F Herrera (2011)* ***Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures****. Information Sciences, Elsevier, cited by 285 (40.71* per year)*

*C Mencar, G Castellano, AM Fanelli (2007)* ***On the role of interpretability in fuzzy data mining****. International Journal of ..., World Scientific, cited by 19 (1.73 per year)*

*J Zurada (2010)* ***Could decision trees improve the classification accuracy and interpretability of loan granting decisions?****. System Sciences (HICSS), 2010 43rd Hawaii ..., ieeexplore.ieee.org, cited by 24 (3.00 per year)*

*CF Juang, CY Chen (2013)* ***Data-driven interval type-2 neural fuzzy system with high learning accuracy and improved model interpretability****. IEEE Transactions on Cybernetics, ieeexplore.ieee.org, cited by 41 (8.20 per year)*

*JM Alonso, L Magdalena (2009)* ***An Experimental Study on the Interpretability of Fuzzy Systems.****. IFSA/EUSFLAT Conf., pdfs.semanticscholar.org, cited by 12 (1.33 per year)*

*H Ishibuchi, Y Nojima (2009)* ***Discussions on Interpretability of Fuzzy Systems using Simple Examples.****. IFSA/EUSFLAT Conf., pdfs.semanticscholar.org, cited by 21 (2.33 per year)*

*H Liu, M Cocea, A Gegov (2016)* ***Interpretability of computational models for sentiment analysis****. Sentiment Analysis and Ontology Engineering, Springer, cited by 9 (4.50 per year)*

*X Zhu, J Li, D Wu, H Wang, C Liang (2013)* ***Balancing accuracy, complexity and interpretability in consumer credit decision making: A C-TOPSIS classification approach****. Knowledge-Based Systems, Elsevier, cited by 21 (4.20 per year)*

*J Søgaard, S Forchhammer, ... (2015)* ***Video quality assessment and machine learning: Performance and interpretability****. Quality of Multimedia ..., ieeexplore.ieee.org, cited by 3 (1.00 per year)*

*A Riid, E Rüstern (2014)* ***Adaptability, interpretability and rule weights in fuzzy rule-based systems****. Information Sciences, Elsevier, cited by 18 (4.50 per year)*

*TA Plate (1999)* ***Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using Gaussian process models****. Behaviormetrika, jstage.jst.go.jp, cited by 34 (1.79 per year)*

*AF Gómez-Skarmeta, F Jiménez, ... (2007)* ***Improving interpretability in***

***approximative fuzzy models via multiobjective evolutionary algorithms**. International Journal of …, Wiley Online Library, cited by 24 (2.18 per year)*

*AG Di Nuovo, V Catania (2009) **Linguistic modifiers to improve the accuracy-interpretability trade-off in multi-objective genetic design of fuzzy rule based classifier systems**. Intelligent Systems Design and …, ieeexplore.ieee.org, cited by 9 (1.00 per year)*

*PK Shukla, SP Tripathi (2013) **Interpretability issues in evolutionary multi-objective fuzzy knowledge base systems**. ... of Seventh International Conference on Bio …, Springer, cited by 8 (1.60 per year)*

*S Tan, KC Sim, M Gales (2015) **Improving the interpretability of deep neural networks with stimulated learning**. Automatic Speech Recognition and …, ieeexplore.ieee.org, cited by 15 (5.00 per year)*

*B Abdollahi, O Nasraoui (2017) **Using explainability for constrained matrix factorization**. Proceedings of the Eleventh ACM Conference …, dl.acm.org, cited by 3 (3.00 per year)*

*F Poursabzi-Sangdeh, DG Goldstein, ... (2017) **Manipulating and measuring model interpretability**. ... Machine Learning …, csel.cs.colorado.edu, cited by 2 (2.00 per year)*

*B Kim, R Khanna, OO Koyejo (2016) **Examples are not enough, learn to criticize! criticism for interpretability**. Advances in Neural Information …, papers.nips.cc, cited by 15 (7.50 per year)*

*JM Alonso, L Magdalena, ... (2009) **Looking for a good fuzzy system interpretability index: An experimental approach**. International Journal of …, Elsevier, cited by 116 (12.89* per year)*

*MJ Gacto, R Alcalá, F Herrera (2011) **A double axis classification of interpretability measures for linguistic fuzzy rule-based systems**. International Workshop on Fuzzy Logic and …, Springer, cited by 6 (0.86 per year)*

*R Florez-Lopez, JM Ramon-Jeronimo (2015) **Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal**. Expert Systems with Applications, Elsevier, cited by 16 (5.33 per year)*

*A Ghandar, Z Michalewicz, R Zurbruegg (2012) **Enhancing profitability through interpretability in algorithmic trading with a multiobjective evolutionary fuzzy system**. Parallel Problem Solving from …, Springer, cited by 5 (0.83 per year)*

*C Strobl, T Augustin (2009) **Adaptive Selection of Extra Cutpoints—Towards Reconciling Robustness and Interpretability in Classification Trees**. Journal of Statistical Theory and Practice, Taylor & Francis, cited by 6 (0.67 per year)*

*JM Alonso, S Guillaume, L Magdalena (2006) **A hierarchical fuzzy system for assessing interpretability of linguistic knowledge bases in classification problems**. Proceedings of IPMU, math.s.chiba-u.ac.jp, cited by 13 (1.08 per year)*

*AS Ross, F Doshi-Velez (2017) **Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients**. arXiv preprint arXiv:1711.09404, arxiv.org, cited by 2 (2.00 per year)*

*V Krakovna, F Doshi-Velez (2016) **Increasing the interpretability of recurrent***

***neural networks using hidden Markov models***. *arXiv preprint arXiv:1606.05320, arxiv.org, cited by 6 (3.00 per year)*

*JV Ramos, A Dourado (2006)* ***Pruning for interpretability of large spanned eTS***. *Evolving Fuzzy Systems, 2006 …, ieeexplore.ieee.org, cited by 6 (0.50 per year)*

*C Mencar, AM Fanelli (2008)* ***Interpretability constraints for fuzzy information granulation***. *Information Sciences, Elsevier, cited by 136 (13.60* per year)*

*GJ Katuwal, R Chen (2016)* ***Machine Learning Model Interpretability for Precision Medicine***. *arXiv preprint arXiv:1610.09045, arxiv.org, cited by 2 (1.00 per year)*

*O Bastani, C Kim, H Bastani (2017)* ***Interpretability via Model Extraction***. *arXiv preprint arXiv:1706.09773, arxiv.org, cited by 1 (1.00 per year)*

*A Hutton, A Liu, CE Martin (2012)* ***Crowdsourcing Evaluations of Classifier Interpretability.***. *AAAI Spring Symposium: Wisdom of the Crowd, aaai.org, cited by 4 (0.67 per year)*

*D Partridge, V Schetinin, D Li, TJ Coats, ... (2006)* ***Interpretability of Bayesian decision trees induced from trauma data***. *Lecture notes in …, Springer, cited by 2 (0.17 per year)*

*C Pereira, A Dourado (2002)* ***On the complexity and interpretability of support vector machines for process modeling***. *… , 2002. IJCNN'02. Proceedings of the …, ieeexplore.ieee.org, cited by 2 (0.13 per year)*

*G Panoutsos, M Mahfouf, GH Mills, ... (2010)* ***A generic framework for enhancing the interpretability Of granular computing-based information***. *Intelligent Systems (IS) …, ieeexplore.ieee.org, cited by 2 (0.25 per year)*

*R Barcellos, J Viterbo, L Miranda, F Bernardini, ... (2017)* ***Transparency in practice: using visualization to enhance the interpretability of open data***. *Proceedings of the 18th …, dl.acm.org, cited by 1 (1.00 per year)*

*M Eftekhari, M Eftekhari, M Majidi (2012)* ***Securing interpretability of fuzzy models for modeling nonlinear MIMO systems using a hybrid of evolutionary algorithms***. *Iranian Journal of Fuzzy Systems, ijfs.usb.ac.ir, cited by 5 (0.83 per year)*

*JM Alonso, L Magdalena (2009)* ***An interpretability-guided modeling process for learning comprehensible fuzzy rule-based classifiers***. *Intelligent Systems Design and …, ieeexplore.ieee.org, cited by 4 (0.44 per year)*

*Y Jin (2003)* ***Interpretability improvement of RBF-based neurofuzzy systems using regularized learning***. *Interpretability Issues in Fuzzy Modeling, Springer, cited by 1 (0.07 per year)*

*JN Foerster, J Gilmer, J Sohl-Dickstein, ... (2017)* ***Input switched affine networks: An RNN architecture designed for interpretability***. *... Machine Learning, proceedings.mlr.press, cited by 2 (2.00 per year)*

*JM Alonso, C Castiello, C Mencar (2015)* ***Interpretability of fuzzy systems: Current research trends and prospects***. *Springer Handbook of Computational …, Springer, cited by 27 (9.00 per year)*

*A Ghandar, Z Michalewicz (2011)* ***An experimental study of Multi-Objective Evolutionary Algorithms for balancing interpretability and accuracy in fuzzy***

***rulebase classifiers for financial prediction**. Computational Intelligence for ..., ieeexplore.ieee.org, cited by 9 (1.29 per year)*

*FA Pasquale (2017) **Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society**., papers.ssrn.com, cited by 1 (1.00 per year)*

*E Lughofer (2012) **Navigating interpretability issues in evolving fuzzy systems**. Scalable Uncertainty Management, Springer, cited by 1 (0.17 per year)*

*F Doshi-Velez, B Kim (2017) **A Roadmap for a Rigorous Science of Interpretability**. arXiv preprint arXiv:1702.08608, arxiv.org, cited by 8 (8.00 per year)*

*Y Dong, H Su, J Zhu, B Zhang (2017) **Improving Interpretability of Deep Neural Networks with Semantic Information**. arXiv preprint arXiv:1703.04096, arxiv.org, cited by 5 (5.00 per year)*

*GS Carpena, JFS Ruiz, JMA Muñoz, ... (2008) **Improving interpretability of fuzzy models using multi-objective neuro-evolutionary algorithms**. Advances in ..., intechopen.com, cited by 2 (0.20 per year)*

*T Laugel, MJ Lesot, C Marsala, X Renard, ... (2017) **Inverse Classification for Comparison-based Interpretability in Machine Learning**. arXiv preprint arXiv ..., arxiv.org*

*M Tulio Ribeiro, S Singh, ... (2016) **Model-Agnostic Interpretability of Machine Learning**. arXiv preprint arXiv ..., adsabs.harvard.edu*

*F Offert (2017) **" I know it when I see it". Visualization and Intuitive Interpretability**. arXiv preprint arXiv:1711.08042, arxiv.org*

*S Pereira, R Meier, R McKinley, R Wiest, V Alves, ... (2017) **Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation**. Medical Image ..., Elsevier*

*S Chakraborty, R Tomsett, R Raghavendra, ... (2017) **Interpretability of deep learning models: a survey of results**., pdfs.semanticscholar.org*

*P Hall, N Gill, M Kurka, W Phan (2017) **Machine Learning Interpretability with H2O Driverless AI**., docs.h2o.ai*

*AM Ghandar, Z Michalewicz (2011) **Considerations of the nature of the relationship between generalization and interpretability in evolutionary fuzzy systems**. ... of the 13th annual conference companion ..., dl.acm.org*

*B Herman (2017) **The Promise and Peril of Human Evaluation for Model Interpretability**. arXiv preprint arXiv:1711.07414, arxiv.org*

*S Sarkar, T Weyde, A Garcez, ... (2016) **Accuracy and interpretability trade-offs in machine learning applied to safer gambling**. CEUR Workshop ..., openaccess.city.ac.uk*

*A Sinha (2017) **Scalable black-box model explainability through low-dimensional visualizations**., dspace.mit.edu*

*A Bibal, B Frénay (2016) **Learning Interpretability for Visualizations using Adapted Cox Models through a User Experiment**. arXiv preprint* *arXiv:1611.06175, arxiv.org*

*H Liu, A Gegov, M Cocea (2016)* ***Interpretability Analysis****. Rule Based Systems for Big Data, Springer*

*H Ponce, ... (2017)* ***Interpretability of artificial hydrocarbon networks for breast cancer classification****. ... 2017 International Joint ..., ieeexplore.ieee.org*

*M Andel, F Masri* ***Sparse Omics-network Regularization to Increase Interpretability and Performance of SVM-based Predictive Models****. radio.feld.cvut.cz*

*M Wu, MC Hughes, S Parbhoo, M Zazzi, V Roth, ... (2017)* ***Beyond Sparsity: Tree Regularization of Deep Models for Interpretability****. arXiv preprint arXiv ..., arxiv.org*

*RBC Evolutionary* ***Accuracy vs. Interpretability of Fuzzy Rule-Based Classifiers-an Evolutionary Approach****. beta.tu.kielce.pl*

*MA Chikh, N Settouti, M Saidi (2012)* ***The fundamental nature of interpretability in diagnosing diabetes using neuro-fuzzy classifier****. Journal of Medical Imaging and ..., ingentaconnect.com*

*P Najaf, VR Duddu, SS Pulugurtha (2017)* ***Predictability and interpretability of hybrid link-level crash frequency models for urban arterials compared to cluster-based and general negative binomial regression …****. International Journal of Injury ..., Taylor & Francis*

*T Assya, L Sebastien, P Claude (2017)* ***Warp: a method for neural network interpretability applied to gene expression profiles****. arXiv preprint arXiv:1708.04988, arxiv.org*

*T Kenesei, J Abonyi (2015)* ***Interpretability of Support Vector Machines****. Interpretability of Computational Intelligence-Based ..., Springer*

*A Moore, Y Cai, K Jones, V Murdock (2017)* ***Tree Ensemble Explainability****., openreview.net*

*K Cpałka (2017)* ***Case Study: Interpretability of Fuzzy Systems Applied to Nonlinear Modelling and Control****. Design of Interpretable Fuzzy Systems, Springer*

*PJ Kindermans, KT Schütt, M Alber, KR Müller, ... (2017)* ***PatternNet and PatternLRP--Improving the interpretability of neural networks****. arXiv preprint arXiv ..., arxiv.org, cited by 6 (6.00 per year)*

*M Anděl, F Masri, J Kléma, Z Krejčík, ... (2015)* ***Sparse omics-network regularization to increase interpretability and performance of linear classification models****. ... (BIBM), 2015 IEEE ..., ieeexplore.ieee.org*

*E Lughofer (2011)* ***Interpretability Issues in EFS****. Evolving Fuzzy Systems–Methodologies, Advanced ..., Springer, cited by 1 (0.14 per year)*

*J Ling, M Hutchinson, E Antono, B DeCost, ... (2017)* ***Building Data-driven Models with Microstructural Images: Generalization and Interpretability****. arXiv preprint arXiv ..., arxiv.org*

*B Letham (2015)* ***Statistical learning for decision making: interpretability, uncertainty, and inference****., dspace.mit.edu*

*IA Tradeoff* ***Improvement Opportunities in the Design of Multi-Objective Evolutionary Fuzzy Classifiers: Handling Rule Selection and Interpretability-Accuracy Tradeoff****. pdfs.semanticscholar.org*

*FEB Otero, AA Freitas (2016)* ***Improving the interpretability of classification rules discovered by an ant colony algorithm: extended results****. Evolutionary computation, ieeexplore.ieee.org, cited by 5 (2.50 per year)*

*T Zhou, FL Chung, S Wang (2017)* ***Deep TSK Fuzzy Classifier With Stacked Generalization and Triplely Concise Interpretability Guarantee for Large Data****. IEEE Transactions on Fuzzy …, ieeexplore.ieee.org*

*Q Shen, JG Marín-Blázquez (2002)* ***Microtuning of membership functions: accuracy vs. interpretability****. ... FUZZ-IEEE'02. Proceedings of the ..., ieeexplore.ieee.org, cited by 4 (0.25 per year)*

*S Guillaume (2001)* ***Designing fuzzy inference systems from data: An interpretability-oriented review****. IEEE Transactions on fuzzy systems, ieeexplore.ieee.org, cited by 659 (38.76* per year)*

*M Virág, T Nyitrai (2014)* ***Is there a trade-off between the predictive power and the interpretability of bankruptcy models? The case of the first Hungarian bankruptcy prediction model****. Acta Oeconomica, akademiai.com, cited by 9 (2.25 per year)*

*LK Senel, I Utlu, V Yucesoy, A Koc, T Cukur (2017)* ***Semantic Structure and Interpretability of Word Embeddings****. arXiv preprint arXiv …, arxiv.org*

*K Cpałka (2017)* ***Introduction to Fuzzy System Interpretability****. Design of Interpretable Fuzzy Systems, Springer*

*C Mencar (2005)* ***Theory of fuzzy information granulation: Contributions to interpretability issues****. University of Bari, academia.edu, cited by 23 (1.77 per year)*

*F Rudziński (2016)* ***A multi-objective genetic optimization of interpretability-oriented fuzzy rule-based classifiers****. Applied Soft Computing, Elsevier, cited by 27 (13.50* per year)*

*E Lughofer (2016)* ***Evolving Fuzzy Systems—Fundamentals, Reliability, Interpretability, Useability, Applications****. ... INTELLIGENCE: Volume 1: Fuzzy Logic, Systems …, books.google.com, cited by 10 (5.00 per year)*

*AA Márquez, FA Márquez, ... (2012)* ***A mechanism to improve the interpretability of linguistic fuzzy systems with adaptive defuzzification based on the use of a multi-objective evolutionary algorithm****. International Journal of ..., Taylor & Francis, cited by 11 (1.83 per year)*

*T Kenesei, J Abonyi (2015)* ***Interpretability of Computational Intelligence-Based Regression Models****., Springer*

*JM Alonso, O Cordón, A Quirin, ... (2011)* ***Analyzing interpretability of fuzzy rule-based systems by means of fuzzy inference-grams****. World Congress on Soft ..., researchgate.net, cited by 19 (2.71 per year)*

*B Bouchon-Meunier (2015)* ***Interpretability, a Silver Lining to a Fuzzy Cloud****. Enric Trillas: A Passion for Fuzzy Sets, Springer*

*S Askari (2017)* ***A novel and fast MIMO fuzzy inference system based on a class*** *of fuzzy clustering algorithms with interpretability and complexity analysis*. *Expert Systems with Applications, Elsevier, cited by 1 (1.00 per year)*

*A Di Nuovo, G Ascia (2013)* ***A fuzzy system index to preserve interpretability in deep tuning of fuzzy rule based classifiers***. *Journal of Intelligent & Fuzzy Systems, content.iospress.com, cited by 3 (0.60 per year)*

*A Fiordaliso (2003)* ***About the trade-off between accuracy and interpretability of takagi-sugeno models in the context of nonlinear time series forecasting***. *Interpretability Issues in Fuzzy Modeling, Springer, cited by 1 (0.07 per year)*

*W Pedrycz (2003)* ***Expressing relevance interpretability and accuracy of rule-based systems***. *Interpretability issues in fuzzy modeling, Springer, cited by 6 (0.40 per year)*

*MA Mortada (2010)* ***Applicability and interpretability of logical analysis of data in condition based maintenance***., *publications.polymtl.ca, cited by 9 (1.13 per year)*

*MJ Gacto, R Alcalá, F Herrera (2010)* ***Integration of an index to preserve the semantic interpretability in the multiobjective evolutionary rule selection and tuning of linguistic fuzzy systems***. *IEEE Transactions on Fuzzy ..., ieeexplore.ieee.org, cited by 134 (16.75* per year)*

*U Bodenhofer, P Bauer (2005)* ***Interpretability of linguistic variables: a formal account***. *Kybernetika, dml.cz, cited by 18 (1.38 per year)*

*C Mencar (2013)* ***Interpretability of Fuzzy Systems.***. *WILF, Springer, cited by 10 (2.00 per year)*

*RJ Kelly, JA Smith, SLR Kardia (2010)* ***8 Providing Context and Interpretability to Genetic Association Analysis Results Using the KGraph***. *Advances in genetics, books.google.com*

*T Kenesei, J Abonyi (2015)* ***Interpretability of Hinging Hyperplanes***. *Interpretability of Computational Intelligence-Based ..., Springer*

*MI Rey, M Galende, MJ Fuente, ... (2017)* ***Multi-objective based Fuzzy Rule Based Systems (FRBSs) for trade-off improvement in accuracy and interpretability: A rule relevance point of view.***. *Knowledge-Based ..., Elsevier*

*RL Marchese Robinson, A Palczewska, ... (2017)* ***Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets***. *Journal of chemical ..., ACS Publications, cited by 2 (2.00 per year)*

*R Cannone, C Castiello, C Mencar, ... (2009)* ***A study on interpretability conditions for fuzzy rule-based classifiers***. *... Systems Design and ..., ieeexplore.ieee.org, cited by 5 (0.56 per year)*

*A Dhurandhar, V Iyengar, R Luss, ... (2017)* ***TIP: Typifying the Interpretability of Procedures***. *arXiv preprint arXiv ..., arxiv.org*

*M Gan, CL Philip Chen, L Chen, ... (2016)* ***Exploiting the interpretability and forecasting ability of the RBF-AR model for nonlinear time series***. *International Journal of ..., Taylor & Francis, cited by 3 (1.50 per year)*

*C Mencar (2009)* ***Interpretability of fuzzy information granules***. *Human-Centric Information Processing Through ..., Springer, cited by 6 (0.67 per year)*

*E Aguirre, A Gonzalez, R Pérez (2003)* ***A description of several characteristics for improving the accuracy and interpretability of inductive linguistic rule learning algorithms****. Accuracy Improvements in ..., books.google.com, cited by 1 (0.07 per year)*

*E Aguirre, A González, R Pérez (2013)* ***improving the accuracy and interpretability of inductive linguistic rule learning algorithms****. Accuracy Improvements in ..., books.google.com*

*JA Jakubczyc (2010)* ***The Interpretability of Contextual Classifier Ensemble****. Prace Naukowe Uniwersytetu Ekonomicznego ..., bazekon.icm.edu.pl*

*V Lužar-Stiffler, C Stiffler (2004)* ***"BOF" Trees Diagram as a Visual Way to Improve Interpretability of Tree Ensembles****. CIT. Journal of computing and information ..., hrcak.srce.hr*

*K Cpałka (2017)* ***Improving Fuzzy Systems Interpretability by Appropriate Selection of Their Structure****. Design of Interpretable Fuzzy Systems, Springer*

*JM Alonso, L Magdalena (2010)* ***Combining user's preferences and quality criteria into a new index for guiding the design of fuzzy systems with a good interpretability-accuracy trade-off****. Fuzzy Systems (FUZZ), 2010 IEEE ..., ieeexplore.ieee.org, cited by 8 (1.00 per year)*

*SM Kia (2016)* ***Interpretability of Multivariate Brain Maps in Brain Decoding: Definition and Quantification****. arXiv preprint arXiv:1603.08704, arxiv.org*

*TR Razak, JM Garibaldi, C Wagner, ... (2017)* ***Interpretability indices for hierarchical fuzzy systems****., westminsterresearch.wmin.ac.uk*

*JG Marın-Blázquez, Q Shen* ***Microtuning of Membership Functions: Accuracy vs Interpretability****. pdfs.semanticscholar.org*

*K Cpałka (2017)* ***Interpretability of Fuzzy Systems Designed in the Process of Evolutionary Learning****. Design of Interpretable Fuzzy Systems, Springer*

*M Fazzolari (2014)* ***Study and design of multi-objective evolutionary fuzzy systems for improving the interpretability accuracy trade off of linguistic fuzzy rule based systems when ….****, digibug.ugr.es*

*BHW Chang (2014)* ***Kernel Machines are not Black Boxes-On the Interpretability of Kernel-based Nonparametric Models****., search.proquest.com*

*R Guha (2005)* ***Methods to improve the reliability, validity and interpretability of QSAR models****., pdfs.semanticscholar.org, cited by 11 (0.85 per year)*

**Query report 02 – “interpreting” AND “explaining”**

# *(intitle:interpreting OR intitle:explaining) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") to 2017*

*Publish or Perish 6.21.6145.6594*

## *Search terms*

***All of the words:*** *(intitle:interpreting OR intitle:explaining) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning")*
***Years:*** *earliest to 2017*

## *Data retrieval*

***Data source:*** *Google Scholar*
***Query date:*** *21/01/2018 12:12:59*
***Cache date:*** *21/01/2018 12:13:31*
***Query result:*** *[0] The operation completed successfully.*

## *Metrics*

***Publication years:*** *1972-2017*
***Citation years:*** *46 (1972-2018)*
***Papers:*** *168*
***Citations:*** *2852*
***Citations/year:*** *62.00*
***Citations/paper:*** *16.98 (*count=9)*
***Citations/author:*** *1270.25*
***Papers/author:*** *87.05*
***Authors/paper:*** *2.61/3.0/1 (mean/median/mode)*
***Age-weighed citation rate:*** *571.22 (sqrt=23.90), 216.45/author*
***Hirsch h-index:*** *23 (a=5.39, m=0.50, 2271 cites=79.6% coverage)*
***Egghe g-index:*** *51 (g/h=2.22, 2648 cites=92.8% coverage)*
***PoP hI,norm:*** *15*
***PoP hI,annual:*** *0.33*

## *Results*

*A Glass (2006)* ***Explaining Preference Learning****., pdfs.semanticscholar.org, cited by 107 (8.92 per year)*

*BR Kowalski, CF Bender (1972)* ***Pattern recognition. Powerful approach to interpreting chemical data****. Journal of the American Chemical ..., ACS Publications, cited by 509 (11.07* per year)*

*MT Ribeiro, S Singh, C Guestrin (2016)* ***Why should i trust you?: Explaining the predictions of any classifier****. Proceedings of the 22nd ACM ..., dl.acm.org, cited by 279 (139.50* per year)*

*E Štrumbelj, I Kononenko (2011)* ***A general method for visualizing and explaining black-box regression models****. International Conference on Adaptive and ..., Springer, cited by 13 (1.86 per year)*

*L Rosenbaum, G Hinselmann, A Jahn, A Zell (2011)* ***Interpreting linear support vector machine models with heat map molecule coloring****. Journal of cheminformatics, Springer, cited by 28 (4.00 per year)*

*M Robnik-Šikonja, I Kononenko (2008)* ***Explaining classifications for individual instances****. IEEE Transactions on ..., ieeexplore.ieee.org, cited by 85 (8.50 per year)*

*R Guha, PC Jurs (2005)* ***Interpreting computational neural network QSAR models: a measure of descriptor importance****. Journal of chemical information and modeling, ACS Publications, cited by 95 (7.31 per year)*

*R Wall, P Cunningham, P Walsh, S Byrne (2003)* ***Explaining the output of ensembles in medical decision support on a case by case basis****. Artificial intelligence in medicine, Elsevier, cited by 29 (1.93 per year)*

*R Ramirez, A Hazan (2006)* ***A tool for generating and explaining expressive music performances of monophonic jazz melodies****. International Journal on Artificial Intelligence ..., World Scientific, cited by 26 (2.17 per year)*

*T Olsson, D Gillblad, P Funk, ... (2014)* ***Case-based reasoning for explaining probabilistic machine learning****. International Journal of ..., search.proquest.com, cited by 6 (1.50 per year)*

*G Montavon, S Lapuschkin, A Binder, W Samek, ... (2017)* ***Explaining nonlinear classification decisions with deep taylor decomposition****. Pattern Recognition, Elsevier, cited by 65 (65.00* per year)*

*YJ Lin (2010)* ***Explaining critical clearing time with the rules extracted from a multilayer perceptron artificial neural network****. International Journal of Electrical Power & Energy ..., Elsevier, cited by 16 (2.00 per year)*

*R Wall, P Cunningham, P Walsh (2002)* ***Explaining predictions from a neural network ensemble one at a time****. PKDD, Springer, cited by 8 (0.50 per year)*

*G Montavon, W Samek, KR Müller (2017)* ***Methods for interpreting and understanding deep neural networks****. Digital Signal Processing, Elsevier, cited by 19 (19.00* per year)*

*BP Knijnenburg, MC Willemsen, Z Gantner, ... (2012)* ***Explaining the user experience of recommender systems****. User Modeling and ..., dl.acm.org, cited by 316 (52.67* per year)*

*W Landecker, MD Thomure, ... (2013)* ***Interpreting individual classifications of hierarchical networks****. ... and Data Mining ..., ieeexplore.ieee.org, cited by 14 (2.80 per year)*

*M Pregeljc, E Štrumbelj, M Mihelcic, ... (2012)* ***Learning and explaining the***

***impact of enterprises' organizational quality on their economic results**. ... Data Analysis for Real …, books.google.com, cited by 7 (1.17 per year)*

*K Främling (1996) **Explaining results of neural networks by contextual importance and utility**. Proceedings of the AISB'96 conference, researchgate.net, cited by 8 (0.36 per year)*

*U Johansson, C Sonstrod, ... (2006) **Explaining Winning Poker--A Data Mining Approach**. Machine Learning and …, ieeexplore.ieee.org, cited by 10 (0.83 per year)*

*S Grossberg, J Markowitz, Y Cao (2011) **On the road to invariant recognition: explaining tradeoff and morph properties of cells in inferotemporal cortex using multiple-scale task-sensitive attentive …**. Neural Networks, Elsevier, cited by 24 (3.43 per year)*

*C Aguilar, H Lipson (2008) **A robotic system for interpreting images into painted artwork**. International conference on generative art, generativeart.com, cited by 15 (1.50 per year)*

*E Štrumbelj, I Kononenko (2014) **Explaining prediction models and individual predictions with feature contributions**. Knowledge and information systems, Springer, cited by 13 (3.25 per year)*

*R Hasan (2014) **Predicting SPARQL query performance and explaining linked data**. European Semantic Web Conference, Springer, cited by 9 (2.25 per year)*

*R Blanco, D Ceccarelli, C Lucchese, R Perego, ... (2012) **You should read this! let me explain you why: explaining news recommendations to users**. Proceedings of the 21st …, dl.acm.org, cited by 11 (1.83 per year)*

*Y Goyal, A Mohapatra, D Parikh, D Batra (2016) **Towards Transparent AI Systems: Interpreting Visual Question Answering Models**. arXiv preprint arXiv:1608.08974, arxiv.org, cited by 13 (6.50 per year)*

*A Palczewska, J Palczewski, ... (2013) **Interpreting random forest models using a feature contribution method**. … and Integration (IRI) …, ieeexplore.ieee.org, cited by 13 (2.60 per year)*

*B Cope, M Kalantzis (2015) **Interpreting Evidence-of-Learning: Educational research in the era of big data**. Open Review of Educational Research, Taylor & Francis, cited by 12 (4.00 per year)*

*W Samek, T Wiegand, KR Müller (2017) **Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models**. arXiv preprint arXiv:1708.08296, arxiv.org, cited by 3 (3.00 per year)*

*J Reilly, K McCarthy, L McGinty, B Smyth (2005) **Explaining compound critiques**. Artificial Intelligence Review, Springer, cited by 29 (2.23 per year)*

*V Kolhatkar, H Zinsmeister, G Hirst (2013) **Interpreting Anaphoric Shell Nouns using Antecedents of Cataphoric Shell Nouns as Training Data.**. EMNLP, academia.edu, cited by 13 (2.60 per year)*

*M Robnik-Šikonja, A Likas, ... (2011) **Efficiently explaining decisions of probabilistic RBF classification networks**. … Conference on Adaptive …, Springer, cited by 4 (0.57 per year)*

*A Nath, K Subbiah (2014) **Inferring biological basis about psychrophilicity by***

***interpreting the rules generated from the correctly classified input instances by a classifier**. Computational biology and chemistry, Elsevier, cited by 6 (1.50 per year)*

*D Sorokina, R Caruana, M Riedewald, ... (2009) **Detecting and interpreting variable interactions in observational ornithology data**. ... , 2009. ICDMW'09 ..., ieeexplore.ieee.org, cited by 6 (0.67 per year)*

*D Westreich, S Greenland (2013) **The table 2 fallacy: presenting and interpreting confounder and modifier coefficients**. American journal of epidemiology, academic.oup.com, cited by 53 (10.60* per year)*

*E Strumbelj, I Kononenko (2008) **Towards a Model Independent Method for Explaining Classification for Individual Instances.**. DaWaK, Springer, cited by 7 (0.70 per year)*

*T Olsson, D Gillblad, P Funk, N Xiong (2014) **Explaining probabilistic fault diagnosis and classification using case-based reasoning**. International Conference on Case ..., Springer, cited by 4 (1.00 per year)*

*Y Yang, C Hinde, D Gillingwater (2003) **A new method for explaining neural network reasoning**. Neural Networks, 2003 ..., ieeexplore.ieee.org, cited by 4 (0.27 per year)*

*Y Nohara, Y Wakata, N Nakashima (2015) **Interpreting Medical Information Using Machine Learning and Individual Conditional Expectation.**. MedInfo, researchgate.net, cited by 1 (0.33 per year)*

*L Arras, G Montavon, KR Müller, W Samek (2017) **Explaining recurrent neural network predictions in sentiment analysis**. arXiv preprint arXiv ..., arxiv.org, cited by 6 (6.00 per year)*

*Y Zhou, G Hooker (2016) **Interpreting Models via Single Tree Approximation**. arXiv preprint arXiv:1610.09036, arxiv.org, cited by 3 (1.50 per year)*

*DL McGuinness, V Furtado, PP da Silva, ... (2010) **Explaining semantic web applications**. Web Technologies ..., igi-global.com, cited by 13 (1.63 per year)*

*M Bohanec, MK Borštnar, M Robnik-Šikonja (2017) **Explaining machine learning models in sales predictions**. Expert Systems with ..., Elsevier, cited by 5 (5.00 per year)*

*P O'Rorke, A Ortony (1994) **Explaining emotions**. Cognitive Science, Wiley Online Library, cited by 98 (4.08 per year)*

*D Yedjour, H Yedjour, A Benyettou (2011) **Explaining Results of Artificial Neural Networks**. Journal of Applied Sciences, researchgate.net, cited by 3 (0.43 per year)*

*S Halim, RHC Yap, HC Lau (2006) **Viz: a visual analysis suite for explaining local search behavior**. Proceedings of the 19th annual ACM ..., dl.acm.org, cited by 25 (2.08 per year)*

*M Green, U Ekelund, L Edenbrandt, ... (2008) **Explaining artificial neural network ensembles: A case study with electrocardiograms from chest pain patients**. ... Machine Learning ..., portal.research.lu.se, cited by 2 (0.20 per year)*

*K Choi, G Fazekas, M Sandler (2016) **Explaining deep convolutional neural networks on music classification**. arXiv preprint arXiv:1607.02444, arxiv.org, cited by 5 (2.50 per year)*

*W Zadrozny, V de Paiva, LS Moss (2015)* ***Explaining Watson: Polymath Style.****. AAAI, aaai.org, cited by 3 (1.00 per year)*

*C Ventura, F Célimene, R Nock, ... (2011)* ***Predicting and interpreting business failures with supervised information geometric algorithms****. ... Conference on Business ..., sta.uwi.edu, cited by 3 (0.43 per year)*

*T Pevný, M Kopp (2014)* ***Explaining anomalies with sapling random forests****. ... Technologies–Applications and ..., webdav.agents.fel.cvut.cz, cited by 4 (1.00 per year)*

*F Sieverink, S Kelders, M Poel, ... (2017)* ***Opening the black box of electronic health: collecting, analyzing, and interpreting log data****. JMIR research ..., ncbi.nlm.nih.gov, cited by 1 (1.00 per year)*

*M Robnik-Šikonja, A Likas, ... (2007)* ***An efficient method for explaining the decisions of the probabilistic RBF classification network****. ... of Ljubljana (FRI) ..., researchgate.net, cited by 1 (0.09 per year)*

*KL O'Halloran, S Tan, P Wignell, ... (2016)* ***Interpreting text and image relations in violent extremist discourse: A mixed methods approach for big data analytics****. ... and Political Violence, Taylor & Francis, cited by 7 (3.50 per year)*

*DL McGuinness (2007)* ***Explaining complex systems****. Semantic e-Science Workshop co-located with ..., vsto.hao.ucar.edu, cited by 1 (0.09 per year)*

*M Robnik-Šikonja, E Štrumbelj, ... (2013)* ***Efficiently explaining the predictions of a probabilistic radial basis function classification network****. Intelligent data ..., content.iospress.com, cited by 1 (0.20 per year)*

*D Martens, F Provost (2013)* ***Explaining data-driven document classifications****., papers.ssrn.com, cited by 46 (9.20 per year)*

*CL Giles, C Omlin (1999)* ***. Understanding and Explaining DRN Behavior****., pdfs.semanticscholar.org, cited by 8 (0.42 per year)*

*NP Da Silva, M Marques, G Carneiro, JP Costeira (2011)* ***Explaining scene composition using kinematic chains of humans: application to Portuguese tiles history****. Proc. of SPIE Vol, cited by 5 (0.71 per year)*

*DL McGuinness, A Glass, M Wolverton, ... (2007)* ***Explaining Task Processing in Cognitive Assistants That Learn.****. AAAI Spring ..., vvvvw.aaai.org, cited by 17 (1.55 per year)*

*CM Stanley, SR Sunyaev, MS Greenblatt, ... (2014)* ***Clinically relevant variants–identifying, collecting, interpreting, and disseminating: the 2013 annual scientific meeting of the human genome variation society****. Human ..., Wiley Online Library, cited by 15 (3.75 per year)*

*D Szafron, R Greiner, P Lu, D Wishart, C MacDonell, ... (2003)* ***Explaining naïve Bayes classifications****., era.library.ualberta.ca, cited by 9 (0.60 per year)*

*J Vig, S Sen, J Riedl (2009)* ***Tagsplanations: explaining recommendations using tags****. Proceedings of the 14th international conference ..., dl.acm.org, cited by 157 (17.44* per year)*

*O Bastani, C Kim, H Bastani (2017)* ***Interpreting Blackbox Models via Model Extraction****. arXiv preprint arXiv:1705.08504, arxiv.org, cited by 1 (1.00 per year)*

*X Morice-Atkinson, B Hoyle, D Bacon (2017)* ***Learning from the machine: interpreting machine learning algorithms for point-and extended-source classification****. arXiv preprint arXiv:1712.03970, arxiv.org*

*M Morin, R Thomopoulos, I Abi-Zeid, ... (2016)* ***Explaining the Results of an Optimization-Based Decision Support System–A Machine Learning Approach****. APMOD: APplied …, hal-lirmm.ccsd.cnrs.fr*

*S Penkov, S Ramamoorthy (2017)* ***Explaining Transition Systems through Program Induction****. arXiv preprint arXiv:1705.08320, arxiv.org*

*P Tamagnini, J Krause, A Dasgupta, E Bertini (2017)* ***Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations.****. HILDA@ SIGMOD, law.nyu.edu, cited by 1 (1.00 per year)*

*M Francis (2017)* ***On Learning Sparse Boolean Formulae for Explaining AI Decisions****. NASA Formal Methods: 9th International Symposium …, books.google.com*

*O Nasraoui (2017)* ***Tell me Why? Tell me More! Explaining Predictions, Iterated Learning Bias, and Counter-Polarization in Big Data Discovery Models****., uknowledge.uky.edu*

*W Samek, G Montavon, A Binder, S Lapuschkin, ... (2016)* ***Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation****. arXiv preprint arXiv …, arxiv.org, cited by 1 (0.50 per year)*

*D Kumar, GW Taylor, A Wong (2017)* ***Opening the Black Box of Financial AI with CLEAR-Trade: A CLass-Enhanced Attentive Response Approach for Explaining and Visualizing Deep Learning-Driven …****. arXiv preprint arXiv:1709.01574, arxiv.org*

*CA Hammerschmidt, Q Lin, S Verwer, ... (2016)* ***Interpreting Finite Automata for Sequential Data****. arXiv preprint arXiv …, arxiv.org*

*S Jha, V Raman, A Pinto, T Sahai, M Francis (2017)* ***On Learning Sparse Boolean Formulae for Explaining AI Decisions****. NASA Formal Methods …, Springer, cited by 1 (1.00 per year)*

*EC Freuder (2017)* ***Explaining Ourselves: Human-Aware Constraint Reasoning.****. AAAI, aaai.org, cited by 1 (1.00 per year)*

*ER Elenberg, AG Dimakis, M Feldman, ... (2017)* ***Streaming Weak Submodularity: Interpreting Neural Networks on the Fly****. arXiv preprint arXiv …, arxiv.org, cited by 3 (3.00 per year)*

*A Henelius, K Puolamäki, A Ukkonen (2017)* ***Interpreting Classifiers through Attribute Interactions in Datasets****., openreview.net, cited by 1 (1.00 per year)*

*CH Chang, E Creager, A Goldenberg, D Duvenaud* ***Interpreting Neural Network Classifications with Variational Dropout Saliency Maps****. cs.toronto.edu*

*H Wu, C Wang, J Yin, K Lu, L Zhu (2017)* ***Interpreting Shared Deep Learning Models via Explicable Boundary Trees****. arXiv preprint arXiv:1709.03730, arxiv.org*

*PK Douglas, A Anderson* ***Interpreting fMRI Decoding Weights: Additional Considerations****. interpretable-ml.org*

*T Furukawa, Q Zhao (2017)* ***Interpreting Multilayer Perceptrons Using 3-Valued Activation Function****. Cybernetics (CYBCONF), 2017 3rd IEEE …, ieeexplore.ieee.org*

*R Takahashi, N Inoue, Y Kuriya, ... (2016)* ***Explaining Potential Risks in Traffic Scenes by Combining Logical Inference and Physical Simulation****. ... of Machine Learning …, search.proquest.com*

*P Hitzler* ***Towards Explaining Neural Networks Through Background Knowledge****. daselab.cs.wright.edu*

*C Stoean, R Stoean (2014)* ***Evolutionary Algorithms Explaining Support Vector Learning****. ... Vector Machines and Evolutionary Algorithms for …, Springer*

*J de Ruiter, T Knijnenburg, J de Ridder (2017)* ***Mining the forest: uncovering biological mechanisms by interpreting Random Forests****. bioRxiv, biorxiv.org*

*S Rüping (2005)* ***Interpreting Classifiers by Multiple Views****. Learning With Multiple Views, stefan-rueping.de*

*K Jana, G Matjaz (2011)* ***Data Mining Techniques for Explaining Social Events****. Knowledge-Oriented Applications in Data …, intechopen.com*

*KR Fleischmann, C Templeton, J Boyd-Graber, ...* ***Explaining Sentiment Polarity****. cs.colorado.edu*

*R Stoean, C Stoean, A Sandita, D Ciobanu, ... (2017)* ***Interpreting Decision Support from Multiple Classifiers for Predicting Length of Stay in Patients with Colorectal Carcinoma****. Neural Processing …, Springer, cited by 1 (1.00 per year)*

*D Kumar, A Wong, GW Taylor (2017)* ***Explaining the Unexplained: A CLass-Enhanced Attentive Response (CLEAR) Approach to Understanding Deep Neural Networks****. arXiv preprint arXiv:1704.04133, arxiv.org, cited by 4 (4.00 per year)*

*B Micenková, XH Dang, I Assent, ... (2013)* ***Explaining outliers by subspace separability****. Data Mining (ICDM) …, ieeexplore.ieee.org, cited by 31 (6.20 per year)*

*A Bustillo, M Grzenda, ... (2016)* ***Interpreting tree-based prediction models and their data in machining processes****. Integrated Computer-Aided …, content.iospress.com, cited by 1 (0.50 per year)*

*W AJMM, B APJ (1999)* ***Interpreting knowledge representations in BP-SOM****. Behaviormetrika, jstage.jst.go.jp, cited by 5 (0.26 per year)*

*JO Jandl (2016)* ***Information Processing in Securitized Real Estate Markets--How Newspaper Content and Online Search Behavior Help Explaining Market Movements****. System Sciences (HICSS), 2016 49th Hawaii …, ieeexplore.ieee.org*

*P De Koninck, J De Weerdt, ... (2017)* ***Explaining clusterings of process instances****. Data Mining and …, Springer, cited by 1 (1.00 per year)*

*D Martens, F Provost (2011)* ***Explaining documents' classifications****. Center for Digital Economy …, pdfs.semanticscholar.org, cited by 6 (0.86 per year)*

*N Dhir, F Wood, M Vákár, A Markham, M Wijers, ... (2017)* ***Interpreting lion behaviour with nonparametric probabilistic programs****., pdfs.semanticscholar.org, cited by 1 (1.00 per year)*

*RM Martins, R Minghim, AC Telea (2015)* ***Explaining neighborhood preservation for multidimensional projections****. EG UK Computer Graphics …, producao.usp.br, cited by 6 (2.00 per year)*

*CK Chan, S Gesbert, AR Masters, C Xu (2012)* ***Interpreting a plurality of M-dimensional attribute vectors assigned to a plurality of locations in an N-dimensional interpretation space****. US Patent 8,121,969, Google Patents, cited by 10 (1.67 per year)*

*V Van Belle, B Van Calster, S Van Huffel, JAK Suykens, ... (2016)* ***Explaining Support Vector Machines: A Color Based Nomogram****. PloS one, journals.plos.org, cited by 2 (1.00 per year)*

*MA ter Hoeve (2017)* ***Explaining Rankings****., pdfs.semanticscholar.org*

*L Li, M Fredrikson, S Sen, A Datta (2017)* ***Case Study: Explaining Diabetic Retinopathy Detection Deep CNNs via Integrated Gradients****. arXiv preprint arXiv:1709.09586, arxiv.org*

*E Lughofer, R Richter, U Neissl, W Heidl, C Eitzinger, ... (2017)* ***Explaining classifier decisions linguistically for stimulating and improving operators labeling behavior****. Information …, Elsevier*

*TC Wong, HK Chan, E Lacka (2017)* ***An ANN-based approach of interpreting user-generated comments from social media****. Applied Soft Computing, Elsevier*

*RD Das (2017)* ***Towards urban mobility-based activity knowledge discovery: interpreting motion trajectories****., minerva-access.unimelb.edu.au*

*A Palczewska, J Palczewski, RM Robinson, ... (2014)* ***Interpreting random forest classification models using a feature contribution method****. Integration of reusable …, Springer, cited by 21 (5.25 per year)*

*R Hasan (2014)* ***Predicting query performance and explaining results to assist Linked Data consumption****., tel.archives-ouvertes.fr*

*RV Boyd, CE Glass (1993)* ***Interpreting ground-penetrating radar images using object-oriented, neural, fuzzy, and genetic processing****. Ground Sensing, spiedigitallibrary.org, cited by 17 (0.68 per year)*

*L Chen, F Wang (2017)* ***Explaining Recommendations Based on Feature Sentiments in Product Reviews****. Proceedings of the 22nd International Conference on …, dl.acm.org, cited by 3 (3.00 per year)*

*D Chasman, B Gancarz, ... (2009)* ***Explaining Effects of Host Gene Knockouts on Brome Mosaic Virus Replication****. PRE WORKSHOP …, pdfs.semanticscholar.org, cited by 1 (0.11 per year)*

*J Dodge, S Penney, C Hilderbrand, A Anderson, ... (2017)* ***How the Experts Do It: Assessing and Explaining Agent Behaviors in Real-Time Strategy Games****. arXiv preprint arXiv …, arxiv.org*

*Y Liu, RN Horne (2013)* ***Interpreting Pressure and Flow Rate Data from Permanent Downhole Gauges with Convolution-Kernel-Based Data Mining****. SPE Annual Technical Conference and Exhibition, onepetro.org*

*K Benoit, K Munger, A Spirling (2017)* ***Measuring and Explaining Political Sophistication Through Textual Complexity****., papers.ssrn.com, cited by 1 (1.00 per year)*

*L Zintgraf (2015)* ***Explaining Individual Classifier Decisions****., researchgate.net*

*JM Whitacre (2009)* ***Survival of the flexible: explaining the dominance of meta-heuristics within a rapidly evolving world****., cogprints.org, cited by 1 (0.11 per year)*

*G Pant, P Srinivasan* ***Explaining and Predicting Web Page Status****. misrc.umn.edu*

*E Byrne (2012)* ***A logical framework for identifying and explaining unexpected news****. Computing and Informatics, cai.sk, cited by 4 (0.67 per year)*

*SC Chelgani, SS Matin, JC Hower (2016)* ***Explaining relationships between coke quality index and coal properties by Random Forest method****. Fuel, Elsevier, cited by 12 (6.00 per year)*

*D Martens, F Provost (2014)* ***Apparatus, method and computer-accessible medium for explaining classifications of documents****. US Patent App. 14/001,242, Google Patents, cited by 1 (0.25 per year)*

*DJ Redo, TM Aide, ML Clark (2012)* ***The relative importance of socioeconomic and environmental variables in explaining land change in Bolivia, 2001–2010****. Annals of the Association of ..., Taylor & Francis, cited by 22 (3.67 per year)*

*C Manescu, C Starica (2008)* ***The relevance of Corporate Social Responsibility criteria to explaining firm profitability: A case study of the publishers of the Dow Jones Sustainability Indexes1****., economia.uniroma2.it*

*M Kopp, T Pevny (2016)* ***Explaining network anomalies using decision trees****. US Patent App. 14/879,425, Google Patents, cited by 1 (0.50 per year)*

*H Eggels, R van Elk, M Pechenizkiy (2016)* ***Expected Goals in Soccer: Explaining Match Results using Predictive Analytics****. The Machine Learning and Data ..., pure.tue.nl, cited by 2 (1.00 per year)*

*BF Pennington (2014)* ***Explaining abnormal behavior: A cognitive neuroscience perspective****., books.google.com, cited by 10 (2.50 per year)*

*I Tiddi (2016)* ***Explaining Data Patterns using Knowledge from the Web of Data****., oro.open.ac.uk, cited by 2 (1.00 per year)*

*R Guha, DT Stanton, PC Jurs (2005)* ***Interpreting computational neural network quantitative structure− activity relationship models: A detailed interpretation of the weights and biases****. Journal of chemical information ..., ACS Publications, cited by 65 (5.00 per year)*

*C Glass (1993)* ***Interpreting ground penetrating radar images using object oriented, neural, fuzzy, and genetic processing Richard Boyd University of Arizona, Department of Mining ...****. Ground Sensing: 14 April 1993 ..., proceedings.spiedigitallibrary.org*

*MD Gillman (2014)* ***Interpreting human activity from electrical consumption data through non-intrusive load monitoring****., dspace.mit.edu, cited by 1 (0.25 per year)*

*GJ Du (2017)* ***Interpreting and optimizing data****., dspace.mit.edu*

*Y Zhang, A Jatowt, K Tanaka (2016)* ***Towards understanding word embeddings: Automatically explaining similarity of terms****. Big Data (Big Data), 2016 IEEE ..., ieeexplore.ieee.org, cited by 2 (1.00 per year)*

*E Štrumbelj, I Kononenko, MR Šikonja (2009)* ***Explaining instance classifications*** *with interactions of subsets of feature values. Data & Knowledge Engineering, Elsevier, cited by 29 (3.22 per year)*

*T Pevny (2016) **Explaining causes of network anomalies**. US Patent App. 14/331,486, Google Patents*

*T Weijters, A van den Bosch **Interpreting Knowledge Representations in**. pdfs.semanticscholar.org*

*JA Barceló, KF Achino, I Bogdanovic, ... (2015) **Measuring, counting and explaining: an introduction to mathematics in archaeology**. Mathematics and …, books.google.com, cited by 5 (1.67 per year)*

*J Stahnke, M Dörk, B Müller, ... (2016) **Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions**. IEEE transactions on …, ieeexplore.ieee.org, cited by 30 (15.00* per year)*

*G Du, T Zimmermann, G Ruhe (2008) **Explaining Product Release Planning Results Using Concept Analysis.**. SEKE, works.bepress.com, cited by 4 (0.40 per year)*

*MMC Vidovic (2017) **Improving and interpreting machine learning algorithms with applications**., depositonce.tu-berlin.de*

*N Saiya, A Scime (2015) **Explaining religious terrorism: A data-mined analysis**. Conflict Management and Peace Science, journals.sagepub.com, cited by 17 (5.67 per year)*

*F Gedikli, M Ge, D Jannach (2011) **Explaining online recommendations using personalized tag clouds**. i-com Zeitschrift für interaktive und …, degruyter.com, cited by 1 (0.14 per year)*

*DB Coimbra, RM Martins, TTAT Neves, ... (2016) **Explaining three-dimensional dimensionality reduction plots**. Information …, journals.sagepub.com, cited by 4 (2.00 per year)*

*DA Klein (1987) **Explaining and refining decision-theoretic choices**., repository.upenn.edu, cited by 5 (0.16 per year)*

*RC Deo (2016) **Alternative Splicing, Internal Promoter, Nonsense-Mediated Decay, or All ThreeCLINICAL PERSPECTIVE: Explaining the Distribution of Truncation Variants in …**. Circulation: Genomic and Precision Medicine, Am Heart Assoc, cited by 7 (3.50 per year)*

*K Kashin, G King, S Soneji (2015) **Explaining Systematic Bias and Nontransparency in US Social Security Administration Forecasts**. Political Analysis, academic.oup.com, cited by 2 (0.67 per year)*

*J Grainger, T Hannagan (2012) **Explaining word recognition, reading, the universe, and beyond: A modest proposal**. Behavioral and Brain Sciences, cambridge.org, cited by 3 (0.50 per year)*

*M Aubakirova, M Bansal (2016) **Interpreting Neural Networks to Improve Politeness Comprehension**. arXiv preprint arXiv:1610.02683, arxiv.org, cited by 2 (1.00 per year)*

*O Arandjelović (2012) **A new framework for interpreting the outcomes of imperfectly blinded controlled clinical trials**. PloS one, journals.plos.org, cited by*

*12 (2.00 per year)*

*JA Ogden, PB Lowry, KJ Petersen, ... (2008)* ***Explaining the Key Elements of Information Systems-Based Supply-Chain Strategy That Are Necessary for Business-to-Business Electronic Marketplace Survival****. Supply Chain Forum: An …, Taylor & Francis, cited by 1 (0.10 per year)*

*E Wu (2015)* ***Explaining data in visual analytic systems****., dspace.mit.edu*

*K Bulkeley (2017)* ***Explaining religious experiences like dreams****. Religion, Brain & Behavior, Taylor & Francis*

*RL Lewis (2016)* ***Are you thinking what I'm thinking? Explaining the relation between management control systems and managers' causal mental models****., opus.lib.uts.edu.au*

*I Mani, J Pustejovsky (2012)* ***Interpreting motion: Grounded representations for spatial language****., books.google.com, cited by 47 (7.83 per year)*

*RM McDonnell* ***Explaining the determinants of Foreign Policy voting behaviour in the Brazilian Houses of Legislature, with a focus on the Senate****. teses.usp.br*

*JY Sasaki, AS Cohen (2017)* ***Explaining agency detection within a domain-specific, culturally attuned model****. Religion, Brain & Behavior, Taylor & Francis*

*NM Moacdieh (2015)* ***Eye tracking: A promising means of tracing, explaining, and preventing the effects of display clutter in real time****., search.proquest.com*

*M Macha, L Akoglu (2017)* ***X-PACS: eXPlaining Anomalies by Characterizing Subspaces****. arXiv preprint arXiv:1708.05929, arxiv.org*

*D Calitoiu, BJ Oommen, D Nussbaum (2012)* ***Large-scale neuro-modeling for understanding and explaining some brain-related chaotic behavior****. Simulation, journals.sagepub.com, cited by 5 (0.83 per year)*

*Y Chen (2017)* ***Towards Explaining Neural Networks****., dspace.library.uu.nl*

*J Ye, G Stevenson, S Dobson, M O'Grady, ... (2013)* ***Perceiving and interpreting smart home datasets with\mathcal {PI}****. Journal of Ambient …, Springer*

*JTG Holmes (2003)* ***Learning by explaining: the effects of software agents as learning partners****., etd.library.vanderbilt.edu, cited by 1 (0.07 per year)*

*P Underwood (2016)* ***Conflict and Stability in the Neoliberal Era: Explaining Urban Unrest in Latin America****., digital.lib.washington.edu*

*S Äärilä (2017)* ***Species distribution models explaining human-wildlife conflicts in Taita Taveta County, Kenya****., dspace3.hulib.helsinki.fi*

*M Miłkowski (2013)* ***Explaining the computational mind****., books.google.com, cited by 98 (19.60* per year)*

*JM Whitacre (2011)* ***Survival of the flexible: explaining the recent popularity of nature-inspired optimization within a rapidly evolving world****. Computing, Springer, cited by 15 (2.14 per year)*

*J Pound (2013)* ***Interpreting and Answering Keyword Queries using Web Knowledge Bases****., uwspace.uwaterloo.ca*

*GH Chen* ***Explaining the Success of Nonparametric Inference: Forecasting Viral News, Recommending Products to People, and Finding Organs in Medical***

***Images***.

*F Lesaint (2014)* ***Modelling animal conditioning with factored representations in dual-learning: explaining inter-individual differences at behavioural and neurophysiological levels****., tel.archives-ouvertes.fr*

*JM Susskind (2011)* ***Interpreting faces with neurally inspired generative models****., tspace.library.utoronto.ca, cited by 3 (0.43 per year)*

*M Greene (2016)* ***Explaining the underlying psycological factors of consumer behaviour with artificial neural networks****., orca.cf.ac.uk*

*AC Acar (2008)* ***Query consolidation: Interpreting queries sent to independent heterogenous databases****., search.proquest.com*

**Query report 03 – “interpretable” AND “explainable”**

# *(intitle:interpretable OR intitle:explainable) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") to 2017*

*Publish or Perish 6.21.6145.6594*

## *Search terms*

***All of the words:*** *(intitle:interpretable OR intitle:explainable) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning")*
***Years:*** *earliest to 2017*

## *Data retrieval*

***Data source:*** *Google Scholar*
***Query date:*** *21/01/2018 13:15:56*
***Cache date:*** *21/01/2018 12:14:53*
***Query result:*** *[0] The operation completed successfully.*

## *Metrics*

***Publication years:*** *1987-2017*
***Citation years:*** *31 (1987-2018)*
***Papers:*** *267*
***Citations:*** *4711*
***Citations/year:*** *151.97*
***Citations/paper:*** *17.64 (*count=19)*
***Citations/author:*** *2350.51*
***Papers/author:*** *125.40*
***Authors/paper:*** *2.84/3.0/3 (mean/median/mode)*
***Age-weighed citation rate:*** *922.65 (sqrt=30.38), 415.69/author*
***Hirsch h-index:*** *29 (a=5.60, m=0.94, 3556 cites=75.5% coverage)*
***Egghe g-index:*** *65 (g/h=2.24, 4258 cites=90.4% coverage)*
***PoP hI,norm:*** *19*
***PoP hI,annual:*** *0.61*

## *Results*

*R Goodacre (2003)* ***Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules****. Vibrational Spectroscopy, Elsevier, cited by 79 (5.27 per year)*

*A Vellido, JD Martín-Guerrero, PJG Lisboa (2012)* ***Making machine learning models interpretable.****. ESANN, elen.ucl.ac.be, cited by 112 (18.67* per year)*

*C Rudin (2014)* ***Algorithms for interpretable machine learning****. Proceedings of the 20th ACM SIGKDD international …, dl.acm.org, cited by 18 (4.50 per year)*

*H Chen, L Carlsson, M Eriksson, ... (2013)* ***Beyond the scope of free-Wilson analysis: building interpretable QSAR models with machine learning algorithms****. Journal of chemical …, ACS Publications, cited by 23 (4.60 per year)*

*O Cordón (2011)* ***A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems****. International Journal of Approximate Reasoning, Elsevier, cited by 216 (30.86* per year)*

*S Guillaume, B Charnomordic (2011)* ***Learning interpretable fuzzy inference systems with FisPro****. Information Sciences, Elsevier, cited by 83 (11.86* per year)*

*W Xing, R Guo, E Petakovic, S Goggins (2015)* ***Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data …****. computers in Human Behavior, Elsevier, cited by 67 (22.33* per year)*

*D Nauck, R Kruse (1999)* ***Obtaining interpretable fuzzy classification rules from medical data****. Artificial intelligence in medicine, Elsevier, cited by 375 (19.74* per year)*

*BR Bakshi, A Koulouris, ... (1994)* ***Wave-Nets: novel learning techniques, and the induction of physically interpretable models****. SPIE, proceedings.spiedigitallibrary.org, cited by 34 (1.42 per year)*

*V Van Belle, P Lisboa (2013)* ***Research directions in interpretable machine learning models.****. ESANN, pdfs.semanticscholar.org, cited by 11 (2.20 per year)*

*JM Alonso, L Magdalena (2011)* ***Special issue on interpretable fuzzy systems****., Elsevier, cited by 68 (9.71 per year)*

*H Kim, WY Loh, YS Shih, P Chaudhuri (2007)* ***Visualizable and interpretable regression models with good prediction power****. IIE Transactions, Taylor & Francis, cited by 45 (4.09 per year)*

*JM Alonso, L Magdalena, ... (2010)* ***Embedding HILK in a three-objective evolutionary algorithm with the aim of modeling highly interpretable fuzzy rule-based classifiers****. Genetic and Evolutionary …, ieeexplore.ieee.org, cited by 27 (3.38 per year)*

*MMC Vidovic, N Görnitz, KR Müller, G Rätsch, ... (2015)* ***Opening the black box: Revealing interpretable sequence motifs in kernel-based learning algorithms****. … on Machine Learning …, Springer, cited by 5 (1.67 per year)*

*D Gunning (2017)* ***Explainable artificial intelligence (xai)****. Defense Advanced Research Projects Agency (DARPA …, cc.gatech.edu, cited by 24 (24.00* per year)*

*HP Oliveira, A Magalhaes, MJ Cardoso, ... (2010)* ***An accurate and interpretable model for BCCT. core****. ... in Medicine and …, ieeexplore.ieee.org, cited by 24 (3.00 per year)*

*D Hofmann, FM Schleif, B Paaßen, B Hammer (2014)* ***Learning interpretable kernelized prototype-based models****. Neurocomputing, Elsevier, cited by 17 (4.25* *per year)*

*B Letham, C Rudin, TH McCormick, D Madigan (2012)* ***Building interpretable classifiers with rules using Bayesian analysis****. Department of Statistics Technical …, cited by 28 (4.67 per year)*

*P Pulkkinen, H Koivisto (2007)* ***Identification of interpretable and accurate fuzzy classifiers and function estimators with hybrid methods****. Applied Soft Computing, Elsevier, cited by 28 (2.55 per year)*

*SM Zhou, JQ Gan (2008)* ***Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling****. Fuzzy Sets and Systems, Elsevier, cited by 204 (20.40* per year)*

*Y Ye, L Chen, D Wang, T Li, Q Jiang, ... (2009)* ***SBMDS: an interpretable string based malware detection system using SVM ensemble with bagging****. Journal in computer …, Springer, cited by 45 (5.00 per year)*

*B Kim (2015)* ***Interactive and interpretable machine learning models for human machine collaboration****., dspace.mit.edu, cited by 13 (4.33 per year)*

*A Fyshe, L Wehbe, PP Talukdar, B Murphy, TM Mitchell (2015)* ***A Compositional and Interpretable Semantic Space.****. HLT-NAACL, cited by 28 (9.33 per year)*

*Z Wang, V Palade (2011)* ***Building interpretable fuzzy models for high dimensional data analysis in cancer diagnosis****. BMC genomics, bmcgenomics.biomedcentral.com, cited by 18 (2.57 per year)*

*JM Alonso, C Castiello, M Lucarelli, ... (2013)* ***Modeling interpretable fuzzy rule-based classifiers for medical decision support****. Data Mining: Concepts …, igi-global.com, cited by 13 (2.60 per year)*

*R Cannone, JM Alonso, ... (2011)* ***Multi-objective design of highly interpretable fuzzy rule-based classifiers with semantic cointension****. Genetic and Evolutionary …, ieeexplore.ieee.org, cited by 16 (2.29 per year)*

*M Bohanec, M Robnik-Šikonja, ... (2017)* ***Decision-making framework with double-loop learning through interpretable black-box machine learning models****. … Management & Data …, emeraldinsight.com, cited by 2 (2.00 per year)*

*T Wang, C Rudin, F Doshi, Y Liu, E Klampfl, ... (2015)* ***Bayesian Or's of And's for interpretable classification with application to context aware recommender systems****., pdfs.semanticscholar.org, cited by 12 (4.00 per year)*

*J Zeng, B Ustun, C Rudin (2017)* ***Interpretable classification models for recidivism prediction****. Journal of the Royal Statistical …, Wiley Online Library, cited by 28 (28.00* per year)*

*PE Dolgirev, IA Kruglov, AR Oganov (2016)* ***Machine learning scheme for fast extraction of chemically interpretable interatomic potentials****. AIP Advances, aip.scitation.org, cited by 9 (4.50 per year)*

*I Gadaras, L Mikhailov (2009)* ***An interpretable fuzzy rule-based classification methodology for medical diagnosis****. Artificial intelligence in medicine, Elsevier, cited by 92 (10.22* per year)*

*I Sturm, S Lapuschkin, W Samek, KR Müller (2016)* ***Interpretable deep neural networks for single-trial EEG classification****. Journal of neuroscience …, Elsevier, cited by 31 (15.50* per year)*

*C Otte (2013)* ***Safe and interpretable machine learning: a methodological review****. Computational intelligence in intelligent data analysis, Springer, cited by 10 (2.00 per year)*

*Y Jin, B Sendhoff, E Körner (2006)* ***Simultaneous generation of accurate and interpretable neural network classifiers****. Multi-objective machine learning, Springer, cited by 20 (1.67 per year)*

*A Emad, KR Varshney, ... (2015)* ***A semiquantitative group testing approach for learning interpretable clinical prediction rules****. Proc. Signal Process ..., pdfs.semanticscholar.org, cited by 6 (2.00 per year)*

*SY Wong, KS Yap, HJ Yap, SC Tan, ... (2015)* ***On equivalence of FIS and ELM for interpretable rule-based knowledge representation****. IEEE transactions on ..., ieeexplore.ieee.org, cited by 21 (7.00 per year)*

*E Choi, MT Bahadori, J Sun, J Kulas, ... (2016)* ***Retain: An interpretable predictive model for healthcare using reverse time attention mechanism****. Advances in Neural ..., papers.nips.cc, cited by 34 (17.00* per year)*

*D Yu, H Shen, J Yang (2011)* ***SOMRuler: a novel interpretable transmembrane helices predictor****. IEEE transactions on nanobioscience, ieeexplore.ieee.org, cited by 16 (2.29 per year)*

*H Liu, M Cocea (2017)* ***Fuzzy rule based systems for interpretable sentiment analysis****. Advanced Computational Intelligence (ICACI) ..., ieeexplore.ieee.org, cited by 8 (8.00 per year)*

*L Arras, F Horn, G Montavon, KR Müller, W Samek (2017)* ***" What is relevant in a text document?": An interpretable machine learning approach****. PloS one, journals.plos.org, cited by 4 (4.00 per year)*

*HR Marateb, S Goudarzi (2015)* ***A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system****. Journal of research in medical sciences ..., ncbi.nlm.nih.gov, cited by 19 (6.33 per year)*

*H Luo, Z Liu, H Luan, M Sun (2015)* ***Online learning of interpretable word embeddings****. Proceedings of the 2015 Conference on ..., aclweb.org, cited by 11 (3.67 per year)*

*D Arp, M Spreitzenbarth, M Hubner, H Gascon, K Rieck, ... (2014)* ***DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket.****. NDSS, researchgate.net, cited by 526 (131.50* per year)*

*Q Xu, Q Zhao, W Pei, L Yang, ... (2004)* ***Interpretable neural network tree for continuous-feature data sets****. ... Processing-Letters and ..., pdfs.semanticscholar.org, cited by 8 (0.57 per year)*

*L Song, P Langfelder, S Horvath (2013)* ***Random generalized linear model: a highly accurate and interpretable ensemble predictor****. BMC ..., bmcbioinformatics.biomedcentral ..., cited by 38 (7.60 per year)*

*H Wang, S Kwong, Y Jin, W Wei, KF Man (2005)* ***Multi-objective hierarchical genetic algorithm for interpretable fuzzy rule-based knowledge extraction****. Fuzzy sets and systems, Elsevier, cited by 197 (15.15* per year)*

*MF Ghalwash, V Radosavljevic, ... (2013)* ***Extraction of interpretable multivariate***

***patterns for early diagnostics**. Data Mining (ICDM) ..., ieeexplore.ieee.org, cited by 42 (8.40 per year)*

*J Jiakel, L Grioll, R Mikut (2000) **Automatic generation and evaluation of interpretable fuzzy rules**. New Frontiers in Computational ..., books.google.com, cited by 8 (0.44 per year)*

*W Samek, T Wiegand, KR Müller (2017) **Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models**. arXiv preprint arXiv:1708.08296, arxiv.org, cited by 3 (3.00 per year)*

*W Pedrycz, M Reformat, K Li (2006) **OR/AND neurons and the development of interpretable logic models**. IEEE Transactions on neural ..., ieeexplore.ieee.org, cited by 18 (1.50 per year)*

*Y Zhang, G Lai, M Zhang, Y Zhang, Y Liu, ... (2014) **Explicit factor models for explainable recommendation based on phrase-level sentiment analysis**. Proceedings of the 37th ..., dl.acm.org, cited by 124 (31.00* per year)*

*M Hapke, M Komosinski (2008) **Evolutionary design of interpretable fuzzy controllers**. Foundation of Computing and ..., pdfs.semanticscholar.org, cited by 11 (1.10 per year)*

*L Obermann, S Waack (2015) **Demonstrating non-inferiority of easy interpretable methods for insolvency prediction**. Expert Systems with Applications, Elsevier, cited by 6 (2.00 per year)*

*S Wachter, B Mittelstadt, L Floridi (2017) **Transparent, explainable, and accountable AI for robotics**., philarchive.org, cited by 4 (4.00 per year)*

*PK Shukla, SP Tripathi (2012) **On the design of interpretable evolutionary fuzzy systems (I-EFS) with improved accuracy**. Computing Sciences (ICCS), 2012 ..., ieeexplore.ieee.org, cited by 6 (1.00 per year)*

*S Briesemeister (2012) **Interpretable machine learning approaches in computational biology**., bibliographie.uni-tuebingen.de, cited by 3 (0.50 per year)*

*M Eliasson, S Rannar, J Trygg (2011) **From data processing to multivariate validation-essential steps in extracting interpretable information from metabolomics data**. Current pharmaceutical ..., ingentaconnect.com, cited by 30 (4.29 per year)*

*T Wang, C Rudin, F Doshi-Velez, Y Liu, ... (2017) **A bayesian framework for learning rule sets for interpretable classification**. ... of Machine Learning ..., jmlr.org, cited by 3 (3.00 per year)*

*D Hein, A Hentschel, T Runkler, S Udluft (2017) **Particle swarm optimization for generating interpretable fuzzy reinforcement learning policies**. Engineering Applications of ..., Elsevier, cited by 3 (3.00 per year)*

*Y Jin, B Sendhoff (2003) **Extracting interpretable fuzzy rules from RBF networks**. Neural Processing Letters, Springer, cited by 80 (5.33 per year)*

*S Liu, S Dissanayake, S Patel, ... (2014) **Learning accurate and interpretable models based on regularized random forests regression**. BMC systems ..., bmcsystbiol.biomedcentral.com, cited by 6 (1.50 per year)*

*KC Sim (2015) **On constructing and analysing an interpretable brain model for the DNN based on hidden activity patterns**. ... Recognition and Understanding*

*(ASRU), 2015 IEEE …, ieeexplore.ieee.org, cited by 6 (2.00 per year)*

*A Cano, A Zafra, SN Ventura (2013)* ***An interpretable classification rule mining algorithm****. Information Sciences, Elsevier, cited by 21 (4.20 per year)*

*MJ Alhaddad, A Mohammed, M Kamel, H Hagras (2015)* ***A genetic interval type-2 fuzzy logic-based approach for generating interpretable linguistic models for the brain P300 phenomena recorded via brain–computer …****. Soft Computing, Springer, cited by 5 (1.67 per year)*

*SM Zhou, JQ Gan (2009)* ***Extracting Takagi-Sugeno fuzzy rules with interpretable submodels via regularization of linguistic modifiers****. IEEE Transactions on Knowledge and Data …, ieeexplore.ieee.org, cited by 32 (3.56 per year)*

*A Mehrotra, R Hendley, M Musolesi (2017)* ***Interpretable machine learning for mobile notification management: An overview of prefminer****. GetMobile: Mobile Computing and …, dl.acm.org, cited by 1 (1.00 per year)*

*S Bouktif, EM Hanna, N Zaki, EA Khousa (2014)* ***Ant Colony Optimization Algorithm for Interpretable Bayesian Classifiers Combination: Application to Medical Predictions****. PloS one, journals.plos.org, cited by 5 (1.25 per year)*

*T Wang, C Rudin, F Doshi-Velez, Y Liu, ... (2015)* ***Or's of and's for interpretable classification, with application to context-aware recommender systems****. arXiv preprint arXiv …, arxiv.org, cited by 6 (2.00 per year)*

*L Obermann (2016)* ***Interpretable Binary and Multiclass Prediction Models for Insolvencies and Credit Ratings****., d-nb.info, cited by 1 (0.50 per year)*

*MB Gorzałczany, F Rudziński (2017)* ***Interpretable and accurate medical data classification–a multi-objective genetic-fuzzy optimization approach****. Expert Systems with Applications, Elsevier, cited by 5 (5.00 per year)*

*A Shrikumar, P Greenside, A Shcherbina, A Kundaje (2016)* ***Not Just A Black Box: Interpretable Deep Learning by Propagating Activation Differences****., pdfs.semanticscholar.org, cited by 4 (2.00 per year)*

*S Guillaume, B Charnomordic (2010)* ***Interpretable fuzzy inference systems for cooperation of expert knowledge and data in agricultural applications using fispro****. Fuzzy Systems (FUZZ), 2010 …, ieeexplore.ieee.org, cited by 9 (1.13 per year)*

*C Mencar, A Consiglio, ... (2007)* ***DCγ : Interpretable Granulation of Data through GA-based Double Clustering****. Fuzzy Systems Conference …, ieeexplore.ieee.org, cited by 4 (0.36 per year)*

*S Fischer, H Bunke (2002)* ***Automatic identification of diatoms using visual human-interpretable features****. International Journal of Image and Graphics, World Scientific, cited by 5 (0.31 per year)*

*R Fong, A Vedaldi (2017)* ***Interpretable Explanations of Black Boxes by Meaningful Perturbation****. arXiv preprint arXiv:1704.03296, arxiv.org, cited by 14 (14.00* per year)*

*M Sapir, S Sherman (2003)* ***A toolkit for the search of the most general interpretable hypotheses****. Integration of Knowledge Intensive Multi …, ieeexplore.ieee.org, cited by 2 (0.13 per year)*

*P Bermejo, A Vivo, PJ Tárraga, ... (2015)* ***Development of interpretable predictive***

***models for BPH and prostate cancer**. Clinical Medicine …, ncbi.nlm.nih.gov, cited by 3 (1.00 per year)*

*P Hartono, S Hashimoto (2007) **An interpretable neural network ensemble**. … 2007. 33rd Annual Conference of the …, ieeexplore.ieee.org, cited by 2 (0.18 per year)*

*J Kim, J Canny (2017) **Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention**. arXiv preprint arXiv:1703.10631, arxiv.org, cited by 6 (6.00 per year)*

*Z Che, S Purushotham, R Khemani, ... (2016) **Interpretable deep models for icu outcome prediction**. AMIA Annual Symposium …, ncbi.nlm.nih.gov, cited by 6 (3.00 per year)*

*P Ponte, RG Melko (2017) **Kernel methods for interpretable machine learning of order parameters**. arXiv preprint arXiv:1704.05848, arxiv.org, cited by 7 (7.00 per year)*

*D Breuker, P Delfmann, M Matzner, J Becker (2014) **Designing and evaluating an interpretable predictive modeling technique for business processes**. International Conference on …, Springer, cited by 5 (1.25 per year)*

*H Ren, W Xu, Y Yan (2015) **Optimizing human-interpretable dialog management policy using genetic algorithm**. Automatic Speech Recognition and …, ieeexplore.ieee.org, cited by 2 (0.67 per year)*

*J Bien, R Tibshirani (2011) **Prototype selection for interpretable classification**. The Annals of Applied Statistics, JSTOR, cited by 41 (5.86 per year)*

*N Van Linh, NK Anh, K Than, NN Tat (2015) **Effective and interpretable document classification using distinctly labeled Dirichlet process mixture models of von Mises-Fisher distributions**. International Conference on …, Springer, cited by 2 (0.67 per year)*

*Z Wang, V Palade (2010) **Multi-objective evolutionary algorithms based interpretable fuzzy models for microarray gene expression data analysis**. Bioinformatics and Biomedicine (BIBM) …, ieeexplore.ieee.org, cited by 2 (0.25 per year)*

*JM Alonso, L Magdalena, S Guillaume, ... (2008) **Designing highly interpretable fuzzy rule-based systems with integration of expert and induced knowledge**. Proceedings of …, gimac.uma.es, cited by 2 (0.20 per year)*

*D Doran, S Schulz, TR Besold (2017) **What Does Explainable AI Really Mean? A New Conceptualization of Perspectives**. arXiv preprint arXiv:1710.00794, arxiv.org, cited by 1 (1.00 per year)*

*W Landecker (2014) **Interpretable machine learning and sparse coding for computer vision**., search.proquest.com, cited by 1 (0.25 per year)*

*S Mousavi, A Esfahanipour, MHF Zarandi (2015) **MGP-INTACTSKY: Multitree Genetic Programming-based learning of INTerpretable and ACcurate TSK sYstems for dynamic portfolio trading**. Applied Soft Computing, Elsevier, cited by 6 (2.00 per year)*

*M Kirmse, U Petersohn (2011) **Large margin rectangle learning an alternative way to learn interpretable and representative models**. Soft Computing and*

*Pattern …, ieeexplore.ieee.org, cited by 1 (0.14 per year)*

*T Weijters, A Van Den Bosch (1998)* ***Interpretable neural networks with BP-SOM****. Tasks and Methods in Applied Artificial …, Springer, cited by 11 (0.55 per year)*

*G Sánchez, F Jiménez, JF Sánchez, JM Alcaraz (2009)* ***A Multi-objective Neuro-evolutionary Algorithm to Obtain Interpretable Fuzzy Models.****. CAEPIA, Springer, cited by 3 (0.33 per year)*

*Z Deng, L Cao, Y Jiang, S Wang (2015)* ***Minimax probability TSK fuzzy system classifier: a more transparent and highly interpretable classification model****. IEEE Transactions on Fuzzy …, ieeexplore.ieee.org, cited by 23 (7.67 per year)*

*T Tran, W Luo, D Phung, J Morris, K Rickard, ... (2016)* ***Preterm birth prediction: Deriving stable and interpretable rules from high dimensional data****. ... on Machine Learning in …, jmlr.org, cited by 1 (0.50 per year)*

*S Wang, Z Chen, L Zhang, Q Yan, ... (2016)* ***TrafficAV: An effective and explainable detection of mobile malware behavior using network traffic****. Quality of Service …, ieeexplore.ieee.org, cited by 4 (2.00 per year)*

*T Amorgianiotis, K Theofilatos, S Mitra, ... (2014)* ***Integrating High Volume Financial Datasets to Achieve Profitable and Interpretable Short Term Trading with the FTSE100 Index****. ... Conference on Artificial …, Springer, cited by 1 (0.25 per year)*

*T Miller, P Howe, L Sonenberg (2017)* ***Explainable AI: Beware of Inmates Running the Asylum****. ... -17 Workshop on Explainable ..., intelligentrobots.org, cited by 1 (1.00 per year)*

*SM Zhou, J Gan (2006)* ***Multiple Objective Learning for Constructing Interpretable Takagi-Sugeno Fuzzy Model****. Multi-Objective Machine Learning, Springer, cited by 1 (0.08 per year)*

*N Huang, YJ Oyang (2014)* ***Microbial abundance patterns of host obesity inferred by the structural incorporation of association measures into interpretable classifiers****. Bioinformatics and Biomedicine (BIBM) …, ieeexplore.ieee.org, cited by 1 (0.25 per year)*

*H Sun, K Nguyen, E Kerns, Z Yan, KR Yu, ... (2017)* ***Highly predictive and interpretable models for PAMPA permeability****. Bioorganic & medicinal …, Elsevier, cited by 2 (2.00 per year)*

*T Rabenoro, J Lacaille, M Cottrell, ... (2014)* ***Interpretable Aircraft Engine Diagnostic via Expert Indicator Aggregation****. ... on Machine Learning …, hal.archives-ouvertes.fr, cited by 1 (0.25 per year)*

*D Liu, F Peng, A Shea, R Picard (2017)* ***DeepFaceLIFT: Interpretable Personalized Models for Automatic Estimation of Self-Reported Pain****. arXiv preprint arXiv:1708.04670, arxiv.org, cited by 1 (1.00 per year)*

*Y Li, J Song, S Ermon (2017)* ***Infogail: Interpretable imitation learning from visual demonstrations****. Advances in Neural Information Processing …, papers.nips.cc, cited by 1 (1.00 per year)*

*S Milli, P Abbeel, I Mordatch (2017)* ***Interpretable and Pedagogical Examples****. arXiv preprint arXiv:1711.00694, arxiv.org, cited by 1 (1.00 per year)*

*Z Ren, S Liang, P Li, S Wang, M de Rijke (2017)* ***Social collaborative viewpoint***

***regression with explainable recommendations**. Proceedings of the Tenth …, dl.acm.org, cited by 8 (8.00 per year)*

*Q Yong, X Zong-yi, J Li-min, W Ying-ying (2009) **Study on interpretable fuzzy classification system based on neural networks**. ICCAS-SICE, 2009, ieeexplore.ieee.org, cited by 1 (0.11 per year)*

*H Wang, S Kwong, Y Jin, W Wei, ... (2005) **Agent-based evolutionary approach for interpretable rule-based knowledge extraction**. IEEE Transactions on …, ieeexplore.ieee.org, cited by 93 (7.15 per year)*

*X Li **Transparent and Interpretable Machine Learning**. ittc.ku.edu*

*NM Alexandrov (2017) **Explainable AI Decisions for Human-Autonomy Interactions**. 17th AIAA Aviation Technology, Integration, and …, arc.aiaa.org*

*S Ghosal, D Blystone, AK Singh, ... (2017) **Interpretable Deep Learning applied to Plant Stress Phenotyping**. arXiv preprint arXiv …, arxiv.org*

*MP Bonacina (2017) **Automated Reasoning for Explainable Artificial Intelligence**., pdfs.semanticscholar.org*

*H Lakkaraju, E Kamar, R Caruana, ... (2017) **Interpretable & Explorable Approximations of Black Box Models**. arXiv preprint arXiv …, arxiv.org*

*A Dhurandhar, S Oh, M Petrik (2016) **Building an Interpretable Recommender via Loss-Preserving Transformation**. arXiv preprint arXiv:1606.05819, arxiv.org*

*T Tran, W Luo, D Phung, J Morris, ... (2016) **Preterm Birth Prediction: Stable Selection of Interpretable Rules from High Dimensional Data**. Machine Learning …, proceedings.mlr.press*

*HF Tan, G Hooker, MT Wells (2016) **Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable**. arXiv preprint arXiv:1611.07115, arxiv.org*

*D Doran (2017) **“Explainable (?)” Statistical ML**., materials.dagstuhl.de*

*A Emad, KR Varshney, DM Malioutov **Learning Interpretable Clinical Prediction Rules using Threshold Group Testing**. pdfs.semanticscholar.org*

*S Khenissi, B Abdollahi, W Sun, P Sagheb, O Nasraoui (2017) **New Explainable Active Learning Approach for Recommender Systems**., uknowledge.uky.edu*

*DM Malioutov, KR Varshney, A Emad, ... (2017) **Learning Interpretable Classification Rules with Boolean Compressed Sensing**. Transparent Data Mining …, Springer*

*KR Varshney (2016) **Interpretable machine learning via convex cardinal shape composition**. … , Control, and Computing (Allerton), 2016 54th …, ieeexplore.ieee.org*

*B Abdollahi, O Nasraoui (2016) **Explainable Restricted Boltzmann Machines for Collaborative Filtering**. arXiv preprint arXiv:1606.07129, arxiv.org, cited by 6 (3.00 per year)*

*R Ramakrishnan, K Narasimhan, J Shah **Interpretable Transfer for Reinforcement Learning based on Object Similarities**. sites.google.com*

*YF Luo, A Rumshisky (2016) **Interpretable Topic Features for Post-ICU Mortality***

***Prediction**. AMIA Annual Symposium Proceedings, ncbi.nlm.nih.gov, cited by 1 (0.50 per year)*

*V Krakovna, J Du, JS Liu (2015) **Interpretable Selection and Visualization of Features and Interactions Using Bayesian Forests**. arXiv preprint arXiv:1506.02371, arxiv.org*

*T Wu, X Li, X Song, W Sun, L Dong, B Li (2017) **Interpretable R-CNN**. arXiv preprint arXiv:1711.05226, arxiv.org*

*J Zuallaert, M Kim, Y Saeys, W De Neve (2017) **Interpretable convolutional neural networks for effective translation initiation site prediction**. arXiv preprint arXiv:1711.09558, arxiv.org*

*K Sharif (2017) **Building and Evaluating Interpretable Models using Symbolic Regression and Generalized Additive Models**., openreview.net*

*J Clos, N Wiratunga (2017) **Lexicon Induction for Interpretable Text Classification**. International Conference on Theory and Practice of …, Springer*

*SG Kim, N Theera-Ampornpunt, ... (2016) **Opening up the blackbox: an interpretable deep neural network-based classifier for cell-type specific enhancer predictions**. BMC systems …, bmcsystbiol.biomedcentral.com, cited by 3 (1.50 per year)*

*TV Vishnu, N Gugulothu, P Malhotra, L Vig, P Agarwal, ... **Bayesian Networks for Interpretable Health Monitoring of Complex Systems**. zurich.ibm.com*

*LTK Phung, VTN Chau, ... (2015) **Extracting Rule RF in Educational Data Classification: From a Random Forest to Interpretable Refined Rules**. Advanced Computing and …, ieeexplore.ieee.org, cited by 1 (0.33 per year)*

*CH Hsieh, WC Chao, PW Liu, ... (2015) **Cyber security risk assessment using an interpretable evolutionary fuzzy scoring system**. ... Technology (ICCST), 2015 …, ieeexplore.ieee.org*

*H Liu, M Cocea (2017) **Fuzzy information granulation towards interpretable sentiment analysis**. Granular Computing, Springer, cited by 6 (6.00 per year)*

*T Kenesei, J Abonyi (2012) **Interpretable support vector regression**. Artificial Intelligence Research, sciedu.ca*

*GB Goh, NO Hodas, C Siegel, A Vishnu (2017) **SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties**. arXiv preprint arXiv:1712.02034, arxiv.org*

*B Du Boulay, D Hogg, L Steel (1987) **EXPLAINABLE KNOWLEDGE PRODUCTION**. ... ECAI-86, Brighton, UK, July 20 …, Elsevier Science Ltd*

*J Clos, N Wiratunga, S Massie **Towards Explainable Text Classification by Jointly Learning Lexicon and Modifier Terms**. IJCAI-17 Workshop on Explainable AI (XAI), earthlink.net*

*R Ramakrishnan, J Shah (2016) **Towards Interpretable Explanations for Transfer Learning in Sequential Tasks**. 2016 AAAI Spring Symposium Series, aaai.org, cited by 1 (0.50 per year)*

*P Ferreira, I Dutra, R Salvini, ... (2016) **Interpretable models to predict Breast Cancer**. ... and Biomedicine (BIBM) ..., ieeexplore.ieee.org, cited by 1 (0.50 per* *year)*

*N Puri, P Gupta, P Agarwal, S Verma, ... (2017)* ***MAGIX: Model Agnostic Globally Interpretable Explanations****. arXiv preprint arXiv …, arxiv.org*

*P Langley, B Meadows, M Sridharan, D Choi (2017)* ***Explainable Agency for Intelligent Autonomous Systems.****. AAAI, aaai.org, cited by 5 (5.00 per year)*

*S Nagrecha, JZ Dillon, NV Chawla (2017)* ***MOOC Dropout Prediction: Lessons Learned from Making Pipelines Interpretable****. Proceedings of the 26th …, dl.acm.org, cited by 4 (4.00 per year)*

*G Bhanot, M Biehl, T Villmann, D Zühlke* ***Biomedical data analysis in translational research: Integration of expert knowledge and interpretable models****. rug.nl*

*S Mishra, BL Sturm, S Dixon* ***LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS FOR MUSIC CONTENT ANALYSIS****. pdfs.semanticscholar.org*

*M Pereira-Fariña, C Reed (2017)* ***Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017)****. ... of the 1st Workshop on Explainable …, aclweb.org*

*C Otte (2014)* ***Interpretable semi-parametric regression models with defined error bounds****. Neurocomputing, Elsevier*

*U Johansson, C Sönströd, ... (2014)* ***Accurate and interpretable regression trees using oracle coaching****. ... Intelligence and Data …, ieeexplore.ieee.org*

*LS Whitmore, A George, CM Hudson (2016)* ***Mapping chemical performance on molecular structures using locally interpretable explanations****. arXiv preprint arXiv:1611.07443, arxiv.org*

*SCH Yang, P Shafto* ***Explainable Artificial Intelligence via Bayesian Teaching****. shaftolab.com*

*J Clos, N Wiratunga (2017)* ***Neural Induction of a Lexicon for Fast and Interpretable Stance Classification****. International Conference on Language, Data and …, Springer*

*TS Kim, A Reiter (2017)* ***Interpretable 3D Human Action Analysis with Temporal Convolutional Networks****. arXiv preprint arXiv:1704.04516, arxiv.org, cited by 5 (5.00 per year)*

*P Urbanke, A Uhlig, JJ Kranz (2017)* ***A Customized and Interpretable Deep Neural Network for High-Dimensional Business Data-Evidence from an E-Commerce Application****., aisel.aisnet.org*

*B Dolan, K Ocke, E Gross, ... (2015)* ***Interpretable Classifier for Identifying High-Value Child Support Cases****. Machine Learning and …, ieeexplore.ieee.org*

*A Holzinger, C Biemann, CS Pattichis, ... (2017)* ***What do we need to build explainable AI systems for the medical domain?****. arXiv preprint arXiv …, arxiv.org*

*YAO MING (2017)* ***A SURVEY ON VISUALIZATION FOR EXPLAINABLE CLASSIFIERS****., pdfs.semanticscholar.org*

*SE Sorour, T Mine (2016)* ***Building an Interpretable Model of Predicting Student***

***Performance Using Comment Data Mining**. ... Applied Informatics (IIAI-AAI), 2016 5th IIAI …, ieeexplore.ieee.org, cited by 1 (0.50 per year)*

*J de la Torre, A Valls, D Puig (2017) **A Deep Learning Interpretable Classifier for Diabetic Retinopathy Disease Grading**. arXiv preprint arXiv:1712.08107, arxiv.org*

*WY Sit, KZ Mao (2012) **A cognitively inspired rule-plus-exemplar framework for interpretable pattern classification**. Information Fusion (FUSION), 2012 15th …, ieeexplore.ieee.org*

*CH Hsieh, YS Shen, CW Li, ... (2015) **iF2: An Interpretable Fuzzy Rule Filter for Web Log Post-Compromised Malicious Activity Monitoring**. ... Security (AsiaJCIS), 2015 …, ieeexplore.ieee.org*

*O Kuzelka, J Davis, S Schockaert (2016) **Stratified Knowledge Bases as Interpretable Probabilistic Models**. arXiv preprint arXiv:1611.06174, arxiv.org*

*J Hou, TS Kim, A Reiter (2017) **Train, Diagnose and Fix: Interpretable Approach for Fine-grained Action Recognition**. arXiv preprint arXiv:1711.08502, arxiv.org*

*S Wisdom, T Powers, J Pitton, L Atlas (2016) **Interpretable Recurrent Neural Networks Using Sequential Sparse Recovery**. arXiv preprint arXiv:1611.07252, arxiv.org, cited by 1 (0.50 per year)*

*V AUTOENCODERS **Interpretable Classification via Supervised Variational Autoencoders and Differentiable Decision Trees**. pdfs.semanticscholar.org*

*T Miller, P Howe, L Sonenberg (2017) **Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences**. arXiv preprint arXiv:1712.00547, arxiv.org*

*A Holzinger, B Malle, P Kieseberg, PM Roth, ... (2017) **Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology**. arXiv preprint arXiv …, arxiv.org*

*B Baron, M Musolesi (2017) **Interpretable Machine Learning for Privacy-Preserving IoT and Pervasive Systems**. arXiv preprint arXiv:1710.08464, arxiv.org*

*M Tu, V Berisha, J Liss (2017) **Interpretable Objective Assessment of Dysarthric Speech based on Deep Neural Networks**. Proc. Interspeech 2017, pdfs.semanticscholar.org*

*P Hartono (2007) **Interpretable Piecewise Linear Classifier**. International Conference on Neural Information …, Springer*

*E Santana, JC Principe (2016) **Perception Updating Networks: On architectural constraints for interpretable video generative models**., openreview.net*

*R Sheh, I Monteath **Introspectively Assessing Failures through Explainable Artificial Intelligence**. aass.oru.se*

*L Obermann, S Waack (2016) **Interpretable Multiclass Models for Corporate Credit Rating Capable of Expressing Doubt**., goedoc.uni-goettingen.de, cited by 1 (0.50 per year)*

*K Cpałka (2017) **Design of Interpretable Fuzzy Systems**. Studies in Computational Intelligence, Springer, cited by 17 (17.00* per year)*

*D Hein, S Udluft, TA Runkler (2017)* ***Interpretable Policies for Reinforcement Learning by Genetic Programming****. arXiv preprint arXiv:1712.04170, arxiv.org*

*F Jiménez, R Jódar, MP Martín, G Sánchez, ... (2017)* ***Unsupervised feature selection for interpretable classification in behavioral assessment of children****. Expert …, Wiley Online Library, cited by 3 (3.00 per year)*

*T Rabenoro, J Lacaille, M Cottrell, F Rossi (2015)* ***Interpretable Aircraft Engine Diagnostic via Expert Indicator Aggregation****. arXiv preprint arXiv …, arxiv.org*

*S Bouktif, EM Hanna, N Zaki, EA Khousa (2014)* ***Ant Colony Optimization Algorithm for Interpretable Bayesian Classifiers Combination: Application to****., academia.edu*

*W Pedrycz, K Hirota (2007)* ***Uninorm-based logic neurons as adaptive and interpretable processing constructs****. Soft Computing-A Fusion of Foundations …, Springer, cited by 16 (1.45 per year)*

*S Bouktif, EM Hanna, NZEA Khousa* ***Ant Colony Optimization Algorithm for Interpretable Bayesian Classifiers Combination: Application to Heart Disease Prediction****.*

*B Ustun, S Traca, C Rudin (2013)* ***Supersparse linear integer models for interpretable classification****. arXiv preprint arXiv:1306.6677, arxiv.org, cited by 13 (2.60 per year)*

*M Jovanovic, S Radovanovic, M Vukicevic, ... (2016)* ***Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression****. Artificial intelligence in …, Elsevier, cited by 4 (2.00 per year)*

*J Kajornrit, KW Wong, CC Fung (2016)* ***An interpretable fuzzy monthly rainfall spatial interpolation system for the construction of aerial rainfall maps****. Soft Computing, Springer, cited by 1 (0.50 per year)*

*M Last, G Danon, S Biderman, E Miron (2009)* ***Optimizing a batch manufacturing process through interpretable data mining models****. Journal of Intelligent …, Springer, cited by 7 (0.78 per year)*

*T Ito, H Sakaji, K Izumi, K Tsubouchi, ... (2017)* ***Development of an Interpretable Neural Network Model for Creation of Polarity Concept Dictionaries****. 2017 IEEE International …, IEEE*

*WH Dempsey, A Moreno, CK Scott, ... (2017)* ***iSurvive: An Interpretable, Event-time Prediction Model for mHealth****. … Machine Learning, proceedings.mlr.press*

*M Lee, D Mimno (2017)* ***Low-dimensional embeddings for interpretable anchor-based topic inference****. arXiv preprint arXiv:1711.06826, arxiv.org, cited by 22 (22.00* per year)*

*V Krakovna (2016)* ***Building Interpretable Models: From Bayesian Networks to Neural Networks****., dash.harvard.edu*

*B Ustun, C Rudin (2014)* ***Methods and models for interpretable linear classification****. arXiv preprint arXiv:1405.4047, arxiv.org, cited by 21 (5.25 per year)*

*N Van Linh, NK Anh, K Than, CN Dang (2017)* ***An effective and interpretable method for document classification****. Knowledge and Information …, Springer, cited by 4 (4.00 per year)*

*H Wang, S Kwong, Y Jin, CH Tsang (2006)* ***Agent based multi-objective approach to generating interpretable fuzzy systems****. Multi-Objective Machine Learning, Springer, cited by 3 (0.25 per year)*

*G Castellano, AM Fanelli, C Mencar, ... (2006)* ***Classifying data with interpretable fuzzy granulation****. SCIS & ISIS SCIS & …, jstage.jst.go.jp, cited by 12 (1.00 per year)*

*R Meyer, S O'keefe (2013)* ***A fuzzy binary neural network for interpretable classifications****. Neurocomputing, Elsevier, cited by 1 (0.20 per year)*

*S Destercke, S Guillaume, B Charnomordic (2007)* ***Building an interpretable fuzzy rule base from data using orthogonal least squares—application to a depollution problem****. Fuzzy Sets and Systems, Elsevier, cited by 35 (3.18 per year)*

*J Chen, M Mahfouf (2010)* ***Interpretable fuzzy modeling using multi-objective immune-inspired optimization algorithms****. Fuzzy Systems (FUZZ), 2010 IEEE …, ieeexplore.ieee.org, cited by 1 (0.13 per year)*

*KC Chatzidimitriou (2006)* ***Robuts and Interpretable Statistical Models for Predicting the Intensification of Tropical Cyclones****., pdfs.semanticscholar.org, cited by 2 (0.17 per year)*

*N Pappas (2016)* ***Learning Explainable User Sentiment and Preferences for Information Filtering****., infoscience.epfl.ch*

*R Heckel, M Vlachos (2016)* ***Interpretable recommendations via overlapping co-clusters****. arXiv preprint arXiv:1604.02071, arxiv.org, cited by 4 (2.00 per year)*

*CF Juang, TL Jeng, YC Chang (2016)* ***An Interpretable Fuzzy System Learned Through Online Rule Generation and Multiobjective ACO With a Mobile Robot Control Application****. IEEE transactions on …, ieeexplore.ieee.org, cited by 9 (4.50 per year)*

*O Kuzelka, J Davis, S Schockaert (2017)* ***Induction of Interpretable Possibilistic Logic Theories from Relational Data****. arXiv preprint arXiv:1705.07095, arxiv.org, cited by 1 (1.00 per year)*

*M Azmi, A Berrado (2015)* ***Towards an interpretable rules ensemble algorithm for classification in a categorical data space****. Intelligent Systems: Theories and …, ieeexplore.ieee.org, cited by 1 (0.33 per year)*

*MF Ghalwash, V Radosavljevic, ... (2014)* ***Utilizing temporal patterns for estimating uncertainty in interpretable early decision making****. Proceedings of the 20th …, dl.acm.org, cited by 23 (5.75 per year)*

*H Liu, A Gegov, M Cocea (2017)* ***Rule based networks: an efficient and interpretable representation of computational models****. Journal of Artificial Intelligence and Soft …, degruyter.com, cited by 1 (1.00 per year)*

*M Eftekhari, M Zeinalkhani (2013)* ***Extracting interpretable fuzzy models for nonlinear systems using gradient-based continuous ant colony optimization****. Fuzzy Information and Engineering, Springer, cited by 7 (1.40 per year)*

*BL Westra, S Dey, G Fang, M Steinbach, ... (2011)* ***Interpretable predictive models for knowledge discovery from home-care electronic health records****. Journal of Healthcare …, hindawi.com, cited by 12 (1.71 per year)*

*Y Xu, QJ Kong, R Klette, Y Liu (2014)* ***Accurate and interpretable bayesian mars*** *__for traffic flow prediction__. IEEE Transactions on …, ieeexplore.ieee.org, cited by 12 (3.00 per year)*

*A Lahsasna (2016) __An interpretable fuzzy-ensemble method for classification and data analysis__., studentsrepo.um.edu.my*

*S Destercke, S Guillaume, ... (2007) __Using the OLS algorithm to build interpretable rule bases: an application to a depollution problem__. ... , 2007. FUZZ-IEEE …, ieeexplore.ieee.org*

*T Diamantopoulos, A Symeonidis (2015) __Towards interpretable defect-prone component analysis using genetic fuzzy systems__. ... Intelligence Synergies in …, ieeexplore.ieee.org, cited by 1 (0.33 per year)*

*C Mencar, A Consiglio, ... (2007) __Interpretable Granulation of Medical Data with DC__. Hybrid Intelligent Systems …, ieeexplore.ieee.org, cited by 1 (0.09 per year)*

*Y Zhao, IM Park (2016) __Interpretable Nonlinear Dynamic Modeling of Neural Trajectories__. Advances in Neural Information Processing …, papers.nips.cc, cited by 4 (2.00 per year)*

*K Ji, H Shen (2016) __Jointly modeling content, social network and ratings for explainable and cold-start recommendation__. Neurocomputing, Elsevier, cited by 2 (1.00 per year)*

*S Banerjee, T Chattopadhyay, A Mukherjee (2017) __Interpretable Feature Recommendation for Signal Analytics__. arXiv preprint arXiv …, arxiv.org*

*D Arp, M Spreitzenbarth, M Hübner, H Gascon, ... (2013) __Technical Report IFI-TB-2013-02 DREBIN: Efficient and Explainable Detection of Android Malware in Your Pocket__., user.informatik.uni-goettingen.de*

*CF Juang, YC Chang (2016) __Data-driven interpretable fuzzy controller design through mult-objective genetic algorithm__. Systems, Man, and Cybernetics (SMC) …, ieeexplore.ieee.org*

*D Ghosh, R Guha (2011) __Using a neural network for mining interpretable relationships of West Nile risk factors__. Social Science & Medicine, Elsevier, cited by 11 (1.57 per year)*

*B Abdollahi (2017) __Accurate and justifiable: new algorithms for explainable recommendations.__., ir.library.louisville.edu*

*M Ganguly, N Brown, A Schuffenhauer, ... (2006) __Introducing the consensus modeling concept in genetic algorithms: application to interpretable discriminant analysis__. Journal of chemical …, ACS Publications, cited by 22 (1.83 per year)*

*M Peleg (2013) __Computer-interpretable clinical guidelines: a methodological review__. Journal of biomedical informatics, Elsevier, cited by 179 (35.80* per year)*

*A Trott, C Xiong, R Socher (2017) __Interpretable Counting for Visual Question Answering__. arXiv preprint arXiv:1712.08697, arxiv.org*

*Y Zhang, EB Laber, A Tsiatis, M Davidian (2016) __Interpretable Dynamic Treatment Regimes__. arXiv preprint arXiv …, arxiv.org, cited by 2 (1.00 per year)*

*S Rüping (2006) __Learning interpretable models__., eldorado.tu-dortmund.de, cited by 50 (4.17 per year)*

*A Riid, E Rüstern (2011) __An integrated approach for the identification of compact,__*

***interpretable and accurate fuzzy rule-based classifiers from data**. ... 2011 15th IEEE International Conference on, ieeexplore.ieee.org, cited by 13 (1.86 per year)*

*KK Lee (2002) **Interpretable classification model for automotive material fatigue**., eprints.soton.ac.uk*

*D Wang, C Quek, GS Ng (2016) **Bank failure prediction using an accurate and interpretable neural fuzzy inference system**. AI Communications, content.iospress.com, cited by 3 (1.50 per year)*

*CE Keefer, GW Kauffman, RR Gupta (2013) **Interpretable, probability-based confidence metric for continuous quantitative structure–activity relationship models**. Journal of chemical …, ACS Publications, cited by 27 (5.40 per year)*

*WG El-Rab (2016) **Clinical Practice Guideline Formalization: Translating Clinical Practice Guidelines to Computer Interpretable Guidelines**., era.library.ualberta.ca*

*YY Lu, J Lv, JA Fuhrman, F Sun (2017) **Towards enhanced and interpretable clustering/classification in integrative genomics**. Nucleic acids research, academic.oup.com*

*RC Kanjirathinkal (2017) **Explainable Recommendations**., cs.cmu.edu*

*X Xu, A Datta, K Dutta (2012) **Using Adjective Features from User Reviews to Generate Higher Quality and Explainable Recommendations.**. Shaping the Future of ICT Research, Springer, cited by 2 (0.33 per year)*

*C Liu, W Wang (2017) **Contextual Regression: An Accurate and Conveniently Interpretable Nonlinear Model for Mining Discovery from Scientific Data**. arXiv preprint arXiv:1710.10728, arxiv.org*

*Y Zhang (2016) **List-based Interpretable Dynamic Treatment Regimes.**., repository.lib.ncsu.edu*

*M Galende, MJ Gacto, G Sainz, R Alcalá (2014) **Comparison and design of interpretable linguistic vs. scatter FRBSs: Gm3m generalization and new rule meaning index for global assessment and local …**. Information Sciences, Elsevier, cited by 3 (0.75 per year)*

*RN Das, K Roy, PLA Popelier (2015) **Exploring simple, transparent, interpretable and predictive QSAR models for classification and quantitative prediction of rat toxicity of ionic liquids using OECD …**. Chemosphere, Elsevier, cited by 8 (2.67 per year)*

*T Lei (2017) **Interpretable neural models for natural language processing**., dspace.mit.edu*

*MP Gleeson (2008) **Generation of a set of simple, interpretable ADMET rules of thumb**. Journal of medicinal chemistry, ACS Publications, cited by 529 (52.90* per year)*

*A Kong, R Azencott (2017) **Binary Markov Random Fields and interpretable mass spectra discrimination**. ... Applications in Genetics and Molecular Biology, degruyter.com, cited by 1 (1.00 per year)*

*MF Ghalwash (2013) **Interpretable early classification of multivariate time series**., search.proquest.com*

*JC Ho (2015)* ***Clinically interpretable models for healthcare data****., repositories.lib.utexas.edu*

*S Su, Y Chen, O Mac Aodha, P Perona, Y Yue* ***Interpretable Machine Teaching via Feature Feedback****. teaching-machines.cc*

*H Ishibuchi, Y Kaisho, Y Nojima (2011)* ***Design of Linguistically Interpretable Fuzzy Rule-Based Classifiers: A Short Review and Open Questions.****. Journal of Multiple-Valued …, search.ebscohost.com, cited by 9 (1.29 per year)*

*G Fung, C Dehing-Oberije, ALAJ Dekker, ... (2011)* ***Knowledge-based interpretable predictive model for survival analysis****. US Patent …, Google Patents, cited by 2 (0.29 per year)*

*FMCRB Gonçalves (2016)* ***Computer-interpretable guidelines in decision support systems: creation and editing of clinical protocols for automatic Interpretation****., repositorium.sdum.uminho.pt*

*JMA Moral (2009)* ***INTERPRETABLE FUZZY SYSTEM MODELING****., researchgate.net*

*P Zhang (2017)* ***Towards Interpretable Vision Systems****., vtechworks.lib.vt.edu*

*A Moral, J María (2007)* ***Interpretable fuzzy systems modeling with cooperation between expert and induced knowledge (Modelado de sistemas borrosos interpretables con cooperación entre …****., oa.upm.es*

*J Kajornrit (2014)* ***Interpretable fuzzy systems for monthly rainfall spatial interpolation and time series prediction****., researchrepository.murdoch.edu.au*

*RR Sharp (2017)* ***Computational Natural Language Inference: Robust and Interpretable Question Answering****., search.proquest.com*

*JM Bischof (2014)* ***Interpretable and Scalable Bayesian Models for Advertising and Text****., dash.harvard.edu*

*SM Lundberg, B Nair, MS Vavilala, M Horibe, ... (2017)* ***Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery****. bioRxiv, biorxiv.org*

*M Lucarelli* ***DC*: an Algorithm for Automatic Acquisition of Interpretable Fuzzy Information Granules****. researchgate.net*

*R Kamimura (2011)* ***Selective information enhancement learning for creating interpretable representations in competitive learning****. Neural Networks, Elsevier, cited by 7 (1.00 per year)*

*MS Kim, CH Kim, JJ Lee (2006)* ***Evolving compact and interpretable Takagi–Sugeno fuzzy models with a new encoding scheme****. IEEE Transactions on Systems, Man …, ieeexplore.ieee.org, cited by 75 (6.25 per year)*

*MS Kim, CH Kim, JJ Lee (2005)* ***Evolving structure and parameters of fuzzy models with interpretable membership functions****. Journal of Intelligent & Fuzzy …, content.iospress.com, cited by 8 (0.62 per year)*

*J Chen, F Shen, DZ Chen, ... (2016)* ***Iris recognition based on human-interpretable features****. IEEE Transactions on …, ieeexplore.ieee.org, cited by 14 (7.00 per year)*

*A Thomas (2015)* ***Moving Towards Interpretable Mechanisms in Human Systems Biology****., search.proquest.com*

*J Chen, F Shen, DZ Chen, PJ Flynn (2013)* ***Iris Recognition Based on Human-Interpretable****., ieeexplore.ieee.org*

*V Subramanian, P Prusis, LO Pietilä, ... (2013)* ***Visually interpretable models of kinase selectivity related features derived from field-based proteochemometrics****. Journal of chemical …, ACS Publications, cited by 21 (4.20 per year)*

*G Grothaus (2005)* ***Biologically-interpretable disease classification based on gene expression data****., theses.lib.vt.edu, cited by 6 (0.46 per year)*

*S Seo, J Huang, H Yang, Y Liu (2017)* ***Interpretable convolutional neural networks with dual local and global attention for review rating prediction****. ... of the Eleventh ACM Conference on …, dl.acm.org, cited by 3 (3.00 per year)*

*MA Qureshi, D Greene (2017)* ***EVE: Explainable Vector Based Embedding Technique Using Wikipedia****. arXiv preprint arXiv:1702.06891, arxiv.org, cited by 1 (1.00 per year)*

*TW Liao (2006)* ***Mining human interpretable knowledge with fuzzy modeling methods: An overview****. Data mining and knowledge discovery approaches ..., Springer, cited by 5 (0.42 per year)*

*JS Kandola, SR Gunn (2001)* ***Interpretable modelling with sparse kernels****., eprints.soton.ac.uk, cited by 12 (0.71 per year)*

*N Zheng (2008)* ***Discovering interpretable topics in free-style text: diagnostics, rare topics, and topic supervision****., rave.ohiolink.edu*

*A Rebai (2011)* ***Interactive Object Retrieval using Interpretable Visual Models****., tel.archives-ouvertes.fr, cited by 1 (0.14 per year)*

*A Vilamala Muñoz (2015)* ***Multivariate methods for interpretable analysis of magnetic resonance spectroscopy data in brain tumour diagnosis****., upcommons.upc.edu*

*A Li (2017)* ***Towards Robust, Interpretable and Scalable Visual Representations****., search.proquest.com*

*AM Drawid (2009)* ***Physically interpretable machine learning methods for transcription factor binding site identification using principled energy thresholds and occupancy****., search.proquest.com*

**Query report 04 – “interpretation” AND “explaination”**

# *(intitle:interpretation OR intitle:explanation) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") to 2017*

*Publish or Perish 6.21.6145.6594*

## *Search terms*

***All of the words:*** *(intitle:interpretation OR intitle:explanation) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning")*
***Years:*** *earliest to 2017*

## *Data retrieval*

***Data source:*** *Google Scholar*
***Query date:*** *21/01/2018 12:16:07*
***Cache date:*** *21/01/2018 12:20:55*
***Query result:*** *[0] The operation completed successfully.*

## *Metrics*

***Publication years:*** *1972-2017*
***Citation years:*** *46 (1972-2018)*
***Papers:*** *360*
***Citations:*** *11259*
***Citations/year:*** *244.76*
***Citations/paper:*** *31.28 (*count=16)*
***Citations/author:*** *7342.90*
***Papers/author:*** *201.48*
***Authors/paper:*** *2.48/2.0/1 (mean/median/mode)*
***Age-weighed citation rate:*** *1022.16 (sqrt=31.97), 535.09/author*
***Hirsch h-index:*** *36 (a=8.69, m=0.78, 9267 cites=82.3% coverage)*
***Egghe g-index:*** *103 (g/h=2.86, 10754 cites=95.5% coverage)*
***PoP hI,norm:*** *27*
***PoP hI,annual:*** *0.59*

## *Results*

*C Nugent, P Cunningham (2005)* ***A case-based explanation system for black-box systems****. Artificial Intelligence Review, Springer, cited by 38 (2.92 per year)*

*SJ Webb, T Hanser, B Howlin, P Krause, ... (2014)* ***Feature combination networks for the interpretation of statistical machine learning models: application to Ames mutagenicity****. Journal of …, Springer, cited by 15 (3.75 per year)*

*B Goodman, S Flaxman (2016)* ***EU regulations on algorithmic decision-making and a "right to explanation"****. ... in machine learning (WHI 2016 …, pdfs.semanticscholar.org, cited by 34 (17.00* per year)*

*S Minton, JG Carbonell (1987)* ***Strategies for Learning Search Control Rules: An Explanation-based Approach.****. IJCAI, ijcai.org, cited by 73 (2.35 per year)*

*B Goodman, S Flaxman (2016)* ***European Union regulations on algorithmic decision-making and a" right to explanation"****. arXiv preprint arXiv:1606.08813, arxiv.org, cited by 55 (27.50* per year)*

*N Barakat, AP Bradley (2006)* ***Rule extraction from support vector machines: Measuring the explanation capability using the area under the roc curve****. Pattern Recognition, 2006. ICPR 2006 …, ieeexplore.ieee.org, cited by 43 (3.58 per year)*

*K Sparck Jones (1972)* ***A statistical interpretation of term specificity and its application in retrieval****. Journal of documentation, emeraldinsight.com, cited by 3151 (68.50* per year)*

*F Sørmo, J Cassens, A Aamodt (2005)* ***Explanation in case-based reasoning–perspectives and goals****. Artificial Intelligence Review, Springer, cited by 135 (10.38* per year)*

*J Diederich (1992)* ***Explanation and artificial neural networks****. International Journal of Man-Machine Studies, Elsevier, cited by 50 (1.92 per year)*

*A Kehler, D Appelt, L Taylor, A Simma (2004)* ***The (non) utility of predicate-argument frequencies for pronoun interpretation****. ... of the North American Chapter of …, aclweb.org, cited by 95 (6.79 per year)*

*G Montavon, ML Braun, T Krueger, ... (2013)* ***Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment****. IEEE Signal Processing …, ieeexplore.ieee.org, cited by 36 (7.20 per year)*

*WJ Clancey (1993)* ***Situated action: A neuropsychological interpretation response to Vera and Simon****. Cognitive Science, Wiley Online Library, cited by 447 (17.88* per year)*

*I Zelic, I Kononenko, N Lavrac, ... (1997)* ***Diagnosis of sport injuries with machine learning: Explanation of induced decisions****. Computer-Based Medical …, ieeexplore.ieee.org, cited by 6 (0.29 per year)*

*D Leake, D McSherry (2005)* ***Introduction to the special issue on explanation in case-based reasoning****. Artificial Intelligence Review, Springer, cited by 30 (2.31 per year)*

*B Poulin, R Eisner, D Szafron, P Lu, ... (2006)* ***Visual explanation of evidence with additive classifiers****. Proceedings Of The …, ocs.aaai.org, cited by 65 (5.42 per year)*

*JL Rojo-Álvarez, Á Arenal-Maíz, ... (2002)* ***Support vector black-box interpretation in ventricular arrhythmia discrimination****. IEEE engineering in …, ieeexplore.ieee.org, cited by 12 (0.75 per year)*

*DL McGuinness, L Ding, PP Da Silva, C Chang (2007)* ***PML 2: A Modular***

***Explanation Interlingua.**. ExaCt, vvvvw.aaai.org, cited by 102 (9.27 per year)*

*LW Glorfeld (1996)* ***A methodology for simplification and interpretation of backpropagation-based neural network models****. Expert Systems with Applications, Elsevier, cited by 45 (2.05 per year)*

*J Bobbin, F Recknagel (2001)* ***Knowledge discovery for prediction and explanation of blue-green algal dynamics in lakes by evolutionary algorithms****. Ecological Modelling, Elsevier, cited by 44 (2.59 per year)*

*JA Walls, AJ Walton, JM Robertson, ... (1988)* ***Interpretation of capacitance-voltage curves for process fault diagnosis: a machine-learning expert system approach****. ... Test Structures, 1988 ..., ieeexplore.ieee.org, cited by 7 (0.23 per year)*

*M Bohanec, V Rajkovic (1988)* ***Knowledge acquisition and explanation for multi-attribute decision making****. 8th Intl Workshop on Expert Systems and ..., researchgate.net, cited by 81 (2.70 per year)*

*K Morgenthal, W Weckwerth, R Steuer (2006)* ***Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation****. Biosystems, Elsevier, cited by 122 (10.17* per year)*

*MJ Pazzani (1987)* ***Explanation-based learning for knowledge-based systems****. International journal of man-machine studies, Elsevier, cited by 30 (0.97 per year)*

*V Rajkovič, M Bohanec (1991)* ***Decision support by knowledge explanation****. Environments for supporting decision processes, cited by 36 (1.33 per year)*

*ML Vaughn (1996)* ***Interpretation and knowledge discovery from the multilayer perceptron network: opening the black box****. Neural computing & applications, Springer, cited by 29 (1.32 per year)*

*A Navia-Vázquez, E Parrado-Hernández (2006)* ***Support vector machine interpretation****. Neurocomputing, Elsevier, cited by 20 (1.67 per year)*

*IA Taha, J Ghosh (1999)* ***Symbolic interpretation of artificial neural networks****. IEEE Transactions on knowledge and data ..., ieeexplore.ieee.org, cited by 215 (11.32* per year)*

*F Sørmo, J Cassens (2004)* ***Explanation goals in case-based reasoning****. Proceedings of the ECCBR 2004 …, researchgate.net, cited by 30 (2.14 per year)*

*N Gamage, YC Kuang, R Akmeliawati, ... (2011)* ***Gaussian process dynamical models for hand gesture interpretation in sign language****. Pattern Recognition …, Elsevier, cited by 21 (3.00 per year)*

*P Cunningham, D Doyle, J Loughrey (2003)* ***An evaluation of the usefulness of case-based explanation****. Case-Based Reasoning Research …, Springer, cited by 94 (6.27 per year)*

*G Stiglic, M Mertik, V Podgorelec, ... (2006)* ***Using Visual Interpretation of Small Ensembles in Microarray Analysis****. ... -Based Medical Systems ..., ieeexplore.ieee.org, cited by 8 (0.67 per year)*

*K Kumar, GSM Thakur (2012)* ***Extracting explanation from artificial neural networks****. International Journal of Computer Science …, researchgate.net, cited by 8 (1.33 per year)*

*PG Polishchuk, VE Kuz'min, AG Artemenko, ... (2013)* ***Universal approach for structural interpretation of QSAR/QSPR models****. Molecular ..., Wiley Online Library, cited by 21 (4.20 per year)*

*KR Levi, DL Perschbacher, MA Hoffman, ... (1992)* ***An explanation-based-learning approach to knowledge compilation: A pilot's associate application****. IEEE ..., ieeexplore.ieee.org, cited by 12 (0.46 per year)*

*MS Duh, AM Walker, JZ Ayanian (1998)* ***Epidemiologic interpretation of artificial neural networks****. American journal of ..., academic.oup.com, cited by 68 (3.40 per year)*

*G Costa, RQ Feitosa, LMG Fonseca, ... (2010)* ***Knowledge-based interpretation of remote sensing data with the InterIMAGE system: major characteristics and recent developments****. Proceedings of the 3rd ..., isprs.org, cited by 24 (3.00 per year)*

*DW McClune, NJ Marks, ... (2014)* ***Tri-axial accelerometers quantify behaviour in the Eurasian badger (Meles meles): towards an automated interpretation of field data****. Animal ..., animalbiotelemetry.biomedcentral ..., cited by 21 (5.25 per year)*

*N Benamrane, A Aribi, L Kraoula (1993)* ***Fuzzy neural networks and genetic algorithms for medical images interpretation****. Geometric Modeling and ..., ieeexplore.ieee.org, cited by 32 (1.28 per year)*

*M Drobics, W Winiwater, ... (2000)* ***Interpretation of self-organizing maps with fuzzy rules****. Tools with Artificial ..., ieeexplore.ieee.org, cited by 19 (1.06 per year)*

*H Núñez, C Angulo, A Català (2002)* ***Support vector machines with symbolic interpretation****. Neural Networks, 2002. SBRN ..., ieeexplore.ieee.org, cited by 14 (0.88 per year)*

*R Goodacre, DB Kell (2003)* ***Evolutionary computation for the interpretation of metabolomic data****. Metabolic profiling: its role in biomarker discovery ..., Springer, cited by 19 (1.27 per year)*

*D Neagu, V Palade (2003)* ***A neuro-fuzzy approach for functional genomics data interpretation and analysis****. Neural Computing & Applications, Springer, cited by 16 (1.07 per year)*

*K Darlington (2013)* ***Aspects of intelligent systems explanation****. Universal Journal of Control and Automation, hrpub.org, cited by 12 (2.40 per year)*

*C Town, D Sinclair (2003)* ***A self-referential perceptual inference framework for video interpretation****. Computer Vision Systems, Springer, cited by 22 (1.47 per year)*

*S Louis, G McGraw, RO Wyckoff (1993)* ***Case-based reasoning assisted explanation of genetic algorithm results****. Journal of Experimental & ..., Taylor & Francis, cited by 46 (1.84 per year)*

*C Nugent, D Doyle, P Cunningham (2009)* ***Gaining insight through case-based explanation****. Journal of Intelligent Information ..., Springer, cited by 16 (1.78 per year)*

*J Diederich, AB Tickle (1994)* ***Explanation and collective computation****. Complex Systems: Mechanism of ..., books.google.com, cited by 8 (0.33 per year)*

*R Turner (2016)* ***A model explanation system****. Machine Learning for Signal Processing (MLSP) ..., ieeexplore.ieee.org, cited by 11 (5.50 per year)*

*V Cherkassky, S Dhar (2015)* ***Interpretation of black-box predictive models****. Measures of Complexity, Springer, cited by 4 (1.33 per year)*

*L Carlsson, EA Helgee, S Boyer (2009)* ***Interpretation of nonlinear QSAR models applied to Ames mutagenicity data****. Journal of chemical information ..., ACS Publications, cited by 51 (5.67 per year)*

*O Biran, C Cotton (2017)* ***Explanation and justification in machine learning: A survey****. IJCAI-17 Workshop on Explainable AI (XAI), intelligentrobots.org, cited by 3 (3.00 per year)*

*A Bernatzki, W Eppler, H Gemmeke (1996)* ***Interpretation of neural networks for classification tasks****. Proceedings of EUFIT, researchgate.net, cited by 8 (0.36 per year)*

*JA Alexander, MC Mozer (1999)* ***Template-based procedures for neural network interpretation****. Neural Networks, Elsevier, cited by 49 (2.58 per year)*

*KW Darlington (2011)* ***Designing for explanation in health care applications of expert systems****. Sage Open, journals.sagepub.com, cited by 10 (1.43 per year)*

*AM Keuneke (1989)* ***Machine understanding of devices causal explanation of diagnostic conclusions****., etd.ohiolink.edu, cited by 39 (1.34 per year)*

*J Lim (2004)* ***The role of power distance and explanation facility in online bargaining utilizing software agents****. Journal of Global Information Management, search.proquest.com, cited by 25 (1.79 per year)*

*R Hasan, F Gandon (2012)* ***A brief review of explanation in the Semantic Web****. Explanation-aware Computing ExaCt 2012, lirmm.fr, cited by 5 (0.83 per year)*

*ML Vaughn (1999)* ***Derivation of the multilayer perceptron weight constraints for direct network interpretation and knowledge discovery****. Neural Networks, Elsevier, cited by 29 (1.53 per year)*

*MR Wick, WB Thompson (1992)* ***Reconstructive expert system explanation****. Artificial Intelligence, Elsevier, cited by 145 (5.58 per year)*

*S Liu (1998)* ***Strategic scanning and interpretation revisiting: foundations for a software agent support system-Part 2: scanning the business environment with software agents****. Industrial Management & Data Systems, emeraldinsight.com, cited by 34 (1.70 per year)*

*R Turner (2016)* ***A model explanation system: Latest updates and extensions****. arXiv preprint arXiv:1606.09517, arxiv.org, cited by 4 (2.00 per year)*

*A Khelassi (2014)* ***Explanation-aware computing of the prognosis for breast cancer supported by IK-DCBRC: Technical innovation****. Electronic physician, ncbi.nlm.nih.gov, cited by 4 (1.00 per year)*

*F Salazar, MÁ Toledo, E Oñate, B Suárez (2016)* ***Interpretation of dam deformation and leakage with boosted regression trees****. Engineering Structures, Elsevier, cited by 5 (2.50 per year)*

*M Liebmann, M Hagenau, D Neumann (2012)* ***Information processing in electronic markets: Measuring subjective interpretation using sentiment***

***analysis**., aisel.aisnet.org, cited by 14 (2.33 per year)*

*S Dietzen, F Pfenning (1992)* ***Higher-order and modal logic as a framework for explanation-based generalization****. Machine Learning, Springer, cited by 32 (1.23 per year)*

*EW Saad, DC Wunsch (2007)* ***Neural network explanation using inversion****. Neural Networks, Elsevier, cited by 53 (4.82 per year)*

*J Balfer, J Bajorath (2014)* ***Introduction of a methodology for visualization and graphical interpretation of bayesian classification models****. Journal of chemical information and …, ACS Publications, cited by 7 (1.75 per year)*

*SS Ibrahim, MA Bamatraf (2013)* ***Interpretation Trained Neural Networks Based on Genetic Algorithms****. International Journal of Artificial …, search.proquest.com, cited by 4 (0.80 per year)*

*SC Park, MS Lam, A Gupta (1995)* ***Rule extraction from neural networks: Enhancing the explanation capability****. Journal of Intelligence and …, koreascience.or.kr, cited by 4 (0.17 per year)*

*B Li, PK Goel (2007)* ***Additive regression trees and smoothing splines-predictive modeling and interpretation in data mining****. Contemporary Mathematics, books.google.com, cited by 5 (0.45 per year)*

*S Todorovic, MC Nechyba (2004)* ***Interpretation of complex scenes using generative dynamic-structure models****. Computer Vision and Pattern …, ieeexplore.ieee.org, cited by 5 (0.36 per year)*

*I Kononenko, E Štrumbelj, Z Bosnić, D Pevec, M Kukar, ... (2013)* ***Explanation and reliability of individual predictions****. Informatica, informatica.si, cited by 5 (1.00 per year)*

*S Rahnamayan, GG Wang, M Ventresca (2012)* ***An intuitive distance-based explanation of opposition-based sampling****. Applied Soft Computing, Elsevier, cited by 30 (5.00 per year)*

*Y Fukuchi, M Osawa, H Yamakawa, M Imai (2017)* ***Autonomous self-explanation of behavior for interactive reinforcement learning agents****. Proceedings of the 5th …, dl.acm.org, cited by 2 (2.00 per year)*

*M Bohanec, M Robnik-Šikonja, MK Borštnar (2017)* ***Organizational Learning Supported by Machine Learning Models Coupled with General Explanation Methods: A Case of B2B Sales Forecasting****. Organizacija, degruyter.com, cited by 1 (1.00 per year)*

*G Wang, Z Deng, KS Choi (2015)* ***Detection of epileptic seizures in EEG signals with rule-based interpretation by random forest approach****. International Conference on Intelligent …, Springer, cited by 4 (1.33 per year)*

*L Cummins, D Bridge (2012)* ***Kleor: A knowledge lite approach to explanation oriented retrieval****. Computing and Informatics, cai.sk, cited by 4 (0.67 per year)*

*J Balfer, J Bajorath (2015)* ***Visualization and Interpretation of Support Vector Machine Activity Predictions****. Journal of chemical information and …, ACS Publications, cited by 8 (2.67 per year)*

*B Fránay, D Hofmann, A Schulz, ... (2014)* ***Valid interpretation of feature relevance for linear data mappings****. … Intelligence and Data …,* *ieeexplore.ieee.org, cited by 8 (2.00 per year)*

*AM Bell, S Ramachandran (2003)* ***An Intelligent Tutoring System for Remote Sensing and Image Interpretation****. Proceedings of the Interservice ..., stottlerhenke.com, cited by 3 (0.20 per year)*

*W Wiharto, H Kusnanto, ... (2016)* ***Interpretation of clinical data based on C4. 5 algorithm for the diagnosis of coronary heart disease****. Healthcare informatics ..., synapse.koreamed.org, cited by 8 (4.00 per year)*

*V Schetinin, JE Fieldsend, D Partridge, ... (2007)* ***Confident interpretation of Bayesian decision tree ensembles for clinical applications****. IEEE Transactions ..., ieeexplore.ieee.org, cited by 25 (2.27 per year)*

*T Kulesza (2012)* ***An explanation-centric approach for personalizing intelligent agents****. Proceedings of the 2012 ACM international conference ..., dl.acm.org, cited by 2 (0.33 per year)*

*R Hasan, F Gandon (2012)* ***Explanation in the Semantic Web: a survey of the state of the art****., hal.archives-ouvertes.fr, cited by 4 (0.67 per year)*

*E Gentet, S Tourret, K Inoue (2016)* ***Learning from interpretation transition using feed-forward neural networks****. ... Workshop Proceedings of ..., pdfs.semanticscholar.org, cited by 2 (1.00 per year)*

*Y Hechtlinger (2016)* ***Interpretation of Prediction Models Using the Input Gradient****. arXiv preprint arXiv:1611.07634, arxiv.org, cited by 5 (2.50 per year)*

*Z Bosnić, J Demšar, G Kešpret, PP Rodrigues, ... (2014)* ***Enhancing data stream predictions with reliability estimators and explanation****. ... Applications of Artificial ..., Elsevier, cited by 10 (2.50 per year)*

*J Dhaliwal (1993)* ***Design and use of Explanation Facilities****. The Handbook of Applied Expert Systems, books.google.com, cited by 3 (0.12 per year)*

*S Flutura, J Wagner, F Lingenfelser, ... (2016)* ***MobileSSI: Asynchronous fusion for social signal interpretation in the wild****. Proceedings of the 18th ..., dl.acm.org, cited by 3 (1.50 per year)*

*G Du, G Ruhe (2009)* ***Does explanation improve the acceptance of decision support for product release planning?****. ... and Measurement, 2009. ESEM 2009. 3rd ..., ieeexplore.ieee.org, cited by 2 (0.22 per year)*

*S Wachter, B Mittelstadt, L Floridi (2017)* ***Why a right to explanation of automated decision-making does not exist in the general data protection regulation****. International Data Privacy ..., academic.oup.com, cited by 28 (28.00* per year)*

*P Lakhani, AB Prater, RK Hutson, KP Andriole, ... (2017)* ***Machine learning in radiology: applications beyond image interpretation****. Journal of the American ..., Elsevier, cited by 1 (1.00 per year)*

*NH Barakat (2007)* ***Rule-extraction from Support Vector Machines: Medical Diagnosis, Prediction and Explanation****., espace.library.uq.edu.au, cited by 1 (0.09 per year)*

*E COMPUTING (1997)* ***Signal interpretation in two-phase fluid dynamics through machine learning and evolutionary computing****. ... Applications or Artificial Intelligence and Expert ..., books.google.com, cited by 1 (0.05 per year)*

*A Ghorbani, A Abid, J Zou (2017)* ***Interpretation of Neural Networks is Fragile****. arXiv preprint arXiv:1710.10547, arxiv.org, cited by 2 (2.00 per year)*

*M Mejía-Lavalle, A Sánchez Vivar (2009)* ***Outlier detection with explanation facility****. Machine learning and data mining in ..., Springer, cited by 2 (0.22 per year)*

*P Perner (2014)* ***A Method for Supporting the Domain Expert by the Interpretation of Different Decision Trees Learnt from the Same Domain****. Quality and Reliability Engineering International, Wiley Online Library, cited by 2 (0.50 per year)*

*D Berleant, LP Falcone, UM Fayyad (1989)* ***Selective simulation and selective sensor interpretation in monitoring****., arc.aiaa.org, cited by 2 (0.07 per year)*

*M Atzmueller, N Hayat, A Schmidt, ... (2017)* ***Explanation-Aware Feature Selection using Symbolic Time Series Abstraction: Approaches and Experiences in a Petro-Chemical Production Context****. Proc. IEEE INDIN ..., kde.cs.uni-kassel.de, cited by 2 (2.00 per year)*

*X Liang, W Pedrycz (2006)* ***Fuzzy logic-based networks: A study in logic data interpretation****. International journal of intelligent systems, Wiley Online Library, cited by 2 (0.17 per year)*

*B Hayes, JA Shah (2017)* ***Improving Robot Controller Transparency Through Autonomous Policy Explanation****. Proceedings of the 2017 ACM/IEEE International ..., dl.acm.org, cited by 10 (10.00* per year)*

*AD Selbst, J Powles (2017)* ***Meaningful information and the right to explanation****. International Data Privacy Law, academic.oup.com, cited by 1 (1.00 per year)*

*XW Zhu, YJ Xin, HL Ge (2015)* ***Recursive random forests enable better predictive performance and model interpretation than variable selection by LASSO****. Journal of chemical information and ..., ACS Publications, cited by 4 (1.33 per year)*

*M Boronowsky (1998)* ***Automatic Measurement Interpretation of a Physical System with Decision Tree Induction****. Conference of IDEAL, books.google.com, cited by 2 (0.10 per year)*

*ML Vaughn, SJ Taylor, MA Foy, ... (2001)* ***Investigating the Reliability of a Low-back-pain MLP by Using a Full Explanation Facility****. Neural Networks, 2001 ..., ieeexplore.ieee.org, cited by 2 (0.12 per year)*

*V Palade, CD Neagu, RJ Patton (2001)* ***Interpretation of trained neural networks by rule extraction****. Fuzzy days, Springer, cited by 30 (1.76 per year)*

*G Du (2010)* ***Design and evaluation of explanation-based decision support for software release planning****., dl.acm.org, cited by 1 (0.13 per year)*

*M Bilgic, R Mooney, E Rich (2004)* ***Explanation for recommender systems: satisfaction vs. promotion****. Computer Sciences Austin, University of ..., cs.utexas.edu, cited by 15 (1.07 per year)*

*F Doshi-Velez, M Kortz, R Budish, C Bavitz, ... (2017)* ***Accountability of AI Under the Law: The Role of Explanation****. arXiv preprint arXiv ..., arxiv.org, cited by 1 (1.00 per year)*

*TS Sobh (2005)* ***Explanation-based learning to recognize network malfunctions****. Information Knowledge Systems Management, content.iospress.com, cited by 1*

*(0.08 per year)*

*J Diederich (1989)* ***Explanation and Connectionism****. GWAI-89 13th German Workshop on Artificial …, Springer, cited by 1 (0.03 per year)*

*R Hänsch, O Hellwich (2015)* ***Performance Assessment and Interpretation of Random Forests by Three-dimensional Visualizations.****. IVAPP, cv.tu-berlin.de, cited by 1 (0.33 per year)*

*M Mejia-Lavalle (2010)* ***Outlier Detection with Innovative Explanation Facility over a Very Large Financial Database****. Electronics, Robotics and Automotive …, ieeexplore.ieee.org, cited by 2 (0.25 per year)*

*Z Li, D Tate (2010)* ***Patent Analysis for Systematic Innovation: Automatic Function Interpretation and Automatic Classification of Level of Invention using Natural Language …****. International Journal of Systematic Innovation, ns2.sme-edu.org.tw, cited by 1 (0.13 per year)*

*N Chen, IJ del Val, S Kyriakopoulos, KM Polizzi, ... (2012)* ***Metabolic network reconstruction: advances in in silico interpretation of analytical information****. Current opinion in …, Elsevier, cited by 27 (4.50 per year)*

*D Kroening, TW Reps, SA Seshia, A Thakur (2014)* ***Decision procedures and abstract interpretation (Dagstuhl seminar 14351)****. Dagstuhl Reports, drops.dagstuhl.de, cited by 1 (0.25 per year)*

*DL McGuinness, A Glass, M Wolverton, PP Da Silva (2007)* ***A Categorization of Explanation Questions for Task Processing Systems.****. ExaCt, ocs.aaai.org, cited by 11 (1.00 per year)*

*P Polishchuk (2017)* ***Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future****. Journal of chemical information and modeling, ACS Publications, cited by 1 (1.00 per year)*

*T Yarkoni, J Westfall (2017)* ***Choosing prediction over explanation in psychology: Lessons from machine learning****. Perspectives on Psychological …, journals.sagepub.com, cited by 42 (42.00* per year)*

*HC Shin, L Lu, L Kim, A Seff, J Yao, ... (2016)* ***Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation****. ... of Machine Learning …, jmlr.org, cited by 13 (6.50 per year)*

*A Zaeri, MA Nematbakhsh (2012)* ***A framework for semantic interpretation of noun compounds using tratz model and binary features****. ETRI Journal, Wiley Online Library, cited by 4 (0.67 per year)*

*J Kim, J Seo (2017)* ***Human Understandable Explanation Extraction for Black-box Classification Models Based on Matrix Factorization****. arXiv preprint arXiv:1709.06201, arxiv.org*

*SJ Webb (2015)* ***Interpretation and mining of statistical machine learning (Q) SAR models for toxicity prediction.****., researchgate.net*

*N Olofsson (2017)* ***A Machine Learning Ensemble Approach to Churn Prediction: Developing and Comparing Local Explanation Models on Top of a Black-Box Classifier****., diva-portal.org*

*TW Kim, B Routledge (2017)* ***Algorithmic Transparency, a Right to Explanation, and Placing Trust.****, business-ethics.net*

*F Doshi-Velez, R Budish, M Kortz* ***The Role of Explanation in Algorithmic Trust****. trustworthy-algorithms.org*

*G Langs, I Rish, M Grosse-Wentrup, B Murphy (2012)* ***Machine Learning and Interpretation in Neuroimaging: International Workshop, MLINI 2011, Held at NIPS 2011, Sierra Nevada, Spain, December 16-17, 2011 …****, books.google.com*

*D Assouline, N Mohajeri, ... (2017)* ***Random Forests (RFs) for Estimation, Uncertainty Prediction and Interpretation of Monthly Solar Potential****. EGU General …, meetingorganizer.copernicus.org*

*I Rish, MGWB Murphy (2016)* ***Machine Learning and Interpretation in Neuroimaging****., Springer*

*HKB Babiker, R Goebel (2017)* ***An Introduction to Deep Visual Explanation****. arXiv preprint arXiv:1711.09482, arxiv.org*

*C Brinton* ***A Framework for Explanation of Machine Learning Decisions****. IJCAI-17 Workshop on Explainable AI (XAI), intelligentrobots.org*

*K Strandburg* ***Decision-making, Machine Learning and the Value of Explanation****. dsi.unive.it*

*C Nugent, P Cunningham (2004)* ***A Case-Based Explanation System forBlack-Box'Systems****. Artificial Intelligence Review (This issue, Citeseer*

*T Zhao (2017)* ***Machine assisted quantitative seismic interpretation****., shareok.org*

*M Al-Shedivat, A Dubey, EP Xing* ***Personalized Survival Prediction with Contextual Explanation Networks****. cs.cmu.edu*

*F Sieverink, S Kelders, S Akkersdijk, M Poel, ... (2016)* ***Work in progress: a protocol for the collection, analysis, and interpretation of log data from eHealth technology****., doc.utwente.nl, cited by 1 (0.50 per year)*

*H Kuwajima, M Tanaka (2017)* ***Network Analysis for Explanation****. arXiv preprint arXiv:1712.02890, arxiv.org*

*M Imai (2017)* ***Application of Instruction-Based Behavior Explanation to a Reinforcement Learning Agent with Changing Policy****. … , ICONIP 2017, Guangzhou, China, November 14-18 …, books.google.com*

*MC Minnotte, A Cutler (2001)* ***VISUALIZATION AND INTERPRETATION OF HIGH-DIMENSIONAL CLASSIFIERS****. Proceedings of the Annual Meeting of the American …*

*I Kononenko, E Štrumbelj, Z Bosnić, ...* ***EXPLANATION AND RELIABILITY OF INDIVIDUAL PREDICTIONS: RECENT RESEARCH BY LKM****. … DRUŽBA− IS 2012, pdfs.semanticscholar.org*

*Y Fukuchi, M Osawa, H Yamakawa, M Imai (2017)* ***Application of instruction-based behavior explanation to a reinforcement learning agent with changing policy****. International Conference on …, Springer*

*L Edwards, M Veale (2017)* ***Enslaving the Algorithm: From a 'Right to an Explanation'to a 'Right to Better Decisions'?****., papers.ssrn.com*

*EE Wille (2016)* ***Wanted: Transparent algorithms, interpretation skills, common sense****., e-collection.library.ethz.ch*

*DSRGP Lu, DWCMDJ Anvik, BPZLR Eisner* ***TCXplain: Transparent Explanation of Naïve Bayes Classifications****. webdocs.cs.ualberta.ca*

*TW Kim, B Routledge (2017)* ***ALGORITHMIC TRANSPARENCY, A RIGHT TO EXPLANATION AND TRUST****., business-ethics.net*

*G Stiglic, N Khan, M Verlic, P Kokol (2007)* ***Gene expression analysis of leukemia samples using visual interpretation of small ensembles: a case study****. IAPR International Workshop on …, Springer*

*ME Johnson (2004)* ***Explanation Facilities in Neural Nets Research Project for****., matthiasjohnson.com*

*EW Saad (1999)* ***Inversion for explanation capability of neural networks and query-based learning****., ttu-ir.tdl.org*

*B Ustun (2009)* ***Support vector machines: facilitating the interpretation and application****., repository.ubn.ru.nl*

*K Muhammad, A Lawlor, B Smyth (2017)* ***On the Pros and Cons of Explanation-Based Ranking****. International Conference on Case …, Springer*

*L Fan (2017)* ***Deep Epitome for Unravelling Generalized Hamming Network: A Fuzzy Logic Interpretation of Deep Learning****. arXiv preprint arXiv:1711.05397, arxiv.org*

*HKB Babiker, R Goebel (2017)* ***Using KL-divergence to focus Deep Visual Explanation****. arXiv preprint arXiv:1711.06431, arxiv.org*

*J Oramas, K Wang, T Tuytelaars (2017)* ***Visual Explanation by Interpretation: Improving Visual Feedback Capabilities of Deep Neural Networks****. arXiv preprint arXiv:1712.06302, arxiv.org*

*A Fuser, F Fontaine, J Copper (2014)* ***Data Quality, Consistency, and Interpretation Management for Wind Farms by Using Neural Networks****. Parallel & Distributed …, ieeexplore.ieee.org*

*A Augello, I Infantino, A Lieto, U Maniscalco, G Pilato, ...* ***Towards A Dual Process Approach to Computational Explanation in Human-Robot Social Interaction****. researchgate.net*

*Q Gao, L She, JY Chai (2017)* ***Interactive Learning of State Representation through Natural Language Instruction and Explanation****. arXiv preprint arXiv:1710.02714, arxiv.org*

*F Ancien, F Pucci, M Godfroid, M Rooman (2017)* ***Prediction and interpretation of deleterious coding variants in terms of protein structural stability****. bioRxiv, biorxiv.org*

*D Jäckle, F Stoffel, S Mittelstädt, DA Keim, ... (2017)* ***Interpretation of Dimensionally-reduced Crime Data: A Study with Untrained Domain Experts.****. VISIGRAPP (3 …, hci.uni-konstanz.de, cited by 1 (1.00 per year)*

*W Duch, R Adamczak, K Grąbczewski, K Grudziński, ...* ***Understanding the data: extraction, optimization and interpretation of logical rules****. phys.uni.torun.pl*

*CA Miller, R Larson, P Bursch (1992)* ***Winning the explanation game; providing user explanations for a model-based expert system****. AUTOTESTCON'92. IEEE …, ieeexplore.ieee.org*

*C Liberati, F Camillo, G Saporta (2017)* ***Advances in credit scoring: combining performance and interpretation in kernel discriminant analysis****. Advances in Data Analysis and …, Springer, cited by 2 (2.00 per year)*

*ML Vaughn, SJ Cavill, SJ Taylor, ... (2001)* ***A full explanation facility for a MLP network that classifies low-back-pain patients****. … Australian and New …, ieeexplore.ieee.org*

*O Parisot, Y Didry, T Tamisier, ... (2015)* ***Helping predictive analytics interpretation using regression trees and clustering perturbation****. Journal of Decision …, Taylor & Francis*

*EK Bowman, ME Jobidonb, A Bergeron-Guyardb, ...* ***Human Interpretation of Text and Video Analytics****. cradpdf.drdc-rddc.gc.ca*

*MA Kneen, DJ Lary, WA Harrison, HJ Annegarn, ... (2016)* ***Interpretation of satellite retrievals of PM2.5 over the southern African Interior****. Atmospheric …, Elsevier*

*M Fiterau, A Dubrawski (2012)* ***Explanation-Oriented Classification via Subspace Partitioning****., cs.cmu.edu*

*E Ferrari, A Ridi, M Muselli* ***Modeling and interpretation of responses from e-noses in the detection of gases in air****. pdfs.semanticscholar.org*

*K Zor, Ö Çelik, HB Yıldırım, O Timur, A Teke* ***Interpretation of Error Calculation Methods in the Context of Energy Forecasting****. researchgate.net*

*T Quandt, G Shegalov, H Sjøvaag, G Vossen (2016)* ***Analysis, Interpretation and Benefit of User-Generated Data: Computer Science Meets Communication Studies (Dagstuhl Seminar 16141)****. Dagstuhl Reports, drops.dagstuhl.de*

*GS Galloway, VM Catterson, C Love, ... (2017)* ***Modeling and Interpretation of Tidal Turbine Vibration Through Weighted Least Squares Regression****. IEEE Transactions on …, ieeexplore.ieee.org*

*J Jameson, HS Abdullah, S Norul, ... (2013)* ***Multiple Frames Combination Versus Single Frame Super Resolution Methods for CCTV Forensic Interpretation.****. Journal of …, pdfs.semanticscholar.org, cited by 1 (0.20 per year)*

*LB ŽAUCER, B ZUPAN, M GOLOBIČ* ***Formal Interpretation of Preference Maps: A Data Mining Approach****. kolleg.loel.hs-anhalt.de*

*ML Vaughn, E Ong, SJ Cavill (1997)* ***Interpretation and knowledge discovery from a multilayer perceptron network that performs whole life assurance risk assessment****. Neural Computing & Applications, Springer, cited by 29 (1.38 per year)*

*M Sester (2000)* ***Knowledge acquisition for the automatic interpretation of spatial data****. International Journal of Geographical Information …, Taylor & Francis, cited by 74 (4.11 per year)*

*B Han, J Lim (2002)* ***Influence of culture and explanation facility on performance of negotiation agents****. System Sciences, 2002. HICSS. Proceedings of …, ieeexplore.ieee.org, cited by 3 (0.19 per year)*

*E Bresso, R Grisoni, MD Devignes, A Napoli, ... (2012)* ***Formal concept analysis for the interpretation of relational learning applied on 3D protein-binding sites****. … Retrieval-KDIR 2012, hal.inria.fr, cited by 3 (0.50 per year)*

*L Edwards, M Veale (2017)* ***Slave to the Algorithm? Why a 'Right to Explanation'is Probably Not the Remedy You are Looking for****., papers.ssrn.com, cited by 4 (4.00 per year)*

*P Polishchuk, O Tinkov, T Khristova, ... (2016)* ***Structural and physico-chemical interpretation (SPCI) of QSAR models and its comparison with matched molecular pair analysis****. Journal of chemical …, ACS Publications, cited by 7 (3.50 per year)*

*BWJ Surridge, S Bizzi, A Castelletti (2014)* ***A framework for coupling explanation and prediction in hydroecological modelling****. Environmental modelling & software, Elsevier, cited by 8 (2.00 per year)*

*B Hodjat (2006)* ***Interpretation phase for adaptive agent oriented software architecture****. US Patent 6,990,670, Google Patents, cited by 7 (0.58 per year)*

*A Gacek (2011)* ***New frontiers of analysis, interpretation and classification of biomedical signals: a computational intelligence framework****. Journal of Medical Informatics & Technologies, infona.pl*

*N Mohamudally, D Khan (2011)* ***Application of a unified medical data miner (umdm) for prediction, classification, interpretation and visualization on medical datasets: The diabetes dataset case****. Advances in Data Mining. Applications and …, Springer, cited by 6 (0.86 per year)*

*MT Cox (2011)* ***Metareasoning, monitoring, and self-explanation****. Metareasoning: Thinking about thinking, books.google.com, cited by 27 (3.86 per year)*

*M Gajzler (2013)* ***The idea of knowledge supplementation and explanation using neural networks to support decisions in construction engineering****. Procedia Engineering, Elsevier, cited by 13 (2.60 per year)*

*I Kononenko, Z Bosnic, J Demsar, G Kespret, ... (2014)* ***Enhancing data stream predictions with reliability estimators and explanation****., repositorio.inesctec.pt*

*K Hasegawa, K Funatsu (2010)* ***Non-linear modeling and chemical interpretation with aid of support vector machine and regression****. Current computer-aided drug design, ingentaconnect.com, cited by 27 (3.38 per year)*

*MC Hao, LEE Wei-Nchih, A Jaeger, ... (2017)* ***Facilitating interpretation of high-dimensional data clusters****. US Patent App. 15 …, Google Patents*

*S Louisy, G McGrawy, RO Wyckoy (1992)* ***CBR Assisted Explanation of GA Results****. Computer Science, html.soic.indiana.edu, cited by 3 (0.12 per year)*

*M Topalovic, S Laval, JM Aerts, T Troosters, ... (2017)* ***Automated Interpretation of Pulmonary Function Tests in Adults with Respiratory Complaints****. Respiration, karger.com, cited by 1 (1.00 per year)*

*M Al-Shedivat, A Dubey, EP Xing (2017)* ***Contextual Explanation Networks****. arXiv preprint arXiv:1705.10301, arxiv.org, cited by 7 (7.00 per year)*

*R Paredes, PL Tzou, G van Zyl, G Barrow, R Camacho, ... (2017)* ***Collaborative update of a rule-based expert system for HIV-1 genotypic resistance test interpretation****. PloS one, journals.plos.org, cited by 1 (1.00 per year)*

*N Liu, D Shin, X Hu (2017)* ***Contextual Outlier Interpretation****. arXiv preprint arXiv:1711.10589, arxiv.org*

*T Jiang, AB Owen (2002)* ***Quasi-regression for visualization and interpretation of black box functions****., pdfs.semanticscholar.org, cited by 14 (0.88 per year)*

*CS Taber (1998)* ***The interpretation of foreign policy events: A cognitive process theory****. Problem representation in foreign policy decision ..., books.google.com, cited by 30 (1.50 per year)*

*Z Tatjana, S Busayarat (2011)* ***Computer-aided Analysis and Interpretation of HRCT Images of the Lung****. Theory and Applications of CT Imaging ..., intechopen.com, cited by 1 (0.14 per year)*

*E Bresso, R Grisoni, MD Devignes, A Napoli, ... (2012)* ***ILP Characterization of 3D Protein-Binding Sites and FCA-Based Interpretation****. ... Joint Conference on ..., Springer, cited by 2 (0.33 per year)*

*P Nakov (2013)* ***On the interpretation of noun compounds: Syntax, semantics, and entailment****. Natural Language Engineering, cambridge.org, cited by 33 (6.60 per year)*

*J Lim (2005)* ***The role of power distance and explanation facility in online bargaining utilizing software agents****. Advanced Topics in Global Information Management, books.google.com, cited by 1 (0.08 per year)*

*SY Park, D Sargent, UP Gustafsson, W Li, ... (2011)* ***Versatile video interpretation, visualization, and management system****. US Patent App. 13 ..., Google Patents, cited by 16 (2.29 per year)*

*A Glass (2011)* ***Explanation of Adaptive Systems****., stacks.stanford.edu, cited by 5 (0.71 per year)*

*A Gacek, W Pedrycz (2012)* ***Ecg signal analysis, classification, and interpretation: A framework of computational intelligence****. ... Signal Processing, Classification and Interpretation, Springer, cited by 4 (0.67 per year)*

*CK Chan, S Gesbert, AR Masters, C Xu (2012)* ***Interpreting a plurality of M-dimensional attribute vectors assigned to a plurality of locations in an N-dimensional interpretation space****. US Patent 8,121,969, Google Patents, cited by 10 (1.67 per year)*

*J Ganesh, M Gupta, V Varma (2017)* ***Interpretation of Semantic Tweet Representations****. arXiv preprint arXiv:1704.00898, arxiv.org*

*P Polishchuk, E Mokshyna, A Kosinskaya, ... (2017)* ***Structural, Physicochemical and Stereochemical Interpretation of QSAR Models Based on Simplex Representation of Molecular Structure****. Advances in QSAR ..., Springer*

*AH Bunningen, L Feng, MM Fokkinga, PMG Apers (2007)* ***An Answer Explanation Model for Probabilistic Database Queries****., doc.utwente.nl*

*H Kiendl, T Kiseliova, Y Rambinintsoa (2004)* ***Fuzzy interpretation of music****. HT014601767, pdfs.semanticscholar.org, cited by 1 (0.07 per year)*

*M Bohanec, V Rajkovic (1993)* ***Knowledge-based explanation in multiattribute decision making****. Computer aided decision analysis ..., researchgate.net, cited by 13 (0.52 per year)*

*V Vijendra, M Kulkarni (2017)* ***Fuzzy Controlled ID Interpretation Based ECG Diagnostic Systems****. Advanced Science Letters, ingentaconnect.com*

*I Robertson, H Kahney (1996)* ***The use of examples in expository texts: Outline of an interpretation theory for text analysis****. Instructional Science, Springer, cited by 15 (0.68 per year)*

*I Danov, MA Olsen, C Busch (2014)* ***Interpretation of fingerprint image quality features extracted by self-organizing maps****. Biometric and Surveillance …, spiedigitallibrary.org*

*Z Wang, J Yang (2017)* ***Diabetic Retinopathy Detection via Deep Convolutional Networks for Discriminative Localization and Visual Explanation****. arXiv preprint arXiv:1703.10757, arxiv.org*

*B Zheng (1999)* ***False-Negative Interpretation in a CAD Environment****., dtic.mil*

*D Doyle (2005)* ***A knowledge-light mechanism for explanation in case-based reasoning****., scss.tcd.ie, cited by 5 (0.38 per year)*

*N Xu, W Kosters (2007)* ***Explanation interfaces in recommender systems****. Master's thesis, Leiden University, pdfs.semanticscholar.org, cited by 5 (0.45 per year)*

*R Daneshjou, Y Wang, Y Bromberg, S Bovo, ... (2017)* ***Working towards precision medicine: predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges****. Human …, Wiley Online Library, cited by 1 (1.00 per year)*

*R Dommisse, T Isaksen (2011)* ***Method and system for dynamic, three-dimensional geological interpretation and modeling****. US Patent 7,986,319, Google Patents, cited by 22 (3.14 per year)*

*TA Farmer, M Brown, MK Tanenhaus (2013)* ***Prediction, explanation, and the role of generative models in language processing****. Behavioral and Brain …, cambridge.org, cited by 74 (14.80* per year)*

*D Doyle, A Tsymbal, P Cunningham (2003)* ***A review of explanation and explanation in case-based reasoning****., ai2-s2-pdfs.s3.amazonaws.com, cited by 36 (2.40 per year)*

*MT McKenna, S Wang, TB Nguyen, JE Burns, ... (2012)* ***Strategies for improved interpretation of computer-aided detections for CT colonography utilizing distributed human intelligence****. Medical image …, Elsevier, cited by 16 (2.67 per year)*

*S Todorovic, MC Nechyba* ***Interpretation of Complex Scenes Using Dynamic Tree-Structure Belief Networks****. COMPUTER VISION AND IMAGE …, pdfs.semanticscholar.org*

*A Gottschalk, PD Stein, HD Sostman, ... (2007)* ***Very low probability interpretation of V/Q lung scans in combination with low probability objective clinical assessment reliably excludes pulmonary embolism: data from …****. Journal of Nuclear …, Soc Nuclear Med, cited by 50 (4.55 per year)*

*DH Kim, CF Pieper, A Ahmed, ... (2016)* ***Use and interpretation of propensity scores in aging research: a guide for clinical researchers****. Journal of the …, Wiley Online Library, cited by 5 (2.50 per year)*

*M Green, U Ekelund, L Edenbrandt, J Björk, JL Forberg, ... (2009)* ***Exploring new possibilities for case-based explanation of artificial neural network ensembles****. Neural Networks, Elsevier, cited by 20 (2.22 per year)*

*AL Cechin, E Battistella (2007)* ***The interpretation of feedforward neural networks for secondary structure prediction using sugeno fuzzy rules****. International Journal of Hybrid …, content.iospress.com, cited by 1 (0.09 per year)*

*A Masood (2014)* ***Measuring Interestingness in Outliers with Explanation Facility using Belief Networks****., search.proquest.com, cited by 1 (0.25 per year)*

*T Reps, A Thakur (2016)* ***Automating Abstract Interpretation****. … , Model Checking, and Abstract Interpretation, Springer, cited by 2 (1.00 per year)*

*RKSRC Debra, BMVV Govindaraju* ***Use of Collateral Text in Image Interpretation****. pdfs.semanticscholar.org*

*E Roux, AP Godillon-Maquinghen, ... (2006)* ***A support method for the contextual interpretation of biomechanical data****. IEEE Transactions …, ieeexplore.ieee.org, cited by 7 (0.58 per year)*

*DA Klein (1994)* ***Decision-analytic intelligent systems: automated explanation and knowledge acquisition****., books.google.com, cited by 75 (3.13 per year)*

*J Bowden, W Spiller, F Del-Greco, N Sheehan, ... (2017)* ***Improving the visualisation, interpretation and analysis of two-sample summary data Mendelian randomization via the radial plot and radial regression****. BioRxiv, biorxiv.org, cited by 1 (1.00 per year)*

*T Caelli, WF Bischof (1997)* ***The role of machine learning in building image interpretation systems****. ... journal of pattern recognition and artificial …, World Scientific, cited by 13 (0.62 per year)*

*S Michie, J Thomas, M Johnston, ... (2017)* ***The Human Behaviour-Change Project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation****. …, implementationscience …, cited by 2 (2.00 per year)*

*L Auret, C Aldrich (2012)* ***Interpretation of nonlinear relationships between process variables by use of random forests****. Minerals Engineering, Elsevier, cited by 22 (3.67 per year)*

*A Harvey (2016)* ***Multi-Scale Visualization and Interpretation of Geological Relationships****., qspace.library.queensu.ca*

*W Charemza, D Ladley (2016)* ***Central banks' forecasts and their bias: Evidence, effects and explanation****. International Journal of Forecasting, Elsevier, cited by 1 (0.50 per year)*

*GB Breuer, J Schlegel, P Kauf, ... (2015)* ***The Importance of Being Colorful and Able to Fly: Interpretation and implications of children's statements on selected insects and other invertebrates****. International Journal of …, Taylor & Francis, cited by 10 (3.33 per year)*

*I Blanco Guerrero (2013)* ***Design of an explanation engine for recommender systems****., upcommons.upc.edu*

*G Melioli, C Spenser, ... (2014)* ***Allergenius, an expert system for the interpretation of allergen microarray results****. World Allergy …, waojournal.biomedcentral.com, cited by 15 (3.75 per year)*

*H Lin, X Yang, W Wang (2014)* ***A content-boosted collaborative filtering algorithm for personalized training in interpretation of radiological imaging****. Journal of* *digital imaging, Springer, cited by 8 (2.00 per year)*

*H Dawid, J Dermietzel (2006)* ***How robust is the equal split norm? Responsive strategies, selection mechanisms and the need for economic interpretation of simulation parameters****. Computational Economics, Springer, cited by 11 (0.92 per year)*

*M Kelly, S Michie, J Thomas, M Johnston, ... (2017)* ***The Human Behaviour-Change Project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation****., repository.cam.ac.uk*

*MB Lillehaug (2011)* ***Explanation-aware case-based reasoning****., brage.bibsys.no, cited by 1 (0.14 per year)*

*ML Vaughn, SJ Cavill, SJ Taylor, MA Foy, ... (2004)* ***A Full Explanation Facility for an MLP Network That Classifies Low-Back-Pain Patients and for Predicting MLP Reliability****. Innovations in Intelligent …, Springer, cited by 1 (0.07 per year)*

*M Zhe, RF Harrison, S Cross (1996)* ***Explanation by General Rules Extracted From trained Multi-Layer Perceptrons****., eprints.whiterose.ac.uk*

*J Pearl (2014)* ***Interpretation and identification of causal mediation.****. Psychological methods, psycnet.apa.org, cited by 99 (24.75* per year)*

*D Katić (2016)* ***Situation Interpretation for Knowledge-and Model Based Laparoscopic Surgery****., books.google.com*

*J Suchan, M Bhatt, P Wałęga, C Schultz (2017)* ***Visual Explanation by High-Level Abduction: On Answer-Set Programming Driven Reasoning about Moving Objects****. arXiv preprint arXiv:1712.00840, arxiv.org*

*RM Smith (2014)* ***Collection, Interpretation, and Delivery of Information in Mobile Health Informatics****., search.proquest.com*

*KO Pedersen (2010)* ***Explanation Methods in Clinical Decision Support: A Hybrid System Approach****., brage.bibsys.no*

*A Bikakis, P Caire, K Clark, G Cornelius, J Ma, ... (2016)* ***Proactive Multi-Agent Explanation Generation and Evidence Gathering in a Service Robot Inhabited Assisted Living Environment****. 8-th International …, doc.ic.ac.uk*

*R Halvorsen, S Mazzoni, A Bryn, V Bakkestuen (2015)* ***Opportunities for improved distribution modelling practice via a strict maximum likelihood interpretation of MaxEnt****. Ecography, Wiley Online Library, cited by 21 (7.00 per year)*

*DH Addison (2016)* ***Toward Automated Interpretation of LC-MS Data for Quality Assurance of a Screening Collection****. Journal of laboratory automation, journals.sagepub.com*

*B Sevilla-Villanueva, K Gibert, ... (2013)* ***Clustering and interpretation on real nutritional data****. … CAEPIA'13: Madrid …, upcommons.upc.edu*

*S Nagulendra (2014)* ***Providing awareness, explanation and control of personalized stream filtering in a P2P social network****., pdfs.semanticscholar.org*

*M Velikova, N Ferreira, M Samulski, PJF Lucas, ... (2010)* ***An advanced probabilistic framework for assisting screening mammogram interpretation****. ... in Healthcare 4, Springer, cited by 1 (0.13 per year)*

*S Psarra (2014)* ***Beyond analytical knowledge: The need for a combined theory***

***of generation and explanation**. A/Z ITU Journal of the Faculty of Architecture, discovery.ucl.ac.uk, cited by 6 (1.50 per year)*

*AK Willard (2017) **Agency detection is unnecessary in the explanation of religious belief**. Religion, Brain & Behavior, Taylor & Francis*

*JS Almeida, R Terlevich, E Terlevich, ... (2012) **Qualitative interpretation of galaxy spectra**. The Astrophysical …, iopscience.iop.org, cited by 15 (2.50 per year)*

*M Dalla Preda (2007) **Code obfuscation and malware detection by abstract interpretation**. PhD diss.), http://profs. sci. univr. it/dallapre …, iris.univr.it, cited by 27 (2.45 per year)*

*A Aamodt (1993) **Explanation-driven retrieval, reuse, and learning of cases**. EWCBR-93: First European Workshop on …, pdfs.semanticscholar.org, cited by 23 (0.92 per year)*

*S Todorovic (2005) **Irregular-structure Tree Models for Image Interpretation**., engr.oregonstate.edu*

*JH Mayer (1990) **Explanation-based knowledge acquisition of schemas in practical electronics: A machine learning approach**., dtic.mil, cited by 3 (0.11 per year)*

*PI Nakov, MA Hearst (2013) **Semantic interpretation of noun compounds using verbal and other paraphrases**. ACM Transactions on Speech and Language ..., dl.acm.org, cited by 26 (5.20 per year)*

*VE Kuz'min, PG Polishchuk, AG Artemenko, ... (2011) **Interpretation of QSAR models based on random forest methods**. Molecular ..., Wiley Online Library, cited by 28 (4.00 per year)*

*T Miller (2017) **Explanation in artificial intelligence: Insights from the social sciences**. arXiv preprint arXiv:1706.07269, arxiv.org, cited by 4 (4.00 per year)*

*T Teijeiro, P Félix (2016) **On the adoption of abductive reasoning for time series interpretation**. arXiv preprint arXiv:1609.05632, arxiv.org, cited by 1 (0.50 per year)*

*R Maximini, A Freßmann, M Schaaf (2004) **Explanation service for complex CBR applications**. Advances in Case-Based Reasoning, Springer, cited by 5 (0.36 per year)*

*RM Caprioli, B De Moor, R Van De Plas, ... (2017) **System for interpretation of image patterns in terms of anatomical or curated patterns**. US Patent App. 15 …, Google Patents*

*R Guha, PC Jurs (2004) **Development of linear, ensemble, and nonlinear models for the prediction and interpretation of the biological activity of a set of PDGFR inhibitors**. Journal of Chemical Information and Computer …, ACS Publications, cited by 88 (6.29 per year)*

*G Chen, DG Gillard, M Imhof (2013) **Method for geophysical and geological interpretation of seismic volumes using chronological panning**. US Patent 8,447,524, Google Patents, cited by 10 (2.00 per year)*

*D Jiménez, J Cock, A Jarvis, J Garcia, HF Satizábal, ... (2011) **Interpretation of commercial production information: A case study of lulo (Solanum quitoense), an under-researched Andean fruit**. Agricultural systems, Elsevier, cited by 10*

*(1.43 per year)*

*PT Johnson* ***Educational App for CT Interpretation: Head-to-Toe Measurement Guide****. rsna2015stage.rsna.org*

*DJ Rope, JY Shyr, MJ Vais, ... (2014)* ***Interpretation of statistical results****. US Patent App. 13/656,455, Google Patents, cited by 2 (0.50 per year)*

*LE Fisher, KA Lynch, PA Fernandes, ... (2016)* ***Including sheath effects in the interpretation of planar retarding potential analyzer's low-energy ion data****. Review of Scientific …, aip.scitation.org, cited by 6 (3.00 per year)*

*C VASKE, SAM NG, E PAULL, ... (2015)* ***Integration of Cancer Omics Data into a Whole-Cell Pathway Model for Patient-Specific Interpretation****. Integrating Omics …, Cambridge University Press*

*J Ganesh (2016)* ***Discovery and Interpretation of Embedding Models for Knowledge Representation****., researchweb.iiit.ac.in*

*AM Palczewska (2015)* ***Interpretation, Identification and Reuse of Models. Theory and algorithms with applications in predictive toxicology.****., bradscholars.brad.ac.uk*

*M Gwinn, MA Hlatky, H Janes, P Kraft, S Melillo, ... (2011)* ***Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration****. Eur J Epidemiol, JSTOR, cited by 3 (0.43 per year)*

*DP Jäckle (2017)* ***Projections for Visual Analysis of Multivariate Data: Methods for Identification, Interpretation, and Navigation of Patterns****., kops.uni-konstanz.de*

*F Vial, A Tedder (2017)* ***… the vast Potential of the Data Deluge in small-scale Food-Animal Production Businesses: challenges to Near real-time Data Analysis and interpretation****. Frontiers in Veterinary Science, frontiersin.org*

*SA Niederer, NP Smith (2016)* ***Using physiologically based models for clinical translation: predictive modelling, data interpretation or something in-between?****. The Journal of physiology, Wiley Online Library, cited by 4 (2.00 per year)*

*ACJW Janssens, JPA Ioannidis, S Bedrosian, ... (2011)* ***Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration****. Journal of Clinical …, jclinepi.com, cited by 2 (0.29 per year)*

*V Raskin (2015)* ***Automatic interpretation of a declarative cloud service description****., elib.uni-stuttgart.de*

*Z Allen-Zhu, L Orecchia, ... (2014)* ***A novel, simple interpretation of Nesterov's accelerated method as a combination of gradient and mirror descent****. arXiv preprint arXiv …, pdfs.semanticscholar.org, cited by 6 (1.50 per year)*

*長谷川清, 船津公人 (2011)* ***Visualization and Chemical Interpretation of Multi-Target Structure-Activity Relationships Using SOMPLS****. Journal of Computer Aided Chemistry, jlc.jst.go.jp*

*J Zhang, SP Curley (2017)* ***Exploring explanation effects on consumers' trust in online recommender agents****. International Journal of Human–Computer ..., Taylor & Francis*

*T Dodson, N Mattei, JT Guerin, ... (2013)* ***An English-language argumentation***

***interface for explanation generation with Markov decision processes in the domain of academic advising**. ACM Transactions on ..., dl.acm.org, cited by 10 (2.00 per year)*

*R Pollack **An Intelligent E-Learning System for Beginner Programming-Using Analogical Reminder for Error Classification and Explanation**. pdfs.semanticscholar.org*

*D Fabbri, K Lefevre (2014) **System for explanation-based auditing of medical records data**. US Patent 8,745,085, Google Patents*

*S Ma (2016) **Just In Time Assembly (JITA)-A Run Time Interpretation Approach for Achieving Productivity of Creating Custom Accelerators in FPGAs**., search.proquest.com*

*RJ Doyle (1985) **Construction and Refinement of Justified Causal Models Through Variable-Level Explanation and Perception, and Experimenting**., dspace.mit.edu*

*LM McShane, MM Cavenagh, ... (2013) **Criteria for the use of omics-based predictors in clinical trials: explanation and elaboration**. BMC ..., bmcmedicine.biomedcentral.com, cited by 72 (14.40* per year)*

*M Imhof, G Chen, DG Gillard, P Dimitrov (2012) **Method for geophysical and geological interpretation of seismic volumes in the domains of depth, time, and age**. US Patent 8,213,261, Google Patents, cited by 11 (1.83 per year)*

*FMCRB Gonçalves (2016) **Computer-interpretable guidelines in decision support systems: creation and editing of clinical protocols for automatic Interpretation**., repositorium.sdum.uminho.pt*

*CS Sauer (2016) **Knowledge elicitation and formalisation for context and explanation-aware computing with case-based recommender systems.**., researchgate.net, cited by 2 (1.00 per year)*

*P Christoffersen, S Tulaczyk, ... (2006) **A quantitative framework for interpretation of basal ice facies formed by ice accretion over subglacial sediment**. Journal of ..., Wiley Online Library, cited by 34 (2.83 per year)*

*J Balfer (2015) **Development and Interpretation of Machine Learning Models for Drug Discovery**., d-nb.info*

*M Pohl, J Doppler Haider (2017) **Sense-making Strategies for the Interpretation of Visualizations—Bridging the Gap between Theory and Empirical Research**. Multimodal Technologies and Interaction, mdpi.com*

*S Liu (2017) **Visual Exploration of High-Dimensional Spaces Through Identification, Summarization, and Interpretation of Two-Dimensional Projections**., search.proquest.com*

*M Xu, M Petrou (2011) **3D Scene interpretation by combining probability theory and logic: The tower of knowledge**. Computer Vision and Image Understanding, Elsevier, cited by 17 (2.43 per year)*

*長谷川清, 船津公人 (2010) **Advanced PLS Technique Focusing on Visualization and Chemical Interpretation-SOMPLS Analysis of Serine Protease Inhibitors**. Journal of Computer Aided Chemistry, jlc.jst.go.jp*

*M Jiline (2011) **Annotation Concept Synthesis and Enrichment Analysis: a Logic-***

***Based Approach to the Interpretation of High-Throughput Biological Experiments***., *search.proquest.com*

*CL Bruce (2010)* ***Classification and interpretation in quantitative structure-activity relationships***., *eprints.nottingham.ac.uk*

*KW Hanley, G Hoberg (2016)* ***Dynamic Interpretation of Emerging Systemic Risks***., *papers.ssrn.com, cited by 6 (3.00 per year)*

*R Guha, DT Stanton, PC Jurs (2005)* ***Interpreting computational neural network quantitative structure− activity relationship models: A detailed interpretation of the weights and biases***. *Journal of chemical information ..., ACS Publications, cited by 65 (5.00 per year)*

*H Meyer (2011)* ***Varieties of Capitalism and Environmental Sustainability: Institutional Explanation for differences in Firms' Corporate Environmental Responsibility Reporting across ….***, *papers.ssrn.com, cited by 1 (0.14 per year)*

*KM Burjorjee (2009)* ***Generative fixation: a unified explanation for the adaptive capacity of simple recombinative genetic algorithms***., *search.proquest.com, cited by 10 (1.11 per year)*

*M Mikhnevich, P Hebert (2009)* ***Active interpretation of an object from multi-view and multi-lighting***., *Citeseer*

*J Elith, JR Leathwick (2009)* ***Species distribution models: ecological explanation and prediction across space and time***. *Annual review of ecology, evolution, and ..., annualreviews.org, cited by 2715 (301.67* per year)*

*T Hubauer (2016)* ***Relaxed Abduction: Robust Information Interpretation for Industrial Applications***., *books.google.com, cited by 1 (0.50 per year)*

*M Imhof, G Chen, DG Gillard (2010)* ***Method For Geophysical and Geological Interpretation of Seismic Volumes In The Domains of Depth, Time, and Age***. *US Patent App. 12/623,034, Google Patents*

*G Perichinsky* ***EPISTEMOLOGY TO INVESTIGATE THE SCIENCES-EXPLANATION OF THE COMPUTER SCIENCE: CASE OF INTELLIGENT DATA MINING***. *pdfs.semanticscholar.org*

*EB Generalization (1989)* ***Higher-Order and Modal Logic as a Framework for Explanation-Based Generalization Scott Dietzen Frank Pfenning***., *cs.cmu.edu*

*Y Moriya (2012)* ***Image interpretation report generation apparatus, method and program***. *US Patent App. 13/496,690, Google Patents, cited by 4 (0.67 per year)*

*L Zhou, L Wang, L Liu, P Ogunbona, D Shen (2014)* ***Support vector machines for neuroimage analysis: Interpretation from discrimination***. *Support Vector Machines ..., Springer, cited by 2 (0.50 per year)*

*C Loncaric, S Chandra, C Schlesinger, M Sridharan* ***Tech Report: A Practical Framework for Type Inference Error Explanation***. *homes.cs.washington.edu*

*I Albarrán, P Alonso-González, JM Marin (2017)* ***Some criticism to a general model in Solvency II: an explanation from a clustering point of view***. *Empirical Economics, Springer*

*P Varner (2015)* ***Ophthalmic pharmaceutical clinical trials: interpretation***. *Clinical Investigation, Future Science Ltd London, UK*

*C Prud'homme, X Lorca, N Jussien (2014)* ***Explanation-based large neighborhood search****. Constraints, Springer, cited by 8 (2.00 per year)*

*S Berardi, U de'Liguoro (2009)* ***Toward the interpretation of non-constructive reasoning as non-monotonic learning****. Information and Computation, Elsevier, cited by 19 (2.11 per year)*

*D Eidsvåg (2017)* ***Rhythm interpretation using deep learning neural networks****., brage.bibsys.no*

*M Romero (2009)* ***Supporting human interpretation and analysis of activity captured through overhead video****., search.proquest.com*

*G Hullám, A Gézsi, A Millinghoffer, P Sárközy, ... (2014)* ***Bayesian systems-based genetic association analysis with effect strength estimation and omic wide interpretation: a case study in rheumatoid arthritis****. ... Research: Methods and …, Springer, cited by 1 (0.25 per year)*

*P Kockelman (2017)* ***The Art of Interpretation in the Age of Computation****., books.google.com*

*B Singh, TL Crippen, JK Tomberlin (2017)* ***An introduction to metagenomic data generation, analysis, visualization, and interpretation****. Forensic Microbiology, books.google.com*

*MP Pulido, M Vivarelli (2016)* ***The Identity of the Contemporary Public Library. Principles and Methods of Analysis, Evaluation, Interpretation****., books.google.com*

*M Fiume (2015)* ***System for Interpretation of Personal Genomes****., search.proquest.com*

*B Gaonkar (2015)* ***Converting neuroimaging big data to information: Statistical frameworks for interpretation of image driven biomarkers and image driven disease subtyping****., search.proquest.com*

*J Kocoń, M Marcińczuk (2017)* ***Supervised approach to recognise Polish temporal expressions and rule-based interpretation of timexes †****. Natural Language Engineering, cambridge.org, cited by 1 (1.00 per year)*

*KL Von Tish (2012)* ***Interpretation and clustering of handwritten student responses****., dspace.mit.edu, cited by 2 (0.33 per year)*

*ML Jensen, E Yetgin (2017)* ***PROMINENCE AND INTERPRETATION OF ONLINE CONFLICT OF INTEREST DISCLOSURES.****. MIS Quarterly, aisel.aisnet.org, cited by 1 (1.00 per year)*

*M Xu, J Ren, Z Wang (2017)* ***Chapter Five-Component Identification and Interpretation: A Perspective on Tower of Knowledge****. Advances in Imaging and Electron Physics, Elsevier*

*D Thompson (2008)* ***Sensitive information: An inquiry into the interpretation of information in the workplace from an individual's perspective using qualitative methods****., search.proquest.com, cited by 3 (0.30 per year)*

*K VanLehn, RM Jones, MTH Chi (1992)* ***A model of the self-explanation effect****. The journal of the learning …, Taylor & Francis, cited by 396 (15.23* per year)*

*AJ Urbanowicz (1998)* ***Interpretation of anaphoric expressions in the Lolita*** *__system__., etheses.dur.ac.uk*

*M Brown, T Coughlan, G Lawson, R Mortier, ... (2012) __Intergenerational interpretation of the Internet of Things__., eprints.nottingham.ac.uk, cited by 1 (0.17 per year)*

*E Mustafa (2014) __Sign Language Interpretation using Kinect__., researchgate.net*

*R Neches, WR Swartout, ... (1985) __Enhanced maintenance and explanation of expert systems through explicit models of their development__. IEEE Transactions on …, ieeexplore.ieee.org, cited by 211 (6.39 per year)*

*R Maffei, LS Convertini, S Quatraro, S Ressa, ... (2015) __Contributions to a neurophysiology of meaning: the interpretation of written messages could be an automatic stimulus-reaction mechanism before becoming …__. PeerJ, peerj.com*

*A Niroula, M Vihinen (2016) __Variation interpretation predictors: principles, types, performance, and choice__. Human mutation, Wiley Online Library, cited by 26 (13.00* per year)*

*R Kaczorowski, G West, S McArthur (2008) __Automated Interpretation of Boiler Feed Pump Vibration Monitoring Data__., esru.strath.ac.uk*

*R Lehavy (2016) __Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach__., deepblue.lib.umich.edu*

*VA Huynh-Thu (2012) __Machine learning-based feature ranking: statistical interpretation and gene network inference__., orbi.ulg.ac.be, cited by 3 (0.50 per year)*

*D Todorov (2013) __Enhanced interpretation of the mini-mental state examination__., orca.cf.ac.uk, cited by 1 (0.20 per year)*

*WR Foster, DG Robertson, BD Car (2013) __The Application of Toxicogenomics to the Interpretation of Toxicologic Pathology__. Haschek and Rousseaux's Handbook …, Elsevier*

*L Wehenkel (2011) __Machine learning-based feature ranking: Statistical interpretation and gene network inference__., orbi.ulg.ac.be*

*DD Suthers (1993) __An analysis of explanation and its implications for the design of explanation planners__., researchgate.net, cited by 23 (0.92 per year)*

*P Hu (2012) __A Stochastic Approximation Interpretation for Model-based Optimization Algorithms__., search.proquest.com*

*B Micenková (2015) __Outlier Detection and Explanation for Domain Experts__., pure.au.dk*

*A St-Onge (2004) __The interpretation of dairy data using interactive visualization__., pdfs.semanticscholar.org, cited by 1 (0.07 per year)*

*OZ Khan (2013) __Policy Explanation and Model Refinement in Decision-Theoretic Planning__., uwspace.uwaterloo.ca*

*D Addison __EVALUATING DATA MINING APPROACHES FOR THE INTERPRETATION OF LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY__. studentnet.cs.manchester.ac.uk*

*A Vilic (2017) __Vital Signs Monitoring and Interpretation for Critically Ill__*

***Patients***., *orbit.dtu.dk*

*S Ali, S Khusro, I Ullah, A Khan, I Khan (2017)* ***SmartOntoSensor: Ontology for Semantic Interpretation of Smartphone Sensors Data for Context-Aware Applications***. *Journal of Sensors, hindawi.com, cited by 2 (2.00 per year)*

*PA Russell (1990)* ***“Intelligence” as description and as explanation***. *Behavioral and Brain Sciences, cambridge.org*

*S Chaumette, O Ly, R Tabary (2011)* ***Automated extraction of polymorphic virus signatures using abstract interpretation***. *Network and System Security ..., ieeexplore.ieee.org, cited by 8 (1.14 per year)*

*J Wagner (2015)* ***Social Signal Interpretation: Building Online Systems for Multimodal Behaviour Analysis***., *opus.bibliothek.uni-augsburg.de*

*DM DeCoste (1990)* ***Dynamic across-time measurement interpretation: maintaining qualitative understandings of physical system behavior***., *dtic.mil, cited by 9 (0.32 per year)*

*ADAN Selvan (2007)* ***Hierarchical clustering-based segmentation (HCS) aided diagstic image interpretation monitoring***., *search.proquest.com, cited by 5 (0.45 per year)*

*MJF Grimnes (1998)* ***ImageCreek: A knowledge level approach to case-based image interpretation***. *NTNU-PhD Thesis, idi.ntnu.no, cited by 3 (0.15 per year)*

**Query report 05 – “interpret” AND “explain”**

# *(intitle:interpret OR intitle:explain) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning") to 2017*

*Publish or Perish 6.21.6145.6594*

## *Search terms*

***All of the words:*** *(intitle:interpret OR intitle:explain) AND (intext:transparency OR intext:black-box OR intext:"black box" OR intext:blackbox OR intext:opacity OR intext:"deep models" ) AND (intext:"machine learning")*
***Years:*** *earliest to 2017*

## *Data retrieval*

***Data source:*** *Google Scholar*
***Query date:*** *21/01/2018 12:21:40*
***Cache date:*** *21/01/2018 12:22:59*
***Query result:*** *[0] The operation completed successfully.*

## *Metrics*

***Publication years:*** *1997-2017*
***Citation years:*** *21 (1997-2018)*
***Papers:*** *41*
***Citations:*** *1428*
***Citations/year:*** *68.00*
***Citations/paper:*** *34.83 (*count=4)*
***Citations/author:*** *957.99*
***Papers/author:*** *18.46*
***Authors/paper:*** *3.05/3.0/2 (mean/median/mode)*
***Age-weighed citation rate:*** *200.48 (sqrt=14.16), 123.49/author*
***Hirsch h-index:*** *11 (a=11.80, m=0.52, 1381 cites=96.7% coverage)*
***Egghe g-index:*** *37 (g/h=3.36, 1428 cites=100.0% coverage)*
***PoP hI,norm:*** *6*
***PoP hI,annual:*** *0.29*

## *Results*

*D Baehrens, T Schroeter, S Harmeling, ... (2010)* ***How to explain individual classification decisions****. ... of Machine Learning …, jmlr.org, cited by 124 (15.50** *per year)*

*RAV Rossel, T Behrens (2010)* ***Using data mining to model and interpret soil diffuse reflectance spectra****. Geoderma, Elsevier, cited by 373 (46.63* per year)*

*C Lustig, K Pine, B Nardi, L Irani, MK Lee, ... (2016)* ***Algorithmic authority: the ethics, politics, and economics of algorithms that interpret, decide, and manage****. Proceedings of the …, dl.acm.org, cited by 10 (5.00 per year)*

*G Shmueli (2010)* ***To explain or to predict?****. Statistical science, projecteuclid.org, cited by 649 (81.13* per year)*

*Y Guo, B Selman (2007)* ***Exopaque: A framework to explain opaque machine learning models using inductive logic programming****. Tools with Artificial Intelligence, 2007. ICTAI …, ieeexplore.ieee.org, cited by 3 (0.27 per year)*

*J Krause, A Perer, E Bertini (2016)* ***Using Visual Analytics to Interpret Predictive Machine Learning Models****. arXiv preprint arXiv:1606.05685, arxiv.org, cited by 6 (3.00 per year)*

*CR MILARÉ, ACP DE LF DE CARVALHO, ... (2002)* ***An approach to explain neural networks using symbolic algorithms****. International Journal …, World Scientific, cited by 7 (0.44 per year)*

*R Blanco, D Ceccarelli, C Lucchese, R Perego, ... (2012)* ***You should read this! let me explain you why: explaining news recommendations to users****. Proceedings of the 21st …, dl.acm.org, cited by 11 (1.83 per year)*

*U Johansson, C Sönströd, ... (2011)* ***One tree to explain them all****. … Computation (CEC), 2011 …, ieeexplore.ieee.org, cited by 4 (0.57 per year)*

*JAB Tinoco, AG Correia, ... (2012)* ***Application of a sensitivity analysis procedure to interpret uniaxial compressive strength prediction of jet grouting laboratory formulations performed by SVM model****. ISSMGE-TC 2i 1 …, repositorium.sdum.uminho.pt, cited by 4 (0.67 per year)*

*Z Yuanhui, L Yuchang, S Chunyi (1997)* ***Using learning and searching approach to explain neural network with distributed representations****. Systems, Man, and …, ieeexplore.ieee.org, cited by 1 (0.05 per year)*

*DC Richardson, SJ Melles, RM Pilla, AL Hetherington, ... (2017)* ***Transparency, geomorphology and mixing regime explain variability in trends in lake temperature and stratification across Northeastern North America (1975–2014)****. Water, mdpi.com, cited by 3 (3.00 per year)*

*D Baehrens, T Schroeter, S Harmeling, ... (2009)* ***How to explain individual classification decisions****. arXiv preprint arXiv …, arxiv.org, cited by 3 (0.33 per year)*

*M Beillevaire* ***Inside the Black Box: How to Explain Individual Predictions of a Machine Learning Model****. nada.kth.se*

*S Harmeling, M Kawanabe, KR Müller* ***How to Explain Individual Classification Decisions****.*

*S Penkov, S Ramamoorthy (2017)* ***Using Program Induction to Interpret Transition System Dynamics****. arXiv preprint arXiv:1708.00376, arxiv.org*

*S Penkov, S Ramamoorthy (2017)* ***Program Induction to Interpret Transition***

***Systems**., openreview.net*

*R Blanco, D Ceccarelli, C Lucchese, R Perego, ... (2012)* ***You Should Read This! Let Me Explain You Why****., pdfs.semanticscholar.org*

*R Kamimura (2011)* ***Information-Theoretic Approach to Interpret Internal Representations of Self-Organizing Maps****. Self Organizing Maps-Applications and Novel …, intechopen.com*

*S Grant, G Schymik (2014)* ***Using Work System Theory to Explain Enterprise Search Dissatisfaction****. Proceedings of the Information Systems …, proc.edsig.org, cited by 1 (0.25 per year)*

*Y Sushko, S Novotarskyi, R Körner, J Vogt, ... (2014)* ***Prediction-driven matched molecular pairs to interpret QSARs and aid the molecular optimization process****. Journal of …, Springer, cited by 22 (5.50 per year)*

*TY Lee, A Smith, K Seppi, N Elmqvist, ... (2017)* ***The human touch: How non-expert users perceive, interpret, and fix topic models****. International Journal of …, Elsevier, cited by 3 (3.00 per year)*

*P Bercher, S Biundo, T Geier, T Hoernle, F Nothdurft, ... (2014)* ***Plan, Repair, Execute, Explain-How Planning Helps to Assemble your Home Theater.****. ICAPS, aaai.org, cited by 29 (7.25 per year)*

*L Hartert, MS Mouchaweh (2010)* ***A hybrid multi-classifier to characterize and interpret hemiparetic patients gait coordination****. Machine Learning and Applications …, ieeexplore.ieee.org, cited by 1 (0.13 per year)*

*D Li* ***Using agent based method to Explain Forward Discount Bias (FDB) puzzle in FX market—A MATLAB Application****. pdfs.semanticscholar.org*

*SG Baker, E Schuit, EW Steyerberg, ... (2014)* ***How to interpret a small increase in AUC with an additional risk prediction marker: decision analysis comes through****. Statistics in …, Wiley Online Library, cited by 18 (4.50 per year)*

*C Manescuc, C Staricad* ***Do Corporate Social Responsibility scores explain and predict firm profitability? A case study on the publishers of the Dow Jones Sustainability Indexes ab****. pdfs.semanticscholar.org*

*J Lawrence, J Park, K Budzynska, C Cardie, ... (2017)* ***Using argumentative structure to interpret debates in online deliberative democracy and eRulemaking****. ACM Transactions on …, dl.acm.org*

*M Digits* ***DimReader: Using auto-differentiation to explain non-linear projections****. issues, hdc.cs.arizona.edu*

*R Faust, C Scheidegger (2017)* ***DimReader: Using auto-differentiation to explain non-linear projections****. arXiv preprint arXiv:1710.00992, arxiv.org*

*D DeVault, M Stone (2009)* ***Learning to interpret utterances using dialogue history****. Proceedings of the 12th Conference of the …, dl.acm.org, cited by 21 (2.33 per year)*

*NP Gibson, S Aigrain, JK Barstow, ... (2013)* ***The optical transmission spectrum of the hot Jupiter HAT-P-32b: clouds explain the absence of broad spectral features?****. Monthly Notices of …, academic.oup.com, cited by 55 (11.00* per year)*

*BJ Culpepper (2011)* ***Learned Factorization Models to Explain Variability in***

***Natural Image Sequences***., *search.proquest.com*

*K Unger, L Ackerman, CH Chatham, D Amso, D Badre (2016)* ***Working memory gating mechanisms explain developmental change in rule-guided behavior***. *Cognition, Elsevier, cited by 1 (0.50 per year)*

*R Guha, PC Jurs (2004)* ***Development of QSAR models to predict and interpret the biological activity of artemisinin analogues***. *Journal of chemical information and computer …, ACS Publications, cited by 68 (4.86 per year)*

*M Lang, R Kundt (2017)* ***Can predictive coding explain past experiences?***. *Religion, Brain & Behavior, Taylor & Francis*

*Y Tang, A Gupta, S Garimalla, MR Galinski, ... (2017)* ***Metabolic modeling helps interpret transcriptomic changes during malaria***. *... et Biophysica Acta (BBA ..., Elsevier*

*JB Bemmels, CW Dick (2017)* ***Widespread southern forests during the Last Glacial Maximum, not refugia, explain genetic structure of two eastern North American hickory species***. *bioRxiv, biorxiv.org*

*EA Dieckman (2014)* ***Use of Pattern Classification Algorithms to Interpret Passive and Active Data Streams from a Walking-Speed Robotic Sensor Platform***., *search.proquest.com*

*NP Gibson, S Aigrain, JK Barstow, TM Evans, ...* ***The optical transmission spectrum of the hot Jupiter HAT-P-32b: clouds explain the absence of broad spectral***. *inspirehep.net*

*J Moya-Laraño, JR Bilbao-Castro, ... (2014)* ***Eco-evolutionary spatial dynamics: rapid evolution and isolation explain food web persistence***. *Advances in ..., books.google.com, cited by 11 (2.75 per year)*