LearnSphere: A Learning Data and Analytics Cyberinfrastructure

John Stamper
Carnegie Mellon University
Pittsburgh, PA, USA
jstamper@cmu.edu

Steven Moore
Carnegie Mellon University
Pittsburgh, PA, USA
stevenmo@andrew.cmu.edu

Carolyn P. Rosé
Carnegie Mellon University
Pittsburgh, PA, USA
cprose@cs.cmu.edu

Philip I. Pavlik Jr.
University of Memphis
Memphis, TN, USA
ppavlik@memphis.edu

Kenneth Koedinger
Carnegie Mellon University
Pittsburgh, PA, USA
koedinger@cmu.edu

LearnSphere is a web-based data infrastructure designed to transform scientific discovery and innovation in education. It supports learning researchers in addressing a broad range of issues including cognitive, social, and motivational factors in learning, educational content analysis, and educational technology innovation. LearnSphere integrates previously separate educational data and analytic resources developed by participating institutions. The web-based workflow authoring tool, Tigris, allows technical users to contribute sophisticated analytic methods, and learning researchers can adapt and apply those methods using graphical user interfaces, importantly, without additional programming. As part of our use-driven design of LearnSphere, we built a community through workshops and summer schools on educational data mining. Researchers interested in particular student levels or content domains can find student data from elementary through higher-education and across a wide variety of course content such as math, science, computing, and language learning. LearnSphere has facilitated many discoveries about learning, including the importance of active over passive learning activities and the positive association of quality discussion board posts with learning outcomes. LearnSphere also supports research reproducibility, replicability, traceability, and transparency as researchers can share their data and analytic methods along with links to research papers. We demonstrate the capabilities of LearnSphere through a series of case studies that illustrate how analytic components can be combined into research workflow combinations that can be developed and shared. We also show how open web-accessible analytics drive the creation of common formats to streamline repeated analytics and facilitate wider and more flexible dissemination of analytic tool kits.

Keywords: data mining algorithms, data repositories, learning analytics

1. INTRODUCTION

As new educational technologies are employed, learning researchers are increasingly gaining access to large datasets (e.g., Koedinger, 2016; Stamper and Pardos, 2016; Selent et al., 2016; Lohse et al., 2019). New tools and systems are emerging to use big data to improve understanding of the mechanisms and factors involved in human learning. This interdisciplinary effort has led to many novel realizations about the nature of learning (e.g., Koedinger et al., 2012), new paradigms (e.g., Ambrose et al., 2010), and new questions (e.g., Koedinger et al., 2013). LearnSphere enables users to share data and analytics in an intuitive, visual environment for researchers, developers, and instructors of all skill levels. Researchers can begin immediately reviewing existing data analytics or build their own workflows right inside the web application using publicly available datasets and analytics tools contributed by the learning science community. Instructors can utilize pre-made workflow "templates" with their own classroom data to extract useful insights. Programmers can make their software tools available to the broader community without putting the burden of installation and setup on the user. In short, LearnSphere supports data and application sharing among learning researchers within an extensible framework designed to encourage common educational data standards and best practices.

LearnSphere addresses several common issues encountered during research in educational domains, including:

1. **Data sharing** and accommodating the heterogeneous nature of educational data and managing access to data.
2. **Analysis sharing** by supporting the creation and sharing of a wide range of analysis tools and allocating the computing resources required to support such tools.
3. Creating a **distributed** network of data and analytical resources.

In order to support data sharing across separate silos of data-driven education research, LearnSphere takes into account the different kinds/types of data used, the time scale of the data collection and associated phenomenon, and the psychological constructs that are the focus of the research as seen in Figure 1. We have highlighted a number of existing data repositories that have been integrated into the LearnSphere work. As seen in the figure, the focus of DataShop (https://pslcdatashop.org) has been on educational technology interactions from cognitive tutors, educational games, simulations, online courses, including clicks, text and symbolic entries, that are recorded in ranges from 100s of milliseconds to many minutes although the data my extend over full semesters or school years. Analyses have focused primarily on investigating cognitive skills, concepts, metacognition, and, increasingly, motivation. The focus of MOOCdb (http://moocdb.csail.mit.edu/) is on interactions in massive online courses that typically have a longer time scale (e.g., full quiz results after watching an online lecture). Much of the engagement in MOOCs is individuals learning on their own and collaborating with their peers, like discussion board interactions (Yang et al., 2014; Wang et al., 2015). One tool that makes use of such data is DiscourseDB, which is designed for aggregating, organizing, and analyzing diverse discourse data, such as debates and discussions, to support research in fields like linguistics, political science, and communication studies. The primary focus of DiscourseDB is on discussion analysis in formal and informal learning contexts more broadly, as well as analysis of text more broadly (Jiang et al., 2019), both in terms of its data pipeline (Rosé & Ferschke, 2016; Rosé, 2017) and processing techniques (Yang et al., 2016; Howley & Rosé, 2016; Fiacco

& Rosé, 2018; Fiacco, Cotos, & Rosé, 2019). Before LearnSphere, these data analytic tools, DataShop, MOOCdb, and DiscourseDB, were separate and not interoperable. Correspondingly, they have serviced research communities that are somewhat separated (or "siloed") communities, as indicated by corresponding conference communities, Educational Data Mining, Learning@Scale, and Computer-Supported Collaborative Learning. To be sure, these learning data silos and research community silos are not limited to these three as there are other areas of data/analytic sharing at different time scales, such as year-long K12 assessment data (e.g., https://nces.ed.gov/nationsreportcard/data), affect and video data (Datavyu Team, 2014; Warlaumont et al., 2017), and research on other constructs, such as how to select and navigate an undergraduate program of courses (Jiang, et al., 2019).

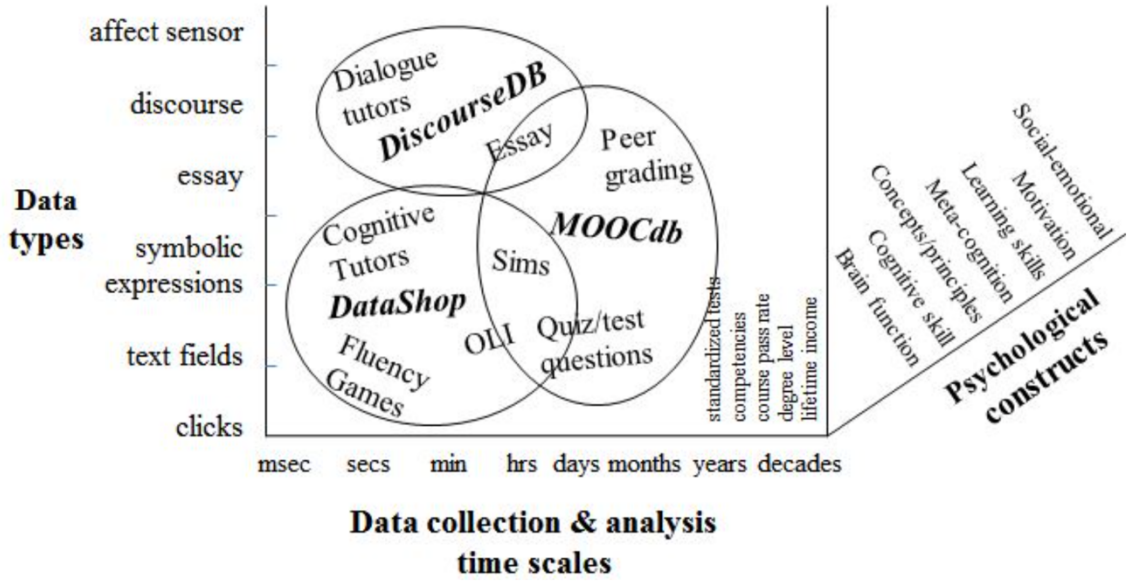


Figure 1: Many paradigms of data-driven education research differ in data types, time scales, and research goals. Disciplinary silos are fostered by differences. The LearnSphere software infrastructure provides analytics across these silos.

LearnSphere was developed to provide avenues for both interoperability across data and analytic silos, in collaboration across research communities in order to share analytics and tools. Given different communities use different analytic software or programming languages (e.g., R, Python, SPSS, C+, etc.) and given some have other forms of expertise (e.g., educational content, psychological constructs) besides programming, we wanted to facilitate analytics sharing and reuse that allows but does not require users to program.

Toward these cross-community engagement goals, the LearnSphere.org website provides pointers to a wide collection of learning data stores, analytics tools, and data-generation methods, pointers to workshops and tutorials at different conferences, and video tutorials on learning analytics. But, most importantly, LearnSphere.org provides an ever-expanding authoring environment for sharing, reusing, recombining, and creating analytic procedures in workflows connected to data. This workflow authoring environment is called Tigris and it facilitates interoperability across different data types and time scales to link existing data silos. Tigris takes advantage of common features between data being collected across different sources while allowing data- and domain-specific properties and processes to exist. It sports a simple drag-and-drop interface, letting users build complex workflows consisting of various

data sources, analysis components, and visualizations with no programming experience. Figure 2 shows a simple workflow (workflowId=135[1]). Using Tigris, a broad range of researchers from different disciplines and with differing technical experience can begin analyzing data immediately.

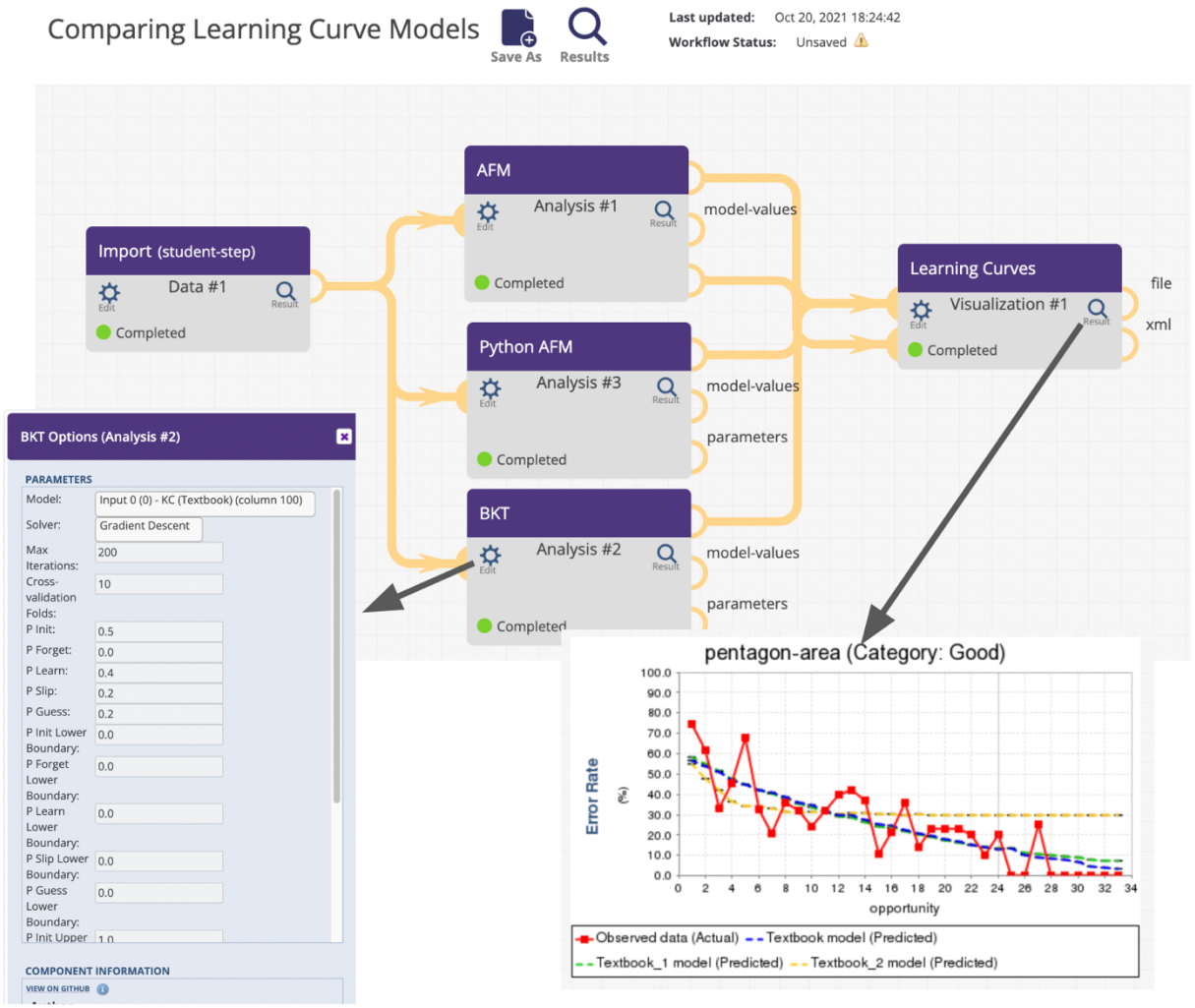


Figure 2: At the top is an example workflow, connecting a single data source (DataShop student-step export) to three different analysis methods (two different AFM components and one BKT) whose outputs are compared using a Learning Curves visualization component. The component-specific parameters can be modified within the workflow (see bottom-left) by clicking on the components Edit icon and each component's output results can be viewed by clicking on the Result icon. The bottom-right shows an example output from the Learning Curves component comparing the predicted error rates generated by the three analyses.

While LearnSphere is primarily targeted to support educational researchers, we have a broad list of use cases targeted to a broad range of users (see Section 4). One key goal is to make learning engineering accessible to instructors, course developers, social science researchers who

[1] This workflow is available publicly at https://pslcdatashop.web.cmu.edu/LearnSphere?workflowId=135.

are not well versed in EDM researchers, as well as non-programmers. Further, the Tigris tool allows for programmers who have expertise in a specific programming language to benefit from having analytics available in other languages they do not know as well. Researchers who wish to contribute their own analyses, classifiers, transformations, and visualizations can minimize time spent on interface implementations and access controls by using the LearnSphere framework. The LearnSphere components are open source and available on github[2], making it possible to build derivative components for reuse in the Tigris workflow environment. Existing tools and visualizations that aid in understanding the data are far from being complete, but a streamlined workflow that facilitates sharing and integration of open-source software combined with a push toward data standardizations could reduce such complications (O'Reilly and Veeramachaneni, 2014).

As such, LearnSphere provides a platform for others to share their work with the broader learning science community without common pitfalls that come with custom software (installing new libraries, finding old libraries, system compatibility issues) and data accessibility (Institutional Review Board (IRB) compliance, access controls, sampling, and storage). The ability to instantly share workflows (data, processes, parameters, results, and notes) addresses concerns of reproducibility, replicability, traceability, and transparency as they relate to data-driven education research. In Tigris, components are made available through an interactive workflow where processes, which we will call components, can be dragged and dropped, connected together, and configured-- essentially, they are the building blocks of a program, which we call a workflow. Workflows can be saved, shared, and copied both with and without access to actual data.

2. BACKGROUND AND RELATED WORK

The primary goal of LearnSphere is to connect the learning community to educational data repositories and analytics software that can be shared. The idea of LearnSphere grew out of the work of the NSF funded Pittsburgh Science of Learning Center (PSLC) LearnLab (https://www.learnlab.org) and the research around Educational Data Mining. One key artifact of the PSLC LearnLab is the DataShop repository (Koedinger et al., 2010), which has become the largest open repository of transactional data collected from educational technologies. DataShop includes a suite of analytic tools for working with data available in the repository. While DataShop continues to be a popular tool within the educational data mining and learning analytics communities, there is an acknowledgment that the repository and its associated tools do not encompass all use cases, particularly regarding the diversity of data types and the range of analyses that researchers in these fields focus on. The primary goal of LearnSphere is to allow data from multiple repositories to be brought together in a way that both data and the analyses can be shared. Figure 1 shows the many types of data both commonly used in these fields and incorporated into the DataShop repository. In this section, we address the targeted groups of researchers that we have identified and engaged with LearnSphere as well as data repositories that have been integrated or could be integrated.

[2] https://github.com/LearnSphere/WorkflowComponents/

2.1. TARGET COMMUNITIES OF RESEARCHERS

LearnSphere envisions supporting a wide range of researchers in the educational research space that make use of educational data and analytics. The communities that can benefit most from LearnSphere are those that utilize educational technology broadly, capture data from these technologies, and have a need or desire to standardize analytics within their communities. Two primary areas of focus have been the Educational Data Mining (EDM) and Learning Analytics (LAK) communities, which have been previously compared (Baker et al., 2012). We have hosted workshops and tutorials at the prime conferences for both communities. Through the LearnLab Summer School[3], we have introduced over 400 young researchers (about 45 per year since 2015) to LearnSphere and many used it as part of an educational data mining project initiated during the week.

2.2. DATASHOP

The largest backbone of the LearnSphere data infrastructure is the LearnLab DataShop. This repository was originally conceived and created as part of LearnLab, a NSF-funded Science of Learning Center started in 2004. DataShop has become a major resource for researchers in educational data mining and the learning sciences and is the largest open repository of educational data collected from educational technologies and learning systems. DataShop is both a repository of learning data and a web application for performing exploratory analyses on those data. DataShop specializes in data on the interaction between students and educational software, including online courses, intelligent tutoring systems, virtual labs, online assessment systems, collaborative learning environments, and simulations. As of January 2024, DataShop offers over 4,330 datasets and across these data sets, there are over 395 million software-student transactions, representing over one million hours of student data. A key feature is DataShop's set of tools for exploring cognitive models both visually and statistically. In DataShop, a cognitive model is a mapping between hypothesized "knowledge components"—a more general term for skill, concept, schema, production rule, misconception, or facet—and steps in the procedural completion of an online activity. A researcher can define a hypothesized model in a spreadsheet and upload it to DataShop, where it becomes available for analyses. Visual analyses include learning curves and an error report, while statistical analyses include a logistic regression model that describes how well alternative cognitive models predict student learning. DataShop has been valuable to both primary and secondary researchers in the learning sciences fueling hundreds of secondary analysis studies and associated papers. For researchers who add their data to DataShop, access controls allow them to keep the data entirely private, share selectively, or make the dataset accessible to all registered users. DataShop enables secondary research by allowing registered users to view public datasets and request access to private ones.

2.3. DISCOURSEDB

We also leverage the existing DiscourseDB as part of LearnSphere to add dialog, forum, and other discourse data into LearnSphere. DiscourseDB is a data infrastructure designed to take data from a wide range of platforms where discourse data is generated, each with their own

[3] https://learnlab.org/simon-initiative-summer-school/

schema highlighting some subset of the full range of aspects of discourse structure and transform the representation into one common representation. The foundation of computational analytic work is representation of data. When expanding beyond simple, flat representations of discourse structure (such as sequences of contributions, as in chat data) to learning in contexts such as threaded discussion forums or MOOCs conducted over a set of loosely integrated platforms, the form that the discussions may take becomes more diverse as they are embedded in a variety of platforms. They may even occur simultaneously through multiple separate streams. DiscourseDB enables translation of data from multiple streams into a common, integrated representation. The interface level representation is translated down into Discourses (e.g., the collection of discussions taking place within a course), with nested Discourse Parts (e.g., separate collections of discussions, or individual discussions), grounding out into a list of Contributions (i.e., a single utterance contributed by a student in some form), which may be related to one another through Relations, and which are associated with content that can be associated with Annotations. This common representation enables combining data across communication streams and applying common modeling technologies.

Once modeling tools and interventions are developed that make use of the DiscourseDB schema, then the same analysis can be applied to different discourse data very easily once it is imported into DiscourseDB (Jo et al., 2016; Jo & Rosé, 2015), and then the interventions can be used in the new context with little adaptation beyond the ability to call the intervention within the new platform (Rosé & Ferschke, 2016). If instead a different type of model is desired, then some work to develop that model using DiscourseDB representations will be necessary, but once that work has been done, it can again be applied to any data set imported into DiscourseDB. In that way, the pipeline can be seen as "plug-and-play", making it easy to include new sources of discourse data or new analytic approaches, and getting the complete many-to-many pairing for free, thus providing an increasing multitude of tools to support the data-to-intervention loop over time.

Discourse data is inherently more sensitive than much of the other forms of data served from LearnSphere's other facilities. And thus, DiscourseDB has fewer public datasets to boast of, though its accompanying text mining tools, TagHelper (Rosé et al., 2008) and LightSIDE (Mayfield & Rosé, 2013) in particular, have been used by tens of thousands of users. The focus of the DiscourseDB work has instead been on sharing the code infrastructure with others interested in large scale analysis of discourse data. Currently there is a research instance of DiscourseDB at Carnegie Mellon University, at the University of Toronto, and Beijing Normal University, and at SUNY Albany, with new collaborations beginning at other universities, and recent interest from industrial partners as well. Shareable data sets so far are all located in the Carnegie Mellon University instance and have all come from semesterly research studies run within a large online Carnegie Mellon University course, focusing on Cloud Computing (Sankaranarayanan et al., 2019; Sankaranarayanan et al., 2018).

2.4. INTEGRATING OTHER DATA SILOS

MOOCdb is a data schema and workflow structure for standardizing and analyzing MOOC data (Veeramachaneni et al., 2015) developed at MIT. The goal of MOOCdb is a developed data model for MOOC data that captures the information in multiple, low level MOOC data streams and expresses it in useful levels of abstraction, such that all the fields in the raw data are preserved, but the data is more structured and concise. The data model has been designed with input from a number of researchers in education, including instructors, platform providers, and data scientists. The data model abstracts on a per student basis, activity related to resources,

submissions, forums and wikis. It provides a high-level view of behavioral activity by multiple sets of data tables. This allows straightforward database queries for data abstraction. It also contributes to improved speed of extraction and software analytics scalability.

DataStage is a repository of course data made available by the Vice Provost Office for Online Learning (VPOL) at Stanford, which facilitates the teaching of online classes. The instruction delivery platforms are instrumented to collect a variety of data around participants' interaction with the study material. Examples are participants manipulating video players as they view portions of a class, solution submissions to problem sets, uses of the online forum available for some classes, peer grading activities, and some demographic data. VPOL makes some of this data available for research on learning processes, and for explorations into improving instruction through DataStage (Lohse et al., 2019).

LearnSphere was specifically designed with use cases around the course datasets available in the described repositories, however, we have also made the import of data robust enough to gather data from alternative data sources as well. Some of these include data from the Department of Education's "What Works Clearinghouse" (2012), other educational architectures such as learning management systems (Canvas, Blackboard, D2L, Moodle, etc.), other frameworks such as GIFT (Sinatra, 2022), and various Learner Record Stores (LRS) in formats such as the Experience API (xAPI) (Sottilare, et al., 2017). Additionally, we can import and combine these data with broad longitudinal datasets from USA Bureau of Labor Statistics (BLS) and Census, or other high level outcome data from sources such as the National Assessment of Educational Progress (NAEP) or the Organization for Economic Co-operation and Development (OECD).

3. LEARNSPHERE ARCHITECTURE

The development of LearnSphere was intentionally designed to allow for maximum flexibility in integrating a wide variety of data sources and analyses, and to be an entry point for the learning community. In order to bridge multiple educational data repositories and analytics software in a persistent, peer-reviewed environment, Tigris is the primary interface for building workflows from data sources, analytics, computational models, and visualizations. Workflows can easily be shared between users, preserving the data and parameters at each step of the process so that studies can maintain transparency and so that replicating or reproducing the study is not an arduous task.

In addition to a lack of transparency and the technical difficulty associated with replicating studies which use advanced frameworks, many existing platforms do not adequately address researchers' demands with big data analytics (Gardner et al., 2018). Platforms might offer data but with insufficient analysis capabilities, or they offer powerful analysis functionality but not data. In Tigris, by allowing users to upload and manage their own code and data, the platform is not restricted to a limited set of analysis tools. Another issue mentioned in Gardner et al. (2018) is that the demand for sufficient computational resources to conduct analyses are often lacking in platforms. Tigris provides ample resources for analysis, as workflow components are executed in parallel using a cluster of nodes which sits behind a load balancer. This allows the system to be manually configured to account for higher user demand or for higher computational demand per component.

A key feature of the architecture is open inputs and outputs for workflow components. We strongly believe that data formats should be flexible and rise from the communities that use them. For this reason, we do not require specific formats in any components, although we have

seen some standardization coalesce naturally in the building of the new tools, for example, many components have adopted the DataShop student-step schema, which is an aggregation of multiple attempts at a step (i.e., transactions) into a single row for each student-step opportunity that summarizes, among other things, the number of incorrect and correct attempts and hint requests the student made at that step.

3.1. INFRASTRUCTURE FOR SHARING AND USING DATA

LearnSphere instances can be run anywhere, giving organizations the ability to directly and physically control the level of access to their data and services. Instances can be connected to the existing LearnSphere server ring where datasets and their meta-data are propagated throughout the participating servers. For example, users at the CMU instance of LearnSphere can see links and information about datasets stored in the Memphis instance and vice versa.

The existing data infrastructures mentioned in section 2 (DataShop, DiscourseDB, MOOCdb, DataStage, etc.) have been integrated with LearnSphere through the Tigris workflow tool interface for control. These data sources maintain their own curation layer, access controls, and ownership rights which are available to use in the Tigris tool. Using the Tigris interface, users can upload data in any form to perform any number of sequential transformations, analytics, visualizations or reporting. Imported data is most typically a file representing a table, with rows indicated by line breaks and cells indicated by typical delimiters like tab or comma. Other import forms include database tables, discourse text, structured data, or any kind of binary file.

Although data sources can be uploaded directly into the system and shared among LearnSphere users, the ability to browse an organized repository that follows a defined structure is preferable to an unstructured file dump. One aim of LearnSphere is to simplify the process of accessing data for analysis using existing learning data repositories. One integrated repository, DataShop (Koedinger et al., 2010), contains data from educational technologies like intelligent tutoring systems, online courses, or educational games. This so-called "clickstream data" is often quite fine-grained recording each user interface action a student takes and corresponding system internal and external actions, such as, an indication of the correctness of the student action and a feedback message provided to the student. Each such "transaction" between student and system is stored with the time it occurred. Links to other data repositories are possible, for example, such as DiscourseDB. Since DiscourseDB exists on a remote server, an "Import" component was created which allows a user to query DiscourseDB from within Tigris. With this configuration, data can be pulled from remote repositories as data sources into any workflow. By adopting LearnSphere, an increasing number of researchers and institutions will be able to support additional learning communities, foster data curation, and facilitate community interactions.

3.2. INFRASTRUCTURE FOR SHARING, USING, AND ADAPTING ANALYTIC ROUTINES

Tigris components implement a generalized function which produces output based on some arbitrary input data and a set of user-defined options. Each component has a user interface which allows the end-user to interact with the options panel, as well as incoming and outgoing connections which allow them to receive or pass meta-data between components. To make component development easy, each is defined by a single definition file, an extensible markup language (XML) Schema Definition (XSD), which provides a structured way to define inputs, outputs, and options relevant to the component. These variable fields are then implemented by the system as a graphical interface, easily accessible to the user. The use of XSD files facilitates

the sharing of components, making specifications more transparent, and analysis more accessible.

The Tigris interface was designed to allow the user to drag-and-drop various analyses, transformations, and visualizations for processing data. Through the Tigris interface, users can create a wide range of simple to complex sequential combinations (see Figure 2). Data connections can be split and recombined, and as each component's input needs are met, its processes are carried out in a breadth-first manner, returning the output results, as well as any errors, warnings, or debugging information that was reported by the component.

The initial workflows in Tigris consist of three main component types: import objects, analysis objects, and visualization objects. The import objects are the way to bring in any form of data. Import objects can connect to a database using security protocols or be as simple as loading a text or csv file. The analysis objects are essentially new or existing analyses wrapped up for use in a workflow. The goal is to support as many types of existing analyses as possible in any programming language. We so far have objects in many languages including python, java, R, C, and C++. Visualization objects are meant to be the user facing output of a workflow. These objects can accommodate any browser-based output. Since the original implementation, we have also added additional types of components including transform objects, that can fit between various other objects to facilitate data connections between different data sources and analyses.

3.3. DISTRIBUTED ARCHITECTURE

While there is a public server that any institution can utilize, organizations can also run their own LearnSphere instance with Tigris and have complete control over their data, users, components, and workflows. If they wish, they can also connect their private instance to the LearnSphere pool, allowing meta-data and links to be shared among them. Doing so allows for institutions to follow their specific privacy, security, and IRB guidelines while maintaining a connection to the broader learning science community. In this way, the data is distributed among the various servers and institutions wishing to take advantage of their own computing resources. Even more, LearnSphere offers a web services Application Programming Interface (API) for interfacing directly with Tigris workflows and the DataShop platform.

LearnSphere instances can take advantage of a back-end cluster to support higher demand for computational resources and ensure program isolation for security. The virtual cluster can be deployed to any number of frameworks (Amazon Cloud, Google Cloud Platform, Citrix Hypervisor Express) and scale automatically with end-to-end encryption. The actual software dependencies for each component are configured, tested, and maintained in a simple Docker image, a task for which Docker is well-suited (Rad, Bhatti, Ahmadi, 2017). As new components are added, they can be thoroughly tested outside of the production system or in tandem with a quality assurance environment. This enables new components to be added or existing ones to be updated with minimal downtime. This process takes a matter of minutes and can be accomplished on a live server. The only time a component cannot be updated is if it is currently running for a workflow.

Component execution is done on a remote cluster for security reasons. Since the nodes which execute the components are separated from the web application server by the use of a web service, the risk of exposing the server to potential attacks is minimized. The way in which components' inputs, options, and outputs are defined by the XSD file also help to mitigate risk. Before component programs ever receive the program arguments (inputs and options), the arguments are tested against the component XSD file. For example, if an arbitrary option defined

as a float is given a string value, then an error is reported and the component will not execute. This error report will also be conveyed back to the executor of the component, so they can implement a fix.

3.4. COMMUNITY-DRIVEN

Tigris allows anyone to contribute their own components for inclusion in the platform. The source code for each component is publicly available via GitHub[4] where anyone can download, modify, and execute components on their own computer. Each component's dependencies can be found on its landing page within GitHub. The current code repository contains over 80 total components for data import, transformation, analysis, reporting, or visualization. Many of these tools enable rich analysis and contain additional parameters that users can change inside of a Tigris workflow for their own needs. Toward supporting a broad community of analytics contributors, we built Tigris so components can be written in any programming language and workflows can combine components flexibly such that the components in a workflow may each be implemented in a different programming language.

4. TIGRIS WORKFLOW CASE STUDIES

As LearnSphere's online workflow authoring tool, Tigris was created to make data analysis more accessible to non-learning scientists and/or users that lack programming knowledge. We target three specific use cases that are explored in each subsection:

4. Sharing components
5. Making analyses broadly available
6. Increasing longevity of older tools

4.1. ILLUSTRATING COMPONENT AND WORKFLOW AUTHORING AND SHARING

In Tigris, the data and programs that make up a workflow, as well as the parameters used to run those programs, can be stored and shared indefinitely. Tigris uses a schema architecture, where inputs for components are scripted using an XML format, to ease and speed development of components and the graphical user interface where input parameters to the component can be adjusted. More ambitious component developers can construct customized scripts so that their components have interactive outputs or visualizations. Overall Tigris is a dynamic system that catalogs data and data-driven analysis methods, provides access controls to data and resources, and therein facilitates secondary research and analysis. By providing capabilities for users to store, share, copy, and modify existing data and analyses, Tigris supports researchers toward greater reproducibility, replicability, traceability, and transparency of the analytics they report in papers. Many researchers have used LearnSphere's Tigris to create workflows corresponding with papers reporting on analytics results (e.g., Bodily, Nyland, and Wiley, 2017; Koedinger et al., 2015; Koedinger et al., 2016; Yudelson et al., 2013; Wiley, 2018; Matz et al., 2017; Beck & Gong, 2013). Some of these are new analyses while others were created based on the previously published work. Several are illustrated in the example case studies below.

This first case illustrates the development of a new component by researchers outside the original core LearnSphere team and utilizing the flexibility in Tigris for component developers

[4] https://github.com/LearnSphere/WorkflowComponents

to implement a customized script that creates more sophisticated interactive visualizations. This component is called RISE (Resource Inspection, Selection, and Enhancement) and it implements a method for analyzing open educational resources to identify areas where course content can be improved (Bodily, Nyland, and Wiley, 2017; Wiley, 2018). The RISE component operates on data from courses where both learning content and assessment outcomes are tagged with learning objectives. An example is shown in Figure 3 where the first row in the Preview table shows data about the learning objective "List the defining characteristics of biological life". The "avg_scores" column indicates students were about 86% correct on assessment items associated with this objective and the "avg_views" column indicates students engaged in associated course content about 1.0 times. Using these values, the RISE analysis creates a four-quadrant plot as shown in the upper right of Figure 3. The interactive rollover shows that the "Classify different types of atomic bonds" objective has a low outcome score (58%) despite a higher level of engagement (1.23 average views). Such information guides course developers to consider improvements in the content or assessments. Other researchers can use this RISE component in Tigris workflows to run this analysis on their course content and provide the interactive visualizations created just like they were in the Bodily et al. (2017) study. Several components, like RISE, use JavaScript libraries to create interactive visualizations right inside the workflow.

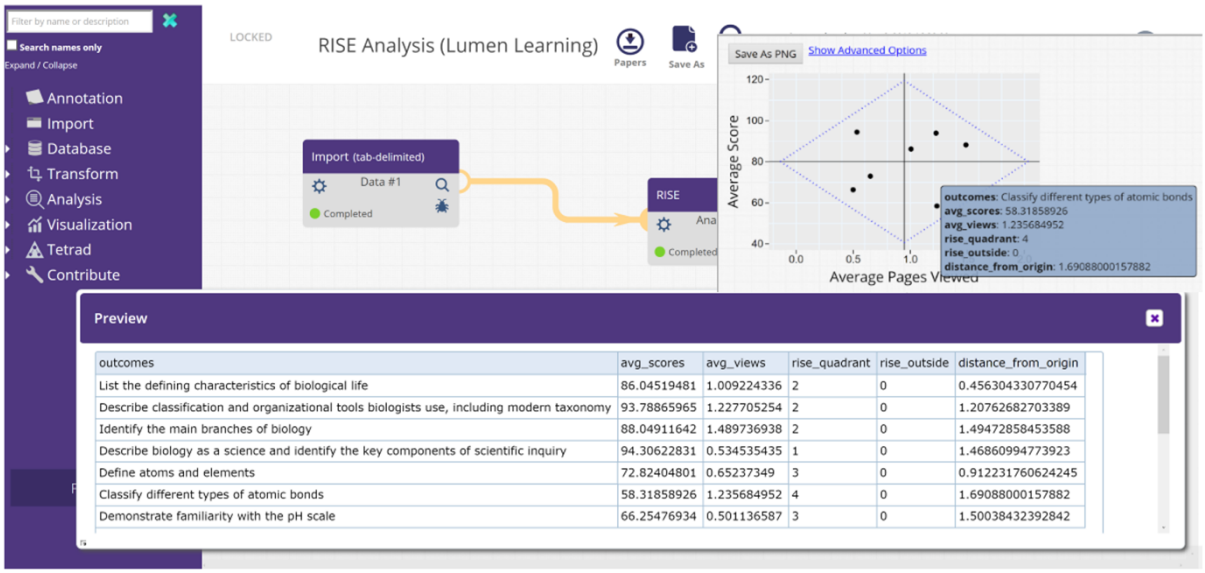
A second case study illustrates one of LearnSphere's aims to support researchers' efforts in secondary studies and derivative works. In an earlier paper, several scientists applied mixed-effect linear regression models to four MOOC datasets (Koedinger et al., 2015). Later, a workflow was created to model their experiments[5]. In a derivative work utilizing newly acquired data, once again, the causal models used in the previous study were used to explore how a student's interests and actions were related (Koedinger et al., 2018). The workflow is pictured in Figure 4 and can be accessed through LearnSphere[6]. Tigris allows users to store and share workflows, which also allows for the parameters and variables used in the analysis to be configured, along with the data used in the experiment. Workflows can be made private or shared with others, available to view, or even modify a copy of an existing workflow. In Tigris, we hope to improve the integrity, replicability, reproducibility, and traceability of all studies so these critical steps can transparently be explored and cross-examined instantly. Just as workflows can be reproduced or used in derivative works, so can the analysis components, themselves. Researchers have already contributed a large number of components, all of which work with publicly accessible and anonymized data directly available through Tigris's interface via "import" components.

While users can explore and experiment with components directly in Tigris, it is also possible to download components to run or modify them, locally. This open-source architecture promotes derivative works, and as a result, a number of related components that work with transaction-level or clickstream data have been created in Tigris in recent years. For instance, the Bayesian Knowledge Tracing (BKT) component (Yudelson et al., 2013) attempts to infer whether a student has mastered a skill given a student's prior attempts (successful or unsuccessful) to apply that skill. Similarly, the Additive Factors Model (AFM) is a specific instance of logistic regression which generalizes the Log-Linear Test Model, as well as Item Response Theory. Performance Factor Analysis (PFA) (Pavlik et al., 2009) is a generalized AFM implementation which performs logistic regression to determine the student's predicted error rate; this has also

[5] https://pslcdatashop.web.cmu.edu/LearnSphere?workflowId=398
[6] https://pslcdatashop.web.cmu.edu/LearnSphere?workflowId=1014

led to the development of a novel component, Temporal Knowledge Tracing, which combines features of PFA with ACT-R, a model of human cognition. Another, iAFM (individualized AFM), includes a per-student slope option for added functionality (Liu and Koedinger, 2017). Yet another derivative, a python implementation called PyAFM, gives the researcher the ability to interact with the model's slipping parameters (MacLellan et al., 2015). The ability of Tigris to support both replicability and reproducibility, particularly in sharing data and analysis methods, is highlighted and enhanced by the creation of these derivative components. This is an invaluable development for the authoring tool, as it complements the practice of reusing data and components. Further, the ability to download and inspect the underlying code adds a level of traceability and transparency that allows for other researchers to truly understand and follow the analysis in a definitive way that is unavailable in traditional research papers.

The need for interoperability between existing analytic tools has become desirable as new and heterogeneous learning data becomes available to researchers. For instance, the collection of MOOC data has led to datasets containing not only transaction-level data, but also discourse data from online discussion boards. As a result, natural language processing algorithms are finding a place among learning researchers who are interested in analyzing student discussions with respect to factors like motivation and engagement. In Tigris, student discourse data can be queried by way of a DiscourseDB import component which works with external DiscourseDB servers (https://discoursedb.github.io). Relatedly, a text classification component exists which utilizes the LightSIDE machine learning and text mining tool bench to detect transactivity in student discussion forums (Wen et al., 2016).



outcomes	avg_scores	avg_views	rise_quadrant	rise_outside	distance_from_origin
List the defining characteristics of biological life	86.04519481	1.009224336	2	0	0.456304330770454
Describe classification and organizational tools biologists use, including modern taxonomy	93.78865965	1.227705254	2	0	1.20762682703389
Identify the main branches of biology	88.04911642	1.489736938	2	0	1.49472858453588
Describe biology as a science and identify the key components of scientific inquiry	94.30622831	0.534535435	1	0	1.46860994773923
Define atoms and elements	72.82404801	0.65237349	3	0	0.912231760624245
Classify different types of atomic bonds	58.31858926	1.235684952	4	0	1.69088000157882
Demonstrate familiarity with the pH scale	66.25476934	0.501136587	3	0	1.50038432392842

Figure 3: The imported data is passed to the RISE component which generates a scatterplot with resource usage on the x-axis and the associated grading on the y-axis. The interactive x-y plot also allows users to drill down to get the details of a selected point.

Data visualizations play a crucial role in understanding experimental results, equal in importance to the analyses themselves. As with all LearnSphere components, visualization components can be generated using any programming language. In addition, visual components can leverage JavaScript and various libraries to produce dynamic, interactive visualizations. These visualization components leverage the Tigris workflow by drawing their data from other components. In the first case study (see Figure 3), the RISE visualization draws its data from a

table import component. In this case, the component author implemented the visualization as part of the analytic routine. Similarly in the second case (see Figure 3) causal graph visualizations are created as part of analytic routines (e.g., Search and Estimator). The visualization in the lower right of Figure 4 is an output of the Estimator component. In other situations, like the one shown in Figure 2, a more general visualization component (labeled "Learning Curves" in Figure 2) is factored apart from the analytic routines such that multiple different analytic components (e.g., AFM, Python AFM, and BKT) can feed the same visualization, either independently or in combination as is shown in Figure 2. Several general visualization components are available that operate on data in a general table format including bar charts, scatter plots, histograms, and force-directed graphs.

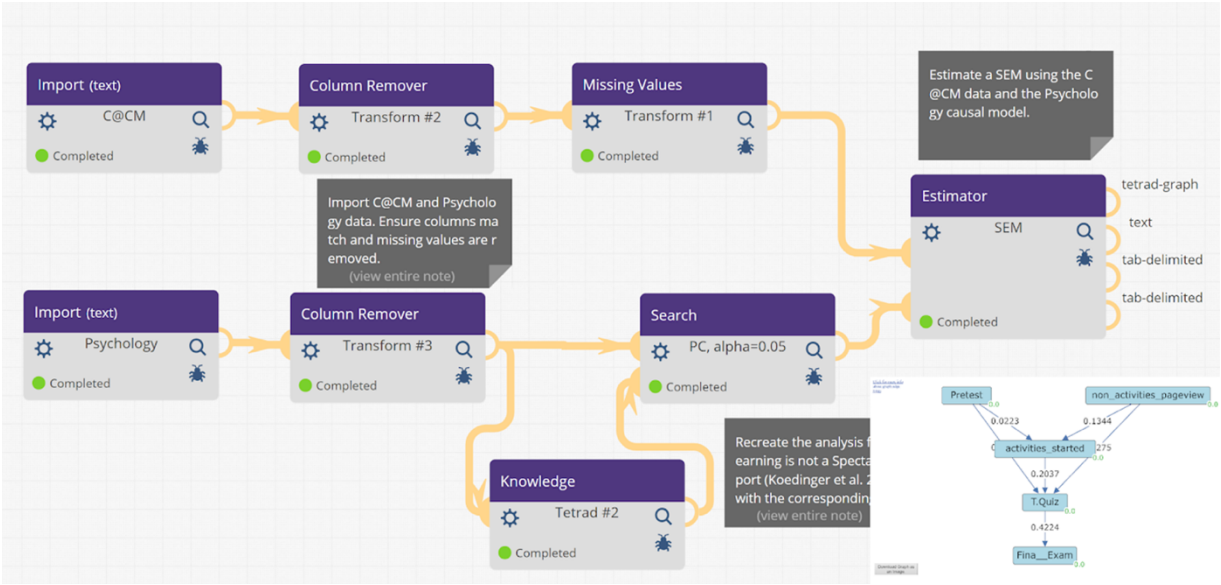


Figure 4: The workflow compares data from two courses (psychology and computing) using causal discovery algorithms available in the Tetrad Search, Knowledge, and Estimator models. The outcome of the Estimator component is pictured, bottom-right and shows the estimated parameters created for Structural Equation Modeling (SEM) using the C@CM data and the Psychology causal model.

4.2. MAKING SPECIALIZED ANALYSIS AVAILABLE BROADLY

4.2.1. Do Students Learn More from Watching, Reading, or Doing?

A number of components and workflows have been developed to analyze student use of online course content and how that use relates to learning outcomes. Our third workflow case study involves related workflows resulting from a line of research that started with an analysis of student interactions in an Introductory Psychology MOOC. This analysis explored whether students who watch more lecture videos, read more online text, or do more online formative assessments have better learning outcomes (Koedinger, Kim, Jia, McLaughlin, & Bier, 2015). The title of the paper, "Learning is Not a Spectator Sport" encapsulates the key finding: active participation, or doing, is six times more highly associated with learning outcomes than passive activities like video watching or online text reading. At least six related LearnSphere workflows have been developed as part of this research, including the three workflows represented in Figures 4-6. The workflow depicted in Figure 4 illustrates the causal inference analysis

originally conducted by Koedinger et al. (2015), and its subsequent application to a different dataset. This workflow begins by importing student process-to-output data tables, which are derived from earlier processing steps shown in Figure 5. A critical component of this workflow is the 'OLI Resource Use' step, which is a component that processes the 100s of course data instances from the Open Learning Initiative (OLI) at Carnegie Mellon University (Bier, Stamper, Moore, Siegel, & Anbar, 2023), encompassing all actions taken by all students in a course, and condenses it into summary statistics that represent each student's use of resources. These summary statistics are then combined with the imported learning outcome data for each student (e.g., final exam or grade data). The amalgamated data is not only used in the workflow in Figure 4, but also serves as input for linear modeling (specifically, the RLM Fitting component) as depicted in Figures 5 and 6. This workflow thus provides a comprehensive analysis that links student resource usage to learning outcomes.

Figure 5: Doer effect workflow that 1) imports course transaction log data of all students' actions, such as video play, online reading access, and online formative assessment question answering (on the left), 2) summarizes the number and timing of actions of each type that each student used (with the 'OLI Resource Use' component), 3) imports learner outcome data such as total quiz score, final exam, final grade (lower middle), 4) joins the student usage and outcome data (Join component), and 5) performs linear regression modeling (RLM Fitting component).

The RLM Fitting component, integral to the workflow showcased in Figure 6, features a user-friendly interface that can be accessed through the gear icon. This interface facilitates the use of the R programming language's function for linear regression modeling. The same component is central to the analysis comparing the 'doer effects' across multiple courses, a key aspect of the workflow in Figure 6. This particular analysis aligns with the findings presented in Koedinger, McLaughlin, Jia, & Bier (2016), which reveals that the 'doer effect' is not limited to the Psychology MOOC. Instead, it extends to a blended format of online course materials used in universities across various disciplines, including Information Systems, Biology, Statistics, and Psychology. This demonstrates the broad applicability and significance of the 'doer effect' in diverse educational settings that researchers can replicate using this workflow.

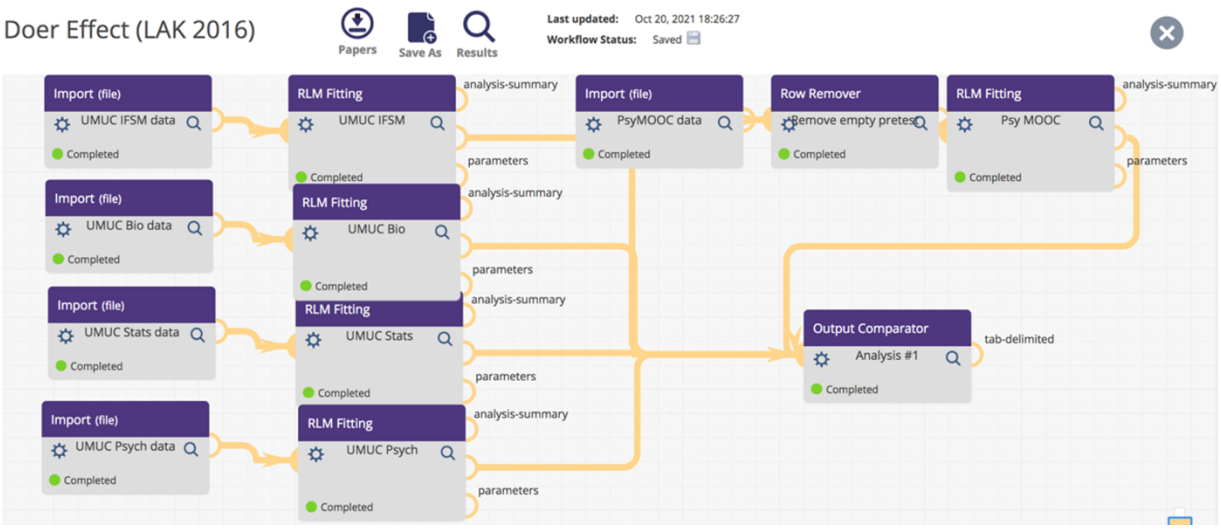


Figure 6: Doer effect workflow analysis across online course data from a MOOC (upper right) and four instead of blended use of online course materials in different university courses. The Output Comparator is a reporting component that combines results from multiple analyses (in this case from the output of 5 different RLM Fitting results).

Researchers outside of CMU have been inspired by this 'doer effect' analysis and have replicated it on their own datasets, including an unpublished MIT course investigation and two industry research investigations (Van Campenhout, Johnson, & Olsen, 2021, Van Campenhout, Jerome, Dittel, & Johnson, 2023).

4.2.2. Use case for MoFaCTS (Mobile Fact and Concept Training System)

The installation of the LearnSphere at the University of Memphis has helped integrate work there with the broader community in the form of shared datasets and tools. One significant project has been to integrate MoFaCTs development with the development of the LearnSphere project (Pavlik, Eglington, & Harrell-Williams, 2021). MoFaCTS is a generic learning tool for simple one-step problems but has a detailed and rich system for scheduling practice according to a rich learner model. MoFaCTS was developed over more than 10 years as part of the Optimal Learning Lab project at the University of Memphis (Pavlik, Olney, Banker, Eglington, & Yarbro, 2020). This integration between LearnSphere and MoFaCTs has occurred in terms of data and learner modeling standards shared between these technologies in a variety of ways.

The integration focuses on standardizing methods and formats to facilitate the immediate import and analysis of MoFaCTS, experimental, and classroom study results into DataShop within the LearnSphere framework. Learner models of students and implications for practice can be taken directly from LearnSphere and used in the MoFaCTS system to close the loop and move new research developments into production quickly. The integration of the two systems incorporates multiple LearnSphere components designed for MoFaCTS but are broadly applicable to most datasets in the DataShop. This connection benefits the University of Memphis's Optimal Learning Lab, as detailed below, but its primary advantage lies in extending the work to the broader community through the use of generalized components.

The most significant component is Generalized Knowledge Tracing (GKT), a sophisticated tool that configures logistic regression for performance analysis with advanced data feature generation. It addresses critical learning effects like forgetting, which are rarely supported in other knowledge tracing systems. Other key components include a graphing tool for visual

comparison of GKT models, an efficiency curve estimator offering pedagogical recommendations for optimal practice scheduling, and a student clustering tool to explore individual differences within the learner models.

These generalizations are promoting coherence in the research with MoFaCTS by creating a shared language for the Optimal Learning Lab research group and the broader community that uses MoFaCTS. For instance, diverse projects, such as categorical perception learning with Mandarin Chinese tones and fill-in-the-blank exercises in Anatomy and Physiology, employ the same flexible GKT methods for learner modeling and pedagogical decision-making. This unified approach enhances team communication and collaboration, as it necessitates more general expertise. Moreover, by generalizing research methods, the specific work on MoFaCTS becomes more accessible to the broader community. The use of DataShop format for tools and models facilitates easier adoption of methods developed in the MoFaCTS context, as they are already in the shared language of LearnSphere.

4.3. ADDING LONGEVITY TO EXISTING ANALYSIS TOOLS

LearnSphere's design that allows for factoring existing code bases into components enhances their maintainability and facilitates future extensions. This modular approach allows for incremental updates and improvements to be made to individual components, without affecting others. Consequently, parts of the code base can be updated independently and more efficiently. Building on this modular approach, LearnSphere streamlines collaboration and knowledge sharing among researchers, as the standardized component structure makes it easier to integrate, understand, and utilize diverse coding efforts across various projects.

The LearnSphere project aims to collaborate with researchers to not only enhance their work and offer a platform for derivative projects, but also to revitalize existing programs and code bases. An example is the Tetrad project (Spirtes et al., 1990) for causal modeling, which includes a suite of data exploration and discovery algorithms and is publicly accessible[7]. Components integrated from Tetrad include a Bayesian classifier and a Search component for identifying causal explanations in data, which can interface directly with a Knowledge and Estimator component to refine search parameters. The library also features Tetrad's dataset Simulator, a robust Regression component supporting linear and logistic regression, and various methods for data manipulation prior to analysis. While Tetrad is already valuable as a standalone tool, its integration within a flexible environment like Tigris opens up endless opportunities for enhanced interoperability and the creation of derivative works. It allows a LearnSphere user to quickly perform analysis using Tetrad with the plethora of datasets available on DataShop. A Tetrad workflow example is illustrated in Figure 4.

5. LEARNSPHERE USAGE INDICATES WIDE ADOPTION

One of the primary goals of the LearnSphere project has been to build a large and diverse community for sharing learning data and analytics. Through outreach at conferences, workshops, and the annual LearnLab Summer School, we have connected to the communities of researchers who can benefit from a platform of shared learning analytics. We have achieved this goal, and the community continues to grow.

[7] https://www.cmu.edu/dietrich/philosophy/tetrad

As of January 2024, there are more than 2,000 existing workflows in Tigris created by over 600 users. Of these, 377 workflows are public and can be viewed by anyone. The scale of import data is represented by over 23 million transactions from 87 unique DataShop datasets and over 1,900 additional imported data files. The number of distinct users continues to grow year to year.

Table 1: Overview of components in LearnSphere's Tigris workflow tool.

Component Type	Count	# of Workflows using
Analysis	36	1455
Tetrad	8	178
Transform	32	671
Visualization	12	822

Using Tigris in learning analytics experiments, researchers have made use of over 700 data transformations, 900 Tetrad components, 3,900 learning analytics output files, and 29,000 visualization files. Since its inception, the Tigris community continues to grow each year, adding users, components as well as shareable workflow and analyses.

Table 2: Tigris Usage by number of workflows and users per year.

Year	# of Workflows	Cumulative Users	Total Users
2016	42	28	28
2017	283	82	124
2018	725	185	315
2019	1424	355	679
2020	1804	452	874
2021	2066	506	1003
2022	2332	559	1107
2023	2538	617	1214

6. CONCLUSION

We created LearnSphere to transform scientific discovery and innovation in education through a scalable data infrastructure that supports many contributing disciplines within the learning sciences. These disciplines include cognitive, social, and motivational psychology, discipline-based education research in Physics, Chemistry, Computer Science, and educational technology innovation research (e.g., intelligent tutoring, dialogue systems, MOOCs). LearnSphere makes data and data analytics available through a web-based application where computer scientists can contribute sophisticated analytic methods and learning scientists can use those methods without additional programming. LearnSphere allows for entire workflows to be shared with data allowing for complete reproducibility of previous work. Workflows can also be shared without data allowing for replicability of studies with new data.

As part of the LearnSphere, we developed a web-based workflow authoring tool, Tigris, to support flexible sharing and easy non-programmer integration of a wide variety of data import, transformation, analytic and reporting functions. With the unique ability to develop analytical

components and use them in workflows that can be shared, LearnSphere allows for models that are the basis of published work to be inspected, replicated or modified by peers. Many researchers have, and we hope will continue, to benefit from these shared workflows, allowing non-programmers access to analytics that were previously only available through computer programming. While at the beginning of the project, we recognized that to be successful, we would need a large number of components and workflows, we now have reached a state where many researchers can access and use the existing tools and integrate LearnSphere into their own research pipelines.

A key part of LearnSphere's success, as well as continued use and growth of DataShop, has been proactive efforts to build a user community around it. Over the span of LearnSphere's NSF funded development from 2015 to 2021, we ran 15 workshops with about 400 participants including nine 1-2 day conference workshops and six week-long summer schools on educational data mining. LearnSphere now has over 15,000 unique user logins. Users have created more than 80 analytic components and over 1,200 workflows that configure these components in novel ways. These contributions come from university and industry users well beyond the university grant participants. LearnSphere also provides for educational data storage and sharing through its direct connection with DataShop. As a reflection of success in building a user community, the number of datasets stored in LearnSphere has shown a 7x increase, from about 600 Datasets at the start of the LearnSphere project in 2015 to over 4300 datasets as of January 2024.

LearnSphere has facilitated discoveries about learning, as represented by the over 120 relevant papers published by the team as well as by publications of other researchers and data-driven educational development projects in industry. For example, one series of papers benefited from LearnSphere capabilities to integrate previously isolated data sources including a) computer-based tutor interaction data, b) MOOC resource use with videos and discussion boards, and c) learning outcome data from quizzes, final exams, final grades, and final projects. An analysis across four online courses involving over five million interactions from more than 12,000 students revealed a striking discovery: Active learning activities (e.g., answering questions with feedback) are associated with 6x greater learning outcomes than passive learning activities (e.g., lecture watching or text reading). Other discoveries include quality discussion board posts predict learning outcomes, second language students show higher positive associations with learning outcomes from online text reading than from video watching (but, as with all students, even higher from active learning), causal models inferences consistently suggest more doing yields more learning across multiple course datasets, and better project outcomes are associated with doing more learning-goal aligned formative assessment questions.

ACKNOWLEDGMENTS

This research was funded through the National Science Foundation award 1443068.

REFERENCES

2012. *What Works Clearinghouse.* Internet site: http://ies.ed.gov/ncee/wwc.

AMBROSE, S. A., BRIDGES, M. W., DIPIETRO, M., LOVETT, M. C., AND NORMAN, M. K. 2010. *How learning works: Seven research-based principles for smart teaching*. John Wiley & Sons.

BAKER, R. S. D., DUVAL, E., STAMPER, J., WILEY, D., AND SHUM, S. B. 2012. Educational data mining meets learning analytics. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK 2012)*. ACM, 20-20.

BECK, J.E. AND GONG, Y., 2013. Wheel-spinning: Students who fail to master a skill. In In *Artificial Intelligence in Education: 16th International Conference, AIED 2013*, Memphis, TN, USA, H. C. Lane, K. Yacef, J. Mostow, and P. I. Pavlik, Eds. Springer Berlin Heidelberg, 431-440.

BIER, NORMAN, STAMPER, JOHN, MOORE, STEVEN, SIEGEL, DARREN, AND ANBAR, ARIEL. 2023. OLI Torus: a next-generation, open platform for adaptive courseware development, delivery, and research. In *Companion Proceedings 13th International Conference on Learning Analytics & Knowledge (LAK 2023)*, ACM, 57-60.

BODILY, R., NYLAND, R., AND WILEY, D. 2017. The RISE Framework: Using Learning Analytics to Automatically Identify Open Educational Resources for Continuous Improvement. *International Review of Research on Distance and Open Learning*, Érudit, 18(2), 103-122.

DATAVYU TEAM. 2014. Datavyu: A Video Coding Tool. *Databrary Project,* New York University. URL http://datavyu.org.

FIACCO, J., COTOS, E. AND ROSÉ, C., 2019, March. Towards enabling feedback on rhetorical structure with neural sequence models. In *Proceedings of the 9th international conference on learning analytics & knowledge (LAK 2019)*, ACM, 310-319.

FIACCO, J. AND ROSÉ, C., 2018, June. Towards domain general detection of transactive knowledge building behavior. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, ACM, 1-11.

GARDNER, J., BROOKS, C., ANDRES, J.M. AND BAKER, R.S., 2018, December. MORF: A framework for predictive modeling and replication at scale with privacy-restricted MOOC data. In *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 3235-3244.

JIANG, W., PARDOS, Z. A., AND WEI, Q. 2019. Goal-based course recommendation. In *Proceedings of the 9th international conference on learning analytics & knowledge (LAK 2019*), ACM, 36-45.

JO, Y., TOMAR, G.S., FERSCHKE, O., ROSE, C.P., AND GAESEVIC, D. 2016. Pipeline for expediting learning analytics and student support from data in social learning. In *Proceedings of Educational Data Mining (EDM 2016),* T. Barnes, M., Chi, and M. Feng, Eds. International Educational Data Mining Society (IEDMS), 59-62.

JO, Y. AND ROSÉ, C. P. 2015. Time Series Analysis of Nursing Notes for Mortality Prediction via State Transition Topic Models, In *Proceedings of The 24th ACM International Conference on Information and Knowledge Management (CIKM 2015)*, ACM, 1171-1180.

KOEDINGER, K., BAKER, R., CUNNINGHAM, K., SKOGSHOLM, A., LEBER, B., AND STAMPER, J. 2010. A Data Repository for the EDM community: The PSLC DataShop. In *Handbook of Educational Data Mining,* C. Romero, S., Ventura, M., Pechenizkiy, and R.S.J.d. Baker, Eds. Boca Raton, FL: CRC Press, 43-56.

KOEDINGER, K. R., CORBETT, A. T., AND PERFETTI, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*. Wiley Online Library, 36(5), 757-798.

KOEDINGER, K. R., BRUNSKILL, E., BAKER, R. S., MCLAUGHLIN, E. A., AND STAMPER, J. 2013. New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, AAAI, 34(3), 27-41.

KOEDINGER, K.R., KIM, J., JIA, J., MCLAUGHLIN, E.A., AND BIER, N.L. 2015. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the Second ACM Conference on Learning at Scale*, ACM, 111-12.

KOEDINGER, K.R., MCLAUGHLIN, E. A., JIA, J. Z., AND BIER, N. L. 2016. Is the Doer Effect a causal relationship? How can we tell and why it's important. In *Proceedings of the Sixth International Conference on Learning, Analytics and Knowledge (LAK 2016)*, ACM, 388-397.

KOEDINGER, K. R., SCHEINES, R., AND SCHALDENBRAND, P. 2018. Is the doer effect robust across multiple data sets? In *Proceedings of the 11th International Conference on Educational Data Mining*, K. Boyer and M. Yudelson, Eds. International Educational Data Mining Society (IEDMS), 369–375.

LIU, R., AND KOEDINGER, K. R. 2017. Towards Reliable and Valid Measurement of Individualized Student Parameters. In *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2017)*, X. Hu, T. Barnes, A. Hershkovitz, and L. Paquette, Eds. International Educational Data Mining Society (IEDMS), 135-142.

LOHSE, J. J., MCMANUS, C. A., AND JOYNER, D. A. 2019. Surveying the MOOC data set universe. In *2019 IEEE Learning With MOOCS (LWMOOCS)*, Meier, R. et al. Eds. IEEE, Madrid, Spain, 159-164.

MACLELLAN, C.J., LIU, R., AND KOEDINGER, K.R. 2015. Accounting for Slipping and Other False Negatives in Logistic Models of Student Learning. In *Proceedings of the 8th International Conference on Educational Data Mining*. O.C. Santos, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, M. Desmarais, Eds. International Educational Data Mining Society (IEDMS), 53-60.

MATZ, R. L., KOESTER, B. P., FIORINI, S., GROM, G., SHEPARD, L., STANGOR, C. G., ... AND MCKAY, T. A. 2017. Patterns of gendered performance differences in large introductory courses at five research universities. *Aera Open*, 3(4), 2332858417743754.

MAYFIELD, E. AND ROSÉ, C. P. (2013). LightSIDE: Open Source Machine Learning for Text Accessible to Non-Experts, Handbook of Automated Essay Grading. Shermis, M. and Burstein. J. Eds. Routledge, New York, 124-135.

O'REILLY, U.M., AND VEERAMACHANENI, K. 2014. Technology for Mining the Big Data of MOOCs. *Research & Practice in Assessment*, 9(2), 29-37.

PAVLIK, P., CEN, H., AND KOEDINGER, K. 2009. Performance Factors Analysis - A New Alternative to Knowledge Tracing. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED 2009)*, V. Dimitrova, R.Mizoguchi, B. du Boulay, and A. C. Graesser, Eds. IOS Press, 531–538.

PAVLIK JR, P. I., OLNEY, A. M., BANKER, A., EGLINGTON, L., AND YARBRO, J. 2020. The Mobile Fact and Concept Textbook System (MoFaCTS). In *Proceedings of the Second International Workshop on Intelligent Textbooks 2020 co-located with 21st International Conference on Artificial Intelligence in Education (AIED 2020)*, S. Sosnovsky, P. Brusilovsky, R. Baraniuk, and A. Lan, Eds. CEUR Vol. 2674, 35-49.

PAVLIK, P. I., EGLINGTON, L. G., AND HARRELL-WILLIAMS, L. M. 2021. Logistic knowledge tracing: A constrained framework for learner modeling. *IEEE Transactions on Learning Technologies,* IEEE, 14(5), 624-639.

PAVLIK JR., P. I., KELLY, C., AND MAASS, J. K. 2016. Using the mobile fact and concept training system (MoFaCTS). In *Proceedings of the 13th International Conference on Intelligent Tutoring Systems*. A. Micarelli and J. Stamper, Eds. Springer, 247-253.

RAD, B. B., BHATTI, H. J., AND AHMADI, M. 2017. An introduction to docker and analysis of its performance. International Journal of Computer Science and Network Security (IJCSNS), 17(3), 228.

ROSÉ, C. P. 2017. Expediting the cycle of data to intervention, *Learning: Research and Practice 3(1), special issue on Learning Analytics*, Taylor and Francis, 59-62.

ROSÉ, C. P. AND FERSCHKE, O. 2016. Technology Support for Discussion Based Learning: From Computer Supported Collaborative Learning to the Future of Massive Open Online Courses. *International Journal of Artificial Intelligence in Education, 25th Anniversary Edition*, Springer, volume 26(2). 660-678.

ROSÉ, C. P., WANG, Y.C., CUI, Y., ARGUELLO, J., STEGMANN, K., WEINBERGER, A., AND FISCHER, F. 2008. Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning, *International Journal of Computer Supported Collaborative Learning,* Springer, 3(3), 237-271.

SANKARANARAYANAN, S., DASHTI, C., BOGART, C., WANG, X., MARSHALL AN, CLARENCE NGOH, MICHAEL HILTON, SAKR, M., AND ROSÉ, C. 2019. Online Mob Programming: Bridging the 21st Century Workplace and the Classroom, In *Proceedings of Computer-Supported Collaborative Learning. Vol.2*, K. Lund, G. P. Niccolai, E. Lavoué, C. Hmelo-Silver, G. Gweo, and M. Baker, Eds. Lyon, FR, 855-856.

SANKARANARAYANAN, S., DASHTI, C., BOGART, C., WANG, X., SAKR, M., AND ROSÉ, C. 2018. When Optimal Team Formation is a Choice - Self-Selection versus Intelligent Team Formation Strategies in a Large Online Project-Based Course, In *the 19th International Conference on Artificial Intelligence in Education (AIED 2018)*, C. Penstein Rosé, R. Martínez Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. M. McLaren, and B. du Boulay, Eds. Springer International Publishing. London, UK. Part I. 518-531.

SELENT, D., PATIKORN, T., AND HEFFERNAN, N. 2016. Assistments dataset from multiple randomized controlled experiments. In *Proceedings of the Third ACM Conference on Learning@ Scale*, ACM, 181-184.

SINATRA, A.M. 2022. Proceedings of the 10th Annual GIFT Users Symposium. Orlando, FL: US Army Combat Capabilities Development Command - Soldier Center. ISBN 978-0-9977258-2-7. Available at: https://gifttutoring.org/documents/

SOTTILARE, R. A., LONG, R. A., AND GOLDBERG, B. S. 2017. Enhancing the Experience Application Program Interface (xAPI) to improve domain competency modeling for adaptive instruction. In *Proceedings of the Fourth ACM Conference on Learning@ Scale*, ACM, 265-268.

SPIRTES, P., GLYMOUR, C., SCHEINES, R. AND RAMSEY, J., 1990. The TETRAD project. *Acting and Reflecting*. 183-207.

STAMPER, J., AND PARDOS, Z. A. 2016. The 2010 KDD Cup Competition Dataset: Engaging the machine learning community in predictive learning analytics. *Journal of Learning Analytics,* 3(2), 312-316.

VAN CAMPENHOUT, R., JOHNSON, B.G., AND OLSEN, J. 2021. The Doer Effect: Replicating Findings that Doing Causes Learning. *The Thirteenth International Conference on Mobile, Hybrid, and On-line Learning,* Think Mind Digital Library, 1-6.

VAN CAMPENHOUT, R., JEROME, B., DITTEL, J. S., AND JOHNSON, B. G. 2023. The Doer Effect at Scale: Investigating Correlation and Causation Across Seven Courses. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, ACM. 357-365.

VEERAMACHANENI, K., DERNONCOURT, F., TAYLOR, C., PARDOS, Z.A., AND O'REILLY, U.M. 2015. Moocdb: Developing data standards for MOOC data science. *MOOCShop at Artificial Intelligence in Education, 2015a.* URL http://ceur-ws.org/Vol-1009/0104.pdf, Madrid, Spain. 17-24.

WANG, X.., YANG, D., WEN, M., KOEDINGER, K. R., AND ROSÉ, C. P. 2015. Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains, *The 8th International Conference on Educational Data Mining*, O.C. Santos, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, M. Desmarais, Eds. International Educational Data Mining Society (IEDMS), 226-233.

WARLAUMONT, A. S., VANDAM, M., BERGELSON, E., AND CRISTIA, A. 2017. HomeBank: A Repository for Long-Form Real-World Audio Recordings of Children. In *Proceedings of INTERSPEECH 2017*, Stockholm, Sweden, F. Lacerda, Ed. International Speech Communication Association (ISCA), 815-816.

WILEY, D. 2018. RISE: An R package for RISE analysis. *Journal of Open Source Software*. The Open Journal, 3(28), 846, https://doi.org/10.21105/joss.00846.

YANG, D., WEN, M., KUMAR, A., XING, E., AND ROSÉ, C. P. 2014. Towards an Integration of Text and Graph Clustering Methods as a Lens for Studying Social Interaction in MOOCs. In *The International Review of Research in Open and Distance Learning* 15(5), Special Issue on the MOOC Research Initiative. 215-234.

YUDELSON, M., KOEDINGER, K.R., AND GORDON, G.J. 2013. Individualized Bayesian Knowledge Tracing Models. In *Artificial Intelligence in Education: 16th International Conference, AIED 2013*, H. C. Lane, K. Yacef, J. Mostow, and P. I. Pavlik, Eds. Springer Berlin Heidelberg. Memphis, TN, USA, 171-180.