Tiia Haverinen

Towards Explainable Artificial Intelligence (XAI)

Master's Thesis in Information Technology

July 13, 2020

University of Jyväskylä

Faculty of Information Technology

Author: Tiia Haverinen

Contact information: tiia.k.haverinen@student.jyu.fi

Supervisor: Jussi Hakanen, Sami Äyrämö

Title: Towards Explainable Artificial Intelligence (XAI)

Työn nimi: Kohti selittävää tekoälyä

Project: Master's Thesis

Study line: Information Technology

Page count: 71+0

Abstract: In the 21st century, the applications of artificial intelligence (AI) have achieved great performance in various tasks. Large datasets, increasing computational power and more complex machine learning models have made it possible. Unfortunately, these complex models are often only black boxes to human users and the user has difficulties to understand and trust the outcomes of AI systems. There has been a great amount of research in the field of explainable artificial intelligence (XAI) to develop methods that increase the explainability of AI systems. In addition to a literature review of the research in XAI, the present thesis includes a small project in which the parameters of an ECR ion source have been surveyed via simple machine learning methods in order to find the optimal parameters for the maximal ion beam intensity.

Keywords: artificial intelligence, explainable artificial intelligence, machine learning, explainability, interpretability, ion sources

Suomenkielinen tiivistelmä: 2000-luvun aikana tekoälysovellukset ovat saavuttaneet erinomaisen suorituskyvyn useissa eri tehtävissä. Suuret datajoukot, kasvava laskennallinen teho sekä yhä monimutkaisemmat koneoppimismallit ovat mahdollistaneet sen. Valitettavasti nämä monimutkaiset mallit ovat usein vain mustia laatikoita ihmiskäyttäjille ja käyttäjällä on vaikeuksia ymmärtää ja luottaa tekoälysysteemin lopputuloksiin. Selittävän tekoälyn osa-alueella on ollut suuri määrä tutkimusta sellaisten menetelmien kehittämiseksi, jotka lisäisivät tekoälysysteemien selittävyyttä. Tämä opinnäytetyö sisältää sekä kirjallisuuskatsauksen selittävän tekoälyn tutkimuksesta että kokeilun, jossa kartoitettiin yksinkertaisilla tekoälymenetelmillä ECR-ionilähteen optimaalisia parametreja maksimaaliselle ionisuihkun intensiteetille.

Avainsanat: tekoäly, selittävä tekoäly, koneoppiminen, selittävyys, ymmärrettävyys, ionilähteet

Preface

The writing process of this MSc thesis reflects in a wonderful way my entire journey at the faculty of Information Technology. Never thoroughly planned but proceeded when the time was right. This thesis was never mandatory, it was just the outcome of my interest. All of my hobbies do not end up to be wrapped up in a MSc thesis, but apparently this can also happen.

I want to express my gratitude to my two patient supervisors, Dr. Jussi Hakanen and Dr. Sami Äyrämö, who provided their excellent professional help but also some peace to my mind whenever needed. Thank you for helping me to complete this piece of work!

A big *thank you* goes also to the faculty of Information Technology. Since my M.Sc. degree was always a serious leisure time activity, it could have been much more difficult to handle my studies, if the faculty did not allow me to study almost without any necessity of physical attendance or other constraints. Thank you for providing me the freedom to study wherever and whenever I wanted.

I want to thank my wonderful colleagues at Gofore for all the support. I want to also thank all my lovely friends, family and especially Miha: Thanks to you all, writing theses is not the only activity I can enjoy outside the office hours.

Jyväskylä, July 13, 2020

Tiia Haverinen

List of Figures

List of Tables

Contents

1 Introduction

The need for explainable artificial intelligence (XAI) is real. Artificial intelligence (AI) has spread everywhere from various applications to wide coverage in media and it plays a significant role in our society. AI has been seen as a powerful and useful tool in many ways, but it has also induced fear as a dangerous weapon to destroy human thinking, steal people's jobs and create mass unemployement (Enqvist 2018; Kissinger 2018; Ford and Colvin 2015; Kaplan 2016). Several world-famous researchers have spoken their mind without hanging back with verbal expressions. In May 2014, English theoretical physicist Stephen Hawking wrote a letter together with other scientists about the risks of AI stating AI can be the best or the last achievement of human race if we are not careful (Hawking et al. 2014). In February 2017, billionaire Elon Musk has stated that humans need to evolve and merge with machines and we have to find new jobs for those people who will lose their jobs for AI - and we have to do it fast, since the changes will be very quick and disruptive (Kharpal 2017).

The risks and fear of mass unemployment are not purely artificial. The capability of AI has been already demonstrated in various applications such as in the form of recognition of speech, playing strategic games, content recommendation (e.g. Facebook, Netflix) and medical diagnosing, to mention some examples. When applications are good enough to replace human work, as a consequence, changes in the job market will take place. The reformation of labor market becomes unavoidable. Furthermore, the job market will not be the only sector of human lives which will undergo big changes. Recent advances in content recommendation and generation of fake content will have a huge social impact by affecting the way people make choices and what kind of content they see in social media. The practical applications have created the need to educate people on artificial intelligence, and provided courses have gathered wide public interest in Finland (Laakkonen 2018; Tiainen 2018).

However, the fear of the new should not be an excuse to explore the unknown. Throughout the history industrial applications have changed the job market and the way of living. Even at the moment there are myriad examples how AI can be a beneficial game changer. AI can be used as an unparalleled tool for fighting famine (Holley 2018), it can help to monitor the food waste in the form of an intelligent bin (Anthony 2019), and it can count the number of

T cells on the digital photograph, thus being a tool in cancer diagnoses as the recent pilot study of University of Jyväskylä suggests (Jyväskylän yliopisto 2018).

Without a doubt, AI and the models associated with it are widely used and their usage is still getting more and more common in science and industry. While these models are becoming more complex and they give predictions with convincing accuracy, transparency is easily lost in the complexity of model, and as a consequence, too often the models are only black boxes to their users. Before the power of these new tools can be released, the models and methods must be known to be reliable. They must be worth of trust.

Trust can be defined or measured in different ways, but trust is always related to the question how much the users understand the model. In order to strengthen the trust on the models, different methods to explain the models and predictions are needed. That is why we need explainable artificial intelligence and we need to understand what kind of techniques have already been implemented on that research field.

On the simplest level the explanation can be external textual or visual information which highlights the facts that lead to the prediction given by the model. Some tools to explanation techniques have already been implemented. In addition to the spontaneously increased interest in explanations, the topic became well-grounded also in the juridical point of view. Thanks to European Union and General Data Protection Regulation (GDPR) (Goodman and Flaxman 2016), the context of a right to explanation has been under wide discussion very recently.

This thesis aims to provide the background of XAI and the summary of latest achievements in the field in the form of literature review. The following research questions are considered: What kind of explanation techniques have already been implemented in the field of XAI? Are the present techniques designed for a specific AI method or can they be applied generally? In addition to the literature review, the parameter space of an electron cyclotron resonance ion source (ECRIS) is studied by applying simple machine learning techniques while paying attention on the explainability of the obtained results, to give a practical example. The main goal of the ECR project is to find out if the parameter space of the ion source can be studied via machine learning methods.

To begin, we wrap-up the history of artificial intelligence and the latest achievements in XAI in Chapter 2. Next, the simple AI methods introduced in Chapter 3 are applied on one case study in Chapter 4. As the practical example we solve the optimization problem of the parameters of an ECRIS. The results of the literature review and the case study are discussed in the end of the corresponding Chapters. Finally, the conclusions of the whole project are provided in Chapter 5.

2 Towards explainable artificial intelligence

In this chapter the development of artificial intelligence (AI) to explainable artificial intelligence (XAI) is discussed. We start from the definition of artificial intelligence and discuss its history briefly. Since machine learning is used almost as a synonym for AI nowadays, the connection between machine learning, neural networks and AI is clarified. In the following, the shortcomings of AI are considered, the key terminology of XAI is introduced and different techniques of explainable artificial intelligence are wrapped up.

2.1 Definition of artificial intelligence

Artificial intelligence refers to intelligence of machines. Sometimes it is called equivalently *computational intelligence*. The research of AI covers the study of intelligent behaviour and intelligent *agents*, devices acting in such a way that the act optimally leads to an achievement of a preconceived goal, while the device observes surrounding (data) environment and learns from the observations (Poole, Mackworth, and Goebel 1998; Kaplan 2016).

The exact definition of AI has varied during the past decades, and it still depends on the context and the person who is defining it. John McCarthy, the father of AI, introduced the term artificial intelligence in 1955 to describe the idea of developing machines that behave as they were intelligent. Nowadays, in addition to the research on the field, the term artificial intelligence can also refer to a computer or a computer program that is capable to make intelligent actions (Wikipedia 2018) or, as B. Copeland (2018) defines it in Encyclopædia Britannica, AI is *the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.* Even though the aim of AI systems has always been to simulate intelligent behavior, many of the applications are working on tasks that are not thought to require vast intelligence from a human. For example, identification of an object on a photo is not considered to be an intellectually demanding mission for a human, in contrast to the computational world (Garnham 1988).

Some references define AI as *whatever computers cannot do yet*. The aforesaid definition is flexible and reflects the fact that some AI problems of the past are not considered to be in the field of AI any longer. One of the recent examples is optical character recognition which was removed from the list of AI things and considered to be a routine technology at present. This means that some applications generally considered to be AI at the moment, such as understanding human speech or autonomously driving cars, will most likely become non-AI one day.

One way to define AI is to do it through the key features. Two main properties characteristic to AI are *autonomy* and *adaptivity*. The first one describes the independence of AI to perform tasks without a human's guidance, the latter describes the capability of AI to enhance its performance via learning. Thus AI could be defined as autonomous and adaptive acting performed by a non-human being.

To conclude, the definition of AI is not fixed in general. Loosely speaking and context-dependently, AI can refer to the intelligence of machines or an intelligent computer program. In this thesis, we refer to AI systems as intelligent computer programs that are capabable to perform tasks without a user's continuous guidance.

2.2 The brief history of AI

One of the first steps towards machine learning and artificial intelligence was taken by British computer scientist and mathematician Alan Turing. During the Second World War, Turing and his collaborators worked on the *Bombe* machines to crack *Enigmas* that were used by the German army to send secured messages. The both machines, Enigma and Bombe, gave a start for sophisticated computers and computer programs. In the mid-20th century Turing (1950) wrote his article *Computing Machinery and Intelligence* that became a classic in the field of machine learning. In the paper Turing proposes a method to test a machine's ability to behave human-likely. The so-called Turing test is based on the idea that a computer is intelligent if it gives responses which cannot be distinguished from the ones given by human beings. The test is inspired by the party game called *the imitation game*.

As illustrated in Figure 1, the original Turing test consists of a human examinator (C), a computer (A) and another human (B). The examinator is trying to find out which one of A and B is a computer and which one is a human by asking questions and receiving answers in a writ-

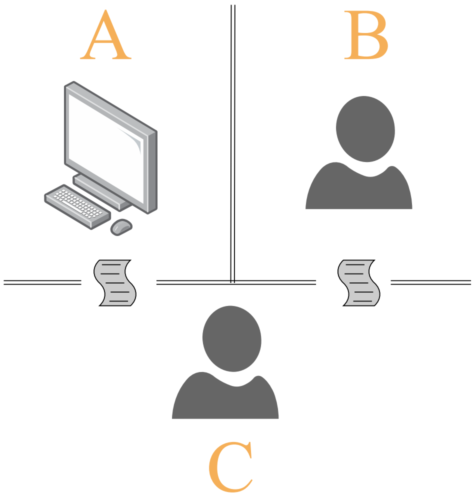


Figure 1: Illustration of the Turing test. Examinator C tries to determine which one of A and B is a computer and which one is a human based on the written responses given by A and B.

ten form. Because the test is performed in a written form, the test does not require (highly) developed communication systems as speech-generating devices. Even though the original test consisted of three attendees, the test is most often performed with one examinator and one answerer in practice.

The scientific collaboration of AI research was founded in the mid 1950s, when John McCarthy invited a group of researchers to develop the concepts around ”thinking machines” at Darthmouth College. Several participants became significant contributors in the research field of AI. The funding proposal written by McCarthy reflects in an outstanding way what kind of expectations and beliefs the first generation of AI researchers had. McCarthy seemed to believe that a computer could simulate basically all the cognitive functions of human beings - for example a computer program could be able to perform self-improvement. On the other hand, McCarthy was also too optimistic when estimating the amount of work intelligent computer programs would need: he wrote that remarkable advancement can be made in a summer if the researcher group is selected carefully (Kaplan 2016; Garnham 1988).

Even though the wild human mind had conceived stories about intelligent artificial beings already in ancient times, these brave ideas of intelligent artificial beings had to wait for programmable digital computers till the 1950s before they could be implemented. In the following decades, after the Dartmouth summer conference, the field of AI research experienced success but also hard financial times called as AI winters. For the sake of scientific interest, but also to gather general interest, various computer programs were implemented to beat humans in chess and other intellectual games. A computer program beat a human in chess for the first time in 1956, and in 1997 the updated version of Deep Blue defeated the grandmaster, Garry Kasparov. Even though Deep Blue - Kasparov games got a great amount of general publicity, there was a shortage of funding from the mid 70s to the mid 90s. At that time, the major problem for AI applications to be successful was the needed amount of data. The existing computers were not capable to handle such large data sets. When the computers were developed enough, the general interest in AI increased explosively after successful machine learning applications in the beginning of 21st century (Kaplan 2016).

AI is used in various different applications, e.g. in search engines, medical diagnoses, e-mail filtering, image recognition, targeted advertisements, face identification of cameras and self-driving cars. One of the latest advances in the AI world is the capability of AI systems to beat real humans in difficult strategic games such as Go (Borowiec 2018). In March 2016 AlphaGo, an AI system of Google DeepMind, beat 18-time world champion Lee Sedol in a five-game Go match 4–1. According to Go professionals, at least one unexpected but successful move was played by AlphaGo during the match, which demonstrates the capability of learning new things. Unfortunately, it is unknown how AlphaGo dediced to play that specific victorious move. In this context it is not crucial to know the logic, but in other applications this lack of transparency may be a substantial hindrance.

2.3 Machine learning and artificial neural networks

Nowadays the term *machine learning* (ML) is used almost as a synonym for AI. However, machine learning is not precisely equal to AI, but it is rather a subset of AI as illustrated in Figure 2. An AI system can be created without machine learning algorithms. Machine learning algorithms need a mathematical model in order to give predictions, and in contrast, massive ruled-based systems predicting outputs are AI systems but without ML, since there is no trained mathematical model.

However, machine learning is a key ingredient of artificial intelligence nowadays. Machine

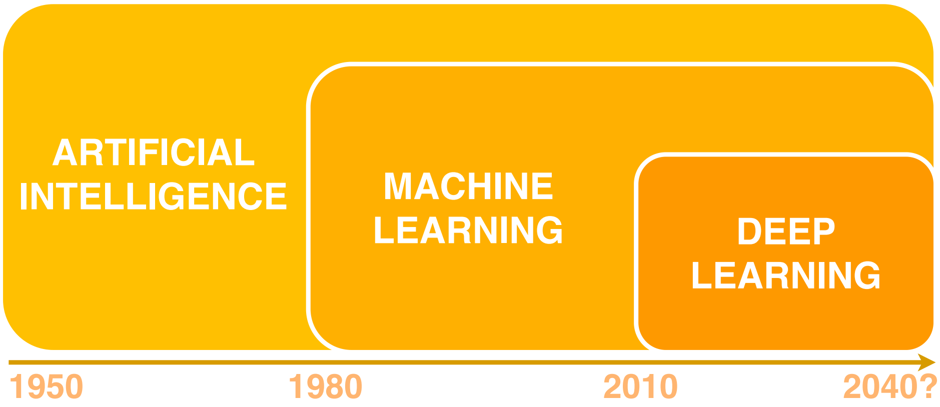


Figure 2: Machine learning and deep learning are subsets of artificial intelligence. The concept of artificial intelligence was launched in the 50s, machine learning methods have been developed since the 80s and deep learning approaches became common in the 2010s. Figure adapted from the blog written by M. Copeland (2016).

learning describes the science of making computer systems to learn and improve their learning autonomously by providing real-world data. In this context, learning means improvement of performance on a certain task, which is achieved by applying statistical techniques. The research field of ML covers the study, construction and implementation of algorithms that are able to learn from data but also make predictions on data. Thus ML is convenient in problems which would be infeasible to solve by explicit rules-based programming. One such an example is email filtering.

Machine learning algorithms can be divided in three subcategories: *(semi-)supervised, unsupervised* and *reinforcement learning*, which are illustrated in Figure 3.

In *supervised* or *semi-supervised learning* algorithms build a mathematical model of a set of labelled data with known inputs and corresponding outputs (Nilsson 1998). The raw data is divided in two parts of which the first part is used to train the algorithm and the other part is used for testing the trained algorithm. Each training example consists of one or more inputs and corresponding desired outputs. However, in semi-supervised learning algorithms some of the training examples do not have a desired output. Supervised learning algorithms are *task driven*, which means that the aim of model usage is to give predictions. There are two main types of supervised learning: classification and regression. The former, classification,

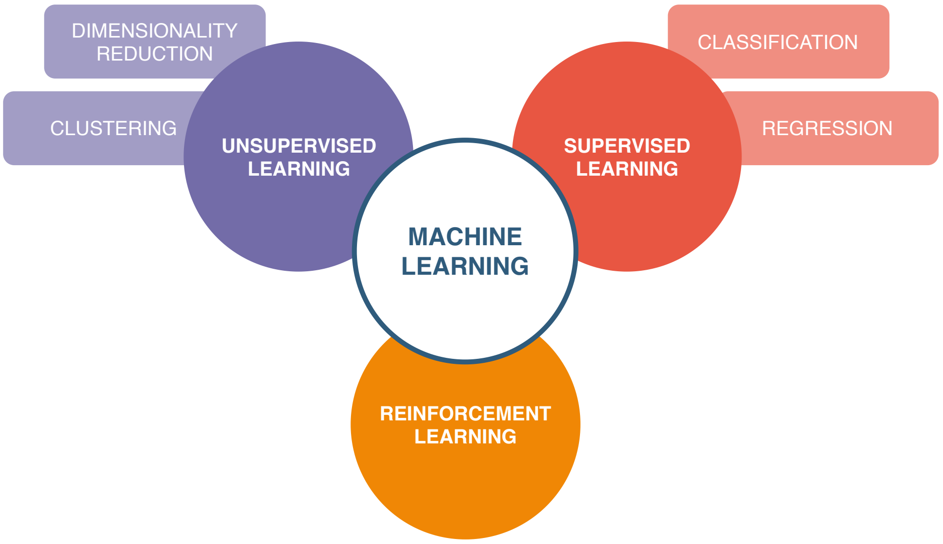


Figure 3: Illustration of machine learning categories. Machine learning algorithms can be subcategorized into three different main groups, namely unsupervised, supervised and reinforcement learning. Unsupervised learning and supervised learning can be divided further into two subsets.

is used when the outputs have a limited set of values (”a class”), and the later, regression, is used when the outputs may have any numerical value that may lay within a range.

The *unsupervised learning* algorithms find a structure in the data by taking a set of data that contains only inputs (Nilsson 1998). The data is not labelled, but the algorithm identifies commonalities in the data and the learned structure i.e. the output can be a grouping or a clustering of the input points. In fact, one of the two main classes in unsupervised learning is cluster analysis. The other main class is principal component analysis, that transforms a set of correlated variables into a set of linearly uncorrelated ones. The method is used for example to visualize relationships between populations.

The third category, *reinforcement learning* algorithms, covers the goal-oriented algorithms that learn from the feedback (Nilsson 1998). Software agents (computer programs) explore the environment and they take actions in order to maximize a reward that is often immediate and is related to the latest transition. The problem environment is typically modelled as a Markov decision process – as a discrete and stochastic process.

Artificial neural networks (ANN), or shortly neural networks (NN), are machine learning algorithms inspired by biological neural networks that can be found e.g. in human brains (Suzuki 2011). Like humans, these artificial systems learn by doing and examining examples. No task-specific rules are implemented in the algorithms. For instance, in image recognition an ANN system may learn to identify handwritten figures on a photo after examining the provided training data. The training data may include a bunch of photos of handwritten figures and the figures in text format. ANNs are widely used in AI applications at the moment.

An ANN system is illustrated in Figure 4. It is a collection of processing units named as *artificial neurons*, and they are transmitting *signals* through connections called *edges*. The strength of the signal is varied by an adjustable *weight* and it reflects the importance of the connection. The signal itself is commonly a real number in ANN implementations. Each neuron is receiving and sending information to several other neurons, thus having several edges.

The neurons of the network are usually categorized in multiple layers. The activity of ANN starts from the first one, *input layer*, and the signals travel through (multiple) *hidden layers* before the signals reach the last layer called *output layer*. Each neuron applies a layer-specific transformation on the input the neuron receives.

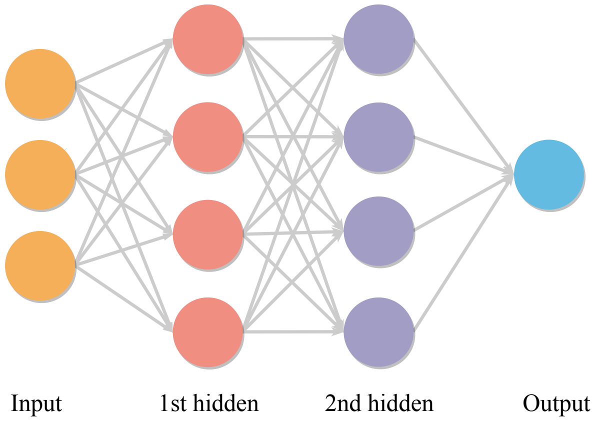


Figure 4: An example of an artificial neural network. ANN is a collection of neurons (circles) that are connected to each other by edges (arrows). The neurons are grouped into layers (colors) that are referred as input, hidden and output layers depending on its position in the process.

2.4 Deficiences in artificial intelligence – why are explanations needed?

Explainable artificial intelligence (XAI) and its importance was already noted many decades ago (Shortliffe and Buchanan 1975; Chandrasekaran, Tanner, and Josephson 1989; Buchanan and Shortliffe 1984; Swartout and Moore 1993). Great success in machine learning has opened many doors for applications of artificial intelligence, but also its shortcomings have become more visible. Recent advances in applications of critical fields, such as medicine and government, pointed out the crucial problems with trust. One of the biggest limitations of AI systems is the lack of ability to explain why a certain decision was made (Došilović, Brcic, and Hlupic 2018). Despite the long history and a significant amount of work, making up good explanations is not trivial. Since the power of AI systems stems from millions of parameters and they become more and more complex, and the models are acting and deciding more and more independently, the actions made by AI systems become easily even less understandable to human users (Gunning 2018; Biran and Cotton 2017). Most often the smartness and transparency of AI are contradictory. However, as long as reasonable explanations are missing, the full capacity of AI applications cannot be reached.

In practice, AI systems are designed to find an optimal model to satisfy a certain goal by using training data. For example, the goal could be to maximize the accuracy of identifying if a tumor is benign or malignant based on the given data. The AI may generate a set of useful and generalizable rules such as ”benign tumors have a slower growth rate”, but the AI may also learn false rules. That can happen especially if the used training data includes some inappropriate connections, for example if the diagnosis of a tumor and its ID number in the training data are connected. If these false rules are learned and then extrapolated in real-life data, consequences may be severe. However, if the AI systems could be able to explain their decisions, these kind of false rules could be more easily spotted and the model itself could be better trusted.

However, the question if there is a need for interpretability is worth of discussion as well. Some researchers regard that the need for interpretability depends on the case. For example, Doshi-Velez and Kim (2018) see that there is no requirement for interpretability of ad servers, postal code sorting systems or even air craft collision avoidance systems, since the output of ML system is not affected by humans, the system is thoroughly tested and unacceptable results will not lead to serious concequences. However, this claim does not take into account the contribution of interpretation to the overall trust on the system.

Instead, Doshi-Velez and Kim (2018) see the necessity of interpretation to be related to incomplete problem formalization. In this context, the incompleteness cannot be quantified: It can be a lack in scientific understanding of the studied problem or impossibility of complete testing of the system, to give a couple of examples. If an incompleteness is present in the problem formalization, interpretation is needed to gather more knowledge.

Incomplete problem formalization is not the only reason, but there are several causes why explainable artificial intelligence is needed (Samek, Wiegand, and Müller 2017). At first, explanations can be used as a verification of the system. Sometimes the data set is biased and a trained AI system will give wrong conclusions, which could be easily detected if the reasoning would be visible (Caruana et al. 2015). Secondly, if one aims to improve the AI system, one must know and understand its weaknesses. The weaknesses of black boxes are not easy to detect. Thirdly, the explanations could provide new knowledge. Since the AI systems are nowadays trained by using enormous data sets that can be inaccessible to humans, AI systems can find new relationships and provide new insights. And in the end, explanations are now also a question of legislation. The new regulation of the European Union states that everyone has a right to explanation, which means that also the decisions made by AI systems must provide an explanation (Goodman and Flaxman 2016).

Explanation is important for a user to accept and be satisfied to the model's output. This has been studied since 1980's (Teach and Shortliffe 1981; Ye and Johnson 1995), and the results of the studies in the 21st century still agree (Herlocker, Konstan, and Riedl 2001; Symeonidis, Nanopoulos, and Manolopoulos 2009). Explanations help a user to critize the model and to consider if a prediction is reasonable or accurate (Kim, Khanna, and Koyejo 2016).

The urgent need for explainability and interpretation has been noticed by research groups but also by funding agencies. For example, the Defense Advanced Research Projects Agency (DARPA) established the Explainable AI (XAI) program with the aim to develop machine learning techniques that, on one hand, produce more explainable models without deteriorating the prediction accuracy, and on the other hand, allow people to understand, trust and manage these AI tools (Gunning 2018). The key goals of the XAI program are wrapped up in Figure 5. The upper panel shows how the present AI applications are implemented and how difficult the interaction between the user and the program is nowadays. The lower panel describes how the future explainable AI system could be implemented and how explanations help the user to understand the AI system. The user understand why a certain output was given, he can know in which situations the program is working and where the possible errors originate (Gunning 2018).

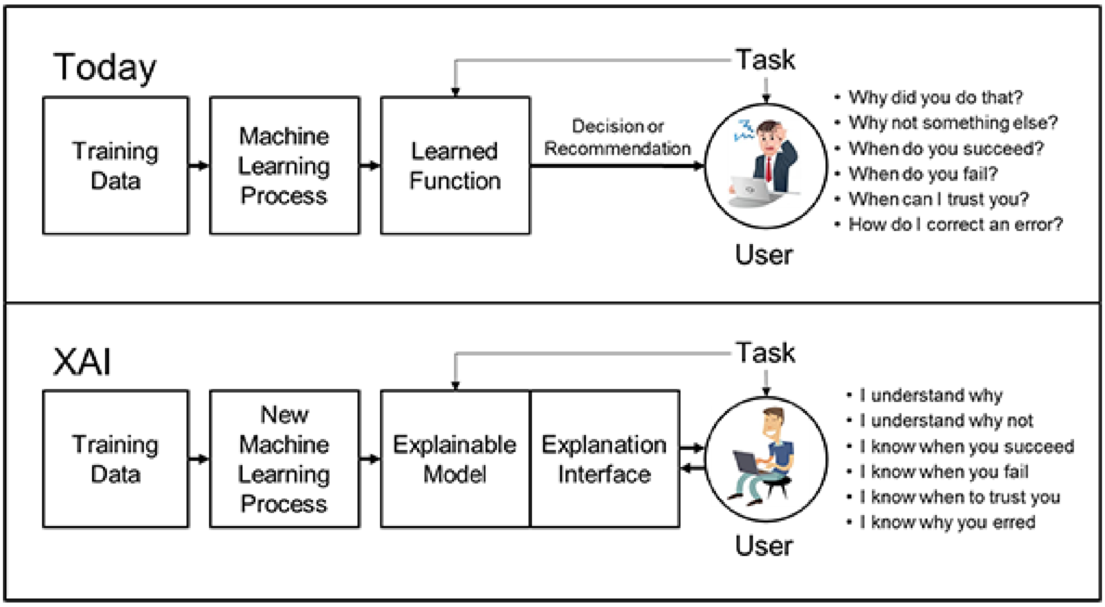


Figure 5: The key goals of the XAI program established by Gunning (2018). The present AI application are black boxes to users and it is difficult to understand when the applications work and when not. In contrast, the future explainable AI applications should be understandable to users. Figure from Gunning (2018).

2.5 Key terminology of explainable artificial intelligence

One of the main building blocks of an artificially intelligent system is the ability to explain why it made certain actions, predictions, recommendations and decisions. There are three key terms in the literature to describe this ability: explainability, interpretability and trust. Unfortunately, the definitions are not fixed in the literature and it clearly complicates the transfer of information, since researchers are describing the same consepts with different names (Došilović, Brcic, and Hlupic 2018).

Interpretability and **explainability** are time to time used as synonyms in literature, but sometimes distinction is made. Doshi-Velez and Kim (2018) define interpretability of ML systems as *the ability to explain or to present in understandable terms to a human*. In contrast, Montavon, Samek, and Muller (2018) define *interpretation as the mapping of abstract concept into a domain humans can make sense of, while explanation is the collection of features of interpretable domain that have contributed for a given example to produce a decision*. Comprehensibility is used as a synonym for interpretability, and transparency is used as a synonym for model interpretability. The latter refers to understanding the working logic of the model (Došilović, Brcic, and Hlupic 2018).

Israelsen (2017) defines **trust** as *a psychological state in which an agent willingly and securely becomes vulnerable, or depends on, a trustee (e.g., another person, institution, or an artificially intelligent agent), having taken into consideration the characteristics (e.g., benevolence, integrity, competence) of the trustee.* On the other hand, the definition of trust can be based on a prediction or a model (Ribeiro, Singh, and Guestrin 2016; Samek, Wiegand, and Müller 2017): trust can be defined as a trust for a prediction or as a trust for a model. A user can trust a individual prediction in order to make decisions, or a user can trust the whole model in order to make decisions. These are related concepts, but they differ as well. A model is trustworthy if a user of the model can trust only a certain prediction enough to use the prediction.

2.6 Explainable artificial intelligence and the connection to cognitive sciences

Much of the research on XAI is paying attention to explaining actions to a human observer. This topic is highly connected to studies in psychology and cognitive science, for instance, in which researchers have studied how humans generate and present explanations and how they employ cognitive biases and social expectations to the explanation process (Miller 2019; Hilton 1990).

Most of the work in the field of XAI seems to be based on the researchers' own intuition and opinion about a good explanation. Research frameworks of social sciences are not applied nor cited, despite the essential need of understanding of how people use and understand explanations. The experts who created the AI model or know the AI model deeply are not the most suitable persons to evaluate what kind of explanations lay users need (Miller 2019).

2.7 Explanations and their characteristics

The required key features of XAI systems and explanation have been studied in the XAI community, as well. According to Miller (2019), trusted anonymous systems shoud

1. generate decisions while having one criterion on how well humans can understand the decision (interpretability/explainability)
2. explain decisions to humans (explanation).

The first feature highlights the importance of explanations. Interpretability and explainability are seen so important that they should be taken into account when anonymous system is finding the solution. The more explainable the solution is, the better. For example, if two different solutions are relatively equal to each other, the solution that can be explained more easily is given as an output. Secondly, an anonymous systems should explain their decision to humans by providing explanations.

What comes to the explanations themselves, Miller (2019) wraps up four major concepts which are common in explanations given by and received by human beings:

1. Explanations are contrastive. People usually ask why a certain event happened instead of another event.
2. Explanations are selected in a biased manner. People rarely expect to get a full list of causes of an event. They tend to select one or two causes to be the explanation.
3. Probabilities are not that important as a part of an explanation. Referring to the probabilities in an explanation is usually less effective than referring to causes.
4. Explanations are social. They are relative to the explainer's beliefs about the explainee's beliefs.

In the first two remarks Miller (2019) wraps up what kind of explanations people are usually looking for. On one hand, they are looking for the reasons why output A was given instead of output B. On the other hand, people are not expecting to get or interested in the full explanation. A couple of main reasons are comprehensive enough.

What comes to the probabilities and causes, in many AI applications, probabilities tend to play a significant role. It is easy to understand that AI experts see probabilities as a good way to explain the output, but according to Miller (2019), providing causes is more effective than providing probabilities. Naturally, it is also user-dependent what kind of explanations are the most effective. People tend to give explanations that are relative to the explainee, as Miller (2019) states.

Ribeiro, Singh, and Guestrin (2016) have discussed the key features of explanations as well. According to them, there are certain desired characteristics for the explanations. At first, the explanations must be interpretable, and the interpretability naturally depends on the users and the problem itself. Explanations should be easy to understand and handle by the users. Secondly, the explanations should be locally faithful. It is usually impossible to ask for complete faithfulness of explanation without going into details of the model, but the behavior of the model near the point of interest is needed to be reasonably explained. In addition, a good explainer should be able to explain any model and required to be model-agnostic and some information about the global fidelity is provided.

Reliability of any model must be evaluated at some level to be useful in any real-life application. For example, classification models are often evaluated based on predictions on some test data. Despite the usefulness of these test in many cases, it can also lead to false estimations of the accuracy of the model, since the real-world data can differ a lot from the data set used in the evaluation.

2.8 Different types of approaches to explain and interpret

The approaches to explain and interpret can be divided in two main categories: integrated (transparency-based) and post-hoc approaches (Došilović, Brcic, and Hlupic 2018). The former, integrated interpretability, covers basically approaches which are aiming for transparency. The level of desired transparency is not cast iron, since even our own human mind is not transparent to us and our explanations can differ from the actual flow of thoughts which led to the decision.

On the simplest level the integrated explanation is the model itself. However, since the explanation must be understood by people, the model can be its own explanation only in the case of the very simplest forms of models, such as linear models and decision trees. Since the simplicity and inflexibility of the model go hand in hand, the approach is limited in somewhat limited models: more complex systems such as artificial neural networks are treated with post-hoc methods.

The post-hoc approach treats the model as a black-box. All the information needed for the interpretability is extracted from the complete, already learned model (Došilović, Brcic, and Hlupic 2018). Since the post-hoc methods treat the model as a black-box, these methods do not have impact on the model and its performance. The post-hoc approaches deal with interpretability and/or explainability. One approach, called as *transparent proxy model approach*, aims to find an approximation model of the more complicated black-box model. Some applications already exist, and the approach has been applied on the ensemble of decision trees to create a single decision tree by Assche and Blockeel (2007). In addition, the method was successfully applied on support vector machines by Martens et al. (2007) and on neural network ensembles by Zhou, Jiang, and Chen (2003).

Indicative techniques, such as visualization techniques, also provide post-hoc explanation but they do not pay so much attention on interpretability. Instead, they highlight some properties of the model. Different kind of visualization techniques have been already applied. Zeiler and Fergus (2014) visualized layers of convolutional neural networks with a visualization technique using deconvolutional networks, and visualization techniques were used to explain recurrent neutral networks by Karpathy, Johnson, and Fei-Fei (2015). Visualization techniques gave valuable insight in the aforementioned cases: in the former, the architecture of the model was improved and in the latter cells which take care of long-range dependencies in text were pointed out. Model-agnostic visualization method based on a sensitivity analysis was proposed by Cortez and Embrechts (2013) and it could be applied e.g. for neural networks and support vector machines. There are visualization methods that are model-agnostic (Cortez and Embrechts 2013; Adler et al. 2016; Tamagnini et al. 2017) and model-specific (Maaten and Hinton 2008; Zeiler and Fergus 2014; Li et al. 2015; Karpathy, Johnson, and Fei-Fei 2015).

Sometimes the explanation approach cannot be directly pointed to any of these two categories, but the used method is more a combination of two categories. Then one may refer to the ”third category”, *hybrid approaches*.

2.9 Implemented XAI applications

The concept of explanations was first introduced in rule-based expert systems in the 1970s (Biran and Cotton 2017; Shortliffe and Buchanan 1975). Rule-based expert systems are considered as the simplest form of artificial intelligence: the humans' knowledge about a specific area is formulated as rules, for example as if-then rules, and following those rules the system ends up in a conclusion. In the following decades, explanations have been studied in other contexts, such as

- Bayesian networks and other probabilistic decision-making systems (Lacave and Diez 2002; Cawsey 1994; Yap, Tan, and Pang 2008)
- Recommendation systems (Herlocker, Konstan, and Riedl 2001; Symeonidis, Nanopoulos, and Manolopoulos 2009; Papadimitriou, Symeonidis, and Manolopoulos 2012)
- Constraint programming (Wallace and Freuder 2001)
- Context-aware systems (Lim and Dey 2010)
- Markov Decision Processes (Khan, Poupart, and Black 2009)
- Case-based reasoning systems (Nugent, Doyle, and Cunningham 2009)
- Causal discovery (Hoyer et al. 2008)

2.9.1 Recommendation systems

Recommendation systems are online services that provide personalized recommendations for products. According to the literature, most of the XAI studies have been made for rule-based expert systems, Bayesian networks and recommendation systems (Biran and Cotton 2017). In recommendation systems, there have been many studies on what kind of justification types people find the most compelling. Herlocker, Konstan, and Riedl (2001) showed that people find rating histograms the most justifying to explain given predictions. Other explanation components that were found to be functional were based on a user's previous performance and similarity of products (Herlocker, Konstan, and Riedl 2001; Symeonidis, Nanopoulos, and Manolopoulos 2009). Papadimitriou, Symeonidis, and Manolopoulos (2012) found out that explanations that combine different types of explanations are the most functional: it is better to justify a recommendation on the user's choices, on similar users' choices and on features, not only lean on one explanation type.

2.9.2 Constraint programming

Constraint solvers are used to solve combinatorial search problems that are represented in terms of devision variables and constraints (Rossi, Beek, and Walsh 2006). There have been studies on explanation generation in systems that are not considered as pure machine learning systems. Wallace and Freuder (2001) discussed how explanations could be given in constraint programming while paying attention to how explanations are organized and presented to the user. It seems that most of perfomed studies were dealing with the explanation of conflicts, that is to say, explaining how the selections made by the user or the set-up of the original problem resulted in a condition for which a complete solution cannot be given. Junker (2001) and Jussien and Barichard (2000) studied how to present the constraints to the user when they are entangled in the conflict, whereas Amilhastre, Fargier, and Marquis (2002) suggested a set of algorithms to restore conflict situations to non-conflict states.

2.9.3 Context-aware systems

Context-aware systems are defined as systems that are able to understand the context of a given situation. In some sense they sense their physical environment and behave accordingly. In the field of context-aware systems, Lim and Dey (2010) presented a toolkit to provide eight different explanation types for the most used decision model types, namely for rule-based models, decision tree classifiers, naive Bayes classifiers and hidden Markov models. These eight types of explanations were categorized as *Inputs, Outputs, What, What If, Why, Why Not, How To* and *Certainty*. *Inputs* explain what kind of input information the application is using (e.g. GPS coordinates or restaurant reviews). *Outputs* inform what kind of outputs and capability the user can expect from the application (e.g. number of result options). *What* explanations signal the users of the previous or current output value, and *What If* explanation type helps the user to understand what would be the result of the application if a certain set of user-set input values were given. *Why* explanations signal why the given inputs resulted in such an output. *Why Not* tells the user whe the result was not a certain alternative. *How To* informs the user how a certain output can be produced. Finally, *Certainty* explanations communicate how certain or uncertain the produced output value is.

Tullio et al. (2007) made a XAI related study in the field of Context-aware systems, as well. They intestigated how users perceive intelligent application and how understanding evolves over time. During the six-week study Tullio et al. (2007) studied how office workers understand the system that predicted their managers' interruptibility and these mental models were compared to the model of the actual predictive system. Higher-level beliefs stayed robust despite the new knowledge provided.

2.9.4 Markov Decision Processes (MDP)

Markov Decision Processes form a stochastic framework for decision making where outcomes are influenced both by a decision maker and randomness. Elizalde et al. (2007) developed an explainable intelligent assistant to help a power plant operator in unusual situations. When an emergency situation occurs, a power plant operator has to analyse a vast amount of information in order to understand the source of the problem and make corresponding actions. The explanaible intelligent assistant explains the commands that were suggested and generated by an MDP planning system, thus leading to the user's better understanding. Despite the work is motivated by power plant operation, the method can be applied in other domains involving people's training or assisting (Elizalde et al. 2009).

Khan, Poupart, and Black (2009) presented a domain-independent technique to explain Markov Decision Processes, as well, and they demonstrated the method in two case problems, namely in course-selection advising for undergraduate students and in handwashing assistance for demented people.

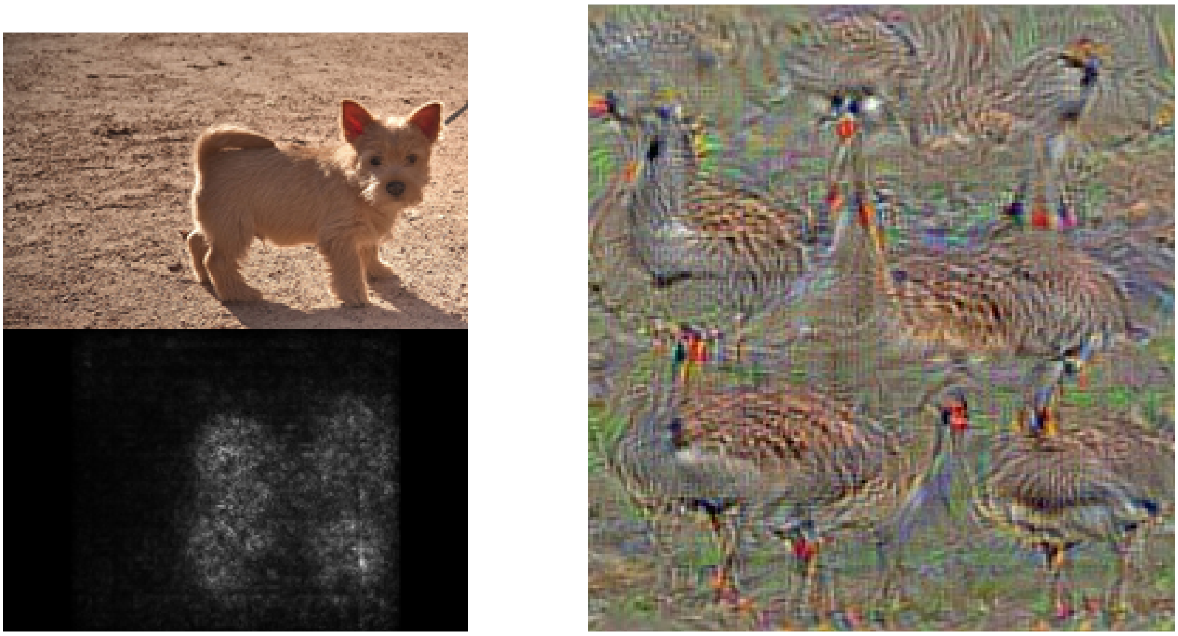
Explanations are important in decision-support systems in general. In a study related to anesthesia medical support systems it was showed that when explanations are provided, users make fewer mistakes than what they would do without explanations, users are more confident about the conclusions they made and they were more critial about the underlying model (Suermondt and Cooper 1992). Explanations were also introduced in legal cases recently (Vlek et al. 2016; Timmer et al. 2017).

2.9.5 Explanations in machine learning

There have been three main approaches to explanation that have been studied in the machine learning literature: *visualization*, *prediction interpretation and justification* and *interpretable models* (Biran and Cotton 2017). Historically, the motivation for creating explanations started from the machine learning experts themselves who wanted to estimate if the model was working correctly. The first step to evaluate the correctness was to visualize the prediction given by the model, and one of the first used tools was a nomogram (Lubsen, Pool, and Does 1978).

Despite nomograms have been used for visualization in various fields in the 21st century as well (Možina et al. 2004; Jakulin et al. 2005; Xu et al. 2015), most of the recent work has been performed on visualizing the hidden states of neural models. Tzeng and Ma (2005) published several visualization designs to explain the underlying dependencies between the input and output data. Simonyan, Vedaldi, and Zisserman (2013) presented two visualization techniques to visualize image classification models based on convolutional networks. They created a method to illustrate the salient pixels of an image that was classified (Figure 6a) and a method that illustrates how the model sees a certain class (Figure 6b). The latter method illustrates in one sense the most optimal image for the given class.

The *prediction interpretation and justification* approach aims to interpret predictions, for example by highlighting contributions of separate features. During the last decades, both model-specific and model agnostic methods have been proposed. Model-agnostic interpretation methods are framework dependent, though. Model-agnostic methods have been proposed for in the fields of classification (Baehrens et al. 2009; Robnik-Sikonja et al.



(a) A puppy.

(b) A goose

Figure 6: Figure (a.) Output of the method that highlights which image pixels were salient for the classification. Figure (b.) Output of the method that illustrates a class model. Here the class of interest is a goose. The both of the methods and Figures are from the article by Simonyan, Vedaldi, and Zisserman (2013).

2011; Kononenko et al. 2013; Martens and Provost 2014) and natural language processing (NLP) (Martens and Provost 2014; Lei, Barzilay, and Jaakkola 2016; Biran and McKeown 2017). There have been studies on model approximation: methods that approximate the complex model globally (Thrun 1994) and locally (Ribeiro, Singh, and Guestrin 2016) have been proposed. The main idea of model approximation is to create a simple or at least simpler model that approximates the solution of the original, more complex, model locally or globally.

Global approximations are often coarse. It is relatively easy to see that a local approximation, that is an approximation near by a point of interest, can reach better accuracy. One such local method is the LIME method (Local Interpretable Model-agnostic Explanations) by Ribeiro, Singh, and Guestrin (2016) which explains predictions given by a classifier by fitting a simpler and interpretable model around the neighborhood area of prediction. The explainable model is on the original data space, e.g. in the case of image recognition, the explanation model space is vectors corresponding to the pixels of the original image of interest. Recently LIME was extended by Peltola (2018) into ”KL-LIME” that is a novel approach combining LIME and predictive variable selection methods.

Another way to pay attention to explanation is the concept of *interpretable models*. Instead of explaining black box models, interpretable models aim to be interpretable themselves: examples of interpretable models are rule-based models such as decision trees. Such interpretable models have been created e.g. in the field of classification (Rudin, Letham, and Madigan 2013). In addition, there have been studies on Bayesian approaches that combine rule lists and probability distributions (Letham et al. 2015; Wang and Rudin 2014; Wang et al. 2015).

2.9.6 Popular techniques for explaining deep learning models

Samek, Wiegand, and Müller (2017) introduce two popular techniques for explaining predictions of deep learning models, namely sensitivity analysis and LRP (layer-wise relevance propagation). As an example, these techniques were applied on image and text document classification, and human action recognition.

A lot of progress have been made especially in image classification, since it is easy to visualize the explanations. Several approaches have been introduced to highlight the most meaningful pixels with respect to the output of the AI system. Thus the aim is to point out the pixels which change the output of the system significantly when they are changed significantly.

One popular method is sensitivity analysis (SA). In this method gradients with respect to input (parameters) are calculated. The most relevant input features are considered to be the ones which affect the most on the output. Samek, Wiegand, and Müller (2017) performed sensitivity analysis with respect to different pixels in a photo, and an output one gets information which pixels affect on the decision the most.

Another popular method is Layer-Wise Relevance Propagation (LRP), which explains the decision by decomposing the prediction into relevance scores by applying certain redistribution rules (Samek, Wiegand, and Müller 2017). LRP differs from most of the other methods since it is not based on gradient evaluation. In addition GradCAM is a popular tool to generate saliency maps representing the relevance of pixels in the studied image (Harradon, Druce, and Ruttenberg 2018).

Causal semantics have been used to explain predictions of deep neural networks, which makes sense since explanations must be causal models in essence. Harradon, Druce, and Ruttenberg (2018) used an auxilliary neural network model to construct consept representations in order to explain the predictions of deep neural networks.

Most of the recent explainable models are unimodal, offering only a visual or textual explanation. First attemps to provide multimodal explanations have also been introduced. Park et al. (2018) were the first ones to provide explanations in the forms of text and image in the contexts of visual question answering and activity recognition. Due to the lack of suitable datasets which human justifications, they collected two datasets to train and test the created model, Pointing and Justification Explanation model.

2.10 Discussion

Despite the interest in explainable artificial intelligence has grown rapidly very recently, there has been long and continuous work on the topic during the last few decades. Much of the gained knowledge can be used to make present AI applications more explainable, but the present models and systems are more complex than ever before in the history.

The aim of the literature review was to answer to the following research questions:

1. What kind of explanation techniques have already been implemented in the field of XAI?
2. Are the present techniques designed for a specific AI method or can they be applied generally?

In the previous chapter, we have seen that numerous techniques have already been implemented. Unfortunately, it is challenging to make a comprehensive list of those techniques and applications. The reasons are manifold. First of all, the scientific community has not fixed the key terminology: There is no fixed definition even for *an explanation*. The unfixed terminology leads to a situation in which the researchers call similar or even identical concepts by different names, and as a result, it is difficult to find similar studies performed by other scientists.

There are a lot of approaches that have been implemented in different AI communities. Naturally, explanations have been designed for the method used in each community, thus being ”community-specific”, but the explanation techniques can be even model-specific. Thus the number of published techniques is relatively great. Both of the method types have been implemented: There are techniques that are designed for a specific AI method and there are methods that can be applied generally in the corresponding subfield, e.g. in Markov Decision Processes or in context-aware systems.

In this chapter we have gathered examples of XAI implementations from various communities. However, the reader should bear in mind that the examples do not cover all the research that has been made since the 70s, but they give a brief overlook instead. Despite the vast research, there are still work to do. In addition to the unfixed terminology, there seems to be a shortage of empirical studies that would measure interpretability from the point of view of a user. As it was seen, most of the work on explanations have been made by research groups of AI specialists, not in collaboration with cognitive scientists.

3 Methods and algorithms

In this chapter the methods that are used to study the parameter space of an ECR ion source are introduced. The building blocks of a machine learning model are the hypothesis about the suitable model, the used penalty function and the chosen learning algorithm. This means that there is always assumptions about the phenomenon that the model tries to mimic. Next, one must choose an indicator for measuring the model error against the data, and the model can be improved via learning algorithms, which are often optimization algorithms (Hastie, Tibshirani, and Friedman 2009).

3.1 Linear regression

Linear regression is one of the simplest methods to model the relationship between a scalar response and one or more explanatory variables. The model is grounded on linear functions, and the unknown parameters of the model are estimated from the data. If there is only one explanatory variable, the method is called *simple linear regression*, whereas a problem involving several explanatory variables is referred as *multiple linear regression* (Hastie, Tibshirani, and Friedman 2009).

A linear regression model assumes there is a linear relationship between an input vector x and an output vector y. if we assume the input vector x to be p-dimensional, the linear regression model can be written as

$$f(x) = \beta_0 + \sum_{i=1}^{p} x_i \beta_i, \tag{3.1}$$

where the coefficients β_i are unknown. The elements of input vectors x_i may be quantitative observables, transformations of observables (e.g. square roots of observed values), expansions (e.g. $x_2 = x_1^2$, $x_3 = x_1^3$) or other kind of combinations of qualitative inputs (Hastie, Tibshirani, and Friedman 2009).

If the relationship between the observable (dependent variable) and the inputs (independent variables) is modelled as a polynomial, e.g.

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n, \tag{3.2}$$

the method is called *polynomial regression*. Even though the fitted polynomial model is not linear, the regression function is linear in the unknown parameters. Thus polynomial regression is categorized as one type of multiple linear regression. However, the coefficients of polynomial model may be more troublesome to interpret, since the different powers of variable x are highly correlated.

Since the parameters β_i are unknown, they must be estimated. The most typical way is to use some training data and estimate the parameters via *least squares method.* In least squares method, the parameters β_i are selected to be the parameters that minimize the residual sum of squares (RSS)

$$
\begin{aligned}
RRS(\beta) &= \sum_{j=1}^{N} (y_j - f(x_j))^2 && (3.3)\\
&= \sum_{j=1}^{N} \left(y_j - \beta_0 - \sum_{i=1}^{p} x_{ji}\beta_i \right)^2 && (3.4)
\end{aligned}
$$

where N represents the size of the training data.

3.2 Ridge and lasso regression

The **lasso** (*least absolute shrinkage and selection operator*) and **ridge regression** are shrinkage methods: they shrink the model coefficients by weighting them with a penalty on their size. These two methods have different penalty functions which leads to the different properties of outcoming regression models.

If we denote the one-dimensional output as y_i and the p-dimensional input vector as x_i, ridge coefficients are determined from the minimization problem (Hastie, Tibshirani, and Friedman 2009)

$$
\min_{\beta} \left\{ \frac{1}{N} \sum_{j=1}^{N} (y_j - \beta_0 - \sum_{i=1}^{p} x_{ji}\beta_i)^2 \right\} \quad \text{subject to} \quad \sum_{i=1}^{p} \beta_i^2 \leq t, \tag{3.5}
$$

whereas the lasso regression model is defined via l1-regularized objective function

$$
\min_{\beta} \left\{ \frac{1}{N} \sum_{j=1}^{N} (y_j - \beta_0 - \sum_{i=1}^{p} x_{ji}\beta_i)^2 \right\} \quad \text{subject to} \quad \sum_{i=1}^{p} |\beta_i| \leq t. \tag{3.6}
$$

The difference between two aforementioned methods is the constraint – one uses the sum of absolute values of β_i or the sum of squared β_i values. When the constant t is small enough, the lasso method will lead to a solution with some regression constants β_j being zero. That is how lasso method performs feature selection, as illustrated in Figure 7.

Figure 7 demonstrates the effects of ridge and lasso penalty functions in two dimensions of β. The constraint regions are marked in blue: the region is a circle for ridge regression, whereas the constraint region of lasso regression is a diamond. The contour lines of a least squares error function are marked in red. As Figure 7 demonstrates, when t is sufficiently small, the result of lasso regression problem is likely one corner of the diamond region, meaning that one of the coefficients β_i is zero, and the corresponding input is not included in the model. In ridge regression getting one zero-valued β_i is not more likely than getting any other solution.

Lasso and ridge regression objectives may be written in Lagrangian form

$$\min_{\beta}\left\{\frac{1}{N}\sum_{j=1}^{N}(y_j-\beta_0-\sum_{i=1}^{p}x_{ji}\beta_i)^2+\alpha\sum_{i=1}^{p}\beta_i^2\right\} \qquad \text{(Ridge)} \tag{3.7}$$

$$\min_{\beta}\left\{\frac{1}{N}\sum_{j=1}^{N}(y_j-\beta_0-\sum_{i=1}^{p}x_{ji}\beta_i)^2+\alpha\sum_{i=1}^{p}|\beta_i|\right\} \qquad \text{(Lasso)}. \tag{3.8}$$

The Lagrangian form is useful when one needs to reformulate a constrained problem into a form for which the derivative test of an unconstrained problem can be performed. The derivative test is used to find the critical points (e.g. a local minimum) of a function.

3.3 Huber regression

Two most used penalty functions are the absolute and squared loss functions, L_{abs} and L_{sq}, respectively. If the loss is calculated on residuals, they are defined as

$$L_{abs}(y_i, f_i(x)) = |y_i - f_i(x)| \tag{3.9}$$

$$L_{sq}(y_i, f_i(x)) = (y_i - f_i(x))^2 \tag{3.10}$$

Ridge and lasso regression are sensitive to outliers, and a few remarkable measurement errors may change the outcoming model. As a consequence, the model may significantly lose

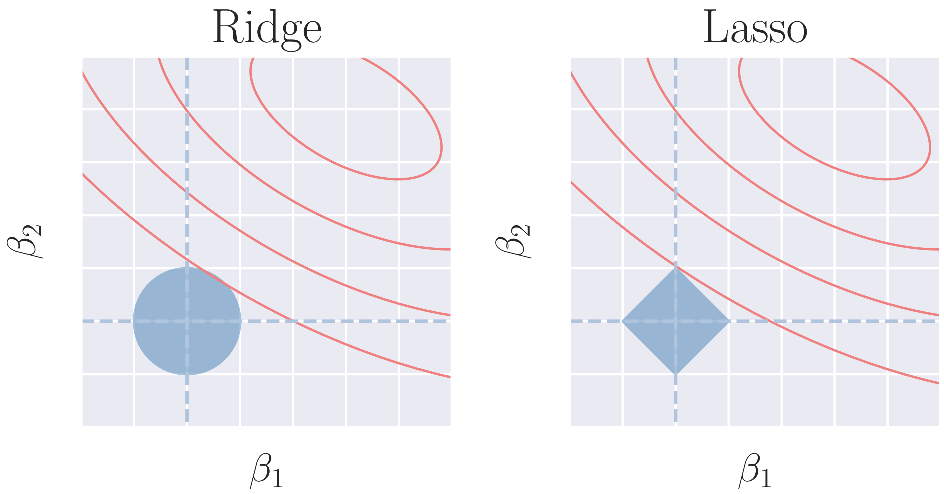


Figure 7: The penalty functions of lasso and ridge regression in two dimensions of β. The constraints of ridge and lasso regression are marked in blue. Ridge constraint region forms a circle, whereas the lasso constraint region is a diamond. The contour lines of a least squares error function are illustrated in red. Figure adapted from Hastie, Tibshirani, and Friedman (2009).

prediction power. Instead of squared or absolute error loss, one may define the loss function in two pieces. The loss function of **Huber regression** is defined as (Hastie, Tibshirani, and Friedman 2009)

$$L_\delta(y_i, f_i(x)) = \begin{cases} \frac{1}{2}(y_i - f_i(x))^2 & \text{for} \quad |y_i - f_i(x)| \leq \delta \\ \delta|y_i - f_i(x)| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases} \tag{3.11}$$

Huber regression combines the loss functions of Lasso and Ridge regression. The hypothesis is the same: It is assumed that the studied phenomenon can be modelled by a linear function. However, the assumptions about the error distribution are different. The squared error loss puts much more emphasis on observations that have large difference to the model output, thus being far less robust method. The absolute loss can handle the outliers much better. In Huber regression one combines the good properties of squared-error loss (non-outliers) and absolute error loss (outliers) (Hastie, Tibshirani, and Friedman 2009).

3.4 Conventional model validation

In **k-fold cross-validation** the data set is randomly divided into k equal sized subsets, from which $k-1$ subsets are used as a training data and the remaining set is reserved for the model validation. This process is repeated k times so that every subset is used as the validation data. In the end, the final estimates and statistics are given based on all the k results, for example by averaging. Commonly used k values are 3, 5 and 10, but the value is not fixed in general (Hastie, Tibshirani, and Friedman 2009).

Leave-*p*-out (LPO) cross-validation method uses p data points as the validation data, and the rest $p-1$ data points are used in training. The difference to the k-fold method is that all the different combinations of p training points are taken into account. The shortcoming of the method is the computational cost: the data set and the coefficient p do not have to be particularly large in order to become computationally infeasible. The LPO cross-validation method with $p=1$ is called **leave-one-out (LOO)** cross-validation (Hastie, Tibshirani, and Friedman 2009).

One statistic that is used to give information about the goodness of fit of a model is the **coefficient of determination**, denoted as R^2 (Hughes and Grawoig 1971). R^2 values normally lie within the interval $[0,1]$, and greater R^2 corresponds to a better fit – the model is able to explain the variation of the output values with different input values.

If we define the mean as $\bar{y}$, that is

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{3.12}$$

in addition to the total sum of squares and the residual sum of squares which are defined as

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{3.13}$$

$$SS_{res} = \sum_i (y_i - f_i)^2, \tag{3.14}$$

the coefficient of determination is defined as

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}}, \tag{3.15}$$

where n is the number of data points, y_i is an observed value and f_i is a prediction given by the model. As we can see, a model with $R^2 = 1$ corresponds to a perfect fit: then $SS_{res} = 0$.

Another estimator used in statistics is the **mean squared error (MSE)** (Lehmann 1983). MSE measures the average squared difference between the predicted values and the actual measured values, thus it measurest the average squared errors. Mathematically MSE is expressed as

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2 . \tag{3.16}$$

Despite its simplicity and usefulness, MSE has a deficiency: it gives a heavy weight for outliers. Since each difference between a prediction and the corresponding observed value is squared, large errors are effectively weighted more than smaller errors.

4 Demonstration: Machine Learning for an ECRIS

In this section we apply simple machine learning algorithms and the methodology of explainable artificial intelligence on a case study related to an electron cyclotron resonance ion source (ECR ion source, ECRIS). ECR ion sources are typical plasma ion sources which use electron cyclotron resonance for generating plasma and obtaining multiply charged ions. In addition to the usage in the research of nuclear and plasma physics, ECR ion sources are used in semiconductor fabricating and cancer treatments (proton therapy), for instance.

The main goal of this study is to find out if machine learning models can be used to study the parameter space of the ion source. All the relationships between the model parameters and the optimal ion beam production are not completely understood, thus it is not known which parameters produce the maximum ion beam intensity for each nuclear charge state. Here we investigate the possibility if machine learning methods could be used to find the optimal parameters.

4.1 Basic principles of an ECR ion source

ECR ion sources are used to create ion beams. An ECR ion source and its operation is illustrated in Figure 8 in a simplified way. Neutral gas atoms of a selected chemical element are transmitted to the plasma chamber (1) that is located in a special magnetic field (2). The magnetic field creates a magnetic bottle that is illustrated as a grey ellipsoid. In the magnetic bottle, moving electrons tend to stay near by the center of the bottle. Sometimes an electron reaches the edge of the magnetic bottle (3), and then it has the same frequency as the microwaves transmitted by the microwave guide have. In this situation the microwaves give energy to the electron. When the energetic electron passes by a gas atom, the electron ionizes the gas atom and the corresponding ion is formed (4). These ions are then extracted and focused for later purposes (5). For the detailed description of ECR ion sources can be found e.g. in the book written by Geller 1996.

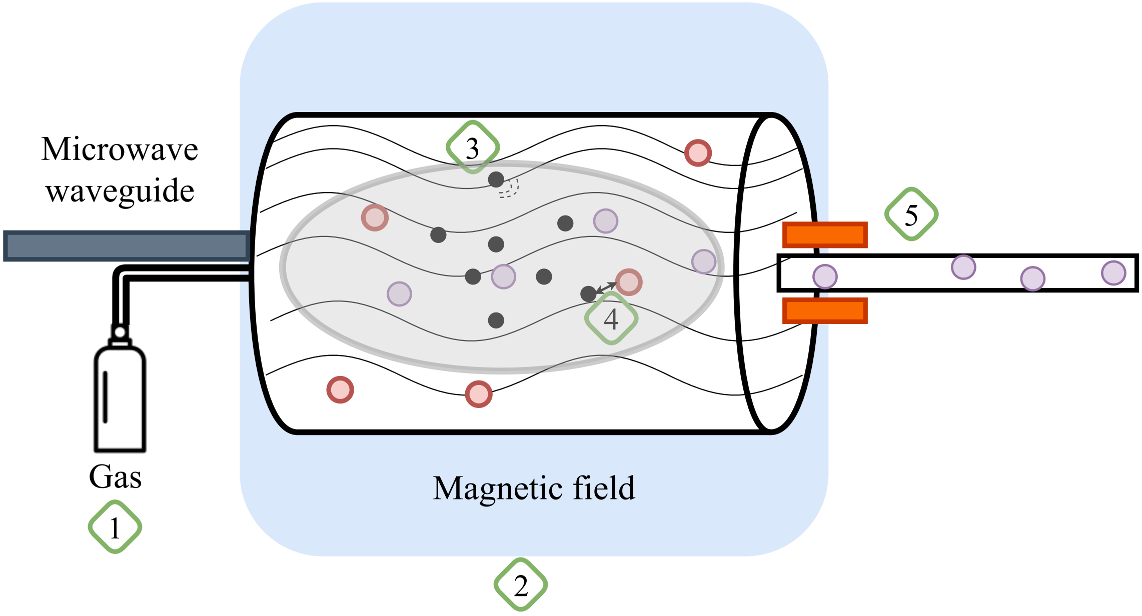


Figure 8: A simple schematic illustration of an ECR ion source.

4.2 Parameters and constraints

The production of ions depends on several input variables and a lot of different kind of physics is involved when ions and an ion beam are created. All the relationships between these variables are not fully understood, thus the optimal parameters for creating the maximum ion beam intensity are not known. In this work we build a machine learning model for an ECR ion source HIISI (Kalvas, Koivisto, and Tarvainen 2018; Koivisto et al. 2014; Kalvas et al. 2014) and optimize the model in order to determine the optimal parameters for a specific ion beam. Five of the input variables were studied as free parameters, and the other inputs and settings were fixed. The free input parameters x_i are listed in Table 1 together with their units and ranges of allowed values.

The bias voltage is used to keep electrons in the plasma chamber. The microwave powers with the frequencies 18 GHz and 14.5 GHz are related to the microwave guide seen in Figure 8. Since there are two microwave emitters, there are actually two microwave frequencies inside HIISI which are transmitting energy to the electrons. By changing the ion gas valve position one can regulate the amount of neutral gas that is transmitted into the plasma chamber. Buffer gas is used to create the plasma in the plasma chamber and one can regulate the

Input parameter	Description	Unit	Range
x_1	Bias voltage	-V	$[0,45]$
x_2	Microwave 18 GHz power	W	$[100,2000]$
x_3	Microwave 14.5 GHz power	W	$[100,1000]$
x_4	Ion gas valve position		$[0,260]$
x_5	Buffer gas valve position		See Eqs.(4.1)–(4.5)

Table 1: Input parameters, their descriptions, the corresponding units and the allowed ranges of the parameters.

amount of buffer gas via the buffer gas valve position.

The possible values of the ion gas valve position (the parameter x_5) cannot be expressed as a simple closed interval, but the variable is constrained by a set of equations depending on the buffer gas valve position (x_4). These constraints were determined experimentally for this study. The allowed values of x_5 as a function of the parameter x_4 are mathematically expressed as

$$250 \leq x_5 < y_1(x_4) \quad \text{when} \quad 0 \leq x_4 < 120, \tag{4.1}$$

$$0 \leq x_5 < \min\{y_1, y_2, y_3\} \quad \text{when} \quad 120 \leq x_4 \leq 260, \tag{4.2}$$

where

$$y_1(x_4) = 390 - \frac{4}{17}x_4, \tag{4.3}$$

$$y_2(x_4) = 690 - 2x_4, \quad \text{and} \tag{4.4}$$

$$y_3(x_4) = \frac{25}{4}(260 - x_4). \tag{4.5}$$

Equations (4.1)-(4.5) confine a closed connected domain in (x_4, x_5) plane. The allowed values of the ion gas valve position x_5 are illustrated in blue in Figure 9 as a function of the buffer gas valve position x_4.

4.3 Design of the measurement points

One week of measurement time was reserved for setting up the measurement, analysizing the ranges of allowed values for each x_i and finally gathering the data. Thus the number

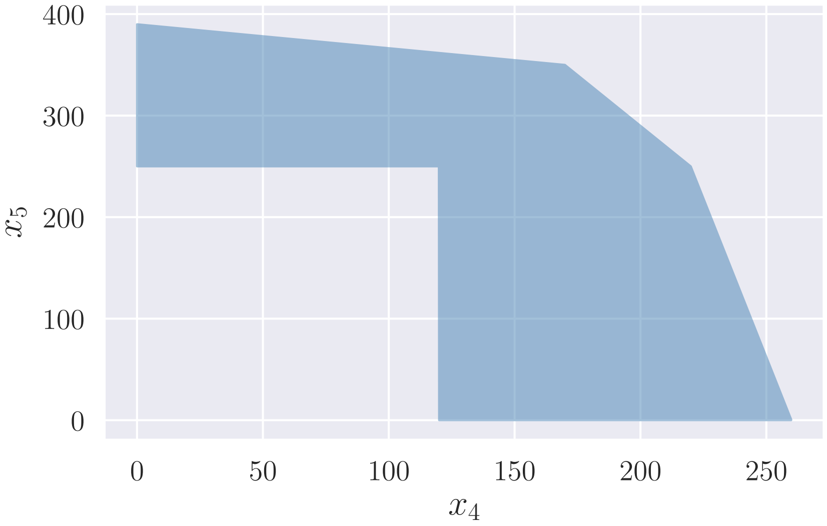


Figure 9: Allowed values of the ion gas valve position (x_5) as a function of the buffer gas valve position (x_4) marked in blue. The variables are unitless.

of data points was limited and approximated to be around 100. Naturally, the measurement points had to be chosen wisely in order to maximize the amount of information.

In random sampling new sample points are created randomly without considering the previous sample points. Despite the uniform probability, it is always possible that samples are not distributed uniformly and a part of the sample space stays unrepresented. On the other hand, a perfect grid with a equal spacing between the points does not provide the maximal information: A measurement point, being different from another only with respect to one parameter, does not provide any additional information on the other parameters and their impact on the output variable. There are some algorithms introduced in literature that create samples so that sample points do not have same coordinates (Latin Hypercube Sampling, LHS) and in addition to that, they cover the subspaces with the equal density (Orthogonal sampling, OS) (Petelet et al. 2009; Cioppa and Lucas 2007). However, there are some hindrances in the application of these two aforementioned algorithms. The LHS method can also leave some subspaces unevenly sampled, and there are no computationally efficient open-source implementation for the OS available.

Thus the design of the measurement points was chosen to be a combination of random sampling and a equal-spaced grid which was easy to implement and perfectly suitable for this study. First the measurement points form a grid with equal-length distances in the input space $\{x_i\}$. In the second step, all the measurement points are moved a bit by a random vector. These two steps are illustrated in Figure 10.

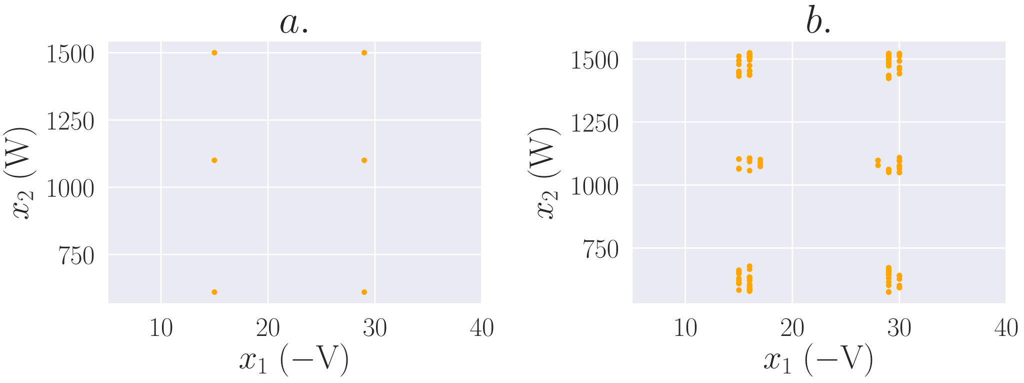


Figure 10: Illustration of the algorithm that was used to create the set of measurement points. Plot *a.* illustrates how the measurement points would locate in (x_1,x_2) plane if there would not be any randomization. Plot *b.* illustrates how the measurement points get separated after relocating them with a small random vector.

Figure 10a illustrates how the multiple measurement points would locate in (x_1,x_2) plane if any randomization would not be applied. Two different values for the variable x_1 and three different values for the variable x_2 were chosen. Since the number of measurement points was limited to 100, it was not possible to include three different values for all the five variables. Thus the ion source group was consulted and asked to priorise the variables x_i.

Without any randomization, all the different measurement points that have different x_3, x_4 and x_5 would not have large variety in x_1 and x_2 coordinates. In order to get even more information from the measurements, the measurement points illustrated in Figure 10a where randomized by moving them by a small random vector. The result of randomization is shown in Figure 10b.

4.4 Exploratory data analysis

Before implementing any machine learning methods, the first look on the data was taken. Every variable x_i was briefly studied with respect to the outputs. As it can be seen in Figure 11, some relationships were already visible from simple figures. For example, the buffer gas valve position (x_5) has a clear relationship with the maximum intensity of extracted $^{16}O^{3+}$ beam current. On the other hand, the buffer gas itself did not seem to have trivial relationships e.g. with extracted $^{40}Ar^{9+}$ beam current.

Figure 11 revealed some unnecessary outputs as well. $^{40}Ar^{18+}$ and $^{16}O^{8+}$ were not measured at all during the experiment (the beam currents were zero all the time), so these variables could already be disregarded. Thus, after eliminating two, 13 different beam currents were left to be studied. In addition to the buffer gas, some clear simple relationships with the extracted beam currents were found also in the ion gas and the microwave (18 GHz).

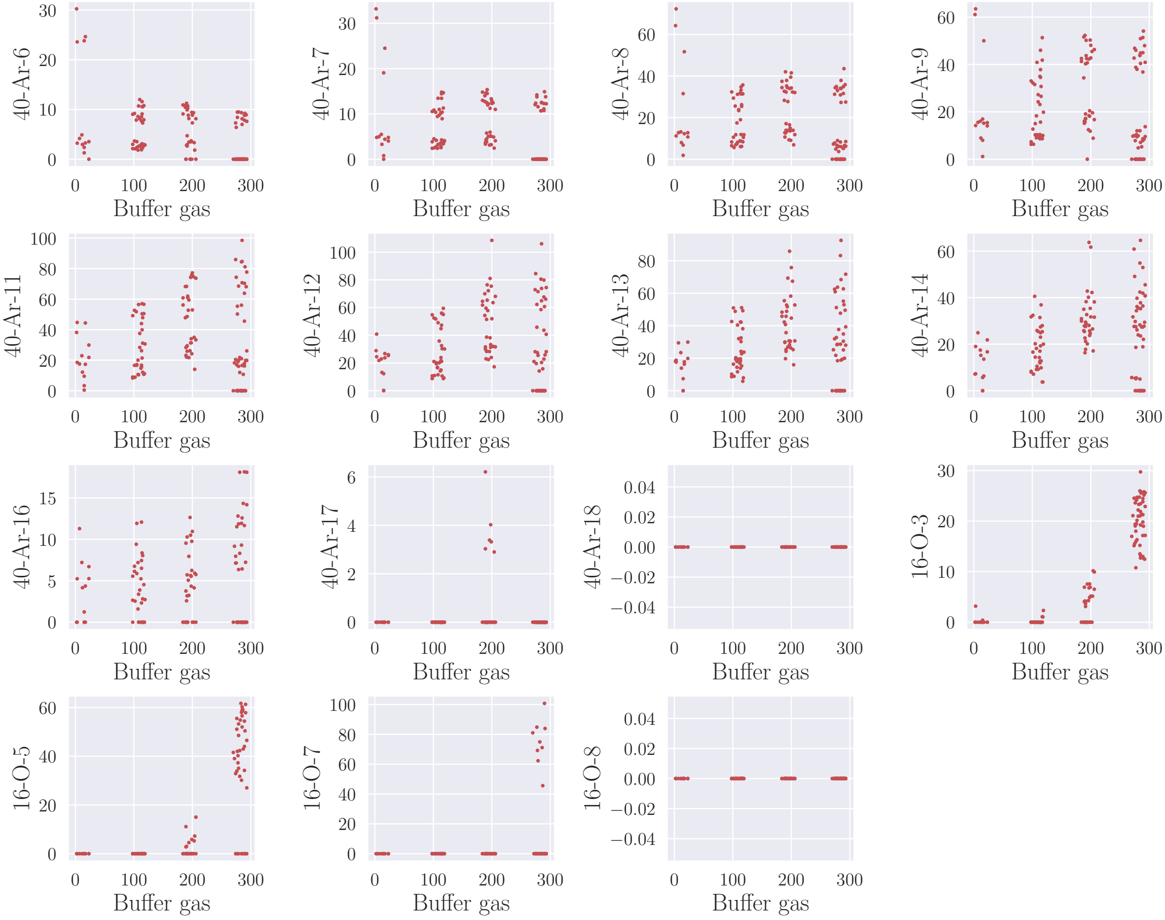


Figure 11: All the different output variables as a function of buffer gas (x_5). The outputs are ion beam intesities for different charge states of ions. Buffer gas is unitless and the ion beam intensities are in μA.

4.5 Multilinear regression

In order to avoid overfitting, it is better to explore the problem by starting from the simplest models. Despite its simplicity, multilinear regression seemed to work surprisingly well in some cases. Multilinear regression models with all five input variables were fitted and tested: bias voltage (-V), 18 GHz microwave power (W), 14.5 GHz microwave power (W), ion gas valve position and buffer gas valve position. Multilinear regression models were created for every ion beam intensity separately. We were interested in modelling and optimizing one certain ion beam intensity at a time, since only one ion beam can be extracted in a dipole at a time.

The LOO method was used for estimating the validity of the fits (Hastie, Tibshirani, and Friedman 2009). During the analysis it was found out that the amount of data points that had a zero ion beam intensity was significant. In addition, zero valued ion beam intensities had a remarkable effect on the models. The outputs having a zero value were problematic, since there were two kind of sources for zeros. One source was meaningful from point of the model: some combinations of input variables give such a small ion beam intensity that it cannot be observed - thus the intensity is more or less zero in reality. However, the experimental data of ion beam intensities included also other kind of zeros that stem from the fact that the intensities are determined from charge state distributions (CSD). The ion beam intensity is read from the corresponding peak in CSD. If two different peaks are overlapping, any of the two ion beam intensities cannot be determined and the corresponding values in the experimental data were marked zeros in this experiment, even though they are non-zero as a matter of fact. The effect of overlapping is illustrated in Figure 12. As a consequence, zero valued ion beam densities were problematic when creating models.

The LOO method was applied in both of the two possible ways, first by including all the experimental data points and then by ignoring data points that had zero ion beam intensity. Figure 13 illustrates the results of the LOO method when all the experimental data was included, whereas the zero valued ion beam intensities were ignored in the models and corresponding results in Figure 14.

As it can be seen in Figure 13, the multilinear regression cannot predict experimentally

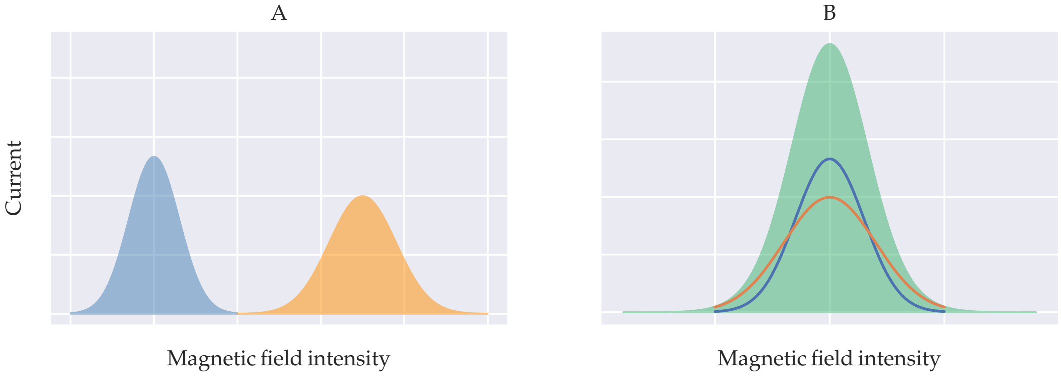


Figure 12: The experimental ion beam intensities are determined from the charge state distributions (CSD). If a peak corresponding to an ion beam is not overlapping, the corresponding ion beam intensity can be determined (A). If any peaks overlap, none of the corresponding ion beam intensities can be determined (B).

measured zero-valued ion beam intensities in general. That is understandable due to the aforementioned two-fold origin of those values. If the data set related to a certain ion beam did not include a significant amount of zero ion beam intensities, those values did not have a significant effect on the predictive power of the corresponding model (Fig. 13, $^{40}\mathrm{Ar}^{11+}$). However, if the experimental data included too many zeros, it could ruin the whole model (Fig. 13, $^{40}\mathrm{Ar}^{16+}$).

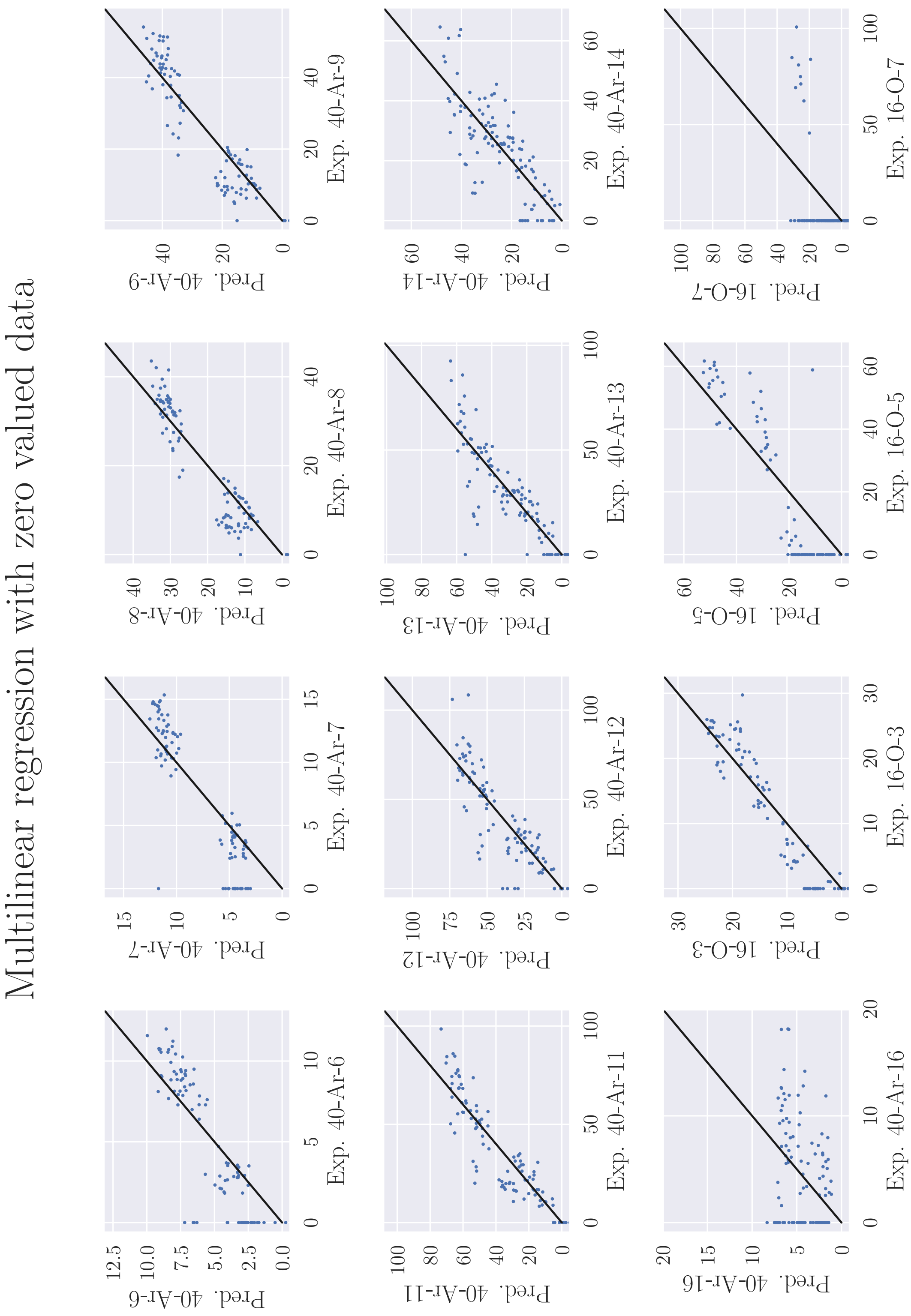


Figure 13: Multilinear regression tested with the LOO method when all the experimental data was taken into account. The predicted output variables (vertical axes) are plotted versus measured ones (horizontal axes). If the model was perfect, all the points in the subplots were on diagonal.

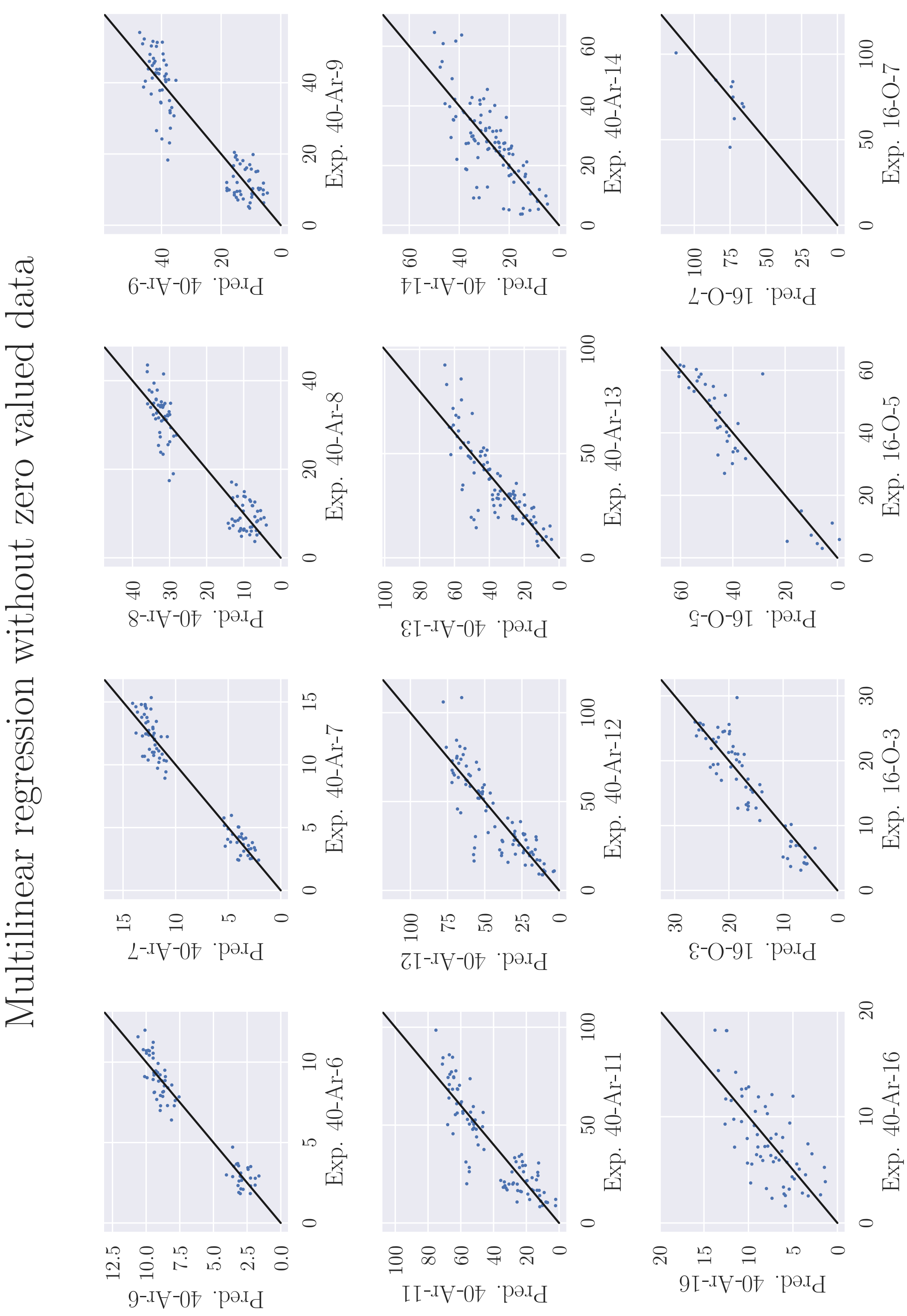


Figure 14: The same as Figure 13 but the zero valued ion beam intensities were ignored.

If the problematic zero-valued ion beam intensities were ignored, the multilinear regression models were surprisingly successful, as seen in Figure 14. However, the removal of the measurement points has its own shortcomings. The amount of experimental data could get too small (Fig. 14, $^{16}O^{7+}$), which leads to the fact that the predictive power of the corresponding model is weak.

Table 2 wraps up the R^2 values of the multilinear regression models which were determined through the LOO method. The full data set consists of 111 measurement points, and naturally those data sets that had zero-valued ion beam intesities removed are smaller. It can be seen from the table that the models succeeded better in general when the zero-valued ion beam intensities were ignored. The biggest change was identified between the models of $^{40}Ar^{16+}$ since the number of zero-valued intensities was so significant. However, some models got better R^2 score when all the available data points were included in the fitting process, e.g. $^{40}Ar^{11+}$. Most likely the complete data set included physically meaningful zero-valued data that was useful in the fitting process.

4.6 Lasso regression

Lasso regression models were fitted to the experimental data and their accuracy was estimated via the Leave One Out method (Hastie, Tibshirani, and Friedman 2009). Since the Lasso objective includes the parameter α, that is determining the amount of penalization and is not fixed in general, it must be chosen in a way or another. In this study, α was determined by a 5-fold cross-validation strategy starting from the default values α of Python scikit-learn library. All the different ion beams were modelled separately, thus leading to a set of different values of α. In order to be more confident with the selected values of α, the average MSEs of models were plotted as a function of the parameter α. One example of those plots is shown in Figure 15 that contains the average MSEs for the ion beam intensity of $^{40}Ar^{12+}$ as a function of α. In this particular case, α was chosen to be approximately 18.8. The shaded region in Figure 15 refers to the error bar.

	All the data		*No zero valued intensities*	
Ion beam	Data points	R^2	Data points	R^2
^{40}Ar^{6+}	111	0.69	70	0.93
^{40}Ar^{7+}	111	0.76	78	0.93
^{40}Ar^{8+}	111	0.86	94	0.89
^{40}Ar^{9+}	111	0.85	94	0.85
^{40}Ar^{11+}	111	0.84	95	0.79
^{40}Ar^{12+}	111	0.76	91	0.73
^{40}Ar^{13+}	111	0.73	93	0.69
^{40}Ar^{14+}	111	0.60	100	0.56
^{40}Ar^{16+}	111	0.06	62	0.47
^{16}O^{3+}	111	0.86	68	0.84
^{16}O^{5+}	111	0.71	41	0.83
^{16}O^{7+}	111	0.22	9	0.33

Table 2: R^2 values for the multilinear regression models. R^2 values are determined via the LOO method. All the models were fitted on data that included all the measurement points (on the left) and the data without zero valued ion beam intensities (on the right). The ion beams were treated separately. The number of all the available data points is reported as well.

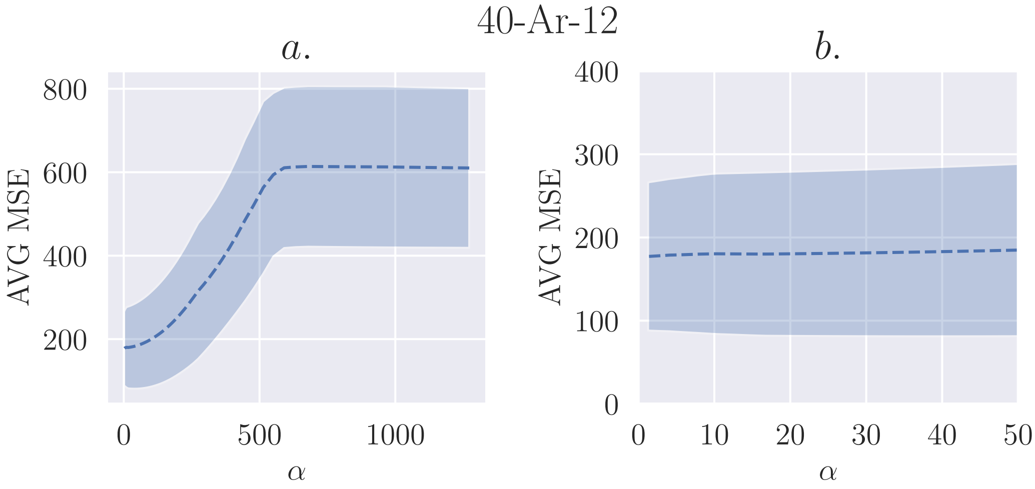


Figure 15: The average MSE of the ion beam intensity of ^{40}Ar^{12+} as a function of the parameter α. The parameter α of the Lasso regression gives the weight for the penalization. The panel a. shows the average MSE for all the used values $alpha$, and the panel b. illustrates that when the parameter α is small enough, the average MSE is relatively constant.

The results of LOO cross validation are shown in Table 3. The R^2 scores are more or less the same as the results of basic multilinear regression models seen in Table 2. However, one clear exception can be found: The R^2 score of the ion beam intensity of $^{16}\mathrm{O}^{7+}$ is negative. This can happen, even though it is not usual – the negative value of R^2 is highlighting the fact that the model is unable to predict the experimental data. This is not a surprise when considering the amount of data points that were used for model fitting. Nine data points for five dimensional input space is clearly far from sufficient.

	All the data		*No zero valued intensities*	
Ion beam	Data points	R^2	Data points	R^2
$^{40}\mathrm{Ar}^{6+}$	111	0.69	70	0.92
$^{40}\mathrm{Ar}^{7+}$	111	0.77	78	0.93
$^{40}\mathrm{Ar}^{8+}$	111	0.86	94	0.88
$^{40}\mathrm{Ar}^{9+}$	111	0.85	94	0.85
$^{40}\mathrm{Ar}^{11+}$	111	0.84	95	0.79
$^{40}\mathrm{Ar}^{12+}$	111	0.76	91	0.72
$^{40}\mathrm{Ar}^{13+}$	111	0.73	93	0.70
$^{40}\mathrm{Ar}^{14+}$	111	0.59	100	0.55
$^{40}\mathrm{Ar}^{16+}$	111	0.11	62	0.44
$^{16}\mathrm{O}^{3+}$	111	0.86	68	0.84
$^{16}\mathrm{O}^{5+}$	111	0.71	41	0.82
$^{16}\mathrm{O}^{7+}$	111	0.21	9	-0.20

Table 3: Same as Table 2 but for the Lasso regression method. Most of the time, R^2 lies in the interval $[0,1]$, but it can be negative if the model is unable to predict the experimental data.

Figure 16 shows the predictions as a function of experimental ion beam intensities (in the units of μA) of $^{40}\mathrm{Ar}^{12+}$ when Lasso regression was used. Zero valued ion beam intensities were ignored and the LOO method was used. Two side panels illustrate the distribution of measured and predicted ion beam intensities. The Lasso regression model can reproduce most of the experimental values relatively well, but the model could perform better as well. If we consider the experimental $^{40}\mathrm{Ar}^{12+}$ ion beam intensities around $20\mu A$, we can observe that they are predicted to lie within the interval $[20,60]\mu A$. This means that the model gives some predictions that are three times greater than they should be, which is a significant hindrance.

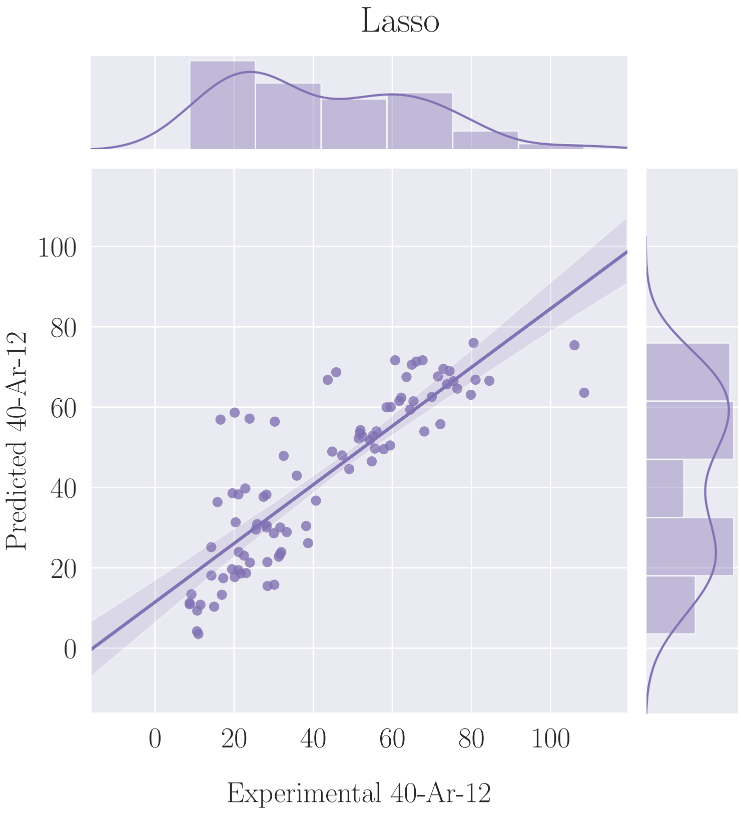


Figure 16: Experimental ion beam intensities of $^{40}Ar^{12+}$ versus predicted ones by the Lasso regression method when zero valued experimental values were ignored and the leave-one-out method was applied. The values are in the units of μA. The side panels illustrate the distribution of the measured and predicted ion beam intensities. The Lasso regression model can produce experimental values relatively well when compared to the complexity of the problem, but the model could perform better as well. For instance, experimental ion beam intensities around $20\mu A$ are predicted to be in the interval of [20,60] μA.

4.7 Huber regression

Since our data included outliers, namely those zero-valued ion beam intensities that should be non-zero, the Huber regression method was believed to be the best option for the study.

Different values of the regularization parameter were tested between the interval $[0.0001, 10.0]$, but they did not make a significant difference to the R^2 score of ${}^{40}Ar^{12+}$ ion beam intensity. Thus the default value for the regularization parameter was chosen, namely 0.0001.

The R^2 scores for different Huber regression models are wrapped up in Table 4. The scores are once again very similar to the scores of the Multilinear and Lasso regression models, which is somewhat surprising, since the Huber regression method should be better in modelling tasks involving outliers. Most likely the result just highlights the fact that there are not many clear outliers in the data sets and the robust techniques are not needed.

	All the data		*No zero valued intensities*	
Ion beam	Data points	R^2	Data points	R^2
^{40}Ar^{6+}	111	0.67	70	0.92
^{40}Ar^{7+}	111	0.75	78	0.93
^{40}Ar^{8+}	111	0.86	94	0.86
^{40}Ar^{9+}	111	0.84	94	0.84
^{40}Ar^{11+}	111	0.82	95	0.76
^{40}Ar^{12+}	111	0.76	91	0.68
^{40}Ar^{13+}	111	0.73	93	0.69
^{40}Ar^{14+}	111	0.57	100	0.56
^{40}Ar^{16+}	111	0.03	62	0.44
^{16}O^{3+}	111	0.86	68	0.83
^{16}O^{5+}	111	0.70	41	0.82
^{16}O^{7+}	111	-0.08	9	0.13

Table 4: Same as Table 2 but for the Huber regression method.

Figure 17 illustrates the predictive power of the Huber model of ^{40}Ar^{12+} ion beam intensity in the very same manner as in Figure 16. The Huber regression can grasp the main features of the data set, but the model seems to have difficulties on predicting ion beams with a high intensity, since there are no predictions above $80\,\mu A$.

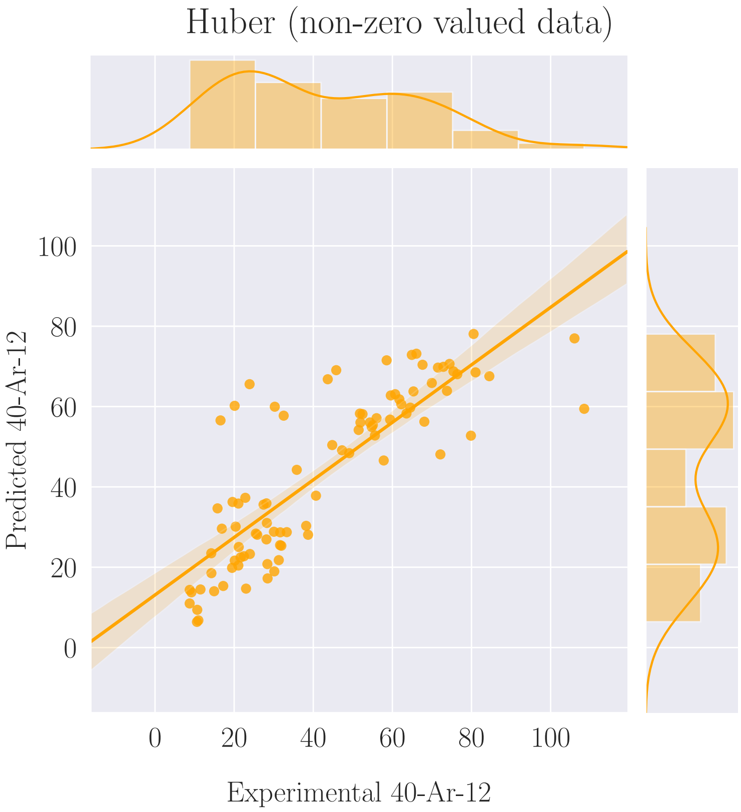


Figure 17: Experimental ion beam intensities versus predicted ones by the Huber regression method when zero valued experimental values were ignored. The values are in the units of μA. The side panels illustrate the distribution of the measured and predicted ion beam intensities.

4.8 Optimization and visualization

Since all the models – multilinear, Lasso and Ridge regression – gave very similar results in the previous sections, there is no model that would be clearly better than the others. According to the R^2 scores tabulated in Tables 2, 3 and 4, it seems that the most simple method, the ordinary multilinear regression, seems to work a bit better than the two other ones. Thus the multilinear regression models were chosen to be studied further.

One main source of motivation for the whole study was to make a model on top of the experimental data and get some hints what are the parameters for the maximal ion beam production with a specific charge state. The multilinear regression models were optimized while ignoring the zero valued measurement points. Since the optimization problem itself was not complicated – the models were linear – the decision about the optimization algorithm to be used was not critical. It was only required that the optimization algorithm can handle constrained optimization problems, since our parameter space was constrained accoring to Table 1 and Equations (4.1)- (4.5). The method Sequential Least SQuares Programming (SLSQP) fulfilled the requirements for this study and it was implemented in SciPy library, thus SLSQP was chosen.

The optimization SLSQP uses Jacobians of inequality constraints to find the optimum. One of our constraints was not differentiable and the easiest solution to handle the problem was to divide the optimization problem in two parts, solve them separately, and choose the optimal solution from the two subsolutions. The division was performed along the variable x_4 and is illustrated in Figure 18.

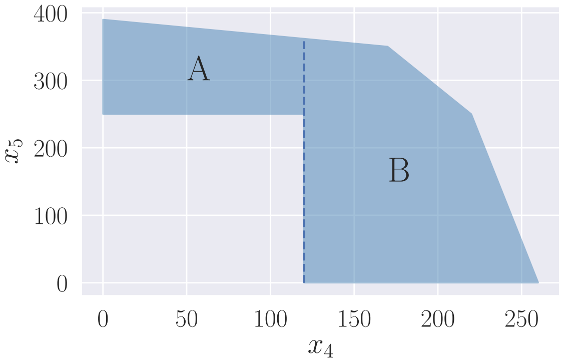


Figure 18: The optimization problem solved in two subspaces A and B, and the final solution was the best out of these two subsolutions. Variables x_4 and x_5 are unitless.

The maximal ion beam intensities and the corresponding optimal parameters for each ion beam can be found in Table 5. The values are rounded up to two decimals. As it was known due to the linearity of the models, optimal solutions are found at the borders of allowed parameter values. Depending on the fact whether the model was ascending or descending as a function of a specific variable, the optimal parameter value was the smallest or the greatest allowed value.

		Optimal parameters				
Ion beam	Max. intensity (μA)	x_1	x_2	x_3	x_4	x_5
^{40}Ar^{6+}	19.50	0.0	100.0	1000.0	260.0	0.0
^{40}Ar^{7+}	23.22	45.0	100.0	100.0	260.0	0.0
^{40}Ar^{8+}	64.64	45.0	100.0	100.0	260.0	0.0
^{40}Ar^{9+}	78.24	45.0	100.0	100.0	260.0	0.0
^{40}Ar^{11+}	107.62	45.0	2000.0	100.0	260.0	0.0
^{40}Ar^{12+}	104.26	45.0	2000.0	100.0	220.0	250.0
^{40}Ar^{13+}	85.25	45.0	2000.0	100.0	220.0	250.0
^{40}Ar^{14+}	64.25	45.0	2000.0	100.0	220.0	250.0
^{40}Ar^{16+}	24.39	0.0	2000.0	1000.0	0.0	390.0
^{16}O^{3+}	48.69	45.0	100.0	1000.0	0.0	390.0
^{16}O^{5+}	130.35	45.0	100.0	100.0	0.0	390.0
^{16}O^{7+}	315.22	0.0	2000.0	100.0	0.0	390.0

Table 5: Maximum ion beam intensity and the optimal parameters when using multilinear regression. Zero-valued data was ignored. Parameter x_1 is in units of $-V$ and parameters x_2 and x_3 are in W. Parameters x_4 and x_5 are unitless.

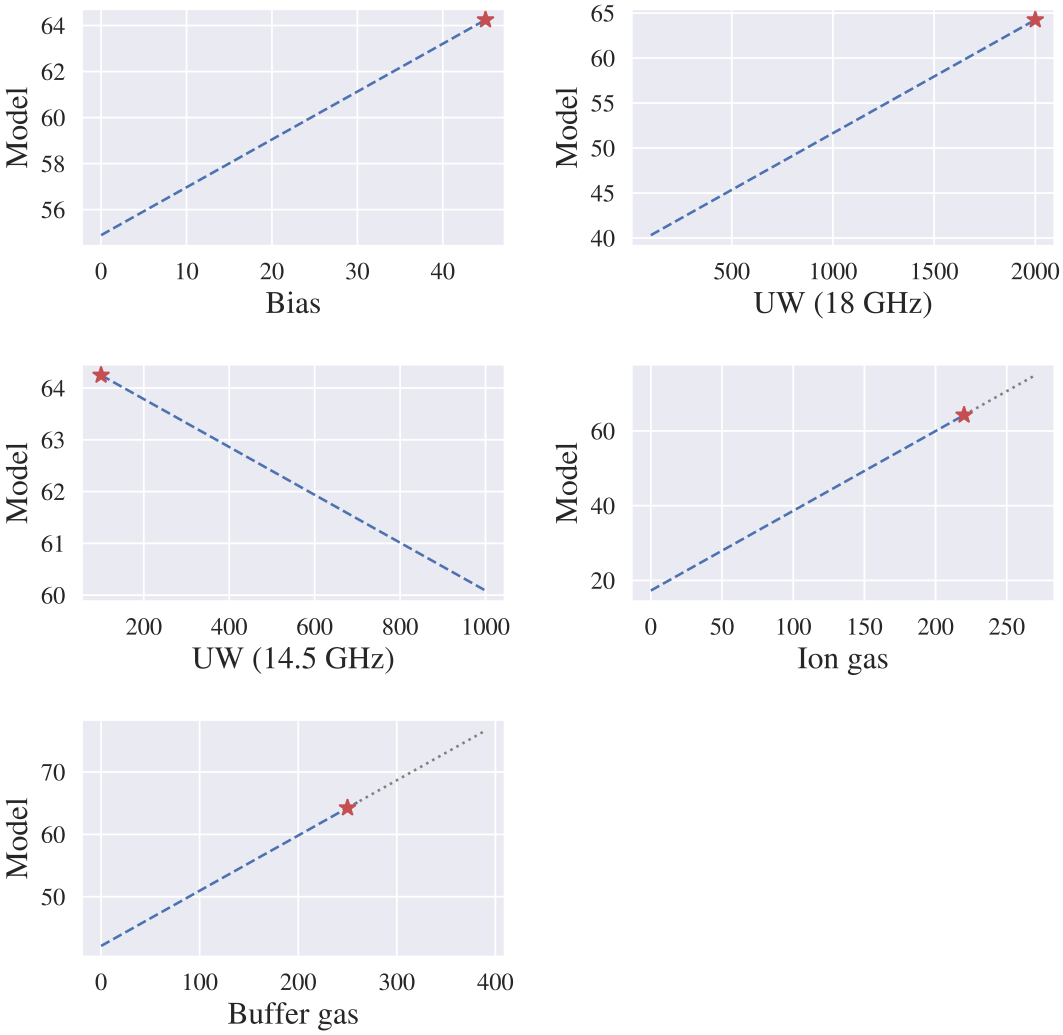


Figure 19: The result of the optimization for the ion beam of $^{40}\mathrm{Ar}^{14}$ in the parameter space. Since the model is linear, it is simple to convince the user why the obtained result is reasonable: The optimal solution must be found in one of the extremes of allowed parameter values. Since the ion gas valve position (x_4) and the buffer gas valve position (x_5) are constrained by the inequalities, the range of allowed values depend on each other.

Finally the obtained results are visualized in the spirit of XAI in Figure 19. Since the studied models were linear, even a simple visualization with respect to different input parameters can illustrate the obtained result well. Since the allowed values of the buffer gas valve position (x_5) and the ion gas valve position (x_4) depend on each other, there are two types of dashed lines in figures. Blue, bold line corresponds to the allowed values and the grey, thinner line illustrates the prohibited values that change when the parameters change. The user can observe that the obtained optimal parameters lie in the extremes of allowed ranges so that the corresponding parameter gives the maximal contribution.

4.9 Discussion

In this chapter the ECR ion source and the maximal ion beam pruduction have been discussed. The main goal of the study was to find out if the machine learning methods could be used to study the parameter space of an ECR ion source. Since all the relationships between the parameters are not understood, the optimal parameters for the maximal ion beam production are not known.

One of the major restrictions in this project was the amount of experimental data points. The measurement time was limited into one week, thus the amount of data points was approximately limited to 100. Since there were five input parameters x_i, it meant that there could be 2-3 data points per each parameter and the models that could be applied reasonably here would be simple linear models. That is why the multilinear regression, Lasso regression and Huber regression were selected.

Another restriction was the zero-valued data points. Some of the data points had zero ion beam intensity. Unfortunately, there could be two kind of zeros in the data: Some of the ion beam intensities were so small that they could not be measured, but there were also some non-zero intensities that were overlapping with other pulses. The ion beam intensity could not be found out for those data points, and they were marked as zero in the measurement log. As a consequence, the regression models faced difficulties to handle the zero-valued data. If the zero-valued data was ignored, the used data set was even smaller.

When the zero-valued data was ignored, multilinear regression models were describing the data set related to the lower charge states (e.g. $^{40}\mathrm{Ar}^{6+}$) relatively well. However, all the regression models, namely multilinear, Lasso and Huber regression, had difficulties to reproduce the data related to the higher charge states such as $^{40}\mathrm{Ar}^{14+}$. This is most likely explained by the fact that the relationships between the input parameters and the ion beam intensities are much more complicated in reality and cannot be modelled linearly. In order to study more complicated relationships, more experimental data is needed.

The Huber regression models were not found to be better than the multilinear and Lasso regression models. This indicates that the data did not include many outliers and robust techniques were not needed. All the three different types of methods – multilinear, Lasso and Huber – performed equally and no significant differences could be found in the results.

What comes to the research question, one can conclude that the parameters of an ECR ion source can be studied by applying machine learning methods. Despite the simplicity of used methods, the results are reasonably good for the lower charge states. If there were more experimental data, more complicated relationships could be studied and the performance of machine learning models would most likely get better.

5 Conclusions

This MSc thesis was dealing with artificial intelligence (AI) and particularly its subfield explainable artificial intelligence (XAI). In the field of XAI, one is aiming to develop tools that would make AI methods more explainable and trustworthy. The thesis started with a brief introduction to the history of AI, which was followed by the introduction to the key terminology of XAI. Next, explanations and their characteristic were discusssed and different types of approaches to explain and interpret were presented. The second part of the thesis consisted of a small project, in which the parameter space of an ECR ion source was studied, and the optimal parameters for the certain ion beam intensities were found out.

The literature review revealed that there has been a lot of published research on explainable artificial intelligence during the last few decades. Unfortunately, the key terminology related to explainable artificial intelligence is not fixed in general, and different communities and research groups use different names for similar or identical concepts. In addition, the goodness of explanations is most often evaluated by AI experts and there seems to be a lack of collaboration with the cognitive scientists. The collaboration would be fruitful, since the experts who created the AI system in question are not the most suitable people to judge the explanations.

In the pragmatic project, the parameters of an ECR ion source were studied via simple linear regression models. The goal of the project was to estimate if the parameter space of an ECR ion source could be approximated by machine learning models. As an outcome one can conclude that the parameter space can be studied via machine learning methods, but more experimental data is needed for the better accuracy. In addition, the better performance could be achieved when the zero-valued ion beam intensities would be subcategorized into zero-valued beams and overlapping pulses. If there were more experimental data, more complicated relationships between the input parameters and the ion beam intensities could be studied. In this study we were limited to examine only linear models due to the amount of the input parameters and data.

As a final conclusion, both of the topics discussed in this thesis can be studied further. The parameter space of an ECR ion source can be studied via machine learning methods if more experimental data is available. In addition, there is a lot of research work on the explanation methods that has been done during the last decades and this knowledge can be used to explain the outcomes of present complex AI systems. However, the present systems are more and more complicated to be explained, and the explanation techniques or the AI systems must be developed further.

Bibliography

Adler, Philip, Casey Falk, Sorelle A. Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2016. *Auditing Black-box Models for Indirect Influence.* arXiv: `1602.07043 [stat.ML]`.

Amilhastre, Jérôme, Hélène Fargier, and Pierre Marquis. 2002. "Consistency restoration and explanations in dynamic CSPs—Application to configuration". *Artificial Intelligence* 135 (): 199–234.

Anthony, Lauren. 2019. "Smart trashcan counts your wasted meatballs". Visited on April 6, 2019. `https://www.reuters.com/video/2019/03/22/smart-trashcan-counts-your-wasted-meatba?videoId=529118442`.

Assche, Anneleen, and Hendrik Blockeel. 2007. "Seeing the Forest Through the Trees: Learning a Comprehensible Model from an Ensemble". *Machine Learning: ECML 2007* (): 418–429.

Baehrens, David, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Mueller. 2009. *How to Explain Individual Classification Decisions.* arXiv: `0912.1128 [stat.ML]`.

Biran, Or, and Courtenay V. Cotton. 2017. "Explanation and Justification in Machine Learning : A Survey Or". *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI).*

Biran, Or, and Kathleen McKeown. 2017. "Human-Centric Justification of Machine Learning Predictions". *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17:* 1461–1467.

Borowiec, Steven. 2018. "AlphaGo seals 4-1 victory over Go grandmaster Lee Sedol". `https://www.theguardian.com/technology/2016/mar/15/googles-alphago-seals-4-1-victory-over-grandmaster-lee-sedol`.

Buchanan, Bruce, and Edward Shortliffe. 1984. *Rule-based Expert System – The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. “Intelligible Models for HealthCare”. *The 21th ACM SIGKDD International Conference* (): 1721–1730.

Cawsey, Alison. 1994. “Developing an Explanation Component for a Knowledge-Based System: Discussion”. *Expert Systems with Applications* 8:527–531.

Chandrasekaran, B., M. C. Tanner, and J. R. Josephson. 1989. “Explaining control strategies in problem solving”. *IEEE Expert* 4 (1): 9–15.

Cioppa, Thomas M., and Thomas W. Lucas. 2007. “Efficient Nearly Orthogonal and Space-Filling Latin Hypercubes”. *Technometrics* 49 (1): 45–55. `http://www.jstor.org/stable/25471274`.

Copeland, B.J. 2018. “Encyclopædia Britannica”. Visited on December 29, 2018. `https://www.britannica.com/technology/artificial-intelligence`.

Copeland, Michael. 2016. “What’s the Difference Between Artificial Intelligence, Machine Learning and Deep Learning?” Visited on May 10, 2020. `https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/`.

Cortez, Paulo, and Mark J. Embrechts. 2013. “Using sensitivity analysis and visualization techniques to open black box data mining models”. *Information Sciences* 225:1–17.

Doshi-Velez, Finale, and Been Kim. 2018. “Towards A Rigorous Science of Interpretable Machine Learning”. `https://arxiv.org/abs/1702.08608`.

Došilović, Filip, Mario Brcic, and Nikica Hlupic. 2018. “Explainable Artificial Intelligence: A Survey.” *MIPRO 2018 - 41st International Convention Proceedings.*

Elizalde, Francisco, Luis Sucar, Julieta Noguez, and Alberto Reyes Ballesteros. 2009. “Generating Explanations Based on Markov Decision Processes”. *MICAI 2009. Lecture Notes in Computer Science* 5845.

Elizalde, Francisco, Luis Sucar, Alberto Reyes Ballesteros, and Pablo Debuen. 2007. “An MDP Approach for Explanation Generation.” *AAAI Workshop - Technical Report* (): 28–33.

Enqvist, Kari. 2018. “Kari Enqvistin kolumni: Tekoäly tulee tuhoamaan ajattelun”. Visited on July 26, 2018. https://yle.fi/uutiset/3-10318808.

Ford, Martin, and Geoff Colvin. 2015. “Will robots create more jobs than they destroy?” Visited on July 26, 2018. https://www.theguardian.com/technology/2015/sep/06/will-robots-create-destroy-jobs.

Garnham, A. 1988. *Artificial Intelligence: An Introduction.* Introductions to modern psychology. Routledge & Kegan Paul.

Geller, R. 1996. *Electron Cyclotron Resonance Ion Sources and ECR Plasmas.* Taylor & Francis. https://books.google.fi/books?id=DtcVZnquQCAC.

Goodman, Bryce, and Seth Flaxman. 2016. “European Union regulations on algorithmic decision-making and a “right to explanation””. Visited on November 18, 2018. https://arxiv.org/pdf/1606.08813.pdf.

Gunning, David. 2018. “Explainable Artificial Intelligence (XAI)”. https://www.darpa.mil/program/explainable-artificial-intelligence.

Harradon, Michael, Jeff Druce, and Brian Ruttenberg. 2018. “Causal Learning and Explanation of Deep Neural Networks via Autoencoded Activations”. Visited on October 7, 2018. https://arxiv.org/pdf/1802.00541.pdf.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Hawking, Stephen, Stuart Russell, Max Tegmark, and Frank Wilczek. 2014. “Stephen Hawking: Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?” Visited on December 29, 2018. https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html.

Herlocker, Jon, Joseph Konstan, and John Riedl. 2001. “Explaining Collaborative Filtering Recommendations”. *Proceedings of the ACM Conference on Computer Supported Cooperative Work* ().

Hilton, D. J. 1990. “Conversational processes and causal explanation”. *Psychological Bulletin* 107:65–81.

Holley, Peter. 2018. “The World Bank’s latest tool for fighting famine: Artificial intelligence”. Visited on April 6, 2019. `https://www.washingtonpost.com/technology/2018/09/23/world-banks-latest-tool-fighting-famine-artificial-intelligence/?noredirect=on&utm_term=.bb46cb3afd1f`.

Hoyer, Patrik, Dominik Janzing, Joris Mooij, Jonas Peters, and Bernhard Schölkopf. 2008. “Nonlinear causal discovery with additive noise models”. *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference* (): 689–696.

Hughes, A., and D. Grawoig. 1971. *Statistics: A Foundation for Analysis.* Addison-Wesley Publishing Company.

Israelsen, Brett W. 2017. “"I can assure you [...] that it’s going to be all right" – A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships”. Visited on November 24, 2018. `https://arxiv.org/abs/1708.00495`.

Jakulin, Aleks, Martin Možina, Janez Demsar, Ivan Bratko, and Blaz Zupan. 2005. “Nomograms for visualizing support vector machines.” *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (): 108–117.

Junker, Ulrich. 2001. “QuickXPlain: Conflict Detection for Arbitrary Constraint Propagation Algorithms”. *In IJCAI-2001 Workshop on Modeling and Solving Problems with Constraint:* 75–82.

Jussien, Narendra, and Vincent Barichard. 2000. “The PaLM system: Explanation-based constraint programming”. *Proceedings of TRICS: Techniques for Implementing Constraint Programming Systems, A Post-conference Workshop of CP 2000* ().

Jyväskylän yliopisto. 2018. “Tekoäly opetettiin analysoimaan syöpäkuvia”. Visited on November 17, 2019. `https://www.jyu.fi/fi/ajankohtaista/arkisto/2018/07/tekoaly-opetettiin-analysoimaan-syopakuvia`.

Kalvas, T., H. Koivisto, and O. Tarvainen. 2018. “Status of new 18GHz ECRIS HIISI”. *AIP Conference Proceedings* 2011 (1): 040006.

Kalvas, T., O. Tarvainen, H. Koivisto, and K. Ranttila. 2014. “Thermal Design of Refridgerated Hexapole 18 GHZ ECRIS HIISI”. *Proceedings of ECRIS 2014 : the 21st International Workshop on ECR Ion Sources:* 114–119.

Kaplan, Jerry. 2016. *Artificial Intelligence - What Everyone Needs to Know.* Oxford University Press.

Karpathy, Andrej, Justin Johnson, and Li Fei-Fei. 2015. *Visualizing and Understanding Recurrent Networks.* arXiv: 1506.02078 [cs.LG].

Khan, Omar, Pascal Poupart, and James Black. 2009. “Minimal Sufficient Explanations for Factored Markov Decision Processes”. *ICAPS 2009 - Proceedings of the 19th International Conference on Automated Planning and Scheduling* ().

Kharpal, Arjun. 2017. “Elon Musk: Humans must merge with machines or become irrelevant in AI age”. Visited on December 29, 2018. https://www.cnbc.com/2017/02/13/elon-musk-humans-merge-machines-cyborg-artificial-intelligence-robots.html.

Kim, Been, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. “Examples are not enough, learn to criticize! Criticism for Interpretability”. In *Advances in Neural Information Processing Systems 29,* edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 2280–2288. Curran Associates, Inc.

Kissinger, Henry A. 2018. “How the Enlightenment Ends”. Visited on July 26, 2018. https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

Koivisto, H., O. Tarvainen, T. Kalvas, K. Ranttila, P. Heikkinen, D. Xie, G. Machicoane, T. Thuillier, V. Skalyga, and I. Izotov. 2014. “HIISI, New 18 GHZ ECRIS for the JYFL Accelerator Laboratory”. *Proceedings of ECRIS 2014 : the 21st International Workshop on ECR Ion Sources:* 99–103.

Kononenko, Igor, Erik trumbelj, Zoran Bosnic, Darko Pevec, Matjaz Kukar, and Marko Robnik-Sikonja. 2013. “Explanation and Reliability of Individual Predictions”. *Informatica (Slovenia)* 37:41–48.

Laakkonen, Johanna. 2018. “Bussikuskit ja kirjastonhoitajat pitäisi saada suunnittelemaan tekoälyä”. Visited on December 28, 2018. https://yle.fi/uutiset/3-10318808.

Lacave, Carmen, and Francisco J. Diez. 2002. “A Review of Explanation Methods for Bayesian Networks”. *Knowl. Eng. Rev.* (USA) 17 (2): 107–127.

Lehmann, E.L. 1983. *Theory of point estimation.* Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley. https://books.google.fi/books?id=YdgwNZJ-YgUC.

Lei, Tao, Regina Barzilay, and Tommi Jaakkola. 2016. *Rationalizing Neural Predictions.* arXiv: 1606.04155 [cs.CL].

Letham, Benjamin, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. “Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model”. *The Annals of Applied Statistics* 9, number 3 (): 1350–1371.

Li, Jiwei, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. *Visualizing and Understanding Neural Models in NLP.* arXiv: 1506.01066 [cs.CL].

Lim, Brian, and Anind Dey. 2010. “Toolkit to support intelligibility in context-aware applications”. *UbiComp’10 - Proceedings of the 2010 ACM Conference on Ubiquitous Computing* (): 13–22.

Lubsen, Jacobus, J Pool, and E Does. 1978. “A Practical Device for the Application of a Diagnostic or Prognostic Function”. *Methods of information in medicine* 17 (): 127–9.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. “Viualizing data using t-SNE”. *Journal of Machine Learning Research* 9 (): 2579–2605.

Martens, David, Bart Baesens, Tony [Van Gestel], and Jan Vanthienen. 2007. “Comprehensible credit scoring models using rule extraction from support vector machines”. *European Journal of Operational Research* 183 (3): 1466–1476. http://www.sciencedirect.com/science/article/pii/S0377221706011878.

Martens, David, and Foster Provost. 2014. “Explaining Data-Driven Document Classifications”. *MIS Quarterly* 38 (): 73–100.

Miller, Tim. 2019. “Explanation in artificial intelligence: Insights from the social sciences”. *Artificial Intelligence* 267:1–38.

Montavon, Grégoire, Wojciech Samek, and Klaus Muller. 2018. “Methods for interpreting and understanding deep neural networks”. *Digital Signal Processing: A Review Journal* 73 (): 1–15.

Možina, Martin, Janez Demšar, Michael Kattan, and Blaz Zupan. 2004. “Nomograms for Visualization of Naive Bayesian Classifier”. *PKDD* (): 337–348.

Nilsson, Nils J. 1998. “Introduction to Machine Learning”. Visited on June 6, 2020. `https://ai.stanford.edu/people/nilsson/mlbook.html`.

Nugent, Conor, Dónal Doyle, and Padraig Cunningham. 2009. “Gaining Insight through Case-Based Explanation”. *J. Intell. Inf. Syst.* 32 (): 267–295.

Papadimitriou, Alexis, Panagiotis Symeonidis, and Yannis Manolopoulos. 2012. “A generalized taxonomy of explanations styles for traditional and social recommender systems”. *Data Mining and Knowledge Discovery* 24 (): 555–583.

Park, Dong Huk, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2018. *Multimodal Explanations: Justifying Decisions and Pointing to the Evidence.* Visited on October 14, 2018. `https://arxiv.org/abs/1802.08129`.

Peltola, Tomi. 2018. “Local Interpretable Model-agnostic Explanations of Bayesian Predictive Models via Kullback-Leibler Projections”. Visited on November 11, 2018. `https://arxiv.org/abs/1810.02678`.

Petelet, Matthieu, Bertrand Iooss, Olivier Asserin, and Alexandre Loredo. 2009. *Latin hypercube sampling with inequality constraints.* arXiv: `0909.0329 [stat.CO]`.

Poole, David, Alan Mackworth, and Randy Goebel. 1998. *Computational Intelligence: a logical approach.* Oxford University Press.

Ribeiro, Marco, Sameer Singh, and Carlos Guestrin. 2016. "“Why Should I Trust You?”: Explaining the Predictions of Any Classifier". *Conference: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (): 97–101.

Robnik-Sikonja, Marko, Aristidis Likas, Constantinos Constantinopoulos, Igor Kononenko, and Erik Strumbelj. 2011. "Efficiently Explaining Decisions of Probabilistic RBF Classification Networks". *Adaptive and Natural Computing Algorithms: 10th International Conference, ICANNGA 2011* 6593 (): 169–179.

Rossi, F., P. van Beek, and T. Walsh. 2006. *Handbook of Constraint Programming*. Elsevier Science.

Rudin, Cynthia, Benjamin Letham, and David Madigan. 2013. "Learning Theory Analysis for Association Rules and Sequential Event Prediction". *The Journal of Machine Learning Research* 14 (): 3441–3492.

Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller. 2017. "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models", 1 (): 1–10.

Shortliffe, Edward, and Bruce Buchanan. 1975. "A Model of Inexact Reasoning in Medicine". *Mathematical Biosciences* 23 (): 351–379.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2013. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". *CoRR* ().

Suermondt, H, and G Cooper. 1992. "An evaluation of explanations of probabilistic inference". *Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care* (): 579–85.

Suzuki, Kenji. 2011. *Artificial Neural Networks - Methodological Advances and Biomedical Applications.*

Swartout, William, and Johanna Moore. 1993. "Explanation in Second Generation Expert Systems" (): 543–585.

Symeonidis, Panagiotis, Alexandros Nanopoulos, and Yannis Manolopoulos. 2009. “MoviExplain: A recommender system with explanations”. *RecSys’09 - Proceedings of the 3rd ACM Conference on Recommender Systems* (): 317–320.

Tamagnini, Paolo, Josua Krause, Aritra Dasgupta, and Enrico Bertini. 2017. “Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations”, (Chicago, IL, USA), HILDA’17.

Teach, Randy L., and Edward H. Shortliffe. 1981. “An analysis of physician attitudes regarding computer-based clinical consultation systems”. *Computers and Biomedical Research* 14 (6): 542–558. `http://www.sciencedirect.com/science/article/pii/0010480981900124`.

Thrun, Sebastian. 1994. “Extracting Rules from Artificial Neural Networks with Distributed Representations”, (Denver, Colorado), NIPS’94: 505–512.

Tiainen, Antti. 2018. “Kaikille avoimista tekoälyopinnoista tuli Helsingin yliopiston kaikkien aikojen suosituin kurssi”. Visited on December 28, 2018. `https://www.hs.fi/teknologia/art-2000005817486.html`.

Timmer, Sjoerd T., John-Jules Ch. Meyer, Henry Prakken, Silja Renooij, and Bart Verheij. 2017. “A two-phase method for extracting explanatory arguments from Bayesian networks”. *International Journal of Approximate Reasoning* 80:475–494. `http://www.sciencedirect.com/science/article/pii/S0888613X16301402`.

Tullio, Joe, Anind Dey, Jason Chalecki, and James Fogarty. 2007. “How it works: A field study of non-technical users interacting with an intelligent system”. *Conference on Human Factors in Computing Systems - Proceedings* (): 31–40.

Turing, A. M. 1950. “Computing Machinery and Intelligence”. *Mind* 49:433–460.

Tzeng, F., and K. Ma. 2005. “Opening the black box - data driven visualization of neural networks”. *VIS 05. IEEE Visualization, 2005:* 383–390.

Vlek, Charlotte S., Henry Prakken, Silja Renooij, and Bart Verheij. 2016. “A method for explaining Bayesian networks for legal evidence with scenarios.” *Artificial Intelligence and Law* 24:285–324.

Wallace, Richard, and Eugene Freuder. 2001. "Explanations for Whom?" *CP01 Work-shop on User-Interaction in Constraint Satisfaction.*

Wang, Fulton, and Cynthia Rudin. 2014. *Falling Rule Lists.* arXiv: `1411.5899 [cs.AI]`.

Wang, Tong, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. 2015. *Or's of And's for Interpretable Classification, with Application to Context-Aware Recommender Systems.* arXiv: `1504.07614 [cs.LG]`.

Wikipedia. 2018. "Tekoäly". Visited on December 28, 2018. `https://fi.wikipedia.org/wiki/Teko%C3%A4ly`.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.* arXiv: `1502.03044 [cs.LG]`.

Yap, Ghim-Eng, Ah-Hwee Tan, and Hwee-Hwa Pang. 2008. "Explaining Inferences in Bayesian Networks". *Applied Intelligence* 29 (): 263–278.

Ye, L. Richard, and Paul E. Johnson. 1995. "The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice". *MIS Quarterly* 19 (2): 157–172.

Zeiler, Matthew D., and Rob Fergus. 2014. "Visualizing and Understanding Convolutional Networks". Edited by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. *Computer Vision – ECCV 2014:* 818–833.

Zhou, Zhi-Hua, Yuan Jiang, and Shifu Chen. 2003. "Extracting Symbolic Rules from Trained Neural Network Ensembles". *AI Communications* 16 (): 3–15.