Chapter 2
Educational Data Mining & Learning Analytics

2.1 Introduction

There has been increasing interest among researchers and practitioners in Learning Analytics (LA) and Educational Data Mining (EDM) in recent years. Through designing computer-aided learning systems and automated processing of educational data, several attempts were made to improve the learning experience (Schroeder, Thüs, & Technologies 2012). In 2011, the Horizon report claimed for a fruitful future of LA (Johnson, Smith, Willis, Levine, & Haywood, 2011): LA is seen as an essential tool to uncover the knowledge and trends concealed from raw data obtained from the educational environment (Siemens, 2012). For this cause, knowledge of LA is raised, and essential ties with data-driven research fields such as data mining and machine learning (ML) are strengthened.

The combined use of LA, a modern field of research that can have a high potential for impacting current educational models (Siemens, 2012), and EDM, a novice growing field of research in the application of data mining techniques for educational data (Bousbia & Belamri, 2014), leads to new inspections of learner behavior, relationships, and learning pathways. In this connection, LA and EDM can give opportunities and great potential to enhance our understanding of learning processes to improve learning through education systems. We should educate learners, teachers, and their institutions and enable them to understand how such useful tools should offer tremendous advantages in learning and progress in educational outcomes by personalizing and adjusting education based on learners' needs (Greller & Drachsler, 2012). Such prospects have been enhanced by a significant change in data re-sources availability. This is a motivating basis for growing research in the area: PSLC DataShop and enriched educational data from MOOCs are examples (Baker & Yacef, 2009). These repositories are also known as benchmarks for advancing current approaches and algorithms in conjunction with other algorithms (Verbert, Manouselis, Drachsler, & Duval, 2012).

S. K G, M. Kurni, *A Beginner's Guide to Learning Analytics*, Advances in Analytics for Learning and Teaching,
https://doi.org/10.1007/978-3-030-70258-8_2

2.2 Educational Data Mining (EDM)

Educational data mining refers to techniques, tools, and research to automatically extract meaning from broad data repositories created by or linked to people's educational activities. Often these data are detailed, fine-grained, and precise. For example, some LMSs monitor details such as the number of times each student viewed the learning object and the number of minutes the learning object was displayed on the computer screen. As another example of intelligent tutoring systems, each time a learner proposes a solution to a problem, they will collect the time of submission, whether or not the solution matches the solution anticipated, time spent since the last submission, the order of solution components entered in the interface, etc. The specific data provides a great deal of process data for review, even in a relatively short session with a computer-based learning environment (e.g., 30 minutes).

In other instances, the results are less fine-grained. For instance, a student's university transcript may include a list of courses taken by the student, the student's degree in each course, and their chosen or changed academic major. EDM uses all data types to identify relevant information about and how different learners learn the domain knowledge structure and the impact of instructional approaches implemented into various learning environments. These analyses offer new information that is hard to distinguish from the raw data. For example, the LMS data analysis will show the connection between students' learning objects and their final grades. Similarly, the analysis of student transcript data can reveal a link between an individual course's degree and its decision to change its academic major. Such knowledge helps students, teachers, school administrators, and educational policy-makers decide how to communicate, deliver, and handle their education resources.

2.2.1 Timeline of Significant Milestones in EDM

Educational data mining can be interpreted in two ways: a research group or a scientific research field. EDM can be used as a sister community for learning analytics as a research environment. In a series of workshops starting in 2005, Educational Data Mining was the first to become an annual conference in 2008 and spawn a journal in 2009 and society, the International Educational Data Mining Society, in 2011. Here you can view a timeline of critical events for the formation of the EDM community (Baker & Inventado, 2014).

- 1995 – Bayesian Knowledge Tracking Paper by Corbett & Anderson – the early primary algorithm still popular today
- 2000 – First workshop related to EDM
- 2001 – Zaiane's theoretical paper on EDM methods
- 2005 – The first workshop using the word "educational data mining."
- 2006 – First EDM book published: "Data mining in E-learning," Romero & Ventura

- 2008 – First Education Data Mining International Conference
- 2009 – First issue of Journal of EDM
- 2010 – EDM first handbook published, Romero, Ventura, Pechenizkiy, and Baker
- 2011 – First conference on Learning Analytics and Knowledge
- 2011 – IEDMS has been established
- 2012 – SoLAR was founded
- 2013 – First Summer Institute for Learning Analytics
- 2022 – Analytics and data mining will be involved in all the educational research.

2.2.2 Goals of EDM

The following four EDM goals were identified by Ryan S. Baker and Kalina Yacef (2009):

1. **Prediction of the future learning behavior of students** through the development of models for students that incorporate detailed information including knowledge, motivation, metacognition, and attitudes;
2. **Discover or develop domain models** characterizing the learning material and optimal instructional sequences;
3. **Studying the impact of various forms of pedagogical support** offered through learning software;
4. **Advancing scientific knowledge about learning and learners** by creating computational models integrating student models, the environment, and the software's pedagogy.

2.2.3 Users and Stakeholders

Four primary users and stakeholders are interested in the mining of educational data (Wikipedia, 2019). Such comprise:

1. **Learners** – Learners want to consider students' needs and approaches to improve learners' understanding and success. For example, learners may also use the information they have gained to recommend activities and tools based on experience with the online learning tool and experiences from the past and the like. Educational data mining can also warn parents about the learning success of their children for younger students. In an online environment, it is also essential to effectively group students. The task is to learn these groups based on complex data and establish models for interpreting these groups.
2. **Educators** – Educators try to understand the nature of learning and the approaches used to develop their teaching methods. EDM applications can be used by educators to assess how the curriculum should be structured and organized, the best approaches for delivering course information, and the resources

to entice students to achieve optimum learning outcomes. In particular, distilling human judgment data offers educators the ability to benefit from EDM, as it allows educators to quickly recognize patterns of behavior, which can promote their learning practices throughout the course or enhance future courses. Educators should identify metrics that reflect student satisfaction and adherence to materials and track success in learning.

3. **Researchers** – Researchers concentrate on designing and testing effective data mining techniques. A yearly international conference started in 2008, followed by the Educational Data Mining Journal in 2009. The broad range of EDM topics includes data mining to increase institutional productivity and student success.
4. **Administrators** – Administrators are responsible for allocating resources in organizations for implementation. As institutions are increasingly accountable for students' achievement, the administration of EDM applications in educational environments is becoming increasingly popular. The faculty and consultants are becoming more involved in recognizing and discussing students at risk. However, often it is a task to provide decision-makers with the knowledge to handle the application quickly and efficiently.

2.2.4 Phases of EDM

With the continued advancement of work in educational data mining, many data mining techniques have been applied to various educational backgrounds. The aim is to turn raw data in each case into concrete knowledge on the learning process to make informed decisions on the design and direction of a learning environment. EDM is thus usually composed of four phases (Romero & Ventura, 2010; (Baker & Yacef, 2009):

1. The first phase of the EDM process (no preprocessing) detects relationships in data. This includes filtering data from an educational environment via a repository to find coherent correlations between variables. Several algorithms have been used to classify these relationships, including classification, regression, clustering, factor analysis, social network analysis, association rules, and sequential pattern mining.
2. To prevent overfitting, discovered relationships must also be validated.
3. Validated relationships are used to predict future events in the learning world.
4. Predictions are used to facilitate strategy and decision-making processes.

During phases 3 and 4, data are also visualized or refined for human interpretation through any other means (Baker, 2010). There has been a significant amount of work into best practices for data visualization.

2.2.5 Main Approaches

In educational data mining, a wide range of current methods are available. These methods fall into the following categories (Baker, 2010).

- Prediction
- Clustering
- Relationship mining
- Discovery with models and
- Distillation of data for human judgment

The first three categories are generally accepted as typical for data mining types (although some have different names). The fourth and fifth groups are especially common in the field of educational data mining.

Prediction In prediction, the aim is to create a model that can infer from a particular combination of certain data aspects (predictive variable) a single aspect of the data. Prediction requires that labels be given for a specific data set in the output variable, where a label reflects some accurate "ground truth" information about the value of the output variable. In some instances, though, it is important to understand how provisional or incomplete these labels can be.

A prediction has two main applications for educational data mining. In some cases, prediction methods may be used to research which characteristics of a model are important for prediction and provide information on the underlying structure. In the second form of application, prediction procedures predict the result value in contexts where the label for that construct cannot be directly obtained.

There are three prediction types:

- Classification
- Regression
- Density estimation

The predicted variable in classification is a binary or categorical variable. Some of the standard classification methods include decision trees, logistic regression, and vector support machines. The expected variable is continuous in the regression. Standard regression methods include linear regression, neural networks, and support vector machine regression in educational data mining. The predicted variable is a probability density function in the density estimation. Several kernel functions can be used, including Gaussian functions. For each prediction form, the input variables can be categorical or constant; depending on the type of input variables used, various prediction methods are more accurate.

Clustering The clustering goal is to find naturally clustered data points, which divide the entire collection of data into a set of clusters. Clustering is particularly useful when the most common categories in the data set are not identified beforehand. If some clusters are optimal within a grouping, every data point is usually more similar than data points in the other clusters. Clusters may be generated in a

variety of potential grain sizes, such as the clustering of schools (e.g., to research similarity and difference between schools), the clustering of students (e.g., to examine similarities and differences between students), or the clustering of student activities (e.g., to study behavior patterns).

Clustering algorithms may either start without any previous hypotheses about data clusters (such as the randomized rebooting of the k-mean algorithm) or start from a particular hypothesis that could be generated with a different dataset in prior research (using the Expectation-Maximization algorithm to move on to a cluster hypothesis for the new data set). A clustering algorithm could support the hypothesis that each data point would belong to one cluster (e.g., k-means) or that other points should belong to more than one cluster or no clusters (e.g., Gaussian Mixture Models).

The quality of a set of clusters is generally measured based on how well the set of clusters match the data in contrast with how many matches the number of clusters could only be predicted by chance using statistical methods, for example, the Bayesian Knowledge Criterion.

Relationship Mining Relationship mining discovers relationships between variables in a data set with a wide range of variables. This could be achieved by attempting to decide which variables are most closely correlated with one variable of particular concern or by trying to determine which relationships are strongest between any two variables.

Four forms of relationship mining occur in general:

- association rule mining
- correlation mining
- sequential pattern mining
- causal data mining

In association rule mining, the goal is to decide if-then rules of the form generally have a particular value if any variable values are found. In correlation mining, the objective is to find linear (positive or negative) correlations between variables. The goal of sequential pattern mining is to find temporal associations between events. Causal data mining aims to evaluate whether an event (or observed construct) has been responsible for another event (or observed construct) by evaluating either the covariance of both events or using knowledge about how one event is initiated.

Relationships found by relationship mining must fulfill two criteria: statistical significance and interestingness. Standard statistical tests, such as F-tests, usually determine the statistical significance. Since several checks are performed, it is essential to verify relationships by chance. One approach is to apply post-hoc statistical methods or modifications to the number of tests performed, such as Bonferroni adjustment. This approach will improve confidence that there is no possibility of an individual relationship. An alternative approach is to determine, using Monte Carlo techniques, the overall likelihood of success. This approach tests how likely the overall pattern of outcomes occurred because of chance.

To minimize the set of rules/correlations/causal relations communicated to the data miner, each finding's interestingness is evaluated. Hundreds of thousands of significant relationships can be found in comprehensive data sets. Interestingness measures aim to decide which outcomes are the most distinctive and well supported by the evidence, and others seek to obtain too many similar results. Various measures of interest, including support, confidence, conviction, lift, leverage, coverage, correlation, and cosine.

Discovery with Models In developing a model, a phenomenon model is built by prediction, clustering, or knowledge engineering. This model is then used in another analysis as a component, such as prediction or relationship mining.

In the prediction case, the generated model's predictions are used in predicting a new variable as predictor variables. The relationship between the predictions of the generated model and additional variables are studied in relationship mining. This allows an investigator to examine the relationship between a complex latent construct and a wide range of observed constructs.

Often, model discovery leverages the validated generalization of an integrated prediction model. Generalization in this way depends on adequate validation that the model generalizes correctly across contexts.

Distillation of Data for Human Judgment The distillation of data for human judgment is another area of interest in educational data mining. In some instances, people may conclude data outside the immediate reach of fully automated data mining methods if addressed appropriately. The tools in this area of education data mining are tools for information visualization. The most commonly used visualizations within EDM are often different from those most often used for other information visualization problems.

Data is refined in education data mining for human judgment for two primary purposes: identification and classification. Data are distilled for identification in ways that allow a person to recognize well-known patterns that are difficult to express formally easily. Alternatively, human labeling data may be refined to enable the subsequent development of a prediction model. In this case, sub-sections of a data set are played in visual or text format with human coders named. These labels are then typically used as the basis for predictor growth.

2.2.6 Main Applications

Many educational data mining applications are in practice; this section addresses four applications that have gained considerable attention in the area (Baker, 2010).

One main field of application is ***improving student models*** that provide detailed information on students' characteristics or conditions such as knowledge, motivation, metacognition, and attitudes. To allow the software to adapt to individual differences, modeling individual differences between students is crucial in educational

software research. In recent years, educational data mining methods have allowed student models' complexity to be dramatically expanded. In particular, educational data mining has allowed researchers to obtain more information on students' actions, such as when a student is gaming in the system, when a student has "slipped" (despite knowing an error), and when a student is engaged self-explanation. These richer models were useful in two respects. Firstly, these models improve the ability to forecast student awareness and future success. Second, these models help researchers to investigate what factors causing students to choose a particular learning environment.

A second primary area of application is **discovering or improving models of the knowledge structure of the domain**. Methods for discovering accurate domain models directly from data have been built in educational data mining. These methods typically combine psychometric modeling frameworks with advanced space-searching algorithms and typically present them as predictive problems for model detection purposes (for example, attempting to predict whether individual behavior using different domain models are correct or incorrect is one common way of developing those models).

A third main application field is the **study of pedagogical support** by learning tools. Modern educational software provides students a range of pedagogical support. The discovery of the pedagogical support is most successful for educational data miners was a key area of concern. The decomposition of learning, a form of relationship mining, is a match for exponential learning curves, which link student performance to the quantity of pedagogical help each student receives (with a weight for each form of aid). The weight of each form of pedagogical support shows how successful it is to enhance learning.

A fourth main area for applying educational data mining is a **scientific discovery** into learning and learners. This takes different forms. In each of the three areas listed before, the application of educational data mining can have broader empirical advantages; for instance, the analysis of pedagogical support can have long-term potential to enrich scaffolding theories. However, within these three fields, multiple analyses were geared directly towards scientific discovery. Model discovery is a crucial tool for scientific exploration through educational data mining. Research into whether status factors or features are stronger predictors of how often a student would game the system is a prominent example of this approach within educational data mining research. Learning decomposition methods are another effective form of scientific study on learning and learners.

2.3 Educational Data Mining & Learning Analytics

Educational data has been increased on a large scale through e-learning tools, instrumental educational software, the use of the Internet in education, and the development of state databases of student knowledge. For several years mainstream educational institutions have used information systems that store lots of exciting

information. Today, web-based education systems have grown exponentially, allowing us to store many possible data from different sources in various formats and granularities (Romero & Ventura, 2017).

Many students' data is also gathered in different education settings, such as blended training (BL), virtual/enhanced environments, mobile/ubiquitous learning, game learning, etc. These systems generate large quantities of high educational value knowledge, which cannot be analyzed manually. Tools are also essential to automatically analyze this type of data as all of this information offers students a wealth of educational knowledge to explore and manipulate to understand how students learn. The rapid increase of education data and the transformation of these data into new insights that can support learners, teachers, and administrators are currently among the most significant challenges facing education institutions (Baker, 2015).

There were two separate communities in the same field with a shared interest in how educational data can be used in education and learning science (Baker & Inventado, 2014):

- **Educational data mining (EDM)** aims to develop methodologies for exploring specific data types from educational environments (Bakhshinategh, Zaiane, ElAtia, & Ipperciel, 2018). The application of data mining techniques (DM) to this specified form of data set from educational environments could also be described to deal with critical educational issues (Romero & Ventura, 2013).
- The measurement/collection, analysis, and reporting of data concerning learners and their contexts for understanding and optimizing learning and the environments in which it takes place (Lang, 2017) can be described as **Learning Analytics (LA)**.

Both groups share a mutual interest in data-intensive approaches to research in education and share the goal of improving education (Siemens & Baker, 2012; Calvet Liñán & Juan Pérez, 2015). LA focuses on the education issue on the one side, and EDM concentrates on the technical challenge. LA focuses on decision-making based on data and combining the technological and social, and pedagogical aspects of learning through established models. On the other hand, EDM typically explores new data patterns and creates new algorithms and/or models. In the end, the discrepancies among the two groups are focused more on emphasis, study questions, and the potential use of models than the methods used (Baker & Inventado, 2014). Notwithstanding the variations between the LA and EDM communities, both the goals of investigators and the approaches and strategies used in the inquiry have substantial similarities.

EDM and LA are interdisciplinary fields, including knowledge collection, advocacy programs, visual data processing, domain-based data mining, social network analysis, psycho-pedagogical research, cognitive science, psychometrics, etc. They can be drawn as a mixture of three main fields (Fig. 2.1): computer science, education, and statistics. Other sub-areas closely linked to EDM and LA, such as CBE, data mining and machine learning, and educational statistics, are also established at the intersection in these three industries.

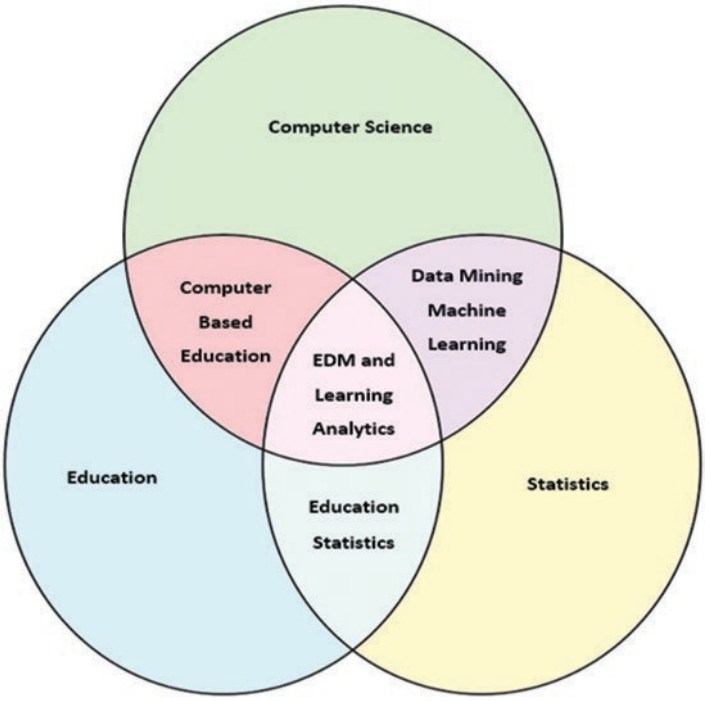

Fig. 2.1 Main areas related to Educational Data Mining/Learning Analytics (Romero & Ventura, 2020)

2.3.1 Benefits of LA and EDM

The advantages of LA and EDM are further clarified in several studies. For example, a UNESCO policy brief defines LA advantages at various micro, meso, and macro levels (Buckingham Shum, 2012). The three key stakeholders are educators, learners, and administrators. Educators are responsible for developing and implementing curriculum programs, and they are most aware of the learning process, the needs, and standard errors of students. The availability of real-time feedback on learners' success allows this community to adapt their teaching activities to students' needs.

Learners seek guidance and input on their learning activities, resources, and paths. The input the students get can be inspiring and encouraging. Finally, administrators handle decision-making and the budget allowance and control the program development and learning process (Vahdat et al., 2015).

Generally speaking, in both fields, the main aim is to enhance learning and get insights into learning processes. LA and EDM are useful for anticipating possible learning patterns to input and adjust implementing methods based on the learner's attitudes. They are also helpful in finding and developing learning domain models

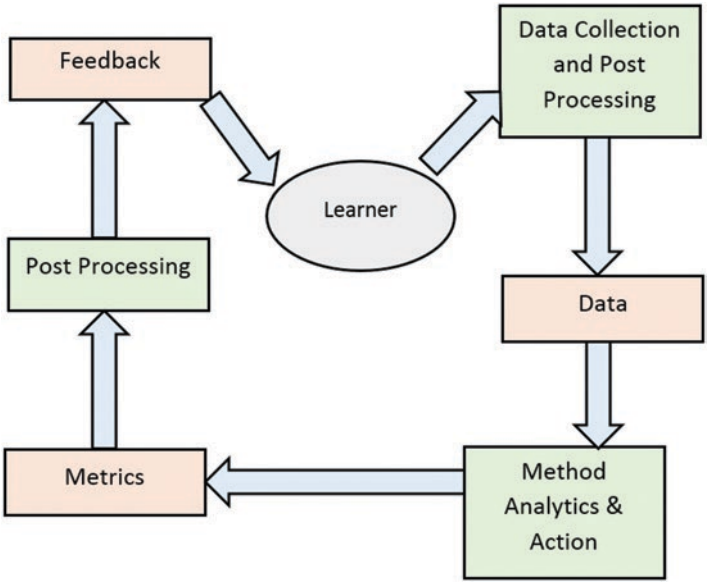

Fig. 2.2 An LA/ EDM process (Vahdat et al., 2015)

and testing learning materials and training tools. They can also advance scientific awareness of students, identify their irregular conduct, and problems and enhance pedagogical support through learning software (Bienkowski, Feng, & Means, 2014; He, 2013). Such study fields are considered complementary, and both present opportunities and challenges (Papamitsiou & Economides, 2014). Figure 2.2 illustrates the LA-EDM process for data collection, processing, and feedback for students to influence and improve the learning outcome.

2.3.2 Similarities and Distinctions

The overlap between the two research fields is important. Nevertheless, there are several variations in the literature. EDM and LA seek to enhance education quality by analyzing vast data quantities to obtain stakeholders' valuable knowledge. Representative organizations in other fields, such as industry, finance, or healthcare, have also adopted statistical, machine-learning, and data-mining technology to boost efficiency through historical data-based decisions. The popularity of these research areas has increased since the beginning of the 2010s, while EDM research began several years earlier. The future gains for students, instructors, administrators, researchers, and society in general, and the importance of current research-based on big data, are expected to continue to be increased in these fields (NMC Horizon Report > 2012 Higher Education Edition, 2012).

LA and EDM have many similarities and common purposes and goals, but there are also significant differences. According to (Siemens & Baker, 2012) following five main differences between EDM and LA are noticeable.

- ***Discovery***: researchers in EDM are interested in automated exploration, and it is an instrument for this to exploit human judgment; it is quite the reverse in LA; the aim is to exploit human judgment.
- ***Reduction and holistic***: EDM reduces component systems and analyzes them and their relationships, while LA needs to grasp systems in their entirety.
- ***Origin***: EDM is rooted in educational software and student modeling; LA roots, on the contrary, have to do with semantic web, "intelligent curriculum," outcome prediction, and systemic interventions.
- ***Adaptation and personalization***: EDM accomplishes automatic adaptation, while LA advises and activates instructors and students.
- ***Techniques and methods***: EDM employ further classification, clustering, Bavarian modeling, relationship mining, model creation, and visualization; while LA focuses on the study of social networks, sentiment analysis, sentiment analysis, influence analysis, discourse analysis, performance prediction of the learner, concept analysis and sensory models.

Such differences reflect large patterns in each group and, therefore, do not determine the entire scope. A similar concept is articulated in (Baker & Inventado, 2014), where it says that "the overlap and differences between the communities are largely organic, developing from the interests and values of specific researchers rather than reflecting a deeper philosophical split."

Bienkowski et al. (2014) assume that LA is more subject to discipline than EDM. LA is linked to information science and sociology and computer science, statistics, psychology, and learning sciences. Therefore, even though both fields' boundaries are complex and their distinctions are based partly on their backgrounds and patterns, they remain important for these authors. Furthermore, both study groups' co-existence, as upheld in (Siemens & Baker, 2012), contributes to a more diverse and significant contribution to society. Communication and competition between the two should, therefore, be promoted. The significant differences between EDM and LA are discussed in Table 2.1 (Calvet Liñán & Juan Pérez, 2015).

2.3.3 *Educational Data Mining/Learning Analytics (EDM/LA) Methods*

EMD and LA have a wide variety of standard methods for solving educational or application problems. The most common EDM / LA methods are as follows (Romero & Ventura, 2020).

Table 2.1 Significant differences between EDM and LA

Differences	EDM	LA
Techniques	Clustering, classification, Bayesian modeling, a discovery with models, and relationship mining	Visualization, statistics, sentiment analysis, social network analysis, discourses analysis, influence analysis, concept analysis, and sense-making models
Origins	Student modeling, educational software, and course outcomes prediction	Intelligent curriculum, semantic web, and systematic interventions
Emphasis	Description and comparison of the data mining techniques used	Description of data and results
Type of discovery	Automated	Making use of human judgment
Data used	Mostly administrative data	Pedagogical, administrative, and other types of data
Goals	Inform education practice	Influence education practice

- ***Causal mining*** – used for causal relationships or for finding causal effects in data. This approach will determine the characteristics of students' behavior that trigger learning, academic failure, drop-out, etc.
- ***Clustering*** – is used to classify related observation groups. Using this approach, we may group related materials or students based on their study and interaction patterns.
- ***Discovery with models*** – Used to employ a previously established model of a phenomenon as a component in another analysis. Using this approach, relationships between student behaviors and characteristics or contextual variables are established.
- ***Distillation of data for human judgment*** – This approach is used for intelligibly representing data using summarization, visualization, and interactive interfaces. This approach can help teachers interpret and evaluate the students' current activities and use of knowledge.
- ***Knowledge Tracing*** – This approach helps students test abilities, using a cognitive model that maps a problem-solving item with the skills needed and records of correct and incorrect responses to prove their knowledge of a given ability. We can track student awareness over time by using this method.
- ***Nonnegative matrix factorization*** – this technique is used to describe a matrix of positive numbers with student test outcomes, which can be decomposed into a matrix of items and a matrix of student mastery of skills. We can test student skills using this method.
- ***Outlier detection*** – This technique is used to denote significantly different people. By using this approach, students with disabilities or abnormal learning processes can be identified.
- ***Prediction*** – uses this technique to infer from a variety of certain variables, the goal variable. With this approach, we can forecast the success of students and detect behaviors of students.

- ***Process mining*** – Used to gain process information from event logs. Using this approach, we can find students' actions across the educational system based on traces of their evolution.
- ***Recommendation*** – Used to predict a user's rating or choice of an item. Using this form, we can give students feedback about their activities or assignments, links to visits, issues, courses to be carried out, etc.
- ***Relationship mining*** – This approach is used for studying relationships between variables and encoding rules. By using this approach, we can recognize relationships in student behavior patterns and diagnose student problems.
- ***Statistics*** – This method is used to compute statistics that are descriptive and inferential. This approach helps us to evaluate, interpret, and draw conclusions from educational data.
- ***Social network analysis*** – This approach is used in networked information to analyze social relationships between individuals. Using this approach, the role and relationship in group activities and experiences with communication tools can be interpreted.
- ***Text mining*** – This method is used for extracting high-quality information from text. Through using this method, forums, chats, web pages, and documents can be analyzed.
- ***Visualization*** – This approach is used to display data graphics. Using this approach, we can create data visualizations that allow educators to communicate EDM/LA research results.

2.4 Educational Data Mining & Learning Analytics Applications

Educational data mining and Learning analytics begin to address increasingly complex questions about what students know and whether they are engaged. Researchers have experimented with new model-building techniques and new learning system data that have proven promising to predict students' performance. This section discusses various applications (Alani, Tawfik, Saeed, & Anya, 2018) using EDM and LA technologies to customize and enhance teaching and learning.

- Student knowledge modeling
- Student behavior modeling
- Student experience modeling
- Student profiling
- Domain modeling
- Learning component analysis and instructional principle analysis
- Trend analysis
- Adaptation and Personalization

Each of these areas of application uses specific data sources; this section briefly discusses questions that these categories address and lists the data sources used in such applications (Table 2.2).

Table 2.2 Data sources

Application area	What can be achieved?	What data needed?
Student knowledge modeling	What content does a student know (for instance, necessary skills and concepts, procedural awareness, and higher comprehension skills)	Student answers (correct, incorrect, partly correct), time spent before answering a prompt or question, hits requested, repeating wrong answers, and errors made. The skills a student mastered and all the realistic possibilities. The performance level of students derived from system function or obtained from other sources such as standardized tests
Student behavior modeling	What do student behavior patterns mean for their learning? Are students inspired?	Student answers (proper, wrong, partly correct), time spent before answering a question or prompt, requested tips, replications of incorrect answers, and errors made. Any changes in the context of the classroom/school during the investigation period.
Student experience modeling	Are users satisfied with their experience?	Survey or questionnaire responses. Choices, actions, or performance in subsequent study units or courses.
Student profiling	What groups do users cluster into?	Student answers (correct, incorrect, partly correct), the time it takes to answer a prompt or query, the requested advice, repetition of wrong answers, and mistakes made.
Domain modeling	What is the right standard for separating subjects into modules, and how should they be sequenced?	Student responses (correct, incorrect, and partial) and results on modules with varying grain sizes compared to an external measurement. A taxonomy of the domain model. Associations between issues, expertise, and issues.
Learning component analysis and instructional principle analysis	Which components facilitate learning effectively? What principles of learning work well? How effective are the whole curricula?	Responses of students (correct, wrong, partly correct) and hierarchical output at various levels of detail instead of external steps. A taxonomy of the domain model. Structure of interaction between challenges and skills and issues.

(continued)

Table 2.2 (continued)

Application area	What can be achieved?	What data needed?
Trend analysis	What changes over time, and how?	Depending on what information is of interest, at least three data points will usually be needed longitudinally to discern a pattern. The collected data contains admissions, grades, completion, student source, and high school data over successive years.
Adaptation and personalization	What next steps should the consumer propose? How can the user experience for the next user be changed? How can the user experience be modified in real-time most often?	Varies according to the particular recommendation given. You can need to collect historical data about the user and relevant product or service to be recommended. Academic performance of the students.

2.4.1 Challenges

Although LA and EDM have beneficial advantages, the researchers and practitioners must also find their disadvantages and challenges (Vahdat et al., 2015). As LA and EDM come from different fields of study, data mining, and statistics, it is difficult for them to create relations with cognition, metacognition, and pedagogy, which are important sources for understanding learning processes. Researchers need to concentrate on learning sciences to facilitate successful pedagogy and improve learning design.

The high costs of software and techniques and the challenges of data interoperability and reliability are also factors listed in many studies. Educational data have been standardized, and movement improved, such as the IEEE standard for Learning Technology (IEEE SLT) and the Experience API. However, the present state of interoperability is not adequate to put together all data levels. Concerning reliability, the way users interpret activity data and make sense of the context through unorganized information poses many challenges. Moreover, ethical standards such as privacy and anonymity are becoming more difficult as data resources and significant resources have been increased.

2.4.2 Tools for EDM or LA

This section seeks to explain to the EDM or LA research practitioner the most commonly used, open, and efficient tools available (Slater, Joksimović, Kovanovic, Baker, & Gasevic, 2017). This discussion's direction will essentially follow the route that might be followed when exploring a study problem or evaluating it. The first big challenge is to turn raw and inchoate data streams into usable variables in

data mining and other data science fields. Data also come in types and formats that cannot be analyzed; the data need to be converted into a more meaningful format and meaningful variables. Further, data must also be cleaned to delete cases and values that are not just outliers but are deliberately incorrect (i.e., timestamps with impossible values, teacher check accounts for learning system results). Microsoft Excel, Google Sheets, and EDM Workbench are widely used to store, clean, and format data and data creation and feature engineering. We will also address Python and database queries play a vital role in this specific mission.

The next question an EDM or LA researchers will ask after data cleaning, transformation in a more workable format, and function engineering is: What experiments can be carried out, what models can be built, what relationships can be mapped and explored, and how can we validate our findings? We mention various resources for this task: RapidMiner, Weka, KEEL, KNIME, Orange, and SPSS. We also define many Python packages that are suitable for testing, analysis, and modeling.

The tools listed so far apply to a variety of data and analysis types. However, some types of data can be analyzed more efficiently with specialized tools adapted to these fields. We will address the most widely used methods for these kinds of specialized data in educational data mining, including information tracing algorithms, text mining, social network analysis, sequence mining, and process mining.

When a researcher has examined and has a validated, functional model, the work is also shared with other researchers, observers, and practitioners at schools and universities or developing curricula. A vital component of the delivery of research is legible and informative visualizations, and we will cover a range of resources in the final section of our debate that enables data scientists to create high quality and informative graphs, maps, models, networks, diagrams, and other types of information visualized. Tableau, d3js, and InfoVis are three visualization tools, and we explore visualization possibilities with a handful of standard Python packages.

The PSLC DataShop is a unique tool to combine data collection, development, analysis, and visualization. DataShop enables researchers to collect popular analyses with cognitive scientists and EDM researchers with one tool.

2.4.2.1 Manipulation of Data and Feature Engineering

Data sets must first be cleaned and processed in their raw state before data mining. Data miners generally work with messier data rather than statistics and psychometricians, although this issue usually exists with any data. Instead of meaningfully collected research or survey data, data mining companies frequently work with log data or learning management systems data reported on formats that cannot be analyzed immediately. Readers with experience dealing with such educational data know that it is unpredictable, often incomplete, often in many parts, and often indifferent or uncomfortable formats. A researcher might be involved in evaluating students, but its data must be systematically tracked behavior. Researchers may wish to use durations between actions to distinguish off-task students (Baker, 2007;

Cetintas, Si, Xin, & Hord, 2010). However, only raw timestamps can be obtained. In this case, a method known as feature engineering must be established to carry out the required analyses (Veeramachaneni, O'Reilly, & Adl, 2015). We present the following tools to clean, organize, and build data. We address the merits of and method, modify and restructure large datasets, and produce new and more useful variables from existing variables.

1. Microsoft Excel/Google Sheets. Microsoft Excel is the most straightforward resource for data scientists interested in analyzing or engineering data and makes it easily accessible when edited. It can be paired with a related web-based application, Google Sheets. These methods are not useful in producing variables in extensive data sets, approximately one million rows and above, but are ideal instruments for developing small features and prototyping new variables in a larger data set subset. One of the main reasons they help evaluate new data (variables) first stage and prototype data is that Excel and Sheets are excellent at clearly displaying data within a completely visual interface. This enables detecting structural or semantic concerns, such as irregular or incomplete values or duplicate entries. These tools also make it simple to create new features, apply these features easily to the entire layer, and visually test the features for proper functionality across various data. Student summaries, problem sets, and other aggregations can be easily determined by filtering and summing or pivot tables, with a feature for linking data sets or aggregation rates. Excel and sheets are not appropriate for all forms of feature design at the same time. Creating applications that require various database aggregations may entail sorting and resorting the data several times, making it difficult to document what has been done and making it easy to alter function semantics by chance. More significantly, the quantity of information that can be prepared, manipulated, and preserved is restricted by Excel and Sheets. Several Excel and Sheets common operators will further reduce efficiency.
2. EDM workbench. The EDM Workbench is an automated distillation and data labeling method (Rodrigo, Ryan, McLaren, Jayme, & Dy, 2012). Many EDM Workbench's automated feature distillation features fix specific Excel and Sheets deficiencies in specific data scientists-relevant tasks, such as generating complex sequential features, sampling and marking data, and aggregating data the subsets of student-tutor transactions, based on user-specified parameters (known as 'clips'). The EDM Workbench helps researchers build features using XML-based authoring and construct a collection of 26 features used in current literature and intelligent tutoring systems. Attributes include (but they are not limited to) the students time spent on the problem (in absolute and relative terms, for example, how much quicker or slower the student was in the same problem phase than other students) as well as the forms, number and amount of acts correctly, wrongly or helpfully for current ability during the final stages, for qualifications and the student. The EDM Workbench has the capability of generating text-replays in data labeling (Baker, Corbett, & Wagner, 2006), pretty print human behavior segments which are labeled in categories of conduct or other labels of interest by researchers or other domain experts. The EDM Workbench

supports the sampling, reliability monitoring, and synchronization of labels and distilled features.

3. Python and Jupyter Notebook. For data scientists with programming skills, a handful of languages are particularly suitable for data processing and functional engineering. For these reasons, Python is regarded by many as an incredibly useful language. In particular, in Python, it is easier to construct context-based or temporal features than in Excel or Google Sheets. Jupyter's Notebook – a server-client app to build and modify the Python code and rich text objects, such as graphs and tables inside a web browser, is another useful Python function. Jupyter Notebook is a tool for preserving order in-user behavior and its outcome, recording analyses performed, and interim results. Despite this benefit, however, data and features generated in Excel or Google Sheets can still be visually reviewed. In particular, data sets can difficult to identify missing details, same cases, or exceptional values, and it can take more time to validate engineered features, especially for inexperienced programmers. Python can also handle other types of uncommon or unique data formats, such as the JavaScript Object Notation (JSON) files provided by various MOOC and online learning platforms. Although Python is computationally stronger than previously covered spreadsheet tools, its capabilities in these areas are not limitless. Although Python can handle larger datasets than previous tools, it remains subject to size limitations, which for these researchers are slower for the range of about ten million rows of data. It should be noted that certain types of programs (for example, those with nested loops) are considerably slower to use the notebook than in standard Python.
4. SQL. SQL (Structured Query Language) is used to organize databases. SQL queries can be a powerful way to retrieve the desired information exactly and sometimes combine ("join") across multiple database tables. Many simple filtering tasks such as selecting a specific student subset or extracting data from a certain date range are significantly quicker than in any of the tools listed above in the database languages such as SQL. SQL can, however, be a very clunky language for constructing complex functions in the system engineering process. In conjunction with other tools listed above, SQL can work effectively: SQL excels at large size sorting and filtering tasks, which in Excel or Python are very slow, while the tools perform better on the kind of small datasets that can be generated by SQL.

2.4.2.2 Algorithmic Analysis

Once features are created, results and ground truth variables have been identified, data collected and organized adequately for analysis, the next step is to initiate data analysis and modeling and validate the resulting models. The tools mentioned in the following section provide a broad range of algorithms and frameworks for modeling and predicting educational data processes and relationships.

1. RapidMiner. RapidMiner is a program to analyze and construct models for data mining. It has restricted flexibility to develop new features from existing features

(for example, multiple interactions) and pick features (based on user interconnections and results measures). RapidMiner, however, has an incredibly wide variety of classification and regression algorithms and clustering algorithms, association rule mining, and other applications. Other algorithms may typically be composed of RapidMiner operators, e.g., for set selection or model bagging. However, support for resampling processes such as bootstrapping is more restricted than in other data mining packages. The graphical programming language of RapidMiner is comparatively more powerful than most other data mining software with comprehensive user specification functions. For example, RapidMiner can be used to perform multi-level cross-validation with the BatchCrossValidation Operator. This support can benefit generalizability analysis and benefit most other data mining packages over the graphics languages. RapidMiner also has a wide range of metrics for model evaluations that can show views such as Receiver-Operating Curves to help users determine a model's fitness. Models can either be generated in terms of the current math models or XML files, running the model using RapidMiner code on new data. The Application Program Interface (API), which can be incorporated into programs written in Java or Python, can perform various tasks that are not possible with RapidMiner's graphical programming language. RapidMiner contains all Weka algorithms discussed below. RapidMiner also features crowd-sourced algorithms and parameter suggestions. RapidMiner has a wide variety of tutorials to learn the graphic programming language easily.

2. WEKA. Waikato Environment for Knowledge Analysis (Weka) is a free, open-source software package that assembles various data mining and model-building algorithms. It does not allow the creation of new features but allows automatic selection of features. Weka has a wide range of algorithms for classification, clustering, and association mining that can be used in isolation or combination with methods such as bagging, boosting, and stacking. Users can invoke command-line data mining algorithms, a GUI, or a Javan API. Command-line Interface and APIs are more comfortable than the Software that does not allow users to access any advanced functions. Weka can create the models that it produces either in terms of the actual mathematical model or in PMML files used to run the model on new data with the Weka scoring plugin to run the model.
3. SPSS. Like Excel, SPSS is not only familiar with the world of data science. It provides various statistical measures, regression models, correlations, and factor analyses, mainly a statistical package. IBM SPSS Modeler Premium complements SPSS, a relatively new data mining software that combines previous analytics and text mining packages. SPSS Modeler can specifically build new features from existing features, filter data, select features, and reduce function space. The tools for transforming the data, selecting features, and the space available in data mining packages with fewer selection methods. There is also an option in the product range to use the target class, which is not included in many other packages. Although SPSS is a comprehensive statistical analysis tool, modeling support is slightly worse than the other tools in this section. SPSS is less versatile, more comfortable to configure, and less documented than other

devices. Procedures that are seen as important by researchers in educational data mining, such as cross-validation, are also lacking in comparison with more data mining-focused tools.

4. KNIME. KNIME is a data clearing and analysis package similar to RapidMiner and Weka ("Naim", KoNstanz Information MinEr, www.knime.org), formerly Hades. It provides many of the same functions as these tools and includes all Weka's algorithms, including RapidMiner. It also provides several advanced algorithms, such as sentiment analysis and social network analysis. KNIME's capacity to incorporate data from many sources (e.g., a .csv of engineered features, word document for answers, and a student demographic database) within the same study is particularly strong. KNIME also provides extensions to Interface with R, Python, Java, and SQL.
5. Orange. Orange is a program for data visualization and analysis. Although the Interface is considerably less algorithmic and more comfortable to understand than RapidMiner, Weka, or KNIME, color-coded widgets differentiate between data entry and cleaning, visualization, regression, and clustering. It provides a wide range of common algorithms, including k-nearest neighbors, random forests, naïve Bayes classification, and supporting vector machines. Orange also has customizable display modules with reasonable documentation for presenting model results. Orange is, however, somewhat limited to Excel in the amount of data it can process. Orange can be better suited as a tool for smaller projects or more advanced researchers based on its easily understood Interface and menu layout.
6. KEEL. KEEL is a tool for data mining used by many EDM researchers. In contrast to some of the tools mentioned above that seek to survey various methods in general, KEEL supports some algorithms and tasks extensively but restricts the support of other algorithms and tasks. For example, KEEL supports discretization algorithms very extensively but has limited support for other techniques for creating new features from existing features. It provides outstanding feature selection support with a wider variety of algorithms than any other method. It also encourages the imputation of missing data and provides substantial support for resampling data. KEEL has a broad collection of classification and regression algorithms for modeling, emphasizing evolutionary algorithms. Its support is more limited than other packages for other types of data mining algorithms like clustering and factor analysis. Association rule mining support is decent but not as comprehensive as some other packages. Although there are help features and a user manual, KEEL has relatively less support for new users than most other data mining packages.
7. Spark. MLLib Spark is a framework for large-scale data processing in a distributed fashion across multiple computer processors. Spark can connect via an API to several programming languages, including Java, Python, and SQL, for distributed processing. The MLLib machine learning platform from Spark offers several popular machine learning and data mining algorithms for implementation. While the functionality of MLLib is still somewhat limited and a purely programmatic tool, its distributed nature makes it a fast and efficient choice.

2.4.2.3 Visualizations

Beyond just mining data, there is a growing awareness that both analysts and practitioners can support useful visualization methods with data meaning (Baker & Siemens, 2014; Duval, 2011; Tervakari, Silius, Koro, Paukkeri, & Pirttilä, 2014; Verbert, Duval, Klerkx, Govaerts, & Santos, 2013). In the next section, we discuss specialized tools for social network analysis applications that can provide sophisticated viewing (e.g., Gephi, SNAPP). Specifically, we want to introduce some general tools and methods for visual analysis, which enable students and instructors to build interactive visual interfaces to acquire data knowledge and insight.

1. Tableau. Tableau presents an interactive data analysis and visualization, product family. Although support for enterprise intelligence is the main focus of the Tableau toolkit, it is commonly used in educational settings to analyze student data, provide actionable insights, increase teaching practices, and streamline educational reporting. Tableau's main advantage is that it needs no programming knowledge to analyze large numbers of data from different sources and make a range of visualizations easily accessible to a broader community. Tableau offers functionality for connecting or importing data from several standardized data storage formats (e.g., databases, warehouses of data, log data). Tableau also provides functionality to build rich and interactive dashboards that allow end-users to display dynamic real-time visualizations. However, Tableau's functionality is limited to this; it does not support predictive analytics or relational data extraction. Also, Tableau is not extendable as a commercial tool and does not support integration with other software platforms.
2. D3js. D3.js (Data-Driven Documents) is a JavaScript library that enables data-driven document manipulation, enabling researchers and practitioners to create complex, interactive visualizations that need the data management and are designed for the modern web browsers. D3.js offers several advantages; it offers considerable flexibility in building a range of data visualization types, requires no installation, supports code reuse, and is open and free. However, the broader adoption of educational research purposes is challenged. D3.js requires extensive knowledge of programming and has problems with compatibility and certain performance limitations for larger data sets. Finally, it provides no way to hide data from visualization users, requiring preprocessing information to ensure data protection and security. In addition to D3.js, many other programmatic data visualization tools offer various visual presentations and interactive dashboards. Chart.js, Raw, JavaScript InfoVis Toolkit, jpGraph, and Google Visualization API are among the most common tools. These tools offer similar to D3.js but are less frequently used by EDM and LA researchers.

2.4.2.4 Specialized EDM and LA Applications

We addressed general-purpose tools for EDM modeling and analysis in the previous section. However, specific data types and particular research purposes also require more complex algorithms not included in these tools. Researchers and practitioners typically use more advanced tools for these cases.

Tools for Bayesian Knowledge Tracing

Bayesian Knowledge Tracing (Corbett & Anderson, 1995) is a standard method for latent knowledge evaluation in which the knowledge of a student is assessed in online learning. This is different from the type of educational assessment typical in tests in that information changes through online education. Bayesian Knowledge Tracing is a Hidden Markov Model and, at the same time, is a simple Bayesian Network (Reye, 2004) that predicts whether or not a student has mastered a special knowledge within an intelligent tutoring system or similar software. BKT models are generally suitable for two algorithms: brute force grid search or expectation maximization (EM). The two algorithms differ in predictive analysis comparably. BKT's tools are BKT-BF, BNT-SM (also Matlab needs to run) and, hmmsclbl.

Text Mining

Text mining is an increasingly growing field of data mining, and there is a broad range of tagging, processing, and recognition systems, applications, and APIs for text data. Text analysis software may analyze text parts of speech, phrase form, and the context of semantical terms. Also, some tools may define symbolic relationships between various words and phrases. Several tools for text mining and corpus analysis are available more than any other set of tools mentioned so far. This is mainly for two reasons: text mining's complexity and the English language are complicated. Developing a full suite of resources for various text bodies and media types is an exceedingly difficult task. The variety of lexical analytical methods represents the nature and sophistication of the language to be analyzed and evaluated. The second explanation is that various linguistic groups often have varying approaches to text definition and analysis and the wide variety of tools available for text mining is a result of many different areas of researchers developing their tools. We consider that the following methods constitute tools that cross the many dimensions of textual processing and analysis and are appropriate for general approaches to text mining and the study of particular structures within text and discourse.

1. LIWC. The Linguistic Inquiry and Word Count (LIWC) tool (Tausczik & Pennebaker, 2010) is a computerized graphical and user-friendly tool for studying vocabulary used to calculate the latent characteristics of a text. LIWC offers more than 80 metrics for various psychological vocabulary categories (e.g., cognitive words, affective words, functional words, analytical words).
2. WMatrix. WMatrix is an online graphical tool for word frequency analysis and visualization for text corpora. While it can be used to complete the study, it is most useful in the function engineering process for extracting linguistic characteristics such as word n-grams and multi-word sentences, such as idioms and similar, part-of-speech tags and semantic word categories. It also allows the text corpora to be visualized in word clouds and simultaneously offers an interface to compare multiple text corpora.
3. Coh-Metrix. The Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004; Graesser, McNamara, & Kulikowich, 2011) a common tool for text analysis,

offers over 100 text measurements in 11 categories. CohMetrix provides a more detailed interpretation and examination of text characteristics and relationships in the data than the WMatrix. Although WMatrix semantically tags terms and multi-word units, CohMetrix has many tags for evaluating deep cohesion text such as narrative steps and/or referential cohesion. With these increases in the research's profound significance, larger data sets are required – CohMetrix essentially appears to need a larger body of text than semantic taggers.

4. Latent semantic analysis (LSA). Latent semantic analysis is another tool frequently used to extract subjects from text corpora (Landauer, Foltz, & Laham, 1998). While LDA and similar probabilistic methods use word co-occurrence to estimate the words which constitute a field, LSA uses a linear algebra technique of matrix decomposition to find words representing various themes. It can also be used by comparing their vectors in the topic space to calculate the semantic similarity of two documents or sections of documents. LSA has been implemented in many programming languages, with one of the most common LSA implementations being a java-based text mining library and the lsa R package (Alves dos Santos & Favero, 2015).
5. NLP toolkits (***Stanford CoreNLP, Python NLTK, Apache OpenNLP***). Text mining systems usually require natural language text analysis, the toolkits for natural language processing (NLP) form a significant part of the text mining toolbox. These methods are usually used in the analytical preprocessing step, for example, (a) split paragraphs into individual phrases, utterances, or words, (b) extraction of syntactic dependency between words (c) assign parts-of-speech categories (word grammatical class) to each word, (d) reduction of derived words (i.e., stemming and lemmatization), (e) named-entity extraction, and (f) co-reference resolution. Several NLP toolkits provide common programming languages (e.g., Java or Python) with programmable APIs. One famous example is the Apache OpenNLP toolkit, which supported most basic NLP activities. Similarly, Python NLTK (Bird, 2006) is an NLP library with somewhat similar python programming language capabilities. The NLP toolkit offers a Java API and standalone GUI on the command line and a collection of wrappers for other programming languages, including C #, Python, R, Ruby, Scala, and JavaScript.
6. ConceptNet. ConceptNet. One of the key reasons why natural language comprehension is a complicated issue is that each statement depends on the listener's meaning and background knowledge. The method adopted by ConceptNet (Liu & Singh, 2004) is to create a vast graph of "common-sense" knowledge and then to be used to understand and process natural text. ConceptNet can use a broad knowledge base to categorize textual articles, extract topical information from corpora, sentiment analysis, and text summarization.
7. TAGME. TAGME is a text annotation tool explicitly developed for short, unstructured, or semi-structured text segments, such as text collected from snippets, tweets, and news feeds (Ferragina & Scaiella, 2010). The text annotation method defines and annotates a series of words with appropriate links to Wikipedia sites. In other words, TAGME assigns each sequence in the analyzed text to a Wikipedia concept. An experimental TAGME (Ferragina & Scaiella,

2010) evaluation demonstrated better performance over short text segments and similar accuracy/recall outcomes for longer text than other solutions. The tool offers an API for text processing and integration with other applications on-the-fly.

8. Apache Stanbol. Apache Stanbol is an open-source text analysis software tool. It is intended primarily to combine semantic technology with existing content management systems and extract text and functions. Like TAGME, it binds keywords extracted from text to concepts from Wikipedia. Apache Stanbol is easy to configure and run in a small set of instances. However, the tool also enables a domain-specific ontology to be integrated into the annotation process. This is particularly useful when dealing with local ideas that are unique to a particular educational setting. Finally, Apache Stanbol supports multi-language text annotation. Multiple content management systems have been built into the application.

Social network analysis

Social network analysis attempts to explain the interactions and relationships most frequently represented as nodes and edge diagrams between individuals and/or societies. SNA is widely used to evaluate interactive social networks, such as social media or student engagement in MOOCs or online courses.

1. Gephi. Gephi is a common and widely used interactive tool to analyze and visualize various social networks. Gephi is commonly used in learning analytics research and supports social networks defined in a broad range of data input formats, which are both direct and undirected. It has various charts for a simple view of social networks and provides the ability, often used as a tool for exploratory research, to color nodes and edges, based on the characteristics of their location in the network. The tool also offers a Java API for manipulating social network graphs, measuring multiple measures (for example, density, average trajectory, and betweenness centrality), and execution algorithms widely used in social network analyses (for example, graph clustering and giant interconnected component extraction). It is licensed under the GPL license and is available on Microsoft platforms such as Windows, Linux, and Mac OSX.
2. EgoNet. EgoNet is a free social network analysis tool that focuses on analyzing egocentric networks that are, in general, social networks developed from particular network actors' viewpoint, usually using survey instruments. Via EgoNet, a researcher specifies the number of network members and gives them a small survey of their relationships with other network members. As participants provide information from their viewpoints about the network structure, EgoNet visualizes the entire network structure and offers various analytical instruments better to understand the network's overall nature and opportunities to ask a network member for further questions.
3. NodeXL. NodeXL, Network Overview Discovery Exploration for Excel, a Microsoft Excel extension that makes it easy to display network data from various data input formats in Microsoft Excel. It is also used for the estimation of the primary network properties (e.g., radius, diameter, density), node properties

(e.g., degree centrality, betweenness centrality, eigenvector centrality), and other network analytics approaches (e.g., cluster analysis for community mining). There are currently two versions, NodeXL Basic and NodeXL Pro. Beyond the primary social network analysis support, NodeXL Pro provides data aggregation functionalities from different social media sites (e.g., Twitter, YouTube, Flickr) and social media text and sentiment analysis.

4. Pajek. Pajek is a free desktop tool for complex analysis of many large networks (thousands and hundreds of thousands of nodes), including social media networks. Pajek is widely used for social network analysis and LA research in academia for network partitioning, group identification, large-scale network visualization, and information flow analysis. There is also a Pajek-XXL version, a specifically built Pajek version for the efficient operation of vast networks (with millions of nodes).
5. NetMiner. NetMiner is a popular graphic tool for the study and visualization of networks. It supports network data import in various formats, network views, and the measurement of standard graphic and node-based statistics, similar to Gephi and NodeXL. NetMiner has a built-in data mining module supporting various data mining tasks such as classification, clustering, reduction, and recommendation) which is also suited for advanced NetMiner network analysis. It also has a Python integrated scripting engine for more complex and personalized analytical forms. It also supports a scripting interface and the graphical user interface, making it ideal for module integration in other software systems. It also facilitates 3D network viewing and network exploration video recording (e.g., for inclusion). Currently, NetMiner is available on Microsoft Windows OS only.
6. Cytoscape. Cytoscape is another open-source framework for the visualization of molecular interaction networks, which is now a fully-functional package for studying different network types, including social networks. Cytoscape consists of a core distribution that uses several user-contributing modules to analyze and view basic network capabilities. Cytoscape is developed and can be used in various operating systems on the Java platform.
7. SoNIA. SoNIA is an open-source framework for longitudinal network data analysis. For the longitudinal network data, in addition to information on relationships (i.e., edges) between network members (i.e., nodes), the time of these relationships occurred, or the order in which they formed is also available. SoNIA can display network change over time, allowing various network architecture algorithms to be defined in multiple timeframes to better visualize network structure changes. The effect is a good, 'smooth' animation of structural changes exported in QuickTime video format over time. SoNIA is developed by Stanford University and can be used in all critical operating systems in the Java programming language.
8. SocNetV. Social Networks Visualizer (SocNetV) is an open-source tool for analyzing and manipulating social networks. This facilitates the loading of data of different network formats, computing traditional graph and node characteristics, and versatile network data visualization (e.g., filtering, coloring, and resiz-

ing nodes based on their characteristics). One of SocNetV's exciting and unique features is the embedded web crawler, which automatically extracts a link structure from a set of HTML documents. It is licensed under GPL and available on Microsoft Windows, Linux, and Mac OSX.

9. NetworkX. NetworkX is a Python Open Source software library for complex network functions, architectures, and dynamics. It is commonly used in academia and has a wide range of advanced features in networked data, including the reduction of graphs through block modeling, group clustering, group detection, link prediction (finding missing links, for example, missing Facebook connections between two friends), analysis of network triads, and others.
10. ***R packages: statnet (network, sna, ergm) and igraph***. In addition to the graphical tools for analyzing social networks, other social network research packages are available in the R programming language. The network package is used to construct and change network objects, extract basic network metrics, and visualize network graphs. Often used together with the network package, the sna package provides a set of features commonly required for social network analysis, including network and node metric measurement, block-modeling graph reductions, network regression, network visualization, and others. The igraph software is another package that is mostly used for social network analysis. It is a library written in C programming language with additional language bindings in R and Python's languages. It can be used to construct and change social networks from a broad range of input formats (e.g., Pajek, Gephi, GraphML, edge list, an adjacency matrix), measurement of network and node properties, visualization of graphs, and various network analysis, including group identification, graph clusters, block modeling, unified blocks measurement, and others. The stat network package focuses on statistical network simulations using exponential random graph models, latent space, and latent cluster models. Another vital package for social network analysis is the stat network package. The statnet package contains network model estimation methods, network model validation, model-based network simulations, and network visualization. It also contains and uses several of the other packages, such as network, sna, and ergm.
11. SNAPP. The Social networks Adapting Pedagogical Practice is a bookmarklet, developed by Bakharia and Dawson (2011), to evaluate student social networks created under popular learning management systems-LMSs (e.g., blackboard, Desire2learn, and moodle)-which is designed to be a bookmark button for the browser bookmark bar. SNAPP extracts a student social network from HTML pages of LMS discussions (formed through student posting and response interactions). The data can then be exported or displayed with a range of graph layout algorithms via SNAPP, or further analysis can be done with other above listed SNA tools. SNAPP can also investigate student social networking trends over time, evaluate extremely active/inactive users, find systemic gaps, and compare analysis for multiple discussion forums.

Process and sequence mining

In addition to more conventional approaches to analyzing education data, such as forecasting learning outcomes or continuing learning, research engaged in monitoring learning sequences to understand learning strategies and processes (Bogarín, Romero, Cerezo, & Sánchez-Santillán, 2014). For this form of application, a distinctive collection of resources has evolved. We will present in this section the process and sequence mining tools ProM and TraMineR that are widely used to support EDM and LA research. These tools are generally used for analyzes, although they often allow some preprocessing levels of data.

1. ProM. ProM is an autonomous, scalable, and open-source Java-based framework supporting several process mining techniques (Van Der Aalst et al., 2009). The new version, ProM 6, supports process mining in a distributed environment or batch processing. ProM supports chaining many process mining algorithms to explicitly define the predicted inputs and outputs for each of the implementations supported. Also, new plugins can be introduced at runtime to allow fast integration into the analysis process. Finally, ProM allows quick integration without programming with current information systems.
2. TraMineR. TraMineR is a free, open-source R package that supports state or event sequences mining and visualization. For analysis and visualizing status sequence data, TraMineR has a variety of primary features: (i) processing various state sequence formats and converting them to and from different representations; (ii) defining longitudinal (i.e., length, complexity, time in each state) and other aggregate sequence characteristics; (iii) accessing a wide range of plotting capabilities (i.e., frequency or density plots, index plot), and (iv) a broad set of metrics for evaluating distances between sequences.

PSLC DataShop

A multifunctional platform is PSLC DataShop (Koedinger et al., 2010). The PSLC DataShop comprises several data sets that can be downloaded and analyzed, and resources to enable exploratory analysis and models. DataShop has a domain structure (knowledge component) comparison functionality on a dataset like q-matrices (Tatsuoka, 1983). It also can visualize student performance over time in terms of correctness, hint use, latent knowledge, response times, and other variables of interest. It also includes visualizations of student performance regularly. The PSLC DataShop is a free, though not open-source, web-based application.

2.5 Conclusion

Two communities have evolved in recent years around the concept of using large-scale educational data to change education research practice. As this field evolves from relatively small and obscure conferences to a subject known across educa-

tional research and affects schools worldwide, the above approaches are available to achieve several goals. Every year, researchers and practitioners use these methods to analyze new constructions and address new research questions, making the application of these methods more understood.

The methods and applications of educational data mining and learning analytics are discussed in this chapter. These approaches can be useful for researchers, teachers, administrators, and ultimately students by evaluating students' attitudes and results. We also discussed variations and similarities between these two methods.

2.6 Review Questions

Reflect on the concepts of this chapter guided by the following questions.

1. Define Educational Data Mining. Illustrate the similarities and differences of Educational Data Mining with Learning Analytics.
2. Trace out the similarities between Data Mining, Educational Data Mining, and Learning Analytics.
3. List the key events that occurred in the formation of the EDM community.
4. Define Educational Data Mining. Identify the goals of Educational Data Mining.
5. Describe the phases involved in Educational Data Mining.
6. Explain the methods of Educational Data Mining with suitable examples.
7. Explain the standard methods used both for Educational Data Mining and Learning Analytics.
8. Describe the broad categories of typical applications of both Educational Data Mining and Learning Analytics.
9. Which are the primary tools for Educational Data Mining and/or Learning Analytics?
10. How Educational Data Mining brings real change in the education system?
11. What are the current trends in educational data mining?

References

Alani, M. M., Tawfik, H., Saeed, M., & Anya, O. (2018). *Applications of big data analytics: Trends, issues, and challenges* (pp. 1–214). https://doi.org/10.1007/978-3-319-76472-6.

Alves dos Santos, J. C., & Favero, E. L. (2015). Practical use of a latent semantic analysis (LSA) model for automatic evaluation of written answers. *Journal of the Brazilian Computer Society, 21*(1), 1–8. https://doi.org/10.1186/s13173-015-0039-7

Baker, R., & Siemens, G. (2014). Educational data mining and learning analytics. *Learning Analytics: From Research to Practice*, 61–75. https://doi.org/10.1007/978-1-4614-3305-7_4.

Baker, R. S. (2015). *Big data and education* (2nd ed.). New York: Teachers College, Columbia University.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics: From research to practice*. https://doi.org/10.1007/978-1-4614-3305-7_4.

Baker, R. S. J. (2010). Data mining for education. *International Encyclopedia of Education (3rd Edition)*, *7*, 112–118. https://doi.org/10.4018/978-1-59140-557-3.

Baker, R. S. J. D. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. *Conference on Human Factors in Computing Systems – Proceedings*, 1059–1068. https://doi.org/10.1145/1240624.1240785.

Baker, R. S. J. D, Corbett, A. T., & Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. In *Proceedings of the educational data mining workshop at the 8th international conference on intelligent tutoring systems*, *2002* (pp. 29–36).

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–16. https://doi.org/http://doi.ieeecomputersociety.org/10.1109/ASE.2003.1240314.

Bakharia, A., & Dawson, S. (2011). SNAPP: A bird's-eye view of temporal participant interaction. *ACM International Conference Proceeding Series*, 168–173. https://doi.org/10.1145/2090116.2090144.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies, 23*(1), 537–553. https://doi.org/10.1007/s10639-017-9616-z

Bienkowski, M., Feng, M., & Means, B. (2014). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. In *Educational improvement through data mining and analytics* (pp. 1–60). U.S. Department of Education, Office of Educational Technology: Washington, D.C., USA.

Bird, S. (2006). NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL on interactive presentation sessions* (pp. 69–72). https://doi.org/10.1017/9781108642408.013.

Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. *ACM International Conference Proceeding Series*, 11–15. https://doi.org/10.1145/2567574.2567604.

Bousbia, N., & Belamri, I. (2014). Which contribution does EDM provide to computer-based learning environments? *Studies in Computational Intelligence, 524*, 3–28. https://doi.org/10.1007/978-3-319-02738-8

Buckingham Shum, S. (2012). *Learning analytics: Policy briefing*. pp. 1–12. http://iite.unesco.org/files/policy_briefs/pdf/en/learning_analytics.pdf

Calvet Liñán, L., & Juan Pérez, Á. A. (2015). Educational data mining and learning analytics: Differences, similarities, and time evolution. *RUSC. Universities and Knowledge Society Journal, 12*(3), 98. https://doi.org/10.7238/rusc.v12i3.2515

Cetintas, S., Si, L., Xin, Y. P., & Hord, C. (2010). Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques. *IEEE Transactions on Learning Technologies, 3*(3), 228–236. https://doi.org/10.1109/TLT.2009.44

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. In *User modeling and user-adapted interaction* (Vol. 4, Issue 4, pp. 253–278). http://www.springerlink.com/index/M50H664760426738.pdf%5Cnpapers3://publication/livfe/id/110900

Duval, E. (2011). Attention please! Learning analytics for visualization and recommendation. *ACM International Conference Proceeding Series*, 9–17. https://doi.org/10.1145/2090116.2090118.

Ferragina, P., & Scaiella, U. (2010). TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities). *International Conference on Information and Knowledge Management, Proceedings*, April 2015, 1625–1628. https://doi.org/10.1145/1871437.1871689.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers, 36*(2), 193–202. https://doi.org/10.3758/BF03195564

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Educational Technology and Society, 15*(3), 42–57.

He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior, 29*(1), 90–102. https://doi.org/10.1016/j.chb.2012.07.020

Johnson, L., Smith, R., Willis, H., Levine, A., & Haywood, K. (2011). *The Horizon report*.

Koedinger, K., Baker, R., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). *A data repository for the EDM community*. April 2016, 43–55. https://doi.org/10.1201/b10274-6.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent SEmantic Analysis. *Behavior Research Methods, 25*(2–3), 259–284. https://doi.org/10.3758/BRM.41.3.944

Lang, C. (2017). *Handbook of learning analytics*. https://doi.org/10.18608/hla17.

Liu, H., & Singh, P. (2004). ConceptNet – A practical commonsense reasoning tool-kit. *BT Technology Journal, 22*(4), 211–226. https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d

NMC Horizon Report > 2012 Higher Education Edition. (2012).

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

Reye, J. (2004). *Student modelling based on belief networks to cite this version: HAL Id: hal-00197306*. pp. 63–96.

Rodrigo, M. M. T., Ryan, R. S. J., McLaren, B. M., Jayme, A., & Dy, T. T. (2012). Development of a workbench to address the educational data mining bottleneck. *Proceedings of the 5th International conference on educational data mining, EDM 2012*, 152–155.

Romero, C., & Ventura, S. (2010). 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. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12–27. https://doi.org/10.1002/widm.1075

Romero, C., & Ventura, S. (2017). Educational data science in massive open online courses. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(1). https://doi.org/10.1002/widm.1187.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), 1–21. https://doi.org/10.1002/widm.1355

Schroeder, U., Thüs, H., & Technologies, I. L. (2012). A reference model for learning analytics Mohamed Amine Chatti*, Anna Lea Dyckhoff. *Int. J. Technology Enhanced Learning, 4*(CiL), 318–331.

Siemens, G. (2012). Learning analytics: Envisioning a research discipline and a domain of practice. *ACM International Conference Proceeding Series*, May, 4–8. https://doi.org/10.1145/2330601.2330605.

Siemens, G., & Baker, R. S. J. D. (2012). Learning analytics and educational data mining: Towards communication and collaboration. *ACM International Conference Proceeding Series*, April 2012, 252–254. https://doi.org/10.1145/2330601.2330661.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics, 42*(1), 85–106. https://doi.org/10.3102/1076998616666808

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tervakari, A. M., Silius, K., Koro, J., Paukkeri, J., & Pirttilä, O. (2014). Usefulness of information visualizations based on educational data. *IEEE Global Engineering Education Conference, EDUCON*, April, 142–151. https://doi.org/10.1109/EDUCON.2014.6826081.

Vahdat, M., Ghio, A., Oneto, L., Anguita, D., Funk, M., & Rauterberg, M. (2015). Advances in learning analytics and educational data mining. *23rd European symposium on artificial neural networks, computational intelligence and machine learning, ESANN 2015 – Proceedings*, April, 297–306.

Van Der Aalst, W. M. P., Van Dongen, B. F., Günther, C., Rozinat, A., Verbeek, H. M. W., & Weijters, A. J. M. M. (2009). Prom: The process mining toolkit. *CEUR Workshop Proceedings*, *489*(May 2014).

Veeramachaneni, K., O'Reilly, U. M., & Adl, K. (2015). Feature factory: Crowd-sourced feature discovery. *L@S 2015 – 2nd ACM Conference on Learning at Scale*, 373–376. https://doi.org/10.1145/2724660.2728696.

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist, 57*(10), 1500–1509. https://doi.org/10.1177/0002764213479363

Verbert, K., Manouselis, N., Drachsler, H., & Duval, E. (2012). Dataset-driven research to support learning and knowledge analytics. *Educational Technology and Society, 15*(3), 133–148.

Wikipedia.(2019).*Educationaldatamining*.Wikipedia.https://doi.org/10.4324/9781351044677-18.