Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Teemu Lehtinen

Bootstrapping Learning Analytics

Case: Aalto Online Learning

Master's Thesis
Espoo, November 22, 2017

Supervisor: Professor Lauri Malmi
Advisor: Ari Korhonen D.Sc. (Tech.)

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

Author:	Teemu Lehtinen		
Title:	Bootstrapping Learning Analytics Case: Aalto Online Learning		
Date:	November 22, 2017	**Pages:**	xii + 74
Major:	Computer Science	**Code:**	SCI3042
Supervisor:	Professor Lauri Malmi		
Advisor:	Ari Korhonen D.Sc. (Tech.)		

The digital transformation of learning brings forth data having unprecedented granularity and coverage of learning activity. The research area of Learning Analytics (LA) uses this data to understand and improve learning. The practice of LA is a cyclic process where learning data is collected from different sources and analytics is developed according to stakeholder objectives. Finally, current results are delivered that lead into action which improves learning and produces new data.

The goal of this thesis is to bootstrap LA in multiple courses that implement different weekly online learning activities. The term bootstrap underlines the aim to support continuity, further development, and expansion of LA. The research questions were: what learning data the courses currently instrument, and what LA objectives the course staff find most important.

This thesis conducts software engineering to construct an LA solution for the research case. Requirements are defined via examination of the case and interviews of the course staff. The developed solution enables real time access to learning data and possibility to integrate data from both Moodle and A-plus learning environments for joined analysis. Novel interactive visualizations are developed according to the user requirements.

The work in bootstrapping LA at course level lead to two general findings. First, the integration of learning data from multitude of sources is a common challenge that requires design. Second, teachers' initial LA objectives include aims to monitor expected progress, improve allocation of learning material, identify problematic areas in learning material, and improve interaction with learners.

Keywords: Learning Analytics, Educational Data Mining, Data Science, Information Visualization

Language: English

Aalto-yliopisto
Perustieteiden korkeakoulu
Tieto-, tietoliikenne-, ja informaatiotekniikan maisteriohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Teemu Lehtinen		
Työn nimi:	Oppimisanalytiikan käynnistäminen Tapaus: Aalto Online Learning		
Päiväys:	22. marraskuuta 2017	**Sivumäärä:**	xii + 74
Pääaine:	Computer Science	**Koodi:**	SCI3042
Valvoja:	Professori Lauri Malmi		
Ohjaaja:	Tohtori Ari Korhonen		

Opetuksen digitaalinen murros synnyttää ennennäkemättömän tarkkaa ja kattavaa tietoa oppimisaktiviteeteista. Oppimisanalytiikan (OA) tutkimusalue käyttää tätä aineistoa oppimisen ymmärtämiseen ja parantamiseen. OA:n soveltaminen käytäntöön on toistuva prosessi, jossa oppimisaineistoa kerätään erilaisista lähteistä ja analytiikkaa kehitetään omistajiensa tavoitteiden mukaisesti. Lopuksi tuotetaan ajantasaisia tuloksia, jotka johtavat toimintaan, joka parantaa oppimista ja tuottaa uutta aineistoa.

Tämän diplomityön tavoitteena on käynnistää OA usealla kurssilla, jotka toteuttavat erilaisia viikoittaisia verkko-oppimisen ratkaisuja. Käynnistäminen pyrkii elinvoimaiseen, kehittyvään ja laajenevaan analytiikkaan. Tutkimuskysymykset olivat, mitä dataa kurssit tällä hetkellä keräävät ja mitkä OA–tavoitteet ovat kurssihenkilökunnalle tärkeimpiä.

Työssä rakennetaan ohjelmistotuotannon keinoin OA–ratkaisu tutkittavalle tapaukselle. Ratkaisun vaatimukset määritellään tarkastelemalla tapausta ja haastattelemalla kurssien henkilökuntaa. Kehitetyn ratkaisun avulla aineisto on saatavilla reaaliaikaisesti. Lisäksi ratkaisu mahdollistaa aineiston yhdistämisen Moodle ja A-plus oppimisympäristöistä yhteistä analyysiä varten. Työssä suunnitellaan uusia interaktiivisia tiedon visualisointeja käyttäjävaatimusten mukaisesti.

Tutkimus OA:n käynnistämiseksi kurssitasolla tuotti kaksi yleistä tulosta. Ensiksi aineiston yhdistäminen eri lähteistä on tyypillinen haaste, joka vaatii suunnittelua. Toiseksi opettajien tavoitteita OA:ta aloittaessa ovat valvoa odotettua edistymistä, parantaa oppimateriaalin mitoitusta, tunnistaa ongelmakohtia oppimateriaalissa ja parantaa vuorovaikutusta opiskelijoiden kanssa.

Asiasanat:	Oppimisanalytiikka, Oppimisen tiedonlouhinta, Datatiede, Informaation visualisointi
Kieli:	Englanti

Acknowledgements

I am immensely grateful for the privilege of my journey to science. I want to thank professor Lauri Malmi and senior university lecturer Ari Korhonen for patient guidance and inspiring discussions. I have shared room and collaborated with many colleagues of whom I name Otto, Lassi, Teemu, Aleksi, Juha, Samuel, Tapio, Kerttu, Petri, Matias, Jaakko, Timi, Markku, Daniel, Mario, Petteri, and Tommi. Thank you and the other great people in Aalto University for your support.

I am also forever indebted, for the same years, to all my ex–neighbors at Vanajantie. My godchildren illuminate the future. I thank you and your families as well as other friends for inviting me in your lives.

My parents have enabled me, despite their odds, cost, worry, shame, and time, to reach anything I could wish for in my life. I am proud of my big brother and little sister whose footsteps I follow. Lastly, I thank my daily listening audience and feline overlords H. and H.

Espoo, November 22, 2017

Teemu Lehtinen

Abbreviations and Acronyms

AA	Academic Analytics. A research field that addresses institutional, national, and international goals to improve learning via analysis of data from educational sources. The thesis considers AA as a field included in LA.
A!OLE	Aalto Online Learning. A development project at Aalto University that seeks to pioneer online learning experiences to improve learning results and to share related knowledge and tools in the university.
API	Application Programming Interface. A definition of how computer programs or parts of them communicate with each other. Another program may exchange data with a service that defines an API to use it.
BI	Business Intelligence. The practice of analyzing data to help businesses make more informed business decisions.
CSS	Cascading Style Sheet. A language that describes presentation, such as color, font, border, or position, of elements in web documents.
CSV	Comma Separated Values. A simplistic file format to store tabular data.
DOM	Document Object Model. A programming API to access and modify elements in web documents.

ECTS	European Credit Transfer System. A standard credit unit of studies that was created to help international studies in Europe. Depending on the course and the student 1 ECTS is estimated to take 25–30 study hours.
EDM	Educational Data Mining. A research field that employs data mining methods to extract value from educational data sources in order to understand and improve learning. The thesis considers EDM as a field included in LA.
GNU	GNU's Not Unix. A project started in 1983 to create a free open–source operating system. The name is a recursive acronym.
GPL	GNU General Public License. A popular open–source software license that requires derivate work to use the same license.
HTTP	Hypertext Transfer Protocol. The definition and rules that enables the internet media and communication known as World Wide Web.
JSON	JavaScript Object Notation. A structured data format that is written in a subset of the JavaScript programming language. It is human readable and writable while having comprehensive and efficient support in different programming languages and environments.
LA	Learning Analytics. Research, development, and practice related to collecting, analyzing, and presenting data from educational sources in order to understand and improve learning.
LLAMA	An animal related to camel or "la lumière à Montagne analytique". The latter one is a visualization client for learning analytics that this thesis contributes.
LMS	Learning Management System. A software system that administrates and delivers educational resources and tools. Virtual Learning Environment (VLE) is a synonymous term.

LRS	Learning Record Store. A data warehouse that stores and retrieves learning activity statements using xAPI standard.
MIT	Massachusetts Institute of Technology. A university mentioned in this thesis in context of MIT license that is a permissive open–source software license. MIT licensed software can be integrated into GPL software but not vice versa.
MVC	Model–View–Controller. A design pattern that separates program modules into a model that stores and accesses data, a view that represents user interface, and a controller that includes application logic.
MOOC	Massive Open Online Course. Educational courses that are available online and accept anyone as a student. Therefore, large number of students is expected.
ORM	Object–Relational Mapping. A solution that maps objects defined in a programming language to a different type of data, such as persistent records in a relational database.
SaaS	Software–as–a–Service. The software vendor is responsible for constantly delivering and maintaining the software for the users. These requirements are satisfied by offering the software using web technologies via a web browser.
SQL	Structured Query Language. A language to create, retrieve, update, and delete data from a database.
URL	Uniform Resource Locator. A system to name or address unique resources in internet.
VLE	Virtual Learning Environment. A software system that administrates and delivers educational resources and tools. Learning Management System (LMS) is a synonymous term.
xAPI	Experience API. An API that defines how learning tool, such as LMS, communicates with LRS.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis discusses *Learning Analytics (LA)* which concerns analysis of data collected from activities where people are learning. The thesis contributes a novel software solution that is designed for the particular research case to start the continuous practice and development of LA.

This chapter introduces the thesis to the reader. First, the current trends of digital transformation in learning are discussed. Second, the state of online learning in Aalto university is summarized. Then, research goals and questions of this thesis are introduced. Presentation of the applied research method follows. Finally, an overview of the thesis structure ends the introduction.

1.1 Digital Transformation in Learning

Our societies are undergoing digital transformation. Computers and internet have become an essential part of our everyday life. People depend on networked mobile computers for communication but also for a growing number of other applications, such as calendar, navigation, or entertainment. Directly or indirectly, our purchases at shops, consuming media, or search for information is enabled via internet.

This digital transformation is a source of ongoing revolution on business models and how we people work in our professions. Today, most professions already require some form of computing but the change is not over. The transformation advances and new ways to use the available computational power, networks, and recorded data emerge.

Education is an important part of society and it has not evaded the digital transformation. Study records and a lot of curriculum information is stored in databases. More interestingly, with or without guidance students adopt digital tools to create content, solve problems, and interact with each other. Educators should enable students to use the new efficient tools in learning. Moreover, new technologies that are specifically designed for learning can further improve learning results. As an example, interactive learning material may provide more personal and timely feedback to a larger number of students than a finite number of educators ever could.

Publishing and accepting learning material online can greatly improve access to educational resources and education itself. Many high profile universities around the world have created courses that are completely online and are open for anyone to enroll. On platforms, such as Coursera[1] or edX[2], these massive open online courses (MOOC) can have thousands of students.

However, a less dramatic development is that educators introduce online components to courses that retain all or some of their face-to-face learning sessions. This mixture of traditional and online learning is known as blended learning. Some disciplines and teaching methods are not as good fit for online learning than others. For example, some disciplines, such as medicine, may require a supervisor that is present to guide the student doing a task.

Today, it is commonplace that courses include online material and assignments. When students open and interact with this material or submit their responses these events are recorded. Such granularity and coverage of learning activity data has not been available before the introduction of online learning. This data along with study records can be analyzed to identify patterns of learning behavior. Such research is known as Learning Analytics (LA).

LA has potential to offer data–driven development to optimize education. Educators can evaluate and detect problems in their material or selection of teaching methods. Institutions may evaluate challenges in study programs. Ideally, LA may improve the very understanding of learning, and give new insight to both students and educators.

1.2 Online Learning in Aalto University

Every course in Aalto university has a representation in an institute controlled online learning management system (LMS). Effectively, every course is enabled online to distribute learning material, collect student submissions, create questionnaires, arrange peer review, and manage discussion boards. However, not that many of the courses are actively using the online learning

[1] https://www.coursera.org

[2] https://www.edx.org

components and most rely on traditional face-to-face lectures and laboratory sessions. In contrast, also several Aalto courses have been arranged as MOOC. Some disciplines, such as computer science or mathematics, have more tradition in online materials and exercises than others.

Aalto university has an ongoing development project known as Aalto Online Learning (A!OLE) [Kauppinen and Malmi, 2017]. It seeks to pioneer online learning experiences to improve learning results and to share related knowledge and tools in the university. The project involves a number of pilot courses in different disciplines. This ensures that new online learning opportunities are currently created in the university.

The online learning activities in Aalto university have great variation. The types and the technical platforms of the activities routinely change from course to course. Currently, access to the data generated in online learning requires deep knowledge of the platforms. Also a common data format is missing. Therefore, the past LA efforts in Aalto university have been individualistic and towards a single course or an exercise type.

1.3 Research Goals and Questions

The goal of this thesis is to *bootstrap LA in multiple courses that implement different weekly online learning activities*. The term bootstrap underlines the aim to support continuity, further development, and expansion of LA. Thus, our research problem includes collection of data from different sources, development of analytics according to stakeholder objectives, and delivery of current results that can lead into action and development to improve learning. This thesis presents the first step into LA, that can be extended to new courses and stakeholders in the future.

The research case in this thesis is A!OLE that includes pilot courses implementing different online learning activities. The pilot courses need to be researched to define requirements for a solution. In addition, the case anchors the evaluation of the presented solution to real educational courses. In order to reach the goal, research questions presented in Table 1.1 must be answered.

Table 1.1: Research Questions.

RQ1	What learning data the courses currently instrument?
RQ2	What LA objectives the course staff find most important?

First (RQ1), we need to examine the pilot courses to identify online learning activities and technical learning platforms that are relevant to our case. These produce the learning data that is currently instrumented. We focus on the currently available data to produce immediate value that increases commitment. Differences on data structure and storage on the different platforms add further requirements for the developed solution.

Second (RQ2), we want to identify the LA objectives that the course staff find most important. LA has a number of stakeholders, such as learners, educators, administrators, and researchers [Korhonen and Multisilta, 2016]. However, if the course staff does not have ownership in LA, it is likely that they neglect to systematically design instruments necessary to produce detailed data of learners in future. The objectives are discovered via qualitative analysis of staff interviews.

Answers to these research questions define software requirements. The thesis then develops software that can bootstrap LA in this research case. The software is the solution to the research problem. Table 1.2 sets three more detailed research goals to direct the design and evaluation of the delivered solution. Accessibility and maintainability are essential for a solution that is only the initial step into the practice of LA.

Table 1.2: Research Goals.

RG1	Course staff and researchers can effortlessly access collected learning data.
RG2	Course staff can efficiently complete their initial LA objectives.
RG3	Software developers can readily maintain and extend the solution to provide further modeling and analysis of learning data in real time.

1.4 Research Methods

The research approach in this thesis is design–science research as described by Hevner et al. [2004]. The research identifies a relevant problem and systematically designs a novel artifact as a solution. The utility, quality, and efficacy of the artifact is rigorously evaluated. In addition to the artifact, verifiable contributions that the design or the design process includes are documented. The development of the artifact can clarify the problem definition and allow evaluation of the designed solution approach.

This thesis conducts software engineering to construct the LA solution for the research case. The software engineering process follows the waterfall model and includes requirements definition, architectural design, implementation and unit testing, and finally validation [Sommerville, 2011]. As a part of the process, we research user requirements using semi-structured interviews [Bernard, 2012]. The interview method is discussed in Chapter 4.

On large scale, we are solving a previously known problem of LA. However, the bootstrapping goal brings forth a specific problem where characteristics of a good solution are not previously defined. The thesis designs a novel solution that improves from previously available solutions to the specific problem which according to Hevner et al. [2004, p. 82] differentiates design–science research from the practice of design.

The structure of the thesis follows the publication schema presented by Gregor and Hevner [2013] with the exception that the research method is defined already here as a part of the introduction to the thesis. Conforming the presentation of the research to an existing schema helps to communicate and establish the contributions of the research. Next, we describe each chapter in this thesis.

1.5 Thesis Structure

This first chapter introduced the reader to the domain, goal, and method of the thesis. The second chapter defines LA and reviews related literature and existing LA solutions. Potential benefits and challenges in the practice and development of LA are evaluated.

The next two chapters define software requirements for the solution. The third chapter explores the A!OLE project to answer the first research question on collected data. Furthermore, it develops focus on the pilot courses where data is currently available. The fourth chapter reports application of user requirements interviews to answer the second research question on analytics objectives.

Then, the fifth chapter presents architectural design and describes software components that comprise the solution to the problem. The design decisions that fulfill the defined requirements are documented.

Finally, the sixth chapter evaluates that the solution is useful and an improvement over previously available alternatives. Completion of each research goal is evaluated. The thesis ends with conclusive discussion on the thesis work and consideration of contributions to domain knowledge.

Chapter 2

Learning Analytics

This chapter presents the related work. First, it defines Learning Analytics (LA). Discussion of different stakeholders who have interest in this domain follows. Next, the complete process of LA is defined and the necessary steps are researched in detail. Finally, the available LA solutions and standards are examined.

2.1 Definition

Journal of Learning Analytics is dedicated to "research investigating the challenges of collecting, analyzing, and reporting data with the specific intent to understand and improve learning" [Gasevic et al., 2014, p. 1]. Journal of Educational Data Mining declares that their research community seeks to use "large repositories of educational data" to "better understand learners and learning, and to develop computational approaches that combine data and theory to transform practice to benefit learners" [Baker and Yacef, 2009, p. 1]. The research areas of Learning Analytics (LA) and Educational Data Mining (EDM) share the same goal of using educational data to understand and improve learning. However, there are different trends and focus between the communities.

Siemens and Baker [2012, p. 253] identify two key differences in research trends. In many cases, EDM leverages human judgement to design automated discovery that directly affects learning environment. In contrast, LA often uses automated discovery to inform humans who make final judgement. The other difference is on holistic vs. reductionistic axis. LA tends to take more holistic approach to understand systems as wholes while EDM often analyses individual components and their relationships.

Ferguson [2012, p. 312] argues that LA and EDM are separate research fields. LA focuses on the challenge of improving education while EDM focuses on the challenge of extracting value from big educational data sets. Thirdly, Ferguson names Academic Analytics (AA) as closely related yet separate research field. AA may use same analytics methods as LA but the former addresses institutional, national, and international stakeholders while, according to the author, LA has focus on course and department level.

Chatti et al. [2012, pp. 321-324] state that "LA concepts and methods are drawn from a variety of related research fields". AA and EDM are included in these related fields by reasoning that is aligned with Ferguson. The authors view that LA builds upon and as a term encompasses the closely related fields.

This thesis uses the term *Learning Analytics (LA)* widely to refer all *research, development, and practice related to collecting, analyzing, and presenting data from educational sources in order to understand and improve learning.* Similarly to Chatti et al. [2012, p. 324], EDM and AA research fields are considered encompassed inside LA and are included in the related work of the thesis.

2.2 Stakeholders

LA has a number of different stakeholder groups, including learners, educators, institutions, policy makers, and researchers [Korhonen and Multisilta, 2016; Romero and Ventura, 2013; Ferguson, 2012; Chatti et al., 2012; Clow, 2012]. This chapter considers stakeholders to form an expectation of possible LA objectives and challenges.

Figure 2.1 presents main stakeholder groups in two different possible relations to LA. Greller and Drachsler [2012, p. 45] present the terms subjects and clients for stakeholders. First, LA may instrument actions and context of stakeholder subjects to collect data. The subjects may have privacy concerns. They may not want to reveal personal information or they may worry that incomplete instrumentation leads to wrong analysis on themselves. Second, LA may inform stakeholder clients using different types of results. Optimally, the clients have objectives that the LA results help to reach. The same stakeholder can also be a subject and a client at the same time.

In the following, we discuss the main stakeholder groups and their potential objectives. In addition, LA involves at least system developers. The thesis assumes that these additional roles take one of the main stakeholder perspectives when they are involved in the LA process. It is also possible that persons move from stakeholder group to another when their role changes.

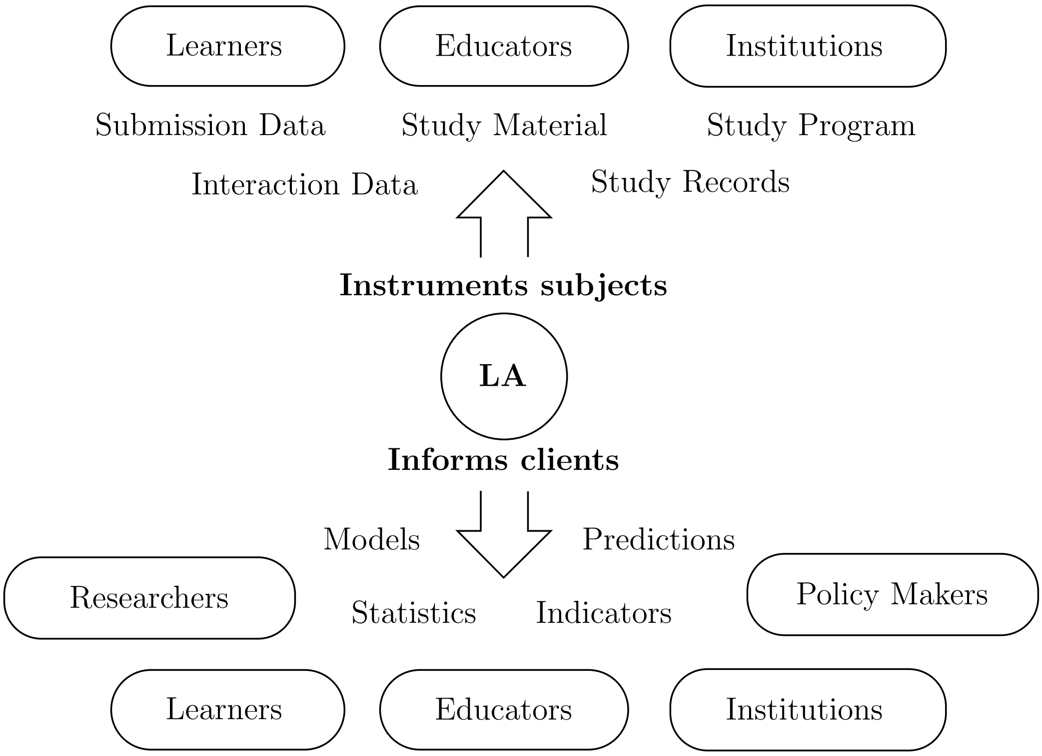

Figure 2.1: Learning Analytics Stakeholders as Both Subjects and Clients.

2.2.1 Learners

The typical objectives of a learner are to improve learning performance, get adaptive feedback or recommendation, and reflect on learning [Romero and Ventura, 2013, p. 18]. From the learner's perspective LA has potential to offer more personalized, and therefore more interesting and effective, learning. Furthermore, LA may help to focus and communicate career goals if lifelong learning data is made available for students. Learners can change their own learning activity in an instant so the potential effect to learning is rapid but limited to one learner at a time [Clow, 2012, p. 136]. Such fast feedback loops require automatic real time analysis.

Most of the data that LA instruments and collects is generated by learners. Course enrollment and grades are stored in study records. Digital exercise submissions and their assessments are a rich source of data. Interactions include everything from posting a message to scrolling down in study material. The finer interaction events are collected the stronger privacy concerns arise. If learners are minors, their guardians are in control of privacy decisions which can further complicate LA [Drachsler and Greller, 2016, p. 89].

Transparency of LA is also important. Learners may be willing to disclose private behaviors in order to improve teaching but are afraid the same data could be used for assessment and grading instead [Chatti et al., 2012, p. 326].

2.2.2 Educators

Educators include teachers and teaching assistants. Also learners can take temporary educator role, e.g. in seminar courses. Educators have objectives to improve teaching performance, understand learning processes, and reflect on teaching [Romero and Ventura, 2013, p. 18]. LA can provide automation that reduces administrative tasks and refines information to highlight the phenomena that educators find interesting in their context. The actions of an educator may address a group of learners but the change in the actual learning of a learner typically has a delay of days [Clow, 2012, p. 136].

Educators may also become LA subjects. Interaction events, such as answers to questions and views of learner profiles, can be interesting for reflection or guidance of efforts. For example, efficiency of educators personal support to different students may be evaluated. Study material may provide contextual input about the course design that the educator is responsible for.

Similarly to learners, transparency is important. Educators are typically employees and they should have right to know what data their employer collects and for what objectives [Chatti et al., 2012, p. 326]. For example, if a teacher designs data collection to develop teaching, the institute must not evaluate the teacher's performance from this data without mutual agreement.

2.2.3 Institutions

Two different institutional stakeholders, administrators and program leaders, have different objectives in focus. Administrative objectives include organization of resources, improvement of student retention, improvement of study progress, and development of student recruitment [Romero and Ventura, 2013; Chatti et al., 2012, p. 326]. Previous objectives have often effect on institutional finance and research of AA has provided tools similar to business intelligence tools to help in decision making [Chatti et al., 2012, p. 319]. Typical actions include staff training, adaptation of new technology, and services, such as healthcare, that support staff and students.

The focus of program leaders is more similar to educators than administrators. They have responsibilities for learning goals and graduate attributes over all courses included in the program. The study program structure can also be a direct contextual input to LA. Changes to resources or study program typically affect multiple educational courses and take semesters to have effect in learning [Clow, 2012, p. 136].

2.2.4 Researchers

Researchers of education are interested in objectives of the other stakeholders in order to find best methods for different education tasks. Ideally, models extracted by LA may help to understand learning and advance learning theory. Researchers of LA have the objective to improve LA methods and process [Romero and Ventura, 2013, p. 18].

As an exception to the other stakeholders, method comparisons often require A/B–test arrangements in data collection and analysis. Also anonymized learning data is beneficial for sharing research material and replicating studies. It can solve the privacy issues which are emphasized in national or international research. Standard formats of educational data can further accelerate evaluating and adapting methods using large studies including multiple courses and institutions. Research does not usually require real time analysis. Effect to learning is slow but can affect whole education discipline.

2.2.5 Policy Makers

Also policy makers on both municipal and national level make decisions that affect education. New policies are best found on well proven research results and align with research objectives. They can enable or accelerate the best known education methods. New policies typically are a result of democratic process. Policies have national or international effect but have years of delay [Clow, 2012, p. 136].

Policies are important to protect privacy and lay the legal framework to conduct LA. Currently, the legal systems are advancing to address privacy, copyright, intellectual property, and data ownership in digital environments. Student exchange may further complicate the legal issues when national laws differ [Siemens, 2013, p. 1394].

2.3 Process

We understand LA as a continuous process where analytics are applied, evaluated and developed to fulfill stakeholder objectives. Siemens [2013, p. 1391] argues that LA is not only a technical challenge:

> The effective process and operation of learning analytics require institutional change that does not just address the technical chal-

> lenges linked to data mining, data models, server load, and computation but also addresses the social complexities of application, sensemaking, privacy, and ethics alongside the development of a shared organizational culture framed in analytics.

Organization starting LA needs to support the introduction, acceptance, and understanding of LA for the whole community including both educators and learners. Greller and Drachsler [2012, p. 43] describe these social or cultural aspects as soft dimensions in contrast to fact-based hard dimensions. Technically LA operations need to continuously adapt to new requirements, such as new data sources, methods or tools.

First, the LA process is described in cyclic steps. Then, each step is investigated in detail to consider relevance, options and potential challenges. This chapter provides a high-level discussion to understand the required steps. Chapter 3 extends to the particular requirements of the case studied in this thesis.

2.3.1 Cycle

The process of learning analytics is described as an iterative cycle [Romero and Ventura, 2013; Chatti et al., 2012; Clow, 2012; Siemens, 2013, p. 1392]. All the descriptions include the following three steps: data collection, analysis, and action in this order. First, data is the target for analysis. Second, if analysis does not lead to action then learning cannot be improved which is part of the accepted definition of LA. Finally, changes in learning are potentially visible in new data which closes the cycle.

Two process descriptions also include a step for refinement and evaluation of the LA itself [Romero and Ventura, 2013; Chatti et al., 2012, p. 324]. In this case, not only the data but also the methods of data collection and analysis or the decided action may change for the next LA iteration.

Clow [2012] adds learners explicitly to the LA process cycle. The data is collected from learners and actions are directed at learners. The author finds similarities to learning theories and feedback loops they describe. A fast feedback loop, similar to discussion, would involve automatic analysis results that are presented to learners in a digital learning environment. At the other end of the spectrum, feedback can be very slow and target other group of learners and content than where the analyzed data was collected from. The latter would be true for e.g. changing government policies.

Chatti et al. [2012, pp. 324-331] propose a reference model for LA based on four dimensions: What, Who, Why, and How. 'What' relates to the data collection step, and 'How' relates to the analysis step. 'Who' is about

LA stakeholders that were discussed in Chapter 2.2. ‘Why’ is an interesting dimension that justifies the whole LA process. This concerns the objectives that are relevant to the stakeholders and that can improve learning.

The research fields of teacher inquiry into student learning and learning design have been linked with LA [Mor et al., 2015]. Teacher inquiry examines teacher’s practice and it’s effects on student learning. Therefore, it produces objectives that LA may be devised to answer. Mor et al. [2015, p. 224] argue that, “We need to be aware that the pedagogical decisions embedded in learning designs affect both the learning analytics process and its outcomes.” These considerations highlight the importance of the two process steps that define objectives for LA and evaluate LA results.

This thesis extends previously presented LA process cycles to explicitly include the objective definition. Furthermore, the concept of learners is expanded to all involved stakeholders. Figure 2.2 presents the proposed extended LA process cycle. A smaller cycle represents the core analytics cycle that is often automatically and continuously executed during the learning process. Next, each step is researched in more detail.

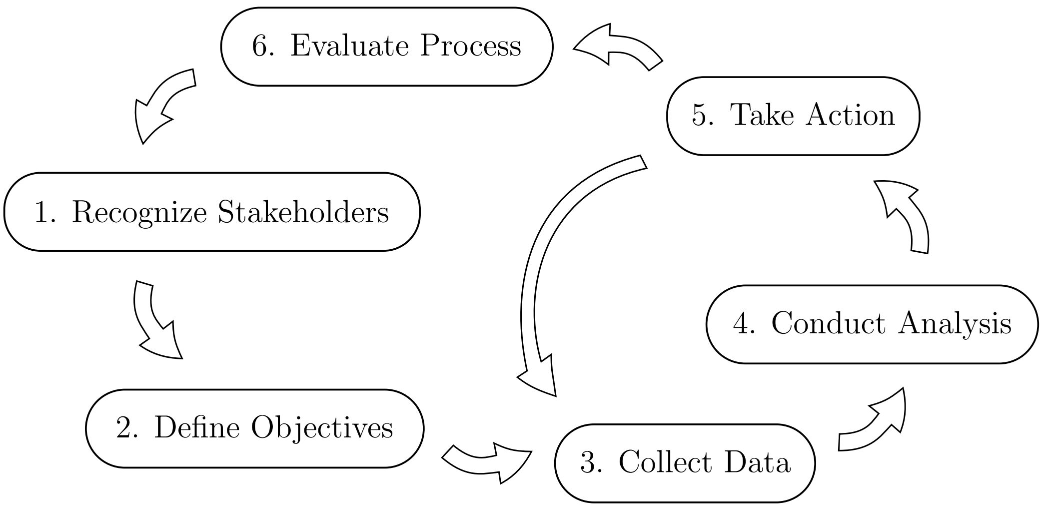

Figure 2.2: Proposed Steps of Learning Analytics Process.

2.3.2 Recognize Stakeholders

New stakeholders enter the LA process at this step. They may appear from opening the process to a new stakeholder group or extending to new educational courses and learners. In order to address the previously described social complexities of LA, it is important to recognize and include all stakeholders of the process. Chapter 2.2 describes both potential subjects and potential clients of LA.

This step should include decision on how stakeholders are represented in 'Define Objectives' and 'Evaluate Process'. The owner of the LA process should design communication towards the stakeholders to share understanding and add acceptance of LA. Siemens [2013, p. 1391] recommends organizations to take "stock of their capacity for analytics and willingness to have analytics have an impact on existing processes." Oster et al. [2016] present an instrument to evaluate learning analytics readiness of an institution. In addition to data management and analysis, the instrument measures culture, communication, policy adaptation, and training. Sclater [2016] defines a comprehensive taxonomy of ethical, legal, and logistical issues.

2.3.3 Define Objectives

LA objectives should be relevant and feasible. The stakeholders can be involved to generate objectives that they are interested in. Additionally, LA research suggests objectives for different stakeholder groups and reports on the required data collection and analysis methods to assess feasibility. Also, Teacher Inquiry research [Mor et al., 2015] or applying Action Research method [Chatti et al., 2012, p. 320] can produce LA objectives, such as testing pedagogical decisions.

Papamitsiou and Economides [2014, pp. 54-56] systematically review LA articles to discover the basic research objectives of LA. They recognize six categories of objectives: student/student behavior modeling, prediction of performance, increase (self-)reflection and (self-)awareness, prediction of dropout and retention, improve feedback and assessment services, and recommendation of resources. Chatti et al. [2012, pp. 327-328] list seven possible categories of LA objectives: monitoring & analysis, prediction & intervention, tutoring & mentoring, assessment & feedback, adaptation, personalization & recommendation, and reflection. The latter categorization is more extensive and can support the cases presented in the former study. Table 2.1 describes these seven objective categories [Chatti et al., 2012, pp. 327-328].

Bakharia et al. [2016, pp. 332-334] approach LA from the point of view of learning design. They develop a framework that includes a very different categorization of LA into five dimensions: temporal analytics, comparative analytics, cohort dynamics, tool specific analytics, and contingency & intervention support tools. In comparison to previous categorization, these dimensions are more closely related to the data and method – 'What' and 'How' in contrast to 'Why'. However, we believe that the five dimensions presented in Table 2.2 are helpful to form concrete LA objectives [Bakharia

Table 2.1: Learning Analytics Objective Categories.

Monitoring and Analysis	Track activities to support decision making and to evaluate learning process.
Prediction and Intervention	Model learners to predict their knowledge and future performance. Optionally suggest actions.
Tutoring and Mentoring	Guide learners within their learning process. Mentoring is more learner initiated and holistic than tutoring.
Assessment and Feedback	Support assessment and provide intelligent feedback according to learner actions.
Adaptation	Organize learning resources and activities according to the needs of an individual learner.
Personalization and Recommendation	Help learners themselves decide and navigate to knowledge. Shape their personal learning environment according to their learning goals.
Reflection	Encourage evaluation of past work and personal experiences to improve learning.

Table 2.2: Learning Analytics Dimensions.

Temporal Analytics	Analyze access times to learning resources, aggregated access numbers, and session durations.
Tool Specific Analytics	Analyze exercise grades, number of attempts, created content, social networks, and discourse.
Comparative Analytics	Compare and correlate metrics from different periods or of different types.
Cohort Dynamics	Apply pattern discovery on available metrics to detect different learner groups.
Contingency and Intervention Support Tools	Enable search and communication using results from analytics.

et al., 2016, pp. 332-334]. Examples of concrete objectives include monitoring time spent in self study before lectures, comparing exercise grades that measure different learning goals in order to recommend learning content, or detecting cohorts of learners whose access and result pattern has previously indicated failing the course.

To address the social concerns the objectives should be transparently communicated to all stakeholders. Trust is easily lost if the previously collected data is used to decide action that was not part of the originally discussed objectives [Chatti et al., 2012, p. 325]. Objectives are also technically critical for the design of the following three process steps: data collection, analysis, and action. Often, many of those three steps are automatized and they decide the feasibility of the set LA objectives.

2.3.4 Collect Data

Online and digital learning environments, such as Learning Management Systems (LMS), record or log many user activities. Therefore, a lot of data often exists that is valuable for LA. However, there are technical challenges to employ the existing data in LA. These challenges are also described as data preprocessing [Romero and Ventura, 2013, p. 20]. Different systems use different data formats and extracting suitable variables to solve a particular LA case may be laborious.

Moreover, ubiquitous online productivity tools and interoperable online material or plugins are commonly integrated into learning material. This presents a big challenge to aggregate and integrate data from multiple data sources with different formats and potentially different granularity [Siemens, 2013; Chatti et al., 2012; Romero and Ventura, 2013, p. 20]. Furthermore, this distributed data set should represent learning process as a coherent whole but in the worst case a single individual may act under different identities in different environments that store heterogenous data [Siemens, 2013, p. 1393].

The previous challenge often exists inside a single educational course. The challenge increases manyfold when we consider all studies in one degree program. A lifelong learning path of a person involves different educational institutions that may cross over national borders and legislation areas. These different parties may have separate educational data that as complete whole would have increased LA potential for the learner. Providing ownership and access to this lifelong data on education becomes a new challenge.

The described problems are largely analogous to the big data concept that revolves around capturing, storing, and analyzing large amounts of data gathered by numerous sources around the world [Swan, 2013]. Current available solutions to the data integration problem in LA are discussed in Chapter 2.4.

Apart from technical problems, giving meaning to numeric data, such as visits or points, requires context [Romero and Ventura, 2013, p. 20]. For example, a particular exercise can be devised to measure a particular learning goal. Without such context information the lack of accessing or mastering the particular exercise has much more limited value. Furthermore, we need to remember that data is just a sample that approximates actual learning processes. While we pursue to quantify learning, e.g. with grades, the actual learning and knowledge are deeply personal and qualitative properties. Achieving sophisticated LA objectives may well require learning design that includes appropriate and tested instruments that can provide suitable data [Gašević et al., 2015].

Increasing the scope of data capture is a recognized challenge in LA research [Siemens, 2013, p. 1392]. Presently, lecture activity or informal learning are sparsely instrumented to data while those are important modes of learning. Such instrumentation could employ RFID tags, mobile devices, new technologies, and new software support. When means to collect data on new situations, such as physical presence at lectures, learning activity in social media, or browsing external internet resources, appear then new level of privacy issues arise. For some objectives anonymized data is adequate and the anonymization can be seen as part of the preprocessing.

Normally, the data collection is an automatic step once it has been designed and developed. On rare occasion, manual input of data that adds value to analysis may be feasible. However, surveys, such as feedback, can also be seen as distributed manual input which is in regular use.

2.3.5 Conduct Analysis

The analysis step seems to be understood as the core of LA. Large part of LA or EDM research investigates different methodology and attainable results. Therefore, the first option to answer an LA objective is to review literature for a presented method and evaluate the transferability to the particular case, including available data and context. Second, if the involved persons have the necessary knowledge, any existing or new analysis methods can be tested to solve a given problem with available data. Next, we discuss the typical and best known methods applied in LA.

Chatti et al. [2012, pp. 137-140] identify four popular techniques in LA literature: statistics, information visualization, data mining, and social network analysis. Romero and Ventura [2013, pp. 21-22] name eleven popular EDM methods: prediction, clustering, outlier detecting, relationship mining, social network analysis, process mining, text mining, distillation of data for human judgement, discovery with models, knowledge tracking, and nonnegative matrix factorization. Chatti et al. makes a higher level categorization that summarizes typical data mining methods to one category. The statistics and information visualization belong to distillation of data for human judgement in Romero and Ventura. Table 2.3 briefly describes the different methodology that has been typical in LA analysis [Romero and Ventura, 2013, pp. 21-22].

Table 2.3: Learning Analytics Method Categories.

Distillation of Data for Human Judgement	Use statistical measures and information visualization to summarize complex systems and to highlight interesting behavior.
Prediction	Use classification, regression, or density estimation to predict attribute, such as learner performance, from other data.
Clustering	Use distance measure to identify groups of similar instances, such as learners that share same learning habits.
Outlier Detection	Detect significantly different instances from the rest to reveal e.g. learners that have learning difficulties.
Relationship Mining	Identify temporal, linear, causal, or other association rules between variables to learn patterns in learning behavior.
Social Network Analysis	Model e.g. learner discussions as social network and use network theory to measure and understand social interaction in learning.
Process Mining	Use model discovery, conformance checking, and model extension to learning process as described by sequence of learning events.
Text Mining	Categorize and summarize natural text from e.g. learner's questions or submissions. Extract concepts and model their relations.
Discovery with Models	Use previously validated model as a component in other analysis to research complex relationships and to approach more general research questions.
Knowledge Tracing	Model learner's mastery of different skills and update relevant estimates after each interaction that is mapped to skills.
Nonnegative Matrix Factorization	Extract new information using a transfer model encoded as a matrix, e.g. model of exam questions to skills maps exam results into a skill matrix.

Korhonen and Multisilta [2016, pp. 305-306] categorize analysis into two groups according to timing and type of the feedback loop. First, analysis can be conducted automatically in real time using the latest data. In this scenario, the general goal is to detect some phenomena and take early action to improve learning. Second, analysis can be conducted postmortem using historical data after the learning activities are finished. The reasons to use historical data include data collection limitations, that make real time data non-feasible, and data completeness requirements, that arise from objectives, such as training models or testing hypothesis.

In the systematic review of LA articles, Papamitsiou and Economides [2014, pp. 60-61] conclude strengths, weaknesses, opportunities, and threats of LA research. Considering the analysis methodology, reported strengths are: ability to use previously refined and validated data mining methods, visualizations that support human interpretation, advancement in more precise user models, ability to reveal critical moments and patterns of learning, and ability to gain insight to learning strategies and behaviors. Weaknesses include likelihood of human misinterpretation and lack of qualitative analysis methods. Reported analysis opportunities are: increased self-reflection and self-awareness, and integration to decision making systems and acceptance models. Threats include over-analysis, contradictory findings, and pattern misclassification. They cause lack of generality and trust issues.

In conclusion, there are big differences between courses and what data they produce. Therefore, knowledge of the context and adjustment of analysis is required regardless of existing solutions to a given objective. Furthermore, transparent analysis methods that form understandable criteria and visualization that supports human interpretation are helpful in the decision to take action [Romero and Ventura, 2013, p. 20].

2.3.6 Take Action

A successful analysis step leads into action so that learning can be improved. Ultimately, any action should affect the learners. However, the directness of action in LA has large variation.

The most direct action is from the analysis to the learner. This typically involves real time analysis. A signal from the analysis may be automatically communicated to the learner inside their digital learning environment or using separate messaging services, such as e–mail or text messages. Alternatively, the results may be used to automatically adapt or personalize learning environments. Another direct approach is to present automatic visualizations to the learner that may improve self-reflection and self-awareness [Auvinen, 2015]. However, the consequences of these direct actions are hard to predict. Beheshitha et al. [2016, pp. 61-62] present that depending on learners' achievement goal orientations the same LA visualization may have positive or negative effect on learning.

Indirect automatic actions often target educators who then practice human judgement. Identically to previous description, signals or visualizations may be presented to the educator. The educator then decides how and when to take action in the learning process. Educators may filter personal signals to learners in a more constructive fashion than automatic system could.

In addition, educators typically evaluate the learning process periodically, and LA can offer many tools to support both real time and postmortem evaluation or reflection. Extremely, LA may be designed to test a single new learning material item or method. The effect to learners may materialize in the next course module or the next course instance with new learners. A possible effect may involve changes in factors, such as material, methods, schedule, grading, or experience of educators. Institutions could have similar actors as the educator described above but with target over several courses.

When researchers or policy makers take action as LA clients the effect to learning typically has years of delay. The actions include scientific publication, media presence, and legislation. When the directness of LA action decreases the amount of human judgement increases. Ideally, this helps to avoid misinterpretation and ill consequence. However, the likelihood of LA leading to any action is reduced, and evaluation of results becomes more time consuming and challenging.

2.3.7 Evaluate Process

Human judgement is part of many LA analysis which thus include evaluation. In addition to that constant evaluation, the whole process should be systematically evaluated. For each LA objective the stakeholders should answer questions related to process steps, such as proposed in Table 2.4.

Table 2.4: Proposed Evaluation Questions.

EQ1	How well did data represent reality?
EQ2	Were analysis true?
EQ3	Was the planned action performed?
EQ4	Did the action improve learning?

Including learners and educators in evaluation helps to detect misinterpretation. If analysis results are open to learners, "misapplications of analytics are more likely to be identified and challenged" [Clow, 2012, p. 137].

Furthermore, Clow [2012, p. 137] makes an important remark: "All metrics carry a danger that the system will optimise for the metric, rather than what is actually valued." Siemens [2013, p. 1395] warns about removing the human and social processes that are essential in learning from the LA. Involvement of all stakeholders and openness seems to be critical. After the evaluation, it is logical to refine and select new objectives with potentially new stakeholders. The process cycle starts from the beginning.

2.4 Software and Standards

This chapter discusses software support for LA. First, features of selected well–known online learning software are examined. Then, applicability of general analytics software is considered. Finally, this chapter discusses software requirements set by LA research and future development.

Investment into LA software depends critically on the previous online learning investments and non–LA features of the available software that have to be considered case–by–case. Integration to existing platforms is a major issue. In some cases, institutional or national policies may limit sharing of the learning data to external services. Therefore, this brief examination does not intend to evaluate the different software options but rather survey the current state of LA in the mainstream products.

2.4.1 Current Learning Analytics Features

Currently, brand-name products in online learning include different learning environments of which we use the term Learning Management System (LMS). Virtual Learning Environment (VLE) is a synonymous term. These LMSs pursue to provide complete learning experience for learners and required support tools for educators. Some of the products are delivered commercially as licensed software while some major LMS are results of open–source software development where people are free to read and contribute to the program source code. Many products are available as Software–as–a–Service (SaaS) where the vendor delivers the application in internet and the acquirer is free from any installation or maintenance work.

Different LMSs typically have some LA features built–in or available as extensions. However, LA is a new addition compared to more traditional educational delivery features, and it requires specific development expertise, such as statistical analysis and machine learning. Therefore, the current LA features in different LMSs may not satisfy all requirements. An option to

Table 2.5: Examined Software.

ID	Product	URL
1	Blackboard Learn	https://www.blackboard.com/
2	Blackboard Moodlerooms	https://www.moodlerooms.com/
3	Canvas	https://www.canvaslms.com/
4	D2L Brightspace	https://www.d2l.com/
5	Sakai	https://www.sakaiproject.org/
6	Open edX	https://open.edx.org/
7	Moodle	https://moodle.org/
7 g	Moodle Plugin: Gismo	https://moodle.org/plugins/block_gismo/
7 i	Moodle Plugin: Inspire	https://moodle.org/plugins/tool_inspire/
8	Intelliboard	https://intelliboard.net/
9	AspirEDU	http://aspiredu.com/

Table 2.6: Current LA Features. The available features for each product are marked with x and available features via plugins are marked with g and i as denoted in the table above.

Feature	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
(LMS capability)	x	x	x	x	x	x	x		
Statistical Summaries	x	x	x	x	x	x	g	x	
Learner vs. Average	x	x	x	x				x	
Learner Self-Reflection	x							x	
Performance Prediction	x	x		x			i		x
Intervention Tools									x
Social Network Analysis		x		x					
Text Mining		x							

the LA features are separate LA platforms that work on data that may be collected by an LMS.

We selected six LMSs that are popular in higher education in North American market [LISTedTECH, 2015]. In addition, we included Open edX which is an open–source platform that powers one of the largest MOOC providers. Moodle, which is a popular LMS, has a plugin architecture to extend features. We searched the plugin library for LA features and selected the two most installed options. In addition to feature plugins, we discovered two integration options to separate LA platforms which were included in the examined products presented in Table 2.5. The LA features of the selected products were examined using current marketing media, recent conference videos, and product trial periods.

Without the exception of one product, that specializes in performance prediction and intervention, all examined software can produce statistical summaries of temporal events and tool specific event data. However, in addition to course and exercise level summaries different products implement different views, such as course activity timelines, student performance quadrants, configurable statistical queries, learner profiles over multiple courses, or video viewers per each second of video. We presume different products are biased toward different types of learning material and course organization which affects the type of statistics the stakeholders are most interested in. Five products currently offer more comprehensive statistical summaries including comparison of individual learner attributes to course averages. Only two products offer an option to present statistical summaries to learners for self–reflection.

In comparison to the distillation of data for human judgement, that includes statistical summaries and information visualization, other analysis methods are currently less available. Prediction is used in five products to predict learner performance according to previous learners that participated on the same course. Social network analysis is used in two products and text mining in one product to summarize learner actions to educators. The discovered LA feature categories are presented in Table 2.6.

The prediction of performance and prediction of student retention are among the most published and researched LA objectives [Papamitsiou and Economides, 2014, p. 53]. Furthermore, software vendors have published success stories on student retention [Blackboard Inc., 2017]. Retention is a measure that institutions typically track and it is in many cases linked to funding. We believe these factors explain the availability and demand for this feature.

SaaS delivered products eliminate technical challenges of LA features activation. However, cultural challenges of conducting LA as discussed in Chapter 2.3 always remain. Software updates and plugin installations that include LA features may pose a significant technical challenge. Some scalable LA systems require a data warehouse that may use advanced database solutions requiring appropriate human and computing resources. The two examined separate LA products are delivered SaaS and they provide LMS plugins that do not require major changes to existing system and can present embedded views inside LMS.

In conclusion, the different well–known online learning software have introduced LA features starting from the statistical summaries. Furthermore, we expect vendors to follow popular demand and develop LA features as research brings forth feasible objectives and new analysis methods mature. However, there is currently a market for separate LA products that can beat the LMS providers in time to market or product quality. Furthermore, LA solutions that depend on one LMS can be problematic for some online courses. As discussed in Chapter 2.3.4, online courses typically integrate learning material from different sources and the activity data is scattered on different platforms.

2.4.2 General Analytics Software

Comprehensive analysis methods and visualizations are provided in general data analytics software. Analytics are available through mathematical software, such as R[1] or MatLab[2], and the various extension packages they support. Other products, such as SAS[3] or IBM SPSS[4], are specifically designed for data analytics. These powerful tools require good understanding of data analytics and are most useful to specialists in mathematical statistics and analytics.

Business Intelligence (BI) is a research area that aims to support businesses to make more informed decisions based on available data. BI has produced a branch of analytics software, such as Power BI[5] or Qlik Sense[6], that offer simplified interface to conduct data analytics and create visualizations. In addition, traditional spreadsheet software, such as Microsoft Excel[7] and the available extension packages, can support many forms of analytics. These tools may be an attractive alternative to the LA stakeholders that are motivated to design new analytics and who lack the resources to use the most scientific statistical software. Indeed, we expect that BI software vendors may ship specific Academic Analytics (AA) and even LA packages in future as LA market grows.

The major challenge with general analytics software is integration of data into the analytics software and integration of analytics views back to the learning environment, e.g. LMS. The data can be imported to these software from LMS data export or in some cases directly from LMS database. An interesting option is to connect analytics software to LMS web service API.

[1] https://www.r-project.org/
[2] https://www.mathworks.com/
[3] https://www.sas.com/
[4] https://www.ibm.com/analytics/us/en/technology/spss/
[5] https://powerbi.microsoft.com/
[6] https://www.qlik.com/
[7] https://www.office.com/

This would remove manual repetitive steps to export and import data. However, it requires specific support from the both analytics software and LMS. It potentially introduces authorization issues including API access tokens and their management.

In addition to technical challenge of data integration, the lack of widely accepted standards for learning data results in custom interpretation. Understanding the data inside the analytics software is a separate effort for every different source of learning data. Furthermore, a deeper level of knowledge on LA, including best methods and models, is required in comparison to applying one of the predefined LA tools described in the previous chapter.

The integration of real time analytics views, that are produced in general analytics software, into LMS requires online cloud features from the analytics software vendor. Similar requirements exist for generating automated report delivery, e.g. weekly email. However, post–mortem or ad hoc analytics may not benefit from such automatic views or reports. It is possible to efficiently test different analytics in external software and later implement the discovered every day analytics methods in the learning environment itself.

Amazon Web Services[8] and Google Cloud Platform[9] both include visualization and analysis tools in their online big data platforms. Their big data warehouses are designed to handle continuous event streams that match the size of the largest MOOC course providers. Furthermore, they include machine learning services that can reduce the implementation effort of advanced analytics methods, such as text mining, speech and image recognition, or raw neural networks. IBM Watson[10] provides similar services and dialog support with artificial intelligence. The LA integration and development efforts for big data platforms are considerable but they can offer unmatched computing services. Cloud analytics can be an interesting approach to institutions that produce vast amounts of learning data and have team of software developers available.

Analytics of web traffic and navigation patterns is a special LA case. Professional products, such as Google Analytics[11], are ready to produce valid and interesting results when applied to web learning environment.

2.4.3 Development and Research

LA research requires access to the data and knowledge of the learning context. From the research point of view, different learning environments, such as

[8]https://aws.amazon.com/
[9]https://cloud.google.com/
[10]https://www.ibm.com/watson/
[11]https://analytics.google.com/

LMS, should include a data export feature or implement a web service API to access data that is required for analysis. Post-mortem data sets, that can be exported after the researched course has finished, are often used to develop data mining methods. However, in order to test early action with learners, a frequent data access during learning is required.

Second challenge is that different systems use different data formats. Therefore, testing the same methods on different or larger data is laborious. This is also critical when a single course integrates materials from different platforms. A standard format to describe learning content and learning actions would help to solve this problem. Kauppinen et al. [2012] define a Teaching Core Vocabulary to encode course content using semantic web technologies. The vocabulary can be extended to include anonymized learning actions. However, privacy concerns and overhead are high if institutions automatically publish actions, such as mouse clicks, in semantic web. Veeramachaneni et al. [2013] design a data base schema where learning actions from different data sources can be collected and converted for standardized analytics access.

Another approach to the data collection challenge reverses the responsibilities. It introduces a concept of Learning Record Store (LRS) where the different learning tools are responsible to transmit learner activities using standard API. In this model, the data for LA is not owned by a single LMS and LRS can integrate and combine data from different environments. Kevan and Ryan [2016] describe the opportunities and challenges of Experience API (xAPI) that defines the LRS using web service standards. According to them, learning software industry has quickly adopted support for xAPI.

IMS Caliper[12] is a recent specification that describes a Sensor API that takes the same role as LRS. Furthermore, the specification aims to define standard metrics to be used in learning.

Learning Locker[13] is an available open–source LRS. Few commercial LRSs, such as Wax[14] or Watershed[15], are available SaaS. The data model in LRS depends on agreed ontology, that is currently kept in a registry[16] controlled by the developers of the specification. It is extendable for the unseen future. Kitto et al. [2015] present a Connected Learning Analytics Toolkit that harvests learner activities from informal environments, such as social media services, and records those into LRS using xAPI.

From the analysis point of view, ability to efficiently implement and test

[12]http://www.imsglobal.org/activity/caliper
[13]https://learninglocker.net/
[14]http://www.saltbox.com/
[15]https://www.watershedlrs.com/
[16]https://registry.tincanapi.com/

new analysis methods and visualizations on current data accelerates LA research. Therefore, learning environments should be configurable and optimally have methods to include result views from external software, such as general data analytics tools or central analytics services as discussed above.

In recent years, the LA research community has collaborated on concept of open learning analytics. Chatti et al. [2017] present a summary of history and goals of such open ecosystem. The ecosystem should support open learning environments where activities are decentralized. It supports lifelong and informal learning. The learning data, analysis methods and models are effectively shared for research. Software and standards are open and participation of all stakeholders is encouraged.

The open xAPI or IMS Caliper are potential solutions to connect different tools and services in such open ecosystem. Currently, Apereo [2017] coordinates a Learning Analytics Initiative that is developing open software platform which offers technology for open LA. Pardos and Kao [2015] present an LA platform that supports the main goals of the open LA in the current technological environment.

Chapter 3

Aalto Online Learning

This chapter addresses the RQ1 in this thesis: What learning data the courses currently instrument? In order to answer the question, we first examine the research case and develop further focus. Four pilot courses included in this case are selected as primary target.

Then, the two most popular LMSs among the pilot courses are studied. Existing functionalities, access to learning data, and structure of learning data, all set requirements for the LA solution. Moodle and A-plus LMSs are described and examined in that order.

3.1 The Case

One of the Aalto University strategic initiatives for 2016–2020 is known as Aalto Online Learning (A!OLE). First, A!OLE goals are researched in relation to this thesis. Second, the different stakeholders in the case are discussed. Finally, the involved pilot courses are inspected using three different criteria and further focus is developed by selecting specific courses.

3.1.1 Project Goals

Kauppinen and Malmi [2017] define the goal of the A!OLE project as "to develop, explore, and evaluate novel advanced technical solutions and pedagogical models for online/blended learning.". They consider digitalization of education as means to improve learning and to support transformation towards more student–centered pedagogies. The project aims to produce new online learning resources that can answer to new and diverse student requirements. Kauppinen and Malmi [2017] see that demand for personalized and flexible distance learning increases and digital resources and tools become new standards in our societies. According to them, pedagogies transform from old teacher–centered approaches, where students were often thought as passive recipients, to new student–centered approaches, where educators support the active learners with novel methods, that require development.

Kauppinen and Malmi [2017] argue that such major change requires rethinking of pedagogical and organizational practices in university. This presents challenge to the whole university staff. Therefore, the A!OLE project aims to change the whole educational culture in Aalto University by building communities and activities that support sound advancement of online learning, including new pedagogies, at a grass root level.

The A!OLE project names LA as a tool they plan to use for providing advanced personal feedback, identifying cases where educator should intervene, improving infrastructure, developing funding and leadership, and ensuring long–term commitment [Kauppinen and Malmi, 2017]. Consequently, we deduct that LA is expected to take an essential role in A!OLE. This is in agreement with general LA and online learning expectations, such as presented in Chapter 1.

Currently, A!OLE recognizes the importance of LA. However, sponsored development projects have not yet included LA in their focus. Therefore, this thesis specifically considers the bootstrapping of LA which aims at a continuous, systematic and developing process that can fulfill many current and future LA objectives as knowledge is built and methods mature. Thus, we largely ignore the large scale and long–term LA vision that was discussed above and focus on the course level objectives that motivate the course staff into LA.

As discussed in Chapter 2.3, conducting systematic LA presents both technical and cultural challenges. Culturally, the stakeholders need to assimilate how to apply and make sense of LA without ignoring privacy and ethics. This thesis aims to ignite LA inside the cultural change that A!OLE is promoting. We support the grass root level approach by introducing LA solution to a small number of educators whose immediate needs we can cover and thus create real value. A!OLE should then facilitate knowledge and experience sharing from these selected educators and pilot courses to larger community.

Furthermore, we promote course staff to take ownership of LA. We believe that involved course staff who can extract value from LA have good reasons to commit to LA development. However, if the course staff is not motivated, it is likely that they systematically neglect to design instruments that are necessary to produce detailed data of the learners. Thus, the most direct opportunity to improve learning via LA would be lost.

This thesis aims at a technical LA solution for online learning. This is an integral part of the main A!OLE goal. Furthermore, we recognize that change happens gradually and focus on introducing viable and evolving solution inside the A!OLE community. We start from the course level and promote ownership of LA for the course staff to grow commitment.

3.1.2 Stakeholders

As discussed in the previous chapter, this thesis has focus on course level LA and it places learners as subject and educators as client. However, the previously researched goals of A!OLE include program level interest for LA. Furthermore, Aalto University has study program leaders and other institutional stakeholders that are potential future LA clients.

Development of institutional level LA specifically requires aggregation and integration of data from multiple courses. In Aalto, such general data consists mostly of the official study records. Generally available course level data is sparse and likely incomparable as the use of online courseware is diverse.

In other words, the institutional decision makers in Aalto take leadership in LA that starts from the study records. We see this as a top–down approach to LA where low fidelity data is analyzed for large trends. In contrast, this thesis advances a bottom–up approach that deals with high fidelity data and analyses course level trends. Both approaches would benefit from having the full data range and at some point in the future we expect these two developments to join at the middle. The top–down may link related larger trends to course level phenomena. The bottom–up advances instrumentation and standardized data access to course level real time data that can provide more accurate and timely picture of larger trends as well.

In addition to institutions and educators, LA research produces opportunities to place learners themselves as LA clients. However, as discussed in Chapter 2.3.6, learners achievement goal orientations are a critical factor in the outcome. Therefore, implementations of LA using learners as client requires careful research and controlled tests.

Finally, Aalto University includes research groups in areas of learning technology, machine learning, statistical analysis, and data science. These groups are seen as potential resources for further development of LA. Researchers require access to structured data and configurability of systems.

This thesis prepares for the expected future requirements of the other stakeholders. However, the immediate value of the solution is aimed towards the educators of the courses in focus.

3.1.3 Pilot Courses

A!OLE has called for idea proposals twice a year from the whole Aalto University to advance the online learning goals. The best ideas have been developed into A!OLE pilots that implement their idea with A!OLE support. At the time of writing, A!OLE has 52 pilots that together involve approximately 60 educational courses or similar units of education that occur in the following academic year [Aalto Online Learning, 2017]. From the LA point of view, all these courses are potential sources of learning data and application targets of LA. However, it is impossible to solve the different technical challenges and communicate the cultural challenges to all of these courses at once. Next, the limits of this thesis are further defined to maximize potential impact using the limited resources.

In their descriptions [Aalto Online Learning, 2017], 17 pilots include development of automatic assessment in their goals. These pilots involve, at minimum, 28 courses. 13 pilots describe different social activities in their focus. Approximately 10 courses are currently subject to these social pilots. Some remaining pilots concentrate on video production, self–learning resources, and guides related to online learning. This superficial analysis produces limited understanding of the pilot courses which regularly combine different learning activities that support each other. However, the automatic assessment courses are attractive for LA as they are already collecting fine–grain and regular data on student attendance and performance. The current commercially utilized LA methods for social learning are less mature and the value for the course staff is harder to estimate than for the structured data from automatic analysis.

Another criteria to categorize pilot courses is the number of students. Many introductory courses have more than 100 enrolled students. In contrast, advanced courses may have as little as 10 enrolled students. On small courses, the educators are likely having discussions with individual learners on weekly basis. This is impossible in large courses and they are likely to gain more immediate value from LA that can summarize data and highlight interesting cases that would else go unnoticed.

Technologically, 25 pilot courses are implemented for the Moodle Learning Management System (LMS). 8 pilot courses are published on the A-plus LMS and 3 courses on the TIM LMS. The previous learning platforms are maintained inside Aalto University. In addition, 8 different externally administrated online learning platforms are used. Furthermore, approximately 10 courses are combining tools from different learning platforms and 7 pilots are developing custom learning applications. This confirms the expected challenge discussed in the Chapter 2.3.4. The learning data is fragmented to different services that are controlled by different owners.

This thesis designs LA solution that includes the two most used platforms in A!OLE. Moodle is one of the most used LMSs in the world and A-plus is an LMS developed in Aalto University. The solution thus builds LA experience with both large software package and more custom application. Both are open–source projects that can accept contributions from this thesis. We will also design for the integration of data from these two sources as a critical future issue.

In conclusion, this thesis is foremost interested in large pilot courses that employ regular automatic assessment and use Moodle or A-plus platforms. In our case study, we focus on the pilot courses named in Table 3.1. Each course is included in a higher education curriculum and has study size of 5 ECTS (European Credit Transfer System). In addition, each course includes weekly online exercise tasks.

Table 3.1: Primary Target Courses.

Code	Name	Students	LMS
MS-A0004	Matrix Computations	113	Moodle
CS-A1101	Programming 1	530	A-plus
CS-A1150	Databases	339	A-plus
CS-C3170	Web Software Development	305	A-plus

3.2 Moodle

Aalto University has provided Moodle LMS as primary online course platform since Autumn 2015. This Aalto platform is branded as MyCourses and each course that student officially enrolls automatically appears in their Moodle. The course staff is responsible for the Moodle content of their course.

According to Moodle[1] community data, Moodle is, at the time of writing, used in over 70 000 institutions, corporations and schools. It is developed as an open–source GPL (GNU General Public License) project which gives credit to 609 individual developers.

First, Moodle's architectural design is discussed. Then, we examine LA related APIs: Activities, Gradebook, Events, Reports, and Analytics in that order.

[1]https://moodle.org/

3.2.1 Architecture

Moodle has a modular architecture where the application core exposes selection of APIs to service plugins that provide actual features on top of the minimum system. Moodle includes a plugin framework that supports 53 different types of plugins that can be implemented in PHP programming language and installed to the server running the software. The project manages a plugins directory that, at the time of writing, includes 1 406 plugins that are available to install.

Studying every available feature of Moodle is out of scope. Therefore, this thesis studies the features that the courses in focus apply. Particular interest is on the data that these features produce. In addition, we study LA capability and development in the Moodle project and plugins.

3.2.2 Activities and Gradebook

In the Moodle architecture, the components that are called *Activities* produce all of the detailed learning data. Such components or plugins implement features, such as forums, wikis, quizzes, and assignments. These extendable activities all define their own database tables where they store the generated data in individual ways. Additionally, they can duplicate learning data as log events that are discussed later.

A direct inspection of activity specific data, such as posted forum messages or submitted quiz answers, requires custom code for each different activity type. Furthermore, *Question API* allows to extend the quizzes and creates further variation on how the stored values can be interpreted. In our focus, the structured learning data is stored primarily from quizzes.

In addition to raw activities, Moodle offers *Gradebook API* that is a shared method to store and retrieve grades for individual activities of different types. A number of attempts on a quizz or other aggregated interaction data, that is interesting for LA, would still need to be implemented separately for each activity type that is going to be supported [Romero et al., 2008, p. 372].

3.2.3 Events

Moodle implements *Events API* to write log entries as events occur. Moodle generates many events, such as a learner viewing a resource, a learner attempting to solve a quizz, or a teacher grading an assignment. An event includes a type, a time, and references to related users and database records. Therefore, the complete learning data of an event can be gathered via decoration of the log event with the related information, such as a submitted answer or an awarded grade. However, querying log entries and related database tables in real time for aggregate data is a heavy operation.

As events are generated, Moodle, by default, stores them in the local database. This event logging design has been harnessed for LA in several plugins. *Logstore xAPI*[2] can decorate and deliver Moodle events to an external LRS that supports xAPI. The LRS solution and few data warehouse options were discussed in Chapter 2.4.3.

3.2.4 Reports

Moodle defines *Reports* and *Quiz reports* plugin types to support creation of report pages that are included in the navigation. By default Moodle includes a report that lists log events and can filter them by type. An example of a visualization plugin is *Events Graphic Report*[3] that provides high level visualizations of event data by user and type. The plugin is documented to exist as an alpha version.

Gismo[4] is a plugin that visualizes student activities on a course. It can visualize the number of accesses to each different course resource by each individual student. The collective numbers can be seen on time scale or alternatively per resource. In addition to accesses, the grade state of both assignments and quizzes per student is available as a visualization matrix.

Technically, Gismo uses a JavaScript library for plotting visualizations and a custom web API resource to feed numeric data to browser. As a nightly task, It computes aggregated numbers of accesses from Moodle event logs. Gismo is available in the Aalto MyCourses installation.

3.2.5 Analytics

Moodle has integrated *Analytics API*, that was originally part of the Inspire plugin, to the core. This API supports definition of *Analysers*, that extract data for analysis, *Indicators*, that calculate more abstract signals, and *Targets*, that define models which predict and notify teachers or learners on results. The analytics component defines a new plugin type for machine learning backends that can be selected for each analytics Target. However, also static models are supported. Such Targets only deduct instead of predicting.

This API seeks to accelerate modular development of LA where different data and machine learning algorithms can be effortlessly integrated together.

[2] https://moodle.org/plugins/logstore_xapi/
[3] https://moodle.org/plugins/report_graphic/
[4] https://moodle.org/plugins/block_gismo/

We expect that different LA predictions and signals become available in Moodle in future versions and plugins. Possibly, the Analysers' output could be used for visualization in reports as it attempts to provide the custom access code that is required for the different Activity types. However, that was not designed in the API.

3.3 A-plus

Computer science educators at Aalto University created an interoperable and extendable LMS that was first implemented by Koskinen [2012] and presented by Karavirta et al. [2013]. The student centered user interface was designed by Krogius [2012]. The A-plus[5] system is also known as A+. However the latter form is problematic in identifiers, such as used in program code or URLs (Uniform Resource Locator), and therefore the former form A-plus is used of the project by itself.

Currently, 18 courses in Aalto University are serviced on A-plus and Tampere University of Technology has adopted it as well. A-plus is developed in open–source under GPL and MIT licenses. The project accepts issues and pull-requests for source code changes.

First, architecture of A-plus is described. Next, we examine the LA related data models: Exercise and Submission. Finally, data integration to A-plus is discussed.

3.3.1 Architecture

The original design allowed to separate the concerns of user session and automatic assessment into different services that communicate over HTTP (Hypertext Transfer Protocol). This idea is aligned with the current web technology trend that employs micro services that are orchestrated together into actual applications [Dragoni et al., 2017]. The development and maintenance of the individual micro parts having their individual responsibilities helps to develop robust, maintainable, and scalable systems. The idea also supports teams and companies to focus and excel in products that have more limited responsibilities.

In the years following the introduction, A-plus design principles have included modular design over HTTP, low effort of implementing custom assessment programs, and controlling learning content via file system and version control software. Today several different services, that follow these design

[5]https://apluslms.github.io/

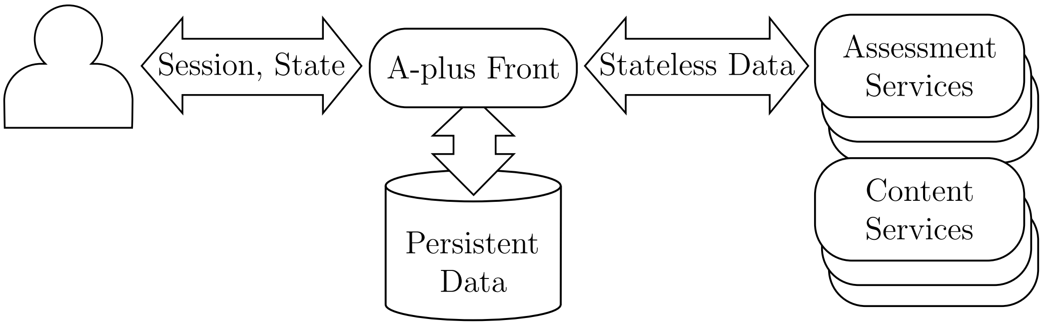

Figure 3.1: A-plus Separation of Concerns.

principles and interoperate together, are considered to be part of the A-plus LMS.

However, course and learner states should be a sole responsibility of the A-plus front[6] service as presented in Figure 3.1. Therefore, that is the critical part for access and control of learning data that we are interested in. The front is implemented on Django[7] framework that implements a Model–View–Controller (MVC) design pattern on Python programming language for web services. Model layer is implemented as Object–Relational Mapping (ORM) that persists data in a database. Template system separates the concerns of view and control. The Django framework offers a complete set of utilities and extendable modular system which support rapid development of advanced web services.

3.3.2 Exercise and Submission

The A-plus front can display different externally produced learning objects to learners and records each display in a database model named *Learning Object Display*. *Learning Objects* form a hierarchy that requires recursive database queries to construct. As virtually every page request requires the hierarchy it is stored in a cached object.

Exercises are a subclass of Learning Objects that is of special interest for LA. Exercises accept learner *Submissions* that include either form or file data. Furthermore, external assessment services can commit feedback to these Submission models. These two Learning Object Display and Submission models describe all available data on learner interaction.

Each Learning Object Display holds an object, a viewer, and a timestamp. Submissions include both structured and unstructured data. For LA the

[6]https://github.com/Aalto-LeTech/a-plus/
[7]https://www.djangoproject.com/

most interesting structured fields are: exercise, submitters, submission time, status, and grade. The unstructured data includes submission and feedback content. In the target courses, the different questionnaires store the names of their fields into the exercise model. Thus, in these questionnaires, also the submission content can be interpreted in structured form.

The application of the A-plus exercises is not limited to graded assignments. On the target courses, the exercise–submission–feedback design is also harnessed for collecting feedback from learners. A-plus is inclined to understand all learner activities as a dialogue between submission and feedback data.

However, our target courses also connect third party services for discussions and service queue that store their interaction data in their separate databases. Integration of data from these third party services to A-plus is out of scope. We consider that the services should either reuse the A-plus exercise–submission–feedback model for storage or implement one of the emerging standard data APIs for LA, such as xAPI, to integrate learning data at a higher level.

3.3.3 Data Integration

The A-plus front includes a web service API implemented using Django REST framework[8]. This API includes secure access to exercise and submission data in JSON (JavaScript Object Notation) and partially CSV (Comma Separated Values) formats. The access can be granted for the standard manually signed in user session or using an API token that is available in user's profile page.

A-plus also supports hooks that can request external web services at specific events. Currently, only one type of hook event is supported. It is triggered after a submission is graded and it posts the identifier of the just graded submission to the configured hook URL (Uniform Resource Locator).

Finally, Riekkinen [2017] developed Astra plugin to access learning content supported by A-plus LMS inside Moodle. The solution bypasses the A-plus front service and replaces it with Moodle. Therefore, in this scenario, Moodle becomes responsible for storing data and conducting LA for the A-plus activities.

[8]http://www.django-rest-framework.org/

Chapter 4

Teacher Interviews

This chapter addresses the RQ2 in this thesis: What LA objectives the course staff find most important? In order to answer the question in our specific case, we ask the staff of the courses in focus. We source their expert knowledge in user requirements interviews. First, this chapter discusses related work and importance of these interviews. Second, the employed interview method is described. Then, the results are presented. Finally, trustworthiness of results is examined.

4.1 Related Work

Two related works interview teachers to discover their wants and needs of LA visualization. Bakharia et al. [2016] create a high level conceptual framework of LA visualization for teachers. They describe LA dimensions that were presented in Table 2.2. In comparison to our focus courses, automatic assessment is rare among their interviewees and the reported results lack detail of specific objectives for our purpose.

Xhakaj et al. [2016] interview only mathematics teachers from elementary school and they focus on particular mathematics learning environment. They describe learning and applying LA in classroom to identify individual learners that need help. In contrast, this thesis focus on large courses in higher education where assignments are typically separate from lectures.

Interviewing teachers from our focus group ensures proper coverage and detail of the issues that are relevant in this specific case. Additional motive for the interviews emerges from the software engineering perspective. The teachers are the users of the software that this thesis develops. They are a natural source of user requirements in this project. Furthermore, the involvement of the course staff is important to build commitment to LA, as previously noted in Chapter 3.1.1.

While the previous interview results are not directly transferable to our case they are used, with other related work, in design of interview themes and critical evaluation of our findings.

4.2 Method

We selected interview as a method that lets us conduct qualitative research of epistemologically subjective knowledge. Maguire [2001, pp. 599–600] recognizes semi–structured user requirements interview as a common technique in human–centered design to gain information on needs or requirements for a new system. Bernard [2012] describes the semi–structured interview method in detail. First, the selection of interviewees is reported. Second, we describe the interview script and how interviews were conducted. Finally, the analysis method is presented.

4.2.1 Interviewees

We decided to interview the responsible teacher from each of the four focus courses. The teacher is likely to have a holistic view of the course issues and if there would be a better person to interview on LA they could delegate. In our case, the teachers indeed were experts in possibilities to apply LA to their course. All of the teachers had entered A!OLE in order to develop online learning and many have previous understanding of LA.

The interviewees include one University Teacher and three Senior University Lecturers. Each interviewee has at least 15 years of teaching experience. One represents department of Mathematics while the others are from Computer Science. The interviewees are above 35 and below 50 years of age. One interviewee is female and the rest are male.

The interviewees are not randomly selected. The selection is determined by the target courses where this thesis focused on. Those courses are selected using the criteria defined in Chapter 3.1.3 that aims to produce immediate value from course level LA. Therefore, these interviews are designed to produce information for the particular research case defined in this thesis.

4.2.2 Script

A semi-structured interview script, that allows free discussion on the selected themes, was designed based on the related research and the current LA software. The interviews were conducted and recorded in audio by single interviewer in Finnish. The duration of one interview was from 50 to 88 minutes, and 64 minutes on average. The English translations of the theme topics and primary questions are presented in Table 4.1.

Table 4.1: The Interview Script.

T1	Teacher's Understanding and Previous Experience of LA
	How do you understand the term learning analytics?
T2	Temporal Analytics
	How could one monitor time use on your course? *Is it important?*
T3	Progress Analytics
	How could one monitor progress on your course? *Is it important?*
T4	Learner State Analytics
	What kind of learner specific analytics could work on your course? *Is it important?*
T5	Social Interaction Analytics
	How important do you consider measuring and analysis of social interactions?
T6	Progress Estimation
	How important do you consider estimates on student success or dropout alerts?
T7	Delivery of Analytics Results
	How would you like to access analytics results? *How important do you consider readability of results, for example naming knowledge areas in addition to exercise or module numbers?*

The definition of LA is broad and the teachers are expected to have varied previous understanding of LA. To open the interview, Theme T1 enables interviewee to express initial thoughts before presenting any question that may lead the answers. In addition, interviewer can adapt to teacher's experience.

The following themes aim to cover the current popular software features, as described in Chapter 2.4.1, as well as different LA dimensions described in Table 2.2. In addition, we consider the relation to LA objective categories presented in Table 2.1.

Themes T2 and T3 cover the most popular LA features in current software which are statistical summaries of temporal events and tool specific event data respectively. In common features, the tool specific events are primarily used to estimate progress in reaching the learning goals. Then, Theme T4 refers to presenting a selected learner in relation to learning content and other learners. This provides opportunities to discuss more on the LA dimensions of comparative analytics and cohort dynamics in comparison to the temporal and tool specific analytics. This may activate also the LA objective categories of tutoring and reflection in addition to the monitoring and analysis.

The next themes broaden the discussion to the outer reach of the current LA features in software. Theme T5 analyzes tool specific data in social interaction context. Theme T6 explicitly moves the LA objective category from monitoring and analysis to prediction and intervention. This topic is likely to cross into contingency and intervention support tools which would cover the last remaining LA dimension. Finally, Theme T7 investigates the usability requirements of an LA solution.

Considering Themes T2–T5, the interviewer prepared to demonstrate screen captures of typical software features on the different themes. They included 16 captures from videos and web pages that are linked in Appendix A. Demonstration can stimulate discussion and suggest concrete visualization methods. However, that introduces a possibility of leading interviewees to conclusions.

The interview aim is to identify specific learning analytics objectives that become requirements in our research case. The interviewer repeatedly encouraged the teachers to think about their own course and name learning analytics objectives they might consider.

4.2.3 Analysis

The interviewer did qualitative deductive content analysis of the recorded interviews. The discussion themes represent our broad understanding of potential LA features and dimensions as discussed above. The interviews aim to discover relevance of the themes and specific needs in our research case. Therefore, we identify LA objectives that the teachers find most important and thus, answer the research question.

First, any requests, objectives, or wishes that teachers could construct were quoted individually and transcribed from the audio recordings. If the same thought reappeared in the same interview only the most detailed formulation was quoted. Then, the quotes were translated into English for reporting.

Finally, the quotes were categorized into the discussion themes that the quotes in their context best belong to. This was not necessarily the discussion theme that was currently active and teachers could freely associate with their previous thoughts. Some quotes shared two discussion themes evenly and they were categorized into combination of the two themes as presented in the results. These themes are then examined and summarized using the included quotes.

4.3 Results

Analysis discovered objectives and wishes as quotes of the interviewees. First, we present the quotes and discuss the findings one theme at a time. Finally, the results are summarized as user requirement statements.

4.3.1 Quotes

The opening discussion in Theme T1 included the interviewee describing LA and their previous experiences. No quotes should be categorized to this theme and it did not produce quotes for the other themes.

Temporal measures were recognized by interviewees and the quotes in Table 4.2 were recorded for Theme T2. Teachers suggested different measures of time usage to aid in allocation of learning material into different units of study or calendar. Attention was also placed in generating proof for communicating typical time requirements and accuracy of learner reported time usage.

Table 4.2: Quotes in (T2) Temporal Analytics.

I want to monitor self–reported time usage in order to allocate amount of material.
I want to monitor time use to allocate amount of material and to generate proof on required work load.
I want to monitor where both students and unregistered visitors spend time in addition to how they report using time during the course.
I would like to see calendar heat map of students all activity including other courses to resolve overlaps.

The majority of the quotes were recorded for progress analytics. These quotes are listed in Table 4.3 for Theme T3. The quotes communicate a general need to monitor that learners are working and proceeding as expected. The possible actions that could follow from monitoring included assistance of learners and improvement of material. However, the actions were not expressed equally strong in comparison to the need of monitoring.

Table 4.3: Quotes in (T3) Progress Analytics.

I want to see on collective level that students are aboard.
I monitor that majority of students achieve full points on exercises they are supposed to.
I want to identify students that have not started in order to push them forward.
I would like to receive weekly activity reports as portions of students who did not answer, answered wrong, and answered correct with different numbers of retries. Thus we could for example improve poor questions.
I want to find cases where exercise is submitted multiple times but no progress is made. The student may need assistance or material could be improved.
I want to find students whose point accumulation rapidly changes. The student may need assistance.
I want to know how many students drop out on each step to identify demanding areas in material.
New teachers may benefit from progress comparison to previous and other courses.
I am interested on solution paths of multiple choice questionnaires to improve interactive feedback.

The quotes indicate interest in deviations and trends over course timeline or course instances. Ratios of learners having different interaction patterns are suggested for summarizing data. The following action would need further design.

Multiple quotes were equally rooted in temporal and progress analytics. Instead of duplicating these quotes to both of the previous tables we separately report the intersection T2 ∩ T3 in Table 4.4. It is evident that the interaction of these two themes or dimensions can provide more information than either of the themes alone.

Table 4.4: Quotes in (T2 ∩ T3) Temporal and Progress Analytics.

I want to find cases where material is studied but related exercise is not submitted. The student may need assistance or material could be improved.
I am interested to see if there are students who read material but do not submit exercises.
I am interested to view animated learning paths à la Hans Rosling in terms of effort and progress.
I am interested to compare [progress and temporal analytics] with previous years to detect changes.

The objectives in this intersection are similar as in the progress theme. However, the analysis target in these quotes is the balance between invested time and progress in learning. In the quotes, a concept of effort was identified that could use different metrics. Both learner reported time usage and number of submissions were suggested.

Table 4.5 lists quotes for Theme T4. Learner state analytics was seen as a tool to improve either benefits or efficiency of interaction with students. Studying individual learner activities is time consuming and large courses do not have resources to routinely view individual learners. A good summary of learner state communicates learner's effort and progress in different units of study in a concise form.

Interest was expressed for comparison with other learners and modeling mastery of different concepts. The quotes in this and previous themes indicate that, on the target courses, a learner typically either completes an exercise with full marks or fails with zero marks.

Table 4.5: Quotes in (T4) Learner State Analytics.

I want to see what exercises student has not finished and may lack knowledge of before answering student's question. Exercises have binary nature.
I want to see number of submissions and deviations from average.
Upon starting an interaction with student, I would like to glance at students' effort, success, and estimate in order to improve the interaction.
We have experimented with online mastery learning model to improve learning and achieved a level of success.

Fundamental interest exists for social interaction analytics or Theme T5. However, only one constructed objective to save time was expressed and recorded in Table 4.6.

Table 4.6: Quotes in (T5) Social Interaction Analytics.

Summary of student discussion topics would help me when I lack time to interact.

Progress estimation is included in both learner state analytics and progress analytics. The discussion often fluctuated between these themes but estimation was commented as separable feature. Table 4.7 lists quotes for Theme T6. Teachers welcomed the addition of estimates. However, many had negative expectation of usefulness of progress estimation and specially drop–out prediction. They either experienced their course as too short to react or previously identified drop–out cases had proven to be beyond salvation for the particular course.

Table 4.7: Quotes in (T6) Progress Estimation.

I am interested on estimates about course results both on collective and individual levels.
Assigning remedial instruction did not produce meaningful improvement.
Proactive intervention has been successful guided by student background. Reactive intervention had poor results before.
Drop–out estimates are very challenging to react to on a six week course.

Delivery of analytics results was a popular theme in the interviews. Quotes for Theme T7 are reported in Table 4.8. Teachers want to interact in real time with the data to test sudden hypothesis and detect new behaviour. They want to filter by units of study and learner demographics once they see the data and also navigate directly into an individual learner state and further into the individual records of the learner activities. However, there was also a wish to export data offline for later research.

In addition, one teacher wished to upload manually created data for comparative analysis. One teacher required access to richer background data on students that is currently only available in official study records and not in LMSs.

4.3.2 Summary

To conclude the interview discussions, the teachers in our focus had previous experience in LA efforts and research. However, continuous development of

Table 4.8: Quotes in (T7) Delivery of Analytics Results.

I want to see total, course module, and exercise statistics separately to monitor both global and local behaviour.
I want to limit statistical views by different student groups in order to test hypothesis and search new phenomena.
I would like to sample different students groups in equal proportions.
Wherever I see individual student, I want the ability open learner state summary and furthermore navigate deeper into their single activities.
I am interested to download learning data to conduct offline analysis.
I want to upload exam score table to the system and study correlation between online and exam problems.
I would like to know how many students of specific program have enrolled to course.

real time LA was currently missing on the courses and definitive interest for such LA exists among the course staff. In the following, we summarize the popular wishes from the interview quotes as user requirements. Each requirement is presented as an objective statement in Table 4.9.

Theme T3 discovered a need for collective monitoring, else educators become easily blinded by the large numbers of learners in our target courses. We consider it natural that educators want verification of and link with learners that only interact with digital platform or disappear into lecture hall filled with hundreds of people (R1). After all, these large courses are as far from the traditional master–apprentice relationship as possible.

Theme T2 identified an objective to allocate learning material so that it best supports learning (R2). Theme T3 suggested deviations and trends as signals for improving learning material (R3). Analysis of these two themes support development of two distinct metrics: learning effort and progress in reaching the learning goals. Theme T7 highlights a need for interactive LA. Ability to filter and navigate interactive analytics results adds value for educators when they inspect different units of study and groups of learners, or interact with students online (R4 and R5).

Theme T4 identified an objective to improve interaction with learners using individual summaries (R6). Theme T6 indicated that estimations were welcome but not requested in the interviews. The courses in focus both encourage learners to group work and offer laboratory sessions. However, Theme T5 suggested, that educators are currently more committed to per-

Table 4.9: User Requirements.

R1	Educators monitor learners' progress and effort to verify learning and learners existence in real time.
R2	Educators measure learners' time usage to allocate learning material into units of study and calendar.
R3	Educators detect both deviations and trends of both learners' progress and effort to improve learning material.
R4	Educators filter learning analytics results by units of study and learner demographics.
R5	Educators navigate from learning analytics results to individual learners and their activity record.
R6	Educators digest real time summaries of individual learner's effort and progress in different units of study to improve interaction with learners.

haps simpler data sources that are available to them.

4.4 Trustworthiness

The trustworthiness of the interview results is evaluated using concepts of credibility, transferability, dependability, and confirmability as described by Lincoln and Guba [1985]. The confirmability is achieved by describing the interview methods and the complete process in Chapter 4. Next, we discuss credibility.

The majority of the communicated objectives were modest and feasible. In part, the modest objectives are explained with the experience of the interviewed teachers. They knew what is feasible with the available data and off-the-shelf methodology. In addition, the themes and demonstrations may have lead the teachers to modest ideas.

The interview did not introduce all LA possibilities. A more complete coverage could be achieved with themes that introduce all LA method categories, as described in Table 2.3, in contrast to LA dimensions and current software as in our case. This would place more focus on the advanced methods and possibly produce more ambitious objectives. Therefore, the interview script is biased towards currently available LA features. The interviews have credibility inside the introduced LA space. However, the interviews can not be used to argue about LA topics that are not included in the interview script.

When considering the transferability of these results, we note that all the interviewed teachers were from courses that have implemented weekly online learning activities and have not yet started systematic LA. Furthermore, the interviewed teachers were part of online learning pilots and had interest and experience in online learning.

To improve dependability, we kept the interpretation of teachers' words to minimum in analysis. However, some interpretation occurs in selection of quotes, transcribing, and translation. During the interviews, we cleared opinions so that they would leave no room for misunderstanding. Another dependability issue is the small number of interviewees. There were only four interviews. However, the analysis of individual interviews revealed similarity in quotes and focus.

Finally, each teacher in our focus has expressed feasible objectives and interest to LA solution that could deliver such value. This allows to design a viable LA solution that can produce immediate value in this research case.

Chapter 5

Solution

This chapter presents software that can bootstrap LA in this research case. This is a solution to the research problem. The design decisions are argued to compliment the technical and organizational environment as described in Chapter 3 and the initial objectives of the course staff as discovered in Chapter 4. Furthermore, the work is considered to align with the consensus of LA research that is discussed in Chapter 2.

First, this chapter develops architectural design that divides responsibilities to different components that together comprise the solution. Then, design decisions of each novel software component are documented.

5.1 Architectural Design

The existing technical environment includes two LMSs that implement different features and data storages. This thesis does not deliver one software application but selection of software components that extend or interact with the LMSs. In following, we argue the need of three different components using previously defined software requirements and our review of related work.

First, real time visualization is considered. Second, steps and benefits of using external analytics tools are discussed. Third, integration of learning data from different sources is examined. Finally, the designed software components and their relations to each other and their environment are presented.

5.1.1 Real Time and Interactive Visualization

Interviews identified user requirements presented in Table 4.9. They include the requirements R1 and R6 to monitor learners' progress and effort in real time. In addition, educators want to interactively filter this data using different course hierarchy and demographic criteria (R4). Furthermore, educators need to navigate from these visualizations further into activity records of individual learners (R5).

First, we discuss how these requirements could be supported in an external system in contrast to implementing visualizations in the LMS system itself. If the LMS learning events are automatically delivered or fetched from an API to an external system, then visualizations can be practically created as real time there as in the LMS. The filtering criteria can be stored to the external system and URLs can offer direct access to the original records in the LMS. However, the transfer of the criteria data, such as demographics, needs design as it is not readily available in the current systems.

The deal breaker in this case is that the analytics should be included in the educators' daily workflow. They may have the motivation to open an external analytics overview in daily basis if a direct link is provided in the LMS. However, our requirements include individual learner summaries that can be glanced when educators are about to interact with a learner in some part of the LMS (R6). We can also imagine that the results of analytics, such as group of discovered students, could lead to action, including communication or assignment of learner labels, that should be another feature available in the LMS.

Considering the previous issues, we believe that critical real time and interactive visualization should be implemented inside the LMS. The primary aim of this visualization inside the LMS is to improve the daily interactions and provide the verification of progress and expected use of time. In a way, the real time visualizations should help to make both learners and educators in the digital system visible and to make sense of the both collective and individual status despite the potentially large number of learners.

Table 5.1: Real Time Visualization Support.

	Effort: R1,R6	Progress: R1,R6	Filter: R4	Navigation: R5
Gismo	chart	color matrix	one by one	-
A-plus	-	table	-	-

Table 5.1 presents existing support regarding real time visualization in the two LMSs: Moodle and A-plus. The *Gismo*[1] plugin for Moodle fulfills some of the requirements. Collective and individual summaries of access are available which can be considered as a measure of effort. Progress is

[1]https://moodle.org/plugins/block_gismo/

visualized as a matrix of achieved grades by learner and activity. The matrix can not provide a collective view on a large course. Data can be filtered by exercise or learner by selecting individual items. The items are not navigable.

A-plus only offers a tabular view of course grades by learner and activity. While this is real time data the lack of summaries makes monitoring progress or effort impossible. Both visualization and interaction are missing.

5.1.2 External Analytics Tools

In Table 4.9, the interviews identified the requirements R2 and R3 that aim to, respectively, allocate and develop learning material. This is not as time critical objective as the previous considerations about real time requirements. Furthermore, Chapter 2 presents LA as a cyclic and developing process. We except that once LA experience on courses grows there will be many new objectives that are less time critical and can manifest in weekly reports or post–mortem analysis.

The previous arguments of including LA inside LMS are not strong in these scenarios. The effort and cost of implementing custom analytics are several times higher in comparison to applying an existing LA tool. Furthermore, general analytics tools offer higher flexibility and rapid experimentation capability. Finding software developers that have required programming skills, knowledge of LMS architecture, and understanding of analytics is not easy.

Aalto University personnel are licensed to use *Microsoft Power BI*[2] that has a low entry barrier to start experimenting with statistics and visualizations. Educators could potentially construct their own dashboards and reports or a resource could be hired to help the different stakeholders to construct analytics and interpret results.

To enable use of external tools this thesis can develop support to fetch data from LMSs into an external analytics software. Power BI can read downloaded CSV, JSON, or Excel files. Furthermore, it can store URLs that provide data in these formats and automatically update the data once analytics are accessed. However, this requires that the access to the URL can be authorized e.g. using a web service API token.

On large courses, downloading all possible data is a heavy operation and conducting simple analysis may require many steps. It would be also useful to have access to aggregate numbers, such as achieved grade by learner and activity, or number of attempts by learner and activity.

Table 5.2 presents existing support for external analytics tools. Moodle by

[2]https://powerbi.microsoft.com/

Table 5.2: External Analytics Tools Support.

	Observations	Aggregate Data
Moodle	download	download
A-plus	API	-

default includes a report plugin that allows data download in Excel format. This is available under *Course reports*. However, only the raw log events can be selected for export in there. In addition, the achieved grades can be downloaded in Excel format from the Moodle *Gradebook*. A-plus includes a web service API that supports CSV and JSON formats. Submission data is available as one resource but aggregate numbers only exists as separate learner resources that would be accessed one by one.

5.1.3 Data Integration

As discussed in Chapter 2.3.4, integration of data from different sources is one major challenge in LA. In Chapter 3.1.2, we recognize integration as an unavoidable challenge in our research case. The expectation is that use of external learning resources becomes more important with new technology and new learning methods, such as problem based learning. We expect that integration of learning data from multitude of sources is an issue that does not disappear, on the contrary, it becomes critical.

In addition to opening new analytics possibilities, solving this issue also provides a single point of access for an external analytics tool as discussed in previous chapter. In Fact, *Learning Locker*[3] markets the connectivity with BI tools. As described in Chapter 2.4.3, it is an open–source LRS that can receive learning events from different applications. Apart from the connectivity, it provides itself an online user interface that supports design and hosting of custom analytics dashboards.

A standard data integration solution has further advantages. It is possible to replace or add an LMS or LRS component and keep the learning data intact and flowing. Furthermore, advanced analytics programs, such as machine learning models, may be developed to communicate with LRS using xAPI. This would be a step towards open learning analytics as discussed in Chapter 2.4.3.

Logstore xAPI[4] is a Moodle plugin that can directly connect with Learning Locker or other LRS implementations. A-plus does not support xAPI.

[3]https://learninglocker.net/

[4]https://moodle.org/plugins/logstore_xapi/

5.1.4 Software Components

This thesis develops missing software components for both Moodle and A-plus that solve the three previously described issues: real time visualization, external analytics tools, and integration of learning data. Figure 5.1 presents four new components and their relations to each other and their environment. In following, we describe the responsibilities of these software components.

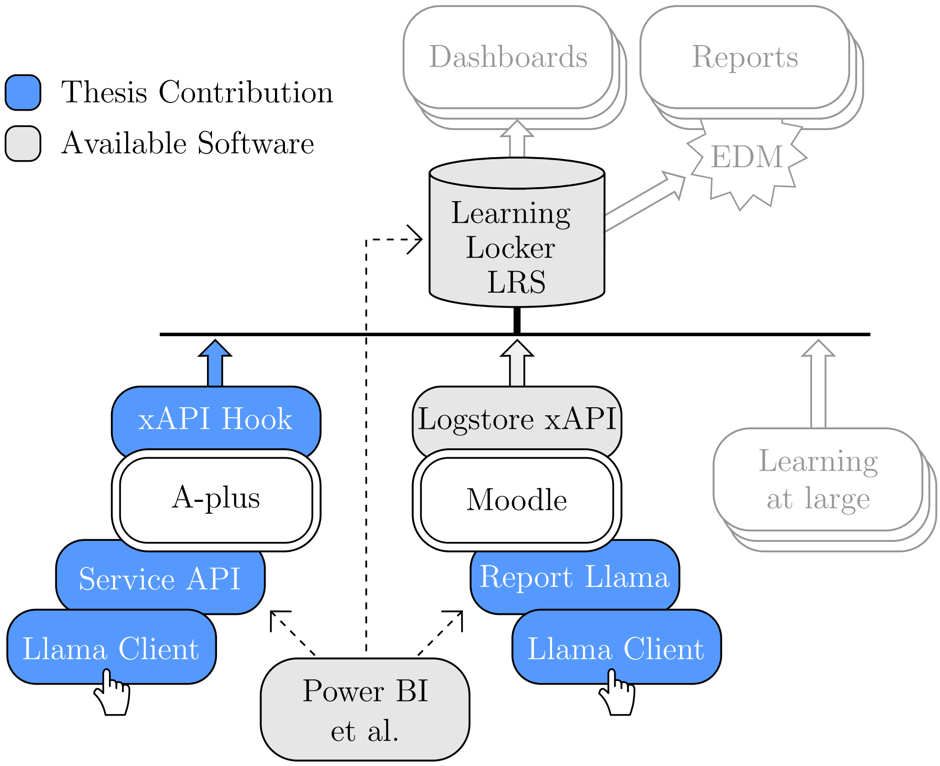

Figure 5.1: Thesis Contributions.

Interactive visualization in web is created with JavaScript program code that is running in the web browser. We develop such interactive program for learning data visualization that connects to a compatible web service API for learning data. This *Llama Client* is used for both A-plus and Moodle in our solution.

The API that Llama requires is provided separately for A-plus and Moodle to support the different learning data structures as discovered in Chapter 3. We extend the existing *Service API* of A-plus with a resource for aggregate data. For Moodle, we provide a *Report Llama* plugin that implements a compatible API for Llama Client that is packaged inside the plugin installation. In addition to servicing Llama, these APIs provide data in format that can be used in external analytics tools.

Finally, we propose that Aalto University would start an LRS where Moodle, A-plus, and other potential learning solutions can deliver learning data. As described in Chapter 5.1.3, this would open up new and exciting possibilities to develop LA at Aalto. In preparation, we extend A-plus design of hook URLs with *xAPI Hook* that supports integration to LRSs. Moodle has an existing plugin for the same purpose.

5.2 Llama Client

This interactive program for learning data visualization can be configured to fetch data from different web service APIs on different LMSs or other learning tools. This adds value to our solution as it is potentially easy to integrate to different platforms where courses may share similar analytics needs. This program is known as Llama Client. In addition to sympathetic pack animal, Llama is an abbreviation of "*la lumière à Montagne analytique*".

We start by describing the principles of system and user interface design. Then, we argument how different views fulfill previously identified user requirements.

5.2.1 Internal Design

From the different visualization libraries D3.js[5] is one of the most flexible and supports visualization that is reactive to changes in data. Our solution includes filters and selections of data that affect the visualization in real time. In addition, D3.js supports reading data in different formats, such as JSON and CSV, and uses Ajax technologies that allow transferring data from server to client without reloading the web page. It is a good match for our purpose and it creates vector visualizations that are accessible and reactive to different screen sizes.

To separate concerns and to support modular structure we created a D3.js support library called *d3Stream* which encapsulates the asynchronous requirements of data transfer and provides chaining of higher order functions for data transformation as well as visualization methods. The implemented transformations include functions, such as map, filter, reduce, cross, repeat, group, and cumulate.

The transformation chain can be split into several displays that will reapply their own transformations if the original data stream is filtered or else

[5] https://d3js.org/

```
var stream = new d3Stream()
  .load('/api/1234/aggregated-submission-statistics', {
    format: 'csv',
  })
  .filter(function (d) {
    return filterTagIds.containedIn(d.tagIds);
  });

stream.display('#learning-trajectories-3-chapters')
  .cross([ '1.1', '1.2', '1.3' ]) // Cartesian cross product
  .mapAsStreams(function (row) {
    return row.map(function (pair, i) {
      return {
        x: i,
        y: +pair[0][pair[1] + '_total'],
        z: +pair[0][pair[1] + '_count'],
        payload: pair[0],
      };
    })
    .cumulate('y');
  })
  .lineChart();

stream.display('#number-of-submitters-3-chapters')
  .repeat([ '1.1', '1.2', '1.3' ]) // Whole set repeated
  .map(function (pair, i) {
    var group = new d3Stream(pair[0]).filter(function(d) {
      return +d[pair[1] + '_count'] > 0;
    }).array();
    return {
      x: i,
      y: group.length,
      z: 0,
      payload: group,
    };
  })
  .barChart();

d3.select('button#remove-filters').on('click', function () {
  stream.reset(); // Displays update automatically
});
```

Figure 5.2: JavaScript Program Using d3Stream Library.

updated. Figure 5.2 presents a sample JavaScript program using d3Stream. Finally, once the data is in both supported and desired format it can be trivially visualized with one of the visualization functions, such as scatterPlot, lineChart, barChart, stackedBarChart, or groupedBarChart. Using default options d3Stream creates clean, ascetic visualization whose appearance is primarily controlled via CSS (Cascading Style Sheet).

The library helps to keep actual application logic cleaner and to avoid the unstructured spaghetti code that event driven JavaScript programs can quickly generate [Mikkonen and Taivalsaari, 2008]. The design is similar to different reactive libraries that use the observer and functional programming patterns to contain asynchronous or user interface processes [Kambona et al., 2013]. We believe the resulting code is, in addition to shorter, more comprehensible than when directly using D3.js library.

On top of the D3.js and d3Stream layer, Llama Client implements configuration of API data sources, filters and visualizations. It employs callback functions supported by d3Stream to implement interactive data selections. We employ jQuery[6] JavaScript library to support the event processing and DOM (Document Object Model) modification that the user interface requires. The jQuery is included in both A-plus and Moodle by default.

Finally, the JavaScript code on both d3Stream and Llama is broken into small files that resemble classes of object oriented programming. This is purely to support development and maintenance of the project code. Tool configuration is included to bundle, test, and minimize the JavaScript libraries as single deployment files for the browsers. Browserify[7] is used for the bundling.

Both of the contributed JavaScript libraries d3Stream[8] and Llama Client[9] are developed in GitHub. They are open–source under MIT license. However, the Llama alone does not finish our task. It has to be complimented with an service API and packaging for the two LMSs at hand. Before discussing that, we present the user interface of Llama.

5.2.2 User Interface Design

In data visualization design this thesis follows a model that Munzner [2014] has presented for visualization design and validation. We have already discussed the visualization domain that is the courses in our research case and their staff who expressed requirements in the interviews.

Next, we consider the 'What' and 'Why' as Munzner's model proceeds. For each visualization we discuss what data is used and why it is viewed. The expected action and target of the viewer leads the selection of visualization

[6] https://jquery.com/
[7] http://browserify.org/
[8] https://github.com/debyte/d3Stream/
[9] https://github.com/Aalto-LeTech/llama-client/

method as different visualizations are better for different action–target pairs, such as comparing trends or locating features.

Finally, for each view we discuss the 'How' in Munzner's model. The visual encoding for data is selected to support the visualization goal and clarity. We follow the inheritance of Edward Tufte in minimizing the visual clutter and focusing on data and integrity [Tufte and Graves-Morris, 2014]. In our case, the interaction is a major part of the 'How' consideration.

5.2.3 Collective Progress

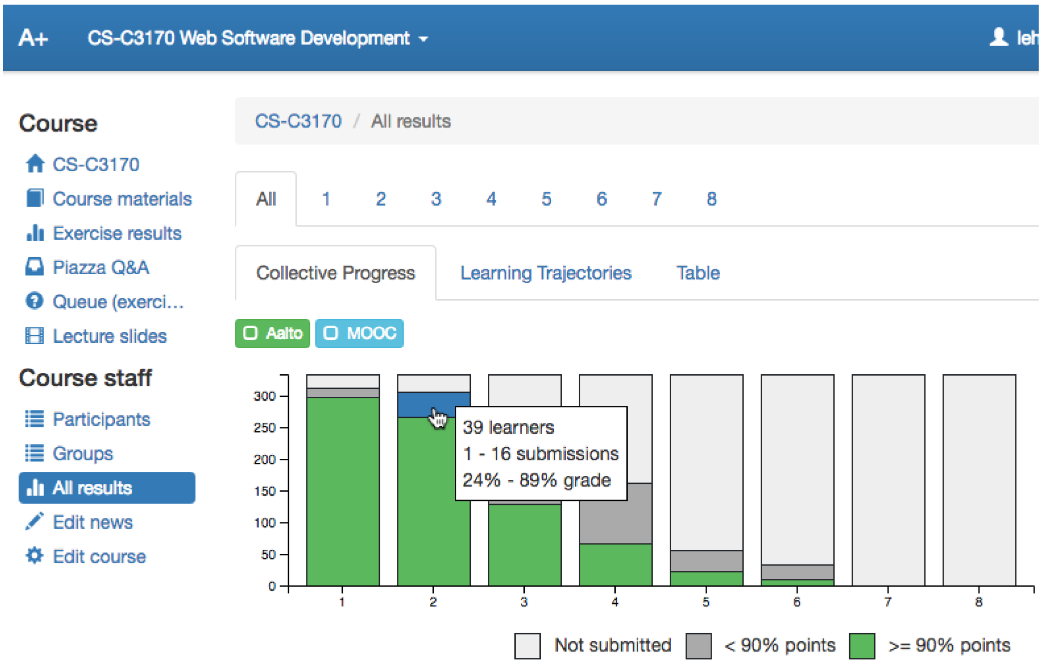

Figure 5.3: Llama: Collective Progress.

The default view of Llama client, that is presented in Figure 5.3, aims to answer the user requirement R1: Educators monitor learners' progress and effort to verify learning and learners existence in real time. Furthermore, it supports the requirements R4: Educators filter learning analytics results by units of study and learner demographics, and R5: Educators navigate from learning analytics results to individual learners and their activity record.

The data is the number of learners who have started or, conversely, not started a given unit of study. The interviews revealed, that in our case the learners typically complete an exercise with full grade or do not start at all. Using this logic, we separate learners who have completed the study unit with near full points, e.g. 90%. This helps to detect the small group of learners who did not really attempt to solve the complete unit or run into serious problems. The level of separation should be later adjusted according to staff experience.

In this view, educators are expected to primarily compare similarity of learner numbers between the study units. A bar chart delivers true and comparable sizes of variable when bottoms of the bars are located at the same level. This principle is broken for the small group of learners who got a low grade. They are stacked on the top of the group who got a high grade. It is still easy to notice if this area grows significantly in relation to the high grade group. In addition, the continued grey bar to the top of the chart presents the inactive learners.

The data can be filtered by selecting different study units or learner labels that are available. When cursor is located on top of a bar area the exact numbers are displayed in text. Furthermore, the visualized bars are selectable and the learners in the selected groups appear as small individual learner beacons below the main visualization. The learner beacons are introduced later.

5.2.4 Learning Trajectories

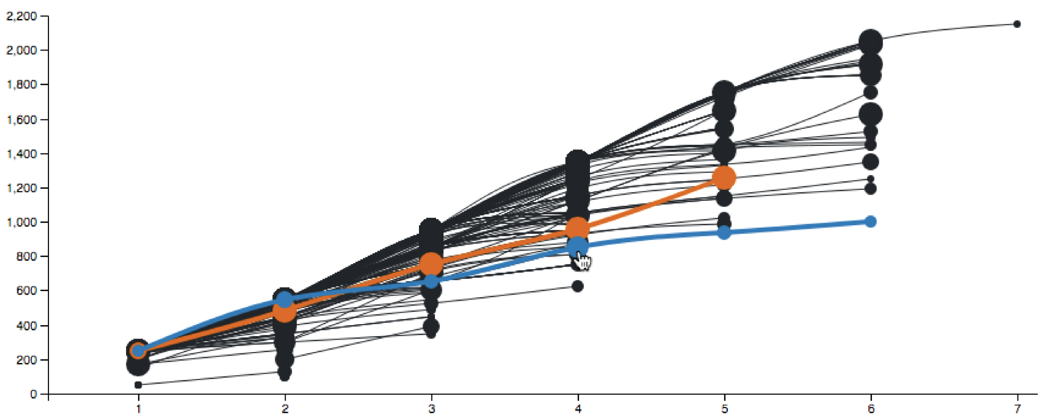

Figure 5.4: Llama: Learning Trajectories.

The trajectory view, that is presented in Figure 5.4, visualizes individual learner's cumulative grade on each unit of study. A line chart is a good and familiar choice to visualize change over time. Human vision is trained at detecting differences of line alignment [Ware, 2012]. In this case, line chart allows to view both collaborative and individual trends in a same view and compare them to each other. This supports the user requirement R3:

Educators detect both deviations and trends of both learners' progress and effort to improve learning material.

This visualization employs position and hue to respectively represent the grade differences and to identify selected cases. Those visual variables are most accurately perceived for quantitative and nominal data [Mackinlay, 1986]. Furthermore, these variables depend on preattentive and separable visual features so human brain makes sense of the visualization elements before conscious thought [Ware, 2012]. Therefore, we believe this visualization is a powerful tool to inspect even large numbers of learners in relation to each other.

In addition to grade, the number of attempts is visualized as size of dots on trajectory lines. This visualization of effort is only able to represent large differences and primarily on selected cases. This sacrifice was made in order to minimize clutter on the main features as described above. The second weakness in the visualization is that it does not represent real time. The x–axis represents the units of study on the course and it does not express at which time an individual learner worked on a given unit. Such time aware visualization sets different requirements for the supporting service API.

Different learners can be selected and deselected by clicking at the trajectories. The same filters are available as in the progress view. Furthermore, the selections including the selected learners remain active from the progress view. This allows to examine learning trajectories of previously selected learner groups in relation to the rest of the learners.

5.2.5 Learner Beacon

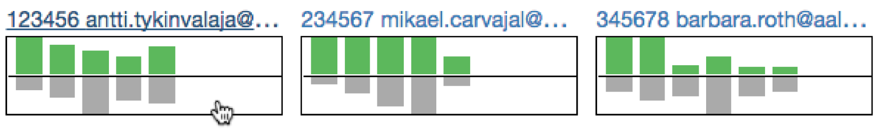

Figure 5.5: Llama: Learner Beacon.

The small visualizations, that are presented in Figure 5.5, aim to answer the user requirement R6: Educators digest real time summaries of individual learner's effort and progress in different units of study to improve interaction with learners. This visualization is like a fingerprint or beacon that has a shape that educators should start to recognize in order to quickly identify different learner behaviors.

The data includes individual learner's grades and attempts per unit of study. The aim of the educator is to both locate deviations and identify trend. The data includes two variables that change over the units of study on the course. A line chart would be the preferred choice of visualization as in the trajectory view. However, the visualization should be so small that it can be integrated into different views where learners are referenced. Two variables in a line chart would become incomprehensible in a small size and when the variables share a similar trend.

As a solution, the variables are presented as two bar charts where grades, or progress, is presented upwards and attempts, or effort, is presented downwards from the base line. The deviations, ultimately a missing bar, are easy to spot even in a small view. This visualization can show interesting imbalances where great effort yields little progress or, conversely, great progress is displayed with minimal effort. The weakness of this visualization is that trend becomes hard to identify if the view is too small.

5.3 Service APIs

Llama Client connects to a web service API that is responsible for providing the data to interactively visualize. Next, we discuss the challenges involved in developing and operating such real time service.

In this case, the largest course has 644 learners. Furthermore, it has 351 different units of study, if we consider exercises, chapters, and modules. One instance of the course includes more than 100 000 submissions. The full data includes all of these timestamped events which potentially takes minutes to query from database and format for transfer in the API.

To solve this problem, we decided to service aggregate data in API. However, if 3 columns are created for each of the 351 study units, there are more than 1 000 columns in a data sheet. That exceeds for example the maximum number of columns that Excel supports. More importantly, just calculating and rendering such table in CSV takes at least similar time to rendering all the individual submissions. Therefore, the study unit filters are critical to implement so that they limit the size of the database queries and maximum number of columns to render. As a result, we define a web service API in Appendix B that is required to support Llama Client.

In implementation for both A-plus and Moodle, we follow closely the standards enforced in these systems. There are no novel ideas. At the time of writing, the implemented API in Moodle supports Quizz and Assignment Activities. Technically, the load of calculating real time aggregates falls on database systems in these LMS installations. Figure 5.6 presents sample code that makes the Django ORM use aggregate functions at the database level.

```
Submission.objects \
  .filter(exercise__course_module=filter_module) \
  .values('submitters__user_id', 'exercise') \
  .annotate(total=Max('grade'), count=Count('id')) \
  .order_by()
```

Figure 5.6: Django Aggregation Queryset.

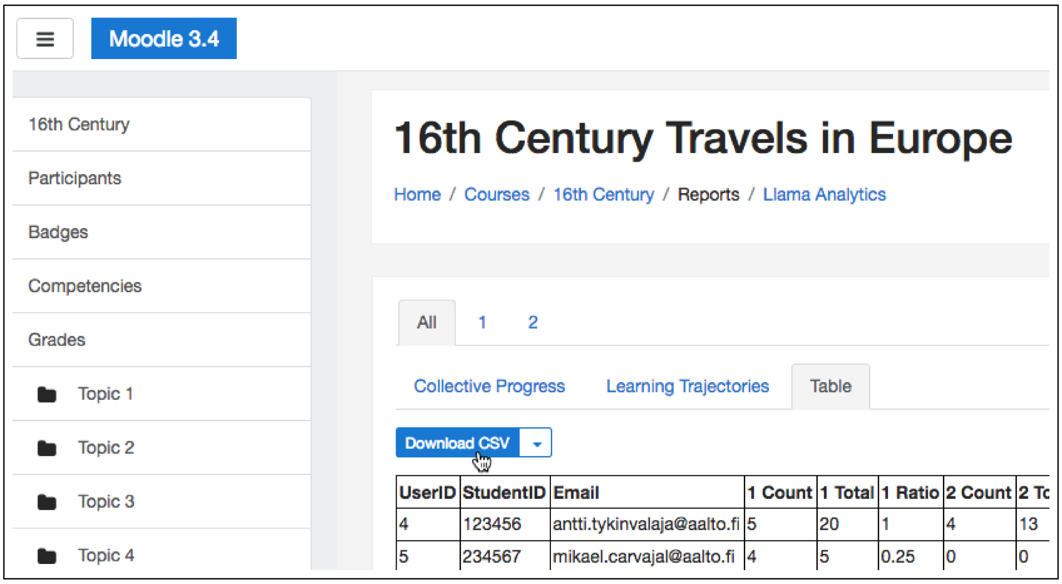

Figure 5.7: Llama: Data Link.

A Moodle plugin *Report Llama*[10] that packages both API and Llama Client is developed in GitHub under MIT open–source license. The contribution to A-plus is delivered in a GitHub pull request[11]. Figure 5.7 presents the Moodle plugin and the link to the raw API data.

5.4 xAPI hook

This thesis extends A-plus hook URLs to support LRS integration. The hook design is extensible to new events. The challenge is what events should be hooked and how they are represented in xAPI.

The correct mapping to xAPI statements ensures that the delivered learning data is not only technically but also conceptually ready for integration.

[10]https://github.com/Aalto-LeTech/moodle-report_llama
[11]https://github.com/Aalto-LeTech/a-plus/pull/306

In our case, this signifies that equivalent actions in A-plus and Moodle deliver same type of events to the common LRS. Therefore, they can be included in analytics that span courses from both A-plus and Moodle.

The xAPI developers manage a recipe registry for this purpose. Appendix C presents the implemented A-plus event mapping to xAPI statements. The xAPI hook feature is proposed to the A-plus in a GitHub pull request[12]. Figure 5.8 presents the xAPI configuration interface in A-plus.

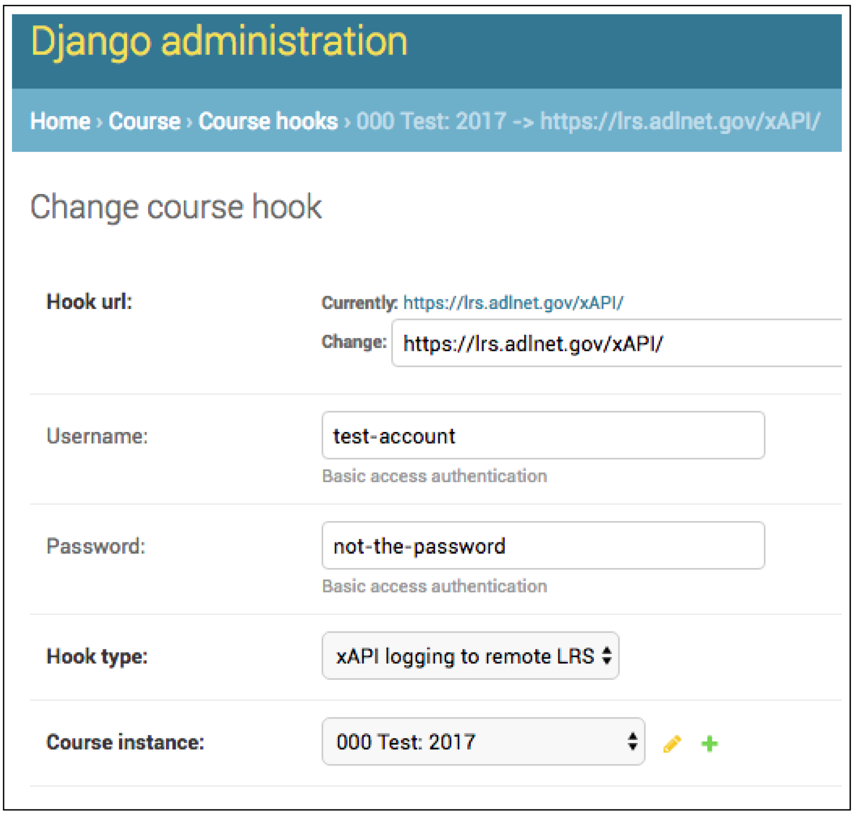

Figure 5.8: A-plus xAPI configuration.

[12] https://github.com/Aalto-LeTech/a-plus/pull/307

Chapter 6

Evaluation

This chapter critically evaluates that the solution is useful and an improvement over previously available alternatives. Completion of each research goal, that were set in Table 1.2, is evaluated. First, the access to learning data is considered. Second, the support for the identified learning analytics objectives is evaluated. Third, the maintainability and extendability of the solution is examined.

6.1 Access to Learning Data

In order to practice, develop, and research LA, access to learning data is a minimum and critical requirement. This issue is enclosed in the first research goal, RG1: Course staff and researchers can effortlessly access collected learning data. This goal describes two stakeholder groups, course staff and researchers, which have different access requirements. We examine this access goal from their different points of view.

The course staff has expressed requirement of interactive analytics that are part of their daily workflow. This thesis designs a real time and interactive visualization tool, Llama Client, to fulfill this purpose. It is available on both LMSs in this research case and thus course staff has effortless access to learning data. The more specific objectives for that tool are evaluated later in LA objectives.

Developers and researches of LA require access to raw learning data. This thesis considers use of an external analytics tool and includes a service API in the solution that supports this purpose. This API can provide both download of data files and programmatic data access. Table 6.1 presents the improvements this solution provides. It adds a novel aggregate data resource for A-plus and Moodle. The programmatic access is mainly a potential improvement for software developers and other stakeholders may prefer download. Generally, ability to use external tools is a useful accelerator for practicing and developing LA as it removes non–LA requirements from a person to work on such tasks.

Table 6.1: Feature Upgrades for External Analytics Tools.

	Observations	Aggregate Data
Moodle	download	download
A-plus	API	-
New	-	API

In addition to previous access considerations, the solution provides support for integrating data from both A-plus and Moodle to one standardized data storage. A novel integration feature is contributed to A-plus. While this feature does not provide immediate value we expect it to become a standard requirement in the future.

6.2 Learning Analytics Objectives

This thesis interviewed course staff to identify LA objectives. The analysis of the interviews encoded the objectives as user requirements. The ability to answer the user requirements is evaluated via inspection of the second research goal, RG2: Course staff can efficiently complete their initial LA objectives.

Table 6.2: Feature Upgrades for Real Time Visualization.

	Effort: R1,R6	Progress: R1,R6	Filter: R4	Navigation: R5
Gismo	chart	color matrix	one by one	-
A-plus	-	table	-	-
New	chart	chart	complete	complete

In architectural design, this thesis evaluated how existing real time visualization features supported the user requirements. Table 6.2 presents the delivered improvements in comparison to that previous support. For A-plus, all these features are novel and enable course staff to interactively study learner behaviour for the first time on A-plus courses. In Moodle, the solution improves the interactive features: filters and navigation.

However, the solution should be evaluated against the user requirements and not only the previous features. Table 6.3 presents which features of the solution answer to which user requirements.

Table 6.3: Features for Different User Requirements.

R1	Educators monitor learners' progress and effort to verify learning and learners existence in real time. *Collective Progress*
R2	Educators measure learners' time usage to allocate learning material into units of study and calendar. *External Analytics Tool*
R3	Educators detect both deviations and trends of both learners' progress and effort to improve learning material. *Learning Trajectories, External Analytics Tool*
R4	Educators filter learning analytics results by units of study and learner demographics. *Collective Progress, Learning Trajectories*
R5	Educators navigate from learning analytics results to individual learners and their activity record. *Collective Progress, Learning Trajectories*
R6	Educators digest real time summaries of individual learner's effort and progress in different units of study to improve interaction with learners. *Learner Beacon*

All of the requirements are supported. However, R2 is only supported in external analytics tools instead of the real time and interactive visualizations provided in Llama Client. The efficiency to fulfill the R2 is questionable as it requires setting up analytics tool and learning to apply it to this objective.

Considering all the other user requirements, the developed solution is useful and efficient. However, it should be noted that the user requirements only expressed the popular initial wishes. We have described LA as a continuous and developing process so new requirements are expected and new features should be developed to the solution. Thus, the next research goal discusses extendability.

6.3 Maintainability and Extendability

Non–functional requirements are included in the evaluation of the third research goal, RG3: Software developers can readily maintain and extend the solution to provide further modeling and analysis of learning data in real time.

Modular and reusable design enables to use the same software component, Llama Client, in different learning environments. It requires a documented service API that separates the concerns of data collection and aggregation from the interactive visualization. This improves maintainability and supports extendability to new systems.

The utilized and developed d3Stream–library provides powerful data transformations through higher order functions. Furthermore, it encloses drawing code so visualization of different variables using the existing chart types is available using minimal and clean JavaScript–code. The organization of client code into class size JavaScript files further improves maintainability. New Llama views can be implemented as new JavaScript files enclosing the view logic.

The contributions to A-plus and Moodle systems follow the design standards these projects establish. Maintaining these parts requires good understanding of these systems. Both of the systems have gathered weight of past and sometimes obsolete decisions that raise the learning curve to contribute in these projects.

We evaluate that the maintainability and extendability of visualizations is good. The service APIs have similar qualities as the two LMSs themselves. In the development of new analytics methods this solution recommends external code that accesses data using the developed API standard or LRS integration. This adds to maintainability and extendability of the new analytics code that remains independent from the LMS platforms.

Chapter 7

Conclusions

This chapter concludes the thesis work. First, knowledge acquired in this research case and contribution to the domain knowledge are considered. Finally, future work is discussed.

7.1 Acquired Knowledge

In design–science research, knowledge on the research problem and the solution is extracted in the creation of an artifact, which is the software solution in this thesis. First, interesting findings in this research case are presented. Then, we discuss the transferability to other cases.

In this case, learning data as stored in the two different LMSs raised different and unique technical requirements. In comparison, the interviews of different teachers found similar objectives that were then derived into user requirements. We believe the following two findings in this thesis are relevant for domain knowledge.

- The integration of learning data from multitude of sources is a common challenge that needs design.
- Teachers' initial LA objectives include aims to monitor expected progress, improve allocation of learning material, identify problematic areas in learning material, and improve interaction with learners.

Retrospectively, these were the guiding principles to the partly novel and partly improved solution in this research case. We bootstrapped LA in four courses that implement different weekly online learning activities. The courses had large number of students and they embraced automatic assessment. Transferability of this knowledge to other cases should be carefully evaluated. However, when starting the practice of LA or developing LA features it is, at minimum, useful to consider these findings.

7.2 Future Work

The goal of this thesis was to bootstrap LA with expectation of future work. We consider LA as a cyclic process that should develop itself on each iteration. Therefore, the LA task is far from finished. However, the future work should be designed with the accumulated experience in each iteration. Ali et al. [2012] used qualitative evaluation to design and confirm improvements on an LA tool.

First, the solution should be evaluated in practice. When the pilot courses start they should use the provided solution and the course staff should report their experience for the next iteration of the LA process. Such iterative process can answer to the expected usability improvement ideas as well as completely new LA objectives.

Some useful work could not be completed in the scope of this thesis. The user requirement to measure learners' time usage was not met in the real time visualizations. In addition, an option to use self reported time as a measure of effort instead of the number of attempts would be useful. These are likely improvement requirements in the future.

Additionally, the interviews included some lonely quotes that were not included in the user requirements. They involved calendar heat maps, comparison to previous course instances, solution paths of multiple choice questionnaires, animated learning paths, summary of discussion topics, estimates, sampling of student groups, uploading grade data, and integration of external study records. These provide possible ideas for future LA development and research.

Alternatively, the presented novel visualization elements, such as the Learner Beacon, can be researched further. Different evaluations with teachers and interactions in LMS can be designed. This allows systematic development of a chosen element. Finally, the LA interviews can be extended to new stakeholders, new methods, and larger populations to improve understanding of LA requirements at different stages of investing into the practice of LA.

References

Aalto Online Learning. A!ole pilots. https://onlinelearning.aalto.fi/pilots/, 2017. Accessed: 2017-09-27.

Liaqat Ali, Marek Hatala, Dragan Gašević, and Jelena Jovanović. A qualitative evaluation of evolution of a learning analytics tool. *Computers & Education*, 58(1):470–489, 2012.

Apereo. Learning analytics initiative. https://www.apereo.org/communities/learning-analytics-initiative, 2017. Accessed: 2017-08-28.

Tapio Auvinen. *Educational technologies for supporting self-regulated learning in online learning environments*. PhD thesis, Aalto University, School of Science, 2015.

Ryan SJd Baker and Kalina Yacef. Editorial welcome. *JEDM-Journal of Educational Data Mining*, 1(1):1–3, 2009.

Aneesha Bakharia, Linda Corrin, Paula de Barba, Gregor Kennedy, Dragan Gašević, Raoul Mulder, David Williams, Shane Dawson, and Lori Lockyer. A conceptual framework linking learning design with learning analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 329–338. ACM, 2016.

Sanam Shirazi Beheshitha, Marek Hatala, Dragan Gašević, and Srećko Joksimović. The role of achievement goal orientations when studying effect of learning analytics visualizations. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 54–63. ACM, 2016.

Harvey Russell Bernard. *Social research methods: Qualitative and quantitative approaches*. Sage, 2012.

Blackboard Inc. Case study: Lewis & clark community college – analytics for student retention. http://bbbb.blackboard.com/LCCCCaseStudy, 2017. Accessed: 2017-08-27.

Mohamed Amine Chatti, Anna Lea Dyckhoff, Ulrik Schroeder, and Hendrik Thüs. A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5-6):318–331, 2012.

Mohamed Amine Chatti, Arham Muslim, and Ulrik Schroeder. Toward an open learning analytics ecosystem. In *Big Data and Learning Analytics in Higher Education*, pages 195–219. Springer, 2017.

Doug Clow. The learning analytics cycle: closing the loop effectively. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 134–138. ACM, 2012.

Hendrik Drachsler and Wolfgang Greller. Privacy and analytics: it's a delicate issue a checklist for trusted learning analytics. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 89–98. ACM, 2016.

Nicola Dragoni, Saverio Giallorenzo, Alberto Lluch Lafuente, Manuel Mazzara, Fabrizio Montesi, Ruslan Mustafin, and Larisa Safina. Microservices: yesterday, today, and tomorrow. In *Present and Ulterior Software Engineering*, pages 195–216. Springer, 2017.

Rebecca Ferguson. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5-6):304–317, 2012.

Dragan Gasevic, Negin Mirriahi, Phillip D Long, and Shane Dawson. Editorial–inaugural issue of the journal of learning analytics. *Journal of Learning Analytics*, 1(1):1–2, 2014.

Dragan Gašević, Shane Dawson, and George Siemens. Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1):64–71, 2015.

Shirley Gregor and Alan R Hevner. Positioning and presenting design science research for maximum impact. *MIS quarterly*, 37(2), 2013.

Wolfgang Greller and Hendrik Drachsler. Translating learning into numbers: A generic framework for learning analytics. *Educational technology & society*, 15(3):42–57, 2012.

Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS quarterly*, 28(1):75–105, 2004.

Kennedy Kambona, Elisa Gonzalez Boix, and Wolfgang De Meuter. An evaluation of reactive programming and promises for structuring collaborative web applications. In *Proceedings of the 7th Workshop on Dynamic Languages and Applications*, page 3. ACM, 2013.

Ville Karavirta, Petri Ihantola, and Teemu Koskinen. Service-oriented approach to improve interoperability of e-learning systems. In *Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on*, pages 341–345. IEEE, 2013.

Tomi Kauppinen and Lauri Malmi. Aalto online learning-a pathway to reforming education at the aalto university. In *Proceedings of EUNIS 2017 – Shaping the future of universities*. Münster, June, 2017.

Tomi Kauppinen, Johannes Trame, and A Westermann. Teaching core vocabulary specification. *LinkedScience. org, Tech. Rep*, 2012.

Jonathan M Kevan and Paul R Ryan. Experience api: Flexible, decentralized and activity-centric data collection. *Technology, Knowledge and Learning*, 21(1):143–149, 2016.

Kirsty Kitto, Sebastian Cross, Zak Waters, and Mandy Lupton. Learning analytics beyond the lms: the connected learning analytics toolkit. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 11–15. ACM, 2015.

Ari Korhonen and Jari Multisilta. Learning analytics. In *New Ways to Teach and Learn in China and Finland: Crossing Boundaries with Technology*, pages 301–310. Peter Lang – Frankfurt am Main · Bren · Bruxelles · New York · Oxford · Warszawa · Wien, 2016.

Teemu Koskinen. Improving interoperability of e-learning systems by using a service-oriented approach. Master's thesis, Aalto University, School of Science, 2012.

Sonja Krogius. A+-palvelu. kohti prosessiorientoitunutta ja opiskelijakeskeistä ohjelmoinnin opetusta. Master's thesis, Aalto University, School of Arts, 2012.

Yvonna S Lincoln and Egon G Guba. *Naturalistic inquiry*, volume 75. Sage, 1985.

LISTedTECH. Lms overview of market share. http://listedtech.com/historical-lms-overview/, 2015. Accessed: 2017-08-27.

Jock Mackinlay. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, 5(2):110–141, 1986.

Martin Maguire. Methods to support human-centred design. *International journal of human-computer studies*, 55(4):587–634, 2001.

Tommi Mikkonen and Antero Taivalsaari. Web applications: Spaghetti code for the 21st century. In *Proc. 6th ACIS International Conference on Software Engineering Research, Management, and Applications (SERA'08)*, pages 319–328. IEEE Computer Society Press, 2008.

Yishay Mor, Rebecca Ferguson, and Barbara Wasson. Editorial: Learning design, teacher inquiry into student learning and learning analytics: A call for action. *British Journal of Educational Technology*, 46(2):221–229, 2015.

Tamara Munzner. *Visualization analysis and design*. CRC press, 2014.

Meghan Oster, Steven Lonn, Matthew D Pistilli, and Michael G Brown. The learning analytics readiness instrument. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 173–182. ACM, 2016.

Zacharoula K Papamitsiou and Anastasios A Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, 17(4):49–64, 2014.

Zachary A Pardos and Kevin Kao. moocrp: An open-source analytics platform. In *Proceedings of the Second (2015) ACM conference on learning@ scale*, pages 103–110. ACM, 2015.

Markku Riekkinen. Integrating stratum and a+ functionalities in moodle: Architecture and evaluation. Master's thesis, Aalto University, School of Science, Finland, 2017.

Cristobal Romero and Sebastian Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1): 12–27, 2013.

Cristóbal Romero, Sebastián Ventura, and Enrique García. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1):368–384, 2008.

Niall Sclater. Developing a code of practice for learning analytics. *Journal of Learning Analytics*, 3(1):16–42, 2016.

George Siemens. Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10):1380–1400, 2013.

George Siemens and Ryan SJd Baker. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254. ACM, 2012.

Ian Sommerville. *Software Engineering: Ninth Edition*. Pearson Education Limited, 2011.

Melanie Swan. The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, 1(2):85–99, 2013.

Edward Tufte and P Graves-Morris. The visual display of quantitative information.; 1983, 2014.

Kalyan Veeramachaneni, Franck Dernoncourt, Colin Taylor, Zachary Pardos, and Una-May O'Reilly. Moocdb: Developing data standards for mooc data science. In *AIED 2013 Workshops Proceedings Volume*, page 17, 2013.

Colin Ware. *Information visualization: perception for design*. Elsevier, 2012.

Françeska Xhakaj, Vincent Aleven, and Bruce M McLaren. How teachers use data to help students learn: Contextual inquiry for the design of a dashboard. In *European Conference on Technology Enhanced Learning*, pages 340–354. Springer, 2016.

Appendix A LA Feature Samples

Screen capture from each of the following media, as accessed 7th November 2017, were prepared for demonstration.

T2 – Temporal Analytics

1. https://youtu.be/qgu8GpQw9F8?t=27s
2. https://goo.gl/KznCBm
3. https://goo.gl/3cTLGy
4. https://goo.gl/3cTLGy
5. https://youtu.be/hcWfSZ_E8P4?t=10m41s

T3 – Progress Analytics

1. https://youtu.be/qgu8GpQw9F8?t=29s
2. https://goo.gl/1NCG9n
3. https://goo.gl/AD7Up9

T4 – Learner State Analytics

1. https://youtu.be/VZv9OCq0TlM?t=16m1s
2. https://youtu.be/VZv9OCq0TlM?t=14m8s
3. https://goo.gl/unkk1q
4. https://goo.gl/ib5kZV
5. https://goo.gl/QXwCHA
6. https://youtu.be/yeJwXhu_bVQ?t=21m19s

T5 – Social Interaction Analytics

1. https://youtu.be/6wTMDpqPg8w?t=24m18s
2. https://youtu.be/6wTMDpqPg8w?t=26m18s

Appendix B Service API for Llama Client

REQUEST	
Authorization	Access by session cookie
Method	GET
URL	Configurable, eg /api/v1/aggregate[/*filter*]
Either URL paths or query parameter must be provided to filter the results by unit of study. If no filter is provided the whole course should be summarized, e.g. at module level.	

RESPONSE	
Content-type	application/json OR text/csv
BODY is either a list of the following JSON objects OR a CSV table of title header and the following rows.	
UserID	Unique identifier
StudentID	Displayed student identifier
Email	Displayed student email
Tags	Optional student tags that can be used to filter data rows
1 Count	The number of submission attempts by the user in the unit
1 Total	The achieved total grade sum by the user from the unit
1 Ratio	The ratio of the total grade from the maximum grade
⋮	
N Count	– ” –
N Total	– ” –
N Ratio	– ” –
Each study unit forms a triplet of columns where the prefix before space character is an identifier for one study unit. The identifier is displayed and should be short, e.g. numeric. Additionally, the same identifier must be supported as a filter in the request. When filter is used, e.g. ”1” for the first study module, the aggregation granularity and the column triplets typically change, e.g. ”1.1”, ”1.2”, . . . ”1.M” as exercises inside the module that were previously summarized in a study unit ”1”.	

Appendix C A-plus Events to xAPI Statements

Learning material viewed

```
{
    "id": "495fcc38-d165-11e7-aa5c-040ccede6c42",
    "verb": {
        "id": "http://id.tincanapi.com/verb/viewed",
        "display": { "en": "viewed" }
    },
    "object": {
        "definition": {
            "type": "https://apluslms.github.io/type/exercise",
            "name": { "en": "Ex. Name" },
            "description": { "en": "Ex. Description" }
        },
        "id": "https://plus.cs.hut.fi/o1/2017/k08/part01/ex1",
        "objectType": "Activity"
    },
    "actor": {
        "mbox": "mailto:nobody@no.zzz",
        "name": "First Last",
        "objectType": "Agent"
    },
}
```

Exercise submitted

```
{
    "id": "b3c94000-d164-11e7-bb22-040ccede6c42",
    "verb": {
        "id": "http://adlnet.gov/expapi/verbs/completed",
        "display": { "en": "completed" }
    },
    "object": INDENTICAL-TO-FIRST,
    "actor": IDENTICAL-TO-FIRST,
    "result": {
        "completion": true,
        "score": {
            "raw": 7,
            "max": 10,
            "scaled": 0.7,
            "min": 0
        },
        "response": "[[\"a\", \"answer\"], [\"b\", \"...\"]]"
    },
}
```