From Analytics to Cognition: Expanding the Reach of Data in Learning

R. Tsoni[1], C. Samaras[1], E. Paxinou[1], C. Panagiotakopoulos[2] and V. S. Verykios[1,a]
[1]*School of Science and Technology, Hellenic Open University, Patra, Greece*
[2]*Department of Primary Education, University of Patras, Patra, Greece*

Keywords: Data Mining, Learning Analytics, Hellenic Open University, Big Data Lab.

Abstract: Education constitutes a rapidly changing and challenging environment, therefore, a shift from reporting to actionable interventions based on data is almost imperative. At the same time, Learning Analytics is incorporating increasingly advanced tools and methods. Artificial Intelligence and cognitive science allow us to study in depth the behavior of students, creating patterns and prediction models. All the above summarize the scope of the newly established Big Data Analytics and Anonymization Laboratory (BAT Lab) in the Hellenic Open University. This paper presents the vision and the work in progress of the BAT Lab in an attempt not only to produce interpretable results but also to organize and present these results in a highly usable way for non-experts. Additionally, PRIME-EDU software, which is the team's latest work, is presented.

1 INTRODUCTION

1.1 The Era of Data Analytics

Higher Education Institutions are facing major difficulties which will, at some point, render them incapable of sustaining their existence. Some of these problems include outdated curriculum, high rates of students' dropouts, and graduates who lack the necessary skills for the modern job market. This leads to unqualified workforce and outdated and irrelevant institutions (DeMillo, and Young, 2015). It is obvious that changes need to be made.

Contemporary methods to data-driven decision making include controlled experiments like A/B testing and the use of questionnaires. However, all these methods present certain problems. Regarding experiments, scalability is a big issue, and as far as questionnaires are concerned, they are not considered reliable as people's answers are not always representative of their true feelings (Lindstrom, 2010). Manual approaches to gathering data incur a very high cost as well.

It appears that there must be a more suitable alternative. Living in the Petabyte age equals rapidly developing technology and more information gathering than ever before. This has led to the creation of Big Data research area which hopefully will provide the answers to some of the problems that society and academia are facing today. In order to do this, data mining and data analytics are employed, by using a well-established iterative process for model building and evaluation. It used to be the case that the focus of this process was on the practical steps, whereas it has now become apparent that the pressing need is to be able to tell a story with the data (Knaflic, 2015). A story that aims to shed light on the issues raised in education.

1.2 Learning Analytics: An Attempt for an Efficient and Innovative Model for Educational Development

Education, as a living and rapidly transforming organism, needs a sophisticated method to collect, analyze and act upon data. Actually, higher levels of complexity demand more sophisticated ways of information processing. Business has benefited and ripped rewards for the insights provided by data analytics and it would not be an exaggeration to say that is has revolutionized the way commerce works today. It would be really useful if education

[a] Work by Vassilios S. Verykios was partially supported by a research grant from New York University Abu Dhabi.

Tsoni, R., Samaras, C., Paxinou, E., Panagiotakopoulos, C. and Verykios, V.
From Analytics to Cognition: Expanding the Reach of Data in Learning.
DOI: 10.5220/0007751904580465
In *Proceedings of the 11th International Conference on Computer Supported Education (CSEDU 2019)*, pages 458-465
ISBN: 978-989-758-367-4

stakeholders could find ways, consistent with pedagogical values and related to the educational demands, to apply these successful models with the same extraordinary results. That is all about Learning Analytics.

Learning Analytics involves analyzing data that has been collected from different educational contexts and environments, and from various levels with the view of discovering patterns. By gaining this knowledge, decisions can be made and changes can be implemented which will bring about favorable outcomes for the students and the institutions.

Hellenic Open University (HOU) is the only Open University in Greece. It offers exclusively distance and blended learning courses to mature students of diverse backgrounds, skills, and qualifications. The unique nature of this institution means that it confronts a unique set of problems as far as meeting the needs of its students is concerned. Finding ways to accommodate such a diverse population and optimize the learning experience is certainly a challenge that led to the creation of the Big Data Analytics and Anonymization Laboratory (BAT lab).

2 BAT LAB PRESENTATION

2.1 The Objective

BAT lab conducts research in the field of large-scale data management and analysis, in conjunction with privacy protection of this data with the view to understand student behavior and interaction both with their peers but also with the teaching staff. Its main purpose is to find ways to delve into the hidden patterns relating to student behavior, communication, engagement, performance and faculty guidance.

2.2 Difficulties and Barriers

Both children and adolescents are familiar with the idea of finding and using digital information. For them, data is ubiquitous and they pursue the creating, using and sharing of data. They own smartwatches to monitor their physical activity, they use social media to state their emotions and they measure their popularity by counting "likes" and "views". Unlike the new generation, a lot of academic members are still using conventional methods to get feedback from their students and theories to predict the efficiency of the educational material they use. Thus, a significant barrier is posed by a pre-existing culture or, in some cases, even a sense of technophobia. So, in order to obtain solutions, it is imperative that the data is analyzed in a way that is intuitive and at the same time more user-friendly, to reveal hidden models or patterns for prediction and summarization, which are not initially obvious (Siegel, 2013). A shift from metrics to analytics and from reporting to actionable interventions is the next generation of the learning environment. However, this transition will require a significant institutional change (Baer, and Campbell, 2012).

Another, more practical, barrier is the fact that almost all the available NLP (Natural Language Processing) tools do not support Greek language. Specific tools for sentiments analysis have been created by Greek researchers (Agathangelou et al., 2018) and are used by the team for analyzing students' fora (see section 3.4).

3 PROJECTS AND ONGOING RESEARCH

3.1 Outlining the Identity of HOU

HOU is University unique at its scope in Greece, as it serves the ideal of openness and inclusiveness. This openness leads to a big diversity amongst the students who choose to be educated through this system. So, it was crucial to be able to gather and analyze successfully all this information that comes from the students' admissions.

For example, data coming from HOU students' admission in the last four years analyzed in BAT Lab showed that there is a greater number of female applicants, although there is a progressing decline in the total number of applicants throughout the years (figure 1). School of Humanities is much more popular among the female applicants, whereas in the School of Science and Technology the male applicants far

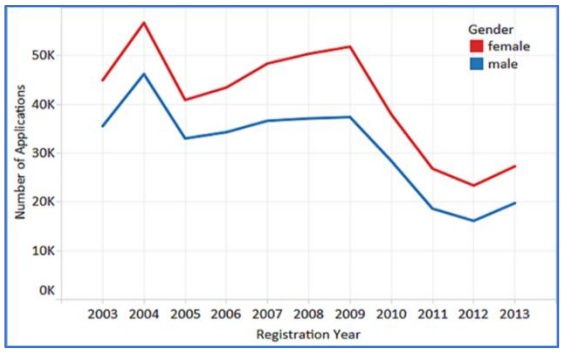

Figure 1: Number of applications per gender during 2003-2013 in HOU (Stavropoulos et al., 2017).

outnumber the female. According to figure 2, the highest percentage of the applicants comes from the regions of Attiki, Achaia, and Thessaloniki although, for the rest of the country, there is not a particular pattern that can be detected.

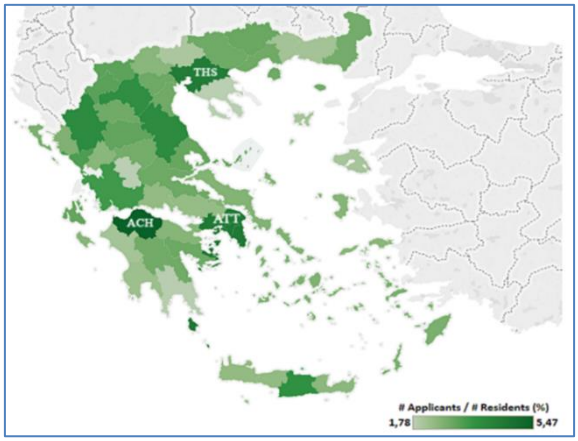

Figure 2: Applications per region (Verykios & Stavropoulos, 2018).

3.2 Assisting Tutors to Catch up with Students' Progress

The percentage of the students who submitted assignments and quizzes, or the level of the students' and the tutors' engagement, or other relevant information is very significant for online courses where the transactional distance is longer than it is in a face to face course (Moore, 2013). For example, data on how long does a student stay connected to the platform or on how many connections does she/he make to the platform on average per day, provides information on student's dedication time. The same information is also available about tutors (figure 3). It is notable that even though some tutors seem to dedicate the least amount of time, in comparison to their colleagues, they still make an adequate number of connections. This sort of information allows the module coordinator to create some guidelines, in a sense of best practices and benchmarks for the tutors in order for the latter to become more efficient in their work and for the coordinator to access more objectively tutors' work.

Another piece of information that could help the tutors become more effective in their roles, is the students' engagement in relation to the time of the day. It appears that students are more active in the evenings (figure 4) which is something that makes sense since the majority of the students in HOU are mature students who have family and job obligations during the rest of the day. Thus, it would be more

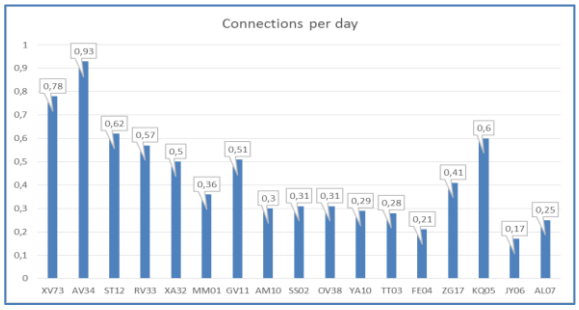

Figure 3: Tutors' connections per day.

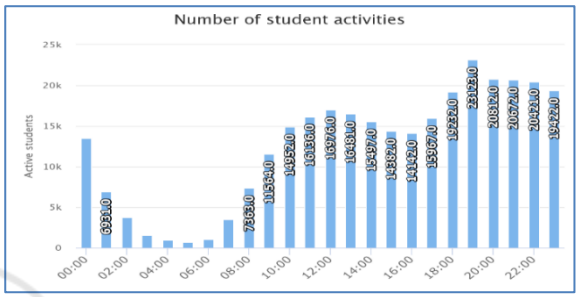

Figure 4: The distribution of active students and students' activities to daily hours (Gkontzis et al., 2017a).

productive for tutors to be also connected to the university platform in the evening rather than in the morning when students' engagement appears to be very low. As a result, every student communicates directly with the tutor, asking for advice and help and the whole educational process becomes more productive for both of them. The student progress bar (figure 5) shows how many of the allocated activities a student completes. From the overview page, the tutor can compare every student's performance to the performance of the whole class.

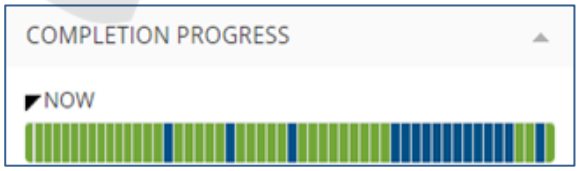

Figure 5: Students' progress bar (Gkontzis et al., 2017a).

All these important facts allow tutors to be supportive to students, provided that these data are easily accessible. Thus, visualization is a major concern. Learning Analytics dashboards had been successfully used, enabling tutors with no previous LA training to detect students' trends in order to provide personalized assistance (Gkontzis, et al., 2017).

Furthermore, students' logins, replies, and quizzes were blended with the average grade of the main written assignments throughout an academic

year. Useful observations from the students' educational online activities were made that can be used as predictive factors for their academic performance (Gkontzis, et al., 2018a). In a related work Gkontzis, et al. (2018b) presented an analytical framework that provides a comparison between students' actual predicted grade, identifying students who have included third-party services in their assignments.

3.3 Evaluating the Efficiency of Methods and Educational Material

While it is true that big data is a field of immense interest, unfortunately, the spotlight is on how to store, index, retrieve and interrogate the data rather than how to analyze and utilize it in an efficient and user-friendly way which is a shame as the answers to most problems are hidden within this ocean of data (Verykios & Stavropoulos, 2018).

In a project run by Paxinou, et al., (2017) the authors presented that the Classical Test Theory (CTT) method is not able to prove successfully the lead of a virtual reality laboratory in preparing Science students in HOU for the experiments in the wet lab, as this method doesn't separate the difficulty of the questions in the written tests from the students' ability. Contrariwise, the applications of more advanced theories and methods, as the Item Response Theory (IRT) allows highlighting the positive impact of a teaching method to the students' learning outcomes. This result poses a significant issue of the importance of choosing a suitable analysis framework in order to achieve a result that best describe reality. There is a great value in distinguishing the skill level of a student from the difficulty level of the questions of each evaluation test. This significant result that takes advantage of the IRT model, was proven from a massive data analysis of a several million recorded online games of chess held by Anderson from Microsoft, Kleinberg from Cornell University and Mullainathan from Harvard University (2016). Their research brought to light unquestionable proof for the players' behavior concerning decision making. The same way Guo et al., (2014) changed the way that educational videos are made by analyzing a vast number of MOOC data concerning students' engagement.

An interesting study highlighted the importance of data analysis for profiling new coming university students. The study was conducted through four successive years in freshmen students of the School of Education at Patras University in a collaborative project between BAT Lab and University of Patras (Panagiotakopoulos et al., 2017). Data analysis revealed the difference between students' opinion and their actual attitudes. Although they stated their acknowledgment for the significance of technology use for educational purposes, data showed that they barely take advantage of technology in their studies.

Concerning HOU students, several sets of data have been analyzed in order to find out the level of engagement with the educational material. The graphs in figure 6 illustrate the engagement students have with the content provided for them on their course. A large number of students who do not use the resources indicates that perhaps the content is problematic and needs to be rethought by the tutors and coordinator.

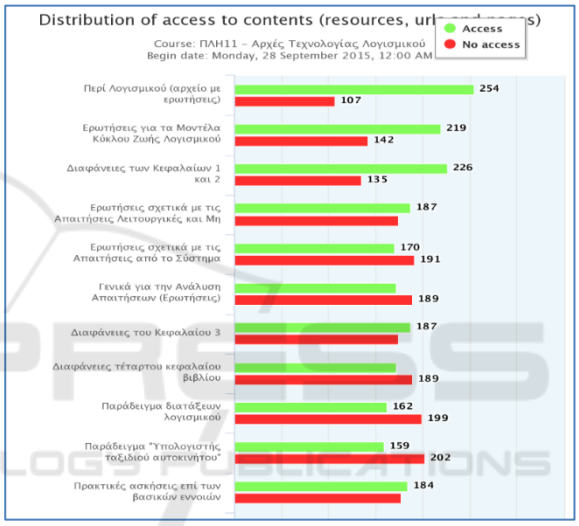

Figure 6: Students' access to content material.

3.4 Gaining Insight into Students' Behavior

In an effort to see beyond logins and submission dates, network analysis was conducted. These graphs (figure 7) show the interaction between students and tutors, and between the students themselves. This gives an overview of the communication that is taking place and it allows us to see how a tutor is interacting with his/her group.

Sentiment analysis of students' fora and emotion classification (figure 8) was held by Gkontzis et al., (2017a) in postgraduate students of the School of Science and Technology in HOU. The study revealed that positive polarity was dominant in students' posts. This fact was more obvious in the most active students, indicating an emotion of satisfaction.

Additionally, this work pointed out some issues for future research: possible correlation of sentiment with

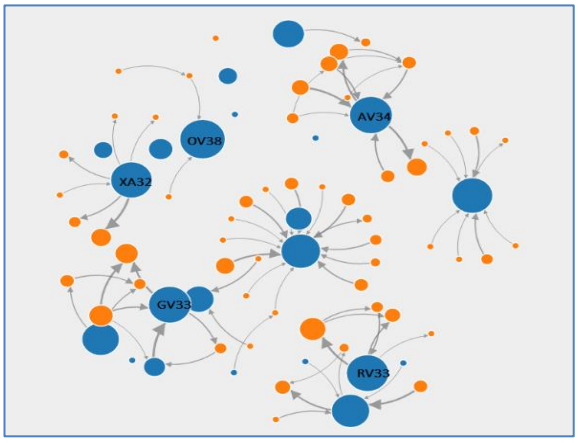

Figure 7: An integrated view of the groups' forum graph.

academic and demographic factors, more active students' sentiment compared with less active ones and the effect of tutors' methods in students' sentiment expressed in fora.

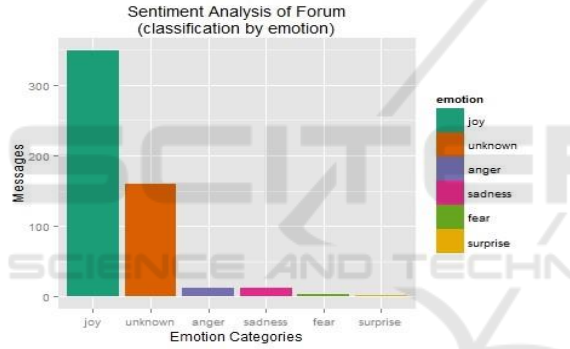

Figure 8: Emotion classification in forum posts (Gkontzis, et al., 2017b).

Text mining in forum data whereby individual terms are picked out of texts sent by students has been used in an attempt to better understand their educational needs. A high frequency indicates that students are having difficulties due to misconceptions or require further clarification. The tutors can intervene by providing more information on the subject or improving their teaching of the subject.

3.5 The Prime-EDU Software

Recently our team has designed and developed a new educational application named PRIME-EDU. Its purpose is to transfer educational data to a Cloud Database. These data are analyzed using Learning Analytics techniques so that the application users can easily make better educational decisions. The above-mentioned Data Base is designed in the form of Cloud

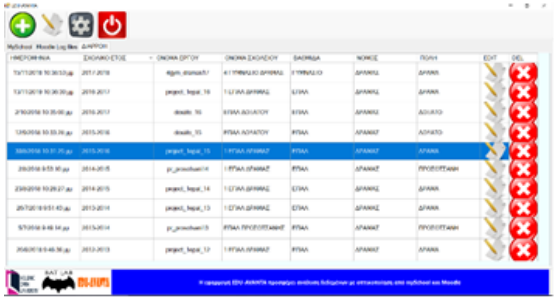

Figure 9: Data visualisation in PRIME-EDU software.

Warehouse Database and its data provide real-time Visualization, Correlation, and Prediction.

The PRIME-EDU application is designed to receive data from "MySchool", "Moodle", the "DIARROI" (STUDENTS DROP OUT) application and the "Classter" of Vertitech. "My School" is the web-based application of the Ministry of Education, that all schools of Secondary Education in Greece use compulsory. Moodle is one of the most established asynchronous e-learning systems that offers many opportunities for analyzing training data. It was also used by HOU. Finally, the DIARROI application is a training software that our team has developed which offers telematics services to reduce early school drop-out (ESL) and educational leakage (Samaras et al., 2018).

The PRIME-EDU application is designed with C# in Microsoft Visual Studio 2017. It makes use of the Microsoft Virtual Machine libraries. NET Framework 3.5 works on various versions of Windows.

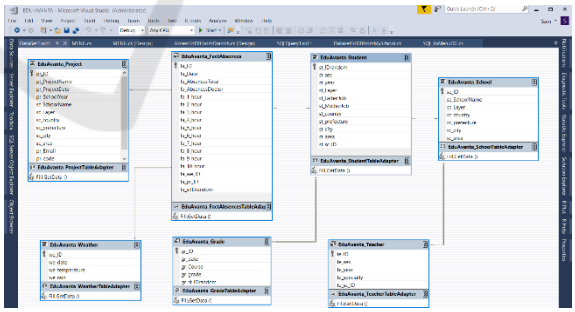

Figure 10: PRIME-EDU Interface.

The application automatically sends learning data to the Warehouse Database from Excel, Access, SQL SERVER. It is also texting information to parents, such as updating for their children's attendance. The basic requirement of the application is that the data sent to the Cloud Warehouse Database must be anonymous in order to comply with the European Data Protection Regulation. Besides that, personal data don't help us make educational decisions.

The PRIME-EDU application transfers data to the Cloud Warehouse Database, that has been developed on SQL Server 2017. The Warehouse Database is designed with a table of events and dimensional tables in a star model. The event table has each student's daily attendance and the dimensional tables contain the most features that affect the student's attendance such as environment, school teacher, family etc.

From the Cloud Warehouse Database, we isolate the data needed to make educational decisions. The technique we use is the business intelligence that companies with a lot of data have been using for many years in order to monitor the behavior of their customers and their organization. Thus, with CUDE and ROLLUS techniques, we are able to isolate an educational phenomenon and handle it per period, per region, per students' age etc. With the data that we isolate with CUBE techniques and the abilities (potentials) of the R Tool Visual Studio (RTVS) tool, we are able more easily visualize an educational phenomenon, to apply Correlation and look for the forecast.

3.6 Privacy and Anonymization

A significant field of concern is privacy protection and data anonymization. This concern also applies to the educational field. In order to protect students that provide data, it is necessary to omit those characteristics that permit identification. Relevant methods and techniques are presented in a series of successive works as the followings: Sakkopoulos et al., (2013) presented a framework that allows conducting research on anonymity techniques in a real-life environment using smartphones. The framework also includes logging mechanisms that facilitate positioning research dataset development in open format. Additionally, Karapiperis, & Verykios, (2016) proposed an efficient scheme for privacy-preserving record linkage by using the Hamming locality-sensitive hashing technique as the blocking mechanism and the Bloom filter-based encoding method for anonymizing the data sets at hand. Highly accurate results were achieved and simultaneously reduced significantly the computational cost by minimizing the number of distance computations performed. Moreover, Bit Vectors (BV), an accurate distance-preserving encoding scheme for representing numerical data values in privacy-preserving tasks, were used by Karapiperis, et al., in their work (2017). Key components of this embedding process were the employed hash functions and the threshold that is required by the distance computations, which they proved that can be specified in a way that guarantees accurate results.

Finally, it is noted that as the level of anonymization rises, the quality of data declines. Consequently, it is important to achieve a balance between tanking advance of data the best possible way and complying with privacy regulations.

4 CONCLUSIONS AND FUTURE WORK

4.1 A Holistic Approach

Overall, it is obvious that Learning Analytics and its application to different contexts and environments has the power to transform the educational system as a whole, as well as to tailor the learning experience to the specific needs of individual students which will lead eventually to their success. Using the resources available in BAT lab, we have been able to conduct research and experiment with the implementation of various Learning Analytics case studies with some promising outcomes. Even though we are at an initial stage we hope to be able to discover ways to revolutionize our learning and teaching environments by offering our students the best possible services.

4.2 From Learning Analytics to Cognitive Science

Since translating experience into words makes a major difference in future actions (Pennebaker et al., 2014), analyzing written text is meaningful in education as it can reveal insights that go deeper than grades or login counts. Therefore, the use of Natural Language Processing methods is crucial in an integrated educational data analytics approach.

Cognitive computing combines neuroscience, supercomputing, and nanotechnology to develop a coherent, integrated, universal mechanism inspired by the mind's capabilities Modha et al., (2011). Learning Analytics combined with State-of-the-art Cognitive Computing can provide new effective ways for content management, especially in Distance Learning Universities where the educational material has to fill the gap in the lack of tutor's physical presence. An example is the work of Dessì et al., (2018) where they used the aforementioned methods to manage the content of micro-learning videos in order to improve their retrieval from students. By using sophisticated systems that provide assistance to

students, at the same time a pool of data is created, allowing the gaining of insight into their behavior and their way of learning. Cognos Analytics provides a set of resources for data analysis that can uncover hidden patterns that could lead to significant conclusions even in cases when the initial hypothesis is not clearly stated. This kind of serendipity is what we consider as the main benefit of the employment of continuously integrated methods and tools.

4.3 The Future Scope

In the future, we hope to be able to create increasingly sophisticated applications and gadgets and to create a kind of augmented reality environment which can communicate in real time how the student is performing. By being able to visualize their progress both the students and the faculty will be able to avoid negative outcomes, such as poor grades or even worse students' dropouts. In a recent study, Brown et al., (2019) provided evidence that large scale data form application users via smartphones can outweigh the noise inherent in collecting data outside a controlled laboratory setting and produce valid results. This conclusion supports the idea of the creation of an online educational application that would gather and present to students selected data in an understandable and useful way in order to enhance their self-regulation and their awareness about their progress. The goal is to make optimal use of all available data. After all, as Schmidt (2011) stated: "*Technology is not really about hardware and software anymore. It's really about the mining and use of this enormous volume of data in order to make the world a better place…"*.

ACKNOWLEDGMENTS

The co-authors would like to acknowledge the support of New York University Abu Dhabi.

REFERENCES

Agathangelou, P., Katakis, I., Koutoulakis, I., Kokkoras, F., and Gunopulos, D. (2018). Learning patterns for discovering domain-oriented opinion words. Knowledge and Information Systems, 55(1), 45-77.

Anderson, A., Kleinberg, J., and Mullainathan, S. (2017). Assessing human error against a benchmark of perfection. ACM Transactions on Knowledge Discovery from Data (TKDD), 11(4), 45.

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine 16.07. Retrieved by: https://www.wired.com/2008/06/pb-theory/

Baer, L., and Campbell, J. (2012). From metrics to analytics, reporting to action: Analytics' role in changing the learning environment. Game Changers: Education and Information Technologies, Educause, 53-65.

Brown, H. R., Zeidman, P., Smittenaar, P., Adams, R. A., McNab, F., Rutledge, R. B., and Dolan, R. J. (2014). Crowdsourcing for cognitive science–the utility of smartphones. PloS one, 9(7), e100662.

DeMillo, R. A., and Young, A. J. (2015). Revolution in higher education: How a small band of innovators will make college accessible and affordable. MIT Press.

Dessì, D., Fenu, G., Marras, M., and Recupero, D. R. (2018). Bridging learning analytics and Cognitive Computing for Big Data classification in micro-learning video collections. Computers in Human Behavior.

Gkontzis, A. F., Karachristos, C. V., Lazarinis, F., Stavropoulos, E. C., and Verykios, V. S. (2017a). Assessing Student Performance by Learning Analytics Dashboards. 9th International Conference in Open and Distance Learning, 9(1A), 101-115.

Gkontzis, A. F., Karachristos, C. V., Panagiotakopoulos, C. T., Stavropoulos, E. C., and Verykios, V. S. (2017b). Sentiment Analysis to Track Emotion and Polarity in Student Fora. In *Proceedings of the 21st Pan-Hellenic Conference on Informatics* (p. 39). ACM.

Gkontzis, A., Kotsiantis, Panagiotakopoulos, C., &Verykios V. (2018a). Measuring Engagement to Assess Performance of Students in Distance Learning. Proceedings of the 9th International Conference on Information, Intelligence, Systems and Applications, Zakynthos, Greece. IEEE

Gkontzis, A., Kotsiantis, S., Tsoni, R., and Verykios, V. (2018b). An Effective LA Approach to Predict Student Achievement. In Proceedings of the 22nd Pan-Hellenic Conference on Informatics. ACM.

Guo, P. J., Kim, J., and Rubin, R. (2014). How video production affects student engagement: An empirical study of mooc videos. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 41-50). ACM.

Karapiperis, D., and Verykios, V. S. (2016). A fast and efficient Hamming LSH-based scheme for accurate linkage. Knowledge and Information Systems, 49(3), 861-884.

Karapiperis, D., Gkoulalas-Divanis, A., and Verykios, V. S. (2017). Distance-aware encoding of numerical values for privacy-preserving record linkage. In Data Engineering (ICDE), 2017 IEEE 33rd International Conference on (pp. 135-138). IEEE.

Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. John Wiley and Sons.

Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. John Wiley and Sons.

Lindstrom, M. (2010). Buy ology: Truth and lies about why we buy. Crown Business.

Modha, D. S., Ananthanarayanan, R., Esser, S. K., Ndirango, A., Sherbondy, A. J., and Singh, R. (2011). Cognitive computing. Communications of the ACM, 54(8), 62-71.

Moore, M. G. (2013). The Theory of Transactional Distance. In Handbook of distance education (pp. 84-103). Routledge.

Panagiotakopoulos, C., Koustourakis, G., Samaras, C., Stavropoulos, I., and Verykios, V., (2017). The Transition from High School to University: Means of Learning and Technology Awareness for Junior Students - Trainee Teachers. 9th International Conference in Open and Distance Learning. 9(2A), 138-153.

Paxinou, E., Sgourou, A., Panagiotakopoulos, C., and Verykios, V. (2017). The item response theory for the assessment of users' performance in a biology virtual laboratory, Open Education: the journal for Open and Distance Education and Educational Technology, 13(2),107-123. doi:http://dx.doi.org/10.12681/jode.14618

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., and Beaver, D. I. (2014). When small words foretell academic success: The case of college

Sakkopoulos, E., Mersini, P., Tsakalidis, A., Sioutas, S., and Verykios, V. (2013). A Novel Mobile Framework for Anonymity Techniques and Services Research. In Mobile Data Management (MDM), 2013 IEEE 14th International Conference on (Vol. 1, pp. 353-355). IEEE.

Samaras, C., Panagiotakopoulos, C., and Verykios, V. (2018). Διερεύνηση της φοίτησης με Learning Analytics: Μελέτη της ετήσιας φοίτησης των μαθητών του ΕΠΑ.Λ. Δοξάτου 2016-2017. 11th Pan-Hellenic & International Conf. ICT in Education

Siegel, E. (2013). Predictive analytics: The power to predict who will click, buy, lie, or die (p. 148). Hoboken: Wiley.

Stravropoulos, E. C., Panagiotakopoulos, G. T., Kagklis, V., Lionarakis, A., Marketos, G., and Verykios, V. S. (2017). Student Admission Data Analytics for Open and Distance Education in Greece. The Journal for Open and Distance Education and Educational Technology, 13(2), 6-16.

Verykios, V. S., and Stavropoulos, E. C. (2018). Exploring the Power of Learning Analytics. The Envisioning Report for Empowering Universities, 9.