An Integrated Framework For Learning Analytics

A thesis submitted for the degree of Doctor of Philosophy

Reza Soltanpoor

Master of Science - Computer Engineering - Azad University, North Tehran Branch, Iran
Bachelor of Science - Computer Engineering - University of Tehran, Iran

School of Science,
College of Science, Engineering, and Health,

RMIT University,

October 2018

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; and, any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed. I also acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Reza Soltanpoor, October 2018
School of Science (formerly Computer Science and Information Technology)
College of Science, Engineering, and Health
RMIT University

To my significant other, the love of my life
***Azadeh** (**Azi**)*
7.9+

To my kind and supporting parents
***Shamsi** (Mom) and **Nemat** (Dad)*

Acknowledgments

I am indebted to my supervisors Professor Timos Sellis, Professor James Harland, Dr. Daryl D'Souza, and Dr. Charles Thevathayan for their continuous guidance and commitment. I was so lucky to get the benefit of having four knowledgeable supervisors. I am very grateful for Timos's ongoing support, effective advice, invaluable feedback, and for being a great human being. It was my privilege to work under his supervision to learn lessons in research and in life. Also, I would like to thank James for his brilliant feedback and impressive supervision. Being a kind and humble human being, Daryl was always there for help and I really appreciate his support in technical, ethical, and proof-reading aspects. I am thankful for Charles's great support in technical matters as well. It was Charles's approval that enabled me to apply the proposed approach in his class and collect real data. I appreciate Dr. Vu Mai's generous help in the development of the web–based application as well.

I also need to thank my friends at RMIT University who made this journey more pleasant and joyful. A special thanks to my parents for their continuous support although thousands of kilometers away. Finally, I am utterly grateful for my kind and loving wife, Azadeh (Azi) for her incredible support, enthusiasm, positive energy and patience throughout my Ph.D. journey.

Credits

Parts of the material in this thesis have previously appeared in, or have been adapted from, the publications listed below.

List of Publications

- ***Publications***
  - Soltanpoor R., Sellis T. (2016) Prescriptive Analytics for Big Data. In: Cheema M., Zhang W., Chang L. (eds) Databases Theory and Applications. ADC 2016. Lecture Notes in Computer Science, vol 9877. Springer, Cham. doi: https://doi.org/10.1007/978-3-319-46922-5_19.
  - R. Soltanpoor and A. Yavari, "CoALA: Contextualization Framework for Smart Learning Analytics," 2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW), Atlanta, GA, 2017, pp. 226-231. doi: https://doi.org/10.1109/ICDCSW.2017.58.
  - Reza Soltanpoor, Charles Thevathayan, and Daryl D'Souza. 2018. Adaptive Remediation for Novice Programmers Through Personalized Prescriptive Quizzes. In Proceedings of 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE'18). doi: https://doi.org/10.1145/3197091.3197097.
  - Reza Soltanpoor, Charles Thevathayan. 2018. Correcting Novice Programmers' Misconceptions Through Personalized Quizzes: (Abstract Only). In Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE '18), Pages 1085-1085. doi: https://doi.org/10.1145/3159450.3162266.
- ***Submissions***
  - Reza Soltanpoor, Timos Sellis. 2018. Learning Analytics: State-of-the-art Review and Framework. WIREs Data Mining And Knowledge Discovery (DMKD)[1]. **Submitted: 20 Feb. 2018**.

[1] http://wires.wiley.com/WileyCDA/WiresJournal/wisId-WIDM.html

Contents

List of Figures

List of Tables

Abstract

Low retention rates have been an ongoing concern, especially among educational institutions amidst expanding their student base and catering to large and diverse student cohorts. Increasing retention rates without lowering academic standards poses many challenges. The traditional teaching techniques using a one–size–fits–all approach appear to be less effective, and the size and diversity of cohorts demand innovative teaching techniques allowing for adaptive and personalized coaching and learning.

In this thesis, we propose a novel, adaptive and integrated analytics framework for learning analytics to address the key concerns of educational institutions. The proposed framework comprises three layers: (1) the conceptual layer which is a context–agnostic and generic analytics layer including descriptive, predictive, and prescriptive techniques; (2) the logical layer or the context–specific learning analytics processes layer that specializes the conceptual layer in the context of education; ten key learning analytics processes are formalized, implemented, and linked to the conceptual layer components; finally, (3) the physical layer that is concerned with education–oriented application implementations and is a context–specific components/algorithmic implementation of the logical layer processes. Our proposed framework, however, is not limited only to the learning and teaching environment. As a proof of concept, we chose the education context and applied our framework on it. The three–layered integrated learning analytics framework proposed allows domain–agnostic elements defined in the conceptual layer to be realized by domain–specific processes in the logical layer, and implemented through existing and new components in the physical layer. Please note that the learning analytics is not confined to the education context alone. The framework, therefore, can be customized for different domains making the approach more widely applicable.

An adaptive and innovative approach in the physical layer named the personalized prescriptive quiz (PPQ) is introduced as a demonstration of education–oriented applications assisting the educational institutions. The novel agile learning approach proposed combines

descriptive, predictive and prescriptive analytics to create a personalized iterative and incremental approach to learning. The PPQ allows students to easily analyze their current problems (especially, identifying their misconceptions), predict future results, and benefit from personalized intervention tasks. The enhanced PPQ incorporating difficulty and discrimination indexes, run–time question selection, and a hybrid iterative predictive model can be more beneficial and effective for personalized learning.

The results demonstrate a significant improvement in student academic performance after applying the PPQ approach. In addition, students claimed that the PPQ helped them elevate their self–esteem and improve student experience which may eventually lead to improved retention rates.

Chapter 1

Introduction

"The beginning is the most important part of the work."

Plato

1.1 Motivation

The prevalence of high failure and attrition rates is now a well–known problem with novice learners of programming in Computer Science and cognate disciplines [1] [Watson and Li, 2014; Watlington et al., 2010; Rumberger, 1987; Akoojee and Nkomo, 2007]. The issue might stem from learners' lower levels of abstract reasoning and problem–solving skills, primarily rooted in the misunderstanding of core concepts taught, and partly because of student cohort diversity and disparate academic backgrounds [Hmelo-Silver, 2004; Schoenfeld, 2009]. Such cohort diversity, especially in the typically large class settings, makes it challenging for instructors[2] to provide personalized attention to students about their understanding of core concepts. Core concepts refer to the foundational concepts of a discipline that are taught in the initial semesters and are necessary to build a solid knowledge framework. The ability to understand and apply core concepts has a significant impact on raising students' self–efficacy, self–esteem, and enhancing student experience. Concepts taught in later courses may rely on the previously taught core concepts. This means that a solid understanding of foundational

[1] Our main focus in this research is the context of face–to–face learning in tertiary education; however, our contribution supports learning in general.

[2] "Instructor" is the most generic term used throughout the thesis. "Teacher" is an alternative term; however, to preserve the consistency, "instructor" is used.

core concepts will help reduce the cognitive gaps of the students as they progress through their following courses in the curriculum. This way, the students' academic success can be guaranteed through more knowledgeable and skillful graduates. Employers also prefer such graduates as they can avoid retraining them in core skills and problem–solving.

Lack of a deep understanding of fundamental concepts leads to a weak, fragile or inadequate knowledge framework, resulting in imparting of unreliable cognitive skills, lower self–esteem, degradation of academic capabilities, and, more importantly, accretion of failure and withdrawal/drop–out rates [Trigwell and Prosser, 1991]. This ultimately compromises the higher educations' goal of high student retention rates. Thus, incorporation of effective approaches in addressing students' misconceptions, and making sure they have properly understood core concepts, is critical for both instructors and educational institutions: instructors, because they play a major role in supporting and maximizing the quality of students' learning experience, and educational institutions, because they need to improve retention rates.

To alleviate the conceptual misunderstanding of core concepts, instructors and educational institutions need to periodically and frequently monitor students' knowledge level of the core concepts taught, especially during the seminal stages of their learning. To address this, several pedagogical approaches have been introduced, including [Best and Kahn, 2016; Cleveland et al., 2018; Michael, 2006; Alters and Nelson, 2002]: (1) Teaching core concepts earlier in the semester. Teaching the fundamental concepts at the start of a semester gives sufficient amount of time to the students to absorb concepts and effectively build their academic knowledge framework more reliably. This way, instructors can promote effective learning. (2) Performing several kinds of assessment [Wiliam and Thompson, 2017; Cleveland et al., 2018]. A number of traditional and state–of–the–art assessment approaches has been enforced to continuously monitor students' progress and perform effective interventions when necessary. To name a few, weekly quizzes, in–class tests, weekly and in–class online assessments, assignments, mid–term tests, and final exams are examples of different assessment approaches. On way to help students improve their help–seeking and problem solving skills based on their previous performance is by deploying the ITS [Hooshyar et al., 2016b;a; Roll et al., 2011; Corbett et al., 1997]. The intelligent tutoring systems (ITS)[3] were introduced to target cohorts of students and provide them with individualized instructions based on their past

[3]Intelligent tutoring systems (ITS) provide individualized instructions tailored for specific students, the key feature being the support of cognitive diagnoses and adaptive remediation. They can be made to supplement traditional teaching and have been shown to outperform traditional teaching for specific cohorts [Ma et al., 2014].

performance. Most recently and with the emergence of personalized learning, institutions of higher education have become more interested in utilizing adaptive and personalized assessment techniques dealing with individual students. (3) Providing a wide range of feedback. Different types of feedback have been constructed to convey the teaching teams' comments regarding students' responses to the performed assessments [Hattie and Timperley, 2007]. Summative and formative feedback mechanisms are the main categories in this area. In the former, little or no feedback is provided to the students. Most of the feedback is given very late (at the end of the semester and after the final exam). In the summative assessment, students' final marks will be published without any reference to students' difficulties. The formative assessment is focused on the fine–tuned instructions for students covering the areas of attainment and promotes active and adaptive learning by providing relevant guidance for improvement in the areas of weakness. The feedback in the latter case is propagated to the targeted student cohorts to improve their grasp of knowledge and positive engagement to the developed assessments. The formative feedback can be conveyed to the students in timely (delayed) or instant (real–time) manners. The recent attention towards personalized learning environments (PLEs) has demonstrated the adoption of more fine–grained feedback mechanisms by focusing on each individual student instead of student cohorts.

Recent advancements in technology–enhanced learning (TEL) field has resulted in the widespread utilization of online learning materials. Consequently, educational institutions are swamped with large amounts of pedagogical data that is generated, exchanged, and sometimes streamed (lecture recordings) via e–learning systems, such as the learning management system (LMS)[4,5]. Typically, such data is held in LMSs and is accessible to instructors, either directly or via analytics tools that provide value–added information. Therefore, robust analytical approaches are the mainstay of educational institutions with vast amounts of educational data collected [Siemens, 2013]. Several types of educational data are generated/collected when students interact with the LMS, such as the history of log-ins, profile and demographic information [6], student credentials, academic records, course discussions and forum posts, multimedia and text submissions of online assessments. Providing analytical solutions makes it necessary to collect a variety of student data embedded within the LMS, cleanse the data (removing redundant or noisy data), reduce its size (based on the important

[4]Learning management systems are "software applications that automate the administration, tracking, reporting, and delivery of the educational resources" [Ellis, 2009].

[5]http://web.csulb.edu/~arezaei/ETEC551/web/LMS_fieldguide_20091.pdf

[6]Students' profile and demographic information are usually collected upon enrollment and stored in the student information systems (SIS).

attributes defined by instructors), unify the data into one standard format to be used within the application, and generate multiple analytical reports such as enrolment history, students' performance during a certain semester and, overall, performing statistical analyses over the students' past data, and categorizing students based on their academic performance.

Analytical approaches focusing on students' past interactions with the LMS to produce analytical reports are usually denoted *descriptive* and *diagnostic* analytics. *Descriptive analytics* mainly reports the past by helping the educational institutions and the instructors to understand what has happened until now. The *diagnostic analytics*, on the other hand, provides the means to extract possible reasons behind past events using certain statistical methods. Therefore, using descriptive and diagnostic techniques on the wealth of historical assessment data is imperative towards understanding various student cohorts' performances. Educational institutions are also interested in making informed decisions through predicting future behavior and choosing appropriate intervention strategies. *Predictive analytics* is a kind of analytics that focuses on projecting student performance trends and notifying the instructors of the likely at–risk students, given students' past records and utilizing particular machine learning techniques [Heffernan and Heffernan, 2014]. As a result, relevant interventions should be applied by the instructors and/or educational institutions in a timely manner. To assist the teaching team to disseminate effective and actionable feedback to targeted student cohorts, specific analytical approaches with the focus on generating adaptive courses of actions should also be adopted. *Prescriptive analytics* was emerged to address this need by producing targeted recommendations and courses of action(s) based on students' past performance and future predictions. Please note as per the above elaboration, the descriptive, diagnostic, and predictive analytics are serving different purposes in responding to analytical needs and thus distinctly explained [Bertsimas and Kallus, 2014; Turban et al., 2013; Soltanpoor and Sellis, 2016; Larson and Chang, 2016].

Given the need for adopting efficient assessment and feedback mechanisms to facilitate active and adaptive learning, researchers have become more interested in introducing several techniques to promote various analytics by identifying and rectifying misconceptions, and at the same time, building a solid and robust knowledge framework. Learning analytics (LA) is a growing area of technology–enhanced learning, which has emerged to address the analytical needs of educational institutions. It takes into account several methods and techniques to collect educational data from disparate sources, unify and analyze the collected data to generate required analytical reports of students' pedagogical activities, project likely trends in students' future behavior and academic performance, and provide means of personalized

learning experiences for students and instructors. Some studies have advocated the blending of learning analytics methods with intelligent tutoring systems (ITS).

This thesis proposes an integrated and layered learning analytics framework to address the learning needs of educational institutions. Our framework incorporates several analytical techniques such as learning analytics, intelligent tutoring systems, and personalized and prescriptive approaches. In the next section, we elaborate on the research questions that are addressed in this thesis, followed by the associated contributions.

Prior to the research question, we will review a motivation scenario in the context of education in Section 1.2.

1.2 Motivation Scenario - Recommending personalized learning material to each individual student

Let's consider the case that during a certain semester, the instructor wants to provide each student with individually selected learning material based on their past performance in prior assessments. We can simply scatter the process in 10 steps as follows:

1. *Question tagging* — tagging each assessment question with their corresponding taught concept(s). This means that when designing each question, the instructor should tag the question with the concept(s) they cover.

2. *Learning material tagging* — tagging each learning material with their relevant concept(s). Similarly, the instructor will tag each learning resources with the concept(s) they cover. In terms of the tagging granularity, the learning material can be tagged in either coarse- or fine–grained manner. In the former case, the whole textbook (the online resource) or chapters of them are tagged, while the latter approach is more concerned with providing the tag for each page, paragraph, or even certain lines of the material.

3. *Previous assessment(s) data collection* — collecting each student's past assessment results. The performance of the students in the tests and assessments since the start of the semester will be recorded. The past assessment results may be in diverse forms and kinds of data (tabular, graphical, text).

4. *Misconception extraction* — extracting the set of misunderstood concepts per student. If the student responded incorrectly to the questions covering particular concepts, those

concepts are considered as the student's misconception. The misconception set will be calculated and constructed for each student based on their past responses to the assessment questions. We assume that all assessment questions were tagged with their corresponding concept(s) from step 1 [Liu et al., 2016].

5. *Performance prediction* — extrapolating each student's likely performance/status towards the end of the semester. For each student, the system can project the pass/fail and likely final mark by the end of the semester.

6. *Adaptive and personalized assessment* — selecting the next assessment questions for each student based on their calculated misconception(s), individually. A personalized set of question(s) for individual students will be generated to cover their misunderstood concept(s). The number and types of questions for one student may be different than others.

7. *Targeted learning resource(s) dissemination* — adaptively recommending learning material to individual students, given their misconception(s) or during their assessment. The system can generate instant, personalized, and adaptive formative feedback to each student based on their incorrect responses to the assessment questions. This means that each student will get a list of recommended learning resources, individually calculated for them, covering the concepts they misunderstood (from their previous assessment results, or during their ongoing assessment).

8. *Notification mechanism* — notifying the instructor and the student of the student's performance during or after each assessment. Having the knowledge of the student's performance, the instructor can plan to perform further interventions if required. The student, on the other hand, can be directed towards further learning material suitable for their status so that they can rectify their misconceptions properly.

9. *Analytical report generation* — generating analytical reports on each student's performance during the semester for the student, the instructor, and the educational institution. The student–targeted report will simply incorporate their current in–semester academic records (assessment history) as well as their current set of misconceptions and the suggested learning material. The instructor–targeted report will include the student–targeted information, their predicted status/performance by the end of the semester, and their performance compared to other students in the class. The institution will also have an analytical report on the student's overall performance status,

their projected status towards the end of the semester, and the whole class performance (current and predicted); an integrated report on all courses and collected surveys from the students and the instructors to evaluate/calculate the student retention rate and the student experience metrics will also be available.

10. *Ongoing follow–up mechanism* — continuous following up on each student's progress during a certain semester or throughout their curriculum. This final step ascertains the benefits of constant formative feedback and adaptive personalized learning material recommendation to individual students and the educational institution. The goal is to improve the learning outcomes in skills required by the academia or the industry, higher levels of self–esteem, and positive student experience. The educational institutions, on the other hand, can aim for a certain student retention rate which in turn satisfies their pedagogical objectives.

Table 1.1 demonstrates the mapping of the mentioned sample scenario's 10–step processes with their corresponding components within the proposed layered framework. Please note that each one of these layers is elaborated in detail in the upcoming chapters.

1.3 Research Questions

There has been a large body of research focusing on the analytical requirements of educational institutions (as mentioned earlier in Section 1.1); however, to the best of our knowledge, there is no extant research on proposing a holistic analytics approach [7] to address the majority of educational institutions' needs. This motivated us to propose a layered, integrated analytics framework which takes into consideration key educational institutions' concerns and provides them with adaptive and actionable solutions. Although our research is capable of being applied to the whole education landscape, we focus on its "tertiary education" subset[8]. The proposed framework is comprised of three different yet related layers similar to the prominent technique in data modeling [Date, 2006]:

1. *The Conceptual Layer* — concerning with the overall and high–level analytical techniques.

[7] By "holistic analytics approach", we meant solutions addressing most of the educational institutions' requirements (kind of one–size–fits–all approach).

[8] Other subsets of the education landscape are "pre-school", "primary", "secondary", as well as the "tertiary" education.

Table 1.1: Mapping Of The Sample Scenario's 10–Step Processes With Their Corresponding Analytical Components In The Proposed Framework.

Steps	**Layers Of The Proposed Framework**		
	Physical	**Logical (LA[1] Processes)**	**Conceptual**
(1) Question Tagging	data annotation (Questions)	-	-
(2) Learning Material Tagging	data annotation (Learning Material)	-	-
(3) Assessment Data Collection	data collection, cleaning, integration, reduction, augmentation, unification	monitoring process	descriptive analytics
(4) Misconception Extraction	data mining, information retrieval techniques	analysis process	descriptive analytics
(5) Performance Prediction	machine learning algorithms	prediction process	predictive analytics
(6) Adaptive Assessment Generation	data processing	assessment, personalization processes	prescriptive analytics
(7) Feedback	data processing, recommendation engine	personalization, feedback, adaptation, intervention processes	prescriptive analytics
(8) Notification	report generation, feedback mechanisms	feedback process	feedback and deliverables
(9) Report Generation	report generation	feedback, intervention processes	prescriptive analytics
(10) Follow–up	data collection, integration, reduction, unification	monitoring process	descriptive analytics

[1] Learning Analytics

2. *The Logical Layer* — focusing on the specializations of the conceptual layer which in our case is translated into the analytics in the context of education (learning analytics).

3. *The Physical Layer* — that takes into consideration the formalization and implementation of the conceptual and logical layers in real–world applications.

In the proposed framework, the conceptual layer is domain–agnostic and acts as a generic solution to most of the analytical scenarios. The logical layer is domain–specific and might vary based on different analytical contexts that the system is specialized for. Finally, the physical layer is flexible in allowing for new requirements within the system, and changes in the components/algorithms.

The main implications and properties of our proposal can be listed as follows:

- *adaptive and automatic remediation of students' misconceptions.*
- *providing and supporting targeted intervention(s).*
- *promoting personalized assessment and feedback mechanisms.*
- *dynamic learning difficulty identification.*
- *supporting adaptive and personalized learning.*
- *fostering high–level cognitive skills*

Let's provide an example of how the proposed layered analytical framework can help educational institutions in achieving their pedagogical objectives in the following section.

To design our proposed integrated analytics–driven framework for the context of education, the following research questions were developed in this thesis.

A federated composite analytics architecture (a prescriptive and not a software architecture) comprising key analytical methods (descriptive, predictive, and prescriptive) is introduced. This architecture is not limited to the domain of education and can be applied to applications of other domains. This is the first step in constructing our analytics framework. Therefore, the first research question is shaped as follows.

Research Question 1)

How do we design an integrated and adaptive analytics architecture?

We propose a novel approach in organizing analytical components with the ability to accept diverse data types and producing dynamic feedback. Descriptive, predictive and prescriptive analytics approaches are incorporated within the architecture to make it serve as a generic analytical model. The introduced analytics architecture is a generic and domain–independent solution to analytical needs of enterprises as mentioned in Chapter 4. Its context–agnostic nature makes it flexible to be applied to a wide range of analytical application scenarios. The details of the architecture and some of its applications are elaborated in Chapter 4. The proposed architecture shapes the conceptual layer of our analytics–driven framework in Chapter 5.

Since the proposed framework is tailored to the field of learning analytics, the next research question is focused on specializing the general composite analytics architecture in the context of education.

Research Question 2)

How do we incorporate the proposed integrated analytics architecture in the context of learning analytics (proposing the analytics framework for learning analytics)?

Learning analytics as the major technology–enhanced learning field is discussed in Chapter 3 along with a learning analytics reference model operating in four dimensions. Moreover, ten key learning analytics processes are extracted based on the main LA needs. Finally, each of the LA processes is implemented in the business process model and notation (BPMN) specification[9]. This step shapes the logical layer of the proposed framework in Chapter 5.

To connect the conceptual and logical layers of the framework together and to justify the introduced analytics architecture, all logical components (here, all ten learning analytics processes implemented in BPMN) must be linked to their corresponding conceptual layer elements. The next research question, therefore, is concerned with the interrelationship between the conceptual and logical layers.

Research Question 3)

How do we formalize learning analytics processes in the proposed framework (connecting learning analytics and prescriptive analytics components)?

[9]http://www.bpmn.org/

The conceptual and logical layers of the proposed analytics–driven framework are connected to address this research question. This step is further elaborated in Chapter 5.

All learning analytics processes in the logical layer are formalized and implemented (in terms of algorithms and coding) to form the physical layer in Chapter 5 to assess the degree to which our proposed approach can address the major concerns of students, instructors, and educational institutions. We also introduce a new approach called the *personalized prescriptive quiz (PPQ)* to provide students with individual questions covering their past misconceptions (concepts they missed or did not understand properly in their previous assessment within a certain semester). Thus, the final research question is developed as follows.

Research Question 4)

How do we devise and link the physical layer components enforcing higher–level processes (linking the physical, logical and conceptual layers altogether)?

The proposed framework takes into consideration real students data collected from their interactions (multiple assessments and their assessment performance history) and is evaluated using well–known qualitative and quantitative techniques. The details and acquired results of the PPQ approach are discussed in Chapter 6. Furthermore, the PPQ components are connected to their corresponding logical and conceptual layers, to make the proposed framework complete. Also, all physical layer components are linked to their corresponding higher level components in the logical and conceptual layers.

To address the mentioned research questions, the next section reviews the major contributions.

1.3.1 Major Contributions

The main contributions of this thesis can be listed as follows:

1. *An adaptive and federated composite analytics architecture incorporating descriptive, predictive, and prescriptive analytics approaches with dynamic feedback mechanisms.* The proposed architecture in Section 4.3, uniquely combines descriptive, predictive, and prescriptive analytics and links them together. The architecture also aims for adaptive and timely generation of courses of actions with the help of certain feedback lines designed within the system. We also introduce a composite design for the prescriptive component which comprises the simulation, optimization, and evaluation parts. This contribution addresses the first research question mention in Section 1.3.

2. *A four–dimensional learning analytics reference model covering key learning analytics processes.* A four–dimensional model is introduced in Section 3.3 that comprising (1) collecting several educational data types in the WHAT dimension, (2) taking into account the main stakeholders (students, instructors, and educational institutions) in the WHO dimension, (3) deployment of certain analytics techniques to analyze collected educational data in the HOW dimension, and (4) capturing and formalizing the top 10 key learning analytics requirements and processes in the WHY dimension. This contribution shapes the ground for the second and third research questions mentioned in Section 1.3.

3. *An integrated analytics–driven framework for learning analytics comprising three layers (conceptual, logical, and physical).* The proposed framework is introduced in Section 5.2. The conceptual layer is generic and domain-agnostic analytics layer. The composite analytics architecture of Section 4.3 constructs the conceptual layer of the framework. The logical layer, on the other hand, is domain-specific and is specialized for the context of education. The 10 learning analytics processes are formalized within the logical layer and linked to corresponding conceptual layer components (linking the logical and conceptual layers). The logical layer is elaborated in 5.4. Finally, the physical layer is designed to be application-specific. It is mainly focused on the implementation of educational applications and their connection to learning analytics processes in the logical layer. This contribution targets the second and third research questions mentioned in Section 1.3.

4. *A novel personalized learning approach implemented in the physical layer called the personalized prescriptive quiz.* The personalized prescriptive quiz (PPQ) approach is introduced in Section 6.3. The PPQ fills key gaps in traditional assessment approaches by providing a form of personalized coaching to students. It helps each student to individually identify and rectify their misconceptions by providing them with individually designed sets of questions covering their misunderstood concepts. The results demonstrate a significant improvement in student academic performance after applying the PPQ approach. Instructors can design more efficient questions covering taught concepts, by taking into consideration student feedback gathered on PPQ performance. By linking the physical layer's components to their corresponding logical and conceptual layers, the integrated analytics framework proposed in Section 4.3 is finalized. This contribution addresses the final research question of Section 1.3 in building the physical

layer of the proposed framework.

To sum up, as a proof of concept, please refer to Figure 7.1 depicted in Chapter 7 in linking all layers together as a whole analytics solution for the education context.

1.4 Thesis Organization

The road–map of this thesis is as follows.

- *Chapter 2) Background* — this chapter is focused on the extensive literature review regarding several analytical methods including descriptive, predictive and prescriptive methods. The chapter continues with reviewing the extant body of research of different analytical approaches in the context of education such as educational data mining, academic analytics, and learning analytics.

- *Chapter 3) Learning Analytics* — this chapter is mainly concerned with the technology and research advancements in the field of learning analytics as a growing area of technology–enhanced learning. After a short introduction to the field and review of the current body of literature in the area, the chapter elaborates on different analytical techniques of the higher education (i.e. educational data mining (EDM), academic analytics (AA), and LA). The chapter then reviews the key requirements every LA solution should support along with recent proposed LA models (LAMs), tools and applications. A 4–dimensional learning analytics reference model covering the mentioned LA requirements is proposed based on the work by [Chatti et al., 2012]. The chapter finally enumerates the field's challenges and future directions and concludes with referring to the proposed analytics framework in Chapter 5 as a solution to address most of LA–related concerns based on the introduced 4–dimensional LA reference model (LARM).

- *Chapter 4) Prescriptive Analytics* — the first research question is addressed in this chapter by proposing a federated composite analytics architecture. At first, a literature review of the current state–of–the–art research in the field is provided. Prescriptive analytics relation to other established analytical approaches like descriptive, diagnostic and predictive analytics is discussed next. Then, three different application scenarios are investigated to emphasize the importance of prescriptive analytics in different sectors of industry and academia. Finally, an adaptive and integrated composite analytics

architecture is proposed to address the research gaps in the field. This architecture is the basis for the analytics–driven framework which is elaborated in Chapter 5.

- *Chapter 5) Analytics–Driven Framework for Learning Analytics* — an analytics–driven framework for the context of education is proposed incorporating the composite analytics architecture in Chapter 4 and the 4–dimensional learning analytics reference model in Chapter 3. The chapter aims to address research questions 2 and 3 with framing a three–layered analytical framework (i.e. the conceptual, the logical, and the physical layers). The conceptual layers' components and their interrelationships are discussed (with the focus on the generic analytics–driven and prescriptive analytics modules). The logical layer is also elaborated in this chapter with the focus on illustrating 10 key learning analytics processes (introduced in Chapter 3) in the business process model and notation (BPMN) specification. Next, each LA process component is connected to their corresponding conceptual layer elements to build the first two layers of the framework. The logical layer covers the third research question in formalizing the LA processes in the proposed framework. Finally, the physical layer is elaborated in Chapter 6 where the framework is applied to one educational application scenario with real student data. Overall, the proposed framework covers the second research question.

- *Chapter 6) Personalized Prescriptive Quiz (PPQ)* — to address the forth research question, an adaptive and personalized approach (the personalized prescriptive quiz (PPQ)) is proposed in this chapter that constructs the physical layer of the framework. Given that the PPQ was applied in several semesters and being used by different core courses, a wealth of real educational data was captured. In the results section, the qualitative and quantitative findings are analyzed. Finally, the future directions of the research and the potential extension points to the PPQ approach are mentioned.

- *Chapter 7) Conclusions and Future Work* — this chapter is mainly focused on revisiting the thesis outcomes to assess the extent to which they addressed the research questions. The chapter concludes by mentioning the future research directions.

Chapter 2

Background

> "Research is to see what everybody else has seen, and to think what nobody else has thought."
>
> *Albert Szent–Gyorgyi*

An extensive literature review in data analytics is provided in this chapter. The general analytics techniques such as descriptive, diagnostics, predictive, and prescriptive analytics are reviewed in Section 2.1. Focusing on the context of education, Section 2.2 reviews the body of research in learning analytics, educational data mining, and academic analytics. Finally, Section 2.3 is concerned with the extant research on analytics frameworks in education.

2.1 Data Analytics and Analytical Techniques

Big business organizations or academic institutions possessing big data are interested to adopt proper analytical solutions to transform data into information and then into insights and process them to elicit business values and act upon them to maximize their objectives [Baker and Gourley, 2014]. Big enterprises need to know what has happened in the past, what is happening now, what is likely to happen in the future, and what are the optimal sequences of actions they can take to satisfy their goals [Kaisler et al., 2014]. They need to extract insights from the wealth of data they own to take advantage of future opportunities and mitigate likely risks [Davenport and Dyché, 2013]. This need translates into operational, adaptive and optimal courses of actions (in the form of some recommendations) based on

the extracted insights [Barga et al., 2014]. Prescriptive analytics has emerged as the new business analytics field to address the mentioned gap and assist the enterprises to meet their objectives. According to the body of research, business analytics is classified into three major categories [Delen, 2014], [Sharda et al., 2013], [Banerjee et al., 2013]: descriptive, predictive, and prescriptive analytics that are elaborated in the following sections.

2.1.1 Descriptive Analytics

"*Descriptive Analytics*" is often called the "data summarization" or the "data reduction" process which is focused on the past (reports the past). It answers the question "What did happen?" and extracts valuable information given the collected data elements from diverse sources [Delen and Demirkan, 2013]. Several analytical reports and the unified and aggregated forms of data (to be utilized by other analytical approaches) are among the key outcomes of this analytical approach. With regard to analyzing the past events, another analytical technique has emerged as an extension to descriptive analytics which is called the "diagnostic analytics". Like descriptive analytics, the diagnostic analytics also reports the past, but it aims at answering questions such as "Why did it happen?". It helps business enterprises to grasp the reasons and causes of the events happened in the past. Diagnostic analytics gives the organizations the knowledge to understand relationships among different kinds of data [Karim et al., 2016; Banerjee et al., 2013].

2.1.2 Predictive Analytics

"*Predictive Analytics*" is also called the "forecasting" process and takes the unified data elements generated by descriptive analytics, and builds accurate predictive models by incorporating proper machine learning techniques [Siegel, 2016; Waller and Fawcett, 2013; Hazen et al., 2014]. By utilizing these models, enterprises can detect future opportunities and risks and plan to face them accordingly. It answers the questions "What will happen?" and "Why will it happen?" in the future [Delen and Demirkan, 2013; Eckerson, 2007; Shmueli and Koppius, 2011]. One key challenge is to feed as much data as possible because more data means more accurate models and predictions. Some well–known techniques in predictive analytics are data mining, text/web/media mining, and forecasting approaches [Shmueli and Koppius, 2011; Waller and Fawcett, 2013]. Predictive analytics produces several future extrapolations along with their corresponding probability scores.

2.1.3 Prescriptive Analytics

"*Prescriptive Analytics*" is called the "recommender or guidance" process which provides business organizations with adaptive, automated, time–dependent, and optimal decisions [Basu, 2013; Adomavicius and Tuzhilin, 2005]. It is mainly focused on bringing the business value through better strategic and operational decisions through relevant recommendations (courses of actions). Generally, prescriptive analytics is one predictive analytics which is expanded to prescribe sequences of actions and illustrate the likely outcome of each action along with their mutual influence. It answers the questions "What should I do?" and "Why should I do it?" using the built–in "what–if" scenarios [Haas et al., 2011]. Core components of a given prescriptive analytics solution are optimization [Liberatore and Luo, 2011; Schniederjans et al., 2014], simulation, and evaluation methods [Bertsimas and Kallus, 2014]. Prescriptive analytics takes predictive analytics outcome into consideration along with the enterprise's business constraints, compliance rules and objectives to generate the optimal courses of actions. It means that prescriptive analytics takes an actionable predictive model into account and generates optimal recommendations to help organizations with their informed decision making processes [Marathe et al., 2014; Apte, 2010]. Prescriptive analytics solutions usually express two major characteristics [Basu, 2013]:

1. *Providing the business organizations with the optimal and actionable recommendations in terms of comprehensible prescriptions*, and
2. *Supporting feedback mechanisms to keep track of the recommendations' effectiveness and occurrence of unprecedented events throughout the system's life–cycle.*

Prescriptive analytics is capable of being applied in a wide variety of use cases and real–world application scenarios, some of which are listed as follows:

- *Transportation* — prescriptive analytics helps active companies in the field to manage transportation capacities, recommend different ticket prices at different rates during times of increased or decreased demand to maximize capacity and profitability (intentional price fluctuations), and so forth.
- *Oil and gas industry* — prescriptive analytics can help the oil and gas companies to locate and produce oil and gas in a cost–effective manner, recommend where and how to drill to maximize production and minimize cost as well as environmental impacts based on the collected data from past drilling processes and production history.

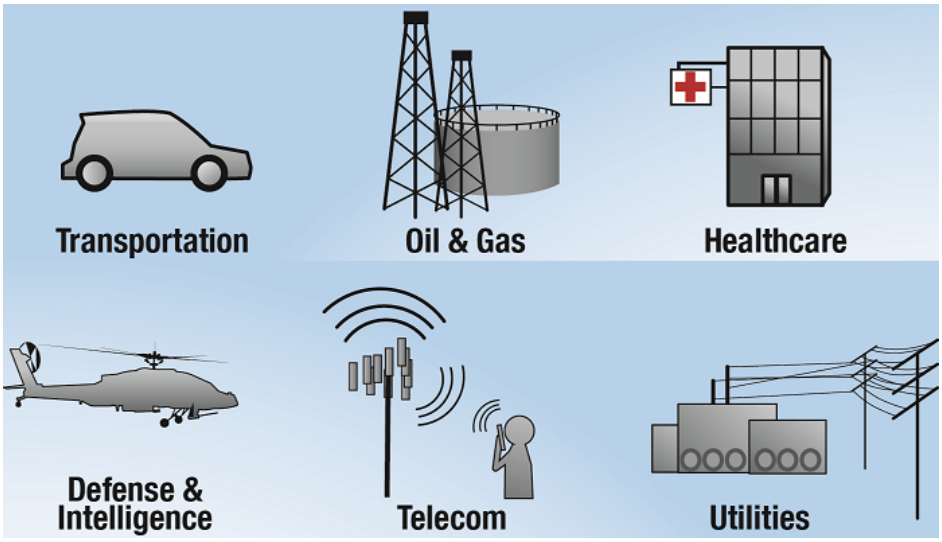


Figure 2.1: Several Prescriptive Analytics Application Varieties

- *Logistics* — prescriptive analytics can help the organizations to reach the most optimal route for their deliveries and freights that can be updated in near real–time (adaptive), given past performances, load information, current conditions, and vehicle specifications.

- *Healthcare* — healthcare providers can utilize prescriptive analytics solutions to leverage operational, demographic, economic and health trends, to plan for investments in new facilities and equipment. Examples in hospitals and health–care related organizations can be providing optimal resource allocation solutions (such as the arrangement of beds in wards, allocation of health professionals to the designated locations, timely and optimal orderings of medical equipment in–line with the predicted future requirements), and so on.

- *Price optimization, inventory management, supply chain optimization, and resource allocation* — prescriptive analytics approaches can be used to make specific offers and modifications to subscriptions or purchasing plans based on the nature and progression of a customer's interaction with the current system.

Figure 2.1 depicts some of the main real–world application scenarios capable of adopting prescriptive analytics solutions.

Figure 2.2 depicts the main business analytics approaches (descriptive, predictive and

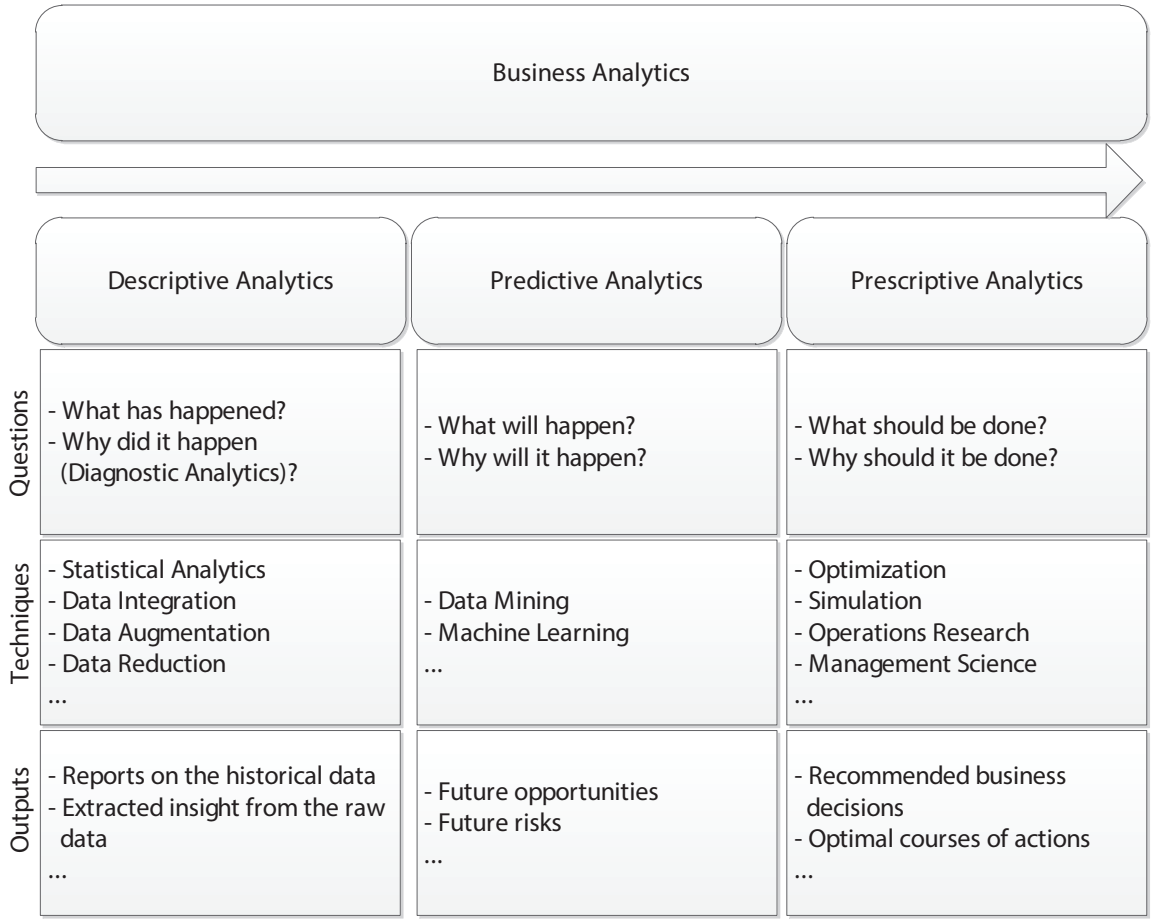


Figure 2.2: Business Analytics Stages

prescriptive analytics) in terms of the questions they answer, the set of techniques they incorporate, and the result(s) of each analytical processes.

Figure 2.3, in addition, illustrates different analytical approaches' spectrum in a successful business analytics value escalator in terms of the provided value of each analytical approach as well as their adoption difficulty levels (according to the Gartner's report on 2017[1]). Starting from left, descriptive analytics tries to answer the question "what happened?" which provides less value to enterprises by generating analytical reports regarding the past. The information is the main player in this stage which means the approaches to transform the collected data into the desired information with the help of statistical methods. This part

[1]https://www.gartner.com/doc/3698935/forecast-snapshot-prescriptive-analytics-worldwide

is called the *"hindsight"* and is the easiest analytical stage to deal with. Moving forward, another analytical approach emerges which is more concerned with the rationale behind the events happened in the past and tries to understand why something has happened (answering the question "why did it happen?"). It is called diagnostic analytics and provides the organizations with in–depth analytical reports to help them understand the relationships among past events and extracting the reasons those events happened. Diagnostic analytics contributes to the inception of the *"insight"* phase in analytics and brings medium–level business value and is still moderate to deal with (in terms of implementation and adoption difficulty within the enterprise). Next, predictive analytics emerges that is mainly focused on the future and tries to complete the chain of the *"insight"* phase started by the diagnostic analytics by adopting proper predictive models and answering questions like "what will happen and why?". Providing the likely future opportunities and risks along with their probabilistic trends is of crucial importance to any organization. Therefore, predictive analytics incorporation provides a considerable amount of business value to enterprises and its adoption process is considered difficult. Finally, prescriptive analytics plays the ultimate role in providing enterprises with optimal and adaptive courses of action(s) to help them make informed decisions and meet their business objectives. It answers the question "how can we make it happen and why?" and has a significant value within the organizations and its adoption difficulty is considered the hardest within the enterprises.

According to the literature, the following lists some of the main research gaps in the area of prescriptive analytics:

- *A concrete definition of prescriptive analytics* — according to the body of research, there is a wide range of definitions associated with the term "prescriptive analytics": a recommender system, an optimization engine, a simulator, etc. We will propose a holistic definition of prescriptive analytics which entails all mentioned analytical components.

- *An extant and valid federated analytics architecture* — to the best of our knowledge, the literature lacks the design and implementation of an integrated analytics architecture incorporating descriptive, predictive and prescriptive components. Furthermore, the relationships among different components should be elaborated clearly. We will address this issue as well.

[2]https://www.smartdatacollective.com/how-risk-management-ecosystem-evolving-data-analytics/

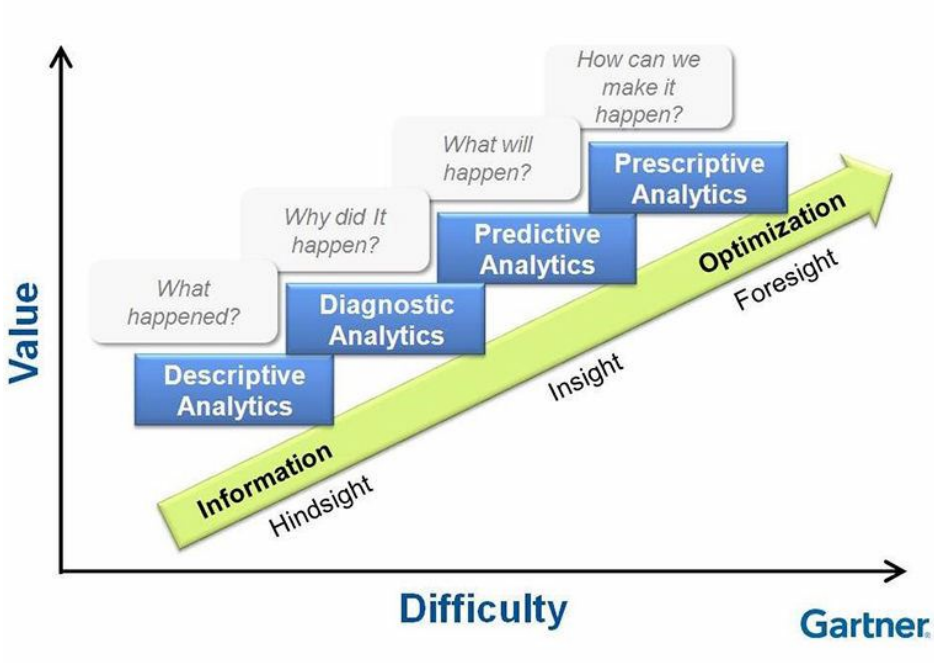


Figure 2.3: Business Analytics Value Escalator (Different Analytics Spectrum) – Gartner's Report[2]

- *An adaptive prescriptive technique* — most of the current proposed prescriptive approaches are incapable of change in terms of the support for back–propagation feedback mechanisms within the system. We will cover this gap by providing and supporting backward feedback lines throughout different analytical components of the proposed architecture.

Given the above–mentioned definitions of descriptive, predictive and prescriptive analytics, a successful analytical solution for current business enterprises seems to be the one incorporating the integrated benefits of all the three approaches. To the best of our knowledge and based on the extant literature, there is no specific study taking into consideration all the mentioned analytical approaches in one framework . This was our main motivation to propose a novel integrated prescriptive analytics to address the mentioned requirements in Section 4.3. Our proposed architecture is capable of handling heterogeneous data from diverse sources, building accurate predictive models based on the acquired data, and generating the adaptive and optimal sequences of actions to help the decision making process

[2]Given that we are discussing the analytics approaches in general, our focus in on any domain (and not just the learning context).

in near real–time with operational recommendations. But prior to elaborating on our approach, some critical application scenarios are reviewed in Section 4.2 to demonstrate the crucial importance of adopting a prescriptive solution in addressing their requirements.

2.2 Analytics in Education

Analytics, in general, is a multi–disciplinary concept and is concerned with data acquisition from diverse sources, performing analytical calculations on the collected data items to extract useful patterns and valuable insights, and distributing the computed outcomes to the corresponding targets [Chen et al., 2012; Power, 2014; Van Barneveld et al., 2012]. Approaches incorporating different data analytics methods have contiguously been evolving over the last decade. An enormous amount of educational data has been produced due to the technological advancements in the digital education, the increasing tendency regarding the generation, sharing and dissemination, and online learning resources utilization (both in traditional on–campus and online educational approaches). Also, learners and instructors' interactions with the learning management systems (LMS) have caused the exponential growth in the volumes of educational data lately [Baer and Norris, 2015; Ferguson et al., 2016; Adams Becker et al., 2017]. Therefore, institutions of higher education need proper analytical tools to process those data elements to improve student experience, increase learners' retention rates, help them pick the optimum learning pathways in accordance to students' objectives, aptitudes and academic records, and provide the institutions of higher education with relevant analytical reports to assist them make strategic and informed decisions [Chen et al., 2012; Van Barneveld et al., 2012; Siemens and Long, 2011; Adams Becker et al., 2017; Siemens et al., 2011]. Given that, educational institutions need to incorporate relevant analytical techniques to process their wealth of educational data [Ferguson, 2012b;a; Freitas et al., 2015].

2.2.1 Learning Analytics

Learning analytics, as a key area of research in the context of education and technology–enhanced learning, has attracted a huge attention recently. A huge amount of research has been conducted in learning analytics definition, its requirements identification, LA proposed models, and its developed tools and applications [Peña-Ayala et al., 2017; Peña-Ayala, 2017; 2018; Baker and Inventado, 2014; Siemens and Long, 2011; Siemens and d Baker, 2012; Ferguson, 2012a; Ihantola et al., 2015; Arnold and Pistilli, 2012; Siemens, 2013; Chatti et al.,

2012].

Learning analytics has emerged to effectively help institutions of higher education to perform desired data analyses over the wealth of students' data and produce actionable outcomes to satisfy their pedagogical objectives. Learning analytics is mostly focused on techniques to collect, unify, aggregate and process educational data from diverse sources (which is called the insight part), predict future trends based on those data elements with the help of proper machine learning techniques (which is called the foresight part), and act upon the produced outcomes to improve the process of learning [Peña-Ayala, 2018; Martin and Sherin, 2013; Pea and Jacks, 2014; Elias, 2011; Ferguson, 2012a; Siemens et al., 2011]. LA has also common interests with other technology–enhanced learning research areas such as educational data mining, academic analytics, personalized and adaptive learning, recommender systems, and action research [Chatti et al., 2014]. The most established definition for LA according to the first international conference on learning analytics and knowledge 2011 (LAK'11)[3] and adopted by the society for learning analytics research (SoLAR)[4] is:

> "The measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and environments in which it occurs" [Siemens and Long, 2011].

According to the Horizon Report [Jaramillo, 2017], learning analytics is currently considered a key trend in the future of teaching and learning in higher education.

E–learning systems – aka as the virtual learning environments (VLE) – are web–based solutions and applications that provide the course resources and materials online. Recently, the increasing tendency in implementing several types of VLEs has drawn further attention towards the adoption of learning analytics techniques [Šumak et al., 2010; Van Raaij and Schepers, 2008]. Learning management systems and courseware management systems (CMS) are among the well–known examples of VLE implementations [Oliveira et al., 2016; Coates et al., 2005; Gibbons, 2005; Romero et al., 2008]. The state–of–the–art VLEs help learners build their own learning pathways by personalizing the published resources and in turn assist them to develop their personalized learning environments (PLE) [Van Raaij and Schepers, 2008; Chatti et al., 2014]. Institutions of higher education are capable of getting benefit from the power of learning analytics and virtual learning environments by taking the emerging concepts of blended and computer–supported collaborative learning (CSCL) into account

[3] https://tekri.athabascau.ca/analytics/
[4] https://solaresearch.org/

[Garrison and Kanuka, 2004; Stahl et al., 2006; O'Malley, 2012].

A large amount of heterogeneous data is generated from learners' interactions with educational systems (including learning management systems) and even the social media [Bienkowski et al., 2012; Shum and Ferguson, 2012; Blikstein and Worsley, 2016]. Also, recent studies have demonstrated a paradigm shift in learning analytics approaches from knowledge–based to empirical experiences and community–based learning which take into consideration the contents produced from mobile, immersive learning, the Internet of Things (IoT), and other online sources (like massive open online courses (MOOCs) [Loizzo and Ertmer, 2016], social network services (SNS), etc.) [de Freitas, 2013; Gibson and de Freitas, 2016; Siemens, 2005; 2014; Jülicher, 2018]. Virtual learning environments provide institutions of higher education with the wealth of learners' usage data such as the number of clicks, posted messages in discussion forums (message analysis), login trends, content usage, attempted assessments and the number of times they were provided with formative feedback (performance analysis and results), and more [Tempelaar et al., 2015; Rienties et al., 2016]. Consequently, recent forms of analytical techniques in the context of education should be considered. Multi–modal learning analytics (MMLA) which incorporate students' text, audio, video, gestures, behavior, biometric figures, and activity logs to analyze learning process in a more realistic ecosystem is an example of such techniques [Blikstein and Worsley, 2016; Schneider and Blikstein, 2015; Worsley, 2014; Worsley and Blikstein, 2015].

Overall, learning analytics as a bricolage of disciplines is specifically focused on education [Ferguson, 2012a; Shum and Ferguson, 2012]. LA also covers a wide range of research areas related to the teaching and learning. Some of which are mentioned in the following [Peña-Ayala, 2018]:

- *Social Learning Analytics* — which is concerned with the techniques and approaches to collect, calculate and evaluate the impact of learning related social media utilization of the students. This category incorporates several social network analysis methods including the social media text and network analysis [Martin et al., 2016].

- *Smart Learning Analytics* — which supports the intelligent educational data processing techniques to extract valuable insights from the collected data from diverse sources [Giannakos et al., 2016].

- *Multimedia and Video Learning Analytics* — which is focused on the techniques to capture useful information from the learning resources with the streaming media (including

the video) [Giannakos et al., 2016].

- *Ubiquitous Learning Analytics* — which acquires rich meta–data from the learners' interactions with the learning management systems and extracts the data and their corresponding contexts to help them connect to their desired relevant contextual learning materials [Mouri and Ogata, 2015].

- *Visual Learning Analytics* — which is mostly concerned with relevant data visualization techniques (from the educational data collection to their interactive visual representation processes) to assist the institutions of higher education with their decision making process [Hillaire et al., 2016].

- *Multimodal Learning Analytics* — which is focused on acquiring and processing heterogeneous data collected from human interactions and activities with the help of sensors and their related technologies [Andrade et al., 2016; Ochoa and Worsley, 2016].

- *Dispositional Learning Analytics* — which supports rapid formative feedback of data at several levels, integrates the process data with the performance data, and produces complex visualizations for both learners and instructors [Tempelaar et al., 2017].

- *Open Learning Analytics (OLA)*[5] — which is concerned with learning analytics solutions that support openness of processes, algorithms and technologies in collecting and processing the educational data, modularized integration of several learning analytics processes, and open technologies for the researchers in the area to get access to different implemented data mining, analytics and adaptive contents [Muslim et al., 2016; Siemens et al., 2011].

Given the abovementioned information, institutions of higher education have become more interested in adopting effective learning analytics techniques with the focus on transforming educational data into useful actions to foster learning processes and help to make better decisions [Chatti et al., 2014].

Learning analytics is capable of covering processes in a wide range of academic levels and stakeholders, including [Ifenthaler and Widanapathirana, 2014; MacNeill et al., 2014; Shum et al., 2012]:

[5] https://solaresearch.org/initiatives/ola/

- *"mega–level" or cross–institutional analytics* — is concerned with identifying patterns across multiple institutions. Such rich datasets and invaluable insights could be of help in policy–making processes within the government.

- *"macro–level" or institution–wide analytics* — is focused on integrating generated data from diverse sources, optimize pedagogical processes, perform desired analyses, and produce relevant visualizations to convey current and future (predicted) performance of student cohorts within the institution of higher education. In a long run, the macro–level analytics can assist institutions to decrease their attrition rates, improve student experience, and make adaptive and coherent academic decisions.

- *"meso–level" or the curriculum and instructor analytics* — which deals with the design process of course materials and resources from the instructors' perspective. Analyses in this level facilitate the learning process by improving the course quality and optimizing the course resources/materials which lead to a more successful experience for both learners and instructors.

- *"micro-level" or the learner–centric analytics* — that is concerned with each individual learner's success in their learning pathways. The analyses in the micro–level collect each student's interaction data with the learning management system and support them through adaptive interventions and relevant recommendations.

A holistic learning analytics solution should be capable of adopting techniques from gaming, educational data mining, computer–supported collaborative learning, recommender systems, intelligent tutoring systems (ITS), social network analysis (SNA), computational linguistics, and information visualization fields to be effective in this level [Shum et al., 2012]. Some studies considered three levels by merging the mega and macro levels as cross–institutional analytics [MacNeill et al., 2014; Shum et al., 2012].

2.2.2 Learning Analytics, Educational Data Mining, Academic Analytics

With the increasing interest of HE institutions in utilizing high–quality analytical techniques to process the wealth of educational data and to meet their academic objectives, several analytical approaches in the context of education have been introduced [Sclater et al., 2016; Siemens and d Baker, 2012]. Academic analytics (AA), educational data mining (EDM), and learning analytics (LA) are examples of such key disciplines [Sclater et al., 2016]. Although LA, EDM, and AA are closely related, the body of research is more focused on the learning

analytics–educational data mining overlap [Siemens and d Baker, 2012; Baker and Inventado, 2014; Papamitsiou and Economides, 2014; Berland et al., 2014; Bienkowski et al., 2012; Liñán and Pérez, 2015; Vahdat et al., 2015; Jülicher, 2018]. The definition of each discipline along with their inter–relationships are elaborated as follows:

- *Academic analytics (AA)* — is a specific field of research concerning with the economic and policy issues of higher education. AA is mostly focused on the administrative processes of educational institutions where admission policies, funding directions, and other relevant processes are taking place. It is referred to as the data–intensive decision–making process at the macro (and in some cases meso) level(s) that aims to improve institutions' effectiveness by using the data and enhancing their processes, resource allocation and evaluation approach [Goldstein and Katz, 2005; Campbell et al., 2007; Baepler and Murdoch, 2010]. The acknowledged definition for AA according to the literature is:

  > "Academic analytics combines select institutional data, statistical analysis, and predictive modeling to create intelligence upon which students, instructors, or administrators can change academic behavior" [Baepler and Murdoch, 2010].

  Academic analytics focuses more on utilizing data for marketing and administrative purposes [Sclater et al., 2016].

- *Educational data mining (EDM)* — is primarily focused on the technical challenges regarding the analysis of a large amount of educational data (utilizing automated methods) to extract valuable insights [Romero and Ventura, 2010; Baker and Yacef, 2009; Romero et al., 2010; García-Saiz et al., 2014; Peña-Ayala, 2014; Sclater et al., 2016; Baker and Inventado, 2014]. EDM is mostly concerned with the technical issues and is categorized as a special area of data mining for HE. The official definition for EDM according to the literature is:

  > "developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist" [Romero and Ventura, 2013; Papamitsiou and Economides, 2014].

Educational data mining was thriving under the hood of intelligent tutoring systems (ITS), artificial intelligence in education (AIED), user modeling (UM), technology–enhanced learning, and adaptive and intelligent educational hypermedia (AIEH) prior to being introduced as an independent area of research [Romero and Ventura, 2013]. The first international conference on EDM was held in 2008 after the first EDM workshop in 2005 [Siemens and d Baker, 2012]. Key acknowledged methods incorporated in educational data mining include but not limited to *prediction* (classification, regression, latent knowledge estimation), *structure discovery* (clustering, factor analysis, domain structure discovery), *outlier detecting*, *relationship mining* (association rule mining, sequential pattern mining, correlation mining, casual data mining), *social network analysis*, *process mining*, and *text mining*. Other methods with great saliency in educational data mining are the distillation of data for human judgment, the discovery with models, knowledge tracing (KT) and nonnegative matrix factorization (NMF) [Romero and Ventura, 2013; Baker and Inventado, 2014].

- *Learning analytics (LA)* — is mostly concerned with educational issues, learner success, and enhancing aspects of learning [Sclater et al., 2016]. Learning analytics utilizes methods in collecting learners' data, analyzing data and extracting valuable information from them, and reporting the results to the learner, educator, and the institute. The ultimate goal of learning analytics is to develop new ways to analyze educational data and constantly improve the learning and teaching processes [Baker and Inventado, 2014]. LA aims at transforming the educational data into useful actions to enhance the quality of learning [Bienkowski et al., 2012].

With regard to LA–EDM related studies, although learning analytics and educational data mining share several analytical objectives and methodologies in higher education, they adopt different perspectives toward their approaches [Baker and Inventado, 2014; Siemens and d Baker, 2012]. A short review over their similarities and differences is elaborated as follows:

- *Similarities* — both are concerned with the data–driven approaches in education [Siemens and d Baker, 2012]. Both are keen to extract valuable insights from the wealth of educational data to help the institutions of higher education with their planning, intervention, and decision–making processes and improve the quality of teaching and student experience [Siemens and Long, 2011].

- *Differences* — according to [Siemens and Long, 2011], several differences between the two can be listed as follows:

    1. Prioritization of discovery types — automated in educational data mining vs. human judgment–based in learning analytics,
    2. Supported types of adaptation and personalization — automated adoption (without human involvement such as intelligent tutoring systems) in educational data mining vs. instructor–learner centric in learning analytics, and
    3. Perspective toward models and frameworks — more reductionistic in educational data mining (reducing the system into its smaller components and analyzing each one) vs. more holistic in learning analytics (viewing and understanding the system as a whole) [Siemens and d Baker, 2012; Baker and Inventado, 2014; Papamitsiou and Economides, 2014; Berland et al., 2014; Bienkowski et al., 2012; Liñán and Pérez, 2015; Vahdat et al., 2015; Jülicher, 2018].

Figure 2.4 illustrates the number of published articles in learning analytics, educational data mining, and academic analytics disciplines along with their mutual studies in the course of eight years (from 2011 to 2018)[6]. Please note that the statistics for 2018 is based on the conducted search on mid–February 2018. As per Figure 2.4, learning analytics has attracted a huge amount of researchers' attention since its establishment in 2011 compared to it other analytical counterparts. According to Figure 2.4, educational data mining seconds learning analytics and the LA–EDM joint study seem to be an appealing topic after learning analytics and educational data mining research areas. Academic analytics also had a spark in 2014, but like LA–AA joint studies, deprived of full attention in the field.

2.3 Analytical Frameworks For The Context of Education

As mentioned in Section 2.2, educational institutions have increasingly become interested in gaining valuable insights using pertinent analytical approaches to help them make informed pedagogical, timely decisions, given their voluminous data [Siemens and Long, 2011; Johnson et al., 2015; Jaramillo, 2017; Siemens et al., 2011], given that: (1) they possess historical and streaming data repositories recording their learners' interactions with the learning management systems (LMS), and (2) they have commonly been inefficient in utilizing the data

[6]Based on the results searched on Google Scholar (https://scholar.google.com.au/)

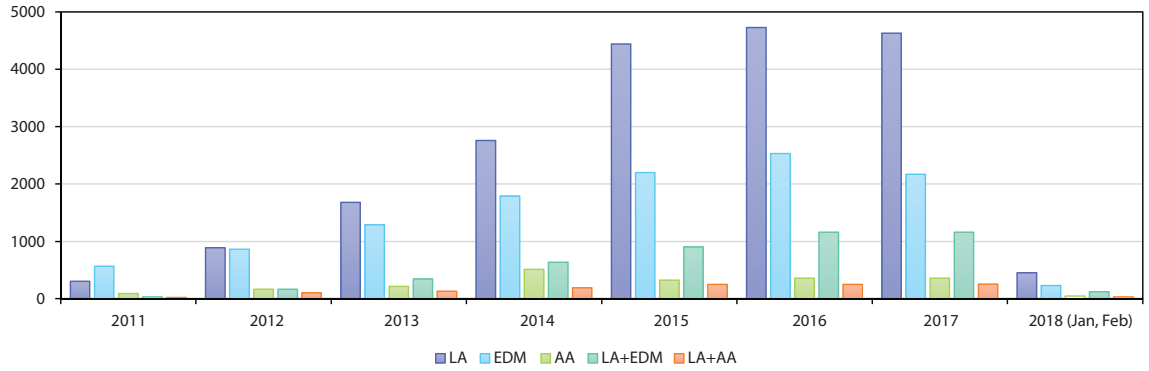


Figure 2.4: LA, EDM, AA, and Their Mutual Publications [2011 – 2018]

in their analytics to generate useful and timely outcomes and feedback [Siemens and Long, 2011]. Therefore, given the students' explicit and implicit activities in learning environments, institutions of higher education can leverage state–of–the–art analytical approaches to extract insights from student data and pursue proper actions [Jaramillo, 2017; Johnson et al., 2015; Siemens et al., 2011]. Learning analytics (LA) has emerged to address these requirements [Martin and Sherin, 2013; Siemens and d Baker, 2012; Chatti et al., 2014; Ferguson, 2012a; Elias, 2011; Siemens, 2013; Siemens and Long, 2011]. Its aim is to help institutions of higher education to make proper decisions, based on their wealth of information using advanced data mining and modeling techniques [Jaramillo, 2017; Johnson et al., 2015].

By considering the learning analytics requirements presented in Section 3.3, it is evident that LA is an excellent use case for utilizing accurate and coherent analytical techniques. One way to address those concerns is to propose robust and sustainable models in the context of LA, which incorporate different analytical techniques in a way that satisfy LA needs. Several proposed models were investigated to be adapted to the learning analytics context. The one that suited most of the requirements was selected as the base LA reference model and was customized to be adjusted to our proposed framework. Given that learning improvement is a key objective for institutions of higher education, utilizing a systematic approach to meet pedagogical requirements is of great importance. Educational institutions need clear learning analytics models (LAM) to satisfy their needs, by analyzing educational data, producing comprehensible insights, assisting them in their decision–making processes and, finally, improving the learning and teaching processes. Considerable research has been conducted in proposing LAMs and researchers have been working towards developing robust and exhaustive LAMs [Bienkowski et al., 2012; Siemens et al., 2011; Elias, 2011; Siemens, 2013; Greller

Table 2.1: Learning Analytics Requirements Coverage By Learning Analytics Models

Learning Analytics Requirements	*Surveyed LA Models* [1]	[2]	[3]	[4]	[5]	[6]
Data Collection and Integration	✓	✓	✓	✓	✓	✓
Learner Profiling	✓		✓	✓	✓	✓
Data Interpretation and Insight Extraction	✓	✓	✓	✓	✓	✓
Prediction	✓	✓	✓	✓	✓	✓
Decision–Making and Intervention	✓	✓	✓	✓	✓	✓
Adaptation and Personalization	✓	✓		✓	✓	✓
Ethics and Privacy					✓	✓
Information Visualization	✓	✓	✓	✓	✓	✓

[1] [Siemens et al., 2011] [2] [Elias, 2011] [3] [Siemens, 2013] [4] [Bienkowski et al., 2012] [5] [Chatti et al., 2012] [6] [Greller and Drachsler, 2012]

and Drachsler, 2012; Chatti et al., 2012]. Table 2.1 summarizes some oft–cited studies in this area and maps proposed LA models to the key learning analytics requirements mentioned in Section 3.3.

According to our literature survey, the body of learning analytics research has agreed on a prominent four–dimensional reference model for learning analytics [Chatti et al., 2012; Greller and Drachsler, 2012]. We picked the proposed model in the work of [Chatti et al., 2012] as the base for our logical module, described in Section 5.4, since it widely covers the main learning analytics requirements in detail, as well as serving as an adequate example of our proposed framework. This model presents a systematic view of learning analytics and its related concepts. Our proposed four–dimensional learning analytics model – named the LA reference model – was elaborated in Section 3.3 and was illustrated in Figure 3.1 of Chapter 3. The mentioned LA reference model frames the scope of the logical layer of the proposed framework in Chapter 5. It accounts for identifying the key learning analytics requirements within the proposed framework.

The utilization of LA, however, entails meeting particular criteria in the context of education which translate into major learning analytics requirements [Siemens and Long, 2011; Siemens et al., 2011; Chatti et al., 2014]. Data collection, insight extraction, prediction, intervention, personalization, adaptation, and visualization are examples of important LA requirements which are elaborated in detail in Chapter 3. One way to address learning analytics challenges is utilizing a clear and generic analytical model [Siemens, 2013; Bienkowski et al., 2012; Greller and Drachsler, 2012]. It is essential for learning analytics to be formalized

in the context of analytics. Fulfillment of LA requirements calls for a novel federated analytics solution which is generic enough to cover most learning analytics needs. An example of why we need an integrated analytics solution to address the main needs of educational institutes can be as follows: during a given semester, students are taught several concepts through a diverse set of learning resources. They also sit for different assessments evaluating their progress in their learning process. At any given point of time, the descriptive analytics will help us generate analytical report on each student's performance in terms of their previous assessment results; the predictive analytics will also help us extrapolate individual student's likelihood of being at risk of failure (based on their past performance); the prescriptive analytics, finally, will give us the opportunity to perform informed and personalized interventions for each student and help them with their productive learning pathways. One integrated analytics solution could give the educational institutions the opportunity of providing each student with personalized and to the point of their needs pedagogical means to improve their student experience and elevate their self–esteem.

Finally, it is worth mentioning that the main novelty of our proposed work compared to ITS could be categorized as follows.

- ITS usually provides students with instant formative feedback based on their previous interactions with the LMS. This is the case in our approach as well. However, our framework goes one level in detail and deals with each individual student as well as cohorts of students (the granularity of the process can be directed towards individuals or groups).

- ITS usually deploys the descriptive (and in rare cases, prescriptive) analytics methods. However, our proposal incorporates descriptive, predictive, and prescriptive approached to provide an enriched set of recommendations to students.

2.4 Summary

The current body of research in data analytics (descriptive, predictive, prescriptive analytics), in the context of education (EDM, AA, LA), and proposed analytics framework in education are reviewed in this chapter. Starting from the next chapter (Chapter 3), the basis for our proposed analytics framework is built by introducing a four–dimensional learning analytics reference model that covers the main learning analytics requirements.

Chapter 3

Learning Analytics

> "Develop a passion for learning. If you do, you will never cease to grow."
>
> *Anthony J. D'Angelo*

This chapter completes the discussion around learning analytics in Chapter 2 and prepares the ground for the analytics framework in Chapter 5. The key learning analytics processes and their interrelationships are elaborated and a four-dimensional learning analytics reference model will be introduced which will be the basis for the logical layer of the proposed framework.

3.1 Introduction

Educational institutions are increasingly relying on data–intensive analytics to make timely pedagogical decisions [Siemens and Long, 2011; Johnson et al., 2015; Australian Government and Training, 2017; Siemens et al., 2011]. They possess large amounts of historical and streaming data generated by students' interactions with the digital learning systems such as learning management systems (LMSs). In addition, Institutions of higher education have traditionally been unsuccessful in taking advantage of the relevant data processing techniques which in turn generates a gap between owning the data and exploiting valuable and meaningful insights based on the data to meet their objectives. Therefore, they have become interested in adopting analytics approaches suitable to their needs to overcome the mentioned gap [Siemens and Long, 2011]. Another critical factor is transforming the avail-

able information into insights and acting upon them to address pedagogical purposes of the educational institutions. Learning analytics (LA), a thriving area in technology–enhanced learning (TEL) research, has emerged to address the aforementioned concerns by combining data collection, insight extraction, prediction, and recommendation techniques [Martin and Sherin, 2013; Siemens and d Baker, 2012; Chatti et al., 2014; Ferguson, 2012a; Elias, 2011; Siemens, 2013; Siemens and Long, 2011].

LA shares some functionalities with other major higher education (HE) data analytics such as educational data mining (EDM) and academic analytics (AA) [Siemens and d Baker, 2012; Baker and Inventado, 2014; Baepler and Murdoch, 2010; Campbell et al., 2007; Goldstein and Katz, 2005]. Learning analytics also emerges in the research fields where the education and computer science disciplines have the majority of the impact on such as sociology, education, learning sciences, statistics, machine learning, and intelligent systems as well as linguistics and philosophy [Sclater et al., 2016; Dawson et al., 2014].

Successful adoption of learning analytics calls for the satisfaction of certain criteria in the context of education which are referred to as the learning analytics requirements [Siemens and Long, 2011; Siemens et al., 2011; Chatti et al., 2014]. Data collection, insight extraction, prediction, intervention, personalization, adaptation, and visualization are examples of key learning analytics requirements. One way of addressing learning analytics concerns is to utilize a clear and generic analytics approach capable of performing data collection and acquisition, storage and retrieval, cleaning, integration, analysis, visualization, and intervention to deploy analytics in educational settings [Siemens, 2013; Bienkowski et al., 2012; Greller and Drachsler, 2012]. According to the literature, the incorporation of learning analytics solutions entails certain benefits and drawbacks that could be taken into consideration by educational institutions. Major factors are summarized in the following [Sclater et al., 2016; Papamitsiou and Economides, 2014].

- *Advantages* — a large amount of educational data is taken into account. Learning analytics is capable of utilizing simple and complex analytical approaches, a wide range of visualization tools are produced for learners, instructors, and institutional staff. LA also promotes adaptive and personalized learning which in turn help students improve their experience, self–esteem, self–assessment, reflection, awareness, and self–efficacy.

- *Disadvantages* — there are several major ethical and data privacy concerns associated with learning analytics applications. There is no generic solution for similar educational problems in current LA solutions. Most LA systems lack proper visualization tools and

proper illustration of analytical results to help the instructors and the institutions of higher education make informed decisions. Finally, learning analytics solutions are usually inefficient in acquiring the educational data from diverse sources, unify, and process several data types.

To construct the basis for the thesis, this chapter aims at providing a systematic literature review of learning analytics, its definition, history, relation with other analytical techniques in higher education, requirements, tools, techniques, applications, models, challenges, and future directions along with a 4–dimensional learning analytics reference model introduction based on the work proposed by [Chatti et al., 2012]. Furthermore, major learning analytics concerns and requirements extracted in this chapter are addressed in the proposed data–driven analytical framework is in Chapter 5 based on the material discussed in this chapter.

For an extensive literature review of several studies in the field of learning analytics, LA definition, different aspects of LA, multiple disciplines of LA involvement, and several academic levels LA–oriented solutions can cover, please refer to Chapter 2. Moreover, a comparative review of different analytical approaches in the context of education along with their definitions, similarities, and differences, and the extent to which the academics have been conducting research on each technique is provided in Chapter 2 as well.

Section 3.2 is dedicated to capturing a clear and comprehensive list of LA needs according to the extant body of research in the field. Key LA requirements are listed in this section. Section 3.3 reviews the top academic literature proposing several models to address the majority of LA needs. By taking into consideration multiple actors within educational systems and key LA requirements extracted in Section 3.2, a 4–dimensional learning analytics reference model is introduced which is capable of addressing most LA processes. Section 3.4 is concerned with several prestigious and active tools and applications introduced in LA along with their functionalities and differences. Section 3.5 is mostly focused on the current issues and challenges most LA solutions face and provides grounds for future research directions in the field. Finally, in Section 3.6, a concise review of the chapter's discussed material is performed along with contribution highlights and the connection to the next chapter.

3.2 Learning Analytics Requirements

Given its definition and objectives in assisting the institutions of higher education, LA needs to ground specific requirements such as data collection, measurement, analysis, reporting and interpretation processes given the educational data repositories [Gašević et al., 2015].

It is critical for a learning analytics solution to be capable of predicting learners' performance and modeling their behavior. The learners' performance extrapolation and modeling have extensively been researched with the help of educational data mining, educational user modeling, and educational adaptive hypermedia communities. The objective is to estimate the unknown value of a variable that describes the learner, such as performance, knowledge, scores or learner grades [Marquez-Vera et al., 2010; Romero et al., 2008]. Such forecasts are for example utilized by intelligent tutoring systems (ITS) [Kulik and Fletcher, 2016] to provide hints, instant comments or any other sort of formative feedback when a student is responding to a question. A plethora of research has also been conducted in the field of dynamic learner models to support the adaptation of the educational hypermedia systems [Brusilovsky and Millán, 2007].

A beneficial learning analytics system could also be able to suggest relevant and proper learning resources/materials. Recommender systems for learning have also gained increased attention recently. A recent survey of technology–enhanced learning recommender systems has been elaborated by [Manouselis et al., 2011]. These systems typically analyze learner data to suggest relevant learning resources/materials, peer learners or learning pathways.

Furthermore, increasing reflection and awareness of the learner is another important attribute. Several researchers are turning their focus on the analysis and visualization of different learning indicators to foster awareness and reflection about learning processes. These indicators include resource accesses, time spending, and knowledge level indicators [Mazza and Milani, 2005].

An effective learning analytics solution could also be capable of enhancing social learning environments. A considerable amount of research has been conducted in the analysis and visualization of the learners' social interactions to make people aware of their social context and to enable them to explore the context [Heer and Boyd, 2005]. In technology–enhanced learning, this is particularly, but not only, relevant for computer–supported collaborative learning (CSCL) [Stahl and Hesse, 2009], where the interactions with peer learners are a core aspect of how learning is organized. In CSCL, much research has focused on the analysis of networks of learners, typically with a social network analysis approach [Reffay and Chanier, 2003].

A perfect learning analytics solution is also able to detect undesirable learner behaviors. The objective of detecting undesirable learner behavior is to discover learners who have some type of problem or unusual behavior, such as erroneous actions, misuse, cheating, dropping out or academic failure [Romero and Ventura, 2007].

Another useful detection attribute of a learning analytics system is learners' affective states detection. Researchers in technology–enhanced learning often refer to the affective states defined by [D'Mello et al., 2007]. These states are classified as boredom, confusion, frustration, eureka, flow/engagement, versus neutral. Among others, the detection of effects is researched to adjust pedagogical strategies during learning of complex material.

According to the body of research, key learning analytics requirements can be classified into the following categories:

- *Data collection and integration* — gathering and unifying educational data from the learning management system, virtual and personalized learning environment (VLE, PLE) sources. Data collection, integration, transformation, and dimensionality reduction using accurate data reduction methods and coherent statistical analysis and data mining techniques are key elements to accomplish this task [Chatti et al., 2012; Jaramillo, 2017; Ferguson, 2012a; Siemens and Long, 2011; Chatti et al., 2014; Brown, 2011; Elias, 2011].

- *Learner profiling* — collecting and processing learners' data from their interactions with the learning management system, utilizing consistent analytics to extract valuable information from the data to build better pedagogies, enrich learning processes, and better educational resource allocation [Siemens and Long, 2011; Jaramillo, 2017; Kay, 2008].

- *Data interpretation and insight extraction* — applying relevant descriptive analytics techniques to understand what has happened until now. It requires special result description and diagnosis approaches to elicit beneficial insights from the educational data [Siemens and Long, 2011; Jaramillo, 2017; Chatti et al., 2012; 2014; Elias, 2011].

- *Prediction* — extrapolating likely scenarios in the future such as student retention rates, students at–risk of failure, resource utilization ratios, and the effects of educational policies by adopting accurate predictive analytics techniques [Siemens and Long, 2011; Jaramillo, 2017; Chatti et al., 2012; 2014; Brown, 2011; Elias, 2011; Siemens et al., 2011; Romero and Ventura, 2013].

- *Decision–making and intervention* — suggesting intelligent courses of action in accordance with the higher education institution's objectives to promote learning processes and academic success. Taking optimal and influential actions can assist educational

stakeholders (students, faculty, staff, and tutors) to meet their goals. Sophisticated and data–driven analytical approaches should be adopted to produce actionable recommendations [Siemens and Long, 2011; Jaramillo, 2017; Chatti et al., 2012; 2014; Brown, 2011; Elias, 2011; Siemens et al., 2011; Verbert et al., 2012].

- *Adaptive and personalized learning technologies* — given the diverse range of learners' needs, objectives and aptitudes, learning analytics should be capable of addressing their dynamic requirements and adapt its educational materials to learners' needs by utilizing sustainable and robust analytical approaches and optimization and recommendation techniques [Siemens and Long, 2011; Jaramillo, 2017; Chatti et al., 2012; 2014; Kay, 2008; Siemens et al., 2011].

- *Ethical issues and privacy preservation* — protecting learners' data in the learning management system. Devising accurate access levels, making data anonymous, and getting learners' consent (as well as institution's ethics approval) prior to processing their data are among the critical issues every system in the context of learning analytics should be concerned with [Siemens and Long, 2011; Ferguson, 2012a; Brown, 2011].

- *Information Visualization* — developing approaches to properly convey processed outcomes to the educational stakeholders. Clear, simple, and yet effective illustrations of the complex analytical results is a critical task. Comprehensible illustration of trends, predictions, recommended actions and extracted insights from the educational data is beneficial to students, instructors and academic institutes [Chatti et al., 2012; 2014; Elias, 2011; Kay, 2008; Siemens et al., 2011; Mazza and Milani, 2005; Mazza and Dimitrova, 2004].

3.3 Learning Analytics Models

One way to address the mentioned requirements in Section 3.2 is to adopt a robust learning analytics model (LAM) that utilizes accurate and coherent analytical techniques. A successful solution covering mentioned concerns could be proposed in the form of one sustainable model which incorporates several analytical techniques (including descriptive, predictive and prescriptive analytics) [Daniel, 2015; Soltanpoor and Sellis, 2016]. An effective LAM is capable of coupling the big data and learning analytics to provide benefits for the following key higher education stakeholders:

- *Administrators* — academic programming, resource allocation, and support ongoing efforts,
- *Students (learners)* — proactive feedback, learning pathways, and plan learning activities, and
- *Instructors* — help students at risk, improve teaching, and instant feedback [Daniel, 2015].

Given the current body of research in proposing learning analytics models and frameworks with the focus on addressing the majority of LA requirements, key studies in this area can be categorized and depicted in Table 3.1.

The body of research has agreed upon a four–dimensional reference model and a six–dimensional framework for learning analytics [Chatti et al., 2012; Greller and Drachsler, 2012]. To build the required ground of our learning analytics reference model (LARM), we picked the work in [Chatti et al., 2012], because it includes all the six–dimensions of [Greller and Drachsler, 2012]. Considering the incoming sources of the educational data (learning management system, virtual or personalized learning environment), the four dimensions of the learning analytics reference model can be elaborated as follows.

1. *What kinds of data, context, and environment does the system collect and utilize (the "WHAT" dimension)?* — Learning analytics is a data–driven approach. It deals with a wide range of educational data from different sources and types. Generally, two main sources of data in higher education are centralized educational systems like learning management systems and distributed learning environments like PLEs.
2. *Who is the stakeholder of the final product (the "WHO" dimension)?* — There are several stakeholders to whom learning analytics can be oriented such as:
   - Students — to improve their grades, enhance student experience, get benefit from adaptive and timely feedback from their lecturers and tutors, make more informed decisions about enrolling in selected courses and building their personalized learning pathways (such as PLEs),
   - Instructors — to augment the effectiveness of their teaching practices and to adapt their teaching offerings with students' needs, provide more enriched learning material, elevate their teaching quality, adopt more advanced pedagogical techniques

Table 3.1: Key Learning Analytics Models and Frameworks

LA Models	**LA Frameworks**
a reference model for learning analytics[1]	an open learner model (OLM) framework to shed light into several forms of OLMs[10]
a maturity LA model to guide practices of students engagement assessments[2]	a holistic learning analytics framework connecting diverse types of educational information which is validated with two case studies[11]
a complexity–grounded learning analytics model for assessment automation in the small–scale[3]	a classroom discussions argument development analysis framework[12]
an infrastructure for learning analytics to adopt ranges of analytical techniques to support real–time summaries along with visual analytics[4]	an evaluation framework (analytics4action) for evidence–based learning analytics interventions[13]
an intervention–oriented learning analytics model for online discussions with embedded analytics[5]	a learning analytics intervention and evaluation framework (LA–IEF)[14]
massive open online courses' discussion forums analysis utilizing an unsupervised technique[6]	a generic framework for LA[15]
a learning analytics model for learner profiling to support personalized learning[7]	a learning analytics framework elaborating on LA implementation with its educational–related processes[16]
a foundational LA model for higher education concerning with stakeholders' dynamic interactions with their data using visual analytics[8]	a conceptual framework which links learning design with LA[17]
a multimodal learning analytics model in complex learning environments to help extract new insights in students' learning trajectories[9]	a learning analytics framework for multiliteracies (new media literacies) assessment[18]

[1] [Chatti et al., 2012] [2] [Clarke et al., 2013] [3] [Goggins et al., 2015] [4] [Shum and Crick, 2012]
[5] [Wise et al., 2013] [6] [Ezen-Can et al., 2015] [7] [Kay, 2008] [8] [Freitas et al., 2015]
[9] [Blikstein and Worsley, 2016] [10] [Bull and Kay, 2016] [11] [Ifenthaler and Widanapathirana, 2014]
[12] [Sionti et al., 2011] [13] [Rienties et al., 2016] [14] [Rienties et al., 2017] [15] [Greller and Drachsler, 2012]
[16] [Colvin et al., 2015] [17] [Bakharia et al., 2016] [18] [Dawson and Siemens, 2014]

to identify at–risk students, and get constructive feedback from students' side to be able to provide them with effective interventions (such as formative feedback), and

- Educational institutions — to support their decision makings, improve students' success, develop student admission policies, adjust course planning, determine hiring needs, and make financial decisions based on the wealth of relevant and comprehensible analytical visualizations and reports.

3. *How does the learning system incorporate proper analytical methods over the collected data (the "HOW" dimension)?* — learning analytics utilizes a wide range of techniques to extract patterns from educational data. According to [Chatti et al., 2012; Greller and Drachsler, 2012], different computational approaches have been used in HE such as data mining, statistical analysis, information visualization, and recently social network analysis.

4. *Why the accumulated data should be processed (the "objectives" or the "WHY" dimension)?* — Any learning analytics solution should focus on key LA objectives such as increasing the retention rates, improving student experience, providing adaptive feedback to learners based on their interactions with the learning management system, helping the institutions of higher education with their critical academic decision making processes, and fostering the administration, teaching and learning processes. To address mentioned goals, a set of learning analytics processes could be implemented. LA processes are the driving force and functional units of the learning analytics framework. Key LA processes, according to the extant body of research, can be categorized as follows. Please note that some of these processes are interrelated as depicted in Figure 3.2. Please also note that all of the following processes are the building blocks of the logical layer proposed in Chapter 5:

   4.1. Monitoring — given students' previous activities and accomplishments in the LMS, the system tracks their digital footprints and provides instructors and educational institutes with students' data. For example, the system can collect each student's assessment results within a certain semester and aggregate them into one unified format to be used by other processes. This process also helps instructors evaluate the learning process in order to improve the learning environment and student experience. The *monitoring* process is explained in Section 5.4.1.

4.2. Analysis — refers to the statistical analysis of the educational data which is available through the learning system (usually LMS). Analysis can help instructors identify patterns and distinguish behaviors of students and produce proper insights to help with the decision–making process. The analysis also provides instructors with proper information to design future learning activities and enhance the student experience. For example, the process can utilize several descriptive and diagnostics techniques on the unified student data to calculate the ratio of correct responses, extract misunderstood concepts, and categorize students in specific performance cohorts. The *analysis* process is described in Section 5.4.2.

4.3. Prediction — builds accurate predictive models to extrapolate students' future performance, behavior and status, given students' activities within the learning management system. Instructors and institutions of higher education can properly intervene in students' tracks and provide them with actionable and effective suggestions and recommendations. For example, incorporating certain predictive algorithms such as Naïve Bayes or Neural Networks to project the students' final marks at the end of the semester, given their assessment results in a semester, and can notify the instructors to take informed actions. The *prediction* process is elaborated in Section 5.4.3.

4.4. Intervention — provides [1] students with proper and actionable suggestions and recommendations based on the effective analysis of activities and accurate prediction of their future performance to improve the academic performanceand enhance student experience. Some examples can be providing students with learning material corresponding to their incorrect assessment responses, or asking them to attend tutorial/mentoring sessions covering the concepts they misunderstood. The *intervention* process is described in Section 5.4.4.

4.5. Tutoring and mentoring — given the analysis results of the students' previous activities and accomplishments, tutors and mentors can provide students with personalized guidance and support. It covers a broad range of activities including learners' orientation, new learning resources (subject–based or interest–based) suggestion, and goal achievement plans. The *tutoring and mentoring* process is explained in Section 5.4.5.

4.6. Assessment — considering students' interactions with the learning management

[1] The interventions are assumed to be generated by the recommendation algorithms.

system and their preferences, the assessment process will help learners to improve their learning processes and enhance their experience using specific assessment and self–assessment techniques to identify learners' strengths and weaknesses in their journey. The communication media with learners is intelligent feedback which is disseminated to both students and instructors/mentors/educational institutes as well. For example, adaptive sets of designed questions covering each student's misconceptions throughout the semester will help them identify and rectify their misunderstood concepts. This process is elaborated in Section 5.4.6 and is one of the major components of the PPQ approach proposed in Chapter 6.

4.7. Feedback — feedback process plays a critical role in the whole learning analytics environment which collects useful information and disseminates them to relevant stakeholders [Pardo, 2018]. Its main objective is to improve the overall learning process, enhance student experience, elevate learning performance, increase retention rates as well as decline the drop–outs, and minimize the number of potential at–risk students. Almost all other processes communicate with the feedback process to deliver their recommended actions to their pertinent targets (mostly students). One example is to provide an instant formative feedback to each student right after they responded incorrectly to one question and explaining why the response was incorrect. The *feedback* process is described in Section 5.4.7.

4.8. Adaptation — by collecting and analyzing students' data as well as their personal preferences, instructors and higher education institutions can trigger specific and effective interventions for learners. Adaptation process provides beneficial learning resources and instructional activities to students based on their requirements, goals, and interests. Adapting the learning material or the suggested actions for each student based on their previous assessment results are examples of the *adaptation* process that is further elaborated in Section 5.4.8.

4.9. Personalization — there has been a shift in learning processes from the learning management systems with the knowledge–push approach to personalized learning environments with the knowledge–pull approach. In the former, information flow is managed by instructors. In the latter, on the other hand, it is the learner who discovers knowledge through the information provided to them based on their personalized and preferred objectives, goals and needs. Well–defined recommender systems can provide the learner with specific learning material/resources (both

implicit and explicit). One example is providing each student with personalized and adaptive sets of questions, specifically designed for that individual student, covering their mistakes from the previous assessment in a certain course. We focused on this process further in Chapter 6 by proposing the PPQ approach. The *personalization* process is also explained in Section 5.4.9.

4.10. Reflection and Self–Reflection — by collecting and analyzing students and instructors' previous activities and experiences in the learning system, the reflection process can help them compare their performance (students and instructors) and teaching approaches (instructors) with other courses, other classes or even other educational institutions. The *reflection* process is further elaborated in Section 5.4.10.

Based on the learning analytics model proposed by [Chatti et al., 2012], the discussion in Section 3.3, and the aforementioned learning analytics processes, we introduce a 4–dimensional learning analytics model which is illustrated in Figure 3.1.

As per Figure 3.1, the first axis represents the "What" dimension which deals with several educational data types from diverse sources; the second axis is concerned with the "Who" dimension which corresponds to the LA solution's stakeholders; the third axis is related to the methods of educational data mining and analytics of the "How" dimension; and finally, the fourth axis is focused on the "Why" dimension and its 10 different LA processes. For each given learning analytics process, we have its corresponding methods, data types, and stakeholders. Later in Chapter 5, we propose one integrated analytics framework that incorporates this 4–dimensional model and relates each LA process of the "WHY" dimension with their corresponding components of other dimensions.

A graphical representation of learning analytics processes and their interrelationships is depicted in Figure 3.2. The *monitoring* process is initiated by collecting the educational data from learners/instructors interactions with the learning management system and provides the input to all other LA processes (except the *feedback* process) for further analysis of the data. The *feedback* process, on the other hand, receives the processed results from all LA processes (except the *monitoring* process), and disseminates them back to the targeted learners/instructors.

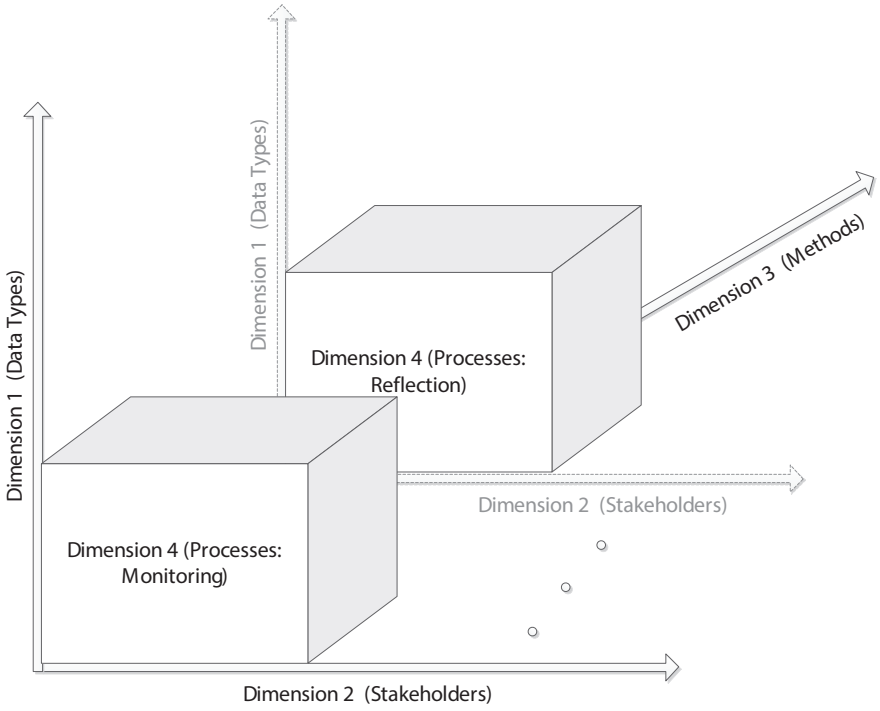


Figure 3.1: The Proposed 4–Dimensional Learning Analytics Model.

3.4 Learning Analytics Tools and Applications

A considerable amount of research has been conducted on developing effective learning analytics tools and applications to address major concerns of the institutions of higher education such as increasing retention rates, precisely identifying at–risk students, providing the learners with supportive interventions, and regulating influential academic policies [Shacklock, 2016; Colvin et al., 2015]. A qualified learning analytics application could be capable of including but not limited to the following features.

- *Performance prediction* — by analyzing students' interactions with the learning management system (which is addressed in the "prediction process" of the proposed analytics framework in the "logical layer" - Chapter 5),
- *Attrition risk detection* — by monitoring students' behavior and analyzing their drop–out patterns (which is addressed in "analysis" and "prediction" processes of the proposed analytics framework in the "logical layer" - Chapter 5),

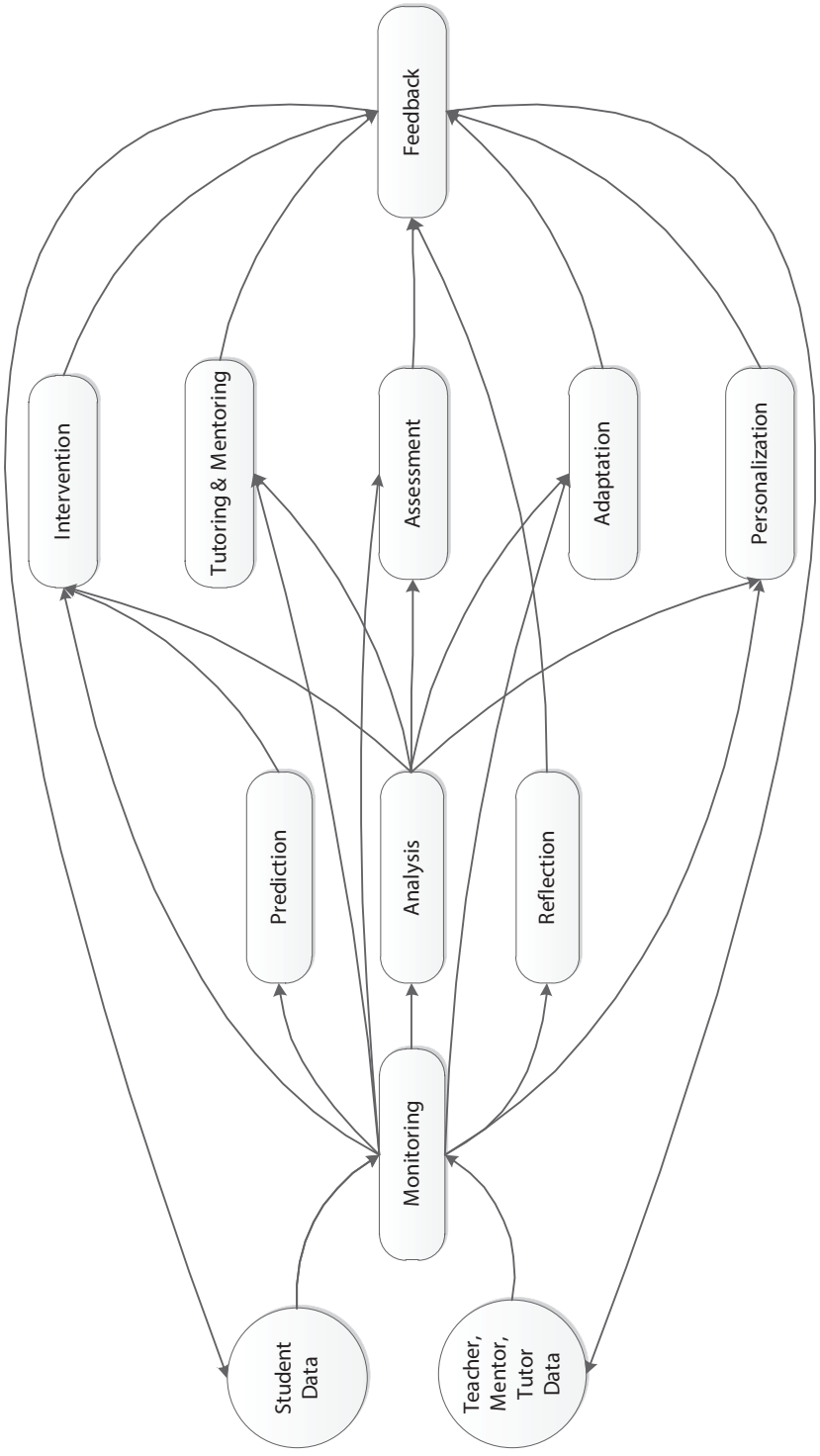


Figure 3.2: Learning Analytics Processes Interrelationships.

- *Data visualization* — by utilizing effective visualization techniques to detect valuable trends in learners' performance and producing visual reports (which is addressed in the proposed PPQ approach in Chapter 6),
- *Intelligent feedback* — by providing near real–time feedback regarding students' interactions with the LMS to help them improve their performance and experience (which is addressed in the "feedback process" of the proposed analytics framework in the "logical layer" of Chapter 5),
- *Course recommendation* by suggesting adaptive courses to students based on their interest, activities, performance, and aptitude (which is addressed in the "intervention process" of the proposed analytics framework in the "logical layer" of Chapter 5) as well as the proposed PPQ approach in Chapter 6 as the "physical layer",
- *Student estimation and behavior detection* — by collecting students' interactions and behavior data from the LMS and other means of pedagogy (like designed educational games) - this is addressed in "monitoring", "analysis", and "prediction" processes of the proposed analytics framework in the "logical layer" Chapter 5, and
- *Social network analysis* — by gathering students' public social data and constructing courseware and concept maps [Charlton et al., 2013; Sin and Muthu, 2015].

Some flagship academic/commercial software and applications (sometimes referred to as "Learning Analytics Dashboards" [Verbert et al., 2013]) are listed in Table 3.2.

3.5 Learning Analytics Challenges and Future Directions

Due to the complex nature of learning analytics and its inherent usage issues with simplified conceptualizations, the institutional adoption of LA has been hampered recently [Colvin et al., 2015]. Furthermore, the learning analytics pedagogy function in data analysis processes has encountered problems in some cases [Dietz-Uhler and Hurn, 2013; Colvin et al., 2015]. It means that there should be pedagogy that drives learning analytics and not the other way. Some key concerns of adopting education–based analytical solutions (including educational data mining, learning analytics, academic analytics) can be categorized as follows [Nunn et al., 2016; Rubel and Jones, 2016; Campbell et al., 2007].

- *Data privacy and security,*

Table 3.2: Well–Known LA Tools and Applications

LA Tools/Applications	**Description**
Course Signals[1,2,3]	improving retention & performance by effective interventions. Student–centric:[b,c]
GLASS[3,6]	visualization of learning performance, comparison. Teacher[α] Student–centric:[a,c]
LOCO–Analyst[3,9]	providing feedback on students' activities & performance. Teacher–centric:[a,c]
Student Success System[3,10]	spotting and treating at–risk students. Teacher–centric:[b,d]
SNAPP[3,11]	evolution visualization of learners' relationships within discussion forums. Teacher–centric:[a,d]
Student Inspector[3]	monitoring students' interactions with the LMS. Teacher–Student–centric:[a,c]
SAM[3,12]	enabling students' self–reflection and awareness. Teacher–Student–centric:[a,c]
StepUp![3]	promoting learners' reflection and awareness. Teacher–Student–centric:[a,d]
Narcissus[3]	improving students' teamwork skills (providing contribution reports). Student–centric:[a,d]
CAMERA[4,5]	promoting self–regulated learning in PLEs by providing students' self–reflection reports.
CourseVis[4,7]	monitoring students' activities in distant courses using Web log data. Teacher–centric:[a]
Moodog[4,8]	monitoring learners' online activities by analyzing CMS logs. Teacher–Student–centric:[a]
BlackBoard Analytics[2]	assisting institutions of HE to optimize student experience.
Civitas Learning[2]	optimizing data, maximizing insights, informing actions, and continuing learning.
D2L BrightSpace Insights[2]	monitoring students' activities in the LMS, real–time interventions.

[α] "Teacher" here is an alternative to "instructor".
[a] *descriptive*: the simple representation of raw data.
[b] *partial–prescriptive*: involvement of prediction algorithms & early warning systems.
[c] *formal*: secured, scalable and traceable from the LMS.
[d] *informal*: more open and social. [1] [Arnold and Pistilli, 2012; Gašević et al., 2015; Barber and Sharkey, 2012] [2] [Jayaprakash et al., 2014] [3] [Park and Jo, 2015; Corrin and de Barba, 2014; Verbert et al., 2014] [4] [Ruipérez-Valiente et al., 2015] [5] [Schmitz et al., 2009] [6] [Leony et al., 2012] [7] [Mazza and Dimitrova, 2004] [8] [Zhang et al., 2007] [9] [Jovanovic et al., 2008] [10] [Essa and Ayad, 2012] [11] [Dawson et al., 2010] [12] [Govaerts et al., 2010]

- *Holistic solutions (entailing descriptive, predictive, and prescriptive analytics),*
- *Multiple stakeholders (including the institution, students, instructors) involvement,*
- *Proper user profiling,*
- *Obligation to act on the produced analytical results,*
- *Proper educational data tracking,*
- *Collection and analysis, and*
- *The academic resource dissemination policy.*

Moreover, building strong connections with the learning sciences, developing methods capable of interacting with a wide range of datasets to optimize learning environments, focusing on learners' perspectives, and regulating and enforcing an explicit set of ethical guidelines are critical factors to be considered in adopting a learning analytics solution [Ferguson, 2012a].

Implementing learning analytics in large–scale (such as the faculty, institutions of higher education, or nation–wide), and performing institutional planning require the involvement of a wide variety of professionals from education, administration and staff [Ferguson et al., 2014]. However, according to [Macfadyen and Dawson, 2012], the process carries multiple shortcomings such as adopting solutions based on the particular culture of institutions of higher education, awareness of the extent to which the institution is inclined to changes and proposing incentive factors to encourage the institutions to adopt changes (behavioral modifications). Also, providing the institutions with adaptive learning analytics solutions based on their evolving requirements is another concern [Ferguson et al., 2014].

Some critical issues (referred to as "gaps") in *scaling–up* the current learning analytics solutions mentioned in [Lonn et al., 2013] can be illustrated as follows.

- *Usability gaps (business objects)* — issues with the graphical user interface (GUI) on how to replicate and illustrate the stored data in a database, in learning management system pages/reports,
- *Calculation gaps (errors in manipulating grade book data)* — inconsistencies among different assessment calculations from the student and the instructor's point of view,
- *Access gaps (two–factor authentication)* — issues with properly securing sensitive student data within the institution's data warehouse,

- *Performance gaps (impact on enterprise systems)* — issues with proper LMS communication between its archival and production data and performing fast scaled (extract, transform, load) ETL processes, and
- *Automatization gaps (manual maintenance of cohort and advisor information)* — issues with converting traditionally persistent manual processes into automatic routines.

The code of practice for learning analytics is provided in [Sclater, 2014] along with a comprehensive literature review and suggestions for adopting LA–based solutions with the focus on the ethical and legal issues. Some key topics/concerns could be listed as:

- *Awareness and consent,*
- *Transparency around algorithms and metrics,*
- *Ownership and control of data,*
- *Usage of publicly available data,*
- *Accuracy of data,*
- *Respecting privacy,*
- *Opting out,*
- *Interpretation of data,*
- *Stewardship,*
- *Preservation and deletion of data,*
- *Interventions and the obligation to act,*
- *Impacts on student bahavior,*
- *Staff awareness and training,*
- *Anonymization*, and
- *Targeting resources appropriately.*

Furthermore, a checklist based on the previous research in the field, critical concerns regarding learning analytics ethics, privacy and legal frameworks challenges, and the LAK15 workshop suggestions on ethics and privacy in LA (EP4LA)[2] is proposed in [Drachsler and

[2] http://www.laceproject.eu/blog/about-todays-ethics-and-privacy-in-learning-analytics-ep4la/

Greller, 2016]. The checklist entails eight action points to be taken into consideration by authorities to establish trusted learning analytics approaches for their institutions [3]. The proposed checklist for a trusted learning analytics solution covering the majority of mentioned ethical concerns is named "DELICATE" and is comprised of the following elements (according to [Drachsler and Greller, 2016]):

- *D–etermination* — Decide on the purpose of learning analytics for your institution.
- *E–xplain* — Define the scope of data collection and usage.
- *L–egitimate* — Explain how you operate within the legal frameworks, refer to the essential legislation.
- *I–nvolve* — Talk to stakeholders and give assurances about the data distribution and use.
- *C–onsent* — Seek consent through clear consent questions.
- *A–nonymise* — De–identify individuals as much as possible
- *T–echnical aspects* — Monitor who has access to data, especially in areas with high staff turn–over.
- *E–xternal partners* — Make sure externals provide highest data security standards.

3.6 Summary

Learning analytics (LA) as a fast–growing field of technology–enhanced learning (TEL) is concerned with the means to collect and analyze educational data from diverse sources (mostly from learning management systems) and produce valuable insights from the wealth of information for the institutions of higher education to help them with their decision making processes. Due to the nature of education and learning that span through multiple dimensions and generation of several data types, LA has attracted researchers' interest in the field. A plethora of research, therefore, has been conducted to define learning analytics and identify its key requirements, models, applications, challenges, and future opportunities.

In this chapter, we aimed at performing a systematic and extensive literature review considering the extant research in LA. A 4–dimensional learning analytics reference model

[3] http://www.laceproject.eu/ethics-privacy/

based on the study of [Chatti et al., 2012] was also introduced, and key learning analytics processes were categorized in 10 different processes to be used in our proposed analytics–driven framework in Chapter 5. This chapter's material is used as an application scenario in Section 4.2.1 of Chapter 4, the construction of the logical layer in Section 5.4 of Chapter 5, the physical layer's construction in Chapter 6, and the physical layer's connection to the logical layer in Section 7.1.2 of Chapter 7.

Chapter 4

Adaptive Composite Analytics Architecture

> "Act as if what you do makes a difference. It does."
>
> *William James*

4.1 Introduction

Big enterprises always seek for the most optimal data–driven analytical solutions relevant to their business situation in order to improve their business values, given the vast amount of data they own [Chen et al., 2012]. They are keen to adopt adaptive analytical techniques in data processing and insight extraction to take advantage of available business opportunities, mitigate likely future risks, and meet their current and future business objectives [Chen et al., 2012]. They need effective tools to help them transform information into insights, extract business values from those insights and act upon them to guarantee their success. Given the above–mentioned facts, the incorporation of relevant business analytics solutions is of crucial importance to big firms to help them satisfy their short– and long–term business objectives [Chen et al., 2012].

Analytics, in general, is a multidisciplinary concept that is defined as the means to collect data (several data–types) from diverse sources (historic, static, streaming), perform relevant data processing operations on them to extract meaningful patterns, trends and insights, and disseminate the outcomes to targeted stakeholders [Power, 2014; Chen et al., 2012;

Van Barneveld et al., 2012]. A considerable amount of research has been conducted in discovering effective methods in collecting, reporting, processing, comprehending and extracting insight from big data. Big data processing solutions assist enterprises with realizing what has happened in the past (analytical reports regarding the past events), and what is likely to happen in the future (likely patterns/trends occurrences and extrapolations in the future) [Kaisler et al., 2014].

To address the mentioned needs in extracting the knowledge of proper hindsight and foresight, two essential types of analytics were introduced: "descriptive analytics" and "predictive analytics" [Schölkopf et al., 2001] which are elaborated in Sections 2.1.1 and 2.1.2, respectively.

Big enterprises (possessing big data repositories) are keen to adopt proper analytical approaches to process the historic/streaming data (descriptive analytics) and extract valuable insights (predictive analytics) to act upon them and meet their business objectives. They need to get the full benefit of business opportunities and taking advantage of them having the knowledge of what has happened in the past (the outcome of descriptive analytics) and what might happen in the future (predictive analytics result). They need to adopt proper analytical approaches to transform the information into insights and then act upon them to satisfy their business objectives [Baker and Gourley, 2014; Chen et al., 2012; Kaisler et al., 2014]. Therefore, there is a gap between the extracted insights and adaptive and operational sequences of actions related to those insights to be recommended to the enterprises to meet their objectives [Barga et al., 2014].

To address this gap, "*Prescriptive Analytics*" as a new frontier in business analytics has emerged [Evans and Lindner, 2012]. It is concerned with the recommendation and guidance that generates optimal, adaptive and near real–time courses of operational actions for organizations based on their predefined constraints and objectives [Basu, 2013]. Prescriptive analytics in its essence is a predictive analytics which prescribes one or more courses of actions and shows the likely outcome or influence of each action. It is purely built based on the "what–if" scenarios. The key components of a prescriptive analytics solution are optimization (acting as the core of prescriptive analytics), simulation (such as Monte Carlo), and decision analysis (evaluation) elements. Prescriptive analytics tries to answer questions such as "What should be done?" and "Why that action should be done?". It also generates comprehensible prescriptions in terms of operational actions for the target organizations. The prescriptive solution will take a solid and actionable predictive model along with the feedback data collected from those actions and recommends decision options to help stakeholders

and decision makers to reach their desired outcomes. Some of the incorporated methods and techniques in prescriptive analytics are as follows: graph analytics, operations research, heuristics and rules engines, complex event processing, neural networks along with the following techniques and algorithms: machine learning, applied statistics, operations research, natural language processing, signal processing, pattern recognition, computer vision, image processing, speech recognition, and so forth. The ultimate goal of prescriptive analytics is to bring business value through better strategic and operational decisions to enterprises based on their objectives and constraints. Also, any prescriptive analytics solution will help improve profitability, increase customer satisfaction, mitigate likely business risks and increase business value by providing the decision makers with strategic, optimal, adaptive, time–dependent and operational recommendations.

In this chapter, an integrated and data–driven composite analytics architecture is proposed to address the analytical needs of big enterprises. The proposed technique is comprised of descriptive, predictive and prescriptive components and is capable of being applied to a wide range of real–world application scenarios with diverse sources of data to facilitate their decision–making processes. The details of the architecture, research gaps, and our contributions are elaborated more in Section 4.3.

In terms of the overall thesis road–map, this chapter addresses the first research question (Section 1.3 of Chapter 1)

Research Question 1)

How do we design an integrated and adaptive analytics architecture?

by proposing a federated analytics architecture. The architecture also constitutes a key module in the conceptual layer of the proposed analytical framework in Chapter 5, by providing the main analytical driving force of the framework.

The rest of this chapter is organized as follows: Section 4.2 mentions a couple of major real–world applications to demonstrate the significance of prescriptive analytics solutions and their application in helping the enterprises make better informed decisions. The integrated analytics architecture covering the major gaps in the field are proposed in Section 4.3. Finally, Section 4.4 will conclude the chapter by reviewing the bullet points of the research and addressing corresponding research questions in the thesis and mentions some future directions. In addition, for a comprehensive review of the extant body of research in descriptive, predictive and prescriptive analytics fields, please refer to Chapter 2.

4.2 Application Scenarios

In this section, we focus on prescriptive analytics applications in the following areas: education and a particularly emerging field named "learning analytics" in Section 4.2.1, the body of the Government and one specific case of "Government building authority" in Section 4.2.2, and finally, the popular industry–related scenario which is the "project planning" that is elaborated in Section 4.2.3.

4.2.1 Use–Case 1: Learning Analytics in Educational Institutions

Institutions of higher education own huge arrays of educational data comprising heterogeneous data (grades, several academic entities' profiles, published courses information, academic resources repository information, and so on) from diverse sources (different e–learning systems such as learning management systems, personalized learning environments, social media, etc). Recently, due to the technological advancements in the e–learning systems area and digital media, the possessed wealth of educational data has grown exponentially. Therefore, making informed decisions based on huge volumes of data becomes a critical requirement for institutions of higher education. Also, educational institutions were not historically capable of analyzing and extracting adequate insights from their growing data repositories. Thus, those institutions became interested in adopting pertinent analytical solutions to address this issue.

Given the above–mentioned facts, institutions of higher education are proper use cases to consider because they possess big educational data and are interested in hiring relevant analytical approaches to extract insights from the wealth of data they own, make proper decisions and act upon them to enhance their learning processes and meet their pedagogical objectives.

"Learning analytics" (LA) as a growing area of technology–enhanced learning (TEL) has emerged to address the mentioned gaps. It is mainly focused on the means to collect educational data from diverse sources, build accurate predictive models based on the acquired data items, and make operational and optimal decisions based on the produced predictions to enhance the quality of learning. For further information regarding the learning analytics, its requirements, models, applications, and challenges, please refer to Chapter 3. A sample learning analytics environment and its key components is depicted in Figure 4.1.

Each one of the three analytical approaches' tasks can be elaborated as follows.

[1] http://epubgeneration.weebly.com/learning-analytics.html

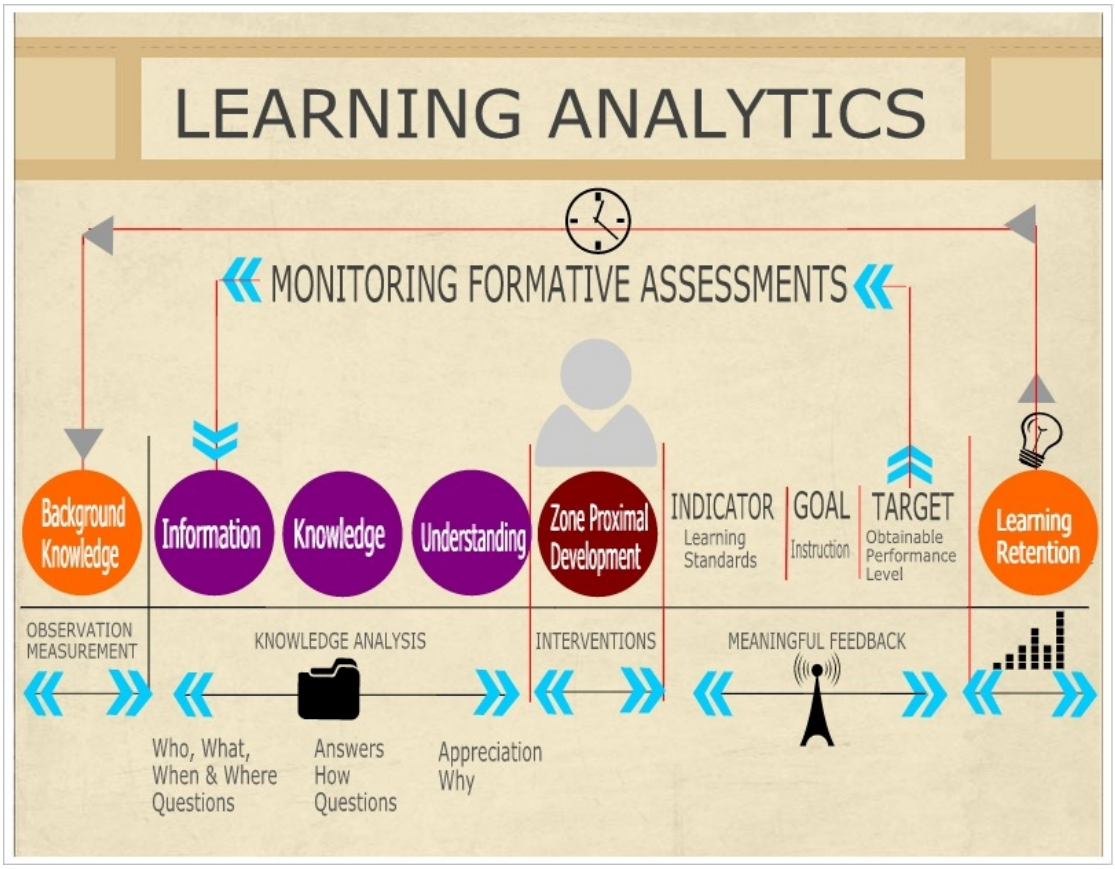


Figure 4.1: Learning Analytics Application Scenario[1]

- *Descriptive Analytics* — collects different types of educational data from diverse sources which are results of learners' interactions with the e–learning systems such as learning management systems (LMSs), reduces those data elements, aggregates and unifies them into one standard format to be used by other analytical approaches, and applies different statistical techniques on the unified data to extract meaningful information and to generate analytical reports to help institutions of higher education understand what has happened in the past and why. The analytical reports can be focused on student level (portraying their academic history and performance throughout a specific course or the academic program), instructor level (providing their teaching performance based on students' feedback and their academic performance), or the educational institution level (reports on allocated resources, financial and fiscal demonstrations, admission and retention rates). Both descriptive and diagnostic analytics were taken into consideration in the mentioned processes.

- *Predictive Analytics* — like descriptive analytics, predictive analytics can be applied to several levels and disciplines. It takes the unified data from the descriptive component and builds relevant predictive models incorporating specific machine learning algorithms to extract valuable insights regarding students' likely behavioral and academic patterns and extrapolated trends of their successes/failures (such as students at risk of failure) in the future. Predictive analytics can also forecast the future students' admission or retention (or attrition) rates as well as likely academic staff recruitments (instructors, tutors, researchers). Furthermore, some future opportunities and risks can be predicted as well (such as putting more focus on generating online learning materials and running more online courses compared to the traditional on–campus arrangements).

- *Prescriptive Analytics* — by taking into consideration the unified data from descriptive part and multiple trends and extrapolations from the predictive analytics unit, prescriptive analytics can help the institutions of higher education to make more informed and adaptive pedagogical and academic decisions in terms of operational and optimal recommendations. These suggestions can be administering specific intervention actions towards at–risk students, notify lecturers on clustered learners' performance in their given assessments throughout the semester, or providing the vision for the educational institutions to refine their admission policies and financial goals to adapt to their objectives, and the list goes on.

4.2.2 Use–Case 2: Government Building Authority

A hypothetical governmental authority is responsible for passing regulations on safety, livability, and sustainability of the built environment by enforcing the building and plumbing industries. The authority runs periodic and regular audits and test routines to ensure that the licensed building organizations meet their high standards. Those organizations who do not follow the enforced regulations are considered as "none–compliance" units. The "none–compliance" term is generally defined in terms of some delays and other factors including delays in building permits, delays in certificate returns, and delays in levy payments. It can be defined as subjective (from "compliance" to "none–compliance") or objective (from 0 to 5 for instance) measurement values. A risk factor can be defined as a likelihood of being non–compliance. The governmental building authority is a proper case study due to the real data they have and the real–world issues that the proposed prescriptive analytics framework

Figure 4.2: Government Building Authority Application Scenario

would help them to resolve. The objective of the authority can be distinguishing the future likely noncompliance organizations given their past historic data and providing a proper set of actions regarding those organizations. The actions can be defined in terms of triage support which constitutes further investigations, more inspections, thorough audit processes and some disciplinary actions (from recording the number of complaints towards a particular organization to registration or license cancellation of non–compliance units). Figure 4.2 represents such application scenario.

Each of the three analytical approaches can be elaborated as follows in this scenario:

- *Descriptive Analytics* — different types of data from diverse sources is unified to be utilized across the organization as the one source of truth (integration and augmentation processes). The unified data can be correlated with additional data sources such as geospatial, social networks and so forth. It also is scalable across several disciplines (like electrical, plumbing, etc.). The combination of descriptive and diagnostic analytics methods can help the authority to extract meaning from the input data and depict connections among them, such as tagging complaints and authority's intervention history per organization for meaning extraction and the connection between the complaints and disputes for the connections. Various analytical reports can be produced as results of this stage like data visualizations in levy payments, actions, work history, outcome status, and connections (who did what and when). In some cases, a dynamic "risk matrix" can be provided as well.

- *Predictive Analytics* — the predictive model can help the authority to calculate the likelihood of organizations to become "none–compliance", reporting on possible groups/clusters of organizations with similar behavior in the future to be "none–compliance", predicting the effect of current "risk' on future "none–compliance" and the effect of current "none–compliance" on future "none–compliance", and forecasting the future interventions' likelihood based on the historical data (according to the HR and other resource allocation limitations of the authority).

- *Prescriptive Analytics* — prescriptive analytics solution in this scenario can address different issues of the authority such as

  - "triage support" — which recommends a sequence of business actions to tackle escalated issues with none–compliance organizations.
  - "license issuance recommendation" which includes both re–issuance of licenses for previously registered organizations or issuing the licenses for the new practitioners considering their history and risk/action profiles.
  - "specific business warnings generation" — to the organizations who might cross the borderline of non–compliance.
  - a "guideline" — which is provided by the authority to the organizations to advise them in following particular sequences of actions to improve their level of compliance and prevent them from becoming none–compliance units.

4.2.3 Use–Case 3: Project Planning

Precise planning of different projects is of crucial importance nowadays. According to the increased complexity in the number of parameters affecting the whole process, enterprises put a lot of effort in proposing accurate plans which take into consideration current resources, their constraints and the organizations' objectives in a more accurate way.

We assume that a default project plan is defined for a hypothetical company to deliver specific results at time t_n. This plan can be of any form: sequential, parallel, or combined (combination of sequential and parallel forms). For the sake of simplicity, we assume that we are dealing with sequential project planning. Given a sequence of tasks (single or composite) to be performed starting from time t_1 and ending at time t_n, the enterprise defines a set of resources which should be allocated to each task. We assume that we are in the middle of the project; then, we have passed t_{i-1} steps already. It is assumed that we have reports

Figure 4.3: Project Planning Application Scenario[2]

regarding the real progress the project has achieved, its KPI (Key Performance Indicator), number of hours per person which has been dedicated to tasks, the offset between the planned tasks times and their actual completion times, types of risks that the project has faced, and resource utilization from the descriptive part. Furthermore, we assume that we have different forecasts according to different types of parameters for the project plan's future which will give the organization different completion times highlighting the ones that go beyond the predefined deadline(s). At each time, the enterprise has the capability to go back and update the prediction model(s) according to the update/feedback requests regarding the model change. By issuing the update/feedback request to the descriptive part, the company can also change the way the reports are generated in descriptive stage (including summarization over the historical data and considering other parameters in the generated reports). Figure 4.3 displays this scenario.

Each of the three analytical approaches can be elaborated as follows in this scenario:

- *Descriptive Analytics* — generates the progress reports, KPIs, number of hours spent, number and types of resources allocated, gaps and risks reports, delays reasons/causes and planned schedule's progress and issues reports.

- *Predictive Analytics* — predicts multiple project completion times according to the future risks (HR risks such as firing or leaving staff, time, budget, organization policy and management changes, framework changes), and previous or historical performance

[2] http://www.projectengineer.net/planning-the-project-schedule/

and reports.

- *Prescriptive Analytics* — prescriptive analytics solution in this scenario helps the enterprise in recommending a precise project plan as follows: given the "variables" (number of current human resources, number of hardware/software resources, etc.), "constraints" (limitation in the budget, time constraints, legal issues, and number of available resources of any kind), and Enterprises' "objectives" (maximizing the profit, minimizing the time to delivery, lead–time, and minimizing the likely risks), the prescriptive module gets the set of predicted futures from the predictive part and the set of predefined $< variables,\ constraints,\ objectives >$ tuples from the data warehouse. Then, its optimization part tries to provide an optimal plan for allocating resources in a way that the project ends before the deadline. Actually, the result of the prescriptive part could be a diverse range of actions including recommending new project plans (to utilize resources more efficiently, to allocate tasks effectively, to reduce production time properly, ...) and prescribing changes in the processes (methodology, framework, ...). The prescriptive part also incorporates a simulation unit to simulate each proposed recommendation and evaluates their effects in accordance with the whole business objective(s). In the prescriptive module, like the predictive module, the system considers updates/feedback to change the predictive model(s), input data and business rules data generator parts.

4.3 Proposed Composite Analytics Architecture

A federated composite analytics architecture is proposed in this section. Our main contributions in addressing the research gaps mentioned in Section 2.1.3 can be listed as follows:

- *Proposing an integrated analytics architecture* — which constitutes all three analytical approaches (descriptive, predictive and prescriptive) with their interrelationships. The unique way of connecting analytical components in the proposed architecture allows every data–driven analytical scenario to get benefit from its outcomes. Also, a holistic prescriptive solution definition is introduced based on the proposed system.

- *Supporting the adaptive and optimal generation of sequences of action(s)* — by incorporating certain feedback lines from each analytical approaches within the system to other components, near real–time recommendations will be generated to adapt the sys-

tem to dynamic changes in the outside world. The architecture can also provide the enterprises with comprehensible and operational outputs in terms of action sets.

To the best of our knowledge, there is no specific study taking into consideration all the mentioned analytical approaches in one framework as mentioned in Section 2.1 of Chapter 2. Apart from the references noted in Chapter 2, we compared our work with others such as [Delen and Demirkan, 2013] and [Deka, 2016] that analyzed each analytical technique separately without considering a consolidated solution incorporating all three, [Bilal et al., 2016] which mainly applied predictive and prescriptive analytics without the usage of the proper descriptive analytics component. Moreover, the technical study of [Bertsimas and Kallus, 2014] is mostly concerned with the migration from predictive to prescriptive analytics with a mathematical perspective without proposing one unified framework. This was our main motivation to propose a novel integrated prescriptive analytics technique to address the mentioned requirements in Section 4.3.

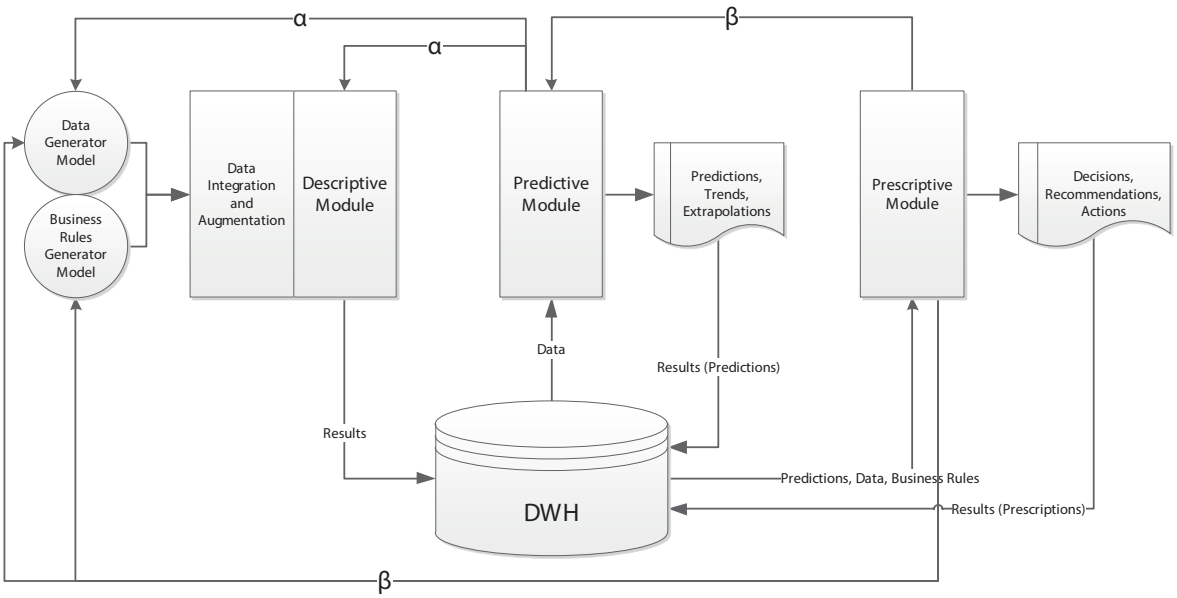
The proposed integrated architecture comprising descriptive, predictive and prescriptive approaches along with the support for diverse data types (by introducing data generator models) and one holistic data storage/retrieval component (the data warehouse) is introduced in this section. The overall architecture is illustrated in Figure 4.4. The high–level analytics architecture is depicted in Figure 4.4a and the prescriptive module's main components are displayed in Figure 4.4b. Please note that Figure 4.4 is a generic (context–agnostic) analytics architecture which is applicable to any domain. We will later apply this architecture in the context of education in Chapter 5 as a proof of concept.

To formally express the proposed architecture, we can formulate each main component as follows.

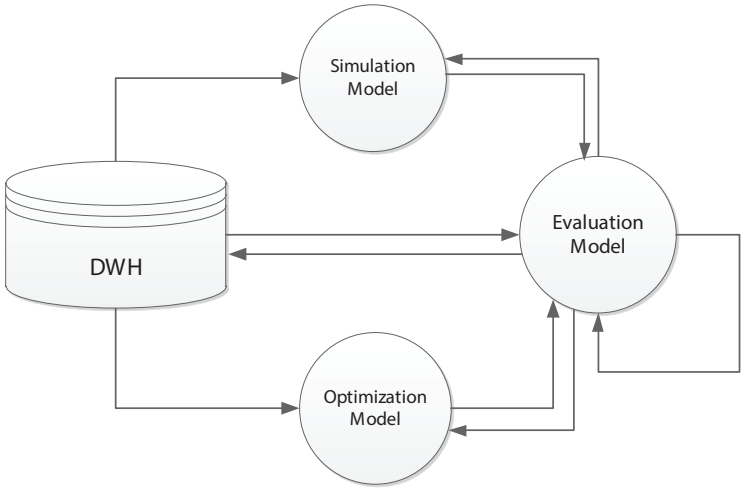
- *Descriptive Analytics result at time t_i (DsA_{t_i}) is calculated by:*

$$DsA_{t_i} = analysis(unify(data_{t_i},\ bizRules_{t_i}))$$
$$where$$
$$unify(data_{t_i},\ bizRules_{t_i}) = f\Big(coll(data_{t_i},\ bizRules_{t_i}),\ integr(data_{t_i},\ bizRules_{t_i}),$$
$$transf(data_{t_i},\ bizRules_{t_i}),\ reduce(data_{t_i},\ bizRules_{t_i})\Big)$$

where, the $data_{t_i}$ and the $bizRules_{t_i}$ correspond to the data and business rules col-



(a) The Federated Composite Analytics Architecture



(b) Key Components

Figure 4.4: The Proposed Composite Analytics Architecture: (a) The Overall Federated Analytics Architecture, and (b) The Prescriptive Module's Main Components.

lected at time t_i, the $coll(data_{t_i},\ bizRules_{t_i})$ refers to the data collection phase at time t_i, the $integr(data_{t_i},\ bizRules_{t_i})$ relates to the data integration at time t_i, the $transf(data_{t_i},\ bizRules_{t_i})$ refers to the data transformation at time t_i, and the $reduce(data_{t_i},\ bizRules_{t}$ corresponds to the data reduction at time t_i. Finally, the $analysis(unify(data_{t_i},\ bizRules_{t_i}))$ generates descriptive analytics results in terms of the analytical reports based on the unified data.

- *The Predictive Analytics result at time t_i (PdA_{t_i}) is calculated by:*

$$PdA_{t_i} = extrapolate(DsA_{t_i})$$

where, the $extrapolate(DsA_{t_i})$ function generates the data projection at time t_i, given the unified data from the descriptive analytics component.

- *The Prescriptive Analytics result at time t_i (PsA_{t_i}) is calculated by:*

$$PsA_{t_i} = action(PsA_{t_{(i-1)}},\ DsA_{t_i},\ PdA_{t_i},\ simul_{t_i},\ opt_{t_i},\ eval_{t_i})$$
$$where\ initially\ (at\ time\ zero\ -\ t_0)$$
$$PsA_{t_0} = action(\emptyset,\ DsA_{t_0},\ PdA_{t_0},\ simul_{t_0},\ opt_{t_0},\ eval_{t_0})$$

where, $simul_{t_i}$, opt_{t_i}, $eval_{t_i}$ correspond to the simulation, optimization, and evaluation components of the architecture at time t_i. Please note that at each time segment, the current PsA_i action list is a function of the previous PsA_{i-1} as well.

According to Figure 4.4, our architecture gets its input from diverse sources that generate heterogeneous data types provide them to data and business rules generator models. The input data elements will be unified within the descriptive analytics unit. Next, the predictive unit will build pertinent predictive model(s) to extrapolate the likely future trends. Finally, the prescriptive technique will act as a predictive unit incorporating certain functionalities like optimization, simulation, evaluation and intervention to produce the optimal and operational actions as the system's output(s). Key elements of the proposed architecture are listed as follows.

- *Data and Business Rules Generator Models* — all the general and domain–specific data for each enterprise will be generated in this model. It is divided into two main models:
  - Data Generator Model — any kind of static or streaming data generated within the system (historical, transactional, derivative, and so forth).
  - Business Rules Generator Model — any domain–specific data such as enterprise's context (organization's objectives, requirements, and interests) and environment data (operations, constraints, and definitions).
- *The Descriptive Analytics Module* — as mentioned earlier in Sections 4.1 and 2.1.2, this module is focused on the events happened in the past and is called the data summarization and reduction unit. All the collected data from the data and business rules generator models will be processed for cleaning, reduction, aggregation, and unification. The descriptive component is also responsible for generating analytical reports based on the fed historical data. In summary, this module will be answering questions like "what has happened?" and "what did that happen?".
- *The Predictive Analytics Module* — is mainly concerned with the future and is called the forecasting component of the architecture. Accurate predictive models are built in this module, given the unified data from the descriptive module. To best adapt with the ongoing trends of the organization and provide more accurate extrapolations, the predictive module is propagating specific feedback lines to certain architecture components (the descriptive module and the data generator model unit) that are illustrated with "α" in Figure 4.4a. The predictive module is responsible for answering the question "what will happen in the future?".
- *The Prescriptive Analytics Module* — is the main component of the proposed architecture that is concerned with the recommendation and guidance. It generates operational sequences of actions based on the unified data of the descriptive module and the forecast trends from the predictive module. The prescriptive analytics module is responsible for answering questions like "what should be done?" and "why should it be done?". The prescriptive module is comprised of certain key elements such as simulation, optimization and evaluation/feedback depicted in Figure 4.4b. To provide the enterprise with adaptive and optimal courses of actions, the module forwards feedback lines to particular components within the architecture which are labeled with "β" in Figure 4.4a. The mentioned key components in Figure 4.4b comprise certain sub–components such

as *decision–making*, *feedback* and *adaptation* which are taken into consideration inside these core elements. The *Simulation* component is responsible for answering the question "what should be done?" by running several what–if scenario simulations. By ranking the simulation component's results based on the system's pre–defined objectives and constraints, the *Optimization* component selects the most optimal result and answers the question "why should it be done?". The *Evaluation* component validates simulation and optimization units' results in the background.

- *The Holistic Data Warehouse Component* — is the data storage/retrieval unit of the architecture that deals with a diverse range of data types and interim and/or final analytical modules' outcomes. The data warehouse units depicted in Figures 4.4a and 4.4b are the same.

The overall work–flow within the architecture can be elaborated in four simple steps elaborated as follows.

Step1) Initially, the data generated in the *Data Generator Model* and the *Business Rules Generator Model* is collected and fed into the *Data Integration and Augmentation* unit. Next, different data cleaning, reduction, and unification techniques are applied to the collected data and the result is stored in the *Data Warehouse* unit. The unified data of the data generator model is utilized in descriptive, predictive and prescriptive modules; where the unified business rules data will be used in the prescriptive module only.

Step2) The *Descriptive Module* will also generate requested statistical and analytical reports over the historical data and stores them into the data warehouse module as well.

Step3) Next, the *Predictive Module* queries the unified data elements from the data warehouse unit and builds accurate predictive models by incorporating a rich set of machine learning algorithms. The module then extracts valuable patterns from the unified data and perform forecast regarding several likely future trends along with their probability scores. By utilizing the produced future trends and patterns, the enterprise can spot future opportunities and risks and consult the prescriptive unit to generate corresponding action plans accordingly. The predictive module then stores the outcomes (the predictions, trends, and extrapolations) into the data warehouse unit. In particular situations, to increase the accuracy of the predictions and getting more data

(more data means more accurate results), the predictive module sends feedback lines towards certain components within the architecture such as the data generator model (to get more data elements or collecting different types of data) or the descriptive module (to perform other unification techniques in terms of the selected final data attributes).

Step4) Finally, the *Prescriptive Module* takes the extrapolations from the predictive module, the unified data (including general data and business rules) from the descriptive module, all of which stored in the data warehouse unit and generates relevant courses of actions as output. The prescriptive module constitutes three key elements (as shown in Figure 4.4b):

1. The Simulation Unit — The simulation unit generates several scenarios and tries to answer the question "What should be done?" by generating a set of operational and actionable recommendations.
2. The Optimization Unit — The optimization unit collects the validated simulated scenarios and applies certain optimization techniques given the system's constraints and objectives to produce the best and optimal courses of action(s) which answers the question "Why should we do it?".
3. The Evaluation Unit — The evaluation unit filters the produced simulated scenarios (from the simulation unit) based on the pre–defined metrics in accordance with the mentioned objectives and business rules. The results of this part will be stored in the data warehouse unit. In addition, the evaluation unit is responsible to assess the optimization unit's outcome compatibility with the enterprise's objectives.

The final output of the system in terms of decisions (yes/no recommendations), scalar or vector values (suggested prices, amounts, fairs, etc.) or a complete production plan will be stored in the data warehouse unit.

4.3.1 An Example – Learning Analytics Application

[3] The proposed composite analytics architecture can be applied to a diverse range of analytics–driven and real–world scenarios (mentioned in Section 4.2). For example, the

[3]Please note that the "Learning Design" is outside the scope of the analytics architecture applications.

learning analytics scenario is selected as one popular analytical and data–driven use–case to utilize the proposed architecture. We also elaborate on how the introduced analytics architecture is capable of addressing most of learning analytics requirements (discussed in Chapter 3) in this section. At first, we bring one pedagogical scenario and then explain what we can get from each component of the architecture.

Imagine that the course instructor is able to perform certain analyses on week 04 to understand the students' performance so far. Descriptive analytics can help to collect students' data (previous assessment results in the course) from week 01 to week 04, unify and aggregate the data, and produce an analytical report on students' performance over the covered concepts. Predictive analytics will project the students' end of the semester performance/results in terms of being passed/failed to raise alarm to the instructor. Given the analytical report and the extrapolated end of the semester result, prescriptive analytics can produce certain interventions (such as recommending particular learning material, attending mentor/consultation sessions, or even come and visit the instructor) to promote adaptive learning. Next, by applying the same procedure on weeks 06, 08, 10, and 12, the instructor is able to keep track of each student's performance and can provide them with informed and personalized recommendations. In each round, the statistical analysis, the predictions, and the generated courses of actions may be different, due to the likely changes in each student's performance as they progress towards the final weeks of the semester.

The following lists examples of each analytical component within the educational context and learning analytics.

- *The descriptive analytics module* — is mainly focused on collecting educational data from diverse sources like students' interactions with the learning management system (LMS) or social activities which provide further digital foot–prints such as learners' activity history on massive open online courses (MOOCs). The acquired data will be cleaned and transformed into one standard and unified format that will be stored in the holistic data warehouse unit. Furthermore, relevant analytical reports/visualizations will be produced as the outcome of this process. The reports are capable of helping institutions of higher education in understanding what happened in the past (in terms of their learning processes, pedagogical trends, the progress of learners, feedback towards instructors, allocation, and utilization of academic resources, and students' retention and success rates) and why.

- *The predictive analytics module* — is mostly concerned with building predictive models

based on the retrieved unified educational data from the descriptive module (retrieved from the data warehouse unit) to extrapolate the institution's likely opportunities and risks in the future. Some examples can be:

  - Targeting at risk of failure students,
  - Forecasting the students' attrition rates based on their academic history and extrapolated patterns/trends,
  - Projecting success/failure patterns for learners, and
  - Student experience and performance forecasting.

  The extrapolated outputs of this stage will be stored in the data warehouse unit. The predictive module also sends "α" feedback lines to the descriptive module and the data generator model to ask them to provide further information required for the predictive model to generate more accurate forecasts.

- *The prescriptive analytics module* — is focused on generating optimal and operational courses of action(s) to assist institutions of higher education to enhance learners satisfaction, produce more adaptive learning materials (more personalized resources for each student given their aptitude, level of knowledge, academic history and preferences), make informed pedagogical decisions and academic policies, given predictive analytics' outputs along with the institutions' pre–defined sets of objectives and business rules (all of which retrieved from the data warehouse unit). The final produced decisions, recommendations, and courses of actions will be stored in the data warehouse unit to be disseminated to relevant targets. Similar to the predictive module, the prescriptive module sends "β" feedback lines to designated system components (the predictive module, and data generator and business rules generator models) to preserve the architecture's adaptiveness to dynamic changes in enterprises' requirements, objectives, or external events influencing the internal behavior of the system.

The more in–depth elaboration and application of the proposed composite analytics architecture in learning analytics is discussed in Chapters 5 and 6.

4.4 Summary

Different analytics techniques along with a couple of data–driven real–world application scenarios were elaborated in this chapter. The new frontier in business analytics – prescriptive

analytics – was defined and the importance of incorporating prescriptive solutions in big enterprises possessing big data was discussed as well as the key research gaps (Section 4.3) in the field.

To address the noted gaps and get the benefit of such analytics–driven approach, an integrated analytics architecture was proposed in Section 4.3 entailing key analytics (descriptive, predictive, and prescriptive). The unique way of organizing the analytical approaches within the architecture along with specific feedback lines from predictive and prescriptive modules to designated targets makes it a novel solution for prescriptive analytics scenarios. The feedback lines were designed to guarantee the adaptive and optimal generation of sequences of actions.

The following lists this chapter's contributions in addressing the first research question.

- *A federated analytics architecture* — comprising three analytics (descriptive, predictive, and prescriptive) along with their interrelationships.
- *A novel way of combining the analytical techniques* — the unique incorporation of three analytical approaches, with breaking down and connecting the building blocks of the prescriptive module is another contribution of this chapter.
- *Adaptive and optimal generation of sequences of action(s)* — by providing certain feedback lines among certain system components, near real–time recommendations will be generated to adapt the system to dynamic changes of the outside world.

The introduced architecture is incorporated as a key element of the conceptual layer in the proposed framework in Chapter 5. The connections among several components in each subsequent (logical and physical) layer with the composite analytics architecture are depicted in Chapter 5 to justify its importance and to verify its capability in addressing the analytical requirements of real–world applications. These interrelationships are elaborated in Chapters 5 (logical to conceptual) and 6 (physical to logical to conceptual).

Chapter 5

Analytics–Driven Framework for Learning Analytics

> "Most decisions are not binary, and there are usually better answers waiting to be found if you do the analysis and involve the right people."
>
> *Jamie Dimon*

5.1 Introduction

The two previous chapters presented a background of learning analytics and a proposed architecture. This chapter proposes an analytics framework for the education context, with the focus on addressing key learning analytics requirements. To provide a quick review, an architecture[1] refers to the abstract design concept of a system and its connected components [Maier et al., 2001]. It encompasses the set of principal design decisions made during its development and any subsequent evolution [Medvidovic and Taylor, 2010]. A framework[2], on the other hand, is a reusable design and building block for a system and/or subsystem [Pree, 1994]. The framework sometimes is comprised of frozen (the architecture which is fixed) and hot spots that are open for extension based on the scenario and application. To sum up, an

[1] https://www.ibm.com/developerworks/rational/library/feb06/eeles/index.html
[2] https://www.igi-global.com/dictionary/software-framework/27680

architecture is more abstract design and is fixed, while a framework is more adaptable to the situation and is extensible.

An integrated three–layered framework is proposed, comprising conceptual, logical and physical [3] components, to support separate yet connected analytical tasks. Each layer is intimately associated with and strengthens others. This chapter addresses the second and third research questions, presented in Section 1.3 of Chapter 1,

> ***Research Question 2)***
>
> ***How do we incorporate the proposed integrated analytics architecture in the context of learning analytics (proposing the analytics framework for learning analytics)?***

and

> ***Research Question 3)***
>
> ***How do we formalize learning analytics processes in the proposed framework (connecting learning analytics and prescriptive analytics components)?***

by linking the logical layer to the conceptual layer (Section 5.6).

In this chapter, we propose a framework capable of being instantiated to the context of learning analytics. The framework models learning analytics and its major functional components. It comprises three key layers: conceptual, logical, and physical as follows:

1. *The Conceptual Layer* — is the generalized, domain–agnostic view of the analytical environment and entails two inner modules: the "generic analytics–driven" module which deals with the higher–level design of the analytics environment and "composite prescriptive analytics" module which is the core analytical engine of the framework and generates intelligent courses of actions based on the institution of higher education's objectives.

2. *The Logical Layer* — is the domain–specific design which is specialized for the context of learning analytics and is adapted for LA's main requirements. Main categories of key learning analytics operational processes is elaborated in the logical module (according to the extracted LA processes in Section 3.3 of Chapter 3) and represented in business

[3] By "Physical", we mean the "Implementation" of the logical layer's constructs in concrete scenarios.

process model and notation (BPMN) specification[4]. The logical layer is related to the conceptual layer via the "IS–A" relation[5] which makes it the specialized case of the conceptual layer.

3. *The Physical Layer* — finally, the conceptual and logical layers' components are formalized, implemented and applied to one particular application scenario (a real–world use–case) in the physical layer. The details of the physical layer are discussed in Chapter 6; however, the way that it gets connected to the conceptual and logical layers is elaborated in this chapter and in Section 5.5.

Given the above–mentioned descriptions, this chapter's key contributions can be listed as follows.

1. *Proposing a generic data–driven design for the context of learning analytics which addresses key educational environment's requirements by introducing the conceptual layer.*

2. *Proposing a specialized analytics–oriented framework to implement 10 learning analytics functional components.*

3. *Implementing the conceptual and logical layers in one specific application scenario and devising a new approach named the personalized prescriptive quiz (PPQ) and forming the final layer of the framework named the physical layer.* The details of the physical layer and the PPQ's detailed algorithm and terminology along with the results are elaborated in Chapter 6.

4. *Combining the conceptual, logical, and physical layers together to form the generic learning analytics–driven framework.*

The rest of the chapter is organized as follows. The proposed analytics–driven framework is elaborated upon in Section 5.2, with the conceptual, logical and physical layers discussed in Sections 5.3, 5.4, and 5.5, respectively. Finally, we the outcomes and the conclusions in Sections 5.6 and 5.7, respectively.

[4] http://www.bpmn.org/

[5] We refer to the "Inheritance" relation (IS–A) in the object–oriented programming. https://www.oracle.com/technetwork/java/oo-140949.html#inh,
http://www.ntu.edu.sg/home/ehchua/programming/java/j3b_oopinheritancepolymorphism.html

5.2 Proposed Integrated Analytics Framework

A generic integrated analytics framework is proposed in this section to address key requirements of learning analytics systems discussed earlier (Section 3.2 of Chapter 3 and Section 2.3 of Chapter 2). The framework is a novel design concept that takes into consideration all 10 key learning analytics functional processes presented in Section 3.3. The proposed framework is illustrated in Figure 5.1 and is comprised of three key elements.

1. *The Conceptual Layer* — this component illustrates a generalized analytics layer which is elaborated in detail in section 4.3 and encompasses two sub–modules.

    1.1 "Generic Analytics–Driven Module" — which is an abstract representation of a data–driven analytics environment. It provides a generic analytical view and its constructing components capable of being instantiated to most of the analytics–oriented application scenarios. The generic module is extended by processes mentioned in Section 3.3 in the context of learning analytics. This element is described in Section 5.3.1.

    1.2 "Integrated Prescriptive Analytics Module" — which is the core analytical engine of the proposed framework. Given the institution's data and objectives, this module generates quality courses of actions to be disseminated to proper destinations. The prescriptive module is discussed in Section 5.3.2. Moreover, the detailed elaboration of the composite analytics architecture is provided in Chapter 4 and Section 4.3.

2. *The Logical Layer* — which is concerned with the representation of key learning analytics processes (the 10 specialized processes described in Section 3.3). The logical layer is elaborated in detail in Section 5.4.

3. *The Physical Layer* — is mainly focused on formalizing, implementing and applying the proposed framework on one specific and real–world application scenario. The incorporation and development of relevant algorithms and techniques take place in this layer. The physical layer is introduced in Section 5.5; its detailed elaboration, corresponding use–case, and the results justifying its validity are provided in Chapter 6.

The relationship among these layers forms the proposed holistic learning analytics framework, where the logical components extend conceptual components according to the "IS–A"

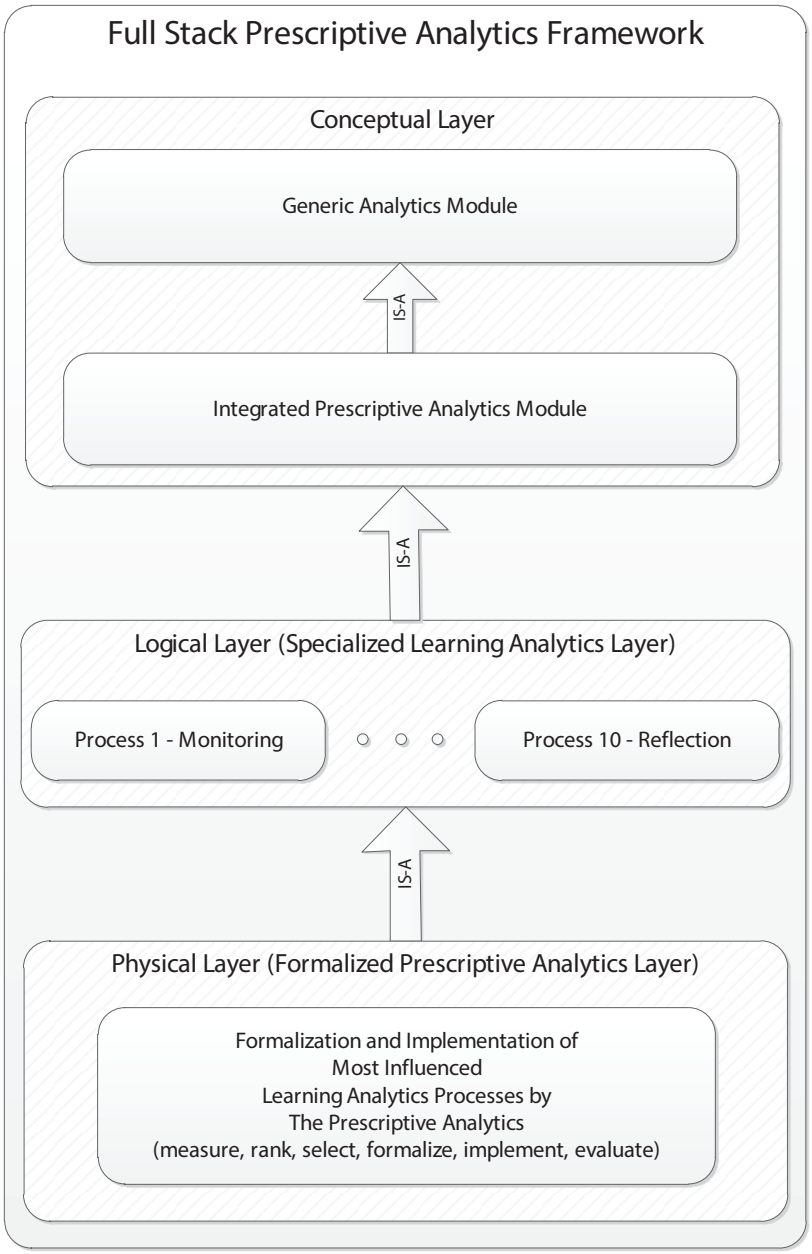


Figure 5.1: Proposed Analytics Framework

subsumption (inheritance) relationship. The same rule applies to the physical layer components interrelationships with their corresponding logical and conceptual components. The specific inheritance relation between the logical layers' processes and abstract conceptual

components justifies the names "Specialized Learning Analytics Module" for the logical module, and "Generalized Analytics Module" for the conceptual module in Figure 5.1. It means that each learning analytics process extends upper–level components (super–classes) from the conceptual module – prescriptive analytics module components in particular. The relation is elaborated in detail in Section 5.4. There are two other "IS–A" relationships in Figure 5.1 between the integrated prescriptive analytics module and the generic analytics–driven module within the conceptual module, and the physical layer's relation to the logical layer processes. Given the former relation, the introduced abstract analytics–oriented design can be instantiated to a wide range of analytical scenarios. Prescriptive analytics module, in our proposed architecture, extends this generic module and forms the generalized prescriptive analytics module. Furthermore, the latter relation between the physical and logical layers illustrates the lower–level relationship (the implementation) between the formalized and implemented algorithm of the physical layer with their corresponding LA processes in the logical layer that sets up the holistic analytics framework depicted in Figure 5.1.

5.3 Conceptual Layer

The conceptual layer is an abstract analytics–oriented module which provides actionable outcomes according to the system's pre–defined objectives. It is an adequately generic data–driven design which is able to be extended to a wide range of analytics–oriented application scenarios. According to Figure 5.1, the conceptual layer is built on top of two main modules which are related to each other using the "IS–A" relationship:

1. The "generic analytics–driven module" which is elaborated further in Section 5.3.1, and
2. The "integrated prescriptive analytics module" which is discussed briefly in Section 5.3.2. The analytics architecture, however, was elaborated in Chapter 4.

The conceptual layer constitutes the top component of the proposed framework as illustrated in Figure 5.1. The "IS–A" relation between the two inner modules demonstrates the prescriptive components as extensions of their abstract analytics–driven elements.

5.3.1 Generic Analytics Module

As depicted at the top component of the framework (Figure 5.1), the generic analytics module is referred to as the "meta" module, because it represents the high–level analytics–oriented

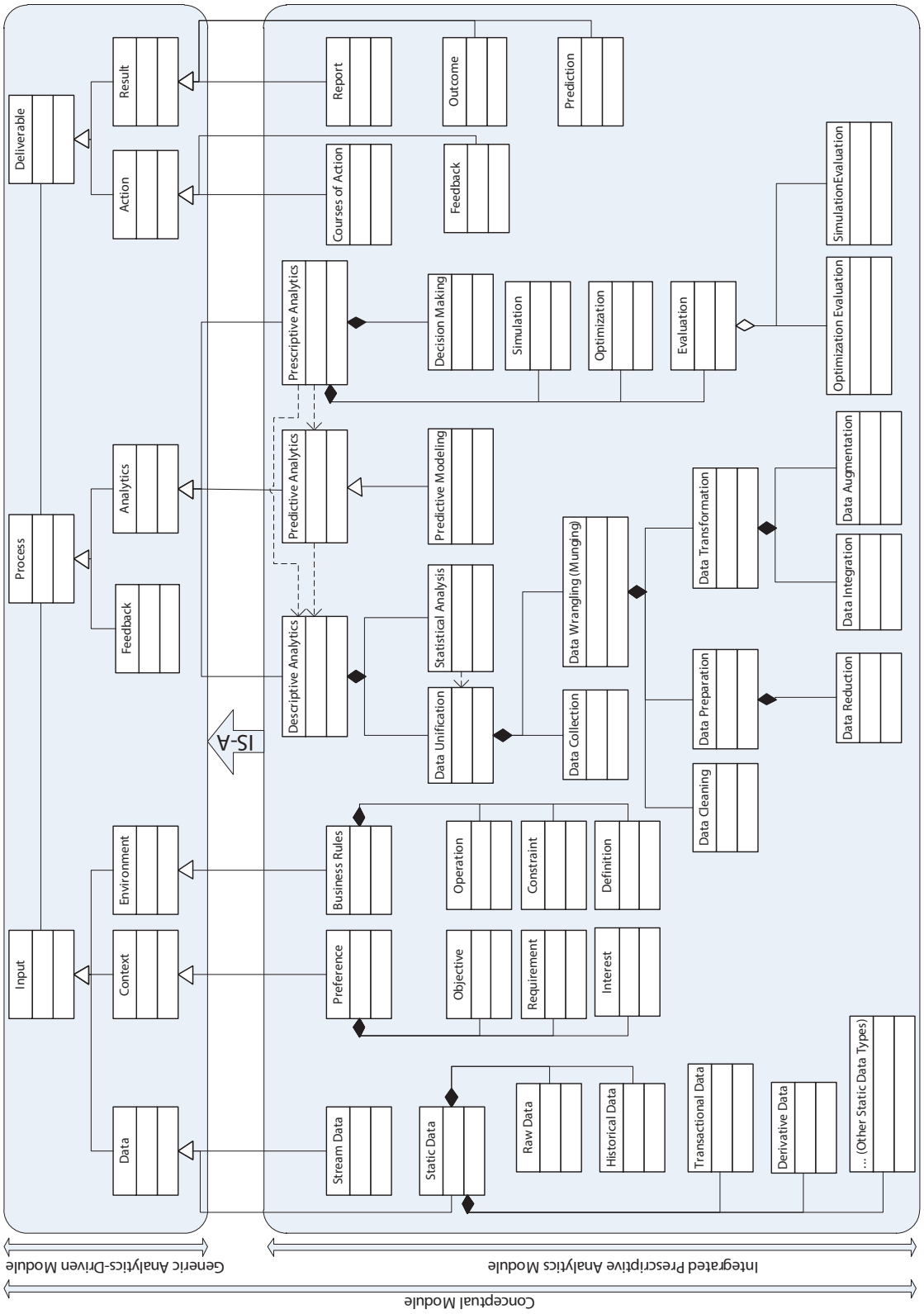


Figure 5.2: Conceptual Layer – Generalized Prescriptive Analytics Module

view encompassing key attributes of any analytics system. This module comprises three sub–components, as follows.

- *Input* — Represents all inputs to the system. This element is capable of collecting any input of any type and comprises three sub–components:
  - The "data" sub–components that capture all types of data, including the *streaming* and *static* data elements. The lecture recording is one example of the streaming data. Some examples of the static data items are raw, historic, transaction, derivative, to name a few.
  - The "context" sub–components, which describe the enterprise's main *preferences* in terms of their objectives, requirements and interests.
  - The "environment" sub–components, which take into consideration all the enterprise's *business rules* including the operations, constraints, and organizational definitions.

- *Process* — Refers to the functional elements of the system and categorizes analytics processes and activities. This is the main focus of our research. Any unit of work in the system extends the *process* component's sub–elements. The process component comprises two sub–components:
  - The "analytics" sub–component, which refers to the top–level view of analytics processes within the system, which, in turn, may be extended into three analytics elements:
    1. *Descriptive Analytics* to analyze past events and apply statistical/diagnostic approaches to generate the desired analytics reports,
    2. *Predictive Analytics* to extrapolate the likely events in the future along with their corresponding probabilistic scores, and
    3. *Prescriptive Analytics* which is concerned with the advice and personalization in terms of the sequences of operational and optimal courses of action(s), based on the enterprise's objectives.
  - The "feedback" sub–component—the key communication component among different units of the framework—to disseminate messages from one unit of work to another.

- *Deliverable* — This demonstrates all likely system outcomes and the processed results. In our design, this element has two main sub–components:
  - The "action" that refers to actionable outcomes of the system, and
  - The "result" which is any general processed results of the system.

In general, a system receives its data from the "input" component, performs computational processes on the gathered data in the "process" component, and generates/distributes the output results to the target "deliverable" component. The UML[6] class diagram[7] of the general conceptual layer is illustrated in Figure 5.2 with the generic analytics module constituting its top–level component. The input, process, and deliverable elements, along with their sub–classes, are also illustrated Figure 5.2.

5.3.2 Integrated Analytics Module

To effectively incorporate analytical methods in producing optimal courses of action(s), an integrated analytics architecture comprising descriptive, predictive and prescriptive analytics components was proposed (Section 4.3, Chapter 4). For a detailed representation of the architecture, please refer to Figure 4.4. To make it more adaptable to the context of education, each prescriptive component is considered a predictive module with aggregated functionalities such as intervention, feedback, assessment, adaptation, recommendation, and personalization. The key building blocks of the integrated analytics module are discussed in Section 4.3.

The *integrated analytics module* illustrated in general in Figure 5.1 and in particular along with its sub–classes in Figure 5.2. Figure 5.2 demonstrates the "IS–A" relationship with the generic analytics module. For example, "stream data" and "static data" elements in the prescriptive module are sub–classes (have the "IS–A" relation to) of the "data" element in the generic analytics module. The same condition applies to other prescriptive elements, such as "preference", "business rules", "descriptive analytics", "predictive analytics", "prescriptive analytics", "courses of action", and "report", to name a few.

[6] http://www.uml.org/what-is-uml.htm

[7] https://www.ibm.com/developerworks/rational/library/content/RationalEdge/sep04/bell/index.html, https://www.lucidchart.com/pages/uml-class-diagram, http://www.agilemodeling.com/artifacts/classDiagram.htm

5.4 Logical Layer

The *logical layer* represents 10 key learning analytics processes, previously mentioned in Section 3.3. As shown in Figure 5.1, the logical layer is specialized for the context of learning analytics. Each of the 10 processes is explained and their corresponding BPMN representations are illustrated in this section.

The "intervention" process is selected and illustrated in Section 5.4.4, for the following reasons.

1. Recommending intelligent and optimal actions (here, in terms of interventions) to the educational stakeholders is of crucial importance in improving the student experience, increasing the retention rates, elevating student self–esteem, and helping the institutions make informed pedagogical decisions, and

2. The proposed composite analytics architecture in Section 5.3.2 is the main driving force of the entire framework, which is at the heart of the intervention process.

The BPMN representation of all 10 learning analytics processes in the logical layer are illustrated in Sections 5.4.1 to 5.4.10. Please note that the detailed elaboration on the BPMN representation as well as its example is explained for the "intervention process" in Section 5.4.4.

Please also note that the connection between each one of the following 10 learning analytics processes with their corresponding analytics approaches (descriptive, predictive, and prescriptive) will be depicted in Table 5.1.

5.4.1 Learning Analytics Process 1 – Monitoring

The *monitoring* process definition (Section 3.3) is summarized thus: given students' previous activities and accomplishments within the LMS, the system tracks their digital footprints and provides instructors and educational institutes with students' data. Students' academic records, assessment history, LMS discussion/forum posts and comments/feedback provided by the system and instructor(s) are examples of such data. This process also helps instructors evaluate the learning process in order to improve the learning environment and student experience, as depicted in Figure 5.3, which illustrates the *monitoring process* as consisting of the *monitoring* and *unification* lanes. The monitoring lane is responsible for collecting the educational data elements from learners' interactions—activities and accomplishments—with the learning management system, along with their stored preferences and interests. The

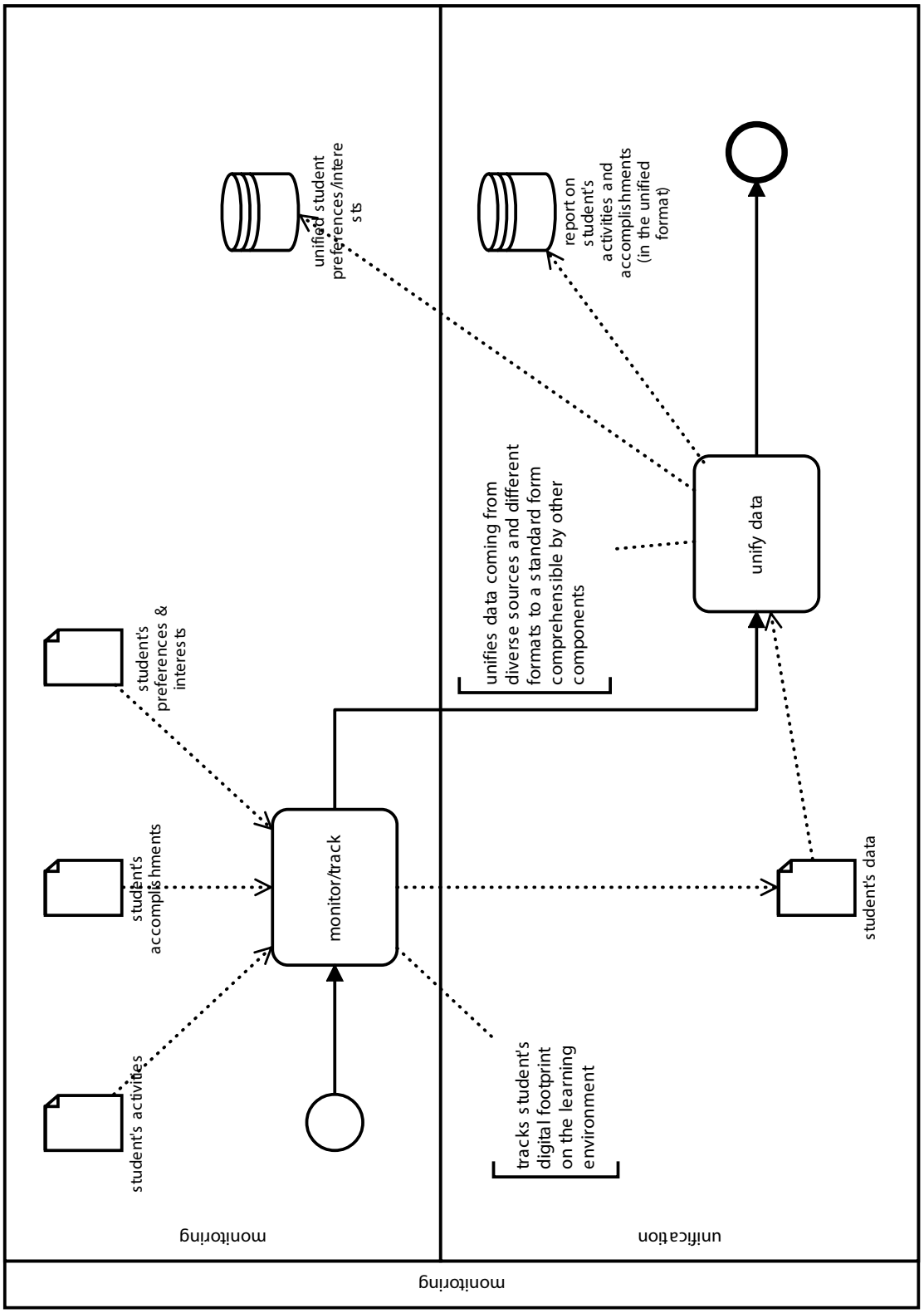


Figure 5.3: Logical Layer – The Monitoring Process [a]

[a] The "Unified Students Preferences / Interests" component is the data storage unit.

unify data component then gathers the collected data and aggregates them into one standard and unified data format, later used by all other LA processes. In fact, the result of the *monitoring process* is utilized in the *analysis*, *prediction*, *intervention*, *tutoring and mentoring*, *assessment*, *adaptation*, *reflection*, and *personalization* processes (Figure 3.2). The resulting format is stored in the *unified learners' preferences/interests* and the corresponding analytics reports are stored in the *report on learners' activities/accomplishments*, both within the holistic data warehouse unit.

5.4.2 Learning Analytics Process 2 – Analysis

The *analysis* process is elaborated in this section. Its main function (Section 3.3) can be defined as follows: analysis can help instructors identify patterns and distinguish behaviors of students and produce proper insights to help with the decision–making process. The Analysis also provides instructors with proper information to design future learning activities and enhance the student experience. This process is illustrated in Figure 5.4.

Based on Figure 5.4 illustration, the *analysis process* gets its data from the unified format from the *Monitoring* process (elaborated in Section 5.4.1), applies relevant analysis techniques on them and stores the interim results in the *analysis report* component.

5.4.3 Learning Analytics Process 3 – Prediction

The *prediction* process builds accurate predictive models to extrapolate students' future performances, behaviors, and status, given their activities within the learning management system. Instructors and institutions of higher education can properly intervene and provide students with actionable and effective suggestions and recommendations. The prediction process is illustrated in Figure 5.5.

The *prediction process* takes the unified data from the *monitoring process* and builds accurate predictive models based on the institution of higher education's objectives, and stores the results in the *predictions on learners' future performance* component.

5.4.4 Learning Analytics Process 4 – Intervention

The *intervention* process, as described in the learning analytics reference model in Section 3.3, aims at elevating learner success and improving student experience by providing them with actionable, intelligent feedback. This process is illustrated in Figure 5.6.

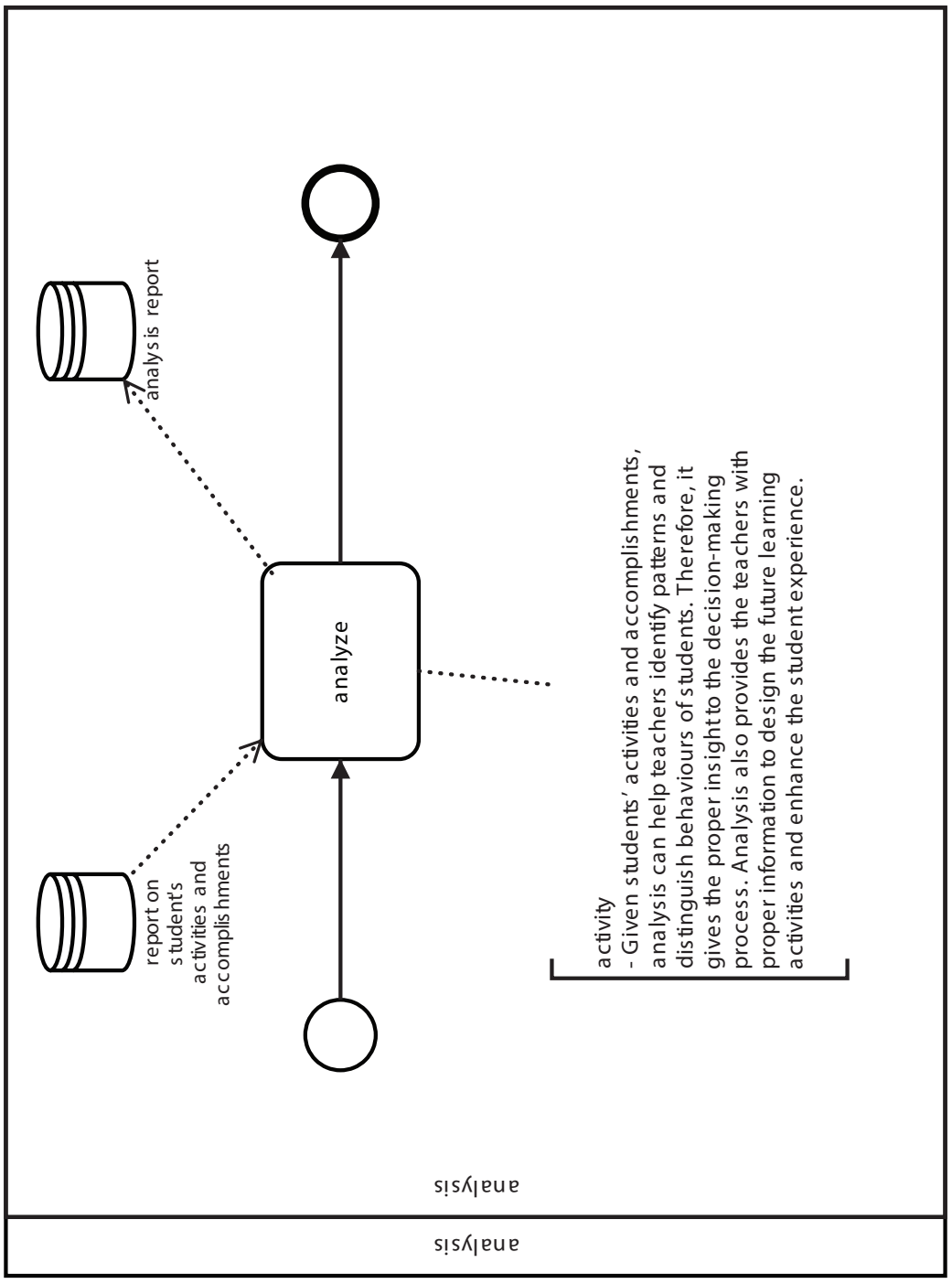


Figure 5.4: Logical Layer – The Analysis Process

Figure 5.5: Logical Layer – The Prediction Process [a]

[a] We have targeted the prediction to "performance" only. Although prediction can take into consideration other factors like strategies, emotional states, engagement, etc., those aspects are outside the scope of this particular work.

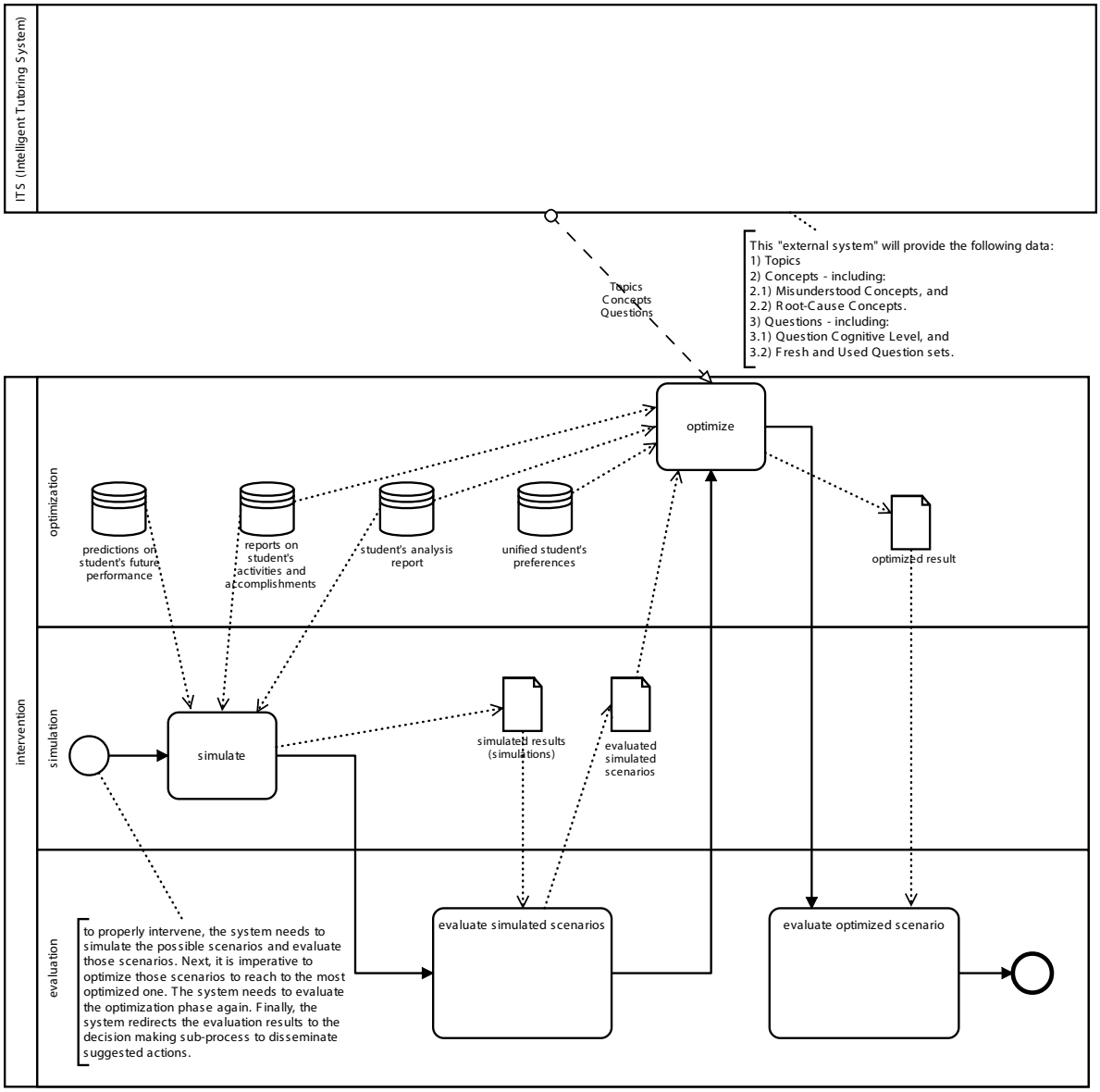


Figure 5.6: Logical Layer – The Intervention Process in BPMN – The Whole Process [8]

Figure 5.6 illustrates the core representation of this process. The entire process is put into one pool named *intervention*, which is divided into four related sub–processes: *intervention information collection*, which accumulates educational data from different sources and, in our case (according to the selected use case), the external "intelligent tutoring system (ITS)" system; *simulation*, which simulates different likely future scenarios based on the historical

data; *evaluation*, which validates the results of both simulation and optimization activities (represented as the sub–processes) in their accordance to the pre–defined objectives; and *optimization*, which provides the best result given the results generated by simulation activity. Figures 5.7 and 5.8 correspond to the expanded *evaluate simulated scenarios* and *evaluate optimized scenario* sub–components of Figure 5.6, respectively. In both cases, the evaluation activity is responsible for verifying the simulation/optimization results, makes decisions on their validity, and sends relevant feedback to pertinent stakeholders. In the final stage of the optimized scenario evaluation (the *feedback* part of the "evaluate optimized scenario" pool) in Figure 5.8, the optimal courses of actions are produced and disseminated to their pertaining stakeholders.

Intervention Process BPMN Elaboration

In this section, the BPMN representation of the intervention process and its example in the education context will be explained further. Given Figure 3.2's LA processes interrelationships, it is evident that the intervention process gets its data from "monitoring", "prediction", and "analysis" processes. This means that we have all the data required to perform an informed intervention per each student. The data regarding the students' interactions with the LMS was collected in the "monitoring" process, was processed and its analytical reports were generated in the "analysis" process, and each student's likely performance in the future (here, at the end of the semester) was extrapolated in the "prediction" process. Therefore, we are in the position to provide individual students with personalized feedback and recommendations. This is the main intention of the "intervention" process.

Next, we will explain each component of the intervention process's BPMN specification. The intervention BPMN is comprised of four lanes (of one Intervention swimlane) according to Figure 5.6 as follows.

1. *The ITS* — which is any external system providing the complementary information regarding students' performance (misunderstood concepts, root-cause misconceptions, the previous assessments attempted and the number of attempts and so forth) as well as learning content information (such as topics, concepts, learning resources, questions and so on). This information will be used in the "optimize" activity.

2. *The Intervention Pool's Simulation Lane* — The whole process starts from this lane. The "simulate" activity gets students' activities and accomplishments data (from the

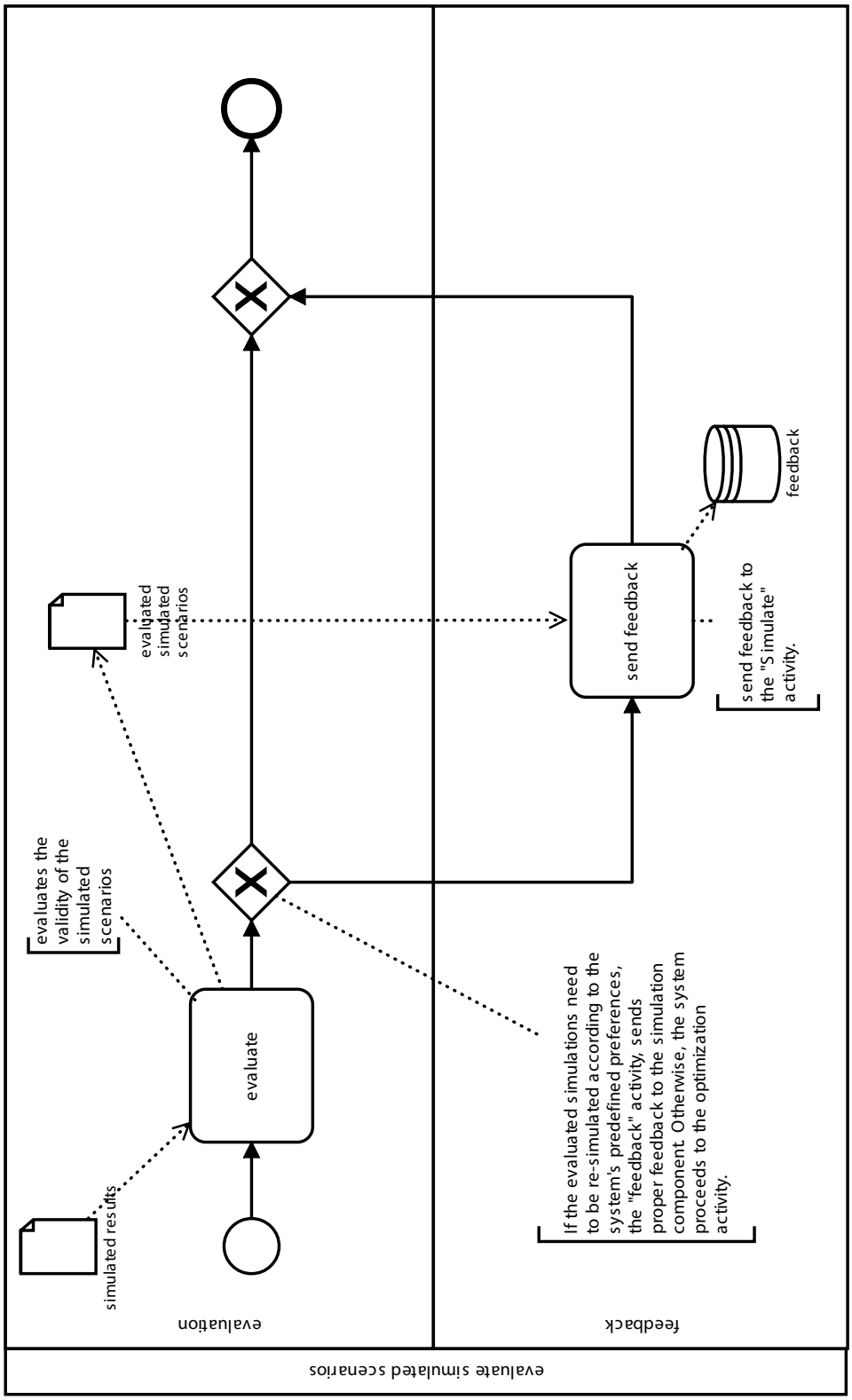


Figure 5.7: Logical Layer – The Intervention Process in BPMN – The Evaluate Simulated Scenarios Sub–Component Expansion

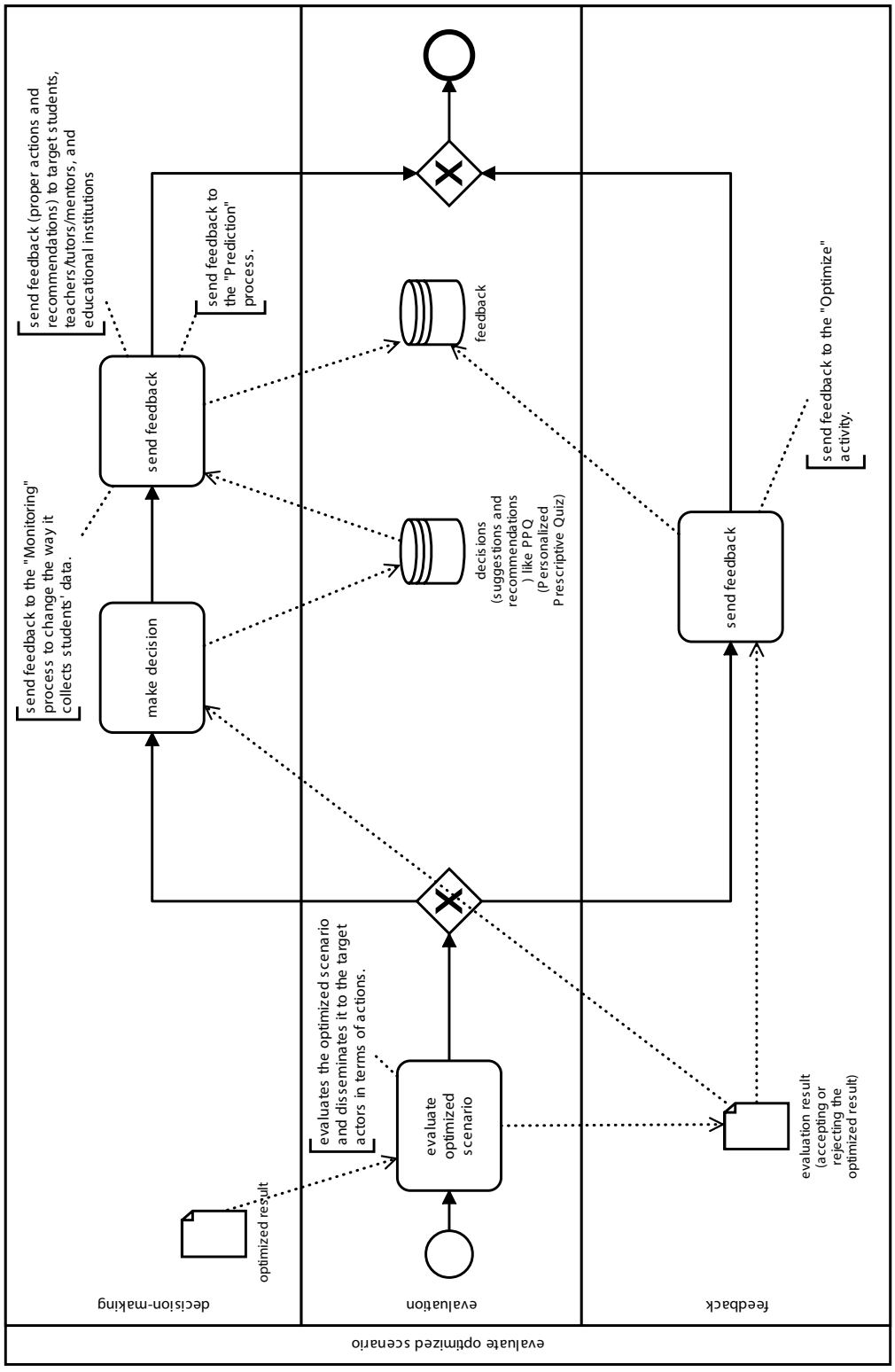


Figure 5.8: Logical Layer – The Intervention Process in BPMN – The Evaluate Optimized Scenario Sub–Component Expansion

"Monitoring" process) as well as their analysis reports (from the "Analysis" process) and performance predictions (from the "Prediction" process) as its input. The simulate activity is responsible in simulating and generating several intervention scenarios based on the student's past data and likely future predictions. The system will produce multiple simulated scenarios (stored in the "simulated results" data object artifact). But, we need to evaluate those scenarios to match their validity. Therefore, the simulated results will be fed into the "evaluate simulated scenarios" activity of the "evaluation" lane. The qualified scenarios will be stored in the "evaluated simulated scenarios" data object artifact and will be fed into the "optimize" activity of the "optimization" lane.

3. *The Intervention Pool's Optimization Lane* — The optimization lane's main task is to produce the best intervention scenario based on the fed input. The "optimize" activity gets the same data elements as the "simulate" activity, plus the filtered simulated scenarios and the student's preferences. The optimization will select the best available scenario in terms of its proximity to the student's preferences. However, this output should be evaluated prior to be disseminated to the student as an intervention feedback. Therefore, the optimized scenario (stored in the "optimized result") will be fed into the "evaluate optimized scenario" activity of the "evaluation" lane. If the selected scenario meets all the criteria, then the intervention is ready to be sent to the student and the whole process will end here (by sending the intervention through the "Feedback" process, according to Figure 3.2).

4. *The Intervention Pool's Evaluation Lane* — As mentioned in the "simulation" and the "optimization" lanes, the "evaluation" lane is responsible for make sure the generated simulated and optimized intervention scenarios are valid and align with students' preferences. The details of each one of the "evaluate simulated scenarios" and the "evaluate optimized scenario" can be found in Figures 5.7 and 5.8, respectively. The "feedback" lane in each one of the simulation and optimization evaluation swimlanes is responsible for distributing the results to the targeted actors (students, teachers, LMS, or any of the internal system components).

5.4.5 Learning Analytics Process 5 – Tutoring and Mentoring

The *tutoring and mentoring* process, described in Section 3.3, is defined thus: given the analysis results of the students' previous activities and accomplishments, tutors and mentors can provide students with personalized guidance and support. It covers a broad range of

activities including learners' orientation, new learning resources (subject–based or interest–based) suggestion, and goal achievement plans [Gidman et al., 2000; Thonus, 2002]. This process is illustrated in Figure 5.9.

The main difference between the tutor and the mentor is that[9]:

- *their definition* — the "tutor" teaches students privately, but the "mentor" is someone who provides advise.
- *their approaches* — the "tutor" is focused in helping students with their learning and has usually a single–dimensional objective, whereas in the "mentor"'s case, it is a multi–dimensional task and goes beyond just academic or learning processes (that tutors do) and is concerned with the student's life.

The discussed Tutoring and Mentoring process is comprised of *mentoring*, *tutoring*, and *feedback* lanes.

5.4.6 Learning Analytics Process 6 – Assessment

The *assessment* process, described in Section 3.3, is as follows. Based on the students' interactions with the learning management system and their preferences, the assessment process will help learners to improve their learning processes and enhance their experience using specific assessment and self–assessment techniques, to identify learners' strengths and weaknesses as they progress through the assessments. The communication with learners is represented as intelligent feedback, which is disseminated to both students and instructors/-mentors/educational institutes as well. This process is depicted in Figure 5.10.

By taking into consideration both the produced analytical reports for learners' activities and accomplishments in the learning management system, and learners' preferences and interest, the *assessment process* generates personalized assessment and self–assessment material in both formative and summative approaches. The results are disseminated to students, instructors, and even the institution of higher education with the help of the *feedback process*.

5.4.7 Learning Analytics Process 7 – Feedback

The *feedback* process, as described in Section 3.3, plays a critical role in the entire learning analytics environment, which collects useful information and disseminates them to relevant

[9] https://dlb.sa.edu.au/mentmoodle/file.php/20/Mentoringvsturoring˙article.pdf

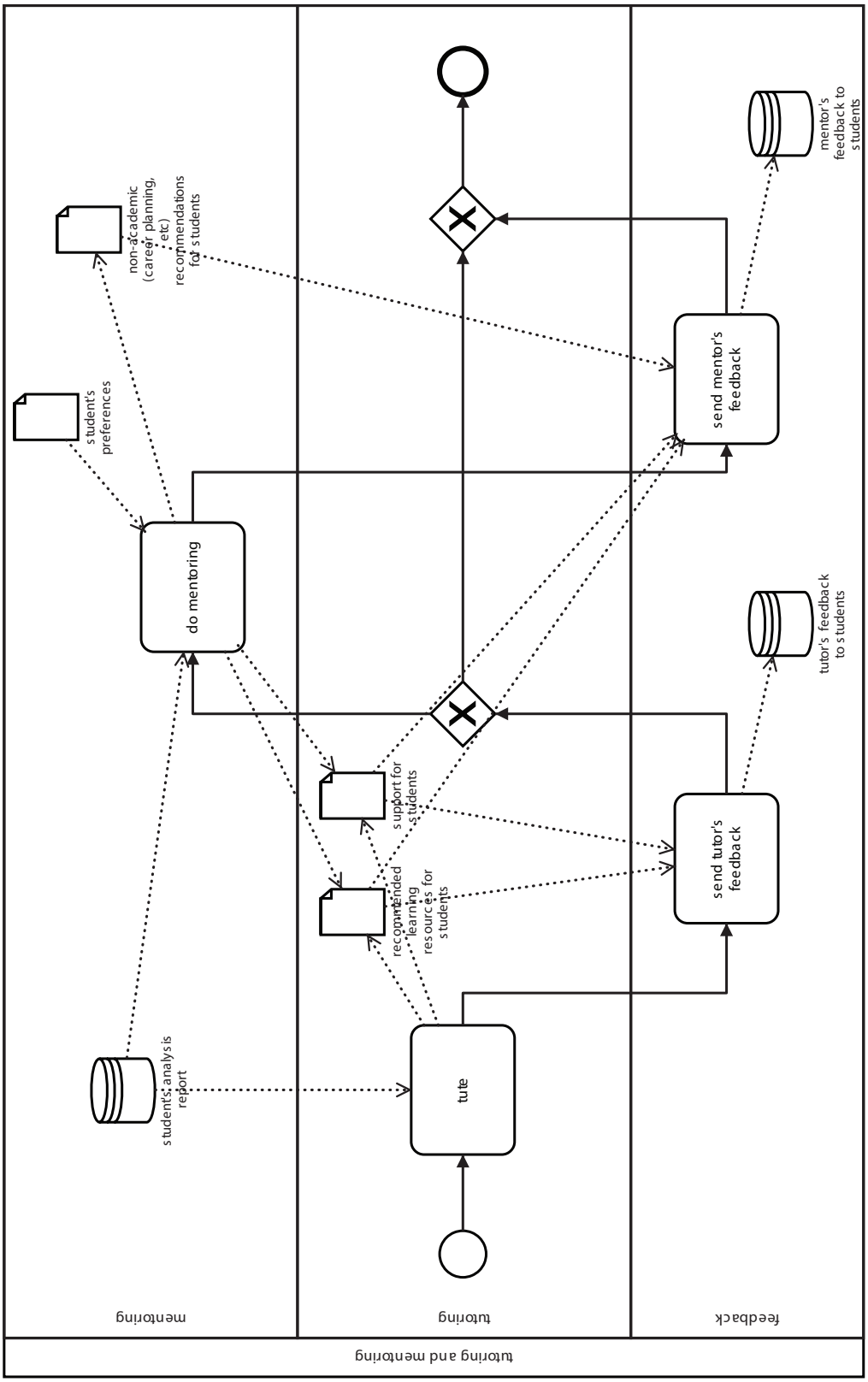


Figure 5.9: Logical Layer – The Tutoring and Mentoring Process

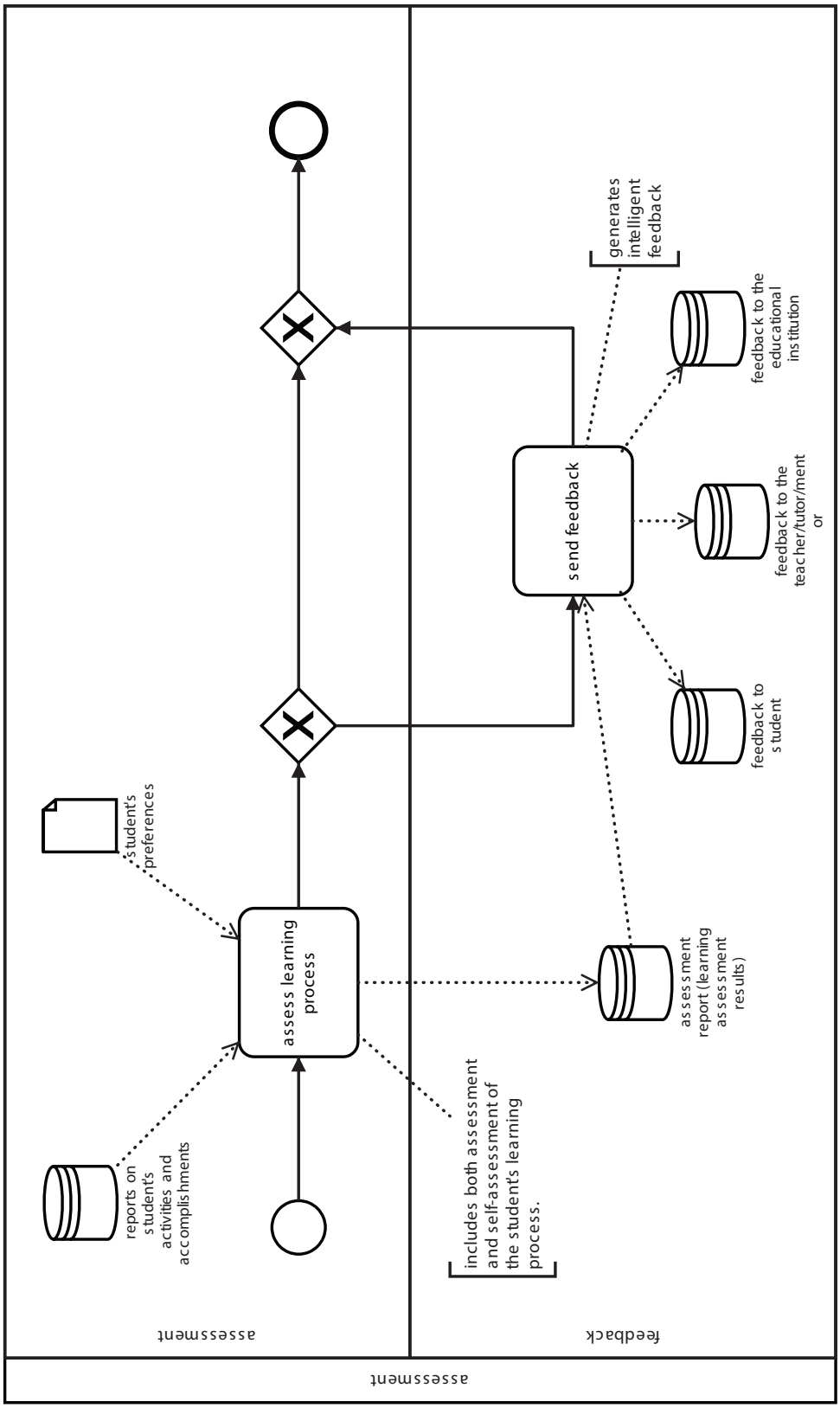


Figure 5.10: Logical Layer – The Assessment Process

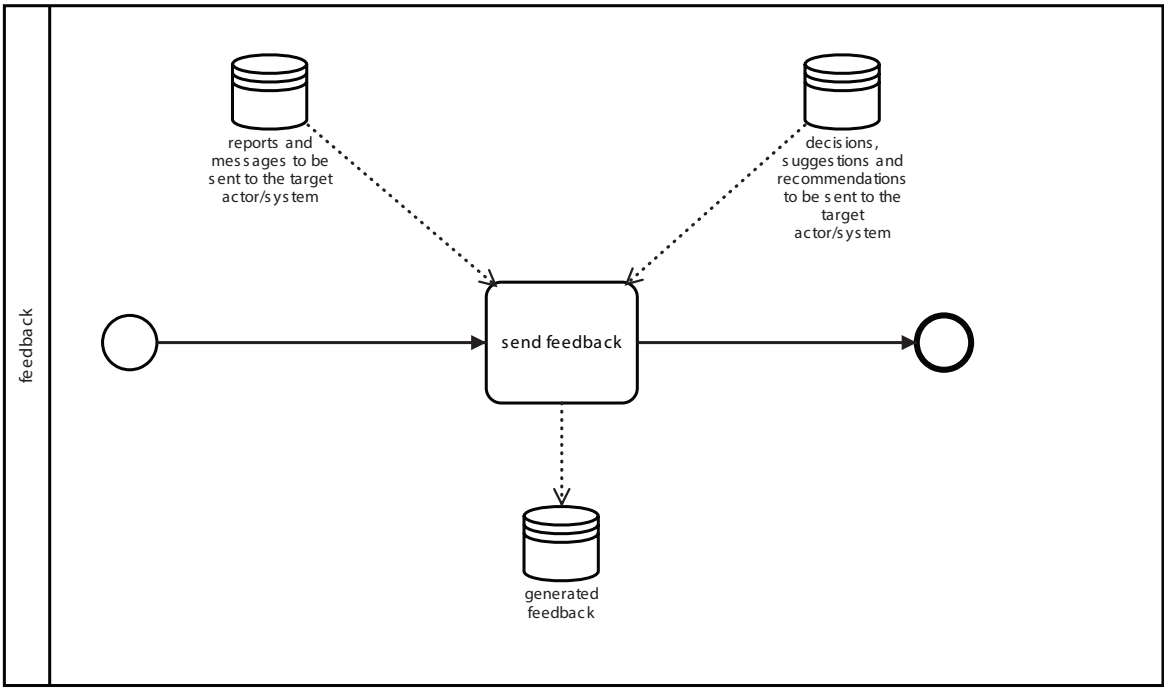


Figure 5.11: Logical Layer – The Feedback Process

stakeholders. Its main objective is to improve the overall learning process, enhance student experience, elevate learning performance, increase retention rates as well as decline the drop outs, and minimize the number of potential at–risk students. Almost all other processes communicate with the feedback process to deliver their recommended actions to their pertinent targets (mostly students). This process is illustrated in Figure 5.11.

The *feedback process* is the core communication element of the architecture, especially in the *logical layer*. All LA processes, except the *monitoring process*, utilize different means of communication through the feedback process (refer to Figure 3.2 for a clearer picture). This process gets the data to be delivered to the tagged targets as input, generates the relevant feedback format for them, and distributes them to their pertinent destinations.

5.4.8 Learning Analytics Process 8 – Adaptation

The *adaptation* process (Section 3.3), by collecting and analyzing students' data as well as their personal preferences, allows instructors and higher education institutions to activate specific and effective learner interventions. The adaptation process provides beneficial learning resources and instructional activities to students, based on their requirements, goals, and

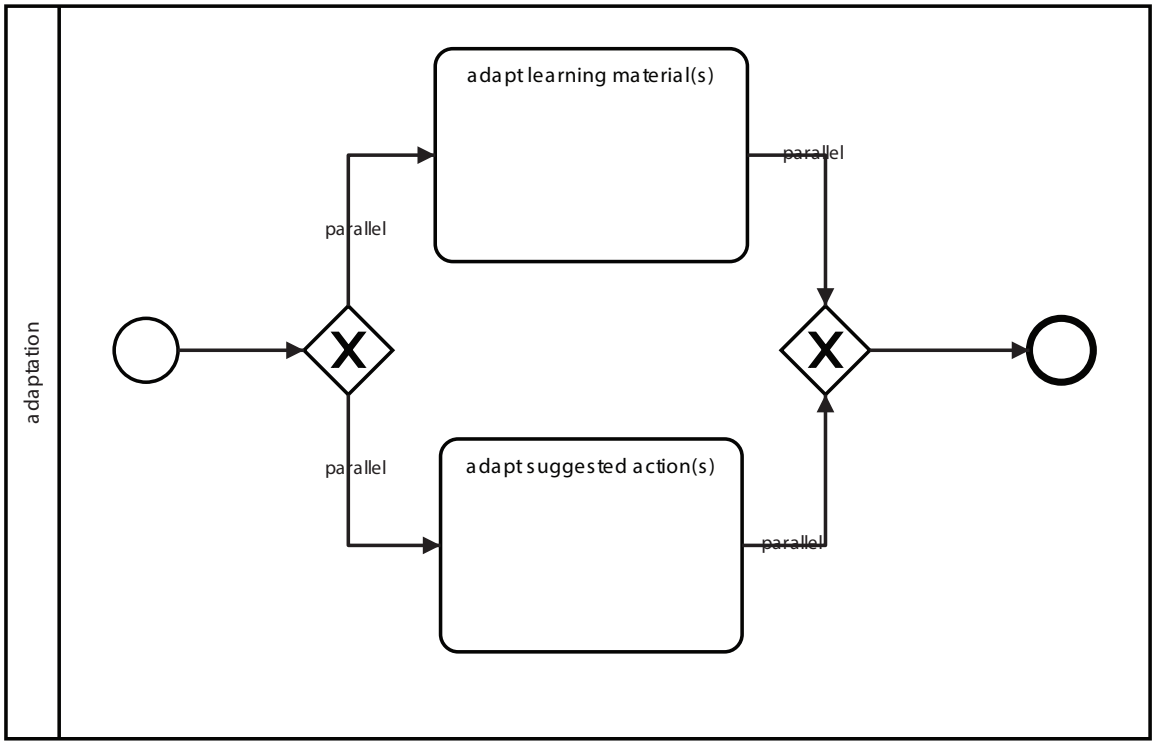


Figure 5.12: Logical Layer – The Adaptation Process – The Whole Process

interests. This process is represented in Figures 5.12.

The *adaptation process*, according to Figure 5.12, is comprised of two parallel sub–components: the *adapt learning material*, and the *adapt suggested action(s)* elements. The *adaptation* process is similar in its core functions to the *intervention process* elaborated in Section 5.4.4 in that it comprises simulation, optimization and evaluation functionalities. It is different from the *intervention process*, however, in that it incorporates the adaptation sub–component to adapt the learning material or the suggested action(s) to the students' needs.

The *adapt learning material* sub–component is illustrated in Figure 5.13, where its *evaluate simulated scenarios*, *evaluate optimized scenario*, and *adapt learning material to the students' needs* sub–components are further expanded and illustrated in Figures 5.14, 5.15, and 5.16, respectively.

Figure 5.17, on the other hand, illustrates the *adapt suggested action(s)* sub–component and its building blocks. Its *evaluate simulated scenarios* and *evaluate optimized scenario* sub–components are pretty similar to the ones depicted in Figures 5.14 and 5.15, respectively. The

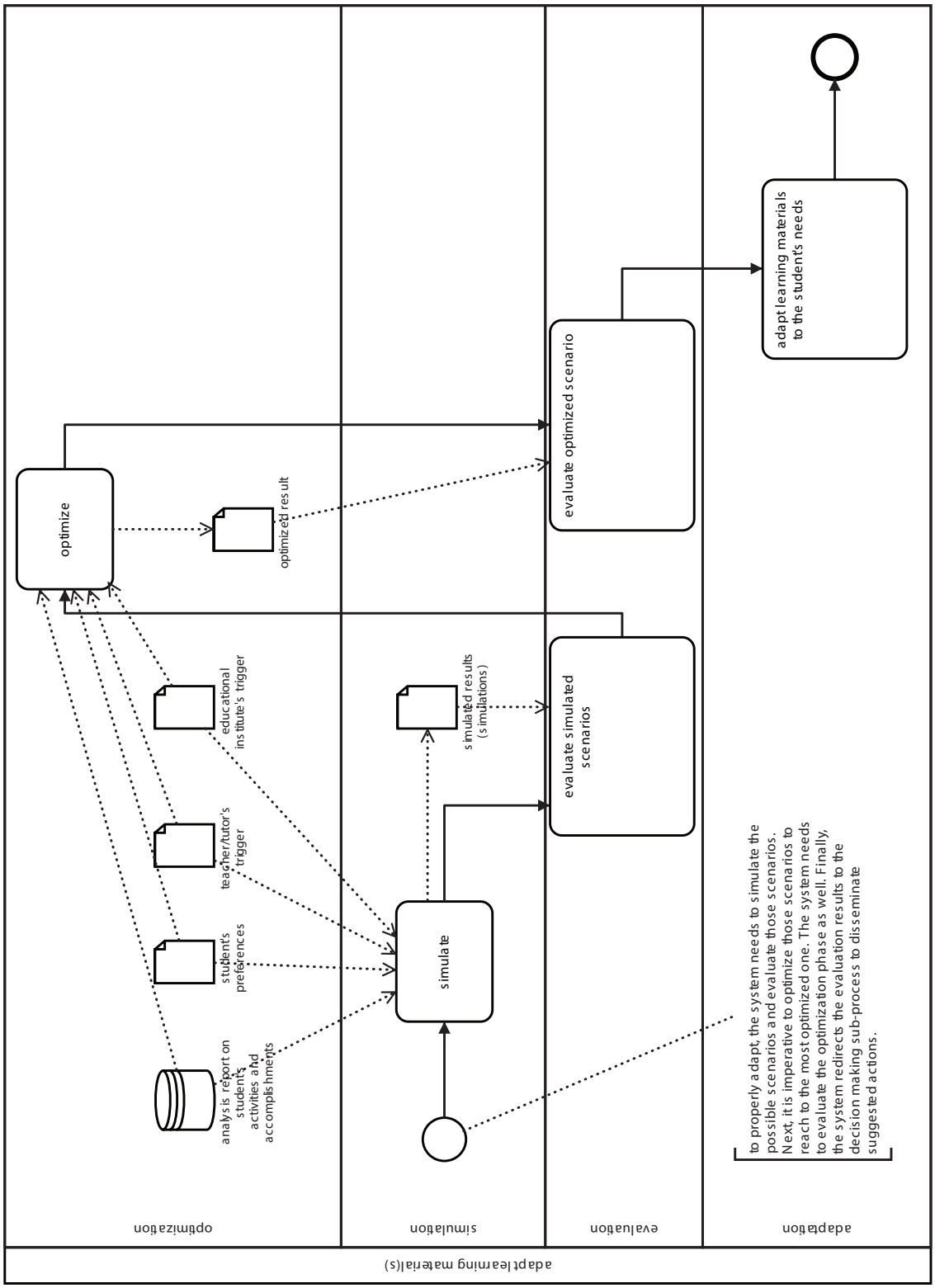


Figure 5.13: Logical Layer – The Adaptation Process – The Adapt Learning Material Sub–Component Expansion – The Big Picture

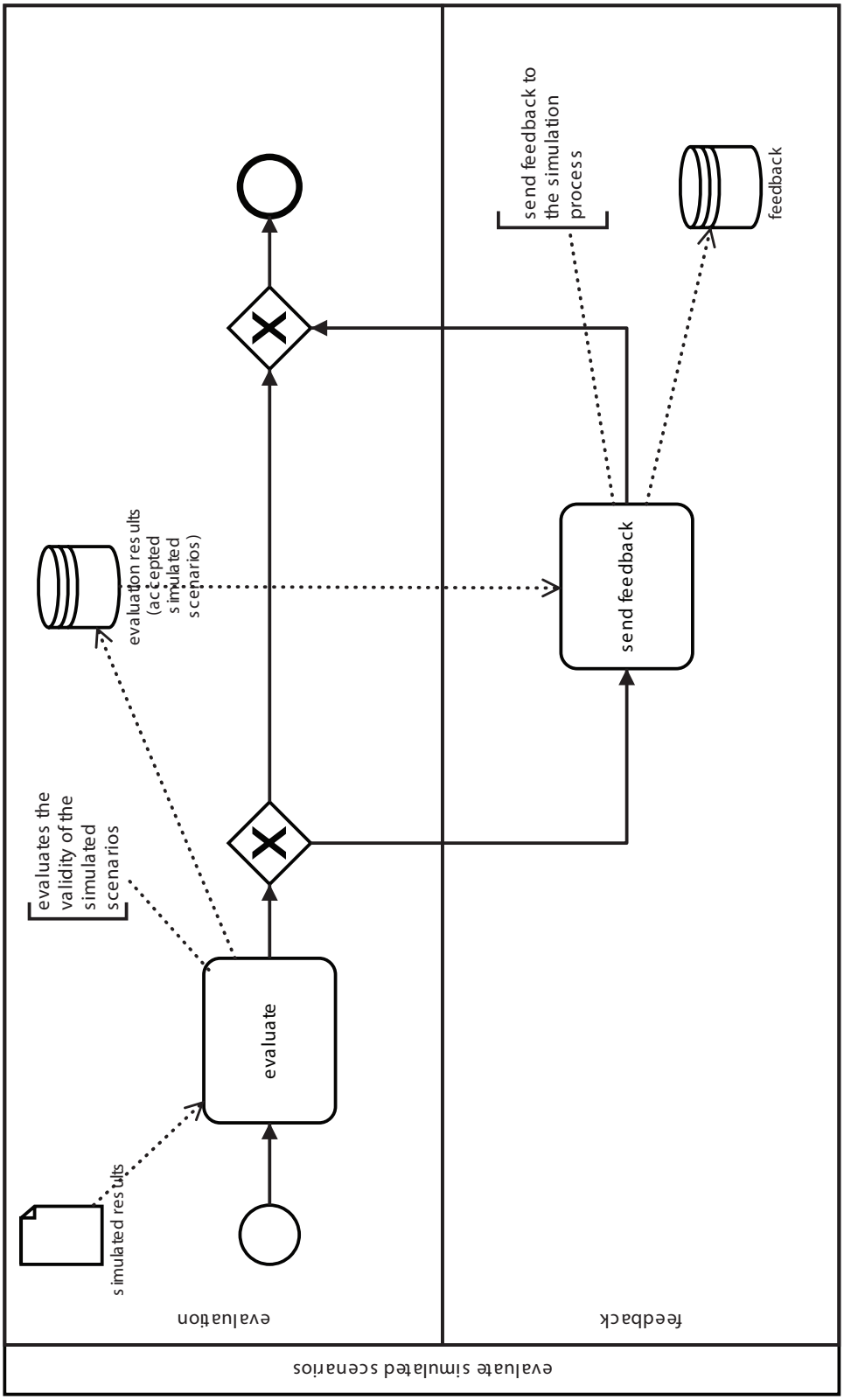


Figure 5.14: Logical Layer – The Adaptation Process – The Adapt Learning Material Sub–Component Expansion – Evaluate Simulated Scenarios Sub–Sub–Component Expansion

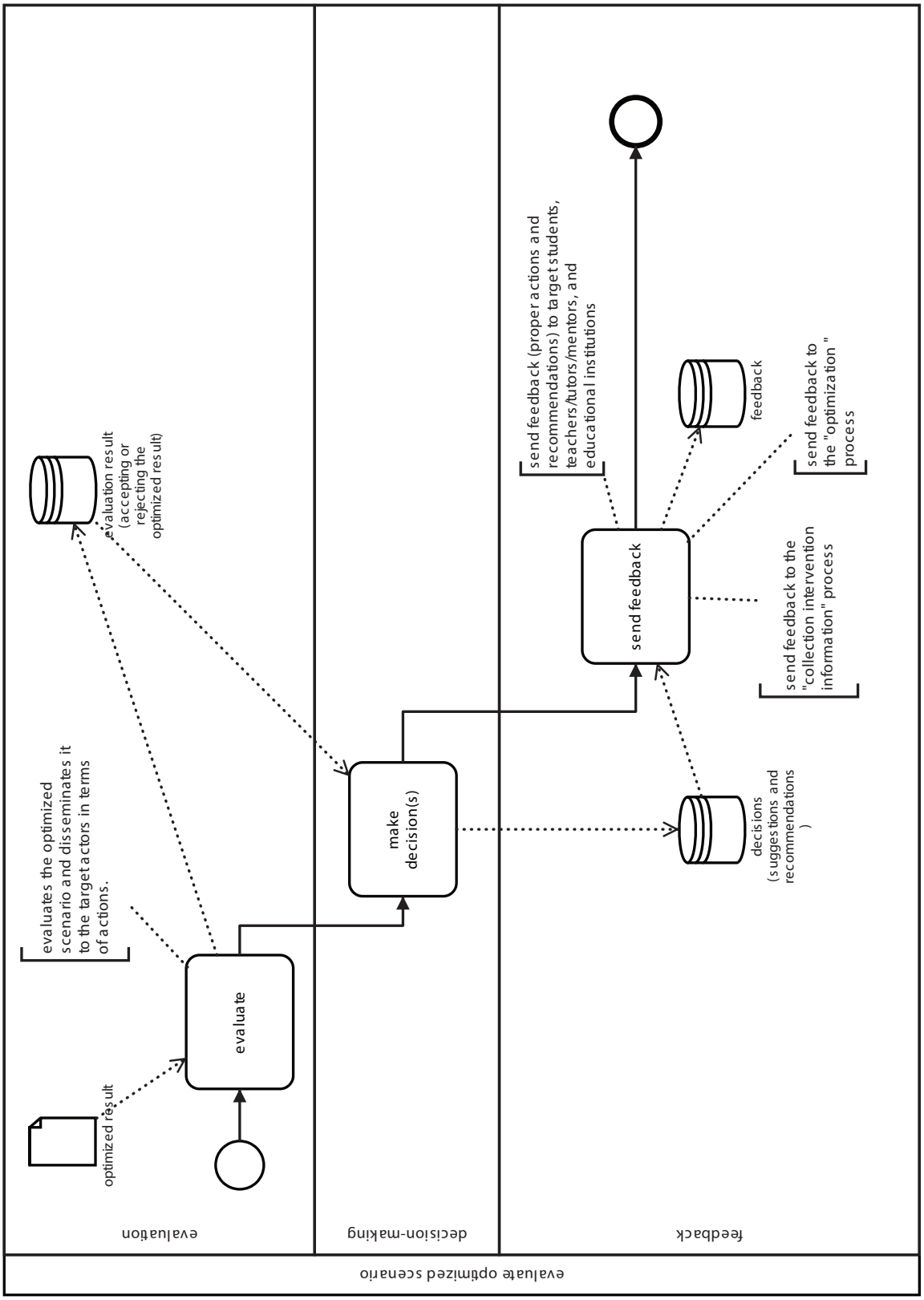


Figure 5.15: Logical Layer – The Adaptation Process – The Adapt Learning Material Sub–Component Expansion – Evaluate Optimized Scenario Sub–Sub–Component Expansion

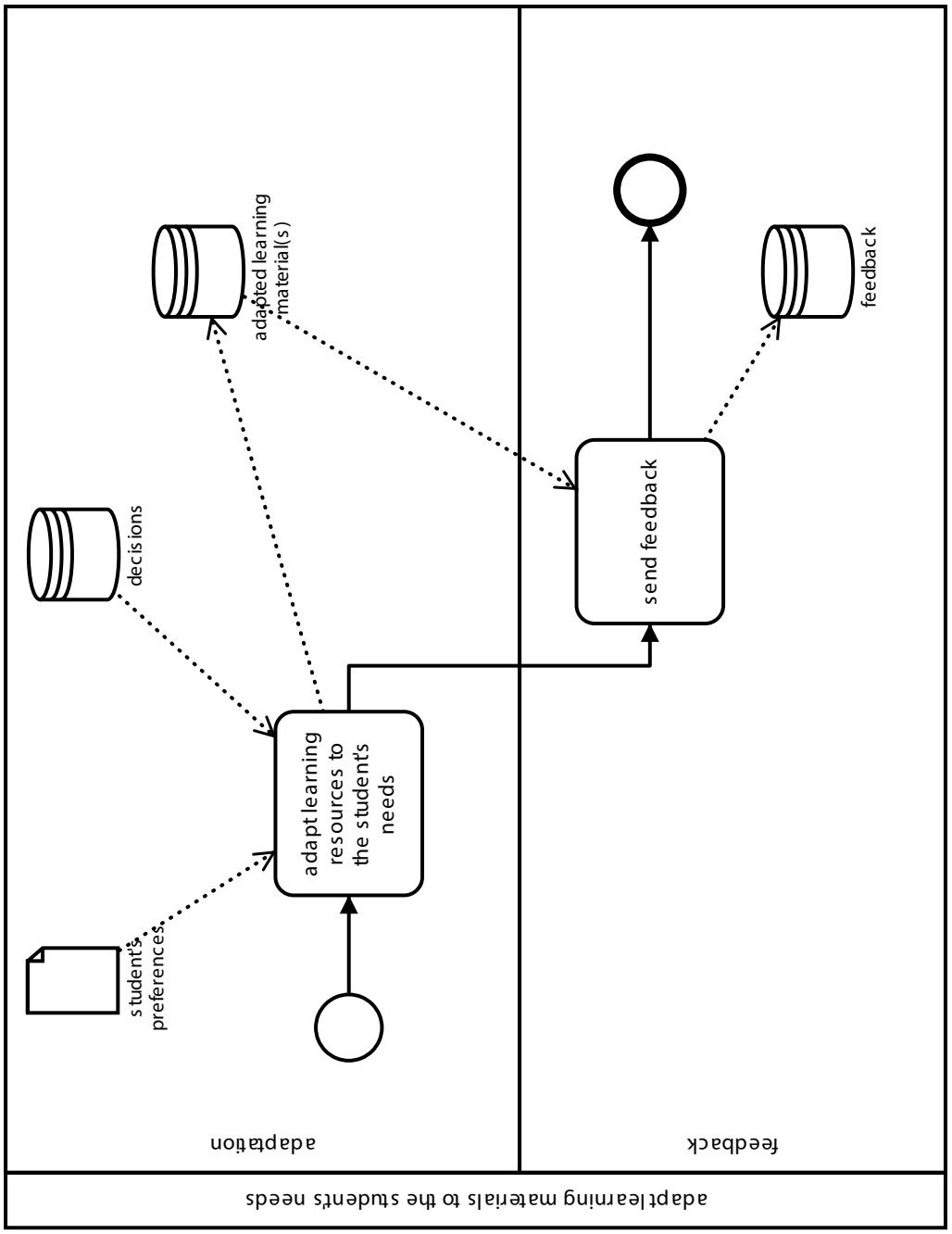


Figure 5.16: Logical Layer – The Adaptation Process – The Adapt Learning Material Sub–Component Expansion – Adapt Learning Material To The Student's Needs Sub–Sub–Component Expansion

adapt suggested action(s) sub–component, on the other hand, is represented in Figure 5.18.

5.4.9 Learning Analytics Process 9 – Personalization

The *personalization* process (Section 3.3) is depicted in Figure 5.19.

According to Figure 5.19, the *personalization process* takes the analytical reports on learners' activities and accomplishments, as well as their preferences and interests from the unified data elements produced by the *monitoring process*, adopts relevant personalization techniques to target each individual student regarding their goals, in order to help them enhance their student experience within the learning environment.

5.4.10 Learning Analytics Process 10 – Reflection

The *reflection* process (Section 3.3) collects and analyzes students' and instructors' previous activities and experiences in the learning management system, and allows the reflection process to help them compare their performances (students and instructors) and teaching approaches (instructors) with other courses, other classes, or even other educational institutions. This process is illustrated in Figure 5.20.

Based on Figure 5.20, the *reflection process* is comprised of two parallel sub–components: the *student reflection* and the *instructor reflection* elements illustrated in Figure 5.21, the detail representations of which are depicted in Figures 5.21a and 5.21b, respectively.

5.5 Physical Layer

The *physical layer*, the final component of the proposed framework in Section 5.2 and Figure 5.1, is the formalized analytical layer that implements key LA processes mentioned in the *logical layer* in Section 5.4. Given that the proposed framework is formalized and evaluated in this layer, a real–world application scenario should be selected to incorporate the framework to assess the extent to which our approach helped them meet their pedagogical objectives. As shown in Figure 5.1, the *physical layer* has the "IS–A" relation with the *logical layer*; therefore, all the implemented elements of the physical layer are specializations of their corresponding logical layer components. Several data mining techniques and machine learning algorithms are adopted in this layer.

Details of the *physical layer* are presented in Chapter 6, where a real–world use case scenario is used to illustrate the proposed framework. A personalized prescriptive quiz (PPQ) approach is introduced in Chapter 6, which quiz assists each student with identifying their

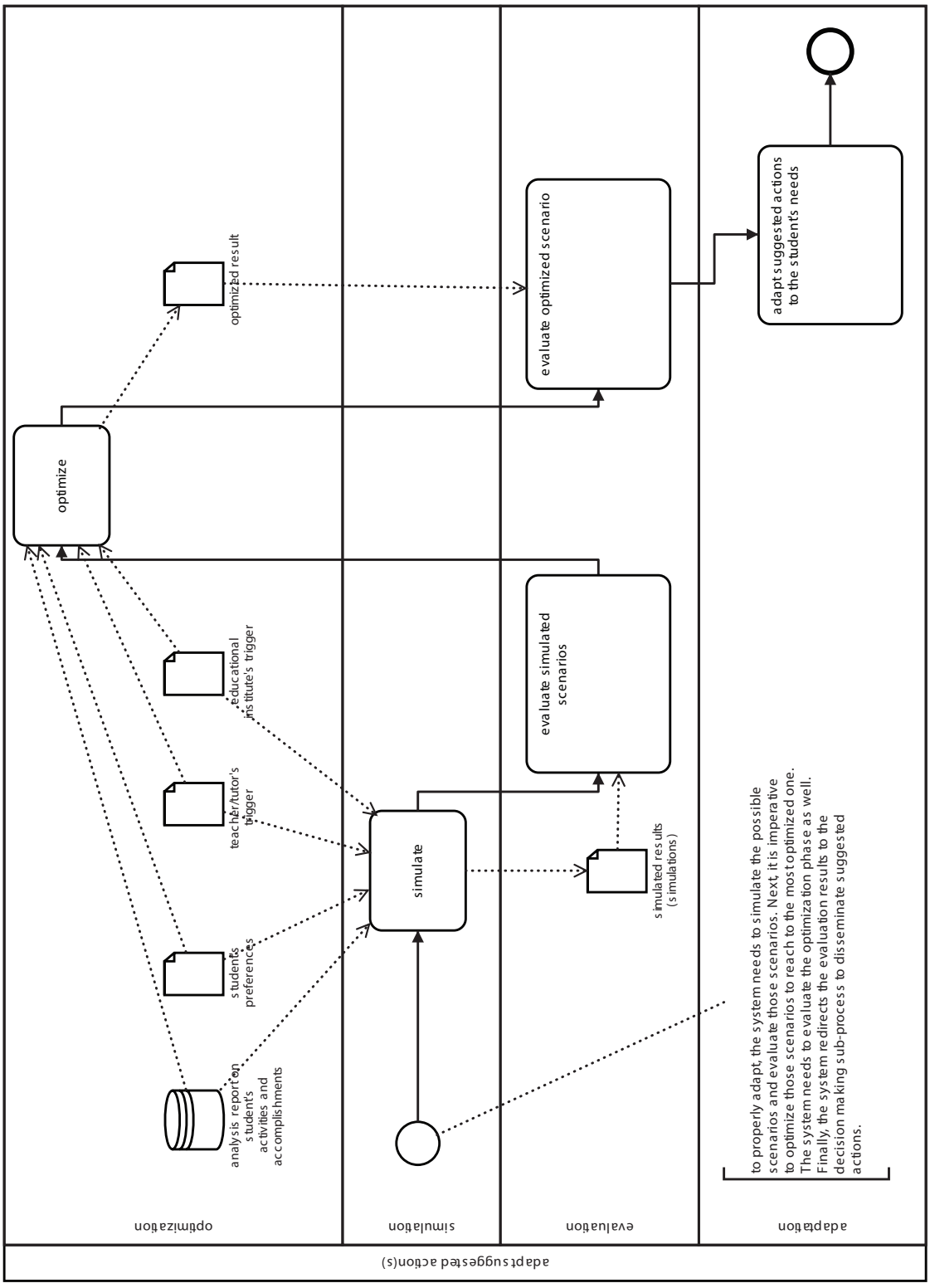


Figure 5.17: Logical Layer – The Adaptation Process – The Adapt Suggested Actions Sub–Component Expansion – The Big Picture

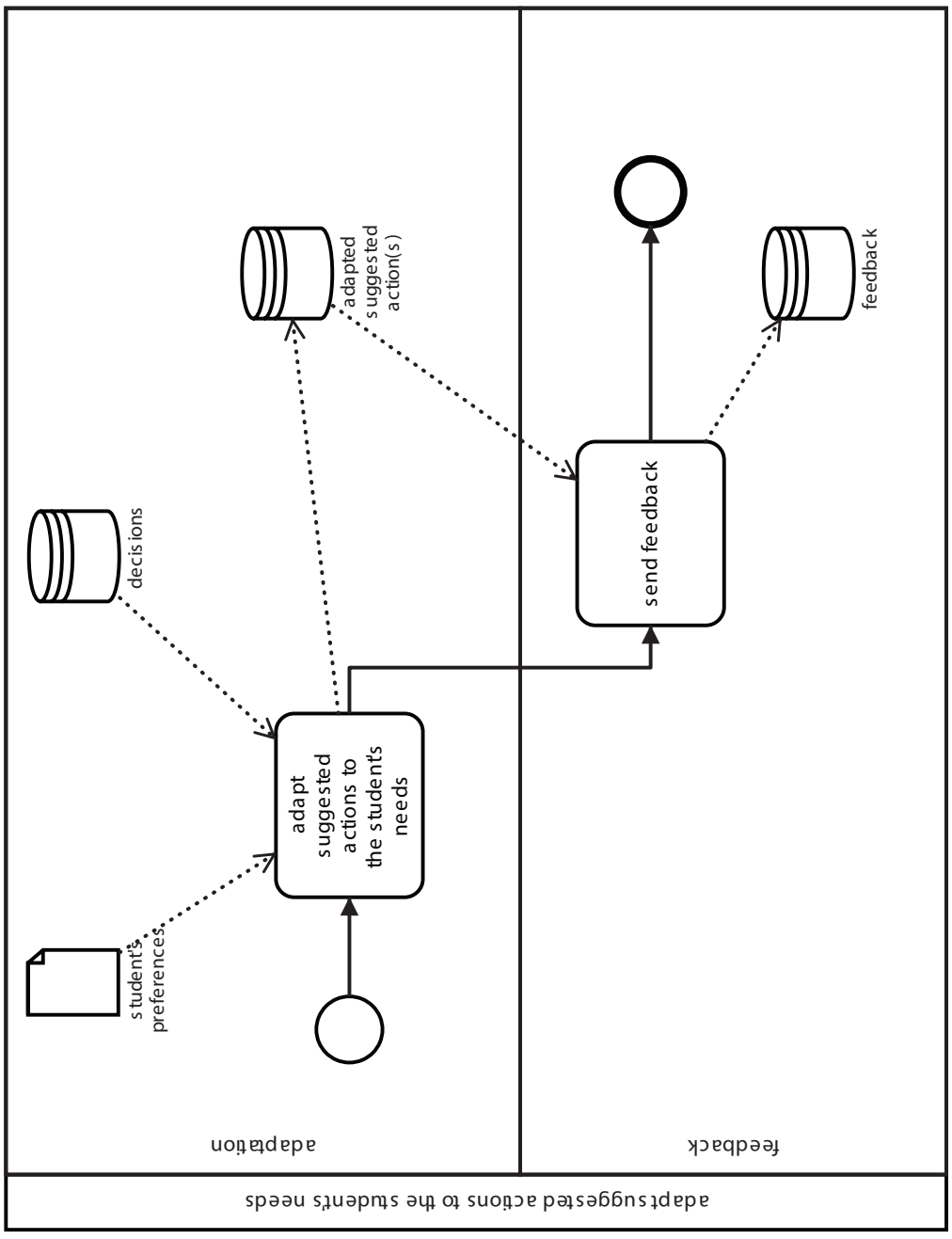


Figure 5.18: Logical Layer – The Adaptation Process – The Adapt Suggested Actions Sub–Component Expansion – Adapt Suggested Actions To The Student's Needs Sub–Sub–Component Expansion

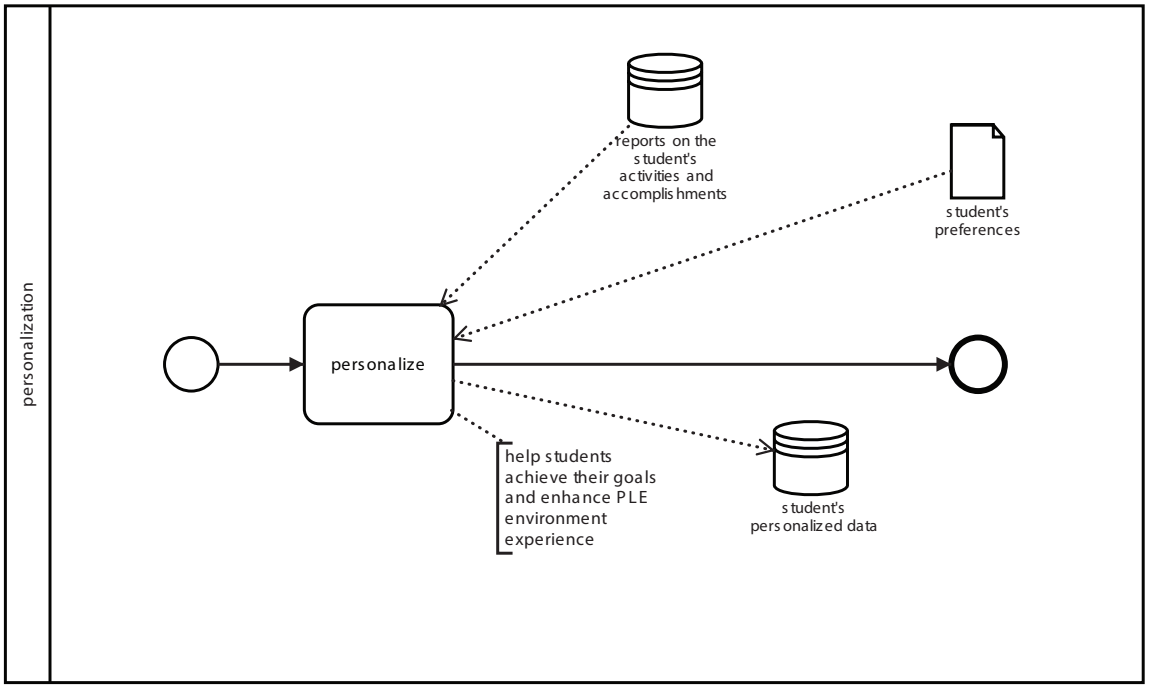


Figure 5.19: Logical Layer – The Personalization Process

misunderstood concepts throughout the semester, and provides them with dynamic and individual sets of questions to rectify their misconceptions. The PPQ algorithm is elaborated upon, along with its building blocks, the research project.Finally, the components of the proposed approach in the physical layer with their relevant logical layer elements are presented, to complete the analytics framework of Section 5.2.

5.6 Discussion

Figure 5.1 illustrates how the proposed analytics framework can be instantiated to the context of learning analytics. We mentioned conceptual (generic), logical (specialized), and physical (formalized) layers to model key requirements of a given learning analytics system. The designs of the conceptual, logical and physical layers were elaborated upon in Sections 5.3, 5.4 and 5.5, respectively. Further discussion of the physical layer appears in Chapter 6. As per the generic analytics architecture (Figure 5.1), each learning analytics process in the logical layer extends a set of conceptual layer elements. Also, three layers are related to each other using the "IS–A" relationship.

The proposed framework is one of the contributions of this work, as it is capable of being

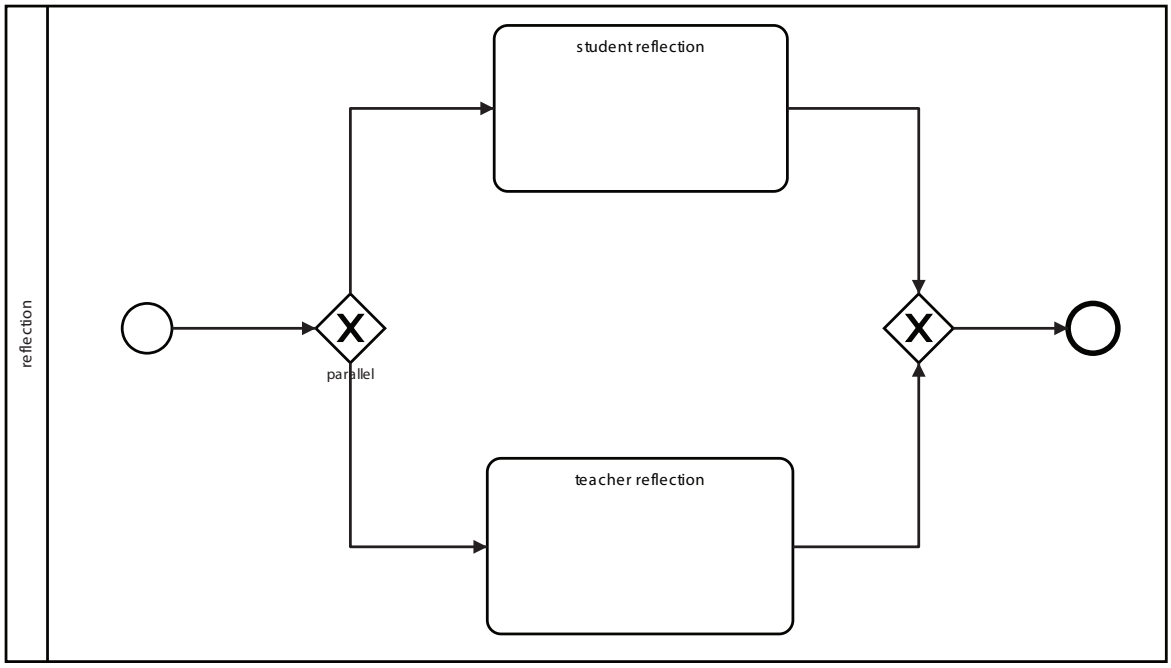
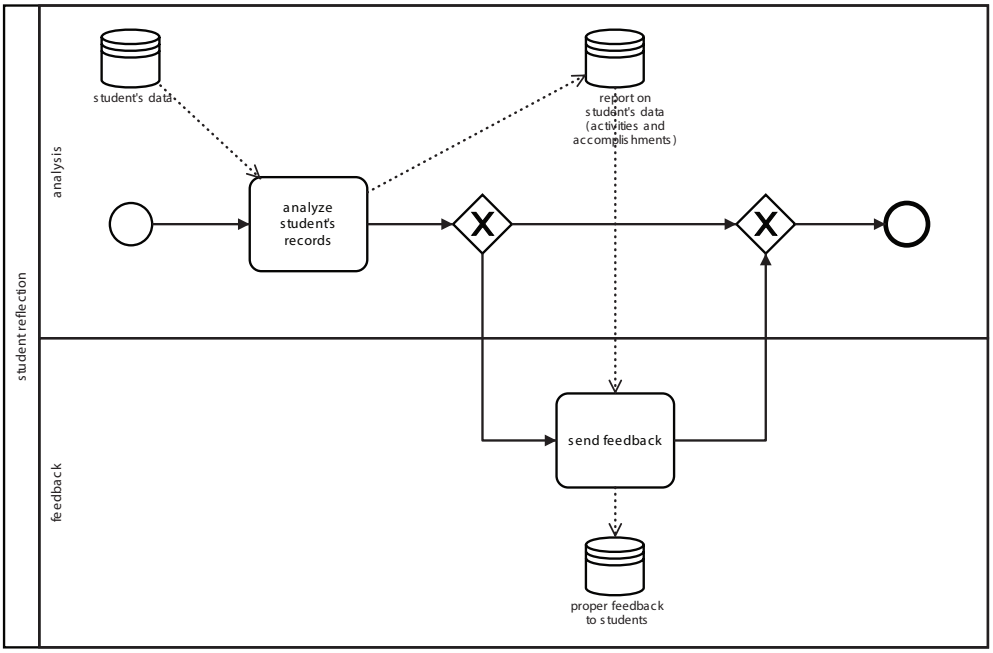


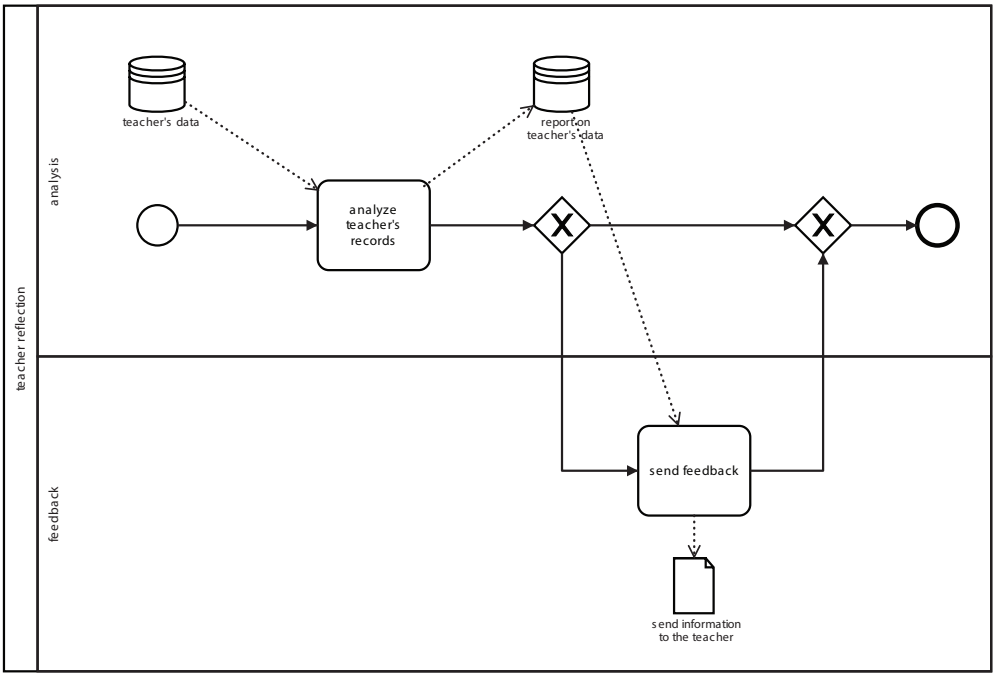
Figure 5.20: Logical Layer – The Reflection Process – The Whole Process

instantiated to the context of learning analytics. Also, the integrated analytics architecture in the conceptual layer and the unique way of combining the conceptual, logical, and physical layers altogether, form the generic learning analytics framework. We implemented the "IS–A" relation representation of all LA processes in the logical layer with their corresponding components in the conceptual layer, one–by–one. Thus, the framework is capable of covering all LA processes (4th dimension) described in Section 3.3, which means that the logical layer is able to represent all LA processes, being specialized for the context of education.

Table 5.1 depicts the mapping of learning analytics processes in the logical layer to their associated analytics components – descriptive, predictive and prescriptive – in the conceptual layer. Table 5.1 shows how the proposed federated analytics approach is capable of addressing all 10 key learning analytics functional processes. For instance, *descriptive analytics* component in the second column addresses the *monitoring, analysis, assessment*, and *reflection process* requirements. *Predictive analytics* component on the third column is responsible for the *prediction process* requirements. Finally, the *prescriptive analytics* component on the last column is concerned with satisfying the requirements of the *intervention, adaptation, personalization, reflection, tutoring and mentoring*, and *feedback processes*.



(a) The Student Reflection Sub–Component



(b) The Instructor Reflection Sub–Component

Figure 5.21: Logical Layer – The Reflection Process – The Expanded Sub–Components: (a) The Student Reflection Sub–Component, and (b) The Student Reflection Sub–Component

Table 5.1: Learning Analytics Processes' Coverage Using The Integrated Analytics Architecture in The Conceptual Layer.

Learning Analytics Processes	**Analytics Components**		
	Descriptive Analytics	*Predictive Analytics*	*Prescriptive Analytics*
Monitoring	✓		
Analysis	✓		
Prediction		✓	
Assessment	✓		
Intervention			✓
Adaptation			✓
Personalization			✓
Reflection	✓		✓
Tutoring and Mentoring			✓
Feedback			✓

As mentioned earlier, all LA processes are represented in BPMN in the logical layer and their elements are mapped into their corresponding superclasses in the conceptual layer. In particular, all logical layer processes extend specific components (superclasses) from the conceptual layer. As per section 5.4, we continue our elaboration with focusing on the *intervention process*.

The relationship of the *intervention process* to its conceptual layer components is presented in Figure 5.22, which is divided into two sections:

- *the conceptual layer* — which is corresponding to the *conceptual layer* depicted in Figure 5.1. The simplified representation of the conceptual layer is illustrated in Figure 5.22. By simplification, we meant that only conceptual layer's classes (Figure 5.2) corresponding to their related logical layer components (Figures 5.6, 5.7, and 5.8) were depicted.

- *the logical layer* — that refers to the *logical layer* represented in Figure 5.1. The logical layer section represents the intervention process illustrated in Figure 5.6.

Figure 5.22 represents several components of the conceptual and logical layers' components and their interrelationships.

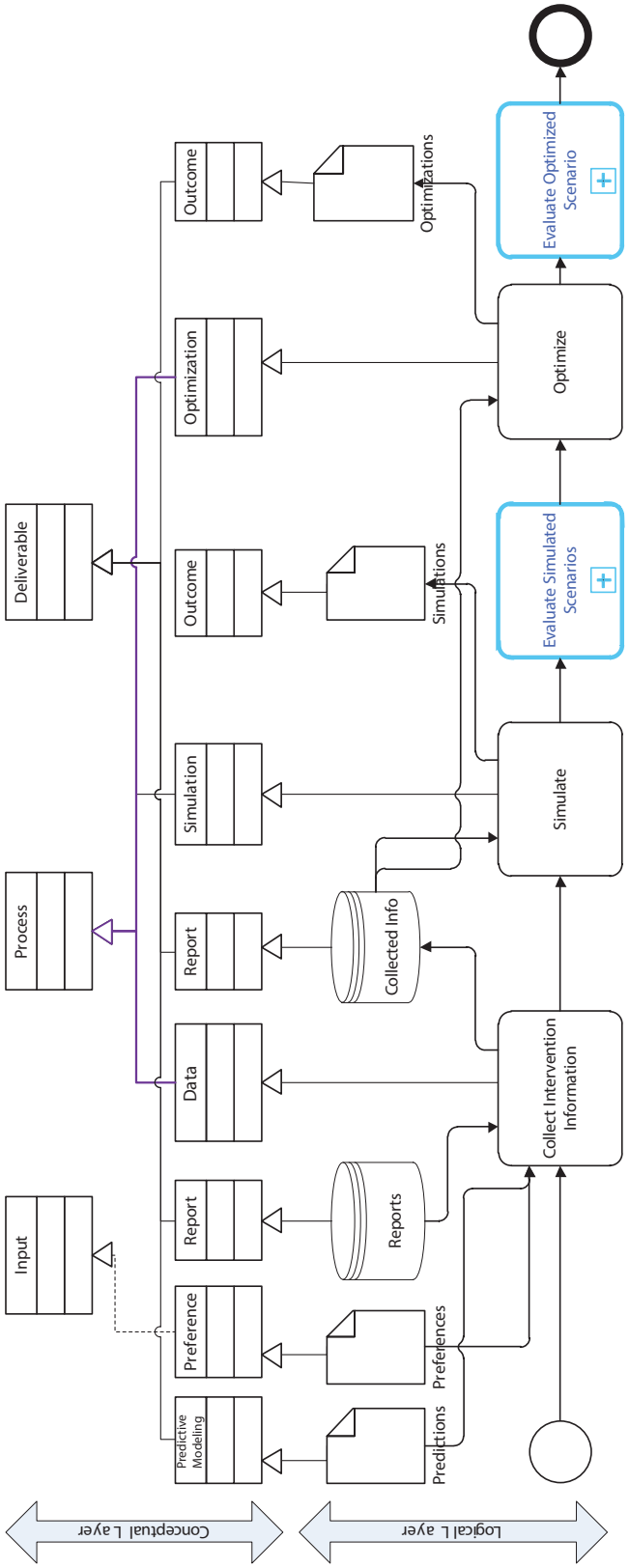


Figure 5.22: Logical Layer To Conceptual Layer Relation – The Intervention Process – The Whole Process

- *the dashed line — indicating the relation of the "preference" component in the logical layer with the "input" element of the conceptual layer,*
- *the purple line — to illustrate the relation of the "data collection" and the "optimization" components of the logical layer with the "process" element of the conceptual layer,*
- *the black line — to represent all other relations among other components of the logical layer with the "deliverable" element of the conceptual layer, and*
- *the blue sub–components — that represents the "evaluate simulated scenarios" and the "evaluate optimized scenario" components of the logical layer and are further expanded in Figures 5.23 and 5.24, respectively.*

Figures 5.23 and 5.24 illustrate the interrelationships between the conceptual and logical layers in a simplified manner. Each logical layer component is connected to their corresponding conceptual layer class. This justifies the "IS–A" relation between the conceptual layer and the logical layer.

The same rule applies to all other LA processes (elaborated in Section 5.4) and their constructing elements are mapped into their corresponding conceptual layer components. The *monitoring process*'s logical to conceptual layer connection is depicted in Figure 5.25 to give an idea about the remaining LA processes representation.

By specializing the proposed framework for LA required processes and illustrating their relation with the generic analytics–oriented architecture, we demonstrate that the proposed approach can cover all key learning analytics requirements and justify its validity by representing those processes' relations to the generic analytical architecture.

In summary, as outlined in the Section 5.1, the proposed analytics framework in Section 5.2, allows an educational institution to address the problem of developing decision support systems in a systematic way. To achieve this, our framework models and implements LA functional processes by monitoring, analysis, and assessment of learner activities in the learning system (centralized or distributed), predicting future learning trends and optimally intervening when necessary, adapting and personalizing the learning design according to learner preferences, capacity and aptitude, and giving intelligent feedback to elevate the student experience and improve the learning environment. The proposed framework can assist institutions of higher education to fulfill their analytical gaps, towards making intelligent decisions in–time, and by improving the teaching and learning quality in practice.

5.7 Summary

A generic analytics framework was proposed in this chapter to address the key issues and requirements in the context of learning analytics in Section 5.2. The framework is composed of three key layers: conceptual, logical, and physical. The *conceptual* (the abstract analytics) layer comprises a generic analytics–oriented module, and a prescriptive analytics module (comprising descriptive, predictive and prescriptive analytics components). It is the conceptual module's task to provide a given analytics–based scenario with proper courses of actions according to the pre–defined system objectives (Section 5.3). The *logical* (the specialized learning analytics) layer, on the other hand, is composed of 10 key learning analytics processes which extend the conceptual layer's components in their building blocks.

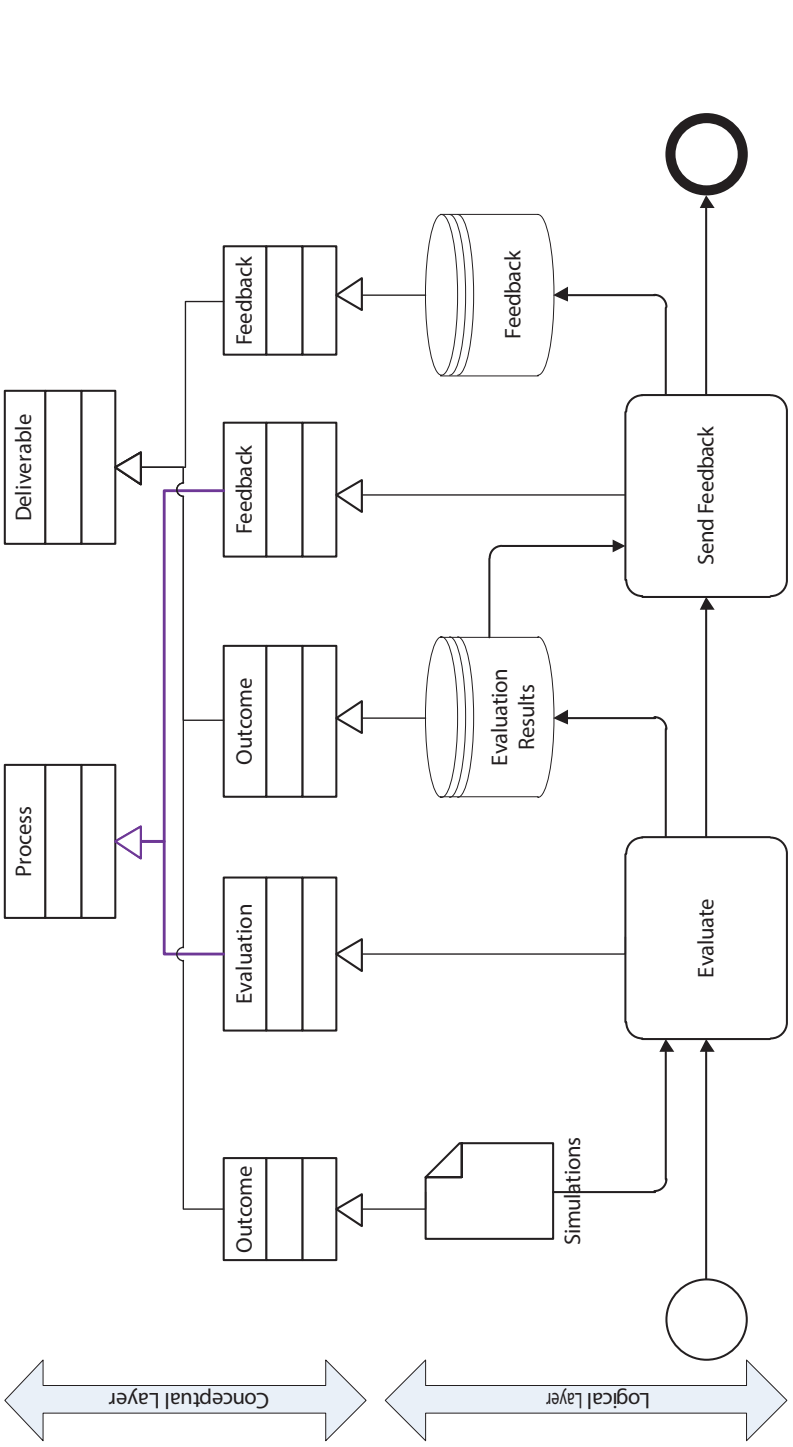


Figure 5.23: Logical Layer To Conceptual Layer Relation – The Intervention Process – The Evaluate Simulated Scenarios Sub–Component

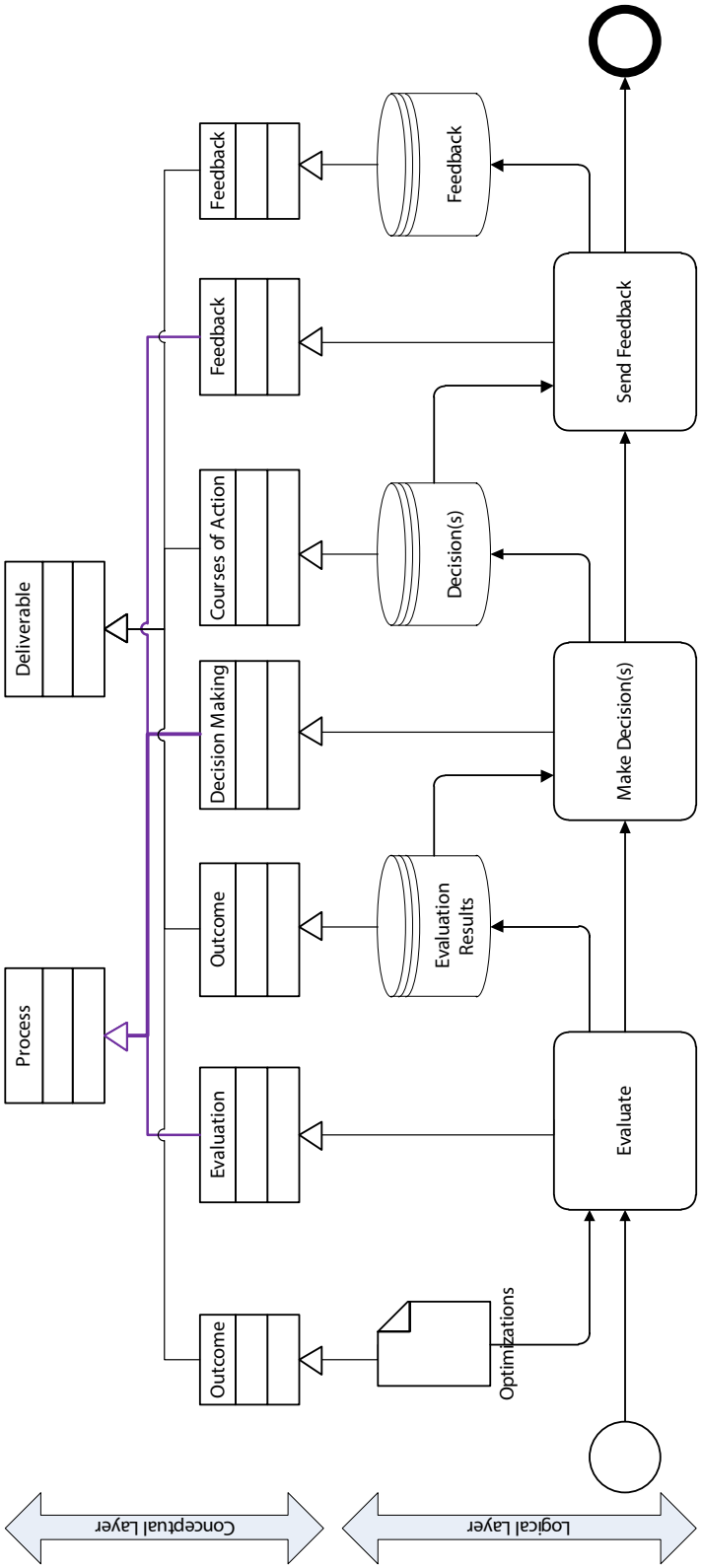


Figure 5.24: Logical Layer To Conceptual Layer Relation – The Intervention Process – The Evaluate Optimized Scenario Sub–Component

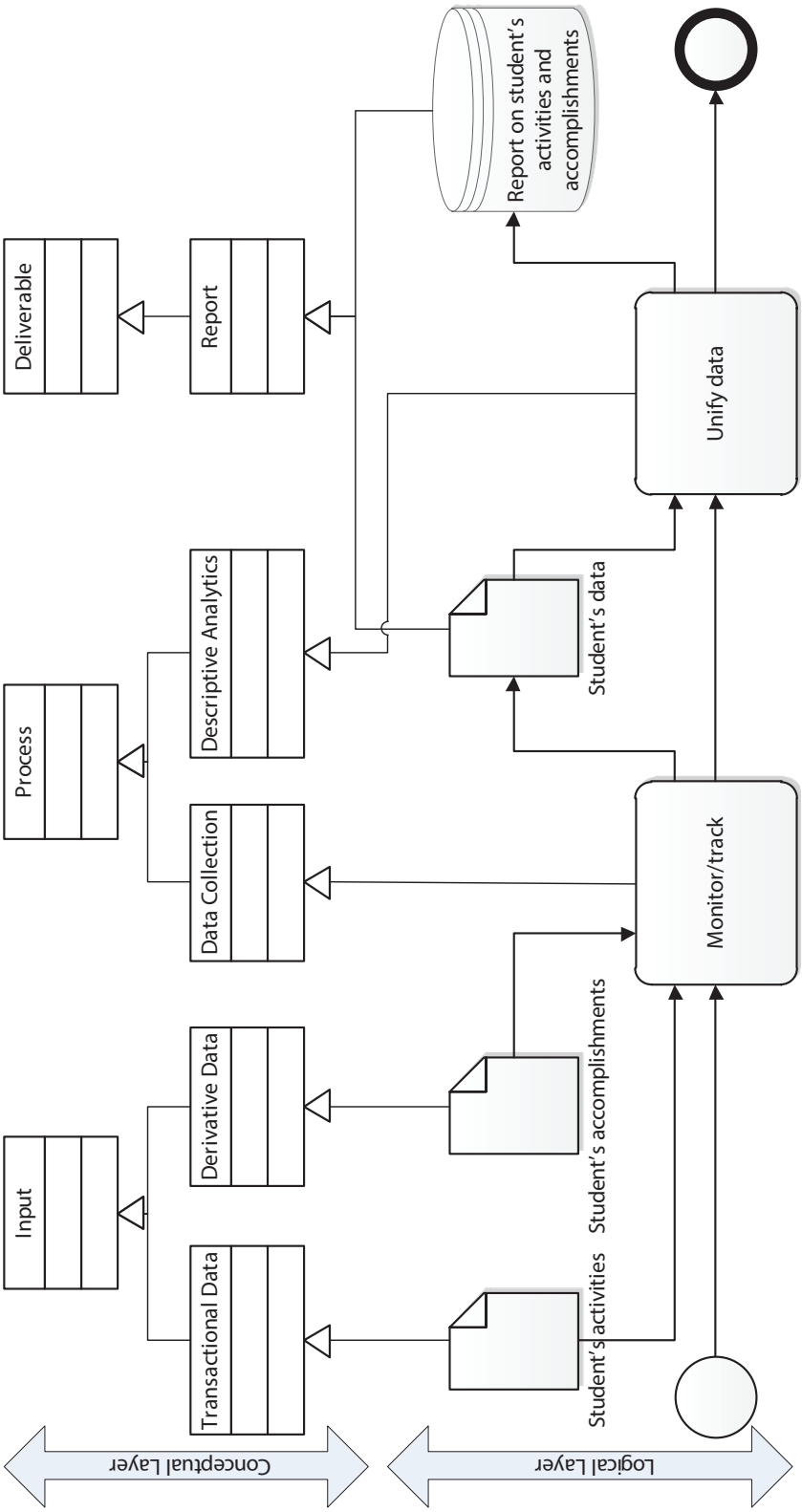


Figure 5.25: Logical Layer To Conceptual Layer Relation – The Monitoring Process

Its goal is to represent learning analytics functional processes (Section 5.4). The *physical* (the formalization) layer is focused on implementing the proposed framework in one real–world application scenario which is elaborated in detail in Chapter 6 along with its algorithm and results (Section 5.5). We will experimentally validate the effectiveness of the proposed framework in Chapter 6 using one real use case. Finally, two sample learning analytics processes (the intervention process and the monitoring process) representations are investigated in detail and their logical layers' "IS–A" relations to the conceptual layer are illustrated.

This chapter addressed the following two research questions mentioned in Section 1.3:

- *RQ2* — is covered in Section 5.6 by connecting the logical and conceptual layers together.
- *RQ3* — is addressed in Section 5.4 by representing 10 learning analytics processes in BPMN.

In the next chapter (Chapter 6), the framework is incorporated in one real–world scenario to validate its usability in analytical use cases. The physical layer is implemented and a new approach is proposed to analyze learners' data and produce desired results.

Chapter 6

Personalized Prescriptive Quiz (PPQ) — Enhanced with Descriptive and Predictive Analytics

> "Fair does not mean giving every child the same thing, it means giving every child what they need."
>
> *Rick Lavoie*

6.1 Introduction

Several methods have been proposed to collect, report, process, comprehend, and extract insight from big educational data [Baer and Norris, 2015], most to assist institutions with the understanding of what happened in the past and what might happen in the future through descriptive and predictive Analytics [Delen and Demirkan, 2013; Eckerson, 2007; Kaisler et al., 2014]. However, having the insight on what has happened in the past and what might happen in the future does not necessarily improve the outcomes. There is a need to transform information into insights and act upon them to meet predefined objectives [Chen et al., 2012; Baker and Gourley, 2014; Kaisler et al., 2014]. Therefore, Prescriptive Analytics

has emerged as the next frontier in business analytics. It is concerned with recommendations and guidance, which provides institutions with adaptive, automated and time–dependent sequences of operational actions [Evans and Lindner, 2012]. The integrated architecture presented in Chapter 4 combines descriptive, predictive and prescriptive analytics to allow near–optimal decisions to be made in real–time [Soltanpoor and Sellis, 2016]. Furthermore, an analytics framework capable of being instantiated in the context of LA was proposed in Chapter 5. The framework comprises conceptual, logical and physical layers to cover all major LA requirements and provides students with intelligent academic feedback [Soltanpoor and Sellis, 2016; Soltanpoor and Yavari, 2017].

In this chapter, we propose a novel approach called personalized prescriptive quiz (PPQ), which is based on the proposed composite analytics architecture presented in Chapter 4. The aspects of our contribution (which to the best of our knowledge, not found in the similar studies) can be listed as follow.

- taking into account the concepts of self–esteem and self–efficacy of students. Students lacking the foundational concepts will experience lower levels of self–confidence and eventually end–up failing or drop–out. The PPQ approach gives the students the opportunity to shape their learning pathways and perform assessments and self–assessments to identify and rectify their misconceptions which may lead to the elevated self–esteem.
- calculating the root–cause conceptual problem(s) for individual students that helps them target their misunderstood concepts.
- incorporating the question's difficulty and discrimination indexes which assists the instructors to better design the qualified questions assessing students' knowledge.
- creating the concept graph where and tagging each designed questions with their corresponding concept(s).

All together, our contributions are not limited to the proposed algorithms for PPQ, but helping students improvde their self–efficacy with incorporation of the conceptual dependency, root–cause analysis, difficulty–level designation and providing students with adaptive and dynamic personalized sets of questions to help them target their misconceptions (and eventually addressing them).

To shape the fabric of the proposed analytics framework this chapter addresses the fourth research question:

Research Question 4)

How do we devise and link the physical layer components enforcing higher–level processes (linking the physical, logical and conceptual layers altogether)?

The Intelligent Tutoring System (ITS), an in–house web–based application, provided the testbed for our approach. In the first phase, descriptive analytics were used by labeling all quiz questions with underlying concepts. Students were able to view concept descriptions whether or not they had problems. In the second phase, the quiz questions were used to predict how well they would perform in the exam. In the final phase, the framework was extended to incorporate prescriptive analytics through pedagogical interventions (such as recommending relevant learning resources or taking tests covering their misconceptions) to address the misunderstood concepts.

PPQ can help improve students performance by correcting their misconceptions early. According to [Robins, 2010], an effective pedagogical tools needs to address the students' misconceptions issue before it is too late, especially in typical introductory programming course because of its atypical rates of both failure and high scores! This leads to a bimodal grade distribution as the nature of such courses characteristic [Robins, 2010]. As students are struggling with the fundamental concepts in core courses (such as introductory programming), approaches such as PPQ identify and rectify students' misconceptions and help with their self–efficacy can address the key gaps in traditional assessment approaches [Robins, 2010]. PPQ provides a form of personalized coaching. It can help each student to individually identify and rectify their misconceptions (acquisition and transfer of critical concepts) by providing them with individually designed sets of questions. Our results demonstrate a significant improvement in student academic performance after applying the PPQ approach. Instructors can design more efficient questions covering taught concepts, by taking into consideration student feedback gathered on PPQ performance.

The remainder of this chapter is organized as follows. Section 6.2 forms a base for introducing the PPQ approach by providing issues and challenges with the current available didactic solutions and approaches. Section 6.3 elaborates on the details of the proposed approach by discussing PPQ terminology and the PPQ algorithm. Section 6.5 is dedicated to the qualitative and quantitative results of applying the approach to first–year programming courses and the impact on student marks. Section 6.6 presents modifications and expansions to the PPQ approach, based on student and instructor feedback, in order to make it more adaptable to their needs. Finally, Section 6.8 summarizes the chapter findings and lists the

contributions that address the relevant research question.

6.2 Problem Statement

Introductory programming courses are experiencing high failure and attrition rates[1] (up to 40%) partly reflecting incoming student diversity and background [Beaubouef and Mason, 2005; Biggers et al., 2008; Soh et al., 2007; Lang et al., 2007; Denning and McGettrick, 2005]. One reason for poor performance is that a standard assignment common to all students is not effective in identifying or correcting the misunderstood concepts of each individual student [Pears et al., 2007; Lang et al., 2007; Venema and Rock, 2014]. The problem is exacerbated the student cohort diversity as each subsequent test assumes that every student has somehow mastered earlier foundational concepts [Harlen and James, 1997; Pears et al., 2007; Lang et al., 2007]. Given that most concepts are interdependent, there is a need to explicitly capture dependencies among them. For example, to understand the "array" concept in programming languages, students need to know the "loop" concept. To figure out the "loop" concept, students need to know the "operator" and the "variable" concepts. Therefore, there is a dependency between "array" and "loop" concepts, and so forth. Consequently, if a student misunderstands the "array" concept, we can determine whether they previously understood the "loop concept. We can continue this process until we find the root misconception.

Furthermore, some critical drawbacks of traditional teaching and assessment strategies are listed as follows:

- *Lack of effective feedback.* Little or no feedback is provided to students regarding their performance on earlier tests. In some "summative" assessment approaches, students are notified of their feedback very late (after the final exam) or no such feedback is presented to them at all. In the latter case, the final marks are published without any descriptive explanation on exam questions [Lang et al., 2007; Harlen and James, 1997]. In contrast, "formative" assessment approaches are mostly concerned with improving students' understanding through conducting multiple diagnostic tests and learning activities, administered over several weeks [Lang et al., 2007; Harlen and James, 1997]. However, even formative approaches are comprised of fixed sets of tests covering complex concepts. This means that they deal with all students uniformly [Harlen and

[1] https://www.education.gov.au/news/release-higher-education-standards-panel-s-discussion-paper-improving-completion-retention-and (accessed on 10 Apr. 2018)

James, 1997; Taras, 2005]. To address this concern, a personalized approach should be designed to give each student the opportunity to overcome their past misconceptions through appropriate feedback.

- *Lack of personalized assessment techniques.* Traditionally, all students do the same quizzes, regardless of their past performance in foundational concepts, which might have the effect of weaker students falling behind in terms of their understanding of later concepts [Daempfle, 2003; Baer and Norris, 2015]. For the weaker students, this has a compounding effect, that is, misunderstandings of even early foundational concepts propagate into future assessments, and more significantly so. A personalized testing approach would allow for such compounding effects to be minimized. To address this issue, personalized question sets must be generated dynamically, based on each individual student's past performances. Such question sets may be interleaved between regular tests to encourage weaker students to catch up, while allowing above average students to be challenged through more demanding tasks. Such an approach could improve the teaching outcomes, especially where diverse student cohorts are involved.

The proposed PPQ approach aims at addressing afore–mentioned concerns and fills the gaps in order to achieve a more efficient and customized pedagogical process. It helps each individual student rectify their misconceptions during the semester, by providing novel personalized quiz sets covering misunderstood concepts.

6.3 PPQ Design

Introductory programming courses impart a range of fundamental programming concepts. Several compulsory interleaved tests and optional quizzes and assignments are designed to assess students' acquired knowledge in those fundamental concepts [Pears et al., 2007; Venema and Rock, 2014]. All test sets (compulsory or optional) are fixed and uniformly designed for all students, without accounting for each student's level of knowledge and understanding of the taught concepts. To address the gaps mentioned in Section 6.2, we propose a novel approach–Personalized Prescriptive Quiz (PPQ)—as an optional assessment context, to provide each student with dynamic and personalized sets of questions, designed to address their misconceptions. The PPQ is an implementation of the intervention process of the *physical layer*, introduced as a part of the analytics framework presented in Chapter 5.

The PPQ approach is a generic technique that may be applied to any course for which

the questions are tagged with the associated references to concepts and topics covered, the cognitive levels of the question (discussed in Section 6.3.1) [Anderson et al., 2001], and other question–dependent meta–data. A dependency graph illustrating all taught concepts and their relations is constructed and stored. Dependencies among concepts allow the instructor to probe the root–cause problem of each student's misconceptions, by traversing the edges in the graph that represent dependencies. For each concept in the graph, the system computes whether the student has responded correctly. If not, the system checks the student's responses in parent concepts, iteratively, until the root causes are identified. Some examples of dependencies among concepts were provided in Section 6.2.

The PPQ approach was applied in the course *Introduction to Programming* offered to first–year Information Technology students at RMIT University, with 274 enrolled undergraduate students. The rationale behind selecting the introductory programming course was influenced by the following:

- *Technical courses have been acknowledged to be among the most challenging for first–year undergraduates* [Venema and Rock, 2014; Wiedenbeck et al., 2004],
- *Introductory programming courses are cornerstones of computer science majors, in terms of the fundamental concepts taught and the skills developed* [Pears et al., 2007; Denning and McGettrick, 2005], and
- *Introductory programming courses have continually been experiencing high dropout and failure rates, which supports that learning to program is challenging for novices* [Beaubouef and Mason, 2005; Wiedenbeck et al., 2004].

In the subsequent sections, the terminology of PPQ approach is presented first. Next, the pre–processing phase of the process is discussed. Finally, the PPQ algorithm is introduced.

6.3.1 Terminology

The PPQ approach was applied between weeks 9 and 12 of the semester when almost all concepts have been taught (in accordance with the syllabus). For the sake of simplicity, all the assessments (tests/quizzes) prior to the PPQ are called "pre–test(s)". Similarly, all assessments completed after the PPQ are denoted "post–test(s)". By analyzing their performances by students in pre–tests, the list of misunderstood concepts for each student is determined. Given the set of extracted misconceptions so determined, the PPQ algorithm

generates a set of personalized questions covering those misunderstood concepts, packaged as one prescriptive quiz per student (denoted by ppq_i). The intention is to help students to improve their understanding of the taught concepts (by assisting them to identify and rectify their misconceptions) before the final exam, by applying the PPQ.

Figure 6.1 illustrates our approach in terms of its intervention process (discussed in Section 5.4.4). Given each student's misconceptions, the system generates individualized sets of questions per student, before the upcoming post–test(s) (including the final exam). As noted from Figure 6.1, not only are different sets of questions (aka quizzes or ppq_is) provided to each student, but the number of questions in each quiz (the quiz size, denoted $|ppq_i|$) also varies. Consequently, stronger students are provided with more challenging sets of questions with smaller ppq_i (fewer questions) and conversely, more question sets (larger ppq_i) are for weaker students. The intervention process occurs between weeks 9 and 12, because:

- *All concepts are taught by week 9*, and
- *The system determines students' misconceptions more accurately having their enriched results from previous pre–tests.* By getting access to the collective student assessment results by week 9, compared to the limited early weeks' assessment results, the system can analyze and calculate each student's misconceptions more effectively.

The following list presents the terminology for the PPQ approach:

- *Topic (T)* — refers to the fundamental programming topics covered in the IoP course. The course covers 5 major topics in programming: *Sequences*, *Data types and Operators*, *Selection*, *Loops*, and *Arrays*. The set of topics of the course is displayed as $T = \{t_1, t_2, t_3, t_4, t_5\}$.
- *Concept (C)* — each topic comprises several programming concepts. More than 30 different concepts were captured in the system. Not all concepts are covered in each topic. A concept may be covered in more than one topic. For example, the *Logical Operator* concept is covered in both data types and Operators, Selection and Loops topics, and the list goes on. The set of all taught concepts are denoted as $C = \{c_1, c_2, ..., c_n\}$.
- *Question (Q)* — refers to a regular question which is designed to assess students' knowledge in their taught concepts. Each question can assess one or more concepts and links to the corresponding topics. The set of question Q is represented as in $Q = \{q_1, q_2, ..., q_n\}$.

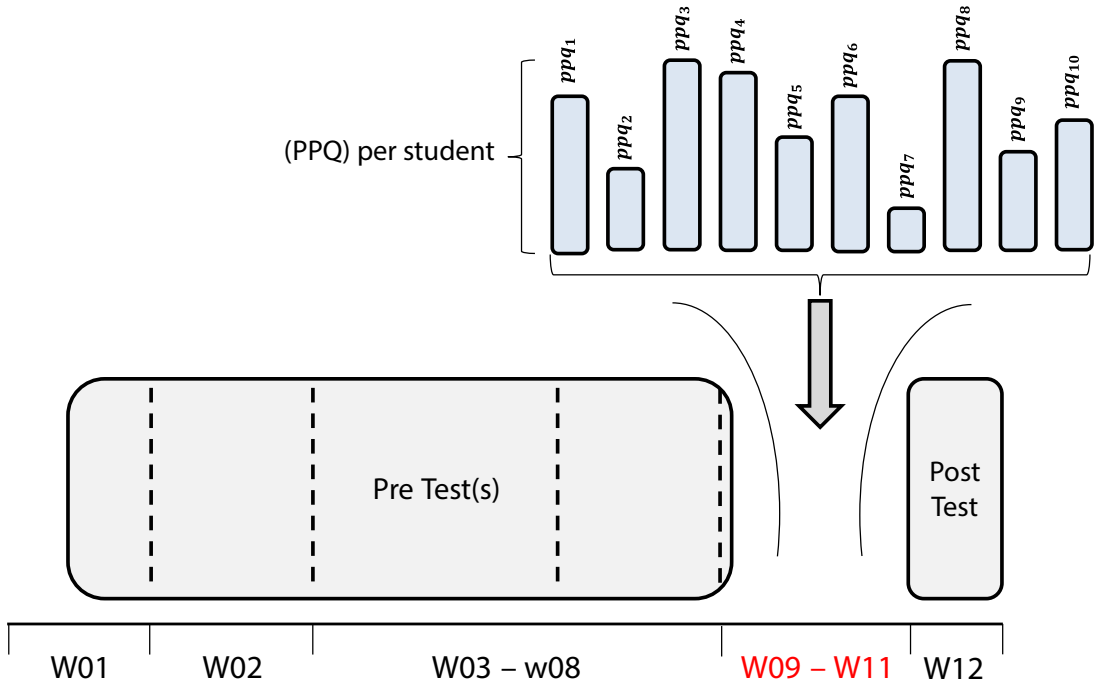


Figure 6.1: Personalized Prescriptive Quiz (PPQ) Approach – The Intervention Process

- *Analysis and Application Level Questions (An_q, Ap_q)* — According to Bloom's Taxonomy [Anderson et al., 2001], educational learning objectives are categorized into three main classes: Cognitive, Affective, and Psychomotor. We focused on four levels of the cognitive category: Knowledge, Comprehension, Application, and Analysis. The Knowledge (Remembering) level is concerned with recalling or retrieving previously learned information. For example, the number of bytes in the integer data type, or the type of true or false values. The Comprehension (Understanding) is concerned with understanding the meaning and interpretation of instructions and problems. For example, explain how Java garbage collection works in your own words, or what is the error in the code segment shown in the test. The Application (Applying) level applies what was learned to real situations. For example, write a recursive program to calculate the factorial of a given input number. The Analysis (Analyzing) level distinguishes between facts and inferences by separating the concepts into component parts, to understand their organizational structure. For example, a question might ask: What does the given code segment do? Students need to understand the concepts and analyze the situation to realize the purpose of the code provided. In this research, we are mainly focused on the "Analysis" and "Application" levels of Bloom's Taxonomy, considered to the more

challenging levels. This gives us a solid understanding of the concepts that students did not understand properly. Henceforth, for the sake of brevity, we denote the Analysis and Application level questions as An_q, Ap_q, respectively.

- *Question Pool (QP)* — the finite set of designed questions (Q) covering the taught concepts for this course. Each question may cover one or more concepts and is linked to their corresponding topics. The question level (Knowledge, Comprehension, Application, and Analysis) is also defined by the designer. The set of all questions in the question pool is denoted as $QP = \{q_1, q_2, ..., q_n\}$.

- *Used Question (UQ)* — a subset of questions in the question pool (QP) that have been asked/answered so far. It is denoted as $UQ = \{uq_1, uq_2, ..., uq_i\}$ and $UQ \subseteq QP$. Initially, $UQ = \emptyset$.

- *Fresh Question (FQ)* — a subset of questions in the question pool (QP) which have not been asked/answered yet. It is depicted as $FQ = \{fq_{i+1}, fq_{i+2}, ..., fq_n\}$ and $FQ \subseteq QP$. Also, $UQ \cup FQ = QP$ and $UQ \cap FQ = \emptyset$. Initially, $FQ = QP$.

- *Student (S)* — the set of all students who enrolled in the course. The students set is defined as $S = \{s_1, s_2, ..., s_n\}$.

- *Misunderstood Concept (MC)* — the subset of concepts (C) with more than half of their covering questions were answered incorrectly. It is represented as $MC = \{mc_1, mc_2, ..., mc_k\}$, where $k \leq n$ and each mc_i represents the set of misunderstood concepts for student s_i.

  Please note that in discussions about "misconceptions", we typically have two main categories: (1) misunderstanding certain concepts, and (2) having the lack of understanding in particular concepts [Cetin and Ozden, 2015; Yadav et al., 2016; Caceffo et al., 2018; 2016; Kaczmarczyk et al., 2010; Ben-Ari, 2001]. This means that there is a difference between students who did not understand a concept and the ones who have formed wrong mental models for that concept. In this study, we are targeting those with incorrect mental models. For example, students may know the math operations properly, but when it comes to the computer programming, they might be puzzled when encounter statements such as $x = x + 1$, or $x++$ which do have meaning in computer programming, but are not correct in math (if you deduct x from both sides of the equation $x = x + 1$ then, it yields $0 = 1$ which is incorrect in math)!

- *Personalized Prescriptive Quizzes (PPQ)* — a subset of dynamically generated questions in the question pool (QP) for each individual student s_i, given their misunderstood concepts mc_i. It is defined as $PPQ = \{ppq_1, ppq_2, ..., ppq_n\}$ and $PPQ \subseteq QP$.

6.3.2 Pre–Processing

Prior to implementing the PPQ algorithm, some derived data elements must be determined, as follows.

- *The maximum number of questions in the system per quiz (n)* — in our tests, n is initialized to 15 ($n = 15$) but may be defined by the question designer.

- *The maximum number of questions per ppq_i (n_i)* — note that the number of questions per ppq_i for each student may vary. The number of questions generated by the PPQ algorithm per ppq_i is n_i, where $n_i = |ppq_i|,\ n_i \leq n$. Note too that not all n_i question slots will be populated for the student s_i, because we are proposing a personalized approach to dynamically generate the number of questions per student and each question set (ppq_i) might be different from others.

- *The number of questions per topic (q_t)* — given that we have 5 different topics and 15 different questions per ppq_i, there will be $q_t = \lfloor \frac{n}{t} \rfloor = \lfloor \frac{15}{5} \rfloor = 3$ questions per topic. However, for the situations where the $\lfloor \frac{n}{t} \rfloor$ fraction is not a natural number, according to the Pigeonhole Principle [West et al., 2001], we distribute q_t questions based on the given number in $\lfloor \frac{n}{t} \rfloor$ and will distribute the rest by choosing the An_q, Ap_q level questions for the student s_i. The number of remaining question slots will be calculated as $(n - \lfloor \frac{n}{t} \rfloor \times t)$.

- *The set of misunderstood concepts for each student (mc_i)* — technically, if the student s_i responded incorrectly to more than half of the questions covering a particular concept c_i, then the concept c_i will be considered as the misunderstood concept mc_i. The process to calculate the set of misunderstood concepts per student is as follows: The q_i is defined as the number of questions covering the concept c_i which have been taken by the student s_i so far. Also, $wrong_i$ is defined as the number of incorrectly answered (among q_i) questions by the student s_i. Then, the concept c_i is added as a misunderstood concept

mc_i to the MC set, if:

$$\begin{cases} q_i = 2k \quad , & \text{and } wrong_i \geq \frac{q_i}{2} \\ q_i = 2k \pm 1, & \text{and } wrong_i > \frac{q_i - 1}{2} \end{cases}$$

- *The set of fresh questions covering the misunderstood concepts for each student (mfq_i)* — the set of FQ covering the misunderstood concepts for student s_i is defined as mfq_i. In generating mfq_i, we filter $mfq_i \in FQ$ with the An_q, Ap_q level. Also, the questions covering more than one concept (including the misunderstood concepts) are prioritized. Consequently, the process guarantees that only harder questions in the Analysis and Application levels are selected.

- *The set of used questions covering the misunderstood concepts for each student (muq_i)* — the set of UQ covering the misunderstood concepts for student s_i is defined as muq_i. In generating muq_i, we filter $muq_i \in UQ$ with the An_q, Ap_q level. Again, the questions covering more than one concept (including the misunderstood concepts) are prioritized. The process guarantees that only harder questions in the Analysis and Application levels are selected.

- *The set of used questions with An_q, Ap_q levels for each student (uuq_i)* — this is the set of UQ taken by the student s_i and answered correctly. These questions generally cover more than one concept including the misunderstood concept and are designed as An_q, Ap_q levels. The set uuq_i is generated and provide to the student s_i because: first, they are among the toughest questions with the Analysis and Application levels. Second, the uuq_i set questions are covering more than once concept including the misunderstood concept mc_i and there is a chance that the student has accidentally answered them correctly. The logic behind this hypothesis is that the student has already answered all other similar questions (covering the misunderstood concept mc_i) incorrectly. Therefore, it is worth providing them with the uuq_i set to assess their knowledge more accurately.

- *The set of used questions with An_q, Ap_q levels which were answered incorrectly by most of the students ($allmuq_i$)* — the set of questions in UQ which were answered incorrectly by most of the students is defined as $allmuq_i$ and is filtered based on the following criteria:

1. The descending order of the total number of wrong answers to any particular question q_i by all students,
2. Prioritizing q_is that cover more than one concept, and
3. Picking the top 90% percentile of the generated list of questions after performing the first and second criteria.

6.3.3 Algorithm

The steps towards generating the prescriptive quizzes for each individual student is specified by Algorithm 1.

Algorithm 1 PPQ Generator

procedure PERSONALIZED-PRESCRIPTIVE-QUIZ
  $PPQ \longleftarrow \emptyset$
  for each $s_i \in S$ **do** {
    input: $s_i,\ mc_i,\ n_i,\ mfq_i,\ muq_i,\ uuq_i,\ allmuq$
    output: ppq_i for s_i
    $ppq_i \longleftarrow \emptyset$
    while $n_i \geq 0$ **do** {
      $ppq_i \longleftarrow mfq_i$
      $n_i - = |mfq_i|$
      $ppq_i \longleftarrow muq_i$
      $n_i - = |muq_i|$
      $ppq_i \longleftarrow uuq_i$
      $n_i - = |uuq_i|$
      $ppq_i \longleftarrow allmuq$
      $n_i - = |allmuq|$
    $PPQ \longleftarrow ppq_i$
    }
  return PPQ
  }

The inner loop of the algorithm is iterated per student s_i, given their set of misunderstood concepts mc_i, the number of total questions per quiz per student n_i, the set of fresh questions covering the misunderstood concepts for the student mfq_i, the set of used questions covering the misunderstood concepts for the student muq_i, the set of correctly answered used question covering the misunderstood concepts for the student uuq_i, and the set of used questions which were answered incorrectly by most of the students $allmuq_i$. The result is returned as a set of personalized prescriptive quizzes (ppq_i) which are individually designed for the student s_i.

By aggregating the generated ppq_is for each student, the algorithm will generate the PPQ set incorporating the personalized prescriptive quizzes for all enrolled students.

Initially, the ppq_i is null. In the first step, the PPQ algorithm adds the list of available fresh questions covering the misunderstood concepts (mfq_i) to the ppq_i set. Then, if there is room to add further questions, the system adds the list of used questions covering the misunderstood concepts (muq_i). Next, if there is still room, the list of Analysis and Application level questions covering the misunderstood concepts (uuq_i) will be added to the list. Please note that these are questions that the student has answered correctly in their previous tests, but due to their importance, we provide them again. Finally, if there is still room for further questions, the list of top 90% percentile questions that most of the students have answered incorrectly ($allmuq$) will be added to the ppq_i list. In the end, the student s_i will be provided with the personalized list of prescriptive quizzes. For each student s_i, the following equation applies in the algorithm termination: $n_i = |ppq_i|$. By iterating the algorithm for all students, the system will provide the list of ppq_is for all students $PPQ = \{ppq_1, ppq_2, ..., ppq_n\}$.

6.4 Pedagogy and Course Design for Programming Fundamentals

Most of the experiments were carried out in the introduction to programming course, a problematic course, which went to major revamp in 2017. This course, a core for the IT degree is taken by around 300 students with little or no programming skills. These students come with varying academic performance, but the IT enter scores are much lower than CS and software engineering degrees. Many students find abstract reasoning difficult and often fare poorly in the semester end exams. The failure rates in exams that test mainly problem–solving skills had been as high as 50% though overall failure rates were less than 30% because they fare better in assignments done in the class. However, most students proceeding to subsequent programming courses without passing the exam component fared poorly. Student feedback revealed many found paper–based exams difficult as they have done all their previous programming tasks on a computer. Most students also skipped the lectures, which focused mainly on programming constructs and syntax. The exam performance revealed a substantial number of students did not even master the foundational topics including selection, repetition, and methods though they fared well in assignments. There was a need to innovate and come up with new teaching methods.

Hence the pedagogy used in 2017 attempted to address all these difficulties using a multimodal approach combining class programming tests, online class tests, quizzes, prescriptive

quizzes, YouTube videos and incremental visual constructivist assignments in addition to the traditional lectures, tutorials, and programming tasks. These components are described briefly in Section 6.5 below.

6.5 Results

The qualitative and quantitative implications of the work are elaborated in this section. First, the students' responses to the survey questionnaires are classified and discussed in Section 6.5.1. Next, the quantitative data analytics regarding the collected numerical results are provided in Section 6.5.2. Prior to conducting the research, we had our ethics application approved by RMIT University College Human Ethics Advisory Network (CHEAN). Before proceeding with the following sections, some information regarding the course and the assessments within the course is provided. Table 6.1 represents the teaching schedule break–down for the Introduction to Programming course in Semester 1, 2017.

Table 6.1 can be simplified in Table 6.2, where:

- W_i — indicates the i-th Week, and for each week.
- L_i — indicates the i-th Lecture. There are 12 Lectures throughout a semester, one for each week. Lectures adapted a problem–solving approach using increasingly complex and authentic problems while introducing new constructs in context. Hence the focus was no longer on programming constructs or syntax but on problem–solving.
- Tt_i — indicated the i-th Tutorial. There are 11 Tutorials during a semester, starting from the second week.
- OLT_i — indicated the i-th in–class Online Test. Five different OLTs are designed for students to be taken during a semester. Online tests administered a week after the release of quizzes were assessed and provided the incentive for students to do the quizzes regularly. The results of the class tests were made available as soon as students submitted their responses.
- PrT_i — indicated the i-th Programming Test. There are four different in–class Programming Tests throughout a semester. These tests are paper–based and written and the results of which will be available in one week after the tests were taken (delayed result). Paper–based programming tests administered in the class required students to

Table 6.1: Teaching Schedule Break–down for Introduction to Programming Course

Week	**Topic (Lectures)**	**Tests (In–Class)**	**Quizzes (In–Class)**	**Assignments (Tute–Lab)**
1	Variables, Sequence, Operations			
2	Objects and Methods / Input Output / String manipulation / introduction to selection		Quiz 1	
3	Selection and Operators	Online Test Test (2%)	Quiz 2	
4	Repetition	Prog. Test 1 (5%)		
5	Methods and Argument Passing	Online Test Test (2%)	Quiz 3	
6	Arrays & Debugging	Prog. Test 2 (5%)		Assignment 1 Part (1/2) (5%)
7	Problem solving and Collaboration	Online Test Test (4%)	Quiz 4	
8	Arrays, Selection, Repetition, Methods & Problem Solving		Assignment 1 Part (2/2) (15%)	
9	Classes	Prog. Test 3 (5%)	**PPQ**	
10	Class Design 1	Online Test Test (2%)	Quiz 5, **PPQ**	
11	Class Design 2	Prog. Test 4 (5%)	**PPQ**	
12	Revision	Final online Test (10%)	Quiz 6	Assignment 2 (10%)

Table 6.2: Vertical Break–Down Per Week

W1	**W2**	**W3**	**W4**	**W5**	**W6**	**W7**	**W8**	**W9**	**W10**	**W11**	**W12**
L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12
	Tt1	Tt2	Tt3	Tt4	Tt5	Tt6	Tt7	Tt8	Tt9	Tt10	Tt11
		OLT1		OLT2		OLT3			OLT4	OLT5	
			PrT1		PrT2			PrT3		PrT4	
	Q1	Q2		Q3		Q4			Q5		Q6
								PPQ	**PPQ**	**PPQ**	
					A1 (1/2)		A1 (2/2)				A2

solve simple problems using control structures, methods, and arrays. To prevent plagiarism 6 different tests at the same standard were created. These tests were intended to get students to start writing their programs within the first three weeks. These tests were marked and returned to students with a one–week turnaround. The following depicts one sample in–class programming test:

Test — You are required to complete the program below to compute the gross salary for all the hourly rated employees. These employees are paid 1.5 times the normal rate for hours exceeding 40. You are required to print two separate salary summaries as described below:

(a) Salary Summary for those earning $20.0 or less per hour, and

(b) Salary Summary for those earning more than $20.0 per hour.

Both summaries should display the first name of the employee, the hours worked, the rate of pay and the gross–pay.

Array Processing Required — The partially completed program has initialized three separate arrays names, rates and hours. You are required to use them and introduce another array named grossPays to store all the gross salaries computed. You are not required to introduce other methods.

Program to be Completed

```
import java.util.*;
  public class SalaryProcessing2A
  {
   public static void main(String args[])
      {
         String names[] =
             {"Bill","Chew","David","Ravi","Smith","Teo"};
          double hours[] = {45,60,60,44,38,45};
          double rates[] = {23.0, 18.0, 32.0, 22.0, 46.0, 34.5};
      }
  }
```

- Q_i — indicated the i-th in–class Quiz. Six different kinds of quizzes are designed to be taken during a semester. Given that quizzes are online, the results will be available right after the student submits the quiz (instant result). Please note that students needed to take quizzes at the end of the week they had taken the in–class Online Test. Class quizzes were designed for most topics and students were encouraged to assess themselves regularly. The class quizzes made use of multiple–choice, multiple–selection, fill in the blanks and Parson's puzzle type of questions. The sample quiz below requires the student to specify the exact output from the program below. Students were allowed to view the answers and the explanations immediately after answering the questions. If results from previous semesters are available, the system allowed the final results to be predicted. These predicted results become more accurate as the semester progresses and more student–specific data become available. However, this feature was not used in 2017 as there was no prior data.

  The following illustrates an example of the in–class quiz:

```
public class PrintStars2
  {
   public static void main(String args[])
      {
         int n[] = {1, 3, 7, 17};
          int j = 0;
          for (int i=1; i<=15; i++)
          {
            System.out.print("*");
              if (i == n[j])
              {
               System.out.print(" ");
                  j++;
              }
          }
      }
  }
```

- *PPQ* — indicated the Personalized Prescriptive Quiz designed for each individual student. Prescriptive quizzes were made available in the school for the first time to identify and address learning difficulties through personalized coaching. Students were allowed

to repeat these quizzes between weeks 9 and 11. The quizzes were generated based on the diagnosis carried out in their performances in standard weekly quizzes. PPQs are personally designed for each student based on their previous assessment results. The results of the PPQ will be available right after students submit their quiz (instant result).

- A_i — indicated the i-th Assignment (for Assignment 1, it was divided into two sub–assignments in weeks 6 and 8). We also refer to them as "Incremental Visual Constructivist Assignments and Demos". Incremental visual constructivist assignments were devised as many average students developed better self–efficacy after implementing simpler tasks. The first assignment was the Snakes and Ladders game where students implemented the standard game logic in the first part and created a customized board in the second. The second assignment was a visual 3–armed robot, which was required to perform increasingly complex tasks allowing students to gradually develop the problem–solving skills. Students were asked to demonstrate their progress in the labs to ensure they are making steady progress. Students are interviewed for their submitted assignments to demonstrate their running codes and answer to the asked questions by markers. The assignments' result availability depends on the lecturer's policy, but it is not instant!

Apart from the items mentioned above, extra pedagogical components were also considered such as:

- *YouTube*[2] *Videos and Lecture Recordings* — YouTube videos of 10 to 15 minutes duration focusing on specific problem–solving activities were created. These videos were made available before the lectures to foster more interaction during class time. In addition, all lectures were recorded allowing average students to learn at their own pace.

- *Incremental Visual Constructivist Assignments and Demos* — Incremental visual constructivist assignments were devised as many average students developed better self–efficacy after implementing simpler tasks. The first assignment was the Snakes and Ladders game where students implemented the standard game logic in the first part and created a customized board in the second. The second assignment was a visual 3–armed

[2]https://www.youtube.com/

Table 6.3: Assessments Result Availability

Assessment	**Assessment Result Availability**
In–Class Online Test	Instantly
Programming Test	One week after the test
Quiz	Instantly
PPQ	Instantly
Assignment	Depends on the policy (not instant)

robot, which was required to perform increasingly complex tasks allowing students to gradually develop the problem–solving skills. Students were asked to demonstrate their progress in the labs to ensure they are making steady progress.

- *Exam* — The exam was made up of three sections to measure the performance of students. The multiple choice and fill in the blank in the first section (30%) covered the breadth of the course. The second section required students to develop simple algorithms using control structures and arrays. The third section required students to develop a complete program made up classes and methods.

Given the above–mentioned activities, the following table (Table 6.3) depicts the likely time during which each assessment result will be ready (availability of the results):

6.5.1 Qualitative Results

At the end of the semester, all enrolled students were asked to complete one survey questionnaire after taking part in the PPQ research project. The survey was designed in two parts concerning both groups of students: (1) part A for those who took PPQs (the test group [3]), and (2) part B for those who opted not to participate (the control group [4]). According to the collected survey responses, 43.7% of the students in the test group responded to the survey questionnaire as well as 24.5% of the control group.

A short summary of the designed survey questionnaire is depicted as follows:

- *Part A* — is intended for those who took the personalized prescriptive quiz that comprises 11 questions (7 scale selection and 4 descriptive response questions). The first

[3] From now on, by the "*Test Group*", we mean the group of students who self–selectively chose to participate in the PPQ approach.

[4] Similar to the note mentioned on the "Test Group", from now on, any reference to the "*Control Group*" corresponds to the student cohorts who freely chose to opt out taking the PPQ.

seven questions provide a five–scale range of "strongly agree", "agree", "neutral", "disagree", and "strongly disagree" per question. These questions are listed as follows:

1. Instant feedback on my response to each question was effective.
2. I found the personalized prescriptive quiz (PPQ) approach beneficial.
3. The PPQ approach helped me clarify concepts I had failed to understand earlier.
4. The PPQ approach presented questions in a logical order.
5. The PPQ approach helped me to proceed to advanced concepts more confidently.
6. After taking the personalized prescriptive quiz (PPQ), my performance in the normal test improved substantially.
7. I like to see such an approach in other courses.

The remaining four descriptive response questions are:

8. What aspects of the system did you find more beneficial?
9. What aspects of the system could be improved?
10. What other types of personalized quizzes do you recommend to be considered in the future?
11. How can the overall system be improved?

- *Part B* — is applicable for those who opted not to take the PPQ and is comprised of four descriptive response questions as follows:

1. What are your reasons for choosing not to take the Personalized Prescriptive Quizzes (PPQs)?
2. Do you think taking the PPQ may have impacted your score in the final test (after the PPQ)?
3. Which concepts of the course did you experience the most difficulties?
4. How can the course learning outcomes be improved?

Among the collected responses to the key question in part A "What aspect(s) of the system did you find more beneficial?" (question 8, part A), the top five responses being instant feedback and explanation, personalization, identifying misunderstood concepts, the

variety of questions, and ease of use. Interestingly, even the majority of students in part B who opted out (63.5%) also answered "yes" to the question "Do you think taking the PPQ might have impacted your score in the final test?".

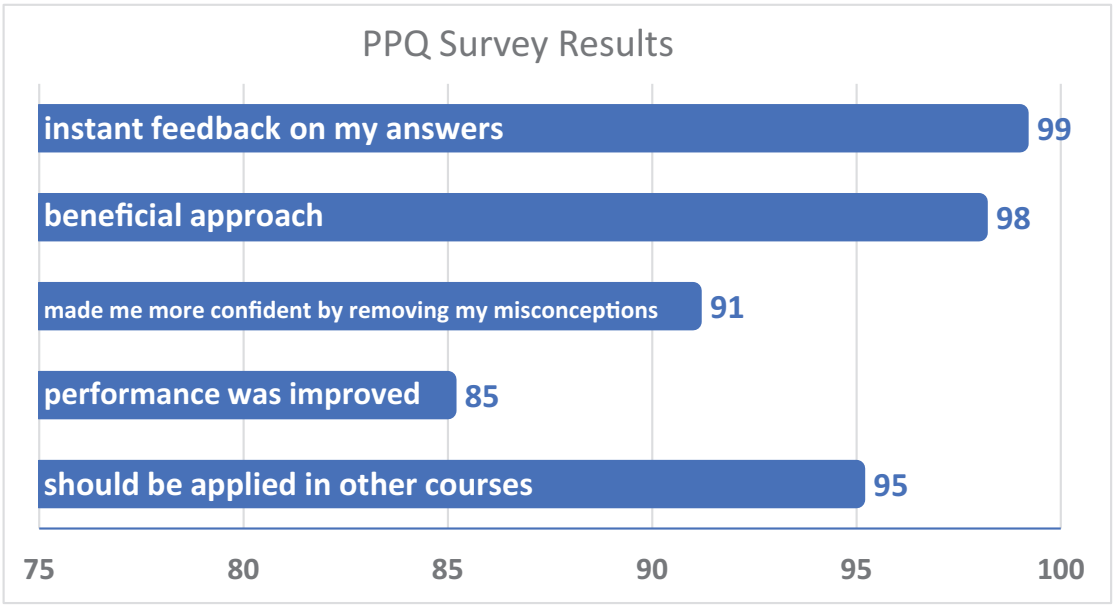


Figure 6.2: PPQ Survey Results

Figure 6.2 illustrates the responses to the five Likert scale questions, where:

- *99% found instant feedback to their responses useful,*
- *98% benefited from the PPQ intervention,*
- *91% found that the PPQ helped them to correct their misconceptions and made them more confident of their progress,*
- *85% acknowledged PPQ has impacted their performance positively*, and
- *95% of the novices surveyed wanted the PPQ to be extended to other courses, mainly because they found the personalized assessment was enjoyable and beneficial.*

6.5.2 Quantitative Results

The impact of applying the PPQ approach on students' performance in their final exam and post–tests is elaborated in this section. Nearly 64% of enrolled students opted to participate in the PPQ research project and are labeled the *Test Group*, and 36% opted not to participate

and are referred to as the *Control Group*. Please kindly note that as per the ethics approval process in getting access to students' data, it was ethically unfair to choose students deliberately for test and control groups. Thus, we asked students to freely choose whether or not they want to be part of the PPQ experiment (forming two self–selecting groups). Therefore, both of the *test* and *control* groups refer to the self–selecting student groups who chose to take PPQ, and opted out, respectively[5]. The two–tailed *t–test* was selected because we had two groups and one independent variable (the impact of applying the PPQ approach) to be assessed. The $p - value$ was calculated with the standard confidence of $\alpha = 0.05$. The null hypothesis was "H_O: *There is no significant difference between the control and test groups in presence of the PPQ approach*" which means the improvement from PPQ in the control group is the same as that of the test group. The improvement to student performance was computed based on the difference between pre and post–class tests.

Table 6.4 demonstrates the results for the control and the test groups in pre and post–tests, both carrying a total of 10 marks. The test group took into consideration (a) all PPQ takers (regardless of their PPQ marks), and (b) those with PPQ scores ≥ 70. The test group students with PPQ ≥ 70 were awarded a bonus (2 marks) to encourage weak students to repeat the PPQs. The *# of Students* column refers to the number of participants in each group. The *Improvement Difference* and the *Improvement %* columns illustrate the actual improvements and the percentage improvements for the control and test groups. In addition, the $p - value$ for the improvements in the two test groups compared to the control group is computed with the confidence of $\alpha = 0.05$. As per the last two rows of the Table 6.4, both cases developed $p - values$ less than 0.05 that demonstrate PPQ's strong significance on students' performance, and therefore the null hypothesis (H_O) is rejected. This result justifies the positive impact of adopting the PPQ approach on students' post–test results.

Next, other analytical implications of this study are elaborated. First, each one of the control and test groups' performance in their final exam responses (broken down into smaller ingredients) are discussed, the correlation between the test group and their overall grades is investigated to further justify the positive impact of adopting the PPQ approach on different clusters of students (from strong to weak), and finally, the performance of undergraduate and postgraduate students is compared and their academic behaviors are analyzed.

Table 6.5 depicts the comparative performances of the test and the control groups in the three equally weighted sections of the final exam (ie. the *Multiple Choice Questions*, the *Short*

[5] "*Test Group*" – *those students who chose to be part of the PPQ experiment, and "Control Group" – those who opted out.*

Table 6.4: PPQ's Impact on Class Test Results

	Pre–Test (10)	Post–Test (10)	# of Students	Improvement Difference	Improvement %
Control Group[1]	4.6	5.08	96	0.48	10%
Test Group (a)[2]	6.26	7.42	175	1.16	19%
Test Group (b)[3]	6.41	7.52	151	1.11	17%
% Difference[4]	**36.09%**	**46.06%**			
Significance (p-*value*)[5]					
Control[1] vs. Test Group (a)[2]	**0.003516***				
Control vs. Test Group (b)[3]	**0.007592***				

[1] Non–PPQ Takers [2] All PPQ Takers [3] PPQ Takers with ≥ 70 Marks on Their PPQ
[4] Control Group vs. Test Group (a) [5] $\alpha = 0.05$ * $p - value < 0.05$

Table 6.5: Control and Test Groups' Performance in Final Exam

	MCQ (10)[1]	SAQ (10)[2]	OOP (10)[3]	# of Students
Control Group	4.35	4.93	1.91	99
Test Group	6.95	7.74	4.12	175
Difference	**59.77%**	**56.99%**	**> 115%**	–
Statistical Significance[4]	$1.13091E-11$	$4.73681E-10$	$4.50746E-09$	–

[1] Multiple–Choice Questions. [2] Short–Answer Questions (Code Fragments).
[3] Object–Oriented Programming (Problem–Solving) Questions.
[4] $p - value$ of the t–test

Answer Questions, and the *Object–Oriented Programming Questions*), which carried 30% of the overall course marks. Please note that both student groups performed poorly in the object–oriented programming (OOP) questions which were mainly focused on the students' problem–solving skills, where they were asked to develop a complete Java program based on the specified requirements. Nevertheless, as the *Difference* row of the Table 6.5 represents, the difference between test and control groups is the highest in the OOP section which demonstrates that the PPQ approach can contribute to improvement in problem–solving.

Next, the correlation between the PPQ performance and the overall grade for the test group (63.8% of the total enrolled students) who took part in the PPQ at least once is calculated. We were mainly interested in identifying the different categories of students in our diverse cohort, and how best we could fine–tune the PPQ to meet their learning needs.

Specifically, we aimed to study whether promoting engagement through repeated attempts to get bonus marks was an effective strategy. Figure 6.3 shows a positive correlation of 0.49 between the PPQ and the final marks, which reveals engaging in PPQs and getting student specific feedback has a positive overall impact on learning outcomes. Figure 6.3 also reveals three main clusters: (1) the red cluster to the right shows that the majority of students who got high marks in PPQs also did well overall in the course, (2) the students in the blue cluster on the left, however, reveals that there are a number students who performed poorly in the exam despite their success in PPQs. These students may be the ones repeating the PPQs blindly to get the bonus marks without making any attempts to clarify their misconceptions. Hence, we may limit the permissible number of repeated PPQ in future, and (3) the third yellow cluster, in the bottom–right of the graph, shows a handful of students doing well in the course despite their low PPQ scores. These are probably the more confident students from the high band who may have attempted the PPQ once or discontinued midway perceiving it to be of little value. Though such students were not our main target group, it revealed the need to develop more PPQ options to challenge our top students.

We also sought to analyze the impact of PPQ on the lower bands of our diverse student cohort, our main target. We did this by comparing the distribution of marks in the pre and post–tests. Figure 6.4 clearly shows the greatest impact on the two lower quartiles. The lowest pre–test value of 24 (left) improved to 36 in the post–test (right). Also, the lower and the median quartiles in the pre–test (56 and 68) increased to 64 and 76, respectively, in the post–test. Figure 6.4 also reveals changes in the upper bands though not to the same extent. These results suggest PPQ is an effective instrument for dealing with diverse student cohorts.

Also, to compare the performance of the students who took PPQ (depicted in Figure 6.4) and those who opted out, lets provide the non-PPQ takers results as well. According to our data-set, the non–PPQ takers results are as per Table 6.6.

6.5.3 Discussion

The student feedback and performance improvements have shown adaptive prescriptive quizzes generated by such a framework can help boost the confidence of stragglers and help narrow the differences in diverse student cohorts. The analysis of exam results suggests a framework explicitly capturing cognitive levels can help novices improve their program–writing and problem–solving skills. Improving learning outcomes, however, requires delving into

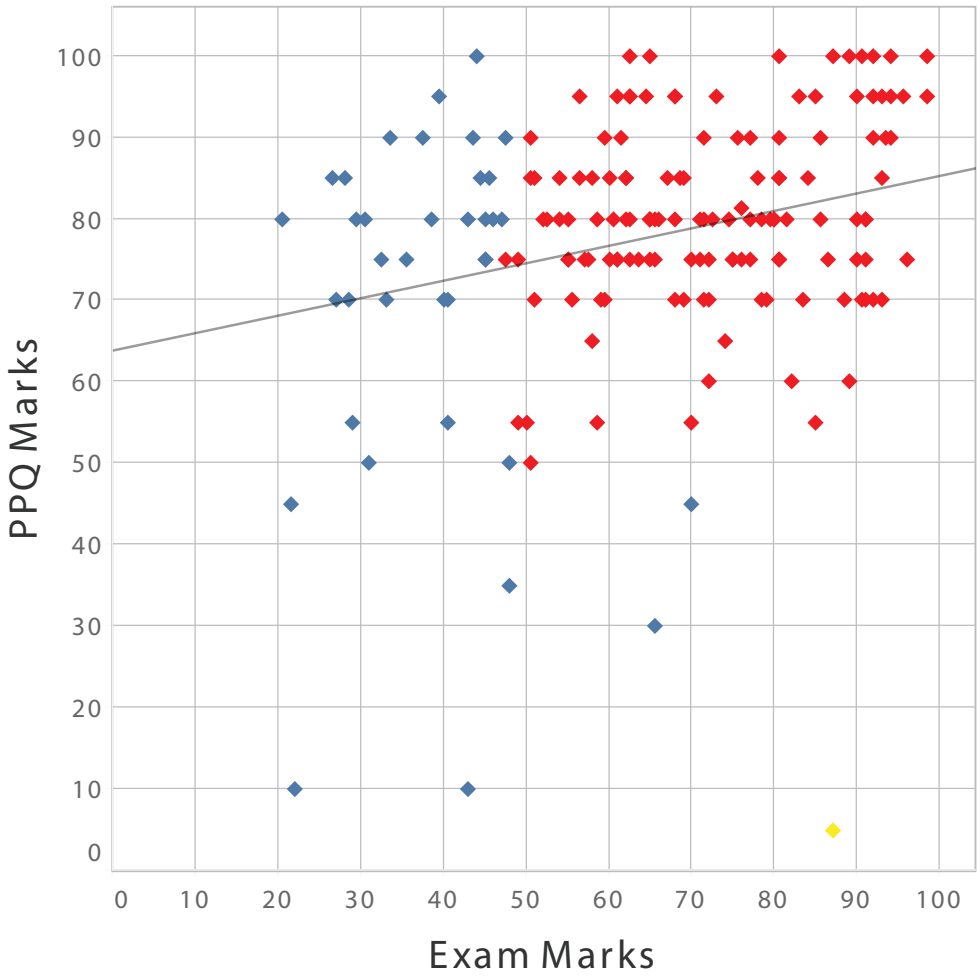


Figure 6.3: Exam Marks vs. PPQ Marks Correlation

pedagogical issues, reasoning about misconceptions, crafting and measuring the effectiveness of new tasks capable of effecting cognitive changes. Our web–based approach demonstrates such efforts can be reduced by capturing and sharing misconceptions and effective tasks between institutions. The courseware, anonymized student data, and the tool for generating PPQ will be made available to instructors on request, thus facilitating a multi–institutional study.

Delay in ethics approval prevented us from offering PPQ early in the first run. In the future, PPQs will be offered in the initial, middle and latter parts of the course offering, which is likely to further improve the learning outcomes. However, learning patterns identified in previous semesters (and not individual student data) will be the main basis for the initial

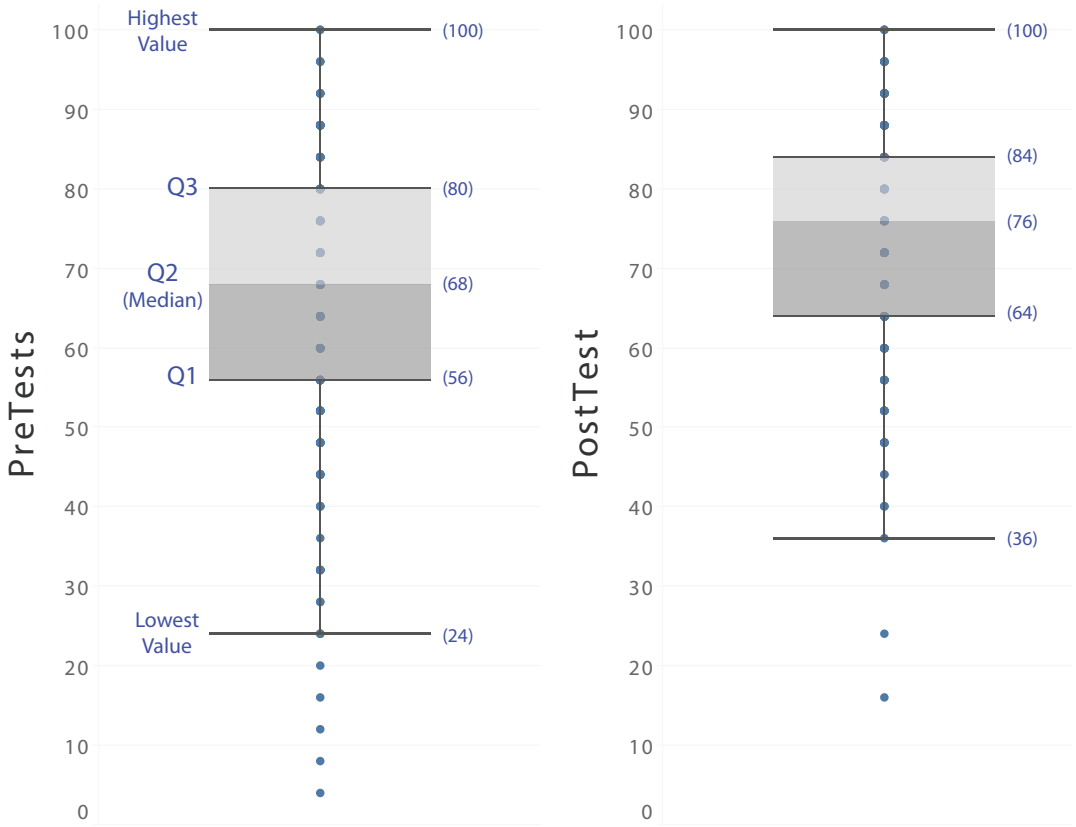


Figure 6.4: Pre–Test(s) Marks vs. Post–Test Marks

PPQ. Another problem was that PPQ criteria of improving self–esteem gradually by varying questions from familiar and unfamiliar concepts did not appeal to some students. To better cater to our student diversity, we intend to give students greater control in customizing their own learning pathways, by specifying ranges for discrimination and difficulty indexes. One limitation of our study was that PPQ and non–PPQ groups were not formed randomly, with self–selecting PPQ students starting with a much lower base in the pretest. However, PPQs appear to be effective in lowering the gap between the two groups, which usually widens with time. We also note that the PPQ, and not simply repeating the quiz, was the main reason for improvement as both groups were allowed to repeat the standard quizzes without any limit. In the future, we will analyze aspects of PPQs such as the time spent per attempt, between attempts and on the number of attempts, as well as the impact of post–test outcomes. We also plan to improve PPQ question selection by considering the impact of specific tasks in bringing about cognitive changes.

Table 6.6: Non–PPQ Takers Performance – Pre– and Post–Test Results (Compared To PPQ Takers)

Sample	Lowest	Q1	Q2 – Median	Q3	Highest
PPQ Takers (Pre–Test Results)	24	56	68	80	100
PPQ Takers (Post–Test Results)	36	64	76	84	100
Non–PPQ Takers (Pre–Test Results)	2	17	46	56	72
Non–PPQ Takers (Post–Test Results)	14	36	56	68	94

6.6 PPQ Extensions

Given students and instructors' feedback towards the first conducted experiment of applying the PPQ, we incorporated several modifications to the PPQ approach to make it more adaptable and responsive to their needs. These amendments can be elaborated in three main categories as per the subsequent sections.

1. Taking into consideration major *Item Analysis* measures for each question designed such as the *Item Difficulty Index* and the *Item Discrimination* to better design quiz questions and effectively address students' capabilities in Section 6.6.1.

2. Design the adaptive and dynamic version of the PPQ approach called the *Adaptive PPQ* which dynamically designs the next question(s) in real–time. Instead of providing the student with a static set of pre–designed questions per quiz in Section 6.6.2, the adaptive PPQ will generate next questions based on the student's response(s) to the previous question(s).

3. Expanding the PPQ approach to incorporate further analytical insights for each student – by plugging–in one novel combined predictive analytics model – and facilitate the future intervention processes in Section 6.6.3.

6.6.1 Incorporating The Difficulty and Discrimination Indexes

To ensure a multiple choice question (MCQ) quiz is well designed, several metrics can be used. *Item Analysis Measure* is one of the most prominent metrics [Gajjar et al., 2014; Hingorjo and Jaleel, 2012; Considine et al., 2005; Sarin et al., 1998]. Using item analysis, our approach incorporated a diverse range of metrics, but we take into account the two main

measures: the difficulty index and the discrimination index. The following provides a short introduction to each one of these critical metrics.

- *Item Difficulty Index (p)* — the proportion of students who answered the question correctly. As it calculates the proportion of examinees' right answers, sometimes it is called the Item Easiness Index. It ranges between 0% – 100% and is represented in the form of percentages. The higher the percentage, the easier the question. The optimal difficulty range is between 30% – 70% as the questions with lower than 30% difficulty index are considered too hard, and questions with higher than 70% difficulty index are categorized as too easy [Linn, 2008; Pajares and Miller, 1994; Hingorjo and Jaleel, 2012]. For each question, the difficulty index can be calculated as the following:

$$p(q_i) = (\frac{U_c(q_i) + L_c(q_i)}{n}) \times 100 \tag{6.1}$$

  where:

  $p(q_i)$ — is the difficulty index (in terms of the percentage) of the i-th question q_i,

  $U_c(q_i)$ — is the number of students in the upper group (high–performing students) who answered correctly to the i-th question q_i,

  $L_c(q_i)$ — is the number of students in the lower group (low–performing students) who answered correctly to the i-th question q_i, and

  n — is the total number of students within a certain class (the total number of examinees)

  Please note that the high– and low–performing student groups can be easily calculated by having the results of all students' responses to the exam (i.e. all questions within the exam), sorting them in the descending order, and dividing them in half (two groups), where the top group includes the high–performing students and the bottom group entails the low–performing students. The high– and low–performing groups usually contain 50% of students each.

- *Item Discrimination Index (D)* — is the measure of how well a question is able to distinguish among those students who are knowledgeable and those who lack parts of the knowledge [Pyrczak, 1973; Kehoe, 1995]. There are several approaches to calculate the discrimination index for a given question, but the most significant one is called

the *point–biserial correlation (PBC)* which takes into consideration the relationship between the student's response to the question (either correct or incorrect), and their performance in the exam (exam mark/score) [Attali and Fraenkel, 2000; Hingorjo and Jaleel, 2012]. For a question with higher degrees of discrimination, we expect that students who performed well in the exam responded correctly to the question, while those who performed poorly in the exam are expected to respond incorrectly to the question. The item discrimination index ranges between [-1.0, 1.0]. In other words, it is expected that the students in the high–performing group answer a particular question correctly more often compared to the students in the low–performing group. If the result was as expected, then the question is considered the one with the *positive discrimination index* which translates into the $D(q_i)$ to be in the range [0, 1]. In case of the results to be the reverse, meaning the low–performing students respond correctly to particular questions more often than the high–performing students, then the question is considered to have a *negative discrimination index* which will be a figure in the range of [-1, 0]. There are three simple steps to calculate the discrimination index $D(q_i)$ for each question described as follows.

Step 1) — grade students exam; then rank them based on their performance in the exam from the lowest to the highest. then differentiate the bottom 27% and the top 27% of students to form the low–performing and high–performing student groups (the middle–group students and those who have not participated in the test will be excluded).

Step 2) — for each question q_i, calculate its discrimination index $D(q_i)$ for the low–performing and high–performing student cohorts (excluding the students in the middle–of–the–road), given the following formula:

$$D(q_i) = \frac{U'_c(q_i)}{n_u{}'} - \frac{L'_c(q_i)}{n_l{}'} \tag{6.2}$$

where:

$D(q_i)$ — is the discrimination index (in terms of the percentage) of the i-th question q_i,

$U'_c(q_i)$ — is the number of students in the upper group (high–performing or the top 27% students) who answered correctly to the i-th question q_i,

$L'_c(q_i)$ — is the number of students in the lower group (low–performing or

the bottom 27% students) who answered correctly to the i-th question q_i,

$n_u{}'$ — is the number of students in the high–performing group,

$n_l{}'$ — is the number of students in the low–performing group, and

Please note that in cases where the number of students in both high– and low–performing groups is exactly the same, the formula is simply converted to:

$$D(q_i) = \frac{U'_c(q_i) - L'_c(q_i)}{n'}; \quad n_u{}' = n_l{}' = n' \tag{6.3}$$

Please also note that this process is similar to calculating the *item difficulty index* for the question q_i for the high–performing (the top 27%) and low–performing (the bottom 27%) students and then subtracting the results as in:

$$D(q_i) = p_{U'}(q_i) - p_{L'}(q_i). \tag{6.4}$$

where:

$p_{U'}(q_i)$ — is the *difficulty index* of the high–performing students for question q_i, and

$p_{L'}(q_i)$ — is the *difficulty index* of the low–performing students for question q_i

Step 3) — analyze the calculated $D(q_i)$ for question q_i. For a good question in terms of discriminating between the high– and low–performing students, the $D(q_i)$ will be in the range [0.4, 0.6]. This means that the question is doing a reasonable job in differentiating between the high–performers and the low–performers. The closer the $D(q_i)$ to 1, the more discriminating the question q_i is between the high– and low–performing students, and vice versa. Questions with $D(q_i)$ values between 0 and 0.2 are considered poorly discriminating. For example, the $D(q_i) = 0$ means that all student cohorts are performing the same (high– and low–performing students are getting that question right or wrong similarly)! This could mean that the particular question is either very difficult (that everybody gets it wrong), or is too easy that all students could answer it correctly. In situations like these, we should investigate the objective of the question.

We incorporated both the *item difficulty index* and the *item discrimination* within the recent version of the PPQ application to help instructors easier distinguish certain student

cohorts' performance on each question and update the question if necessary. Furthermore, students can be allowed to select questions based on represented difficulty or discrimination indexes.

6.6.2 Adaptive PPQ

Another improvement to our base PPQ approach, given students and lecturers' feedback, is the introduction of the *Adaptive PPQ* approach which aims at satisfying the following objectives:

- *Providing each student with the real–time generated questions within a certain quiz, based on their responses to the very recent question, instead of providing each student with the pre–defined sets of individualized questions.* Although the base PPQ approach generated personalized sets of questions based on each student's past performance within a particular subject (by identifying their misconceptions and providing them with questions covering those misunderstood concepts), the solution still lacked dealing with students' instant responses to the quiz questions to adaptively generate next questions based on their answers within that particular quiz. With the introduction of the adaptive PPQ approach, this gap is addressed and the subsequent questions within the same quiz will be adaptively designed based on the student's real–time response to the current question.
- *Identifying and demonstrating each student's "Root–Cause Concept(s)".* Root–cause concepts are critical to developing solid knowledge as subsequent concepts are usually constructed based on them. Therefore, targeting and rectifying root–cause concepts become of crucial importance.

The *adaptive PPQ* approach relies on the base PPQ approach to generate individual questions per student, but its policy to target the misconceptions and providing the student with the next question within a given quiz is more adaptive and novel. As mentioned above, the adaptive PPQ approach is based on the previously proposed standard PPQ approach in Section 6.3. It means that the "adaptive PPQ" is another implementation of the proposed framework's "Logical Layer" processes in the "Physical Layer". The overall connection among several layers and processes within the proposed framework (from "Physical" to "Logical" to "Conceptual" layers) was also depicted in Figure 7.1.

Algorithm 2 Adaptive PPQ

```
1: procedure ADAPTIVE-PERSONALIZED-PRESCRIPTIVE-QUIZ
2:     AdaptivePPQ ⟵ ∅
3:     RootCause ⟵ ∅
4:     for each s_i ∈ S do {
5:         input:  s_i, mc_i, n_i, mfq_i, ppq_i, CG
6:         output:  dppq_i, rootCause_i for s_i
7:         topConcept ⟵ ∅
8:         nextConcepts ⟵ ∅
9:         responseCorrectness ⟵ FALSE
10:        dppq_i ⟵ ∅
11:        rootCause_i ⟵ ∅
12:        while |dppq_i|< n_i do {
13:            while mc_i ≠ ∅ do {
14:                topConcept ⟵ pop(mc_i)
15:                dppq_i ⟵ ppq_i(s_i, topConcept, 1, mfq_i, ∅, ∅, ∅)
16:                responseCorrectness ⟵ collectedResponse(q_i, s_i)
17:                if (responseCorrectness == TRUE) then
18:                    continue
19:                else{
20:                    if (isRootCause(CG, topConcept)) then {
21:                        rootCause_i ⟵ topConcept
22:                    }
23:                    nextConcepts ⟵ parentConcepts(CG, topConcept)
24:                    nextConcepts ⟵ prioritizeConcepts(nextConcepts)
25:                    push(mc_i, nextConcepts)
26:                }
27:            }
28:        RootCause ⟵ rootCause_i
29:        AdaptivePPQ ⟵ dppq_i
30:        }
31:    return AdaptivePPQ, RootCause
32:    }
```

According to the Algorithm 2, the system is capable of generating and disseminating the *adaptive PPQ* per student. The steps towards this goal is elaborated in the following:

- The *Adaptive PPQ* gets the following data as its input:

  S — the list of all enrolled students in a given subject. Each individual student is targeted as s_i, where $s_i \in S$.

  mc_i — the list of misunderstood concepts per student s_i.

  n_i — the number of questions per quiz ppq_i per student s_i. The default is 15 as per Section 6.3.2.

  mfq_i — the set of fresh questions covering the misconceptions of student s_i.

  ppq_i — the set of questions specifically designed for student s_i, according to Algorithm 1.

  CG — the concept graph which entails all thought concepts within a particular subject in a certain semester along with their interrelationships (such as which concept(s) is(are) the parent(s) of a given concept – the "IS–A" relationship).

  and the following as its output:

  $RootCause$ which is a set of $rootCause_i$s per each student s_i, demonstrating their root concept(s) causing their misconceptions.

  $AdaptivePPQ$ which is a set of $dppq_i$s for each student s_i, representing the updated sets of questions within a particular quiz to be asked based on the student s_i's response to the current question. The $dppq_i$ guarantees the instantaneous and real–time generation of personalized questions for each student.

- The algorithm simply iterates over each student s_i and generates the dynamic subsequent questions based on their current responses as follows:

  – The *adaptive PPQ* pops the top element (concept) from the student's misunderstood concept stack (i.e. the mc_i).

  – Having the top concept (referred to as the *topConcept*, the *adaptive PPQ* algorithms calls the base *PPQ* algorithm (Algorithm 1) to generate only one question for the student to be asked by calling the ppq_i method
  $ppq_i(s_i,\ topConcept,\ 1,\ mfq_i,\ \emptyset,\ \emptyset,\ \emptyset)$,
  where:

s_i — is the student taking the PPQ.

topConcept — is the concept that has just been popped up from the mc_i stack.

1 — is the size of the question set to be retrieved (only one question should be returned in this case).

mfq_i — the set of fresh questions covering student s_i's misconceptions.

the first $\emptyset$ — refers to the muq_i (please refer to Section 6.3.2), and in this case means we do not have any of those questions.

the second $\emptyset$ — refers to the uuq_i (please refer to Section 6.3.2), and in this case means we do not have any of those questions.

the third $\emptyset$ — refers to the $allmuq$ (please refer to Section 6.3.2), and in this case means we do not have any of those questions.

Altogether, the *adaptive PPQ* seeks for only one question from the PPQ algorithm to ask the student and stores it in the $dpppq_i$ set.

– The *collectedResponse*(q_i, s_i) method evaluates the student s_i's response to the question generated by $dpppq_i$ (which is q_i). If the result is true, meaning that the student s_i answered to the question correctly, the system will proceed with generating the next question by popping up the next misconception from the mc_i stack. In this case, the *adaptive PPQ* and the *base PPQ* approaches are performing the same. However, if the student s_i responds to the question q_i incorrectly, the following steps are followed:

1. First of all, the system will check whether the misconception – here, the *topConcept* – is a root–cause, by calling the ***isRootCause*** method
   $isRootCause(CG,\ topConcept)$,
   where CG is the concept graph, and *topConcept* is the concept that the student s_i just responded to its covering question q_i incorrectly. The mechanism to detect whether a given concept is root–cause relies on two main conditions, given the concept graph CG: (1) is the concept *topConcepts* the root concept in the concept graph? This means that if a concept does not have any parents in the CG, then it is considered as the *root concept*. (2) if the concept is not a root concept, has the student s_i responded to its parent concept(s) correctly? This guarantees that the corresponding concepts in the higher levels were understood properly by the student, and the current concept in the CG is the

highest level concept that the student misunderstood. Please note that for each student there could be more than one root–cause concept (if any).

2. Next, all the parent concepts of the misunderstood concept (*topConcept*) will be retrieved, given the concept graph CG, by calling the ***parentConcepts*** methods
   $parentConcepts(CG,\ topConcept)$.
   The result is stored in *nextConcepts* to be used in the next step.
3. Now that we have the list of parent concepts of the misconception, we need to prioritize them based on the concepts that are residing the highest within the concept graph. This task is accomplished by calling the ***prioritizeConcepts*** method
   $prioritizeConcepts(nextConcepts)$.
   The prioritization is performed by considering the level of each retrieved concept within the given concept graph. The higher the concept level, the more important the concept is. The method takes into consideration concepts interrelationships (the "IS–A" relation) in that it prioritizes concepts that are parents of others.
   At the end, the list of prioritized concepts will be stored in *nextConcepts* set.
4. The final step is pushing the calculated concepts to the mc_i stack by calling the ***push*** method
   $push(mc_i,\ nextConcepts)$.

- Now that we have all the ingredients, we can build the *RootCause* and *AdaptivePPQ* sets by collecting the information provided per each iteration (for each student s_i). The system will return both sets to the students to clearly identify which concepts were among the misconceptions and where they did wrong.

6.6.3 Enhancing PPQ Intervention Incorporating Descriptive and Predictive Analytics

In the introductory programming, we incentivized the use of non–assessed PPQs and quizzes by interleaving them with assessed tests covering the same topics. We also awarded bonus marks to encourage students to attempt the quizzes until they attained the required standard (70%). The qualitative feedback revealed 91% of the students using the PPQs felt more confident. In this section, we describe the techniques we have devised to enhance the

self–efficacy of students by combining PPQs with descriptive and predictive analytics. In our system, a fine–grained performance summary is provided in individual concepts, a form of descriptive analytics. Such students were then able to focus on topics and concepts where most of the misconceptions occurred. The predictive analytics allowed students to project their overall performance at any stage of the course. The more PPQ attempts a student made the better the expected performance in the overall progress occurs. While descriptive, predictive, and prescriptive analytics provided students the ability to self assess their performance, our research was also interested in how students perceive the tasks and their own progress. The enhanced model, therefore, allows students to give feedback on any task, which included the quality of questions and the explanation of solutions allowing instructors to improve the tasks. Students were asked to rank the tasks as very useful, useful, neutral and confusing. These data also allowed the instructors to create more useful questions for specific concepts and student cohorts. We have also incorporated an assessed reflection task in the system, which is enabled after students complete the required PPQs. The students are asked to respond to their strengths, weaknesses, their main conceptual difficulties, and how they plan to address their conceptual difficulties. These tasks provide instructors with qualitative feedback in addition to promoting student self–reflection. Our proposed iterative and incremental composite analytics model is shown in Figure 6.5. Each round of the iterative process represents 2 weeks period and incorporates the descriptive, prescriptive, and predictive analytics which ends up with students' feedback and self–reflection. The upcoming rounds are built on top of previous rounds as the semester proceeds. The techniques for predictive analytics extending prior work are presented in the rest of this section.

We incorporated the following predictive analytics algorithm based on their significant impact in the body of research in the context of education [Larusson and White, 2014; Papamitsiou and Economides, 2014; Berland et al., 2014; Martin and VanLehn, 1995; Storey et al., 2003; Baker and Inventado, 2014; Siemens and d Baker, 2012]: *Naïve Bayes (NB)*, *Neural Network (NN)*, and *Random Forest (RF)*. Involving the predictive component within the framework provides more benefits to both lecturers and students. For example, by analyzing the early performance of each student, the system can extrapolate the likelihood of passing/failing of that student prior to the end of the semester and helps the lecturers to devise relevant intervention mechanisms accordingly.

In this part, we mention two different experiments conducted on students' data using predictive methods:

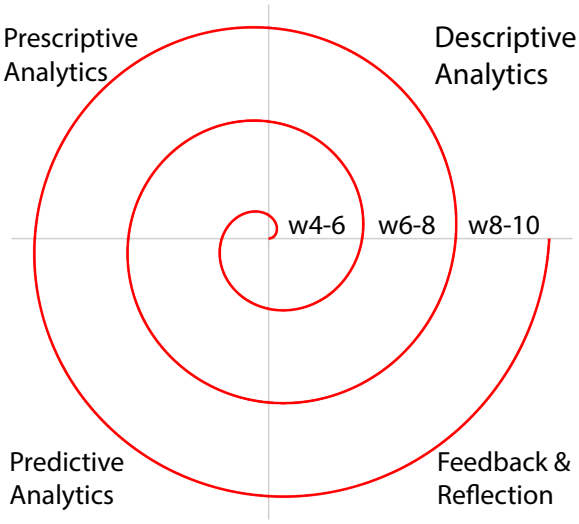


Figure 6.5: Enhanced PPQ Incorporating Descriptive, Predictive, Prescriptive Analytics In An Iterative And Incremental Manner

1. *Applying each predictive algorithm separately on the student data to assess the effectiveness and accuracy of them in projecting their pass/fail likelihood* — As mentioned earlier, we adopted three predictive algorithms for this purpose – the naïve bayes, the neural networks, and the random forest. For this part, we collected several data–sets in terms of students interactions with the learning management systems and the provided web–based assessments. These data–sets included 69 (ITS takers), 185 (PPQ takers), and 241 (ITS, PPQ, in–class Online Test, and Written Tests) students, respectively. The first two data–sets took into account students' test scores resulted by their interaction with the online (web–based) assessment applications such as ITS and PPQ. The only difference was regarding the number of students participated in those assessments. The third data–set, however, considered more data elements including the online assessment tools (ITS and PPQ), the in–class online tests, and the written in–class tests. Given that the majority of students participated in those assessments, the number of students in the last data–set was the highest compared to the other two. Table 6.7 represents the results of each applied predictive approach on the first (69 students) and second (185 students) data–sets. Table 6.8, on the other hand, represents the results of applying different predictive models on the more reach data–set with 241 students. It seems that as the number of elements per data–set increases, so does the accuracy of the random forest approach. Also, one can find the accuracy of the neural nets

Table 6.7: Applied Different Predictive Analytics Results – ITS and PPQ Only

Predictive Analytics Algorithm	**Number of Students**	**Accuracy (%)**	
		10–fold CV[1]	**90% Train[2]**
Naïve Bayes	69[4]	98.52	**100**
Neural Networks	69	95.58	**100**
Random Forest (100–DTs)[3]	69	97.05	**100**
Naïve Bayes	185[5]	95.67	**100**
Neural Networks	185	98.91	**100**
Random Forest (100–DTs)	185	99.45	**100**

[1] 10–fold cross–validation technique on the predictive model accuracy estimation [Refaeilzadeh et al., 2016; Kohavi et al., 1995; Zhang, 1993; Fushiki, 2011]. [2] 90% Training Set technique [Foody et al., 2006; Jaworska et al., 2005; Boser et al., 1992].
[3] The Random Forest technique with 100 Decision Trees [Pal, 2005; Svetnik et al., 2003]. The 100 decision trees were picked based on the system's performance in converging to the optimal number. According to our conducted experiments, 100 trees were sufficient to prevent outliers such as false positive and false negative.
[4] The number of students who took ITS.
[5] The number of students who took PPQ.

Table 6.8: Applied Different Predictive Analytics Results – ITS, PPQ, In–Class Online Test, and Written Test

Predictive Analytics Algorithm	**Number of Students**	**Accuracy (%)**	
		10–fold CV	**90% Train**
Naïve Bayes	241	97.51	**100**
Neural Networks	241	98.34	95.83
Random Forest (100–DTs)	241	**100**	**100**

approach is increasing with growing the data–set size as well. Altogether, almost all three predictive techniques perform well with the 90% training set technique with 100% accurate projections.

2. *Combining the three mentioned algorithms in one hybrid predictive model and apply it on different aspects of student data to predict more complex sets of data such as students' pass/fail likelihood, their knowledge level estimation, and the likelihood of students to withdraw the subject* — is further elaborated as one future work in Section 7.2.

6.7 Tailoring Quizzes to Specific Cohorts

Educational institutions are catering to increasingly diverse cohorts in terms of age group, culture, background, and experience, which are known to impact how they learn and progress through the course. These groups may exhibit different learning patterns, with different groups responding differently to different types of intervention. Studying and classifying the learner behaviors through analytics can help to tailor the learning instruments to improve the overall learning outcomes. While weekly quizzes and personalized quizzes appear to have benefited the vast majority of students the progress has not been uniform for different groups. This section attempts to identify different learning patterns with the aim of improving the design of personalized quizzes in the future.

We chose the software engineering course, as it is a core for all the students in the computer science and IT discipline. Over 500 students from different levels (postgraduate and undergraduates) and different disciplines (games, computer science, software engineering and IT) take this course offered in two semesters. Another reason for choosing this course is because it imparts many of the soft and technical skills needed for employability. The quizzes in this course were therefore classified into different concepts and cognitive levels defined by the Blooms taxonomy. The pedagogy exploited intrinsic and extrinsic motivation to improve learning outcomes; the former by ungraded quizzes while the latter by class design tests and the final exam. Moreover, quizzes were made to cover the same concepts as the class tests that followed them and our analysis attempted to study whether combining intrinsic and extrinsic motivation lead to improved learning outcomes.

The learning outcomes at synthesis–level constructivist tasks were measured through volunteer participation, where students were asked to take a pre– and post–test together with design tasks through the active learning tool. Both tests covered Comprehension, Application and Analysis level questions with similar difficulty levels. The pre– and post–test results were released only after the completion of the active learning tasks. Only 68 of the 94 students who volunteered initially completed both design tasks and tests.

Our results demonstrated that constructivist tasks can improve the performance of all students. The number of students in the 90–100% range went up from 0 to 6, while those failing dropped sharply from over 30 to less than 15. Based on these values, we reject the null hypothesis "No significant improvement after the active task" as the *p-value* of 0.000269 is less than 0.01 (with the standard confidence of $\alpha = 0.05$).

To study whether active learning has a long–term impact, Figure 6.6 compares the per-

formance of the study group with the control group (the remaining 172 students) in the 4 manual tests (use–case, class, activity, sequence), and the exam (sequence). Note that the study group, which had 13% weaker performance than the focus group prior to the intervention, performed 4% better in the subsequent test and the final exam.

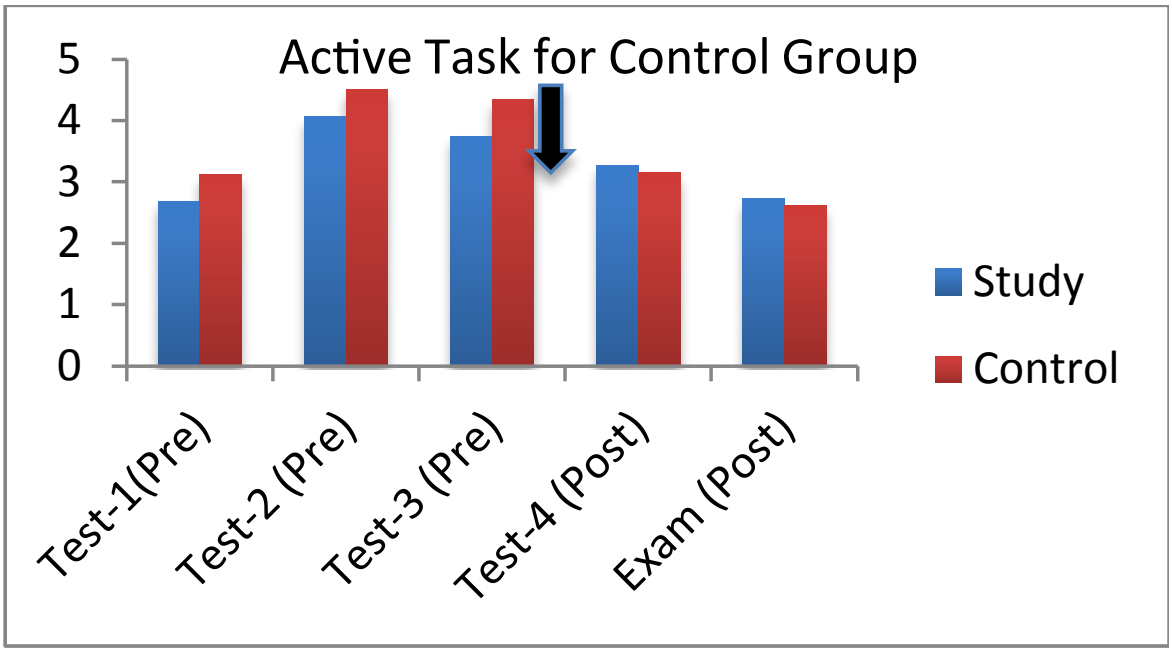


Figure 6.6: Long–term Impact of Constructivist Tasks

Learner characteristics is the main factor in determining the success of formative feedback [Shute, 2008]. Initially, we isolate and analyze engagement and performance levels for groups with major differences in background and prior experience. Specifically, we study the differences between undergraduates (UGs) and postgraduates (PGs). In the second part, we use learning analytics to analyze engagement and performance patterns to extract different student clusters. Figure 6.7 compares the performance of students in their exams and quizzes. The X–axis represents the exam marks for design, the Y–axis the sum of all quiz marks adding up to 45, and the Z–axis is set to either 1 for UGs or 2 for PGs. PGs (2 in the Z–axis) evidently performed better than UGs with the most quiz and exam marks in the top right quadrant. Moreover, many UGs have less than 15 for design in the exam while PGs have none.

Figure 6.8 is color–coded to visually depict the differences between UG and PG engage-

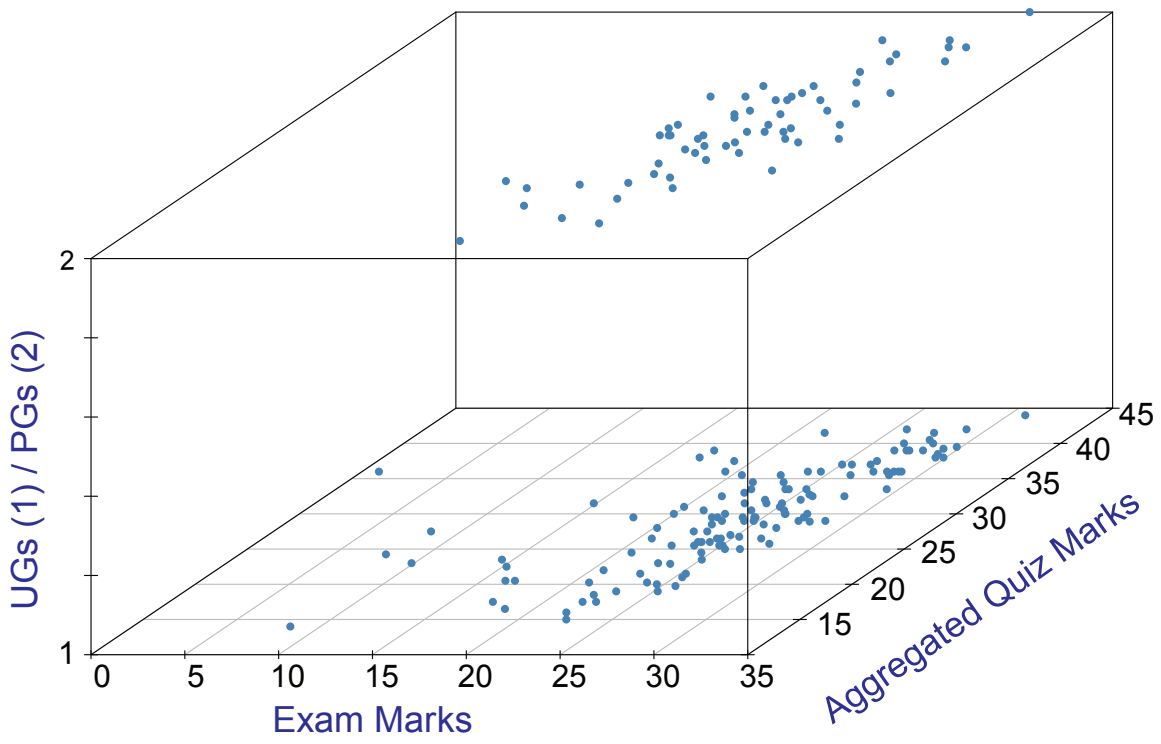


Figure 6.7: The Overall Undergraduates vs. Postgraduates' Performance in Quizzes and The Exam

ment and performance, where UGs and PGs are plotted in blue and red, respectively. PGs have only 5 students with exam marks below 20, while UGs have 18. It is evident that PGs perform better in tests with only two students getting less than 50%. We posit the main reason was their level of motivation, exemplified by over 90% doing the quizzes regularly. UG performance in quizzes, tests, and exams appears to be very diverse. The strong correlation between poor coursework (quizzes and tests) and the final exam makes it possible to identify at–risk students early.

Next, we investigate students' performance in multiple assessments to arrive at different clusters, using the k–means algorithm. We attempted a different number of clusters for the k–means but found 5 to be the optimal number, as with others using different data–sets [Nyroos et al., 2016]. Figure 6.9 illustrates the relation between the quizzes and the exam marks, along with their trend lines. The student cohort in cluster 1 performed below average in both the exam and class quizzes. These students are probably the UGs who are lacking motivation and are not engaging in class and project activities. A common reason cited by such students is the high level of abstraction in software engineering. Such students are likely

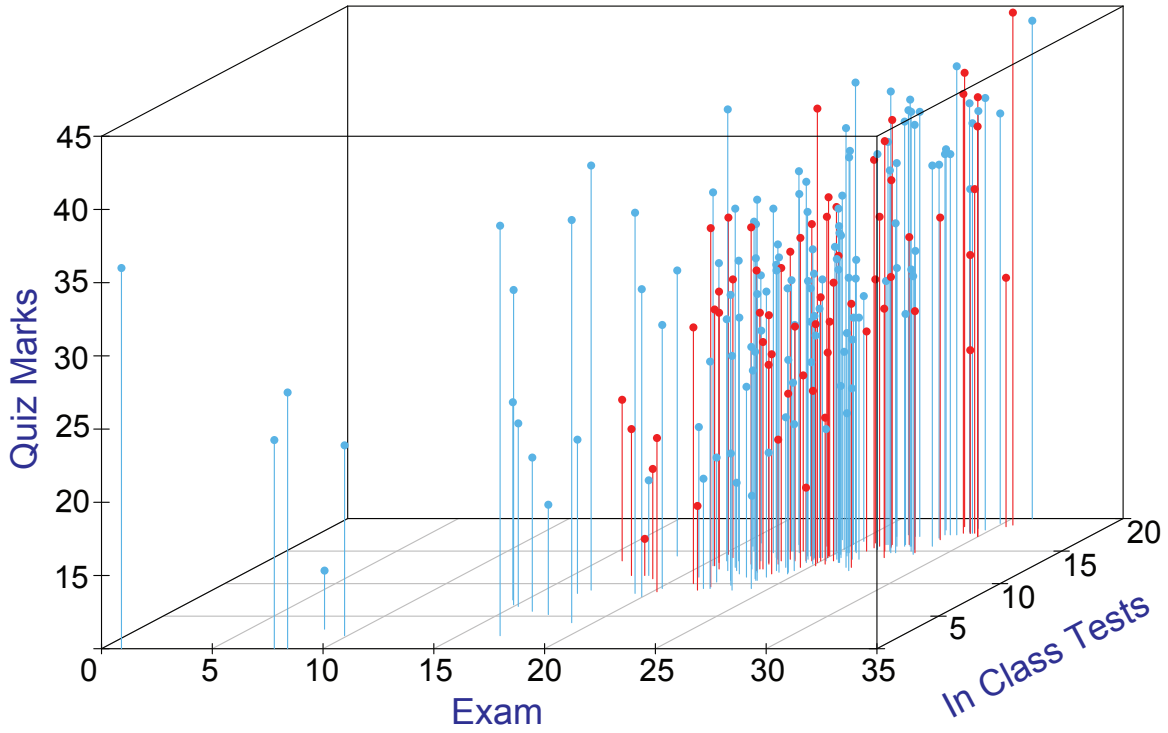


Figure 6.8: UG vs. PG – Performance Comparison in Quizzes, In–class Tests, and the Exam

to benefit from enrolling in additional programming courses prior to software engineering.

Students in cluster 5 performed very poorly in their exam (as per the sharp plummeted green trend line) though they managed to do well in their quizzes when compared to cluster 1 students. These students are likely to have taken the quizzes several times to get the answers correct, but without paying attention to feedback. Students in cluster 4 performed above average in their exams though faring poorly in their quizzes. These are probably the students who are not disciplined enough to do the quizzes regularly and therefore not reaching their full potential. Cluster 2 students are likely to be the student group who are well motivated to complete all the quizzes and performed above average in the exam. Cluster 3 represents the highly motivated group who benefit the most from all the formative feedback while suggesting improvements and offering their own questions. These students naturally performed the best in the exam.

The correlation between the exam and in–class tests are depicted in Figure 6.10 along with their trend lines. The students were grouped into 5 clusters as before and the clusters were similar though a stronger correlation was observed between the exam and tests. Generally,

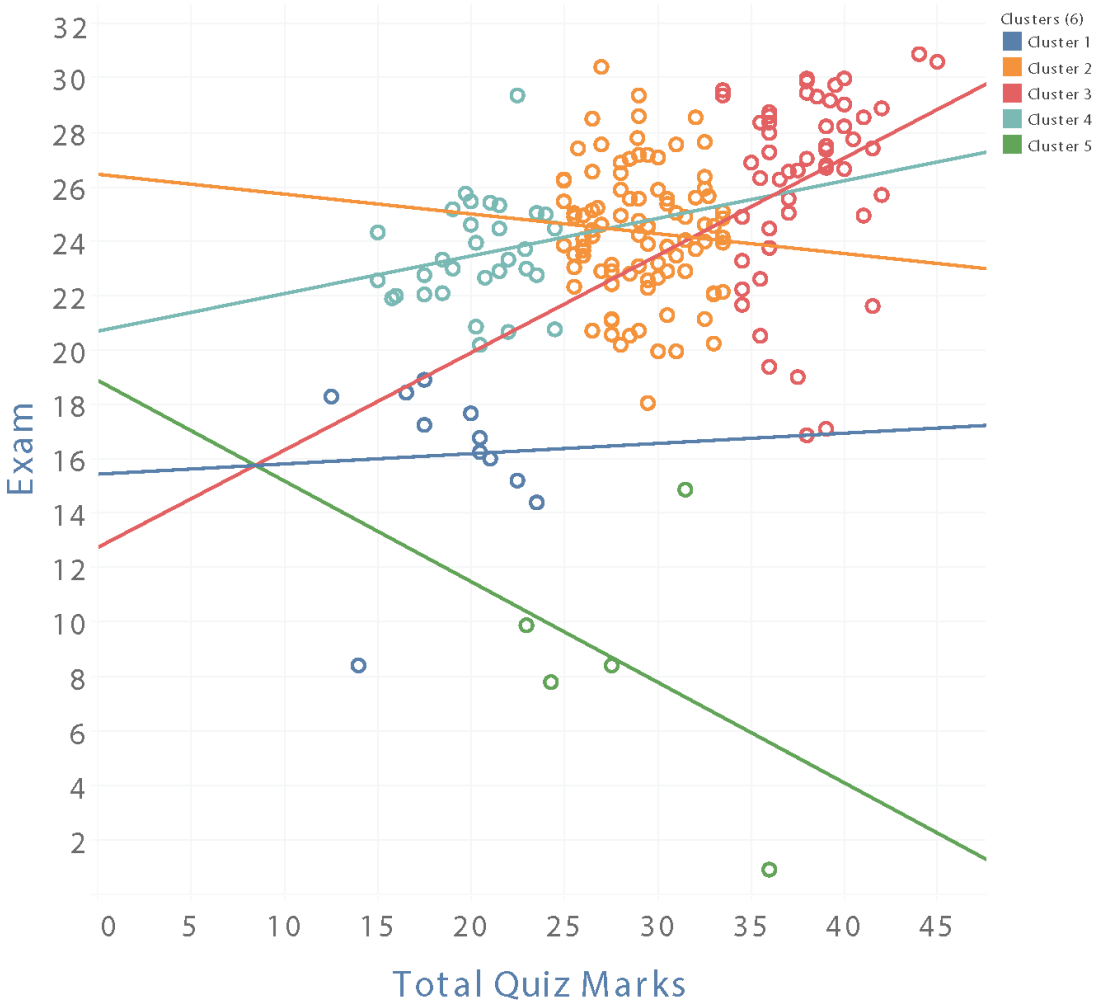


Figure 6.9: The Exam Marks vs Quiz Marks – Cluster of 5 Student Cohorts

students who did poorly in their in–class tests did poorly in their exams and vice–versa. The stronger correlation could be the result of many factors. Firstly, students were allowed to take the test only once, and under exam conditions thus the closer similarity is consistent. Secondly, most students took the tests only after doing the quizzes and therefore the design activities also benefited from the many cognitive tasks.

To sum up, learning analytics has been used to predict student results, identify areas of misconceptions and track student engagement levels. In our work, learning analytics helped to identify the extent to which formative assessments foster students' design skills. It is evident, manual tests, with high correlation with exam performance, play an important role in building up good design skills. However, formative quizzes appear to play an equally

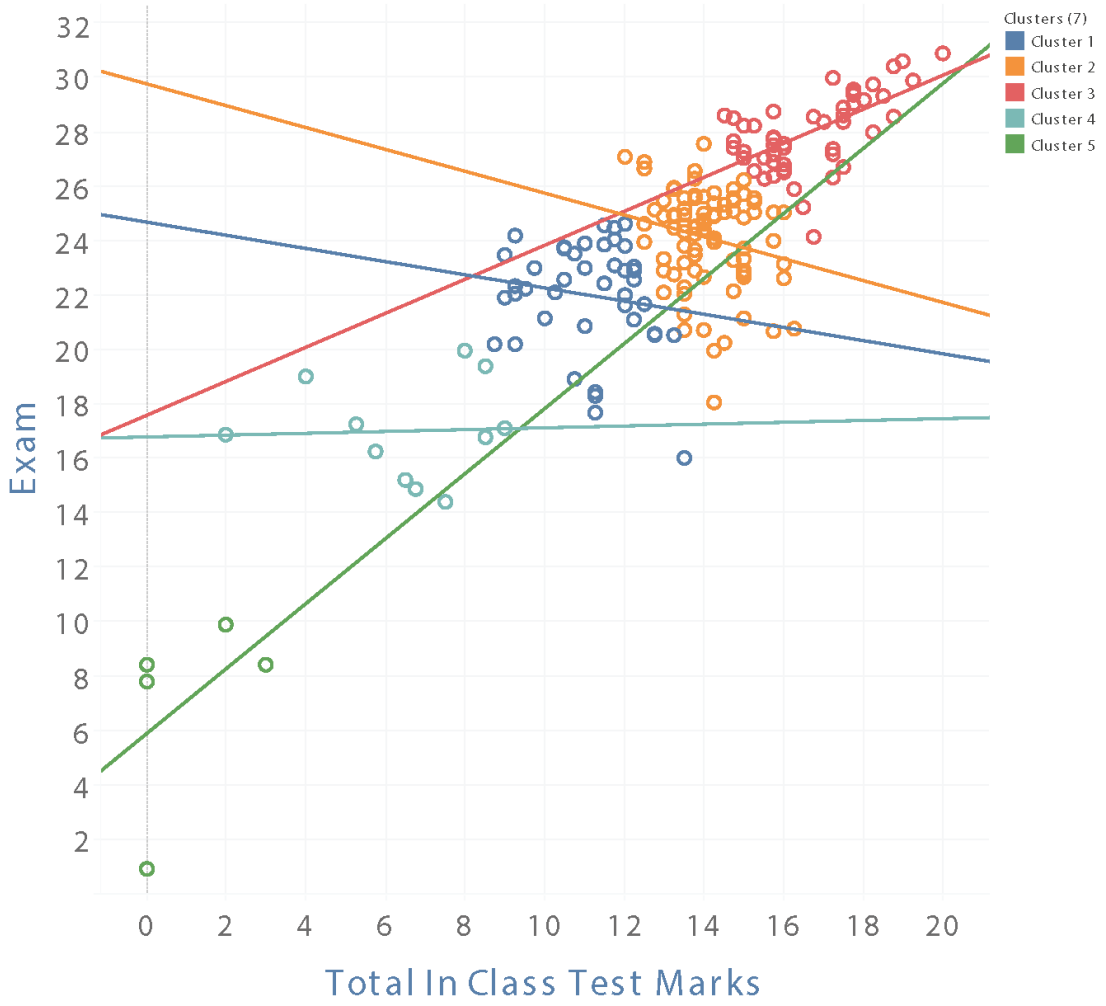


Figure 6.10: The Exam Marks vs In–Class Test Marks – Cluster of 5 Student Cohorts

vital role in preparing students for manual tests and the exam. For example, a formative quiz highlighting coupling and cohesion or contrasting distributed and centralized design can help students come up with better designs. Our postgraduate students with much higher quiz completion rates have also scored 10% higher in the exam. Our analysis also reveals performance in exam correlates well with formative quizzes at higher cognitive levels. It is therefore beneficial to develop and share such tasks in software engineering. Many such tasks span across multiple domains, common with most authentic design problems in software engineering.

Active learning tools combining a constructivist approach with immediate and holistic feedback appear to improve both short and long–term learning outcomes. Automated forma-

tive feedback has many advantages over manual marking. Firstly, the marking of paper–based designs can be very subjective as marker perceptions may vary. Other benefits include instant feedback and reduced marking costs. Our active learning tools, though effective, need further work to make them more open–ended. Crafting pedagogically sound active learning tasks require much initial effort, but these tasks can be reused and shared without any additional costs.

Formative feedback can be more beneficial when tasks are designed considering the type of learners. Our analysis of student performance across quizzes, tests, and the exam reveals a number of student clusters with common characteristics. From our analysis, class tests appear to be more important for our diverse undergraduate students, as they are less motivated to self–learn through quizzes. With increasing diversity of incoming undergraduates, learning analytics is likely to play a vital role in matching student needs to educational resources.

6.8 Summary

In this Chapter, a personalized prescriptive quiz (PPQ) approach was introduced as one implementation of the *physical layer* to put together all the required ingredients of the proposed analytics–driven framework of Chapter 5. By doing so, the fourth research question

> ***Research Question 4)***
>
> ***How do we devise and link the physical layer components enforcing higher–level processes (linking the physical, logical and conceptual layers altogether)?***

was addressed by collecting students test results within a particular semester using the web–based PPQ approach, and providing the students with the individually designed set of questions (quizzes) to help them identify and then rectify their misconceptions.

In addressing the 4th research question, we made a number of contributions and devised a new integrated framework for learning analytics highlighted as follows.

- *The personalized prescriptive quiz approach* — we made automatic remediation for large cohorts of novice programmers possible which resulted in both short and long–term gains and greatly improved student self–efficacy. The PPQ approach also made it possible to identify and correct invalid mental models blocking further progress. The novelty of our model is that it allows dependencies between concepts to be embedded

into the system allowing questions to be generated based on student responses. Incorporating difficulty and discrimination indexes allow diverse students to specify the most appropriate learning pathway. Average students may initially choose questions with low difficulty and discrimination index before proceeding to more difficult ones. The learner characteristics identified allow PPQs to be adjusted for specific cohorts based on their study patterns and motivational levels.

- *The iterative and incremental enhanced PPQ combining descriptive, prescriptive, predictive, feedback and reflection components* — the enhanced PPQ approach incorporating main analytics techniques provides both instructors and students with invaluable feedback and analyses to improve their performance throughout the semester.

- *The physical layer combining prescriptive, predictive and descriptive analytics components provides a systematic approach to translate logical layer processes into physical layer modules in the learning and teaching domain.*

Later, by taking into consideration the valuable feedback from the students and lecturers participated in PPQ experiment, and to address further requirements of a dynamic and adaptive learning analytics solution, the base PPQ approach was expanded incorporating the following features:

- *Incorporating the item difficulty and discrimination indexes* — to provide more enriched sets of questions to each student.

- *Adaptive PPQ* — to make the base PPQ approach more adaptive, dynamic and further adapted to the real–time responses of the students to the previous questions.

- *Combined predictive model* — to incorporate the capacity of several predictive analytics techniques and help both students and lecturers project their likely performance at the end of a certain semester and provide the ground for effective and informed interventions.

- *Finalizing the overall analytics–driven framework of Chapter 5* — to frame the whole analytical solution to address the major requirements of learning analytics.

In the next chapter (Chapter 7), we revisit the research questions of this thesis and our main contributions, along with the future research directions of the work.

Chapter 7

Conclusions and Future Work

> "A story really is not truly a story until it reaches its climax and conclusion."
>
> *Ted Naifeh*

This thesis proposed a three–layered integrated analytics framework to address the key needs requirements of educational institutions towards adopting learning analytics solutions. The framework incorporates the conceptual, logical, and physical layers. The conceptual layer includes a federated composite analytics architecture (comprising descriptive, predictive, and prescriptive analytics) to cover the high–level analytics requirements of educational institutions. The logical layer incorporates the 10 key learning analytics processes. For the physical layer, we demonstrated one specific formalization/implementation of the framework (the PPQ approach) and linked it to other layers. We further evaluated our framework using real student data in different course offerings, and the results demonstrated the positive impact of our approach on improving student knowledge, self–esteem, and academic experience.

The following section (Section 7.1) revisits the key research questions and presents the contributions which address the research questions. Section 7.2 presents some related future improvements and research directions.

7.1 Research Objectives Revisited

- *Research Question 1) How do we design an integrated and adaptive analytics architecture?*

An adaptive and composite analytics architecture incorporating key analytical techniques (descriptive, predictive, and prescriptive) with dynamic feedback mechanisms was introduced to address this research question. The architecture is elaborated upon in Section 4.3 of Chapter 4. The novelty of the work can be depicted as: (1) incorporating three analytics techniques in one seamless architecture working together with their interrelationships, (2) constructing the prescriptive analytics part with simulation, optimization, and evaluation components and linking them together, and (3) producing the adaptive and timely courses of actions by providing certain feedback lines among analytics components.

- *Research Question 2) How do we incorporate the proposed integrated analytics architecture in the context of learning analytics (proposing the analytics framework for learning analytics)?*

  Key learning analytics processes were captured and a four–dimensional learning analytics reference model (LARM) was introduced in Chapter 3, to address this research question. The LARM's dimensions were designed as (Section 3.3): (1) what? – that accounts for collecting, recognizing and analyzing diverse data types, (2) who? – which reflects the main stakeholders of the system (students, instructors, and educational institutions), (3) how? – that refers to certain analytics techniques deployed to analyze educational data, and (4) why? – which mentions the main learning analytics requirements and the processes responsible for them. We captured 10 major learning analytics processes and illustrated their interrelationships in Section 3.3.

- *Research Question 3) How do we formalize learning analytics processes in the proposed framework (connecting learning analytics and prescriptive analytics components)?*

  By specializing the composite analytics architecture in the context of education, an integrated analytics framework comprising the conceptual, logical, and physical layers was proposed to address this research question. The framework is elaborated in Section 5.2 of Chapter 5. The proposed composite analytics architecture in Chapter 4 constructs the conceptual layer of the framework. All key captured learning analytics processes (10 processes) are implemented in the business process model and notation (BPMN) specification and put within the logical layer. Each component of learning analytics processes in the logical layer was also linked to their corresponding conceptual layer components.

- *Research Question 4) How do we devise and link the physical layer components enforcing higher–level processes (linking the physical, logical and conceptual layers altogether)?*

  To address this research question, the physical layer of the proposed framework was constructed as a result of formalizing and implementing the components of the conceptual and logical layers in real application scenarios. A novel adaptive learning approach, the personalized prescriptive quiz (PPQ), was also proposed (Chapter 6) and later was enhanced by incorporating descriptive, predictive and prescriptive analytics components. The PPQ accounted for identifying students' misunderstood concepts, projecting their likely performance at the end of the semester, and providing certain intervention mechanisms to help them rectify their misconceptions. By linking the physical layer's components to their corresponding logical and conceptual layers, the integrated analytics framework proposed in Chapter 5 was finalized.

As a final remark regarding each layer of the integrated analytics framework, the *conceptual layer* is designed to address the high–level analytical requirements and therefore is kind of a fixed layer which is specialized for different application scenarios (in our case, the learning analytics and the educational institutions). This means that the *conceptual layer* will not change and serves as an abstract domain–agnostic analytics component which may be applied in a wide range of analytics scenarios. The *logical layer*, on the other hand, is domain–specific and is focused on the learning analytics processes and aims to cover as many requirements as possible. This layer might be adapted to a particular analytics context (in our case, the learning analytics). Finally, the *physical layer*, which is concerned with the implementation and formalization of multiple learning analytics processes in certain application scenarios, may frequently change depending on the emerging requirements of the application.

Given the implementation of the physical layer, Section 7.1.1 brings together the various ingredients of the proposed analytics framework of Chapter 5 to construct one integrated learning analytics solution for use by educational institutions.

7.1.1 Finalizing The Framework Layers

By putting together the formalized implementation of the *physical layer*, we have all the required ingredients to finalize the proposed analytics framework, which helps us to build our learning analytics–based solution. Sections 7.1.2, and 7.1.3) are dedicated to constructing the entire framework, incorporating the conceptual, logical, and physical layers.

7.1.2 Linking The Physical Layer To The Logical Layer

The last step to finalize the framework is to connect the implemented *physical layer* to the remaining layers of the framework. This task is accomplished as depicted in Figure 7.1 for the *Intervention Process* described in Section 5.4.4. For the sake of simplicity, only the major components of the intervention process were illustrated.

As shown in Figure 7.1, the intervention process was implemented within the *physical layer* along with its major components – the *simulation*, the *evaluation*, the *optimization*, and the *PPQ generation* elements. Each one of these components is related to their corresponding components in the *logical layer* using the "IS–A" relationship. The interrelationships among several *logical* to *conceptual* layer components were further elaborated upon in Chapter 5. Altogether, all learning analytics processes were implemented within the *physical layer* and connected to their corresponding *logical layer* components. Figure 7.1 also represents the "IS–A" relationship that occurs between the *physical* to *logical*, and the *logical* to *conceptual* layers.

7.1.3 The Overall Analytics–Driven Framework

The whole thesis and its contributions are summarized in Figure 7.2.

Figure 7.2 shows how the proposed framework is capable of covering the major requirements of learning analytics, by marrying the two worlds of analytics and education. The *conceptual layer* of the framework is responsible for modelling the key analytical approaches of the analytics world, that is, the *descriptive*, *predictive*, and *prescriptive* analytics. In addition, the *logical* and *physical* layers are focused on addressing the education world requirements, especially the learning analytics, by mapping the proposed composite analytics architecture of the *conceptual layer* into the education world scenario. The 10 learning analytics processes were represented in BPMN specification, and their corresponding components were implemented and formalized within the *physical layer*, such as the PPQ.

7.2 Future Research Directions

While the proposed integrated framework for learning analytics is able to support adaptive learning, there remain several opportunities for future improvements. The following list highlights some key promising future research directions in relation to our proposed framework.

- *Addition of learning analytics stakeholders.* LA stakeholders, also referred to as system

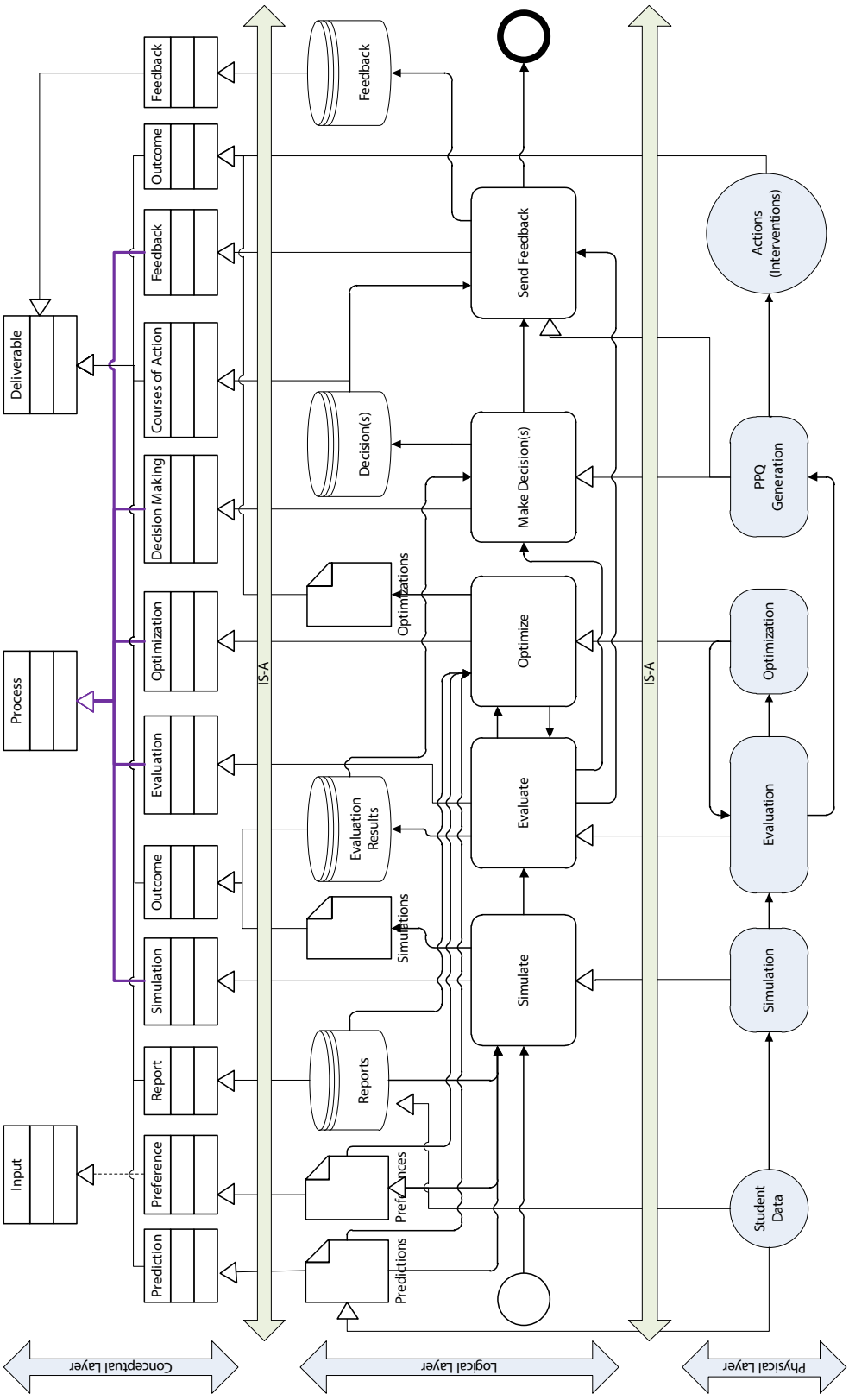


Figure 7.1: Physical To Logical To Conceptual Layers Relationships

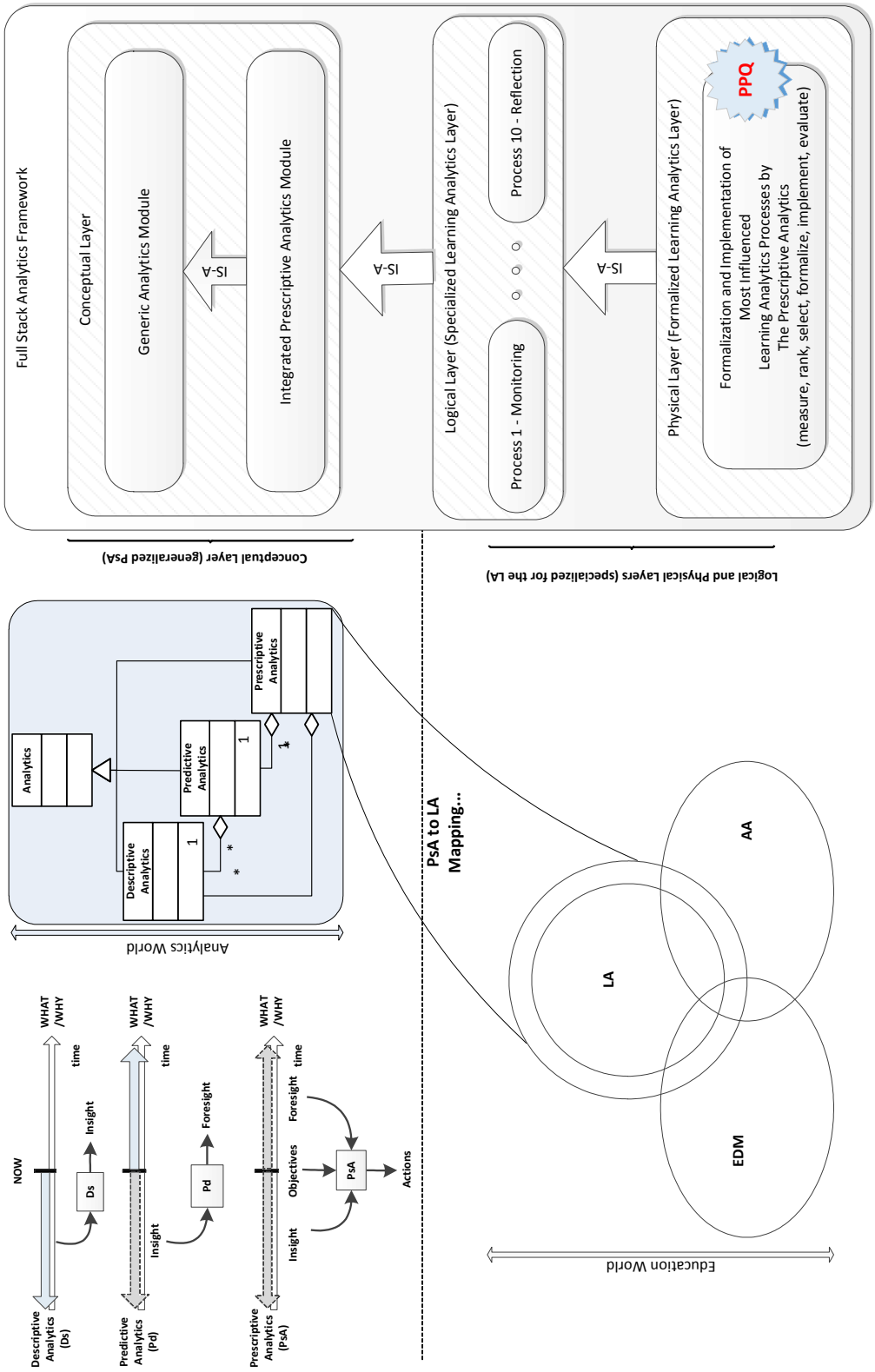


Figure 7.2: The High–Level View of The Overall Analytics–Driven Framework and Its Analytical Techniques Coverage Within The Analytics and Education Worlds

actors (such as instructors, students, educational institutions), were investigated as part of the second dimension of the learning analytics reference model, introduced in Section 3.3. Their representation within each analytics layer, as well as their contributions towards the entire framework, may be elaborated upon.

- *Covering more learning analytics requirements within the logical layer.* As an interesting extension to this work, further learning analytics processes and techniques could be covered within the proposed framework. Some of these extensions could ostensibly address ethics and privacy concerns in the area of learning analytics, within the *logical layer*. Although there is no one–size–fits–all solution for all learning analytics requirements, taking the ethics and privacy issues into account could make the framework more adaptable to the ongoing needs of educational institutions. Depending on any emerging learning analytics requirements, the corresponding modifications are expected to be applied within the *logical* and the *physical* layers, accordingly.

- *Further physical layer modification/expansion.* Given the high diversity of educational application scenarios, several expansion directions to the physical layer can be considered. The following lists some major feasible extensions or modifications with respect to the *physical layer*.

  - The proposed hybrid predictive analytics approach (Section 6.6.3, Chapter 6) can be applied on more courses and additional student data. By collecting more data, the accuracy of the predictive model can be improved in terms of projecting students' performance at the end of each semester.
  - We applied the PPQ approach in weeks 9-11. It can be offered in the initial, middle and final weeks of the course offering. Given that the PPQ approach is enriched with analytical power (incorporating descriptive, predictive, and prescriptive analytics), it is more likely to promote adaptive learning by helping both students and instructors to achieve their objectives. Furthermore, applying the PPQ approach in the earlier weeks of the semester will allow students to manage and rectify their misconceptions in a timely and stress–free manner. It also enables instructors to identify at–risk students earlier, to perform relevant interventions.
  - One issue with the PPQ in improving self–esteem gradually was that varying questions from familiar to unfamiliar concepts did not appeal to some students. To better cater to our student diversity, we intended to give students greater

control in customizing their own learning pathways. One step done within our research was to specify ranges for discrimination and difficulty indexes. The other step, the future work, is to give the students the option to choose from multiple PPQ approaches (such as standard and adaptive PPQs) instead of just enforcing them to participate in the standard or the adaptive PPQ experiments.

  - In future, other aspects of the PPQ, such as the time spent per attempt, between attempts, and on the number of attempts, as well as the impact of post–test outcomes, should be incorporated and analyzed. This way, the instructors can get access to each student's history of attempts, along with the number of times they participated with the corresponding score per quiz. This provides a more enriched data set, to effectively design the next sets of personalized quizzes and perform more focused and relevant interventions pertaining to the students' needs.
  - Currently, the quiz questions are tagged with their associated concept(s). One useful extension is to annotate and tag the questions with their corresponding learning resources as well. Having the question meta–data would allow the PPQ to instantly provide formative feedback, linking relevant study materials to each student, in the cases where incorrect answers are presented to the quiz questions.

- *Personalized course management system for future recommendations.* By taking into consideration each student's performance within a given semester or throughout their educational journey, the framework could recommend them which courses they might select in future. This way, the system acts as an intelligent course management system (CMS). With more accurate predictions and the wealth of students' academic records and aptitudes, the system could provide individual students with a personalized curriculum to foster and promote adaptive learning. The personalized CMS can also contribute to better student retention rates and improved student experience.

Bibliography

S. Adams Becker, M. Cummins, A. Davis, A. Freeman, C. Hall Giesinger, and V. Ananthanarayanan. Nmc horizon report: 2017 higher education edition. *Austin, Texas: The New Media Consortium*, 2017.

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

S. Akoojee and M. Nkomo. Access and quality in south african higher education: The twin challenges of transformation. *South African journal of higher education*, 21(3):385–399, 2007.

B. J. Alters and C. E. Nelson. Perspective: Teaching evolution in higher education. *Evolution*, 56(10):1891–1901, 2002.

L. W. Anderson, D. R. Krathwohl, P. Airasian, K. Cruikshank, R. Mayer, P. Pintrich, J. Raths, and M. Wittrock. A taxonomy for learning, teaching and assessing: A revision of bloom's taxonomy. *New York. Longman Publishing. Artz, AF, & Armour-Thomas, E.(1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. Cognition and Instruction*, 9(2):137–175, 2001.

A. Andrade, G. Delandshere, and J. A. Danish. Using multimodal learning analytics to model student behavior: A systematic analysis of epistemological framing. *Journal of Learning Analytics*, 3(2):282–306, 2016. doi: http://dx.doi.org/10.18608/jla.2016.32.14.

C. Apte. The role of machine learning in business optimization. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1–2, 2010.

K. E. Arnold and M. D. Pistilli. Course signals at purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 267–270. ACM, 2012.

Y. Attali and T. Fraenkel. The point-biserial as a discrimination index for distractors in multiple-choice items: Deficiencies in usage and an alternative. *Journal of Educational Measurement*, 37(1):77–86, 2000.

D. o. E. Australian Government and Training. Release of the Higher Education Standards - Panel's Discussion Paper on Improving Completion, Retention and Success in Higher Education. https://www.education.gov.au/news/release-higher-education-standards-panel-s-discussion-paper-improving-completion-retention-and, 2017. [Online; accessed 31-August-2017].

P. Baepler and C. J. Murdoch. Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning*, 4(2):17, 2010.

L. L. Baer and D. M. Norris. What every leader needs to know about student success analytics. *White paper developed for Civitas Learning. Retrieved June*, 17:2016, 2015.

P. Baker and B. Gourley. *Data Divination: Big Data Strategies*. Delmar Learning, 2014.

R. S. Baker and P. S. Inventado. Educational data mining and learning analytics. In *Learning analytics*, pages 61–75. Springer, 2014.

R. S. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. *JEDM— Journal of Educational Data Mining*, 1(1):3–17, 2009.

A. Bakharia, L. Corrin, P. de Barba, G. Kennedy, D. Gašević, R. Mulder, D. Williams, S. Dawson, and L. Lockyer. A conceptual framework linking learning design with learning analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 329–338. ACM, 2016.

A. Banerjee, T. Bandyopadhyay, and P. Acharya. Data analytics: Hyped up aspirations or true potential? *Vikalpa*, 38(4):1–12, 2013.

R. Barber and M. Sharkey. Course correction: using analytics to predict course success. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 259–262. ACM, 2012.

R. Barga, V. Fontama, and W. H. Tok. *Predictive Analytics with Microsoft Azure Machine Learning: Build and Deploy Actionable Solutions in Minutes*. Apress, 2014.

A. Basu. Five pillars of prescriptive analytics success. *Analytics Magazine*, pages 8–12, 2013.

T. Beaubouef and J. Mason. Why the high attrition rate for computer science students: some thoughts and observations. *ACM SIGCSE Bulletin*, 37(2):103–106, 2005. doi: 10.1145/1083431.1083474.

M. Ben-Ari. Constructivism in computer science education. *Journal of Computers in Mathematics and Science Teaching*, 20(1):45–73, 2001.

M. Berland, R. S. Baker, and P. Blikstein. Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, 19(1-2): 205–220, 2014.

D. Bertsimas and N. Kallus. From predictive to prescriptive analytics. *arXiv preprint arXiv:1402.5481*, 2014.

J. W. Best and J. V. Kahn. *Research in education*. Pearson Education India, 2016.

M. Bienkowski, M. Feng, B. Means, et al. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *US Department of Education, Office of Educational Technology*, 1:1–57, 2012.

M. Biggers, A. Brauer, and T. Yilmaz. Student perceptions of computer science: a retention study comparing graduating seniors with cs leavers. In *ACM SIGCSE Bulletin*, volume 40, pages 402–406. ACM, 2008. doi: 10.1145/1352322.1352274.

M. Bilal, L. O. Oyedele, O. O. Akinade, S. O. Ajayi, H. A. Alaka, H. A. Owolabi, J. Qadir, M. Pasha, and S. A. Bello. Big data architecture for construction waste analytics (cwa): A conceptual framework. *Journal of Building Engineering*, 6:144–156, 2016.

P. Blikstein and M. Worsley. Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, 3(2):220–238, 2016.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

M. Brown. Learning analytics: The coming third wave. *EDUCAUSE Learning Initiative Brief*, 1(4), 2011.

P. Brusilovsky and E. Millán. User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web*, pages 3–53. Springer, 2007. doi: https://doi.org/10.1007/978-3-540-72079-9_1.

S. Bull and J. Kay. Smili: a framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education*, 26(1):293–331, 2016.

R. Caceffo, S. Wolfman, K. S. Booth, and R. Azevedo. Developing a computer science concept inventory for introductory programming. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pages 364–369. ACM, 2016.

R. Caceffo, G. Gama, and R. Azevedo. Exploring active learning approaches to computer science classes. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pages 922–927. ACM, 2018.

J. P. Campbell, P. B. DeBlois, and D. G. Oblinger. Academic analytics: A new tool for a new era. *EDUCAUSE review*, 42(4):40, 2007.

I. Cetin and M. Y. Ozden. Development of computer programming attitude scale for university students. *Computer Applications in Engineering Education*, 23(5):667–672, 2015.

P. Charlton, M. Mavrikis, and D. Katsifli. The potential of learning analytics and big data. *Ariadne*, (71), 2013.

M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs. A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5-6):318–331, 2012.

M. A. Chatti, V. Lukarov, H. Thüs, A. Muslim, A. M. F. Yousef, U. Wahid, C. Greven, A. Chakrabarti, and U. Schroeder. Learning analytics: Challenges and future research directions. *E-learn Educ (Eleed) J*, 10:1–16, 2014.

H. Chen, R. H. Chiang, and V. C. Storey. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36(4), 2012.

J. A. Clarke, K. J. Nelson, and I. D. Stoodley. The place of higher education institutions in assessing student engagement, success and retention: A maturity model to guide practice. 2013.

L. M. Cleveland, T. M. McCabe, and J. T. Olimpo. A call for programmatic assessment of undergraduate students' conceptual understanding and higher-order cognitive skills. *Journal of microbiology & biology education*, 19(1), 2018.

H. Coates, R. James, and G. Baldwin. A critical examination of the effects of learning management systems on university teaching and learning. *Tertiary Education & Management*, 11(1):19–36, 2005.

C. Colvin, T. Rogers, A. Wade, S. Dawson, D. Gašević, S. Buckingham Shum, and J. Fisher. Student retention and learning analytics: A snapshot of australian practices and a framework for advancement. *Sydney, NSW: Australian Office for Learning and Teaching*, 2015.

J. Considine, M. Botti, and S. Thomas. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, 12(1):19–24, 2005.

A. T. Corbett, K. R. Koedinger, and J. R. Anderson. Intelligent tutoring systems. In *Handbook of Human-Computer Interaction (Second Edition)*, pages 849–874. Elsevier, 1997.

L. Corrin and P. de Barba. Exploring students' interpretation of feedback delivered through learning analytics dashboards. In *Proceedings of the ascilite 2014 conference*, pages 629–633, 2014.

P. A. Daempfle. An analysis of the high attrition rates among first year college science, math, and engineering majors. *Journal of College Student Retention: Research, Theory & Practice*, 5(1):37–52, 2003. doi: 10.2190/DWQT-TYA4-T20W-RCWH. URL http://dx.doi.org/10.2190/DWQT-TYA4-T20W-RCWH.

B. Daniel. Big data and analytics in higher education: Opportunities and challenges. *British journal of educational technology*, 46(5):904–920, 2015.

C. J. Date. *An introduction to database systems*. Pearson Education India, 2006.

T. H. Davenport and J. Dyché. Big data in big companies. *May 2013*, 2013.

S. Dawson and G. Siemens. Analytics to literacies: The development of a learning analytics framework for multiliteracies assessment. *The International Review of Research in Open and Distributed Learning*, 15(4), 2014.

S. Dawson, A. Bakharia, E. Heathcote, et al. Snapp: Realising the affordances of real-time sna within networked learning environments. Networked Learning, 2010.

S. Dawson, D. Gašević, G. Siemens, and S. Joksimovic. Current state and future trends: A citation network analysis of the learning analytics field. In *Proceedings of the fourth international conference on learning analytics and knowledge*, pages 231–240. ACM, 2014.

S. de Freitas. *Education in computer generated environments*. Routledge, 2013.

G. C. Deka. Big data predictive and prescriptive analytics. In *Big Data: Concepts, Methodologies, Tools, and Applications*, pages 30–55. IGI Global, 2016.

D. Delen. *Real-World Data Mining: Applied Business Analytics and Decision Making*. FT Press, 2014.

D. Delen and H. Demirkan. Data, information and analytics as services. *Decision Support Systems*, 55(1):359–363, 2013.

P. J. Denning and A. McGettrick. Recentering computer science. *Communications of the ACM*, 48(11):15–19, 2005. doi: 10.1145/1096000.1096018.

B. Dietz-Uhler and J. E. Hurn. Using learning analytics to predict (and improve) student success: A faculty perspective. *Journal of Interactive Online Learning*, 12(1):17–26, 2013.

S. D'Mello, R. W. Picard, and A. Graesser. Toward an affect-sensitive autotutor. *IEEE Intelligent Systems*, 22(4), 2007. doi: 10.1109/MIS.2007.79.

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

W. W. Eckerson. Predictive analytics. *Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report*, 1:1–36, 2007.

T. Elias. Learning analytics: The definitions, the processes, and the potential, 2011.

R. K. Ellis. Field guide to learning management systems. *ASTD Learning Circuits*, page 2009, 2009.

A. Essa and H. Ayad. Student success system: risk analytics and data visualization using ensembles of predictive models. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 158–161. ACM, 2012.

J. R. Evans and C. H. Lindner. Business analytics: the next frontier for decision sciences. *Decision Line*, 43(2):4–6, 2012.

A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth. Unsupervised modeling for understanding mooc discussion forums: a learning analytics approach. In *Proceedings of the fifth international conference on learning analytics and knowledge*, pages 146–150. ACM, 2015.

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

R. Ferguson. The state of learning analytics in 2012: A review and future challenges. *Knowledge Media Institute, Technical Report KMI-2012-01*, 2012b.

R. Ferguson, D. Clow, L. Macfadyen, A. Essa, S. Dawson, and S. Alexander. Setting learning analytics in context: Overcoming the barriers to large-scale adoption. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, pages 251–253. ACM, 2014.

R. Ferguson, A. Brasher, D. Clow, D. Griffiths, and H. Drachsler. Learning analytics: visions of the future. 2016. doi: https://doi.org/10.1145/2883851.2883905.

G. M. Foody, A. Mathur, C. Sanchez-Hernandez, and D. S. Boyd. Training set size requirements for the classification of a specific class. *Remote Sensing of Environment*, 104(1): 1–14, 2006.

S. Freitas, D. Gibson, C. Du Plessis, P. Halloran, E. Williams, M. Ambrose, I. Dunwell, and S. Arnab. Foundations of dynamic learning analytics: Using university student data to increase retention. *British Journal of Educational Technology*, 46(6):1175–1188, 2015.

T. Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21(2):137–146, 2011.

S. Gajjar, R. Sharma, P. Kumar, and M. Rana. Item and test analysis to identify quality multiple choice questions (mcqs) from an assessment of medical students of ahmedabad, gujarat. *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine*, 39(1):17, 2014.

D. García-Saiz, C. Palazuelos, and M. Zorrilla. Data mining and social network analysis in the educational field: An application for non-expert users. In *Educational Data Mining*, pages 411–439. Springer, 2014.

D. R. Garrison and H. Kanuka. Blended learning: Uncovering its transformative potential in higher education. *The internet and higher education*, 7(2):95–105, 2004.

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

M. N. Giannakos, D. G. Sampson, and Ł. Kidziński. Introduction to smart learning analytics: foundations and developments in video-based learning. *Smart Learning Environments*, 3 (1):12, 2016. doi: https://doi.org/10.1186/s40561-016-0034-2.

S. Gibbons. Course-management systems. *Library Technology Reports*, 41(3):7, 2005.

D. Gibson and S. de Freitas. Exploratory analysis in learning analytics. *Technology, Knowledge and Learning*, 21(1):5–19, 2016.

J. Gidman, A. Humphreys, and M. Andrews. The role of the personal tutor in the academic context. *Nurse Education Today*, 20(5):401–407, 2000.

S. P. Goggins, W. Xing, X. Chen, B. Chen, and B. Wadholm. Learning analytics at" small" scale: Exploring a complexity-grounded model for assessment automation. *J. UCS*, 21(1): 66–92, 2015.

P. J. Goldstein and R. N. Katz. *Academic analytics: The uses of management information and technology in higher education*, volume 8. Educause, 2005.

S. Govaerts, K. Verbert, J. Klerkx, and E. Duval. Visualizing activities for self-reflection and awareness. In *International Conference on Web-Based Learning*, pages 91–100. Springer, 2010.

W. Greller and H. Drachsler. Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society*, 15(3):42, 2012.

P. J. Haas, P. P. Maglio, P. G. Selinger, and W. C. Tan. Data is dead... without what-if models. *PVLDB*, 4(12):1486–1489, 2011.

W. Harlen and M. James. Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, 4(3):365–379, 1997. doi: http://dx.doi.org/10.1080/0969594970040304.

J. Hattie and H. Timperley. The power of feedback. *Review of educational research*, 77(1): 81–112, 2007.

B. T. Hazen, C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, 154:72–80, 2014.

J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 32–39. IEEE, 2005. doi: 10.1109/INFVIS.2005.1532126.

N. T. Heffernan and C. L. Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014. doi: https://doi.org/10.1007/s40593-014-0024-x.

G. Hillaire, G. Rappolt-Schlichtmann, and K. Ducharme. Prototyping visual learning analytics guided by an educational theory informed goal. *Journal of Learning Analytics*, 3(3): 115–142, 2016. doi: https://doi.org/10.18608/jla.2016.33.7.

M. R. Hingorjo and F. Jaleel. Analysis of one-best mcqs: the difficulty index, discrimination index and distractor efficiency. *JPMA-Journal of the Pakistan Medical Association*, 62(2): 142, 2012.

C. E. Hmelo-Silver. Problem-based learning: What and how do students learn? *Educational psychology review*, 16(3):235–266, 2004.

D. Hooshyar, R. B. Ahmad, M. Yousefi, M. Fathi, A. Abdollahi, S.-J. Horng, and H. Lim. A solution-based intelligent tutoring system integrated with an online game-based formative assessment: development and evaluation. *Educational Technology Research and Development*, 64(4):787–808, 2016a.

D. Hooshyar, R. B. Ahmad, M. Yousefi, M. Fathi, S.-J. Horng, and H. Lim. Applying an online game-based formative assessment in a flowchart-based intelligent tutoring system for improving problem-solving skills. *Computers & Education*, 94:18–36, 2016b.

D. Ifenthaler and C. Widanapathirana. Development and validation of a learning analytics framework: Two case studies using support vector machines. *Technology, Knowledge and Learning*, 19(1-2):221–240, 2014.

P. Ihantola, A. Vihavainen, A. Ahadi, M. Butler, J. Börstler, S. H. Edwards, E. Isohanni, A. Korhonen, A. Petersen, K. Rivers, et al. Educational data mining and learning analytics in programming: Literature review and case studies. In *Proceedings of the 2015 ITiCSE on Working Group Reports*, pages 41–63. ACM, 2015.

S. G. Jaramillo. Horizon report-2017 higher education edition. *CUADERNO ACTIVA*, 9(9): 171, 2017.

J. Jaworska, N. Nikolova-Jeliazkova, and T. Aldenberg. Qsar applicability domain estimation by projection of the training set descriptor space: a review. *ATLA-NOTTINGHAM-*, 33 (5):445, 2005.

S. M. Jayaprakash, E. W. Moody, E. J. Lauría, J. R. Regan, and J. D. Baron. Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1):6–47, 2014.

L. Johnson, S. Adams Becker, V. Estrada, and A. Freeman. *The NMC Horizon Report: 2015 Higher Education Edition.* ERIC, 2015.

J. Jovanovic, D. Gasevic, C. Brooks, V. Devedzic, M. Hatala, T. Eap, and G. Richards. Loco-analyst: semantic web technologies in learning content usage analysis. *International journal of continuing engineering education and life long learning*, 18(1):54–76, 2008.

T. Jülicher. Education 2.0: Learning analytics, educational data mining and co. In *Big Data in Context*, pages 47–53. Springer, 2018.

L. C. Kaczmarczyk, E. R. Petrick, J. P. East, and G. L. Herman. Identifying student misconceptions of programming. In *Proceedings of the 41st ACM technical symposium on Computer science education*, pages 107–111. ACM, 2010. doi: 10.1145/1734263.1734299.

S. H. Kaisler, J. A. Espinosa, F. Armour, and W. H. Money. Advanced analytics–issues and challenges in a global environment. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 729–738. IEEE, 2014.

R. Karim, J. Westerberg, D. Galar, and U. Kumar. Maintenance analytics–the new know in maintenance. *IFAC-PapersOnLine*, 49(28):214–219, 2016.

J. Kay. Lifelong learner modeling for lifelong personalized pervasive learning. *IEEE Transactions on Learning Technologies*, 1(4):215–228, 2008.

J. Kehoe. Basic item analysis for multiple-choice tests. eric/ae digest. 1995.

R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

J. A. Kulik and J. Fletcher. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of Educational Research*, 86(1):42–78, 2016. doi: https://doi.org/10.3102/0034654315581420.

C. Lang, J. McKay, and S. Lewis. Seven factors that influence ict student achievement. In *ACM SIGCSE Bulletin*, volume 39, pages 221–225. ACM, 2007. doi: 10.1145/1269900.1268849.

D. Larson and V. Chang. A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, 36(5):700–710, 2016.

J. A. Larusson and B. White. *Learning analytics: From research to practice*, volume 13. Springer, 2014.

D. Leony, A. Pardo, L. de la Fuente Valentín, D. S. de Castro, and C. D. Kloos. Glass: a learning analytics visualization tool. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 162–163. ACM, 2012.

M. Liberatore and W. Luo. Informs and the analytics movement: The view of the membership. *Interfaces*, 41(6):578–589, 2011.

L. C. Liñán and Á. A. J. Pérez. Educational data mining and learning analytics: differences, similarities, and time evolution. *International Journal of Educational Technology in Higher Education*, 12(3):98–112, 2015.

R. L. Linn. *Measurement and assessment in teaching*. Pearson Education India, 2008.

R. Liu, R. Patel, and K. R. Koedinger. Modeling common misconceptions in learning process data. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 369–377. ACM, 2016.

J. Loizzo and P. A. Ertmer. Moococracy: the learning culture of massive open online courses. *Educational Technology Research and Development*, 64(6):1013–1032, 2016. doi: https://doi.org/10.1007/s11423-016-9444-7.

S. Lonn, S. Aguilar, and S. D. Teasley. Issues, challenges, and lessons learned when scaling up a learning analytics intervention. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 235–239. ACM, 2013.

W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu. Intelligent tutoring systems and learning outcomes: A meta-analysis., 2014.

L. P. Macfadyen and S. Dawson. Numbers are not enough. why e-learning analytics failed to inform an institutional strategic plan. *Journal of Educational Technology & Society*, 15 (3):149, 2012.

S. MacNeill, L. M. Campbell, and M. Hawksey. Analytics for education. *Reusing Open Resources: Learning in Open Networks for Work, Life and Education*, page 154, 2014.

M. W. Maier, D. Emery, and R. Hilliard. Software architecture: Introducing ieee standard 1471. *Computer*, (4):107–109, 2001.

N. Manouselis, H. Drachsler, R. Vuorikari, H. Hummel, and R. Koper. Recommender systems in technology enhanced learning. In *Recommender systems handbook*, pages 387–415. Springer, 2011. doi: https://doi.org/10.1007/978-0-387-85820-3_12.

M. V. Marathe, H. S. Mortveit, N. Parikh, and S. Swarup. Prescriptive analytics using synthetic information. *Emerging Methods in Predictive Analytics: Risk Management and Decision-Making: Risk Management and Decision-Making*, page 1, 2014.

C. Marquez-Vera, C. Romero, and S. Ventura. Predicting school failure using data mining. In *Educational Data Mining 2011*, 2010.

C. K. Martin, D. Nacu, and N. Pinkard. Revealing opportunities for 21st century learning: An approach to interpreting user trace log data. *Journal of Learning Analytics*, 3(2):37–87, 2016.

J. Martin and K. VanLehn. Student assessment using bayesian nets. *International Journal of Human-Computer Studies*, 42(6):575–591, 1995.

T. Martin and B. Sherin. Learning analytics and computational techniques for detecting and evaluating patterns in learning: An introduction to the special issue. *Journal of the Learning Sciences*, 22(4):511–520, 2013.

R. Mazza and V. Dimitrova. Visualising student tracking data to support instructors in web-based distance education. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 154–161. ACM, 2004.

R. Mazza and C. Milani. Exploring usage analysis in learning systems: Gaining insights from visualisations. In *Workshop on usage analysis in learning systems at 12th international conference on artificial intelligence in education*, pages 65–72, 2005.

N. Medvidovic and R. N. Taylor. Software architecture: foundations, theory, and practice. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 2*, pages 471–472. ACM, 2010.

J. Michael. Where's the evidence that active learning works? *Advances in physiology education*, 30(4):159–167, 2006.

K. Mouri and H. Ogata. Ubiquitous learning analytics in the real-world language learning. *Smart Learning Environments*, 2(1):15, 2015. doi: https://doi.org/10.1186/s40561-015-0023-x.

A. Muslim, M. A. Chatti, T. Mahapatra, and U. Schroeder. A rule-based indicator definition tool for personalized learning analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 264–273. ACM, 2016. doi: https://doi.org/10.1145/2883851.2883921.

S. Nunn, J. T. Avella, T. Kanai, and M. Kebritchi. Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. *Online Learning*, 20 (2), 2016.

M. Nyroos, I. Schéle, and C. Wiklund-Hörnqvist. Implementing test enhanced learning: Swedish teacher students' perception of quizzing. *International Journal of Higher Education*, 5(4):1–12, 2016.

X. Ochoa and M. Worsley. Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics*, 3(2):213–219, 2016. doi: http://dx.doi.org/10.18608/jla.2016.32.10.

P. C. d. Oliveira, C. J. C. d. A. Cunha, and M. K. Nakayama. Learning management systems (lms) and e-learning management: an integrative review and research agenda. *JISTEM-Journal of Information Systems and Technology Management*, 13(2):157–180, 2016. doi: http://dx.doi.org/10.4301/S1807-17752016000200001.

C. O'Malley. *Computer supported collaborative learning*, volume 128. Springer Science & Business Media, 2012.

F. Pajares and M. D. Miller. Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of educational psychology*, 86(2):193, 1994.

M. Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005.

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

A. Pardo. A feedback model for data-rich learning experiences. *Assessment & Evaluation in Higher Education*, 43(3):428–438, 2018. doi: https://doi.org/10.1080/02602938.2017.1356905.

Y. Park and I.-H. Jo. Development of the learning analytics dashboard to support students' learning performance. *J. UCS*, 21(1):110–133, 2015.

R. Pea and D. Jacks. The learning analytics workgroup: A report on building the field of learning analytics for personalized learning at scale, 2014.

A. Pears, S. Seidman, L. Malmi, L. Mannila, E. Adams, J. Bennedsen, M. Devlin, and J. Paterson. A survey of literature on the teaching of introductory programming. *ACM SIGCSE Bulletin*, 39(4):204–223, 2007. doi: 10.1145/1345375.1345441.

A. Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4):1432–1462, 2014.

A. Peña-Ayala. *Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning*, volume 94. Springer, 2017. doi: https://doi.org/10.1007/978-3-319-52977-6.

A. Peña-Ayala. Learning analytics: A glance of evolution, status, and trends according to a proposed taxonomy. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018. doi: 10.1002/widm.1243.

A. Peña-Ayala, L. A. Cárdenas-Robledo, and H. Sossa. A landscape of learning analytics: An exercise to highlight the nature of an emergent field. In *Learning Analytics: Fundaments, Applications, and Trends*, pages 65–112. Springer, 2017. doi: https://doi.org/10.1007/978-3-319-52977-6_3.

D. J. Power. Using 'big data' for analytics and decision support. *Journal of Decision Systems*, 23(2):222–228, 2014.

W. Pree. Meta patterns—a means for capturing the essentials of reusable object-oriented design. In *European Conference on Object-Oriented Programming*, pages 150–162. Springer, 1994.

F. Pyrczak. Validity of the discrimination index as a measure of item quality. *Journal of Educational Measurement*, 10(3):227–231, 1973.

P. Refaeilzadeh, L. Tang, and H. Liu. Cross-validation. *Encyclopedia of database systems*, pages 1–7, 2016.

C. Reffay and T. Chanier. How social network analysis can help to measure cohesion in collaborative distance-learning. In *Designing for change in networked learning environments*, pages 343–352. Springer, 2003. doi: https://doi.org/10.1007/978-94-017-0195-2_42.

B. Rienties, A. Boroowa, S. Cross, C. Kubiak, K. Mayles, and S. Murphy. Analytics4action evaluation framework: A review of evidence-based learning analytics interventions at the open university uk. *Journal of Interactive Media in Education*, 2016(1), 2016.

B. Rienties, S. Cross, and Z. Zdrahal. Implementing a learning analytics intervention and evaluation framework: what works? In *Big Data and Learning Analytics in Higher Education*, pages 147–166. Springer, 2017.

A. Robins. Learning edge momentum: A new account of outcomes in cs1. *Computer Science Education*, 20(1):37–71, 2010.

I. Roll, V. Aleven, B. M. McLaren, and K. R. Koedinger. Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21(2):267–280, 2011.

C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146, 2007. doi: https://doi.org/10.1016/j.eswa.2006.04.005.

C. Romero and S. Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40 (6):601–618, 2010.

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

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

C. Romero, S. Ventura, M. Pechenizkiy, and R. S. Baker. *Handbook of educational data mining*. CRC press, 2010.

A. Rubel and K. M. Jones. Student privacy in learning analytics: An information ethics perspective. *The Information Society*, 32(2):143–159, 2016.

J. A. Ruipérez-Valiente, P. J. Muñoz-Merino, D. Leony, and C. D. Kloos. Alas-ka: A learning analytics extension for better understanding the learning process in the khan academy platform. *Computers in Human Behavior*, 47:139–148, 2015.

R. W. Rumberger. High school dropouts: A review of issues and evidence. *Review of educational research*, 57(2):101–121, 1987.

Y. Sarin, M. Khurana, M. Natu, A. G. Thomas, and T. Singh. Item analysis of published mcqs. *Indian pediatrics*, 35:1103–1104, 1998.

H.-C. Schmitz, M. Scheffel, M. Friedrich, M. Jahn, K. Niemann, and M. Wolpers. Camera for ple. In *European Conference on Technology Enhanced Learning*, pages 507–520. Springer, 2009.

B. Schneider and P. Blikstein. Unraveling students' interaction around a tangible interface using multimodal learning analytics. *Journal of Educational Data Mining*, 7(3):89–116, 2015.

M. J. Schniederjans, D. G. Schniederjans, and C. M. Starkey. *Business Analytics Principles, Concepts, and Applications: What, Why, and How*. Pearson Education, 2014.

A. Schoenfeld. Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. *Coleccion Digital Eudoxus*, (7), 2009.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

N. Sclater. Code of practice for learning analytics: A literature review of the ethical and legal issues. *Jisc, November*, 5, 2014.

N. Sclater, A. Peasgood, and J. Mullan. Learning analytics in higher education. *London: Jisc. Accessed February*, 8:2017, 2016.

X. Shacklock. *From bricks to clicks: the potential of data and analytics in higher education*. Higher Education Commission, 2016.

R. Sharda, D. A. Asamoah, and N. Ponna. Business analytics: Research and teaching perspectives. In *Information Technology Interfaces (ITI), Proceedings of the ITI 2013 35th International Conference on*, pages 19–27. IEEE, 2013.

G. Shmueli and O. R. Koppius. Predictive analytics in information systems research. *Mis Quarterly*, pages 553–572, 2011.

S. B. Shum and R. D. Crick. Learning dispositions and transferable competencies: pedagogy, modelling and learning analytics. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 92–101. ACM, 2012.

S. B. Shum and R. Ferguson. Social learning analytics. *Journal of educational technology & society*, 15(3):3, 2012.

S. B. Shum, S. Knight, and K. Littleton. Learning analytics. In *UNESCO Institute for Information Technologies in Education. Policy Brief*. Citeseer, 2012.

V. J. Shute. Focus on formative feedback. *Review of educational research*, 78(1):153–189, 2008.

E. Siegel. *Predictive analytics: The power to predict who will click, buy, lie, or die*. John Wiley & Sons Incorporated, 2016.

G. Siemens. Connectivism: A learning theory for the digital age. *International journal of instructional technology and distance learning*, 2(1):3–10, 2005.

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

G. Siemens. Connectivism: A learning theory for the digital age. 2014.

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

G. Siemens and P. Long. Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, 46(5):30, 2011.

G. Siemens, D. Gasevic, C. Haythornthwaite, S. P. Dawson, S. Shum, R. Ferguson, E. Duval, K. Verbert, R. Baker, et al. Open learning analytics: an integrated & modularized platform. 2011.

K. Sin and L. Muthu. Application of big data in education data mining and learning analytics–a literature review. *ICTACT journal on soft computing*, 5(4), 2015.

M. Sionti, H. Ai, C. P. Rosé, and L. Resnick. A framework for analyzing development of argumentation through classroom discussions. *Educational technologies for teaching argumentation skills*, pages 28–55, 2011.

L.-K. Soh, A. Samal, and G. Nugent. An integrated framework for improved computer science education: Strategies, implementations, and results. *Computer Science Education*, 17(1): 59–83, 2007.

R. Soltanpoor and T. Sellis. Prescriptive analytics for big data. In *Australasian Database Conference*, pages 245–256. Springer, 2016. ISBN 978-3-319-46922-5. doi: https://doi.org/10.1007/978-3-319-46922-5_19.

R. Soltanpoor and A. Yavari. Coala: Contextualization framework for smart learning analytics. In *Distributed Computing Systems Workshops (ICDCSW), 2017 IEEE 37th International Conference on*, pages 226–231. IEEE, 2017. ISBN 978-1-5386-3292-5. doi: 10.1109/ICDCSW.2017.58.

G. Stahl and F. Hesse. Practice perspectives in cscl. *International Journal of Computer-Supported Collaborative Learning*, 4(2):109–114, 2009. doi: https://doi.org/10.1007/s11412-009-9065-9.

G. Stahl, T. Koschmann, and D. Suthers. Computer-supported collaborative learning: An historical perspective. *Cambridge handbook of the learning sciences*, 2006:409–426, 2006.

J. D. Storey et al. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

B. Šumak, G. Polancic, and M. Hericko. An empirical study of virtual learning environment adoption using utaut. In *Mobile, Hybrid, and On-Line Learning, 2010. ELML'10. Second International Conference on*, pages 17–22. IEEE, 2010.

V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

M. Taras. Assessment - summative and formative - some theoretical reflections. *British Journal of Educational Studies*, 53(4):466–478, 2005. ISSN 1467-8527. doi: 10.1111/j.1467-8527.2005.00307.x. URL http://dx.doi.org/10.1111/j.1467-8527.2005.00307.x.

D. T. Tempelaar, B. Rienties, and B. Giesbers. In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, 47:157–167, 2015.

D. T. Tempelaar, B. Rienties, and Q. Nguyen. Towards actionable learning analytics using dispositions. *IEEE Transactions on Learning Technologies*, 10(1):6–16, 2017. doi: https://doi.org/10.1109/TLT.2017.2662679.

T. Thonus. Tutor and student assessments of academic writing tutorials: What is "success"? *Assessing Writing*, 8(2):110–134, 2002.

K. Trigwell and M. Prosser. Improving the quality of student learning: the influence of learning context and student approaches to learning on learning outcomes. *Higher education*, 22(3):251–266, 1991. doi: https://doi.org/10.1007/BF00132290.

E. Turban, D. King, R. Sharda, and D. Delen. *Business intelligence: a managerial perspective on analytics*. Prentice Hall, New York, 2013.

M. Vahdat, A. Ghio, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg. Advances in learning analytics and educational data mining. *Proc. of ESANN2015*, pages 297–306, 2015.

A. Van Barneveld, K. E. Arnold, and J. P. Campbell. Analytics in higher education: Establishing a common language. *EDUCAUSE learning initiative*, 1(1):l–ll, 2012.

E. M. Van Raaij and J. J. Schepers. The acceptance and use of a virtual learning environment in china. *Computers & Education*, 50(3):838–852, 2008.

S. Venema and A. Rock. Improving learning outcomes for first year introductory programming students. *FYHE 2014*, 2014.

K. Verbert, N. Manouselis, H. Drachsler, and E. Duval. Dataset-driven research to support learning and knowledge analytics. *Journal of Educational Technology & Society*, 15(3):133, 2012.

K. Verbert, E. Duval, J. Klerkx, S. Govaerts, and J. L. Santos. Learning analytics dashboard applications. *American Behavioral Scientist*, 57(10):1500–1509, 2013. doi: 10.1177/0002764213479363.

K. Verbert, S. Govaerts, E. Duval, J. L. Santos, F. Van Assche, G. Parra, and J. Klerkx. Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing*, 18(6):1499–1514, 2014.

M. A. Waller and S. E. Fawcett. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2):77–84, 2013.

E. Watlington, R. Shockley, P. Guglielmino, and R. Felsher. The high cost of leaving: An analysis of the cost of teacher turnover. *Journal of Education Finance*, pages 22–37, 2010.

C. Watson and F. W. Li. Failure rates in introductory programming revisited. In *Proceedings of the 2014 conference on Innovation & technology in computer science education*, pages 39–44. ACM, 2014.

D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

S. Wiedenbeck, D. Labelle, and V. N. Kain. Factors affecting course outcomes in introductory programming. In *16th Annual Workshop of the Psychology of Programming Interest Group*, pages 97–109, 2004.

D. Wiliam and M. Thompson. Integrating assessment with learning: What will it take to make it work? In *The future of assessment*, pages 53–82. Routledge, 2017.

A. F. Wise, Y. Zhao, and S. N. Hausknecht. Learning analytics for online discussions: a pedagogical model for intervention with embedded and extracted analytics. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 48–56. ACM, 2013.

M. Worsley. Multimodal learning analytics as a tool for bridging learning theory and complex learning behaviors. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 1–4. ACM, 2014.

M. Worsley and P. Blikstein. Leveraging multimodal learning analytics to differentiate student learning strategies. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 360–367. ACM, 2015.

A. Yadav, M. Berges, P. Sands, and J. Good. Measuring computer science pedagogical content knowledge: An exploratory analysis of teaching vignettes to measure teacher knowledge. In *Proceedings of the 11th Workshop in Primary and Secondary Computing Education*, pages 92–95. ACM, 2016.

H. Zhang, K. Almeroth, A. Knight, M. Bulger, and R. Mayer. Moodog: Tracking students' online learning activities. In *EdMedia: World Conference on Educational Media and Technology*, pages 4415–4422. Association for the Advancement of Computing in Education (AACE), 2007.

P. Zhang. Model selection via multifold cross validation. *The Annals of Statistics*, pages 299–313, 1993.