# Does Machine Learning Need a Theory of Everything?

**Tim Mondorf**

Bachelor Project presented for the degree of
BSc in Machine Learning and Data Science


Department of Science Education (IND) and
Department of Computer Science (DIKU)
Supervisor: Professor Henrik Kragh Sørensen, Ph.d.
(IND)
22 October 2023

# Abstract

*Explainable Machine Learning (ML) is a property not just of models but of ML's scientific identity. This identity, as illustrated by a case study of one curriculum and 4 seminal articles, is disparate and reflects a preference for probabilistic rather than general statements. This bachelor project challenges this preference and discusses a number of ways to potentially render ML less disparate and more explainable. A Standard Model, originally proposed by Hu, is amended to facilitate a structured discussion of accuracy, model desiderata, context and ethical and legal requirements of all ML models.*

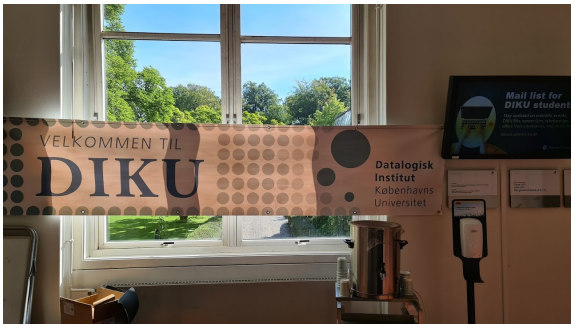


Figure 1: DIKU, 28 August 2021

## Volume

Chapters 1-5 of this bachelor project comprise 77.709 characters, equivalent to 32 normal pages of 2.400 characters.

## Title in Danish

The title of this project in Danish is 'Har ML behov for en Teori om Alting?'

# Contents

# Preface

I would like to thank my supervisor, professor Henrik Kragh Sørensen, Ph.d., (IND), for many fruitful conversations. Henrik taught the course Philosophy of Computer Science which informs this project. After submission of this project and prior to the presentation, I will have the opportunity to discuss the topic at the workshop 'Verifiable and Robust AI', organized by Digital Research Center Denmark (DIREC) in Sønderborg, 6 to 10 November 2023. I would also like to thank my fellow student Malte Ro Buchwald for reading and commenting on a draft on 17 October 2023. I would finally like to thank teachers, staff and not least all students for an exciting time at the Bachelor of Science (BSc) in ML and Data Science at the Department of Computer Science of the University of Copenhagen (DIKU).

# Chapter 1

# Introduction

The EU General Data Protection Regulation (GDPR) grants individuals a right to 'meaningful information (about) automated decision making' involving themselves. (EU, 2018, par. 13.2.f, [1]). At the same time, such meaningful explanations of ML models are not always available. Therefore, Explainable ML is becoming an ever more active research field.

Where previous work has focussed on explainability or lack thereof as a property of models (Søgaard, 2022, p. 1, [2]) or of projects (Jiang et al., 2021, p. 1, [3]), this bachelor project will study explainability as a property of ML's scientific identity.

This perspective on Explainable ML is inspired by my own personal learning process. I have an MSc in economics and have been a part-time assistant professor ('ekstern lektor') of economics at KU and CBS. When I enrolled as a student of ML, I found the teaching of ML to be less intuitive than the teaching of social sciences with which I was familiar. When I reflected further on this subjective impression, it seemed very much related to differences

in scientific identity. Economics operates with comprehensive categories: macroconomics versus microeconomics, demand versus supply, static versus dynamic, real versus monetary, private sector versus public sector or exogeneous model variables versus endogeneous model variables. Each side of such dichotomies, described further in Appendix 6.1, is equally valid and approximately equally well researched. To a beginner, ML did not appear to provide the same comprehensive categorizations. In particular, there were few frameworks that comprised all use cases in the way that the broad categorizations listed above comprise almost all use cases of economics.

I then came across Hu's 2022-article "Towards a Standard Model of ML" on X. The article also had a comparative perspective, discussing the scientific identities of ML and physics. Hu described 'disparate, narrowly focussed methods [in ML]... that made development difficult' (Hu et al., 2022, p. 2, [4]). The article proposed a potential remedy in the form of one model that would comprise all ML models as special cases. I decided to use that article as the starting point for my bachelor project.

I was born in 1968, studied economics between 1988 and 1994 and enrolled at DIKU in 2021. It is outside the scope of this project to examine whether I remember my study of economics as being less complicated than it actually was at the time. It is not uncommon for university students to feel confused during their studies. The differences in learning experience at the age of 20 versus the age of 55, or in the historical periods of the late 1980s and early 1990s versus the 2020s, are also not studied. The comparison between ML and economics would not be fair without addressing the issue of accuracy which is done in Appendix 6.1.3.

## 1.1 Research philosophy, process and strategy

This bachelor project is the study of an idea, Systematic Methods in ML, that presumably will have the consequence of ML becoming more Explainable, leading to the fulfilment of the right to an explanation, along with other consequences for society. My project therefore reflects a pragmatist research philosophy, i.e., the belief that the meaning of ideas can be clarified by tracing their 'practical consequences' as expressed by Peirce cited by Legg (Legg, 2021, par. 5, [5]). Anoter interpretation of pragmatism is that 'if a philosophical theory does not contribute directly to social progress then it is not worth much' (Legg, 2021, par. 1, [5]). The research philosophy is therefore pragmatism.

The project originates in my personal learning experience in ML which I attempt to formalize and confront with observations. The work of Popper is well suited to research processes that begin with ideas that involve 'a leap of the imagination' even at the risk of those theories later being proven incorrect (Thornton, 2022, par. 1, [6]). Popper proposes to begin with defining a statement, then performing a formal step verifying whether the statement is 'internally consistent or contradictory', a semi-formal step verifying whether the statement is 'a theory or simply, e.g., tautological', verifying whether the statement is truly innovative, and finally testing the empirical application of the statement. In fact, Popper argued that corroboration of existing theories did not increase human knowledge. Human knowledge could only be increased by an iterative process of formulating, rejecting and adapting 'bold conjectures'. The original definition by Popper refers to 'theories' even

in the initial stages of conjecture without distinction to 'hypotheses' and this project will do the same (Popper, 2002, Part 1., chap. 1, sect. 3, p. 31, [7]). Both the iterative rejection and the eventual corroboration of innovative theories would improve human knowledge. Popper also argued that science was structured around problems and that scientists were 'problem-solvers' rather than observers of fact. Explainable ML is a recognized problem to be solved. Popper's research process as described by Thornton is therefore a good fit for this bachelor project and will be referred to as the Popper-Thornton process (Thornton, 2022, par. 3-6, [6]). This research process is deductive.

Since the inspiration is my personal learning process, an obvious choice of research strategy is to perform a case study of some of the seminal work that I have encountered. This requires a framework that allows the understanding of a phenomenon more deeply in its context at the expense of being able to generalize the conclusions beyond this context, and case study is exactly such a framework. Most case studies represent an inductive approach, i.e., the researcher immerses himself in a context with few pre-defined theories, but deductive case studies also exist. The chosen research strategy is therefore case study.

Chapter 2 performs the first four steps of the Popper-Thornton process. The first step is accomplished by defining two Theories, called the Theory of Disparate Property (2.1.1) and the Theory of Impact (2.1.2). These Theories refer to concepts, ML, Systematic Methods and Explainable ML, that are then defined in Definitions (2.2.1), (2.3.1) and (2.4.1). The second and third steps are completed by verifying that with these definitions, the statements can be considered Theories. The fourth step is completed in Section

2.5 by confirming that the Theories are innovative relative to existing ML literature. For the purpose of testing the Theories, their empirical application is discussed in Section 2.6.

Chapter 2.6 performs the fifth step of the Popper-Thornton process, conducting the case studies and discussing whether they corroborate the two Theories.

The Popper-Thornton process is iterative, producing new knowledge through defining and testing more specific versions of the initial theories. Chapter 4 attempts such a next iteration, proposing a set of Systematic Methods and illustrating the consequences of these in the form of a mock scientific article in Appendix 6.10. These proposals again reflect the philosophy of pragmatism, i.e., that ideas must attempt some form of progress.

Chapter 5 presents the conclusions.

# Chapter 2

# Is ML difficult to explain because it is disparate?

## 2.1 Defining two Theories

The purpose of defining a Theory is to express the above subjective impression in a way that can be confronted with observations. The theory should reflect Hu's view of 'disparate and narrowly focussed methods'. The literal definition of disparate is 'made up of parts that are very different from each other' (Oxford Learner's Dictionary, year not indicated, entry for 'disparate', [8]). Alternatively, the theory could have referred to various stages of maturity for ML as a new discipline. It could also have allowed for the possibility that, while certain elements of ML might appear disparate, they form part of a comprehensive structure that is not immediately visible to a beginner. In the end, definition (2.1.1) is selected.

**Definition 2.1.1 (The Theory of Disparate Property)** *ML is disparate.*

The idea that reducing this disparate property will benefit explainability

is expressed in the second Theory in Definition (2.1.2).

**Definition 2.1.2 (The Theory of Impact)** *Systematic Methods that reduce the disparate property of ML will benefit Explainable ML.*

Before the next steps of the Popper-Thornton process can be performed, the relevant terms must be defined.

## 2.2 Defining ML

Bringsjord writes that 'ML is concerned with building systems that improve their performance on a task when given examples of ideal performance on the task, or improve their performance with repeated experience on the task.' (Bringsjord et al., 2018, sect. 4.1., par. 1, [9]). Where economics is founded on the notion of optimality, such as discussed in Appendix 6.1.4, ML here represents passively observing examples of ideal performance that are somehow external to ML. ML means being handed the data from somebody else, being assigned the problem to be analyzed by somebody else, starting from an arbitrary initial condition and being content to see the error become a little smaller for each iteration.

Most views of ML are more volontarist. Turing Award winner Gray called data science 'the fourth paradigm of science' (Gray, cited by the ACM, 2021, p. 1, [10]). The ACM quotes Gray and lists ML as a 'knowledge area' under data science, further defined as 'algorithms for identifying patterns in data' (ACM, 2021, p. 94, [10]). Raschka has a similar definition of ML as 'the application and science of algorithms that make sense of data... and the

most exciting field of all the computer sciences' (Raschka, 2022, chap. 1, par. 1, [11]). ACM's and Raschka's definitions are not much different from Naur's famous but much earlier 'Datalogy: the science of the nature and the use of data' (Naur, 1966, p. 1, [12]). Attempts to define ML therefore lead to discussions of its relationship to computer science.

Understanding this relationship requires analysis of ontology, i.e, the objects being studied, and epistemology, i.e., the nature of the knowledge being produced.

Hansen defines the ontology of computer science as 'data, representations of objects and ideas' (Hansen cited by Kragh Sørensen et al., 2023, chap. 1, p. 3, [13]). Along with the above quotes by the ACM, Raschka and Naur, this would indicate that the ontology of ML is not too different from the ontology of computer science.

The difference seems larger when it comes to epistemology. Hansen (cited by Kragh Sørensen et al., 2023, chap. 1, p. 3, [13]) states that computer science produces 'general principles for data processing'. A computer scientist can produce a scientific claim that the mergesort algorithm has a lower time complexity bound than the insertion sort algorithm and then corroborate this claim empirically (Sedgewick et al., 2011, chap. 2, p. 270, [14]). This is different from Bringsjord's reductive view where ML produces claims that are not required to be correct, accompanies them by probabilistic statements on how large the error may be and iteratively reduces the error. In Bringsjord's view, ML does not provide the tools to determine when this error has become sufficiently small. A margin of error that is suitably small for proposing the next movie on Netflix may be too large when applied to

medical diagnostics. ML, unlike computer science, therefore seems unable to define 'general principles for data processing'. The epistemology of ML would then be confined to probabilistic statements and heuristic knowledge of which application domains would tolerate which margin of error.

The traditional explanation for this difference would be the No Free Lunch theorem which was originally identified by Wolpert (Wolpert, 1996, p. 1, [15]) and also discussed by Goodfellow (Goodfellow et al., 2016, sect. 5.2.1., p. 113, [16]) and by Mavuduru (Mavuduru, 2020, par. 10, [17]). Based on these references, I have elaborated the following illustrative example. Consider two ML algorithms A and B and a dataset of one labelled observation $(y_A, x)$. A and B take the feature variables x as input and return a predicted label. A returns the correct label, $y_A$ and B returns an incorrect label $y_B$. Now construct an artificial dataset $(y_B, x)$, i.e., the features x are the same but the correct label is now $y_B$. A will now incorrectly predict $y_A$ and B will correctly predict $y_B$. Since such an artificial dataset can always be constructed, no two algorithms A and B exist for which A is more accurate than B across all datasets. The No Free Lunch theorem also means that complex ML algorithms will never be more correct than simple algorithms such as static predictions or samplings from uniform distributions, measured across all datasets that are theoretically possible. The only way to determine that A is better than B is to determine that $(y_A, x)$ is a more relevant representation of empirical reality than $(y_B, x)$. This again can only be determined through empirical observation. The No Free Lunch theorem is therefore an example of the induction problem described by Hume (Hume cited by Kragh Sørensen et al., 2023, chap. 1, p. 12, [13]), i.e., that all theories that can be tested in a

deductive process have their origin in inductive observations of reality. This induction problem applies to all sciences, not specifically to ML. The computer scientist who determined that mergesort was faster than insertion sort may have originally been inspired by inductive observations, but based on those observations he was able to define a theory and corroborate it through a deductive process. In the same way, the No Free Lunch theorem should not be a valid reason for ML to be unable to produce formal, generalizable statements. This would mean that the epistemology of ML could evolve to more closely ressemble the epistemology of computer science.

Appendix 6.2 confirms that ML is consistent with both the theoretical and technological paradigms of computer science as defined by Kragh Sørensen (Kragh Sørensen et al., 2023, chap. 3, p. 5, [13]). It also confirms that ML is both a formal and an empirical science, something that also applies to computer science, again referring to Kragh Sørensen (Kragh Sørensen et al., 2023, chap. 1, p. 6, [13]). This leads to the question of whether ML is simply computer science or whether it can meaningfully be defined as a 'proper subset' thereof. If ML is the study of certain algorithms, these algorithms would need to be distinguishable from the algorithms generally studied in computer science. Hill defines algorithms as as a 'finite, abstract, effective, combined control structure that is given imperatively and which obtains a given purpose under given circumstances' (Hill, cited by Kragh Sørensen et al, 2023, chap 3, p. 10, [13]). ML algorithms are consistent with Hill's definition but cater to uncertainty and ambiguity in a different way. Seldin titled his lecture note 'ML: The Science of Selection under Uncertainty' (Seldin, 2023, p.1, [18]). ML is often coloquially referred to as 'a mix of

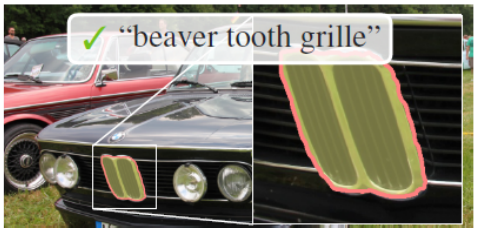


Figure 2.1: Segmentation returned at the natural language prompt 'beaver tooth grille' (Kirillov, 2023, p. 11, [19])

computer science and statistics'. Figure (2.1) shows the segmentation of the picture of a BMW car returned from the verbal prompt 'beaver tooth grille' by the Segment Anything-model of Kirillov (Kirillov, 2023, p. 11, [19]). The algorithm allows ambiguity in a way that is different from most algorithms in computer science. While the Segment Anything-model is still an algorithm in the sense of Hill, the first words that come to mind are not 'finite', 'imperative' and 'a given purpose under given circumstances'.

I therefore conclude that ML is a science and that it is a computer science subdiscipline, leading to Definition (2.2.1).

**Definition 2.2.1 (ML)** *ML is the computer science subdiscipline that concerns algorithms which infer from data while allowing uncertainty and ambiguity.*

With this definition, the Theory of Disparate Property defined in 2.1.1 is a relevant, consistent, non-tautological and non-contradictory statement as required by step two and three of Popper's deductive process.

## 2.3 Can ML become less disparate through Systematic Methods?

The idea that ML is disparate, and that reducing this disparate property will improve Explainable ML, leads to the question of how. This bachelor project will discuss a number of potential solutions under the heading of Systematic Methods which must now be formally defined.

Such a definition can begin with Helge Kragh who stated that 'Science is a common term for systematic methods to produce, organize and disseminate knowledge and skill' (Helge Kragh, 2003, par. 1, [20]). Having argued that ML is a computer science subdiscipline, and as such is a science, it is easy to make a tautological argument that ML must employ Systematic Methods. To avoid this pitfall, two additional concepts will be applied.

### 2.3.1 Discovery and justification

The first is the distinction between the scientific context of discovery and the context of justification. This distinction is described by Schickore (Schickore, 2022, sect. 5, par. 1, [21]). The context of discovery is the 'de facto thinking processes' and the context of justification is the 'de jure defense of the correctness of these thoughts'. Applying this distinction, Systematic Methods would not play much of a role in ML's context of discovery. Systematic means 'done according to a system .. in a complete, efficient or determined way' (Oxford Learner's Dictionary, year not indicated, entry for 'systematic', [22]). Discovering new 'systems' is not easily reconciled with following existing 'systems' in a complete way. If discovery requires less formalism, and if ML in recent years has seen a disproportionate amount

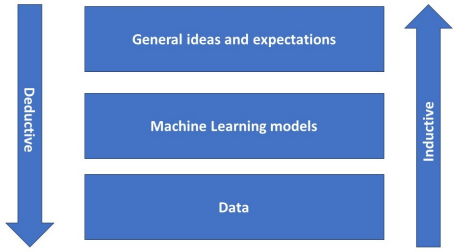


Figure 2.2: Deductive and inductive processes in ML, inspired by figure 4 in Kragh Sørensen (Kragh Sørensen et al., 2023, chap. 1., p. 11 and figure 4, [13])

of discovery, this would be a straightforward explanation for the Theory of Disparate Property.

One overall structure for scientific discovery does however exist in the form of the distinction between deductive processes, i.e., predictions and explanations, and inductive processes, i.e., inferring from observations (Kragh Sørensen et al., 2023, chap. 1., p. 11 and figure 4, [13]). Applying this distinction to ML, a deductive process is the motion from a general idea and expectation via an implementation in the form of an ML model to the test of the model on data. An inductive process is the motion that learns from data, trains a model and infers a general conclusion or idea. This bachelor project will study how ML discoveries articulate deductive and inductive processes.

When it comes to justification, Helge Kragh stated the importance of scientific claims being tested 'systematically' (Helge Kragh cited by Kragh Sørensen et al., 2023, chap. 1, p. 5 [13]). This would indicate that Systematic Methods have a different role in ML's context of justification than in ML's context of discovery. If Systematic Methods are not followed, it

becomes more difficult to reproduce and justify results. Kuhn rejected the distinction between contexts of discovery and justification, saying that even the justification of scientific claims was a social construct (Kuhn, 1995, p. 265, [23]). Kuhn's argument appears valid since the justification of ML today is expressed not least by the number of consumers who are willing to interact with ML functionality on social media or online services. Nevertheless, the use of Systematic Methods in ML's context of justification is relevant for the Theory of Disparate Property and will therefore be studied here.

### 2.3.2 Symbolic generalizations and exemplars

The second additional concept for the analysis of Systematic Methods is Kuhn's scientific matrix, most notably the concepts of symbolic generalizations and exemplars. Kuhn's scientific matrix was related to his proposition that science was a set of beliefs that a group was committed to. A scientific paradigm in Kuhn's view was structured around a scientific matrix, comprising 4 elements. Symbolic generalizations are the expressions that can be 'deployed without question or dissent from others in the group' (Kuhn, 2012, 1969 postscript, p. 182 [24]). Metaphysical paradigms are the models in which the group members believe. Values are the beliefs that are shared beyond the group. For instance, only economists share in the metaphysical paradigm of utility maximation, but both economists and non-economists can understand the value of a private business being profitable as opposed to loss-making. Exemplars are 'concrete problem-solutions that students encounter' but Kuhn later expanded the concept to innovative problem-solutions by adding 'technical problem-solutions found in periodical literature' (idem). This is today best translated to seminal ML publications such as the 4 ar-

ticles in Chapter 2.6. Kuhn's notions of symbolic generalizations and exemplars are relevant for this project. If ML is disparate, it will be observable in the use of symbolic generalizations and exemplars. If the solution is Systematic Methods, symbolic generalizations and exemplars may be the form that these Systematic Methods may take.

The symbolic generalizations of ML would include notation, words and definitions. Where Kuhn said that they should be 'deployed without question', other literature, however, emphasizes the ability to select just the appropriate notation in the given innovative context. Hadamard interviewed a number of accomplished mathematicians regarding their thought process. Most informed him that they avoided the mental use of words or algebraic or other signs. A notable few, however, including George Polya, made active use of words in their reflection process, claiming that 'a well-turned word or sentence.. enlightens the situation and gives things .. a physiognomy... The right word help us recall the mathematical idea'. Polya stated that something similar applied to notation (Hadamard, 1945, chap. 7, p. 84, [25]). Hadamard's perspective was that it was important for a researcher to be able to select the appropriate words or notation in a given context and less important to be able to deploy these 'without question'. Symbolic generalizations such as notation, words and definitions will be considered part of ML's Systematic Methods and their role in both discovery and justification will be analyzed.

The exemplars of ML would range from problem-solutions taught to students to articles that announce innovative models and their test results. The beaver tooth grille in Figure 2.1 could be considered a defining problem-

solution of ML. Kirillov demonstrates the problem (ambiguous prompts) and the solution (allow several model predictions, both literal and figurative, and select the prediction with the highest likelihood).

This leads to the following definition:

**Definition 2.3.1 (Systematic Methods)** *Systematic Methods is a set of symbolic generalizations, exemplars and systems applied in a complete, efficient or determined way in ML.*

## 2.4 Can ML become more explainable through Systematic Methods?

After Definition (2.2.1), a definition of Explainable ML requires defining the concept of explanation. As mentioned in the introduction, legislators have in recent years elevated the explanation of ML models to a right. The GDPR states that a data subject, i.e., an individual, who is subject to decisions by a data controller, e.g., a commercial company such as a bank, shall be informed of '.. the existence of automated decision-making ... and ... (receive) meaningful information about the logic involved' (EU, 2018, par. 13.2.f, [1]). The GDPR, along with the EU Digital Services Act (DSA) (EU, 2022, par. 26.1.d, [26]) and their American cousin the California Consumer Privacy Act (State of California Department of Justice, 2023, par. 2, [27]) entitle millions of individuals to explanations in a multitude of scenarios. Over time, these individuals, along with lawyers and consumer organizations will become increasingly competent in exercising this right. If technology providers are to provide all these various explanations in a high volume, this will in itself imply Systematic Methods.

Also ML researchers, e.g., Doshi-Velez (Doshi-Velez et al., 2017, p.3, [28]) have connected the disparate property of ML to insufficiently Explainable ML. It is one thing if researchers do not provide explanations because they are too deeply involved in ML's context of discovery and want to work unencumbered by formalism. If instead the reason explanations are not being provided is that researchers themselves do not understand the model well enough to express it formally, this lack of formal accuracy reflects insufficient actual understanding. This insufficiency will lead to inability to perform the Systematic Method of enumerating all use cases which again will lead to inability to identify potentially unsafe use cases. Insufficiently Explainable ML in this view becomes a canary in the coal mine for unsafe ML. While Explainable ML has thus been connected to Systematic Methods, this bachelor project cannot define Explainable ML in those terms since that would render the Theory of Impact tautological: 'Systematic Methods equal explanation so more Systematic Methods will benefit explanation'.

A consistent definition of Explainable ML must therefore not just require explanations to be systematic, but delve more deeply into the nature of such explanations. The deductive-nomological model of scientific explanation as described by Woodward (Woodward et al., 2022, sec. 2, [29]) states that an explanation must comprise the phenomenon to be explained ('explanandum') and a series of statements that account for the phenomenon ('explanans'). Explanandum must follow logically from explanans. Each statement in explanans must be true. At least one statement in explanans must be a law of nature, hence the term nomological. This statement must be essential in the explanation, i.e., if it is removed explanandum no longer follows deduc-

tively from explanans. If the deductive-nomological model relies on 'laws of nature', the next question would be which such laws could be relevant for Explainable ML. Section 2.2 discussed a view of ML as producing 'probabilistic statements and heuristic knowledge' rather than general statements. If confirmed, this would be an example of a relationship between ML's scientific identity and the fact that ML models are often considered difficult to explain. Other models of scientific explanation must therefore be considered.

Woodward (Woodward et al., 2021, sect. 3, 4.4. and 5.1, [29]) and Kragh Sørensen (Kragh Sørensen et al., 2023, chap. 6b, p. 13, [13]) also describe a number of alternatives to the deductive-nomological model. Statistical relevance-explanations are explanations such as the following statement: when an individual has the property A it is often because they also have the property B. This can be seen from the probability of A given B being much higher than the probability of A $P(A|B) >> P(A)$. Statistical-relevance explanations are well suited for ML because they allow uncertainty, cf. Definition (2.2.1). Consider, however, a more complex conditional probability where A is explained not by a single factor but by a complex correlation $P(A|X_1 \cap ... \cap X_n) >> P(A)$. As n, the numbers of factors involved in this correlation, increases, the explanation becomes less understandable to humans. In order to understand conditional probabilities involving a large number of factors, humans rely on models, for instance ML models. ML models have then become their own explanation. Therefore, statistical relevance explanations will not be applied as a definition of Explainable ML.

Something similar applies to the unificationist account of explanation. Kitcher cited by Woodward wanted to 'derive descriptions of many phenom-

ena, using the same pattern of derivation again and again ... reducing the number of facts we have to accept as ultimate' (Woodward et al., 2021, sect. 5.1. par. 5, [29]). Processing a large number of phenomena (observations) with the same pattern of derivation (algorithm) to arrive at a sparser set of facts (predictions) is ML. Applying the unificationist approach would again make ML models their own explanation. Therefore, the unificationist approach will not be applied as a definition of Explainable ML.

A separate category called outcome-based explanation models is discussed in Appendix 6.3 demonstrating that the value proposition explanation model that is very popular within business economics also cannot be applied.

I therefore decide to apply the first explanation model, the deductive-nomological model. This is done with the expectation that this project will identify ML statements that go beyond probabilistic statements and heuristic knowledge and which can serve as part of explanans. For now, having defined ML as a computer science subdiscipline in Definintion (2.2.1), explanations must refer to general principles of computer science. According to Søgaard, such model explanations can be either global, i.e., concerning properties such as bias or distribution of predictions, and local, i.e., explaining specific decisions (Søgaard, 2022, chap. 2, p. 13, [2]). This leads to Definition (2.4.1).

**Definition 2.4.1 (Explainable ML)** *Explanations that are consistent with the deductive-nomological model are a property of ML models. All global model properties as well as all local model decisions must follow deductively from a set of statements that are declared explicitly along with the model. These statements must contain references to general principles of computer*

*science.*

With Definitions (2.3.1) and (2.4.1), the Theory of Impact in Definition (2.1.2) realizes the second and third step of Popper-Thornton's deductive process.

## 2.5 Innovation

The fourth step of Popper-Thornton's deductive process is to ensure that the Theories do in fact represent innovation, Popper would say a 'scientific advance' (Popper, 2002, Part 1., chap. 1, sect. 3, p. 31, [7]). As mentioned in the introduction, most literature considers explainability a property of ML models or processes. The novelty of this project would then be the discussion of how ML's scientific identity relates to the fact that models are often considered difficult to explain. It was also mentioned in the introduction that Hu (Hu et al, 2022, p. 2, [4]) are a key reference for the observation that ML currently represents '... a multitude of learning paradigms and methodologies ... disparate, narrowly focused methods...make...development difficult'. Hu states that a unified way of thinking in ML would be 'incredibly valuable'. They refer to the 'Theory of Everything' of physics which is defined as a 'the end goal of physics ... that fully explains and links together all physical aspects' (Hu et al., 2022, p. 50, [4]). The term originated from Weinberg who defined it as '...principles .. that cannot be explained in terms of deeper principles' (Weinberg quoted from (Horgan, 2015, par. 2, [30]).

Doshi-Velez (Doshi-Velez et al., 2017, p.3, [28]) recognize an 'incompleteness in the problem formalization' and link to Explainable ML. Goldberg

(Goldberg, 2015, p. 54, [31]) recognizes 'a number of challenges...inferring the exact model form from reading its description in a research paper can be quite challenging. Many aspects of the models are not yet standardized, and different researchers use the same terms to refer to slightly different things [hiding] behind ambiguous figures or phrasing'.

Of the mentioned articles, only Hu (Hu et al, 2022, p. 2, [4]) have Systematic Methods as the main topic, Goldberg (Goldberg, 2015, p. 54, [31]) devotes one subsection and Doshi-Velez (Doshi-Velez et al., 2017, p.3, [28]) has a passing reference to the topic. From a quantitative perspective, it can therefore be argued that this bachelor project improves upon existing knowledge. Furthermore, this bachelor project attempts to specify the notion of Systematic Methods in ML beyond the three works cited here. For example, Definition (4.4.1) adapts Hu's Standard Model of ML to an even more general and transparent notation.

To the extent that ML is today insufficiently Explainable, few sources other than Doshi-Velez (Doshi-Velez et al., 2017, p.3, [28]) have discussed the relationship to Systematic Methods.

The Theory of Disparate Property and the Theory of Impact are therefore innovative.

## 2.6 Empirical application

The fifth step in Popper-Thornton's process is empirical testing of the 'conclusions derived ... and their empirical application' (Thornton, 2022, par. 3-6, [6]). This empirical application is the 'practical consequences' of ideas

that is key to the pragmatic research philosophy.

The empirical application of the Theory of Disparate Property would be that a case study of ML would not encounter many symbolic generalizations, examplars and systems or that these, when encountered, were not applied in a complete, efficient or determined way' as per Definition (2.3.1).

The empirical application of the Theory of Impact would be that the occurence of Systematic Methods would be positively correlated with Explainable ML such as defined in Definition (2.4.1), all other factors being equal. Due to limitations, these other factors are not identified here. Several possible scenarios could be consistent with the Theory of Impact, i.e., 1) that Systematic Methods were absent while ML was at the same time considered to be insufficiently Explainable, 2) that Systematic Methods were present while ML was at the same time recognized as being fully Explainable and 3) from a comparative static perspective that efforts to improve Systematic Methods would be accompanied by progress with regard to Explainable ML.

# Chapter 3

# Case study of one curriculum and 4 seminal articles

The fifth and final step in the Popper-Thornton deductive process, empirical application of the Theories, is now performed through case studies. These case studies will show the extent to which ML is disparate. Identifying concrete examples will also further a discussion of how this disparate property might be reduced and how that might impact Explainable ML. As the objects of the case study is selected one curriculum and four seminal articles.

## 3.1 The ACM data science task force

Since the 1960s, The Association for Computing Machinery (ACM) has published curricula recommendations that inform university programmes globally. Between 2017 and 2021, an ACM task force undertook to define the organization's first curriculum for data science. ACM defined ML as one of 11 knowledge areas of data science. By virtue of the ACM's role, this publication (ACM, 2021, p.1, [10] is an obvious candidate for a set of

Systematic Methods in ML. The deliberations of the task force also more generally reflect data science's and ML's scientific identity.

The task force emphasizes that data science is multidisciplinary, involving computer science, statistics and mathematics. The task force also celebrates the role of application domains, i.e., industries that generate data for analysis and which apply the resulting ML functionality.

The task force does not define a clear hierachy among these disciplines. Discrete mathematics is already considered the most important auxiliary discipline of computer science, so when the task force enumerates mathematics along with computer science the structure is not clear. The focus on the multidisciplinary identify of data science, and by extension ML, may have a political explanation. The task force notes that 'departments are being encouraged by administration (with pressure from trustees, in some cases) to start Data Science programs [but] ... may already be facing staffing challenges'. The task force opted for a 'big tent', almost a Danish approach to involving all relevant stakeholders, perhaps at the expense of assigning clear ownership.

Metaphorically speaking, the task force commits the ML sin of defining too rich a hypothesis space leading to the risk of overfitting data. It is a well-recognized result in ML that models that have too many parameters can be trained to predict the observable reality very well. Models that are only allowed to have a restricted range of parameters will not reflect observable reality as well, but their sparsity may force them to learn the salient features of data and they will therefore be more applicable to future scenarios. Apply this logic to an academic curriculum. An eclectic approach will often provide

a relevant solution or explanation to a diverse range of scenarios. A more hierachical approach will take longer, be more limited but also force the researcher to define the problem in scientific terms.

The task force nods to the disparate property of ML by stating that the scope of ML must be 'a set of principled algorithms .. rather than a "bag of tricks"'. ACM enumerates four subdomains: supervised learning, unsupervised learning, mixed methods and deep learning. This categorization seems at best very preliminary. LeCun has argued that the supervised learning is of limited relevance due to the relative unavailability of labelled data and that unsupervised learning should be renamed self-supervised learning, representing methods where models are trained on information and structure that is inherent in the data rather than have the form of separate labels (LeCun, 2021, Sect. 2, [32]). Mixed methods is a residual category and deep learning is then an advanced category of the concepts already introduced. The four subdomains are overlapping, e.g., supervised learning in a deep neural network falls in more than one category. The task force reflects the ideas of Bringsjord cited above, where the importance of application domains is celebrated at the expense of defining a strong scientific identity for ML proper. The work of the task force is therefore not a Systematic Method for ML. Given the eminent role of the ACM, the broad scope and extensive time allocated, one conclusion could be that it is not straightforward to formulate more general principles for ML. Therefore, the case study corroborates the Theory of Disparate Property. The case study does not lend itself to a study of the Theory of Impact.

## 3.2 Vaswani et al. Attention is all you need (2017)

According to Kuhn, publishing innovative articles is in itself an exemplar (Kuhn, 2012, 1969 postscript, p. 182 [24]). This case study will consider four articles that were all part of the course of Advanced Deep Learning taught by Sommer (Sommer et al., 2023, p. 1, [33]) and which have each been important in the ML revolution of recent years. The articles are key evidence of how ML discoveries are made and justified. Although the results of case studies cannot be generalized, the articles have played such a role in ML that any findings will at least justify further research. If Systematic Methods can benefit Explainable ML, the articles are a relevant place to start. The analysis is structured around how the articles employ symbolic generalizations, how they can be interpreted as exemplars and how they articulate deductive versus inductive processes .

### 3.2.1 Symbolic generalizations

Vaswani (Vaswani et al., 2017, p. 1, [34]) is an example of an article that presents several new terms and appropriates existing terms. The term transformer is the name of the overall model category. Its novelty depends on a specific algorithm called an 'attention mechanism'. 'Attention head' is derived from 'attention'. The article also appropriates the well-know computer science terms 'query', 'key' and 'value' and assigns them new meaning. This novel terminology may have been selected because these very words help reflection the best, as suggested by Hadamard, or it may reflect lack of interest in Systematic Methods. In order to translate a word, the transformer

model correlates it with all other words in the text, i.e., a global approach. The transformer model analyzes the word's position, i.e., a local approach, but does so as a secondary measure. A logical name for the transformer model could therefore be global-before-local model, a term that will be used in this project. At the time of its publication, the global-before-local model was innovative relative to traditional models that translated each word as a function of the word and a state vector incorporating all previous words and which used attention as a secondary measure. These models could be called local-before-global models, a term that will be used in this project. The attention mechanism itself is an algorithm that calculates vector products and their relative weights. Attention mechanism could therefore be called vector-to-vector correlation algorithm, a term that will be used in this product. When Vaswani uses the terminology of 'transformer' or 'attention' or when they misappropriate the 'query-key-value' terms, they may have selected the right terminology in the context of discovery, but in so doing they created symbolic generalizations for ML that lacked clarity and consistency.

### 3.2.2 Deductive and inductive processes

Vaswani's article represents a deductive process as described in Figure (2.2). Vaswani expects a global-before-local model to be more accurate than previous local-before-global-models and to not have higher computational complexity. This expectation is implemented in a model. Some model parameters are invariable throughout the experiment, e.g., the neural network should be rectangular, with the width of all layers matching the 512-element length of the input vectors.

The article also represents an inductive process, where the model weights are learned from data in a neural network forward-backward-propagation algorithm. A number of supporting mathematical functions are subject to more unstructured and undocumented experimentation. For instance, language sentence structure, what Vaswani et. al. calls positional encoding (Vaswani et al., 2017, p. 5, [34]) is represented by a sine function but could alternatively have been represented by a modulus function. This choice is made in a way that can not be reproduced. In that sense, both model weigths and supporting mathematical functions such as sentence structure are learned inductively from data, but only the learning process for the model weights is clearly articulated. The choice of supporting mathematical functions has as much impact on the model predictions as the model weights and therefore both should have been subject to a structured and documented inductive learning process. Vaswani's article is an example of an unclear articulation of deductive and inductive processes.

## 3.3 Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale (2021)

### 3.3.1 Symbolic generalizations

Dosovitskiy (Dosovitskiy, 2021, p.1, [35]) is one of several ML articles that assign a meaning to the term 'inductive bias' that is so specific that it tends to be misleading. In general, inductive bias, as stated by Mitchell, cited by Henderson, means 'an a priori assumption about the domain the algorithm

is employed upon' (Henderson, 2022, sec. 3.4., par. 2, [36]). In ML image analysis, the term relates to the type of models that were used before Dosovitskiy. These models will again be called local-before-global models instead of their original name, convolutional models. The local-before-global models divided images into patches that were analyzed separately. When detecting part of a segment, e.g, a sleave, the search for the other sleave began with the neighboring patches. If a second sleave was identified, the image was categorized as containing a jacket. Local-before-global models were able to identify objects irrespective of global position, i.e., a jacket would be classified as such whether it was at the bottom or the top of the suitcase. ML literature has no less than four different names for this property: spatial invariance, translation invariance, translational equivariance and inductive bias. The term inductive bias is applied because a local-before-global model already assumes that local search is most efficient and therefore does not have to learn this approach first. It is a very general term applied to a very specific context and it makes a virtue out of necessity since the local search strategy was originally due to computational limitations.

Along comes Dosovitskiy who applies Vaswani's global-before-local approach to image analysis. The global-before-local model has no a priori assumption of whether local or global search is better, so it correlates all patches to all other patches. In the first layers of its neural network, the model then learns that the vector products of patches close to one another (local) should have approximately equal weight to vector products of patches far from one another (global), but that in the deeper layers, global relationships should have the highest weights. These deep, global relationships allow

the model to be more accurate than local-before-global models as long as training data sets are sufficiently large. Dosovitskiy jubilantly writes 'large-scale training trumps inductive bias' (Dosovitskiy et al., 2021, p.2, [35]). That statement translates to 'as long as we have enough data we do not need to know anything beforehand'. As pointed out by Kragh Sørensen, 'data devoid of theory' does not exist (Kragh Sørensen et al., 2023, chap. 6b, p. 3, [13]). Employing a global-before-local model is also an expression of inductive bias, i.e., the knowledge that this model will produce a relevant result. Dosovitskiy, and more broadly the use of the term inductive bias in image analysis, are therefore an example of how symbolic generalizations are applied in ways that lack clarity and consistency.

## 3.4 Kirillov et al. Segment anything (2023)

### 3.4.1 Exemplars

It is discussed above how Kirillov (Kirillov et al., 2023, p. 1, [19]) present a new problem-solution in the form of handling ambiguous prompts in a new way. The article also reflects another problem-solution, i.e., if relevant data is scarce then such data can be generated as part of the project.

### 3.4.2 Deductive and inductive processes

Kirillov's bold conjecture is to take an existing innovation, the global-before-local model of Vaswani, apply it to a new purpose, image segmentation instead of image classification, and expect to obtain good accuracy and com-

putational efficiency. In that sense, Kirillov's scientific project represents the deductive process. However, the body of general ideas and expectations that Kirillov applies is very loosely defined ('we build on transformer models ... In general, the image encoder can be any network') (Kirillov, 2023, p. 5 and 16, [19]). Kirillov therefore does not present a clear, falsifiable scientific theory. Kirillov's project clearly is also an inductive one, learning model parameters from data. In order to build Facebook's Segment Anything-model Kirillov realizes that labelled data does not exist at sufficient scale. His team therefore builds a large test set (1 million images with 1.1 billion identified segments). This is done by first using human annotators and then by having the ML model learn how these annotators worked and apply that learning to segmenting additional images. This generates a data set that the final model can learn from, but this dataset is itself the product of a highly managed process. Kirillov's article is a good illustration that this use of ML as part of the development of new ML requires very clear distinction between deductive and inductive processes as part of Systematic Methods.

## 3.5 Brown et al. Language models are few-shot learners (2020).

### 3.5.1 Exemplars

The article by Brown introduces in-context learning (Brown et al., 2020, p. 1, [37]). ML models generally learn by updating a specific data structure, i.e., the weights of the neural network. Brown's innovation is the ability to improve model performance without updating these weights. Brown commu-

nicates this result without providing a scientific explanation involving general principles of computer science. One such general principle is that programs can be illustrated by flow diagrams of data structures and how they are related. Such a flow diagram would have shown that Brown's model involves two data structures, the model weights and the sequence of prompts. The model does not update the weights but still 'learns' because it reads the sequence of prompts auto-regressively. This is what enables a pre-trained language model to perform well on the sequence of prompts 'Oslo is the capital of Norway. Stockholm is the capital of Sweden. Copenhagen is the capital of '___'. The model weights are the data structure that has already stored the general geographic knowledge. The input sequence is the data structure that stores the information from which the model can infer that it should continue a sequence of capital - country pairs. If explaining a complicated model in terms of its data structures and algorithms is an exemplar of computer science, Brown's article is an illustration of how this exemplar could have improved explainability.

In-context learning is also more widely a novel problem-solution. A model is pretrained on a very large general dataset and then learns a more specific task from the history of the queries that are input by the user. This property first arises when the general training dataset and the number of parameters become very large or as Brown says 'larger models are more proficient at in-context learning' (Brown et al., 2020, p.2, [37]). Neither Brown nor ML literature in general have so far been unable to reduce this property to simpler principles, for instance by demonstrating a simple example of in-context learning on a small model. Emergent abilities such as in-context learning

have been cause for concern as expressed by Bengio (Bengio et al., 2023, par. 1, [38]) and Sommer (Sommer, 2023, par. 6, [39]). Bengio even called for a 6 months pause to giant AI experiments. Emergent abilities therefore underline the need for Explainable ML Systematic Methods must be defined to help analyze these emergent abilities by reducing complex models to simpler principles.

## 3.6 Summary

Each of the four articles constitute exemplars, i.e., novel problem-solutions within ML. The articles do not apply symbolic generalizations that can be met without dissent and they do not clearly articulate the distinction between deductive and inductive processes. The four articles therefore corroborate the Theory of Disparate Property.

The four articles are consistent with the Theory of Impact because they do not employ Systematic Methods, at least not in the form defined here, while at the same time the models they present such as large language models are recognized as being difficult to explain.

The four articles' context of justification is often comparative, i.e., model accuracy is discussed not just in absolute terms but in comparison to 'state-of-the-art models' (Vaswani et al., 2017, p.1, [34]). The state-of-the-art models are in turn selected based on their accuray but also based on their level of acceptance. The articles therefore seem to corroborate Kuhn's view that the context of justification is a social construct.

# Chapter 4

# Proposing Systematic Methods for ML

## 4.1 Why propose Systematic Methods?

As defined in Chapter 1, this bachelor project adopts a research philosophy of pragmatism, concerned with understanding ideas through their practical consequences, perhaps even the progress they produce. Systematic Methods are an idea the consequences of which can be better analyzed if it takes the form of a concrete proposal. The curriculum ('Studieordning') for the BSc in ML and Data Science at DIKU requires students to 'solve and reflect upon the solution of a problem of relevance for ML and Data Science' and to 'describe a ..specific solution emphasizing own contributions' (Det Natur- og Biovidenskabelige Fakultet, Københavns Universitet, 2019, p. 37, [40]). Therefore, a set of Systematic Methods will be proposed in Section 4.5. To ensure that this proposal is sufficiently specific, it is further illustrated by a mock scientific article drafted accordingly ('Vector products are all you need' Appendix 6.10).

While ML may be disparate, it does contain several elements that each in their own way may form part of Systematic Methods. The discussion here will focus on continual development models and a Standard Model of ML. Other elements such as the data science pipeline, MLOps, the Universal Approximation theorem, the use of graph models, ablation and Occam's Razor are discussed in Appendix 6.4, 6.5, 6.6, 6.7, 6.8 and 6.9.

## 4.2 Continual development models

Definition (2.3.1) refers to Systematic Methods as being applied in a 'complete, efficient or determined way'. This naturally brings up the potential role of automation. Gesmundo describes how stages of the ML process can be automated and repeated, allowing for the systematic testing of a large number of model components (Gesmundo, 2022, p. 1, [41]). Where Section 3.2.2 described how Vaswani followed a heuristic process to find the right mathematical function to represent sentence structure, Gesmundo's approach means first defining a search space for such functions, then sampling that search space and testing the result. This process, called 'extended hyper-parameter search', ressembles the general ML process of iteratively learning model weights. Automation allows this search to cover a larger space of possible hyper-parameter values. Gesmundo demonstrates that this larger space increases the probability of chosing the optimal combination of design and hyper-parameters and results in higher model accuracy without higher computational complexity. Systematic Methods in ML can therefore include the automated optimization of not just model weights but 'hyper-parameters'

such as supporting mathematical functions or model design.

Gesmundo cites the discussion of Artificial General Intelligence (Gesmundo, 2022, p. 2 and 17, [41]). If Artificial Intelligence can perform an ever larger number of tasks, why should it not also be able to initiate the development of new ML models while systematically testing and learning from the largest possible number of configurations and datasets. The reference is obvious to Anderson, who as cited in Kragh Sørensen spoke of 'The End of Theory', where machines would learn from data, inductively iterating over all possible interpretations and never explicitly formulating any testable theories (Kragh Sørensen et al., 2023, chap. 6b, p. 4, [13]). To avoid this scenario, there must be a clear distinction between the variables that will be learned from data, also called the endogeneous variables of the model, and the variables that will not change during the process, also called the exogeneous variables. This clear distinction will be part of Systematic Methods.

## 4.3 Hu et al.'s Standard Model of ML

As mentioned in Chapter 1, it was the article by Hu which originally inspired this bachelor project (Hu et al., 2022, p. 1, [4]). Hu proposes 'a general formulation for learning a target model via a constrained loss minimization program' (Hu et al., 2022, sect. 3, p. 13, [4]). This involves a minimization problem that is specificed in the following way (Hu et al., 2022, p. 13, equation (3.1), [4])

$$\min_{q,\theta,\xi} \ell = -\alpha H(q(\theta)) + \beta d(p, q(\theta)) + U(\xi) \tag{4.1}$$

subject to a set of secondary conditions

$$\text{s.t.} \quad -E[f_k^{(\theta)}] \leq \xi_k, \, \forall k \tag{4.2}$$

$\ell$ is the loss function to be minimized. $\alpha, \beta > 0$ are relative weights. $H(q)$ expresses entropy. Stone defines how in information theory, entropy, also called surprise or information value, is seen as an attractive model quality in prediction (Stone, 2015, p. 31, [42]). A surprising item of information makes the audience better informed than an unsurprising item. Entering entropy with a negative sign means that if two models have the same error d, the model which makes the most surprising predictions is preferred. It is more interesting to study the model's accuracy when it predicts a blue pixel to come after 1000 red pixels or when an attention-based model suggests a non-literal translation, rather than when it predicts the next blue pixel in an already long sequence or when it predicts a literal translation. p is the target function that the model is trying to learn, and q is the prediction function which depends on parameters $\theta$. d is the error or distance between prediction and observed target. d can accomodate both supervised learning, where p is explicit, and self-supervised learning where p is a function that takes part of the data as input in order to predict other parts of the data, i.e., predicting ('Copenhagen is the capital of ___'). The third term $U(\xi)$ is related to the secondary conditions. The more lax the secondary conditions are, expressed by a higher value of $\xi$, the higher the value of the loss function. For two models with the same error, the model with the tighter secondary conditions is preferred. The secondary conditions can express context that the model developer for any reason does not expect the model to learn from the available

data. Imagine a model of upselling of cruise passengers on a short cruise where no passenger has time to exceed his credit card limit and therefore the credit card limit cannot be learned from data. The model developer can then wish to still include this context in the model as a secondary condition. Secondary conditions can also be ethical or legal. The specification of 4.1 means that the secondary conditions are soft. Between two iterations, the model developer can relax a specific secondary condition by increasing $\xi_k$ and then paying a penalty since the $U(\xi)$-term of the loss function will increase.

ML in the model occurs by randomly sampling an initial state of parameter values $\theta$ within the secondary condition in 4.2, calculating the value of the loss function 4.1 and then iteratively minimizing that value through a mathematical process such as gradient descent, constantly ensuring compliance with 4.2. Iterations will stop when a local minimum is attained but, as per the discussion in Section 2.2, there is no requirement to attain a global minimum.

Standard libraries such as Pytorch or TensorFlow do not easily support this type of loss function with secondary condition. A search in the Pytorch documentation for 'secondary condition' returned zero hits and a search in the TensorFlow documentation returned no relevant hits.

Hu concludes that the model in 4.1 has the advantage of representing a 'vast algorithmic space governed by a few components' such as entropy, error and secondary conditions (Hu et al., 2022, p. 49, [4]).

## 4.4 Amending Hu et al.'s Standard Model of ML

I now respecify Hu's model to become even more general, somewhat challenging its claim to be a 'general formulation'. While entropy is a model desideratum, ML literature comprises several other such desiderata. Sparsity, i.e., preferring a simple model to a complex model with the same accuracy, follows from the original Occam's Razor principle also discussed in Appendix 6.9. Goodfellow introduces smoothness as a model desideratum, i.e., for a given error term, a model that avoids large updates to its weights in each single iteration will be preferred (Goodfellow et al., chap. 20, p. 645, [16]). I therefore replace the entropy term $-\alpha H(q(\theta))$ in (4.1) with a more general term, $MD(\theta, q(\theta))$, reflecting all model desiderata such as entropy, sparsity or smoothness. Due to this general specification, the minus becomes redundant.

Goodfellow also applies secondary conditions to express model desiderata, e.g., forcing the gradient to have a norm less than a given size (Goodfellow et al, 2016, sect. 14.2.3, p. 499, [16]). For clarity, I will not apply the secondary conditions this way.

Optimization under secondary conditions, the way it is practised in, e.g., economics, leaves secondary conditions constant, so Hu's model is updated accordingly. The context that Hu expresses in the secondary conditions is then divided into soft context and hard context. Soft context may be the above example of known factors such as credit card limits that are not observed in data. In contrast to Hu's original model, the soft context now enters directly into the loss function rather than as a secondary condition.

The model desiderata and soft context are combined to the residual term $MDSC(q(\theta), \theta)$. Hard context may be ethical or legal restrictions, such as preventing predictions that do reflect training data but which should be excluded from the range of possible model predictions, e.g., a self-driving car can see speeding violations in the training data while being prevented from speeding via a secondary condition. The hard context is now the only factor in the secondary conditions $HC(q(\theta), \theta)$. The $U(\xi)$-term in equation (4.1) is eliminated.

A Standard Model of ML can now be defined this way

**Definition 4.4.1 (Standard Model of ML)** *A Standard Model of ML is a model that minimizes the loss function*

$$\min_{q,\theta} \ell = \alpha d(p, q(\theta)) + \beta MDSC(q(\theta), \theta) \tag{4.3}$$

*subject to a set of secondary conditions*

$$s.t.\ \theta \leq HC(q(\theta), \theta) \tag{4.4}$$

*where $\alpha, \beta$, $p$, $q$, $\theta$ and $d$ are defined as above, MDSC is a residual term representing model desiderata and soft context, and HC is hard context. MDSC and HC are functions of model parameters $\theta$ and of the prediction function $q$.*

Hu's Standard Model of ML, adapted as in Definition (4.4.1), distinguishing between error, model desiderata, soft context and hard context relating to ethical or legal requirements, is a relevant component of the Systematic Methods.

## 4.5 Enjoy your free lunch: Proposed elements of Systematic Methods

Based on the above case study and inspirations, the following is proposed

### 4.5.1 Apply the Standard Model to defining general statements

It was discussed in Section 2.2 that, although the No Free Lunch theorem is often cited as the reason for the relatively few general statements in ML, this theorem, in the form of Hume's induction problem, is a general statement on all sciences and as such should not be a restriction.

The specification of the Standard Model of ML in Definition (4.4.1) will be helpful in the formulation of such general statements. The model will make it model desiderata and their relative importance more transparent. It will drive a discussion of whether sparsity can be a general desideratum, i.e., a complex model with an accuracy of 0.99 will never be preferred to a simple model with the same accuracy. Entropy can be another such potential general principle. Consider surprising predictions. Consider model A that has an accuracy of 0.8 on surprising predictions and 0.9 on unsurprising predictions, and model B that has the inverse accuracy ratio. Assume A and B have the same overall accuracy. A conjecture could be that A will never be preferred to B.

The hard secondary condition in Definition (4.4.1) is inspired by similar conditions in optimization problems of economics. These secondary conditions have strong intuitive interpretation. If professors are paid five times more than teaching assistants (TAs), any cost-minimizing computer science department will organize itself so that professors are five times more pro-

ductive than TAs. Similar intuition should be sought for the secondary conditions in ML models. The secondary condition reflects ethical and legal restrictions. It will be interesting to see which intution can be derived regarding the trade-off between accuracy, ethics and law.

Some ML models will be 'better', i.e., more computationally efficient, more transparent, less biased or more compliant than others, the same way that the mergesort algorithm mentioned above is 'better' than the insertion sort algorithm. That ML will be able to make such general statements is a free lunch waiting to be enjoyed.

### 4.5.2 Define a clear hierachy of sciences

ML will always be multidisplinary and rely on data and context from the various application domains. To avoid what could metaphorically be called 'overfitting', it is all the more important that a clear hierachy be defined. Comte defined a succession of sciences (Comte described by Bourdeau, 2022, Par. 1, [43]), but here the term will have a more specific meaning. Solutions and explanations should first be sought within computer science, secondly within statistics, thirdly within mathematical disciplines other than those that are already auxiliary disciplines of mathematics and statistics, and finally within the application domain. Each progression on the hierachical ladder should be motivated and documented. This sparsity may be restrictive but should ideally drive a clearer identification of the salient features of the relevant problem and dataset. At each step of the ladder, researchers will be invited to realize that they cannot explain the whole problem referring only to, e.g., computer science principles. They will then define the dimensions of the problem that can in fact be explained within computer science,

before progressing to statistics. Imposing a well-defined hierachy of sciences reflects the general ML discussion of the advantages of sparsity.

In the example in Figure (2.1), a research strategy could be a) it is attempted to define the segment anything model as an algorithm, b) since computer science does not have strong representation of uncertainty, and since prompts are ambiguous, it is decided to allow the algorithm to output more than one potential solution and a statistical model is then introduced in order to select among these potential solutions. Such a hierachy would be more cumbersome but bring to the fore the novelty of the algorithm, i.e., its ability to handle uncertainty and ambiguity.

### 4.5.3 Consider problem-solution-explanation an ML exemplar

Where Kuhn spoke of problem-solutions that defined a science, the focus on Explainable ML means that this idea should be expanded to defining problem-solution-explanations. An ML problem-solution should not be 'justified' without an explanation. The explanation should be provided according to the deductive-nomological model as per Definition (2.4.1) and involve both global model properties, such as bias or distribution of predictions, and local properties such as specific decisions.

Section 4.5.2 proposed the definition of a hierachy. This means that Definition (2.4.1) can be updated.

**Definition 4.5.1 (Explainable ML)** *Explanations that are consistent with the deductive-nomological model are a property of ML models. All global model properties as well as all local model decisions must follow deductively*

*from a set of statements that are declared explicitly along with the model. These statements must contain references to general principles from a well-defined hierachy of computer science, statistics, novel elements of mathematics and application domains.*

### 4.5.4 Define intuitive exemplars for ML beginners

The use of the data science pipeline as a framework for teaching or performing ML, described in Appendix 6.4 should be discontinued and MLOps discussed in Appendix 6.5 , is not a good replacement. As discussed in Appendix 6.7, neural networks should be learned first as computer graph objects, then as graphical models, then as probabilistic graphical models and as a final step as graph models where the functions are learned inductively. The individual layer of a neural network should be interpreted as a line of computer code or alternatively as a mathematical function that takes input from a previous function and returns output to a subsequent function. This will be one of several areas where concepts from functional programming should be used more actively, bringing the code closer to the underlying mathematical functions. The Standard Model should be implemented in popular libraries such as Pytorch or TensorFlow.

### 4.5.5 Create clearer and more consistent symbolic generalizations

The case studies have demonstrated several examples of words and notations that are related to specific contexts of discovery and contribute to the disparate property of ML. These could be more general and transparent

and be derived from the above-mentioned hierachy of sciences. Appropriating existing terms should be avoided. When defining new terminology, the agreed 'Definition, result, example' structure from mathematics should be followed. Above, proposals for specific terminology has been put forward regarding global-before-local models, local-before global models, vector-to-vector correlation algorithms and the discontinuation of the use 'inductive bias' in image analysis models. Discussion of model performance should be articulated into accuracy and computational efficiency. The term 'hypothesis space' should be replaced by 'endogeneous variable search space'. The term Occam's Razor, discussed in Appendix 6.9, should be replaced by 'endogeneous variable search strategy' to avoid any confusion with sparsity as a model desideratum. The separation into supervised and unsupervised learning should be replaced by self-supervised learning as a general category.

Appendix 6.10 shows a mock scientific article 'Vector products are all you need' drafted in accordance with this proposal

### 4.5.6 Clearly articulate deductive and inductive processes

Models should be defined as Standard Models as per Definition (4.4.1). Expectations should be defined in terms of model accuracy and computational complexity. All variables and design choices that will not be modified in the course of the experiment should be declared as exogeneous variables. The MDSC-term in the Standard Model of ML will make transparent all design choices that are not directly related to error minimization. All variables that will be modified in the course of the experiment should be declared as as

endogeneous variables. This will support safe automation. The distinction between exogeneous and endogeneous variables will allow the discontinuation of the distinction between parameters, hyper-parameters and model design. Ethical and legal requirements should be specified as hard context as per Definition (4.4.1). Ablation, discussed in Appendix 6.8 will support comparability and quantification of gains in accuracy and computational efficiency.

# Chapter 5

# Conclusion

This project has illustrated a clear relationship between ML's scientific identity and the fact that ML models are often considered difficult to explain.

Popper contended that contradiction produced more knowledge than corroboration. This project has not contradicted the Theory of Disparate Property. The ACM curriculum emphasized multidisciplinarity but was not able to provide a consistent operational formula. The four seminal articles are very different, apply different symbolic generalizations and do not clearly articulate their scientific process in terms of deduction or induction. In the introduction I suggested that this might be a question of scientific maturity, but I have not observed any work outside of the ACM on general frameworks. I also allowed for the possibility that such comprehensive structures might exist without being immediately visible to an ML beginniner. Few such latent or tacit structures have become visible during the project. One driver for future standardization may be the legal right to an explanation awarded to a wide public audience.

The project has also not contradicted the Theory of Impact. It has identified a number of areas where ML could attempt to produce general statements that could form the basis of better scientific explanations. Hu's Standard Model has been amended to an even more general form which can support a structured discussion of accuracy, model desiderata, context and ethical and legal restraints.

The project may have contradicted one theory that was not formulated explicitly at the outset, i.e., that defining Systematic Methods for ML would be realistic, let alone easy.

Popper's process is iterative, and the proposals in this project allow for such iterations. For example, future scientific articles could be elaborated in two versions, the original and a re-draft according to the proposed Systematic Methods in (4.5). The two drafts could be read by two groups of students or other relevant audience, and it could be measured which version was seen as easiest to understand. Such a study would, unlike the present case study, produce generalizable conclusions.

Inviting researchers to communicate differently, or inviting technology companies to declare their model use differently, will require time and money. These resources could alternatively be applied towards developing new ML functionality. The project did not identify any social mechanism to introduce Systematic Methods, nor did it quantify the cost. It is possible that this cost can only be justified in the context of the suggested pause to large model development. In the unlikely event that a global pause is agreed, the time that becomes available can be used for implementing the above proposals.

Where this bachelor project has covered two of the 4 elements of Kuhn's

scientific matrix, symbolic generalizations and exemplars, a next iteration would focus on the role of the third element, i.e., values. Systematic Methods will only promote Explainable ML if they reflect the values in the ML community. This bachelor project has touched upon values such as innovation, transparency and safety. These values cannot be explained in terms of deeper principles, they constitute, in other words, a Theory of Everything.

# Chapter 6

# Appendices

A number of appendices are included in order to control the overall page count and pursue strands that are orthogonal to the two Theories. The appendices are Appendix 6.1 on the reference to economics, Appendix 6.2 on MLin relation to the theoretical and technological paradigms of computer science, Appendix 6.3 on outcome-based explanation models, Appendix **??** on mathematical functions to represent neural networks and Appendix 6.10 with a mock scientific article.

## 6.1 The science of economics

### 6.1.1 References to social sciences in Machine Learning

This account of MLas a science allows the student to revisit the comparison to economics that began this bachelor project. Jiang et al.[**?**] list 'Incorporate

insights from social sciences' as one of four explainability recommendations. Miller et al. [44] recommend to 'infuse more results from the social and behavioural sciences... and present some key results from these fields'. So there is academic support for comparing MLto economics as two empirical sciences which are both required to be explainable.

### 6.1.2 The role of dichotomies

Consider two economists discussing why eggs have become so expensive, with one saying that it just a reflection of products generally becoming more expensive, and the other saying that this specific retailer just saw his chance to earn more money on the eggs he already had in stock. The two economists will quickly realize that one is taking a macro-economic approach while the other is taking a microeconomic approach. This will not help the two agree, but the disagreement will quickly become a discussion of selecting the most appropriate model.

Demand versus supply represents the frameworks for analyzing the demand for goods of services in the form of consumption, determined by decisions by governments, households and investors, versus the supply of goods and services, determined by decisions by companies and workers.

A model that does not describe time is called static and other models are called dynamic.

Real refers to the market for goods and services whereas monetary refers to the financial markets.

The private sector is households, companies and investors whereas the public sector is national or local governments.

One older dichotomy is that between closed economy and open economy but today most analyses consider only economies with access to international markets.

Exogeneous variables are determined and explained outside of a model whereas endogeneous variables are determined and explained by the model. A model in the first-year curriculum will show how an exogeneous increase in public spending will lead to an endogeneous increase in employment.

Economists often take a partial analysis of one of the above categories applying the classic assumption of Ceteris Parisbus, i.e., 'assuming all else is equal'.

### 6.1.3 Scientific accuracy of economics versus Machine Learning

In the early 1990s, it was considered a rule of thumb that economic forecasting for Denmark was generally off by 1 percentage point, i.e., if the consensus estimate in August was that average unemployment for the following year would be 5 per cent, it could be safely assumed that without major shocks, once the year was concluded actual unemployment would have been between 4 and 6 per cent of the active population. This academically recognized 'one percentage point'-rule seemed to apply whether forecasting was done based on the ADAM model or more heuristic projections. It also applied to economic growth and the current account balance, corresponding to a variance of thousands of people and billions of Danish kroner. 5 January 2023, Politiken could report that the Danish Ministry of Finance had found an additional 52,3 billion Danish kroner in the government cash balance for

2022, relative to what had been estimated in August of the previous year, and that the year before the same variable had been underestimated by 33 billion Danish kroner [45].

### 6.1.4 The notion of optimality

The science of economics ('the study of the distribution of finite resources to meet insatiable human needs') is structured around the idea of optimality. Subjective agents choose the most satisfying consumption, the most lucrative investment or the most productive physical asset, selecting not from all options that have empirically been observed to be available, but among all options that are theoretically conceivable, subject only to the hard secondary condition that the agent must be able to afford his choice. The optimum is defined within the science of economics itself. Economists study reality, but they do so deductively, working from an expectation that economic behavior reflects the quest for a theoretically defined optimum.

The student has studied generative models, such as described by, e.g., Goodfellow [16]. Given a large dataset of pictures of dogs, the instinctive reaction of machine learners is to seek to generate a dog that, without being a copy of any one specific sample, represents the most 'salient' features of the dataset. Machine learners first seek to learn the explicity probability distribution of each pixel, and when this becomes too complicated, they train a neural network. The student reflected that there surely there must be qualitative differences among the dogs in the dataset, in terms of health, physical performance, loyalty, obedience, cuddliness, retail price or whatever 'utility' the dogs provided to their owners. Given a large dataset, why would

machine learners instinctively seek to generate the most 'probable' dog rather than the 'optimal' dog. Machine learners would respond that they would be ready to train on qualitative features and ask for two additional dataset, one of winners of dog shows and one of dogs which were healthy when the picture was taken but are known to not have lived beyond their second year. Prodded as to whether the existing dataset itself would not contain some information about what an optimal dog would look like, machine learners would entertain a discussion of self-supervised learning.

## 6.2 ML and computer science

The theoretical paradigm for computer science reflects the work by Turing. A computer with a finite number of states or 'm-sequences' is able to perform any 'operation ... used in the computation of a number'(Kragh Sørensen et al., 2023, chap. 3, p. 5, [13]). MLis consistent with the theoretical paradigm for computer science.

The technological paradigm for computer science reflects the work of von Neumann. A computer is defined by having 'memory, control unit, arithmetic unit, input unit and output unit', i.e., is able to store both data and instructions in its memory (Kragh Sørensen et al., 2023, chap. 2, p. 5, [13]). MLis consistent with the technological paradigm for computer science. MLis a succesful application of these two paradigms rather than a new paradigm in itself.

A formal science is defined by Kragh Sørensen et al. as the study of 'formal systems posited by ourselves' (Kragh Sørensen et al., 2023, chap. 1,

p. 6, [13]). This would apply to several central concepts in Machine Learning, such as the loss function and its minimization. MLis therefore consistent with the definition of a formal science.

An empirical science is defined by Kragh Sørensen et al. as the study of phenomena 'related to an external reality which may be physical or social' (Kragh Sørensen et al, 2023, chap. 1, p. 6, [13]). Machine Learning, as the study of data, is therefore consistent with the definition of an empirical science. MLcan therefore be considered both a formal and an empirical science, and as such is not different from computer science.

## 6.3 Outcome-based explanation models

There are also models of explanation that are outside of the scientific explanation category and could be called outcome-based, i.e., they emphasize the outcome of providing a certain explanation. An example is Ankarstad who defines explainable AI as methods that ensure that 'results of the solution can be understood by human experts' (Ankarstad, 2020, par. 3, [46]). This description is intuitively appealing but also has a high potential for being tautological, i.e., an explanation is defined as something a specific audience can understand. Ankarstad's definition will therefore not be applied as a definition of Explainable Machine Learning.

Ankarstad goes on to say that explainable AI 'is an implementation of the social right to explanation' (Ankarstad, 2020, par. 4, [46]).

The value proposition explanation type was developed by Lanning et al.(Lanning et al. 1988 p. 1 [47]). Lanning was a consultant tasked with

understanding why IBM could consistently sell personal computers at higher prices than the competition. The value proposition transformed the way explanations were provided in strategic management, in fact throughout a large number of disciplines from marketing to business economics. The value proposition explanation model explains complex technological or social phenomena by the value they provide, e.g., the explanation of the phenomenon of personal computer became the value it provided to the user, discarding any technical description as a mere 'delivery mechanism' for that value that was not just irrelevant but detrimental to the purpose of doing business. MLapplies the value proposition model excessively when phenomena such as ChatGPT are explained exclusively by their functionality without explaining the underlying moddel or 'delivery mechanism'. Therefore, value proposition-based explanations will not be applied as a definition of Explainable Machine Learning.

## 6.4 The data science pipeline

The data science pipeline as described by Boomsma (Boomsma, 2022, p. 32, [48]) and presented in Figure (6.1) is widely applied in the teaching of Machine Learning. The model has strong intuition and structure in the sense that it describes a consecutive set of activities, however it does not support qualitiative decisions such as model choice, nor does it distinguish MLfrom simpler prediction techniques. The data science pipeline therefore can not be part of the Systematic Methods.

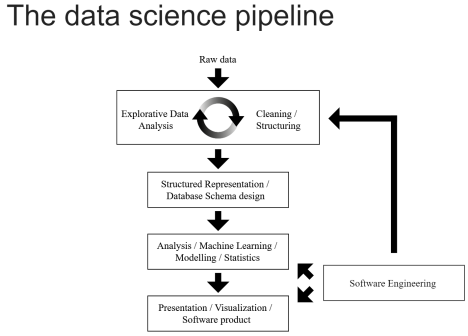


Figure 6.1: The pipeline model for Data Science as shown by Boomsma (Boomsma, 2022, p. 32, [48])

## 6.5 MLOps

Microsoft described MLOps as a framework comprising the MLLifecycle of training, packaging, validating, deploying, monitoring and retraining a 'model' (Microsoft, 2021, p. 5, [49]). MLOps, similar to the Data Science Pipeline, does not comprise the important initial choices of a MLprocess and therefore is not relevant for Systematic Methods.

## 6.6 The Universal Approximation theorem

The Universal Approximation Theorem was defined by Hornik (Hornik, 1990, p. 556, [50]) and described by Goodfellow (Goodfellow, 2016, chap. 6, p. 192, [16]). The Universal Approximation Theorem proves that a wide class of mathematical functions can be approximated by a neural network with only one layer. The intuitive example is the XOR-function, a function which cannot be approximated by linear regression but which could in fact be learnt

by a very simple neural network. Due to its generality, the Universal Approximation Theorem is a system that should be reflected in the Systematic Methods.

## 6.7 Graphs

The attempts to express MLin the form of deeper, simpler principles must lead to graphical models since some of the most advanced and least easily explained models are neural network models. Sedgewick et al. define graphs as 'pairwise connections between items' (). At the bottom of every large and opaque MLmodel lies a long sequence of pairs. If Systematic Methods could be defined to illustrate these pairs in a new way, it could help explainability.

The simplest graph model has binary 'pairwise connections', i.e., two nodes in the graph are either connected or they are not. More complex models allow the connections to take the form of mathematical functions and even stochastic functions. The neural network can be interpreted as the final stage in this evolution where functions are not just stochastic but endogeneous, i.e., defined during iterations of the model itself. Presenting neural networks as an evolution from simpler graph models would support intuition.

A layer of a neural network is a logical grouping of a number of cells. In a fully connected rectangular network, a layer can be defined as all the cells that can be reached from the initial layer in the same number of steps. Goodfellow discusses how a deep neural network, i.e., a neural network with several layers, can be interpreted as '...a hierachy of concepts that enables computers

to learn complicated concepts by building them out of simpler ones' (Goodfellow, 2016, chap. 1, p. 3, [16]). A deep neural network represented the intuition that a complex target function could be expressed as a sequence of simpler functions, the same way a long computer programme consisted of many lines of code each of which performed a simpler subfunction.

Alternatively, a layer in a neural network can be interpreted as a mathematical function that takes the output of the previous layer as input and returns its output to the subsequent layers. This brings up the role that functional programming could play as part of Systematic Methods. Mondorf discusses how MLin general is dominated by a Python-based, object oriented ecosystem which continues to grow not least due to the network effects while functional programming (F# in this case) has the merit of a more intuitive implementation of neural networks, bringing the code closer to the underlying mathematical function (Mondorf, 2022, p. 1, [51]).

Intuitive interpretations of neural networks in the form of considering the layers as a hierachy of concepts, as lines of code or as mathematical functions, can be relevant for Systematic Methods.

## 6.8 Ablation

Dosovitskiy (Dosovitskiy, 2021, p. 16, [35]) applied the ablation technique. This technique is applied when a new model achieves greater accuracy than other models. The technique allows the quantification of the contribution from various model components to this overall increase in accuracy. This is done by successively removing aspects of the model and comparatively

measuring accuracy. For example, if model B is 10 percent more accurate than model A, and if it has 20 percent more layers, and if each layer is 30 percent wider than the layers of model A, then a model C can be defined which has the same number of layers of model A. Comparing the accuracy of B and C will clarify the contribution played by layer width to the overall improvement of B relative to A. The impact of network depth can be analyzed in a similar way.

Ablation is a therefore an intuitively appealing way of accounting for MLprogress. It is applied in the articles studied in the case, but not systematically. Furthermore, results will not always be intuitive or together sum to 100 percent.

Liu et al. perform a similar experiment, demonstrating that a large part of the progress in model accuracy realized by the global-before-local revolution was in fact not due to the vector-to-vector correlation algorithm but to other improvements in modelling technique (Liu et al, 2022, p. 9, [52]). Ablation or other systematic growth accounting techniques can be part of Systematic Methods.

## 6.9 Occam's Razor

MLliterature, e.g., Seldin (Seldin, 2023, p. 26, [18]), appropriated the term Occam's Razor without either reiterating its general meaning or defining a meaning specific to Machine Learning. The general concept of Occam's Razor, such as defined by Aristotle and cited by Baker, refers to 'the superiority of ... the demonstration ... which derives from fewer ... hypotheses' (Baker,

2022, sect. 1, par. 2, [53]). The MLversion of Occam's Razor was the idea that when presented with several alternative parameter values (Seldin calls these hypotheses), it is allowed to spend more time and effort on some alternatives than on others. Greater statistical confidence can be obtained for the alternatives to which greater effort is allocated. This prioritization is done within a fixed, total 'confidence budget', i.e., it is not possible to increase confidence for all alternative parameter values. The notion of simplicity in MLwas not included in this version of Occam's Razor. Goodfellow discussed the issue of sparsity in models several times, e.g., in (Goodfellow, chap. 7, p. 229, [16]), but sparsity cannot be said to be a general principle of MLthat comes before other model desiderata discussed below in section 4.5. Since the original Occam's Razor has two elements, sparsity and prioritization, and the MLOccam's Razor has only one element, prioritization, the two cannot be said to represent the same idea. Therefore, Occam's Razor is not a symbolic generalization for Machine Learning.

## 6.10 Vector products are all you need: A mock scientific article by Tim Mondorf loosely based on Vaswani et. al. Attention is all you need (2017)

The following short text exemplifies the Systematic Methods as a paraphrase of Vaswani [34]. For practical purposes, it is limited in scope and to demonstrate that it is a mock text with no scientific value, italics have been applied.

*A global-before-local model of translation will have higher accuracy without higher computational complexity, compared to a local-before-global model. This statement or theory will be tested on the WMT English - German and English French datasets, applying X computational resource and devoting Y MLengineers over a period of Z months.*

*In Standard Model format, the model becomes*

$$\min_{q,\theta} \ell = \alpha d(p, q(\theta)) \tag{6.1}$$

*subject to a set of secondary conditions*

$$s.t.\ \theta \leq HC(q(\theta), \theta) \tag{6.2}$$

*d is the BLEU-score of overlap between model translation and existing human-generated translation. There are no specified Model Desiderata or Soft Context. The Hard Context is that the model must be global-before-local and that neither input nor output must include words on the EU Commission Blacklist.*

*The model has the following exogeneous variables: Model architecture will consist of 6 subprogrammes which each consist of two subsubprogrammes. The first such subsubprogramme in a subprogramme is a vector-to-vector-correlation algorithm and the seoond subsubprogramme is a graph data structure where nodes are connected by functions that are learned inductively. Each subsubprogramme will process a vector that is 512 elements long, corresponding to the length of the input vector.*

*The model has the following endogeneous variables: The weights of the*

*graph data structure. The subalgorithm for representing sentence structure. The search space for this subalgorithm is modulus functions and sine functions.*

*Computational complexity is quadratic since the vector-to-vector correlation algorithm is quadratic in the number of vectors. This is however mitigated by parallelization by dividing the vectors into 'Heads'. The Systematic Methods for MLstipulate that before a novel concept such as 'Head' can be introduced, it must first be confirmed that no existing computer science, mathematical or MLterms are appropriate. While 'subvector' is in fact recognized in mathematics, cf. Encyclopedia of Mathematics [54], the author has decided to cause the extra effort for his audience that comes with the introduction of a new term. This is because subsequent research will show that 'Heads' have interesting new interpretations such as correlation with semantic meaning of the words represented by the vector, and this innovation would be less clear by using the term subvector.*

***Definition 6.10.1 (Head)*** *A Head is a subvector of a given vector that embeds MLinput.*

***Result 6.10.1 (Head)*** *The separation of vectors into Heads allows for parallelization of vector product calculations. The individual Heads become interpretable in language models.*

***Example 6.10.1 (Head)*** *Consider the vector* $x = \{0, ..., 511\}$. $h_0 = \{0, ..., 63\}$ *will be the first head and* $h_7 = \{447, ..., 511\}$ *will be the eight head of* $x$.

*In the same way, the algorithm introduces three new data structures current-vector (in previous drafts this was called 'query'), other-vector ('key') and*

*vector product ('value').*

*The closed-form mathematical represenation of the model is*

$$y = F(X) \tag{6.3}$$

*It is concluded that the above statement has been corroborated since it was indeed possible to achieve higher accuracy, testing on the WMT English - German and English French datasets, applying X computational resource and devoting Y MLengineers over a period of Z months.*

*Ablation studies indicate that around half of this higher accuracy was due to the vector-to-vector correlation algorithm and the other half due to the sentence structure representation algorithm.*

*Future iterations of this theory would include applying global-before-local models to image recognition.*

### 6.10.1 Explanations derived from the model

*This section will apply the principles for Explainable MLaccording to Definition (2.4.1). One global and one local explanation will be provided.*

#### Example of global explanation

*Explanandum:*

- *The model, applying global-before-local, outperforms traditional models that were local-before-global without being less computationally efficient.*

*Explanans:*

- *It is known from linguistics that long-range relationships between words can determine context.*

- *Encoding positions as a cyclical function allows approximation of the place of a word in a sentence, replacing the state vector representation of local-before-global models.*

- ***Statement containing a reference to a recognized principle of computer science and required for the explanandum to follow from the explanans:*** *In its simplest form, the global-before-local algorithm is quadratic in time complexity since it correlates each word with all other words. For a corpus of the given size, parallelization made computations feasible.*

**Example of local explanation**

*Explanandum:*

- *The model correctly translated 'fish' to 'the new inmate', something that traditional models have failed to do.*

*Explanans:*

- *The training history of the model involved several crime novels (forward interpretation model).*

- *Over 1.000 iterations, the model will translate fish to its two literal meanings in 93 per cent of cases, to 'new inmate' in 6.5 per cent of cases and will translate incorrectly 0.5 per cent of cases (forward interpretation model).*

- ***Statement containing a reference to a recognized principle of computer science and required for the explanandum to follow from the explanans:*** *The algorithm sought a local minimum of the loss function defined in* **??** *subject to* **??**. *As long as the model translated 'fish' incorrectly, it was possible to reduce the loss so the algorithm continued increasing the multiplicative constant applied to the vector product of 'fish' and 'prison', in other words 'fish' and 'prison' attended to each other. Once the model started predicting correctly, the model concluded that a local minimum of the loss function had been attained (backward explanation model).*

# Chapter 7

# Bibliography

# Bibliography

[1] European Union. Eu general data protection regulation (gdpr). Official website of the European Union, 2018.

[2] Anders Søgaard. Explainable natural language processing. Springer Verlag, 2022.

[3] Helen Jiang Erwen Senge. On two xai cultures: A case study of non-technical explanations in deployed ai system. (pre-print) Accepted to Human Centered AI (HCAI) workshop at NeurIPS 2021., 2021.

[4] Zhiting Hu Eric Xing. Toward a standard model for machine learning. Harvard Data Science Review, 2022.

[5] Catherine Legg Christopher Hookway. Pragmatism. Stanford Encyclopedia of Philosophy, 2021.

[6] Stephen Thornton. Karl popper. Stanford Encyclopedia of Philosophy, 2022.

[7] Karl Popper. The logic of scientific discovery. Routledge, 1935.

[8] Oxford Learner's Dictionary. Systematic. Oxford Learner's Dictionary, undated.

[9] Selmer Bringsjord Naveen Sundar Govindarajulu. Artificial intelligence. Stanford Encyclopedia of Philosophy, 2018.

[10] Andrea Danyluk Paul Leidig. Computing competencies for undergraduate data science curricula acm data science task force. Association for Computing Machinery, 2021.

[11] Sebastian Raschka Yuxi Liu Vahid Mirjalili. Machine learning with pytorch and scikit-learn: Develop machine learning and deep learning models with python. Packt Publishing, 2022.

[12] Peter Naur. The science of datalogy. Communications of the ACM, 1966.

[13] Henrik Kragh Sørensen Mikkel Willum Johansen. Invitation til de datalogiske fags videnskabsteori. Lærebog til brug for undervisning ved Institut for Naturfagenes Didaktik, Københavns Universitet, 2023.

[14] Robert Sedgewick Kevin Wayne. Algorithms. Addison-Wesley, 2011.

[15] David H. Wolpert. The lack of a priori distinctions between learning algorithms. Neural Computation, 1996.

[16] Ian Goodfellow Yoshua Bengio Aaron Courville et al. Deep learning. MIT Press, 2016.

[17] Amol Mavuduru. What "no free lunch" really means in machine learning. Towards Data Science, 2020.

[18] Yevgeny Seldin. Machine learning the science of selection under uncertainty. Lecture note for Machine Learning A, 2023.

[19] Alexander Kirillov Eric Mintun Nikhila Ravi Hanzi Mao Chloe Rolland Laura Gustafson Tete Xiao Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollar Ross Girshick et al. Segment anything. Meta AI Research, FAIR, 2023.

[20] Helge Kragh. Videnskab (entry in den store danske encyklopædi). Gyldendal, 2001.

[21] Jutta Schickore. Scientific discovery. Stanford Encyclopedia of Philosophy, 2022.

[22] Oxford Learner's Dictionary. Systematic. Oxford Learner's Dictionary, undated.

[23] Thomas S. Kuhn. Objektivitet, værdidom og valg af teori. Forlaget Fremad, 1995.

[24] Thomas S. Kuhn. The structure of scientific revolutions - 50th anniversary edition. The University of Chicago Press, 2012.

[25] Jaques Hadamard. The mathematician's mind. Princeton University Press, 1945.

[26] European Union. Eu regulation on a single market for digital services (digital services act). Official website of the European Union, 2022.

[27] State of California Department of Justice. California consumer privacy act. State of California Department of Justice, 2020.

[28] Finale Doshi-Velez Been Kim. Towards a rigorous science of interpretable machine learning. Submitted directly to arXiv, 2017.

[29] James Woodward Lauren Ross. Scientific explanation. Stanford Encyclopedia of Philosophy, 2021.

[30] John Horgan. Nobel laureate steven weinberg still dreams of final theory. Scientific American, 1966.

[31] Yoav Goldberg. A primer on neural network models for natural language processing. Unpublished draft, 2015.

[32] Yann LeCun et al. Self-supervised learning: The dark matter of intelligence. Meta Research website, 2021.

[33] Stefan Sommer et al. Advanced deep learning. Course taught at DIKU, spring 2023, 2023.

[34] Ashish Vaswani Noam Shazeer Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Łukasz Kaiser Illia Polosukhin. Attention is all you need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

[35] Alexey Dosovitskiy Lucas Beyer Alexander Kolesnikov Dirk Weissenborn Xiaohua Zhai Thomas Unterthiner Mostafa Dehghani Matthias Minderer Georg Heigold Sylvain Gelly Jakob Uszkoreit Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. Published as a conference paper at ICLR 2021, 2021.

[36] Lea Henderson. The problem of induction. Stanford Encyclopedia of Philosophy, 2022.

[37] Tom B. Brown Benjamin Mann Nick Ryder Melanie Subbiah Jared Kaplany Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray Benjamin Chess Jack Clark Christopher Berner Sam McCandlish Alec Radford Ilya Sutskever Dario Amodeiet al. Language models are few-shot learners. 34th Conference on Neural Information Processing Systems (NeurIPS 2020) Vancouver, 2020.

[38] Yoshua Bengio et al. Pause giant ai experiments: An open letter. Future of Life Institue, 2023.

[39] Stefan Sommer. Professor om det værste ai-scenario: Tænk ikke på robotter med våben. Berlingske Tidende, 2023.

[40] Københavns Universitet Det Natur-og Biovidenskabelige Fakultet. Den uddannelsesspecifikke del af studieordningen for bacheloruddannelsen i machine learning og datavidenskab ved det natur- og biovidenskabelige fakultet. Det Natur- og Biovidenskabelige Fakultet, Københavns Universitet, 2019.

[41] Andrea Gesmundo. A continual development methodology for large-scale multi-task dynamic ml systems. Submitted directly to arXiv November 2022, 2022.

[42] James V. Stone. Information theory - a tutorial introduction. Sebtel Press, 2015.

[43] Michel Bourdeau. Auguste comte. Stanford Encyclopedia of Philosophy, 2022.

[44] Tim Miller Piers Howe. Explainable ai: Beware of inmates running the asylum. IJCAI-17 Workshop on Explainable AI (XAI), 2017.

[45] Ritzau. Optælling finder 52,4 milliarder kroner ekstra i statskassen. Politiken, 5. januar 2023.

[46] Nicklas Ankarstad. What is explainable ai (xai)? Towards Data Science, 2020.

[47] Michael Lanning et al. A business is a value delivery system. McKinsey Staff Paper, 1988.

[48] Wouter Boomsma. Lecture slides for data science. DIKU February 2022, 2022.

[49] Microsoft Azure (Scott Guthrie Executive Vice President of Cloud and AI is quoted on the first page). Mlops with azure machine learning accelerating the process of building, training, and deploying models at scale. Microsoft, 2021.

[50] Kurt Hornik. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Networks, 1990.

[51] Tim Mondorf. A functional journey. Project outside of Course Scope on the BSc-programme on Machine Learning and Data Science, supervised by professor Jon Sporring, Ph.d., DIKU, 2022.

[52] Zhuang Liu Zhuang Liu1 Hanzi Mao1 Chao-Yuan Wu Christoph Feichtenhofer Trevor Darrell Saining Xie1. A convnet for the 2020s. arXiv, 2022.

[53] Alan Baker. Simplicity. Stanford Encyclopedia of Philosophy, 2022.

[54] C. Birkenhake. Abelian surface - entry in encyclopedia of mathematics. Encyclopedia of Mathematics, 2011.