MASTER THESIS | ORGANIZATIONAL INNOVATION AND ENTREPRENEURSHIP| MAY 2018

THE POTENTIALS OF ARTIFICIAL INTELLIGENCE IN THE EDUCATIONAL SECTOR

SERVICE AUTOMATION OF EXAMINING AT COPENHAGEN BUSINESS SCHOOL

STUDENTS
Candida Gravili (107084) & Amanda Smit (107387)

SUPERVISOR
Jonas Hedman

PAGES | CHARACTERS
100 pages | 205,604 characters including spaces

ABSTRACT

Artificial intelligence is revolutionizing the way in which technology is conceived in society. While previously, its purpose was to simplify rule-based activities, it is nowadays a mean to aid humans in complex and unstructured data intensive decisions and to better understand human preferences and traits. As the business sector is already taking steps towards an intelligent digitalization of their activities and commercial offerings, the educational sector has foreseeable potential. The human examination process of written assignments, for instance, is an activity that can be improved through artificial intelligence as it consists of data intensive decisions, and at the same time, requires the teaching professor to focus on the individuality of the performance of each student.

This research paper uses a case study approach to look at the opportunity of automating the human examination process with an already available artificial intelligence technology called automated essay scoring. As the current literature on automated essay scoring converges on the use of this tool within the context of primary and secondary education, this paper aims at expanding the topic to higher education. The case study of Copenhagen Business School is looked upon by analyzing how much resources are currently being spent on the examination process and how the stakeholders (teaching professors, students and university management) are experiencing it. In light of the human aspect of artificial intelligence, it was considered relevant to adopt a human centered design approach to investigate the case study. Given the technological challenges of developing an efficient solution while embracing the additional features suggested, a practical roadmap is presented at the end of the paper that directs and aids Copenhagen Business School.

The result of this research brings to light several points. From the case study findings, it came out from teaching professors, students and university management that there is a need of finding a new way of handling the examination and feedback provisioning activity. Concerning the already available tools in the market, this research paper concludes with the additional features and capacities that automated essay scoring tools need in order to be implemented at higher education examinations.

INDEX OF FIGURES AND TABLES

FIGURES

TABLES

APPENDICES

TABLE OF CONTENTS

Definition of key terms used in this research paper:

Key Terms	Definition
Algorithm	Algorithm is a term used in programming language to describe a rule or instruction that is used to teach a computer or software to manipulate data in order to independently carry out an operation or solve a problem (Technopedia, 2018c).
Artificial Intelligence	Artificial intelligence is a term that defines a new type of technology that is able to replicate human thinking and processes. Similarly, as humans, artificial intelligence learns from experience by examining large amounts of data and is able to adjust to new circumstances by improving the way in which they carry human-like tasks (Sas.com, 2018a).
Automated Essay Scoring	Automated essay scoring is a technology that is able to automatically assess written exams and provide feedback related to the students' performances assessed. It aims at replicating the human examination process and reducing the time that is required to assess exams (Hubert.ai, 2017).
Examination	Examination is a process that consists of teaching professors examining submitted assignments by students. During the examination process, two activities are carried out. The first activity is the assessment of the exam, which entails checking for wrong and correct answers as well as weaknesses and strengths of the exam being assessed. The second activity is the final provision of a grade that reflects the overall performance of the student (Ramsden, 2003).
K-12 Education	K-12 Education is an educational technology term used mainly in the United States and Canada. It refers to all of the educational levels below higher education: from 1st to 12th grade (also known as primary and secondary education) (Rouse, 2005).

Machine Learning	Machine learning is a computer science term used within the field of artificial intelligence. Through the use of a specific algorithm, it applies data analysis techniques to scan for patterns of data that can be then exploited as a basis and model for automated decision-making processes (Sas.com, 2018b).
Natural Language Processing	Natural language processing is a field of computer science that uses linguistic principles to allow a computer to understand and analyze human language. Through natural language processing, a computer can perform automatic summarization, translation, relationship extraction, sentiment analysis, speech recognition and topic segmentation. With respect to the act of interpreting numerical data, human language understanding is a harder activity to perform for a computer, as it uses different and similar words that are linked to create a deeper meaning than the one that is behind each single word (Kiser, 2016).
Servitization	Servitization is a word that has been coined exclusively to indicate the purpose of this thesis- that of providing a service to professors and students in order to help them in the examination process by implementing automated essay scoring. The concept of providing a solution by the means of a service comes from the theory of service dominant logic.

See Appendix 1 for glossary and acronym directory.

1. INTRODUCTION

This chapter aims to present the context of the research paper by providing an overview of the technological landscape emerging in today's world. Specifically, it looks beyond adopting new technologies in the corporate context, into the public sector within higher education. This provides the scope of the research field, emphasizing on the knowledge gap and introducing the research question.

Technology is transforming the world, while human capabilities are along the ride. Currently Artificial Intelligence (AI) is decreasing bureaucracy and increasing automation, resulting in the simplification of time-consuming activities to make business processes more fluid and adaptive. This type of technology is crucial at the moment, as real-time data provided by the Internet of Things (IoT) is making companies' products and service offerings more centered on the individuality of the single customer (Daugherty and Wilson, 2018). This new era reflects the innovation of major industries according to the digital transformation that is occurring, a mean for people and machines to collaborate. This collaboration and shift towards a personalized world is, and will continually be, heavily influenced by AI. According to PwC (2018), 45% of economic gains will occur by 2030 reflecting enhancements to products and services, with the adoption of AI.

Besides the corporate context, the public sector has foreseeable potential of enhancing the quality of services delivered to citizens by the use of AI. This includes examples such as detecting fraud, planning new infrastructure, making welfare payments and immigration decisions, as well as answering citizen queries (Martinho-Truswell, 2018). Moreover, AI is continually increasing in the educational sector, to enhance the value in classrooms. The learning and teaching ecosystems are continually changing, whereby a "one size fits all" model is no longer compatible within the same environment and content. As students and professors learn and teach differently, the focus of individualized and personalized learning and teaching is moving beyond the classroom. This shift blends technology and physical lectures to implement engaging activities with students, as professors have the power to enhance the learning experience (Ramsden, 2003).

As AI is exploited whenever there is an opportunity of cutting out a tedious, time consuming, decision making activity and increasing the value of another aspect, a use case application would be the examination system in higher education. This is because, on the one hand, examining takes a large portion of professors' working hours and, on the other, it cuts out other valuable activities. An example of these valuable activities is feedback provisioning after students have submitted their assessments, which requires professors to focus on the individual students' performances. Given the high number of students attending a course at higher education, technology has indeed already found a way to overcome this matter. Thus far, multiple choice and fill in the blank responses are possible to correct by a scanner machine, while Automated Essay Scoring (AES) is in the works and increasingly becoming known within the industry (Shermis et al, 2016). Until now, on the market AES tools have proven an accuracy rate with human graders of approximately 80%. This backs up the belief that these tools, after implementing additional improvements and as AI rises to achieve a gold standard, are fast approaching the replication of a human examiner (Mark Shermis 2018, personal communication, 6 March).

That being said, this field is crucial to analyze by looking at the blue-ocean opportunity of transforming the examination system into a service. A case study is used in this research paper to observe the phenomena from an inside perspective at Copenhagen Business School (CBS) in Copenhagen, Denmark. More specifically, an analytical approach is taken to understand the technological implications of AES within higher education, leading to the research question of:

What are the potentials of AI in examinations at CBS?

To this end, we, as inside researchers, investigate the examination system of CBS by looking at the amount of resources spent on this activity and by studying how different stakeholders, more specifically, professors, students and university management are experiencing it. Besides the case study conducted, an analysis of the available AES solutions in the market will be presented to evaluate the innovative opportunities and the processes involved in adapting the servitization of examination process.

1.1 STRUCTURE OF THE THESIS

This thesis will synthesize the analysis conducted on the examination landscape at CBS with an analytical approach. An overview of the papers structure will be provided to guide the reader systematically throughout this research study.

First, an introduction of the thesis has described the topic of the technological landscape that is continually emerging in the world today. It emphasized the importance of how adapting to this new era is crucial to stay innovative, specifically within the educational sector. As AI is being adopted within the sector, the research gap of the servitization of the examination system at CBS is introduced. After identifying the gap within the subject field and stating the purpose for the research, an overview of the existing theoretical literature is provided in Chapter 2 in accordance to the correlation between the topics of AI technology and higher education.

Going further, Chapter 3 introduces an initial theoretical framework to illustrate how the data can further be analyzed to answer the research question. Given the structure for the data analysis, an analytical approach is designed in Chapter 4 to provide the reader with a descriptive overview of this research methodology. Thereby describing the methods of the data collection activity, including the strategy and quality of the research. Additionally, a description and benefits of conducting a case study is defined.

The findings are presented in Chapter 5 in several different forms, providing the analysis of the research, with a reflection upon the theoretical foundation previously described within the literature review and theoretical framework. Based on the data collected, Chapter 6 presents a proposed roadmap for CBS that of which emerged from the research, including a detailed description of the proposition.

Continuing, Chapter 7 concludes the research findings, reiterates the recommendations for CBS, reflects upon the limitations of the research study, and highlights the academic contribution of this research study to expand the knowledge within the literature. Lastly, recommendations for future work is outlined for further research to be conducted and developed within the field.

2. LITERATURE REVIEW

This chapter reviews the existing literature on AI technology and higher education to analyze the past and future opportunities within both contexts. This reflects the evolution of the automation of work processes and seeks to identify ways in which this technology can be utilized within the higher education setting- introducing automated essay scoring tools.

2.1. HIGHER EDUCATION TEACHING AND LEARNING

In the educational sector today, specifically higher education in public institutions, class enrollments are on the rise. As a consequence, occupational satisfaction rates of professors are decreasing (Ramsden, 1996) as they have to spend more time preparing for class and examining assignments. Governments are cutting budgets on public expenses, including the educational sector. With these dynamics in place, the impacts on the quality of higher education is negatively affected. Students have become pre-occupied with advancing their grades rather than mastering the content of the subject matter (Ramsden, 2003). This leads to duality of what learning is- one for professors and one for students. Learning is a conversation, based on a specific subject, between different experience level learners, whereby knowledge sharing is essential (Ramsden, 2003). Within this conversation, an exchange of knowledge, beliefs, behaviors or attitudes are relayed, known as a process (Ambrose et al, 2010).

The concept of learning, from the students' perspective, is not only what is being taught in the classroom by the professor, but it is also strongly influenced by the classroom environment itself and the educational context. Swedish psychologist, Roger Säljö (1979), conducted a study, including 90 participants, to establish the understanding of learning for adult students. Five key conclusions were drawn, as follows:

1) "Learning as a quantitative increase in knowledge. Learning is acquiring information or 'knowing a lot'.
2) Learning as memorizing. Learning is storing information that can be reproduced.
3) Learning as acquiring facts, skills, and methods that can be retained and used as necessary.

4) Learning as making sense of abstracting meaning. Learning involves relating parts of the subject matter to each other and to the real world.
5) Learning as interpreting and understanding reality in a different way. Learning involves comprehending the world by reinterpreting knowledge" (Ramsden, 2003, p.27).

It can be seen that there are highly different approaches to students reflecting their own learning. Interestingly, points one, two and three reflect the external environment, meaning that students gain this knowledge just by being present. Whereas points four and five suggest the internal environment, where learning is something a student gains from understanding the real world (Ramsden, 2003). This internal environment reflects more those students studying at a higher education level, where emphasis is put more on critical thinking.

The distinction between how students learn and teaching from the professor's perspective is critical. Students perceptions of the educational system are based on three main criteria: the curricula, teaching methods, and assessment procedures. Rowntree (1977, p.1) states, *"if we wish to discover the truth about an educational system, we must look into its assessment procedures".* To assess this critical point and for the sake of this thesis, the assessment procedures will be explored further in depth.

According to Ramsden (1996), assessments are: a method to help students learn, a method to analyze students' progress, and a method for professors to alter ways of teaching to better effect students. In order to help students to learn and evaluate their progress, there is an inevitable interlink between the two forms of assessments: *formative* and *summative,* explained in the following section. As Ramsden (1996) argues, these assessment "manuals" are not separated in reality- there is only one world out there that ultimately evaluates student achievement and the relevant measures to adopt changes to learning.

2.1.1. Assessments

Furthermore, assessments are defined as the activity of collecting information on the knowledge depth of a student that has attended an educative and formative course. This process is carried out by examiners or automated software solutions, later discussed, and entails the evaluation of the learner's performance and instructional outcomes. The evaluation is communicated through the provision of a score, or a simple pass or fail statement that reflects the student's knowledge depth (Ramsden, 1996).

The assessment process includes several activities that professors and examiners have to carry out before and after the actual examination event. The first activity is a priori selection of the examination criteria done by, defining and specifying the technical requirements and learning goals on which students will be graded. To test these technical requirements, the professor selects, or creates an exam format where the students will have to apply theories that have been discussed throughout the course (Ramsden, 1996).

After the assessment, the professor and/or examiner will evaluate each assignment reflecting upon the nature of the measures: specific criteria, rubrics and learning objectives. Depending on the scoring rubric, and governed by the laws of each institutions and countries, the examiner may also compare student's submissions with other students' performances, and eventually determine a grade. The provision of feedback, either before and/or after the assessment, depends on the type of the assessment, the possibility of retaking the exam, the willingness of the professor (given their time constraints), and the rules and common procedures of each educational institution (Biggs, 2003).

Concerning the assessment, as known within this paper as an examination, activity itself, there are two general categories of assessment activities: *formative assessments* and *summative assessments*. *Formative assessments*, also called classroom assessments, are examined assignments that students have to submit throughout the entire period of a course and are used to communicate the learning progress, as well as the achievement of curricular goals of each individual student. This type of assessment benefits both the student and the professor. On the one hand, the student has the opportunity to exploit what they are learning and test whether they have any knowledge gaps or

doubts regarding specific topics or areas. Inherent in this form of assessment is the opportunity to receive feedback. Moreover, the professor has the possibility to verify their teaching effectiveness and identify whether there is any method or activity that needs to be changed or customized according to students' interest and learning progress (Shermis and Di Vesta, 2011). *Summative assessments,* also called *after learning assessments* (ibid.) are performed at the end of an entire course block and have the purpose to test the knowledge or performance of students regarding material that have been taught throughout the course. With *summative assessments,* students receive a final grade that is based on an achievement standard. The final grade can be expressed in two ways: *norm-referenced* and *criterion-referenced*. A *norm referenced* grade tells the student what their position is with respect to the rest of the students in the class. As Biggs (2003) states, *norm-referenced* grade is also called "grade on the curve" where, for example, the top 10 percent students are graded with high distinction, the following 15 percent with distinction, the next 25 percent credit and 45 percent pass. *Criterion referenced* grades express what students have learned in reference to a set of behavior objective or standards. With this type of assessment, examiners evaluate students' performances based on a set of learning objectives regardless of how other students have performed. With respect to *norm-referenced assessments,* the quality requirements of the assessment are defined at the end of the assessment activity and depend directly on the performance of all the students, while *criterion referenced assessment* standards are always formulated and defined before the actual examination process (Biggs, 2003).

2.1.2. Learning Objectives

Assessments are directly connected to the learning objectives that professors formulate to construct the entire teaching course. Learning objectives clearly state what students are expected to learn and achieve by the end of the course period. As Arreola (1998) states, learning objectives consists of three components: a definition of the skills/behavior that a student will be able to show at the end of the course, a definition of the conditions under which the student will be able to apply these new skills and how these skills will be assessed.

Learning objectives provide professors with an overview of what content and instructional material to select and how to organize the teaching course, reflecting upon the basis on which the assessment

will be developed. These objectives also aid students in the learning process, as they can be utilized as a guideline for class participation and preparation for the final exam.

In order to construct valuable learning objectives, a widely used framework named after Benjamin Bloom, known as the Bloom's Taxonomy of Educational Objectives for knowledge-based goals (UNA Charlotte, 2018) can be employed. According to Shermis and Di Vesta (2011) the different levels in the Taxonomy of Educational Objectives for knowledge-based goals from lowest to highest are: *knowledge*, *comprehension/understanding*, *application*, *analysis, synthesis*, and *evaluation*. In order to assess *knowledge*, students are asked to define and describe specific terms, ideas and theories. At the level of *comprehension/understanding*, students are expected to be able to connect different pieces of information and definitions with each other and be able to classify concepts. At the *analysis level*, students are asked to describe the structure of different ideas and be able to compare them describing the differences. Moving to the *synthesis* ability, students should show that they can put together different ideas and come up with their own theories. The highest and last level of learning, *evaluation,* is achieved when the student is able to apply specific concepts and theories to concrete cases and are evaluated based upon these.

2.1.3. Bologna Accord

Standardization and harmonization in the quality of higher-education within the European Union (EU), known as the Bologna Accord, can be viewed as general learning outcomes. This educational reform was approved in 1999 and moved into full implementation in 2010 as a way to organize student's movement between different countries within the EU. The accord simplifies the higher education system within EU by establishing clear distinctions between bachelor and master studies. Prior to this agreement, higher education in the EU was complicated because of the unique matrix of individual agreements among universities, concerning admissions offices across the union, as well as looking for jobs where employers had no standardization to base their applicants on (Gmac.com, 2005).

Given the declaration, not only did this affect the EU, but also education overseas relating to:

- "New degree requirements and transcripts- within the educational institution applicant pool
- More bachelor's graduates, and consequently, more potential master's students

· More willingness to study abroad
· More competition for students" (Gmac.com, 2005)

With these clearer distinctions in a pedagogical view and with the learning objectives previously discussed, the understanding of these different levels of knowledge makes it easier for professors to come up with clear guidelines for learning objectives and, at the same time, be used to design evaluation criteria for assessments, known as rubrics.

2.1.4. Rubrics

Rubrics consist of "quantifiable declarations of what human raters are instructed to score as being important, typically on a scale ranging from 'poor' to 'excellent' ", in order to define writing standards of the submission (Shermis and Burstein, 2013, p.222). The criteria within rubrics can be determined either by evaluators creating a customized rubric, or by using a standardized model, which can aid evaluators to grade more objectively. Relating to the previously described two approaches of criterion referenced and norm-referenced assessments, there are three classifications of rubrics: *holistic*, *analytic* and *trait*.

The *holistic* scoring approach reflects an overall assessment of the writer's performance on the entire essay. As the criteria is grouped as a whole, the disadvantage lies when the 'end-user' receives the score with little to no details on the strengths and weaknesses of the submission. In turn, this reflects the lack of specific feedback, resulting in not understanding the grade given for the submission. In contrast, the *analytic* scoring approach reflects the different components of the essay submitted, such as the ideas and wording. Additionally, the *trait* scoring approach entails greater feedback relating to the writer's ability, outlining strengths and weaknesses.

Interestingly, when put into the context of AES, researchers Page, Poggio, Keith (1997) and Smith (1993) found that a holistic approach is rather more reliable and efficient than the trait approach. A study conducted by Page et al. (1997) that included 500 essays for National Assessment of Educational Progress concluded that when human raters are using a holistic approach, there is greater consistency than that of a trait approach.

Learning objectives are another aspect that evaluators can utilize where statements are produced that define the course objectives, outcomes and goals. Rubrics can therefore be used in order to assess the learning objectives of a particular course or program.

2.1.5. Essay Exams and Prompts

In order to introduce all the concepts behind the topic of AES, it is important to define what an essay prompt is. Within high-level education (university), written exams take different forms according to the specification of the topic(s) that the student has to be examined on. These include problem or case-based exams that can also require a numerical explanation, list of questions with short answers, questionnaire or *essay exams*.

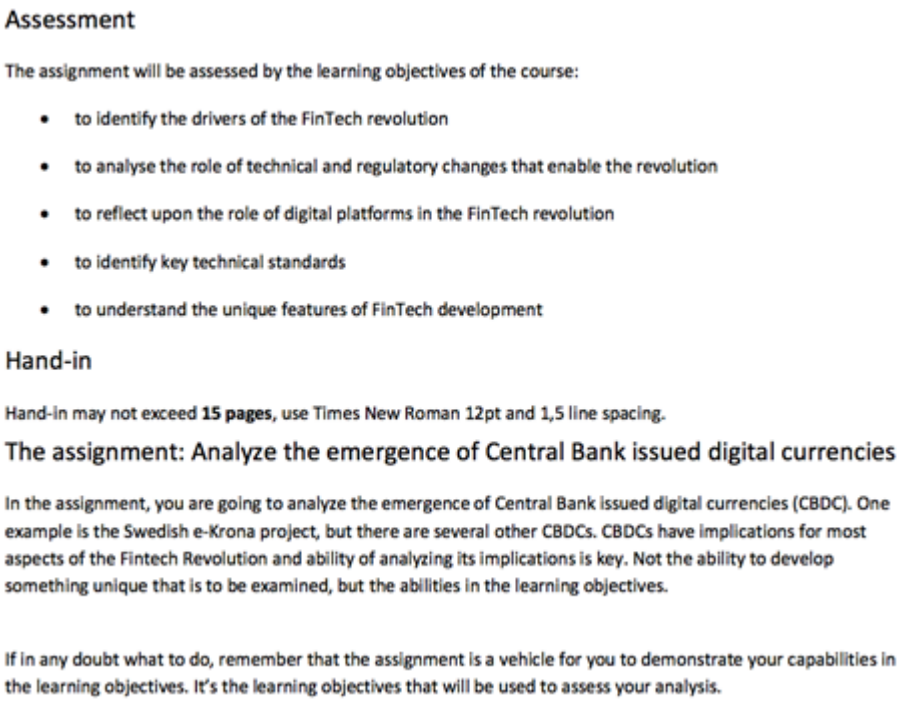
Essay exams are used to assess the ability of the student to summarize information and produce their own theory and argumentation about a specific topic. Through an *essay exam*, professors want students to show that: they understood the concepts that were explained during a course, they can apply those concepts to interpret other topics or events; they can connect different ideas and compare them to each other, they can use the information they learned and justify their argument against a specific topic and that they can criticize and analyze different ideas and theories (The Writing Center, 2018).

An *essay exam* consists of a prompt that is given to the student as a stimulus to start writing the essay. Prompts are sentences that refer to a topic or an issue that can also contain open ended or more specific questions. Besides university course exams, essay prompts are also used in English compositions and literature classes, as well as at entry exams for college. Typically, a prompt opens up a discussion on a determined topic or issue and asks students to communicate their point of view. Within English composition *essay exams*, essay prompts can also push the student to write a persuasive composition on a debated issue (Study.com, 2018).

In order to give an idea of what an essay-prompt based exam is, and to show how learning objectives and assessments are connected, the below figure will provide an example of a summative essay-based written exam.

2.1.6. Example of a Master's Degree Course Final Assessment[1]

Figure 1: CBS exam prompt example



Assessment

The assignment will be assessed by the learning objectives of the course:

- to identify the drivers of the FinTech revolution
- to analyse the role of technical and regulatory changes that enable the revolution
- to reflect upon the role of digital platforms in the FinTech revolution
- to identify key technical standards
- to understand the unique features of FinTech development

Hand-in

Hand-in may not exceed **15 pages**, use Times New Roman 12pt and 1,5 line spacing.

The assignment: Analyze the emergence of Central Bank issued digital currencies

In the assignment, you are going to analyze the emergence of Central Bank issued digital currencies (CBDC). One example is the Swedish e-Krona project, but there are several other CBDCs. CBDCs have implications for most aspects of the Fintech Revolution and ability of analyzing its implications is key. Not the ability to develop something unique that is to be examined, but the abilities in the learning objectives.

If in any doubt what to do, remember that the assignment is a vehicle for you to demonstrate your capabilities in the learning objectives. It's the learning objectives that will be used to assess your analysis.

This example shows the final exam of the Master's Degree elective course on the "Fintech Revolution" that was given to students at CBS in 2018. It is a take-home written assignment where students were given 72 hours to write a maximum of 15 pages in response to the provided prompt.
As shown above, the professor has listed the learning objectives of the course stating that these five points will be the basis of what the students will be graded on. Each learning objective sentence starts with an action verb that represents the type of behavior that students are expected to show in the assessment: to identify, to analyze, to reflect and to understand. Each skill has to then be applied on a specific feature of the general topic of the "FinTech Revolution": the drivers, the role of the new tools, key technical standards and the unique features of the Fintech phenomenon.

With respect to the prompt type, it does not include a question, but it clearly describes what the content is regarding the topic that students are expected to write about: the emergence of Central Bank digital currency which is a "sub-phenomenon" of the FinTech Revolution. A hint is also given

[1] From the course Fintech Revolution taught by Jonas Hedman, February 2018

to spur students on discussing the consequences of it: “CBDCs have implications for most aspects of the Fintech Revolution and ability of analyzing its implication is key”. This type of prompt gives the students the opportunity to apply all the skills described in the learning objectives by using them also as a guideline on how to structure the paper and what to discuss within the paper (drivers, technical standards etc.).

2.1.7. Feedback

At the end of each formative and summative assessment, students *should* receive feedback in order to improve their performances and to better appreciate why they received a specific score or evaluation. Feedback has a twofold purpose: it is the consequence of performance and, at the same time, it is an integral part of learning. As Knight (1995, p.158) defines, feedback is the “verbal and nonverbal responses from others to a unit of behavior provided as close in time to the behavior as possible, and capable of being perceived and utilized by the individual initiating the behavior”. From this definition, it can be interpreted that feedback is hence a constructive and valuable comment that, if provided with responsiveness, has the capability of helping an individual to correct their mistakes or increase their skills in performing a specific action.

Evans (2013) underlines two different views of feedback: the *cognitive* view and the *socio-constructivist* view. The *cognitive* view interprets feedback as a corrective approach that is provided by an expert to a passive recipient. The *socio-constructivist* view sees feedback as a stimulus for the student to gain the information and knowledge independently that is needed to correct the mistakes and improve skills. In relation, feedback in a higher education setting is seen as a way to help the student become independent by monitoring, evaluating and regulating their own methods of learning.

There are different types of feedback and a distinction can be made based on the individual that is providing it. First, there is *expert feedback* that, as explained previously, entails a person at a higher working level or with a higher level of knowledge or expertise than the feedback recipient, to give technical insights on complex topics that require a solid knowledge base. Feedback can also come from an individual that is at the same level of the feedback recipient and who is performing the same activity, known as *peer feedback*. It involves peers rating each other against a performance

rubric and provide constructive feedback to each other. The purpose of *peer feedback* is to teach students or employees to constructively criticize the work of others to help them improve. By evaluating others on specific rubrics and guidelines, simultaneously, they learn to understand what quality standards they have to follow themselves are in order to achieve excellence (Evans, 2013). Similarly, *self-assessment* helps students be critical of their own work and to individually learn how to correct themselves. Lastly, there is *e-assessment feedback* which is delivered through information communication technology, either from a web-based platform or software application. E-assessment feedback is the fastest solution among all the others and it has the additional advantage of being able to reach a large number of recipients and, hence, provide personal feedback to each one of them- no matter the size of the individuals being assessed. Electronic feedback systems are, however, developed by experts who will predict the different types of behaviors of a student in an assessment and will produce in advance feedback on each different type of behavior.

It is important to stress the fact that feedback has to be an integral part of the learning process as it is based on learning objectives and has the purpose of communicating them to students. The interest in receiving feedback makes the student engage and participate in formative assessments and exercises, as feedback cannot be received if the student does not complete the work and exercise in advance. Moreover, feedback is said to support learning, instead of merely giving a final score on the performance and indicating what is right and wrong, by focusing on explaining to the student the what, the how and the why of their mistakes and poor performances (Evans, 2013).

2.2. AUTOMATION

2.2.1. The History of Automation

Gottfried Wilhelm Leibniz, creator of the first calculating machine in 1670s, described its value to astronomers by stating that: "*It is unworthy of excellent men to lose hours like slaves in the labor of calculation which could safely be relegated to anyone else if machines were used*" (Willcocks & Lacity, 2016, p: 35). This quote was regarded as the starting point of a new type of thinking that supports the idea that human time-consuming and repetitive activities have to be replaced and eased by automation and machines.

Highly competitive industries like telecommunications, utilities, financial services and healthcare have always been characterized as having to carry back office and bookkeeping activities that represent a consistent portion of the company's costs. On top of that, these industries incur additional costs related to other types of activities like business management, security and compliance, marketing, innovation and service excellence- all requiring additional human skills. It is here that the topic of work automation comes to hand by suggesting the use of technology to replace back office operations to let employees focus on more critical and strategic thinking activities. Over the last century, the way of handling business has shifted from an era where automation was replacing computational activities to today where machines are able to understand the individuality of customers and communicate in a sensitive way with them (Yonck, 2017).

To better analyze the impact that the rise of computers and work-related technology have had on the work of humans, Willcocks and Lacity (2016) defined four eras of computing and IT investment cycles on automation. The first one is called "*The System-Centric Era*" that arose between the years 1964 and 1981. This era was characterized by companies investing in large computing systems as influenced by the Grosh Law which states that the power of computers increases as the square of the cost. Previously, automation was only employed centrally for operations such as finance. During this era, mainframe computers were beginning to be used for other non-finance functions such as engineering, production departments and even to service suppliers. This was made possible by *time-sharing* computing capabilities that allowed different employees to use the same computing system at the same time (Dictionary of Information Technology, 2002). Next came the "*PC-Centric Era*", between 1981 and 1994, during which personal computers were first introduced. The advent of the PC and its availability that was extended to private individuals implied a shift from the use of computers as a corporate tool, to the use of computers also as a commodity product. The period between 1994 and 2005, the "*Network-Centric Era*", was designated by the birth of the World Wide Web (WWW) that started to be regarded not only as a computational tool, but also as a communication tool that made it possible for anyone to get in contact with anyone all over the world. Virtual communities and a new type of economy based on "the Network" cropped up out of this new form of communication. As explained by Metcalfe's Law, the network economy introduced a new way of value creation that was dependent on the number of individuals that joined the network, where the cost of the network increases linearly as new nodes (and hence new

individuals) are added but its value increases exponentially. The last and still present era (2005-2025) is the "*Content-Centric Era*" where the technologies employed by companies are all geared towards the WWW, which completely revolutionized the way people used computers. Computers provide the individual customer to offer a customized and valuable product and/or service. Thus far, this period has seen the introduction of six technological developments: mobile internet access, the automation of knowledge work, big data, the IoT, robotics and digital fabrication (Willcoks and Lacity, 2016).

These technologies are designed for improving the quality of production (digital fabrication), understanding and connecting with customers (mobile internet access, big data, the IoT) and augmenting humans working capabilities (the automation of knowledge work, robotics). As the focus of this thesis is on the automation of human capabilities, the following section will dig deeper on the topic of work automation.

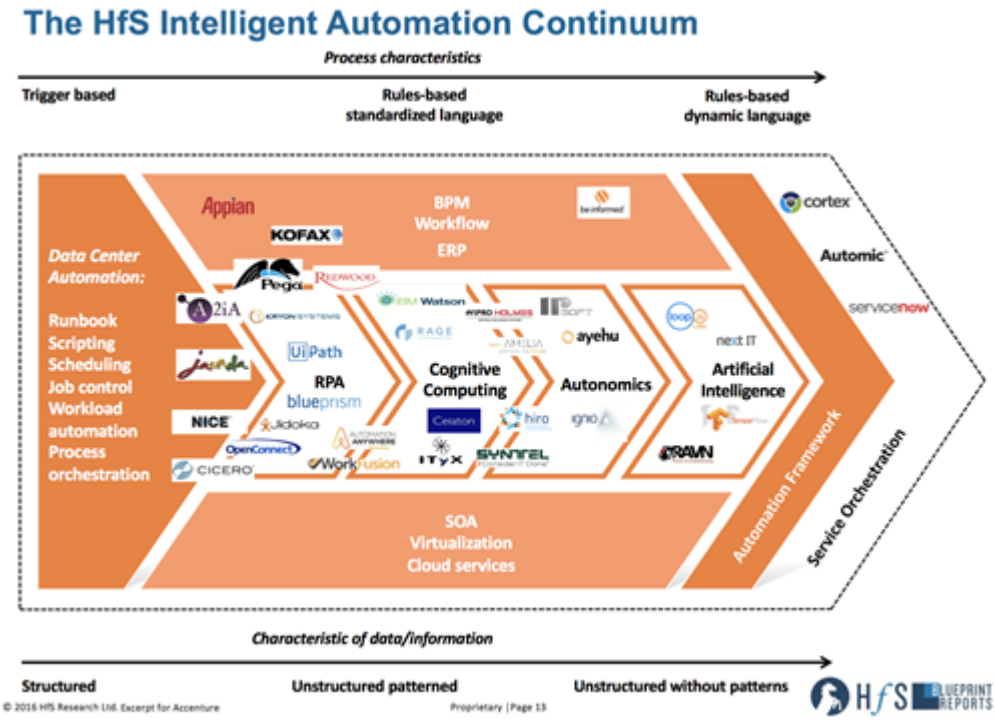
2.2.2. Automation of Work and Business Processes

Work automation is the process of creating cost efficiency by using machines to perform processes, tasks and business functions within a company or organization. Besides cost reduction, firms use automation to achieve service excellence, business enablement, scalability, flexibility, security and compliance. Another word for work automation is *intelligent automation*, which is an umbrella term that includes all the different types of machines and technologies that have the power of scaling, increasing the speed and decreasing the complexity of specific activities by acting as a complement to human skills (Accenture, 2016).

The Intelligent Automation Continuum (IAC), developed by HfS Research (shown below), maps out all of these different technologies on a spectrum that goes from the easiest to the most complex to implement: *Robotic Process Automation (RPA*), *Cognitive Computing (CC)*, *autonomics* and *AI.* Following the framework, a firm that wants to automate a specific process has to firstly analyze the characteristic of the data behind the process and the characteristics of the process itself in order to select the right technology. As shown in the picture below, on the lower and top arrows, the data behind a specific process can be structured, unstructured patterned and unstructured without patterns and the process can be trigger based, show rules-based standardized language and rules-

based dynamic language (Reuner, 2016). The complexity of an activity to automate increases as the activity lies more and more to the right side of the IAC (Willcoks and Lacity, 2016).

Figure 2: The HfS Intelligent Automation Continuum



Source: Reuner, 2016

The technology that is positioned to the extreme left of the spectrum is *RPA*. *RPA* is a software-based solution that is used to automate operational procedures that are backed by structured data and where employees take data from one set of systems, apply rules to them and then add the results into a record. This type of activity is defined as having a "drag-and-drop" modus that captures, schedules and follows process steps. The other three solutions in the spectrum are all connected to each other to some extent. The reason for this is that they all represent a new way of computing that has the purpose of reproducing human thinking and analytical skills, relying on unstructured data. Furthermore, in addition to processing numbers and values, they are also able to understand human language. The first technology within this group is *CC, w*hich lies one-step further to the right of *RPA*. *CC* is a technology that uses self-learning, Natural Language Processing (NLP), data mining and human computer interaction to solve uncertain and ambiguous problems. It is able to deal with

dynamically shifting situations by adapting to new information (without continuous manual intervention) and weight conflicting information to suggest a solution to the problem (Technopedia, 2018a). With respect to *AI*, which is able to autonomously provide the solution to a problem, *CC* is a tool helping the decision maker to understand which solution has the highest chance of success (Evans, 2017). Hence, the human will still be the last one to make a decision. *Autonomics*, as the name suggests, is a type of computing that is able to self-configure, heal and optimize. This type of solution can be implemented in a company when there is a lack of qualified IT professionals. Autonomics technologies require human intervention only at the configuration of their systems (Technopedia, 2018b). The last and most complex technology in the IAC is AI, which will be briefly introduced here and explained more in-depth in the following section. AI technologies are used to automate decision activities that do not involve routine but that, rather, manage other processes like moving self-driving cars (Reuner, 2016). In a self-driving car, AI does two things at the same time: it analyzes the condition of the street and makes decisions on where and how to move by triggering the final movement of the car.

The IAC framework has the goal of providing an overview on how problem-solving activities are beginning to be tackled differently by these new types of technologies. In comparing the two extremes, *RPA* and *AI*, their difference lies within the way they address limitations. As *RPA* is rule-based and relies on structured or semi-structured data, it has the advantage of being highly deterministic and helps overcome existing limitations within the activity that was previously carried out entirely by humans. Conversely, *AI* has the additional attribute of being able to work with any type of data: structured, semi-structured and unstructured. The only drawback, which can be solved by developing a good prediction model, is that it is more probabilistic than deterministic compared to *RPA*. Nevertheless, its ability to learn from the data, change behavior and mimic human decision-making makes it able to work with the limitations of an activity and convert them into relevant output (Everest Group, 2018).

2.3. A NEW ERA: TECHNOLOGY DISRUPTOR

2.3.1. Artificial Intelligence and Machine Learning

As previously touched upon, new technologies in everyday human life are employing intelligence and automation that is transforming the world; a new era. Buzzwords of AI and Machine Learning (ML) are accelerating within all industries and are resulting in unprecedented reach, power, and influence. As sectors are advancing in technologies, it is important to note that the definition of AI is continually changing, along with the technology itself. The disruptive force of AI technology is commonly defined and utilized interchangeably with ML. However, it is important to understand the distinction between the two, where AI is the broad "head" category that has a variety of subfields, including, but not limited to: ML, deep learning, NLP, data at scale, and simulation (PwC, 2018).

According to several researchers' definitions in the AI field, it can be concluded within four different categories that could describe the definition; thinking humanly, thinking rationally, acting humanly, and acting rationally. These definitions outline two dimensions based on thought processes and reasoning, as well as behavior.

Table 1- Various definitions of AI described within four dimensions

Thinking Humanly	Thinking Rationally
"The exciting new effort to make computers think… *machines with minds*, in the full and literal sense." (Haugeland, 1985)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985)
"[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning…" (Bellman, 1978)	"The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)

Acting Humanly	Acting Rationally
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990)	"Computational Intelligence is the study of the design of intelligent agents." (Poole et al., 1998)
"The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"AI… is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

Source: Russell and Norvig, 2010, p:2

For the purpose of this paper, relating to the service automation of the examination system at CBS, the human aspect of these technologies is the major focus. Acting Humanly involves a computer/software to hold the following proficiencies: NLP, knowledge representation, automated reasoning and ML. With these four aspects, the machine is able to act in similar ways as, or even better than, human beings. Knowledge and information is stored within the software whereby communication in English is carried out and the machine is able to capture what it hears, as well as sees, to create new data and development. This process creates the identification of foreseen patterns to elicit further outcomes (ibid.).

Besides these proficiencies, it is important to note that several myths have been connected to AI, first: assuming that a single AI solution can solve all problems. This is not the case; different types of problems or areas require different types of AI techniques in order to establish a *solution* or *aid* for humans. This links back to the "thinking humanly" in the above table. In order to aid a human, it is important for technology to have a way of "determining" how humans think; described by Russell and Norvig (2010) as the cognitive modelling approach. To determine this, researchers need to understand and observe client thoughts, actions, and reasoning. This is done through introspection, psychological experiments, and brain imaging.

This is followed by the second myth: AI is replacing humans in different industries, in saying that ML learns from data without having any humans involved. Again, at this day in age, it is seen that

ML needs the involvement of humans to acquire and arrange the vast amount of data, and besides, to select, train and guide the machine (PwC, 2018).

In order to show that AI is an umbrella of "human in the loop" with the processes or "no human in the loop", the below table has been established by PwC.

Table 2:Umbrella of AI

	Human in the loop	No human in the loop
Hardwired/specific systems	Assisted Intelligence	Automated Intelligence
Adaptive systems	Augmented Intelligence	Autonomous Intelligence (self-driving cars for example)

Source: PwC, 2018

Assisted intelligence and augmented intelligence are the main drivers of today's economy. Analysis and details are being specified by the robot, an aid to the human, resulting in an information circle, where both are informing each other constantly (ibid.).

Today is thought to be in the third cycle of AI: the first cycle began in the 1990's and was based on narrow AI such as rule based and speech, the second cycle in the 2000's based on narrow AI with an incorporation with big data such as B2C and e-commerce. Although some of these categories are already being developed and underway, it is predicted that within the next five years, AI will be focused on democratization and available to data scientists, home and service robots, and self-driving cars. During the next 20 years, a prediction of collaborative AI and new AI hardware will be developed such as man-machine collaboration, neuromorphic computing and brain-computer interfaces. Finally, AI will be seen in quantum computing, explained in building computers differently, and emotional robots (PwC, 2018).

AI is rapidly growing and by 2030, it is estimated that the contribution of AI technologies will affect global GDP to increase by 14%, roughly US$15 trillion (PwC, 2018). Industries that adopted the AI technology earlier in the "first cycle" include bank and retailers, whereas today's emphasis is

in health-care and manufacturing companies. One industry that has seen a rapid pace of technology is the education sector.

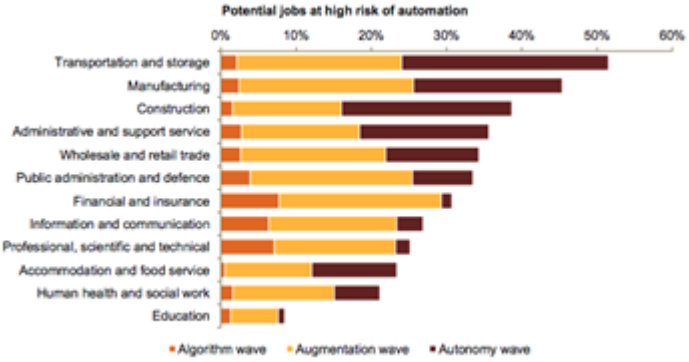
Pearson (2016) refers to this latter trend as Artificial Intelligence in Education (AIEd), which has been around for more than 30 years. Several researchers have been investigating AIEd in regard to the learning aspect, in order to continue looking at the foundation of formal education as well as lifelong learning for the future. This conveys AI with learning sciences: education, psychology, neuroscience, linguistics, sociology, and anthropology, together to support the development of adaptive learning environments along with other AIEd tools (Luckin et al., 2016). These tools can be seen to be flexible, inclusive, personalized, engaging, and effective, because every student is different in one way or another; their environment, learning needs, and emotional state.

On the other hand, professors typically calibrate their teaching to the "average" student in face-to-face classes, which can result in a disengagement of those ahead or behind average students. However, AIEd technologies are currently being adopted and combined with educational data mining, a research field relating to data mining, ML, and statistics from educational aspects, techniques to track behaviors of students. Additionally, novel user interfaces are being examined in order to analyze speech, gesture recognition, eye tracking, and other physiological sensors (Luckin et al., 2016).

The ability of AI customize learning environments creates a more personalized experience for students as AI collects data regarding learning patterns, success patterns, emotional states, and several more in order to create a so called "blueprint" for students (Medium, 2018). Additionally, it can be seen that Pearson and other vendors are currently supporting software applications that focus on learning itself; personal tutors, intelligent support for collaborative learning, and intelligent virtual reality.

However, when researching and implementing these software applications, it is important to note that the role of the professor will continue to evolve, but AIEd will not replace the professor. As shown in Figure 3, the education industry will not be as affected as most all other industries.

Figure 3: Potential rates of job automation by industry across waves



Waves	Description and impact
Wave 1: Algorithmic wave (to early 2020s)	Automation of simple computational tasks and analysis of structured data, affecting data-driven sectors such as financial services.
Wave 2: Augmentation wave (to late 2020s)	Dynamic interaction with technology for clerical support and decision making. Also includes robotic tasks in semi-controlled environments such as moving objects in warehouses.
Wave 3: Autonomous wave (to mid-2030s)	Automation of physical labour and manual dexterity, and problem solving in dynamic real-world situations that require responsive actions, such as in transport and construction.

Source: PwC, 2018

The continuous developments of AI can be categorized by three waves: wave 1 known as the algorithm wave (to early 2020s), wave 2 known as the augmentation wave (to late 2020s), and wave 3 as the autonomous wave (to mid-2030s). It can be seen that the education sector will be affected mostly within the augmentation wave, in the late 2020s. When using AI for decision making tasks and problem-solving tasks that will require a responsive action, it is important to note that professors will be able to allocate their time more effectively and efficiently, while their expertise areas will be better deployed, leveraged, and augmented (Luckin et al., 2016). This will ultimately affect higher educational standards and usage rates for educational institutions.

2.4 AUTOMATED ESSAY SCORING

Currently, AI in education is implemented and focused on the purpose of freeing professors from routine-based activities, described in wave 1 and 2. This can reflect tasks such as the examination process, to let professors focus on more valuable and productive activities like teaching, researching and assisting in individual students' needs. In 1966, even before the time when the concept of AI

was firstly introduced, and even before students used computers to write essays, such a solution as AES was already being tested by Ellis B. Page (Potts, 2005).

Page (1996) came up with the idea of using a computer program to examine essays, as he realized that there was a lack of English writing evaluation on the essays, where professors were not promoting writing quality as they preferred to focus on the learning objectives of their own subjects instead. He was reflecting on the multiple-choice test, popular way of testing subject-matter knowledge in a cheaper and objective way than essay exams. It was, however, a weak knowledge test as it only implied the recognition of information by the student and, as Page (1966) argued, could not test the ability of students to synthesize theories in their own words and analyze facts. To address the skeptical comments of other colleagues on letting a machine to correct exams, Page responded that his solution was "a way to measure essay quality with the same reliability, validity and generalizability - with the same "objectivity" - which they enjoy multiple-choice items" (Page, 1966, p. 239).

AES tools are computer programs that are able to analyze the text of an essay on the basis of several writing quality and content variables that are defined a priori by a human rater. AES tools are nowadays already implemented for the examination of high-stakes written tests. In addition to examining summative assessments, they are also used in formative assessments and, hence, as an instructional tool that is able to provide feedback to students. These tools are typically web-based and include two components: an *electronic portfolio* and an *AES engine*. The *electronic portfolio* component is the platform and graphical interface where students: assess essay prompts, use specific writing tools, upload their essays and receive feedback. The feedback they receive are in two forms: *qualitative* and *quantitative. Qualitative* feedback is given as suggestions for the students regarding improvements of their writing, in order to meet specific qualities. *Quantitative feedback*, on the other hand, either takes the form of a single numeric score, or of different scores that rate the essay on specific traits such as content, creativity, style, mechanics (spelling, capital letters and punctuation) and organization (essay structure quality) (Shermis, 2010). The *AES engine* is the component that scans the essays through statistical algorithms that are built on the concepts of ML and NLP and then evaluates them. As Mark Shermis stated in an interview, the *AES engine* is the component in charge of providing summative evaluation while, the *electronic portfolio* is the

part that is able to come up with qualitative feedback for the student (personal communication, 6 March).

AES tools provide *qualitative* feedback through discourse analysis by scanning a paper and identifying the main points that the writer made within the paper. If for instance, a writer is making three points on a specific argument, the tool is able to determine how much information the writer is giving for each of the three points. If the writer has not given enough information for a specific point as they did for the other two, the AES tool will point this out, suggesting to add more information. However, it will not be able to tell the student which critical argument they are missing. This is because the software is only able to draw on the essays that were used to formulate the statistical model of the specific prompt topic (ibid.).

Even though there is a variety of different AES tools available in the market, a general procedure that is followed to develop these programs. In relation to the exam where AES will be applied, the initial step is to primarily design the methodology on how essays will be evaluated. This will imply human evaluators to develop a map, or a rubric, as previously defined, that will explain in detail the different levels of performance (scores) and the specific features that are characteristics of an essay within each score level typically. With the purpose of additionally creating the feedback structure of the software, human evaluators will also have to identify typical errors and come up with feedback that will help the student address them. As a next step, a sample of 300-500 essays has to be collected. To build the quality evaluation model of the software, each of the collected essay has to be already evaluated by two human evaluators. The more evaluations on the collected essays are rated, the higher the probability that the AES software will come up with the same scores of the human evaluators. Within this sample set, around 300 essays have to be randomly selected and scanned through various text computational analyses. These will then be evaluated according to different features of the essay quality and against which the human ratings will be regressed. This will allow the creation of a regression equation that will be the basis of the model used to evaluate the essays. In order to evaluate the accuracy of the model, the final step will entail the cross-validation of the regression equation on the remaining set of essays that were not used for the regression (Shermis, 2010).

An AES software evaluation model can be *generic* or *prompt-specific*. A *generic* model is developed to score a particular genre of writing or developmental level. It is not designed to evaluate the content of an essay, but rather general writing ability. Hence, this type of model does not require professors to continually update the scoring algorithm, and it is easier to implement even without asking for the help of a computer science expert. A *prompt-specific* model, on the other hand, can also evaluate the content of an essay and say whether the student has made a weak argument depending on the amount of information they are including within the paper. In this case, the model has to change each time for an exam type, and hence the exam prompt changes, and its efficacy increases when there is a large set of already evaluated past exams that are based on the same prompt (Mark Shermis 2018, personal communication, 6 March).

However, it has to be considered that, so far, these types of softwares have mostly been used to assess and improve English writing and that most of them are hence built with a *generic* evaluation model. For this reason, most of these tools follow evaluation criteria's that are based on writing quality standards as the 6+1 Trait® and the Common Core State Standards (CCSS) initiative (Shermis et al, 2016).

The 6+1 Trait® is a rubric model developed by Education Northwest which establishes that the writing quality of a paper depends on the following traits: *ideas* (the main message)*, organization* (the structure of the paper), *voice* (the personal tone)*, word choice* (the vocabulary used)*, sentence fluency* (the flow of the language)*, conventions* (mechanical correctness) *and presentation* (how the writing looks) (Education Northwest, 2018). CCSS is a set of high-quality academic standards that outlines the learning goals that each student should achieve through their K-12 education before going to college. These standards are used by AES tools to evaluate the organization and development of an essay by considering the presence, or absence, of relevant discourse units. These include an introduction, thesis statement, main ideas, supporting details, and conclusion (Shermis and Burstein, 2013). Besides the organization and development, as described on corestandard.org (2018), the standards are based on the following:

- Research- and evidence-based
- Clear, understandable, and consistent,
- Aligned with college and career expectations,
- Based on rigorous content and application of knowledge through higher-order thinking skills

- Built upon the strengths and lessons of current state standards, and
- Informed by the other top performing countries in order to prepare all students for success in our global economy and society.

Concerning the validity of these tools, different studies have been conducted to analyze the percentage of agreement between the AES and human raters scores. To evaluate the agreement percentage, the Cohen's kappa coefficient (k) is used, a statistic that measures agreement by taking into account the possibility of agreement by chance (Shermis, 2014): the kappa statistic ranges from zero to 1. The closer it is to 1, the higher the agreement there is with the human scores. A quadratic weighted kappa score of 0.81407 was achieved by the winning team of the "Automated Student Assessment Prize" sponsored by the Hewlett Foundation (Hubert.ai, 2017) and a team at Carnegie Mellon University that built the AES engine *LightSIDE*, which achieved a kappa score of 0.833. It is hence considered that these types of software's can now replicate human evaluators scores and can be used in high-stakes assessments (Shermis, 2014).

Notwithstanding the high agreement rate, AES tools still has a number of limitations with regard to its overall acceptability. For example, they do not evaluate an essay in the same way that a human rater reads and understands a piece of writing. They are not able to evaluate the cognitive, interpersonal and intrapersonal aspects in a piece of paper (Shermis, 2014). With respect to an automated engine, a human rater when examining a paper brings background knowledge and expectations, in order to get complex points and to evaluate the writer's knowledge depth of the writer. Additionally, AES engines cannot reward the writer for mentioning specific points of the expected literature and using an ironic and humoristic writing style (Shermis, 2013).

Bennett (2011) suggests a list of features on which these tools should make improvements in the future to better enhance the effectiveness of their use. The author points out a modification of the design of AES tools that allows an integrated assessment process where the automated engine and the human scorer perform interrelated roles. To increase the agreement with human evaluator scores, it is of high importance to enhance the understanding of human scoring processes. To this extent, an extensive disclosure of examining approaches by professors is required and this thesis has the purpose of contributing to this research.

It is relevant to note that AES has been implemented in some US state-wide high-stakes examinations. For example, scoring engines have been used in the state of Utah, Louisiana, West Virginia, and are being considered for the state of Ohio. Moreover, the e-rater AES engine, further described in Chapter 5.2.2.2, was firstly used in 1999 for the writing section of the Graduate Management Admission Test (GMAT). Nowadays, it is used as a check score for the Graduate Record Examination (taken by business school applicants and prospective graduates) and in the computer-based Test of English as a Foreign Language (TOEFL iBT) test (Shermis et al, 2016).

At the higher education level, other universities are looking at personalized solutions that improve writing. University of Michigan, for instance, is using AES to solve "the feedback gap" that was experienced at their M-Write program. Here a team was established to develop course-specific algorithms that could signal whether the student has not understood a specific topic or concept (Brown, 2016). Massachusetts Institute of Technology (MIT) is currently working on a software system that can help automatically examine assessments that are done on the EdX platform, which is a Massive Open Online Courses (MOOC) provider that has been developed by MIT in collaboration with Harvard University. According to MIt is Computer Science and AI Laboratory, AES would help solve the issue of MOOC courses of having to examine a high number of exams (Hardesty, 2013). Another researcher, Peter Vitartas, associate professor in Marketing at La Trobe University in Australia, is carrying out research on the development of an AES system that aims at supporting students with their writing skills, which has also already resulted successful in the provision of feedback to students. According to Vitartas, AES is an efficient solution for Australian universities as classes can range over 1000 - 2000 students. Vitartas's research is at the moment looking at how to automatically grade critical thinking and students' own research insights by looking at the validity of the references provided by the students in the papers (Peter Vitartas 2018, personal communications, 6 March). Last but not least, in the Scandinavian area, two researchers of the Department of Linguistics at the University of Stockholm, Robert Östling and Andre Smolentzov have tried to develop an AES system for examining high school essays in Swedish, which has to however improve efficiency in order to be implemented in a practical setting (Stockholm University, 2013).

Having touched on the research field of available AES tools in the world today, it is important to analyze the potential of AI opportunities within the case study of CBS. By doing so, a description of

the grading scale system in Denmark will be provided, as well as a description of the different types of exams to provide an overview of the Danish higher education system.

2.5. INTRODUCTION TO CBS CASE STUDY

Established in 1917, CBS is an international business school teaching over 21,000 students and employing 1,500 employees. Since 1917 until 1971, there were no standard marking schemes, thus individual departments created their own. Introduced in 1971, a 00 to 13 grading scale was used, whereby grades could be placed within 4 different groups according to the performance of students: (1) Where 13,11,10 are excellent (2) 9,8,7 are average (3) 6 are just acceptable and (4) 5, 03, 00 are hesitant.

As of August 2007, Denmark enforced a new 7-step grading scale, shown in Table 3 below, to create more compatibility within an international context, specifically the European Credit Transfer System (ECTS) grading scale (Eng.uvm.dk, 2018).

Table 3: 7-Step Danish grading system

Grade	Description	ECTS
12	For an excellent performance displaying a high level of command of all aspects of the relevant material, with no or only a few minor weaknesses.	A
10	For a very good performance displaying a high level of command of most aspects of the relevant material, with only minor weaknesses.	B
7	For a good performance displaying good command of the relevant material but also some weaknesses.	C
4	For a fair performance displaying some command of the relevant material but also some major weaknesses.	D
02	For a performance meeting only the minimum requirements for acceptance.	E
00	For a performance which does not meet the minimum requirements for acceptance.	Fx
-3	For a performance which is unacceptable in all respects.	F

Source: Cbs.dk, 2016

This 7-step scale is based on the overall performance of a student and on the academic requirements. The grade of 02 is the lowest to receive in order to pass. Students are graded at the

end of each course through oral and written exams. Written exams can either be sit-in-exams where the student has to write an exam at the university, home-taken assignments where the student has a limited period to write the exam outside of the university (24, 48 or 72 hours) and projects where the student works individually or in groups on a theoretical problem (Copenhagen Business School, 2018b).

Regarding the teaching formats, the academic year 2018/2019 will allow all professors to pick one of these three formats: *face-to-face teaching, blended learning* and *online teaching*. *Face-to-face teaching* is performed through standard classes with students on campus where the professor has the possibility of using online materials and tools. *Blended learning* gives the professor the possibility of mixing face-to-face lectures with online activities that can take the form of video lectures, quizzes, discussion forums, online peer assessment and the use of MyEconLab platform (Vice Dean of Education 2018, personal communication, 20 April). Online teaching takes place mainly online through online teaching and virtual classrooms, but some activities might take place at the university, as the introductory lecture of the course or Q&A lectures at the end of the course (Blog.cbs.dk, 2018).

Drawing upon the literature available in the different topic areas described, it becomes clear how AI has the potential to enhance of the quality of higher education. In particular, due to the increasing number of students participating to the same courses, higher education is nowadays demanding the use of automation technologies that are intelligent enough to overcome the individuality of both each student's learning process and professor's teaching style. As previous technology could be easily applied to repetitive and rule based activities that did not require human judgement, there is currently the need of developing digital tools that are able to take complex decisions as, in the context of this research case study, the examination activity. In line with this need, the following section will build a framework that will help build an AI prototype by understanding the individuality of the users of the examination activity.

3. THEORETICAL FRAMEWORK

This chapter aims to review the relevant theories of service dominant logic and value proposition design, to establish an initial theoretical framework. This will enable the reflection on the servitization of the examination field, by involving technology and the opinions of stakeholders to design a proposed roadmap for the case study of CBS. This will aid with analyzing the data and answering the research question.

3.1. A DESCRIPTIVE REVIEW OF SERVICE DOMINANT LOGIC

AI is decreasing bureaucracy and increasing automation, which results into the simplification of time consuming activities. As the automation of work and business processes is continually on the rise, as previously described, a Service Dominant Logic (SDL) approach can be viewed. With this approach, a general tendency view on the demand side of economics can be analyzed, whereby growth is generated through high demands of products and services, that are becoming more and more service oriented. This entails a world that is customer oriented, whereby the logic approach focuses on the servitization of products.

3.1.1. Evolving to a SDL Perspective

The emergence and evolution of SDL changed in 2004, when Vargo and Lusch (2004) first introduced the new perspective based on shifting the idea of the role of service, in regard to exchange and value creation. For decades' past, dominant logic was based on the exchange within a Goods Dominant Logic (GDL) view, focusing on tangible resources, embedded value, and transactions (Vargo and Lusch, 2004), in particular manufactured goods. However, Vargo and Lusch sought the opportunity to view a new perspective that focused on the economic exchange of a more service-oriented offering, that is embedded within intangible resources, the co-creation of value, and relationships (ibid.).

Although the perspective has emerged from Vargo and Lusch, it has been indicated in research that there are ways in which services are different from goods. Looking back two decades ago, research from Gummesson (1995, p.250-51) indicates that:

> "Customers do not buy goods or services: They buy offerings which render services which create value… the traditional division between goods and services is long outdated. It is not a matter of redefining services and seeing them from a customer perspective; activities render services, things render services. The shift in focus to services is a shift from the means and producer's perspective to the utilization and the customer perspective" (Vargo and Lusch, 2004).

As this shift of services is more focused on the customer perspective and utilization, it is essential to view resources, as they are key to understanding the new SDL logic approach. Looking back, Edith Penrose (1959), one of the first economists, recognized the shifting role and view of resources.

Although Penrose studied the growth of firms, known as *The Theory of the Growth of the Firm*, it can be interlinked with the new logic approach. Penrose's (1959) basic assumption for the definition of a firm is a bundle of resources, build up over time and managed by administrative unite. Through this bundle of resources, Penrose (1959) states that the uniqueness of every firm is based on the distinction between resources and the services that these resources provide: "it is never resources themselves that are the "inputs" in the production process, but only the services that the resources can render" (Penrose, 1959, p.24-25).

As the world economy has shifted to a service orientation, Constantin and Lusch (1994) classified two types of resources: operand and operant resources. Operand resources are those resources that have been produced through an operation or act, such as a physical tangible good. Whereas, operant resources are often invisible and intangible, action is normally taken to create operand resources such as skills and knowledge.

3.1.2. Distinction between SDL and GDL

In a GDL centered view, operand resources are the primary source of factors of production. A firm, as previously described through Penrose, has operand resources, that of the factors of production, along with operant resources, the technology behind the production. With both resources, the firm can create value, reflecting upon the new technological era that the world faces, whereby generating a focus on digital "things" and ultimately deleting specific work tasks. By doing so, "customers, like resources, become something that needs to be acted or looked upon in order to penetrate the market" (Vargo and Lusch, 2004, p.2), by obtaining more and more customers: ultimately, a world that focuses on operand resources. In contrast, in a SDL centered view, operant resources are the primary source of producing effects. This creates a world in which humans can create additional operant resources, by adding value to natural resources. Skills and knowledge of humans are utilized and are the primary source of resources within exchange processes, markets, and customers (Vargo and Lusch, 2004).

With contrasting views between a SDL approach and a GDL approach, Vargo and Lusch (2006) have outlined the primary concepts of the two, in conjunction with the transitional concepts that occur between the two.

Table 4: Conceptual transitions

GDL concepts	Transitional concepts	SDL concepts
Goods	Services	Service
Products	Offerings	Experiences
Feature/attribute	Benefit	Solution
Value-added	Co-production	Co-creation of value
Profit maximization	Financial engineering	Financial feedback/learning
Price	Value delivery	Value proposition
Equilibrium systems	Dynamic systems	Complex adaptive systems
Supply chain	Value-chain	Value-creation network/constellation
Promotion	Integrated marketing communications	Dialogue
To market	Market to	Market with
Product orientation	Market orientation	Service orientation

Source: Vargo and Lusch, 2006

3.1.3. Definition of SDL

Not only are humans utilizing their full skill and knowledge potential in a SDL centred view, but also in a market driven economy where marketing is customer centric (Sheth, Sisodia, and Sharama, 2000). A market that is customer-centric involves collaboration with customers in order to listen and learn from their dynamic needs and wants. This reflects back to the world being more customized and personalized for individuals, in order to create additional value. The value that emerges from this dominant logic can be viewed in the 9 Foundational Premises (FP) that Vargo and Lusch (2006) describe, which have been revised in *Service Dominant Logic Reactions, Reflections, and Refinements* (2006), following their first published paper:

FP1: The application of specialized skills and knowledge is the fundamental unit of exchange
FP2: Indirect exchange masks the fundamental unit of exchange
FP3: Goods are distribution mechanisms for service provision
FP4: Knowledge is the fundamental source of competitive advantage
FP5: All economies are services economies
FP6: The customer is always a co-creator of value
FP7: The enterprise can only make value propositions
FP8: A service centered view is customer oriented and relational
FP9: Organizations exist to integrate and transform micro-specialized competences into complex services that are demanded in the marketplace

3.1.4. How to Apply SDL to the Servitization of the Examination Activity

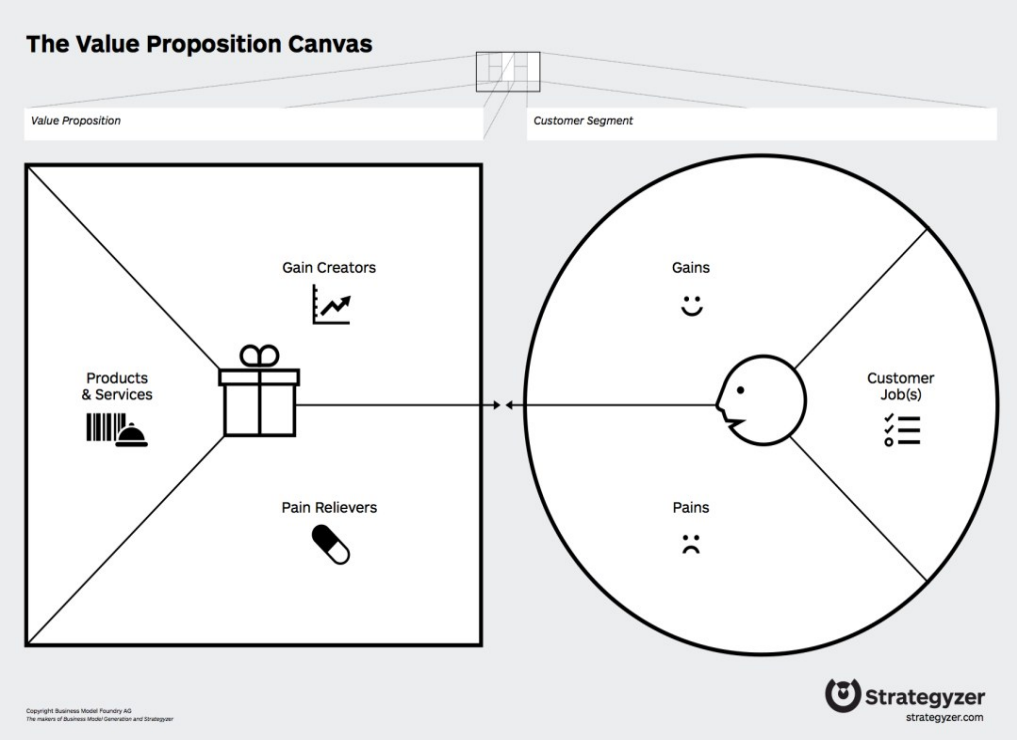
Research within SDL has been conducted in many areas, but there are several areas that are yet to be explored. For the purpose of this thesis, the discussion and analysis on the link between AI and education will be explored using a SDL approach. Nevertheless, clarifying the theory does not provide a full understanding of how this approach can be applicable to the servitization of examining in a higher education setting. Thus, the imperative task is to apply that knowledge with the Value Proposition Design (VPD) framework (Osterwalder et al, 2014) in order to understand how the potentials of AI, within the context of CBS, can benefit all stakeholders.

3.2. VALUE PROPOSITION DESIGN CANVAS

Osterwalder et al (2014) VPD framework is used in this paper as a way to understand the two most important stakeholders of the examination process, professors and students, and to design a solution that collaborates the two. As professors and students have very different needs and experience the examination process in two different ways, following the VPD framework, two different customer profiles are created. Even though two different customer profiles are created, it can be observed that both of them participate in the examination process: professors as "active" users as they are the ones that evaluate the performance, and students as "passive" users as they receive the grade and feedback. Therefore, a single value proposition will be created throughout the paper by looking at the interconnection between these two different stakeholders.

The VPD canvas, developed by Osterwalder et al (2014), is a theory that is used in human-centered innovation processes for improving or developing new products and services. Its main contribution lies with the idea that an organization can create real value only after understanding the individuals to whom they are offering a unique solution. This is accomplished by creating a solution that fits perfectly with their profile. As shown in the picture below, the framework consists of two distinct parts: the *customer profile* and the *value map*. These parts have to be mapped out for every distinct customer segment that an organization or firm wants to serve.

Figure 4: The Value Proposition Canvas



Source: Strategyzer.com, 2018

The *customer profile*, shown on the right-hand side of the picture, is the first part of the framework that has to be completed. It entails the observation of customers and the precise description of the specific customer segment that a solution is tailored to. Osterwalder et al (2014) suggests different and multiple techniques that can help fill out the blanks in the customer profile canvas. An innovator can start off with the analysis of user data and draw upon marketing research insights to find specific inputs on where to focus the research. Following, customer interviews can help innovators to answer questions on missing patterns and to understand what matters most and least to the customer. The author clearly specifies that these types of interviews have to be conducted with a "beginner's mind" (Osterwalder et al, 2014, p: 112) and without having too many expectations on what the answers will be. By posing open questions and not asking for mere opinions, the customer will have the possibility to open up and not be biased on the answers they will provide.

After collecting the data from first-hand and second-hand research, the customer segment has to be mapped out by distinguishing between *gains*, *pains* and *jobs*. Customer *jobs* are the tasks that a customer wants to get done, the problems of which they are trying to find a solution, and the needs

they are seeking to satisfy. Customer *pains* are related to customer jobs and refer to the undesired outcomes, obstacles and risks that bother the customer while performing a specific job. The last part of the customer profile are customer *gains*, which includes all of the positive outcomes that a customer wants to see coming out from the performance of a job. At the end of each section, all *pains, gains* and *jobs* should be ranked from the most to the least important, in order to gain an overview of what the real priorities are of the specific customer segment and to create a solution that originates out of them (Osterwalder et al, 2014).

After drawing the *customer profile* of the customer segment that a solution wants to serve, the *value map*, shown on the left-hand side of the framework has to be filled out. The goal of mapping out the value proposition canvas is to guide the creation of value as a response to the customer profile. As a result, the entries of the value map are named *gain creators*, *pain relievers* and *product* and *services*. *Pain relievers* explain how the product or service solution aim at resolving and reducing specific customer *pains*. *Gain creators*, on the other hand, tackle the outcomes that are already mentioned in the *gains* part of the customer profile that a customer will get out of the solution offered. Lastly, the *product and services* section helps with disclosing the final solution and it includes all of the different products and/or services that are included within the entire value proposition.

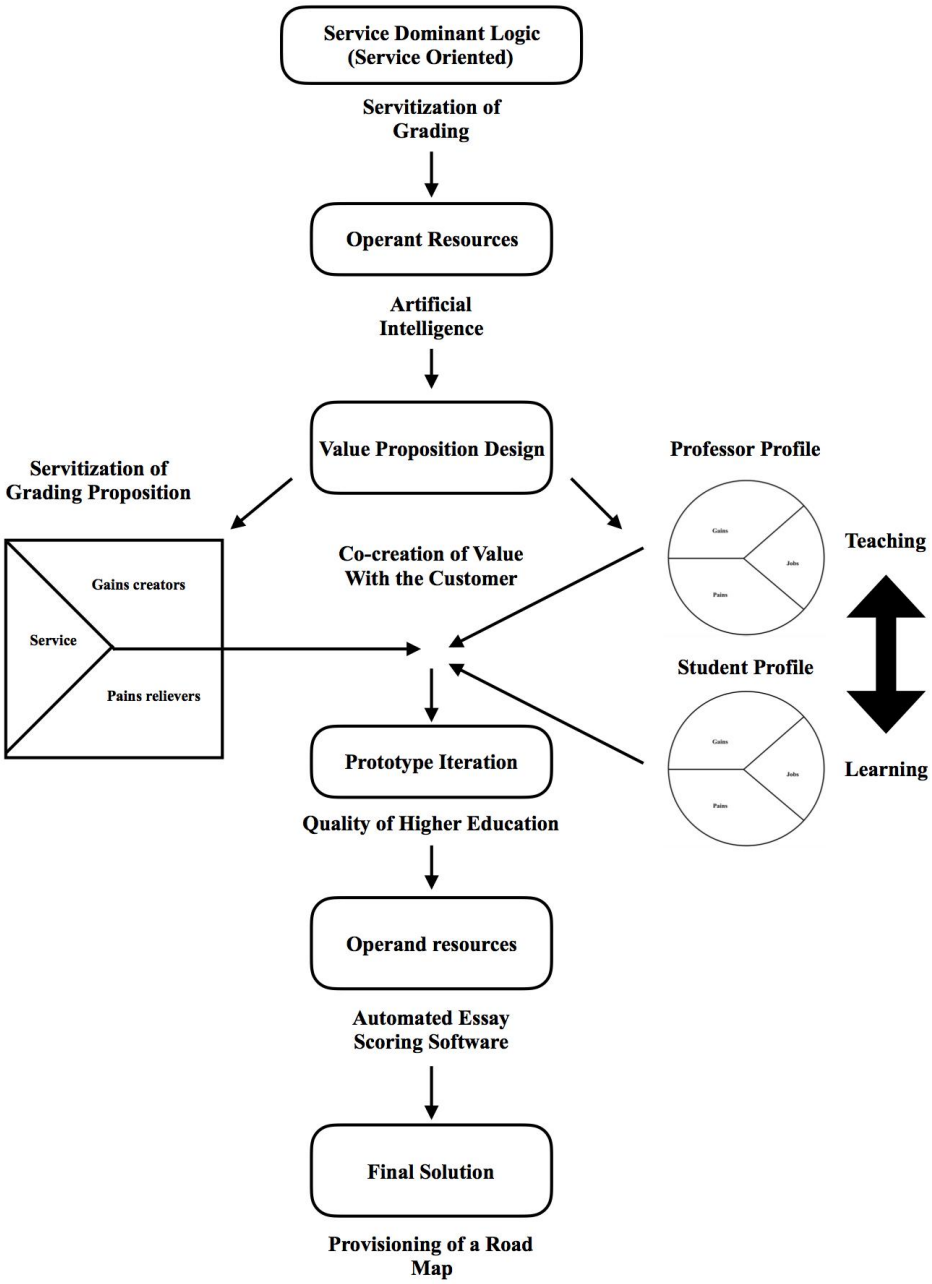
The last stage is to find a "fit". Osterwalder et al (2014) suggests that there are three stages of fit that are related to the level of maturity of the solution from prototype to final product/service: *problem-solution fit*, *product-market fit* and *business model fit*. *Problem-solution fit* is achieved when it is proven that customers' most relevant jobs, pains and gains are the ones tackled by the solution, even though at this stage it is not yet proven that they will in fact use the solution. *Product-market* fit is created when customers start showing interest of using and buying the designated solution and see the real value they can get out of it. The final stage of success is then reached when the solution has a *business model fit* and, hence, when it is proven that there is a stable business model that can be profitably sold, in a sustainable way.

3.3.INITIAL THEORETICAL FRAMEWORK

When looking at an SDL approach, for the purpose of this research on *the potentials of AI in examinations at CBS,* it is important to understand this view in a customer centric world. In order to do so, the selection of the VPD canvas was chosen to analyze the nature of the process of the innovation that is being created- the servitization of the examination system. The examination process of assessments has been, so far, an extremely human related activity that has always been a subjective process for the experts, teaching professors, performing it and the different performances of students. Furthermore, as each educational institution has a different approach to examining and as CBS has been selected as the case study of this paper, prior to this research, it was not clear how CBS professors carried out this activity. Consequently, there was a need to find a framework that could have been used to gather an in-depth understanding of the specific customer segment, aligning it with the service era of today's economy.

Likewise, as the initial intention of this research was to focus on the examination process, it was also necessary to test whether this was really an activity that CBS felt the need of improving, reflecting on the stakeholders involved. In this case, the SDL theory and VPD canvas is a way to analyze the shift of the role of services, as well as identify what the real needs of the specific stakeholders are. Finally, yet importantly, this integrated framework will be applied in the analysis of this thesis to understand whether an AI solution is a feasible solution for the servitization of the examination process at CBS. Furthermore, an analysis of the conceptual transitions, previously described in Table 3, will be analyzed to see how the innovation process can transition from a GDL approach to a SDL approach, in connection to the pains and gains that stakeholders defined.

Figure 5: Initial theoretical framework



Based on: Ramsden (2003), Shermis and Di Vesta (2011), Willcocks and Lacity (2016), Page (1966), Shermis (2010), Shermis et al (2016), Shermis and Burstein (2013), Vargo and Lusch (2004), Constantin and Lusch (1994), and Osterwalder et al (2014).

In accordance to the SDL approach being primarily focused on the operant resources, the skills and knowledge that are to be utilized, further connected to the VPD, can be observed where co-creation of value happens. To outline the operant resources needed for the solution, the first step consists of mapping out the value proposition along with the professor and student profile. A fit between the two is achieved by mapping out features of the solution in the value proposition canvas that respond to the insights collected in the professor and student profile canvases. The final solution will consist of provisioning a roadmap that reflects the goal of pursuing the solution in the future. The roadmap involves a prototype, namely the co-creating activity between the customers, professors and students, the users of the service, to gain an understanding of their dynamic needs. After the process of continually iterating, aiming to achieve improvements of the value proposition, the operand resources that are needed for the creation of the final solution can be outlined.

In essence, as the general tendency view on economics is about becoming service oriented, a SDL approach is taken to analyze the servitization of examination. This aspect focuses on the world becoming more service oriented, while also focusing on digitalization. This approach is conducted through interlinking the VPD canvas to analyze the opinions of the different stakeholders: professors, teachers, and management. Together, the initial framework illustrates a strong bond between both theories that will be utilized to further explore this research thesis.

4. METHOD- IN THE SHOES OF CBS STUDENTS

This chapter aims to present how the theoretical framework is used with an analytical approach. It describes the different phases and steps that were taken to conduct the research study, outlining the different data gathering techniques. Further, it identifies the quality of the research by being reliable and valuable, with an evaluation of the process thereafter.

4.1. METHODOLOGICAL CONSIDERATIONS

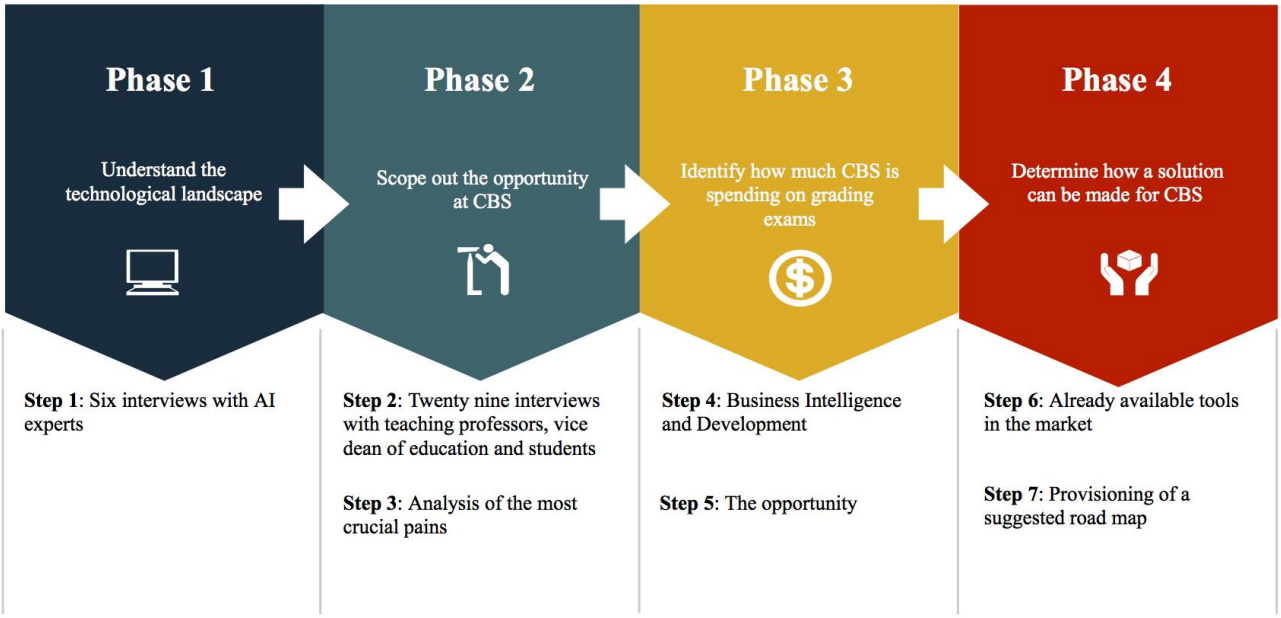
After reviewing the relevant literature and theories, it became apparent that the research field between AI and education, using a SDL approach, has yet to be explored. Thus, based on the research question of *What are the potentials of AI in examinations at CBS*, the research method of a single case study will be employed. As identified by Yin (2013), a case study examines an existing experience whereby the analysis of a real-life context is conducted in order to identify and analyze the knowledge gaps that are unexplored, with no clear evidence. Hence, being two students of CBS, we, as researchers, wanted to understand in depth, the real-life phenomenon of the examination system at CBS.

Pettigrew (1990) suggests that choosing a case study is advantageous when the situation is a vivid one where the process is "transparently observable". As CBS students, this was seen as an observation when receiving grades, as well as receiving limited feedback. Hence, the examination process was of interest to observe and later, provide a suggested innovative road map to enhance students and professors' experience at CBS. By understanding the different variables within the proposed innovation and by looking at the integrated framework in Figure 5, the research question can be broken down into four components: (1) How is AI technology emerging in the world today, (2) How is the examination system structure seen by CBS stakeholders, (3) How much is CBS spending on examining exams, and (4) How can the grading system be improved by applying new AI technology.

4.2 ANAYLTICAL APPROACH

Figure 6 illustrates an analytical approach to describe the different phases that were conducted throughout the process.

Figure 6: Analytical approach



Given these four components of the research strategy, it was important that the researchers, as Yin (2013) states, are familiar and understand the dynamics of the broader scope of theories and terms that reflect the research that is to be conducted, known as *theory development*. Through theory development, the primary task was to analyze the AI aspect and understand how the world today is emerging in a whole new era with AI technology. This can be reflected upon operant resources that are emerging based on the skills and knowledge of humans today. At the same time that operant resources emerge through operand resources, it was essential to connect the education industry, with the emerging AI technology, to analyze the examination process at CBS, known as the case study. Additionally, the topic was discussed with fellow students, to gain the learning aspects of education, and in turn provided ourselves with challenging questions about the topic, in order to overcome the barriers of the theory development (Yin, 2013).

Through the challenging questions that arose in the beginning of the research, it is essential to describe the method of findings.

4.2.1. Phase 1- Understanding the Technological Landscape

Step 1- Six interviews with AI experts

The first step was to gather a broader understanding of the topic of AI and AES. Six interviews with experts in this field were conducted, via Skype and face-to-face. Interviewing experts, as IDEO.org (2015, p. 43) states, allows to get a system-level view of the project area that is being researched. This was indeed crucial to extend the knowledge, known as the operant resources, on specific applications of this technology, gathering opinions on the efficacy of such an application, guidance on how to structure the innovation process and ideas on how to tackle the AI opportunity while co-creating with stakeholders. A list of all the experts that collaborated in this process is provided in section 4.3.3.

4.2.2. Phase 2 - Scope out the Opportunity at CBS

Step 2- Twenty-nine interviews with professors, management and students

After having gathered extensive knowledge on the topic, and moving to the case study, the following step consisted of going deeper into the activity of how humans examine assignments to understand how the stakeholders: professors, management, and students, were experiencing it. The VPD canvas, explained previously in Chapter 3, was used as a guide to collect and map out stakeholders' insights (pains, gains and jobs). The collection of stakeholders' insights was carried out by conducting 29 face-to-face interviews with professors and students. This process is backed up by the premise that interviews are an effective way to appreciate personal insights of the subjects interviewed, as interviewees have the possibility to explain their own experience and opinions in their own words. The structure of the interview questions will be explained in Section 4.3.2.

Step 3- Analysis of the most crucial pains

It was noted that there was a common severe pain for students: the lack of feedback on their writing. This pain had to be investigated further, as according to De Vaus (2014, p.9) "observations require explanation but equally explanations need to be tested against the facts". For this reason, a survey was conducted with the scope of: 1) quantifying the issue by determining the severity of the pain felt by students, 2) determining the event in which it was mostly occurring (ex: type of examination) and 3) analyzing whether this pain was mostly common among students with a specific trait (ex: higher grades vs lower grades). The data that comes out of a survey is hence structured in a way where every subject analyzed, in this case single students, can be then compared to the others in a structured manner (De Vaus, 2014). The technique chosen to generate the data for the questionnaire is further described in Table 7 of this chapter.

4.2.3. Phase 3- Identify How Much CBS is Spending on Examining

Step 4- Business Intelligence and Development

As the opinions of stakeholders outlined that the examination process was an area that could improve, evidence behind the pains was an important part of the research strategy. As Eisenhardt (1989) states, it is typical that a case study combines multiple data collection methods, including qualitative and quantitative data, to provide a stronger substantiation behind the research. Therefore, as a qualitative approach had been conductive thus far, obtaining quantitative data was key to create a combination for this study to become "highly synergistic".

As follows, research was conducted to determine that the Business Intelligence and Development (BID) team at CBS is responsible for overviewing the coordination and business development at CBS. Through data analysis, the team utilizes internal and external sources in order to sustain the processes (Copenhagen Business School, 2018a). Contact was made to BID, requesting:

- How much money and time, represented in DKK and hours, do teaching professors utilize to different activities based on their profession, for all bachelor and master programs (refer to Table 9 and Figure 11 in section 5.2.1.

    - Why: To acknowledge how resources are consumed by the different professor activities at CBS.
- How much expenditures are spent on thesis (refer to Table 10 in section 5.2.1).
    - Why: Given the results above, fluctuations occurred due to the Danish Government's study reform for Danish universities reflecting on thesis hand-ins. This data confirmed the fluctuations shown in the results.
- How much money is provided by grants from the Ministry of Higher Education and what are the breakdowns of the grants (refer to Table 11 in section 5.2.1).
    - Why: To gain an overview of each course and gain an understanding of how much money is budgeted based on different factors.
- How much money, in year 2017, did CBS spend on all professors' activities (refer to Figure 12 in section 5.2.1).
    - Why: To observe the total consumption that CBS spent, of their grant provided from the results above.

Data was extracted from a cross-table for the "targit-extract" solution that was requested from the business intelligence and analytics software, Targit. In the software, a query was submitted, in order to obtain the desired results for the specific data requested. Data was obtained through Excel spreadsheets and was collected from the past four years, to determine any fluctuations within the different years (BID 2018, personal communication, 2 March).

Step 5- The opportunity

After collecting the data from BID, a manual extraction of the Excel data was executed, which had confirmed our hypothesis that a significant amount of costs was associated with the examination activity at CBS, refer to Chapter 5.2, for further explanation.

4.2.4. Phase 4 - Determine How a Solution can be Made for CBS

Step 6 - Already available tools in the market

As the initial hypothesis of this study was to innovate the examination system by using AI to automatically examine written papers and, as there was no knowledge on whether the market had already a viable alternative, an internet-based research was initially conducted. The keywords used to browse the web were: examining exams with AI and AI in education. This initial web-research brought up the discovery of the term "AES" which proved that solutions were existent and was the base of further study. Following, was then a research on how these tools functioned and what were the different solutions already offered as a product/service. For this purpose, extensive reference was made to the work of Mark Shermis and his colleagues who have studied and published in the AES domain for the past twenty years. Besides, a review was conducted of the different solutions features, that was retrieved from each producer companys' proprietary website.

As several AES solutions are on the market today, it was essential to outline the commonalities and differences shown within section 5.2.2. After the creation of a list of the already available tools in the market, the outcome of Phase 6 was an AES table, refer to Table 12, that outlined all of the different and common characteristics of these tools, later described in depth in section 5.2.3.

Step 7- Provisioning of a suggested roadmap

The AES table was used as a starting point to highlight which features these softwares lack and that could have then established a viable fit with the pains, gains and jobs of professors and students that came out from the study conducted. Accordingly, Phase 7 consisted of the development of a prototype solution that was iterated to different versions on the base of the already available tools' limitations and the different stakeholders' insights gained during the different phases of this research. A suggested roadmap is presented and explained in detail in Chapter 7 to guide CBS in the adoption of the servitization of the examination system in the near future.

4.3. DATA COLLECTION

4.3.1. Primary and Secondary Data

Both primary and secondary data has been collected for the purpose of this study, an overview shown in Table 5. Primary data as Salkind (2010, p.2) explains, consists of "firsthand, unmediated information that is closest to the object of study". In this research project, primary data hence included all of the collected data on the selected case study: that of interviews with CBS teaching professors, management, and students, a student survey and experts' interviews.

Secondary data, on the other hand, is other sources of data that have been created by other authors and are also available for purposes that are different than the original one for what they have been produced (Salkind, 2010). In this group falls the data on CBS' examination system that was provided by the BID department, AES tools specifications data obtained from the producer companies' proprietary website, the Danish regulations on examinations and several authors and firms' reports, books and articles founded both online and on paper.

More specifically, the data can then be divided among the one that was used to build general knowledge on the topic of AI and AES (reports, books and articles), and the one used for the case study itself.

Table 5: Primary and secondary data

Primary Data	Secondary Data
Interviews with CBS teaching professors	CBS BID data
Interviews with CBS students	AES tools specifications
Survey with CBS students	Danish examination regulations
Experts interviews	Existing literature

4.3.2. Interviews Structures

A qualitative research has been carried out in order to get an insight perspective of the social phenomena that was the subject of the case study: the examination process of exams at CBS. As Kvale (2007) explains, qualitative research helps to understand the views and experiences of the stakeholders of the social phenomena studied in their natural context. For this research, two types of interviews have been conducted: interviews with experts and interviews with the stakeholders. Most interviews were conducted face-to-face, to create a more personal dialogue, but also to observe the interview in a different perspective by respondents' body language and emotions throughout the responses.

The structure of the expert interviews was held as a branching conversation and as each expert had a different background, each interview questions varied. On the other hand, interviews with management and professors had a similar structure, as the aim was to find out similar insights and patterns on which the final suggested roadmap should be based on. In order to pursue this goal, semi-structured life-world interviews were used. These consisted of a list of open questions regarding the examination topic, and, depending on the insights that the interviewee introduced, other unplanned questions followed up directly during the interview event (Kvale, 2007). When talking to students, a short dialogue was held by using an unstructured interview model on their experiences concerning their satisfaction (or dissatisfaction) with different aspects of CBS while observing their behaviors when talking to gain answers in divergent ways. This was done in order to obtain both a factual and meaning level (Kvale, 2007). The factual level was achieved when the students talked about something that has happened and what they thought about a specific topic. The meaning level, on the other hand, consisted of us, interviewers digging deeper in the whys of the facts introduced by the students and asking them to explain more in details without letting us show personal beliefs that could have created bias in the following answer (Kvale, 2007).

Professor and AI expert interviews were transcribed in order to extrapolate the most important insights. For the professors' interviews, a table was used to categorize the insights, shown in Appendix 6, according to: professor's examination style and strategy, feedback provision activity, personal motivation in human examination and activities that they felt to be needed to be improved.

Students interviews were not transcribed as they were not recorded, however, all key findings for the interviews can later be seen in Chapter 5.1.2, Figure 10.

4.3.3. Interviews Conducted

By addressing and constructing multiple sources of data collection to obtain strong evidence through analytical data, a total of 42 interviews were conducted: 39 via face-to-face, 1 via phone, and 2 via Skype, as shown in Table 5. The interviews lasted approximately 40 to 60 minutes and most interviews were transcribed, shown in Appendix 3 and 6. When selecting individuals, it was important to seek experts from different areas of the research question when it came to the technological aspect. Respondents consisted of IBM Watson users, top management from IT, founders of AI and ML startups, as well as authors and creators of AES.

Secondly, 13 interviews were conducted, face-to-face, with individual professors at CBS to reflect upon their profession as a professor, and to hear their opinions of the suggested solution. These professors were chosen randomly, from a wide variety of departments at CBS. Additionally, as there is more behind the research question in regard to legal aspects, the Ministry of Higher Education was interviewed.

Table 6: Interviews

Interviewee	Relation to Education Sector or Technological Sector	Interview Date	Interaction
Senior lecturer CPHBusiness	Senior lecturer at CPHBusiness, used IBM Watson Software for three years (analyzed data from quantitative market research)	2017-10-25	Phone
CPHBusiness Students (5)	Working with IBM Watson	2017-11-01	Face-to-face
Co-founder of AI startup	AI expert	2017-11-09	Face-to-face
IBM Watson expert	Transformation Architect CTO team, IBM	2017-11-14	Face-to-face
Professor A	Professor at CBS	2017-12-21	Face-to-face
Professor B	Professor at CBS	2018-01-26	Face-to-face
Professor C	Professor at CBS	2018-01-30	Face-to-face
Professor D	Professor at CBS	2018-01-31	Face-to-face
Professor E	Professor at CBS	2018-02-08	Face-to-face
Professor F	Professor at CBS	2018-02-09	Face-to-face
Professor G	Professor at CBS	2018-02-22	Face-to-face
Ministry of Higher Education	Employees attended included: Consultants, analysts, and researchers	2018-02-22	Face-to-face
CBS Students (15)	Students	2018-03-01 - 2018-03-31	Face-to-face

Business Intelligence and Development	Budget, analyst, Business Intelligence at CBS	2018-03-02	Face-to-face
Professor H	Professor at CBS	2018-03-05	Face-to-face
Professor I	Professor at CBS	2018-03-05	Face-to-face
Mark Shermis	Author, AES Expert	2018-03-06	Skype
Peter Vitartas	AI Expert, University of LaTrobe, Australia	2018-03-06	Skype
Co-founder	Co-founder and CEO of Peergrade, Machine learning expert	2018-03-08	Face-to-face
Professor J	Professor at CBS	2018-03-09	Face-to-face
Professor K	Professor at CBS	2018-03-19	Face-to-face
Professor L	Professor at CBS	2018-03-20	Face-to-face
Professor M	Professor at CBS	2018-03-20	Face-to-face
Management	Vice Dean of Education	2018-04-20	Face-to-face

4.3.4. Questionnaire

When designing the student questionnaire, it was important that the questions were asked in an unbiased way. As Bradburn et al. (2004) states in *Asking Questions*, there is no standard way in which questions are asked when it comes to education for example. This is due to the way in which the answers are intimately tied to the purpose of the response. Besides asking background and open-ended questions, each question had a 1 to 5 response category, 1 being that the respondent is not satisfied, while 5 being that the respondent is very satisfied. The objectives of each question are explained below in Table 6, to understand and explain how the questions relate to the research question underlining the survey: how satisfied are students with the examination system at CBS?

Table 7: Questionnaire to students

Questions	Reasoning
Are you studying a bachelor or masters at CBS?	To determine if there is a correlation between those with less education experience in bachelors and those with more education experience in masters (Bradburn et al., 2004).
What year are you currently enrolled in?	To determine how much experience the student has with CBS's examination system (Bradburn et al., 2004).
What is the title of your program?	To see whether some programs have more forms of exams, resulting in different feedbacks (Bradburn et al., 2004).
Are you a Danish or International student?	To evaluate if there are discrepancies between respondents from Denmark or International (Bradburn et al., 2004).
What is your Grade Point Average currently at CBS, based on your grades from a 7-point	To analyze whether there is a correlation between students' GPA and their satisfaction

Danish scale? (if >0.5 round to the higher number, <0.5 round to the lower number)	with their grades and feedback received (Bradburn et al., 2004).
How satisfied were you with the examination system of oral exams in terms of fairness? Fairness meaning the effort you put in relation to the grade you received. The examination system meaning the way your exam was evaluated.	To determine the students' satisfaction rate based on the different types of exams. These results will allow the researcher to analyze these differences (Ramsden, 2003).
How many times have you received oral exam feedback?	To determine how accurate and weighted the previous question should be taken into consideration ((Ramsden, 2003).
How satisfied were you with the examination system of written sit-in exams in terms of fairness? Fairness meaning the effort you put in in relation to the grade you received. The examination system meaning the way your exam was evaluated.	To determine the students' satisfaction rate based on the different types of exams. These results will allow the researcher to analyze these differences (Ramsden, 2003).
How many times have you received written sit-in exam feedback?	To determine how accurate and weighted the previous question should be taken into consideration (Ramsden, 2003).
How satisfied were you with the examination system of take-home exams in terms of fairness? Fairness meaning the effort you put in in relation to the grade you received. The examination system meaning the way your exam was evaluated.	To determine the students' satisfaction rate based on the different types of exams. These results will allow the researcher to analyze these differences (Ramsden, 2003).
How many times have you received take-home exam feedback?	To determine how accurate and weighted the previous question should be taken into consideration (Ramsden, 2003).
How satisfied are you with the amount of feedback you have received thus far in your program?	To determine an overall satisfaction rate of the student in terms of feedback (Ramsden, 2003).
How important do you think receiving feedback is? For example, you get a 7 on an exam and receive feedback stating, "The theoretical readings could have been reflected further in depth in relation to your case study". Although you cannot retake that exam, how much will receiving feedback help you to improve on writing future exams?	To identify whether students feel the need to receive more feedback in the future (Ramsden, 2003).
Are there any experiences you had with CBS's examination system that you want to share with us?	To see students' emotions by letting them have the chance to elaborate on any experiences they once had regarding the examination system (Bradburn et al., 2004). To gain a better understanding of how students are feeling, by asking an open-ended question (Bradburn et al., 2004).

As shown above, the questionnaire consisted of 15 questions by which 5 questions were background questions, 8 questions were provided with scale responses, and concluded with 2 open-ended questions.

4.4.DATA ANALYSIS

In order to understand the data that was derived from the questionnaire, it was crucial to conduct a regression analysis. This is a form of statistical modeling whereby an analysis is done to identify whether there is a relationship between the dependent variable, that of what we wanted to test shown in Chapter 5, and the independent variables, that cannot be changed, known as the individual student's Grade Point Average (GPA).

Through this analysis, a statistical approach was required to view the strength of the relationship between student's GPA and their satisfaction rate with the examination system at CBS, known as the correlation. Additionally, it was important for us to determine if the results differed between Bachelor and Master student's, Danish and International student's, and the different types of exams (oral, written, sit in). Thus, a scatter plot was designed for each of the different variables that could have an influence on the results. Furthermore, a linear regression was conducted on the individual 12 graphs. This is a linear approach where the relationship between variable x and y can be displayed, in order for us to determine whether there is a direct relationship between GPA and satisfaction rate.

Besides the student questionnaire, reflecting back to the interviews with the main users of the suggested solution: professors, and students, the responses of this data collected was analyzed and conducted in a way that the teaching professor and student profiles in the VPD canvas was completed. The most important insights from all interviews were added to their respective customer profile graph. By outlining the pains, gains and jobs, the researchers can understand the features that should be incorporated into the suggested road map for CBS. The gain creators and pain relievers outline how the value proposition will be done and built around, to create value for all stakeholders. This value was co-created, with a SDL approach, by viewing the conceptual transitions of a GDL concept to a SDL concept, further explained within the following chapter.

4.5.1. Reliability and Validity

The quality of the research conducted was guaranteed by pursuing the concepts of validity and reliability. As Yin (2013) says, validity can be broken down into three parts, construct validity, internal validity and external validity.

Construct validity consists of laying down the right strategies to reach the research purposes and this was achieved by firstly defining the phenomena of examining at CBS and finding out operational measures that were based on already existing studies that could have been applied to the phenomena. For the latter, multiple sources of evidence coming from different authors on AES and AI were indeed used. This created the validity that the suggested option of using AES was backed up by already published studies, Shermis' publications, AES tools that have already been implemented. Concerning *internal validity*, which aims at explaining the relationships between causes and effects, this study investigated the situation at CBS by utilizing BID data that outlined how much the university is allocating on the examination activity in terms of time and money. Additionally, interviews with professors and students helped to understand their experiences with examining, as well as a student questionnaire that clarified how students perceived the feedback activity at CBS. The final one, *external validity* consists of making sure that the results of a study can then be applicable to other cases. This was met as the final suggested roadmap can potentially, given many factors, be applied to other universities that want to implement the servitization of the examination process, while incorporating feedback. Last but not least, in order to explain the reliability of this research it had to be proven that the procedures that were followed to produce this study could have been replicated in the future. This was done by disclosing the analytical approach that had been followed in order to develop the prototype idea and final suggested roadmap.

4.6. METHODOLOGICAL EVALUATION

4.6.1. Methodological Limitations

As the data collected (test scores and interviews) is considered valid and reliable, it is also worth noting that there are specific limitations of the method that was used. The observations and interpretations that are presented within this thesis will be different than that of, for example, an independent researcher coming to CBS. This could cause we, as students, to not be as objective as one could have wished, such as an outsider.

Looking more in depth, regarding the 13 interviews with professors, this can be seen as a smaller sample size, given a number of 100+ teaching professors at CBS. Therefore, the data analysis insights for these interviews should not carry much weight on the results as such a small sample size can result in a margin of error (Kvale, 2007). Nevertheless, several key takeaways were taken to complete the teaching professor profile.

Besides only conducting interviews with professors, relating back to the cognitive modelling approach by Russell and Norvig (2010) in Chapter 1, it is important to observe humans by getting inside of their minds. As we as researchers are students, to understand the teaching professors' world, an observation of the professor examining papers could have been done to observe their actions and the process that is taken for individual professors. These results could influence the final prototype in a sense of what professors specifically look for when examining, further discussed in Chapter 7.

During the 15 interviews with students, it was observed that the more interviews were conducted, students were continuously stating similar pains and gains, affirming mainly to limited feedback. Due to time constraints and availability of students, focus group interviews could have been conducted in order to assess respondents' attitudes and beliefs about the topic. This would enable us as researchers to gain further insights in the students' profile, as well as see the significance of their experiences (Kvale, 2007). Although a survey was conducted, the sample was based on those that responded via a social media link. This influenced the proportion of respondents being Master and Bachelor, as well as Danish and International, shown within the findings of Chapter 5.

5. FINDINGS

This chapter presents the data findings of the research, reflecting back to the theoretical framework. This provides an overview of how much CBS is spending on different teaching professor activities, the opinions of stakeholders, as well as a description of the already available AES tools in the market. Additionally, the limitations, benefits and drawbacks of implementing AES at CBS is looked upon.

5.1.PRIMARY DATA FINDINGS- STAKEHOLDERS

5.1.1.Survey Outcome

A total of 124 respondents was obtained for the questionnaire, previously shown in Table 7. Respondents were selected as a random sample size of students currently attending CBS. Out of the 124 respondents, 84% are Master students of which 53% are enrolled in first year and 47% in second year. In addition, 16% of respondents are Bachelor students of which 30% are enrolled in first year, 25% in second year, and 45% in third year. As there is a significant difference between Master and Bachelor students, this will be taken into consideration when analyzing the data and discussing the results. Considering the years that the students are enrolled in, there is a relatively even distribution of respondents. Additionally, 79% of respondents are international students, 18% of respondents are Danish students, and 3% of respondents preferred to be anonymous. To gain an accurate representation of the data set and to observe correlations, the 3% of anonymous results were disregarded (Bradburn et al., 2004). As previously explained in section 4.4, a scatter plot was conducted to determine whether there are any correlations between the different variables. The scatter plot results are shown in Appendix 5, whereby a summary of the plots is shown in the table below.

Table 8: Regression analysis data results

Legend

B	Bachelor
M	Master
I	International
D	Danish

Graph Title	Y	R2
BI Oral	0,8017x + 6,2828	0,2555 = 25.5%
BD Oral	1,5229x + 3,1835	0,6258 = 62.5%
BI Written	0,4819x + 6,9746	0,0572 = 5.7%
BD Written	1.6x + 3,4	0,5614 = 56.1%
BI Take Home	0,5787x + 6,3661	0,076 = 7.6%
BD Take Home	1,5556x + 2,4444	0,4298 = 43.0%
MI Oral	0,6647x + 7,3357	0,2779 = 27,8%
MD Oral	0,3786x + 7,3398	0,0617 = 6.2%
MI Written	0,2753x + 8,7013	0,0363 = 3.6%
MD Written	0,3786x + 7,3398	0,0617 = 6.1%
MI Take Home	0,2871x + 8,6003	0,0473 = 4.7%
MD Take Home	0,4474x + 6,8158	0,0333 = 3.3%

The results of the linear approach come in the simplest form of the regression equation by the formula y= c + b*x, where y= estimated dependent score, c= constant. This equation is the equation of the linear line that is shown in the middle of the graphs, Appendix 5.

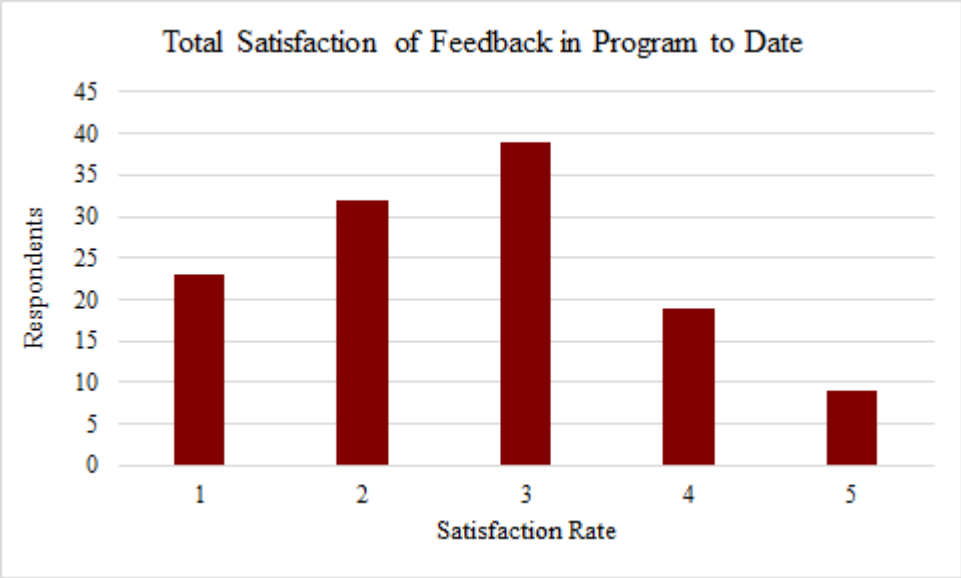
Going further, the R^2 has been calculated, which is the basis for our questionnaire results. This number is a measure to show how close the data is to the linear line. In conclusion, R^2 is always between 0 and 100%, whereby:

- 0% indicates that the model explains none of the variability of the response data (Regression Analysis, 2013).
    - The lower the R^2 is for this data set, the lower the correlation there is between GPA and students' satisfaction.
- 100% indicates that the model explains all the variability of the response data (Regression Analysis, 2013).
    - The higher the R^2 is for this data set, the higher the correlation there is between GPA and students' satisfaction.

Looking at the data set results with Master students, five R^2 results, MD Oral, MI Written, MD Written, MI Take Home, MD Take Home, are below 6.2%, and one, that of MI Oral, correlates 28%. This shows there is significantly little to no correlation between GPA and students satisfaction for those in Master programs. Additionally, based on the similarity of the R^2 results, there is no distinct difference between the results of Danish and International students.

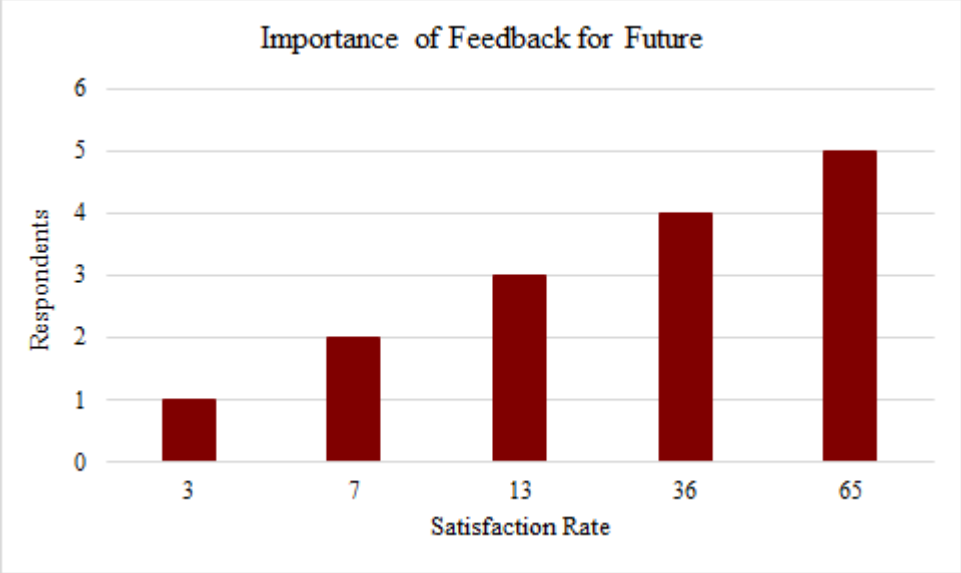
Considering the Bachelor student results, BD Oral, BD Written, and BD Take Home, all have R^2 results above 43%. This affirms that there is a higher correlation between GPA and students' satisfaction. In contrast, BI Oral, BI Written, and BI Take Home have R^2 results below 26%, affirming that there is little to no correlation between the variables. It is interesting to note that within the data set of Bachelor students, Danish students have a higher correlation between the variables than the international students. This could be made on an assumption that Danish students could be cultural bias in a sense that CBS is in their country and they are brought up with the Danish examination system. Additionally, no correlation between the programs were observed, as well as no correlation between how satisfied students were with the feedback they received in regard to how many exams of that type they had. Besides evaluating the regression analysis between GPA and the examination system satisfaction rate, the feedback satisfaction rate will be analyzed.

Figure 7: Total satisfaction of feedback in program to date



Assuming that satisfaction rate 1 and 2 show an unsatisfied rate, 45.1% of respondents feel unsatisfied with the amount of feedback they have received this far in their program and 32% of respondents are neutral or in the middle of unsatisfied and satisfied. In contract, 23% of student are satisfied. There are many factors to take in consideration when analyzing these results, such as the different programs at CBS, as well as the specific courses that respondents have taken. As a whole of all respondents, shown below in Figure 20, students feel that it is valuable to receive feedback in order to enhance their performance in the future for other courses. A total of 81.4% felt it was 4 and 5 on the satisfaction rate, with 2.4%, 5.6%, and 10.5% felt it was 1, 2 and 3, respectively.

Figure 8: Importance of feedback for future



In conclusion, looking at the data results from the questionnaire, students would like to receive more feedback as they feel it would be valuable for future courses. Besides analyzing the survey and to indicate that this was a major pain for students, the student and professor interview results will follow.

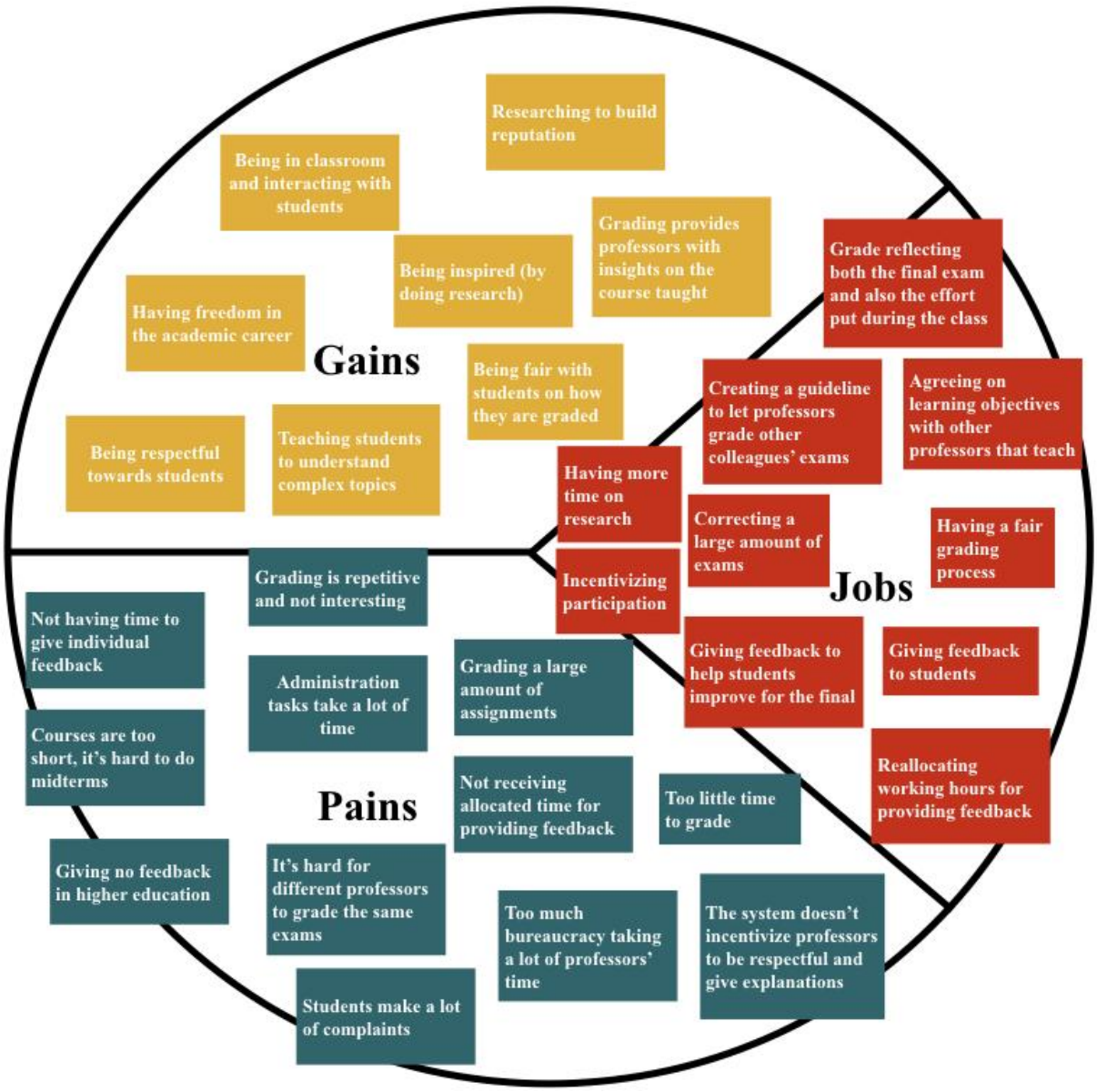
5.1.2. Interview Results

The insights from the interviews with teaching professors and students have then been analyzed and grouped by using the customer profile canvas (Osterwalder et al, 2014), as described in the research theoretical framework in Chapter 3.3. It has to be noted that only the insights that were mostly

recurrent among professors and students are represented. As shown in Figure 9 and Figure 10, there are two patterns that connect the pains, gains and jobs of students with the ones of teaching professors. These patterns are *time* and *feedback*. From the teaching professors' perspective, the lack of time and management is a major issue as it dictates their availability in being part of determined activities, such as teaching. From the professors' interviews the lack of time seems to be connected with the fact that there some activities that are repetitive and time-consuming such as administrative tasks and examining exams. This lack of time from professors has an echo on the students' side as having to wait for one month to receive a grade is considered a pain. The lack of time also triggers the pain of students of not receiving any feedback and explanations on their final performances (assessments), which ultimately results into the decrease of the quality of teaching. This pain is also reflected on the professors' profile as they feel they are not being respectful towards students by not providing students with feedback. Even though professors admit that, according to the university rules, they have to be open to provide feedback anytime a student asks for it. Professors state that, however, they are not given an allocated time for this activity and this leads to arbitrary decisions on whether, and to what extent, to provide feedback. Referring back to the literature review of this study and looking at the data that came out of the interviews with professors and students, it is evident that CBS is missing out an important part of the learning process. This is that of regarding feedback as a support of learning that explains the student the reasons for their mistakes and performance results (Evans, 2013) and creates the basis for increasing their knowledge depth in a specified subject, while improving future performances.

By looking closely at the teaching professors' profile (Figure 9), some additional considerations have to be made. A common point in the costumer job section was the one of "creating an exam format that allows other professors to grade the exam". This is due to 5 out of 13 professors saying that, in some cases, they have to either create a solution guide to the exam so that it can be graded by another professor (or external examiner) and to collaborate with other professors with whom they teach in the same course on how to grade the exam in the same way. It follows that, even though the examination activity of professors is very subjective, according to what the professor themselves say, there is anyway a precondition of having to design an exam that can be graded in an objective way as possible. In addition, "incentivizing in class participation" was also a shared customer job for 6 out of 13, where also some of them said they would like to give a percentage of the final grade based on participation.

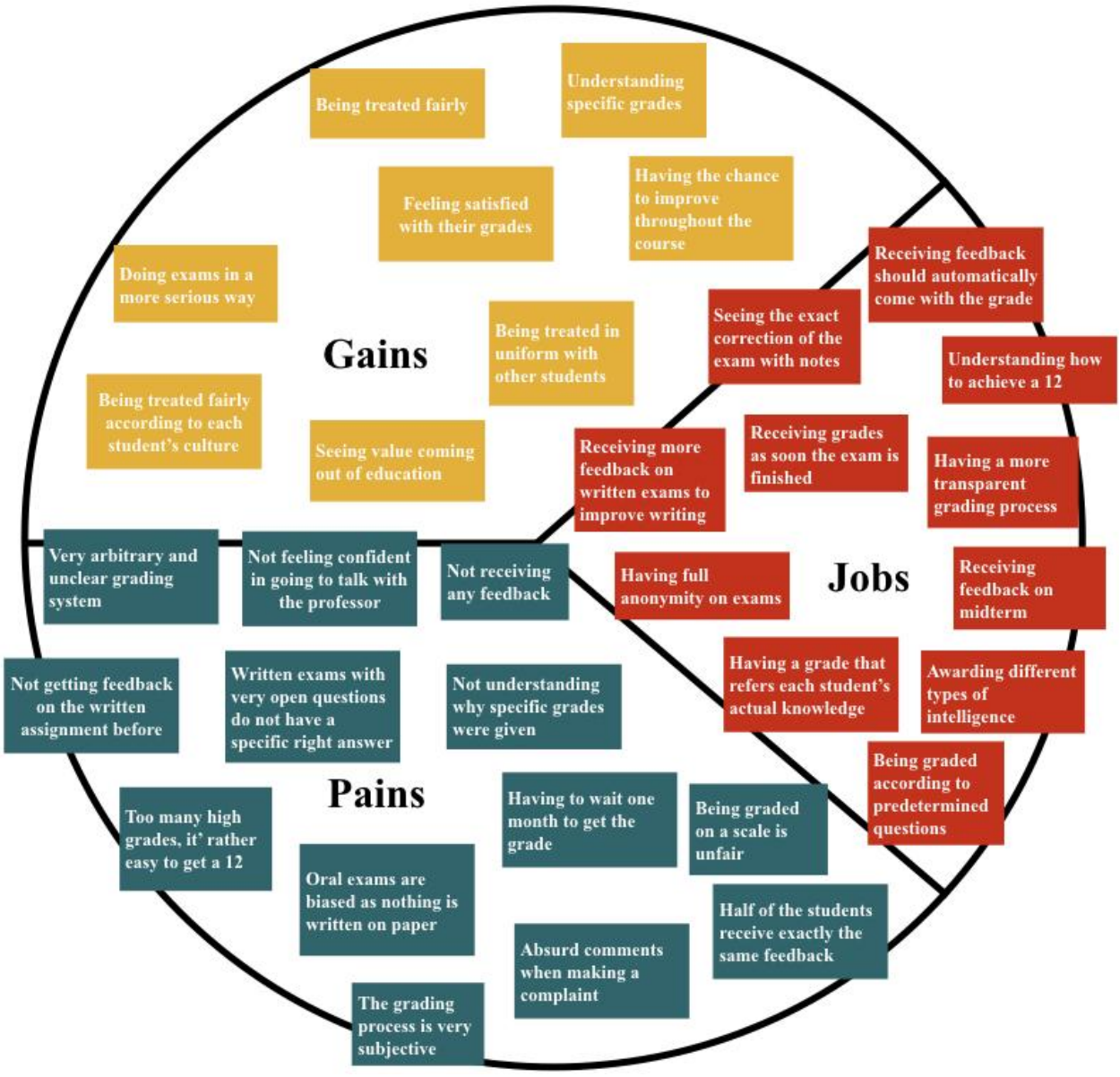
Figure 9: Teaching professor profile



Moving to the students' profile (Figure 10), besides receiving feedback, a shared request was having a more transparent examination process that is based on answers that are partially decided a-priori even for essay based exams, as standardized solutions are already being used for quantitative exams. This would mean that when a grade is received, it will be easier for a student to understand how different their answers are in comparison to what the teaching professor is looking for. Connected to this point there is the need of a standardization of the exam for the professors. Another relevant observation that was made, after analyzing the open answers that were provided through the student survey, was the desire of students of having an examination process that can

take into account the different cultural educational background of each student. For instance, according to a respondent, Danish students are very good at presentations and oral exams as they are being taught public speaking and argumentation techniques since the beginning of their education, while in other countries the knowledge depth of a student is the most important point in assessments.

Figure 10: Student profile



A common point that came out both from professors and students was the inconsistency of the disproportion of the 7- point Danish grading scale. From the professors' side, the scale does not allow them to award the excellence, which was previously possible with the grade 13, previously explained in Chapter 1. The scale also does not help the examination activity, as it is a pain for professors to decide between the grades 4-7 and 7-10 as there is a large gap that makes students with very different performance levels (for instance a very poor 7 and a very high 7) receiving the same grade. This is one of the main reasons why students make complaints as they feel their better performance is not awarded fairly compared to the others. Additionally, as one professor said referring to essay exams, this type of grading scale does not allow exceptions and students' extra effort or knowledge as, according to the description of each grades in the scale, professors have to look for how many mistakes the students did in order to decide which grade to give. This makes the examination process an activity that is more based on negatively scanning exams than looking for what is positive. Similarly, from the students' perspective, the scale divides the students among good and bad performing ones and the weight of a low grade has a larger negative impact than a high one to the final GPA. This is because there are 3 points among the grades 4-7 and 7-8 and only two between 10-12. Even though this is a highly shared pain, this point cannot be the focus of this study as the grading scale is determined by the law, but it was however relevant to bring to light.

Following the two most important points of this analysis, an investigation of the time and money that is allocated to professors and examiners at CBS was considered relevant in finding clear figures that would support what came out from students and professors' insights. At the same time, the capability of AI technologies in reproducing and automating human activities seemed a valid opportunity to help with decreasing the time spent on examining exams and finding a suitable way of providing every student tailored feedback. For this reason, a research on whether the market was already proving a solution to this, was conducted and the outcomes are presented in section 5.2.2.

5.2.SECONDARY DATA- BUSINESS CASE

5.2.1.Business Intelligence and Development

As previously mentioned, quantitative data was obtained from the BID department to gain a better understanding of how the budget at CBS is being spent throughout the different educational activities that professors encounter on their day to day profession. As the title of "professor" is quite broad, for this research, the two categories of professors, known as VIP and DVIP, will be explored in detail. These terms are used in educational settings in Denmark with the understanding that VIP refers to permanent staff and PhD's, in this case those that are employed by CBS, while DVIP refers to part-time professors such as external lecturers, teaching assistants and external sensors. Additionally, the education year for this data is from September to September.
The data shows that the three main educational activities for all categories of professors: education (referring to teaching), exams and guidance, and other activities. The hours that the professors spend are actual hours calculated from the teaching coverage in Prophix, a performance management software used at CBS.

For the purpose of this research while putting a focus on the examination system at CBS, analyzing the data within four years provides an overall representation to view any changes or inconsistencies and examine why. Shown in yellow, it is notable that the percentages within the exams and guidance category is quite steady throughout the four years.

Table 9: VIP/DVIP ratio of day studies divided by educational activities (represented in thousands of hours)

	2014		2015		2016		2017	
	Hours	Share %	Hours	Share %	Hours	Share %	Hours	Share %
Total number of hours (VIP and DVIP)	507,314	100%	504,159	100%	557,330	100%	496,458	100%
That of VIP	295,958	58%	277,362	55%	287,686	52%	251,327	50%
Education (teaching)	133,592	26%	114,916	23%	100,533	18%	99,092	20%
Exams and guidance	115,976	23%	111,323	22%	124,270	24%	99,899	20%
Combined & others	46,390	9%	51,123	10%	52,882	10%	52,336	10%
That of DVIP	211.356	42%	226.797	45%	269.644	48%	245.131	50%
Education (teaching)	81,901	16%	86,272	17%	91,477	16%	92,771	19%
Exams and guidance	123,460	25%	132,797	26%	170,019	31%	144,515	29%
Combined & others	5,995	1%	7,728	2%	8,149	1%	7,845	2%

Source: Business Intelligence and Development CBS, 2018

One fluctuation can be seen in year 2016, with an increase of 2% for VIP and 5% for DVIP, a total of 50,172 thousand more hours spent than in 2015 for exams and guidance. This is due to the Danish Government's study progress reform for Danish universities where the government sought out those students who were not completing their thesis and extending their student status into the school system. Therefore, these students were required to submit their thesis within a given period, resulting to a higher submission rate, increasing the number of hours and share percentage that professors had to spend (BID 2018, personal communication, 2 March). As illustrated in Table 10, this correlates to how much money was spent on thesis' alone in these four years according to the category of the professor.

Table 10: Expenditure on thesis' by category of professors (million DKK)

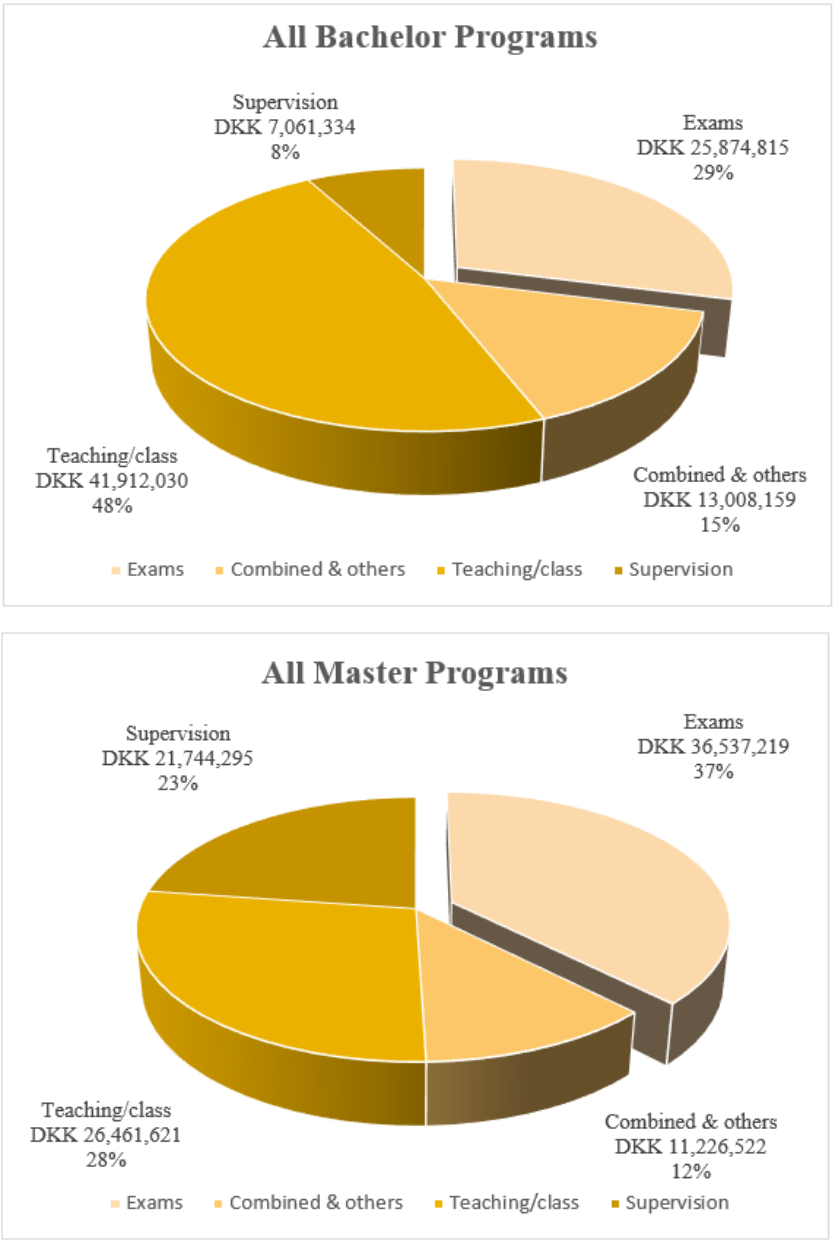
	2014	2015	2016	2017
Total	**28**	**27,9**	**57,1**	**38**
External Censor	5,8	5,9	13,4	10
External Lecturer	5,3	5,8	13,4	9,8
Full Time Lecturer	14,7	13,7	25,9	15,8
Department Managers	0,1	0,1	0,2	0,1
PHD	0,5	0,4	0,5	0,3
Teaching Assistants	1,6	2	3,8	1,9

Source: Business Intelligence and Development CBS, 2018

As 2016 was a distinctive year, to further understand how much is being spent on the exam process (that being from the time students submit their exam until the grades are distributed to the students), at CBS, January to December, known as fiscal year 2017 will be analyzed. This data was developed more in depth and divided the three main educational activities into four: supervision, exams, teaching, and combined and others.

It is important to note that the below tables relate to specific study activities and all general CBS activities are not included. These other activities are called overhead, being anything from i.e. administration, rent and research (ibid.).

Figure 11: Salary spent on education in fiscal year 2017 for both all Bachelor Programs and Master Programs



Source: Business Intelligence and Development CBS, 2018

As shown, professors, in all bachelor programs and master programs in 2017, spent 29% and 37%, respectively, of the resources examining exams.

Additionally, given the context of CBS and the vast number of programs that are being taught, it is important to note how much CBS receives in grants for education from the Ministry and how much

of this is actually being spent within the specific courses. This data provides an overview to analyze how different each course is and how much money can be budgeted for specific activities, such as exams and research, shown with further details in Appendix 4.

Table 11: Grants given by ministry 2017

Study Program	Grant 2017 (Million DKK)	Total Consumption 2017 (Million DKK)	Of which teaching (Million DKK)	Of which operation (e.g. printing, transport, etc) (Million DKK)	Difference than grant provided	STÅ 2017	Costs per STÅ
Total of all studies	**207,4**	**189,5**	**184,2**	**5,3**	**17,959,954**	**12,567**	**16,507**
Total Bachelor	100,9	91,5	88,1	3,5	9,362,535	6,244	16,161
Total Masters	106,5	97,9	96,1	1,9	8,597,419	6,323	16,849

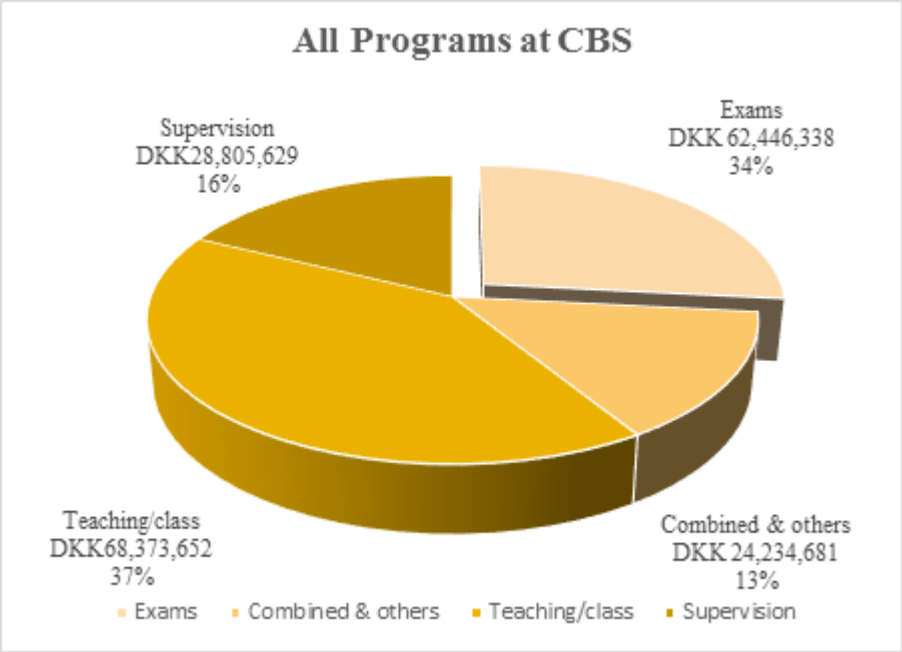
Source: Business Intelligence and Development CBS, 2018

It is important to understand the grant system and how the educational sector works in Denmark. The university's grant from the Ministry is provided dependent on the amount of passed exams submitted every year and converted into student year work, known as studenterårsværk (STÅ) in Danish. A student's year workload, that of 60 ECTS, is equivalent to 1 stå (Ufm.dk, 2018). Each stå at CBS receives a grant of approximately DKK 45,000, but ultimately depending on the program (BID 2018, personal communication, 2 March). Other institutions receive different amounts depending on the rate of the particular education. There are three tariff levels that typically receive the highest grant which are subjects within natural science, technical science and health science. This is due to the large amounts of resources that institutions need in order to complete these courses.

All money at CBS that is collected from the Ministry, gets split between teaching the specific program and the administration for that program, approximately a 50/50 split (ibid.).

Besides the resources that professors spent examining, a total of all programs can be analyzed. Looking at the total consumption, given in Table 10, that was spent on all professor's activities in year 2017, DKK 184,153,023, it is important to see the costs of all activities for all programs at CBS, shown in the below figure.

Figure 12: Total spending DKK at CBS on all activities



Analyzing the data extracted above, it is shown that CBS spent **DKK 62,446,338** on exams in 2017, equaling to **34%** of their expenditures.

5.2.2. Already Available Tools in the Market

The research on the already available tools in the market yielded the discovery of the existence of AES tools that are nowadays sold by different firms that operate in the educational services market, in particular: Measurement Incorporated, ETS, Pearson and Vantage Learning. In this section, the solutions that are already sold and that can be applied to different contexts are presented. It is important to note, however, that there are also other institutions and companies that are working within the field of AI and education providing tailored solutions to different aspects of education. One of these, worth mentioning, is IBM Watson which is a range of AI solutions that IBM consults to apply and offers to different industries such as advertising, customer engagement, financial

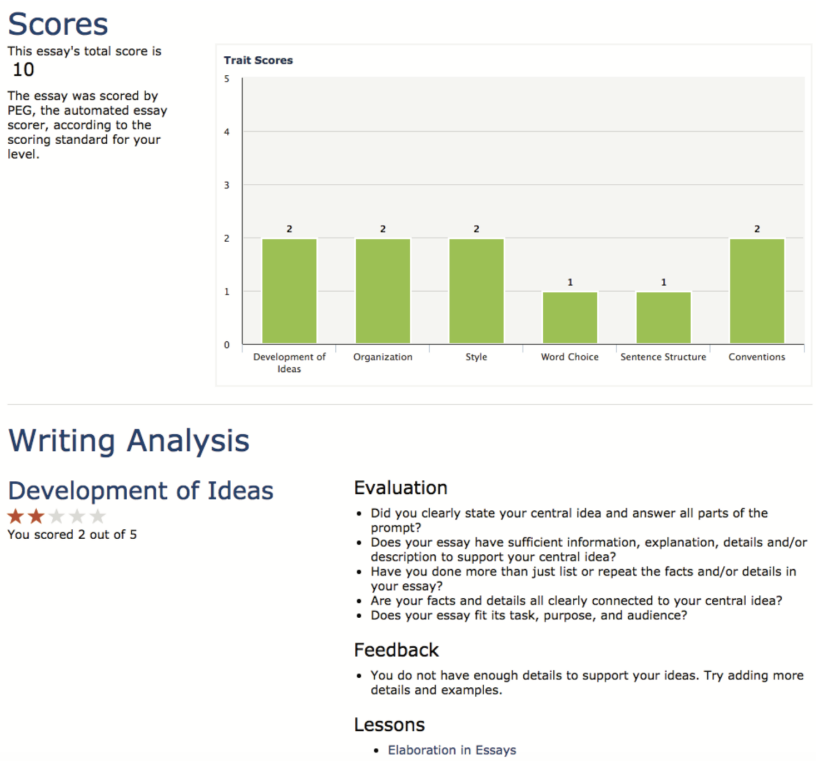
services, health, Internet of Things (IoT), media, talent, work and education (IBM.com, 2018a). In the field of education, according to Knight (2016) IBM is making deals with the previously cited Pearson. At the moment, it has two solutions for educators: IBM Watson Element for Educators and IBM Watson Enlight for Educators. The first one is a platform where educators can get insights from their students, track academic progress and gives motivation to educators on creating meaningful interactions with students (IBM.com, 2018b). The second one is a planning tool for creating tailored activities that are based on each students' strengths and areas of growth (IBM.com, 2018c). After having considered the broader market of AI solutions applied to education, a more detailed analysis of the AES tools will follow, that tackle the examination process of assessments.

5.2.2.1. PEG Writing

PEGwriting (Project Essay Grade) is a web-based learning and AES platform for *formative writing* for students in grade 3-12. It enhances student's writing skills by providing guided support and immediate feedback to students during writing exercises (Pegwriting.com, 2018). PEGwriting is an electronic portfolio that employs PEG ® AES engine that was initially developed by Ellis Page, the inventor of AES, and further sold by Measurement Incorporated.

PEGwriting offers four types of feedback to students: *targeted feedback*, *holistic feedback*, *professor's feedback*, *peer review feedback*. *Targeted feedback* is displayed on the writing platform when the student is writing the essay and it entails spelling and grammar error messages that contain an instructional explanation. *Holistic feedback*, is communicated only after the engine has read the entire paper and then displays suggestions to attributes of the essay (i.e., detail specificity and student's word choice) (Shermis et al, 2016). Regarding professor's feedback, professors, having access to an overview of all students' performances, can decide to send electronic sticky notes and instant messages to the students that need help the most (Measurement Incorporated, 2017a). *Peer review*, allows students to write constructive feedback to their peers and learn from each other by using learning guidelines and assessment criteria when assessing other peers' papers (Measurement Incorporated, 2017b). After receiving feedback, students can make improvements to their paper and upload it to receive a new score. This way, students are motivated to make staged improvements by trial and error (Measurement Incorporated, 2017b).

Figure 13: PEG Writing example



Source: Measurement Incorporated (2018c)

When selecting the type of assessment, professors can access and use existing writing prompts that are already stored within PEG and also create their own custom prompts based on other stimulus material like websites, passages and videos (Measurement Incorporated, 2016). The automated scoring engine PEG ® scores essays by using the 6+1 Trait® writing model which evaluates the following traits of each student's essay quality: ideas, organization voice, word choice, sentence fluency, conventions and presentation. Additionally, by using a three-point scoring rubric in PEGwriting, professors can add two other scores to assess the content of the paper: text evidence and content accuracy. These scores provide guidance to students with combined professor feedback that highlight the topic areas they should expand on more in the paper.

PEG ® is an AI engine that applies concepts of NLP, semantic and syntactic analysis and classification methods. The scoring engine is trained on a set of already human-scored essays that provide the basis on how to score future exams. In order to provide an accurate prediction, a statistical and linguistic model are built by analyzing the already human-scored essays through the

use of dictionaries, word lists and more than 500 variables including fluency, diction, grammar and construction (Measurement Incorporated, 2017b).

PEG® is able to identify implicit points that a human examiner looks at when assessing an essay by using a diverse set of features that can be divided as: *explicit, similarity-based* and *implicit features.*

Explicit features are those features that are clearly noticeable in a paper (i.e., grammar errors) and are evaluated through a rule-based statistical method, while *similarity-based* features are the ones that allow the AES engine to evaluate how a paper is similar to the essays graded by human experts in the training set. *Implicit features* influence the scoring of the quality of an essay. An implicit feature can be a specific topic expansion within a paper that is calculated through a Latent Dirichlet Allocation (LDA) (Shermis et al, 2016). LDA is a probabilistic model that is able to understand the different topics in a paper by analyzing groups of words belonging to the same topic. For instance, the words "school", "students" and "professors" all belong to the same topic of education. The more words presented that are related to education, the higher is the percentage of that topic (Blei, Ng, Jordan, 2003). As a result, PEG ® is able to provide a score that is similar to the one that would be given by a human expert by applying ML algorithms on these three types of features.

5.2.2.2. Criterion

Criterion is a web-based electronic portfolio developed by Educationa Testing Service (ETS), that through its AES engine e-rater, helps students in the production and revision of written essays and professors in saving time on examining and focusing on higher level writing skills of students.
The use and application of this tool is based on the CCSS which is a chart of high-quality academic standards used in the US. CCSS outline the learning goals that each student should achieve through their K-12 education before going to college. These standards are evaluated by the software that analyzes the organization and development of an essay by considering the presence (or absence) of `relevant discourse units`. These include an introduction, thesis statement, main ideas, supporting details, and conclusion (Shermis & Burstein, 2013).

Besides the organization and development, as described on corestandard.org (2018) the standards are based on the following: 1) research- and evidence-based, 2) clear, understandable, and

consistent, 3) aligned with college and career expectations, 4) based on rigorous content and application of knowledge through higher-order thinking skills, 5) built upon the strengths and lessons of current state standards, 6) informed by the other top performing countries in order to prepare all students for success in our global economy and society.

The AES engine e-rater allows Criterion to produce both holistic scores and real-time diagnostic trait feedback on 5 traits: 1) grammar, where the engine checks for grammatical mistakes and sentence construction, 2) usage, that evaluates the correct use of words in a text, 3) mechanics, that checks the correct use of spelling and punctuation, 4) style, which analyzes the use of specific words and the complexity structure of sentences, 5) organization and development that looks at the way the paper is structured and the main ideas and thesis that are supported (Vitarta, 2017). In addition to holistic scores and real-time diagnostic trait feedback, Criterion offers a peer review tool where students exchange feedback among each other and includes dialog boxes that allows interaction between professors and students (Ets.org, 2018a).

With Criterion, professors can use already existing prompts available on the platform or can chose to develop customized ones. The already existing prompts are built on a prompt-specific model and can hence receive both holistic scores and real-time diagnostic trait feedback. When using new prompts, professors can decide whether to use a *generic* model or build their own *prompt-specific* model. With the *generic* model, professors do not have to spend time on training a new model as they can use the same one that is used for the already available prompts in the platform. The weakness of relying on a generic model is that students will only be able to receive holistic scores and not tailored feedback. However, a professor can build a *prompt-specific* model for new prompts by using custom rubrics and creating a new scoring model (Ets.org, 2018b).

E-rater uses human-assigned holistic scores to build the scoring model. In order to create the prompt-specific model, developers have to collect a randomly selected set of 250-300 human scored essays and apply NLP techniques to process the text on different language features (i.e., style weaknesses, essay-based discourse elements, grammatical errors etc.). These features are then converted into a vector matrix and scanned through a regression modeling approach that determines the proper weight for each feature and is able to then compute the final score prediction model (Shermis et al, 2016).

5.2.2.3. Open Mark

Open Mark is a web-based and Computer-Assisted Assessment (CAA) system that is used for formative assessments and it has been developed for distance university students by Open University. The system has an interactive design: it gives hints to students throughout the exercise in order to guide them to reach the correct answer. For this reason, Open Mark gives students the possibility to take several attempts in the case the first answer was wrong (The Open University, 2018a). It provides individualized targeted feedback that is shown on the platform immediately after an answer to a prompt is submitted (The Open University, 2018b).

OpenMark is able to assess numeric responses, text responses, multiple choice exercises and 2D (graphical content) responses. The efficacy of the feedback system depends on the quality of the questions that professors decide to ask in an assessment. This type of tool suggests professors to use Bloom's taxonomy to formulate the questions. The questions are required to have four elements. First, the question itself that has to clearly state the problem through different formats (i.e., images, text, numbers etc.). Second, the professor has to list a set of predicted responses against which students' responses will be scored. Third, the professor should also create specific feedback, in advance, that aims at correcting predicted wrong responses and, lastly, a full explanation that will be provided at the end of all of the attempts (The Open University, 2018c). In addition, professors are expected to create questions whose responses and mistakes are good to predict, and design a feedback system that suggests students to refer to determined study materials (The Open University, 2018d).

OpenMark is able to predict whether the content of an answer to a question is correct. However, in contrast to PEGwriting and Criterion, OpenMark is designed for short responses and is not used to assess essays and evaluate the quality of each student's writing skills. The following image shows two examples of questions that can be asked through OpenMark: the one to the left requires a short qualitative answer and the one to the right requires a quantitative answer.

Figure 14: Open Mark example

The photograph shows a selection of igneous rocks. How are igneous rocks formed?

*Please give your answer as a **short** phrase or sentence.*

Check

A catalyst speeds up a reaction. What other single change will make each of these reactions go faster?

A $Cl_2(g) + H_2(g) \rightarrow 2HCl(g)$ $\Delta H = -184$ kJ

B $CH_4(g) + NH_3(g) \rightarrow HCN(g) + 3H_2(g)$ $\Delta H = +256$ kJ

C $HClO(aq) \rightarrow H^+(aq) + ClO^-(aq)$ $\Delta H = +13.8$ kJ

D $N_2(g) + 3H_2(g) \rightarrow 2NH_3(g)$ $\Delta H = -93$ kJ

E $S_2O_8^{2-}(aq) + 2I^-(aq) \rightarrow I_2(aq) + 2SO_4^{2-}(aq)$ $\Delta H = -341$ kJ

F $2SO_2(g) + O_2(g) \rightarrow 2SO_3(g)$ $\Delta H = -198$ kJ

G $2H_2(g) + O_2(g) \rightarrow 2H_2O(g)$ $\Delta H = -482$ kJ

Enter answer

Source: The Open University (2018e), The Open University (2018f)

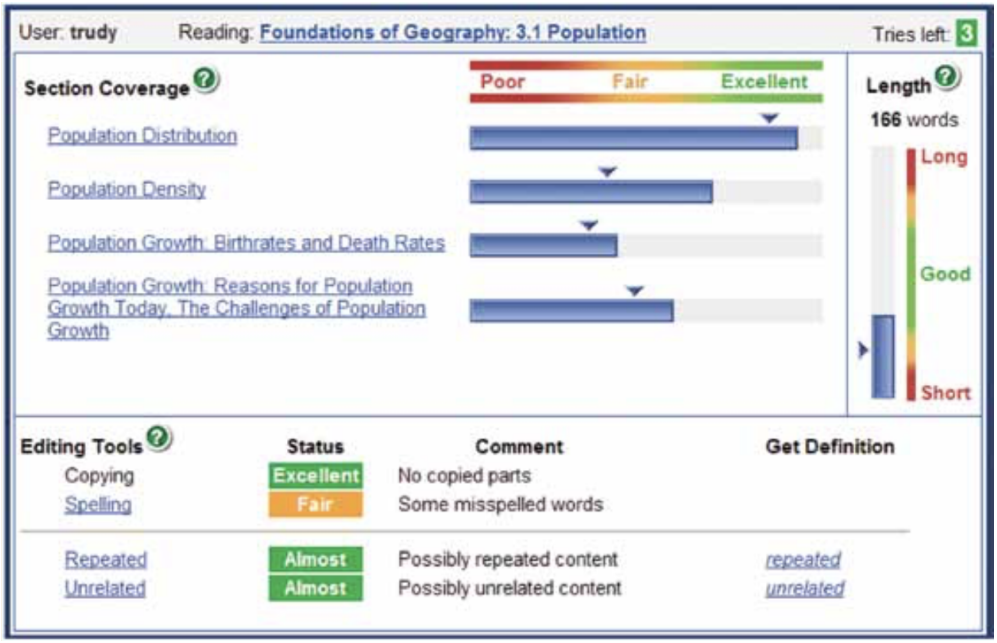
5.2.2.4. WriteToLearn

WriteToLearn is a web-based electronic portfolio and AES engine developed by Pearson that follows the US national CCSS guidelines. It helps both students in the comprehension of readings through the exercise of writing summaries and in the learning of writing skills through writing response exercises to prompts. Students receive immediate holistic scores on a 4- or 6- point scale and specific trait score feedback. Through the score of each trait, students get explanations on how to improve their writing and they can also get examples of good and poor writing (Shermis et al, 2016).

WriteToLearn is made up by two distinct components: Summary Street and Intelligent Essay Assessor (IEA). IEA is the AES engine whose component evaluates essays under six trait scores: ideas, organization, conventions, sentence fluency, word choice and voice. Summary Street is the web-based electronic portfolio of WriteToLearn that gives feedback on content and writing style. This component uses the Knowledge Analysis Technologies (KAT) engine, which is able to analyze the content coverage of each essay section. KAT uses Latent Semantic Analysis (LSA)

techniques, which measure the semantic similarity of words, and is able to evaluate whether the right information is communicated within the text (Pearson, 2007a). The following picture shows how Summary Street rates content coverage. Per each topic, it gives a rate that goes from poor to excellent which suggests which topics the student should expand more on.

Figure 15: WriteToLearn example



Source: Pearson, 2007b

5.2.2.5. My Access

My Access is an electronic portfolio and AES platform developed and sold by Vantage Learning. Its functionality is based on the CCSS, the State of Texas Assessments of Academic Readiness (Texas) and Standards of Learning (SOL) (Virginia). With MyAccess, students enrolled in grade four through higher education can exercise their writing skills and receive immediate, detailed feedback and have the possibility to communicate with the professor directly on the platform. More specifically, thanks to the component MYTutor, students get feedback on how to revise their paper and through MyEditor they are assisted in correcting writing errors (Vantage Learning, 2018). Students are also provided with scoring rubrics to evaluate the quality of their paper on their own and a writer's checklist that they can use to organize their writing process (Shermis et al, 2016).

My Access uses the AES engine IntelliMetric to assess students' essays on more than 1,500 prompts already present on the platform, covering topics within math, science, language arts and social studies. Professors also have the possibility to create their own prompts that include narrative, persuasive, informative, literary and expository genres.

IntelliMetric produces both a holistic and an analytical score on the following traits: focus and meaning, content and development, organization, language and use and style. In order to perfectly mimic the exact same performance of human scorers, it uses six virtual judges that are mathematical models that look for different features in a paper in order to predict the final score (Shermis et al, 2016).

5.2.3. Overview of Already Available Tools

By comparing all of the previously described tools, a synthesis table is further provided. In general, these tools all target formative writing assessments, apart from Open Mark, which is more for exercise assessments that are based on short answers. It can be noticed that PEG Writing, Criterion, Write To Learn and My Access all have the capability of assessing the style of writing by looking at different features. In the table under the column "evaluates", all of the features that each software uses to evaluate the assessments are listed. The features in black evaluates the writing style, while the ones in red evaluates the content of the paper. Even though PEG Writing, Write to Learn and MyAccess all have one or two features that address content, it can be stated that PEG Writing is the best one at assessing content as it does not only scan for content accuracy but also for text evidence, which means that it is able to analyze the points that are made by the reader. From the feedback feature point of view, both PEG Writing, Criterion and MyAccess have a good range of feedback types that can aid the student during and after the writing process.

Table 12: Overview of available tools in the market

Type of Tool	Producer	Components	Use Type	Types of Feedback	Evaluation Criteria	Standards Used
PEG Writing	Measurement Incorporated	Electronic portfolio & PEG Automated Scoring Engine	Formative Writing	Targeted feedback, holistic feedback, teacher's feedback, peer review feedback	**Ideas, organization voice, word choice, sentence fluency, conventions and presentation** & 3-point scoring rubrics that can assess content by evaluating **text evidence** and **content accuracy**	Common Core Standards
Criterion	ETS	Electronic portfolio & E-rater Automated Scoring Engine	Formative Writing	Holistic scores, real-time diagnostic trait feedback, peer review, teacher's feedback	**Grammar, usage, mechanics, style, organization, and development**	Common Core State Standards
Open Mark	Open University	Computer-assisted assessment	Formative short answer assessments	Instantaneous hints, individualized targeted feedback	**Numeric responses, text responses, multiple choice exercises and 2D graphical responses**	Bloom's Taxonomy
Write To Learn	Pearson	Electronic portfolio, Summary Street & IEA Automated Scoring Engine	Formative Writing	Immediate Holistic Scores, specific trait scores	**Ideas, organization, conversations, sentence fluency, word choice and voice, content coverage**	Common Core State Standards
My Access	Vantage Learning	Electronic Portfolio & Intellimetric AES	Formative Writing	Holistic feedback, analytical score, teachers' feedback, self-assessment rubrics	**Focus and meaning, content and development, organization, language, use and style**	Common Core State Standards, State of Texas Assessments of Academic Readiness, Standards of Learning (Virginia)

As several already available tools are on the market, it is important to analyze what features of these tools have potential within the context at CBS. As 34% of their expenditures are spent on the examination activity, as well as insights from professors and students outlining their pains and gains, it can be seen that AI technology, AES engines, has the potential to be implemented at CBS, which will further be reflected upon within the following section.

6. DISCUSSION

6.1. General Considerations and Solution

Physicist Stephen Hawking, engineer and inventor Elon Musk, and philosopher Nick Bostrom disclosed their concern on AI exceeding and taking over human intelligence (Yonck, 2017). The digital world is coming closer and closer to humans, a new era where technology is replicating human thought processes, which previous machines were incapable of. These beliefs are, however, still unproven in contrast to the challenges of developing such machines.

As the world today is seen to be in the *third loop of AI* from decades ago, the educational sector is rapidly growing when it comes to implementing forms of AI technology to enhance the learning experience for students, as well as the teaching experience for professors (Yonck, 2017). Reflecting back to Gartner and the waves in AI, it can be seen that we are in the "*algorithm economy*". Within this economy, innovation is key to the success of educational institutions for today and for the future. If the definition of AI is looked upon in Bellmans's (1978) perspective, where thinking humanly is key, "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning...)", AI can turn into the servitization of the examination system.

Within the case study of CBS, it has been shown that the examination system is not viewed favorably by the stakeholders- professors, students, and management. Not only is this based on interview responses, but as well as the findings from BID that outlines there are extensive amounts of resources, money and hours, being spent. When reflecting back to Penrose, there is an opportunity for reallocating these resources to provide all stakeholders with an increased quality of education. By doing so, it is suggested that CBS implement an AES system.

The first step will consist of implementing automatic examining for formative assessments that will provide feedback for students when preparing for the final exam and testing their knowledge. The second step will be implementing automatic examining both for formative and summative assessments which will at the same time, help provide students feedback (both during the class and

after the exam) as well as help the professor in speeding up the examination process of written hand ins.

This shift of resources does not mean that this tool will be a replacement for professors, but rather an *aid* to professors. As AI technology is still underway, there will still be the need of a "human in the loop", discussed previously, whereby the professor and the software collaborate together in the result of the examination and grade- leading to a more efficient and effective process with higher value added.

By creating this added value, examining would be in the approach of a SDL world. This value can reflect back to the nine fundamental units of exchange that Vargo and Lusch (2004) outline, stated within Chapter 3.1.3. Three of these units will be looked upon further. It is stated that FP3: "Goods are distribution mechanisms for service provision", which is seen in the sense that the direct services that professors are involved with, can be replaced by computer and application softwares (Vargo and Lusch, 2004). When looking at FP6: "The customer is always a co-creator of value", the suggestion of the servitization of examining can be directly linked to the research and findings within this thesis paper. When creating such a service, the stakeholders: professors, students, and management, all have an opinion that was involved in the production of this value- outlined in the VPD canvas, Figure 9 and 10, and was continually iterated to come up with a final suggested road map. Finally, considering FP9: "Organizations exist to integrate and transform micro-specialized competences into complex services that are demanded in the marketplace" can be related back to Penrose (1959), whereby the value creation is a process of integrating and transforming services, requiring interaction between customers and networking, which will be analyzed more in depth within the suggested roadmap for CBS in Chapter 7.

With the suggested AES approach, it is important to break the approach down into two parts. As the empirical findings outlined: (1) Students crucial pain was receiving little to no feedback, while (2) Professors are spending too much time on the examination activity which is being taken away from other activities- teaching and researching. The suggested AES approach will hence consist of a two-step plan that will gradually aim at covering both the aspects of feedback and examining time.

When looking at the first empirical finding, learning for students can differentiate within the quality of learning and understanding their own difficulties, creating a balance between structure and freedom (Ramsden, 2003). This balance must be related to the method of feedback in which an educational institution invests. As Ramsden (2003, p.76) states, "if the university is to remain an education center and not become just a degree machine assessing the 'pass-fail' of students, the usefulness of feedback cannot be ignored". After speaking with the Vice Dean of Education of CBS, within the past 2-3 years, there have been and are several initiatives being introduced to improve feedback, focusing on the needs within specific departments. These initiatives can be anything from introducing logbooks, creating online discussions, using quizzes, as well as utilizing rubrics for students to gain standardized feedback on specific issues with their assignments (Vice Dean of Education 2018, personal communication, 20 April).

As discussed with the Vice Dean of Education, in addition to individual student interviews, it is important that feedback be also incorporated throughout the entire course. By implementing an AES system for formative assessments, this will enable students to reflect upon their work in relation to their learning and utilize this feedback to improve future work, besides thoughts, effectively. According to the Danish law, programs are allowed to have a certain share of exams (also some that have only that assessment of the students throughout the course) that are pass-fail, along with a grade within the 7-step scale for the final assessment (Ministry of Higher Education 2018, personal communication, 22 February). Incorporating standardized response assignments in the middle of the semester could be integrated in a way that is pass-fail, and mandatory to complete in order to attend the final exam.

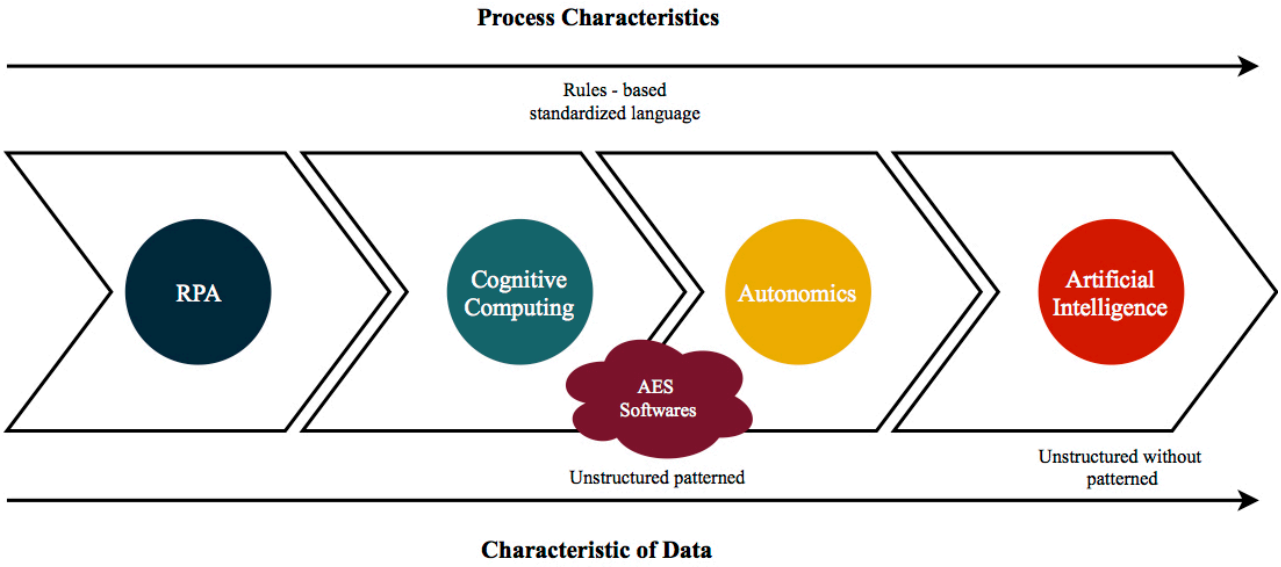
Touching upon the learning context of students and reflecting upon the added value that professors can bring to the classroom, thereby in correlation to the quality, more resources are required for professors to put a greater focus on teaching and researching. If such an AES software is implemented for summative assessments, resources could be reallocated from the examination activity of approximately 34%, to other areas as an integral part of teaching, in the practice to improve teaching and enhance the researching value at CBS. This shifting of resources can be seen in a SDL world, whereby the software itself can be viewed as operant resources. These resources are observed as primary whereby they enable humans to increase the value by utilizing their skills and knowledge on different types of resources (Vargo and Lusch, 2004). By relocating resources

from the salary of examining exams to teaching and researching, core competencies can be identified and developed for CBS. These core competencies will be outlined by the benefits, drawbacks and limitations that such a software will enable.

6.2. THE AUTOMATION LEVEL OF AES

The implications of implementing AES are many. It is important to start off by explaining where this new type of automation technology lies in the previously cited IAC framework developed by HfS Research (Willcocks and Lacity, 2016). According to the framework, the position of the technology on the continuum is determined by the process and data/information characteristics, AES capabilities have been analyzed against these two variables. Analyzing the process characteristics of AES tools, it should be noted that they rely on a process that follows *rules-based standardized language* as they grade new essays by comparing them to the training set of around 300 already evaluated past essays (Shermis, 2010). Regarding the type of data that AES scans for, it can be said that they rely on unstructured patterned data. This is because AES tools grade essays by scanning for already predefined patterns of words to provide a score and feedback. As a result of this analysis, it comes out that AES lies in the middle of CC and the *Autonomics* automation level. This is because AES, as *CC* uses self-learning, NLP and data mining to examine assessments but after the software is trained on a set of past exams, human intervention is not or rarely required as in *Autonomics.* According to the IAC framework, it can be hence inferred that AES not yet at the full *AI* level, as it is not yet able to entirely analyze dynamic and unstructured language. In order to have a better picture of this, the graph below represents a synthesis of the IAC and plots where AES is positioned.

Figure 16: AES in the Intelligent Automation Continuum framework



With regard to the automation level of AES, a discussion on whether the previously analyzed already available tools in the market that can be directly implemented at the higher education level, and more specifically, at CBS, has to be made.

6.3. ALREADY AVAILABLE TOOLS' LIMITATIONS

Referring back to the different types of exams that are used at CBS, AES can be applied to written exams. These tools are, however, at the moment limited in analyzing the content, testing strategic and analytical thinking of the student. This is because every student has a completely different way of thinking and expressing themselves. For instance, according to both a professor interviewed and Shermis (2013), a student might use an ironic way of commenting facts and theories or, have a very strong opinion against a specific author or framework that makes them write about it in a negative way. In light of how these AES algorithms are trained or limited access to training essays, it may result that these tools cannot take into consideration all points of view and have a "restricted mind" that compares the students' performances with only the ones of the students that wrote the essays that have been used for the training set.

Higher education exams are different than the one in K-12 education where AES is mostly used. K-12 education is a preparation to the higher education life where students are expected to be more independent and have learnt the basis of the different general knowledge subjects. AES tools are, indeed, primarily built, as the inventor Ellis Page (1966) said, with the purpose of scanning the writing abilities of students as this was something that not all professors were always evaluating in subjects different than English writing. While K-12 students are expected to be able to synthesize theories and analyze facts (ibid.), higher education demands for additional abilities. For instance, referring to the case study, the Danish education system awards excellence when the students meet a list of skills called "The great learning objectives". According to this list the students have to be able to:

- "Account for selected theories,
- discuss the strengths and weaknesses in those theories,
- apply the correct theory on a given issue,
- present argumentation that supports a given action oriented conclusion on a given case problem, and
- reflect on the consequences of applying theories on a given issue" (Brøchner Nielsen, 2013, p.3).

This considered, assessments at the higher education level take for granted that a student knows how to properly write a written composition, and look more for skills in using the theories and coming up with brand-new insights that are not yet written in scientific papers and books. For this reason, the already existing tools in the market seem to lack, at the moment, the ability of scanning for higher level skills. Notwithstanding this aspect, a distinction of the limitations and capabilities of these tools can be done by also looking at the different types of subject courses. With regard to the subject material, overall, quantitative courses, such as Microeconomics, Statistics, Econometrics, Math, (etc.), could be easier to examine automatically. Based on assumptions, this is because they have standardized solutions, that, from what emerged from professors' interviews, even now professors have to develop prior to the assessment. Following the logic where AES works best on standardized answers assignments, the exam format also has an impact. For instance, professors that use written exams formats that consist of a list of several questions and short answers, usually have in mind what the right answer is. Points are awarded according to whether

the clear answer is given and whether the student introduces new points- for instance real life case examples. On such type of an exam, AES would be able to grade how accurate the answer provided by the student matches the one provided in advance (and then trained on the software) by the professor. However, considering the moment state of the art AES tools, they are not yet able to evaluate brand new insights provided by the student because it is not yet clear how to predict them.

Despite the limitations that AES reveals, Shermis (2018 personal communications, 6 March) suggests that AES tools performs best on content-related answers. The tool is not able to tell the student whether they made a good argument or came to a right conclusion in the paper, but it can give a probability on the strength of their arguments. According to the author, the ability of examination content related answers depends on whether a specific or a generic model is established. Specific models are best for content scanning as they are trained on specific topics that the student is expected to address in the paper. It follows that, if applied to CBS assessments context, AES should take the form of a specific model. After having considered the current state of AES tools, the next section will touch upon the advantages and drawbacks of introducing such a technology at the higher education level that came out from the interviews with the examination system stakeholders at CBS.

6.4. BENEFITS AND DRAWBACKS OF IMPLEMENTING AES

Implementing automatic examining will allow CBS to increase efficiency in the examination system but, at the same time, there are drawbacks that have to be acknowledged. The points that are cited in this section align with the insights collected from professors and students, hence it might be useful here for the reader to refer back to the professors and students' profile canvases mapped out in Chapter 5.1.2.

Most importantly, as already mentioned before and in relation to the most important professors' pain, automatic examining will help reduce the time spent on examining and for students to receive a grade. In particular, besides being faster at examining summative assignments, professors will have the possibility of doing more formative assignments during the course without having to spend additional time on examining them. From the interviews with management and professors, it came out that many professors did not do formative assignments because they stated, no time is available

to correct them. At the same time, however, as it resulted that some departments are starting to look for new ways of increasing the use of formative assessments as "pass or fail" exams or requiring the students to send a minimum number of completed ones in order to be admitted to the final exam. Additionally, as one professor stated, the university has a contract with the Ministry of Higher Education that aims at increasing the study intensity of students as they are said to show a low study effort during the course of the classes. By cutting out time spent on examining, professors will have the opportunity of dedicating their time on more valuable activities like teaching and researching. Reflecting back to the topic of higher education in the literature review and to the students' gain pinpointed in the canvas as "seeing value come out of education", this will have a positive impact on the quality of teaching (Ramsden, 2003). From the extra time saved from examining, professors will be able to add additional valuable lectures that would incentivize students' participation. For instance, in-class participation can be stimulated by following Säljö (1979)'s point four and five outlined in Chapter 1: creating more in class students' discussions that incentivize students to compare different types of learnings and assigning in class practical or case study exercises to make students interpret knowledge and understand reality in a different way.

The other most important point, is that the use of these tools will also allow students to receive rapid feedback on formative assignment so that they will have the time to improve their knowledge in advance for the final exam. Besides feedback on formative assessments, with the same system, they will also be able to get feedback on their final exams which is one the "jobs" that students want the examination system to perform.

By using AES, the examination process will become more standardized as professors will have to design their assignments by developing a standard solution beforehand or a specific rubric on which the software will be trained to grade the exam and provide feedback. In relation to this, the examination process will become more objective. Having a more objective examination process will enhance transparency and fairness of the examination method: students will all be evaluated against the same and predetermined criteria.

Last but not least, AES will provide feedback to each single student, as well as a general overview of the students' performances and knowledge level to the professor. This will help the teaching professor investigate whether there are determined topics that have to be explained further during

the rest of the course, and at the end, after the final exam, understand whether their teaching measures have to be improved or changed.

There are some potential drawbacks that might result as a consequence of the implementation of automatic examining, some concerning the examination process itself and others concerning ethical matters. The first one is that students, after learning how to use these tools and knowing what these tools exactly look for when examining, may figure out how to deceive and understand the system in order for them to receive a higher grade, if automatic examining is used for final exams. Another point that has to be examined in order to be in accordance with the law, if such a solution was to be implemented, to ensure that in the formal examination process of summative assessments, professors will still have the role of signing off the grades, whereby the software is only an aid to a professor and not replacing the professor (Ministry of Higher Education 2018, personal communication, 22 February). Tightly related to the aspect that professors will not read all of the written assignments, there is the issue of using AES for assessing the written productions that students are usually required to submit at CBS as a part and preparation for the oral discussion. When the examination is of this type, at the oral exam, the professor usually starts off by asking questions related to the students' written production and that requires the student to further reflect on what he has written. On the basis of the students' answers, the professor will then move to other topics in the syllabus. When a written production is assessed by an AES tool, it may be harder for the professor to come up with questions that are related to the student's paper as he was not able to read it fully. Looking at the work that has to be done in order to set the software for each different course, professors need to put time in, firstly, finding out the content on which to train the software and, secondly, developing the solutions that they want their students to come up with together with listing predictable mistakes in the system.

Issues may arise when students get low grades through AES. As, Wind (2018, personal communication, 8 March), co-founder and CEO of Peergrade, states, such a new technology will take time to meet the trust of students and hence students might be very satisfied when they get a good grade but feel angry and not treated fairly when they get a low grade. This trust issue means that students may end up questioning their grades more by a machine than the ones provided by a professor and hence apply for complaints even more frequently than before. In such a case, the

professor will have to come into place to evaluate the complaint and establish whether the grade given was the right one.

One of the ethical concerns has to do with the difference between humans and machines. By reflecting on the use of automatic examining, a professor brought up the pain of facing the tradeoff of having to give up authenticity in order to achieve standardization. This is because, on the one hand, AES would increase the fairness and objectivity of examinations creating an important benefit to the students that feel that they are not treated equally with respect to the others and on the other hand, it will decrease the value of authenticity. Three professors during the interviews said that, even though examining more than 100 papers is a struggle, they actually learn new things from what the student write in their papers and this is something that amuse and excite them. Writing essays that have to be graded by a scoring engine might turn out as a low value activity for students as there will no longer be a human behind the process that would be able to appreciate the pieces of writing. In relation to this, Yonk (2017) said that no matter how much human-like these tools becomes, the question will still be on how authentic its thinking processes and responses will be.

Notwithstanding the ethical drawback brought up by professors, it is important here to consider that the value of education is however not lost. While students will be graded by a machine, as already stated among the benefits of automatic examining, professors will spend more time in class interacting with them through engaging exercises where the students have to bring in their own perspectives and thoughts. It can be hence stated and argued that, with automatic examining, the value of education and teaching will shift from the moment after the final exam and after receiving the grade to the moment in which the class is actually being taught. If students were to see this value added throughout the course of a class, they would be more incentivized to participate and, by actually participating, might thereupon increase the chances of getting a higher grade in the final exam.

Table 13: Benefits and drawbacks of AES solution

Benefits	Drawbacks
• Less time spent on examining • Less time for students to receive a grade • Possibility of doing both formative and summative assessments • Increasing students' study effort during courses by using formative assessments • More time for professors for focusing on teaching and researching • Students will receive feedback both on summative and formative assessments • Standardized and objective examination process • More transparent and fair examination process • The professor will have an overview of the level of the entire class	• Students might learn how to game the system to get higher grades • Professors have to trust the software giving the grades to students • The professor will not be able to know enough about the students' written composition that is prepared for the oral examination • Professors need to spend time on setting up the software • Students not trusting the grades that were given through automatic examining and filing more complaints • Losing human authenticity • Losing the value of written composition

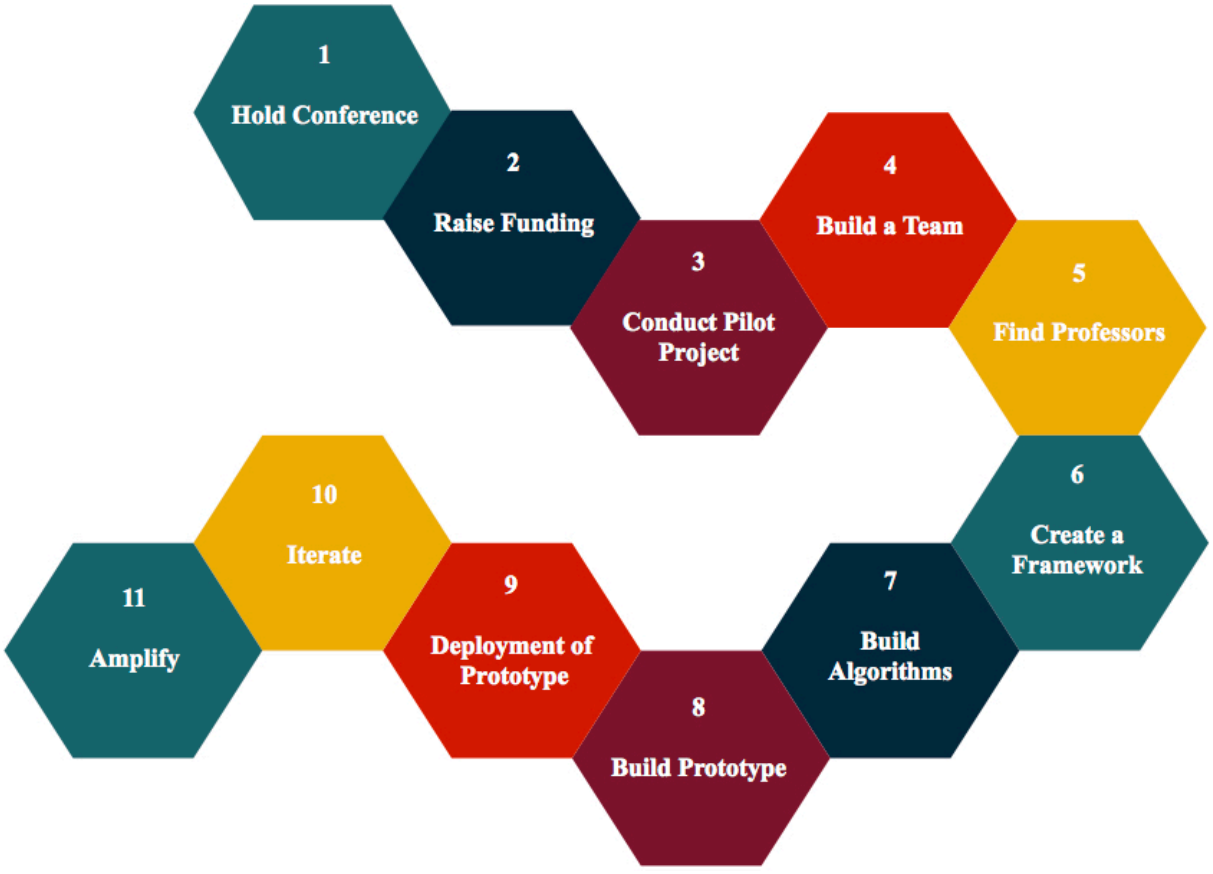
7. MOVING CBS TOWARDS ARTIFICIAL INTELLIGENCE

This chapter proposes a suggested roadmap for CBS to implement an AES software, based on the key insights taken from the research. A detailed description follows to outline the different steps and objectives of the project, provided with a pilot project plan timeframe.

7.1.SUGGESTED ROADMAP

As the benefits, drawbacks and limitations have been explained, to implement such a solution at CBS, a suggested roadmap will be described. The roadmap is based on assumptions, as well as suggestions given within *Service Automation: Robots and the Future of Work* related to the Service Delivery Automation (SDA) deployment roadmap (Willcocks and Lacity, 2016), shown in Appendix 2. This combination of knowledge will enable us, as researchers, to select the right processes and overcome conceptual barriers to automation.

Figure 17- Suggested roadmap



Step 1: Hold Conference

To begin, a conference should be held to attract professors and students to understand the opportunity that arises when combining AI technology within the examination system. The four learning objectives for the conference will be:

1. How technology is transforming the educational sector
2. Our findings: There are issues with the current examination system, where opportunities can arise
    - Describe how a school can re-allocate its resources to become more efficient
    - AES
3. Guest Experts (AES experts, companies)
4. Pitch the project

Referring to point one, technology as a whole can be discussed in a broad way for participants to understand the higher educational sector and what opportunities can arise ahead of it. The approach of blended learning at CBS could be used as a broad concept example of how using technology can enhance students learning and professors teaching. This pedagogic approach is seen as an individualized learning experience, and therefore can reflect and collaborate with other technological approaches for the future of education.

After describing the landscape of the future of education and technology, the findings of this thesis will be presented for participants to recognize and observe the current examination system, as well as the amount of resources that are being spent on this activity at CBS. Within the findings, the idea of the servitization the examination system by transforming the process to a SDL world will be explored. The AES software's will be introduced as a way of re-allocating resources for teaching professors' activities to create more value for higher learning, as previously described.

The conference will be held by inviting guest experts, including AES experts and organizations utilizing AI technology, to gain a technological perspective behind the idea of the servitization of the examination system.

Step 2: Raise Funding

After holding the conference and attracting interested professors and potential students, teaching and learning applications for research grants will be sought, which will later be touched upon within Chapter 9.

Step 3: Conduct Pilot Project

By raising funds, this will enable the process of conducting a pilot project to evaluate the feasibility and time, as well as evaluate how well the different tools can perform on formative assessments.

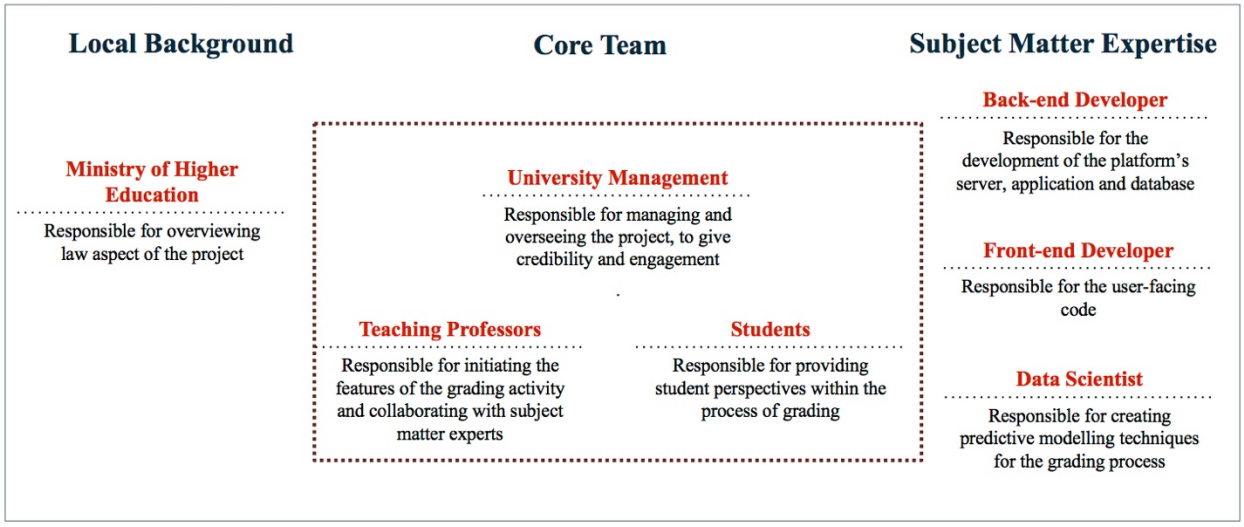
Step 4: Build a Team

Once receiving funding, a business analytics group will be formed, based on an interested team. With the purpose of building the digital structure of the AES platform, three technology experts are needed in this research project. The first one is a back-end developer that has competences in the fields of NLP and AI and who will be in charge of building the "skeleton" of the server, the application and the database of the online AES platform (Wales, 2017). The second expert needed is a front-end developer who will develop the user-facing code of the platform (ibid). The user-facing code includes both the design of the actions that will arise when the user (professor or student) will interact with the platform and the architecture and graphics that will enhance the functionality of it. Considering the variety of unstructured data on which the AES tool will have to be trained on and the variety of written assignments that it will have to grade, the third technology expert needed is a data scientist. The data scientist will apply ML on data to improve predictive modelling techniques for examining and analyzing patterns and relationships in students' papers for examining (Rouse, 2017). In addition to technology experts, the team will demand the presence of an individual that belongs to the management board of CBS in order to enhance credibility of the project, and student interns to gain student perspectives within the development of the AES platform as they are themselves part of the examination process.

Step 5: Find Interested Professors

Besides organizing a main team, it is essential to find interested professors that want to research on this topic. This includes those who are positive about improving the current examination system, as well as help the department with the process. By including professors, as they are better at pedagogical practices, this will help technological experts understand where these systems are coming from and how this can aid professors within the examination activity (Dr. Peter Vitartas 2018, personal communication, March 6). This includes collaborating with two to five professors and starting small, in order to create ambassadors for this project to take responsibility and promote while experimenting with it, to grow in the future.

Figure 18: Proposed team- structure and role



Step 6: Create a Framework

Once a concrete team is in place, the first step is to develop a framework that is wanted by co-creating with the different departments to analyze the effectiveness for each course of where the servitization of the examination system could be incorporated. Such a framework could include for example, when looking at Dr. Peter Vitartas who has implemented such a project in La Trobe University in Australia, evaluating assignments based on five different areas: assignment statistics, readability, concept coverage, critical thinking and discipline theory.

Figure 19: A framework for evaluating assignments

Source: Vitarta, 2017

As Benett (2011) suggests enhancing the understanding of the human scoring processes, these different areas will be based upon the findings that professors collaborate within and study regarding: the professor's teaching style, the learning objectives of the course and the readings that make up the list of the course syllabus. This will be done in connection to the professors examining style based on criteria that can be used to develop a rubric, explaining what each and every professor looks for within the specific exam that is being tested (Dr. Peter Vitartas 2018, personal communication, March 6).

Step 7: Build the Algorithms

For each of the areas within the above framework that will be developed, algorithms have to be developed for each unique component. In order to make sure that each of these algorithms will best imitate the human examination process, it is important that the professors involved in the project collaborate also at this stage to oversee that their examination method is properly translated.

Step 8: Build the Prototype

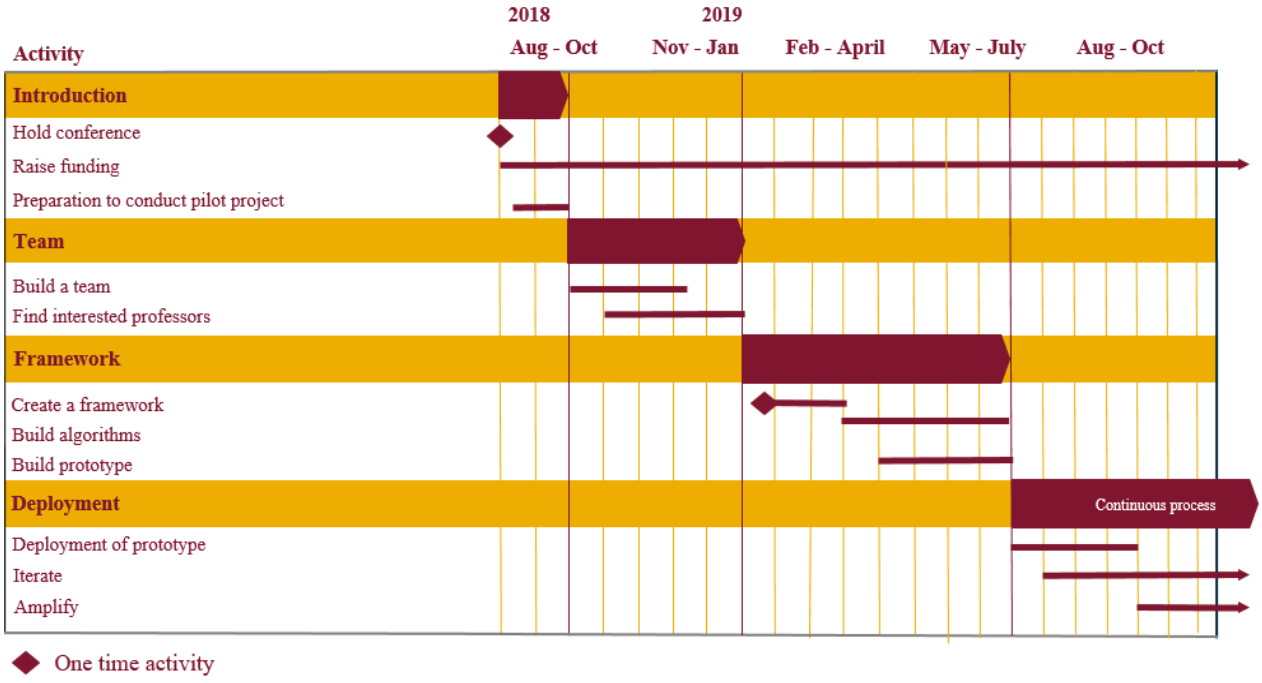
The first prototype model will be developed for examining formative assessments and hence it will include a limited number of features. For instance, while in a summative assessment a student is

tested on more developed skills, as they are expected to be able to acquire all of them by the end of the course. In a formative assessment students are usually tested only partially on learning objectives. An example could be to evaluate at the formative assessment event only on the knowledge of specific themes and theories, and at the final exam evaluate the application of that knowledge. Once the features of the prototype are determined, a web-based interface will be built. This will include a dashboard, whereby the technological process will include how to get students to access the program, for example connected to Learn or Digital Exam at CBS, and how these results can be relied back from the software to the professors (Dr. Peter Vitartas 2018, personal communication, March 6).

Step 9: Deployment of the Prototype

The first prototype will be tested on two to five professors' courses on which the study has been conducted. As the number of professors will gradually increase over time, a proposed tentative plan for the project will be outlined, ensuring approximate dates, as well as a transparent process for all stakeholders involved. Made with several assumptions, the plan is established to represent the first 15 months, which can be tailored as the process progresses and whereby further subject knowledge by all parties enhances.

Figure 20: Pilot project plan



These assumptions are based on two benchmarks: that of the time frame that blended learning initiatives are currently being in place at CBS (as previously described), and that of the implementation of AES softwares process based on Peter Vitartas pilot project being conducted in Australia.

Step 10: Amplify

In conclusion, as a lot of assumptions were made for the suggested roadmap and pilot project plan, it is important to continue with the pilot project in the future, after the 15 months, to improve the process and take this proof of concept further by continually refining the algorithms. Relating back to the initial framework, a fit will be achieved with the proof of concept, once professors and students pains, gains and jobs are undertaken. This is known as a *problem-solution fit,* previously described in Chapter 3 that is determined even if both sides of the customers are not using the product yet (Osterwalder et al, 2014). As trust issues may arise along the development process from students, as well as professors, in the case where such a tool could result in having inconsistencies and errors, the developed AES software will not be fully utilized on a large scale until it will be proven that a goal standard is achieved. Once the development of the software is tested on a small scale at CBS and a goal standard is achieved, a fit known as a *product-market fit,* will be established where more professors utilize and gain value in the software (ibid.).

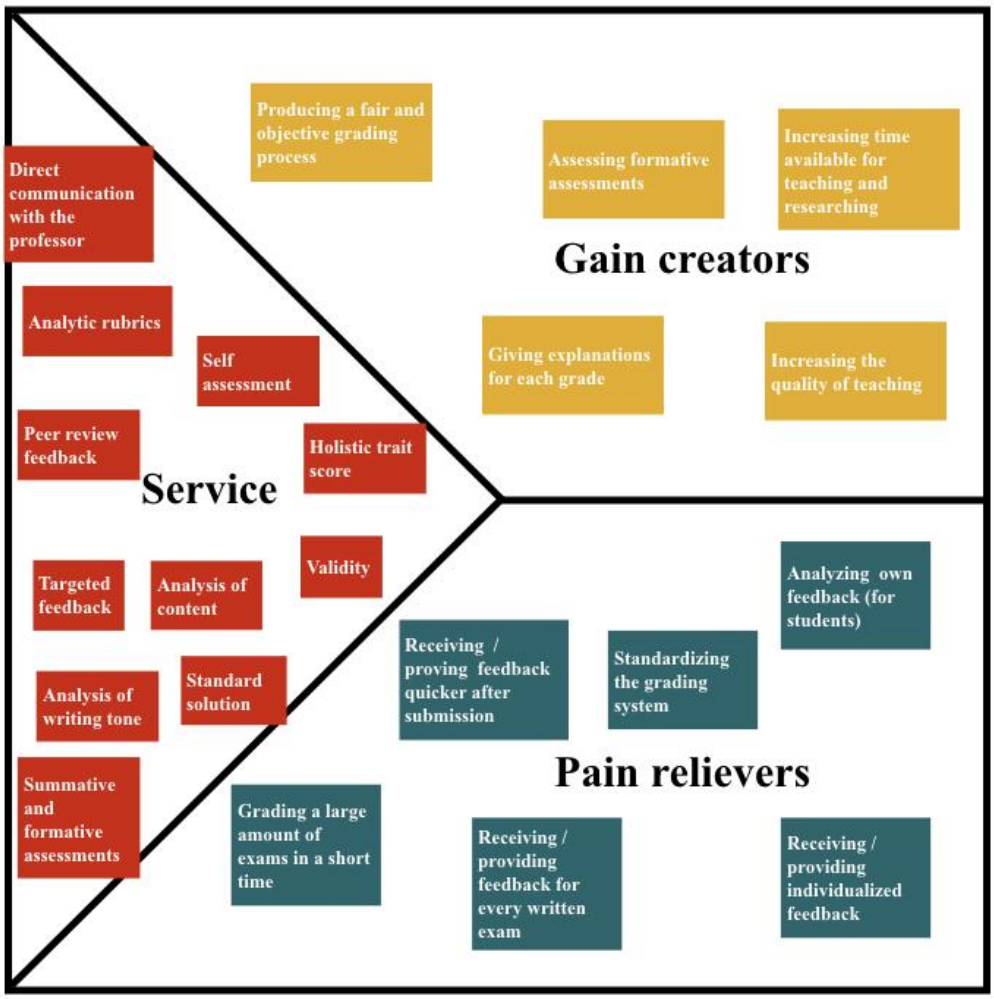
7.2 THE VALUE PROPOSITION THROUGH THE SERVITIZATION OF EXAMINING

As the future of technology will continually develop, along with human minds, operant resources will continually enhance operand resources. Referring back to the literature review on AI, it is said that at the moment automation is facing the *third cycle of AI*, soon AI solutions will pick up the architecture of man-machine collaboration and enter into the fourth cycle (PWC, 2018). Similarly, AES is going towards this new cycle by helping professors overcome the repetitive and time-consuming activity of examining. By following Osterwalder et al (2014) framework, once the solution of a customized AES tool creates a fit with the professors and students' profiles, a value proposition fit is achieved and can be finally implemented in the CBS examination system. Following, the value proposition canvas that came out from the servitization of examining is

depicted, shown in Figure 21. As a result of the interconnection between the professor and the student, the first one setting up and correcting the exam and the second one receiving and submitting the assessment, the proposed solution is a single one. To this extent, the solution of creating an AES platform for CBS can be seen as a "two-sided platform" that serves in different ways both of the "customer segments".

By looking at the value proposition canvas it can be noticed that both gain creators and pain relievers have been created against the most important insights of professors and students that were pinpointed in each customer profile. The features of the service, represented on the left triangle of the canvas, were established by taking as a starting point the compelling features of the already available tools and, additionally, creating new ones as a response to the two different stakeholders' profiles.

Figure 21: Value proposition



The features that were directly taken from the already available tools are the feedback capabilities. The proposed solution should be able to include all of the different types of feedback capabilities leaving then the option to the single professors of picking up the ones they want to provide their students with in the assessments. *Targeted feedback* and *holistic trait scores* should be included in order for the student to get an overview of the final performance and, at the same time, an explanation of the specific errors and parts that have to be improved in the assignment. This feature, according to Knight (1995), has the purpose of helping students improve their knowledge and skills by looking at and correcting their own mistakes. The option of having a *communication channel* between professors and the single students is of high relevance too. In the case where students were not able to understand or agree with a particular score or comment received through the software, they would have the opportunity to *flag* (David Wind 2018, personal communication, 8 March) the specific comment to the professors and directly communicate with them. Moving to a different concept of feedback, the new tool should also comprise of a component that allows for *peer-review feedback* and *self-assessment*. This is because *peer review* gives students the advantage of understanding quality standards by criticizing and evaluating other peers' performances (Evans, 2013) and *self-assessment* represents a purpose for the students to be critical on their own performance by using rubrics (ibid.).

By looking at the future, as already cited in the roadmap, some additional features will have to be added in order to achieve the gold standard and start using AES officially for formative and summative assessments. In particular, a feature that, by looking at the already available tools in the market, has to be improved, is the ability of testing out the content and validity of the points made by the student in the written paper. Even though nowadays ML and AI capabilities are still limited, it is key here to find a way to predict new content that may be brought in by the student. Last but not least, in relation to the future of AI where emotional intelligence will be eventually reproduced by machines (Yonck, 2017), AES will have to be able to understand the tone of writing of the student, in order to better analyze whether the arguments that they make have a valid reasoning or present some errors. This would benefit students that have an elevated writing style that a software might, at the moment, not be able to fully appreciate and award.

8. RESEARCH LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORK

Reflecting on the level of expertise that was brought behind this research, it is important to acknowledge that the most important limitation of this research was the lack of practical computer science know-how. Notwithstanding this limitation, the entrepreneurial and innovative approach taken allowed to look strategically at the opportunities and features that AI technology, such as AES, could bring to CBS and adopt in the future.

The suggested roadmap for implementing an AES system at CBS can be considered as a starting point for future research. As noted throughout this research project, scientific literature on the use of AES for higher education summative assessments is still limited. However, this paper has shown that the already existing tools in the market are already commercialized and have already been implemented in several state-wide and high-stake examinations. This means that this technology has the potential of overtaking other examinations formats at different educational levels. Since AI and AES experts have yet to find the perfect algorithms that best imitate the activity of human examining, specifically on other aspects different from formative writing, it remains unclear which format these tools will take in the future.

First, as this research was based on the case study of CBS, it is important for future research to analyze other educational systems before directly applying it to other learning realities. In particular, compliance with other educational systems' regulations and law should be further looked upon together with the structure of written examinations and grading scales. Along with the examination customs of other countries, their feedback system should be additionally studied, by carefully analyzing whether and how any sort of feedback is provided. Speaking of both other countries and CBS case study, prior to the development of a prototype, supplementary research should be conducted on the differences between bachelor's and master's level written assignments. This is because master's students are expected to be more experienced in academic writing than bachelor's and, hence, examinations might give different weights to the writing abilities and styles of students.

Concerning the roadmap and raising funds for such a pilot project, an estimation of the value of price or percentage has not been able to be determined for the purpose of this thesis. Going further, looking at the AES solution itself, as already suggested in the roadmap, further technical developments have to be made. Potential questions arise in the context of developing new algorithms. The first one concerns how to test strategic, analytical and critical thinking in a paper accounting for the different ways in which teaching professors test these skills when correcting a written assignment. The second one is geared towards finding a way for the software to understand the student's tone of writing and award for using an elevated tone of writing. The last research question that was brought up throughout this research, and found yet unanswered reflecting nowadays literature, was how to analyze and grade the new knowledge that is brought by the student within the paper in forms of personal experiences, new examples and case studies. Last but not least, speaking of already AES existing features, further improvements should be done on how the software analyzes the accuracy of the argumentations and points made by the student as well as how to evaluate the content level, specificity of topics, and theories mentioned in the paper.

9. CONCLUSION

This chapter concludes the research findings, reiterates the recommendations for CBS, reflects upon the limitations of the research study, and highlights the academic contribution of the research study.

Together with additional innovations as, for instance, additive manufacturing (3D printing), AI is considered one of major breakthrough of the *Content - Centric Era* (2005-2025) where technology is said to be conceived to enhance customization of products and services (Willcocks and Lacity, 2016). Being part of the AI transformation cycle, AES is a technology that grades written assessments through a standardized and objective process, but the way through which it is set up and sends grades and feedback to students is customized. Customization comes in place so that professors have the possibility of setting up their own evaluation criteria, according to the specific subject on which the assessment is regarding. Likewise, students, instead of receiving solely a standard solution of the exam, receive customized feedback cloesly related to their own specific performance.

Having an examination tool that functions as a service for professors and students, implies a standardization of the examination process. By observing the EU context, in line with The Bologna Accord 2005 (Gmac.com, 2005) which aims at standardizing the educational system of all the different EU countries,
AES, if implemented, would help achieve harmonization between different higher education systems and make it easier for a member country's student to move from one system to another while having the same examination process.

Our research contributed to fill in the gap in the literature of AES in the specific context of higher education and summative assessments, as the AES academic discourse was mainly concentrated on K-12 education and formative assessments. Moreover, looking at how the examination process was originally, legally and formally conceived, we adopted an innovative approach. This consisted of applying human centered design through the VPD framework (Osterwalder et al, 2014) and by investigating knowledge and value creation with the SDL approach of the servitization of the examination process (Vargo and Lusch, 2004). The central question posed by our research was:

What are the potentials of AI in examinations at CBS? To address this case study, we adopted an analytical approach built on four different phases, comprising of:

1) Understanding the technological landscape,
2) scoping out the opportunity at CBS,
3) identifying how much resources are spent on examining, and
4) determining how an AI solution can be adopted and created by conducting an analysis of the existence of available tools in the market.

Interestingly, we came across three findings that show relevant potential for adopting AI technology within the examination system at CBS. *Firstly*, the examination activity is defined as a cumbersome and time-consuming activity by teaching professors. *Secondly*, students are not fully satisfied by the process through which they are examined and feel the need of increased feedback provisioning activity. *Thirdly*, CBS management is at the moment looking at new ways of improving the feedback system and enhancing the quality of teaching and learning through technology. As a result, notwithstanding the additional improvements that AES tools need in order to fully imitate human graders, AES resulted as the perfect match to meet the insights collected from the stakeholders of the examination process at CBS.

As CBS spent 34%, a total of DKK 62,446,338 of their expenditures in 2017 on the examination activity, this strength the idea of the servitization of the examination process to reallocate resources, money and time, from the activity of examining to teaching and researching. The outcome of this research is hence the provisioning of a roadmap that would guide CBS in a customized development and implementation of AES that has the subsequent purpose of enhancing the value of the institution's educational innovation.

REFERENCE LIST

Accenture (2016). *Intelligent Automation: The essential new co-worker for the digital age*. Technology Vision 2016. [online] Accenture.com. Available at: https://www.accenture.com/t20160125T111718Z__w__/dk-en/_acnmedia/Accenture/Omobono/TechnologyVision/pdf/Intelligent-Automation-Technology-Vision-2016.pdfla=en#zoom=50 [Accessed 22 Mar. 2018].

Ambrose, S., Bridges, M., Lovett, M., DiPietro, M. and Norman, M. (2010). *How Learning Works: Seven Research-Based Principles for Smart Teaching.*. San Francisco: Jossey-Bass Publishers.

Arreola, R. (1998). *Writing Leaning Objectives - A Teaching Resource Document from the Office of the Vice Chancellor for Planning and Academic Support*. [ebook] Memphis: The University of Tennessee. Available at: https://www.uwo.ca/tsc/graduate_student_programs/pdf/LearningObjectivesArreola.pdf [Accessed 20 Mar. 2018].

Biggs, J. (2003). *Teaching for quality learning at university*. 2nd ed. Maidenhead: McGraw-Hill/Society for Research into Higher Education/Open University Press.

Blei, D., Ng, A. and Jordan, M. (2003). Latend Dirichlet Allocation. *Journal of Machine Learning Research*, [online] 3. Available at: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf [Accessed 15 Mar. 2018].

Blog.cbs.dk. (2018). *Teaching formats for 2018/2019 – CBS TEACHING & LEARNING*. [online] Available at: http://blog.cbs.dk/teach/teaching-formats-for-2018-2019/ [Accessed 25 Apr. 2018].

Bradburn, N., Sudman, S. and Wansink, B. (2004). *Asking Questions*. Hoboken: John Wiley & Sons, Inc.

Brown, J. (2016). *How U of Michigan Built Automated Essay-Scoring Software to Fill 'Feedback Gap' for Student Writing - EdSurge News*. [online] EdSurge. Available at: https://www.edsurge.com/news/2017-06-06-how-u-of-michigan-built-automated-essay-scoring-software-to-fill-feedback-gap-for-student-writing [Accessed 29 Apr. 2018].

Brøchner Nielsen, N. (2013). *The principles for "The great learning objective"*. Copenhagen: Copenhagen Business School, p.3.

Cbs.dk. (2016). *The 7-point grading scale used at Danish Universities*. [online] Available at: https://www.cbs.dk/files/cbs.dk/danish_grading_scale_2017.pdf.

Constantin, James A. and Robert F. Lusch (1994), *Understanding Resource Management.* Oxford, OH: The Planning Forum.

Copenhagen Business School. (2018)a. *Business Intelligence and Development | CBS - Copenhagen Business School*. [online] Available at: https://www.cbs.dk/en/about-

cbs/organisation/administrative-units/financial-analysis/business-intelligence-and-development [Accessed 6 Apr. 2018].

Copenhagen Business School. (2018)b. *Studying at CBS | CBS - Copenhagen Business School*. [online] Available at: https://www.cbs.dk/en/study/bachelor/studying-cbs [Accessed 25 Apr. 2018].

Corestandards.org. (2018). *About the Standards | Common Core State Standards Initiative*. [online] Available at: http://www.corestandards.org/about-the-standards/ [Accessed 17 Mar. 2018].

Daugherty, P. and Wilson, H. (2018). *Human + Machine: Reimagining Work in the Age of AI*. 1st ed. Boston: Harvard Business Review Press.

De Vaus, D. (2014). *Surveys in social research*. 6th ed. London: Routledge, p.9.

Dictionary of Information Technology (2002). In: 3rd ed. London: Peter Collin Publishing.

Education Northwest (2018). *What are the Traits?*. [online] educationnorthwest.org. Available at: http://educationnorthwest.org/traits/trait-definitions [Accessed 27 Mar. 2018].

Eisenhardt, K. (1989). Building Theories from Case Study Research. *Academy of Management Review*, 14(4), pp.532-550.

Eng.uvm.dk. (2018). *7-point grading scale*. [online] Available at: http://eng.uvm.dk/general-overview/7-point-grading-scale.

Ets.org. (2018)a. *Criterion: About the Criterion Service*. [online] Available at: https://www.ets.org/criterion/about/ [Accessed 17 Mar. 2018].

Ets.org (2018)b. *About the e-rater Scoring Engine*. [online] Ets.org. Available at: https://www.ets.org/erater/about/ [Accessed 18 Mar. 2018].

Evans, C. (2013). Making Sense of Assessment Feedback in Higher Education. *Review of Educational Research*, [online] 83(1), pp.70-120. Available at: http://journals.sagepub.com/doi/pdf/10.3102/0034654312474350 [Accessed 18 Mar. 2018].

Evans, D. (2018). *Cognitive Computing vs Artificial Intelligence: what's the difference? - iQ UK*. [online] iq.intel.co.uk. Available at: https://iq.intel.co.uk/cognitive-computing-vs-artificial-intelligence/ [Accessed 23 Mar. 2018].

Everest Group (2018). *RPA & AI: Different Approaches to Problem Solving | Market Insights™*. [online] Everest Group. Available at: https://www.everestgrp.com/2018-01-rpa-ai-different-approaches-problem-solving-market-insights-43514.html/ [Accessed 25 Mar. 2018].

Gmac.com. (2005). *The Bologna Accord: A European Revolution with Global Implications*. [online] Available at: https://www.gmac.com/why-gmac/gmac-news/gmnews/2005/january-february/the-bologna-accord-a-european-revolution-with-global-implications.aspx.

Gummesson, Evert (1994), "Broadening and Specifying Relationship Marketing," *Asia-Australia Marketing Journal*, 2 (August), 31-43.

Hardesty, L. (2018). *Automatically grading programming homework*. [online] MIT News. Available at: http://news.mit.edu/2013/automatically-grading-programming-homework-0603 [Accessed 29 Apr. 2018].

Hedman, J. (2018). *Fintech Revolution - Home Assignment Exam at Copenhagen Business School*. [image].

Hubert.ai (2017). *AI In Education — Automatic Essay Scoring*. [online] Medium. Available at: https://medium.com/hubert-ai/ai-in-education-automatic-essay-scoring-6eb38bb2e70 [Accessed 27 Mar. 2018].

IBM.com. (2018)a. *IBM Watson*. [online] Available at: https://www.ibm.com/watson/ [Accessed 22 Apr. 2018].

IBM.com. (2018)b. *IBM Watson Element for Educators - Overview - United States*. [online] Available at: https://www.ibm.com/us-en/marketplace/education-insights [Accessed 22 Apr. 2018].

IBM.com. (2018)c. *IBM Watson Enlight for Educators - Overview - United States*. [online] Available at: https://www.ibm.com/us-en/marketplace/personalized-learning [Accessed 22 Apr. 2018].

IDEO.org (2015). *The Field Guide to Human - Centered Design*. IDEO.org, p.43.

Kiser, M. (2016). *Introduction to Natural Language Processing (NLP) - Algorithmia Blog*. [online] Algorithmia Blog. Available at: https://blog.algorithmia.com/introduction-natural-language-processing-nlp/ [Accessed 27 Mar. 2018].

Knight, P. (1995) *Assessment for Learning in Higher Education.* London: Kogan Page

Knight, W. (2016). *IBM's Watson is everywhere lately—but what is it?*. [online] MIT Technology Review. Available at: https://www.technologyreview.com/s/602744/ibms-watson-is-everywhere-but-what-is-it/ [Accessed 22 Apr. 2018].

Kvale, S. (2007). *Doing interviews*. London: SAGE.

Lagi, M. (2018). *Natural Language Processing – Business Applications*. [online] TechEmergence. Available at: https://www.techemergence.com/natural-language-processing-business-applications/ [Accessed 27 Mar. 2018].

Luckin, R., Holmes, W., Griffiths, M. & Forcier, L. B. (2016). Intelligence Unleashed. An argument for AI in Education. London: Pearson.

Martinho-Truswell, E. (2018). *How AI Could Help the Public Sector*. [online] Harvard Business Review. Available at: https://hbr.org/2018/01/how-ai-could-help-the-public-sector [Accessed 9 May 2018].

Measurement Incorporated (2016). *Get The Most Out of PEG Writing Features*. [PDF] pegwriting.com. Available at: https://www.pegwriting.com/sites/default/files//PW-Features.pdf [Accessed 15 Mar. 2018].

Measurement Incorporated (2016). *Get The Most Out of PEG Writing Features*. [PDF] pegwriting.com. Available at: https://www.pegwriting.com/sites/default/files//PW-Features.pdf [Accessed 15 Mar. 2018].

Measurement Incorporated (2017)a. *The Engine Driving Automated Essay Scoring*. [PDF] pegwriting.com. Available at: http://www.pegwriting.com/sites/default/files/peg-Info-report.pdf [Accessed 14 Mar. 2018].

Measurement Incorporated (2017)b. *PEG Writing Information Packet*. [ebook] pegwriting.com. Available at: http://pegwriting.com/sites/default/files/PW-Info-ePacket-Full.pdf [Accessed 15 Mar. 2018].

Measurement Incorporated (2017)c. *Score Report with Holistic and Trait Scores, and Suggested Lesson Link*. [image] Available at: http://pegwriting.com/sites/default/files/PW-Info-ePacket-Full.pdf [Accessed 16 Mar. 2018].

Measurement Incorporated (2018). *PEG Writing Information Packet*. [ebook] Measurement Incorporated. Available at: http://pegwriting.com/sites/default/files/PW-Info-ePacket-Full.pdf [Accessed 7 May 2018].

Medium. (2018). *Immersive Futures in Education and Design – Method Perspectives – Medium*. [online] Available at: https://medium.com/method-perspectives/immersive-futures-in-education-and-design-4fcad20522ab [Accessed 27 Mar. 2018].

Open.ac.uk (2018). *OpenMark Examples* |. [online] Open.ac.uk. Available at: http://www.open.ac.uk/openmarkexamples/ [Accessed 7 May 2018].

Osterwalder, A., Pigneur, Y., Bernarda, G., Smith, A. and Papadakos, T. (2014). *Value proposition design*. Hoboken, N.J.: John Wiley & Sons.

Page, E. (1966). The Imminence of ... Grading Essays by Computer. *The Phi Delta Kappan*, [online] 47(5). Available at: http://www.jstor.org/stable/pdf/20371545.pdf?refreqid=excelsior%3A4aefb9ef310d70b996f3804981326dea [Accessed 26 Mar. 2018].

Page, E. B., Poggio, J. P., & Keith, T. Z. (1997). *Computer analysis of student essays: Finding trait differences in the student profile.*

Pearson (2007)a. *General Overview of WriteToLearnTM and Its Components*. [ebook] Available at: https://cdn2.hubspot.net/hubfs/559254/WTL/resources/WTL_WhitePaper_GeneralOverview_WTL_Components_r1.pdf?t=1520621384043 [Accessed 18 Mar. 2018].

Pearson (2007)b. *Summary Street feedback example*. [image] Available at: https://cdn2.hubspot.net/hubfs/559254/WTL/resources/WTL_WhitePaper_GeneralOverview_WTL_Components_r1.pdf?t=1520621384043 [Accessed 18 Mar. 2018].

Pegwriting.com. (2018). *Home | PEG Writing*. [online] Available at: http://www.pegwriting.com [Accessed 14 Mar. 2018].

Penrose, Edith T. (1959), *The Theory of the Growth of the Firm*. London: Basil Blackwell and Mott.

Pettigrew, A. (1990). Longitudinal Field Research on Change: Theory and Practice. *Organization Science*, 1(3), pp.267-292.

Porath, C. (2016). *Give Your Team More-Effective Positive Feedback*. [online] Harvard Business Review. Available at: https://hbr.org/2016/10/give-your-team-more-effective-positive-feedback [Accessed 28 Mar. 2018].

Potts, M. (2018). *Ellis Page, 81, a Developer of Computerized Grading, Dies*. [online] Nytimes.com. Available at: https://www.nytimes.com/2005/05/23/us/ellis-page-81-a-developer-of-computerized-grading-dies.html [Accessed 25 Mar. 2018].

PwC (2018). *Will robots really steal our jobs?*. An international analysis of the potential long term impact of automation. [online] Available at: https://www.pwc.com/hu/hu/kiadvanyok/assets/pdf/impact_of_automation_on_jobs.pdf [Accessed 7 Mar. 2018].

Ramsden, P. (1996). *Learning to Teach in Higher Education*. USA and Canada: Routledge.

Ramsden, P. (2003). *Learning to Teach in Higher Education*. 2nd ed. USA and Canada: RoutledgeFalmer.

Regression Analysis: How do I interpret R-Squared and Assess the Goodness-of-Fit?. (2013). [Blog] *The Minitab Blog*. Available at: http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit [Accessed 12 Apr. 2018].

Reuner, T. (2016). *Intelligent Automation 2016 - Excerpt for Accenture*. HfS Blueprint Report. [online] HfS Research Ltd. Available at: https://www.accenture.com/t20170411T110624Z__w__/us-en/_acnmedia/PDF-39/Accenture-HfS-Blueprint-Intelligent-Automation-2016-Excerpt.pdf [Accessed 23 Mar. 2018].

Rouse, M. (2005). *What is K-12? - Definition from WhatIs.com*. [online] WhatIs.com. Available at: https://whatis.techtarget.com/definition/K-12 [Accessed 12 May 2018].

Rouse, M. (2017). *What is data scientist? - Definition from WhatIs.com*. [online] SearchEnterpriseAI. Available at: https://searchenterpriseai.techtarget.com/definition/data-scientist [Accessed 8 May 2018].

Rowntree, D. (1977) *Assessing Students*, London: Harper & Row.

Russell, S. and Norvig, P. (2010). *Artificial intelligence: A Modern Approach*. 3rd ed. Pearson.

Salkind, N. (2010). Primary Data. In: *Encyclopedia of Research Design*. [online] Thousand Oaks, CA: SAGE Publications Ltd. Available at: http://methods.sagepub.com/reference/encyc-of-research-design/n333.xml [Accessed 15 Apr. 2018].

Sas.com (2018)a. *Artificial Intelligence – What it is and why it matters*. [online] Sas.com. Available at: https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html [Accessed 13 May 2018].

Sas.com (2018)b. Machine Learning: What it is and why it matters. [online] Sas.com. Available at: https://www.sas.com/en_us/insights/analytics/machine-learning.html [Accessed 12 May 2018].

Shermis, M. (2010). Chapter 10: Automated Essay Scoring in A High Stakes Testing Environment. In: V. Shute and B. Becker, ed., *Innovative Assessment for the 21st Century*. [online] Springer Science+Business Media, pp.167 - 185. Available at: https://link.springer.com/content/pdf/10.1007/978-1-4419-6530-1_10.pdf [Accessed 27 Mar. 2018].

Shermis, M. and Di Vesta, F. (2011). *Classroom assessment in action*. Lanham, Maryland: Rowman & Littlefield Publishers.

Shermis, M. and Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York: Routledge.

Shermis, M. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, [online] 20, pp.53-76. Available at: https://www.sciencedirect.com/science/article/pii/S1075293513000196.

Shermis, M., Burstein, K., Elliot, N., Miel, S. and Foltz, P. (2016). Automated Writing Evaluation. An Expanding Body of Knowledge. In: C. MacArthur, S. Graham and J. Fitzgerald, ed., *Handbook of Writing Research*, 2nd ed. Guilford Press, pp.395 - 405.

Sheth Jagdish, Rajendra S. Sisodia, and Arun Sharma (2000), "The Antecedents and Consequences of Customer-Centric Marketing," *Journal of the Academy of Marketing Science*, 28 (Winter), 55-66.

Smith, W. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M.M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press, pp. 142-205.

Stevens, D. and Levi, A. (2013). *Introduction to rubrics: an assessment tool to save grading time, convey effective feedback, and promote student learning*. 2nd ed. Sterling, Va.: Stylus.

Stockholm University (2013). *Automated Essay Scoring - Stockholm University - Department of Linguistics*. [online] Ling.su.se. Available at: https://www.ling.su.se/english/nlp/tools/automated-essay-scoring [Accessed 29 Apr. 2018].

Strategyzer.com (2018). *Value Proposition Canvas*. [online] Strategyzer.com. Available at: https://strategyzer.com/canvas/value-proposition-canvas [Accessed 6 May 2018].

Study.com (2018). *Essay Prompt: Definition & Examples - Video & Lesson Transcript | Study.com*. [online] Study.com. Available at: https://study.com/academy/lesson/essay-prompt-definition-examples-quiz.html [Accessed 20 Mar. 2018].

Säljö, R. (1979) 'Learning in the learner's perspective. I. Some common-sense conceptions', *Reports from the Institute of Education, University of Gothenburg*, 76.

Technopedia (2018)a. *What is Cognitive Computing? - Definition*. [online] Techopedia.com. Available at: https://www.techopedia.com/definition/32037/cognitive-computing [Accessed 23 Mar. 2018].

Technopedia (2018)b. *What is Autonomic Computing? - Definition*. [online] Techopedia.com. Available at: https://www.techopedia.com/definition/191/autonomic-computing [Accessed 23 Mar. 2018].

Technopedia (2018)c. *What is an Algorithm? - Definition*. [online] Techopedia.com. Available at: https://www.techopedia.com/definition/3739/algorithm [Accessed 27 Mar. 2018].

The Open University (2018)a. *OpenMark Examples*. [online] Open.ac.uk. Available at: http://www.open.ac.uk/openmarkexamples/ [Accessed 18 Mar. 2018].

The Open University (2018)b. *Promoting learning with instant feedback | OpenMark Examples*. [online] Open.ac.uk. Available at: http://www.open.ac.uk/openmarkexamples/overview/promoting-learning-instant-feedback [Accessed 18 Mar. 2018].

The Open University (2018)c. *Notes for question authors | OpenMark Examples*. [online] Open.ac.uk. Available at: http://www.open.ac.uk/openmarkexamples/overview/notes-question-authors [Accessed 18 Mar. 2018].

The Open University (2018)d. *Guidelines for question authors | OpenMark Examples*. [online] Open.ac.uk. Available at: http://www.open.ac.uk/openmarkexamples/overview/guidelines-question-authors [Accessed 18 Mar. 2018].

The Open University (2018)e. *OpenMark - Free Text answer*. [image] Available at: http://www.open.ac.uk/openmarkexamples/text-response/free-text [Accessed 18 Mar. 2018].

The Open University (2018)f. *OpenMark - Simple text entry*. [image] Available at: http://www.open.ac.uk/openmarkexamples/text-response/simple-text-entry [Accessed 18 Mar. 2018].

The Writing Center (2018). [online] Writingcenter.unc.edu. Available at: https://writingcenter.unc.edu/tips-and-tools/essay-exams/ [Accessed 20 Mar. 2018].

Ufm.dk. (2018). *Tilskud til uddannelse — Uddannelses- og Forskningsministeriet*. [online] Available at: https://ufm.dk/uddannelse/videregaende-uddannelse/universiteter/okonomi/uddannelsestilskud [Accessed 1 Mar. 2018].

UNC Charlotte (2018). *Bloom's Taxonomy of Educational Objectives | The Center for Teaching and Learning*. [online] Teaching.uncc.edu. Available at: https://teaching.uncc.edu/services-programs/teaching-guides/course-design/blooms-educational-objectives [Accessed 19 Mar. 2018].

Vantage Learning (2018). *MY Access!® Writing and Assessment Solution*. [online] Vantagelearning.com. Available at: http://www.vantagelearning.com/products/my-access-school-edition/ [Accessed 18 Mar. 2018].

Vargo, S. and Lusch, R. (2004). Evolving to a New Dominant Logic for Marketing. *Journal of Marketing*, 68(1), pp.1-17.

Vargo, S. and Lusch, R. (2006). Service-dominant logic: reactions, reflections and refinements. *Marketing Theory*, 6(3), pp.281-288.

Vitarta, P. (2017). *Using Learning Analytics to Guide and Manage Learning in Large Classes*. Copenhagen Business School, Copenhagen, Denmark.

Wales, M. (2017). *Front-End vs Back-End vs Full Stack Web Developers*. [online] Udacity. Available at: https://blog.udacity.com/2014/12/front-end-vs-back-end-vs-full-stack-web-developers.html [Accessed 8 May 2018].

Willcocks, L. and Lacity, M. (2016). *Service Automation: Robots and the Future of Work*. Steve Brookes Publishing.

Yin, R.K. (2013). *Case study research: Design and methods.* Sage publications.

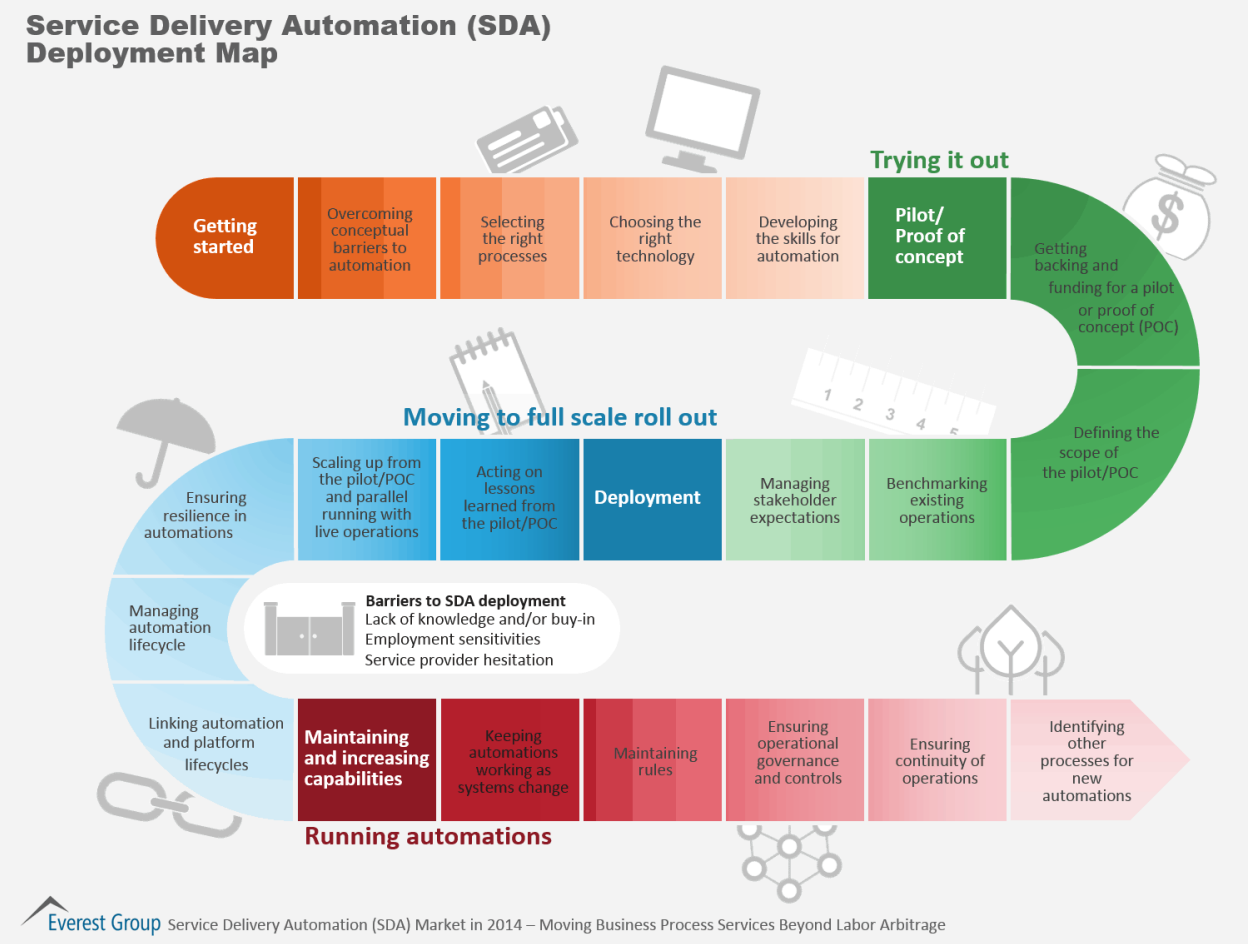
Yonck, R. (2017). *Heart of the Machine: Our Future in a World of Artificial Emotional Intelligence*. Arcade.

APPENDICES

Appendix 1- Glossary and acronym directory

Acronym	Definition
AES	Automated Essay Scoring
AI	Artificial Intelligence
AIEd	Artificial Intelligence in Education
BID	Business Intelligence and Development
CAA	Computer-Assisted Assessment
CBS	Copenhagen Business School
CC	Cognitive Computing
CCSS	Common Core State Standards
ECTS	European Credit Transfer System
ETS	Educationa Testing Service
FP	Fundamental Premises
GDL	Goods Dominant Logic
GMAT	Graduate Management Admission Test
GPA	Grade Point Average
IAC	Intelligent Automation Continuum
IoT	Internet of Things
KAT	Knowledge Analysis Technologies
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
ML	Machine Learning
MOOC	Massive Open Online Courses
NLP	Natural Language Processing
PEG	Project Essay Grade
RPA	Robotic Process Automation
SDA	Service Delivery Automation
SDL	Service Dominant Logic
SOL	Standards of Learning
TEO	Taxonomy of Educational Objectives
TOEFL iBT	Test of English as a Foreign Language
VPD	Value Proposition Design
WWW	World Wide Web

Appendix 2: Service delivery automation deployment map



Appendix 3: Interviews with AI experts

A.Peter Vitartas

Us- Intro to project

It is interesting that you are taking your project in the students point of view. In Australia I don't think the students, maybe because we use rubrics and we provide students with as much information as possible about what the marking criteria is and try to stick to this criteria as much as possible. Maybe it is a little bit different between the two countries.

What is your current role now, are you still researching this topic?

At the point I just received another grant, leaching and learning grant, to continue the work we are doing but it is interesting because we developed our system and we applied it and tested it last semester, last year, and it is been very successful in terms of providing students with some feedback. What we did for this year we sought some research funds to develop the program a little bit further, mainly with the idea to support students with their writing. We have gone away from the idea of marking to more idea of trying to support students in improving their writing skills. This was mainly because the marking has still a fair bit of work to develop this and we will continue developing the marking side but what we want to do is focus more on helping the students and providing feedback. Because in Australia, we have very large classes with over 1000+ students in the same subject in the same semester. So they may not be in the same class but they may do it across multiple locations, online, different ways. Some subjects are 2000 students. As you get larger and larger with your class, what happens is there is less and less interaction with the students and the professor. What we want to do is to be able to supplement that support and provide students with better feedback, especially in the development and preparation in their assignments. We are looking at how we can improve and support the students and provide more formative feedback for their assignments. We can do that quite effectively with our program and we are trailing it on a number of subjects. The difficulty we face is that every time we change the subject, you have to reprogram the software. Which is not difficult because we got it set up that it can do that, but it does require a fair bit of time and effort to do that. We are working on it. We want to develop schemas for each subject so that we can say ok if you look at this particular words or terms and concepts,

they are more apparent in this particular subject than say another subject. We are starting to develop algorithms for those particular different subjects. At the moment we have one for marketing and sociology. But each time there's a different assignment, we need to set it up and train the program. One other thing doing this year, how different software/programs can develop different algorithms. We are trying to refine the algorithm process.

Currently at this conference https://latte-analytics.sydney.edu.au/ . Talks about these types of programs. There's work being done in the US, through Milincamp, open university in UK, UTS (sydney) and they all have their own programs doing different things. UTS is working with program to help support research students to develop their research writing skills. So the feedback side of this AI is actually being used a lot. The open university program, they are doing feedback to students as well.

Seems like there is two sides to this 1)marking side- being consistent because when we put two markers beside each other, they invariably come up with different marks. The AI systems can solve that problem 2)Feedback side.

Early in the development stage

How did you get involved with AI?

We have a business analytics group and I talked to them and I basically put the grant application together with them for our internal uni grant. That provided funding for some PhD students to help develop the program. They have expertise in analytics and AI and developing these algorithms and computer programs. My background is education and higher education so it was a combination of these two things. For any student who is going into this, the computer science people can be great at developing the programs, etc but they need a reason to do this stuff. They need to team up with other people to develop these projects. I don't have a technical background- i can analyze the results and the data that we are getting but we need the programs to pull in all the assignments and data.

Did you look at other programs when you were developing your own, or strictly on what you wanted specifically?

In this space there is a lot of open source software which means it is available for computer programmers to download other peoples programs. For example, in python, there's a thing called NLP that has their own special programs that are available within python. Other people have open source software as well, some of it is purchased and some of it is not that expensive either, and some people are sharing ID's and programs. This area and space is developing quite a lot as well.

When you use it for different courses, is it the professors telling the computer softwares how to adapt it to that course or who is actually changing the software?
The way we have established it is that we start from the bottom up. We start with the rubric and say what is it that is required in the assignment. Generally, in most assignments there are some common requirements in terms of grammar and in terms of what the professor is trying to get the students to do. Then it is about breaking that down and seeing what common programs we can use for that, and then what programs we need to develop specifically all that assignment. That's where the algorithm is used to identify key terms or phrases or particular wording. That then provides us with a good idea of how well the student has answered the question or is talking about relevant concepts that are being covered in the assignment. At this stage, we haven't perfected it so that the computer can do all the marking, but we have got it sufficiently well to be able to group assignments such as saying these are good assignments, these are not so good assignments. The computer program was able to predict 80% of the grades for the assignment- which is pretty good.

Can the software also detect critical thinking with open assignments or is it more these students listed four out of five points, or how does it work this way?
We have critical thinking as one of our key areas that we look at so that is based on research of critical thinking and what is critical thinking and understanding how critical thinking is described. We are able to seperate it out by saying these assignments are good in terms of critical thinking, or it says it is higher with critical thinking in other examples or evidence of critical thinking. So critical thinking is an important characteristic that we try to get students to develop this as part of their learning and that is expected, so there should be evidence in their assignments of some aspects of a reflection, commentary, or opinion that comes across in the assignment.

The other one is around research so we look at what the student has undertaken in terms of their research, this is usually evidence from their bibliography or through their reference list, that type of

thing. That gives us an indication of how well their research has been. It is not only counting the references, it is also understanding whether the references are good or not.

How did you start the process? In terms of proposing it to your university or what was the road map that you went through?

It was about putting together a proposal and with that proposal we put that to our learning and teaching department, which is centralized, and each year they have grants available. So we put in for an initial grant and then as a result of that grant we got some funding which allowed us to employ some programmers and also a research assistant. We started with the literature to see what has already been done and then we started building on that. The first step was to develop the framework of what we wanted (in slides), five different areas. From those areas, we then worked out that we wanted to use the rubric and we brainstormed how academics go about marking and what they look at, and we used the rubric as the basis. We then broke that down into each of the components and said let's get some programming done for each of these different areas. That then formed the basis for our program. We started putting it together and then said how do we then present our results back to the students and back to the academic. That's when we started to build an interface, a web based interface with a dashboard, and then it becomes a computer process to try to get students to be able to access it and get the results back from the programming.

When you wrote the proposal, was it more looking at giving students feedback or saving time for professors? What was your main category that you pitched ot the learning department?

Initially it was about the marking side of things and we did it as an exploratory type of project. This was the main driver for the first grant application. The second grant application, was around trying to improve the process and to take a proof of concept (first set up), and try to refine that and improve and develop it. And now we are at the stage where we are taking it to the next level, moving from marking to feedback because we identified this as a value as well. At the same time, we are continually refining our algorithms and computer programs to make it more sophisticated as well to make it be able to inform the professors better as well.

How is the acceptance by the professors and students for these solutions? Their reactions of this technology.

Their reaction is always very positive. The students like the feedback and the idea that they can get feedback immediately and quickly. They like the idea that they can test and see if they can improve their results so that they can try to make improvements. They like if they get these improvements, they can get a higher mark. Students always after higher mark.

Some students don't see value in it, as they think they haven't gotten any improvement, they might be able to be good students or they haven't invested the time to use the software program. There's always going to those people for and against.

The professors like it because they like the idea of the consistency of marking but also they are hopeful they don't have to do any marking. Of course that is still a way off, as we also will have to do the marking but for them it gives them something to check against to see how their feedback or grade aligns with the computer program and gives them confidence as well that they are on track.

From my point of view, as I deal with a lot of markers, i can see how each of them are performing. If one looks a bit wrong, I can go back to the assignments and I can talk to the marker and say this seems to be different, or your marking is higher or lower than the rest of the group and we can talk about why and look into the results. One example was that one marker was relying more on the research side and less on the written style, we were looking for a balance of both. Another marker was looking at style and less interested in research. You can detect the natural bias that they may have. Then you can correct for that or at least have a discussion about that.

Do professors think it is more work that they need to use these softwares to detect what they really are looking for in the assignments?

I think we have been working with people who are very positive and really want to improve systems and help with the development of this process. We tend to work with those that are more technical and are interested in making these developments. We are staying away from the critics. I guess this is a natural thing because we want the positive support and want people who are interested in this. Those people are better at pathological practise so that they understand where we are coming from, how this can help them, so they are positive.

How many people are working on the project/software?

At the moment we have 7 people working on the project and make contributions on the project. This semester we are hoping to recruit 3 or 4. We will end up with about 5 different subjects that we are testing using this system. If we have a successful outcome this semester or year, and based on the results we are hoping that we can get more funding to continue the project and expanding even further.

How did you get involved with speaking to CBS?

I first had a personal contact there working in the credential office, as that's another part of my role, and she put me to head of marketing department. We corresponded and they put me to the information technology area, who were doing some work in this space in terms of analytics side of things. As a result, as i was on study leave and was travelling to europe, so I came to talk.

Who were you presenting to?

Small group of people because of the timing but I spoke to a lot of people individually. Some were very positive and a little okay that sounds interesting but they didn't want to do anymore with it.

I think if you are going to make recommendations, selecting people who are more likely to be more positive or innovative, they are the ones that will lead the development of this type of project along. This is probably true to most diffusions and innovations. They are the ones that are testing and working things out ,and then as things progress, better programs, then people will adopt it more often.

Any suggestions to give us or books?

There are opportunities for commercialization, opportunities for pedagogical input. Looking at what are the approaches. There are a lot of questions that you can ask as well around what is the best practice? How do you use this and bring it into the classroom? There is a lot of unanswered questions in this space as well.

A. Mark Shermis

How to implement these types of softwares?

Formal language is more along the lines of English and there are already existing software packages that will do most of what you are asking for and it will require us the creation of some **statistical models** and it may not provide most of the technical feedbacks that you might want.

How do these software work?

There is this **automated scoring component** which is the evaluation component that will take a look at the essay and go through a statistical algorithm and make a prediction to what score you might actually have. And it doesn't provide any qualitative feedback and it is more for **summative evaluations** like high-stakes assessments.

There are other systems that are called "**electronic portfolio systems"** where you would write an essay, the software would go through it and it would score it and rates you on several dimensions and then attempt to provide a **qualitative feedback.** Ex: "Amanda I noticed that this is the thesis of your essay, you appear to be making 3 points, you appear to have a lot of information on point 1 and 2, but nothing on the last point." What the system will not do is that they will not annotate and say you missed this point be able to say "There is a critical argument you didn't include". The software cannot really understand what you are writing about but it can draw on many essays that were used to formulate the statistical model and then provide feedback based on that and some of that it can be topical based but it depends on the statistical model and the software you choose

If a professor wants to use automated scoring to grade a specific course's exam does he have to train the algorithm on the previous exam?

That's what we call a **Prompt specific model:** → this would be good when this questions of the all exams and future exams will be similar. But if you will change the exam type you will have to recreate the model.

You can create a **generic model** that evaluates your abilities to write but it really will not have a lot to say on content as much as the prompt specific model. The generic model can be used forever.

You can use a generic model when you do not have enough essays for one prompt to create a prompt specific model so you can combine the responses of 4 prompts.
The generic model based on these prompts will be able to evaluate an exam that is based on those prompts but is based on **Narrative writing:** your ability writing level to address narrative writing and it will provide you feedbacks on narrative writing. Such a model will not however be able to tell you that you made a bad or weak argument on a particular topic because the topics are basically converged.

Do you think there is a specific subject that is easier to evaluate through automated scoring than any other type of exam?
We did some research a few years ago and our hypothesis was that essays had a lot of content associated with them and would be harder to grade than more open ended essays like: " What did you do last summer?" and it turned out that the technology worked better on content intense essays. And there is work going on by ETS in a stem area that really does address some of the arguments that people do. So this machine can't tell you if you made a great argument or if you made the right conclusion, what it can tell you is that there is a low probability that you made the wrong conclusion.

Which already available tool in the market do you think is best?
E-rater is the best at providing qualitative feedback, and it can look at the structure of your writing and it can at least analyse that and it has a lot of NLP features that allow it to provide you some feedback on various dimensions. If you want to get a rating on how you've addressed content, it will give you such a rate if you are using a prompt specific model, but that rating will not be that precise if you are using a generic model. The weakness of e-Rater is that it really doesn't do a great job on content per se. So if you look at the model, there are 10 elements that make up the model and only 2 of them are related to content.

For some kinds of writing content is not really important so if you are doing an argumentative writing , content isn't really part of your writing ability

Considering the diversity of the grading process across different university, would it be best to develop a customized software or implement one of the already available tools

KLightside is available for download. It uses a different approach than the other vendors that use NLP. It creates a vector that mimics quite well what NLP does but it doesn't differentiate. So once you create the statistical model with Lightside is that you are only getting a score. You could actually use this tool and add additional features to start developing your own solution.

If you want something that you can use right away, you can use Criterion and apply it not on high-stakes assessments, but on formative writing which is a kind of writing that you would do if you had to write a report and want to get feedback.

What are the missing features of this softwares to achieve the 100% agreement of human raters?

Now it is around 80%. You could create a:
GOLD STANDARD: to configure a model that addresses a particular prompt a a particular style of writing where you use no human ratings but a preconfigured weightings on your model.

What was your role as a GRE Technical Advisor at ETS
I was contributing with my expertise in writing assessments and at that time their software was used for the GRE exam that ETS administers.

Is there any exam/course that has used AES already?
GRE exam is one.In this type of exam they have one human rater and the machine rating each exam and if there is a difference in the grades they both provide, they have to send it to a third human adjudicator. Most of the times the grades are more or less the same.

The GMAT exam is using AES as well.

It is also used for several licensing exams. For example CPA: Certified Public Accountant

Six Sigma Designation certificate

The technology is used to evaluate writing performance.

What's the best way to address the development of a project like ours?

I would suggest you to conduct a Pilot project by finding a group of professors that are amenable on contributing to such a project, and then do it on a small range for example with a couple of classes. Business writing is terrific for this technology.

1. You should start with a conference where the technology will be explained to the participants and then you should find out where you would like to use it.
2. Apply the solution firstly on lower stakes exams, for example formative assessments.
3. Allow for iteration and test within a small group of people don't roll it out until the system is perfect otherwise people will stop use it as soon as they see a flaw in it.

How did you become interested with AES?

I was working as a testing center director in Indiana and met up with Ellis Page who was "inventor" of AES and he and I worked on bringing AES to the world wide web and we used it for placement test. They worked very well on placement tests and were very good as human beings. I've been working on this technology since 1997- 96 and I really enjoy this field of research.

Do you see this coming more and more in the future? How long do you think it will take for it to be used it?

Some people will never use it. I think there will be a wide spread use of it in the next 5 years especially if Nape in the US accepts it as a grading technology. I think you will see a really good implementation in the next 15 years. People's expectations are shaped by the interaction with other kinds of technology so if you turn on your tv and the color of your tv is off you go mad because your expectation is that it is going to give you perfect color. The technology is improving at the same rate as our expectations are increasing

C.CEO of Peergrade

How did you come up with this idea?

I am a PHD student at DTU in machine Learning and I was working there as a teacher as well and I wanted to run my own course and convinced someone to do that and then I had this small class of 30 students on ML at DTU and the it got the department exited because it was on AI and ML so I decided to also include Big Data and make it as an official course and then 150 people signed up instead of 30. Then I way I decided to create 6 open ended assignments that students had to hand it throughout the course and that made up the entire grade for the course.

I've taken some courses on coursera where they had 5000 students and were the grading was done through AES or peer grading. So to grade my course's assignment I decided to build my own peer review platform where students rate each other and give feedback. I ruled it out this idea on my class and then I built the team.

For me it was more of a technical issue: how do you ensure the grading quality of students? Can you make any system that could make sure that students are giving fair grades and feedback on each other? So I did a research on Eurovision Song Contest and see how they could find which countries were cheating and favoring each other and you can use the same model in Peergrade.

How do you deal then with this problem of cheating?

Let's take the Eurovision Example where every song has some kind of "quality" and then you see how many points does every company get on average. So let's say that Cyprus get 4 points on average but then Greece gets Cyprus 12 points so then obviously there is some kind of bias between Cyprus and Greece and then you calculate generally how accurate are the different countries, some of them are always off the charts and then when you know how good the countries are (ex when Denmark always rates everyone according to the average) then you can make a better guess at the song quality and what's the trustworthiness of the countries.

So to provide the final grade on Peergrade do you look at the median or at the average of the grade?

With Peergrade we realized that we can make a lot of cool mathematical models to make the grade more accurate but people didn't like it because no one was understanding the technique. Let's take the example of Self Driving Cars: they are always more accurate and precise than people driving but when someone is going to be bitten by a self-driving car, this person gets really mad thinking that it was a robot that did it and not a human . So people are happy when they get a good average but when they fail they are super angry and this is the same with throwing some mathematical models, it might be better on average but that one student that gets tricked by the robot, he will get very mad.
So we took out a lot of smart features from Peergrade.

What was the reaction of your students of having to be graded by other peers?
At the beginning they were really skeptical of the quality of others feedback. So we built this feature of "Flagging " each other. When a students doesn't like or understand a feedback that a peer gave to him he can flag it and that feedback is sent to the professor and reviewed by him. —> this builds additional trust on the students. For my exam I had to only check 10 % of the feedback.

Are then the students graded according to the feedback coming from Peergrade?
Yes. It was possible for me to do that because DTU is a very flexible university. The grade was hence 70% based on the result of the feedback that each student got, and 30 % on the quality of the feedback this student gave to other peers. (Students were also rating each other's feedback). And I check every flags.
We have another system: it is called "**flagbot"** that is a system that automatically flags feedback to the teacher. So if a student gets too nice or too bad feeback, and gets a feedback that is very distant from the others, the system automatically flags it.

When do you review the flags do you think the feedback are good?
Usually it is very accurate. If people know they get spot checked then they tend to behave fairly. So with this flag system they have more incentives to give good feedbacks.

Aren't you curious of what your students write on their assignments?
I have a course where I had some relatively simple closed problems which are easier for peers to evaluate, and then I have some open ended problems where there is not a single one right solution

(where students had to build a machine learning model and test how good it is), which made the students compete on the best model and then there is a price for the winner. Those tasks I find it very interesting to read because usually the top 10% of the students are actually better than me so I see a lot of good work where I could learn from and students can do that as well and then I usually publish the ones with the best solution

How did you become interested in this?

I started programming when I was a kid so by high school I was already very good on that so at university instead of starting coding courses I decided to go to math courses and took the advance courses. At the time when I was finishing my bachelor Coursera was launched and there was a big course in Machine Learning and I took that course over summer and realize that that was what I wanted to do. So when he started his bachelor he found a way to skip that course as he already took the summer course.

What do you feel about a system that automatically grades students?

I think it is going to work at some point but it will be there in the future. In 5-10 years it will be up to a level where you can have quite decent feedback. Until then I think a lot of p companies have been very successful with this thing called "human-computer symbiosis" where people are good at something at computers are good at some other things and if you merge them together you get a really effective work and that's why I think peer grade has.

We cannot replace teachers and we do not want that to happen, but students can learn from each other and do some of the review work and the system can work on the automation of the activity like flagging. Then the teacher can come in a spend time on only 10% of the grading work (ex when students disagree)

Does the professor have to input something at the beginning?

You set up a feedback rubric which is a sort of guideline that peers will follow to give feedback and some people do that very open handed like completely text based. —> on this types of rubrics is very hard to do machine learning.

While some people create very numeric rubrics (ex on a scale from 1 to 5). The really good rubrics are a sort of mix: how good is this introduction and they explain on a scale from 1 to 5 what each score actually means.

We think that the use of rubric is what really makes students learning from reviewing.

Is it easy to build a rubric?
It is as easy as creating a google form.

Have you tested the grading method you adopted for the course you were telling us about on other courses? (Using Peergrade to provide students with their final grades)

We have many users around the world and there is an economy course at the University of Copenhagen that has been using Peergrade for a few years. Students of this course have to submit 5 assignments per semester which are graded through Peergrade and then there is an unknown component at the exam as well. And to qualify to the exam you have to have provided good feedbacks. In this class there are around 200 students. They get the people that gave the worst feedback and do not qualify them and then they look at the second worst and say that that is actually good enough to qualify for the exam and then the rest gets automatically qualified. This motivates students to provide good quality feedback and also does not scare them that they have to be graded finally by Peergrade.

Do you pitch your solution to university or do they come to you?
It is a mix of both, now that we are more established we also have many institutions directly approaching us.

Who is paying for this solution?
We do not want nor students nor professors to pay for this solution, it is the institution that pays for it.

Does the price go by course or by program?

So for the universities in Denmark we are selling the product for the whole institutes. Also at CBS, they are trying to roll it out but they are a little bit slow.

I think the biggest misunderstanding about Peergrade is people always see it as a replacement of professors. It is not a replacement, it is another thing you can do. We are not pitching Peergrade such as “have a better weekend and click on the Peergrade button”. That’s not the idea. It is the idea that you have a lot of different classes in schools and universities where there are some things you would love to do but you can’t do it because you don’t have the resources. Now you can do that! You can have an assignment every week. That was the biggest problem I saw, at least when I talked to CBS, they said we have all the students that go to the exam and they haven’t learned anything. But it is way too late and we don’t have the money to introduce assessments during the course because we don’t have money to grade it. Also don’t want to have students submit things if it will not be reviewed by anybody.

What is your pricing model?

It depends upon how large is the institution and how did they get on. It is very open and we are just releasing a new pricing model today. It doesn’t scale linearly, as if a really big institution wants to go, there is discounts. Because if you multiply any number by 40,000, such as University of Copenhagen, you get a very large number. The very first invoice for Peergrade was my own course and my supervisor found it and wanted to use it so I added two users to the system. Then he said you should totally sell it to the university. He took me by the hand and took me to the department head. I came up with a number of 800$ a course, per semester and take 2 TA less in my course. They made a deal as they were saving 2 TA, say 4000$ per semester per course. So two courses was 16,000$ per course. I hired a friend to be my TA, now CTO, to help build Peergrade. Back then there was a clear economical incentive for him, as if he can replace one TA per year, then it is worth the entire department. If you actually use it for replacement, there would be massive time reductions. Many people do not, but most important thing is that students learn a lot and teachers don’t get too stressed and spend their time on something else. If students spend 30-40 minutes per paper, you get 2 hours of focused attention on your paper. They might not be professors, but they will clearly give more feedback to students than I can. Then I can spend that 30 hours a week on something else.

Systems like universities are slow and scared and conservative. We can slowly see that it is loosening up in some ways. People are finding holes in the system, where they can put Peergrade in. As I am not actually grading them, but I am officially giving the grade because that's the only way I can fit it into the legal requirements. University of Copenhagen has a law school and they are sceptical about everything. They found ways because it is not grading as it is a pass/fail requirement, and then they snuck peergrade in behind the curtains with these pass/fail requirements. Suddenly its becoming widespread and adopted but it is a big process.

What was you roadmap in developing the service?

There was no roadmap. We launched September 1, when my course started and there was nothing that worked. You could only login. So we had one week to build submission (when first submission was to be handed in). And then when that was working the day before, we had all students sitting next to us so they would all come and complain and tell us what wasn't working. We took that and fixed it and then we put a gap of 3 days from submission to giving feedback. So we gave ourselves 3 days to build the feedback part, work was intense, building it as we needed it.

Now, we have been working on it for 2 years now and kind of realizing that we need to start over. It is a little bit messy to say the least. But we learned a lot about from the early days. What do the students actually think about this. Testing with our own students, it is always a good idea to solve your own problems because then your intuition is probably okay. So if I want that feature, it is probably not the best feature but its not the worst. So when we have middle school teachers teaching 9 year olds with peergrade, that's clearly a different world- it doesn't necessarily work for them.

We moved from DTU to Uni of Cop to sit at the economics department. Then we sat at CSE at CBS to talk to professors, universities, and users which was really valuable.

We came with this cool idea and perspective that we need to put a lot of machine learning in here and people reacted surprisingly negative towards that. They didn't want that but they wanted transparency. It doesn't matter if they get feedback, it is much better to get it from the professor than the robot. So what we did was make a lot of machine learning, find the problems, and ask the teacher to verify it. So for example Rasmus gave Morten this feedback, we think it should be this

and instead of that. The teacher can click yes or no and if they click yes, the feedback gets sent to the student, but coming from the teacher and not the student. This seems more valid from students perspective. But we make it as easy for the professor as possible.

That's the way to get machine learning into the system but without scaring people too much.

Universities think the company is very young. As it is two years, different perceptions. We have 2000 institutions around the world. Schools are not the fastest moving creatures in the world.

Any suggestions on how we can make our road map?

I think it is interesting to look at, if you assume AI will go from where it is not to a point where it will be able to replace everything a teacher does. Then there is a curve of quality along the way, as it gets better and better and better until it reaches that point. What's the right approach along that curve? What's the right approach for when you have no AI at all, manual grading, and when the AI is better than the teacher, you use AES clearly. I think at least, but maybe you don't. Maybe it is not about quality maybe it is about human interaction. Then what do you do when its good but not good enough. What's the roadmap for the universities. Maybe there is complete teacher grading, maybe complete robot grading, or maybe it never gets there. Maybe there is this human interaction where it helps teachers be more efficient but it doesn't replace them. For us, we have this flagbot that tells you what to look at and you as a professor can look at it. Maybe the next step for that is that it recommends the specific thing you should be doing, like should I click this for you. And maybe the next step is that it gives automated feedback, where you can check it and then robot takes over.

Think it is interesting to plot out the different steps of automation. Then I think teachers are scared of AI so everyone should be scared! But I think it is also about what do people really come to university for. Is it feedback on a piece of paper or is it study with other people they get to know, is it interaction with the teacher, and these massive online courses have been pretty dramatically blown up to be big things. Assume they become more popular, what becomes the role of the university. My perspective of that is the most important part of university is the people there, the network they get form students, and mentoring. Like one to one mentoring with a teacher is formative. But this is not happening anywhere except if you go to PhD programs. Also there are

some universities that are rich, like Oxford and Harvard, where classes could be 4 students, then it is basically mentoring. This is one of the reasons they can be so good.

What if technology can help take away the easy parts? I don't need to learn about linear algebra from a professor at DTU. I can go to coursera, from best professor in the world, do at home. Then I can go to DTU and they don't need to do the same lecture all the time over and over again. I can go to DTU and talk to my professors. This is one part of the way towards the end- I don't know exactly.

It is going to be awhile and the question is will we ever get complete automation. What if we get half way, what do we do then, can we do something still cool.

Big question is what is the incentive structure of university. I am not getting paid to teach, it is all about research. This is a whole other problem that is broken. This is why if you start using Peergrade to replace TA, etc, its a lot of time to save and this time will be done on research.

I would hate it if we don't go to lectures anymore, do it all online, grade each others work or get it AES, and we never see a teacher. How can we give teachers time to do this? In my course we did lectures for 10 minutes, we will do peer review, not going to read the papers, and the rest 3 hours and 50 minutes is group work. I will be there with the TA and will walk around and help people. Every student has the option to book me on Skype if they had an questions at any point. The very good students called me a lot worrying about the grades. I could have done more but I wouldn't have taken any skype calls if I still had to do the grading. If professors can free up on grading time, they can give this to office hours- one of the most valuable.

I think you made a good point about when you get the feedback. As if you get it too late, why would you care?

I know at CBS there is this internship thing, and I know there's at least been talk about and set in practice, that students write this internship report. In the middle of it they get feedback and can continue on it. It is a good case because you cannot plagiarize someone's internship so it is okay to see other peoples work and get inspired by others. For CBS purposes, people get started a lot earlier as they have an incentive to have the feedback.

D. Cofounder of Collektive

Collektive start up- One project going right now with machine learning where several companies can train the same algorithm without sharing the data with everyone. One thing with this project is to predict stress for employees. Having one company on board is probably not enough data, therefore better to have more companies on board together to have more data and better model of stress indicators and create a better solution. Working with another startup that has an app. Pension and insurance companies that can work together on single algorithm but also algorithm that you can put on phone so that you don't have to connect personal data in a central place- helps prevent privacy concerns. AI is kind of this big umbrella turning that covers anything. A lot of ethical questions. Machine learning is very specific in algorithms that basically uses statistics of big data set to learn something

Writing style- needs to know about the curriculum of the course. Need to specialize- we look at this specific course or something and then depending on the course you choose, very writing heavy, there are these branches of machine learning one called natural language processing. There you can do a lot of work on sensitometer analysis where it is basically trying to see if sentence is positive or negatives- used to analyze reviews

Access to previous exams is a good baseline. Years back of exams would be good to have as more data. Depends on the exams and structure. Look at more theoretical based exam where professor has specifics for the response. Simple thing you could do is, assuming text based, bag of words, that counts the words and you get a list of counts for each document. Then you apply machine learning with this. IBM and Microsoft has these standard frameworks that should be easy way to set up. Talk to professors and see what they say. If the algorithm can detect like these ones are definitely good a 10 or 12 but these ones might miss something- another way to see the problem. This may cut 20 percent of their work

Do students mention this word in the same paragraph as this word- simple rules like that to create a simpler model

Machine learning basic thing- say you have some documents, in order to do machine learning you need to convert this document into some kind of list of numbers vector because that's how machine

learning algorithms work. One way to do this, standard way, you simply count the words, say there is critical as one word, and then this creates a dictionary for all words that you have and then you count such as this word has been used 10 times, 40 times, etc. For each document you then have a list of numbers, basically comes as a matrix of numbers, with words on the vertical axis and document on the horizontal axis. This is the training set. Then you input this into machine learning algorithm, which could be Microsoft service. You can look up Naïve Bayes Classifier. You would also have another matrix with all the grades, for each document you have a grade

Perhaps you can do simpler one, saying pass or not pass (binary classification). Simplest thing to test and see if it works, have full set up and create more and restrict specific words for those references to specific things

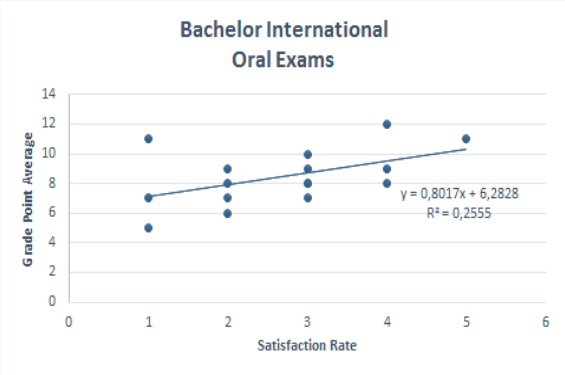
Small thing up and running on own machine that helps. Different algorithms for different courses- different words for the texts about the subjects. Always the problem with this- as you want to specialize with these data sets, it shrinks but the more data you have is better. Tradeoff to think about. Will become very complex with all the rules that you are looking for- first ask professors how they grade

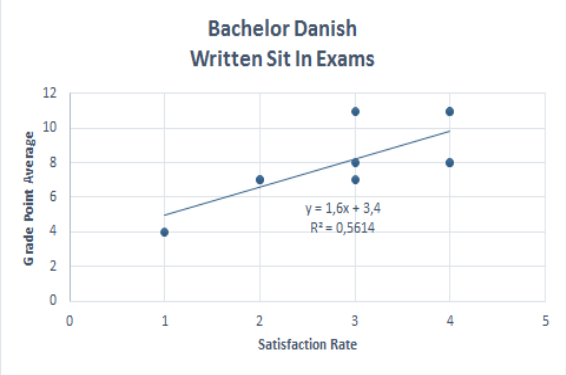
Appendix 4: Business intelligence and development data

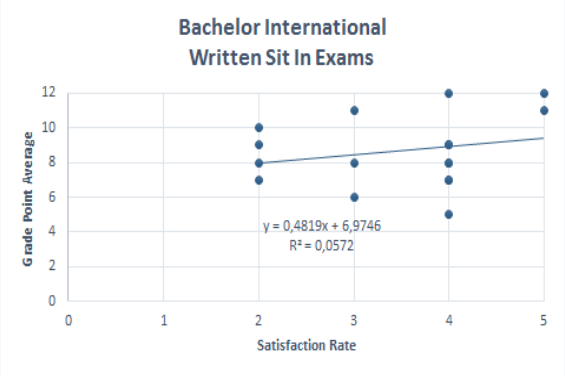
Study Program	Grant 2017 (Million DKK)	Total Consumption 2017 (Million DKK)	Of which teaching (Million DKK)	Of which operation (e.g. printing, transport, etc) (Million DKK)	Difference than grant provided	STÅ 2017	Costs per STÅ
Total of all studies	**207,4**	**189,5**	**184,2**	**5,3**	**17,959,954**	**12,567**	**16,507**
ASP	3,9	3,3	3,3	0,1	524,491	205	18,851
BAEUB	6,1	3,8	3,7	0,1	2,271,474	235	25,757
BIN	2,1	1,7	1,6	0,1	448,577	111	19,040
BSc BLC	5,9	5,7	5,5	0,2	181,870	331	17,924
BSc SEM	5,5	4,3	4,1	0,2	1,148,716	346	15,834
EOK	4,4	4,0	3,9	0,1	396,985	217	20,394
HA	22,4	21,3	20,5	0,7	1,143,723	1,723	13,004
HA EB	0,2	0,9	0,8	0,0	-652,090		
HA FIL	3,1	2,7	2,6	0,1	400,764	143	21,250
HA Shipping	2,0	1,9	1,7	0,2	67,299	99	20,000
HA IT	3,9	3,4	3,2	0,2	424,093	220	17,596
HA MAT	3,7	3,6	3,5	0,1	96,443	189	19,521
HA Pro	3,1	3,2	3,0	0,1	-36,307	172	18,263
HA SOC	2,6	2,5	2,4	0,1	174,035	129	20,578
HAI	5,4	4,8	4,7	0,2	561,578	411	13,151
HAK	4,5	4,6	4,4	0,2	-56,737	314	14,401
HAP	4,7	4,0	3,8	0,3	625,397	314	14,841
HJ	5,5	4,8	4,6	0,2	772,333	423	13,079
IMK	7,5	7,3	7,1	0,2	232,137	406	18,532
POL	4,4	3,8	3,7	0,1	636,755	259	17,090
Total Bachelor	100,9	91,5	88,1	3,5	9,362,535	6,244	16,161

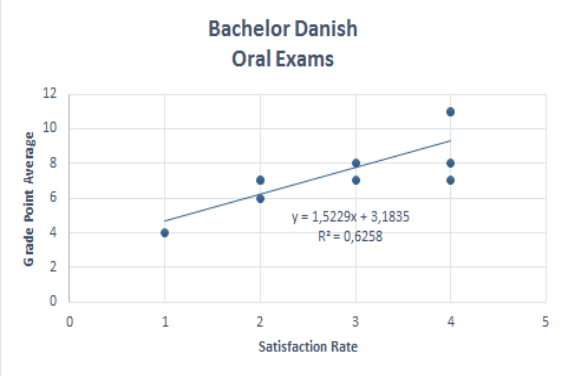
Cand Oecon	1,8	1,7	1,6	0,0	151,864	75	24,362
Cand Soc	8,4	8,4	8,3	0,1	48,888	487	17,325
CLM Samlet	1,9	1,5	1,5	0,0	403,734	110	17,013
CM IHC	1,3	1,0	1,0	0,0	297,854	43	30,716
CM IT	3,2	3,2	3,2	0,0	-42,634	178	17,842
CM POL	3,3	3,1	3,1	0,0	227,365	161	20,453
CM Samlet	48,2	42,4	41,6	0,7	5,879,908	3,096	15,581
CMA	6,3	5,3	5,2	0,1	940,235	404	15,540
CM Bio	0,9	0,8	0,8	0,0	85,821	26	35,018
CMF	1,8	1,4	1,3	0,1	408,254	63	28,700
CMJ	3,5	3,3	3,2	0,0	213,365	239	14,612
CMK	4,8	4,8	4,7	0,0	16,030	359	13,388
CM MAT	1,7	1,8	1,8	0,0	-97,876	76	22,089
CMP	2,3	2,5	2,4	0,0	-133,864	151	15,463
MA IMS	0,3	0,8	0,8	0,0	-476,732		
MAIBC	5,7	5,1	5,0	0,1	551,083	314	18,001
MCO		0,1	0,1		-146,543		
MSc BLC	4,8	4,9	4,9	0,0	-137,054	290	16,443
Msc EBUSS	5,0	4,9	4,9	0,0	104,755	251	19,829
CEMS	1,4	1,1	0,7	0,4	302,966		
Total Masters	106,5	97,9	96,1	1,9	8,597,419	6,323	16,849

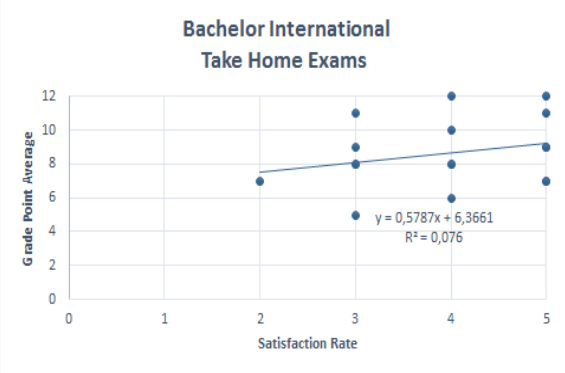
Appendix 5: Scatter plots for questionnaire results

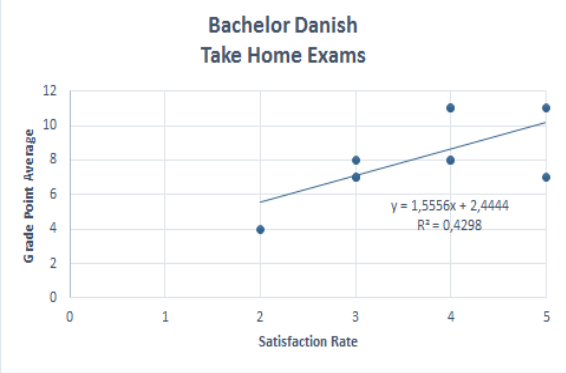


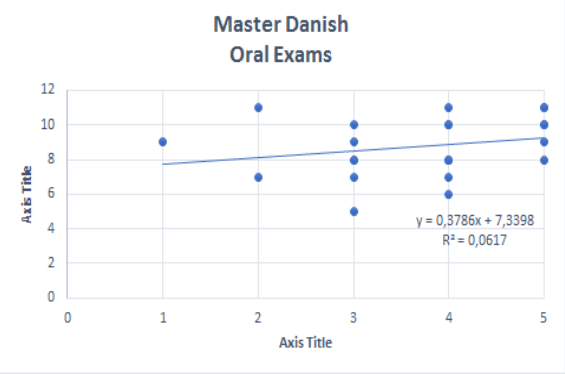


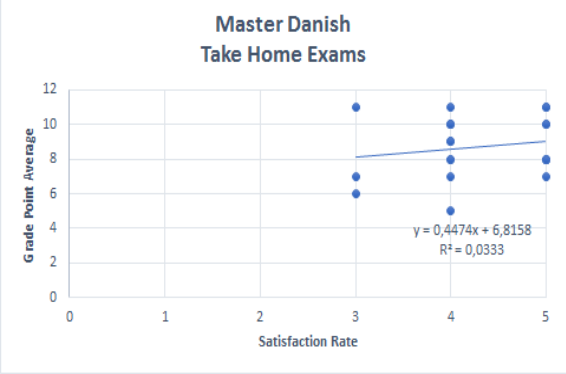


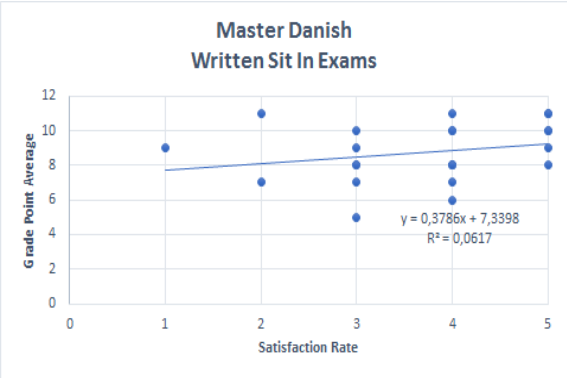


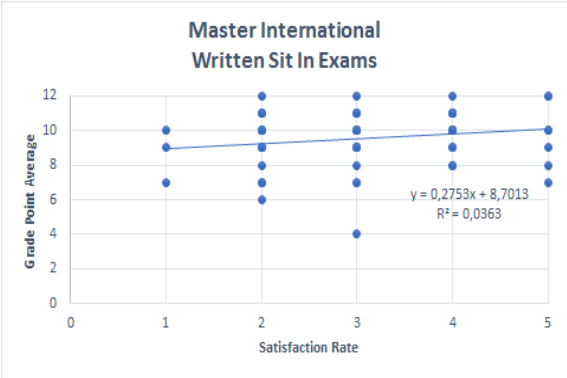


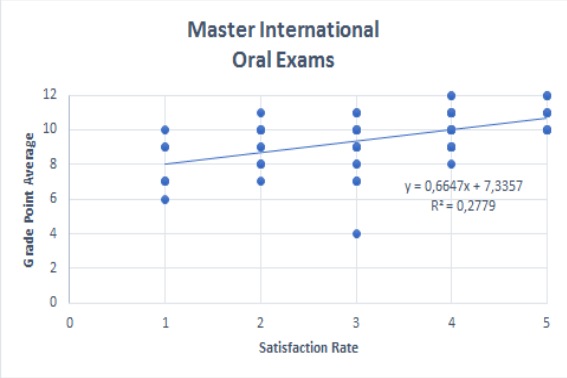


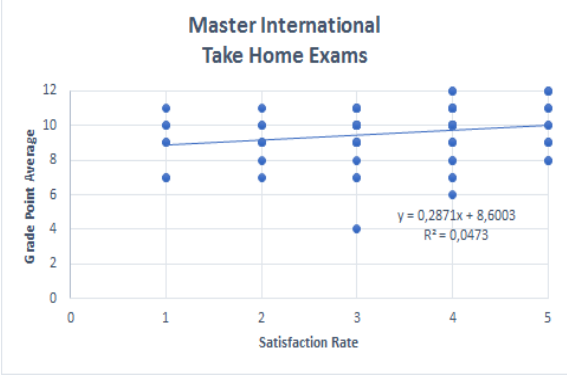





Master Danish
Written Sit In Exams
Grade Point Average
Satisfaction Rate
y = 0,3786x + 7,3398
R² = 0,0617


Master International
Written Sit In Exams
Grade Point Average
Satisfaction Rate
y = 0,2753x + 8,7013
R² = 0,0363


Master International
Oral Exams
Grade Point Average
Satisfaction Rate
y = 0,6647x + 7,3357
R² = 0,2779


Master International
Take Home Exams
Grade Point Average
Satisfaction Rate
y = 0,2871x + 8,6003
R² = 0,0473

Pains	Gains	Jobs	Exam structure	Grading process	Insights	Thoughts on the automatic grading
Administration tasks fulfil a lot of time to coordinate with other departments	Being in classroom and interacting with students	A lot of time is taken away from researching due to a lot of other admin tasks and such	The exam has to cover the entire syllabus so that students have to have a complete study	Bullet key theories or arguments that should be addressed	Would like to see the final grade reflect in class participation, or midterm exam to provide feedback for the final exam	Would have to write a really good answer in order to compare students answers with it. Hard with case analysis as there is no right answer
Grading exams: very repetitive and not most interesting relationship to have with students	Academics is fascinating and inspiring job as it can align with interests	Clash between research and teaching- "The reason I am hired and the university builds great reputation is research"	I start with the questions. You reflect upon that there should be different levels of analysis in the question: there is a DESCRIPTIVE PART, A COMPARATIVE PART and an ANALYTICAL PART. I know when I grade for instance, that the foundation is in the descriptive part so that would provide the student with up to some kind of medium grade. Then I know if the student has been able to handle the comparative part that would give him a strong grade. If the student has been able to use the theories in relation to the analysis then we can talk about highest grades	Remind of learning objectives	Positively surprised with the Danish teaching style	I wouldn't trust a machine to do it as I would want to compare if the technology gave a 10, I would want to make sure it's a 10

Don't have time to grade and give individual feedback on all exams	Like the new technologies- using Peergrade and blended learning resources	Always give feedback but normally in course itself. Give an assignment and provide feedback to improve for final. Also include 3 blind reviews from peers to gain valuable feedback	I usually give students cases. To interpret the case the students have to use theory	Start reading and make notes (+ or -, circle things) to refer back to	Even though it's not communicated in an effective way, students can contact professors to receive feedback	A lot of opportunities with technology such as blended learning approach
To me it's problematic that there is sort of physical separation between the performance during the grading situation and the performance during the learning process	Writing your exam based on whatever you want more or less	150 - 180 exams to correct each time	We go through some theories from course x, the role of organizations and different authors etc… and then I go through a number of cases and then I ask for example if you had another institutional approach how would you analyze it? If you have a more organizational approach based on gender, strategy and structure, how would you then explain the different developments of that company. So, I have very open question but I basically ask for a theory and how to apply it. And this is because this course in the first semester of first year bachelor and is a kind of introduction to the whole program.	Look at ability to communicate clearly, but the structure of the argument is more than the way it is written	"Point where CBS needs to decide if they want to be a mass institution or a great institution. I think most of this AI would be implemented at universities that don't consider themselves as being the top 1% in the world. Like mass institutions that make a lot of money from that"	CBS isn't about efficiency and money exercise but needs to be a pathological argument that it improves learning

I don’t like to read 120 assignments	Getting paid to learn	Another professor from the same department has to read my exam text and say if it is understandable before I send it to student (It's a specific procedure of our department)	The questions are quite open ended and usually the question in known because there is not any topic but they have to apply a number of theories in the analysis. So there is no big reveal of the exam question.	If names are written, cannot unsee the names. Might have a student that does really well in the exam but I never seen them in class, and then start to get skeptical. Or if they are a 10 or a 12 and they are right between and I haven’t seen them in class, then I give them a 10	Part of CBS system is cause its national system and mass institution with relatively few staff	Problem I see if you have such an open exam analyzed by AI, you have to specify the learning objectives that the AI can understand. Would really limit professors to be open with their exams
Doesn't like midterm pass or fail assignments as a grade is very important for motivation and being appraised to let someone know they are doing well	Don't love or dislike grading, but like that it’s a good way to understand how the teaching has worked because you see what comes back	The structure of the exam has to accommodate	They can choose from available theories within the syllabus but there is no prompt in using anything specific. They can include original material in it if they wish and they can go more in details than someone else and that’s ok but the fact that it is original material versus something else does not change their grade, it does not give them bonus points	Start with questions and reflect upon the different parts of analysis: a descriptive part, a comparative part, and a analytical part	Don't look at how much time is allocated for grading exams, but assume get more than what is spent	Could work for specific exams such as law where there is a definite answer. Also, bachelor level exams that do basics

Don't get time allocated for feedback and if you openly tell students you'll provide in office feedback, you'll be taking about 5-10 hours	Freedom in career	The grade has to reflect not only the final result but also the effort that the student has put in solving the exam, whether he followed the right procedure	For oral exams, we don't have a question until the course is finished, the main oral exam that I've had for some years is called a "student conference" so they have a short abstract they have to produce based on a choice of questions and then a 10 mins presentation not just in front of their examiners but also in front of some of their peers.	Already design the course with the exam in mind. Very strict, especially as gives students feedback during the course, so expect a lot more from them at the final	Cannot give feedback because there is no facility to do so, but if students ask, its given (2-5%)	AI technology would definitely help in blended learning to assist in tasks such as: moderating a form, answering questions, moderating discussion, handing in online
Classes are so big don't know students' names	Being interactive in class to understand what the students are actually learning	With the Vice Dean of Education we decided to allocate 6 additional hours 2 weeks before an exam to have a lecture where students can ask me to solve exercises during the class and ask for doubts	Oral exams: For some of my courses students would get the questions some time in advance and in order to have them to calculate something they would have a small assignment to calculate in preparation of the exam.	Adhere to learning objectives with room for subjective interpretation	In DK, would prefer to see modest fees and good loan system to better fund the university to employ more staff	Spend more than 20% of time on admin tasks, why can't AI do some of this
Give too many options for retakes, makes professors life hard	I want to be fair towards the student on how he is graded	I want to have a fair grading process with my students that also look at the effort they put during the class and before the final end result		Critical thinking skill	Doesn't give feedback to students as they use against for appeals	May take up more time for me to input data into the system

Courses are so short at CBS, that cannot implement a midterm as it takes too much time from the course	Feeling proud and happy when I get the cahnce to meet the students in person during oral exams	I would like to grade in class participation like (10 - 20 %)		Go through exams until completely convinced what the quality is. If student between 7 and 10, due to Danish scale, need to look a lot carefully	Uses 1/3 of allocated time to grade papers- efficient and in business for long time	In courses like accounting, would be sensible to use AI
Giving no feedback in higher education is ridiculous and a big problem	I want to be part of the academic discourse	I want my students to be prepared beforehand on the case study that we discuss in class		Have to look at learning objectives	No midterm exams as don't want to give feedback when students don't get a concrete grade because they try way less than if it was graded	Not entirely fair for the student. There are things that these softwares won't be able to do. For example, let's pretend that I ask the student to talk about a specific theory in an essay and that the student knows the theory well but has a really strong opinion against it. In this case this student might write an essay that is arguing against this theory while at the same time explaining it. If I would have to correct such an exam I would be able to appreciate it because I would be able to understand his point of view and see that he knows the topic and give him a very good grade anyways. But a software might get that argumentation as a negative point and it

						would result in a poor grade for that student
Too little time to grade exams- 4 weeks not enough	I want my students to understand something complex and hence spend more time on the learning process	I don't want to spend my working days on taking decisions		Remain skeptical about peergrade as those that don't attend classes opinions wouldn't really be valid	Doesn't give feedback to students	In many qualitative courses, I don't really think that's actually an option. But I would love to try it but I still want the job in the field. I still want the lectures to be part of the assessments. But such a software could be a good help in those situations when I'm not sure about the whether to give a grade or not
Bureaucracy, we have to do a lot of nonsense things to do and that take a lot of our time	Want to be respectful towards my students	4 professor teaching in the the same course and then we all grade it but we will need to pick out some examples, test grade it and discuss the level and then say why we graded like this and not like that and then we can take our bunch and use our sort of common level. I give them the CBS definition of essay and tell them how they should structure it and address the different learning objectives.		Normally based on conceptual models and theoretical framework, applying to specific case. No correct answer	When you have to apply a new technology a lot of time has to be spent on making students understand the changes and the new procedures as they might be probably used to the old system	Well students become extremely concerned about the grade that they get, only at the moment of the examination and just before. Professors, on the other hand, are less concerned. Students study for the exams but they should study for life. I think that if we could minimize the time spent on grading and use that for learning, then it would be a very good opportunity. I think that most professor would

						welcome a situation where part of the grading process would be done by a machine
The grading scale is very bad. I think it's poor, unreflective and the jumps in the higher grades are too high		Grading needs to be a fair process		Bring a sheet of paper that outlines the Danish grading scale	I like to use new technologies, as I've been using the blended learning resources. We've been using a programme called Peergrade which has been very helpful: this was used in the internship program when students had to upload their interviews with their mentors on Peergrade and then all students had to correct 3 other interviews (assess them and grade them). It wasn't really an ordinary grade (it was only from 1 to 5)	I can tell you that having a software that is able to grade multi choices and having a system like peer grading, it would help us getting to read some of that. And that would be wonderful. But one of the reasons why we have oral exams also is that it is very often a learning experience, so the interaction creates something. Sometimes students don't talk at all with the professor and at the oral exam students have the opportunity of finally directly talking with the professor. Then you get a grade and you get explanations on why you get that grade → and there I don't really see the role of a machine or artificial intelligence so I would say, by all means, if we are going to have the right of decreasing the grading activity, then I would say of course.

Having to deal with students' complaints		I would like to have a more bottom up approach at CBS in terms of innovation, meaning that they should come from also us		Absolute process so that start fresh with every student, don't compare	With our new program that starts in august we are experimenting something new. We'll have the first year without grades and just pass or fail and what's very important for us is that it doesn't get easier for students to pass. We also have a contract with the Ministry that we should increase the study intensity for the student throughout the semester as students usually do not study during the semester and go crazy during the period before this exam. So we have this obligation to create more experiences and engagement throughout learning so what we are doing at the moment is to designing exams that (for exams like statistics, microeconomics, etc) - instead of having a 4 hour sit in written exams - we have 6 or 7 hand ins throughout the semester that is anyways within the	To be honest I don't have any opinion because I don't know how they work. I would use it if I would feel confident that it would work. And I would use it under one condition: students should always have an opportunity for a second opinion from the professor. I think that that should be still an option. I would prefer to take some time of control of it, at least in the beginning because I will take some time to implement a system like that but if I feel confident I would not have anything against it

					time allocated, it's not extra time. They have their lectures, exercises classes and then every once in a while, they have to sit in exercises classes and work on these exercises and then submit them and to pass the exam they have to submit a certain number of these exercises. Then we also have other types of courses where they do multiple choices exams and they have some peer grading technologies where the big point of it is that you learn from getting feedback and only when you have given 3 to other students you can get yours.	
Having to grade 100 students		I meet up with students that want to receive feedback		Start from the top and state these weaknesses and factors push it down	It depends on what you think feedback is. Some students think that feedback is something that somebody is teaching, but it also comes in class when we discuss cases which we prefer to do. The main issue here is that students don't read the cases beforehand. Feedback requires a student activity ahead so there's nothing here	AI can do a lot of analytical work already. If they can do that I would definitely try. I would be very intrigued on what it can do, but I can see how it can follow human thinking, but I cannot imagine exactly how it would do that.

					to give feedback on. I teach a subject where we use a lot of case studies and the main issue is that they don't have anything for us to give them feedback on!	
Staying at university is distracting because of all the meeting that we need to have		I want to make my students more clever		When I look for excellence I refer to the online definition of the grade and I try to pick the one that reflects the exam		No, because I think that the hard parts are the grey areas and I don't think that the software would be good at discussing those grey areas because you cannot talk to a software. What you do when you grade - once you've done like 50 students - you talk with the censor and figure out " I value that part, I value this one.." so you do a lot of things that you are probably not supposed to do, but you do think about how hard was the exam actually and nobody could do that question.. So the question was probably too long.. All of these kinds of things you cannot discuss with the computer

Preparing for exams take a lot of time		Have more money for feedback, not only after the exam but also before		When I don't know which grade to give I go back to the answer of the assignment and then you have to discuss this higher complexity. Then I go back to the learning aims and I look at whether the student is following all of them		I would say, as long as I have a censor, it doesn't make sense to invoke any more complication. We've already spent a lot of resources in Denmark compared to what everybody else does. There are no other countries in the world that do that.
Student evaluations have a negative impact in our work environment		I want to have the chance to see the students' progress ideas		First of all, concerning the questions I ask, the student should explain at least the theory, they should not reproduce it (and this is quite challenging for a first-year bachelor) then I can see if they got it as they have to explain it and give examples. Secondly, when I give them a case and they don't know which one of them it is and what the angle would be, can they use it for a analysis of a specific company and say something about that. And if they can do that, and there are only minor mistakes then they get a 12. But I know that they have been under a 4 hours time pressure and I also		

				take that into account that.		
I hate doing examinations		Among professors in the same course we have to agree on what are the learning objectives of a course		When I'm grading I have two piles, ones about which I'm pretty sure whether it is a 2 or 12, and the other one that I have to read once more. The reason for this is that again, in the good old days we had 6-7-8-9-10 and 13 so jumps were not very large with the previous system. But now btw 7 and 10 it's really a great jump in grading and I think that the result of this is that 7 could be both a good answer or a poor answer because it's good but not a 10 but it's poor but not a 4. That's why I really have to think about it. And then the guidelines are not really clear, what is a substantial fault? And if you read the guidelines they operate with many mistakes or few mistakes and substantial and not so important mistakes.		