Lecture Notes in Educational Technology

J. Michael Spector
Vivekanandan Kumar · Alfred Essa
Yueh-Min Huang · Rob Koper
Richard A. W. Tortorella · Ting-Wen Chang
Yanyan Li · Zhizhen Zhang *Editors*

Frontiers of Cyberlearning

Emerging Technologies for Teaching and Learning

Springer

Lecture Notes in Educational Technology

The series *Lecture Notes in Educational Technology* (LNET), has established itself as a medium for the publication of new developments in the research and practice of educational policy, pedagogy, learning science, learning environment, learning resources etc. in information and knowledge age, – quickly, informally, and at a high level.

More information about this series at http://www.springer.com/series/11777

J. Michael Spector · Vivekanandan Kumar
Alfred Essa · Yueh-Min Huang
Rob Koper · Richard A. W. Tortorella
Ting-Wen Chang · Yanyan Li
Zhizhen Zhang
Editors

Frontiers of Cyberlearning

Emerging Technologies for Teaching and Learning

Springer

Editors

J. Michael Spector
Department of Learning Technologies
University of North Texas
Denton, TX
USA

Vivekanandan Kumar
School of Computing and Information System
Athabasca University
Athabasca
Canada

Alfred Essa
Analytics and R&D
McGraw-Hill Education
Boston, MA
USA

Yueh-Min Huang
Department of Engineering Science
National Cheng Kung University
Tainan
Taiwan

Rob Koper
Open University in the Netherlands
Heerlen
The Netherlands

Richard A. W. Tortorella
School of Computing
University of Eastern Finland
Joensuu
Finland

Ting-Wen Chang
Smart Learning Institute
Beijing Normal University
Beijing
China

Yanyan Li
School of Educational Technology
Beijing Normal University
Beijing
China

Zhizhen Zhang
School of Educational Technology
Beijing Normal University
Beijing
China

ISSN 2196-4963 ISSN 2196-4971 (electronic)
Lecture Notes in Educational Technology
ISBN 978-981-13-0649-5 ISBN 978-981-13-0650-1 (eBook)
https://doi.org/10.1007/978-981-13-0650-1

Library of Congress Control Number: 2018943385

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

Learning Any Time, Anywhere: Big Educational Data from Smart Devices

Mark A. Riedesel and Patrick Charles

Abstract For many people, especially young people, a smartphone is a constant companion. Mobile apps which allow individuals to use a smart device to enhance their learning have the potential to be very useful for mastering basic educational material. In order to evaluate and enhance the effectiveness of such applications when deployed at large scale, an infrastructure designed specifically for the collection of educational analytics data from such mobile apps is required. We detail here a set of applications and their associated infrastructure which was developed to allow students in courses using digital textbooks to enhance their knowledge of the basic course content anywhere and anytime by using their smart device to do spaced practice of the knowledge components of a course. The power of current smart devices allows the entire application, including content and adaptive algorithm to be hosted and run locally on the user's smart device, so it functions fully even when no network connection is available. The infrastructure for the collection and analysis of the educational analytics data is entirely cloud-based, using AWS S3 for data collection and storage, and the Apache Spark parallel computing framework for data analysis. Thus, the entire system requires only laptop computers for the mobile developers who create the applications and this is also sufficient for the learning scientists who analyze the data. Both the data collection system and the data analysis system can scale to handle the data from many millions of users with no modification to their architecture. Similar architectures are now used for the Internet of Things (IOT) but have not yet been widely used for educational applications. These applications have currently been deployed to thousands of users' smart devices and analytics data is being received from these users' smart devices from a wide range of locations on several continents. In our highly connected world, this type of application will become much more common. We describe here the type of infrastructure, security, and analytic methods needed to use these apps to advance learning and learning science.

M. A. Riedesel (✉) · P. Charles
McGraw-Hill Education, 281 Summer Street, Boston, MA 02210, USA
e-mail: mark.riedesel@gmail.com

P. Charles
e-mail: patrick.charles@mheducation.com

J. M. Spector et al. (eds.), *Frontiers of Cyberlearning*, Lecture Notes in Educational Technology, https://doi.org/10.1007/978-981-13-0650-1_1

Keywords Spaced practice · Adaptive flashcards · Mobile learning Messaging systems · Parallel computing · Cloud computing

1 Introduction

For many subjects, memorizing basic facts is an important first step in learning and mastering content. In learning a foreign language, for example, basic vocabulary must be memorized as an essential part of learning syntax and grammar. In medicine, it is still considered essential for doctors, nurses, medical assistants, and EMTs to have basic knowledge of anatomy and physiology, pharmacology, and medical terminology memorized for immediate recall in critical situations. Even for subjects which involve higher levels of cognition, retention of basic knowledge is commonly the foundation upon which higher levels of abstraction are built (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956).

Applications designed for long-term memorization are often conceptually based on models for human memory which grew out of early work on how memories decay with time but can be reinforced by repetition spaced in time (Ebbinghaus, 1885). Further research, particularly in the past 20 years, has shown that three related methods can help optimize the memorization of material: spaced practice (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008), active retrieval (Roediger & Karpicke, 2006; Karpicke & Roediger, 2008), and interleaving (Brown, Roedinger, & McDaniel, 2014).

Spaced practice is a method where material one wants to learn is repeated on a schedule designed to reinforce decaying memories. Ideally, an item (an atomic piece of knowledge) would be repeated just before it would otherwise be forgotten. Using very specific models for the forgetting process allows mathematical optimization methods to be employed to produce optimal schedules for spaced practice (Pavlik & Anderson, 2005; Pavlik & Anderson, 2008; Mozer & Lindsey, 2016; Settles & Meeder, 2016).

Active retrieval is the principle that requiring a learner to recall material in response to a challenge is more effective than re-reading material. This is true even when there is some active involvement when reading, such as highlighting the most important parts of a passage. Thus, requiring a learner to respond to a question is significantly more effective than having the learner re-read a text containing the same basic piece of knowledge. Questions can be in several modes such as multiple-choice single-answer, multiple-choice multi-answer, fill-in-the-blank, or matching.

Interleaving is a method where questions on different topics or subtopics are intermixed. It has been shown in learning arithmetic problems, for example, that mixing up different types of problems in practice sessions is more effective than concentrated drill on just one type (Rohrer & Taylor, 2007).

A learning program which can easily incorporate all three of these methods is through the use of flashcards. Flashcards always require active retrieval and can also be timed and sequenced to incorporate spaced practice and interleaving. Manual methods of doing this have been used at least since the 1970s (Leitner, 1972) but to implement a system which can optimize the user experience, a computer program is an especially effective way to accomplish this. In a computer-based flashcard system, algorithms can be employed to create schedules for spaced practice which incorporate the user's record of correct and incorrect responses, the subject matter of each question, and the timing of the appearance of each item.

Several such systems have been deployed over the past 25 years or so. These are typically designed for learning hundreds or thousands of facts overtimes extending over weeks, months, or years. Such applications include SuperMemo, Anki, Duolingo, Brainscape, FireCracker, and Memorang (http://www.supermemo.com, http://ankisrs.net, http://www.duolingo.com, http://www.brainscape.com, http://www.firecracker.me, http://www.memorangapp.com). Most of these applications require a web browser with an active network connection. Those which have self-contained mobile versions which can operate offline do not allow a centralized collection of user interaction data. More academically based applications which do allow this have not yet been widely deployed (Kam, Kumar, Jain, Mathur, & Canny, 2009; Pavlik, Kelly, & Maass, 2016).

With the smart devices currently available, it is possible to develop self-contained mobile applications which include all of the educational content and which have the software needed to adaptively schedule spaced practice completely independent of a central computational service. Such applications can generate user interaction data messages which can be sent to a central collection point. Such data is needed for learning scientists to evaluate student performance and to further refine scheduling algorithms. This allows users to make optimal use of their time in learning the material with these applications. As large datasets accumulate, such data will also be very valuable for advancing basic understanding of the learning process.

2 Mobile Practice of Course Content

The Higher Education division of McGraw-Hill Education (MHE) was interested in a way for students in a course using one of MHEs online, interactive textbooks, called SmartBooks (http://www.mheducation.com/highered/platforms/smartbook.html), to be able to use their smartphones to memorize some of the declarative knowledge presented in a title. They quickly realized that a mobile phone application which had been developed to allow candidates in India to study for the U.S. Medical Licensing Exam (USMLE), called StudyWise, could be adapted for this purpose.

This application was developed to utilize the existing homework questions from medically related titles and present them as flashcards on a smart mobile device. The application was entirely self-contained, with the content, the user interface, and the scheduling algorithm all entirely contained on the smart device. This allowed the application to function even when a network connection was not available, a requirement for use in areas with spotty WiFi or cellular coverage.

It was recognized that this software could be re-targeted to provide optimized practice of course material within the time frame of a college semester by modifying the scheduling algorithm to optimize for practice of dozens to hundreds of items over a few weeks or months rather than the USMLE application which was designed for study of thousands of items over the course of 1 to 2 years. By leveraging published research on learning and memory, a new algorithm was successfully developed which fit this use case (Riedesel, Zimmerman, Baker, Titchener, & Cooper, 2017).

2.1 Smart Device Mobile Applications

As currently deployed, each of eight separate content titles has its own IOS and Android app. Each app presents the homework questions in a title as an electronic flashcard, grouped by topic. The questions, known as probes or items, come from MHE's existing SmartBook database of probes. There are currently about 1500 SmartBook titles on a wide range of subjects. For all of these titles combined, there are some 2,350,000 distinct probes with almost three billion recorded answers for them from the web-browser-based SmartBook system.

In SmartBook, each probe is associated with a knowledge component called a Learning Objective (LO). The LOs are organized by topic, which in turn are related to the title's subject. In a course that uses a SmartBook title, the instructor creates an assignment for a specified set of LOs. Students see only probes associated with the LOs for that assignment, which they do online through a web browser.

StudyWise was designed as a mobile application which would allow students to practice all of the LOs in a title, organized by topic. The mobile nature of the applications allows users to learn and master material whenever they have a few spare moments and wherever they happen to be. The algorithm is designed to allow the learner to master each LO by repeated practice. Once an LO's associated probes have been answered correctly three times, it is considered learned.

The fact that it is entirely self-contained on the mobile device means that it can be used whenever the user has a few spare minutes. The pattern of session durations suggests shows that these applications are indeed being used primarily for a few minutes at a time, as can be seen in Fig. 1. This is rather different from the originally envisioned usage model, which was for dedicated 30 min sessions each day.

At present, there are separate IOS and Android apps for each of eight titles in subjects including Anatomy and Physiology, Medical Terminology, Psychology, Human Resources Management, American History, Majors Biology, Human Anatomy, and Medical Assisting Certification Prep.

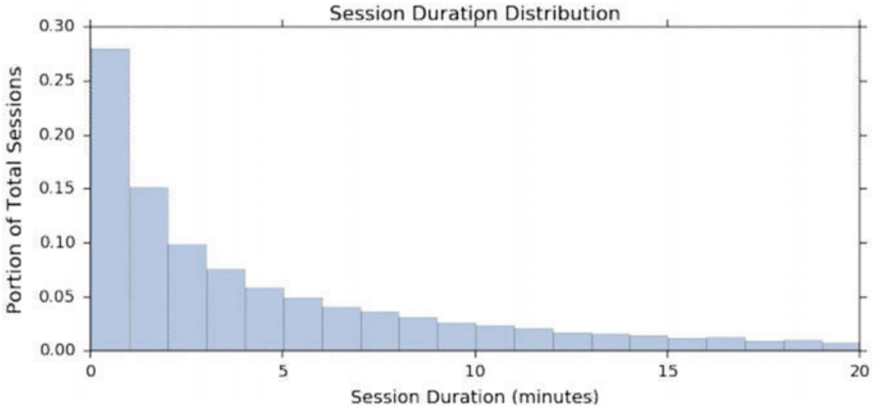


Fig. 1 A histogram of session duration times in minutes. More than one-fourth of all sessions are just one minute or less

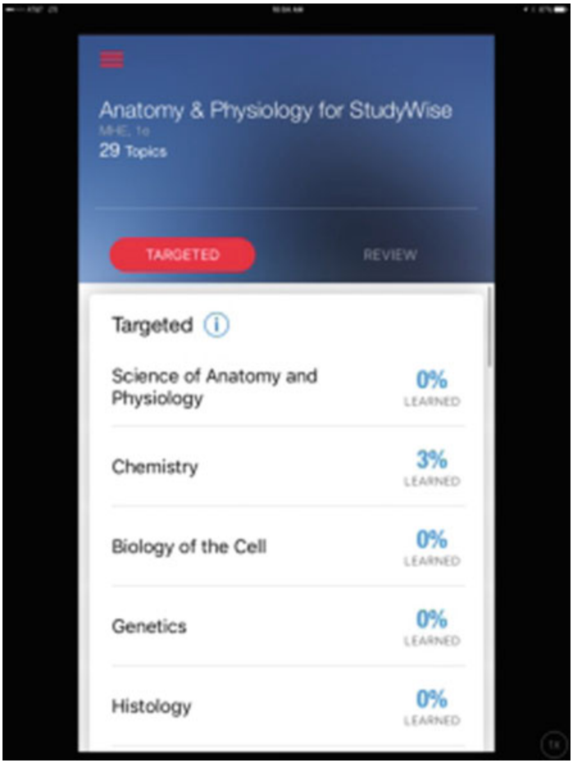
2.2 *User Interface*

The StudyWise apps are all native applications specifically written for smartphones, with an IOS and an Android version available for each title. When a user opens an app for the first time, they are asked if they want to register with an email address. This information is used only to allow the synchronization of a user's progress in the app between two different devices, such as, for example, an iPhone and an iPad or between an Android phone and an iPhone. There is a hamburger menu in the upper left-hand corner of the screen which allows the user to (a) study, (b) get help, or (c) create an account for synching between devices and to also check for updates to the app's content.

To start a learning session, the user goes to the "Study" page, which gives the app's name at the top of the page and has two modes: "Targeted" which presents a list of Topics from which the user then chooses a topic from which questions will be drawn or "Review" which presents an overall measure of progress through the app and then will present questions from the set of Topics a user has previously studied (Fig. 2).

Once a topic is selected, the algorithm selects the first LO and one of that LO's probes is then chosen at random. This first probe is then displayed on the next screen. The question presented will be one of several types found in SmartBook: multiple-choice single-answer (as shown in Fig. 3), fill-in-the-blank, multiple-choice multi-answer, matching, matching-rank, or deconstruction (a special type for medical terminology).

At the bottom of the page, a question about the learner's confidence about their knowledge of the question appears: "Do you know the answer?", with the two possible choices "Yes, I know it" or "I'm unsure". The user must answer this question in



Anatomy & Physiology for StudyWise
MHE, 1e
29 Topics

Targeted

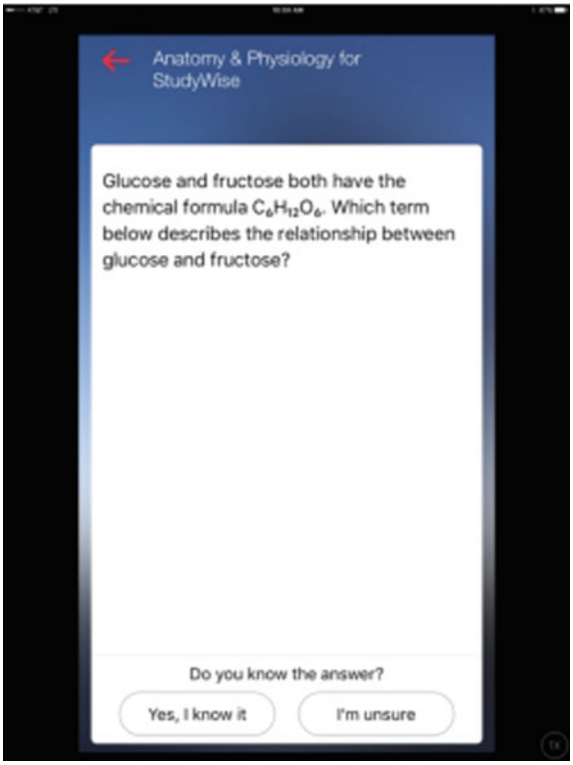
Science of Anatomy and Physiology	0% LEARNED
Chemistry	3% LEARNED
Biology of the Cell	0% LEARNED
Genetics	0% LEARNED
Histology	0% LEARNED

Fig. 2 The Topic selection page for the IOS anatomy and physiology for StudyWise app, with targeted mode selected

order to move on to the next page. For fill-in-the-blank, matching, matching-rank, and deconstruction questions, answering the confidence question submits the answer for grading.

For multiple-choice single-answer and multiple-choice multi-answer, the next page is the answer selection page (Fig. 4). For this case, the user can then either scroll back to the question page and update their confidence after seeing the possible answers (after which the answer page is again displayed) or simply select their answer(s) from those shown.

Once the user's answer is submitted, the answer is checked against the correct one and an answer response page is then shown (Fig. 5). This page reiterates the user's answer and indicated confidence, whether or not the answer was correct, and then gives additional background information on the question's content. The level of this additional information varies from question to question depending upon how much such background the author of the SmartBook probe included when it was originally created.



Glucose and fructose both have the chemical formula $C_6H_{12}O_6$. Which term below describes the relationship between glucose and fructose?

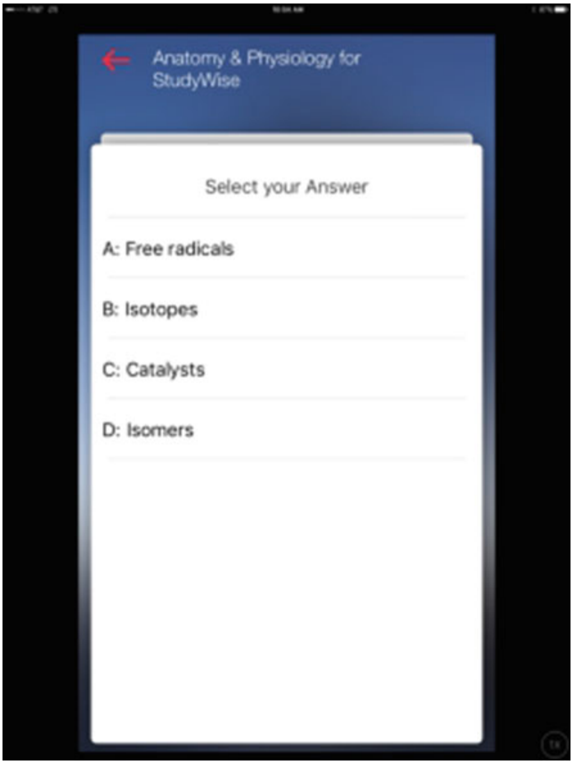
Do you know the answer?

Fig. 3 The question presentation page

The user then can select the "NEXT" button at the bottom of the page to move on to a new LO's probe chosen by the algorithm or the user can hit the back arrow in the upper left-hand corner to return to the Topic selection page, ending that session. At that point, the user can again choose a Topic or can hit their device's "Home" button to exit the application.

2.3 Algorithm Self-contained Within the Smart Device

A key element of effective memorization is repeating an item often enough for it to be remembered. However, we do not want to waste the learner's time by repeating items already well learned. The algorithm used in StudyWise is a proprietary one that uses a mathematical algorithm which has adjustable parameters that vary the appearance time of the questions associated with each LO. This spacing also depends on whether or not an LO's previous question was answered correctly or not, if this was not the first appearance of that LO. For an LO to be considered learned, a total of three correct answers to that LO's associated questions is required.



Anatomy & Physiology for StudyWise

Select your Answer

A: Free radicals

B: Isotopes

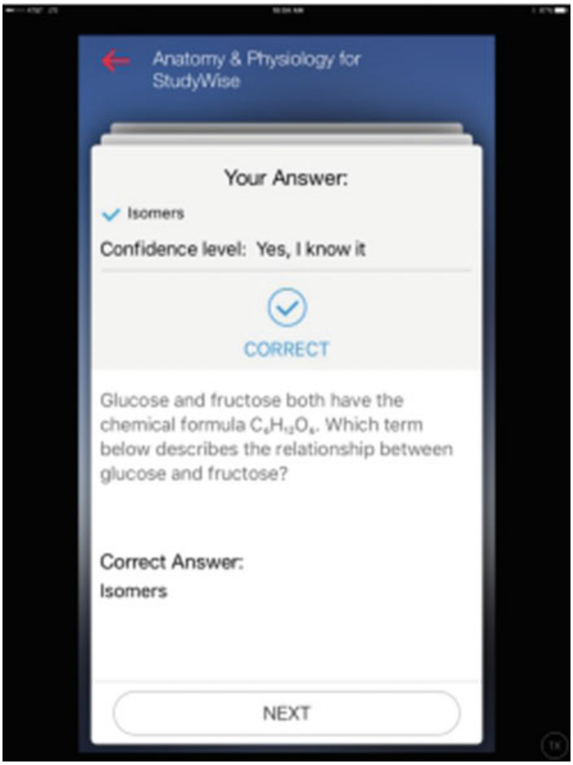
C: Catalysts

D: Isomers

Fig. 4 The answer selection page for multiple-choice questions

When an LO first appears, it will be repeated frequently until the first correct answer is given. After this, the LO is repeated somewhat less frequency, and after a second correct answer, the repetition interval is lengthened even more. The spacing also depends on the number of LOs included in a given topic. The desired practice schedule was specified by subject matter experts, who wanted topics of a particular size to fit within a specified practice schedule. A sample pattern of LO appearances with time is shown in Fig. 6, for a Topic with 100 LOs (Riedesel et al. 2017).

This algorithm is compact in both size and computational complexity and is easily self-contained within a learner's smart device. The learner's complete record of progress through each topic within a title is also stored locally. This means that the full adaptive experience can be presented to the learner even when not connected to the Internet. A learner can practice for hours, days, weeks, or even months without a network connection. This was originally motivated by the desire for medical students in rural India to be able to use this application but it also means that it can be used seamlessly even on an airplane flight or other locations where cellular data service is not available.



Anatomy & Physiology for StudyWise

Your Answer:

✓ Isomers

Confidence level: Yes, I know it

CORRECT

Glucose and fructose both have the chemical formula $C_6H_{12}O_6$. Which term below describes the relationship between glucose and fructose?

Correct Answer:

Isomers

Fig. 5 The answer response page

3 Data Messaging System

The availability of cloud-based computing and storage allows a smart device to return information on user interactions with StudyWise apps without the need for a database to present the user interface or to store data messages produced by user interactions. In this way, smartphones hosting StudyWise operate in the same way as devices which are part of the Internet of Things (IOT). The particular architecture used for StudyWise is based on Amazon Web Service’s “Serverless Computing” technology (http://aws.amazon.com/serverless/), which is frequently employed for IOT systems.

This architecture works very well for applications which are designed to stand alone on a user’s smart device, and it is also at the heart of how Internet-connected sensors and devices which are part of IOT return the data used in machine learning and other predictive analytics. Using this type of architecture, as noted, there is no database behind the application. This means that the usual methods of doing both business and educational analytics by querying relevant data from an OLTP database are not available. All of the data that business analysts and learning scientists use must come from a custom-designed system of messages which are sent as learners interact with the system.

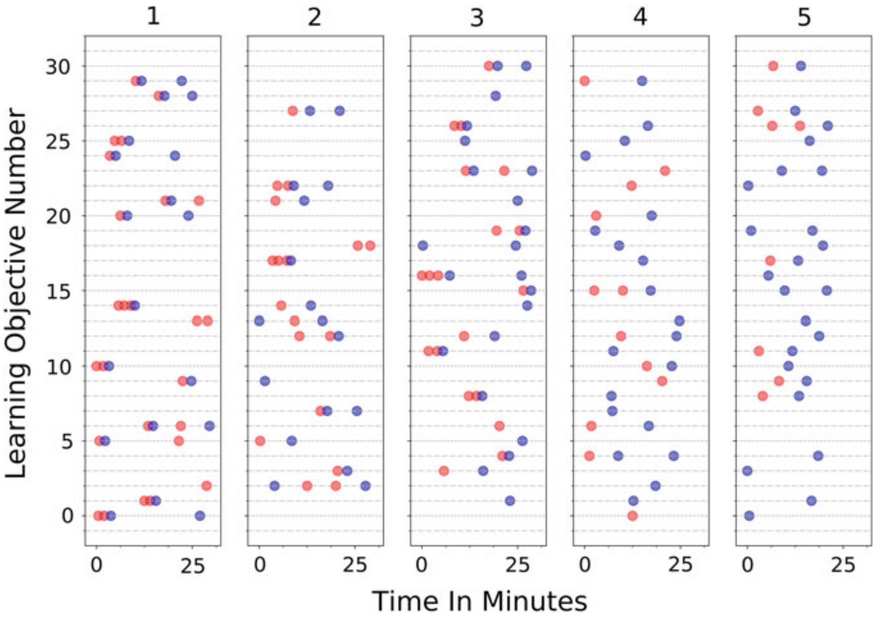


Fig. 6 Plots of the first 30 LOs from a 100 LO deck. Five 30-min sessions were needed to complete 100 LOs. Red dots indicate incorrect answers, as computed by the model, and blue dots are for correct answers. The start times for the sessions are separated by 24 h

When designing the messaging system, we solicited input from both the business owners of the application and learning scientists. We wanted to ensure the availability of the data needed for an understanding of the frequency and patterns of use needed for business analytics and for doing the Learning Science related to spaced practice and active retrieval. Both types of information can be used to improve the user experience and the educational effectiveness of this suite of applications.

By design, as little personally identifiable information (PII) is ever solicited from the users as possible and none is transmitted to the data collection system. The only PII information StudyWise is the email address used to synchronize a user's progress between devices. Even this information is contained in the JSON messaging data only as a non-reversible hash of the input email address.

3.1 Questions for Business Analytics

To determine what data might be needed for business analytics, we asked the Higher Education group for a list of the type of reports they might want to be able to construct from the messaging data. They provided a list which guided our design.

Here are some examples of the types of questions the business analysts wanted to be able to address:

- How much was each application used in terms of questions answered per unit of time (hour/day/week/month)?
- What is the pattern of usage in terms of time of day, in local time?
- How is usage correlated with the time of year, such as the start, middle, or end of a semester?
- How many learning objectives have been answered for each topic for each title?
- What is the average user progress through each topic?
- What is the usage of IOS versus Android for each title?

One field added specifically to address some of the business questions was the local time zone, in terms of offset from UCT.

3.2 Questions for Learning Science

In addition to data for business analysis, we also wanted to be able to measure the efficacy of the apps and to be able to gain insight into the learning science related to spaced practice and active retrieval. To be as comprehensive as possible, we consulted with the MHE Data Science team and with academic learning scientists while defining the messaging fields.

Among the types of questions we would like to be able to address in this area include:

- Which app (i.e., which Title) is the learner using?
- Which question (including its Topic and Learning Objective) has just been answered?
- Which answer was selected and was it the correct answer?
- How long did it take to answer the question?
- How long was spent looking at the explanation of the answer?
- What do the learning curves look like for each learning objective and or each student?
- What confidence level was chosen and how does it correlate with the correctness of the answer?
- How do performance and confidence vary as a given LO is repeated?
- How far through a given topic has the user progressed?

3.3 Data Fields and JSON Schema for the Messaging System

To be able to answer all of the above questions, the data fields in Table 1 were created. There are also several fields to identify the version of each app overall and the version of the adaptive algorithm which are not shown in this table. As the application is in

Table 1 Data messaging schema

Item	Type	Example	Source
IOS version number	String	9.0.1	Mobile OS
Device type	String	iPhone 6+	Mobile OS
IP address	String	10.10.10.10	Mobile OS
StudyWise application	String	A&P	App
Software version	String	1.0.0	App
Algorithm version	String	1.0.0	App
Topic IDs	String	[224562,135283,252034]	App
Topic titles	List string	[“Countries and capitals”, “Present tense irregular verbs”, “Past tense of verbs”]	App
StudyWise topic size	Number	68	App
User ID	String	315352FD-44B1-406E-A785-B74100A0B2A9	App
Session ID	String	3B20792D-F46B-48B4-8F45-AA79AE620EA	App
Date/time session start	String	2015-08-01T06:00:00.000Z	App
Event type	String	One of “targeted” or “review”	App
Learning objective ID	String	CL323778e1-8d23-425f-8be4-996f20ae4933	App
Question identifier	Number	589823749	App
Question type	String	Multiple-choice/MCQ	App
Date/time presented	String	2015-08-01T06:00:00.000Z	Mobile OS
Date/time answered	String	015-08-01T06:00:00.000Z	Mobile OS
Answer selected	String	CL323778e1-8d23-425f-8be4:243565606:0	App
Success/failure	Number	0 for failure, 1 for success	App
Confidence	List	1 or 2 [1, 2]	App
User interrupted details	List or string	[“phone”, “other”]	Mobile OS
Session progress	Number	53%	App
Product title	String	Anatomy and physiology	App
Product Id	Number	151514	App
Local time zone	String	−5	Mobile OS
Date/time session end	String	2015-08-01T06:00:00.000Z	App

use, each time an answer to a question is submitted, information is sent. Fields which are directly linked to a user’s practice session come from the StudyWise “App” and those which give information on the user’s device, time zone, and other information on the session’s context come from the mobile devices “Mobile OS” and hardware.

Although both IOS and Android mobile devices can transmit a user's location, the business owners were concerned that asking the user for permission to include their location would be seen as unnecessarily intrusive and might also raise student privacy issues, so this information is not collected. Local time zone information is collected for determining the time of day an app is being used.

3.4 JSON Schema

JSON was specified by MHE's data engineering group for encoding the messages in order to be compatible with a company-wide standard for data messaging systems for future educational software products. IMS Caliper was considered as the basis for this API, but at the time this API was designed (in mid-2016). Caliper lacked a number of fields needed to satisfy both the educational and business requirements, and MHE's data architects approved the schema shown here.

One great advantage of using JSON is the ability to query a collection of JSON documents using standard SQL query commands. This can be done either through systems such as Apache Spark (http://spark.apache.org), jQuery (https://jquery.com), or other similar tools. Even quite complex queries have worked successfully and an experienced database analyst can fully leverage their background in the SQL language while analyzing this type of data. In analyses done thus far, essentially all SQL queries attempted using Apache Spark SQL have worked on this dataset.

3.5 Online and Offline Modes

The first version of StudyWise was created to allow medical students overseas to study for the United States Medical Licensing Exam (USMLE). Such students could possibly be in areas where network coverage was spotty, so StudyWise was designed to be fully usable when the user's mobile device had no cellular data connection. This means that all of the logic and code needed for a fully adaptive experience needed to be contained in JavaScript code on the mobile device. Similarly, all of the questions, answers, explanations for incorrect answers, and ability to determine if an answer was correct or incorrect are included in the mobile app.

However, it was still desired to have a record of user interaction data, even for sessions done offline. In order to accommodate this, the JSON analytics data message for each interaction is either (1) buffered and then sent to the collection server in MHE's AWS virtual private cloud (see below) if the user is connected or (2) stored locally on the mobile device for later transmission if a network connection is not currently available. Stored data from offline sessions is later sent to the collection server when a user is running the StudyWise app and has a network connection. At present, this is the only widely available mobile educational application of which we know that has this capability.

3.6 Receiving System

As an app is in use, each time a learner answers a question and a JSON record is generated. This record is stored locally in a buffer and when this buffer is full or the session ends, the buffer of records is securely transmitted to a receiving service running on an AWS instance. The receiving service validates the data, as described below, and stores it in a flat file in a secure location in AWS S3 file storage.

The storage is a simple directory hierarchy organized by year, month, day, and hour. Many utilities exist to read JSON data stored in this way, including in Apache Spark. As noted above, this type of JSON file system can also be directly queried using SQL just as if it was a relational database.

This method of collecting and storing the data is limited in speed by the transmission time of the Internet but otherwise is capable of very high data throughput and can be set up in parallel to scale almost arbitrarily, if necessary. Total storage available is essentially limited only by the ability to pay for it.

This means that this data system could handle potentially millions of users daily by simply scaling out the receiving and storage systems, with no change in architecture. Using the capabilities of AWS means that no server or storage hardware need be purchased to set up such a system.

4 Security and Privacy

As a system which is used in education, it is very important to maintain the privacy of each user's data. This requires an architecture which protects the integrity and security of the data at each step of the collection and analysis. An important first step, as noted above, is to store no personally identifiable information in the analytics data stream.

4.1 Data Encryption

Data encryption is a critical element of security. Strong encryption of data both at rest and in transit prevents an attacker who might gain access to a storage system or communication channel from obtaining sensitive information. In StudyWise, this means that all communications between the mobile application, the data collection end point, and the processing pipeline are encrypted using HTTPS/SSL.

Data stored for analysis, and the derived datasets that comprise the output of those analyses, are encrypted in cloud storage. Amazon S3 supports multiple mechanisms including SSE-S3 (transparent server-side encryption), SSE-KMS (server-side encryption using AWS key management), and SSE-C (server-side encryption using customer-provided keys).

Furthermore, the pipeline runs in a virtual private cloud on infrastructure not directly accessible from public networks.

4.2 Access Policies and Controls

User authentication controls who can access the analytics environment. Integrated identity management can simplify user management by leveraging a service to authenticate users. Examples include SAML2.0 and Active Directory.

In order to ensure that only authorized users can access data stored and processed in the analytics pipeline, role-based access provides fine grain controls on storage (who can access data sources in the processing environment), on clusters (who/what is authorized to configure/launch/terminate/restart clusters), on processing jobs (who/what can attach and run processing on clusters) and the environment (configuration and settings).

Auditing and logging provide the ability to alert on, monitor, and review key events in the environment.

4.3 Data Integrity

In the compute layer, within Apache Spark, the RDD (see below) provides data immutability and fault tolerance by design. In the storage tier, AWS S3 provides not only an extremely high level of durability (99.999999999%) but also the ability, via the optional Content-MD5 request header, to verify the integrity of data stored there.

4.4 Certifications

An end-to-end data analytics pipeline, especially one reliant on third-party managed or cloud services, is a system based on the integration of many components. No matter how strong the data encryption and access policies in place, the system is only as secure as its weakest link. Managed service and cloud providers certify their platforms according to documented compliance standards.

The FERPA (Family Educational Rights and Privacy Act) standard governs access to educational information and records. Other standards, many originating in the financial and health industries, are also relevant in education.

AWS documents their compliance at (https://aws.amazon.com/compliance/). Certifications include FERPA, HIPPA, GLBA, FISMA, RFR, and PCI DSS.

Databricks, the computing environment used here for data analysis, documents their platform security and compliance at http://go.databricks.com/. Their security measures are based primarily on SOC2 Type-1 Certification. SOC2 Certification encompasses five key areas: security, availability, processing integrity, confidentiality, and privacy.

5 Data Processing and Analysis

In order to use the data collected for either educational or business insights, we need to set up a system to read, clean up, and analyze the JSON data produced originally on each user's device which has been sent to the AWS S3 collection system. It is possible to do all of this entirely in the cloud using AWS and other services, and this is the approach that we have taken with StudyWise.

In particular, we would like a system which has the following characteristics: (1) straightforward to access, use, and maintain and (2) powerful and scalable enough to handle increasing amounts of data as StudyWise is more widely deployed. In the past 10 to 15 years, several systems have been developed which have been designed to satisfy these types of needs. All are based on horizontally scalable clusters of servers to allow computing to be done in parallel.

Many high-performance computing systems in use by the academic and government research communities use the Message Passing Interface (MPI) to do parallel computing. This requires writing code in either C or Fortran. This is scalable to the very largest systems in existence and can be used for the most complex types of parallel computing but it requires a very high level of specialized knowledge and expertise to use. MPI can be used not just for processing large amounts of data in parallel but also for doing very large parallel theoretical calculations for such applications as global climate models and modeling the interiors of stars.

A system designed to make parallel computing more accessible to a wider range of users is the Apache Hadoop/MapReduce system of software developed for the parallel processing of large amounts of data. This system is simpler than MPI but still requires the use of the Java programming language and also requires writing intermediate results in its processing pipeline to disk, which can make it rather slow.

To overcome the limitations of MPI and Hadoop/MapReduce, a system called Apache Spark (http://spark.apache.org/) has been developed by the open-source community which allows all of the computation to be done in a compute node's memory and which also combines some of the steps in map/reduce processing, making the writing of code to process large amounts of data in parallel considerably easier than previous systems. Apache Spark has come into widespread use in just the last 3 to 4 years but is becoming the dominant method of processing very large amounts of data.

Spark also is available with APIs for programming in Scala, Python, R, and SQL, making it accessible to a much wider audience. On a practical level, a Spark program can be many, many times faster (x100, in some cases) than an equivalent Hadoop/MapReduce program, as long as all of a given subset of data in a parallel processing job can be fit into the memory of one of the servers in a Spark computing cluster.

In this section, we show how Apache/Spark can be used to process and analyze large-scale data with examples of its use in StudyWise.

5.1 Apache Spark

Apache Spark is both an engine and API for large-scale data processing, making very large-scale parallel computing accessible to a much wider range of users than previous parallel computing methods. Large organizations have deployed Apache Spark on clusters consisting of thousands of nodes, processing petabyte-sized datasets which grow on the order of terabytes per day (Tsai, 2017).

The main abstraction in Apache Spark is the "RDD", or resilient distributed dataset.

"Resilient" refers to redundancy of data across compute nodes, Apache Spark's fault tolerance and ability to continue processing when compute nodes fail.

"Distributed" refers to the fact that data is partitioned across many multiple cluster nodes. More importantly, computations are moved to the data, rather than the traditional approach of moving data to central location(s) for processing. This eliminates I/O bottlenecks and allows processing to scale in a linear fashion as the amount of data grows.

"Dataset" refers to large-scale collections of data. An important attribute of these datasets in Apache Spark is immutability. Immutability drastically simplifies the cost and complexity of coherently processing distributed data. Spark represents processing steps in a directed acyclic graph. This traversable graph, combined with the immutability of data at each intermediate step in the process, allows datasets to be easily reconstructed on failure, contributing both to the resiliency and distributed characteristics of the system.

More recently, Apache Spark has added an additional abstraction layer on top of the RDD called a dataframe. The dataframe makes the manipulation and processing of the data conceptually simpler than is the case for RDDs and it is very similar to dataframes found in Python Pandas and in R (where the concept was first developed).

Data streaming Apache Spark supports both batch-oriented and stream processing using a single unified API. The use of streaming allows essentially up to the second processing to be done on the analytics data as it arrives from the user's mobile devices, if desired.

The Apache Spark Structured Streaming API mimics the Apache Spark batch API, but behind the scenes it utilizes the SparkSQL engine to continuously update data as new information arrives. The same SparkSQL queries and operations that work on batch-loaded information can also be performed on streaming dataframes.

In the concrete examples below, Spark Structured Streaming is used to perform event-driven processing.

APIs and Analysis Tools Apache Spark's processing API supports multiple languages, including Scala/Java, Python, and R as well as a dialect of SQL. In the following example, data is ingested into a Spark dataframe. Here are brief examples of how to read JSON flat files into a Spark dataframe for the different Spark languages:

Table 5.1. Scala

```
val eventSampleDf = spark.read.json ("pathBatch / *. json")
```

Table 5.2. Pyspark (Python)

```
eventSampleDf = spark.read.json ("pathBatch / *. json")
```

Table 5.3. SparkR (R)

```
eventSampleDf = read.df (sqlContext, "pathBatch/* . json", "json")
```

The Apache/Spark API also allows SQL-like operations to be performed on dataframes.

A simple select/filter operation in Scala using the SparkSQL API:

Table 5.4. SQL query using Scala

```
val  apDf = eventsBatchDf .
                    filter ($"session.studywiseApplication" === "A&P")
```

The same operation expressed in Spark SQL syntax:

Table 5.5. SQL query directly in Spark SQL

```
%sql select *  from apDf where session.studywiseApplication = 'A&P'
```