Ville Heilala

Framework for Pedagogical Learning Analytics

Master's thesis of mathematical information technology

April 17, 2018

University of Jyväskylä

Faculty of Information Technology

Author: Ville Heilala

Contact information: ville.s.heilala@student.jyu.fi

Supervisor: Professor Tommi Kärkkäinen

Title: Framework for Pedagogical Learning Analytics

Työn nimi: Pedagogisen oppimisanalytiikan viitekehys

Project: Master's thesis

Study line: Educational Technology

Page count: 61

Abstract: Learning analytics is an emergent technological practice and a multidisciplinary scientific discipline, which goal is to facilitate effective learning and knowledge of learning. In this design science research, I combine knowledge discovery process, concept of pedagogical knowledge, ethics of learning analytics and microservice architecture. The result is a framework for pedagogical learning analytics. The framework is applied and evaluated in the context of agency analytics. The framework contributes to the practical use of learning analytics.

Keywords: pedagogical learning analytics, pedagogical knowledge, student agency, knowledge discovery, ethics, GDPR, microservice

Avainsanat: pedagoginen oppimisanalytiikka, pedagoginen tieto, opiskelijan toimijuus, etiikka, GDPR, mikropalvelu

Glossary

ALA	Automated Learning Analytics
AUS	Agency of University Students
DSRP	Design Science Research Process
EDM	Educational Data Mining
FEDS	Framework for Evaluation in Design Science
FPLA	Framework for Pedagogical Learning Analytics
GDPR	General data Protection Regulation
IEDMS	International Educational Data Mining Society
LA	Learning Analytics
LAK	Learning Analytics and Knowledge Conference
PPEDD	Protection, Privacy and Ethics by Design and by Default
PPK	Pedagogical and Psychological Knowledge
SOC	Service-Oriented Computing
SoLAR	Society for Learning Analytics Research
UML	Unified Modelling Language

List of Figures

List of Tables

Contents

1 Introduction

Several authoritative organizations have listed important global issues that humanity is facing already in the present (e.g. United Nations 2015; WCED 1987). Especially teachers face complex challenges (e.g. UNESCO 2017). Teachers have to help students to achieve their full potential and to become members of 21st century society in a complex and uncertain environment. Many recent reports and studies claim, that educational systems and the nature of teaching profession are in the midst of a major change (Davis 2017; Day 2017; Guerriero 2017; Krokfors et al. 2015), not least because of rapid evolution of technology. Technological development is one of the most challenging change agent teachers have to take over (Villegas-Reimers 2003).

But what are the possibilities of technology in teaching and learning? How could technology support educators and learners in a volatile, uncertain, complex and ambiguous world? Can technology contribute beneficial change through education? In this thesis I begin outlining my own solution space and reasoning by starting with an area called learning analytics.

1.1 Motivation

Learning analytics is an emergent technological practice and a multidisciplinary scientific discipline, which goal is to produce effective learning and knowledge of learning. Despite recent efforts, learning analytics has not yet managed to redeem its promises (e.g. European Comission 2016; Ferguson and Clow 2017). There exists a significant gap between learning analytics and evidence of its effectiveness (Ferguson, Brasher, et al. 2016). Hoel and Chen (2016) comment also on the fact that there is a gap between concerns and challenges of ethical implementations of learning analytics and proposals for design to solve these important issues.

In my research, I combine traditional knowledge discovery process, concept of pedagogical knowledge, ethics of learning analytics, and microservice architecture. The conceptual basis for this research is what I call as pedagogical learning analytics (Figure 1). The concept of pedagogical learning analytics is new and only one article by Wise (2014) was found and it discusses about “pedagogical learning analytics intervention design”. Also, Greller and Drachsler (2012) examine the place of pedagogy on learning analytics.

1.2 Research questions

RQ1: What kind of useful knowledge a teacher could obtain using learning analytics?

RQ2: What are the ethical challenges in learning analytics process?

RQ3: Is it possible to automate learning analytics process?

1.3 Objectives of the solution

The objective of my research is to sketch a framework for providing novel and meaningful pedagogical knowledge for teacher in automated and ethical way. The framework is applied and evaluated in a scenario of analyzing university student agency. Figure 1 describes the conceptual model of this system. In the center of the conceptual model is our understanding of learning processes. Human agency is a fundamental part of learning (Jääskelä, Poikkeus, Vasalampi, Valleala and Rasku-Puttonen 2016). Thus, it is applied as a core concept for analysis.

Designing automated and ethical learning analytics consists of solving ethical, analytical and automation related issues. Automated and ethically conducted learning analytics could provide novel and meaningful knowledge for teachers, when applied using relevant knowledge about learning processes. I call this kind of analytics as *pedagogical learning analytics*. It can be presented as a process cycle (Figure 1), which is synthesized in this research and forms the basis for the framework. The result design artifact of this research is a learning analytics framework for providing pedagogical knowledge to teachers.

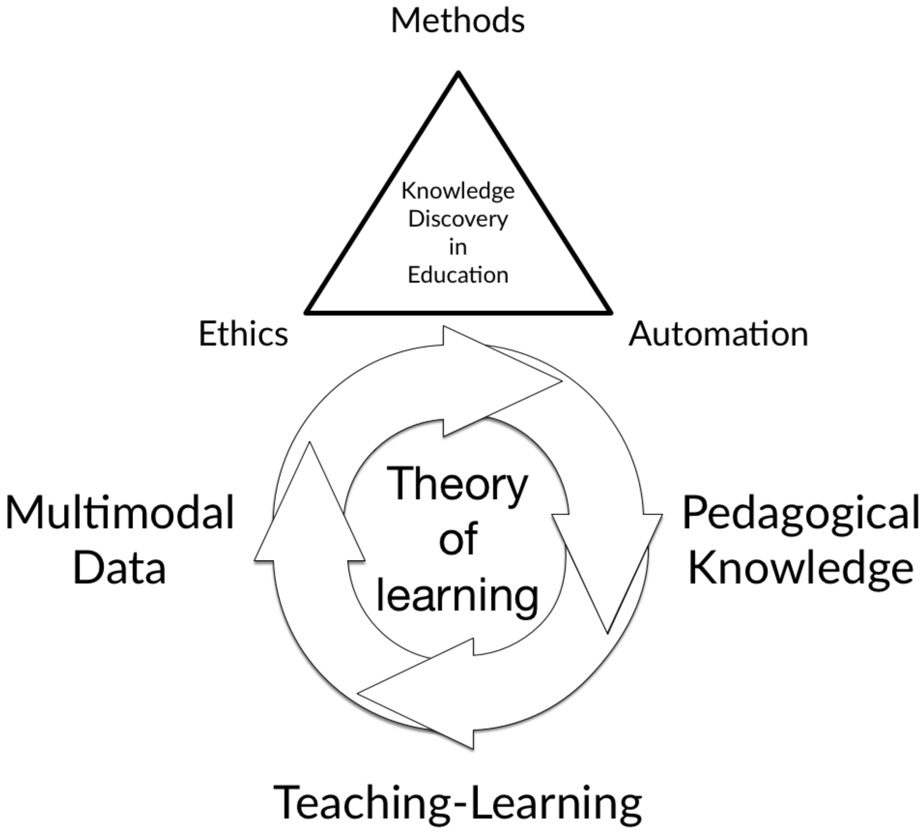


Figure 1. Conceptual model of pedagogical learning analytics cycle for providing novel and useful knowledge about learning processes.

2 Research method

The purpose of this research is to develop an information technology artifact, which in this case is a framework. Thus, the appropriate research method for this research is design science research. Design science research is a problem-solving process, which purpose is to derive novel knowledge and understanding of a design and its solution by designing and building an artifact (Hevner, March, Park and Ram 2004).

The design science research process guidelines define the methodological framework. The design science research guidelines and how they are applied in this research is summarized in Table 1. As Hevner et al. (2004) define, the creation and description of an innovative and purposeful artifact is the main goal of the design science research. Literature suggest a few conceptualizations of information systems (IS) artifacts and information technology (IT) artifacts. Lee, Thomas and Baskerville (2015) unpack the general term IS artifact into three separate classes: information artifact, technology artifact and social artifact. Offermann, Blom, Schönherr and Bub (2010) classify one important IT artifact typology, a guideline, which provides general suggestions about how the system should be developed. It's similar to artifact called framework, which is a metamodel (Peffers, Rothenberger, Tuunanen and Vaezi 2012). A metamodel is "model which is intended to give an all-inclusive picture of a process, system, etc., by abstracting from more detailed individual models contained within it" ("metamodel, n.". OED Online). The artifact created in this research is a framework, which provides general suggestion about the pedagogical learning analytics system.

The objective in design science research is to find knowledge and understanding in order to build technology-based artifacts that solve important problems. Thus, the problem relevance is important, and the created artifact has to be a sound solution to the presented problem. Solution needs to be evaluated based on the initial requirements. (Hevner et al. 2004.) The requirements for the framework are derived from the research questions. First of all, the framework should provide useful information to teachers (RQ1). The framework has to address the ethical issues (RQ2) and has to be automated (RQ3).

Evaluation in design science research can be observational, analytical, experimental, testing based, or descriptive. Case and field studies are observational methods, where the artifact is observed in a real business setting. In the analytical evaluation methods, one examines the static, dynamic, architectural, or performance related properties of the artifact. Experimental evaluation methods make use of controlled experiments and simulations. Evaluation by testing can be functional or structural and the purpose is to discover defects or values of chosen key metrics. Informed arguments based on background theory or construction of detailed scenarios are descriptive evaluation methods. (Hevner et al. 2004.)

An illustrative scenario is one of the most commonly used method for evaluating design science research (Peffers et al. 2012). The artifact in this research is evaluated by applying it to a scenario and evaluating it by using informed arguments based on the background theory and design objectives. Venable, Pries-Heje and Baskerville (2016) propose a Framework for Evaluation in Design Science (FEDS) for evaluating design process in design science research. In this research FEDS is used to guide the design science evaluation process.

Hevner et al. (2004) propose three kinds of research contributions that a design science research can provide, and at least one contribution must exist in a design science research project. The first kind of research contribution is the design artifact itself. Artifact must be implementable, and it has to solve the important previously unsolved problem. The second possible contribution is foundational knowledge, which improves and extends the existing knowledge base. The third contribution is the development of new methodologies for evaluation and new evaluation metrics.

Research rigor in design science research is derived from the proper use of theoretical foundations and research methods. Design science process is also an iterative search process, where the goal is to find the most effective solution. At the starting point, some factors of the design process can be simplified and then refined on later iterations. In communicating the design science research results, both the technology-oriented and managerial-oriented audience must be taken into account. (Hevner et al. 2004.) This research addresses both technical and pedagogical foundations. The design science research guidelines and how they are applied in this research are presented in Table 1.

Guideline	**Applied in this research**
Design as an Artifact	The artifact created in this research is a framework.
Problem Relevance	The solution provides pedagogical knowledge for teachers.
Design Evaluation	The framework is evaluated using an illustrative scenario.
Research Contributions	The research contribution is the artifact itself.
Research Rigor	The research uses comprehensive knowledge base and the artifact is evaluated using an evaluation framework.
Design as a Search Process	The result describes the first iteration of the design process.
Communication of Research	Research is documented in a form of a thesis.

Table 1. The design science research guidelines (Hevner et al. 2004) and how they are applied in this research.

Design Science Research Process -model (DSRP) (Peffers, Tuunanen, Rothenberger and Chatterjee 2007) is a mental model how the design science research can be conducted, presented and documented. This research follows the basic sequential activities of DSRP, which are (Peffers et al. 2007):

- problem identification and motivation
- objectives of a solution
- design and development
- demonstration
- evaluation
- communication.

In this research problem identification and motivation are presented in the introduction. Conceptual model of pedagogical learning analytics (Figure 1) represents the scope of the solution. Design and development are based on comprehensive theoretical knowledge base, which is derived from research literature. The designed artifact, framework, is applied in a scenario of analyzing university student agency.

3 Pedagogical learning analytics

Teaching and learning are actions, which produce a vast amount of different kinds of data. Data are stored in educational institutions but still rarely utilized by educational practitioners. This chapter outlines the conceptual model of pedagogical learning analytics.

3.1 Defining data

In the context of computing, data are “quantities, characters, or symbols on which operations are performed by a computer … information in digital form” (“data, n.”, OED Online). Data are also representations, “symbols that represent the properties of objects and events” (Ackoff 989, 3). The existence of large amounts of data leads to data-intensive computing (e.g. Gorton, Greenfield, Szalay and Williams 2008) and data-intensive science, which is sometimes referred as the fourth paradigm of science (e.g. Hey, Tansley and Tolle 2009; Kitchin 2014) or data-intensive scientific discovery (e.g. Philip Chen and Zhang 2014).

At the current time, the term big data is a popular buzzword (Figure 2), although Fan and Bifet (2013) summarize, that there is no need to separate big data analytics from data analytics. While the data used in this research is not in the scale of big data, it is still important to define the concept as it relates closely to the other concepts.

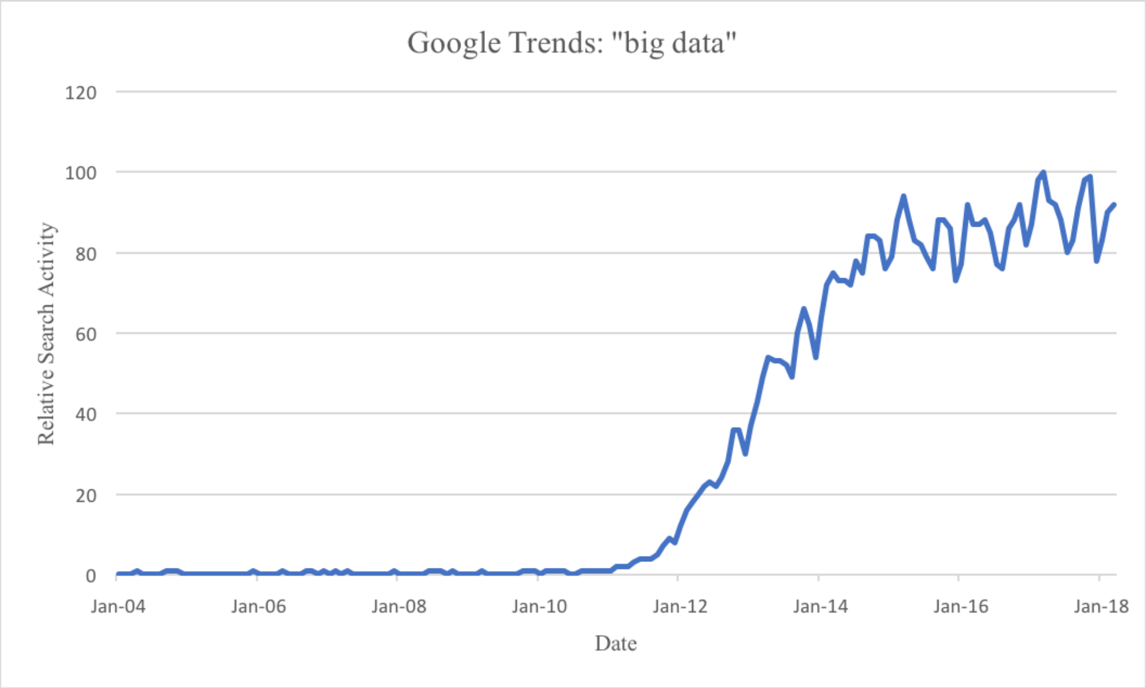


Figure 2. Relative search activity for keyword “big data” in Google Trends -service. The term was added to Oxford English Dictionary in mid 2013.

Due to the development of information technology, the amount of available data is increasing rapidly. This development has given rise to many buzzwords like big data, data mining and data science. The term big data was added to Oxford English Dictionary in June 2013 along with other technology-related terms like crowdsourcing, e-Reader, mouseover, redirect, and stream (Simpson 2013). According to the dictionary definition, big data means “data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges” (“big, adj. and adv.", OED Online). However, the first occurrence of the term was in sociologist Charles Tilly’s working paper (Tilly 1980), where he writes: “...that none of the big questions has actually yielded to the bludgeoning of the big-data people…”. While his article is not at all about the current big data concept, few decades later the “big-data people” have formed a whole new group of professionals. The current concept probably started to exist in lunch-table conversations in the middle of 1990s (Diebold 2012).

Big data is associated with specific challenges that are typically described with words starting with the letter “V”. The first original three words were *volume*, *velocity* and *variety*

(Laney 2001). Volume describes the vast size of the data sets. Velocity represents the frequency at which data are generated, stored and processed. Variety refers to different types of data, which can be highly unstructured. Later fourth word *veracity* was added, which means the reliability of the data. These four words are commonly used in the big data context, but other combinations and amounts of words are also used (i.e. Demchenko, Grosso, de Laat and Membrey 2013; Gandomi and Haider 2015; van der Aalst 2011). Oracle (2014), e.g. an enterprise cloud service provider, presents in their white paper an additional fifth big data definition word *value*. Big data might have a significant economic value. In the context of learning analytics and educational data mining the value of the big data depends on the utilization of the discovered knowledge.

3.2 Epistemology of data-intensive science

Due to the big data -phenomenon, it is undoubtedly worth to consider the epistemological foundations of data-intensive science. Does big data really represent a paradigm shift in science? Leonelli (2014) argues that the novelty of big data emerges from two changes in scientific practice: *data handling* and *data prominence*. There have been invented new efficient ways and methods to handle and analyze data. Prominence relates to the data as commodities with high value. Data is collected, recorded, and used constantly and to an increasing extent. Data are seen widely as an asset, and already the division between data-rich and data-poor countries has risen concerns (e.g. Melamed, Morales, Hsu, Poole, Rae, Rutherford and Jahic 2014).

Floridi 2004 explores the open questions in philosophy of information. One important question among them is whether nature can be informationalized? John Wheeler formalizes the idea in his famous conceptualization "it from bit", which otherwise stated means that "every it — every particle, every field of force, even the spacetime continuum itself — derives its function, its meaning, its very existence entirely — even if in some contexts indirectly — from the apparatus-elicited answers to yes or no questions, binary choices, bits" (Wheeler 1990, 310). The big data phenomenon and data intensive practice in science might not be a paradigm shift. Kitchin (2014) further argues that while big data causes disruption across disciplines, there is no need to declare the end of theory (i.e. Anderson 2008), but to critically review emerging epistemologies.

3.3 The Knowledge Discovery Process

Various data are currently being collected continuously. Databases are common places to store this information. The need to produce relevant information from the different datasets has led to the development of information processing methods, workflows, and processes.

Fayyad, Piatetsky-Shapiro and Smyth (1996a, 1996b, 1996c, 1996d) define knowledge discovery in databases (KDD) concisely as “the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data”. They break the definition further into smaller details. A Pattern is an expression describing some subset of the attribute values in the data. It includes the model or structure in data. The validity of the pattern means that the discovered patterns should apply to some extent to the new data. The found patterns should be novel and potentially lead to some useful actions. The novelty can be measured by comparing new values or knowledge to old ones and usefulness depends on the application domain. Lastly, they state that the patterns found must ultimately be comprehensible to human beings.

The knowledge discovery process (Fayyad et al. 1996c, 1996d) involves multiple interactive and iterative steps from understanding the problem domain to the utilization of the new knowledge (Figure 3).

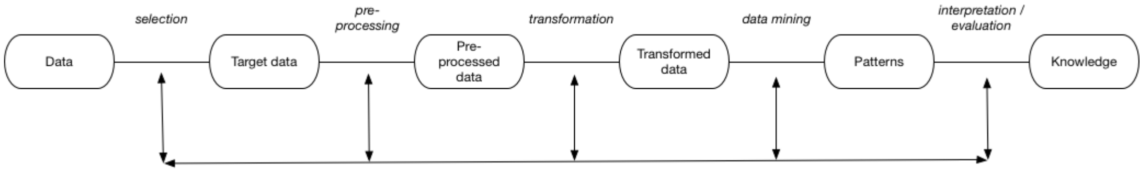


Figure 3. The knowledge discovery in databases process (Fayyad et al. 1996c, 1996d).

The knowledge discovery process starts with goal setting and learning the application domain. Next, the dataset required for the process is created. The target dataset can be the whole data or a subset of variables or data samples. Raw data from the real world is often untidy and poorly formatted. Preprocessing involves operations to convert data into a tidy form. Problems with the real-world data occurs when there is too much data, too little data or the data are fractured (Famili, Shen, Weber and Simoudis 1997).

Once the data are cleaned and preprocessed, it is ready for transformation. Transformation means methods to reduce data dimensions, number of variables, and to find invariant variables. The overall goal of data transformation is to find the optimal number of features to represent the data. The transformation phase of the knowledge discovery process is followed by the actual data mining. This step involves selecting the purpose and method of data mining as well as the implementation and execution of the mining algorithm. Thus, data mining is one part of knowledge discovery process (Zaki and Meira 2014).

In the interpretation phase, the relevant patterns are selected and changed into a form that is understood by users. This includes possible visualization of the results. In the last step the new knowledge is evaluated, reported and implemented (Fayyad et al. 1996b, 1996d).

3.3.1 Data selection

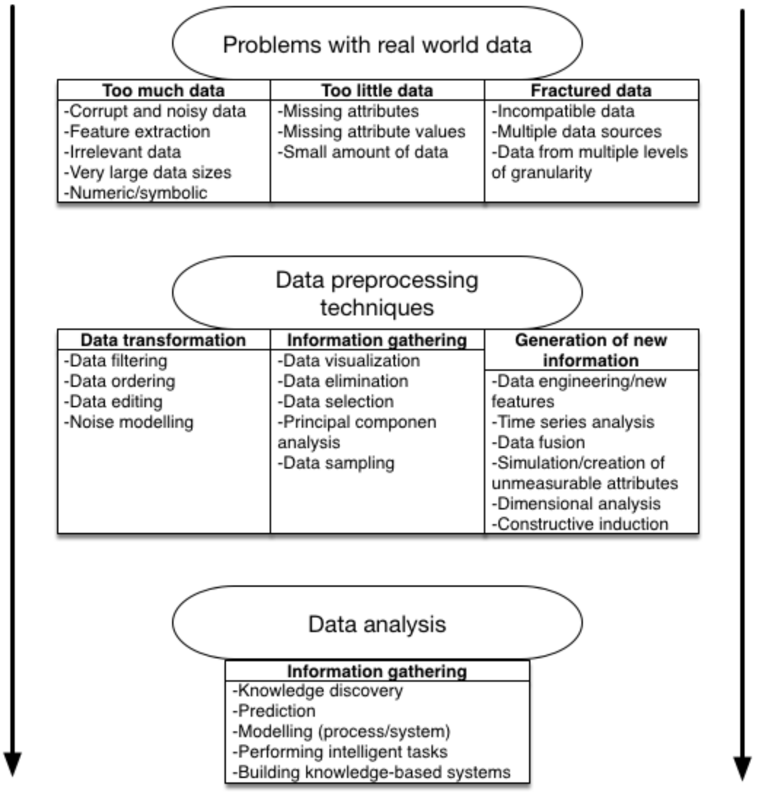
Data come in various forms and are stored in different places. Data can be structured or unstructured and it can be stored in various data repositories, databases, data warehouses or on the Web (Han, Pei and Kamber 2011). Different devices and sensors are continuously collecting new data. Chen and Zhang (2014) argue that the capacity to store information has doubled every three years since the 1980s.

When considering the rate of which data are generated and the possibilities to store it, data are often available more than enough. From the knowledge discovery point of view, it is not necessary nor practical to use all available information. Some form of data selection is often needed in order to make the whole process more efficient. Fayyad et al. (1996c) emphasize the importance of the relevance of the attributes and data flawlessness. They call for strong domain knowledge, prior knowledge, which can help in determining the important attributes and the potential relationships. Äyrämö (2006) emphasizes the significance of a domain analysis, which is a prerequisite for a successful knowledge discovery.

3.3.2 Data preprocessing

Data preprocessing is a step in the knowledge discovery process, and according to Famili et al. (1997, 5) it "consists of all the actions taken before the actual data analysis process starts". The purpose of preprocessing is to transform the raw data into a more usable form while preserving the "valuable information". Comparing to the knowledge discovery process, they group together the preprocessing and the transformation steps.

Famili et al. (1997) divide the problems with the real-world data into three categories: 1) too much data 2) too little data, and 3) fractured data. They present a detailed but not exhaustive description of possible techniques to address these issues (Figure 4). Data preprocessing is needed if the data contains problems that prevent any type of analysis, if more understanding of the nature of the data is needed in order to perform better analysis, if extracting more meaningful information is needed, or any combination of the previous reasons.



Problems with real world data

Too much data	Too little data	Fractured data
-Corrupt and noisy data -Feature extraction -Irrelevant data -Very large data sizes -Numeric/symbolic	-Missing attributes -Missing attribute values -Small amount of data	-Incompatible data -Multiple data sources -Data from multiple levels of granularity

Data preprocessing techniques

Data transformation	Information gathering	Generation of new information
-Data filtering -Data ordering -Data editing -Noise modelling	-Data visualization -Data elimination -Data selection -Principal componen analysis -Data sampling	-Data engineering/new features -Time series analysis -Data fusion -Simulation/creation of unmeasurable attributes -Dimensional analysis -Constructive induction

Data analysis

Information gathering
-Knowledge discovery -Prediction -Modelling (process/system) -Performing intelligent tasks -Building knowledge-based systems

Figure 4. Problems with real word data and possible preprocessing techniques (Famili et al. 1997).

Data preprocessing also often involves cleaning the data. Data cleaning means, for example, removal of noise and handling missing values and outliers (Maimon and Rokach 2009). Noise is meaningless information, which needs to be removed. Missing data are a data points, which have no stored value. Outlier is an abnormal value, which does not belong to the data. Maletic and Marcus (2009) describe data cleaning as a three-phase process. The first step is to determine and define error types. When the error types are known, the second step is to search and identify these erroneous data points. The last step is to correct the uncovered errors.

Kantardzic (2011) presents two common data preprocessing tasks, which are outlier detection and feature transformation. Outliers can be dealt by detecting and removing them or by using robust data mining methods, which are not so sensitive to outliers. Feature scaling, encoding and selecting are transformations that need to be executed in particular cases.

3.3.3 Data transformation

Real world data are often multidimensional and contains invariant variables. This kind of multidimensional data brings with it challenges related to data mining methods and computing resources. These challenges can be addressed using various data transformation and dimension reduction methods. The purpose of the data transformation is to further prepare the cleaned data in order to enable efficient data mining.

Fayyad et al. (1996a, 1996b, 1996c, 1996d) present data transformation as a step in knowledge discovery process, where amount of variables can be reduced and invariant representations of the data can be found. Dimensionality of the data can be reduced, for example, by finding the best features to represent the data, which is called feature extraction. Another popular way to transform data and reduce the dimensionality is to project the data into lower dimensional space. Making new variables and combining existing ones can also reduce the number of variables.

The data transformation step is important for the whole knowledge discovery process to succeed. On the other hand, the process is often project-specific and requires some degree of knowledge of the problem domain (e.g. Äyrämö 2006; Maimon and Rokach 2009).

3.3.4 Data mining

In some cases, the actual data mining step is used in a broader sense synonymously with knowledge discovery process (Han et al. 2011), but Fayyad et al. (1996a, 1996b, 1996c, 1996d) describe it as a separate step in the knowledge discovery process executed after data has been transformed into suitable form. In the later view, it involves fitting models to or finding patterns from target data. Selecting and executing a proper data mining algorithm is fundamental part of this steThe actual data mining phase consists of three parts: choosing the proper data mining task, choosing the data mining algorithm, and, lastly, implementing and executing the data mining process (Maimon and Rokach 2009).

Based on the primary goal of the data mining outcome and considering the function of the mining algorithm, data mining algorithms can be divided into two categories: *descriptive* algorithms and *predictive* algorithms. Descriptive data mining describes the data in a meaningful way and produces new and nontrivial information. Predictive data mining examines the system and produces the model of the system based on the given data set. (Kantardzic 2011.)

Fayyad et al. (1996a, 1996b, 1996c, 1996d) define that generally, every data mining algorithm can be presented as composition of three general principles. These principles are the *model*, the *preference criterion*, and the *search algorithm*. Model is a description "of the environmental conditions, both overt and hidden, for an experimental or observational setting" (Shrager and Pat 1990). The data mining model has a representation in some language and a function, which is a description of the intended use of the model.

The preference criterion or the model evaluation criteria of the data mining algorithm is a quantitative function, which measures how well the goals of the knowledge discovery process are met. The search algorithm is the last step of the data mining algorithm, and it contains two parts: parameter search and model search. Parameter search is used to find model parameters which optimize the preference criterion. The purpose of the model search is to loop over the fixed parameters in order to find the preferred model representation. (Fayyad et al. 996c, 1996d.) The search algorithm is often a trade-off between time used in searching the result and optimality of the model, because finding of the optimal model might be computationally too expensive (Cheeseman 1990).

3.3.5 Interpretation and evaluation

The previous data mining step eventually returns some mining results. The data mining result is the model induced from the data. In this step the usefulness of the model is evaluated, and visualization and documentation are important tasks of the interpretation and evaluation process (Maimon and Rokach 2009). Fayyad et al. (1996a, 1996b, 1996c, 1996d) define interpretation and evaluation as a step where the results are evaluated with respect to the defined goals and all previous steps. The knowledge discovery is an iterative process and all steps can be revisited if necessary.

3.4 Learning analytics and educational data mining

Learning analytics (LA) and educational data mining (EDM) are both fairly recent scientific fields and research communities which exploit data gathered in an educational setting. They both have their own societies, conferences and journals. The practitioners of learning analytics have their Society for Learning Analytics Research (SoLAR) founded in 2011, Journal of Learning Analytics first published in 2014, and Learning Analytics and Knowledge Conference (LAK) first held in 2011. The main international authorities in the field of educational data mining are International Educational Data Mining Society (IEDMS) founded in 2011, Journal of Educational Data Mining (JEDM) first published in 2009 and International Conference on Educational Data Mining first held in 2008. Both publications are classified as Class 1 in JuFo (Julkaisufoorumi, Publication Forum) classification in 2018.

The first International Conference on Learning Analytics and Knowledge defined learning analytics as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (LAK11 2010). International Educational Data Mining Society defines educational data mining as a discipline, that is “concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in” (educationaldatamining.org, n.d.).

In other words, learning analytics is the analysis of educational data of all sizes, both big and small, with a goal of producing effective learning and knowledge of learning in general. Educational data can also be obtained using different methods. Blikstein and Worsley (2016) describe several computational technologies for measuring complex learning tasks. They call those methods as multimodal learning analytics, which include methods like text and speech analysis, handwriting and sketch analysis, action and gesture analysis, affective state analysis, neurophysiological markers and eye gaze analysis.

Learning analytics make use of knowledge discovery process (Fayaad et al. 1996a, 1996b, 1996c, 1996d) applied in educational context. This process is called educational knowledge discovery (Saarela and Kärkkäinen 2017; Romero and Ventura 2013) and educational data mining is essential part of the process. Both learning analytics and educational data mining consider the actions of a learner at the micro level (Piety, Hickey and Bishop 2014). Siemens and Baker (2012) compare that there is both overlap and key distinctions between these two separate disciplines, while they share similar goals. They state that learning analytics community has emphasis on systemic understanding and intervention, while educational data mining community has more reductionist approach. LA has focus on empowering and informing learners and educators and EDM concentrates more on adaptive automation. In the context of higher education, there exists also a concept called academic analytics. Academic analytics is “a process for providing higher education institutions with the data necessary to support operational and financial decision making” (Van Barneveld, Kimberly and Campbell 2012, 8). It is targeted more to the institutional decision-making level.

Both learning analytics and educational data mining can be seen as outcomes of a shift towards data intensive sciences applied in educational setting. Learning analytics is utilizing big data to an increasing extent (Saarela and Kärkkäinen 2017). In Kuhnian sense, learning analytics has a promise of better understanding of learning and providing more efficient ways to learn in the future. Despite recent efforts, learning analytics has not yet managed to redeem its promises (Ferguson and Clow 2017).

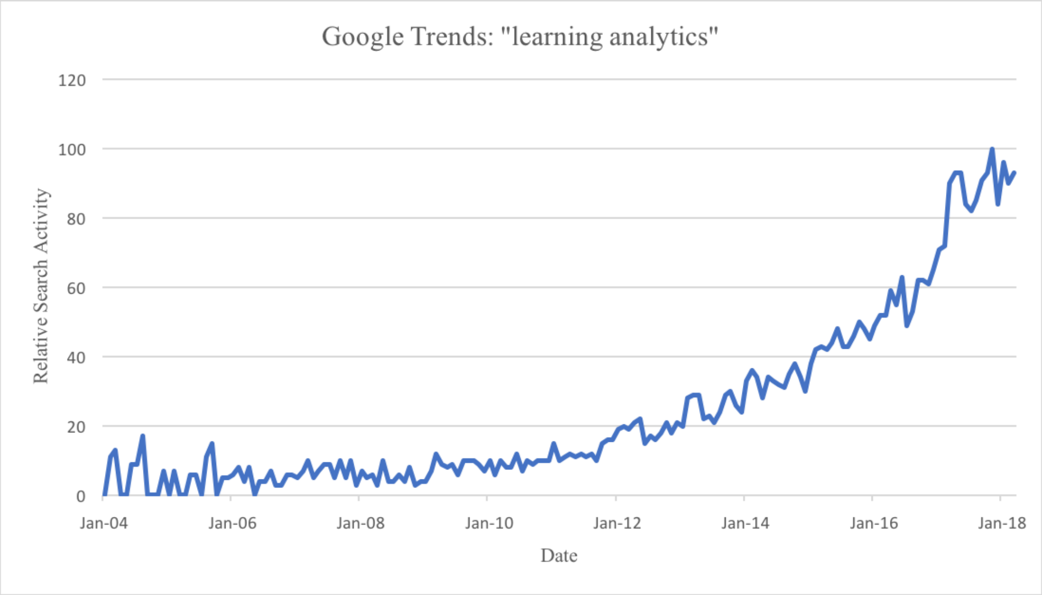


Figure 5. Relative search activity for keyword “learning analytics” in Google Trends - service.

Learning analytics is currently a popular search term according to Google Trends (Figure 5). Keywords “learning analytics” in Google Scholar returns 21 500 results in early 2018 and about half of them are dated from 2016 onwards. It is justified to say that learning analytics is a hot topic in education. Therefore, it is crucial that proper evidence can prove that learning analytics is useful. There exists a significant gap between learning analytics and evidence of its effectiveness (Ferguson, Brasher, et al. 2016). There is a need for pedagogical learning analytics, which combines the concept of pedagogical knowledge and learning analytics.

Baker (2010) presents five primary categories of educational data mining methods, which are prediction, clustering, relationship mining, discovery with models and distillation of data for human judgement. Prediction involves developing a predictive model, which can infer a variable based on predictor variables. Han et al. (2011, 443) define clustering as a data mining “process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters”. Clustering is an unsupervised method, which means that there is no need for labeling the data. Labels are assigned based on the clustering result. Saarela and Kärkkäinen (2017) conclude that hierarchical clustering, k-means, and expectation-maximization are the most common clustering methods in educational data mining.

Relationship mining is a data mining method for discovering relationships between variables. In discovery with models, this kind of model can be used as a further source for educational data mining. Educational data mining can also provide information for human judgement. (Baker 2010.)

Methods used in educational data mining is one way to generate new information. The new information can be then used in learning analytics. Clow (2012) describes the Learning Analytics Cycle (Figure 6), which has four linked steps. In the cycle learners generate data which is used to generate metrics, analytics and, visualizations in order to make interventions, which influence learners.

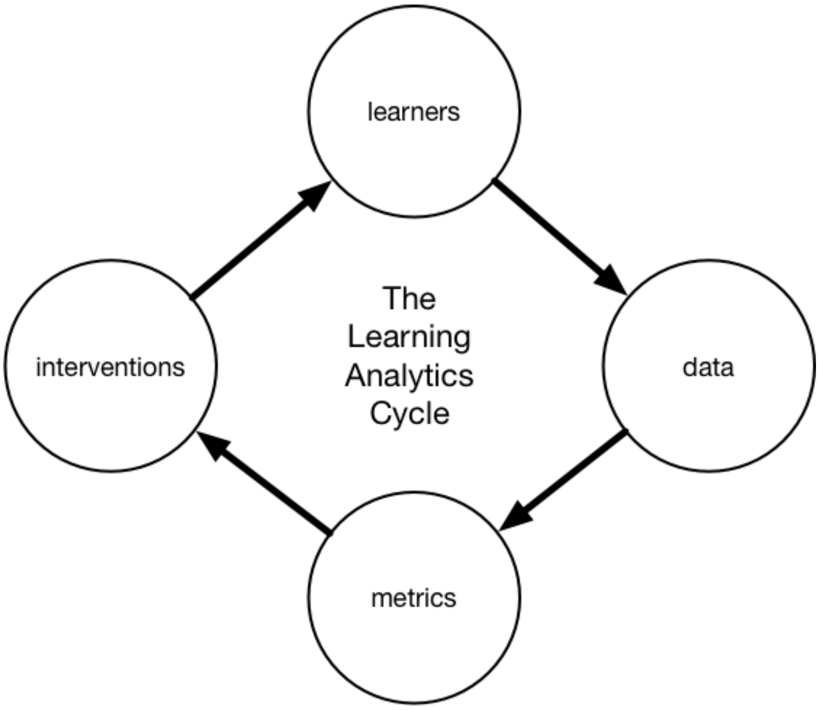


Figure 6. The Learning Analytics Cycle (Clow 2012).

The learning analytics cycle (Figure 6) is a feedback loop. There are four stakeholder groups involved in the process: learners, teachers, managers and policy makers. Learners are the central agents in the loop. Teacher is a person who is directly involved with the learning process. Managers and policymakers are also included in the loop and they are responsible for organizational administration and setting policies in any level. (Saarela and Kärkkäinen 2017; Clow 2012.) The learning analytics cycle doesn't take a stand on what kind of information these stakeholder groups would benefit.

Different stakeholders need different kind of information. Learners benefit from personalized information, while policymakers need information that supports their decision making. Teachers operate on the basis of their knowledge base. The knowledge base includes all profession-related insights, which affect teacher's activities in teaching and learning situations (Verloop, Van Driel and Meijer 2001). Pedagogical learning analytics addresses the needs of teachers by contributing relevant information to their knowledge base.

3.5 Pedagogical knowledge

Many studies suggest that one of the most important contributing factor to student achievement in school is the quality of teaching and teachers (e.g. Canales and Maldonado 2018; Darling-Hammond 2000; Muñoz, Prather and Stronge 2011). Teacher quality is also suggested to generate a significant economic value (Hanushek 2011). Thus, improving and investing in teacher quality is a good way to get better educational results (Akiba, LeTendre and Scribner 2007). Pedagogical knowledge is one, though less researched, indicator of quality of a teacher (Guerriero 2013). Thus, through contributing to teacher's pedagogical knowledge it might be possible to get better learning outcomes.

Shulman (1987) was one of the first researchers trying to define categories of teacher's knowledge base, which includes notions of content knowledge, pedagogical content knowledge, and general pedagogical knowledge. According to him, general pedagogical knowledge involves "broad principles and strategies of classroom management and organization that appear to transcend subject matter" (ibid., 8). Later on, other scholars have developed the concept further. Voss, Kunter and Baumert (2011, 953) define general pedagogical and psychological knowledge (PPK) as "the knowledge needed to create and optimize teaching–learning situations across subjects". They constructed a factor model and a questionnaire to assess general pedagogical and psychological knowledge. The overall PPK consists of four factors representing teacher's knowledge about teaching methods, classroom management, classroom assessment, and students' heterogeneity. (Voss et al. 2011.) The definition of general pedagogical and psychological knowledge has similarities with the definition of learning analytics: their purpose is to understand and optimize learning across subjects. One example of pedagogical knowledge is the knowledge about student agency.

3.5.1 What is human agency?

An agent is a being having a capacity to act, and agency means the manifestation of this capacity. Due to the broad definition, it is natural to say agency is practically everywhere. In a narrower sense, agency often denotes the performance of intentional actions. It has a long history in philosophy, and in recent years agency has also been growing interest in other fields of research such as social science, psychology, cognitive neuroscience, and anthropology. It has also gained popularity in education, working-life studies and gender research. (Eteläpelto, Vähäsantanen, Hökkä and Paloniemi 2013; Schlosser 2015)

Social sciences are largely responsible for the theorizing of agency and the roots date back to Talcott Parsons (1937) and Anthony Giddens (1984) Despite the efforts and prevailing appeal in many research fields, agency is still a misunderstood concept that is not evaluated systematically, and is missing an explicit definition of its core meaning, and has inconsistent definitions across different theoretical frameworks (Emirbayer and Mische 1998; Eteläpelto et al. 2013; Hitlin and Elder 2007). Agency is even argued to be a "red herring" without any sociological merit (Loyal and Barnes 2001).

Hitlin and Elder (2007) try to clarify the concept of agency and suggest dividing it into four analytical types. Existential agency is a universal human potential. It is a basis for "free will" and it also takes place in social action and all circumstances through temporal horizons. Pragmatic agency is associated with new situations in the present, where a routine way of doing things fail. Identity agency is linked to everyday routine situations, and it characterizes a capacity to act according to social role expectations. Life course agency extends the temporal horizon to cover life pathways, and it defines decisions made at turning points and transitions.

Emirbayer and Mische (1998) argue in favour of redefining human agency. They propose a triadic and temporally embedded definition of agency and describe it as (ibid., 970):

> *"...the temporally constructed engagement by actors of different structural environments — the temporal-relational contexts of action — which, through the interplay of habit, imagination, and judgment, both reproduces and transforms those structures in interactive response to the problems posed by changing historical situations"*

In the previous definition the primal elements of agency are iteration, projectivity, and practical evaluation. Iterative element implies routine and practical activity and can be compared to the identity agency proposed by Hitlin and Elder (2007). It draws meaning from the past and brings stability and order to social structures. Projectivity orients toward the future and is a capacity to imagine alternative possibilities. Practical evaluation is the capability to make rational and normative judgments among alternative trajectories of action. (Emirbayer and Mische 1998.) According to this definition, agency originates from the past through the present to the future.

In the field of psychology, Bandura (2006) identifies four core properties of human agency. The first is intentionality, which means briefly that people form intentions and plans for realizing them. The second property is forethought and it brings temporal dimension to human agency. People make plans for the future, set goals, and anticipate likely outcomes. Self-reactiveness is the third property of human agency and it states that people are also self-regulators. After having an intention and action plan, agents have an ability to construct motivational courses of action. The fourth property, self-reflectiveness, provides means to evaluate thoughts and actions and make corrective adjustments.

One way of describing human agency is the notion that humans have a sense of agency. The sense of agency is defined as "the ability to recognize oneself as the agent of a behavior" (Jeannerod 2003) or "a sense of control and of being the agent or owner of the action" (Schlosser 2015). There is no clear consensus on the origin of the sense of agency. However, the human motor control system is suggested to have an essential role in the generation of the sense of human agency (Schlosser 2015).

The recent developments of neuroscience have made it possible to explore more complex cognitive functions like the sense of agency. Recent brain imaging studies have identified particular brain regions that have been linked to the human sense of agency and also motor control system (Haggard 2017; Renes, van Haren, Aarts and Vink 2015; Spengler, von Cramon and Brass 2009).

3.5.2 Agency of University Students Scale

Jääskelä, Poikkeus, Vasalampi, Valleala and Rasku-Puttonen (2016) have constructed a factor model of university student agency and a questionnaire for measuring it. The questionnaire, Agency of University Students (AUS) Scale, contains 60 propositions in a five step Likert scale. (Jääskelä et al. 2016.) The individual agency profile can be extracted from the questionnaire response using the factor model.

The agency of university students consists of three resource domains. Individual resource domain is, according to its name, dependent on the individual and contains dimensions of self-efficacy, competence beliefs, and participation activity. However, agency is also relational and context-bound. Relational resource domain consists of dimensions like power relationships and peer support. Contextual resource domain has three dimensions, which relate to different kinds of perceived opportunities in the learning context. The AUS scale is a tool to develop university teaching, it can reveal course-specific knowledge, and be a basis for pedagogical implementations. (Jääskelä et al. 2016.)

3.6 Pedagogical learning analytics

Learning analytics is, as mentioned, “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs” (LAK11 2010). The beginning part of the definition, “measurement, collection, analysis and reporting”, is a direct reference to the knowledge discovery process (i.e. Fayyad et al. 1996b, 1996c). As the process is applied in the context of teaching and learning, the process can be called as educational knowledge discovery (e.g. Saarela and Kärkkäinen 2017).

General pedagogical knowledge is independent of the subject and its purpose is “to create and optimize teaching–learning situations across subjects” (Voss et al. 2011, 953). This equals with the later part of the learning analytics definition: the purpose of both is to facilitate more effective learning in different educational environments. By synthesizing the findings presented in this chapter about knowledge discovery, learning analytics, educational data mining and pedagogical knowledge, I present the following definition: Pedagogical learning analytics makes use of educational knowledge discovery process in order to provide valid, novel and useful knowledge, which teachers can utilize when creating and optimizing teaching–learning situations and environments across subjects. Combining this definition with the idea of learning analytics cycle (i.e. Clow 2012) and expanding the meaning of educational data with multimodality (i.e. Blikstein and Worsley 2016), I sketch the conceptual model of pedagogical learning analytics cycle (Figure 7).

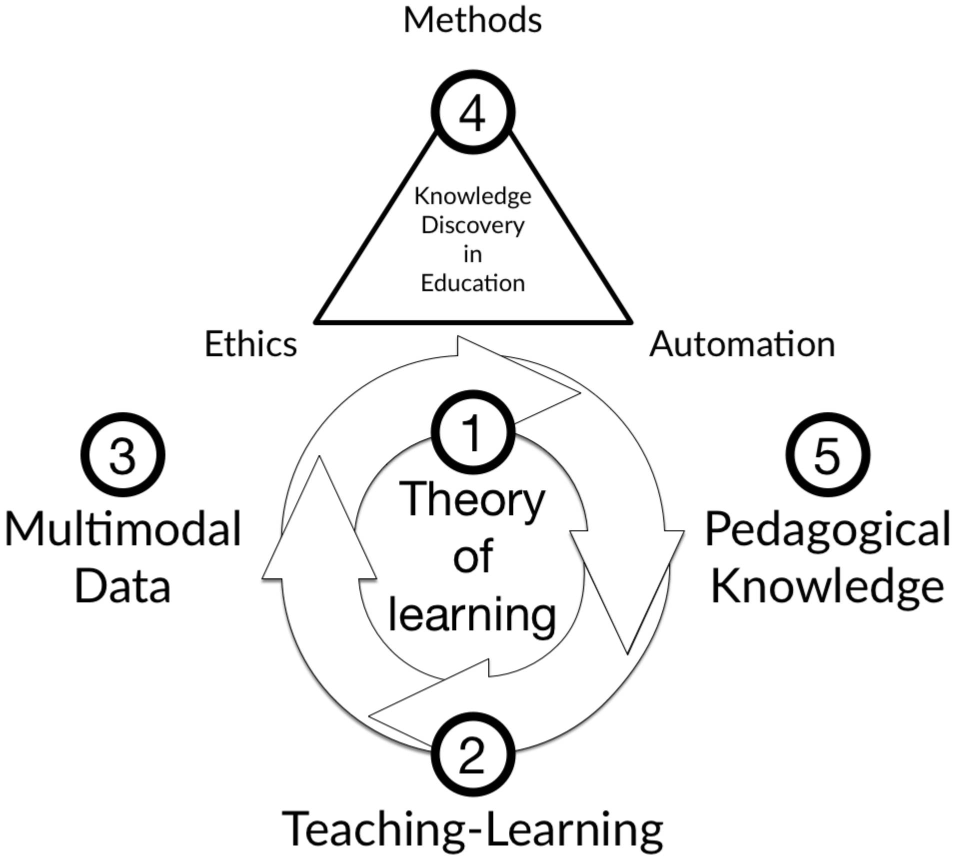


Figure 7. Pedagogical learning analytics cycle.

In the center of the pedagogical learning analytics cycle (Figure 7) are the scientific theories and knowledge about learning (1). Latest knowledge about how humans learn provide the foundation for the pedagogical learning analytics. For example, theories of learning might provide guidelines of what kind of data are needed. The actual learning happens, when learner and teacher make actions in teaching–learning situation in order to produce effective learning (2). These actions produce different kind of multimodal data (3), which are collected and recorded. Ethical and automated information processing system makes use of knowledge discovery process and appropriate data mining methods (4). The output of the knowledge discovery process is pedagogical knowledge (5), which contributes to the teaching–learning situation. Pedagogical learning analytics could be a positive feedback loop, as the knowledge acquired from the knowledge discovery process might contribute new knowledge about learning in general.

4 Ethical learning analytics

Compliance with ethical principles is one of the most fundamental requirements of the automated learning analytic services. First of all, automated learning analytic services have to be in compliance with the respective law. From the European perspective, the General data Protection Regulation (GDPR) lays significant requirements to learning analytics systems. This chapter examines privacy aspects in relation to ethical considerations of learning analytics and key concepts of GDPR.

Different ethical aspects have to be considered in a wider scope than merely from the legal point. Dahl (2015) points out contradiction in the recent reports about learning analytics. The students are somewhat comfortable with gathering of information about them in order to facilitate better learning. They are already used to deal with impaired privacy when using different commercial services. On the other hand, regulation and ethical concerns make it necessary to focus on privacy, security, and individual rights. He concludes that learning analytics is impossible to implement unless these concerns aren't addressed properly.

Educational institutions need to implement proper learning analytics policies, which specifically address the issues of ethics and privacy in learning analytics. Existing policy frameworks seem to be insufficient in addressing these issues (Prinsloo and Slade 2013). Data privacy is also a major concern for data mining in case any type of personal data is handled. Two fields of research and practice relate to data privacy in data mining: Privacy-Preserving Data Mining (PPDM) (Aggarwal and Yu 2008) and Statistical Disclosure Control (SDC) (Willenborg and de Waal 2012).

4.1 Ethics of learning analytics

In learning analytics, ethics, privacy and data protection are closely related. Ferguson, Hoel, Scheffel, and Drachsler (2016) suggest, that it would be useful to first consider these topics separately. After presenting 21 different challenges in ethics of learning analytics, they provide nine ethical goals for learning analytics (Ferguson, Hoel, et al. 2016.):

1. student success
2. trustworthy educational institutions
3. respect for private and group assets
4. respect for property rights
5. educators and educational institutions that safeguard those in their care
6. equal access to education
7. laws that are fair, equally applied, and observed
8. freedom from threat
9. integrity of self.

The goals are open to interpretation and they are dependent on context (Ferguson, Hoel, et al. 2016). However, they provide a starting point for exploring different policy implementations and frameworks. The DELICATE checklist (Drachsler and Greller 2016) is examined for addressing these ethical goals.

Determination	Why you want to apply learning analytics? What is the added value (Organizational and data subjects)? What are the rights of the data subjects? (e.g., EU Directive 95/46/EC)
Explain	Be open about your intentions and objectives What data will be collected for which purpose? How long will this data be stored? Who has access to the data?
Legitimate	Why you are allowed to have the data? Which data sources you have already (aren't they enough?) Why are you allowed to collect additional data?
Involve	Involve all stakeholders and the data subjects Be open about privacy concerns (of data subjects) Provide access to the personal data collected (about the data subjects) Training and qualification of staff
Consent	Make a contract with the data subjects Ask for a consent from the data subjects before the data collection Define clear and understandable consent questions (Yes / No options) Offer the possibility to opt-out of the data collection without consequences
Anonymize	Make the individual not retrievable Anonymize the data as far as possible Aggregate data to generate abstract metadata models (Those do not fall under EU Directive 95/46/EC)
Technical	Procedures to guarantee privacy Monitor regularly who has access to the data If the analytics change, update the privacy regulations (new consent needed) Make sure the data storage fulfills international security standards
External	If you work with external providers Make sure they also fulfill the national and organizational rules Sign a contract that clearly states responsibilities for data security Data should only be used for the intended services and no other purposes

Table 2. The DELICATE checklist (Drachsler and Greller 2016). Checklist refers to an old directive: EU Directive 95/46/EC is superseded by General Data Protection Regulation.

DELICATE (Drachsler and Greller 2016) is an eight-point checklist (Table 2) and it's based on legal texts, literature reviews, and workshop discussions. The authors emphasize that learning analytics should follow a value-sensitive design process and the checklist is a tool to facilitate discussion between stakeholders. The checklist addresses issues of power-relationship, data ownership, anonymity, data security, privacy, data identity, transparency and trust.

When DELICATE checklist is reflected towards aforementioned ethical goals, the results show that the checklist seems to cover all ethical goals (Table 3). While the list of ethical goals nor the DELICATE checklist are exhaustive interpretations of ethical issues, they seem to provide a reasonable starting point for evaluating learning analytics implementations and facilitating discussion. The result of this discussion is usually a written document, learning analytics policy, which is the guideline for using learning analytics in educational institution.

DELICATE	**What ethical goals are covered?**
Determination	(1) student success, (2) trustworthy educational institutions, (4) respect for property rights, (7) laws that are fair, equally applied, and observed, (8) freedom from threat, (9) integrity of self
Explain	(1) student success, (2) trustworthy educational institutions, (9) integrity of self
Legitimate	(1) student success, (2) trustworthy educational institutions, (5) educators and educational institutions that safeguard those in their care, (9) integrity of self
Involve	(2) trustworthy educational institutions, (6) equal access to education, (7) laws that are fair, equally applied, and observed
Consent	(2) trustworthy educational institutions, (7) laws that are fair, equally applied, and observed, (8) freedom from threat, (9) integrity of self
Anonymise	(2) trustworthy educational institutions, (3) respect for private and group assets, (7) laws that are fair, equally applied, and observed
Technical	(2) trustworthy educational institutions, (3) respect for private and group assets, (5) educators and educational institutions that safeguard those in their care, (7) laws that are fair, equally applied, and observed
External	(2) trustworthy educational institutions, (4) respect for property rights

Table 3. The DELICATE checklist (Drachsler and Greller 2016) is reflected towards ethical goals (Ferguson, Hoel, et al. 2016).

Creating a learning analytics policy is one step in utilizing ethical learning analytics in the institutional level and in practice outside academic research projects. A policy is “a principle or course of action adopted or proposed as desirable, advantageous, or expedient … method of acting on matters of principle, settled practice” (“policy, n.”, OED Online). Applying this definition, learning analytics policy describes the principles for ethical use of learning data. Staalduinen (2015) summarizes the consensus that there is a need for a separate learning analytics policy in educational institutions. Policy needs to cover areas like ethics, privacy, legal context, data governance, data usage, purpose of usage, transparency, student consent and stakeholders.

Institution	**Purpose**	**Principles covered**
The University of Edinburgh	improve retention **enhancement** of student experience (quality, equity, personalized feedback, coping with scale, student experience, skills, efficiency)	“not be used to inform significant action”, “not ... only at supporting students at risk of failure”, transparent about: collect, use, share, consent, ethical use, “data and algorithms can contain and perpetuate bias”, minimize negative impact, good governance, focus on development, “will not be used to monitor staff performance”
University of West London	help students **succeed** and achieve their study goals	clarity of purpose, individuals, openness, consent, responsibility, quality, access, partnership, appropriate use, compliance
University of Gloucestershire	provides new opportunities to **support** learners and to **enhance** educational processes **assist** current students in achieving their study goals and to help the institution to **improve** aspects of education for future learners	responsibility, transparency, consent, confidentiality, sensitive data, validity, access, interventions, minimizing adverse impacts

Table 4. Brief summary of learning analytics policies of The University of Edinburgh (2017), University of West London (2017) and University of Gloucestershire (2016).

Several learning analytics policies of different institutions are openly accessible in the web (Table 4). A brief overview reveals that helping students to succeed is the major goal of learning analytics (e.g. Ferguson, Hoel, et al. 2016) in sample universities (Table 4). Wide range of principles are covered. The University of Edinburgh also mentions staff: learning analytics is not used for monitoring staff performance. However, while Staalduinen's (2015) list of coverable aspects in learning analytics policy is not exhaustive, there is still gaps in sample policies compared to it. For example, other stakeholders in the context were mostly omitted. Prinsloo and Slade (2013) conclude that many institutions concentrate on academic analytics for research purposes and there seems to be challenges for wider institutionalized use of learning analytics.

The purpose of learning analytics policy is important. It might affect learner's disclosure of private information concerning their learning. Communication Privacy Management (CPM) theory is about how people manage their privacy and make decision what to reveal and what to conceal (Petronio 2012). Chang, Wong and Lee (2015) use CPM to construct a model how people manage their privacy when organizations are asking their data. They call the three-phase model as Cognitive Process Model of Privacy Boundary Management. In the first *institutional boundary identification phase* a person decides and makes an opinion how well and effectively an organization follows its existing privacy policy. In the second phase of *mutual boundary rule formation* a person compares the privacy boundary of an institution with their own need for privacy protection. In the last *individual boundary decision phase,* a person reaches a self-assessed state where others can have a limited access to personal information. (Chang et al. 2015.)

Privacy boundary evaluation might be a situation when a leaner assesses a learning analytics policy of an institution. A learner makes decision what information to disclose based on learning analytics policy and potential benefits and negative effects. In learning analytics it's not always possible to disclose only some information as learning management systems often collect automatically wide range of information. The importance of a credible policy is important. Carelessly and unethically drafted policy might lead to minimized use of analytics. Most of all, it might lead to illegal activity. Thus, in learning analytics it's important to acknowledge and comply with relevant legal regulation.

4.2 General Data Protection Regulation (GDPR)

The General Data Protection Regulation (GDPR) is a regulation adopted within the European Union (EU), which aims to improve the privacy of individuals and their personal data. The law imposes significant requirements for the processing of personal data. There are several important and essential organizational and legal obligations that must be taken into account when dealing with data containing personal information. From the researcher point of view, the main objective of the GDPR is to protect "fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data", while still enable researchers to use personal data for scientific research (Regulation (EU) 2016/679 2016).

GDPR is published in Official Journal of the European Union (OJ), which is the main source for European Union legislation. The journal is published in all official EU languages daily on weekdays and only in urgent cases on weekends and public holidays. As of the 1st of July 2013, only the electronic versions of the Official Journal (e-OJ) are legally binding, however all issues since the first edition in 30th of December 1952 are available online. The journal has two series: L-series is for legislation and C-series is for information and notices. GDPR is published in OJ number L 119/1.

A legislative act starts with a title, which is followed by a preamble. A preamble contains everything between the title and the enacting terms of the act (i.e. citations and recitals). Citations indicate the legal basis and the preparatory acts. Recitals start with a word "Whereas:" and they introduce the reasons for the contents of the enacting terms. The normative part of an act, the enacting terms, are divided into articles. Articles can be arranged in groups and subdivisions. In the end of an act are the mention of compulsory character of regulation, concluding formulas, and annexes.

4.2.1 The scope and application of GDPR

First consideration that must be done, is to find out what is the scope of GDPR and when the regulation has to be applied. The Article 2(1) in Regulation (EU) 2016/679 of the European Parliament and of the Council states, that:

> *"This Regulation applies to the processing of personal data wholly or partly by automated means and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system."*

According to the definition, GDPR applies to any kind of operation that is performed on personal data, whether short or long-term use or large amounts or small subset of data. Personal data are defined in Article 4(1) (Regulation (EU) 2016/679 2016) as follows:

> *"'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;"*

It is clear that GDPR applies to data, which is stored information and can be related to an identified or identifiable person. Data identifies the person if the person can be detected directly or indirectly using any kind of characterlike identifiers or a combination of different information. Identifiability, the possibility of identification for example using additional information, is enough to make data personal. However, there is suggestions based on the interpretations of previous legislation that the data are not personal, and person is not considered identifiable, if the data controller or processor could not possibly gain access to missing information that would make identification possible (Voigt and von dem Bussche 2017).

4.2.2 Controller and processor

As Article 4(7) defines, "'controller' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data" (Regulation (EU) 2016/679 2016). The controller can determine the purposes and means of the processing. Thus, the controllership depends on who makes the decisions. To identify the decision maker, Voigt et al. (2017) suggest asking the questions: "why does the processing take place, and who initiated it?".

Processor is another entity defined in GDPR, which "means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller" (Regulation (EU) 2016/679 2016). The controller decides who is processing the data on behalf of it. Processor has to be a separate legal entity or individual, which is processing personal data on behalf of the controller. (Voigt and von dem Bussche 2017.)

A joint controller is defined in Article 26(1) (Regulation (EU) 2016/679 2016) as controllers who jointly and transparently determine the purposes and means of processing and their respective responsibilities. A processor will become controller if processor takes a role in determining the essential means and purpose of processing (Voigt and von dem Bussche 2017).

4.2.3 Anonymization and pseudonymization

Anonymization and pseudonymization of the data are two important concepts in the respect of General Data Protection Regulation. Anonymization is a “process that removes the association between the identifying data set and the data subject” (ISO/TS 25237:2008 2008). Anonymous data are according to Recital 26 (Regulation (EU) 2016/679 2016) “information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable”. According to Recital 26 (Regulation (EU) 2016/679 2016), the GDPR does not apply if the data are anonymized.

El Emam and Arbuckle (2013) present two general types of anonymization exists: masking and de-identification. Masking distorts the data in such a way that identification is not possible. In de-identification the information is removed or generalized preventing identifying individuals. (El Emam and Arbuckle 2013.)

Another technique used to protect the privacy of the individuals when dealing personal data are pseudonymization. Pseudonymization is a “particular type of de-identification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms” (ISO/TS 25237:2008 2008).

Article 4(5) (Regulation (EU) 2016/679 2016) defines pseudonymization as a way to present data in a way that it’s not attributed to a person without additional information. The additional linking information has to be kept separately and both technical and organizational measures must be applied to keep the data secure. Pseudonymized data are personal data and still under the scope of GDPR as there is a greater risk of re-identification compared to anonymous data, but pseudonymization might help processors and controllers to meet the data protection obligations as pseudonymization has the potential to guarantee data privacy if applied correctly (Voigt and von dem Bussche 2017).

However, there is a possibility that pseudonymized data might be anonymous under certain circumstances. Mourby, Mackey, Elliot, Gowans, Wallace, Bell, Smith, Aidinlis and Kaye (2018) argue that pseudonymized data can be rendered anonymous and pseudonymized data in some organization can be anonymous for another organization. They appeal to the statement in Recital 26 (Regulation (EU) 2016/679 2016), which states that "account should be taken of all the means reasonably likely to be used" in determining whether a person is identifiable. According to GDPR, pseudonymization is a way of processing data rather than a way of determining if data are personal (Mourby et al. 2018).

An interpretation of a legal case Patrick Breyer v Bundesrepublik Deutschland (2016) gives some implication that pseudonymized data could be anonymous. The key point is whether the relationship between two parties, the controller of the pseudonymized data and a third party, is such that the third party has according to Recital 26 (Regulation (EU) 2016/679 2016) any means reasonably likely to be used to identify the individual. It might be possible that the pseudonymized data are personal for some entity and anonymous for another entity, if the later has no means reasonably likely to be used to access the identifying information. (Mourby et al. 2018.)

4.2.4 Data protection by design and by default

The concepts of data protection by design and by default is defined in Article 25 (Regulation (EU) 2016/679 2016). According to the definition, "appropriate technical and organizational measures" has to be implemented in order to protect the rights of data subjects. Article 25(1) specifically mentions pseudonymization and data minimization as ways to protect data. The principle of data minimization in Article 5(1c) (Regulation (EU) 2016/679 2016) states that personal data shall be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed". In other words, only necessary data should be collected and pseudonymized in a proper way. Protective measures should be taken into account already when developing and designing products, services and applications.

Data protection by default in Article 25(2) (Regulation (EU) 2016/679 2016) emphasizes the fact that “only personal data which are necessary for each specific purpose of the processing are processed”. The measures have to meet by default the privacy requirements and obligations of the GDPR that apply to the “amount of personal data collected, the extent of their processing, the period of their storage and their accessibility”.

4.2.5 Implications of GDPR on learning analytics

The General Data Protection Regulation has many important implications on learning analytics and other handling of personal data within educational institutions. Hoel and Chen (2016) present in their seminal paper some implications of GDPR for learning analytics design. They conclude openness, transparency and continuous negotiation between data subjects and data processors are the key principles for further research.

The requirements of the GDPR have to be taken into account by design and by default. A person can give consent to use personal data in an ethically conducted scientific research purpose even the research purpose is not clear at the data collection time (Recital 33 GDPR). In that case the educational institution may use the data to conduct own learning analytics research. However, in addition to the already mentioned requirements of the GDPR, the processor has to design the learning analytics systems to comply also the following requirements:

1. *Lawfulness of the processing* (Article 6 GDPR): educational institution has to ask student to give permission to use private information in learning analytics. In this case the legal basis is the consent given by the student. Student can give also a consent to use data in ethically conducted learning analytics research.
2. *Right to erasure* (Article 17 GDPR): Student has a right to ask removal of personal data. Students may deny access to their data for learning analytics, as the legal basis is the student's permission.

3. *Data minimization* (Article 5(1c) GDPR): Only minimum viable amount of data should be collected in order to conduct the learning analytics.
4. *Purpose limitation* (Article 5(1b) GDPR): Student data is used only for the learning analytics purposes (i.e. to enable more efficient learning and better learning results).
5. *Security of Processing* (Article 32 GDPR): Learning analytics systems have to implement appropriate technical and organizational measures to ensure the privacy of personal data.
6. *Special protection of children's rights* (Recital 38 GDPR): Children are seen as less aware of the risks in relation to handling their personal data and thus they have special protection.
7. *Automated individual decision-making, including profiling* (Article 22 GDPR): Student can't be subject of a decision, which is based solely on automated processing including profiling, and if it significantly affects him or her. This doesn't apply if the automated decision-making is based on student's consent and special categories of data (Article 9 GDPR) are not used.
8. *Right to data portability* (Article 20 GDPR): Student can receive the personal data, which he or she has provided to the controller, in a structured format.
9. *"Right to explanation"* (Article 13-15 GDPR): As there is no direct mention about "Right to explanation" in GDPR, the data subject is still entitled to receive "meaningful information about the logic involved". At the current time there is arguments both in favor (e.g. Goodman and Flaxman 2016; Selbst and Powles 2017) and against (e.g. Wachter, Mittelstadt and Floridi 2017) about whether this right exists and what does it mean.

The aforementioned list is non-exhaustive and future interpretations of the regulation might reveal new legal information. However, in order to facilitate scientific research, Recital 157 GDPR enables researchers to use personal data and registries for research purposes:

> *"Within social science, research on the basis of registries enables researchers to obtain essential knowledge about the long-term correlation of a number of social conditions such as unemployment and education with other life conditions. Research results obtained through registries provide solid, high-quality knowledge which can provide the basis for the formulation and implementation of knowledge-based policy, improve the*

> *quality of life for a number of people and improve the efficiency of social services. In order to facilitate scientific research, personal data can be processed for scientific research purposes, subject to appropriate conditions and safeguards set out in Union or Member State law."*

The processing has to comply with GDPR and member state laws and as Recital 159 GDPR lays it down, where "personal data are processed for scientific research purposes, this Regulation should also apply to that processing". The objective of the GDPR is to protect the rights of individuals and their personal data while still allowing to use personal data for scientific research (Chassang 2017). However, some critical claims have been presented about how GDPR could restrict research (e.g. Nyrén, Stenbeck and Grönberg 2014; Kerr 2014).

4.3 Protection, privacy and ethics by design and by default

The main ethical goal of learning analytics is the learner success (Ferguson, Hoel, et al. 2016). The planning and designing of the learning analytics systems must follow the prevailing legislation and support the ethical goals of learning analytics. Country-specific legislation and especially within European Union the General Data Protection Regulation lays foundations for ethical design of the learning analytics systems from the perspective of data privacy and security. However, organization wide learning analytics policy is needed to complement the legal requirements in order to meet the ethical goals.

GDPR incorporates the idea of "protection by design and by default". In the process of designing learning analytics systems, there is also approach called "ethics by design" (e.g. Steiner, Kickmeier-Rust and Albert 2016). In addition, there exists a concept of "privacy by design" (e.g. Hustinx 2010). System designers need to address these issues from the technological point of view (Pardo and Siemens 2014) and already in the designing phase. By combining aforementioned concepts, I propose, that *protection, privacy and ethics by design and by default* (PPEDD) should be the guiding principle in learning analytics system design and policy formation. In addition to legal constraints, PPEDD is a design principle and it should function as the default functionality of a learning analytics system. Following summarizes the PPEDD principles:

- Protection by design and by default: e.g. Organizational and technical measures are implemented in order to protect learner's data and rights.
- Privacy by design and by default: e.g. Learner gets to evaluate and decide what data are given to the data controller (i.e. educational institution), and how and by whom data are used.
- Ethics by design and by default: e.g. Learning analytics aims for the success of a learner while respecting the rights of all stakeholders.

Within educational institution (Figure 8), PPEDD is the guiding principle in system design and should be applied in every step of the learning analytics process. Learning analytics policy is a course of action for institutionalized use of learning analytics in practice. Ethical principles and legal restrictions of data controller affect to the policy formation and lay down the facts how PPEDD is applied. Learners evaluate the learning analytics policy and how effectively the educational organization is implementing it.

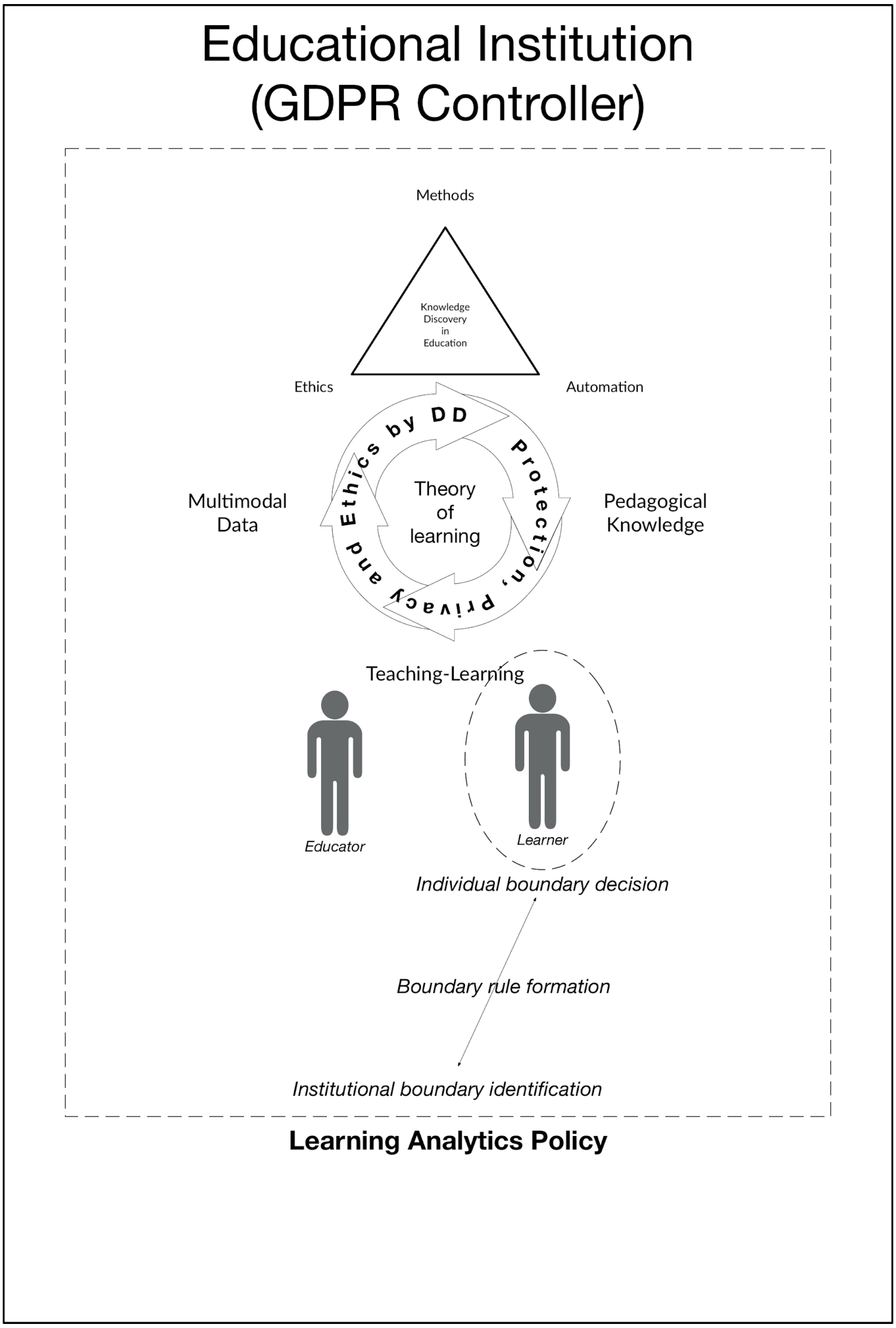


Figure 8. Legal regulation, PPEDD and learning analytics policy are guidelines for executing ethical learning analytics. Learner evaluates how successfully the institution applies these guidelines from the personal privacy perspective.

5 Automated learning analytics

5.1 Service oriented computing and architecture

The service-oriented computing (SOC) is a software design paradigm, whereby applications are composed of multiple networked services. A service is a reusable programmatic implementation of business functionality, which is wrapped in a documented interface and accessed via the network. One common example of service-oriented computing is a web service. A web service is identified by URI, accessed through Internet and implemented using Internet standards and protocols. (Papazoglou 2003; Papazoglou, Traverso, Dustdar and Leymann 2007.)

The networked services are the basic constructs of the applications, and they are used to perform wide range of various functions from a single task to more complex processes. A simple service performs a single functionality and a composite service combines many services into a more complex service. (Papazoglou 2003; Papazoglou et al. 2007.) In service-oriented computing applications use services by putting them together and composing larger constructs. It can provide means to simplify software development and create new value by reusing existing services. (Huhns and Singh 2005.)

The service-oriented architecture (SOA) is the key concept in realizing the service-oriented computing. The services in service-oriented computing follow particular design principles, which are a standardized service contract, loose coupling, abstraction, reusability, autonomy, statelessness, discoverability and composability (Erl 2008). The service-oriented architecture is an architectural style and a logical way of designing applications, which make use of the service-oriented computing design principles. The service-oriented architecture also depends on the relationship of the service provider, the service client, and the service discovery agent. The service provider publishes the service via the service discovery agent. The service client finds the service by using the service discovery agent and then binds with the service provider by using the service description. (Papazoglou 2003; Papazoglou et al. 2007.)

In educational institution there are several user groups and stakeholders. They all need different kinds of services and information. Service oriented approach could provide an efficient way to construct educational information systems that serve all user groups. Services can be inside the educational institution or external service providers. Next, microservice architecture is explored as a way to design and build systems.

5.2 Microservice architecture

A microservice architecture is a software design paradigm, which has gained interest in recent years (Figure 9). Many technology giants including Amazon, Netflix, Uber and Zalando are utilizing microservice architecture (Richardson 2017). Microservice architecture is an opposite design paradigm to monolithic architecture. Monolithic application is designed as a single, standalone application enclosing various program components. Microservice architecture consists of several modular software components and consists of multiple separate services (Figure 10).

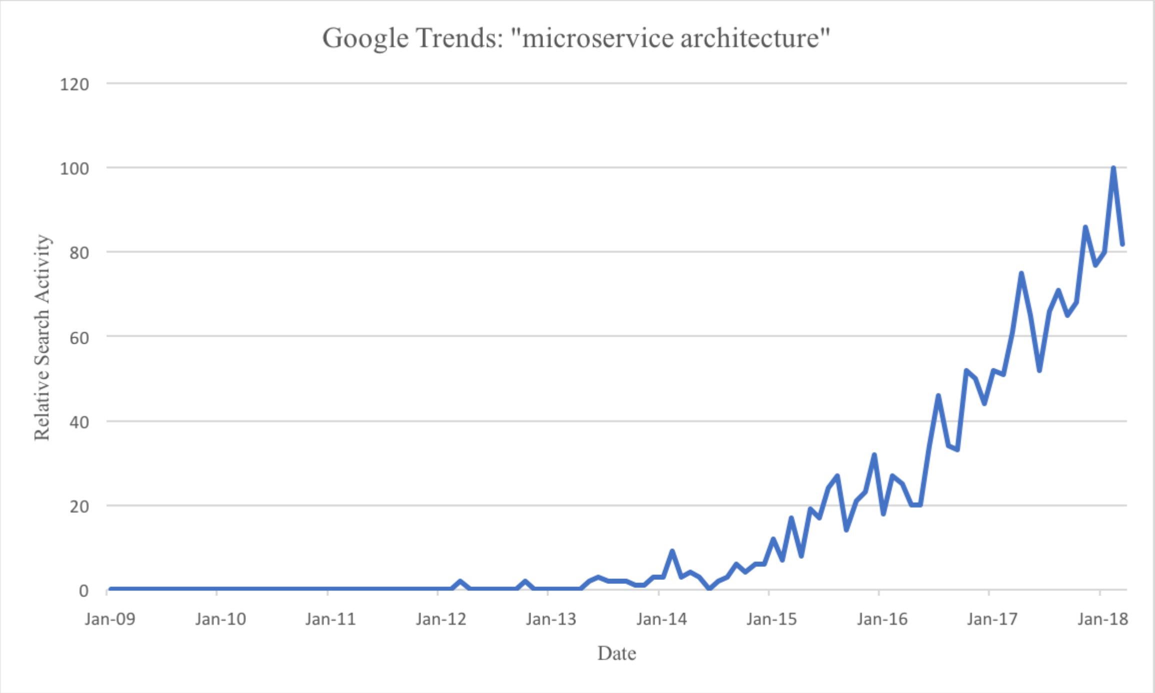
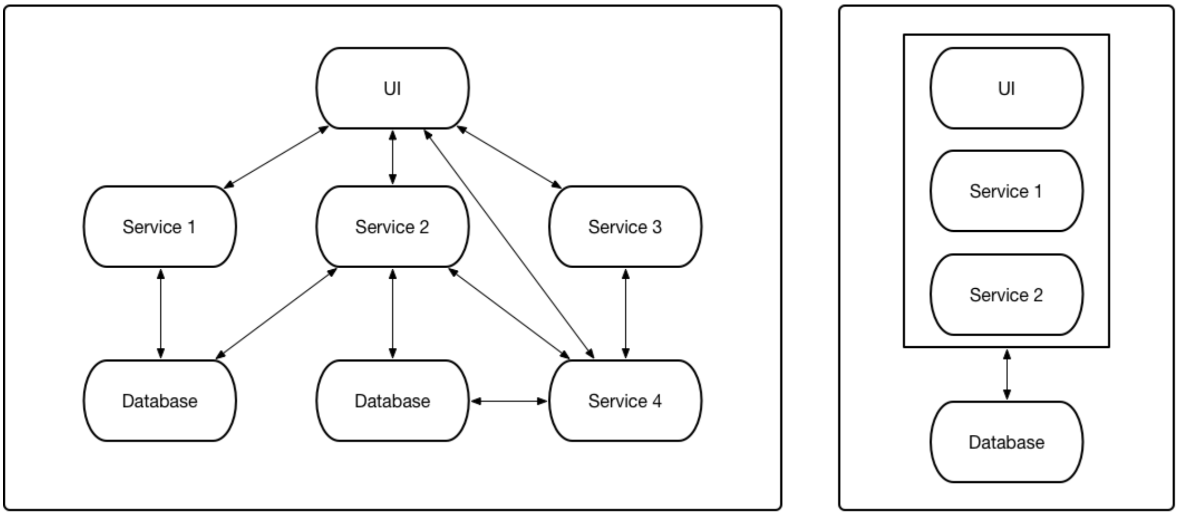


Figure 9. Relative search activity for keyword “microservice architecture” in Google Trends -service.



Microservices Architecture

Monolithic Architecture

Figure 10. Differences of microservice architecture and monolithic architecture.

Namiot and Sneps-Sneppe (2014) define microservice as a service that "is a lightweight and independent service that performs single functions and collaborates with other similar services using a well-defined interface". Nadareishvili, Mitra, McLarty and Amundsen (2016) describe microservices similarly but adding also the architectural dimension (ibid., 6):

> *"A microservice is an independently deployable component of bounded scope that supports interoperability through message-based communication. Microservice architecture is a style of engineering highly automated, evolvable software systems made up of capability-aligned microservices."*

Lewis and Fowler (2014) state that microservices have emerged from service-oriented computing. They also point out that some consider microservices as a subset of SOA while others reject the whole SOA categorization. They continue describing microservice architecture by listing a set of common characteristics of architectures that can be considered as microservices:

- Componentization via Services: Components of a software are independently deployable out-of-process services communicating, for example, via web service request or remote procedure call.

- Organized around Business Capabilities: The development work is organized as small cross-functional teams around business capabilities instead of siloed technical functionalities.
- Products not Projects: Microservice application development tries to avoid common project-based development model. Instead, the development team takes responsibility of the software for the whole life cycle. This enables "personal relationships between service developers and their users". As Amazon vice president Werner Vogels states it: "you build it, you run it" (Gray 2006).
- Smart endpoints and dumb pipes: Microservices are decoupled, cohesive, and intelligent endpoints, which receive a request, apply logic, and produce a response. Communication between services is handled, for example, using simple Representational State Transfer (REST) protocol rather than complex messaging protocols.
- Decentralized Governance: Decentralized approach lets developers to choose most suitable tools for the job. Instead of developing standards and other rules that strictly guide the development process, developers focus on producing tools that other developers can use to solve similar problems.
- Decentralized Data Management: Microservice can use own databases with different technology or shared database.
- Infrastructure Automation: Getting software safely and quickly into production is highly automated using software engineering techniques like Continuous Delivery.
- Design for failure: Microservice architecture has an additional layer of complexity comparing to monolithic architecture. Service-based systems have to be resilient to all kinds of situations where some service is unavailable.
- Evolutionary Design: Microservice is independently replaceable and upgradable, which enables better handling of change in the system.

Nadareishvili et al. (2016) describe the microservice design process in four steps. The first step is to identify the optimization goals. The microservice system is functioning properly if it helps to meet the optimization goals. The second step is to develop general design principles that address policies, constraints and ideals of the intended system design. The third step is the actual sketching of the system design. This step is highly iterative as all information might not be available from the start. Last, the fourth step is to implement, observe, and adjust the system to achieve the goals.

In learning analytics, the goal of a microservice is to optimize a specific task for analysis. Pedagogical learning analytics and related design principles, including legal restrictions and protection, privacy and ethics by design and by default, guide the design process. Empirical research should be used to validate the implemented system.

5.3 Representational State Transfer (REST)

As mentioned, microservices can communicate using Representational State Transfer. Representational State Transfer (REST) is developed by Roy Fielding (2000) in his doctoral dissertation. It is an architectural style, which is based on constraints and design decisions behind World Wide Web. The REST architecture style is derived from six constraints (Fielding 2000):

- Client-server: Client and server and their concerns are separated, e.g., user interface concerns are separated from data storage concerns. This improves portability of user interface and scalability of server components.
- Stateless: Client-server communication is stateless. Only client keeps the session state. Every request to the server contains the state information needed to complete the request.
- Cache: Data within a response is labeled as cacheable or non-cacheable. Cacheable data can be reused to improved performance.
- Uniform interface: Implementations and services they provide are decoupled by adding architectural constraints.
- Layered System: The overall architecture is composed of hierarchical layers.
- Code-On-Demand: This optional constraint allows client to request locally executable code.

Uniform interface is defined by four architectural constraints: identification of resources, manipulation of resources through representations, self-descriptive messages, hypermedia as the engine of application state. Clients identify and access resources in a networked system using identifier mechanism like Uniform Resource Identifiers (URIs). (Fielding 2000.) REST agents use uniform and predefined operations. In the case of HTTP protocol, common operations are CRUD operations like GET, POST, PUT, and DELETE (Booth et al. 2004). Client manipulates resources using representations, which are sequences of bytes including describing metadata. All messages between client and server must be self-descriptive, meaning requests and responses need to include metadata. Server responses contain hypermedia links, which guide the client how to use the service. (Fielding 2000.)

5.4 Automating learning analytics

Learning analytics and educational data mining uses wide range of different analytics methods and algorithms (e.g. Saarela and Kärkkäinen 2017; Baker 2010). The variety of available learning data are also vast (e.g. Blikstein and Worsley 2016). Considering the diversity of the both fields, developing valid learning analytics products takes a lot of time, effort and resources. It would benefit both practitioners and researcher, if these products could be accessed easily. A modularized view of learning analytics systems might enable the development and scientific collaboration between researchers (G. Siemens et al. 2011), practitioners and industry. Service oriented architecture and specifically microservices might provide an architectural approach for developing modular learning analytics systems. In this approach, the knowledge discovery process in education is automated using network of multiple learning analytics microservices (Figure 11).

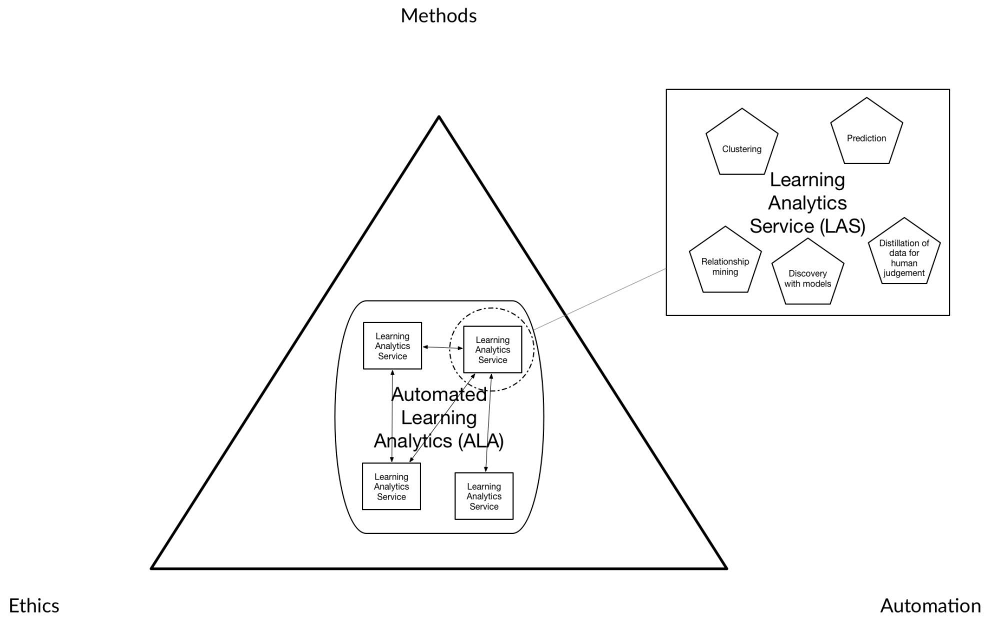

Methods
Ethics
Automation
Learning Analytics Service (LAS)
Clustering
Prediction
Relationship mining
Discovery with models
Distillation of data for human judgement
Automated Learning Analytics (ALA)
Learning Analytics Service
Learning Analytics Service
Learning Analytics Service
Learning Analytics Service

Figure 11. Automated Learning Analytics (ALA) is executed using microservice architecture. Learning Analytics Service (LAS) uses one or more learning analytics methods (e.g. Baker 2010).

Learning analytics service (LAS) is responsible of particular method or combination of methods (e.g. Baker 2010). It provides analytics about a particular learning activity. Learning analytics services can form a layered structure. They can be combined to provide more complex aggregate services. Services communicate for example using representational state transfer. Because of the benefits of REST and microservice architecture, services can be designed, developed, implemented and updated independently. Architecture enables also the use of external analytics services, which are provided by a separate service provider.

6 Framework for pedagogical learning analytics

Up to this point, I have constructed a knowledge base about pedagogical learning analytics, legal and ethical issues of the learning analytics, and one possible way of automating learning analytics services. The design artifact of this design science research is a framework for learning analytics in order to provide pedagogical knowledge to teachers. The framework combines the partial solutions presented in the knowledge base. The design artifact is then applied to a scenario of analyzing university student agency.

6.1 The framework artifact

The basis of the framework (Figure 12) is the concept of pedagogical learning analytics. Pedagogical learning analytics is an analytics cycle, which provides pedagogical knowledge to teachers. Teachers can use this knowledge as a building blocks for their own pedagogical knowledge base.

The pedagogical learning analytics starts from teaching-learning interaction. This interaction generates different kind of multimodal data traces, which are then collected and recorded. Automated educational knowledge discovery process is grounded in theory of learning. The analytics produces pedagogical knowledge, which teacher can utilize in the teaching-learning interaction.

Legal regulation (e.g. GDPR) and ethics of learning analytics constitute foundations for LAP and whole system design. The system design follows the principle of protection, privacy, and ethics by design and by default. Learning Analytics Policy describes the principles of the use of learning analytics within the educational institution. Learner evaluates these principles in relation to his or her own individual need of privacy.

Educational institution can complement its own analytics repertoire using external service providers. Service provider can be considered as data processor under GDPR if the data are used only processing on behalf of educational institution and not any other purpose. In any case, the processor has to comply with legal regulation, learning analytics policy and other contracts done with educational institution.

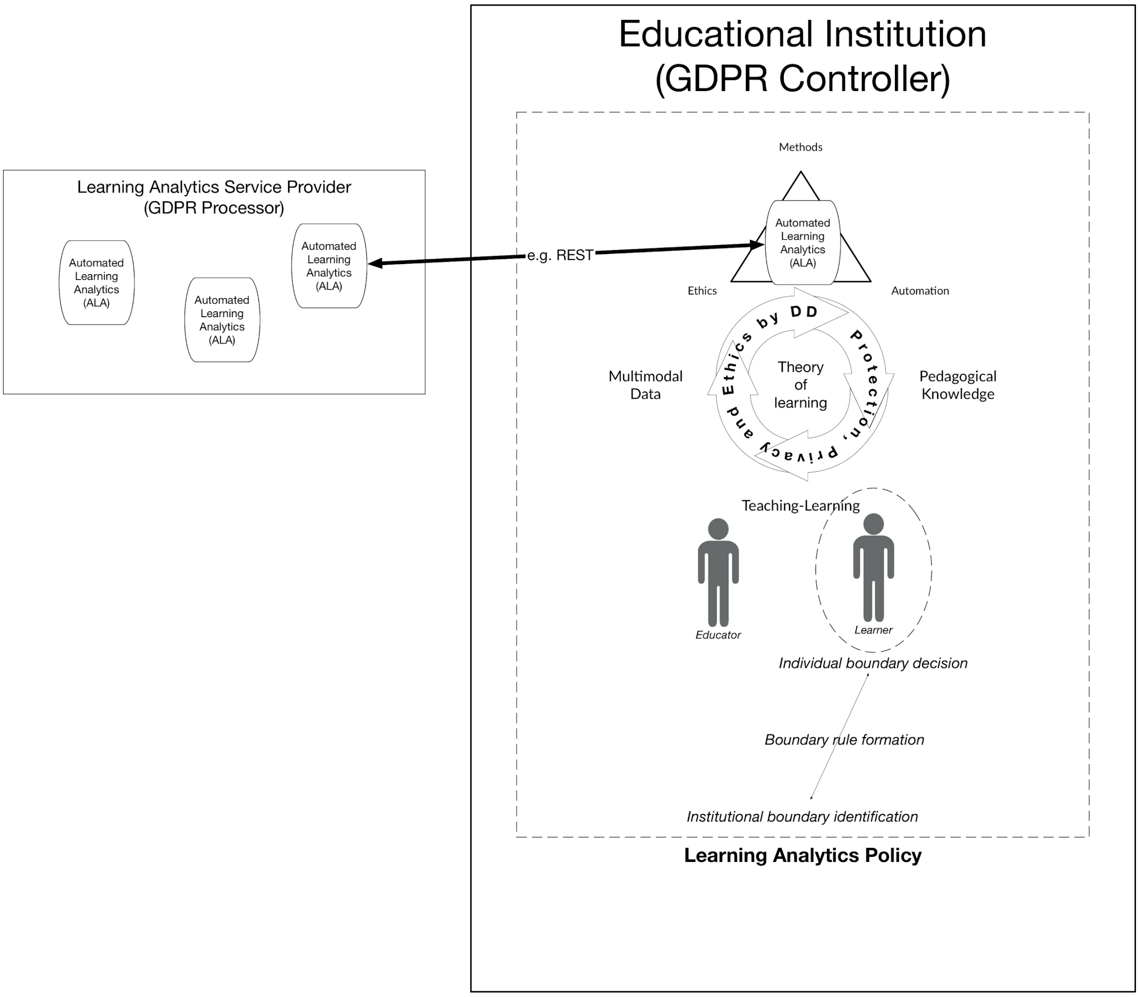

Educational Institution
(GDPR Controller)
Learning Analytics Service Provider
(GDPR Processor)
Automated Learning Analytics (ALA)
Automated Learning Analytics (ALA)
Automated Learning Analytics (ALA)
e.g. REST
Methods
Automated Learning Analytics (ALA)
Ethics
Automation
Ethics by DD Protection, privacy and
Multimodal Data
Theory of learning
Pedagogical Knowledge
Teaching-Learning
Educator
Learner
Individual boundary decision
Boundary rule formation
Institutional boundary identification
Learning Analytics Policy

Figure 12. Framework for pedagogical learning analytics (FPLA)

Automation takes advantage of microservice architecture. Learning analytics services inherit the benefits of the architecture. As independent and decoupled services they can be programmed with different programming languages. Analytical tasks can be divided as separate services, which can then form more complex services. Services communicate using Representational State Transfer, which enables client-server architecture and uniform interface. Thus, services are also replaceable and upgradable as they are dependent of each other and client implementations.

6.2 Scenario: Agency analytics

Learner agency is an important goal in education (e.g. OECD 2018). An interdisciplinary agency analytics group studied university student agency in the University of Jyväskylä under eEducation-project in 2017. The basic idea behind the scenario is that there are students in a university course and the teacher wants to collect and analyze the agencies of the students in order to amend teaching and provide individual counseling. Students also get individual agency results in comparison to all other students in the course. A simple web-based version of the analytics software was developed in order to demonstrate the agency analytics workflow from the student perspective.

The scenario presents the general idea behind the demo version. System design for pedagogical learning analytics is then applied to the presented agency analytics workflow and the result is evaluated using design science research evaluation framework. The goal of the agency analytics project was to demonstrate the agency analytics workflow and the use of a suitable educational data mining method, in this case robust clustering (Kärkkäinen and Äyrämö 2005; Äyrämö 2006). The AUS Scale questionnaire (Jääskelä et al. 2016) was transformed into a webform version using LimeSurvey, an open source online survey tool.

The analytics service was implemented as a microservice REST application programming interface written in Python. Individual agency profile of a student was calculated using the factor model. A robust clustering of all student responses provided the profiles of four different student agency groups. The individual agency and the group profiles were visualized and presented to the student as a web page. Teachers got the visualizations of different group profiles. This provided information about what kinds of agency groups they were having in their courses.

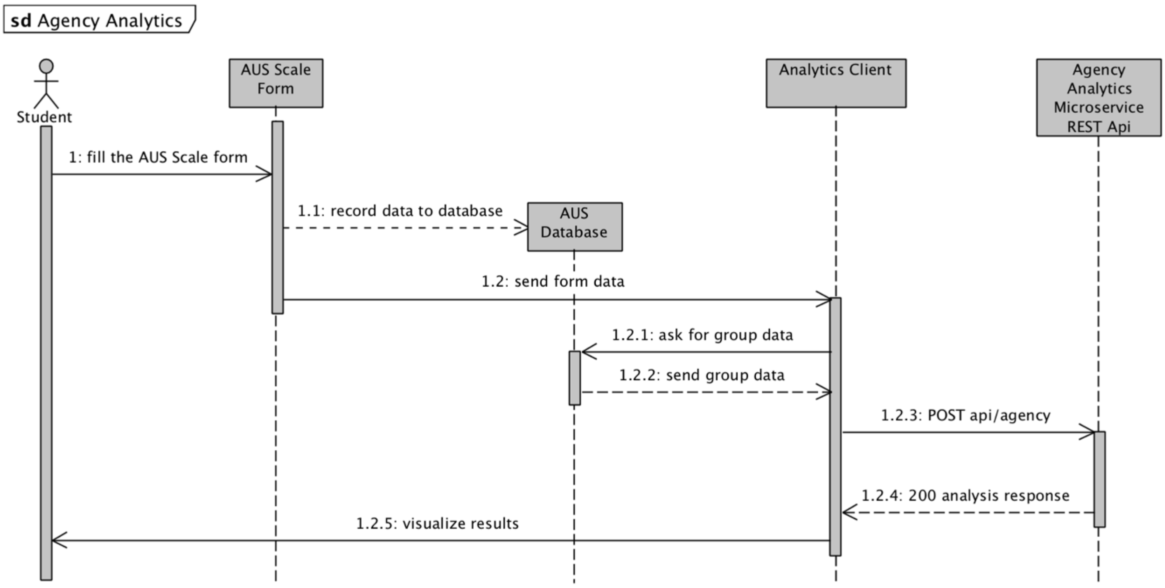


Figure 13. Agency Analytics UML sequence diagram

A sequence diagram is used in Unified Modeling Language (UML) to represent and model interaction between objects (Rumpe 2016). Figure 13 represents the UML sequence diagram of the agency analytics system in case of a one student. The students in the course first fill the AUS Scale web form (1) and the data is stored in a database (1.1). Agency data is represented as Likert-scale values between 0-5, where 0 indicates a missing value. The web form sends the data to analytics client (1.2), which is used as intermediary between the web form application (LimeSurvey) and analytics service. The analytics client also handles the data visualization. Analytics client can also be a Learning Management System (LMS) or any other service needing agency analytics services and results.

AUS database contains the responses of other students and in order to provide comparative agency information between individual student and the whole group of students, analytics client retrieves the results of other students (1.2.1 and 1.2.2). Analytics client pre-processes the data to comply with the predefined Application Programming Interface (API) in agency analytics service. It then makes a REST call (1.2.3) with POST method to analytics service. POST request is made to the agency resource address, and it contains the pre-processed individual agency data and the group data. Analytics service analyses the data and sends the response back to the client (1.1.4). The response contains the individual agency results of the student and clustering results of the whole course data.

The rest call is made with POST request. However, the analytics service does not store the data in database, only processes it. The benefit of this is that it might make the analysis safer from the student point of view. This way the data is stored in the database, which is owned by the educational institution and external services can be used only to process the data. Fielding and Reschke (2015, 24) present in RFC 7231 that POST method "requests that the target resource process the representation enclosed in the request according to the resource's own specific semantics". The POST method can be used for "providing a block of data, such as the fields entered into an HTML form, to a data-handling process". The response code 200 means that the request has succeeded. It is used in this agency analytics case instead of response code 201, which would mean that the data would have been recorded. With status code 200, the POST method can return the analytics result. According to the previously mentioned restrictions in REST style, the system has to conform to the uniform interface. As a result of these design decisions, the agency analytics service conforms with the interface using HTTP protocol.

The analytics client receives the analyzed data and makes the visualization of the results. The results represent the individual agency results in comparison to the group results. The illustrative example (Figure 14) represents the of agency analysis of a student. As it can be seen, all agency factors are close to the group average, except Factor 5 and Factor 6, which represent lower agency compared to the group.

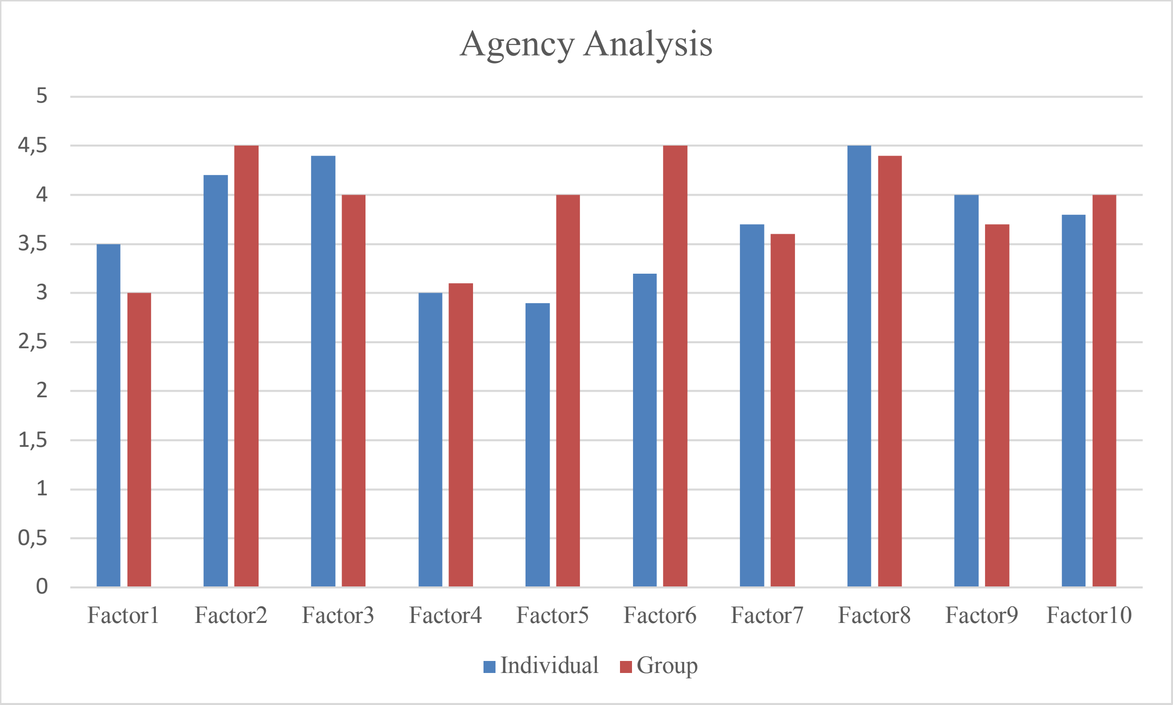


Figure 14. An illustrative example of agency analytics results of a student.

Robust clustering (Kärkkäinen and Ärämö 2005; Äyrämö 2006; Saarela, Hämäläinen and Kärkkäinen 2017) provides overall results of the whole group of students in the particular course (Figure 15). The service could be integrated into a learning management system (LMS) or a student information system (SIS). As it can be seen in the illustrative example of the agency group profiles, the Profile 1 has in general higher agency score than the other profiles (high agency group). The Profile 2 has average agency and Profile 4 is the low agency group in the example course. The Profile 3 is interesting group as it has varying agency across the profile. These agency profiles could provide novel and meaningful pedagogical knowledge to teachers.

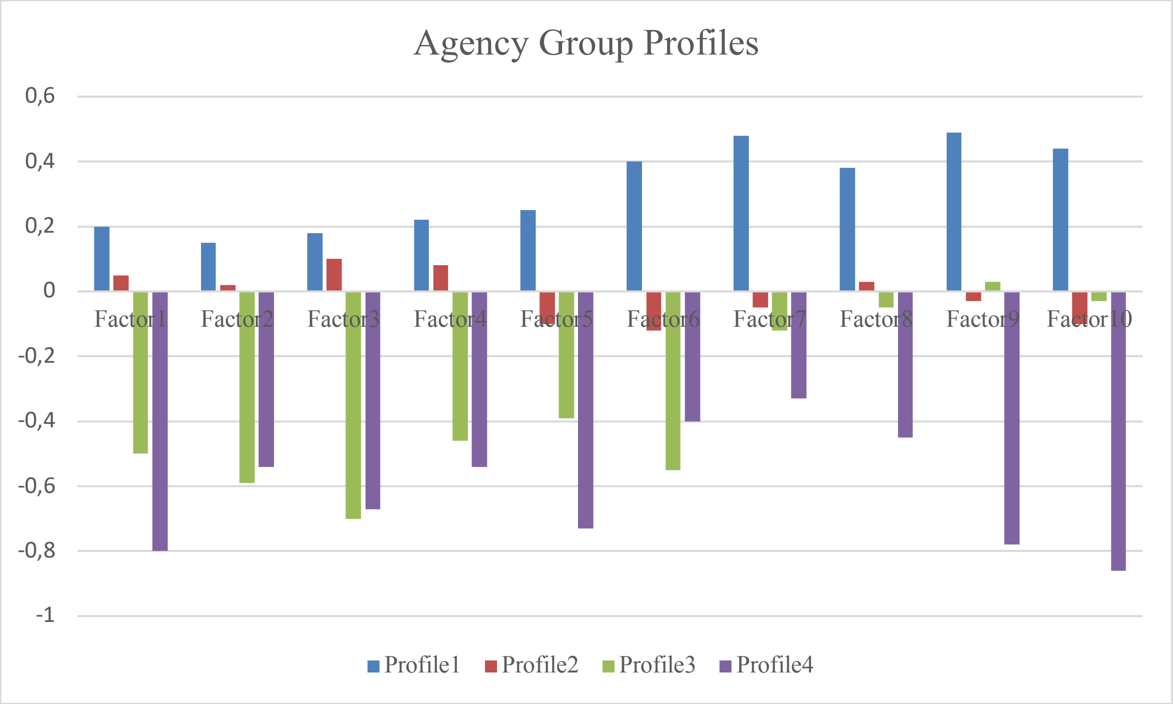


Figure 15. An illustrative example of four agency group profiles.

6.3 Evaluation

This research is a design science research and its result is a framework. Venable et al. (2016) propose a framework for evaluating artifacts in design science research. Based on their framework, they propose a four-step process for choosing an evaluation approach for design science research. The first step is defining the goals of the evaluation. The second step is choosing the evaluation strategy. The third step is determining what properties should be evaluated. The fourth step is designing the evaluation episodes. (Venable et al. 2016.)

In this research, the goals are derived based on research questions. The artifact where the framework for pedagogical learning analytics is applied, should provide pedagogical knowledge to teacher, it has to address the ethical issues and the analytics process has to be automated. The evaluation strategy is to show successful application of the framework in the scenario context. Prat, Comyn-Wattiau and Akoka (2014) present criterias for artifact evaluation. The criterias for evaluation of the agency analytics are efficacy, harnessing of recent technologies, correspondence with another model and robustness.

Efficacy relates to the goals (Prat et al. 2014) and how well the artifact produces it desired effects (Venable, Pries-Heje and Baskerville 2012). Wang and Wang (2010) argue that a valuable design science artifact is built on other new artifacts. Prat et al. (2014) call this as harnessing of recent technologies. Correspondence with another model can be characterized by construct redundancy. Robustness is an ability of an artifact to respond changes in an environment. (Prat et al. 2014.) The artifact framework is applied in the agency analytics scenario context and evaluated using aforementioned goals and criterias when applicable (Table 5).

	Pedagogical knowledge	**Ethical issues**	**Automation**
Efficacy	Needs further empirical research.	Research ethics was applied, and a research permit was asked. Learning analytics policy was not applied.	Microservice architecture and representational state transfer was successfully applied.
Harnessing of recent technologies	Latest knowledge and research about university student agency was applied.	GDPR effects were considered. Learning analytics policy was not applied.	Agency analytics process was based on recent technologies (e.g. microservices and robust clustering).
Correspondence with another model	No other model was applied.	No other model was applied.	Microservice architecture model was applied.
Robustness	Agency analytics process is relatively independent from environment but designed for higher education use.	Learning analytics policy was not applied.	Microservice architecture enables quick changes in analytics.

Figure 16. Evaluation of the framework for pedagogical learning analytics in agency analytics context.

The evaluation based on the pedagogical learning analytics framework seems to successfully reveal useful information about agency analytics process. It shows which areas in system design are covered and which need to be improved. The case example of agency analytics shows that automation corresponded with goals. The case example also utilized the latest technologies. However, the efficacy of pedagogical knowledge needs further research and learning analytics policy needs to be applied.

7 Discussion

Education is shifting rapidly towards intensive use of data. Learning analytics can be seen as one manifestation of this so called fourth paradigm of science in the context of education. As Thomas Kuhn (1970, 12) puts it simply, “the successive transition from one paradigm to another via revolution is the usual developmental pattern of mature science”. The paradigm shift is criticized, but still at this very time, education could be on the verge of data-intensive revolution. In Kuhnian sense, learning analytics give us a promise of acquiring higher understanding and knowledge of learning. Data-intensive education has a mythological “aura of truth, objectivity and accuracy” (Boyd and Crawford 2012, 663). Education plays a major role in human history and in the possible futures. It is at the forefront of every agenda. For this reason, it’s even our moral responsibility to investigate the possibilities of learning analytics.

Unfortunately, learning analytics has not yet managed to redeem its expectations as cornucopia of educational knowledge. There exist several fundamental issues that need to be solved. It’s a common stance to presume that methodological tools can be transferred from one field to another without ontological and epistemological assumptions (Perrotta and Williamson 2018). In the context of learning analytics and reformulating Wheeler (1990), can learning be inferred from digital information, bits? Learning analytics tries to reconstruct a learner as a “data double” and “it assumes that the learner can be perceived and understood scientifically as data, whilst also implying that the data construct itself is ontologically symmetrical with the person being represented” (Perrotta and Williamson 2018, 7). Ferguson et al. (2016) comment on the fact that there exists a gap between learning analytics and evidence of its effectiveness. Hoel and Chen (2016) also remind us about a gap between concerns and challenges of ethical implementations of learning analytics and proposals for design to solve these issues.

In the first research question, I pore over learning analytics from the teacher perspective. What kind of useful knowledge a teacher could obtain using learning analytics? The question seeks to find out how a teacher could benefit from learning analytics. Teachers operate among others on the basis of their pedagogical knowledge base and learning analytics could be one important source of this knowledge. In this research, I present the concept of pedagogical learning analytics. In order to do this, I combine learning analytics with the concept of pedagogical knowledge (e.g. Shulman, 1987; Voss et al. 2011). Clow (2012) presented a learning analytics cycle, which describes it as a cyclical process. I then apply the cyclical process to pedagogical learning analytics. The result is pedagogical learning analytics cycle, which aims to provide meaningful information to teachers. Pedagogical learning analytics makes use of educational knowledge discovery process in order to provide valid, novel and useful knowledge, which teachers can utilize when creating and optimizing teaching–learning situations and environments across subjects. In other words, pedagogical learning analytics is a tool for a practitioner. For researchers it can give a starting point for knowledge discovery. However, the concept has to be grounded on theory of learning. Further rigorous empirical research and practical implementations are needed to prove the effectiveness pedagogical learning analytics.

The second research question addresses the ethical concerns of learning analytics. What are the ethical challenges in learning analytics process? To summarize, the challenges are privacy, protection and ethical use. All learning data are personal and even if it's anonymous, learning data are still produced by a real person. In European Union, the General Data Protection Regulation lays down the legal responsibilities of data controllers and data processors. However, in learning analytics, everything that could be done legally might not be ethically justified. It is also important to realize, as Orlikowski and Iacono (2001, 131) argue, that "IT artifacts are designed, constructed, and used by people, they are shaped by the interests, values, and assumptions of a wide variety of communities of developers, investors, users". In case of learning analytics, it's essential to acknowledge these interests, values and assumptions for the sake of all stakeholders. The overall purpose of learning analytics is to benefit the learner. Ethics has the key role in promoting learning analytics in real world settings. Many educational institutions have introduced their own learning analytics policies (e.g. Tsai and Gasevic 2017) and few policy frameworks exist (e.g. DELICATE). Key aspect in the framework for pedagogical learning analytics is the implementation of a transparent learning analytics policy. Learners can evaluate the policy against their own need for privacy. In Europe, the upcoming interpretations of GDPR will affect to the formation of learning analytics policies. As a result, I add the idea of protection, privacy and ethics by design and by default to the pedagogical learning analytics cycle.

Last research question concerned about automation of learning analytics. Is it possible to automate learning analytics process? One possible way to automate learning analytics is by using microservice architecture. Siemens et al. (2011) have previously suggested a modularized construction of learning analytics system. Recently popular microservice architecture (e.g. Lewis and Fowler 2014) provides a modular, extendable and upgradable architectural solution for learning analytics services. Services communicate using Representational State Transfer (Fielding 2000). These allow the building of decoupled service, which also makes possible the use of external analytics services. Microservice architecture can also be used to build systems, which only process data instead of also storing the data. This can help in building more ethical learning analytics systems. The application of the framework in agency analytics context shows, that microservice architecture is suitable design choice for learning analytics systems, and it can be used to automate the analytics process.

Finally, all the findings are combined as a unified framework for pedagogical learning analytics (FPLA). Pedagogical learning analytics cycle, protection, privacy and ethics by design and by default and microservice architecture form the basis of FPLA. Knowledge about student agency is pedagogical knowledge, which teachers can use to design learning situations and interventions. The framework for pedagogical learning was applied to agency analytics workflow in order to find out what kind of information it could provide about the design. The selected criteria for evaluation of the framework were efficacy, harnessing of recent technologies, correspondence with another model and robustness (Prat et al. 2014). The application of the framework in the scenario context revealed that there is a need to validate a pedagogical model for agency analytics and to implement protection, privacy and ethics using secure system design and learning analytics policy. Thus, the framework is useful, and it can provide information about the scenario. However, more different kinds of evaluations and empirical validation should be done in the future, as only one scenario was used. Aforementioned is also the basis for the further development of the framework.

This research and the developed framework contribute to the practical use of learning analytics. The framework for pedagogical learning analytics could be used to fill the research gap between theory and practice. The framework can guide research towards pedagogically meaningful practice and provide starting points for a leaning analytics system design. This is one of the possibilities how technology might help us meet the goals of the better future and contribute beneficial change through education.

8 Conclusion and future work

8.1 Conclusion

The goal of my design science research was to find out what kind of useful knowledge a teacher could obtain using learning analytics in automated and ethical way. Firstly, I started by describing general knowledge discovery process, learning analytics and pedagogical knowledge. From this knowledge base I derived the new concept of pedagogical learning analytics. It provides valid, novel and useful knowledge, which teachers can utilize when creating and optimizing teaching–learning situations and environments across subjects.

Secondly, I explored the ethical issues from the legal and learning analytics policy perspective. Protection, privacy and ethics by design and by default is a guiding principle in learning analytics system design. Thirdly, I studied the possibility of using microservice architecture as an architectural approach. Microservice architecture could provide a flexible way to implement learning analytics systems.

Finally, I combined the three knowledge bases to a unified framework for pedagogical learning analytics (FPLA). The framework was applied and evaluated in a scenario, where university student agency was analyzed. The framework for pedagogical learning analytics could be used to guide learning analytics system design and future research. Hopefully results would someday support educators and learners in the future challenges.

8.2 Future work

Further research would be needed especially relating to learning analytics policies in European context. Rigorous and relevant empirical research would be needed to develop further and validate the idea of pedagogical learning analytics. There is also a need for more detailed architectural model of ethical and automated learning analytics.

Bibliography

Aalst, Wil M. P. van der. 2011. “Data Mining.” In *Process Mining*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-19345-3_3.

Ackoff, Russell L. 1989. “From Data to Wisdom.” *Journal of Applied Systems Analysis* 16 (1): 3–9.

Aggarwal, C., and P. Yu. 2008. “A General Survey of Privacy-Preserving Data Mining Models and Algorithms.” In *Privacy-Preserving Data Mining: Models and Algorithms*, edited by C. Aggarwal and P. Yu, 11–52. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-70992-5_2.

Akiba, M., G. LeTendre, and J. Scribner. 2007. “Teacher Quality, Opportunity Gap, and National Achievement in 46 Countries.” *Educational Researcher* 36 (7). American Educational Research Association: 369–87. https://doi.org/10.3102/0013189X07308739.

Anderson, Chris. 2008. “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” *Wired Magazine* 16 (7): 16–07. Accessed March 20, 2018. https://www.wired.com/2008/06/pb-theory/.

Äyrämö, Sami. 2006. *Knowledge Mining Using Robust Clustering*. University of Jyväskylä.

Baker, R. 2010. “Data Mining.” In *International Encyclopedia of Education*, edited by P. Peterson, E. Baker, and B. McGaw, 3rd ed., 112–18. Amsterdam: Elsevier Science.

Bandura, Albert. 2006. “Toward a Psychology of Human Agency.” *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 1 (2): 164–80. https://doi.org/10.1111/j.1745-6916.2006.00011.x.

"big, adj. and adv.". OED Online. 2018. Oxford University Press. Accessed April 17, 2018. http://www.oed.com.ezproxy.jyu.fi/view/Entry/18833?redirectedFrom=big+data.

Blikstein, Paulo, and Marcelo Worsley. 2016. “Multimodal Learning Analytics and Education Data Mining: Using Computational Technologies to Measure Complex Learning Tasks.” *Journal of Learning Analytics* 3 (2): 220–38. https://doi.org/10.18608/jla.2016.32.11.

Booth, David, Hugo Haas, Francis McCabe, Eric Newcomer, Michael Champion, Chris Ferris, and David Orchard. 2004. “Web Services Architecture.” World Wide Web Consortium. Accessed April 15, 2018. https://www.w3.org/TR/ws-arch/.

Boyd, D., and K. Crawford. 2012. “Critical Questions for Big Data.” *Information, Communication & Society* 15 (5): 662–79. https://doi.org/10.1080/1369118X.2012.678878.

Patrick Breyer v Bundesrepublik Deutschland. 2016. Case C-582/14 Patrick Breyer v Bundesrepublik Deutschland [2016] ECLI: EU: C: 2016:779.

Canales, Andrea, and Luis Maldonado. 2018. “Teacher Quality and Student Achievement in Chile: Linking Teachers’ Contribution and Observable Characteristics.” *International Journal of Educational Development* 60 (May): 33–50. https://doi.org/10.1016/j.ijedudev.2017.09.009.

Chang, Y., S. Wong, and H. Lee. 2015. “Understanding Perceived Privacy: A Privacy Boundary Management Model.” In *PACIS 2015 Proceedings*. Accessed March 15, 2018. http://aisel.aisnet.org/pacis2015/78.

Cheeseman, P. 1990. “On Finding the Most Probable Model.” In *Computational Models of Scientific Discovery and Theory Formation*, edited by Jeff Shrager and Pat Langley, 73–95. The Morgan Kaufmann Series in Machine Learning. California: Morgan Kaufmann Publishers.

Clow, Doug. 2012. “The Learning Analytics Cycle: Closing the Loop Effectively.” In *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge - LAK ’12*, 134–38. New York, NY: ACM. https://doi.org/10.1145/2330601.2330636.

Dahl, M. 2015. “Notat: Læringsanalyse.” Senter for IKT I Utdanningen. Accessed April 29, 2015. https://iktsenteret.no/ressurser/notat-laeringsanalyse.

Darling-Hammond, Linda. 2000. “Teacher Quality and Student Achievement.” *Education Policy Analysis Archives* 8 (0): 1. https://doi.org/10.14507/epaa.v8n1.2000.

"data, n.". OED Online. 2018. Oxford University Press. Accessed April 17, 2018. http://www.oed.com.ezproxy.jyu.fi/view/Entry/296948?rskey=lkSGV8&result=1.

Davis, Niki. 2017. *Digital Technologies and Change in Education: The Arena Framework*. Abingdon-on-Thames: Routledge.

Day, Christopher. 2017. *Teachers' Worlds and Work: Understanding Complexity, Building Quality*. Abingdon-on-Thames: Routledge.

Demchenko, Y., P. Grosso, C. de Laat, and P. Membrey. 2013. "Addressing Big Data Issues in Scientific Data Infrastructure." In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 48–55. https://doi.org/10.1109/CTS.2013.6567203.

Diebold, F. 2012. "On the Origin(s) and Development of the Term 'Big Data.'" *St. Louis: Federal Reserve Bank of St Louis*. https://doi.org/10.2139/ssrn.2152421.

Drachsler, Hendrik, and Wolfgang Greller. 2016. "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*, 89–98. LAK '16. New York, NY: ACM. https://doi.org/10.1145/2883851.2883893.

educationaldatamining.org. n.d. "Educationaldatamining.org." Educationaldatamining.org. Accessed March 20, 2018. http://educationaldatamining.org/.

El Emam, Khaled, and Luk Arbuckle. 2013. *Anonymizing Health Data: Case Studies and Methods to Get You Started*. Sebastopol, CA: O'Reilly Media.

Emirbayer, Mustafa, and Ann Mische. 1998. "What Is Agency?" *American Journal of Sociology* 103 (4): 962–1023. http://www.journals.uchicago.edu/doi/abs/10.1086/231294.

Erl, Thomas. 2008. *Soa: Principles of Service Design*. Upper Saddle River, NJ: Prentice Hall.

Eteläpelto, Anneli, Katja Vähäsantanen, Päivi Hökkä, and Susanna Paloniemi. 2013. "What Is Agency? Conceptualizing Professional Agency at Work." *Educational Research Review* 10 (Supplement C): 45–65. https://doi.org/10.1016/j.edurev.2013.05.001.

European Comission. 2016. "Learning Analytics - Key Messages." Edited by ET 2020 Working Group on Digital Skills and Competences. European Comission. Accessed March 20, 2018. https://ec.europa.eu/education/sites/education/files/2016-pla-learning-analytics_en.pdf.

Famili, A., Wei-Min Shen, Richard Weber, and Evangelos Simoudis. 1997. “Data Preprocessing and Intelligent Data Analysis.” *Intelligent Data Analysis* 1 (1): 3–23. https://doi.org/10.1016/S1088-467X(98)00007-9.

Fan, Wei, and Albert Bifet. 2013. “Mining Big Data: Current Status, and Forecast to the Future.” *SIGKDD Explor. Newsl.* 14 (2). New York, NY, USA: ACM: 1–5. https://doi.org/10.1145/2481244.2481246.

Fayyad, Usama M., Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996a. “Data Mining and Knowledge Discovery in Databases.” *AI Magazine* 17 (3). Association for the Advancement of Artificial Intelligence (AAAI): 37–54. https://doi.org/10.1145/240455.240463.

Fayyad, Usama M., Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996b. “From Data Mining to Knowledge Discovery: An Overview.” In *Advances in Knowledge Discovery and Data Mining*, edited by Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, 1–34. California: AAAI Press / The MIT Press.

Fayyad, Usama M., Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996c. “Knowledge Discovery and Data Mining: Towards a Unifying Framework.” *KDD-96 Proceedings*. Accessed January 15, 2018. https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf.

Fayyad, Usama M., Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996d. “The KDD Process for Extracting Useful Knowledge from Volumes of Data.” *Communications of the ACM* 39 (11). New York, NY, USA: ACM: 27–34. https://doi.org/10.1145/240455.240464.

Ferguson, Rebecca, Andrew Brasher, Doug Clow, Adam Cooper, Garron Hillaire, Jenna Mittelmeier, Bart Rienties, Thomas Ullmann, and Riina Vuorikari. 2016. “Research Evidence on the Use of Learning Analytics: Implications for Education Policy.” Edited by Riina Vuorikari and Jonatan Castaño Muñoz, Science for Policy Reports, . Seville, Spain: Joint Research Centre. https://doi.org/10.2791/955210.

Ferguson, Rebecca, and Doug Clow. 2017. “Where Is the Evidence? A Call to Action for Learning Analytics.” In *LAK ’17 Proceedings of the Seventh International*

Learning Analytics & Knowledge Conference, 56–65. ACM International Conference Proceeding Series. New York, USA: ACM. https://doi.org/10.1145/3027385.3027396.

Ferguson, Rebecca, Tore Hoel, Maren Scheffel, and Hendrik Drachsler. 2016. "Guest Editorial: Ethics and Privacy in Learning Analytics." *Journal of Learning Analytics* 3 (1): 5–15. https://doi.org/10.18608/jla.2016.31.2.

Fielding, Roy. 2000. "Architectural Styles and the Design of Network-Based Software Architectures." Edited by R. Taylor. Phd, University of California, Irvine. Accessed February 4, 2018. https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf.

Fielding, R., and J. Reschke, eds. n.d. "Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content." IETF. Accessed April 16, 2018. https://doi.org/10.17487/RFC7231.

Floridi, Luciano. 2004. "Open Problems in the Philosophy of Information." *Metaphilosophy* 35 (4): 554–82. https://doi.org/10.1111/j.1467-9973.2004.00336.x.

Floridi, Luciano. 2011. *The Philosophy of Information*. Oxford: Oxford University Press.

Gandomi, Amir, and Murtaza Haider. 2015. "Beyond the Hype: Big Data Concepts, Methods, and Analytics." *International Journal of Information Management* 35 (2): 137–44. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Giddens, A. 1984. *The Constitution Of Society: Outline of the Theory of Structuration*. Cambridge: Polity Press.

Goodman, Bryce, and Seth Flaxman. 2017. "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation.'" *AI Magazine* 38 (3): 50. https://doi.org/10.1609/aimag.v38i3.2741.

Gorton, I., P. Greenfield, A. Szalay, and R. Williams. 2008. "Data-Intensive Computing in the 21st Century." *Computer* 41 (4): 30–32. https://doi.org/10.1109/MC.2008.122.

Gray, Jim. 2006. "A Conversation with Werner Vogels." *ACM Queue: Tomorrow's Computing Today* 4 (4): 14–22.

Guerriero, Sonia. 2013. "Teachers' Pedagogical Knowledge and the Teaching Profession." OECD. Accessed April 17, 2018.

http://www.oecd.org/education/ceri/Background_document_to_Symposium_ITEL-FINAL.pdf.

Guerriero, Sonia. 2017. "Teachers' Pedagogical Knowledge: What It Is and How It Functions." *Educational Research and Innovation*. Organisation for Economic Cooperation and Development (OECD), 99–118. Accessed 17 February, 2018. http://www.oecd-ilibrary.org/education/pedagogical-knowledge-and-the-changing-nature-of-the-teaching-profession/teachers-pedagogical-knowledge-what-it-is-and-how-it-functions_9789264270695-6-en.

Haggard, Patrick. 2017. "Sense of Agency in the Human Brain." *Nature Reviews. Neuroscience* 18 (4): 196–207. https://doi.org/10.1038/nrn.2017.14.

Han, Jiawei, Jian Pei, and Micheline Kamber. 2011. *Data Mining: Concepts and Techniques*. Burlington, MA: Elsevier.

Hanushek, Eric A. 2011. "The Economic Value of Higher Teacher Quality." *Economics of Education Review* 30 (3): 466–79. https://doi.org/10.1016/j.econedurev.2010.12.006.

Hevner, Alan R., Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. "Design Science in Information Systems Research." *The Mississippi Quarterly* 28 (1). Management Information Systems Research Center, University of Minnesota: 75–105. https://doi.org/10.2307/25148625.

Hey, T., S. Tansley, and K. Tolle, eds. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Second. Redmond, Washington: Microsoft Research. Accessed April 5, 2018. https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf.

Hitlin, Steven, and Glen H. Elder. 2007. "Time, Self, and the Curiously Abstract Concept of Agency." *Sociological Theory* 25 (2). SAGE Publications Inc: 170–91. https://doi.org/10.1111/j.1467-9558.2007.00303.x.

Hoel, Tore, and Weiqin Chen. 2016. "Implications of the European Data Protection Regulations for Learning Analytics Design." In *Presentation at The International Workshop on Learning Analytics and Educational Data Mining (LAEDM 2016) in Conjunction with the International Conference on Collaboration Technologies (CollabTech 2016), Kanazawa, Japan-September*, 14–16. Accessed March 24,

2018. http://www.hoel.nu/files/LAEDM_Kanazawa_Sep2016_Hoel_Chen_final_w_header.pdf.

Huhns, M. N., and M. P. Singh. 2005. “Service-Oriented Computing: Key Concepts and Principles.” *IEEE Internet Computing* 9 (1): 75–81. https://doi.org/10.1109/MIC.2005.21.

Hustinx, Peter. 2010. “Privacy by Design: Delivering the Promises.” *Identity in the Information Society* 3 (2): 253–55. https://doi.org/10.1007/s12394-010-0061-z.

ISO/TS 25237:2008. 2008. “Health Informatics: Pseudonymization.” ISO. 2008.

Jääskelä, Päivikki, Anna-Maija Poikkeus, Kati Vasalampi, Ulla Maija Valleala, and Helena Rasku-Puttonen. 2016. “Assessing Agency of University Students: Validation of the AUS Scale.” *Studies in Higher Education*, February. Routledge, 1–19. https://doi.org/10.1080/03075079.2015.1130693.

Jeannerod, Marc. 2003. “The Mechanism of Self-Recognition in Humans.” *Behavioural Brain Research* 142 (1): 1–15. https://doi.org/10.1016/S0166-4328(02)00384-4.

Kantardzic, Mehmed. 2011. *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons.

Kärkkäinen, T., and S. Ayrämö. 2005. “On Computation of Spatial Median for Robust Data Mining.” In *Proceedings of the EUROGEN 2005 Conference,* edited by R. Schilling, W. Haase, J. Periaux, H. Baier and G. Bugeda. Münich: FLM.

Kerr, D. 2014. “Policy: EU Data Protection Regulation—harming Cancer Research.” *Nature Reviews. Clinical Oncology* 11 (10), 563-563. London: Nature Publishing Group.

Kitchin, R. 2014. “Big Data, New Epistemologies and Paradigm Shifts.” *Big Data & Society* 1 (1). SAGE Publications: 2053951714528481. https://doi.org/10.1177/2053951714528481.

Krokfors, L., Marjaana K., Kopisto, K., Rikabi-Sukkari, L., Salo, L., and O. Vesterinen. 2015. “Learning. Creatively. Together. Educational Change Report 2016.” University of Helsinki, Faculty of Behavioural Sciences. Accessed April 17, 2018. https://helda.helsinki.fi/bitstream/handle/10138/156502/LearningCreativelyTogether_Educational_Change_Report_2016.pdf?sequence=4.

Kuhn, Thomas S. 1970. *The Structure of Scientific Revolutions*. Second. Chicago: University of Chicago Press.

LAK11. 2010. “About.” LAK11 1st International Conference on Learning Analytics and Knowledge 2011. July 22, 2010. Accessed April 2, 2018. https://web.archive.org/web/20101220223648/https://tekri.athabascau.ca/analytics/.

Laney, Doug. 2001. “3D Data Management: Controlling Data Volume, Velocity and Variety.” File: 949. Stamford: META Group Inc. Accessed January 5, 2018. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Lee, Allen S., Manoj Thomas, and Richard L. Baskerville. 2015. “Going back to Basics in Design Science: From the Information Technology Artifact to the Information Systems Artifact.” *Information Systems Journal* 25 (1): 5–21. https://doi.org/10.1111/isj.12054.

Leonelli, S. 2014. “What Difference Does Quantity Make? On the Epistemology of Big Data in Biology.” *Big Data & Society* 1 (1). Thousand Oaks: SAGE Publications. https://doi.org/10.1177/2053951714534395.

Lewis, James, and Martin Fowler. 2014. “Microservices: A Definition of This New Architectural Term.” Martinfowler.com. March 10, 2014. Accessed February 16, 2018. https://martinfowler.com/articles/microservices.html.

Loyal, Steven, and Barry Barnes. 2001. “‘Agency’ as a Red Herring in Social Theory.” *Philosophy of the Social Sciences* 31 (4), 507–24. Thousand Oaks: SAGE Publications Inc. https://doi.org/10.1177/004839310103100403.

Maimon, Oded, and Lior Rokach. 2009. “Introduction to Knowledge Discovery and Data Mining.” In *Data Mining and Knowledge Discovery Handbook*, 1–15. New York: Springer. https://doi.org/10.1007/978-0-387-09823-4_1.

Maletic, Jonathan I., and Andrian Marcus. 2009. “Data Cleansing: A Prelude to Knowledge Discovery.” In *Data Mining and Knowledge Discovery Handbook*, 19–32. New York: Springer. https://doi.org/10.1007/978-0-387-09823-4_2.

Melamed, C., L. Morales, Y. Hsu, J. Poole, B. Rae, I. Rutherford, and A. Jahic. 2014. “A World That Counts - Mobilising the Data Revolution for Sustainable

Development." United Nations. Accessed March 12, 2018. http://www.undatarevolution.org/wp-content/uploads/2014/12/A-World-That-Counts2.pdf.

"metamodel, n.". OED Online. March 2018. Oxford University Press. Accessed March 3, 2018. http://www.oed.com.ezproxy.jyu.fi/view/Entry/245262?redirectedFrom=metamodel.

Mourby, Miranda, Elaine Mackey, Mark Elliot, Heather Gowans, Susan E. Wallace, Jessica Bell, Hannah Smith, Stergios Aidinlis, and Jane Kaye. 2018. "Are 'pseudonymised' data Always Personal Data? Implications of the GDPR for Administrative Data Research in the UK." *Computer Law & Security Review* 34 (2): 222-233. https://doi.org/10.1016/j.clsr.2018.01.002

Muñoz, Marco A., Joseph R. Prather, and James H. Stronge. 2011. "Exploring Teacher Effectiveness Using Hierarchical Linear Models: Student- and Classroom-Level Predictors and Cross-Year Stability in Elementary School Reading." *Planning and Changing* 42 (3/4), 241–73.

Nadareishvili, Irakli, Ronnie Mitra, Matt McLarty, and Mike Amundsen. 2016. *Microservice Architecture: Aligning Principles, Practices, and Culture*. Sebastopol: O'Reilly Media.

Namiot, D., Sneps-Sneppe, and M. 2014. "On Micro-Services Architecture." *International Journal of Open Information Technologies* 2 (9): 24-27.

Nyrén, Olof, Magnus Stenbeck, and Henrik Grönberg. 2014. "The European Parliament Proposal for the New EU General Data Protection Regulation May Severely Restrict European Epidemiological Research." *European Journal of Epidemiology* 29 (4): 227–30. https://doi.org/10.1007/s10654-014-9909-0.

OECD. 2018. "The Future of Education and Skills Education 2030." OECD. Accessed January 15, 2018. http://www.oecd.org/education/2030/OECD%20Education%202030%20Position%20Paper.pdf.

Offermann, Philipp, Sören Blom, Marten Schönherr, and Udo Bub. 2010. "Artifact Types in Information Systems Design Science--a Literature Review." In *International*

Conference on Design Science Research in Information Systems DESRITS 2010, 77–92. New York: Springer.

Oracle. 2014. “Oracle: Big Data for the Enterprise.” Oracle. Accessed February 12, 2018. http://www.oracle.com/technetwork/database/bigdata-appliance/overview/wp-bigdatawithoracle-1453236.pdf.

Orlikowski, Wanda J., and C. Suzanne Iacono. 2001. “Research Commentary: Desperately Seeking the ‘IT’ in IT Research—A Call to Theorizing the IT Artifact.” *Information Systems Research* 12 (2): 121–34. https://doi.org/10.1287/isre.12.2.121.9700.

Papazoglou, M. P. 2003. “Service-Oriented Computing: Concepts, Characteristics and Directions.” In *Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003. WISE 2003.*, 3–12. https://doi.org/10.1109/WISE.2003.1254461.

Papazoglou, M. P., P. Traverso, S. Dustdar, and F. Leymann. 2007. “Service-Oriented Computing: State of the Art and Research Challenges.” *Computer* 40 (11): 38–45. https://doi.org/10.1109/MC.2007.400.

Pardo, Abelardo, and George Siemens. 2014. “Ethical and Privacy Principles for Learning Analytics.” *British Journal of Educational Technology* 45 (3): 438–50. https://doi.org/10.1111/bjet.12152.

Parsons, Talcott. 1937. *The Structure of Social Action*. New York: Free Press.

Peffers, Ken, Marcus Rothenberger, Tuure Tuunanen, and Reza Vaezi. 2012. “Design Science Research Evaluation.” In *Design Science Research in Information Systems. Advances in Theory and Practice*, 398–410. NewYork: Springer. https://doi.org/10.1007/978-3-642-29863-9_29.

Peffers, Ken, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. 2007. “A Design Science Research Methodology for Information Systems Research.” *Journal of Management Information Systems* 24 (3): 45–77. Abingdon: Routledge. https://doi.org/10.2753/MIS0742-1222240302.

Perrotta, Carlo, and Ben Williamson. 2018. “The Social Life of Learning Analytics: Cluster Analysis and the ‘performance’of Algorithmic Education.” *Learning, Media and Technology* 43 (1): 3–16.

Petronio, Sandra. 2012. *Boundaries of Privacy: Dialectics of Disclosure*. Albany: SUNY Press.

Philip Chen, C. L., and Chun-Yang Zhang. 2014. “Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data.” *Information Sciences* 275 (Supplement C): 314–47. https://doi.org/10.1016/j.ins.2014.01.015.

Piety, Philip J., Daniel T. Hickey, and M. J. Bishop. 2014. “Educational Data Sciences: Framing Emergent Practices for Analytics of Learning, Organizations, and Systems.” In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, 193–202. LAK ’14. New York: ACM. https://doi.org/10.1145/2567574.2567582.

Prat, Nicolas, Isabelle Comyn-Wattiau, and Jacky Akoka. 2014. “Artifact Evaluation in Information Systems Design-Science Research-a Holistic View.” In *PACIS 2014 Proceedings*, 23.

"policy, n.1." OED Online, Oxford: Oxford University Press, Accessed 28 March 2018. www.oed.com/view/Entry/146842.

Prinsloo, Paul, and Sharon Slade. 2013. “An Evaluation of Policy Frameworks for Addressing Ethical Considerations in Learning Analytics.” In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, 240–44. LAK ’13. New York: ACM. https://doi.org/10.1145/2460296.2460344.

Regulation (EU) 2016/679. 2016. “Regulation (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation).” *Official Journal of the European Union (OJ)* 59 (L119): 1–88. Accessed January 23, 2018. http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN.

Renes, Robert A., Neeltje E. M. van Haren, Henk Aarts, and Matthijs Vink. 2015. “An Exploratory fMRI Study into Inferences of Self-Agency.” *Social Cognitive and Affective Neuroscience* 10 (5): 708–12. https://doi.org/10.1093/scan/nsu106.

Richardson, C. 2017. “Who Is Using Microservices?” Microservice Architecture. 2017. Accessed February 7, 2018. http://microservices.io/articles/whoisusingmicroservices.html.

Romero, Cristobal, and Sebastian Ventura. 2013. “Data Mining in Education.” *WIREs Data Mining and Knowledge Discovery* 3: 12–27.

Rumpe, Bernhard. 2016. “Modeling with UML.” *Language, Concepts, Methods. Springer International* 4. New York: Springer.

Saarela, Mirka, and Tommi Kärkkäinen. 2017. “Knowledge Discovery from the Programme for International Student Assessment.” In *Learning Analytics: Fundaments, Applications, and Trends*, edited by Alejandro Peña-Ayala, 229–267. Studies in Systems, Decision and Control 94. New York: Springer. https://doi.org/10.1007/978-3-319-52977-6_8.

Saarela, M., Hämäläinen, J., & Kärkkäinen, T. (2017). Feature Ranking of Large, Robust, and Weighted Clustering Result. In J. Kim, K. Shim, L. Cao, J.-G. Lee, X. Lin, & Y.-S. Moon (Eds.), *Advances in Knowledge Discovery and Data Mining* (Vol. 10234, pp. 96–109). Cham: Springer.

Schlosser, Markus E. 2015. “Agency.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Selbst, Andrew D., and Julia Powles. 2017. “Meaningful Information and the Right to Explanation.” *International Data Privacy Law* 7 (4): 233–42. Oxford: Oxford University Press. https://doi.org/10.1093/idpl/ipx022.

Shrager, Jeff, and Langley Pat. 1990. “Computational Approaches to Discovery.” In *Computational Models of Scientific Discovery and Theory Formation*, edited by Jeff Shrager and Pat Langley, 1–25. California: Morgan Kaufmann Publishers.

Shulman, Lee. 1987. “Knowledge and Teaching: Foundations of the New Reform.” *Harvard Educational Review* 57 (1): 1–23. Cambridge: Harvard Education Publishing Group.

Siemens, George, and Ryan S. J. d. Baker. 2012. “Learning Analytics and Educational Data Mining: Towards Communication and Collaboration.” In *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, 252–54. LAK ’12. New York: ACM. https://doi.org/10.1145/2330601.2330661.

Siemens, G., D. Gasevic, C. Haythornthwaite, S. Dawson, S. Buckingham Shum, R. Ferguson, E. Duval, K. Verbert, and R. Baker. 2011. “Open Learning Analytics: An Integrated & Modularized Platform.” SoLAR. Accessed January 21, 2018. http://www.elearnspace.org/blog/wp-content/uploads/2016/02/ProposalLearningAnalyticsModel_SoLAR.pdf.

Simpson, John. 2013. “A Heads Up for the June 2013 OED Release.” Oxford English Dictionary. June 2013. Accessed January 21, 2018. http://public.oed.com/the-oed-today/recent-updates-to-the-oed/previous-updates/june-2013-update/a-heads-up-for-the-june-2013-oed-release/.

Spengler, Stephanie, D. Yves von Cramon, and Marcel Brass. 2009. “Was It Me or Was It You? How the Sense of Agency Originates from Ideomotor Learning Revealed by fMRI.” *NeuroImage* 46 (1): 290–98. https://doi.org/10.1016/j.neuroimage.2009.01.047.

Staalduinen, J. 2015. “Policy & Policy Recommendations for Learning Analytics – A Literature Survey.” STELA Erasmus+ project. Accessed February 20, 2018. http://stela-project.eu/files/O2-part2-literatureSurvey-LearningAnalyticsPolicy&Recommendations.pdf.

Steiner, Christina M., Michael D. Kickmeier-Rust, and Dietrich Albert. 2016. “LEA in Private: A Privacy and Data Protection Framework for a Learning Analytics Toolbox.” *Journal of Learning Analytics* 3 (1): 66–90. https://doi.org/10.18608/jla.2016.31.5.

The University of Edinburgh. 2017. “Learning Analytics Principles and Purposes.” The University of Edinburgh. Accessed March 13, 2018. https://www.ed.ac.uk/files/atoms/files/learninganalyticsprinciples.pdf.

Tilly, C. 1980. “The Old New Social History and the New Old Social History.” *CRSO Working Paper N.o 218*. Accessed December 12, 2017. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/50992/218.pdf.

Tsai, Yi-Shan, and Dragan Gasevic. 2017. “Learning Analytics in Higher Education --- Challenges and Policies: A Review of Eight Learning Analytics Policies.” In *Proceedings of the Seventh International Learning Analytics & Knowledge*

Conference, 233–42. LAK ’17. New York: ACM. https://doi.org/10.1145/3027385.3027400.

UNESCO. 2017. “Unpacking Sustainable Development Goal 4 Education 2030.” Accessed January 13, 2018. http://unesdoc.unesco.org/images/0024/002463/246300E.pdf.

United Nations. 2015. “Transforming Our World: The 2030 Agenda for Sustainable Development.” Accessed January 13, 2018. https://sustainabledevelopment.un.org/post2015/transformingourworld/publication .

University of Gloucestershire. 2016. “Learning Analytics Policy.” University of Gloucestershire. Accessed March 13, 2018. http://www.glos.ac.uk/docs/download/Key/learning-analytics-policy.pdf.

University of West London. 2017. “Learning Analytics Policy.” University of West London. Accessed March 13, 2018. https://www.uwl.ac.uk/sites/default/files/Departments/About-us/Web/PDF/policies/uwl_learning_analytics_policy_final.pdf.

Van Barneveld, Angela, Kimberly E. Arnold, and John P. Campbell. 2012. “Analytics in Higher Education: Establishing a Common Language.” *EDUCAUSE Learning Initiative* 1 (1): l – ll.

Venable, John, Jan Pries-Heje, and Richard Baskerville. 2012. “A Comprehensive Framework for Evaluation in Design Science Research.” In *Design Science Research in Information Systems. Advances in Theory and Practice*, 423–38. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-29863-9_31.

Venable, John, Jan Pries-Heje, and Richard Baskerville. 2016. “FEDS: A Framework for Evaluation in Design Science Research.” *European Journal of Information Systems* 25 (1). Palgrave Macmillan UK: 77–89. https://doi.org/10.1057/ejis.2014.36.

Verloop, Nico, Jan Van Driel, and Paulien Meijer. 2001. “Teacher Knowledge and the Knowledge Base of Teaching.” *International Journal of Educational Research* 35 (5): 441–61. https://doi.org/10.1016/S0883-0355(02)00003-4.

Villegas-Reimers, E. 2003. “Teacher Professional Development: An International Review of the Literature.” Accessed April 17, 2018. http://unesdoc.unesco.org/images/0013/001330/133010e.pdf

Voigt, Paul, and Axel von dem Bussche. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. New York: Springer.

Voss, Thamar, Mareike Kunter, and Jürgen Baumert. 2011. “Assessing Teacher Candidates’ General Pedagogical/psychological Knowledge: Test Construction and Validation.” *Journal of Educational Psychology* 103 (4). American Psychological Association: 952.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.” *International Data Privacy Law* 7 (2): 76–99. Oxford: Oxford University Press. https://doi.org/10.1093/idpl/ipx005.

Wang, Shouhong, and Hai Wang. 2010. “Towards Innovative Design Research in Information Systems.” *Journal of Computer Information Systems* 51 (1): 11–18 Milton Park: Taylor & Francis. https://doi.org/10.1080/08874417.2010.11645445.

WCED. 1987. “Report of the World Commission on Environment and Development: Our Common Future.” Edited by G. Brundtland. United Nations. Accessed March 20, 2018. http://www.un-documents.net/our-common-future.pdf.

Wheeler, J. A. 1990. “Information, Physics, Quantum: The Search for Links.” *Complexity, Entropy, and the Physics of.* Accessed April 17, 2018. http://cqi.inf.usi.ch/qic/wheeler.pdf.

Willenborg, Leo, and Ton de Waal. 2012. *Elements of Statistical Disclosure Control*. Edited by P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg, I. Olkin, N. Wermuth, and S. Zege. Lecture Notes in Statistics, Vol. 155. New York: Springer.

Wise, Alyssa Friend. 2014. “Designing Pedagogical Interventions to Support Student Use of Learning Analytics.” In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, 203–11. LAK ’14. New York: ACM. https://doi.org/10.1145/2567574.2567588.

Wolfgang Greller, and Hendrik Drachsler. 2012. “Translating Learning into Numbers: A Generic Framework for Learning Analytics.” *Journal of Educational Technology & Society* 15 (3): 42–57.

Zaki, Mohammed J., and Wagner Meira Jr. 2014. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. New York: Cambridge University Press.