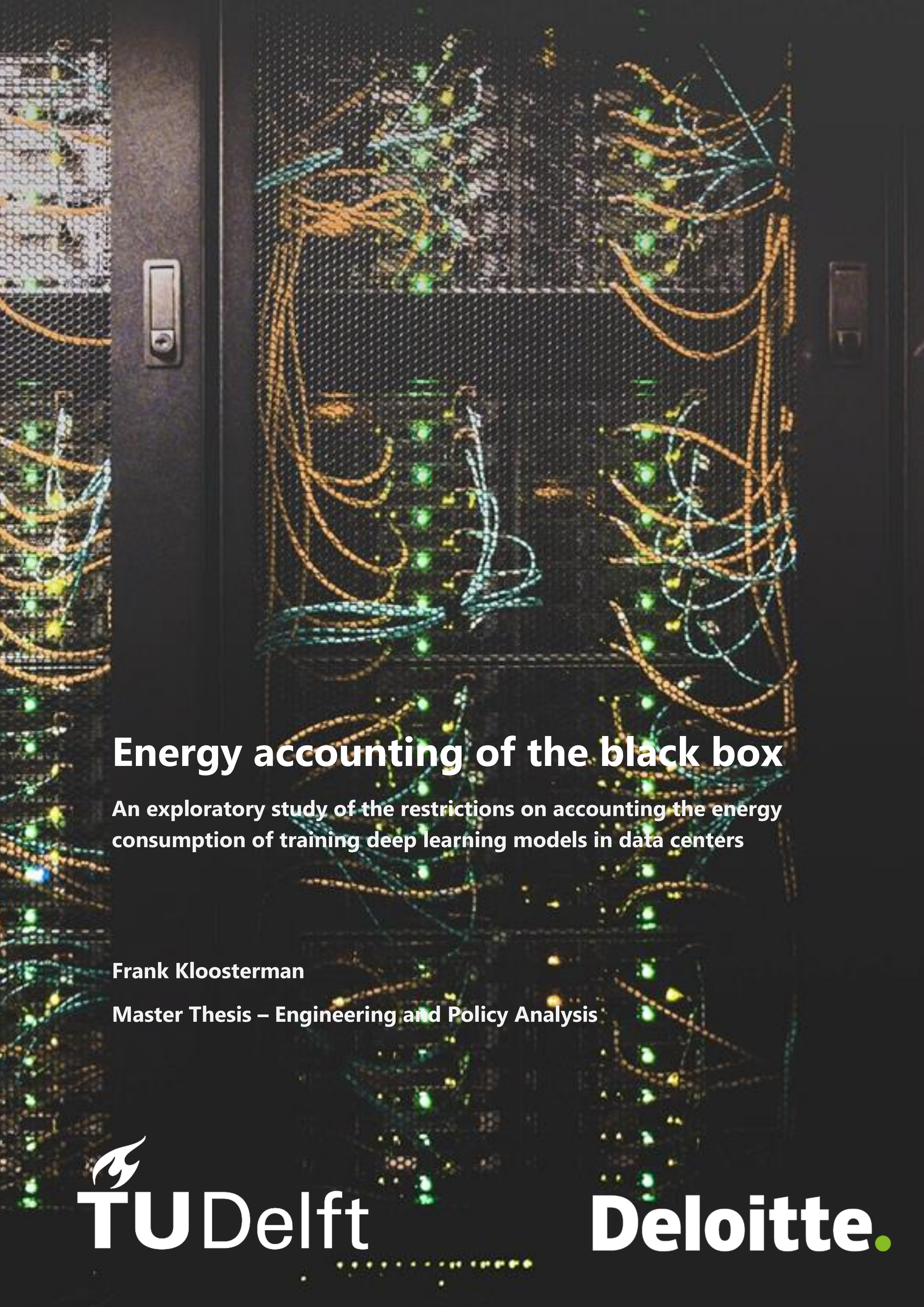

Energy accounting of the black box

An exploratory study of the restrictions on accounting the energy consumption of training deep learning models in data centers

Frank Kloosterman

Master Thesis – Engineering and Policy Analysis

TU Delft

Deloitte.

ENERGY ACCOUNTING OF THE BLACK BOX

An exploratory study of the restrictions on accounting the energy consumption of training deep learning models in data centers

A thesis submitted to the Delft University of Technology in partial fulfillment of the requirements for the degree of

Master of Science in Engineering and Policy Analysis

by

Frank Kloosterman

4373863

To be defended in public on April 9 2021

Graduation committee

Chairperson	Prof.dr. Martijn Warnier	Multi-Actor Systems
Daily Supervisor	Dr. Roel Dobbe	Engineering Systems and Services
First Supervisor	Dr. Aaron Ding	Engineering Systems and Services
First External Supervisor	Bas van Boven	Deloitte
Second External Supervisor	Tony Damen	Deloitte

Frank Kloosterman: *Energy accounting of the black box: An exploratory study of the restrictions on accounting the energy consumption of training deep learning models* (2021)

The work in this thesis was made in the:

Faculty of Technology, Policy, and Management
Delft University of Technology

As a graduate intern at:

Tech Strategy and Operating Model
Deloitte The Netherlands

Supervisors:	Prof.dr. Martijn Warnier	Faculty of Technology, Policy, and Management
	Dr. Roel Dobbe	Faculty of Technology, Policy, and Management
	Dr. Aaron Ding	Faculty of Technology, Policy, and Management
	Bas van Boven	Deloitte The Netherlands
	Tony Damen	Deloitte The Netherlands

PREFACE

In the last six months I've dived into the world of model developers, deep learning, service providers, processing units, and data centers. A world completely new to me, but it interested me from the moment I started. In these months, I've learned a lot about the opportunities and limitations of deep learning and especially the complexity of it.

First of all, I would like to thank my supervisors for the guidance and support during this process, especially with COVID that made graduating even more isolated than usually. Martijn, I would like to thank you for your help in the past year(s). I really appreciate the meetings we had before my thesis, when you supervised my internship project and helped me with my thesis preparations. Roel, thank you for the massive amount of hours you put into my thesis and for providing this very specialised but interesting topic. I enjoyed our weekly meetings, although we often got completely off topic. It was an honour to be your first graduation student! Aaron, I would like to thank you for your specialised feedback. Although you mainly took on a formal role during the meetings, I appreciate your feedback and IT expertise.

Second, I would also like to thank my supervisors from Deloitte for their guidance, expertise, and support. It was strange to do a full internship from home, but I have virtually met some great colleagues! Everybody from TS&T told me that it should be frustrating to graduate during the pandemic, but the interaction with the team often gave me the energy to continue. Bas, I would like to thank you for all your critical notes and hours you spent on my thesis. Your computer science expertise helped a lot in finding and investigating the right concepts and asking the important questions. Tony, thank you for the efforts you put into my thesis and the connections you brought. The energy consumption of training deep learning models might not be your first expertise, but you really helped me with structuring my thesis and putting all the parts into the right places.

Thirdly, I would like to thank my parents and my sister for all the support in the past years. Last year, I had some tough moments with my canceled exchange in Santiago and slightly isolated graduation. However, you helped and supported me to keep on going and finish my thesis in time! Moreover, I would like to thank my friends from study and outside my study for the amazing years, whether it was during studying, for mental support, or just for having fun. You gave me many great memories and hopefully many more to come.

Before you lies the closure of my student-era. The last 6,5 years was an amazing period at the faculty of Technology, Policy and Management, with the friends I've met, my student house *'Huize van God Los'*, my committees and board year at S.V.T.B. Curius, the Wijnhaven building in The Hague, my master-student house *'De WK'*, my internship at Arcadis, my six-week exchange adventure in Santiago de Chile, and graduation internship at Deloitte. I would like to thank everybody who I've met and with whom I have enjoyed these years. Let's start my next adventure!

Frank Kloosterman
The Hague, 26 March 2021

EXECUTIVE SUMMARY

Scientists believe Artificial Intelligence (AI) will play a dominant role in solving global warming, by e.g. tracking greenhouse gas emissions and optimizing energy markets [Rolnick et al., 2019; Vinuesa et al., 2020]. AI can also be implemented for other future problems, such as reducing healthcare cost and improving the quality of education [Kalis et al., 2018; Dwivedi et al., 2019; Königstorfer and Thalmann, 2020; Rangaiah, 2020]. AI has already been implemented in most sectors and has promising application in many more. AI is a concept with a varying definition, but a broadly accepted definition is that a system is intelligent when a person cannot distinguish the different between another human and the system [Haenlein and Kaplan, 2019].

However, broadly applying AI also has a negative impact on society. Recent studies reported on the growing carbon footprint of AI applications and researcher are expressing their concerns about the footprint of AI [Vinuesa et al., 2020; Hoa, 2020]. Especially deep learning (DL) is a field of study within AI that requires a lot of computational power, so consumes a lot of energy, and therefore produces a lot of carbon dioxide [Li et al., 2016; Pouyanfar et al., 2018; Strubell et al., 2019]. The advantages of DL are that it can process raw input data and can identify complex patterns based on deeper layers in the data [Lecun et al., 2015]. These advantages enable faster and more accurate problem solving. The application of DL is expected to grow, since it is expected to become the dominant big data analysis method in many industries and the computational demand of the state-of-the-art DL models grows exponentially [Amodei and Hernandez, 2018; Choudhary and Linden, 2020; Sicular and Vashisth, 2020].

Ideally, the energy consumption of DL models would be traceable to determine which models and applications consume most energy and how this can be reduced. However, in reality the problem is that the energy consumption of DL models is very hard to track. Consequently, there is no awareness about the energy consumption of these models and therefore no direct incentive to limit the energy consumption. The objective of this research is to identify the restrictions that arise when trying to account the energy consumption of developing deep learning models in data centers. The main research question is therefore:

What are the restrictions on accounting the energy consumption of building, training, and maintaining Deep Learning models in data centers?

To answer the main research question, four exploratory case studies have been compared with a cross-case analysis, to make robust conclusions on the availability of information [Creswell, 2003; Yin, 2018]. Next, the qualitative data of these case studies has been analysed by coding restrictions and by grouping them, to identify the restrictions on accounting the energy consumption [Auerbach and Silverstein, 2003]. Then, these findings have been validated with expert in-depth interviews and the qualitative data of these in-depth interviews have been analysed to identify additional restrictions.

The cross-case analysis revealed that there is little information available across the cases to account the energy consumption of training DL models. The cross-case analysis consisted of four cases, in which a DL models have been built and trained. The process to find and train the final and best model architecture was often unstruc-

tured and poorly documented. So, the minimal required information to determine an estimation of the energy consumption of the process was available in only two of the four cases. The outcomes of these energy consumption estimations were relatively low, partly because the method underestimates the energy consumption of the cases and partly because the DL applications in the cases were relatively simple. Moreover, the cross-case analysis identified the stakeholders and their current roles in energy accounting the training of DL models. This revealed that most stakeholders do not take an active role in the energy accounting, besides the scientific community.

The qualitative data analysis of the case studies revealed nine groups of restrictions, aggregated from more detailed restrictions shared in the interviews. These restrictions can be classified into three categories based on the causes of the restrictions in the interviews. These categories describe the causes of the restrictions, but can also be used to formulate solution directions to overcome the restrictions. The three categories are Organizational, Social, and Technical.

From these nine restrictions, eight have been validated by the expert in-dept interviews. The experts have been divided into three perspectives, namely the governmental, the scientific, and the service provider. However, these restrictions are not validated. The eight validated restrictions from the case studies and corresponding categories are:

- Complexity of Deep Learning *(Social & Technical)*
- Innovative stage of Deep Learning *(Social & Technical)*
- Lack of incentive to determine energy consumption *(Organizational & Social)*
- Lack of model developers' energy accounting knowledge *(Organizational & Social)*
- Lack of societal awareness (*Organizational & Social)*
- Lack of systematic evaluation of models *(Organizational)*
- Long and diverse training time *(Organizational & Technical)*
- No hardware details available *(Organizational & Social)*

To conclude, building, training, and maintaining Deep Learning models proved to be an unstructured process, which resulted in scattered information regarding the energy consumption of these models. This makes it really hard to account the energy consumption of training these models. Also, the stakeholders pay little attention to the energy consumption of the models. They have no direct incentive to account or reduce the energy consumption and/or they are not aware that remote servers consume significant amounts of energy. The restrictions on accounting the energy consumption of training Deep Learning models can be overcome by providing options for the stakeholders to educate themselves, stimulating interaction among stakeholders, and creating a structure for the stakeholders to organize themselves and the information required for energy accounting. This provides the stakeholder with the means they need to cope with the technical complexity of Deep Learning.

Concrete policies to overcome the validated restrictions are to (i) set standards on what and how to communicate the energy consumption of service providers to the model developers, (ii) set standards on what is high, normal, and low energy consumption for certain DL architectures and applications, (iii) develop and issue certificates that require logging of all training hours, and (iv) develop a knowledge sharing platform for best practices of (DL) technologies and the energy consumption of these technologies.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI	Artificial Intelligence
CPU	Centralised Processing Unit
DL	Deep Learning
DRAM	Dynamic Random-Access Memory
FLOPS	Floating Point Operations per Second
FPO	Floating Point Operations
FPGA	Field Programmable Gate Arrays
GPU	Graphics Processing Unit
ML	Machine Learning
NLP	Natural Language Processing
PUE	Power Usage Effectiveness
TDP	Thermal Design Power
TPU	Tensor Processing Unit

1 INTRODUCTION

1.1 MOTIVATION FOR DEEP LEARNING

In 2018, the world's leading climate scientists presented a report with a clear message; We need urgent change to cut the risk of extreme heat, drought, floods, and poverty caused by climate change [Watts, 2018]. If the world does not reduce its emissions, global warming is likely to reach 1,5 degrees of Celsius by 2030 what will cause risks to health, livelihoods, food security, water supply, human security, and economic growth [IPCC, 2018]. Many scientists believe artificial intelligence (AI) can play a prominent role in mitigating global warming [Rolnick et al., 2019; Vinuesa et al., 2020]. It might prevent climate change by, inter alia, tracking the source of greenhouse gas emissions and optimizing the fluctuating energy market with new sustainable energy sources.

The definition of AI varies between scientists and over time, but a broadly accepted test to determine whether a system is intelligent is the Turning Test [Haenlein and Kaplan, 2019]. This test states that a system is intelligent if a human cannot distinguish the difference between another human and the system. Section 3.1.1 elaborates on the definition of AI. Besides reducing greenhouse gas emissions, AI is applied in many different sectors in the recent years. It has promising application in sectors such as healthcare, education, finance, manufacturing, retail, supply chain, and utilities [Kalis et al., 2018; Dwivedi et al., 2019; Königstorfer and Thalmann, 2020; Rangaiah, 2020]. AI can assist in surgeries, design a customized learning profile for students, and determine the creditworthiness of banks' customers in minutes. In other words, it will be involved in almost every part of our future lives. However, facilitating all these changes with AI comes at a cost.

Recent studies reported on the growing carbon footprint of AI applications and more researchers are expressing their concerns that AI is not only for the good [Vinuesa et al., 2020; Hoa, 2020]. In 2019, Strubell et al. published a report about the life cycle assessment for training several common state-of-the-art AI models. They concluded the whole process of training one model could emit up to 284.000 kg of CO2, which is the equivalent of the combined emissions of 5 US cars in their lifetimes (including fuel and production) or the sum of 315 single-person round-trip flights between New York and San Francisco (approximately 6 hours) [Strubell et al., 2019; Hoa, 2019b].

Within the field of AI there is one field of study that requires most computational power and therefore consumes most energy and produces most emissions, namely deep learning (DL) [Li et al., 2016; Pouyanfar et al., 2018]. DL is often used interchangeably with machine learning (ML), but ML is a broader field of study containing DL and ML is in turn a field of study within AI (see Figure 1.1). Where ML 'learns' from carefully selected input and output, DL can process raw data and can identify complex patterns based on deeper layers in the data [Lecun et al., 2015]. This reduces the engineering by hand, but demands more data and computational power. Moreover, the neural networks of ML models contain some hidden layers and DL models in general more than 3, up to thousands of layers Dimiduk et al. [2018]. AI and DL are explained and defined in more detail in Section 3.1.2.

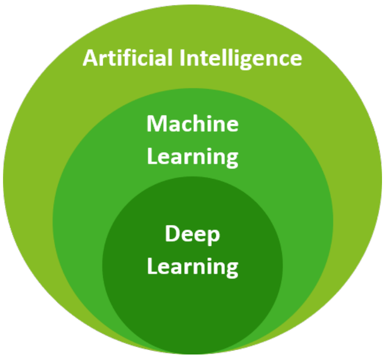

Figure 1.1: A visualisation of DL being a field of study within ML and ML in turn being a field of study in AI.

Moreover, DL might contribute significantly to climate change, since the amount of application in different industries and the demand for computing power of new DL models is growing rapidly. Firstly, Choudhary and Linden [2020] expect DL to be the dominant big data analysis method over other ML methods by 2022. Industries impacted by DL include healthcare, transportation, national security, military, criminal justice, cities, finance, and social media [Sicular and Vashisth, 2020]. Secondly, the demand for computational power of state-of-the-art DL models is rising faster than ever before [Hoa, 2019a]. Amodei and Hernandez [2018] reported that the demand for computational power in the largest DL training runs increased exponentially with a doubling time of 3.4 month. This resulted in an increase of a factor 300.000x in the years between 2012 and 2018.

To put this doubling time in perspective, Moore's law was the norm for years and stated that the number of transistors on a integrated circuit doubles every two years. The number of transistors does not relate linearly to the computational power of CPUs or GPUs, but it does affects the computational power. More important, the computational demands of state-of-the-art DL models develop a lot faster than the hardware, resulting in a growth in the number of data centers and energy consumed by these data centers.

1.2 PROBLEM STATEMENT AND RELEVANCE

Ideally, the energy consumption of DL models would be traceable to determine what models and applications consume most energy and how this can be reduced. This provides insights for modellers into what effects their model decisions have, insides for product owners into what the impact of their product is, and provides decision makers with handles to limit the energy consumption of these models. However, in reality the problem is that the energy consumption of DL models is very hard to determine, due to many restrictions. Consequently, there is no awareness about the energy consumption of these models and therefore no incentive to limit the energy consumption.

1.2.1 Societal and scientific relevance

The problem, as stated above, is relevant to society, since researchers recently reported on the significant contribution of DL models to CO2 emissions [Strubell et al., 2019]. This contributes to global warming with many significant societal risks, as described in section 1.1. It is therefore relevant that the energy consumption of DL models and the CO2 emissions linked to it can be accounted.

The scientific relevance of the problem described above is to formulate and validate a method that can be used to determine the energy consumption of DL models, because there is no straight-forward way to determine it [García-Martín et al., 2019]. Moreover, since there is no straight-forward method, it is scientifically relevant to understand what restrictions limit the energy accounting of DL models in practice.

1.3 RESEARCH OBJECTIVE AND MAIN RESEARCH QUESTION

The objective of this research is to identify the restrictions that arise when trying to account the energy consumption of developing deep learning models in data centers. The main research question corresponding with this research objective is:

What are the restrictions on accounting the energy consumption of building, training, and maintaining deep learning models in data centers?

1.4 EPA SUITABILITY

Within the master program of Engineering and Policy Analysis (EPA) is a clear focus on how decision makers can tackle grand challenges. When thinking of grand challenges, one is often referred to the Sustainable Developments Goals (SDGs) of the United Nations, since these goals represent the biggest challenges of today's society [Vinuesa et al., 2020]. One of these goals (Goal 13: Climate Action) is take urgent action to combat climate change and its impacts [Department of Economic and Social Affairs, 2019]. This is related to the aim of this research, since this research strives to contribute to the accountability and transparency of the energy consumption of training Deep Learning models in data centers. The goal of increasing the accountability and transparency is to create more awareness about the environmental impact of training Deep Learning models to eventually reduce its impact.

1.5 OUTLINE OF THE THESIS

To answer the main research question, Chapter 2 presents the scientific approach that is followed. Next, Chapter 3 explains the literature about what Artificial Intelligence and Deep Learning are and what metrics can be used to express the computational power of the two. Chapter 4 also elaborates on literature about the methods to account the energy consumption of training deep learning models. Chapter 5 describes the case studies that are used in this thesis and analyses the information in the different cases what is available to account the energy consumption. Chapter 6 presents the identified restrictions from the case studies and shows what restrictions are identified in what case studies. Chapter 7 validates the identified restrictions with in-depth interviews with experts and presents additional restrictions brought up by in the in-depth interviews. Chapter 8 discusses the thesis results by interpreting the results in the literature, discussing the impact of the results, and explaining the limitations. Chapter 9 eventually answers the sub-research question, concludes on the main research question, suggests policy considerations, explains the contribution to society and science, and recommends on further research.

2 RESEARCH APPROACH

This chapter presents the approach that is used to answer the main research question as presented in section 1.3. Also, this chapter describes the sub research questions that are formulated to answer the main research question and the methods that are used to answer these sub research questions. Finally, this chapter presents how the approach and research questions relate to each other in the research design.

To determine what research approach is appropriate for a research project, Creswell [2003] described three different elements of the project. These elements are:

1. What knowledge claims are being made by the researchers, i.e. what assumptions have the researchers about how and what they will learn?
2. What strategies of research fits the knowledge claims?
3. What methods of data collection and analysis will be used for the research strategy?

For the knowledge claims, this research mainly has characteristics of constructivism as described by Creswell [2003], which is a perspective that focuses on the subjective view of the participants. So, for this research project the focus is on what kind of information is available for the participants to account the energy consumption, rather than what is the actual content of the information. Also, the focus is on what the participants perceive and name as restrictions on accounting the energy consumption, rather than what actually were the restrictions. The next section describes the strategy or approach that fits this knowledge claims and will describe the methods of data collection according with the strategy.

2.1 QUALITATIVE APPROACH

The strategy in-line with constructivism is the qualitative research approach [Creswell, 2003]. First, multiple case studies are investigated to explore what information is available in practice when DL models are developed and to determine to what extent the energy consumption of the cases can be accounted. Next, the interviews of the case studies and the in-depth interviews are analysed to identify and validate the restrictions on accounting the energy consumption for training DL models.

2.1.1 Exploratory case studies

Yin [1984] described a case study as an empirical research project, which:

- investigates a contemporary phenomenon within its real-life context: when
- the boundaries between phenomenon and context are not clearly evident; and in which
- multiple sources of evidence are used.

These characteristics described by Yin [1984] are applicable on this research project, since (i) this case study investigates the energy consumption of DL models that is actually consumed in data centers, (ii) it is unclear what part of the energy consumption of data centers can be attributed to DL models, and (iii) part of the information can be found in literature but this information is insufficient, as section 2.2 describes.

The case study approach is typically used to gain insight into a program, an event, an activity, a process, or (an) individual(s) that are bound by time and activity [Creswell, 2003; Verschuren and Doorewaard, 2010]. Case studies can be explanatory, descriptive, and exploratory or a combination since they are not mutually exclusive [Schell, 1992; Yin, 1998]. Explanatory case studies focus on explaining the occurrence of a phenomena, descriptive case studies on describing certain phenomena after observing and analysing it, and exploratory case studies on exploring a new field of studies and opening up research possibilities.

Case descriptions

The case studies in this research project investigation is an exploratory case study, since it aims to reveal what knowledge is known about the process of training a DL model in practise. First, four different cases will be investigated to discover what information is available in the different cases. These cases all used DL models to solve a problem and trained the DL models in different ways.

The cases of this cross-case analysis were selected based on availability within the time-span. Since the multiple case study design is only a part of the research project with a limited time, all cases that were at-hand are used. Chapter 5 describes the different projects and problems that were solved with DL.

Cross-case analysis

The aim of this research project is to identify restrictions on the energy consumption of various DL models, so the method needs to be robust, i.e. applicable for many cases. The evidence of a multiple-case study design is considered to be more robust than the evidence of a single-case study design [Yin, 2018]. Therefore, this research project examines multiple-case studies.

The case studies in this research project are conducted and reported individually and later on compared with each other. The case study interviews are semi-structured, so it is able to compare the different cases and ask questions outside the line of questioning if it is relevant. The multiple case study design is based on Figure 2.5 of Yin [2018] and presented in Figure 2.1. Before starting with the case studies, a theory needs to be developed based on the existing literature. Once this theory is developed, the cases can be selected and the data collection protocol can be designed. Next, the interviews of the different case studies can be conducted and the case studies need to be reported individually, to identify patterns within each case study. After conducting all the case studies, cross-case conclusions can be drawn, the initially theory can be modified, and a cross-case report can be written.

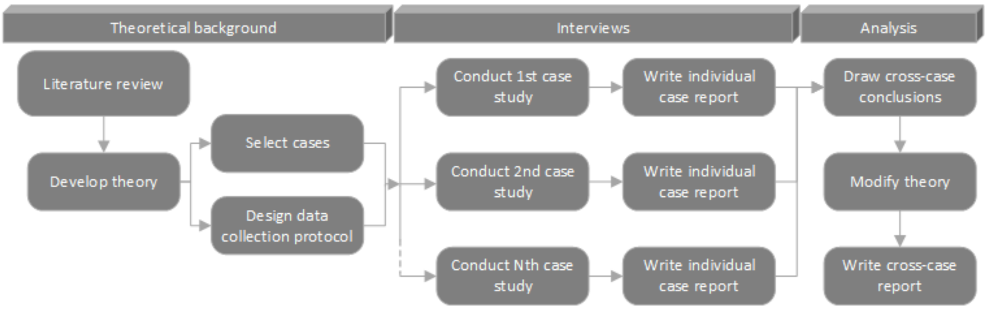

Figure 2.1: Multiple case study design, adjusted from Yin [2018]

2.1.2 Qualitative data analysis

To analyse the interviews of the case studies and the in-depth interviews, the qualitative data analysis process of Auerbach and Silverstein [2003] is used. They describe three phases and six steps to analyse the data by coding the interviews and constructing a theory. The three phases of Auerbach and Silverstein [2003] are (I) Making the Text Manageable, (II) Hearing what was said, and (III) Developing theory. In total these three phases consist of six steps for constructing a Theoretical Narrative from text, these steps are listed below:

1. Explicitly state your research concerns and theoretical framework
2. Select the relevant text for further analysis, by reading through your raw text with Step 1 in mind and highlighting relevant text.
3. Record the repeating ideas by grouping together related passages of relevant text.
4. Organize themes by grouping repeating ideas into coherent categories.
5. Develop theoretical constructs by grouping themes into more abstract concepts consistent with your theoretical framework.
6. Create a theoretical narrative by retelling the participant's story in terms of theoretical constructs.

Data analysis process

In this research process, the steps as describes above are executed multiple times. First, the case study interviews are coded with the research concerns and theoretical framework in mind. Second, these codes are grouped into coherent categories and more abstract themes. Third, these themes are validated with in-depth interviews from different fields. The found restrictions are validated explicitly by asking the experts' opinions about them and implicitly by coding the in-depth interviews. This process validates or invalidates the restrictions that are initially identified in the case studies and compliments the list with additional restrictions.

In-depth interviews

The interviews of the case studies are semi-structured and based on a line of questioning to identify the available information about the energy consumption of the projects. The in-depth interviews use the identified restriction from the case studies as a structure. Therefore, the in-depth interviews are less structured than the case studies, but not unstructured.

The selected interviewees are stakeholders that are identified in the case studies. The persons eventually approached is based on the network of the researcher and is extended with the snowball-approach. This means that every interviewee is asked for somebody else who might add information to the research project. Ideally this process is continued until the restrictions are validated from different perspectives and no new restrictions are identified. However, a complete analysis according to this setup turned out to be too time consuming and can therefore not be fulfilled within the limited time.

2.2 SUB-QUESTIONS AND METHODS

This section presents the different sub-questions that fit the methods, as presented above. This section discusses for each sub-question the method and data required to answer the question.

The first chapter briefly touches upon the lack of information to determine the computational power of AI and DL models. And, since the computational power of a model is related to its energy consumption, the first step of this research project is to determine what metric can be used to determine the computational power. To do so, a literature review is conducted to present the different metrics that are already used, what assumptions are made to use these metrics, and eventually what metric defines the computational power of DL models best. Chapter 3 describes the literature review to answer the first sub-research question:

1. What metrics can be used to define computational power of Deep Learning models?

The second research question focuses on the methods that are already available to account the energy consumption of AI and DL models. Also, the answer to this question reveals the limitations of existing methods. A literature review is conducted to identify these accounting methods and limitations. Chapter 4 discusses this literature review and the corresponding sub-question is:

2. What methods are available to account the energy consumption of training Deep Learning models?

Once the information from literature provides for the energy accounting of DL models has been identified, it needs to be determined whether that information is useful in practice. So, the next step is to discover what information is available in real case studies, compare this information with the literature, and determine how the energy consumption of DL models can be accounted, what the limitations are of the energy accounting, and what other stakeholders are involved. To discover this, multiple case studies are conducted, analysed, and compared with literature. Chapter 5 discusses the cases studies and the analysis. The corresponding sub-question is:

3. To what extent can the energy consumption of training Deep Learning models be accounted in practice?

To discover what factors restrict the energy accounting of developing DL models, the interviews of the case studies are analysed and coded. The output of this process is a list with restrictions that are identified explicitly and implicitly. This list is then validated with experts via in-depth interviews and afterwards again coded to identify again the restrictions that are named explicitly and implicitly. The coding of these interviews is conducted with the software ATLAS.ti 9. This software is a powerful workbench for the qualitative analyse of large quantities of textual data [ATLAS.ti, 2020]. The corresponding sub-question is:

4. What are the restrictions on the energy accounting for developing Deep Learning models?

2.3 RESEARCH DESIGN

Below presents in Figure 2.2 is the research design of this research project. It visualises the different methods and sub questions in order of execution to give a full oversight of the research.

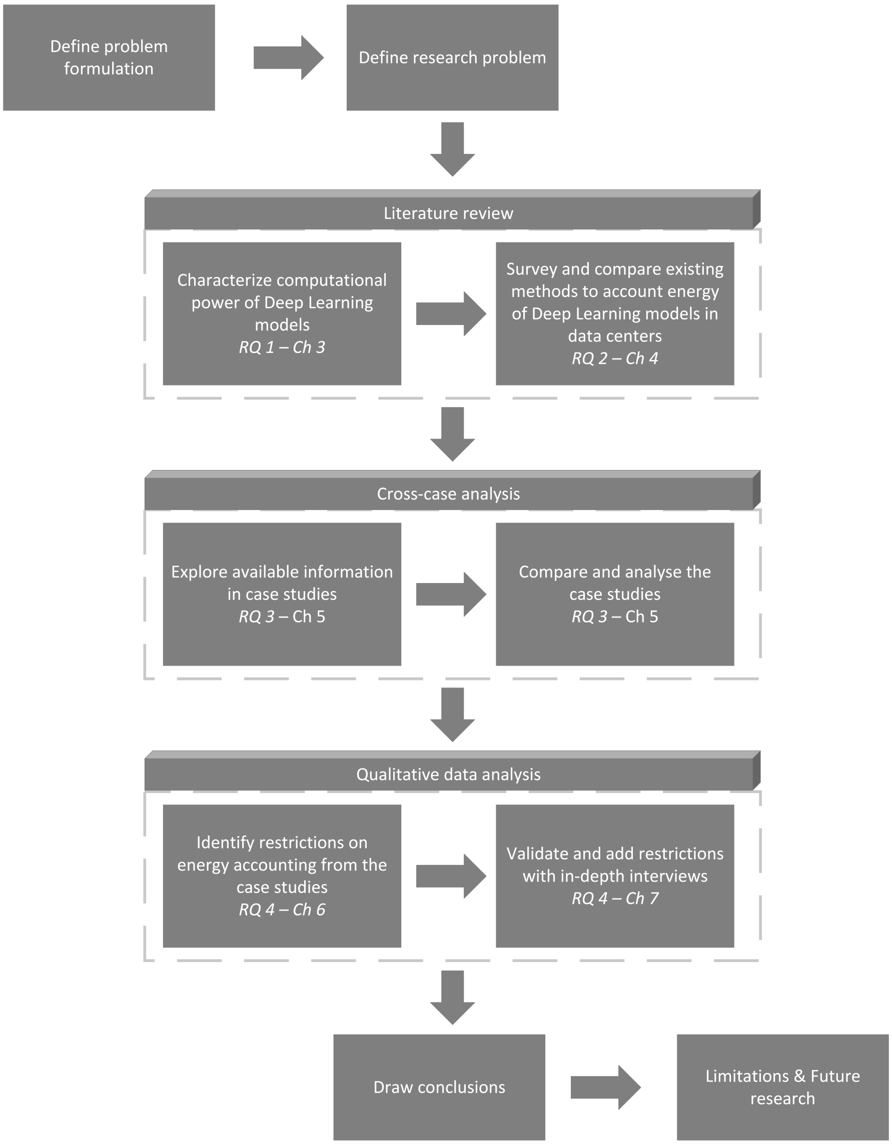
Define problem formulation
Define research problem
Literature review
Characterize computational power of Deep Learning models
RQ 1 – Ch 3
Survey and compare existing methods to account energy of Deep Learning models in data centers
RQ 2 – Ch 4
Cross-case analysis
Explore available information in case studies
RQ 3 – Ch 5
Compare and analyse the case studies
RQ 3 – Ch 5
Qualitative data analysis
Identify restrictions on energy accounting from the case studies
RQ 4 – Ch 6
Validate and add restrictions with in-depth interviews
RQ 4 – Ch 7
Draw conclusions
Limitations & Future research

Figure 2.2: Full research design of the research project.

Once the research problem has been identified, the research consists of three main components, namely the literature review, the cross-case analysis, and the qualitative data analysis. First, the literature is used to explore the computational power of DL models (Chapter 3) and to explore how this can be used to account the energy consumption of these models (Chapter 4). Next, multiple case studies are executed and compared to explore what information is available at the case studies and how this can be used to account the energy consumption of DL models (Chapter 5). Then, these case study interviews are used to identify the restrictions on what makes it so hard to account the energy consumption in practice (Chapter 6). Next, these findings are validated and supplemented with in-depth interviews (Chapter 7). Finally, conclusions can be drawn, limitations can be listed and future research can be suggested (Chapter 8 and 9).

3 COMPUTATIONAL POWER OF DEEP LEARNING

This chapter presents computation power metrics for DL and therefore answers the first sub-question as presented in section 2.2. The correlated sub-question is: *What metrics can be used to define computational power of Deep Learning models?* First, the chapter presents a definition of DL and how it differs from AI. Next, the section describes different metrics as defined in literature to determine computational power of AI models and DL models. Finally, this chapter answers the sub-question as stated above.

3.1 WHAT IS DEEP LEARNING?

Chapter 1 already briefly discusses how DL is a field of study within AI. To elaborate on the differences between AI and DL, this section first presents how literature defines AI. Next, DL is defined to stress the differences.

3.1.1 Artificial Intelligence

The first notion of Artificial Intelligence (AI) dates back to over 200 years ago, when a chess playing machine named "the Turk" was invented [Buchanan, 2005]. This machine fooled people to let them think it had a mind of its own. Modern AI concepts found its origin during the second world war, when computers were invented. After the second world war, Alan Turning introduced the Turning Test to determine whether an artificial system has intelligence [Haenlein and Kaplan, 2019]. This test stated: "If a human is interacting with another human and a machine and unable to distinguish the machine from the human, then the machine is said to be intelligent." After those years, scientists made tremendous improvements in the development of AI and in the 1980s the first commercial parties got involved [Kaplan, 1984]. Back in the 1980s, AI was defined by applications such as pattern analysis, computer science, and cognitive psychology [van den Besselaar and Leydesdorff, 1996]. However, it was nearly impossible to extensively apply these techniques, due to a lack of computational power and available data. In recent years, these two bottlenecks improved tremendously and scientists and commercial parties were able to practise and improve the AI techniques.

The fact that the meaning of AI changed so much over time, already reveals how subjective the concept is. Within the scientific community is therefore no general agreement about the definition of AI [Kok et al., 2009; Vinuesa et al., 2020]. In 2018, the European Commission defined AI with a broad term, namely:

> *Systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals.*

However, the above definition is too broad to be used as an overarching definition, since concepts as *intelligent* and *some degree of autonomy* are subject of many discussions and can cause a lot of confusion [Kok et al., 2009]. Therefore, literature often describe capabilities or skills a computer or model should posses to be called AI. Kok et al. [2009] defined four skills, based on logic that these skills are necessary to pass the Turning Test. So, what skills are necessary for a person to not be able to distinguish a computer from a human [Haenlein and Kaplan, 2019]. The four skills that AI would need at least according to Kok et al. [2009] are:

- Natural language Processing (NLP): able to communicate in a natural language like English.
- Knowledge representation: able to have knowledge and store knowledge.
- Automated reasoning: able to reason based on stored knowledge.
- Machine learning: able to learn from its environment.

Vinuesa et al. [2020] elaborated on these skills and broadened the skills to some extent. They specified six capabilities of which software should have at least one to be called AI. These capabilities are:

- Perception (audio, visual, and tactile)
- Decision-making
- Prediction
- Automatic knowledge extraction and pattern recognition
- Interactive communication
- Logical reasoning

These capabilities touch a large amount of sub-fields in different research areas, including machine and deep learning [Vinuesa et al., 2020]. This last list of capabilities is the definition of AI as used in this report.

3.1.2 Definition of Deep learning

As described in the introduction, DL is a field of study within the field of study of machine learning and machine learning in turn is a field of study within AI (Figure 1.1). DL can be perceived as one of the technique to realise the AI capabilities.

DL basically means that a model is provided with the input and output, but determines the relation between the input and output itself [Pouyanfar et al., 2018]. These relations are determined by different parameters or nodes in different layers between the input and output and together form the neural network (see Figure 3.1). These layers between the input and output are the hidden layers, since the model determines these parameters and are often unknown to the modellers. This is also where Dimiduk et al. [2018] differ ML and DL. They state that ML neural networks use one-to-three hidden layers and DL neural networks have tens-to-thousands.

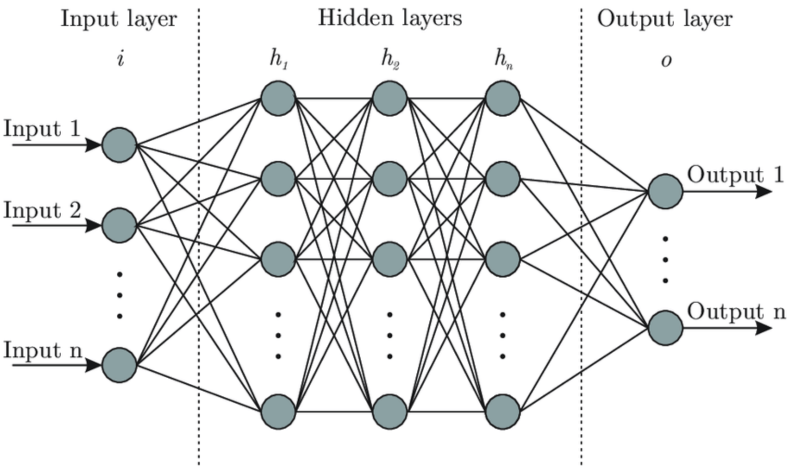

Figure 3.1: Visualisation of a neural network from Bre et al. [2018]

Scientifically, Pouyanfar et al. [2018] state that DL uses multiple layers to represent different structures of data to build computational models. In other words, DL discovers complex structures in big data sets to indicate how a model should change its internal parameters, which are used to calculate the outcomes of each layer based on the outcomes of the previous layer [Lecun et al., 2015]. There are different DL networks that can be applied to discover the complex structures in the data, i.e. the architecture of the neural network. In Table 3.1 is a list of different common DL network architectures presented with related descriptive key points, reviewed and presented by Pouyanfar et al. [2018].

Table 3.1: Different types of common Deep Learning Network Architectures with key points, presented by Pouyanfar et al. [2018]

Deep Learning Network Architectures	**Descriptive Key Points**
Recursive Neural Network (RvNN)	Uses a tree-like structure, Preferred for NLP
Recurrent Neural Network (RNN)	Good for sequential information, Preferred for NLP & speech processing
Convolutional Neural Network (CNN)	Originally for image recognition, Extended for NLP, speech processing, and computer vision
Deep Believe Network (DBN)	Unsupervised learning Directed connections
Deep Boltzmann Machine (DBM)	Unsupervised learning Composite model of RBMs Undirected connections
Generative Adversarial Network (GAN)	Unsupervised learning Game-theoretical framework
Variational Autoencoder (VAE)	Unsupervised learning Probabilistic graphical model

There are many different architectures developed over recent years for specific application, but convolutional and recurrent neural networks are the most widely adopted architectures and most easily applicable by industries [Li et al., 2016; Sicular and Vashisth, 2020]. Industries impacted by DL include healthcare, transportation, national security, military, criminal justice, cities, finance, and social media.

3.2 HOW TO DEFINE COMPUTATIONAL POWER?

This section is divided into two parts. The first part describes some metrics that can be used to express computational power of AI models. This knowledge is required to understand why it is hard to determine the computational power of DL models. The second section will explain different metrics to express the computational power of these models.

3.2.1 Computational power metrics

Literature provides different metrics to express computational power of AI. Henderson et al. [2020] summarizes a list with different computational metrics that can be used to account the energy consumption of AI models. The section below describes the computational metrics and explains what influences these metrics.

Metric A: Floating Point Operations

The computational power of computational models is typically measured and reported in the number of floating point operations (FPO), what provides an estimation of the work required to generate a result and can be used to estimate the energy consumption of a model [Howard et al., 2017; Sandler et al., 2018; Schwartz et al., 2019]. A FPO (i.e. Mult-Adds or Madds) is computed analytically by an addition or a multiplication in the model. Advantages of using FPO are (i) that it directly computes the amount of work done by a machine and this can be linked to the amount of energy that is consumed, (ii) it facilitates fair comparison between different approaches, and (iii) it is strongly correlated with the running time of the model.

To determine how much computational power is used by a model at a moment in time, the computational power is expressed per unit of time. Therefore, it is not reported in FPO, but in FLOPS that stands for floating point operations per second, i.e. FPO per second [Sun et al., 2020]. Amodei and Hernandez [2018] expand this measure for AI models to petaFLOPS-day, since AI models can run for days, weeks or months with different peaks in computational demand. Therefore, this metric reports the average number of FLOPS during one day.

There are several methods to determine the amount of FPOs that is consumed by an AI models. Schwartz et al. [2019] describe the first by presenting a method that uses several software packages to compute FPO in various neural network libraries. However, these are not broadly available and none of the packages contain all building blocks required to determine the FPOs of all modern AI models. Amodei and Hernandez [2018] describe two methods to calculate the number of peta-FPOs and petaFLOPS-day. The first method is based on the counted operations in the model and the second on the running time of the Graphical Processing Unit (GPU). Important to note is that these two methods are not intended to be precise, but are meant to be correct within a factor of 2 to 3. Counting the operations in a model is particular easy to apply when the number of operation in a forward pass is provided, it can then be calculated by multiplying the following variables:

- Number of add-multiplies per forward pass
- 2 FLOPS per add-multiply (fixed variable)
- 3 for forward and backward pass (fixed variable)
- Number of examples in data set
- Number of epochs, i.e. number of times the weights of the network are changed [Vijay, 2019].

The amount of add-multiplies per forward pass depend on several characteristics of the AI models as presented by Amodei and Hernandez [2018]. Important characteristics to consider are the architecture of the models, the number of nodes in the hidden layers, and the required accuracy of the model [Lottick et al., 2019; Henderson et al., 2020]. The precise relations between these variables and the add-multiplies per forward pass are not defined. However, Howard et al. [2017] show that the accuracy of a DL model is correlated with the logarithm of the total number of FPOs. So, a higher required accuracy results in an exponential increase in total number of FPOs.

Calculating the number of FPOs based on the GPU time can be done by multiplying the following variables [Amodei and Hernandez, 2018]:

- The number of used GPUs
- The FLOPS per GPU
- Total run-time of the model
- Estimated utilization of the GPUs

However, this method requires a lot of information about the hardware that is used, which is not always fully available. The FLOPS per GPU and the estimated utilization of the GPU are constantly changing during a training run and are hard to fully map [Amodei and Hernandez, 2018]. Moreover, the number of FPOs is not dependent on all variables above, but the variables above are dependent on each other. For example, the number of FPOs does not increase when the utilization increases, instead the run-time probably decreases. Therefore, these variables do not determine the number of FPOs.

To conclude, the following non-exhaustive list of concepts determines the amount of FPOs executed by a considered model :

- Add-multiplies per forward pass
- Architecture of the neural network
- Number of nodes in the hidden layer
- Required accuracy
- Size of the example data set

Metric B: Number of parameters

Howard et al. [2017] and Sandler et al. [2018] express computational power by a combination of the number of parameter and the number of FPOs. In both research projects, the number of parameters was defined by the input image resolution and a width multiplier defined by Howard et al. [2017]. This multiplier thins the network at each layer and reduces the number of hidden layers, to simplify the network and let it consume less computational power while maintaining accuracy. However, Lottick et al. [2019] present that different amounts of parameters for training different types of architectures show different and seemingly unpredictable accuracy. Some architectures increase in accuracy when being trained with more parameters, but most show a varying accuracy when adding parameters.

Therefore, it is hard to compare the computational power of different types of architectures based on solely the number of parameters. However, the section describes that the number of parameters is required to determine the number of add-multiplies per forward pass, which in turn is requires to determine the number of FPOs that are required for training a model. Concluding, the number of parameters on itself is a poor metric to express computational power, but can be used to determine Metric A.

Metric C: CPU/GPU utilization

To compare the performances of different approaches to train AI models, Assran et al. [2019] report the utilization of the GPU and CPU together with the power drawn by the two processing units. The objective of the research project is to optimize the hardware usage. So, the computational power in this research project is expressed in how efficient the model uses the hardware at hand. Dalton et al. [2019] conducted a research with a comparable objective; optimizing the utilization of the GPU. The computational power was therefore also expressed in the utilization of the GPU. The biggest limitation of this metric is the hardware dependency. To compare different models with this metric would require the models to run on the exact same hardware. This is inconvenient for comparing different models with different architectures, since types of models benefit from different hardware [García-Martín et al., 2019].

Related to the CPU or GPU utilization is the metric defined as CPU or GPU-hours. Soboczenski et al. [2018] reported their computational power in CPU or GPU-hours to compare the results of applying different methods. This metric has limitations reporting the computational power of different models due to two reasons. First, like the utilization of CPUs and GPUs this metric is highly dependent on the hardware that the model uses. Second, CPUs and GPUs do not have a linear efficiency, i.e. 10 servers running at 10% do not produce the same computational power as 1 server at 100% [Barroso and Hölzle, 2007]. The metric CPU or GPU-hours does therefore not provide enough information to properly compare different models.

3.2.2 Implications for deep learning

For a variety of fields of study within AI, including the field of DL, there is a distinction between the computational power consumed by inference and training. Training a DL model consists of finding the right weights to the nodes in the neural network and inference is the forward propagation in the network once the weights have converged [Sebastian et al., 2019]. However, as mentioned in the introduction, this report only focuses on the training of DL models. Because, inference is a less demanding in computational power, but harder to track Schwartz et al. [2019].

Tensor Processing Unit

Another implication for DL models is the development of another processing unit, namely the Tensor Processing Unit (TPU). This unit is specially developed for DL models [Google, 2020]. Section 3.2 only describes the utilization of the CPU and GPU, but for some application a TPU can outperform them. Wang et al. [2019] report a speedup between the 3x and 6,8x when using TPU instead of a GPU for commonly used DL models.

However, the TPU is highly optimized for large batches and Convolutional Neural Networks, where the GPU is more flexible and more programmable for irregular and often smaller computations [Wang et al., 2019]. TPUs should be used for models with large to very-large batch sizes that train for weeks or months and do a lot of matrix computations [Google, 2020]. So, it might be possible that the utilization of the TPU should be considered when defining the computational power of DL models, but TPUs are beneficial for only a particular type of DL models and therefore used less.

Field Programmable Gate Arrays

Besides TPUs, Field Programmable Gate Arrays (FPGA) offer a solution to speed up DL training, when comparing to GPUs [Hwang, 2018]. The difference between FPGAs and CPUs or GPUs is that FPGAs do not run programs in stored memories. They are a collection of connected logic blocks that can be adjusted by a programmer [Xilinx, 2020]. This enables FPGAs to be 2.3x to 3x faster than GPUs in training DL models, which is not faster than the TPUs [Simon, 2017; Wang et al., 2019]. Moreover, the high level of customizability makes it hard to compare and evaluate different neural network architectures with different FPGAs [Hwang, 2018]. Finally, FPGAs are less programmable, since the FPGA are customized for specific operations [Hwang, 2018]. This makes them less attractive for data centers or service providers. Therefore, FPGAs are less applied for modern DL training and are TPUs preferred. FPGAs are therefore not taken into account in this research project.

3.3 SECTIONAL CONCLUSION

This section aims to answer the first sub-research question as mentioned in section 2.2. This question is:

What metrics can be used to define computational power of Deep Learning models?

To fully understand the concept of DL, the first part of this chapter describes the difference between AI and DL. AI can be summarized by a set of capabilities systems can have that simulate human skills. DL is a field of study within AI, which imitates the human brain by simulating a neural network. This technique enables computational models to discover complex structures in data.

As mentioned in section 3.2.1, training DL model requires a vast amount of computational power with different peaks during a run-time. Therefore, it is less convenient to report the number of FPOs required for a result, as it gives no indication the hardware required to process the peak demands. A more convenient metric would be FLOPS-day in combination with the total run-time, since it provides insight in the computational power demand at a moment in time and the total computational power demand to acquire a result. Also, CPU, GPU, and TPU utilization could be a metric to express the computational power of DL models combined with the run-time of the training, to provide information about the capacity of the hardware.

To conclude, there is no one straightforward metric that completely defines the computational power of DL models. It might therefore be useful to report on FLOPS as well as on GPU utilization for a full image of the model training. However, FLOPS are hard to retrieve when programming, so the utilization of the processing unit(s) is the preferred metric.

4 ENERGY CONSUMPTION OF DEEP LEARNING IN DATA CENTERS

To determine the energy consumption of DL models, literature provides different methods. The corresponding sub-research question is: *What methods are available to account the energy consumption of training Deep Learning models?* The first part of this chapter will describe different methods for accounting the energy consumption of DL models. The second part describes complications that arise when accounting the energy for models that run in data centers.

4.1 ENERGY ANALYSIS METHODS

To compare the energy consumption of different products and services, Blok and Nieuwlaar [2021] present the life-cycle energy analysis. They make a distinction between the energy that is directly consumed in the process and energy that is consumed before the process. For example, in case of a server they make a distinction between the energy that is consumed by the server and the energy it cost to produce the server. Dayarathna et al. [2016] concluded that the energy consumption of producing IT-equipment can be neglected when comparing it to the total energy consumption of the IT-equipment's life-span. So, this research project focuses on the energy that is consumed by the IT-equipment.

Blok and Nieuwlaar [2021] propose four methods to determine the direct energy consumption of various processes. These methods are:

- It can be derived from data provided by companies;
- It can be derived from statistical data;
- It can be calculated based on the equipment that is used for the process and the data of this equipment;
- It can be measured directly.

The first and the second methods are hard to execute, since there is a lack of data and a lack of modelling capabilities for energy analysis of data components [Lei, 2020]. IT equipment and data center cooling and power provisioning infrastructure are not adopted consistently or in high dimensions in current energy analysis models. However, calculating the energy consumption based on the equipment and measuring the energy consumption directly is possible. Calculating the energy consumption can be done afterwards with retrieved information or estimations from the training runs, but measurement is only possible while training.

4.1.1 Literature selection

To investigate the different methods that are available for calculating and measuring the energy consumption of training DL models a literature reviews is conducted. The literature used in this review derives from relevant papers that tried to account the energy consumption of DL models. The papers provide methods with different level of details and therefore different levels of accuracy. Section presents and describes two methods with different levels of detail to calculate the energy consumption of training DL models. Section 4.3 presents and describes two related methods to measure the energy consumption from interfaces of the hardware.

4.2 CALCULATING ENERGY CONSUMPTION

Calculating the energy consumption is the first method to determine the energy consumption of DL models. This section presents two methods from literature. The first focuses on how software programs consume energy and the second focuses on the peak performance of the main processing unit.

4.2.1 Method A: Software energy consumption

García-Martín et al. [2019] summarize a general method to determine the energy consumption of software and this section presents this method. To determine the energy consumption of the software, they make a difference between the static power and the dynamic power. The static power, i.e. the leakage power, is the power consumed when there is no software activity on a circuit. The dynamic power is the power consumed by a circuit when the software is active. The dynamic power can be calculated with the formula:

$$P_{dynamic} = \alpha \times C \times V_{dd}^2 \times f \tag{4.1}$$

With dynamic power ($\mathrm{P}_{\mathrm{dynamic}}$) in Watt, the activity factor (alpha) as representing percentage of the hardware that is active, the capacitance of the capacitor (C) in Farad, the voltage (V_{dd}) in Volt, and the clock frequency (f) in Hertz. Then, to calculate the energy consumption of the dynamic power of the circuit, the integral of the dynamic power over a period of time can be calculated. This results in the energy to perform a task in joules (J). This is considered to be the main variable, since it relates directly to money spend on computations.

Besides the energy consumption of the circuit it is also possible to determine the energy consumption of a program. To determine this, first the total execution time of a program needs to be calculated using the following formula:

$$T_{execution} = IC \times CPI \times T_c \tag{4.2}$$

With the number of instructions (IC), the average number of clock cycles per instruction (CPI), and the machine cycle time (T_c). The total energy consumption of a program can be calculated with the formula:

$$E = IC \times CPI \times EPC \tag{4.3}$$

With the energy per clock in EPC and EPC is a proportional constant of C X $\mathrm{V}^2{}_{\mathrm{dd}}$ in formula 4.1. This means that the discharging and charging of the capacitor on the circuit is not equal to but in proportion with the energy per clock.

This method gives an oversight of the variables and calculations that need to be available and executed to determine the energy consumption of a program. So, this could also be used to calculate the energy consumption of a DL model. However, this is also a limitation of the method, since it would require a lot of detailed information about the hardware and the instructions of the programs. Moreover, García-Martín et al. [2019] explain that measuring the execution time does not give a realistic view of the energy consumption, since some instructions take more time to execute and other consume more computational power, i.e. more FPOs.

4.2.2 Method B: Peak performance of main processing unit

Lacoste et al. [2019] reported a method to determine the approximate CO2 footprint of training or running a model, with input variables the run-time, the type of GPU used, and geographical zone of the server. Beside these input variables, they collected public information available on the energy consumption of the hardware, the location of the providers' region of compute, the region's CO2 equivalent emissions per kWh, and the potential offset bought by the provider.

As part of calculating the carbon footprint, the method of Lacoste et al. [2019] can be used to calculate the energy consumed by the model in kWh. This calculation is based on run-time of the model and the processing power of the selected processing unit as input variable. The processing power is based on the theoretical peak performance of the processing unit, considering the Thermal Design Power (TDP). The TDP is the highest power consumption of the processing unit that can sustain over a longer period, without damaging the CPU or GPU [Intel, 2007]. It is not the absolute maximum power of the hardware, since processor units can have a higher peak performance during a short period of time. Lacoste et al. [2019] used a basic formula to calculate the energy consumption of a model, namely:

$$E_{total} = \frac{P \times T_{run}}{1000} \tag{4.4}$$

With total energy ($\mathrm{E_{total}}$) in kWh, Power (P) in Watt considering the TDP of the main processing unit, and the run time ($\mathrm{T_{run}}$) in hours.

However, formula 4.4 only provides an estimation of the energy consumed by a DL model. Lacoste et al. [2019] state themselves that the method is only a starting point, with the main limitation that the estimation is based on the theoretical peak performance of the processor units. While in reality the performance of the processor units varies during a model run [Amodei and Hernandez, 2018]. Another limitation is the assumption that energy is only consumed by the hardware to compute, while in reality a significant part of the consumed energy is converted to waste heat [Strubell et al., 2019]. This is expressed in the Power Usage Effectiveness (PUE), which represents the total energy consumption divided by the energy consumed of the IT-equipment.

4.3 MEASURING POWER DRAW OF HARDWARE

In the section above, Lacoste et al. [2019] only considers the energy consumption of the main processing unit at its theoretical peak performance. This is because GPUs can provide more computational power than CPUs and most DL models rely mainly on GPUs for fast training [Li et al., 2016]. However, Li et al. [2016] reported that even in idle state the CPUs consume a significant amount of power. In general, 22% to 40% of total energy consumption is used by the CPU. Besides, the Dynamic Random-Access Memory (DRAM) consumes a relatively small amount of energy (11%) when a DL model is based on CPU-framework. Therefore, the GPU, CPU, and DRAM need to be considered when determining the energy consumption of trainingDL models in more detail. The measurement methods provided by literature also take these hardware components into account.

4.3.1 Method C-1: Hardware power draw

Strubell et al. [2019] and Lottick et al. [2019] present similar methods to calculate the energy consumption of DL models with the power related information of the CPU and GPU. Strubell et al. [2019] aim to quantify the approximate financial and environmental costs of successful DL models for NLP and to reduce these costs. Lottick et al. [2019] aim to provide individual computer science researchers with industrial level analyses on measuring the energy consumption and carbon footprint of their model use.

Both research projects use the Running Average Power Limit (RAPL) interface of Intel to calculate the average power consumption of the CPU and DRAM. To sample the average GPU's power draw, both research projects use the NVIDIA Systems Management Interface (SMI). The sum of these power draws is then multiplied by the running time of the model to derive the model energy consumption. Also, both take into account the PUE of the power draw of the hardware. The formula for this methods is formulated as:

$$E_{total} = PUE\frac{T_{run}(P_{gpu} + P_{cpu} + P_{dram})}{1000} \tag{4.5}$$

With total energy ($\mathrm{E_{total}}$) in kWh, the run time ($\mathrm{T_{run}}$) in hours, and the average Power ($\mathrm{P_{resource}}$) in Watt per hardware resource. The PUE is dimensionless, but Strubell et al. [2019] consider a coefficient of 1,58 and Lottick et al. [2019] of 1,25.

Wolff Anthony et al. [2020] use a similar method to determine the energy consumption. They split the training time of the model into specific time block, to determine the actual carbon intensity with the time block. Such a time block is called an epoch. The given formula by Wolff Anthony et al. [2020] is presented below. Note that the formula is very similar to formula 4.5.

$$E_{total} = PUE \sum_{epoch} \sum_{device} \frac{P_{device} T_{epoch}}{1000} \tag{4.6}$$

With total energy ($\mathrm{E_{total}}$) in kWh, the run time per epoch ($\mathrm{T_{epoch}}$) in hours, and the average Power per device ($\mathrm{P_{device}}$) in Watt per hardware resource. The PUE is ddimensionless and assumed to be 1,58, like Strubell et al. [2019].

One difference between two research projects is that Lottick et al. [2019] differentiate between total power draw and the extra power draw to process a model. This is determined with the average base line power of the hardware, i.e. the idle power draw. The average power to process a model is therefore defined as *Average total power - Average baseline power*.

The method described above has several limitations. First of all, the interfaces that are used to determine the power draw of the different hardware components are only available for two manufactures, namely Intel and NVIDIA. Both are considered to be dominant players in the market for respectively CPUs and GPUs, but there are many more and especially when considering customized hardware [Hwang, 2018]. The second limitation is the required access to the hardware's interfaces, which is relatively easy to access if owned but might face difficulties if not. Finally, this method is limited as it only considers the average power draw of the hardware over a period of time. Although Lottick et al. [2019] pay attention to the idle power consumption, both research project do not touch upon the utilization of the hardware.

4.3.2 Method C-2: Utilization of the hardware per process

Henderson et al. [2020] present a method that not only take into account the different hardware components, but also the utilization of the hardware per process. They present this method and a related framework to provide a simple interface for tracking real-time energy consumption and carbon emissions. The goal of their paper is to propose strategies to mitigate carbon emissions and reduce energy consumption. Similar to the previous approach by Strubell et al. [2019] and Lottick et al. [2019], this approach uses the RAPL interface of Intel and NVIDIA's SMI to retrieve information about the power draw of the GPU, CPU, and DRAM.

The difference between this research projects and the previous is what information is retrieved from the interfaces. Henderson et al. [2020] track the energy consumption based on the utilization of the CPU and GPU per process. The energy consumption per process is determined by the average power draw of the hardware components, multiplied by the run-time of a process and a percentage for the utilization of the hardware that was dedicated to this process. Next, the energy consumption of all processes are added and multiplied by the PUE. The formula is presented below.

$$e_{total} = PUE \sum_{p} \frac{T_p(U_{gpu}P_{gpu} + U_{cpu}P_{cpu} + U_{dram}P_{dram})}{1000} \tag{4.7}$$

With total energy (e_{total}) in kWh, the utilization of the resource per process ($U_{resource}$) in percentages, the average power draw of the resource ($P_{resource}$) in Watt, and the running time of the process (T_p) in hours. Henderson et al. [2020] assume the same coefficient for PUE as Strubell et al. [2019], namely 1,58.

The limitations of this method is partly similar to the previous method, since this method is also dependent on Intel and NVIDIA and it requires full access to the hardware. Another limitation is the assumption that the computational power of the hardware increases linearly with the energy consumption of the hardware. Barroso and Hölzle [2007] reported that hardware becomes more efficient when approaching the maximum computational power.

4.4 METHOD OVERVIEW

The previous sections describe various strengths and limitations of accounting the energy consumption for training DL models. Here, these considerations are summarized in the table below. The sections above show a division between the calculation approach (Method A & B) and the measurement approach (Method C-1 and C-2). This division of approaches corresponds with the distinction of techniques made by García-Martín et al. [2019], named respectively the simulation technique and the performance monitoring counters (PMC) technique. The table below presents an overview of the methods from the literature and the strengths and limitations of the four methods.

Table 4.1: The methods described by literature to account the energy of training DL models

Method	Literature	Characteristics
A: Software energy consumption	García-Martín et al. [2019]	**Strengths** • Detailed calculation of different parts of the hardware **Limitations** • The required information is very detailed; • The measuring of some variables is not realistic
B: Peak performance main processor	Lacoste et al. [2019]	**Strengths** • Only basic information required about the hardware and training runs **Limitations** • Strong overestimation of the main processing unit; • Neglection of varying processor performances; • Only main processing unit accounted; • Idle power is not accounted

C-1 Hardware power draw	Lottick et al. [2019];
	Strubell et al. [2019];
	Wolff Anthony et al. [2020]
		Strengths
		• Multiple processing units are accounted;
		• the idle power can be accounted
		Limitations
		• No retrospective measurement;
		• Access to hardware interface is required

C-2 Utilization per process	Henderson et al. [2020]	**Strengths** • Multiple processing units are accounted; • The utilization of the hardware components per process can be accounted **Limitations** • No Retrospective measurement • Access to hardware interface is required; • Idle power cannot be accounted

4.5 SECTIONAL CONCLUSION

This section aims to answer the second sub-research question as mentioned in Section 2.2. This question is:

What methods are available to account the energy consumption of training Deep Learning models?

This chapter presents the four methods for energy accounting that are identified in literature. The first method is a detailed calculation about the energy consumption of programs on circuits. The second is an estimated calculation that only requires basic input data about the hardware and the training. The third method is measuring the power draw of the different processing units. The fourth is measuring the utilization of different processing units per process.

The two methods that measure the energy consumption are very detailed and complete approaches, but are not applicable after the model has been trained. The Method A also calculates the energy consumption in detail, but proved to be unrealistic and requires a lot of information about the hardware and programs. So, it is most realistic to use Method B to determine the energy consumption of DL models, although this only provides an estimation.

But, the method as presented in Formula 4.4 can be extended by adding the PUE. Since this value might be unknown the default value of the PUE is 1,58, like Strubell et al. [2019] and Henderson et al. [2020] assume. The correlating formula is:

$$e_{total} = PUE \frac{P \times T_{run}}{1000} \tag{4.8}$$

With total energy (e_{total}) in kWh, Power (P) in maximum Watt considering the TDP of the main processing unit, PUE dimensionless with a default of 1,58, and the run time (T_{run}) in hours.

5 MULTIPLE CASE STUDY ANALYSIS

This chapter aims to answers the following sub-research question: *To what extent can the energy consumption of training Deep Learning models be accounted in practice?* To answer this question, the first section describes the different cases for which this question is considered, the model that is used in the case, and additional remarks made by the interviewees. The next section analyses the cases and presents the information that is available to calculate the energy consumption of DL models.

5.1 CASE REPORTS

This section describes each case study individually. The description of the case study is structured based on the semi-structured questions in Appendix A. So, the case studies are divided into the case and model description. The description below are based on the interviews that are transcribed in Appendix E.

5.1.1 Case study 1: Situation recognition in slaughterhouse

The description below is based on the interviews in section E.1. The two interviewees have the roles of project manager and model developer.

Case description

The goal of the Deloitte project was to improve the animal welfare in slaughterhouses in the Netherlands. The project aimed to improve animal welfare by analyzing camera footage from the slaughterhouse and to train a model to recognize undesired situations. These situations where then viewed by employees, who then had to decide who did what wrong.

The project originated from one of the partners at Deloitte, who was the client of the product. The old process was inefficient, since employees had to watch random and often useless camera footage. The first application of the model was in a pig slaughterhouse and the client was an alliance between the slaughterhouse and an animal welfare organization with a common goal increase the animal welfare.

Model description

The goal of the model itself was to identify undesirable situations in slaughterhouses. To achieve this goal, the model consisted of three layers. The first layer focused on the general identification of objects on the camera footage. This was partly done with pre-trained models, but needed additional training. This layer is mainly used to distinguish what is useful and to identify animals from other objects in the footage. The second layer was used to identify what were undesired situations. So, what have to happen with the identified objects to be labeled as an undesirable situation. For example, when one of the objects, in this case pigs, was left behind from other objects. The final layer was used to upload the footage of the undesired situation to a platform for employees to check it.

To find the best architecture for the model, the performance of different architectures was evaluated based on the mean average precision with intersection over union evaluation with a threshold of 0.5. So, from the predicted box by the model, at least half of the predicted box had to overlap with actual footage of the object. To find the best performing architecture, 6 or 7 different architectures where tested with each 5 or 6 different configurations. This means that in total 30 to 42 different configurations were tested and each configuration ran until a number of configurations were completed, what took between the 12 and 16 hours. Eventually, Faster R-CNN ResNet101 proved to be the best performing architecture.

These configurations were all executed by the service provider Azure of Microsoft in a virtual instance, named Standard_NC6. The hardware underneath this virtual machine was half a NVIDIA Tesla K80, which contains 2 GPUs. The selected region for the this virtual instance was West-Europe and for Azure this is in Belgium. But there was no information available about the PUE of the data center.

Logically, most information about the model could be found at the technical model developer. The project manager had basic information about the model and described its process as: ”The model has been thought to recognize objects on image and to distinguish what is happening on image. Then it has been taught what is an good situation and what is potentially a wrong situation.” All other information about the model was provided by the model developer.

Additional remarks

The project is perceived by the model developer to be efficient, since the cost for training the model was already minimized and therefore is the GPU not excessively used. The virtual machine was not constantly occupied, so the GPU was only reserved during training. Also, it is stated by the developer that information from the data center or service provider about the energy consumption would not have reduced the energy consumption. He stated that all configuration runs were necessary for the final result. Finally, a structure is proposed to compensate for the carbon footprint of the models and charge these costs to the customer to increase awareness among clients and modellers.

5.1.2 Case study 2: Bank's Credit

The description below is based on the interviews in Appendix E.2. This project contained sensitive information, therefore not all details are provided. The interviewees had the roles of IT architect and model developer.

Case description

The goal of the project is for a bank to offer a proposition for a loan to the client within 10-20 minutes that is a good and appropriate offer. Normally, clients would have to apply for a loan at a bank, the bank would have to take a look at the application and would do a proposition for a loan. This process would take some days.

This project originated from a changing need of the client that the bank identified. Clients wanted to receive the information about their loan faster. Speed was important for the project, since clients want clarity about their credit rapidly. Also, the information a client had to feed to the model needed to be hands-on available. So, basic information about who you are and the bank should be able to fill in the rest with your permission.

Model description

The goal of the model itself was to determine whether a loan could be offered to the client and if so, what kind of loan and how much can be loaned. The input data were long series of millions of historic loan offers of many years. To offer the loan, the model used two applications of DL. These applications are NLP and XG Boost, the latter is a sub-form of random forest that enables the model to learn and improve its own performance. Advantages of using XG boost are its speed, flexibility for input data, possibility to customize, and good performance [Chen and He, 2021].

The model was evaluated based on technical and non-technical key performance indicators. The most important evaluation parameter for the model was again the speed, since the result was requested within nanoseconds. Moreover, the technical operation can be evaluated based on several parameters, e.g. distribution of false-positive. However, the model should not be over-fitted to the data, since it is costly in training and an undesired outcome. But to determine what is sufficient, more non-technical people need to be involved to judge the outcomes. Also, in the market of banks economic parameters are used to determine whether outcomes are representative for an economy. Such less model-technical parameters and the judgement of experts then determine a threshold for the technical parameters.

To eventually reach the determined threshold, the model trained for 5 to 10 weeks. What architecture was used for the model cannot be elaborated on, but there are only a few tested to find the best performing. The architecture was chosen before training. The model is trained on internal servers with the use of an intermediate platform and what hardware is used can be guessed, but not said with complete certainty. The reason for internal training was confidentiality.

Additional remarks

Due to the nature of the bank's market the model required a high level of explainability, which is extraordinary for DL models. This explainability was required to justify the model towards regulators of the bank and authorities. Eventually, they succeeded to trace changes in outcomes to a small set of parameters.

Energy wise, the training is expected to be inefficient, since the energy consumption is not taken into account in the evaluation. Also, taking into account the energy consumption might limit the potential of AI and DL. The focus is still much more on can we do it. Often when something seems impossible, people want to use DL. But it turns out to complex to create a good model, with data of high quality that is representative for reality. People first need to trust the technique before next steps such as the energy consumption are considered. Nevertheless, energy accounting is perceived as very relevant and interesting to present to the bank.

Another remark is that there are several movements in the world of DL models, especially when looking at innovation. One focuses on personal data rich models to identify patterns. Another focuses on privacy protection of personal data. However, the sustainability story linked to the Paris Agreement is one that nobody is paying attention to and nobody is linking sustainability and data yet. This is perceived as unique in the world of DL and very relevant for clients.

5.1.3 Case study 3: Ericsson Product information assistant

The description below is based on the interview in Appendix E.3. The interviewee was one of the model developers.

Case description

The goal of the Ericsson project was to develop a tool that provides all information for field engineers that have to do repairs on inconvenient and dangerous places. Normally, field engineers would climb up a mast for installation, troubleshooting, exchanging, or upgrading of software or hardware and were in need of some additional information. Before, it was a time consuming activity to climb down, so this project started. The discussion about the project started 5 years ago and back then were only high hopes about the technology. Only in recent years the project really flourished due to new techniques. Especially the speech-to-text and text-to-speech models of Google were important developments.

Model description

The goal of the model itself was to provide, information for the field engineers. So, a knowledge based model was constructed to structure the distributed data. For this Knowledge Base model are newer DL models used, but cannot be elaborated on. Models similar to Transformer, XLNet, and BERT. Questions could be linked to text in documents. To handle the dialogue pipeline, a classifier was trained on publicly available conversational data, as well as their own conversational data. This enables the model to follow the steps, be flexible, and react on the user if it states things in differently. Different models were used to test this process. Eventually one was chosen that seemed to perform best, with a test data set of 20 to 50 different sections and a few questions.

The training time is perceived as little. Small parts of the training were partly executed on laptops. Bigger parts were trained on internal servers of Ericsson. There is no clear overview of all training. A virtual machine was set-up and allocated to the team within an internal cloud of Ericsson, but the data about hardware underneath the virtual machine is unavailable. The energy consumption is not accessible, but there is some control on it since teams only have limited access to computational power, what is linked to the energy consumption. So, there is some control on the energy that is consumed by the models, but no direct monitoring. However, the trained models are no huge models, e.g. no full transformer model with billions of data points.

Additional remarks

First, it is stressed that energy consumption is normally a requirement for all Ericsson projects, but it has no special priority in training of models. Second, it was important for Ericsson to keep the model on internal servers, since the data that was used could not leave the premises. Since this information is classified.

5.1.4 Case study 4: Asphalt damage detection

The descriptions are based on the interview in Appendix E.4. The interviewee was one of the model developer.

Case description

The goal of the Arcadis project was to automate the damage detection of the asphalt of Dutch roads. Normally, road inspectors had to watch video of roads and had to visit locations to inspect the asphalt. This could be dangerous for high-speed roads, labor-intensive, and subjective for different inspectors, since it can be hard to detect small damages. The output of the project for the clients is a map that indicates the road quality per segment of road. This project orginitated from 2018 and started internally at Arcadis, since they did jobs for various clients on roads and noticed the problems stated above and the fact that road inspectors already used video footage. Image and video recognition were upcoming by then and it was easier to use on-location supercomputers over cloud.

Model description

The goal of the model was to identify damages in the asphalt of roads, based on video footage of vehicles driven over these roads and filming the road. The model consists of a pipeline that loads the data of the client into the model. Next, the image recognition model identifies the damages in the asphalt. These damages then need to be translated into the standard and methodology of the client. Finally, the status of the roads need to be visualized on a maps per segment of road.

To find the best architecture, it is estimated that a little under 100 configurations with different architectures were tested. Eventually is decided to use a Tensorflow Mask RCNN. Every configuration ran about 2,5 day and 175.000 steps. To evaluate what architecture and configuration worked best, the configurations were tested with a validation set of images of different roads and different damages. This set was also evaluated by different road inspectors, to erase the subjectivity between them. These configurations were all trained on one of the two supercomputers located at Arcadis. Both have two GPUs, which is the GeForce RTX 2080. So, in total there were 4 of these GPUs available. The model training only used 1 of these.

Additional remarks

First, the supercomputers used for the model training might consume a lot of energy, but Arcadis almost certainly compensates the produced carbon of the building. Second, they chose a on-location supercomputer over a cloud based service, since it was easier to realise in 2018. But, they now work partly in the cloud or at least its in the test phase. Thirdly, there is a pretty good understanding of how the energy consumption of the model training can be accounted. The guess is to retrieve data about the server and combine it with run-time of the model.

5.2 CROSS-CASE ANALYSIS

This section analyses the cases and compares the cases with each other. The first section compares the information that is available in the different cases for energy accounting. The second section implements the energy accounting method that is available with the available information. The third section describes the limitations and restrictions that occur from this method.

5.2.1 Cross-case overview

The table on the next page presents an overview of the available information per case about the project, the model, and the information that can be used for energy accounting.

Table 5.1: Overview of the information available in the different case studies.

	Case study 1: **Slaughtery**	**Case study 2:** **Bank's credit**	**Case study 3:** **EPIA**	**Case study 4:** **Asphalt damages**
Purpose of the model	Recognize animal unfriendly situations	Provide a fast loan offer to clients	Provide information for engineers & interact with them	Recognize damages in road asphalt with more accuracy and more consistent than road inspectors
Input data	Camera footage	Long series of historic loan offers	Text documents and instructions of different products	Video-footage of roads
Process to decide on architecture	6 to 7 architectures were tested with 5 to 6 different configurations per architecture	From the beginning one architecture is chosen and built	Several neural networks were tested, but not one that tremendously outperformed the others	Tested little under 100 configurations
Decided architecture	Faster R-CNN ResNet101	Known, but confidential	Known, but confidential	Tensorflow masked Faster RCNN
Deep learning applications	Image/video recognition	Knowledge Base model with NLP	Knowledge Base model with NLP	Image/video recognition
Training time	12 to 16 hours per configuration, so 360 to 672 hours	5 to 10 weeks	Few days of training	2,5 days per configuration, so little under 250 days.
Service provider	Azure	Internal	Internal	Internal
Hardware	Half a NVIDIA Telsa K80	Unknown	Uknown, but can be retrieved	GeForce RTX 2080
Location of hardware	Data center in West Europe	At the Bank	At Ericsson	At Arcadis
Energy consumption of the model	Uknown, but run-time as the best proxy to determine energy consumption	Uknown, but perceived to be inefficient as it has no priority	Uknown, but limited amount of computational power is allocated per team	Unknown, but run-time as the best proxy to determine energy consumption combined with GPU details

5.2.2 Remarks about the case studies

Below are the most important remarks listed from the table above with the overview of the case studies.

Division in type of cases

It can be concluded that the cases roughly consist of two types of DL models, namely Image/video recognition (Cases 1 & 4) and a Knowledge Base model that uses NLP (Cases 2 & 3). When continuing this division of case studies, it is noteworthy that the Knowledge Base models consist of more confidential and unknown information. An explanation for this separation in confidentiality could be that the input data of the Knowledge Base models is more confidential. This input is in case 2 loan offers of previous bank customers and in case 3 detailed documents and instructions of Ericsson products.

Ambiguous process to final architecture

In 3 of the 4 cases, the process to decide on the final architecture was ambiguous and not well documented. Often, model developers just tried several possibilities that might work with the given input data and desired outcomes. Therefore, the number of tested architectures in the three case studies are only an estimation Besides, it is possible that in case study 2 smaller models were tested, but were not named in the interviews as the training time was only a fraction of the final training time. Also, the final architecture was often chosen on best performance, however in none of the project was a notion of the consideration that an architecture could be good enough for a given purpose. Defining this 'good enough' performance could significantly reduce the number of tested architectures.

Internal over service provider

A third remark is that most (3 of the 4) cases still use internal servers over a service provider. The two Knowledge Base projects explained that the reason for internal servers was confidentiality of the input data and project. The reason for the third case study was the practicality of a internal server over service provider, since they had more experience on the internal server .

Little hardware knowledge

Fourth, in two of the four project the used GPU was unknown and in all cases it was unknown what the CPU and DRAM were. Also, only case study 1 had some details about the hardware locations, since they used a service provider to train the DL models. For the other three case studies, the location of the hardware was intern, but there was no information available about the energy efficiency of the hardware location.

Unknown energy consumption

Finally, none of the case studies had numbers about the energy consumption of the DL models or the used hardware. Two case studies (Cases 1 & 3) could determine the run-time as a proxy for the energy consumption. At case 2 the main remark was that the training was probably inefficient, but had no knowledge on how to improve the efficiency. Case 4 was the only to mention that the energy consumption could be determined based on the run-time of the model and the details of the GPU.

5.3 ENERGY ACCOUNTING OF THE CASES

This section strives to determine the energy consumption of the case studies for which at least some information was available. Important to note is that the calculations are based on the limited available information and only based on the energy use of the GPU. Therefore it would be more accurate to state that the calculation is an estimation of the energy use of the GPU. The first section below describes the calculations that are executed and the second section discusses the implications of the outcomes.

5.3.1 Calculations

In Section 4.5 is formula 4.8 presented to calculate the energy consumption of the peak performance of the main processing unit. This formula requires the PUE, training time, and type of hardware. The former is not required from the case studies, since literature provides an approximate default value and the latter can be used to search the peak performance online. With the approximate numbers from the case studies it is possible to calculate an estimation of the energy consumption of case studies 1 and 4. The calculation is executed in Appendix B.

Appendix B elaborates on the calculations with the available knowledge and some additional sources to retrieve data about the processing units and the data center. Case studies 1 and 4 consumed an estimated energy of respectively between 122 to 226 kWh and 2.038 kWh. This is the average yearly energy consumption of respectively one LED-television and 8,5 fridges.

5.3.2 Implications

The energy consumption of the case studies are relatively low compared to the research of Strubell et al. [2019]. However, this does not mean it is not relevant to reconsider. Because, as mentioned before, these numbers are a strong underestimation of the actual consumed energy. Moreover, although the energy consumption of the case studies is less than expected, the case studies show that energy consumption is not considered in any of the cases and therefore can be assumed that the energy consumption could be reduced. This energy saving might be low on one case study, but when considering that there are many similar cases, this saving can become a significant share.

5.4 STAKEHOLDER IDENTIFICATION

This section provides an overview of the different stakeholders that are identified and their current role in accounting the energy consumption of DL models. The stakeholders are identified in the case study interviews and the in-dept expert interviews, see Appendices E and F. Important to note is that the overview of stakeholders is not a complete overview, but merely those identified in the interviews. The stakeholder with an asterisk (*) are stakeholders that are not questioned personally, but are discussed by other stakeholders.

Table 5.2: Overview of the identified stakeholders in the interviews.

Stakeholder	Current role
Client	Little to no attention to minimizing energy consumption of their project.
Customer*	Little to no attention about the energy consumption of the services they consume.
Model developer	Little attention or incentive to minimize the energy consumption, except for cost for compute.
Service providers	Double incentive to maximize rented server-hours and to maximize the usage of the servers.
Data centers*	Maxime the number of servers to rent within the limited power supply.
Ministry of Economic Affairs and Climate	Focussed on minimizing energy leakage in data centers.
Frans Timmermans' council of European Green Deal	Focussed on making data centers as sustainable as possible and finding ways to limit the energy consumption of the IT-sector.
Machine learning scientific community	Finding ways to increase efficiency of machine learning and to account energy based on variables that are unavailable in the case studies.

The current roles of the identified stakeholders reveal the lack of focus on the energy consumption of training DL models. The machine learning scientific community is the only stakeholder who is actively increasing efficiency of ML and has knowledge on how to account the energy. Overall it shows that most stakeholders are not fulfilling an active role in accounting the energy consumption of training DL models.

Customers and client could create an incentive by adding it to the requirements for services and product, but they have little attention to it. Model developers, service providers and data centers could create some awareness by adding the energy consumption as an additional service. But, the model developers have no incentive since it is not a requirement by the client and the service providers and data centers have an opposite incentive to maximize the rented server-hours and not publish the . And, governmental agencies could create an incentive by setting up specific policies regarding the publication of the energy consumption of services. However, these agencies mainly focus on policies to the energy consumption of the hardware rather than the software that runs on the hardware.

5.5 SECTIONAL CONCLUSION

This section aims to answer the second sub-research question as mentioned in Section 2.2. This question is:

To what extent can the energy consumption of training Deep Learning models be accounted in practice?

This chapter presents the different projects, which are used to explore the information that is available to account the energy consumption of DL models. The case studies present how limited the available information is. The energy consumption can be determined for only two of the four case studies, since crucial information is unknown or confidential. For the projects that have the required information, the energy consumption can only be estimated based on the peak performance of the GPUs. The estimated energy consumption of the case studies was relatively low, but is still considerable since the consumption is a strong underestimation and there is a big saving potential.

Finally, this chapter describes the stakeholders that are identified in the interviews with their current roles. These roles present the lack of awareness among the stakeholders about the energy consumption of DL models and the lack of incentive these stakeholders have to account the energy consumption of training DL models.

6 IDENTIFYING CASE STUDY RESTRICTIONS

This chapter aims to answers the following sub-research question: *What are the restrictions that limit the energy accounting for Deep Learning models?* To answer this question this chapter first discusses the research concerns and theoretical framework that are necessary to understand the identification of the restrictions. Second, this section presents the restrictions that are identified in the case studies. Third, the restrictions are categorized based on the causes and possible solution directions. Finally, this chapter analyses how the restrictions are divided in the different case studies.

6.1 CODING PREPARATIONS

As stated in Section 2.1.2, the researcher needs to explicitly state his or her research concerns and theoretical framework before coding the interviews. Auerbach and Silverstein [2003] explain the research concern as what it is that you want to learn about and why you want to learn it. Stating the research concerns helps the reader understand the coding of the text and for the researcher to keep the same focus when evaluating the different interviews. The research concern of this coding process is:

The factors that restrict energy accounting of training deep learning models, to eventually lift these restrictions and reduce the energy consumption of these models.

Stating the theoretical framework explains the subjectivity of the researcher to make the coding process as a whole more objective [Auerbach and Silverstein, 2003]. Because, explaining the subjectivity of the researcher provides an understanding of the decision to code certain parts of the text and this understanding provides other researchers the opportunity to evaluate these coding decisions. This eventually strengthens the objectivity of the process. The theoretical framework is explained as:

The set of beliefs about processes with which you approach your research study.

The theoretical framework of this research project is based on the literature and the outcomes of the case study analyses of respectively Chapters 4 and 5. Literature showed that methods exist to account the energy consumption, but the case studies proved that there is little to no information available to apply these methods. Also, the identified stakeholders and their roles in the previous chapter present that most stakeholders have no active role in the energy accounting of training DL models.

So, the belief to approach this research study is that there is a lack of information in these case studies and the stakeholders do not have an active role in the energy accounting. By coding the interviews, qualitative data analysis is used to identify the restrictions that are at the roots of these beliefs.

6.2 CASE STUDY RESTRICTIONS

This section presents the restrictions that are identified by coding the case study interviews. This section discusses nine restrictions individually that are grouped from a long list of restrictions with examples from the interviews.

The coding process was executed per case study with the research concern in mind, as Auerbach and Silverstein [2003] suggest. First, all possible restrictions in the transcribed text were highlighted in the interview(s). Next, the top 'blank' restriction was labelled with a code. Then, all blank restrictions were compared whether they could be labelled with the same code. This process repeated until all codes were labelled. Then, all the codes were grouped into the restrictions discussed in next section. Appendix C provides a full overview of all coded restrictions that are underlying the grouped restrictions described in this chapter.

6.2.1 Explanation of the identified restrictions

This section discusses the restrictions that are identified in the case studies in alphabetical order. The list of these restriction is:

- Complexity of Deep learning
- Innovative stage of Deep learning
- Lack of incentive to determine energy consumption
- Lack of societal awareness
- Lack of systematic evaluation of models
- Long and diverse training time
- Model developers' energy accounting knowledge
- No hardware details available
- Updating model over time

Below is per restriction explained what the restrictions are and from what aspects they are constructed. Also, each restriction has two quotes and these quotes elaborate on what factors have caused these restrictions to appear in the case studies.

Complexity of Deep learning

DL is a relatively new and complex technique to solve problems, and therefore not always fully understood. Different DL architectures show varying behaviour with different configurations when training the models. Also the combination of architecture and hardware can influence the efficiency of the training. This is often not well understood by the model developers and therefore, many different architectures and configurations are tested before the best performing model is found.

"Different architectures have been tested to check which one delivered the best results." - Appendix E.1

"It's actually all still quite new and the models are not always well understood" -Appendix E.2

The quotes above are from a model developer and a IT architect in two different case studies. These quotes show that the DL is a complex technique that is often not fully understood by the stakeholders that use it. The complexity is partly caused by characteristics of the technology and partly caused by the lack of knowledge of the stakeholders. So, the restriction has social and technical causes.

Innovative stage of Deep learning

A second restriction is the innovative stage of DL models. Most projects indicated that some part of the model or project was confidential and could not be published. This is attributed to the innovativeness of the technique that creates a value to not share insights. This restricts the ability to account the energy consumption as it limits the available information, as seen in Section 5.2.1. Also, within projects the available DL techniques changes and improve, which provides new opportunities but makes the training process less transparent. Finally, since the technique is so new the focus is more on understanding and realising the technique, instead of mapping and mitigating its side effects.

"I think that people first need to have some trust in the techniques before one will do the next step and will ask what does it do with the energy." - Appendix E.2

"The discussion about the project started around 5 years ago and back then they had some high hopes about the technology, but only in recent years it really kicked of due to new technologies." - Appendix E.3

The quotes above are from two model developers from two different case studies. The first quote shows that innovative techniques need to be understood by stakeholder before they will consider negative side effects of the techniques. The second quote shows that the DL techniques changes in only a few years, which influences the projects. So, the restriction has social and technical causes.

Lack of incentive to determine energy consumption

This restriction derives from the lack of awareness about the power cost and the lack of priority to reduce energy consumption. Projects often have incentives to make processes efficient to minimize costs for the servers, but there is no incentive to determine or minimize the energy consumption of the training. Although the increase in efficiency and minimization in energy are often linked, it is an restriction on the energy accounting that this is no stand-alone incentive.

"The consideration to scale down (...) was the efficiency of the whole process." - Appendix E.2

"(...) the supercomputers might consume a lot of energy, but they are located in the building (...) and the CO2 emissions of the building are compensated." - Appendix E.4

The quotes above are from an IT architect and a model developer from two different case studies. The quotes show that there are other incentives within the organization or personally, which have higher priority when developing DL models, than the energy accounting.

Lack of model developers' energy accounting knowledge

This restriction appears in all investigated cases and arises from the lack of knowledge of model developers about the consumed energy of their DL model training. Also, it covers the lack of knowledge on how to reduce the energy cost of the models they develop and train.

"I would also find it difficult to act on it. I would not directly know how to reduce it, since I don't have to tools for that. I can figure out a lot about the model, but I can't monitor the energy consumption. - Appendix E.2

"How much energy is consumed is unknown, but its probably a lot." - Appendix E.4

The quotes above are from two model developers of two different case studies. It shows the lack of knowledge about the energy consumption of the models and the lack of tools to account the energy consumption. This lack of knowledge and tools can be caused by the organization or individuals that do not pay attention to the energy consumption.

Lack of societal awareness

This restriction mainly derives from the lack of awareness of the clients. Clients are often not aware of what kind of technique DL is while asking for it and/or they are not aware of the energy consumption of the training of DL models. It also covers the lack of awareness of society in general, with the assumption that societal awareness will eventually increase the client's awareness and creates an incentive for the model developers to reduce the energy consumption.

"The sustainability story (...) is one that nobody is paying attention to and nobody links sustainability and data yet." - Appendix E.2

"In general, energy consumption is a requirement for projects, but for this project it was not a big deal."- Appendix E.3

The quotes above are from two model developers of two different case studies. The quotes show the general lack of interaction among the stakeholders about the link between sustainability and data. And, it shows the lack of attention within organizations to account the energy consumption of DL models, while energy consumption is relevant in other parts of the organization.

Lack of systematic evaluation methods

This restriction arises from different evaluations by different people in different projects. A good result of the models is often not just determined by technical KPIs, but also by human judgement. This can cause subjective evaluation of the models, since one or more human(s) has/have to decide what is good and bad result of the model. Also, there is no overarching systematic evaluation method to determine what a good result is for training a DL model.

"It turns out to be quite complex (...) to collect data to train and to say with the training outcomes that a model is a good idea." - Appendix E.2

"We had different models in place and we just picked the one that seemed to got the best results of understanding their questions and test data." - Appendix E.3

The two quotes above are from two model developers from two different case studies. The quotes show that there is a lack of organization to systematic evaluate the outcomes of the models and to determine whether a model is good (enough).

Long and diverse training time

This restriction derives from the training process of DL models. Often many different architectures with many different configurations are tested before the best one is found and this process is often not carefully documented. Also, once the best configuration is found, many models need additional training to keep the model up-to-date with new input data. This can create an ambiguous process of many forgotten training hours.

"5 to 10 weeks for the initial training and biggest start (...) training is never finished. So, you're going to keep updating after that." - Appendix E.2

"(...) when they tried fine tuning the models, they spent a few days training (...) that wasn't really enough to improve the models." - Appendix E.3

The quotes above are from two model developers from two different case studies. The first quote shows that the technique of DL requires continuous training for the models to be up to date. The second quote shows that the training process can contain many training-hours without any results. This requires a structured organization to collect all the data about the energy consumption of training.

No hardware details available

This restrictions derives from the lack of details available for the model developer. This could be caused by an organizational structure with distributed knowledge or by the hardware data simply not being available for an organization. Also, the use of development software, service providers, or virtual machines can create an additional layer of infrastructure that makes it harder to identify the underlying hardware.

"An intermediate platform is used for training the model and storing the data. (...) Exactly what hardware is underneath is hard to say." - Appendix E.2

"It would be difficult to find that information even if we had a week or two and that was part of the requirements."- Appendix E.3

The quotes are from two model developers from two different case studies. The quotes show that a lack of interaction among stakeholders and lack of organization or too much layers in an organization can results in a lack available hardware details for the model developers and therefore also for other stakeholders.

Updating model over time

In two case studies the interviewees indicated that the DL model itself, the scope of the model, and the complexity of the model changed over time. This can create a complex history of training runs for different parts and products of the model for different clients. Also, input data can change over time, which require more effort to retrieve valuable outcomes. Eventually this can create a non-transparent image of the energy that is consumed by training a model.

"And sometime in the last six months, the model was updated to support a little more products and sectors, and you also see now the model has become a little bit more complex."
- Appendix E.2

"Some preconditions depend on which customer (...). One wants it delivered this way and the other on those methodologies. They differ. The input images can also be different. "
- Appendix E.4

The quotes are from an IT architect and a model developer of two different case studies. The quotes show that the restriction is caused by a lack of organization among the stakeholders to use similar outcomes or standards and by changing demands of customers, what leads to a more ambiguous training process to account the energy.

6.3 CATEGORIZATION OF THE RESTRICTIONS

The restrictions can be classified into three categories, namely organizational, social, and technical restrictions. These three categories derive from the seven points of view that are used in the RISMAN method to perceive projects and identify risks [van Well-Stam et al., 2011]. The other four described categories are areal, financial, legal, and policy.

The previous section discussed the restrictions and the underlying quotes, which show what causes the restrictions and how they are caused. The table below presents an overview of the categorization of the restrictions, which are briefly discussed in the previous section. The sections below the table explains the three categories that cause the restrictions and discusses briefly a possible solution per category and per restriction. These solution direction are based on the causes explained above.

Table 6.1: Overview of the categorization of the restrictions.

Restrictions	Organizational	Social	Technical
Complexity of deep learning	0	1	1
Innovative stage of deep learning	0	1	1
Lack of incentive to determine energy consumption	1	1	0
Lack of modeler developers' energy accounting knowledge	1	1	0
Lack of societal awareness	1	1	0
Lack of systematic evaluation methods	1	0	0
Long and diverse training time	1	0	1
No hardware details available	1	1	0
Updating model over time	1	1	0
Totals	7	7	3

6.3.1 Organizational

This category contains the restrictions that are caused by organizational structures or a lack of organization. To solve the restrictions in this category, the structure of the organization should be changed or a structure needs to be set-up. Within this category seven restrictions with possible solutions exist:

1. **Lack of incentive to determine energy consumption**; provide an incentive top-down in the organization to stimulate the energy accounting.

2. **Lack of model developers' energy accounting knowledge**; develop a structure for education and tools within the organization to account the energy.

3. **Lack of social awareness**; report as an organization the energy consumption of the DL that are used and stress the importance of the energy accounting.

4. **Lack of systematic evaluation methods**; develop a general approach to evaluate outcomes of models within the organization.

5. **Long and diverse training time**; set-up protocols to document all training-hours or server-hours of DL models and tools to account the energy.

6. **No hardware details available**; use standardized layers in an organization that provides hardware information to the users.

7. **Updating model over time**; create predefined outcomes by a model in a organization that can be modified for different customers or stakeholder to limit the amount of updates.

6.3.2 Social

This category contains the restrictions that are caused by (lack of) social interactions or lack of knowledge of the stakeholders. The restrictions in this category are relatively easy to solve on the short term, since the actors can be educated and processes can be adjusted. Within this category seven restriction with possible solutions exist:

1. **Complexity of deep learning**; educate stakeholders about how to cope with the complexity to help them understand DL.

2. **Innovative stage of deep learning**; educate stakeholders in an early stage about the negative side effects of existing DL techniques and new techniques.

3. **Lack of incentive to determine energy consumption**; stimulate interaction among stakeholders about the energy consumption of DL models and stress the importance of energy accounting.

4. **Lack of model developers' energy accounting knowledge**; provide stakeholder the opportunity to learn about energy accounting and let them share experiences.

5. **Lack of societal awareness**; educate stakeholders about the energy consumption of training DL models and stimulate interaction about the subject..

6. **No hardware details available**; stimulate interaction among stakeholders to share hardware details and experiences of best practices.

7. **Updating model over time**; stimulate interaction between stakeholders about requirements of models and align these requirements.

6.3.3 Technical

This final category consists of restrictions that are caused by technical characteristics of DL models. These restrictions are relatively hard to solve and therefore the emphasis is more on mitigating the restrictions rather than solving them. Within this category three restriction with possible solutions exist:

1. **Complexity of deep learning**; gather information of best practices to map and reduce the complexity of training DL models.

2. **Innovative stage of deep learning**; map the functionalities of existing and new DL techniques to keep track of new opportunities.

3. **Long and diverse training**; reconsider what training is required for a desired result and schedule and document these hours.

6.4 OVERVIEW PER CASE STUDY

This section presents how the different restrictions that are described in the previous section are distributed over the case studies and discusses remarks about the distribution of the restrictions. The table below provides an overview of what restrictions are coded in the case studies. The results are binary, so 1 if the restriction was named in an interview and 0 if not. The three lower rows of the table provide the sums of the categories of the restrictions in the case studies. This can help identified if a category is over represented in case studies.

Appendix D.1 provides an overview of how often the restrictions are named per case study. This table is not used here, since it gives a distorted presentation of the results. Some case studies had more and longer interviews, which results in more notions of some restrictions.

Table 6.2: Binary overview of what restrictions are identified in what projects.

Restrictions	O	S	T	Case study 1: Slaughter house	Case study 2: Bank’s credit	Case study 3: EPIA	Case study 4: Asphalt damages
Complexity of deep learning	0	1	1	1	1	0	1
Innovative stage of deep learning	0	1	1	1	1	1	0
Lack of incentive to determine energy consumption	1	1	0	0	1	0	1
Lack of modeler developers’ energy accounting knowledge	1	1	0	1	1	1	1
Lack of societal awareness	1	1	0	1	1	1	0
Lack of systematic evaluation methods	1	0	0	0	1	1	1
Long and diverse training time	1	0	1	1	1	1	1
No hardware details available	1	1	0	1	1	1	0
Updating model over time	1	1	0	0	1	0	1
Totals	7	7	3	6	9	6	6

Totals				
Organizational	4	7	5	5
Social	5	7	4	4
Technical	3	3	2	2

Remark about the case studies

The first remark is that case study 2 ticks the boxes of all restrictions, while the other case studies all have 6 restrictions spread differently. Explanations for this remark are that (i) the case study had two interviewees, (ii) that the two interviewees both were technical involved in the model, and (iii) that the two interviewees exchanged more information. Case studies 3 and 4 had only one interviewee. Case study 1 had only 1 interviewee who was technically involved. All together it explains the high total restrictions, but important to note is that there is no restrictions that is only named by case study 2.

The categorization of the restrictions over the case studies shows an equal distribution over the different case studies, comparable with the total number of restrictions identified in the case studies.

Remarks about the restrictions

The two restrictions *Long and diverse training time* and *Model developer's energy accounting knowledge* have been identified in all case studies. This already indicates that these restrictions are important in case studies, since they are identified separately in all case studies. The restriction *Lack of incentive to determine energy consumption* is one of the two restrictions that is only named twice in the case studies. This might indicate that the restriction is less valid for cases in general. However, it should be noted that this restriction is linked to the restriction *Lack of social awareness* and some of the underlying restrictions from the analysis are very similar. Therefore, this restriction is assumed to be significant. The restriction *Updating model over time* is the other restriction that is only identified in two of the four case studies. This restrictions is not linked to another restrictions and might therefore be less valid for cases in general.

6.5 SECTIONAL CONCLUSION

This section aims to answer the second sub-research question as mentioned in Section 2.2. This question is:

What are the restrictions on the energy accounting for Deep Learning models?

This chapter first provides an overview of the nine restrictions that have been identified in the case studies and explains what the different restrictions are. For each restriction, there are two examples directly from the interviews and an explanation what the causes for that restriction are. Second, this chapter categorized the restrictions into three categories, namely organizational, social, and technical. Next, a binary overview is presented of the restrictions identified in the case studies. Special remarks about this overview is that one case study identifies all restrictions, two restrictions have been identified in all case studies, and two restrictions are identified in only 2. At each restrictions the categories are listed, these categories are organizational, social, and technical. Below an overview of all the restrictions and the categories of the restrictions is given:

- Complexity of Deep learning (Social, Technical)
- Innovative stage of Deep learning (Social, Technical)
- Lack of incentive to determine energy consumption (Organizational, Social)
- Lack of model developers' energy accounting knowledge (Organizational, Social)
- Lack of societal awareness (Organizational, Social)
- Lack of systematic evaluation of models (Organizational)
- Long and diverse training time (Social, Technical)
- No hardware details available (Organizational, Social)
- Updating model over time (Organizational, Social)

The next chapter validates with experts whether these restrictions can be generalized.

7 VALIDATION IN-DEPTH INTERVIEWS

This chapter presents an overview of the experts of the in-depth interviews and the validation of the restrictions that are found in the previous chapter. Also, this chapters presents additional restrictions that are found outside the case studies.

7.1 VALIDATION OF THE CASE STUDY RESTRICTIONS

To validate the restrictions, this section first explains the interviewees. Then, it presents how the different perspectives validated the restriction that were found in the previous chapter.

7.1.1 Categorization of interviewees

The experts who are interviewed are listed by role and organization in Table 7.1. The table also shows that the interviewees can be categorized into three perspectives, namely the governmental, scientific, and service provider. Important to note is that the categorization is added after the interviewees were identified. This categorization improves the clarity of the validation of the restrictions. Appendix F provides more details about the interviewees and the transcription of the interviews. Below the table is explained what the perspective is and why they are relevant.

Table 7.1: Categorization of the roles of the interviewees.

Perspective	Role	Organization
Governmental	Senior Advisor ICT	Netherlands Enterprise Agency (RVO)
	Member of Cabinet	Frans Timmermans' Team of the European Commission
Scientific	Assistant professor	Carnegie Mellon University at Language Technologies Institute
	Associate professor	Delft University of Technology at Faculty of Electrical Engineering Mathematics and Computer Science
	Assistant professor	Delft University of Technology at Faculty of Electrical Engineering Mathematics and Computer Science
Service provider	Program manager	Machine learning team at Service Provider
	Cloud Architect	Technology Strategy & Transformation team at Deloitte Netherlands

Governmental perspective

This perspective derives from the responsibility the government has to reduce global warming. Chapter 1 explains the societal relevance of the problem and therefore it is relevant to examine the perspective of the Dutch government and European Commission.

Scientific perspective

This perspective derives from the literature that proved in Chapter 4 and 5 to provide insufficient tools to calculate the energy consumption of DL training. Therefore, it is valuable to evaluate the restrictions with the scientific community to confirm whether the literature lacks tools for model developers in practice.

Service provider perspective

This perspective derives from the dominant role service providers proved to play in accounting the energy consumption of training DL models. More and more training is executed on cloud services and without the perspective of the service providers, no complete image can be drawn.

7.1.2 Validated initial restrictions

This section discusses the validation of the restrictions that have been identified in the case studies. The table below presents an overview of what restrictions were validated in total by the different perspectives. The results are binary, so 1 if one of the interviewees validates the restriction and 0 if not. The lower rows of the table sum the categories per perspective.

Appendix D.2 provides a full overview of how often the restrictions are validated in the individual interviews. This table in Appendix D.2 is not used in this section, since some interviews resulted in more text and therefore more validations of the same restrictions.

Categories of the case study restrictions

The categories in the table below are the categories that have been identified in the previous chapter. This does not mean that the identified restrictions do not have areal, financial, legal, or policy causes, but these other categories are not identified in the case studies and the in-depth interviews focused on validating the findings from Chapter 6.

Table 7.2: Binary overview of what restrictions are identified in what projects.

Restrictions	**O**	**S**	**T**	**Governmental perspective**	**Scientific perspective**	**Service provider perspective**
Complexity of deep learning	0	1	1	0	1	1
Innovative stage of deep learning	0	1	1	1	1	1
Lack of incentive to determine energy consumption	1	1	0	1	1	1
Lack of modeler developers' energy accounting knowledge	1	1	0	0	1	1
Lack of societal awareness	1	1	0	1	1	1
Lack of systematic evaluation methods	1	0	0	0	1	1
Long and diverse training time	1	0	1	0	1	0
No hardware details available	1	1	0	1	1	1
Updating model over time	1	1	0	0	1	0
Totals	7	7	3	4	9	7

Totals			
Organizational	3	7	5
Social	4	7	7
Technical	1	3	2

Opposed perspectives

The first remark about the validation by the different perspectives is the relatively low number of validated restrictions by the governmental perspective. The scientific and service provider perspective validate 9 and 7 of the restrictions, but the governmental only 4. This can be attributed to to the varying knowledge among the stakeholders about the technique of DL. This is confirmed by the varying sums of the technical category per perspective.

In contrary to the previous remark, the second remark is about the full validation of the scientific perspective. Although the service provider perspective only validated two restrictions less, it is remarkable that the scientific perspective validated all. An explanation for this validation is the extensive knowledge of the scientific interviewees about DL, which is opposing the DL knowledge of the governmental interviewees.

Less valid restrictions

Almost all restrictions are validated by at least two perspectives, but the *Long and diverse training time* is one of the restrictions that is only validated by one perspective. Not being validated by the governmental perspective is not that remarkable since it requires basic knowledge about how DL is trained, but not being validated by the service provider is remarkable. An explanation can be that the service provider perceive more training time not as a restriction, but as positive since it might provides more revenue. It is extra remarkable that this restrictions is validated by only one perspective, since the restriction is coded in every case study. The second restriction that is only validated by the scientific perspective is *Updating model over time*. This is less remarkable as it was only named in two case studies.

Fully validated restrictions

Three restrictions found in the case studies are validated by all perspectives. *Lack of societal awareness* and *No hardware details available* are not very remarkable, since they were identified in three of the four case studies. However, *Lack of incentive to determine energy consumption* is identified in only two case studies and with a very low count in number of notions (see Appendix D.1). An explanation could be that the case study interviewees were less willing to elaborate on missing incentives as it might affect their daily work and reputation. Nevertheless can be concluded that it is a valid restriction.

7.2 IDENTIFICATION OF THE ADDITIONAL RESTRICTIONS

Besides the restrictions that have been identified in the case studies, additional restrictions have been identified in the in-depth interviews. First, these restrictions have been explained in more detail. Second, the restrictions have been categorized, similar as in the previous chapter. Finally, an overview is provided with the additional restrictions and the perspective in which they are identified.

7.2.1 Explanation of the additional restrictions

The six additional restrictions identified in the in-depth interviews are not validated, but independently named by the interviewees. Below presents these six additional restrictions.

Conflicting interests at service provider & data center

This restriction refers to two conflicting interests of service providers and to some extent data centers. One interest is to train DL models or execute functions as efficient as possible, so they can process the requests of as many customers as possible. However, another interest for them is to let users use compute as much as possible, so they can rent out more servers. This creates an incentive for the service providers and data centers to not provide insight into the energy consumption of the rented server, so customers will not limit their use.

"(...) then they (data centers that lease spaces for servers) have an interest in having as many customers as possible who want that. It's basically up to the customer to say: I don't need a whole corridor (of servers) anymore" - Appendix F.1

"I agree there is a conflicting interesting (...) some people in the business simply want our customers to run massive training jobs all day, every day." - Appendix F.3

The two quotes above are from the governmental perspective and the service provider perspective. The quotes provide the insight that the conflicting interest is caused by the earning model of the organization, so the structure of the organization. And, that the restriction is caused by the method of interaction between the different stakeholders.

Lack of energy accounting tools at service provider

This restriction derives from the interviews with the service provider perspective, where is indicated that the service providers do not have a tool to properly measure the energy consumed by services they provide. They stress that the necessary data probably is available within the organization, but that it needs to be structured and combined. This process of developing a tool has started. Without this tools, energy accounting at service providers remains estimations, since most of the information is confidential.

"There must be some data somewhere, (...) where you get insight into the consumption of such a server in AWS or Azure." - Appendix F.3

"I have actually the equivalent of a dissertation on ways that we can do it (accounting the energy consumption). We just need to measure it first." - Appendix F.3

The two quotes are from two different interviewees from the service provider perspective. The quotes show that the restriction is not caused by technical difficulties, but by organizational. The quotes show that the information is somewhere available, but just need to be structured.

Lack of governmental enforcement tools

This restriction derives from the lack of tools by the government to create incentives for model developers or clients to reduce the consumed energy by training the DL models. The Dutch government mainly focuses on running the existing servers as efficiently as possible, but not on the services that run on the servers. Controlling this input is nearly impossible and undesired by the Dutch government.

"(...) it's also very difficult to enforce, because you have the recognized measures list that the technology (...) pays for itself within five years and that companies have to apply it. That's very difficult to legislate that in terms of software." - Appendix F.1

"So, there are people in Digi-connect working with us, with the sector also, to look at: How can we arrive at the same measurement method, because every week some tech company comes up to me (...) and I can do very little with it, because everyone uses a different methodology (...)" - Appendix F.1

The two quotes are from two interviewees from the national and European governmental perspective. The former quotes shows that it is hard to create policy for the technology. So, this restrictions has policy and technical causes. The latter shows a lack of organization in the measurement methods at the governmental agencies and the lack to translate it into a policy.

Lack of scientific tools to account energy consumption

This restrictions derives from the limited availability of scientific tool to calculate and measure the energy consumption of training DL models. Although there are several methods to calculate or measure the energy consumption, there is not one straight-forward methods to apply in science.

"You could take a relatively simple model and then the energy consumption would still be complex. But, that's complex for almost all models." (- Appendix F.2

Usually the averages are used to determine the capacity of the hardware over a whole bunch of iterations and then you average the estimations. So, you get more or less stable estimates, hopefully. But yeah, it does vary per batch. - Appendix F.2

The two quotes above are from two interviewees in the same interview of the scientific perspective. The quotes show that the energy accounting is technically complex, regardless of the model you are using. And, they show that the best calculations are still estimations of the average energy consumption, so that there is no precise technical method to calculate it.

Limited information provided by service providers

This restriction derives from the confidentiality of much of the information that is required to account the energy consumption of training DL models. Different service providers compete on different levels with each other and the energy consumption and energy cost is one of these levels to gain advantages over competitors.

"I have come across very little from service providers in terms of dashboards and reports about how, as a customer, you have insights into what energy consumption underlies (..)" - Appendix F.3

"if you publish results about the AWS energy consumption is dramatic (...) I can imagine they want to do their own research first and if they like it, they publish it very nicely and if they don't like it, they don't publish it." - Appendix F.3

Both quotes above are from one interviewee with the service provider perspective. The quotes show that there is little information sharing or social interaction among the stakeholders and there is no organization to report the information among the stakeholders.

Separation between science and society

This restriction derives from the different focus of science and society on DL models and energy accounting. Many of the tools and input metrics used in the scientific community is not known or not relevant for the model developers in practice. Science is likely to focus on FLOPS, where the model developers in society mainly focus on the GPU utilization. This restricts the energy accounting in society, since many of the tools from the scientific community cannot be used.

" I found some very shallow way of computing that from looking at the number of floating operations. And then, you know, at certain GPU support a certain number of floating operations per Watt and then you can compute an approximate estimate of how much it would take to train a certain network, if you know the number of floating operations the network takes." - Appendix F.2

"That is useful, because we sit in our ivory tower smoking our pipe, thinking a lot. But, it's nice to talk to people because sometimes interesting problems appear that you don't think of staring outside of the ivory tower." - Appendix F.2

Both quotes above are from one interviewee with the scientific perspective. The first quote is an explanation of their scientific method to calculate the energy consumption, but the case studies showed that this method is not applicable in practice. Although the second quote is intended jokingly, it confirms the restrictions that there is a social division between science and society. Both quotes show the restriction is caused by a lack of interaction of the stakeholders.

7.2.2 Categorization and overview of the restrictions

The additional restrictions are categorized similar to the categorization in Section 6.3. However, one additional restrictions is categorized into a category besides the organizational, social, and technical categories, namely policy. This category contains restrictions that are directly caused by a lack of proper policies by governmental institutions. The table below present an overview of what perspectives identified what additional restrictions and the categories of the restrictions.

Table 7.3: Binary overview of what additional restrictions are identified by what perspective.

Restrictions	O	P	S	T	Governmental perspective	Scientific perspective	Service provider perspective
Conflicting interests at service provider & data center	1	0	1	0	1	0	1
Lack of energy accounting tools at service provider	1	0	0	0	0	0	1
Lack of governmental enforcement tools	0	1	0	1	1	0	0
Lack of scientific tools to account energy consumption	0	0	0	1	0	1	0
Limited information provided by service providers	1	0	1	0	0	0	1
Separation between science and society	0	0	1	0	0	1	0
Totals	3	1	3	2	2	2	3

The sparse table above shows that each perspective added two or three potential restrictions. Also, it show little overlap over the different perspectives. There is only one restriction that is identified in two instead of one perspective, namely *Conflicting interest at service provider & data center*. The categorizations are also sparse, since the restrictions had causes in only one or two categories.

7.3 SECTIONAL CONCLUSION

This chapter focuses on the question whether the restrictions identified in Chapter 6 are validated by experts from different perspectives and it describes the additional restrictions found in the in-depth interviews.

To validate the restrictions from the case studies, in-depth interviews were conducted with experts from three perspectives, namely governmental, scientific, and service provider. From these perspectives the governmental was less familiar with DL, which noticeable in the number of validated restrictions. All restrictions from the case studies were validated at least once. Only two restrictions were validated only once. The first one was identified in all case studies and is therefore assumed to be valid with a validation from one perspective. The other restrictions was only identified in two case studies and is therefore less valid. This restriction is *Updating model over time*. All other restrictions are validated.

In the in-depth interviews, six other restrictions have been identified. One restrictions was identified separately in two perspectives, but all other restrictions were identified in one perspective. These restrictions cannot (yet) be marked as valid restrictions for the case studies. Therefore, they are not taken into account in further research. The restrictions were categorized in the four categories of which one was new, compared to the previous chapter, namely policy.

8 DISCUSSION

8.1 REFLECTION OF THE RESULTS ON THE LITERATURE

Chapters 3 and 4 provide an overview of the literature about metrics to express computational power and about methods to account the energy consumption of DL models. The literature presents to what extent it is possible to determine the energy consumption. However, besides the methodological limitations the literature describes little about what makes it so hard to account the energy consumption of DL models. The literature review in this thesis describes extensively the different metrics to express computational power and energy accounting methods. It also provides a clear overview of the strengths and limitations of the energy accounting methods that can be used.

Also, the methods in the literature for energy accounting all focus on nearly ideal cases. This research projects provides insights into real world cases where a lot of the information is not available for the model developers. The energy accounting methods in literature spend little attention to the training of DL models at service providers and the roles of the different stakeholders in the energy accounting process. This research projects examines what information is available when training these models at service providers, which stakeholders are involved, and which restrictions play a role in the energy accounting.

Moreover, this research project provides to some extend the restrictions on the application of scientific accounting methods in real world cases. The results of this thesis are an extension to the literature on what problems can be perceived for stakeholders outside the scientific community. For example, the scientific perspective validated that the complexity of DL can be a restriction for the stakeholders, but also stressed it was less relevant in the scientific community.

To conclude, there are three main points to interpret the results in the literature. First of all, this thesis reviews the literature of different energy accounting methods for training DL models and discusses the strengths and limitation of the methods. Second, the results of this thesis reveal that the data available in cases from literature possess a lot more information than real world cases. Third, the results of this thesis add to literature problems that stakeholders can encounter, when applying the energy accounting methods.

8.2 REFLECTION ON THE RESULTS

Chapter 5 reveals that there is little information available in the case studies to account the energy consumption of DL models. These findings say something not only about the case studies, but also about the complexity of the information gathering and the lack of transparency of information in the industry. Many stakeholders are afraid to share information because it might contain sensitive information. Moreover, this research project presents the stakeholder identification, in which most stakeholder have no active role in energy accounting. Many stakeholders point to others within or outside their organization, when being asked about the energy consumption of the training of DL models. This results in a situation where nobody takes responsibility and no action is taken. It is up to the governmental institutions to set the rules of the game and to stimulate the interaction and knowledge sharing, up to the clients to demand information about the energy consumption of the product or services they pay for, up to the service providers to provide straightforward numbers about the energy consumption to the model developers, and up to the model developers to document and summarize the energy consumption of the product or services.

Chapter 6 presents the identified restrictions on the energy accounting of training the DL models from the case studies. These restrictions reveal the dispersion of the information to account the energy consumption across the stakeholders and the unawareness of most stakeholders. The stakeholders do not give each other incentives to account the energy consumption and for some of them it saves time and money to not name it.

Chapter 7 validates the restrictions identified in the case studies with three different perspectives, namely governmental, scientific, and service provider. The governmental perspective reveals the difficulty for governmental institutions to form policies on the energy accounting of training DL models. For them, the focus is on the efficiency of the supply side of computational power, i.e. the servers in data centers. Logically, since the energy consumption of these servers can be clearly measured and optimized. It is harder and more subjective to find and measure the efficiency of DL models. The scientific perspective revealed the gap between society and science. Most of the research studies use metrics that are irrelevant for the stakeholder and cost a disproportionate amount of effort compared to the benefits. The service provider perspective revealed that there is some attention to the energy accounting of the services they offer, but it is still in its infancy. Also, it revealed that it is an organizational and social challenges, rather than a technical challenges to account the energy consumption. This suggests that the service providers are mainly afraid to share the information, but it would benefit the stakeholders to share this information.

8.3 POLICY CONSIDERATIONS

There are several policy consideration to educate stakeholders, stimulate interaction among stakeholders, create a structure for the stakeholders to organize themselves, and mitigate the technical complexity. The policy advise provides a number of policies that can be applied on a national or European scale. It is up to the decision makers to decide what policies are desirable to implement.

The first policy to consider is setting standards for the service providers on what and how to communicate the energy consumption of the services they offer. This policy stimulates the interaction between service providers and model developers and structures the interaction. It is valuable to divide the energy consumption into the energy consumption of the training itself and the overhead energy cost. This separation provides transparency about the energy consumption and the efficiency of the services. However, the service providers are reluctant with sharing this data, since they claim that the data needs to remain a confidential. Therefore, the decision makers should assess what interest is more important. This policy can (i) improve the model developers' energy accounting knowledge, contribute to a systematic evaluation method, (ii) make the energy consumption of long and diverse training time more opaque, and (iii) provide more details about the hardware and the energy consumption of the hardware.

The second policy to consider is to provide standards about what energy consumption is perceived to be 'high', 'normal', and 'low' for certain applications of DL. This policy contributes to the stakeholders' knowledge about energy consumption and stimulates interaction among stakeholders about the topic. Decision makers should decide with the stakeholders how the energy consumption should be perceived as 'high', 'normal', and 'low'. This policy can (i) provide an incentive to determine the energy consumption for model developers, clients, and customers, (ii) improve the model developers' energy accounting knowledge, and (iii) increase the social awareness about the energy consumption of these models.

The third policy to consider is to develop a certificate for green computing for model developers and clients. This certificate would require the model developers to keep a log of the total hours of training and the total energy consumption of these training hours. This policy structures the documentation of training hours and its energy consumption between the stakeholders and enhances the interaction among the stakeholders. Decision makers should decide on the conditions to receive the certificate. The conditions could be to just log the training hours and energy consumption of these hours or could be more strict and could demand more measures to reduce the energy consumption of the models. It would be convenient to issue the certificates to companies that applied the required protocols, instead of assessing individual cases. This policy can (i) provide an incentive to determine the energy consumption for model developers and clients, and (ii) improve the structure and documentation of the long and diverse training hours.

The final policy to consider is to develop a knowledge sharing platform where model developers and clients can share best practices about new (DL) technologies and the energy consumption of these technologies. This policy contributes to stimulating the interaction among the stakeholders, enhancing their knowledge, and documenting the different and new DL techniques and the corresponding energy consumption. Decision makers should decide on the level of generalizability of the the platform and should stimulate stakeholders and organization to participate on the platform. The certificate of the previous policy could be linked to the platform to stimulate or oblige stakeholders to participate. This policy can (i) reduce the complexity of DL by documenting the different techniques, (ii) enhance the knowledge of new and innovative DL techniques, (iii) improve the model developers' energy accounting knowledge,(iv) increase social awareness, and (v) provide more information about the energy consumption of hardware and what combination of hardware and model architecture perform best.

To conclude, to overcome all validated restrictions revealed in this thesis, there are four policies that could be implemented. These four policies are:

- to set standards on what and how to communicate the energy consumption of service providers to the model developers;
- to set standards on what is high, normal, and low energy consumption for certain DL architectures and applications;
- to develop and issue certificates that require logging of all training hours, and;
- to develop a knowledge sharing platform for best practices of (DL) technologies and the energy consumption of these technologies

8.4 LIMITATIONS OF THE RESEARCH

There are several limitations to this thesis. The first limitation is that the cases studies only consisted of one or two interviewees, where two is assumed to be the absolute minimum for a clear image of the cases by Yin [1984]. This might have caused the case studies in this thesis to be a weak representation of reality. However, the case studies with two interviews revealed that the second interviewee knew little about the energy consumption of training DL models. Despite of the this second interviewee being a project manager or an IT architect. Moreover, finding additional participants for the case studies turned out to be difficult, since many potential interviewees were unwilling to participate. They perceived their own knowledge about DL, training DL, and/or the energy consumption of the training insufficient to participate. Besides the case studies, this problem also occurred with the in-depth interviews. Potential interviewees argued that they had little to no knowledge about the energy consumption of DL or the energy consumption of any service in data centers. Other potential interviewees were not willing to participate, since they were not allowed to be interviewed about the subject due to confidentiality.

The second limitation is the selection of case studies and interviewees. Ideally these case studies would represent a variety of cases and the interviewees would represent a variety of experts. In this research project, the case studies and interviewees used were at hand within the limited time for this thesis. For the case studies, this resulted in only two types of DL applications, namely Knowledge Base and Image/video recognition. For the in-depth interviews, this resulted in only one interview with a service provider and for the governmental perspective interviewees that were not specialised in DL models. Therefore, the case studies and interviewees are not a complete representation of all cases or all involved experts. So, the validation of the case study restrictions could be more deepened with more interviewees and more perspectives.

The third limitation is that the interviews and coding of the interviews are conducted by only one investigator. Ideally, the interviews and its coding would have been executed by multiple investigators to make it less biased. Auerbach and Silverstein [2003] explain how multiple investigators can significantly decrease the bias of coding the interviews. However, this was not possible due to the individual nature of this thesis.

9 CONCLUSIONS AND RECOMMENDATIONS

In this chapter, we summarise the sub-research questions and conclude on the main research questions. The table below presents the sub-research questions and in which chapters these questions were addressed.

Table 9.1: Sub-research questions.

Question	Chapter
1. What metrics can be used to define computational power of Deep Learning models?	3
2. What methods are available to account the energy consumption of training Deep Learning models?	4
3. To what extent can the energy consumption of training Deep Learning models be accounted in practice?	5
4. What are the restrictions on the energy accounting for developing Deep Learning models?	6 & 7

First, we answer the sub-research question from the table. Next, we conclude on the main research question and explain the relevance of the research project. Finally, the recommendations for further research are presented.

9.1 ANSWERS TO SUB-RESEARCH QUESTIONS

Question 1 - What metrics can be used to define computational power of Deep Learning models?

There is no straight-forward metric to define the computational power of Deep Learning models. Training Deep Learning models requires a vast amount of computational power with a varying demand during the run-time, which can be days or weeks. Because of this long run-time, it is less convenient to only report the total number of computations that is required to produce a result, the so-called floating point operations.

The hardware must be able to process peaks in the computational demand, so it is valuable to express the average speed of the computations over a period of time. It is common to measure this in floating point operations per second-day, which is the average over a day of the number of floating point operations that is processed by the hardware per second. Since this metric is very dependent on the hardware that is used to process the floating point operations, it is also common to express the utilization of the processing units. A benefit of this metric, compared to the previous, is that it can be directly linked to the energy consumption of the computations and is easier to retrieve.

To conclude, the average number of floating point operations per second over one day combined with the utilization of the processing units can be used to define the computational power of Deep Learning models. However, the utilization of the processing is a more convenient metric to define computational power, since it is more accessible and directly linked to energy consumption.

Question 2 - What methods are available to account the energy consumption of training Deep Learning models?

Four methods have been investigated that can calculate or measure the energy consumption of training Deep Learning models. The methods that measure the energy consumption can be applied in future cases and use the utilization of different processing units to account the energy consumption of training Deep Learning models. However, this utilization can not be retrieved after training the models.

The calculation methods have two different approaches, one very detailed and one general estimation. The detailed method requires very detailed information about the hardware, such as the capacitance of the capacitor on the hardware circuit. So, this method is inconvenient to apply on the training of Deep Learning models. The second method is to calculate the energy consumption based on the peak performance of the main processing unit and the total run-time. This method creates an incomplete indication of the energy consumption, but it can be adjusted with adding the Power Usage Effectiveness.

To conclude, the peak performance of the main processing unit and the total run-time is the best applicable method to account the energy consumption of training Deep Learning models. This method does not fully account the energy consumption, but is the best available method to account the energy in practice.

Question 3 -To what extent can the energy consumption of training Deep Learning models be accounted in practice?

Four case studies have been investigated to explore what information is available to account the energy consumption of training Deep Learning models. Generally, there was very little information available to account the energy consumption of the models that were developed. The process to find and train the final model architecture was often unstructured and not well documented.

Only two of the four cases provided the crucial information to account the energy consumption of the training process. This crucial information was the graphical processing unit that was used for the training and total training time of the different model architectures. This information can be used to provide an (under) estimation of the energy consumption of training the models. In the other two cases, this information was unknown or confidential. The energy consumption of the case studies was relatively low, but this is an underestimation of the actual energy consumption. Also, the case studies present that the energy consumption of training Deep Learning models is often not considered and there are steps to be made.

Moreover, the case studies identify the stakeholders, which can or should play a role in accounting the energy consumption of training Deep Learning models. However, the stakeholder identification presents that none of the stakeholders take an active role in the energy accounting of training Deep Learning models, besides the scientific community.

To conclude, the ability to perform the accounting of the energy consumption of the training in practice is hampered by various factors. For all four cases, detailed energy accounting turned out to be impossible due to missing information. Also, the case studies identify the different stakeholders and their (lack of) current role in energy accounting the training of Deep Learning models.

Question 4 - What are the restrictions on the energy accounting for developing Deep Learning models?

The case study interviews revealed nine restrictions. To arrive at these restrictions, we aggregate more detailed restrictions from the transcripts of the interviews. These restrictions are classified into three categories, based on the causes of the restrictions in the transcripts. These categories describe the causes of the restrictions, but can also be used to formulate solution directions to overcome the restrictions. The table below presents the restrictions and corresponding categories.

Table 9.2: Categorization of the restrictions.

Restrictions	**Organizational**	**Social**	**Technical**
Complexity of deep learning	0	1	1
Innovative stage of deep learning	0	1	1
Lack of incentive to determine energy consumption	1	1	0
Lack of modeler developers' energy accounting knowledge	1	1	0
Lack of societal awareness	1	1	0
Lack of systematic evaluation methods	1	0	0
Long and diverse training time	1	0	1
No hardware details available	1	1	0
Updating model over time	1	1	0
Totals	7	7	3

The restrictions and categories of the restrictions are relatively equally distributed over the different case studies. Each restriction is identified in two to four case studies and there is no category over or under represented in any of the case studies.

To validate the findings of the case studies, in-dept interviews have been conducted with experts of three different perspectives, namely governmental, scientific, and service provider. These experts validated all restrictions, except for *Updating model over time*. Also, these experts identified six additional restrictions classified into four categories instead of three. The table below presents the additional restrictions and corresponding categories.

Table 9.3: Categorization of the additional restrictions.

Restrictions	**Organizational**	**Policy**	**Social**	**Technology**
Conflicting interests at service provider & data center	1	0	1	0
Lack of energy accounting tools at service provider	1	0	0	0
Lack of governmental enforcement tools	0	1	0	1
Lack of scientific tools to account energy consumption	0	0	0	1
Limited information provided by service providers	1	0	1	0
Separation between science and society	0	0	1	0
Totals	3	1	3	2

Most of the additional restrictions are identified by only one of the expert perspectives. Therefore, these additional restrictions have not been validated.

9.2 CONCLUSION ON MAIN RESEARCH QUESTION

In this section, we conclude on the main research question. This question is:

What are the restrictions on accounting the energy consumption of building, training, and maintaining Deep Learning models in data centers?

Eight restrictions have been identified and validated the case studies and in-dept interviews of this research project. These restrictions can be categorized, based on the causes of the restrictions and possible solution direction to overcome them. The restrictions that have been identified in the case studies and validated by the experts are mainly categorized as organizational and social. This does not mean that the restrictions cannot be caused by other factors, but no other factors were identified in this research project. These identified restrictions and corresponding categories are:

- Complexity of Deep Learning *(Social & Technical)*
- Innovative stage of Deep Learning *(Social & Technical)*
- Lack of incentive to determine energy consumption *(Organizational & Social)*
- Lack of model developers' energy accounting knowledge *(Organizational & Social)*
- Lack of societal awareness (*Organizational & Social)*
- Lack of systematic evaluation of models *(Organizational)*
- Long and diverse training time *(Organizational & Technical)*
- No hardware details available *(Organizational & Social)*

To conclude, building, training, and maintaining Deep Learning models proved to be an unstructured process, which resulted in scattered information regarding the energy consumption of these models. This makes it really hard to account the energy consumption of training these models. Also, the stakeholders pay little attention to the energy consumption of the models. They have no direct incentive to account or reduce the energy consumption and/or they are not aware that remote servers consume significant amounts of energy. The restrictions on accounting the energy consumption of training Deep Learning models can be overcome by providing options for the stakeholders to educate themselves, stimulating interaction among stakeholders, and creating a structure for the stakeholders to organize themselves and the information required for energy accounting. This provides the stakeholder with the means they need to cope with the technical complexity of Deep Learning. Concrete policies to overcome the validated restrictions are to (i) set standards on what and how to communicate the energy consumption of service providers to the model developers, (ii) set standards on what is high, normal, and low energy consumption for certain DL architectures and applications, (iii) develop and issue certificates that require logging of all training hours, and (iv) develop a knowledge sharing platform for best practices of (DL) technologies and the energy consumption of these technologies.

9.3 RELEVANCE OF THE RESEARCH

9.3.1 Scientific contribution

First of all, this research project provides a systematic approach to analyse what information is available about the energy consumption of new technologies in real cases and what factors make it so hard to determine the energy consumption of these new technologies. This research project focused on training of DL models, but the combination of case studies with few interviews and validation by experts with different perspectives can be applied to other technologies as well. Normally, a case study with a low amount of interviews is perceived to not describe the case study very well and to give a biased view of the case study. However, when investigating the energy consumption of a very specialised new technology, there may not be many interviewees with knowledge about the energy consumption or interviewees that are willing to be participate. So, this thesis offers a systematic approach with multiple case studies and a multi-perspective validation to cope with this problem.

Second, this research project provides an overview of the literature on the different methods to account the energy consumption of training DL models. The literature review provides a clear overview of the different available methods and the strengths and limitations of these methods to calculate and measure the energy consumption of training DL models. It also provides guidance to decide on which method can be applied given a certain amount of information in a case study.

Third, this research project explores the difficulties of energy accounting on real world cases with a multi-stakeholder perspective. Where literature focuses on the technical feasibility of energy accounting methods, this research project also takes into account the different stakeholders, how the stakeholders are organized, and what restricts these stakeholders from accounting their energy consumption. This strongly contributes to lifting the energy accounting of training DL models out of experimental environments of scientists into the real world with clients, customers, service providers, data centers, and governmental institutions to take into account.

Fourth, this research project provides an overview of the restrictions that limit the energy accounting of training DL models. The restrictions provide clear and grounded starting points for the scientific community to further investigate and enable the energy accounting. Also, the research project suggests the categories of solutions to overcome these restrictions. These categories are: to develop or strengthen the organizational structure between stakeholders, stimulate the social interaction and knowledge sharing among stakeholders, and document and mitigate the technical (im)possibilities of accounting the energy consumption of training DL models. Concrete possible solutions for these categories are to develop standards on what and how to communicate the energy consumption of service providers to the model developers, set standards on what is high, normal, and low energy consumption for certain DL architectures and applications, develop and issue certificates that require logging of all training hours, and develop a knowledge sharing platform for best practices of (DL) technologies and the energy consumption of these technologies.

9.3.2 Societal contribution

The identification of the restrictions on accounting the energy consumption of training deep learning model is a first step to create awareness among model developers. Some of the restrictions can already be solved if model developers are more aware of the restrictions and actively fight them. By overcoming these restrictions it will be possible to account the energy consumption and reduce the energy consumption of training deep learning models. If training DL models consumes a significant amount of energy in the (near) future, this research project will contribute to society by reducing the carbon dioxide produced by training these models. With a significant contribution to the energy demand from training DL models, this contribution will reduce global warming.

Moreover, this thesis provide three categories of solution directions to overcome the identified and validated restrictions. These categories are Organizational, Social, and Technical. The thesis even presents four concrete policies, which can be implemented by national or continental institutions to overcome the restrictions. These policies are:

- to set standards on what and how to communicate the energy consumption of service providers to the model developers;
- to set standards on what is high, normal, and low energy consumption for certain DL architectures and applications;
- to develop and issue certificates that require logging of all training hours, and;
- to develop a knowledge sharing platform for best practices of (DL) technologies and the energy consumption of these technologies

9.4 RECOMMENDATIONS FOR FURTHER RESEARCH

This research project provides many directions for further research. Below are six recommendations for the most relevant further research, but many more recommendations are possible. The first recommendation is to investigate case studies with other types of DL application and to validate the findings of this thesis. The case studies in this research project only consisted of knowledge base and image/video recognition applications. Most relevant case studies with other types of DL application to be investigated are NLP and speech recognition. The literature and interviews showed that both application can consume a lot of energy, especially NLP.

The second recommendation is to investigate other case studies about training ML application and case studies about the inference of both ML and DL, instead of only case studies about training DLapplications. ML and inference were out of the scope for this research, but relevant ML applications to investigate are optimization and statistical analysis. Relevant research about inference investigates the restriction on energy accounting in data centers, on device, and in the edge. The literature showed that ML consumes in general less energy than DL, but that ML is applied more often. The literature and the interviews revealed that the total energy consumption of inference is higher than the energy consumption of training the models, but it is also a lot harder to fully account than the training.

The third recommendation is to examine the categories of the restrictions in more detail. The categories of the validated restrictions are based on the causes identified in the case studies. However, the restrictions can have additional causes from those identified in the case studies. For example, it is possible that *Lack of incentive to determine energy consumption* also has policy causes while this is not identified in the case studies, probably because the government has little policy on the reduction of the energy consumption of software. Other possible categories are areal, financial, and legal. It is therefore relevant to further investigate these categories of restrictions to create a complete image of the causes and solution directions per restriction and to explore what are the (best) solutions to overcome the restrictions.

The fourth recommendation is to investigate the effects of the policy considerations on the restrictions and eventually on the energy consumption. Firstly, the effects of the policies on the restrictions should be determined and what the effects are of overcoming the restrictions on the energy consumption. Secondly, the policies can be modeled to explore the effects of the policies on the long term and to compare the effects of the different policies. These policies can be modeled with an agent-based model to emphasize the heterogeneity of the stakeholders. Also, these policies can be modeled with system dynamics to aggregate the heterogeneity of the stakeholders and emphasize the delayed information feedback in the system. However, it is important to gather the required information before modelling the system, since a lot of the data is unknown about how often a model trained, how much energy it consumes, and how much energy is consumed in data centers.

The fifth recommendation is to prioritize the restrictions to determine what factors restrict energy accounting most and to determine what solution will have most impact. The research for this thesis started prioritizing the restrictions. However, these results are not published in this thesis, since the data collection process was too time consuming due to time limitations. Some of the experts were already asked to rank the restrictions from not relevant to very relevant. It would be very useful to further investigate this ranking of restrictions by experts.

The final recommendation is to validate the six additional restrictions identified in the in-depth interviews. It is interesting to investigate whether these restrictions are relevant for the case studies and whether the restrictions are visible for the case studies. It is also relevant to investigate what effect they have on the case studies. And, if the restrictions are not identified or validated in the case studies, to explore who is responsible for these restrictions. The validation can be executed by validating the restrictions with the interviewees of the used case studies and in-depth interviewees in this thesis or other case studies and experts.

BIBLIOGRAPHY

Amodei, D. and Hernandez, D. (2018). AI and Compute.

Assran, M., Romoff, J., Ballas, N., Pineau, J., and Rabbat, M. (2019). Gossip-based Actor-Learner Architectures for Deep Reinforcement Learning. *Advances in Neural Information Processing Systems*, 32.

ATLAS.ti (2020). What is ATLAS.ti.

Auerbach, C. F. and Silverstein, L. B. (2003). *Qualitative data : an introduction to coding and analysis* . New York University Press, New York.

Barroso, L. A. and Hölzle, U. (2007). The case for energy-proportional computing. *Computer*, 40(12):33–37.

Blok, K. and Nieuwlaar, E. (2021). *Introduction to energy analysis*. Routledge, London - New York, third edition.

Bre, F., Gimenez, J. M., and Fachinotti, V. D. (2018). Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings*, 158:1429–1441.

Buchanan, B. G. (2005). A (Very) Brief History of Artificial Intelligence. *AI Magazine*, 26(4):53–53.

Chen, T. and He, T. (2021). xgboost: eXtreme Gradient Boosting. Technical report.

Choudhary, F. and Linden, A. (2020). Innovation Tech Insight for Deep Learning. Technical report, Gartner.

Creswell, J. W. (2003). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Sage Publications, Inc, Thousand Oaks, London, New Delhi, second edi edition.

Dalton, S., Frosio, I., and Garland, M. (2019). Accelerating Reinforcement Learning through GPU Atari Emulation. *arXiv*.

Dayarathna, M., Wen, Y., and Fan, R. (2016). Data center energy consumption modeling: A survey. *IEEE Communications Surveys and Tutorials*, 18(1):732–794.

Department of Economic and Social Affairs (2019). Goal 13.

Dimiduk, D. M., Holm, E. A., and Niezgoda, S. R. (2018). Perspectives on the Impact of Machine Learning, Deep Learning, and Artificial Intelligence on Materials, Processes, and Structures Engineering.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L. C., Misra, S., Mogaji, E., Sharma, S. K., Singh, J. B., Raghavan, V., Raman, R., Rana, N. P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., and Williams, M. D. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, page 101994.

European Commission (2018). Artificial Intelligence for Europe. Technical report, EU, Brussels.

García-Martín, E., Lavesson, N., Grahn, H., Casalicchio, E., and Boeva, V. (2019). How to Measure Energy Consumption in Machine Learning Algorithms. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11329 LNAI, pages 243–255. Springer Verlag.

Google (2020). Cloud Tensor Processing Units (TPUs).

Haenlein, M. and Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, 61(4):5–14.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020). Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning.

Hoa, K. (2019a). The computing power needed to train AI is now rising seven times faster than ever before.

Hoa, K. (2019b). Training a single AI model can emit as much carbon as five cars in their lifetimes.

Hoa, K. (2020). AI researchers need to stop hiding the climate toll of their work.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., and Andreetto, M. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Technical report.

Hwang, T. (2018). Computational Power and the Social Impact of Artificial Intelligence. *SSRN Electronic Journal*.

Intel (2007). Dual-Core Intel ® Xeon ® Processor 5100 Series Datasheet. Technical report.

IPCC (2018). *Global warming of 1.5°C An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change,* .

Kalis, B., Collier, M., and Fu, R. (2018). 10 Promising AI Applications in Health Care. Technical report, Harvard Business Review.

Kaplan, S. J. (1984). The Industrialization of Artificial Intelligence: From By-Line to Bottom Line. *AI Magazine*, 5(2):51–51.

Kok, J. N., Boers, E. J. W., Kosters, W. A., van der Putten, P., and Poel, M. (2009). Artificial Intelligence: Definition, Trends, Techniques and Cases-Joost N ARTIFICIAL INTELLIGENCE: DEFINITION, TRENDS, TECHNIQUES, AND CASES. Technical report.

Königstorfer, F. and Thalmann, S. (2020). Applications of Artificial Intelligence in commercial banks – A research agenda for behavioral finance. *Journal of Behavioral and Experimental Finance*, 27:100352.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the Carbon Emissions of Machine Learning.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning.

Lei, N. (2020). A robust modeling framework for energy analysis of data centers. In *ACM International Conference Proceeding Series*, pages 177–180. Association for Computing Machinery.

Li, D., Chen, X., Becchi, M., and Zong, Z. (2016). Evaluating the energy efficiency of deep convolutional neural networks on CPUs and GPUs. In *Proceedings - 2016 IEEE International Conferences on Big Data and Cloud Computing, BDCloud 2016, Social Computing and Networking, SocialCom 2016 and Sustainable Computing and Communications, SustainCom 2016*, pages 477–484. Institute of Electrical and Electronics Engineers Inc.

Lottick, K., Susai, S., Friedler, S. A., and Wilson, J. P. (2019). Energy Usage Reports: Environmental awareness as part of algorithmic accountability. *arXiv:1911.08354 [cs, stat]*.

Microsoft (2015). Microsoft's Cloud Infrastructure - Datacenters and Network Fact Sheet.

Microsoft Azure (2021). Kies de juiste Azure-regio voor u.

Milieu Centraal (2021a). Grote energieslurpers.

Milieu Centraal (2021b). Koelkasten en vriezers: Koop- en bespaartips.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M. L., Chen, S. C., and Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications.

Rangaiah, M. (2020). 4 Major Applications of Artificial Intelligence in Education Sector .

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Körding, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., and Bengio, Y. (2019). Tackling Climate Change with Machine Learning. Technical report.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4510–4520.

Schell, C. (1992). The Value of the Case Study as a Research Strategy. Technical report.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2019). Green AI. *arXiv:1907.10597 [cs, stat]*.

Sebastian, A., Boybat, I., Dazzi, M., Giannopoulos, I., Jonnalagadda, V., Joshi, V., Karunaratne, G., Kersting, B., Khaddam-Aljameh, R., Nandakumar, S. R., Petropoulos, A., Piveteau, C., Antonakopoulos, T., Rajendran, B., Gallo, M. L., and Eleftheriou, E. (2019). Computational memory-based inference and training of deep neural networks. In *Digest of Technical Papers - Symposium on VLSI Technology*, volume 2019-June, pages T168–T169. Institute of Electrical and Electronics Engineers Inc.

Sicular, S. and Vashisth, S. (2020). Hype Cycle for Artificial Intelligence, 2020.

Simon, J. (2017). Building FPGA applications on AWS — and yes, for Deep Learning too — by Julien Simon — Medium.

Soboczenski, F., Himes, M. D., O'Beirne, M. D., Zorzan, S., Baydin, A. G., Cobb, A. D., Gal, Y., Angerhausen, D., Mascaro, M., Arney, G. N., and Domagal-Goldman, S. D. (2018). Bayesian Deep Learning for Exoplanet Atmospheric Retrieval.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. In *57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sun, Y., Agostini, N. B., Dong, S., and Kaeli, D. (2020). Summarizing CPU and GPU Design Trends with Product Data. Technical report.

TechPowerUp (2020). NVIDIA Tesla K80 Specs.

TechPowerUp (2021). NVIDIA GeForce RTX 2080 Specs.

van den Besselaar, P. and Leydesdorff, L. (1996). Mapping change in scientific specialties: A scientometric reconstruction of the development of artificial intelligence. *Journal of the American Society for Information Science*, 47(6):415–436.

van Well-Stam, D., Lindenaar, F., van Kinderen, S., and van den Bunt, B. (2011). *Risicomanagement voor projecten*. Unieboek, Houten, third edition.

Verschuren, P. and Doorewaard, H. (2010). *Designing a ReseaRch PRoject*. Eleven International Publishing, The Hague, second edition.

Vijay, U. (2019). What is epoch and How to choose the correct number of epoch.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., and Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals.

Wang, Y., Wei, G.-Y., Brooks, D., and Paulson, J. A. (2019). Benchmarking TPU, GPU, and CPU Platforms for Deep Learning. Technical report, Harvard University.

Watts, J. (2018). We have 12 years to limit climate change catastrophe, warns UN.

Wolff Anthony, L. F., Kanding, B., and Selvan, R. (2020). Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. Technical report.

Xilinx (2020). What is an FPGA? Field Programmable Gate Array.

Yin, R. K. (1984). *Case study Research: Design and Methods*. SAGE Publications, Newbury Park, 1st editio edition.

Yin, R. K. (1998). The abridged version of case study research: Design and method. In *Handbook of applied social research methods.*, pages 229–259. Sage Publications, Inc, Thousand Oaks, CA, US.

Yin, R. K. (2018). *Case Study Research and Applications: Design and Methods*. SAGE Publications, Los Angeles, sixth edit edition.

A SEMI-STRUCTURED QUESTIONS

This appendix gives an overview of the questions for the semi-structured interviews. The interviews themselves will deviate, as other or additional questions might occur more relevant during the interviews. Below is the list with questions first presented in English and followed by a Dutch version for Dutch interviewees.

A.1 ENGLISH QUESTIONS

Project related

- What was the question of the client?
  - What was the problem the client addressed?
  - What were the requirements of the client?
  - Where there any additional wishes?
  - Were there any conflicting interests in the request?
  - To what extend did the client mention the energy consumption of the solution?
- How would you describe the client?
  - What is the sector of the client?
  - How much experience had the client with deep learning models?
- What stakeholders were involved in the problem?

Model related

- What was the purpose of the model?
  - What deep learning applications are applied? (e.g. NLP, speech to text, or object recognition)
- What were the design considerations?
  - What was the neural network architecture, and why this one?
  - Is the model designed from scratch or with pre-trained models?
- What was the input data?
  - What were dimensions of the input data?
- What was the desired result of the model?
  - What was the desired accuracy?
- What was the training time of the model?
- What information was available from the service provider for training the model?
  - Which service provider is used?

    - What information was known about the data center that was used?
        * Was the Power Usage Effectiveness known?
        * Was the location known?
    - What information was known about the hardware that was used?
        * What hardware was used?
        * What was the computational power of the hardware? (E.g. in TFLOPS)
        * What was the efficiency of the hardware? (E.g. in GFLOPS/Watt)

- To what extent is known how much energy is consumed by the model?

A.2 DUTCH QUESTIONS

Project gerelateerd

- Wat was de vraag van de klant?
    - Wat was het probleem dat de klant aandraagde?
    - Wat waren de eisen van de klant?
    - Wat waren additionele wensen van de klant?
    - Waren er tegengestelde belangen in de aanvraag?
    - In hoeverre was de klant geïntresseerd in de energy consumptie van het model?
- Hoe zou u de klant beschrijven?
    - In welke sector is de klant actief?
    - Hoeveel ervaring had de klant met deep learning modellen?
- Welke stakeholders waren betrokken bij het probleem?

Model gerelateerd

- Wat was het doel van het model?
    - Welke deep learning vormen zijn toegepast?
- Wat waren de ontwerp overwegingen?
    - Wat was de neurale netwerk architectuur en waarom is voor deze gekozen?
    - Is het model ontworpen vanaf niks of gebasseerd op een voor getraind model?
- Wat was de gebruikte data?
    - Welke dimensies had de gebruikte data?
- Wat was het gewenste resultaat van het model?
    - Wat was de gewenste preciesie van het model?
- Wat was de training-tijd van het model?
- Welke informatie is beschikbaar vanuit de serviceprovider voor het trainen van het model?
    - Welke serviceprovider is gebruikt?
    - Welke informatie was bekend over het gebruikte datacenter?

        - Was de gebruikte Power Usage Effectivess bekend?
        - Was de locatie van het datacenter bekend?
    - Welke informatie was bekend over de gebruikte hardware?
        - Welke hardware is gebruikt?
        - Wat was de computer kracht van de hardware? (bijv. in TFLOPS)
        - Wat was de efficiëntie van de hardware? (bijv. in GFLOPS/Watt)
- In hoeverre is bekend hoeveel energie is verbruikt door de model training?

B ENERGY ACCOUNTING OF THE PROJECTS

The method as formulated in Chapter 4 to calculate the energy consumption of DL models is:

$$e_{total} = PUE \frac{P \times T_{run}}{1000} \tag{B.1}$$

With total energy (e_{total}) in kWh, Power (P) in maximum Watt considering the TDP of the main processing unit, PUE dimensionless with a default of 1,58, and the run time (T_{run}) in hours.

This Appendix calculates the energy consumption of the slaughtery and asphaly damage project, since only these two projects have sufficient information to calculate an estimation.

B.1 SLAUGHTERY

The location of the hardware is a Data Center of Azure in West Europe. This data center is located in The Netherlands and build in 2010 [Microsoft Azure, 2021]. In 2015, Microsoft published a fact sheet where it stated that the average PUE of data centers of Microsoft is 1,125. Although the construction and the determination of the average PUE differ 5 years and it has been over 5 years, it is assumed to be an indication of the correct PUE. The hardware used for the training was a NVIDIA Telsa K80. The TDP of the GPU is 300Watt [TechPowerUp, 2020]. This is assumed to be the constant power draw of the GPU. The total training time of the different architectures is 360 to 672 hours, so these two extremes are calculated. The estimation of the energy consumption of this model training is:

$$e_{total} = 1,125 \frac{300 \times (360 - 672)}{1000} = 122 - 226 kWh \tag{B.2}$$

The energy consumption of 122 to 226 kWh is relatively low, and comparable with the average yearly energy consumption of a LED television [Milieu Centraal, 2021a].

B.2 ASPHALT DAMAGE

The location of the hardware is at Arcadis, so the PUE is assumed to be the default, namely 1,58. The hardware used for the training was the GeForce RTX 2080. The TDP of the GPU is 215Watt [TechPowerUp, 2021]. This is assumed to be the constant power draw of the GPU. The total training time of the model was assumed to be little under 250 days, so little under 6000 hours. The estimation of the energy consumption of this model training is:

$$e_{total} = 1,58 \frac{215 \times 6000}{1000} = 2.038 kWh \tag{B.3}$$

The energy consumption of 2.038 kWh is significantly higher than the energy consumption in the slaughtery. It is the equivalent of the average yearly energy consumption of 8,5 fridges [Milieu Centraal, 2021b].

C OVERVIEW OF THE GROUPED RESTRICTIONS

The restriction in all the interviews are coded and analysed with ATLAS.ti. This first section briefly describes the restrictions that are identified in the case studies and presents the identified codes that are underlying to the grouped restrictions. The second part shows a table with all identified codes and the grouped codes that resulted into the restrictions presented in Sections 6.2 and 7.2.

C.1 CASE STUDIES' RESTRICTIONS

In the table on the next page are the 37 codes found in the four case studies. These 37 codes are clustered into 'Code Groups'.

Table C.1: Overview of the restrictions found in the case studies with descriptions.

Restriction	Description
Complexity of deep learning	Deep Learning is a complex technique that is not always fully understood, which makes it hard to trace its energy consumption.
Innovative stage of deep learning	Deep learning is still a new technique and is still changing over time. New techniques are continuously developed.
Lack of incentive to determine energy consumption	Different stakeholders have a lack of incentive to account the energy consumption of deep learning models.
Lack of societal awareness	Lack of awareness in society and at the client about energy consumption of training DL models.
Lack of systematic evaluation method	There is no clear to goal to train the deep learning models and to determine once the models are done training.
Long and diverse training time	The training process is a long and often never-ending process, resulting in a changing energy consumption.
Model developers' energy accounting knowledge	Model developers have no or insuffent knowledge about how to account the energy consumption or on how to reduce the energy consumption.
No hardware details available	Lack of details about the hardware that is used to train the models.
Updating model over time	The scope of the model changes over time, which can make it non-transparent what the energy consumption is of what part of the model.

Table C.2: The codes identified in the case study interviews, grouped into 9 categories. Codes that appear in multiple groups are marked with *.

Identified code	**Code Groups**
Lack of knowledge about what architecture to use Explainability for regulator Understanding of the DL technique Complex model landscape Lack of knowledge about DL of project manager	Complexity of Deep learning
Confidentiality of the project Fast evolving DL technique Too innovative technique	Innovative stage of DL
Lack of priority at energy consumption	Lack of incentive to determine energy consumption
Lack of society's awareness Lack of knowledge of the client about DL Lack of client awareness Client's anxiety about new technology	Lack of societal awareness
Unknown goal for model Lack of evaluation tools for model results Distributed information in the project Evaluation by different (kind) of people	Lack of systematic evaluation method
Extra training for changing model Pre-trained models often needs additional training Training of many different configurations Training tries to improve Long initial training time Never-ending training Proper training data	Long and diverse training time
Lack of modeler's awareness Lack of knowledge about energy accounting Lack of knowledge about energy efficiency of models	Modeler developers' energy accounting knowledge
Lack of knowledge about hardware Non-transparancy of Virtual machines Lack of knowledge about datacenter Layers of infrastructure Distributed information about hardware	No hardware details available
Different functionalities with different compute Different needs of different customers Changing input data Changing model complexity	Updating model over time

C.2 IN-DEPTH INTERVIEWS' RESTRICTIONS

Table C.3: Overview of how all identified codes are grouped into final restrictions.

Identified code	**Code Groups**
Understanding of the DL technique Lack of knowledge about DL of project manager Lack of knowledge about what architecture to use Experts are better at choosing hyperparameters Complex model landscape Flexibility of deep learning makes the technique complex Explainability for regulator	Complexity of deep learning
Need for justification within service provider Incentive of datacenters to rent as much servers as possible Conflicting incentives within service providers to account No incentive to reduce CO2 footprint at Alibaba Other pricing model of service provider that only provides run-time Lack of priority about energy accounting at service providers Conflicting interest of service provider to rent many servers and optimize server use Mixed interest of serviceproviders to optimize servers and use as much Basic pricing model of service provider is only server per hour	Conflicting interests at service provider & data center
Too innovative technique New more complex deep learning techniques are applied Confidentiality of the project Fast evolving DL technique Lack of information about energy consumption service provider indicates thats in a young stage	Innovative stage of deep learning
Energy accounting data exists just not clear where it is Incapabel of measuring energy consumption Lack of existing method for energy accounting at service providers Lack of reserach into energy accounting by service provider	Lack of energy accounting tools at service provider
Governmental focus is on energy efficiency measures* Lack of tools for government to enforce measures Subsidies do not take into account the side effect of energy consumption Little awareness at government about the energy costs No laws or subsidy to reduce the energy consumtpion of software Lack of consistent method for government to compare energy consumption	Lack of governmental enforcement tools

Identified code	**Code Groups**
Lack of awareness of power cost in general Lack of awareness about energy consumption of server use No incentive from hardware restrictions to limit lines of code Lack of priority at energy consumption No priority at green software Lack of awareness about importance of energy consumption	Lack of incentive to determine energy consumption
Lack of awareness about energy cost of more accuracy Lack of knowledge about energy accounting Lack of modeler's awareness Lack of knowledge about energy efficiency of models Reporting of numbers that do only partly relate to energy consumption	Lack of modeler developers' energy accounting knowledge
Corrupted data is processed different, but unclear how much energy that consumes Implementation of the model has a big effect on the energy consumption, but is often not reported Even simple model have complex energy accounting Lack of tool to account energy of DL models Existing scientific tools only provide estimations GPU optimizations are different on different GPUs Lack of understanding of easier estimates Tool of Lacoste as a best estimation Energy consumption is very dependeing on hardware and implementation Input data combined with memory and architecture all have its effect on the energy consumption Implementation of the framework has a big and non-transparent effect on energy consumption Only average energy consumptions used.	Lack of scientific tools to account energy consumption
Governmental focus is on energy efficiency measures* Little attention about the effect of efficient software in Amsterdam economic board Little societal attention for green software No societal awareness about energy consumtpion of computers Lack of knowledge of the client about DL Smaller clients have less priority at energy accoutning Lack of society's awareness Lack of client awareness Client's anxiety about new technology The scientific community is not really aware of the energy consumption of deep learning No current demand for energy accounting of cloud services	Lack of societal awareness

Identified code	**Code Groups**
No clear agreements on who reports what energy numbers Evaluation by different (kind) of people Lack of evaluation tools for model results Unknown goal for model Distributed information in the project Different invoices from service providers for different services Lack of standardized energy performance	Lack of systematic evaluation methods
Lack of transparency on power cost of cloud services by serviceproviders Biased information from service providers about energy consumption Service providers not willing to share data about energy consumption Very little information from service provider about energy consumption	Limited information provided by service providres
Pre-trained models often needs additional training Never-ending training Different re-training needs for different models Extra training for changing model Long initial training time Training of many different configurations No reporting of previous experiments Proper training data Training tries to improve	Long and diverse training time
Non-transparancy of Virtual machines Distributed information about hardware Lack of knowledge about hardware Limited access to energy accounting data Lack of information on higher scale Increased complexity of using cloud provider Organizational constraint of dispersed data at service provider Incomplete information about training due to confidentiality Layers of infrastructure Lack of knowledge about datacenter The ICT infrastructure is non-transparent about energy consumption. Only limited information is available from the hardware Increased complexity at a higher level	No hardware details available
Science uses numbers that are unavailable in practice Difference in focus and jargon between science and practice Seperation between the scientific world and practice	Seperation between science and society
Changing model complexity Different needs of different customers Changing input data Different functionalities with different compute	Updating model over time

D QUANTITATIVE OVERVIEW OF RESTRICTIONS

This appendix present the quantification of the identification of the restrictions in the case studies and the validation of the restrictions in the in-depth interviews.

D.1 CASE STUDIES' RESTRICTIONS

This section presents the quantification of how often each restriction is identified in the different case studies. Most of the restrictions were identified implicitly, since the case study interviews aimed to discover what information is available to account the energy consumption of training DL models (see Appendix A for the questions).

Table D.1: Overview of the restrictions in absolute numbers per case study.

Restrictions	Project 1: Slaughter house	Project 2: Bank's credit	Project 3: EPIA	Project 4: Asphalt damages	Totals
Changing model over time	0	4	0	2	6
Complexity of Deep learning	4	9	0	2	15
Innovative stage of DL	1	10	4	0	15
Lack of incentive to determine energy consumption	0	1	0	2	3
Lack of societal awareness	4	2	1	0	7
Lack of systematic evaluation of models	0	10	2	2	14
Long and diverse training time	2	8	3	1	14
Modeler developers' energy accounting knowledge	6	4	2	1	13
No hardware details available	2	6	10	0	18
Totals	19	54	22	10	105

D.2 ADDITIONAL RESTRICTIONS

This section presents the quantification of how often each restriction is validated by the different interviewees. This could be implicitly or explicitly. The first two interviewees are the governmental perspective, the second two interviewees are the scientific perspective and the final two interviewees are the service provider perspective. More details about their roles and organization can found at the transcripts of the interviews, in Appendix E.

Table D.2: Quantification of how often the different restrictions were validated explicitly and implicitly by the interviewees.

	1. Hartkamp	2. Mes	3. Strubell	4. Van Gemert & Pintea	5. Anonymous	6. Van den Bosch	Totals
Complexity of deep learning	0	0	2	2	3	0	4
Conflicting interests at service provider & data center	2	0	0	0	3	4	2
Innovative stage of deep learning	0	3	0	1	2	1	4
Lack of energy accounting tools at service provider	0	0	0	0	5	1	0
Lack of governmental enforcement tools	5	2	0	0	0	0	7
Lack of incentive to determine energy consumption	4	0	2	0	1	0	6
Lack of modeler developers' energy accounting knowledge	0	0	3	0	1	0	3
Lack of scientific tools to account energy consumption	0	0	3	11	0	0	14
Lack of societal awareness	5	2	1	1	1	2	9
Lack of systematic evaluation methods	0	0	1	0	0	2	1
Limited information provided by service providres	0	0	0	0	1	3	0
Long and diverse training time	0	0	3	0	0	0	3
No hardware details available	2	0	6	1	1	1	9
Seperation between science and society	0	0	0	3	0	0	3
Updating model over time	0	0	1	0	0	0	1
Totals	18	7	22	19	18	14	66

E | TRANSCRIPT OF THE CASE STUDIES

This appendix contains the transcribed reports of the interviews that are conducted for the case studies. The table below presents an overview of the different interviewees, their roles, and the projects they worked on.

Table E.1: Oversight of the interviewees of the case studies.

Name	Case study	Role
Vincent Bolwerk	Situation recognition - Slaughterhouse	Model developer
Anne-Louise Meijer	Situation recognition - Slaughterhouse	Project manager
Dirk Boersma	Credit provider - Bank	Technical project manager
Ernst Fluttert	Credit provider - Bank	IT architect
Andreas Soderlund	Product Information Assistant - Ericsson	Model developer
Jasper Keij	Asphalt damage detection - Arcadis	Model developer

The interviews are transcribed individually and in this appendix grouped per case study. At the individual interviews is noted whether the original interview was conducted in English or Dutch and translated afterwards.

E.1 SITUATION RECOGNITION – SLAUGHTERHOUSE

E.1.1 Model Developer

The original interview was conducted in Dutch and is translated to English.

The interview:

The project was of a customer who wanted to monitor the animal welfare of the animals in slaughterhouses using video cameras that hang in the slaughterhouses. These cameras hang in the slaughterhouses for an animal welfare certificate of the meat by the slaughterhouse to customers. However, in general the camera footage is not used, since it takes too much man hours for the client to check all footage. Therefore, the client asked for the project weather this camera footage could be monitored more efficient with the use of Artificial Intelligence. An example of an undesired situation that needed to be detected is when an animal gets excluded from the group and experiences a lot of stress. So, this needed to be identified.

The customer was a special alliance between a slaughterhouse and an animal welfare organization. Because of the nature of the two parties there were some conflicting interests, but it was a joined project. The customer had no experience whatsoever with Artificial intelligence or Deep Learning and therefore asked Deloitte to help.

The aim of the model was to identify undesirable situations in the slaughterhouses, as mentioned before. The model consisted of several layers of models. First of all, different objects were recognized and with the use of transfer learning the model was 'thought' what objects needed to be filtered out of the video footage. This was not fully possible with a pre-trained model, since not all objects could be filtered with the pre-trained model. However, it was used to filter what was and was not interesting for the model. Next, a model is trained to distinguish the animals from other objects in the footage. Second, The model filtered undesired situations from normal situations by recognizing that, for example, one of the objects (or animals) was les behind from the other objects. Finally, the undesired images were uploaded to a platform to check for employees.

For the architecture, different architectures have been tested to check which one delivered the best results. This have been 6 or 7 different architectures until the 'best' one was found. Eventually the best performing architecture was Faster R-CNN ResNet101 within TensorFlow of Google. For each architecture, 5 or 6 different configurations have been tested. So, eventually about 40 different versions have been tried before was decided to go with the Faster R-CNN. Each version or configurations ran between 12 and 16 hours and the models ran until a number of iterations were completed. So, the length of the training was based on the number of iterations and the performance was evaluated based on the Mean Average Precision with Intersection over Union (IoU) evaluation with a threshold of 0.5.

For training the model, a virtual instance of Azure is used. The instance used is: STANDARD_NC6. It was possible for them to select the region of the datacenter in which the hardware of the virtual machine was located. They selected West-Europe and after checking it online, this data center is located in Belgium. It is also known what hardware was used to power the virtual machine. They used a NVIDIA Tesla K80. This equipment consists of two GPU's, but for the training they and the virtual machine, they only used half a NVIDIA Tesla K80, so only 1 of the 2 GPU's. About the energy consumption or the energy efficiency of the datacenter was no information available. Also not the Power Usage Efficiency (PUE) of the used datacenter. He indicates that this information might be accessible online, but he did not noticed it anywhere or searched for it online.

Besides the run-time of the model is not a lot known about the energy consumption of the model, since it was not noticed anywhere. Eventually they received a bill from Azure which indicated the hours of compute and this number of hours provided some indication of the order of magnitude for the computational power, but no other details.

To evaluate on the decisions that have been made during the project. During the project there were not a lot of aspects that could have been more efficient to safe energy. The costs for training the model was minimized and therefore is the GPU not excessively used. For the virtual machine, the hardware and GPU were not constantly used, occupied or reserved. Only when a new architecture or configuration was ready, the virtual machine was turned on and the model was trained.

When the data center would have stated more clearly what the energy consumption was, for example per hour, then they would probably not have executed less iterations, less runs or less configurations, since these were all necessary to achieve the final result. There was no short-cut to retrieve the information.

An option that might be possible is to compensate the energy consumption and carbon footprint of the model and to charge these costs to the client to increase the awareness of the client about the question or project they requested and the effects of this question or project. It can also be used to make other model developers more aware about the effects of using Deep Learning or Artificial Intelligence models.

E.1.2 Project manager

The original interview was conducted in Dutch and is translated into English.

The interview:

During the project a tool was developed, an AI tool, which Deloitte developed themselves. So that's an asset that didn't exist yet of which the assignment manager saw that there was a gap in the market and that there wasn't yet a smart way of using cameras from slaughterhouses. So, these cameras record images, but are randomly viewed and checked. Now she wants to make a smart camera out of that, which didn't exist yet. And since you obviously have to develop a product before you go to customers and try to sell it, you have to develop it first. So, first is within Deloitte impact foundation that asset tool developed. It was checked several times within the mill that you go into. Adjusted again, checked, and when minimum variable product came out of that, the partner, who helped us develop the tool at the time, then became the customer. So they were happy with the product that was developed and they said, okay, now I would like to see it implemented. So we implemented AI-driven surveillance cameras at the customer and made sure that the adoption of the tool went well, that employees actually use it and that above all the goal is also achieved, which is ultimately improving animal welfare within a slaughterhouse. So you not only want the tool to be used or people to understand how the dashboard works. But you also want them to make the necessary changes afterwards if they see that they are not complying with animal welfare rules. So that was also an important point in the project, because it's really nice that it all works. But how do we ensure that it actually improves animal welfare? Because that is the whole reason we are doing this.

There was no question from the customer, since there was no customer. This was an idea that a partner had. The partner of the project in question is vegan himself and very involved in animal welfare and animal rights and so on. He suddenly got the idea. Why don't we do something with that? Because images are regularly leaked in the media about what goes on in slaughterhouses and often quite violently. So you see that people working in slaughterhouses are mistreating animals or treating them in a brutal way. And then he looked into it. He looked into: How does the whole process work now? And what is the control that is done by the government? And it was just minimal. So he actually saw more of an opportunity on the market to develop a tool like that.

And how the first process goes; you do need to have a partner within Deloitte for the development, because it's not just the camera, the camera also needs to be taught through machine learning when a deviation occurs, when does animal welfare come into play, and for that you do need a partner to actually make room available to develop it and that partner eventually became the customers.

There were in general no additional requirements for the project, because it was really in the development phase. The idea though is that this project and this tool can be used on a lot of different things and at least animal welfare. I know that also an idea is emerging to maybe implement this at the airport to see what the queues are on the runway. It's just more actually the broad deployment of smart cameras and that can be for any purposes. And if we're going to exploit that then I think customers themselves will come up with a requirements list of what they want it to meet.

The implementation at the slaughterhouse went well. As with many projects, but I think this is really the classic textbook example of workers who are afraid that technology is going to take over their jobs, because it's also a lot of practically trained people who are doing those jobs. On the other hand, it's really just another example of AI, the machine thinks for itself and takes work off your hands. There was some resistance in that respect, so you have to introduce it very well. Mainly, we put the focus on animal welfare and showed a lot of understanding and indicated to that employee: we understand that you have to watch half an hour of camera footage a day then, and this will not change. They're still watching a half hour of footage a day, but instead of randomly grabbing a half hour clip. They get seven times eight minutes of footage, where maybe something animal-unfriendly happened. So you make much more efficient use of the time they were spending on that task. Implementation in any project you have to apply and tailor to your target audience. Those are people who are very used to routine work. And haven't studied very long, so you just have to explain very little by little and in the end that went well. But we've actually all figured it out now.

But we have now actually found out in the final phase of the project that the most exciting part is whether they are going to use it in the right way. Such a training is very nice, but still there is the doubt about is it arrived right, are we going to have them use it the right way. Because at some point you can then assign actions if you see something that's not okay that happens and you see person A doing that. So then you can kind of put tag there. And then people sign of: I see here that this and this is going wrong and then you tag for example a manager and say take action. They started to see it as a game. Now I'm going to tag that, now I'm going to tag that. It's very difficult to control when you deliver a project like that. The adoption of a new tool just takes a very long time. It's not a matter of a couple of training sessions and then you can't assume that it's going to do well. In fact, we're working on that right now with after the track design where we're checking in so maybe monthly or weekly to see if everything is going well with used and the use of the tool is also good.

The slaughterhouse was involved in making the model, in the sense that they had the camera footage.They are mainly involved in mapping out the process that is being filmed and also explaining to us, where there might be some critical points, things that you might not think about yourself, because you don't know the process well enough. I don't know exactly what you know about the context, but so it's about pig slaughterhouses. I was honestly surprised at how many pigs are involved, so many. Daily that trucks are unloading pigs on the assembly line. So things where you think, the pigs you want to get from A to B and then you want to get them there as easily and as well as possible. But in practice: in ten minutes the next load will be ready and another hundred pigs will be unloaded from a truck through a narrow corridor, so in ten minutes that corridor just has to be clear. We wish we had time to supervise all the pigs properly and with respect for animal welfare. But we don't have that, because in 10 minutes there will be another truck. So you really need the customer very much to understand the process properly. Because you have to understand it completely, from A to Z, if you really want to develop an efficient tool for this.So we were also very closely involved and in daily contact with people like Vincent, who created the model.

When the project was set up, there were many stakeholders involved, especially with animal welfare from their perspective. So animal protection. IKEA hooked up at one point because IKEA Food solutions is so big with their food branch. That's just super big. So for example, all the Swedish meatballs that they have on the shelves really come from their own slaughterhouses. They just really have clear suppliers, are quite a frontrunner on that, and they did have ideas. IKEA is more like a sparring partner to exchange knowledge and Eyes on Animals is also a nonprofit organization that deals with animals. Specific more on slaughterhouses. Because animal protection is of course very broad and Eyes on Animals is more concerned with animals in the food supply chain. Those were still important stakeholders at the time it was developed. But when you enter the implementation phase - and you start dealing with how does regulation fit into the picture, then you're with the NVWA, so that's the food and commodity authority. And then you're looking at the government as well. You also need to have contact with those people in order to be able to say "we're working on this" and to see if it's possible to ensure that the NVWA carries out more frequent inspections or carries out inspections in a different way at slaughterhouses. So government, especially from a regulatory point of view.

It was the partner's initial idea to use deep learning for this project, but he didn't know anything about it. The funny thing was, he read an article that was about the future of the food chain and consumer behavior when it comes to buying food. In it, someone outlined in that article: what if we applied AI surveillance in slaughterhouses later on. Then he took that concrete idea and thought doesn't it already exist then and then he did a lot of research. Then he said: if nobody is doing it, then why don't we do it. So actually the idea was kind of proposed, and only after he started doing research after that did he find out that it just didn't exist at all yet.

Probably the energy consumption of the model is not mentioned by the partner in the initial idea, no probably not. I do think it's a good question. He does think things through very well, so I can hardly imagine that that didn't occur to someone somewhere. That may have gone through his head very quickly. Maybe it did, and then he ended up switching with people from Deloitte technology. I don't dare say no and I don't dare answer for him either. Let's put it this way, if it was even a little known that it is a thing. I speak for myself for me it is not known, maybe for other people. Then I can't imagine, that that doesn't go through your mind. That you don't take that into account somewhere.

To the extent that I know what the model does is that model it has been taught to recognize objects on image and distinguish between what is happening on image and then it has been taught what is an image that you want to see so is a good situation and what is potentially a wrong situation. Based on animal welfare, because of course you can set that, or make that different.

My role within the project is project manager. The project consists mainly of people from analytics and cognitive working on the model, so also someone working on the dashboard, a few more senior partners. And what my role in that is actually to coordinate everything. So I lead the weekly stand-up meeting of the whole team and all the things that come along that are not IT/tech related, just to call it very broad. So press release, communication, a bit of adoption you could also call change management. I also created the training kit for the end users and facilitated the training sessions for the end users, so basically all the content that we have that Vincent and other colleagues have created. Translating that into something that is understandable for an end-user who has no knowledge at all about the model and drawing up a bit of an adoption plan of how are we going to ensure that they make good use of it and a bit of aftercare as well. More of the soft skills so to say and mainly contact with the client.

About the energy consumption of the model I find it hard to say anything, how it works I don't really know exactly. In my mind it's something that all takes place on the cloud, so it's not in a box somewhere. I actually wouldn't know what the energy consumption of software is, I'm sure it's a lot. But I wouldn't know exactly how much that is and how that translates. I do know that, for example just in terms of storage space for images and maybe also for the model, that we do pay something for that and that's really a monthly fee. So you, you do pay for that space.

E.2 BANK'S CREDIT

E.2.1 Technical project manager

The interview was conducted in Dutch and is translated to English.

The interview:

I'll just see if I can summarize the project nicely. Actually the question from the party where this came from, and that's a bank, says the customers' need changes. Customers want to get in touch with their bank more often from a phone or an iPad or something similar and do their banking from there. What this was about: Actually, customers also want very easy access to credit, to a loan. And we, as a bank, want to go a step further and realize the wishes of today's customers, but also with a glance at the future. That actually means that you, as a bank, have to ensure that you get into the environment of that customer, so really get on his or her phone, get on his or her tablet. That it is possible for customers to know this credit themselves without too much fuss. If you order something online, you're also happy when you're not asked what you want, but just asked a few questions. Yes, that's actually what the landscape is all about now, that you say as a customer at a bank you just want to (i) build a relation with your bank, (ii) assume that the banks know a lot about you, so you don't want to be asked twice, and (iii) you want to receive a good offer, an appropriate offer. That has actually been the goal of this project to realize that for that bank.

Specific requirements of the customer were speed. Within fifteen minutes the client had to know whether he or she could borrow money, which is different from having money within fifteen minutes, but in fifteen minutes the client wanted clarity, can I have this money, yes or no, that was actually the single dimension. The most important one.

About the other requirements that came into play in the project. You have to imagine that it is a very large project on which about 100 people work at its peak. And if we move on to the model, which is actually a subsection of such a project. You have to imagine that part of the project also involves a lot of designers designing a website with the orange button or the yellow button on the left, at the top or right below, etc. But when we look to this piece of relevant model, you actually look at: What is relevant in such a type of trajectory is? It is relevant to serve a purpose with a model, a purpose that can serve the client and the bank as well obviously. You want the model not to be a delay because you only have 15 minutes. The model has to be quick to run and 'quick' means we talk about Nano seconds.

So, the turnaround of that model has to be super-fast. It also has to be easy to implement in the architecture, but ultimately the most relevant is the data collection that the model needs. This data collection has to be able to come naturally from your solution.

To give an example, if the model needs you to fill in your 12 ancestors, it doesn't help you. You probably have to search for that information for half a year before you can run this. So, the model has to be able to run on data that is hands-on available to you. That's either because you have it, or because the bank can already use it, with your permission.

Unfortunately, I can't say more about the bank that set up the project, and I can't say how much experience the bank has with deep learning models. However, they did need additional knowledge. That's why Deloitte was approached for the assignment. You see anyway that a lot of deep learning models are currently created in collaboration and you see deep learning models very much still in living lab environments. Anyway, this was one of the first times that such a model was used in a real process.

In the client's request, energy consumption was not mentioned. However, as indicated in the previous conversation he thinks the idea is very interesting, although he never thought about it.

If we would have had to take it into account, I would also find it difficult to be able to act on it. I would not directly know how to reduce it, since I don't have the tools for that. I can figure out a lot about the model, but I cant monitor the energy consumption. I don't know what adjustments affect the energy consumption in what direction.

The project involved the customers, the bank, and Deloitte. Other than that, there are no additional external parties for information or anything similar. I think that's also one of the advantages of Deloitte, there is so much knowledge available. Unbeatable for other parties.

About the model, before was indicated that the model needed to be able to offer credit to the customers within a short time. To achieve this, the model consists of different forms of Deep Learning, a combination of different aspects. One aspect is natural language processing. Another is called XG Boost, which is a sub-form of random forest. Those two are the main drivers of the model.

For the design considerations, there were a couple of key things. As mentioned, one of them is that the model has to fit into an environment and another is that the model content has to meet quality standards of which one of the most important is the regulatory side. Also, an important factor is that the model needs to be explainable and understandable. It also has to be flexible, so it can be adapted over time.

About the explainability, we know how the model works. We cannot explain quickly the model for 100%, but we can see in the results that it's probably caused a bit more by one part and a less be another part. This already helps tremendously. For example, when we see the results of the model with X parameters, we can state that the result is mainly caused by a set of two or eight parameters. Because you want to work in a big process, of course there are many more stakeholders and the model is the engine of the solution. So, you have to be able to trust the model at all times and in order to trust it, explainability helps. If that is not possible, the environment will trust it less and if the model is not trusted, a model is not used. That was a condition of the client.

I think the most important thing in terms of data usage is that it should not contain an infinite amount of variables, which in turn means that you will probably end up with variables that are not available in your total solution. So you also want to work only with data that is available or expected to become available in the short term.

Unfortunately, nothing can be said about the architecture used.

For the input data, different data sources play a role. So, you have the bank's own data, you now have the transaction data available for the bank with open banking and some data that is collected elsewhere. However, elsewhere cannot be explained in more detail. To see if a model produced a good result, so when does that model actually work. To determine a good result, you have different possibilities of course. Just from the real technical operation you have a number of parameters that you can look at. For example how the distribution of False positives is divided.

However, again it is important that you come up with the people around the model to a score that is understandable by several people. Because you also want to avoid that the model keeps developing if an expert doesn't think it's good enough. If you're going to build a model in a total solution, it means that the model has to score just more than sufficient. But the difference between just more than sufficient and a ten is in work hours really a lot. So you actually want to look for that just more than sufficient or slightly better than that, but not for a ten. To do so, you need to involve people into the end-solution that can judge what is sufficient. If you only focus on the technical parameters, you won't get there. What actually is used a lot in the market for these models: you start looking at a parameter, e.g. the Gini coefficient, and then you're together going to determine some sort of threshold. What should be the minimum and that is what you are going to strive for. Eventually you're then going to determine based on this KPI and some technical KPIs whether the model is good. But when you look at how that model works, when you are going to work with models and certainly with those models. Those are never finished. So, you're going to keep updating after that. You create a job or multiple jobs and that job is to maintain and improve the model. But you want to keep an eye on whether it continues to perform and you do that on those KPIs. As mentioned. So its also a kind of risk analysis for the bank, what is an acceptable margin of error and the KPIs are adjusted to it. Also by testing on all kinds of datasets, but that's how you get there.

Regarding the training time of the model so far. The training time was about several week. A little more than 5 weeks eventually. 5 to 10 weeks for the initial and biggest start. But from there, smaller adjustments over time.

There is a protocol for the additional training of the model. However, this is related to the application of the model. So, on day one the model can be used for the initial idea, but on day 90 you might want to use it for more things. So, that triggers the need for another training, but purely due to change of scope and not because the model required additional training. That gets mixed up a little bit. So I can't say exactly when, in fact we always took advantage of the improvement moments that were scheduled. But that was more because the improvement moments were already there than because we had to train.

And finding the optimum training strategy is also quite difficult, because you don't want to train too much. Well, that's very easy to achieve, by not training. You don't want to train too little and that's very difficult to say what is too little, so to speak. The training of the model was done internally at the bank and of course, the data is stored in a solution. That solution ultimately belongs to an external party. What hardware is used to store that data, if I had to guess, I think I'd come pretty close. But I can't say for sure what hardware was used for training.

An intermediate platform is used for training the model and storing the data. Which platform it is cannot be mentioned. Exactly what hardware is underneath it is hard to say, but it can probably be retrieved. So, I do think it can be found. I thought I'll look it up briefly, but I can't find data right now.

But at least a service provider was not directly used for training the model. And of course, the model is trained in one environment and the end-solution is eventually stored in a different environment. So, the location of the end-solution is not the same as the intermediate platform. In training most data is consumed, since it used long series of millions of records of many years. So, that has been heavy, interesting to discover the different findings, and to combine the datasets. But that's in a different environment and les familiar to me.

However, I'm pretty sure that there is inefficient training there, if you look from an energy consumption perspective. That's not taken into account and I think that is in general very little done. You're the first person who brought it to my attention and I've been in this type of model for about 3.5 years now, so I've been in the larger companies since the beginning, but energy... I think the focus right now is still very much on can we do it. There's a lot of promise in AI and in those types of models. And people say, if something is impossible, they use those kinds of models. However, it turns out to be quite complex to do that; to collect the data to train and to say with the training outcomes that a model is a good idea. Because, the data you have often varies in quality over the years. So, is the training outcome representative or will reality be better or worse than the training runs. However, I think that people first need to have some trust in the techniques before one will do the next steps and will ask what does it do with the energy? But when we now say energy is one of the parameters to consider, that comes at the expense of development.

Its certainly interesting to look at energy consumption, because I think it runs out of control in the cloud. Also, I think it is also super interesting to talk again in a year and then also talk to some of my customers, to go a little deeper into that. I also think that my customers really like it to have this conversation. Especially with the tech heavy people, as they're not really thinking about the energy consumption. I did hold back a little bit on the details about the neural network. So if you're really totally stuck on these things and it was too high over, let me know. Then I might be able to go into more detail.

You see a lot more coming up now on the sustainability side as well. So you have some moving streams now, especially when I look in innovation. The one side is very much about searching a lot of information about people and that happens with the heavy processing. I call it the fun models, since really a lot of data is thrown at the model to see; this is good, this is bad, and lets try something else. The other stream focusses on the privacy part of the information and what can and cannot be shared. However, the sustainability story with the Paris energy agreement is one that nobody is paying attention to and nobody links sustainability and data yet. What you're doing is pretty unique in my world. I also think that it can be much more interesting for our customers as it is now.

E.2.2 IT Architect

The interview was conducted in Dutch and is translated to English.

The interview:

Interview Ernst Fluttert – Bank 2 So I've been working on the project for almost two years. That's a forward-thinking, innovative way of getting financial products to customers as quickly as possible. But then that's mainly about getting clarity as quickly as possible. How much can I borrow? What kind of product can I borrow? At what interest rate can I borrow it? This was very new, so the concept of doing it really very quickly was new. Before that it was manual. That could take number of days to weeks and we do it in 15 minutes. So super cool project. One of the important component in this is of course how do you calculate that and that's where a piece of AI comes in. So we use trained models. Well which one is that? I think it's gradient Boosted Forest. The exact name sometimes escapes me, but the Boosted gradient forest used for that. Then that model deployed. So it's not a model that continues to learn in production. Because it's only trained offline.

My role is: I am the Solution Architect on the project where I actually ensure that the entire flow can be carried out, so from front to back a piece of front-end, piece of back-end, integrations. How does the data flow, how do you make sure you are compliant, how do you make sure you remain secure? We did this by means of an event driven, micro services architecture, which is quite a mouthful. Basically what that means is we only work with messaging internally. Its not new, but special and cool way to work together. Basically, you can imagine it as, a long line of mailboxes. All you do is throw bills in and you don't know who's going to read those bills. It could also be that several people are going to read the bills or several services in this case and in this way you make sure that you get a kind of transactional system that is entirely based on events. In this way, we can follow very well what is happening and also react on certain input we get at the front-end and the second piece is we are cloud native, or cloud first I should say. For example, apply horizontal scaling. We also do that for the models for example.

How it goes with input from the customers. So you come as customers on the website and you just say "I need working capital". We don't say what products are connected and say hey. What's the reason? The borrowing reason that you want something, what you need money for. So it's also completely geared towards what the customer needs. So a customer doesn't have to decide for themselves: I want a business mortgage. But it could also be that they want a credit or a loan. So when you look at that, you're looking at, dear customer, what do you want to borrow it for, how much do you want to borrow, and then you can move on. So our whole model is also trained on transaction data, so that's a piece for the input, the models input and that's also where it's all bright and shiny new, and where that adds nice value. So that's how we make sure, that we can do pricing based on what people have done over the last few years.

In this way, a risk is determined and the corresponding prices. And then what kind of input can you expect. Well, what is your borrowing capacity, what do you have? What was your profit last year? Your depreciation last year, but also your bank account, upload so that can be done with PSD2 or MTf 40s. Do you want to put in collateral, for example, or not. And that could be if you're in a certain sector, or have a certain construction of your business. That is used to determine whether more questions are needed, to cover additional risks.

All that goes as input, we make nice summary of that and then that goes into the models. Models then say: We have a whole number of options, all the options are calculated through, so you can imagine that in some cases that you can choose five terms and within that term you can also shift something a little bit, that it can produce quite a lot of options. That's really exactly what we're using for as well. As a precondition, within the banking world compliance is super important. So everything we do also has to be validated, revalidated again and after you've done that you also have to prove once again that you were compliant. That's roughly how you can see it. Another precondition is just making sure you market well.

Look, the people themselves who take the customers don't see what we do at the back. Something that works very well for us. It's not just AI and then you get financial product. No, it's an offer, so I'm giving you an indicative offer. From this offer, we assume that we can offer and you have five days to contact us. And so then you also have the bank employee still contacting that person to validate that what they've filled out, that that's correct an accurate. There are, of course, a number of additional checks that are going to happen after that before you get any money in your bank account at all. So I think that bit of run-up, we've sped that up very much, but there are always a number of extra steps behind it, including face to face meeting. So that you're working with the right people, with the right information to deal with, but the offer remains. If that's all right.

On the considerations for the architecture, we looked at it from okay. We just want to have some meta data very simply. Exactly what data cannot be mentioned in detail.

And what impact that has on architecture. The architecture trade-off is to make sure that you have the most efficient journey for your customer and so you collect the right information to run models as quickly as possible. Because actually what we're doing is kind of, you think of it as a filter or a funnel where at the beginning you allow everything. And actually as quickly as possible you want to filter out companies that don't qualify. Because, those are obviously less interesting. That's also where the models help, because if you can say, "you may not be eligible," that you also return that as quickly as possible. And then that's often from go to the employee or something like that: architecture, wise. We do see that the models do tend to be heavier for certain products, for example, and so we scale those horizontally.

So we currently have four instances running of the models and they run in parallel in the cloud. What cloud cannot be elaborated on. It is also known what instance is used, but by naming is the cloud service revealed.

Because these are Python scripts they run on Python, of course all the way on a stripped down secured container and those are called via APIs. That's basically it.

I don't know what the training was done on and I do know the data underneath, but how exactly they train this. I don't know that.

This linked to the subject of the thesis. How do I make sure that with as few containers as possible, In our case, because we have everything containerized- as few services as possible run as much as possible. And when you run a service that you run it preferably with one instance instead of 16 instances, so to speak. And I think that's where you can find your gains in particular. That would be my first lead. If you really looked at it like hey - I want to know how this works, that would be it.

The main consideration to scale down the containers was the efficiency of the whole process. If you just have an efficient process. And this is also where you actually come up with what do you want to achieve in your journey. That you shouldn't build for the sake of building and I think we're looking very closely at What are really? The mandatory steps in the process that you have to go through to run a good process. If you do that well, then you actually need very few instances. For example, we have a functional domain, still specified in the journey, and per domain we have one, two, three, four, five micro services running, and they do a certain functionality together. And then you can think about, for example. I'm looking for a company because that's important for when you want to start in the Journey. You don't want to have tons of instances of that.

Because of course that would be a waste, so in that way I look much more at: how to make sure that the footprint of our application landscape, because we have a pretty complex landscape. But how can you make sure that it runs few containers, but offers maximum functionality. That's always a very difficult trade-off. For example, we've seen the models that were for us the slowest component.

And because that's the slowest component, we also said, well, we should have more instances of that that we run simultaneously. Because they are stand alone, it doesn't matter either. At which instance the input arrives, because the output is captured, and then we have the advantage of event driven architecture? That it also doesn't matter, because it's then packaged again as a message on the event bus and then that's picked up again by then subsequent service so. Precisely because we have decoupled that, you can also scale horizontally much more easily.

What I know about training the model. Training, what is done based on transaction data or basically a summary of your transaction data. So different scores are calculated. what we've seen is when we started the project - it was a very simple model and the first thing that happens is can you make sure that we can get all the input and output from production as well, so that we can train the model offline, where they can also take outliers themselves, but of course you don't want the model to change in production while you're working on financial products. It's not the most convenient thing, so offline training is just an important component in this. In addition to that, they've also helped at times. We do a kind of summary of the transaction data, and the models use that summary. We also looked at pieces of data together with them so that we knew what to look for. What do we have to pay attention to? That the summary is also correct. You can see that this is also done throughout the chain, everywhere, in order to ensure that what goes into the models is correct and subsequently what comes out is also correct.

And sometime in the last six months, the model was updated to support a little bit more products and sectors, and you also see now the model has become a little bit more complex. It's become a little more complex than it was and I think one of the challenges that you're going to see now is of course, what effect does corona have on those models, precisely because it's very based on transactions of the past year, in a normal year, that's much better. Then if you now for example if you are hospitality industry, I don't want to know what the model says then. I can't help you and then that's a consequence. And then I go a little bit too far on what the model should do. But you actually see continuous through the development of the model itself and a continuous offline through training as well. And I think that's necessary to stay sharp of what's happening? What are the trends in the market? Because if you train something in 2019 and you suddenly apply it now, you have a different picture of what is actually the reality. So we also have a continuous actually update to the models which we continuously implement. So when we get a new trained model - and then it's also immediately okay, that's then put live.

About updating models, there is a dedicated team on those models and of course there are more than just the model itself, so also a bit of input and output. Sometimes there is a mistake in that the wrong field is returned. that kind of thing actually I can say: every two weeks we make a change to the risk models, but because we also have three and in addition it is just an important topic within our whole engine, it is not so strange that that also just continuously ripples through. The same thing is with our cadence. Our cadence is at releasing every two weeks, so that fits in there just fine.

About evaluating the models, I know there has been a formal validation, also of those models. Of course, it was first a hypothesis for them when they started this two years ago. We're going to try something new - try it with AI see how that turns out. In addition, those models were also validated by the bank itself, an external team, and that was good and they had gotten the go from that. And that's actually as far as I know.

I just know that so certainly in addition to validating themselves, continuing to train the model, there's definitely an external there. I don't want to say auditors busy. But continuous validation has to take place, also because it's actually all still quite new and the models are not always well understood. But you also need quite specialized knowledge if you want to train good AI model and then also apply it.

Yes, so here you have another piece, so that offline validation and offline through training. That does get done. So from project we run, product we run online, because we literally have their model in the engine. What we do is make sure that they can get to that data. All the inputs and outputs from production are captured and those are then shared with that team, which makes those models and they train on that. So that validation, it does sort of semi training, semi validation, but just how they shoot it themselves.

About the information from the service provider about energy consumption. There's always the settlements of course, at the end of the day it just comes down to what credit card did you link and is there enough credit on there to pay. I think there's undoubtedly something on there somewhere about how many minutes which instance they ran, but basically for the landscape we say: we have twenty four seven always running one service so they're not all ladas, and we've applied that sometimes for really things that only occur very infrequently or just periodically, but because it's actually a website with that engine, you don't want to warm it up when a customer comes. When the customer comes, you have to be running already. So what we have done is: you always have one instance of a service running and where necessary screw up the instances and you could also apply dynamic scaling there.

To understand what I'm talking about I also have a picture of all the services that are running and how they are connected. But this is confidential and cannot be added. The point is that many processes can be scaled up with additional instances, but something always has to run because it is connected to the website and constant running works better.

E.3 ERICSSON PRODUCT INFORMATION ASSISTANT

E.3.1 Model developer

The interview was conducted in English.

The Interview:

Interview Andreas The project is EPIA, product information assistant, of Ericsson, which is recently to be released globally. It's a product that uses a lot of information available at Ericsson, when it comes to installation, troubleshooting, exchanging or upgrading the different software or hardware they have in the field. These can be radio transmitters, cable units, base band boxes, these things. Often, the engineers that change these things or work with them, work on heights. So, they don't want to bring their laptop up, where they have the PDF instructions available with instructions. They want to scrape and retrieve all the information stored in the database and want to make it available by speech or by typing on their phones (Android and iPhone). With the speech-to-text and text-to-speech but also different inputs, you can get lists, step by step instructions, and similar products to it.

The original problem they had was that the information was spread out. The technicians that worked for a long time knew their stuff very well, but the newer technicians may need to look up these details while far up in the mast. So, even with a senior technician available they should climb down the mast, check the information and climb back up. The discussion about the project started around 5 years ago and back then they had some high hopes about the technology, but only in the recent years it really kicked of due to new technologies. So, the language models, speech-to-text and text-to-speech they use now, were not available back then. What they required was that it had to be on premise within Ericsson and for clients to use it, it also had to be on premise and not on the Infrastructure of Google or Apple. 2 years ago Google released self-contained speech-to-text and text-to-speech module for English and that one is used. Now only English is available.

In general, energy consumption is a requirement for Ericsson projects, but for this project it was not a big deal. Normally, it is a requirement for the hardware they sell. For EPIA, the front-end works on a cellphone and the back-end on Ericsson servers. It hasn't really been a big thing. Response time for sending something to the server and getting something back is an important requirement.

The stakeholders. - The engineers were involved as tester in the feedback process to adopt their preferences to refine and develop new features to the product. - The Product owner is within his group. - Team lead - Some testers - Some experts on the data sources that are used.

For the deep learning applications that are used, they used a broad a set of NLP technologies. For the speech-to-text and text-to-speech they were as hands-off as possible on the models and relied on implementation of Android or iPhone and overall it worked really well. They mainly built a Knowledge Base model with the knowledge about the products and how they related to each other. By extracting a lot of information from the text documents and instructions of the different products on how to install them. That is more of a traditional implementation of NLP, the good thing is that they can be certain about the variety of the questions they got. As long as they put true statements in it, there will be true answers out of it. With the newer DL models, such as transformer, XLNet, BERT. Then they are able to ask questions about the text and then the result is a part of text in a document. So, then a section can be linked to a question. It took some time to train or fine tune the newer models and in the end it didn't really improved the results.

They tried several neural networks, but there is not 1 that worked best. Transformers, classifiers, etc. Classifier is more standard NLP. Take some text, train a classifier on it and get some output and then just send text through it. For the dialogue they use a combination of public available data in conversational data as well as their own conversational data to train and that is used to handle the dialogue pipeline. They have that in parallel with a final state machine going on. So, the combination means that they can both follow the steps well, but also be flexible if the user asks or states things in a different way, they can still response. For the DL part they trained and tried different models. We had different models in place and we just picked the one that seemed to got the best results of understanding their questions and test data. So, we set up a test data set of 20 to 50 different sections and a few questions for each and we just ran it through the model and see what the results were.

There definitely are ways of standardizing this. But then you would also need to create an expected response from that situation. And usually you would need an expert in the field, given the responses. And they just didn't have that. Well, at least they didn't have it for their dataset. There are test data sets out there, but for their own data, we just want to see what's was the best and it works really well.

They are on the second iteration. They used the previous model up until this summer. He thinks that model may have been from 2017 or something. A new family of NLP models came into place, so stepping up from fast text to transformer generation, they used one of the first of the transformer family. It worked well, but it was a bit large and slow and it didn't get the best results. So they switched to a second generation one from late 2018, early 2019 or something like that. That's the one they have in place right now.

They spend a few days on training the models.

The type of service was probably some kind of a cloud solution where they allocated a bit of C.P.U and a bit of speed for G.P.U. He can be quite sure that it was internal server. So it could have been run on a local machine, so like a laptop. But my feeling is that it would have been internal Ericsson cloud solution because that data couldn't leave the premises. Therefore there is no information available about the data center that was used by them, its was quite far away from them.

It's difficult to even know where to ask these questions for them in the project, they can be fairly sure about the G.P.U. and C.P.U of the allocated server or most likely a subset of a server. So like a virtual server or a virtual machine or not, they can be very sure about how much they have been allocated. But even if they would ask the one who allocated the virtual machine for them about these things. So what are the hardware underneath and these things, they would have to send it off to someone else to answer. So for them, it would be difficult for him to find that information even if they had a week or two and that was part of the requirements. Definitely in the end, they would be able to find if these are the hardware things going on and they run approximately do these things, maybe they have measured how much energy they consume. Maybe not. But there's definitely like there's an IBM or it's some Intel or something like that machine going, that kind of operating system on it. And they run this much according to the hardware information from the provider. But because they never run it on their own technology, well, in the sense of their own hardware, so Ericsson units, it's not as big of a concern.

So, one of the restrictions is the limited information and that the information is spread out. Also, when it comes to energy consumption in. It's quite tied to the processing time that they do so they have a model in place where it's possible to scale it to its dimensions. So it's using up most of the allocated processing power for it. And then it's quite tight. To the shorter time you use it, the less energy to use. So from their side, they would be able to say, OK, it's more energy efficient. And of course, the server standing somewhere, they're going to measure how much energy they use, how much electricity they use. And if they use way too much, they will start allocating differently and they will start asking people who run stuff on the servers. But that's that is those people are not the same ones that allocate the servers for us. They would need to start talking with each other there and saying, OK, but they have allocated these ones there. And then they would start talking to them. But in the end. Right. They have a certain allocated server space, and if they use too much of that, they either can ask for more. And if they say no, then they come back and say, OK, how can we make it more efficient? So he guesses that's where they can make it somewhat energy efficient in a roundabout way. They are in control of the server space, the virtual machines, and they, of course, have restrictions on there, even from like higher up in the company. They do have energy constraints, even internally. They have energy. And what they need to meet when it comes to energy consumption, once electricity they use, they‘re mainly towards their clients, but they also have it for themselves. So, if they would have servers that from really inefficient and these things, they would need to see how can we lower energy consumption? And then that could be a way. Then they would say, OK, your team, you cannot actually have 50 percent of the global server capacity, just your small project. You need to have just a small one. And if you want more than that, you have to pay for it. And that payment has to come from somewhere. They don't have the budget for that so that they are not restricted by the energy consumption, but they are restricted by what they are allocated and what they as a project can afford to use. So in that case, in that sense, they're concerned with energy, but as a project. They are more concerned about how efficiently that model can from what they're allocated and when it comes to energy consumption.

It is hard to make an estimation about the energy consumption of the training in the project. And cant really name numbers. About the server occupations: So that difference between production and training and testing these things, they have a production and testing at their well, production is allocated to us. I think testing is always allocated as well. Training, I think. What they have done is that they have used some allocated virtual machines that are allocated to our section rather than something that they have for their own project, because a lot of the training and testing of these things are small testing, especially since they never train a full transformer model on billions of data points.

E.4 ASPHALT DAMAGE RECOGNITION

E.4.1 Model developer

The interview was conducted in Dutch and is translated to English.

The interview:

The customer's question is: what is the status, the condition status of my roads in acreage? How good are my roads? Historically, you can determine that in a number of ways. Usually you just go out and see what the status of the road is. How good your road is. I can ask an inspection company to do that, but of course a road authority can also do that, they just want to know what a road looks like, but they usually ask an inspection company to map out the entire area. It used to be done by hand, by measuring and then looking. And I don't know when, a few years ago that was also done more by camera inspections, among other things, for national roads, provincial roads. Why camera inspections? Don't have to look along the road of what does the road look like? Is it good enough or not? and why not stand along the road? That's just obviously huge safety risk, not in the neighborhood, but on the highway. And then the customer was, and that's internal in this case. We actually looked internally of yes, can't that be even easier? Can't that be done better? Couldn't that be done more uniformly by means of automatic image recognition? And then we ourselves, if frames or problems come up and we looked at whether that could be solved with automatic image recognition.

Requirements for the project were: the model to be good enough and what is good enough of course? Problem with road defect is large cracks you can recognize. But fraying, that's one of the fun examples then, because those are loose pebbles from the asphalt, is hard to distinguish because it's just harder to see. Don't tie me down on it, because I'm not an inspector. Especially on camera footage. But one road inspector may find it fraying and another inspector may not find it fraying and then image recognition offers a huge addition because it always says the same thing.

Also, of course, an inspector can miss things. Yes, it doesn't always work out, you can't always see everything and an image recognition model does. Well, then basically in its looked at is of yes, how good should it be and actually that should be better than the inspectors. Or about the same.

It was difficult because of those different looks from inspectors to then determine what a fray is, for example. That was dealt with by having as many inspectors as possible, as many people as possible, and then outlining an average.

About the different functions of the model. The current model makes sure, that pipelines are automatic and somewhat clumsy retrieving images from a third party. Running an image recognition model, so looking at what the effects on those images are present. Then also positioning the images and ultimately for the customer. With the right methodologies, that's then based on (?) among other things. So the inspection must meet certain requirements and that on the basis of different methodologies, to convert that as well. And in addition, a customer also wants to know at segment level how good his road is. So ultimately you get as output a map with all the defects, but also the segments with their goods. The municipality can then, for example, immediately see where the really bad segments are and immediately filter them out. In an online environment.

About other preconditions. Some preconditions depend on which customer you're sitting with. One who wants it delivered this way, and the other on those methodologies. They differ. The input images can also be different, so these are also preconditions. They all have to take that into account.

As architecture, we use Tensorflow Mask RCNN models and these models are built from scratch.

The evaluation of models are twofold. In the sense of do you mean the evaluation of model itself or really the results. A validation set of images is used to evaluate this. For example with all types of damages and all types of roads as well.

About the different architectures and cofigurations used for the latest version. We've been around for a while and different models have been trained, with different configurations. We once started with a Faster RNN, jolo, those are all the image recognition model types and we're looking of yes, which one is the best qua and what is the best treshold you should use.

For an estimate on the different options that have been tried. The training set is also extensive of course, so let's say we've already made 100 models. About maybe that's a bit of an over estimate. They ran for about 12 hours and then was evaluated how far they were. We have two supercomputers standing around with two times the Geforce GTX 2080 as the GPU. So in total 4. And for the training, often one is used for the training runs. The energy consumption of the supercomputer or the GPUs is not known.

Correction about the run-time: The run-time is 2,5 day and about the 175000 steps.

How much energy is consumed is unknown, but its probably a lot.

There was no use of cloud services because when we had started it was 2018. Then there was cloud, but it was easier to just do it with these computers than in the cloud and now we do work partly in the cloud now. However, its still in the test phase.

A final notion is that the supercomputers might consume a lot of energy, but they are located in the building of Arcadis and the CO2 emissions of the building are compensated. So, eventually it will not be a lot.

F TRANSCRIPT IN-DEPT INTERVIEW

This appendix contains the transcribed reports of the in-depth interviews that are conducted for the validation. The table below presents an overview of the different interviewees per perspective and their roles in the organization.

Perspective	**Interviewee**	**Role**
Governmental	Frank Hartkamp	Senior Advisor ICT at Netherlands Enterprise Agency (RVO)
	Daniel Mes	Member of Cabinet of Frans Timmermans' Team of the European Commission
Scientific	Emma Strubell	Assistent professor at Carnegie Mellon University at Language Technologies Institute
	Jan van Gemert	Associate professor at Delft University of Technology at Faculty of Electrical Engineering Mathematics and Computer Science
	Silvia Pintea	Assistent professor at Delft University of Technology at Faculty of Electrical Engineering Mathematics and Computer Science
Service provider	Anonymous	Program manager at Service Provider within the Machine Learning Team
	Sander van den Bosch	Manager at Deloitte Netherlands at Technology Strategy & Transformation

The interviews are transcribed individually and in this appendix grouped per perspective. At the individual interviews is noted whether the original interview was conducted in English or Dutch and translated afterwards.

F.1 GOVERNMENTAL INSTITUTIONS

F.1.1 Frank Hartkamp - Netherlands Enterprise Agency (RVO)

The interview was conducted in Dutch and translated to English.

The interview:

When you look at energy consumption, you have to distinguish between implementing the policy and what our own consumption is. Let's leave the latter out of it. At the moment the policy is mainly focused on energy efficiency with a payback period of five years at the level of taking measures. But the impact of how you deal with the use of that service, otherwise that power management, then you're at the level of software. In fact and I have started and for years kept a knowledge network in the air, we call that knowledge center green software.

I know off the top of my head WIRTH's law. Also something about just like Moore's law that that also applied to software itself. That there is a huge growth in the number of lines of code and you name it. That there's a world to be won there and if you look in the energy efficiency area is. Actually the most biggest promise is something you can do there all around green software. Because in the past when computers were very large switch boxes and the capacity of the space was decisive, you had programming languages that took that very much into account. That you gave as few commands as possible to get to a solution, for example in Cobalt. At a certain point, with the amount of hardware, when that no longer gave the restriction, that was completely abandoned. And then you got all kinds of programming languages that play a much higher level and have become less and less efficient, that serve a whole bunch of functions where it is often not necessary at all, but that the developers who work with that software. They are not judged at all. They have no knowledge of it at all. Have no awareness. The whole word energy is not mentioned in the average ICT training. All those sorts of things. The chairman of the knowledge network at the time worked at the software improvement group, which is primarily concerned with the quality of software.

I am now working with parties around Amsterdam economic board to look at what a sustainable ICT infrastructure system, in 2030 and beyond might look like, what actions are needed there, to accelerate that and then one of the insights there is also that the role of software is really very dominant actually in the potential. but very little dominant now in the attention that it gets.

It is still very much in its infancy that these design choices matter. To provide insight into the extent to which it matters and what choices you have to make to steer it. One of Jas Visser's conclusions from his work in Leiden is that if you have a programmer program for an extra month, it may cost 5000 or 10000 euros more, but you recoup those costs by reducing energy costs and management costs. Because the moment you have something done with less code, you also need less hardware, you also need less management around it.

With AI then I sit myself as a civil servant at a consultation of the Directorate of Digital Economy, where the whole coalition artificial intelligence has started, also with the policy field. And then I say: yes, you also have to look at how much energy all these ICTs cost, also earlier with blockchain. The beautiful discovery of the cryptocurrency and you name it. I say guys do you know how much energy goes into that and do you want that. Then they say yes, is not our responsibility. I say well it is. But so with everything around AI and AI coalition and all the millions and billions that are budgeted for that there is very little there about sustainability. It does say that sometimes you can use AI for beautiful sustainable solutions, because then you can calculate it better. But I also read once: Yes, that's very nice, but the moment you use that technology to further explore and extract even more oil reserves, you are also using the technology for unsustainable purposes and then you disqualify yourself by saying that it helps sustainable solutions.

The silly thing is that you don't see the gigantic emissions from models. Because the average researcher or in a company. They're just sitting at their computer. Then a huge flame doesn't suddenly start burning somewhere, because that just goes via the cable to a data center and in a data center you don't see that a server suddenly starts running faster or something. And the only one who sees that is ultimately the energy manager, and if he hasn't measured it properly, then he won't see it either.

There are several anecdotal examples where people were made aware of how much a computer can consume. But it hasn't landed enormously yet, and it's also very difficult to enforce, because you have the recognized measures list that the technology that you have pays for itself within five years and that you have to apply it. That's very difficult to legislate that in terms of software. It's a very tricky one, but just making it transparent is, I think, a first step. If you come up with a great AI application and it turns out to be equal to the energy consumption of a medium-sized village or town, you can find out whether you are doing humanity a huge favor. Ultimately it's about if you find AI or blockchain or whatever application very relevant and important for the future. Then start thinking: How can that application be as sustainable as possible?

As the RVO, we can focus on creating awareness, but in the context of policy implementation it must ultimately become a law or a subsidy scheme. In that sense, other than that knowledge network, nothing further has ever come of it. And we're at the level of those recognized measures and we're not getting any further than thou shalt apply visualization in an environment. And you have to apply power management, but there is still no question of if you write a program, then you should not write more rules than necessary or if you think of a program for something know that it may cost much more energy, and that you better go there yourself by cab to bring the answer.

The whole energy issue is caused by the current scarcity of renewable energy, but the moment the whole country and the sea are built with windmills and there is no scarcity anymore. Yes, then everyone can do their thing and use all the crazy AI they want and it doesn't matter how much energy it takes. But that situation is not here. In fact, we are investing heavily as a government to keep it all a bit affordable and to make that renewable energy. So anything that you don't need extra of that, that would help. And in the growth of ICT you see that there is quite a bit of unbridled growth there and with that unbridled energy use, which in part would just not be needed.

In part, the data centers that lease the spaces to organizations that put a their servers there. And then they have an interest in having as many customers as possible who want that. It's up to the customer to say: I don't need a whole corridor anymore, but next time I'll do it with half a corridor, because I've optimized my software. Basically, that goes against their business model of data centers. Only at the enterprise centers, which only have their own things to run. So the hyper scalers of Microsoft and Google, who only run their own things, do have that interest. The idea is that they would also be receptive to this, in part. But I honestly don't have a good view on that. But with all the novelty of AI and the fact that we can and do pluck data from everywhere. And that they have to come from all sorts of different sides for one job that data has to come from all sides. If you think about that, actually a very weird concept, because everywhere there is the need for those servers to be on and running because that data can be requested there at that time plus all the transport you need for that and the infrastructure. The moment you make it explicit to everyone, that that triggers something. Yes, perhaps at a certain point people will say I'll wait or I'll do it less or differently. But it's mainly about different, maybe not less, but better and not different.

Automating good behavior would help tremendously. So the moment you're working in software and you get a warning of this line of code is redundant or makes it unnecessarily complicated; It can be simpler, would you like to go on for a while? Then it would also influence your choices. So just having a counter of what you do from has so many consecutive commandos as a result.

On average, in ICT, communication is about functionality and whether that is realized and whether it is error-free without interruptions. For the time being everyone has good money for that. It all has to be three times safe and redundant and once it's running well you especially shouldn't reset it, because then it can go wrong. So there are all reasons to do it less green than could be done.

That's very dominant, because in the considerations of cost what you can earn and what you can save or need less, I just mentioned, with AI there are really billions being pumped in. Because it's all important for the knowledge developments in the future, but there's no consideration there of what kilowatt hour price or whatever is. But with all the renewable energy targets in the Netherlands, of 75 percent electric from renewable sources by 2030. The moment this takes off with AI and with ICT in general, we really won't get to that 75 percent in 2030. Because then the energy consumption will simply be a lot higher and then the share of sustainably generated energy will be much lower.

In the past, with consolidation in data centers, companies often said: no, the server must remain on, because it runs a database. But which one they didn't really know, but let's not turn it off, because later someone will complain. That turned out to be an almost unused small database.

With a study of power management, you could see that there was a huge difference in visualization rates between different servers from different organizations. And then I know that KPN was very good about having the server utilization high, a CPU utilization of 60 percent, but other organizations often stayed below one percent CPU utilization or below one percent. So that means that compared to KPN. That then such a party has 60 times too many servers hardware on. So that's yes and in these times of attention to climate and what they all have to do differently, In these times of climate attention it's actually very strange that you come across that. Even stranger is when you say that it has to change, that they then say: well that's not our business, so it doesn't have that much attention. So we're not going to do anything with it. That is in principle forbidden. In the Netherlands you are not allowed to open the tap and let the water flow if there is no purpose behind it.

F.1.2 Daniel Mes - Frans Timmermans' Council

The original interview is conducted in Dutch and translated to English.

The interview:

It is interesting to stay In touch with science, because when you look at AI and hyper computing, it is for Europe very important to stay in the race with China, Asia, and America.

Most things are still in their infancy. People know what we are talking about with artificial intelligence, but in practice it is still in its infancy. That now gives an opportunity in that rollout itself to already think about it. All those solutions, we can't already think about energy consumption. What is important for us is to avoid a situation where we say with the green deal, for example, we're all going to use renewable energy, and you get stuck in a situation where Amazon buys half of the wind capacity for data centers. That's for the data centers that we have now, not even for high performance computing that we also want to have. It's important with that rollout, to look directly at that. The way we look at it is that we are already looking at what we can do with the sector itself in terms of technical solutions. I think the sector itself also sees a bit of this problem. They are also really afraid of ending up in the new dark corner in terms of energy consumption. I notice that in all the conversations I have with the sector. They are also really interested in seeing how we can roll out the system so directly that energy use doesn't deteriorate.

That's why it's interesting to talk to independent scientists to see what's technically possible. What remains then is for Europe for legislation and regulation, because that is not excluded. But because it is still in the rollout phase and it is important for Europe to have these solutions, we choose these two tracks.

What we are of course in the situation that it is not possible like with aviation, like you already have it and you are going to look at cleaning it up. We've moved on with the data centers and expressed the ambition to have them climate neutral by 2030. So there will be a package there eventually as well. But where we're looking now, artificial intelligence and high performance computing is interesting to take those two tracks.

On transparency of energy consumption of services also for smaller companies is indeed also being considered. I'm curious about what you can do with it at the end, once it's transparent. I have to say, we are now looking at transparency mainly from the point of view of what is ultimately needed from the supply side. People can make an informed choice of course, but many of those companies that really choose these solutions. Everyone has an interest in keeping their energy bills low, so everyone has an interest in knowing that this is energy efficient and I also think that society is changing and everyone wants it to be sustainable. You're right that we're looking a bit at the supply side, but we're really looking at it in a very broad sense, because it's both a question of green procurement, so buying the energy, but also thinking about it in advance: When you build a data center, from Europe then, that we also think about. How can I connect it to the heat network so that all excess heat can be used to heat up the neighborhoods next door? That's what we're doing in our energy strategy, and we're really ambitious about that, and in addition to that you also have all the things we're doing with the circular economy. We are also looking at how data centers can contribute to that. We're also looking at how data centers can contribute to this, that you build these things in a good way, but also that 40 percent of the equipment is not thrown away as it currently is. Part of this is also in technological solutions, because you can look at how you can spread the use of data, for example. You can often do bank transfers at night. So you can leave the system alone during the day. These are all of those things that we do in one package with the offering.

Indeed, we don't do much on the demand side and transparency is something we do work on and not only transparency, but also that the apples can be compared to apples. So there are people in Digi-connect with us working with the sector also to look at. How can we arrive at the same measurement method, because every week some tech company comes up to me with a super nice climate-story and I can do very little with it because everyone uses a different methodology to say what they are doing. The only thing that is consistent is: Everyone is buying the wind energy, but then I say that's important, but I also don't want all the wind energy to go to the data centers and digital infrastructure. I also want to know about energy use and then it does become difficult because then people are measuring different things.

I think it's very interesting to also think about how can you already make sure that that market develops in such a way that you can also offer these kinds of solutions to people, so actually when do you train the model? And what is the best to avoid an energy shock later on? It's a little easier with the things we already have, the data centers. We all know what they look like what they do, it also with the cloud services by the way. Also the infrastructure, so there we know what we need to do there. We know what we need to do and want to adapt - that's the eco design. The eco design is much more than the little label that we as consumers know. With that leaf that something is eco. The eco design is really a standard that every manufacturer of equipment must take into account when making equipment-and we're going to extend the eco design this year to all the things that are already in data centers, so all the equipment and all the data infrastructure will be covered by the eco design. But indeed, AI is nothing but a model. We're also moving to a world where everything is connected, so in any case everything is going to be edge computing and a lot of it is going to take place at the edges. And then, of course, it's going to be much harder to sit down and agree on an eco-design centrally, because it's everywhere. It's a bit like your dishwasher, you have the eco mode and then you automatically choose the right one. We have to think about that for a while.

It's interesting and also in our digital strategy, that we can focus on all the things that make the energy efficient choices, is not quite the same as what you're saying, but it comes close. Because the commitment for the existing infrastructure is very clear there, to make that 2030 climate neutral state-of-the-are circular. But the commitment is also there for artificial intelligence and all the high performance computing. To, what we call, Europe wants to be the ecosystem of excellence it and then also the European ecosystem of excellence where it comes to energy efficient technology and this does fit with that. That's how we describe that ambition. I do find it very interesting when you say that we can't push this a bit more concretely.

The bosses are also working on this, so Frans Timmermans from the green deal idea. Energy consumption and emissions are now on par with aviation and you all know how we treat aviation. Those really treat like a black sheep on the climate agenda, so we have to be careful that the technology sector doesn't go the same way. So they need to start working now to prevent them from becoming the energy consumer of the future. Again I do think people are embracing that.

Margrethe Vestager is also really with that agenda. Because she is the digital supreme of Europe, but also really does have a green mindset. If we come up with the right tools, it should be doable, because the tech sector is precisely the sector where there is a lot of creativity to just tackle this in advance and just do it, and the employees in silicon valley and also in the tech sector in Europe who want to have it, these are often the progressive people. Those are often the people who think climate action is very important and they really do demand that of employers. So I think everything is there, but it is indeed a question of: how can we give hands and feet to this in a practical way and in a way that it doesn't sound as if Europe has closed its borders to technology again. I think that dilemma is not a dilemma, that we can indeed come out of it if we indeed think about it concretely now.

We are looking at certifications. How can we make things transparent and then certify them in a certain way, when a certain method is followed, that it is indeed true that it is energy efficient. A bit with that in mind, because ultimately you also have to throw something over the top - that you can compare apples with apples and that it's all right what happens there, without being too heavy handed. This will continue to be an area, were learning by doing is, that's just the way it is and that's not just for energy use, that's also for other big things in society. Also learning by doing in the sense of does a bias come into this or discrimination, does safety come into this, it's learning by doing across the board.

F.2 SCIENTIFIC COMMUNITY

F.2.1 Emma Strubell - Carnegie Mellon University

This interview is conducted in English.

The interview:

Probably, you know, be from, like the the paper I wrote on quantifying the carbon footprint of certain deep learning models and NLP. My background and my actual expertise, like my actual focus is machine learning, like machine learning researcher specifically. I work on machine learning for natural language, and specifically I work on developing more efficient approaches. So which is why it was in my head, the introduction to a lot of my papers was these models use a lot of energy and we care about that. It has high costs of monitoring and for the environment. I was saying this, but there's like no citation, so let me try to quantify this. And it was really hard. It's very hard. Like I would like to do a follow on work, but it's really hard without, obviously, a lot of information at companies and stuff that you don't necessarily have access to. I don't think they even have access to it, they don't necessarily have the tools in place because I talked to people in companies and they said: that would be really hard to compute. Beyond sort of a fairly small scale thing that I did, which is kind of the low hanging fruit. Like easiest possible estimation. I'm an assistant professor at CMU and the Language Technologies Institute. My focus is on developing machine learning algorithms that are as efficient as possible without sacrificing accuracy. So, that's related to caring about the environment.

The lack of information to calculate the energy consumption of machine learning is an important problem. I don't think a lot of people are thinking about it, but people who care about this are thinking about it.

I have not really an idea about the energy consumption of smaller models. It's like a thing that I would love to do, but I haven't had the time and I don't have a student assigned to it. It would be so easy to estimate that if I just sat down and did it. I mean smaller models being basically simpler linear classifier, like SPM or something. Well, so here's the point. There's a number that we don't have. It is to what extent those models are being run and deployment versus the larger neural network models. I do think people have been sort of switching over from simpler linear classifiers to, not necessarily these enormous deep neural networks, but at least smaller networks. Which are still going to be more energy intensive or at least they're going to use more compute, which tends to correspond to requiring more energy.

I don't have the numbers unless I'm trying to think of I could do some back of the envelope estimate. I definitely could. You could maybe estimate, based on the number of parameters in the model. It would vary a lot depending on the actual application, so if you're doing natural language versus computer vision or something. I can't do it off the top of my head, I have to sit down and figure it out. I think I could figure out a ballpark, sort of multiplicative factor. And so there's also this discrepancy between training and inference.

So in our paper we get numbers for these enormous models for training, but we do also have numbers for smaller models, they are all known neural network models. We do have a number for a typical NLP pipeline. So, if you have a company that is running NLP, that's a reasonable number for that. It is still using like a neural network, not just a little classifier. Yeah, I mean, it's hard.

Yeah, there are numbers that people still haven't explored that are sort of easy to get. This is harder estimates. We also don't even have a really clear understanding of the easier estimates, which is really unfortunate.

When thinking about restrictions that make it hard to calculate the energy consumption for the training of deep learning models, lets first start with inference.

Training time is a good place to start, because inference time is impossible to do. It's very hard. But the reason I do care about it is there's some estimates that of AI or machine learning computation in data centers, like 90 percent of it is inference and only 10 percent of it is training. So that's why we could try to better understand the cost of inference that would be important, but it's a lot harder.

So, for training time or for training cost. So depending on at what level you're trying to measure it, it becomes more and more difficult. So, like at the lower level, I think we have the tools. To measure your training. If you're a practitioner or a researcher and you're training a specific model you can measure the energy use of training that model. Typically, I think one challenge is mapping the energy use to, if you want to get a carbon estimate, the actual carbon intensity of the energy that your energy source is using. So, the carbon intensity of your energy source. I think that becomes even more challenging. And there are tools that exist that I haven't used for this, I think it can become more challenging once you're using a cloud provider. So different cloud providers, it seems like they're developing tools so that you can kind of easily estimate this, which is awesome. And I think for Amazon, basically I don't know how much information is actually available. I'm guessing, you know the tool that was developed for if you are using the cloud, you'd say which cloud, which part of the cloud, which part of the world you're using. And it will give you a carbon estimate, based on your use.

One of the things for me is I feel like a lot of the tools are there, but people don't use them. So it's not common to report these numbers in papers. When people do report numbers in papers, they report the wrong numbers. They're not wrong. I mean, I think they're easier to get. So they report floating point operations, flops, and/or they report the training time and the hardware. And so those are both approximations of the actual energy use, but they're not the same as the actual energy use. The actual energy use is tied. It's like a function of the specific algorithm and the hardware that you're running it on. Certain models are not going to efficiently use a GPU. And so, even if the floating point operations are very low, because of a sparse model, that hardware doesn't do those operations efficiently. So you're not actually saving anything even though the flops are lower. So that's why I advocate for reporting actual energy use. And there's also a tool that reports the energy use, a Python library. So that seems useful, but I guess people aren't using it, but also it's pretty new.

One of the things that I'm working on is, within my research communities, trying to develop standards where people should be reporting the stuff. And reviewers will expect that and things like this.

About evaluation metrics to determine when a model is done training and the stopping mechanism, at least in NLP. This might actually be a little bit different than people who are doing stuff in practice, but what we do in research in NLP and I think computer vision, we tend to have a development set. So we measure the accuracy on the development set and then just stop. I think people will just train for a fixed number of steps and then take the model that performed best within that fixed number of steps. So you're typically overshooting. I think you're doing more training, that doesn't need to happen.

But actually, a related issue that you reminded me of also kind of a reporting issue that makes it difficult to estimate the amount of energy that's actually being spent on training is, and this is also something we discuss in our paper. No one reports all the experiments that actually go into this final. So, there is almost no way to estimate that. We had a case study in our paper of what went into that and it was quite substantial compared to if you were to just report the training of this one model. Well actually we had to train many models to get there. And it makes it hard to make decisions if you're not reporting how much of this development was needed. It's like totally unclear based on what we report, how much energy and how much computation is going to have to go into that development because we don't report how sensitive these models are. Are you going to be able to, out of the box, apply it to this new data? Or you're going to have to train a ton of different models to try to find the correct parameters for that model.

You just reminded me of another thing. So going back to the original point about what stage you want to estimate the energy use. So if you want to estimate the energy use, so that's like the lowest level. I have a single model like what does energy use? It immediately becomes much harder if you want to actually try to estimate it on a sector level or something at a higher level. Because we don't know for which models retraining happens.

Facebook has released a couple of papers on what their machine learning workloads look like in their data centers. And it's interesting because they do provide some statistics of which models they retrain and how frequently they retrain them. So, larger models, like the machine translation model is only getting retrained weekly, but then there are models that are getting retrained every hour or every few hours. Most places don't report that. I think it's also not well understood that retraining happens all the time, at a place like Facebook. Like the banks right now, you have all this new data coming in and you want to constantly adjust your model to the new data. So, training is not a static thing, it does happen frequently.

So about the Facebook papers, it's still frustrating because it's very clear that these papers have to go through an internal review where the lawyers don't let them say certain things.

Next, about the model developers that lack a level of knowledge. I would hope that the researchers have more knowledge, but I see the same things happening. I think unfortunately this is the issue with these deep learning models, they're very different from previous models and that makes them way more inefficient to train. Because, we don't really understand very well how they work. The key thing is the amount of computation that actually goes into finding a good model, it's the hyper parameter tuning. And the deep learning models have almost always more hyper parameters than you can possibly even search the space. So even in the research community, there are like machine learning theory people trying to better understand these and trying to develop better ways. Because, if we actually understood how these parameters interacted and we'd be able to say this is what our data looks like, this is how we should set these parameters. We don't really understand how they interact and how they interact with the data. So, that's why it's so inefficient, because we're like let's try every combination of them. So, I think people with more experience, researchers and people in industry, this is a skill you develop of engineering these models and having an intuitions for how to set these things to do less computation. But it's almost black magic or like a dark art.

People often want to use deep learning just so they can say they use it. I think that happens in industry, especially because people who are not as technical want to hear that you're using deep learning. Funders for startups want to hear those keywords. So maybe they're using it but they don't need to use it, I think in research as well.

I think if we had standards for reporting, like the computational requirements of models that we could actually look for a given model at how much accuracy improvement are we actually getting versus a computational requirements. It's a little bit different, but it's actually needed here. Is the two percent accuracy improvement worth 20 times increase in cost. And when we don't report the cost and we only report the accuracy, which is what is standard now, you don't even see that part of the picture. Especially for companies, it costs money to run these models so they don't just have the environmental incentives. There's also the money incentive.

F.2.2 Jan van Gemert & Silvia Pintea - Delft University of Technology

This interview was conducted in English with both interviewees at once.

The interview:

Associate prof van Gemert is more focused on teaching the theory about deep learning and haven't trained a model in long time, which he would like to more if time would allow him.

Assistant Prof Pintea focuses on the computer vision/deep learning and the efficiency of these models. Also a bit at the energy consumption and the efficiency. And I found some very shallow way of computing that from looking at the number of floating operations. And then, you know, at certain GPU support a certain number of floating operations per Watt and then you can compute an approximate estimate of how much it would take to train a certain network, if you know the number of floating operations the network takes. So, that you can actually do it will probably be a rough estimate. But I think that that's as far as my practical advice comes to this.

So, to calculate the energy consumption, one should know this data about the model training and about the hardware.

The first restriction named is if you run a model in parallel, it may have effects on the energy consumption. And the model itself doesn't say anything about how it's implemented. So you can do all the convolutions sequentially on the CPU or the GPU. But this will probably be less efficient, then if you do them all in parallel. which will be also faster accidentally, which is of course the reason why you do it on parallel on the GPU. But that will probably also be more efficient because then there's probably a certain threshold and certain overhead for using the GPU.

And then if you only use 10 percent of the compute, this threshold stays there is fixed for a longer time, whereas you have this initial threshold and then you have 100 percent use. I can imagine that affects efficiency. So you don't know how the implementation will be. So it's not a hardware problem. It is more of a how is it going to be implemented. And that relates to which framework you use if you a Tensorflow, or Pytorch, or Xnet or something. So that you could probably have the same model more efficiently. So then if you just have the the model that doesn't necessarily directly related to the energy. And then in addition, the flops is also an approximation, I would say, because the reason of implementation. Memory is also something that's often forgotten. I don't know if you read papers that they all talk about flops or some measure based on flops, but memory is also often a bottleneck.

What memory you refer to, because a lot of people give a number of parameters, but that is not static? Exactly. Number parameters is only a million or something. Or order of million, but all the feature maps and all the batch sizes. That depends on the training import data. So if you run it on image net, which has two hundred per two hundred, it's a lot more than if you run it on MNIST. So it's twenty by twenty pixels. But, that again depends again on the batch size for example. So does it get hyperparameters. Hyperparameters also determine real efficiency and that's a second thing as implementation. So if you have the model that you also have to choose hyperparameters.

We wrote a small paper about the black magic and deep learning. How hyper parameters effect model accuracy. Experts are better at suiting this hyperparameters than not experts.

It is stressed that the definition of the model developers is different. Van Gemert does not see the model developer as the one who runs the model, or makes it deeper or wider. But, the one who makes architecture. So, on the side of the person who invents the RESNET. So, that is different from the one who deploys the model.

About the complexity of DL that makes it harder to account the energy consumption. It is just multiplication and addition. That's the basic operation, but I think what also explains the popularity of deep learning is that you can do any kind of differential operation. Everything that is differentiable to any kind of function you can put that into your network and then people invent the most crazy stuff that they can put into it as long as it's differentiable. It's very free. So that makes it complex. But the model in itself is not per definition complex. You could take a relatively simple model and then the energy consumption would still be complex. But, that's complex for almost all models. I don't know if the interoperability of the weights really play a role, I would say the interpretability of the training, so the batch size which samples are chosen. The model in parallel, the optimizer does stuff. It seems to be all implementation dependent, always how the framework is implemented.

we have the idea with pen and paper we could write down what happens during training. It's just that we don't know how exactly it's implemented in a framework. So we don't know what additional caching of feature maps and other operations is done in the background. You could write down here's one layer network and these are all the operations that happen and then you're done. With more hidden layers, it is the same thing that happens. So it does not become more complex, but the total sum becomes bigger. The only thing that I can see is that it very much depends on the implementation details and the framework that you're using has its own optimizations in the background that are not clear to the person writing the code. And then you also have GPU optimizations which are done actually on the hardware and then certain operations are done differently on different GPUs, then that can also cause a difference.

The other thing that plays a role that I just thought of is the input data. For example, the size of the images could play a huge role. And the number of parameters stays the same. And I can imagine that even in a practical setting, that you could have corrupted images. And then the model just has to stop. Let's assume it doesn't crash, let's assume it's nicely implemented with a try catch or something, and then it just skips this corrupted data. Data from the case study was also corrupted, but in a different way. The input data during corona was a different trend that might be undesired as input data.

About the different evaluation tools and methods of the models. People already use peta flops, run time, and number of parameters. As I said, this can be used to determine the energy consumption for a architecture, but everything is approximate. But as far as I know, it's the only thing that people report this, because indeed, energy consumption is very much depending on the machine you're running on and the implementation that you're working with. So it's hard to say.

Usually the averages are used to determine the capacity of the hardware over a whole bunch of iterations and then you average the estimates. So you get more or less stable estimates, hopefully. But yeah, it does vary per batch.

This average is calculated from the model, not measured. So we say we have this model, we know that it has to do this many multiplications and additions. We have this input size. We have this a number of weights and then you can just estimate it with pen and paper. So, how many operations would that mean? But as I said, this is a rough estimate because you don't know how many extra operations are done in the background by the framework. And also, you don't if the GPU actually optimized itself because it may actually skip certain operations and then that's also counted. And also, you don't if the GPU actually optimized itself because it may actually skip certain operations and then that's also counted.

About Deep learning being a relatively new technique to businesses. There's all kind of new operations indeed happening, to every paper proposes their own new layer. Now the self-attention layer is getting more attraction. It has very different complexity than CNN's for example.

Another thing that I just thought of is that NVIDIA makes it difficult to extract statistics about the GPU use. So I know that our manager has difficulty getting the information out of it because he says that the API doesn't support it. Their library isn't supported in getting efficiency usage. So how much of the GPU is now being used? I think you can't get it unless it's a runtime, so there is this NVIDIA command that you can run, but then you have to be on the same note that you're running your code and then it shows during runtime. How much of the GPU you're using? Like 90 percent, 10 percent, 20 percent. And of course, there's a lot of variation between a distance when you wait for the next batch to be loaded from the CPU to the GPU. So this is like this big peak and then going down and the peak then going down. So it's fluctuating. It's not constant. So that will also affect the energy, of course, because it's not constantly running on the on the inputs.

About the social awareness about deep learning. I think people are not aware and they're becoming more aware of this because Open AI, for example, is creating these huge models and people are now making some noise about how much it even cost. Tens of thousands dollars and how much Watt so how much energy was used in training this huge model. But that is getting some traction, there was even a paper about this: Green AI.

If we can have maybe a copy of your findings, then maybe we can indeed look at these because it may give us an insight into what should we now optimize for. If you want to do research for efficiency. Yeah. What what's what are the factors that are really important in the real world? That is useful often because we sit in our ivory tower smoking our pipe, thinking a lot. That is useful often because we sit in our ivory tower smoking our pipe, thinking a lot. Yeah, but yeah, it's nice to talk to people because sometimes interesting problems appear that you don't think of staring outside of the ivory tower. You just have to go down in the mud and find it there.

F.3 SERVICE PROVIDERS

F.3.1 Program manager - Service provider A

The original interview was conducted in English.

The interview:

The hype surrounding Deep Learning is an interesting point. It's for many business cases not necessary being the most advanced or having the largest models, I think there's a sweet spot. You can optimize your return on investment for, say, maybe just using a simple statistical model to achieve your goals, rather than training a custom Bird model, text classification.

I am on the Machine Learning Team. So we offer a platform for AI workflows across the entire machine learning lifecycle. So I work on the team that touches on everything from training to testing to deployment to monitor. So I take a very unique view on machine learning, and I'm interested also in performing case studies across a lifecycle so that you have an accounting framework for not only training cost, but potentially even your data and the inference as well. So at *Service provider A* there are a number of broad sustainability initiatives, as you may be aware, and I applied for a grant to build the foundations to measure and mitigate the carbon footprint of AI. so I actually onboarding after this call, onboarding my team who will begin work on that. So we have a number of different things planned and features that are coming down the pipeline.

Most common restrictions found by me (Frank) are the lack of awareness of the client, lack of knowledge of project member not being modelers, modelers only being aware about basic proxies as running time, and modelers lacking knowledge about precise knowledge on when to use what deep learning architecture. He totally agrees with those restrictions. That's a really great point about the different the lack of ability to decide which architecture provides the best solution. "Everything you're saying is exactly what I'm working on." He is developing a methodology for measuring the life cycle, the operational life cycle for machine learning models. So, for training, storage, and inference.

For the training he has not yet a method, but he believes at the end of this grand cycle that will be measured. He has a good understanding of the technical constraints that we're facing and it's going to it's going to take a few months to get to a good state.

It's the end goal for the tool of *Service provider A* to publish the energy consumption. "This feature that we're developing will be useful for machine learning, but we intend to build it so that it can be used across all cloud services." The goal is to provide transparency on the power and carbon costs of cloud services. It's not only to create awareness, but also that behavior change and to save carbon.

Once you measures then you can start to mitigate. By providing recommendations. I've actually got the equivalent of a dissertation on ways that we can do it. We just need to measure it first.

A big part of the business of *Service provider A* is running servers, so that why he thinks that they're in a good position to execute this. For more or less all of the tools we'll have to access some of our internal metrics.

He doesn't see this project conflicting with *Service provider A* or as losing part of the business. "I see it as a way to identify cost savings and opportunities that can then be passed on to customers." I agree there is a conflicting interest with. So some people in the business simply want our customers to run massive training jobs all day, every day. But perhaps a better way to frame it might be why don't we maximize utilization of our servers. Providing the same product with less compute is a win-win for *Service provider A* and the clients.

About the decarbonizing. Reducing the energy is one good step. There are ways to say shift your workloads to different times of day or different locations that are cleaner and potentially explore pricing mechanisms to incentivize sustainable behavior.

About the energy reporting metric. We're still finalizing what we can and can't share of it, but we believe as of right now that it's OK to share power. It's also fairly, somewhat public knowledge of what the regional carbon intensities are.

About the availability of the PUE? It's a good question. I know that we track it. I don't know if we are able to expose, uh. That is maybe a closely guarded secret. The reporting on the energy consumption will be on end-to-end solutions as well as virtual machines that are rented.

The biggest challenge he's facing is. The fact that the data is hidden away in different parts of the organization and no one has brought it together yet. And there are some constraints as well. The data is hidden in a organizational way. It's a vast company with conflicting incentives. Many silos. Its totally a goose chase.

About the link between deep learning and sustainability. Within *Service provider A*, he thinks he is one of the people at *Service provider A* that has done most research into it. To his knowledge they're really early, they're leaders. And none of the other big service providers is working on similar products. Or at least they don't promote it.

He's also looking for projects/case studies that request the energy consumption for underpinning his project within *Service provider A* and justification to his management.

F.3.2 Sander van den Bosch - Deloitte

The original interview was conducted in Dutch and translated into English.

The interview:

My background is: I studied in Enschede, where I studied technical computer science and did a master's in business and IT. Because I found the business side interesting - I started six years ago at Deloitte there also did my graduation research. That was about enterprise architecture and security and how you can actually get those two disciplines to work together more and connect them, and then actually stayed at Deloitte, doing various projects. Most of them in the financial industry, and then I mainly focus on bringing new propositions live, usually at the edge of the organization. So a bank that wants to bring a new proposition to the market. To do that, they set up a subsidiary, for example, that will market that under a new brand, a new product. That's what I focus on. What do you need from that side to realize that, which often involves the following tension: you want to bring something to the market quickly and often on a small scale, but it must be feasible, but on the other hand it must also be attractive enough. And big enough that people say: hey, there's something in it. Cloud is a topic that often comes back, or actually always comes back, because the scaling model of cloud fits that type of proposition very well. And I was one of the initiators of the cloud deep dive, a two-day training course that we organize to bring our colleagues up to speed on what the cloud is, the developments in it and the possibilities. Including a hands-on lab that we give. So I am definitely interested in the subject and the TS&T team is one of the initiators of the cloud topic.

More focused on the question. I do see that parallel between insight into what kind of energy consumption and insight into what costs are involved. And ultimately, with this type of question, you also have to consider what resources you are using and ultimately this leads to an invoice in euros, or an invoice in CO2 emissions. I've never done any in-depth research into it, but I've come across very little from service providers in terms of dashboards and reports about how, as a consumer, you have little insight into what energy consumption underlies this. I don't know if there and that will probably not be a one-to-one translation, but the two do of course increase as you consume more. I don't know if there's simple conversion tool in there. That certainly won't, so far I think, the parallel that goes on there.

Besides aws and azure you are left with google and Alibaba as two other big ones. Where if I look at Alibaba, I don't know if they are very concerned about their carbon footprint. If you look at the other operations that are under that group, but Google is obviously another interesting one could be. The fact that the service providers do not want to say much about their the energy consumption and its calculation indicates at least at what stage they are.

There is one party that we as Deloitte do still have contact with and that is schuberg philis. They do the same thing as aws and Google, but on a much smaller scale have their own data center in Schiphol.

What I could imagine: that such a party is a little more accessible and if you publish results about the AWS energy consumption is dramatic then and then I can imagine that they want to do their own research first and if they like it they publish it very nicely and if they don't like it they don't publish it. Schuberg Philis might be more willing to publish number and details. Ultimately to reduce their own costs, they will also be concerned about energy consumption so there is an in incentive to have that understood. I just don't know if it plays out, but you could look.

Then back to the relationship between you and the service providers. You get an invoice from AWS. Per the type of service you're using, is a completely proprietary pricing model and most work that you turn on a server and you just bill it by the hour. Whether it's been busy or not or used its maximum capacity or not, it's just a cost per hour. That's the most basic model that many people know. And then you have more serverless computing. Then you don't pay per server but per millisecond that a function is running.

And if that function doesn't perform nothing, you don't get charged either. With lambdas from AWS, you pay per millisecond. You write a very small program. You put in that lambda function, it executes it. And if it takes three seconds. So then you get charged for three seconds. The latter might be an even better proxy for energy consumption than a server running, because if it's busy or less busy, it uses more or less energy depending on the load, whereas those microservices there you might get even better insight and be a more accurate proxy.

In addition, on that invoice you will also see what licenses you have used and what you may still have to pay for. But it depends a bit on what type of service you use. The more detail you get in the invoice.

For calculating the energy consumption of services that you purchase. I think that's how I would differentiate it by the type of service that you're purchasing. What the characteristic are of the energy, I would start with a common service and server running of a certain type. Get from that what the average energy cost is on an hourly basis. And then you'll have an insight into that at the time that it's running so longer, and there will probably be running under load at the time that you're training those deep learning models. That's not something where you say just for an hour it does and then it's down for ten hours and you're training it, so you want to get the most out of it. I would make an assumption it's full continuous ballast. So what is the energy consumption of such a server and then you have at least that for the core to deal with. And then you have some additional services that AWS will offer to keep that server running. They have monitoring running, of course, and things like that. You didn't include those then initially. But maybe you need to take a mark-up from that. You can calculate that in detail about saying, I'll take five percent markup.

When calculating for projects about the biggest constraints. The main one: is the lack of that data about your energy, if you don't get it reported by default from the aws, then the question is can I get that. If you were to do that yourself in your own managed environment, it would be much easier. Then you just have to look at the electricity meter then and you'll see that reflected there. If you have your own set up then that's doable. If you buy that from a provider, then the biggest restriction is the access to the data around energy consumption. If you do have access to that data, then you have to look at how much the sun's server is using, but that's what the monitoring does for you. So I think there is data available on what the load is on a CPU. And then I think you should be able to get there. The lack of data seems to me to be a big challenge here.

Is there a demand in the market for the energy consumption? Yes, that is an interesting direction. If you put effort into this, it's an effort you can't put into something else. I think that if you look at the importance of sustainability and energy reporting, you'll see that more and more parties are looking at this. They may be less interested in a small project somewhere that is not part of their core business. But at a certain point you see that they purchase so many services that it becomes a significant part of their IT spend and potentially also of their energy consumption and emissions. I do see that customers would start asking for that more and more. I don't see the demand being current. It's not something I get a lot of questions about today. But the overall reporting also if you look at Deloitte. And how Deloitte is working on that. Making electric driving more attractive or buying off CO2 from air travel that's made? That sort of thing. On the IT side, too, there is of course considerable energy consumption. And in order to do this reporting well, I do think that there will be a greater need for insight. So when I look into the crystal ball, I think that demand is going to increase and that more insight will be expected into that, whether they expect that from Deloitte who reports integrally or whether they expect that from their cloud service providers where they purchase the service, I think it's more the latter that they ask those questions to aws. For example, I'd mention a bank, they send a monthly invoice to the bank saying can you pay me this many euros and I wouldn't be surprised if at some point they said. Can you clarify how much CO2 we emitted before that, or the energy consumption before that? And what are you doing about reducing energy consumption? You can see this particularly in the contracting of third parties. You also see more attention being paid to this. So not today. But in the future, and as Deloitte, we'll also try to be ahead of our clients' questions.

There must be some data somewhere, where is then the question. Where you get insight into the consumption of such a server in aws or Azure. I can hardly imagine, but well, if nobody has ever asked that question, then it may not be there, but there must be something somewhere. That is probably easier to find at a smaller service provider, since they even might publish it in their yearly reports. That you know how much energy there is consumed and how many servers there are in the data center. It is at least possible to calculate an average. That could be an initial step.

It is stressed that service providers have a conflicting interest of on one hand running their servers efficiently, but on the other hand selling as much server-hours as possible.