JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 6(2), 107-119. http://dx.doi.org/10.18608/jla.2019.62.8

Rearchitecting Data for Researchers: A Collaborative Model for Enabling Institutional Learning Analytics in Higher Education

Steven Lonn[1], Benjamin Koester[2]

Abstract

This article presents the case of the Learning Analytics Architecture (LARC) dataset, a collaborative effort at the University of Michigan to develop a common and extensible tool using administrative data and designed primarily for learning analytics researchers to investigate enrolled students' academic careers, demographics, and related teaching and learning outcomes. The institutional context prior to the creation of the dataset and the rationale, design, development, and maintenance involved in creating LARC are all detailed. Also discussed are the procedures for access, documentation, and ensuring the continued usability and relevance of the dataset for a growing learning analytics and data science research community. The authors conclude the case description with recommendations for institutions seeking to replicate this effort.

Notes for Practice

- Administrative data in higher education is not typically ready "out of the box" for learning analytics researchers. How can institutions leverage their own resources and knowledge of their data to reduce time-intensive and repetitive data-cleaning efforts to make the most of their own data?
- This article presents the case study of the LARC dataset, as well as the decision points and process that the committee members undertook to build and maintain the data. This kind of effort remains a major hurdle for many higher education institutions.
- A framework is provided for other institutions interested in building a similar dataset for learning analytics research. Six areas of consideration are discussed: (1) facilitating partnerships across departments, roles, and levels; (2) level setting and arriving at a common level of understanding among team members; (3) obtaining buy-in from relevant stakeholders and data stewards; (4) designing for the needs of the specific institutional context; (5) utilizing national and international standards whenever possible; and (6) understanding the institutional landscape of learning analytics and designing datasets accordingly.

Keywords

Dataset, collaboration, higher education, administrative data

Submitted: 16.07.2018 — **Accepted:** 30.04.2019 — **Published:** 05.08.2019

Corresponding author [1]Email: slonn@umich.edu. Address: Office of Enrollment Management, The University of Michigan, 515 E. Jefferson St., Suite 2300, Ann Arbor, MI, USA 48109-1316

[2] Address: Department of Physics, The University of Michigan, 450 Church St., Ann Arbor, MI, USA 48109-1040

1. Introduction

Postsecondary educational institutions have a wide variety of needs for student data. Otherwise known as "administrative data," this information is used by institutions to carry out key functions such as admissions, financial aid, course registration, student employment, graduation, and development. The operational data systems built to run these functions often apply business intelligence techniques to allow administrators and other decision makers to gather and analyze key metrics for reporting, budgeting, marketing, and other administrative needs. This "practice of mining institutional data to produce 'actionable intelligence'" is commonly referred to as academic analytics and is most often used at institutional, regional, and national or international levels of analysis (Campbell, DeBlois, & Oblinger, 2007). Academic analytics is therefore commonly interested in information at a high, aggregate level of detail that summarizes trends and areas in which administrators can effect change (e.g., student retention). Administrative data is therefore organized, often in large relational databases, to support

the types of reporting and analysis typically conducted by institutional research or business intelligence professionals.

Learning analytics, by contrast, centres on understanding and optimizing the learning process, including the relationships between the learner, the content, the institution, and the educator (Long & Siemens, 2011). The unit of "actionable intelligence" within learning analytics analyses is therefore typically smaller or more fine grained (e.g., segments of learners, specific courses, or departments) than those in academic analytics. While learning analytics research focused on the learning processes in higher education may utilize many of the same general data sources as academic analytics, the tools and techniques can vary greatly. For example, academic analytics often relies on reports and basic visualizations driven by business intelligence tools, while learning analytics commonly uses statistical and network analysis software and sophisticated predictive modelling techniques.

2. Bridging the Data Divide between Academic and Learning Analytics

Prinsloo, Slade, & Khalil (2018) argue that there are various issues and policies inherent in the differences between academic and learning analytics that impact the collection, analysis, and use of student data at the micro, meso, and macro levels of an institution. Learning analytics can therefore demand different types of data, and in different formats, than those commonly employed for academic analytics. While, in practice, academic and learning analytics techniques may be merging or cross-pollinating at certain institutions, the systems and data formats designed for academic and learning analytics remain quite different and are optimized for different purposes.

Reformatting administrative data for learning analytics research can often be cumbersome, time-intensive, confusing, and repetitive across multiple researchers and projects, even within a single institution. For example, the most popular student information systems in use in higher education today (e.g., PeopleSoft, Banner, Workday) typically store student information to facilitate the efficiency of the software. This results in several database tables with multiple rows per student to deliver fine-grained information, such as a student's current major(s). Learning analytics researchers, however, often need simplified versions of this information and in a format more easily interpreted by statistical and machine learning software packages.

This article describes an effort at one institution of higher education, the University of Michigan (U-M), to begin bridging this divide by creating a new dataset from institutional administrative data that could be a common and extensible foundation, designed primarily for faculty, student, and staff researchers. Since its official launch in late 2016, the resulting Learning Analytics Architecture (LARC) dataset has emerged as a model for how institutions can serve both academic and learning analytics priorities through committed collaboration across constituencies. The authors will describe the rationale, design, development, and continued maintenance involved in creating the LARC dataset; point out several examples of how researchers have conducted new analyses made possible by LARC; and conclude with several recommendations for other institutions seeking to create their own version of this dataset.

3. Building LARC: A Collaborative Enterprise

3.1. Institutional Context and Rationale

The rationale for the LARC dataset grew from learning analytics researchers' prior experience of working with administrative data. As detailed by Lonn, McKay, & Teasley (2017), U-M had built a nascent learning analytics community with many research projects being launched to investigate a variety of questions related to student learning and academic success. For example, one project revealed that first-year engineering students who received a 4 on their Math AB Advanced Placement (AP) exam only received an A or B grade 50% of the time when enrolling in Calculus II their first term (Nam, Lonn, Brown, Davis, & Koch, 2014). A different project analyzed the impact of a general chemistry prerequisite on later student achievement and progression in subsequent chemistry courses (Shultz, Winschel, & Gottfried, 2015). A significant number of these projects each independently obtained, transformed, and analyzed institutional student data, replicating such work for each subsequent project. Simultaneously, Benjamin Koester was a key staff member coordinating a multi-institutional effort to examine similar student- and course-level trends. This project struggled to find ways to structure student data in comparable ways and to best conduct parallel analyses of the data, once normalized. We thus began to strategize that by collaboratively constructing a shared dataset, we could eliminate the need for such replication in the future and potentially create a model for other institutions to emulate.

3.1.1. Barriers to Administrative Data for Researchers

Based on the experiences of working with administrative data across several research projects, several themes help describe the kinds of barriers learning analytics researchers experienced while working with administrative data prior to the launch of the LARC dataset. First, it was not always easy to locate or uncover particular data elements or areas of research interest. For example, trying to identify an undergraduate student's intended major began with looking at information from a survey that

students answered at orientation and then investigating information from the application and weighing that data source against an institutional survey for the liberal arts college.

A second theme was that the data was not centralized. This is a common barrier for large research institutions where individual departments or units may control institutional data that may or may not be linked to other administrative data elements. For example, the data for U-M's learning management system was housed in a separate system than course data — these two data sets were governed by two different data stewards, both of whom a researcher would need to get approval from if they needed to combine these data sources to conduct their learning analytics project.

A third theme was that administrative data structures are often complex and can be inconsistent across domains. For example, a "flag" variable in the registrar's dataset indicated whether a student was an "underrepresented minority" based on their race or ethnicity. However, the admissions dataset did not include this flag for a variety of historical reasons. Furthermore, depending on who was using this flag, students who were citizens of another country were included or excluded from the count, regardless of their flagged status. It is important to note, however, that researchers at U-M were not flying completely blind. The Information and Technology Services (ITS) unit regularly updated and publicly published data dictionary documents that detailed the data tables, elements, descriptions, and value examples for each dataset centrally maintained at the institution (https://its.umich.edu/enterprise/administrative-systems/data-warehouse/data-areas/student).

Finally, a fourth theme was that it was often difficult to interpret or understand business processes that defined values in administrative data but were not necessarily documented for researchers. For example, if a student changed majors from one school or college at U-M, the new school or college could choose whether to accept the grades from the originating school or college. If they chose not to accept the grades, students would still have credit toward graduation, but their grade point average (GPA) would effectively reset. Thus, it was nearly impossible for researchers to understand why a second-year student would have a 3.8 GPA (on a 4.0 scale) one semester and a 0.0 GPA at the start of the next semester.

The researchers who strategized for a new common dataset for learning analytics sought not only to reduce these barriers for their own research but also to create a common good for future researchers.

3.2. Constitution and Composition of the Collaboration

The initial call to discuss a new "data structure strategy for learning analytics" was sent in December 2014. Recognizing that in order to successfully construct a learning analytics dataset derived from administrative data they would need a variety of expertise, the organizers invited a wide variety of stakeholders to the initial planning meeting, including faculty, graduate students, staff domain experts, and IT business intelligence professionals. Arnold, Lonn, & Pistilli (2014) posit that a variety of academic and support staff as well as students of all levels need to be involved in creating a culture of awareness and acceptance for learning analytics, including efforts such as constructing a new dataset.

As a first step to defining the scope of the project and the correct project team membership to successfully implement that scope, the attendees from the planning meeting were tasked with listing their most common questions or queries of student administrative data. The questions were then categorized into general data areas (e.g., admissions, courses, human resources). For example, one question listed was, "For a student who received grade X in course Y, what is her/his grade likelihood in course Z?" Another question, more focused on data from admissions, asked, "How well does high school X prepare their students compared to other high schools?"

Following this exercise, the group members met again, discussed the range of questions and queries listed, and agreed on a general scope and strategy for moving forward. Given this refined scope, group members also suggested additional or alternative individuals to join the project team. By the end of January 2015, the membership was solidified and began work in earnest on defining what, exactly, U-M learning analytics researchers needed from administrative data.

3.3. Guiding Principles: Building for an Emergent Domain

As the work of the LARC team began in earnest, three tenets emerged as central pillars to guide the development of the dataset.

First, the researcher would remain the central persona for whom the dataset would be designed. This immediately led to design decisions that could constrain other potential users of the dataset. For example, the data would be stored as "flat" as possible, meaning as few rows as possible to more easily import the data into a statistical analysis software package such as R, STATA, or SPSS. Other potential uses, such as a web-based reporting tool, might perform better with a more traditional, vertically constructed, relational database but would not be considered since they deviated from the central persona's needs. The LARC dataset would also not have the same administrative business intelligence front-end tools as other administrative datasets since researchers did not commonly utilize this tool, which alleviated many of the constraints that other administrative datasets must adhere to.

The second guiding principle reflected the emergent nature of learning analytics. As a field of study, learning analytics is relatively new. The standard definition was published at the First International Conference on Learning Analytics and

Knowledge in 2011 (https://tekri.athabascau.ca/analytics/). As such, the members of the LARC committee were hesitant to declare what the bounds of learning analytics research questions might be. However, the members were also reluctant to design a dataset for every possible groundbreaking research project. This increasingly common challenge to database organization for research was acknowledged by Stonebraker, Frew, Gardels, & Meredith (1993) in the Earth Sciences and later by Szalay, Kunszt, Thakar, Gray, Slutz, & Brunner (2000) in the Sloan Digital Sky Survey (York et al., 2000). The latter advocated a "20 queries" solution to managing multi-terabyte astrophysics databases. The 20 queries were intended in spirit to address basic questions of regular utility to the astronomy research community. Naturally, these queries did not cover the set of innumerable potential complex research questions that might be posed. Instead, they "characterize the kinds of questions astronomers are likely to ask" (Szalay et al., 2000).

Unlike earth science or astronomy, learning analytics is a nascent field where researchers are still developing core questions to inform their research. Thus, the compromise solution was to design the dataset for about 80% of the local learning analytics research community's core questions and allow for extensibility for those researchers on the leading edge so that those who really needed the uncommon administrative data would be able to combine additional administrative data with the core LARC dataset. The LARC committee's experience suggested that four out of every five questions could be considered simple queries. For database architects, this rule of thumb also minimized resource-intensive over-engineering to address every possible research question.

The third guiding principle was strongly linked to the second. The committee and, most important, the ITS members and their supervisors committed to refining and/or enhancing the LARC dataset on a regular basis. Each refinement or enhancement would have to meet the 80% threshold to be considered, and that 80% would be a moving target based on the development of the field of learning analytics. This principle is important for keeping the dataset relevant and "lively," meaning still working and of inherent use (Lupton, 2016) versus being broken or no longer useful for the community in which it was created (Pink, Ruckenstein, Willim, & Duque, 2018).

3.4. Directives for Data Development

Operating under the tenets described above, the LARC committee began the intensive work of identifying data elements that fit into the 80% solution and helped answer many of the initial questions listed by the committee members on the initial brainstorming document. For each data element or set of elements, the four directives in the next four subsections were used to guide the development.

3.4.1. Flatten and Combine

As mentioned in the first principle in section 3.3, the LARC dataset was designed to be as flat as possible and combine as many similar data elements as possible to best facilitate data translation to statistical software. This meant that LARC could be delivered in as few as five tables: Student Info (one row per student), which contains one-time data such as demographics, test scores, and degrees; Student Term Info (one row per student per term), which contains one-time-per-term data such as class standing, GPA, and major(s); Student Course Info (one row per student per course), which contains course-specific data such as course subject and number, course grade, and classroom details; Student Term Transfer (one row per transferred course), which contains details for each course awarded transfer credit from externally attended courses, test credit, or other credit-equivalent programs; and Personally Identifiable Information (one row per individual), which contains the institutional identification number, email address, and birthdate for each student in the LARC dataset.

3.4.2. Simplify

By design, there is a lot of redundancy and nuance in student administrative data. For example, the administrative data captures a student's permanent address, their current address, their campus employment address (if any), and a history of all prior addresses. For the purposes of the LARC dataset, the need for address data boiled down to knowing where a student was "from" — thus, researchers needed only the zip or postal code and country from the student's first known permanent address.

3.4.3. Normalize

Administrative data is often built over time and also often carries values directly from a source system. This results in inconsistent values, where a positive flag could have "y" and "n" values, "yes" and "no" values, or random numerical values (e.g., 500 or 600). For the LARC dataset, the values 0 and 1 would be used as much as possible for valid values, and a null value would be entered for any missing data. This approach, often referred to as "dummy coding" by researchers, would simplify interpretation and eliminate much of the need for data cleaning before statistical analysis.

3.4.4. Autocalculate

This final directive was for values or common calculations that learning analytics researchers found themselves commonly

using in their research projects. In the initial design, this included several GPA-related values, such as the prior-term GPA, and the students' term and cumulative GPA if the current course being examined was removed from the GPA calculation. In a subsequent update to the LARC dataset, the time between enrolment at a prior institution (e.g., high school or previous two- or four-year college) and enrolment at U-M has been autocalculated to more easily identify students with a gap in their higher education career.

3.5. Dataset Development Procedures

3.5.1. A Brief Moment in Time: Developing for Snapshots

Different from most administrative datasets, except perhaps those used for annual reporting, the LARC dataset was designed to be regularly updated but capture the entire known history of students at the institution within each update. The dataset would be updated three times a year; each update would be timed to follow the dominant semesters of the academic year (Fall, Winter, and Summer) but allow time for grades and degrees to be posted to the systems of record. The timing was thus paired with the following semester's census date, three weeks into the following term. This practice emulated the update strategy for the Sloan Digital Sky Survey dataset (https://www.sdss.org), which regularly releases updates to the known universe of astronomical objects but retains previous releases for researchers who may have begun work on a previous iteration of the survey (York et al., 2000).

The scenario posited by the project team was that a researcher could begin their study using a release of the LARC dataset following the Winter 2016 semester. That research project continues for a year. During that year, 14 students in an introductory economics course had their Winter 2016 grades changed from "incomplete" to A and B letter grades, thus affecting their term and cumulative GPAs. The LARC dataset would update and reflect these new grades following the Summer 2016 term, but the researcher could continue using the Winter 2016 snapshot unhindered. The LARC team operated under the perspective that neither the Winter 2016 nor the Summer 2016 data was "right" or "wrong"; each simply reflected the entire known history at the time of that snapshot.

3.5.2. Using Google Sheets for Data Elements and Rapid Development

The majority of the LARC dataset development was conducted using Google Sheets (Figure 1) and collectively discussing data elements or sets of data elements using the three guiding principles described in section 3.3 and the four directives described in section 3.4.

	A	B	C	D	E	F	G	H
1		COLUMN_ID	TABLE NAME	CONVERSION	NOTES	APPROVED		Chunks
3	Gender	SEX	M_SRDW1.PERSONAL_DATA			3/11/2015		Student Demographics
4		SEX_DESCRSHORT	M_SRDW1.PERSONAL_DATA			3/11/2015		Student Demographics
5	Ethnicity	ETHNIC_GROUP	M_SRDW1.PERSONAL_DATA	IPEDS ethnicity		3/11/2015		Student Demographics
6		ETHNIC_GROUP_DESCRSHO	M_SRDW1.PERSONAL_DATA	IPEDS ethnicity		3/11/2015		Student Demographics
7	Ethnicity Detail	ETHNICITY_CODE (1-7)	M_SRDW1.ETHNICITY_DTL	from M_SRDW1.ETHNICITY_DTL, M_ where M_SRDW1.PERSONAL_DATA.I and M_SRDW1.PERSONAL_DATA.EN	1 column each for each ethnicity with 0/1 indicators Recode to match ETHNIC_GROUP numeric values Check on old values -- were they converted (or not) When did Two or More become active? (maybe 2010)	4/8/2015		Student Demographics
8	URM	UNDREP_MNRTY_IND	M_SRDW1.PERSONAL_DATA	Transform to 0/1	Check against Two or More -- are there cases where URM should be flagged and is not?	4/8/2015		Student Demographics
9		MULTI_ETHNIC	M_SRDW1.PERSONAL_DATA	Transform to 0/1		3/11/2015		Student Demographics
10		HISP_LATINO	M_SRDW1.PERSONAL_DATA	Transform to 0/1		3/11/2015		Student Demographics
11	Perm Residence Zip	PERSON_POSTAL	M_SRDW1.PERSON_ADDRESSE	case WHEN PERSON_COUNTRY = 'USA' and ADDR_TYPE = 'PERM' THEN PERSON_POSTAL ELSE NULL END	Want FIRST permanent Zip from Production GEOLOC info instead (or addition?) Source from Admissions?	3/11/2015		Student Demographics
12	Foreign Residence Country	PERSON_COUNTRY	M_SRDW1.PERSON_ADDRESSE	case WHEN PERSON_COUNTRY != 'USA' and ADDR_TYPE = 'PERM' THEN PERSON_COUNTRY ELSE NULL END	Want FIRST	3/11/2015		Student Demographics
13	Foreign Residence Country Description	PERSON_COUNTRY_DESCR	M_SRDW1.PERSON_ADDRESSE	case WHEN PERSON_COUNTRY != 'USA' and ADDR_TYPE = 'PERM' THEN PERSON_COUNTRY_DESCR ELSE NULL END	Want FIRST	3/11/2015		Student Demographics
14	Citizenship	CITIZENSHIP_STATUS	M_SRDW1.PERSONAL_DATA			3/11/2015		Student Demographics
15		CITIZENSHIP_STATUS_DESCF	M_SRDW1.PERSONAL_DATA			3/11/2015		Student Demographics
16		CITIZENSHIP_COUNTRY1	M_SRDW1.PERSONAL_DATA			3/11/2015		Student Demographics
17		CITIZENSHIP_COUNTRY1_DE	M_SRDW1.PERSONAL_DATA			3/11/2015		Student Demographics
18		CITIZENSHIP_COUNTRY2	M_SRDW1.PERSONAL_DATA			3/11/2015		Student Demographics
19		CITIZENSHIP_COUNTRY2_DE	M_SRDW1.PERSONAL_DATA			3/11/2015		Student Demographics
20	Term Started	FIRST_TERM_ATTENDED	M_SRDW1.PERSONAL_DATA			3/11/2015		Student Demographics
21		FIRST_TERM_ATTENDED_DE	M_SRDW1.PERSONAL_DATA			3/11/2015		Student Demographics
22	First Term Admit Type	ENTRY_TYPE	M_SRDW1.STDNT_CAR_TERM	Should correspond to first term in STDNT_CAR_TERM	copy entry type from stndt_car_term where term = personal data First Term Attended	6/25/2015		Student Demographics

Figure 1: Screen capture of a Google Sheets spreadsheet where student demographic elements were being discussed and approved for inclusion in the dataset

For each data element or set, the description, location in the administrative data, conversion elements, notes, and date of the project committee approval were listed. The collaborative nature of Google Sheets allowed for draft structures to be discussed, such as the proposed structure to simplify students' majors and degree programs to a single row in the LARC dataset. In all, the design phase of the LARC dataset took roughly six months through June 2015.

Once the design phase was complete, the building phase began under the direction of the ITS business intelligence team. This team also utilized Google Analytics to share examples of metadata, data logic, and, eventually, sample data elements in a simple, portable format that could easily be shared and commented on (Figure 2).

	E	F	G	H	I	J	K	L	M	N	O
1	Class Number	Term Short Description	Grading Basis Enrollment Code	Grading Basis Enrollment Description	Course Grade Input Code	Course Grade Official Code	Units Taken Number	Units Earned Number	Earn Credit Indicator	Include Grade Point Average Indicator	Grade Points Per Unit Number
2	70573	SU 2009	CNC	Credit / No Credit	C	CR	2	2	1	0	2
3	70617	SU 2009	GRD	Graded	B-	B-	4	4	1	1	2.7
4	70813	SU 2009	GRD	Graded	C	C	2	2	1	1	2
5	30737	FA 2009	GRD	Graded	C	C	4	4	1	1	2
6	33533	FA 2009	NON	Non-Graded Component			0	0	0	0	0
7	33547	FA 2009	GRD	Graded	B-	B-	4	4	1	1	2.7
8	37393	FA 2009	GRD	Graded	W	W	4	4	0	0	0
9	37407	FA 2009	NON	Non-Graded Component			0	0	0	0	0
10	45378	FA 2009	GRD	Graded	A	A	1	1	1	1	4
11	15555	WN 2010	GRD	Graded	C-	C-	4	4	1	1	1.7
12	33631	WN 2010	GRD	Graded	A	A	3	3	1	1	4
13	37845	WN 2010	GRD	Graded	C-	C-	1	1	1	1	1.7
14	38451	WN 2010	GRD	Graded	C-	C-	4	4	1	1	1.7
15	39183	WN 2010	NON	Non-Graded Component			0	0	0	0	0
16	15719	FA 2010	NON	Non-Graded Component			0	0	0	0	0
17	27067	FA 2010	GRD	Graded	C-	C-	4	4	1	1	1.7

Figure 2: Screen capture of a Google Sheets spreadsheet where course grade elements were being prototyped

Through several iterations and prototypes, the initial LARC dataset was finalized and approved in August 2016 and released in September 2016. Informatica was used to design the extract, transform, and load (ETL) processes, and ER/Studio was used for technical data modelling.

3.6. Rules of the Road: Designing Procedures

While the dataset was being developed, a series of conversations were held and decision points made related to dataset delivery, documentation, and researcher access processes.

3.6.1. Dataset Delivery Mechanisms

One of the most difficult decisions in designing the LARC dataset was how to deliver that dataset to the research community that would ultimately be the users of the data. Since business intelligence in higher education typically designs for academic analytics needs, this was a new problem. Keeping the first guiding principle in mind, that the researcher was the central persona being designed for, the ITS members of the project team utilized their user experience colleagues to conduct interviews with a wide range of potential users. They discovered that there were two different types of learning analytics researchers. The first type of researcher was technologically savvy and competent in using SQL queries to pull, manipulate, and export data to a variety of software or programming endpoints. The second type was savvy in statistics or other research areas but was less adept at querying databases — they commonly downloaded flat files from a variety of sources related to learning.

Given the two types of researchers, the ITS members of the project team designed a dual-delivery strategy. The LARC dataset would be delivered both in a simplified Oracle database where users could query the data directly and would be responsible for choosing the correct snapshot, making appropriate joins, and selecting the correct data elements for their research, and in flat comma-separated value (.csv) files stored on a secure file-sharing service. One file would be delivered for each table for each snapshot release, along with an updated data dictionary and release notes describing the changes to the dataset in the new snapshot, if any. The only difference between these two delivery formats is that, for security reasons, the personally identifiable information table, which contains institutional identifiers and individuals' full date of birth, is only available in the Oracle database. Flat file users who need to match additional institutional data elements with the .csv files can request assistance from ITS or the Registrar's Office.

3.6.2. Documentation, Mock Dataset, and User Forum

Similar to administrative datasets, the LARC dataset has a publicly available data dictionary, updated after each snapshot, that details all tables, elements, and sample values for all aspects of the dataset (https://its.umich.edu/enterprise/administrative-systems/data-warehouse/data-areas/teaching-learning#larc). Shortly after the launch of the first LARC snapshot, potential users contacted the LARC team and indicated that they would like to be able to see a sample of the data before proceeding with the access procedures to make sure that the data would meet their needs. The ITS members of the team thus created a publicly available mock dataset that showed values similar to ones found in the real dataset but with made-up data elements

(https://docs.google.com/spreadsheets/d/1h3K_5VAAT3vdFIX1sHsxWR-6PS78W5jTNRwzyO3I4AA). Finally, a user-only email group was created where LARC users could ask questions of other members of the user community and also browse the history of previously asked questions. One user, for example, asked about the different values found in the "high school GPA" data element and whether they were weighted or unweighted depending on AP courses in the high school transcript. The user community was able to respond, explaining the nuances of the data and the different calculation methods before and after the Fall 2010 first-year class.

3.6.3. Access Procedures for the LARC Dataset

The process for researchers to gain access to administrative data was confusing and decentralized in the early days of learning analytics research at U-M (Lonn et al., 2017). However, as this community of researchers began to grow, the Institutional Review Board (IRB) became more able to direct learning analytics researchers to the appropriate approval processes for projects that could typically be considered exempt from ongoing review and those that needed to follow the nonexempt review process. Also, the Office of General Counsel developed a streamlined memorandum of understanding (MOU) document for an institutional learning analytics program that defined how the program researchers were to comply with the Family Educational Rights and Privacy Act (FERPA) and data security standards.

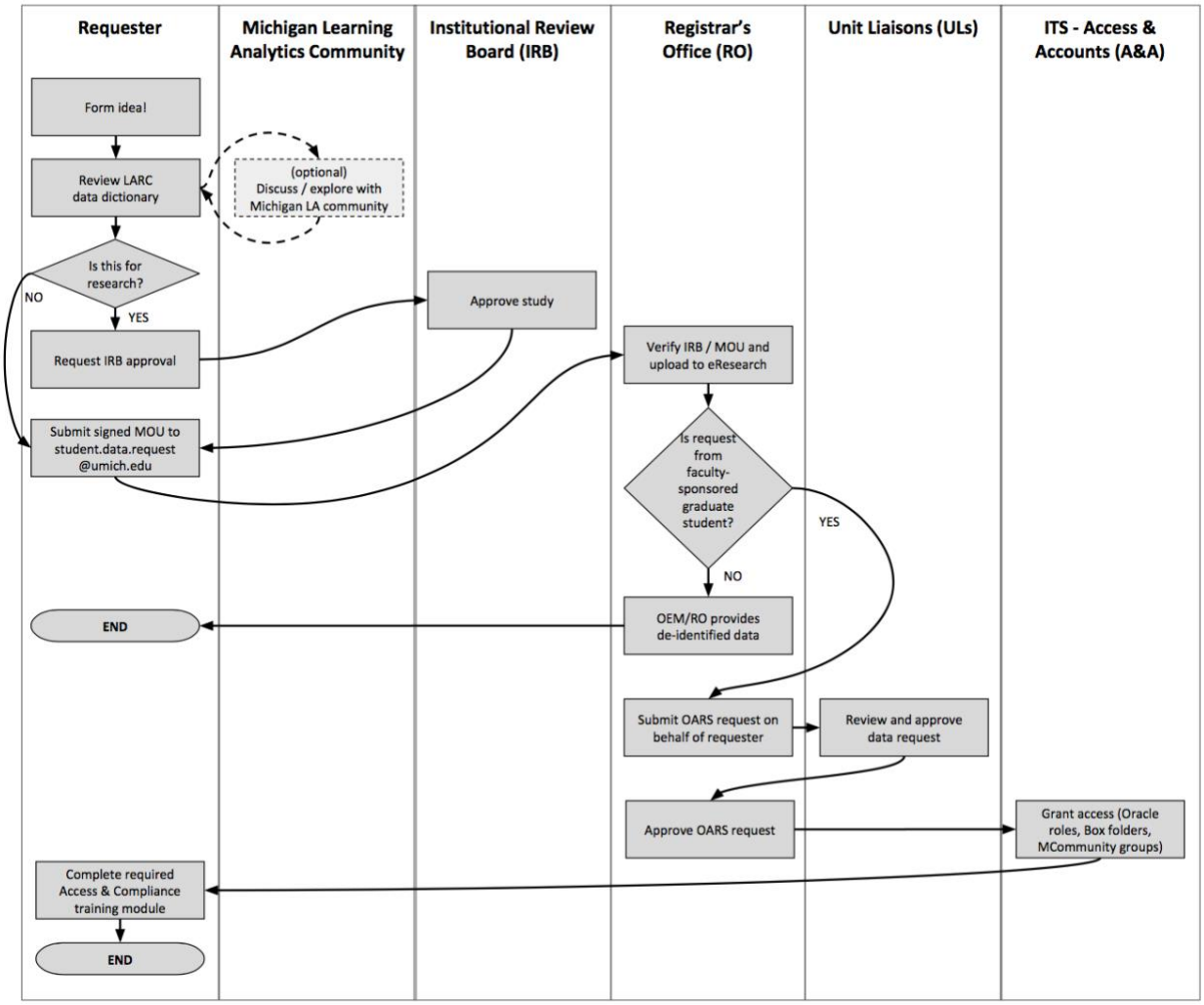

Figure 3: Flow diagram for LARC dataset access approval

Building on these efforts, project lead Steven Lonn worked with the IRB and the General Counsel to create a streamlined process for researchers to gain access to the LARC dataset. First, the streamlined MOU document was modified to reflect the LARC dataset and other data elements the researcher might typically request for learning analytics research. Second, staff members in the Office of Enrollment Management, who approve the MOU documents from researchers, were granted "interested party" status to the IRB system to verify that the researcher had obtained IRB approval and allowed the staff

member to upload the approved MOU document to the IRB file. Finally, all elements of the approval process were documented and shared via the Office of Enrollment Management website (https://enrollment.umich.edu/data-research/learning-analytics-data-architecture-larc). The entire approval process from initial request to dataset access, taking approximately one week in total, is detailed in Figure 3 (OARS refers to the Online Access Request System).

These procedures and standards are important to safeguard the use of this data. While a student's identity is obscured within the LARC dataset, re-identification is certainly possible from the demographic and academic data. Furthermore, given the nascent nature of the learning analytics field, the LARC committee decided not to limit the fields available to researchers, with the exception of the personally identifiable table. The IRB, MOU, and unit liaisons thus work in concert with the individual researcher to ensure ethical use of the data.

3.6.4. Future Enhancement and Refinement of the Dataset

As previously stated in section 3.3, the LARC committee committed to refining and/or enhancing the dataset on a regular basis. The authority for this enhancement process has been delegated to a group of faculty and staff researchers, partially funded by the Provost's Office and the liberal arts college. This committee, which includes the authors of this paper, regularly meets with ITS staff and discusses new elements that could be added to the LARC dataset as well as data elements that might be obsolete or replaced. Some example elements that have been added since the dataset launched in Fall 2016 are new AP, ACT, and SAT test score elements and details for the last undergraduate institution attended (for both undergraduate transfer and new graduate students).

4. Example Analyses and Visualizations

Since the release of the LARC dataset, numerous new research projects have greatly benefited from this normalized and simplified resource. As of April 2019, the dataset has over 50 active users and the user community continues to grow steadily. Provided in this section are a few examples of analyses and visualizations that have been made easier to conduct and generate using the LARC dataset.

The first example is from a research project that asked, "What is the distribution of majors that take Physics 140 (General Physics 1), and what are the average Physics 140 grades of those majors?" The tree map (Figure 4) shows the relative fraction of the total Physics 140 student body that comes from various majors. The tree map itself is hierarchical, so that College of Engineering majors are grouped and labelled E, science majors S, and humanities majors H. Particular majors underperform, which led the researchers to open new investigations about what else we know about students in these majors that might explain their performance.

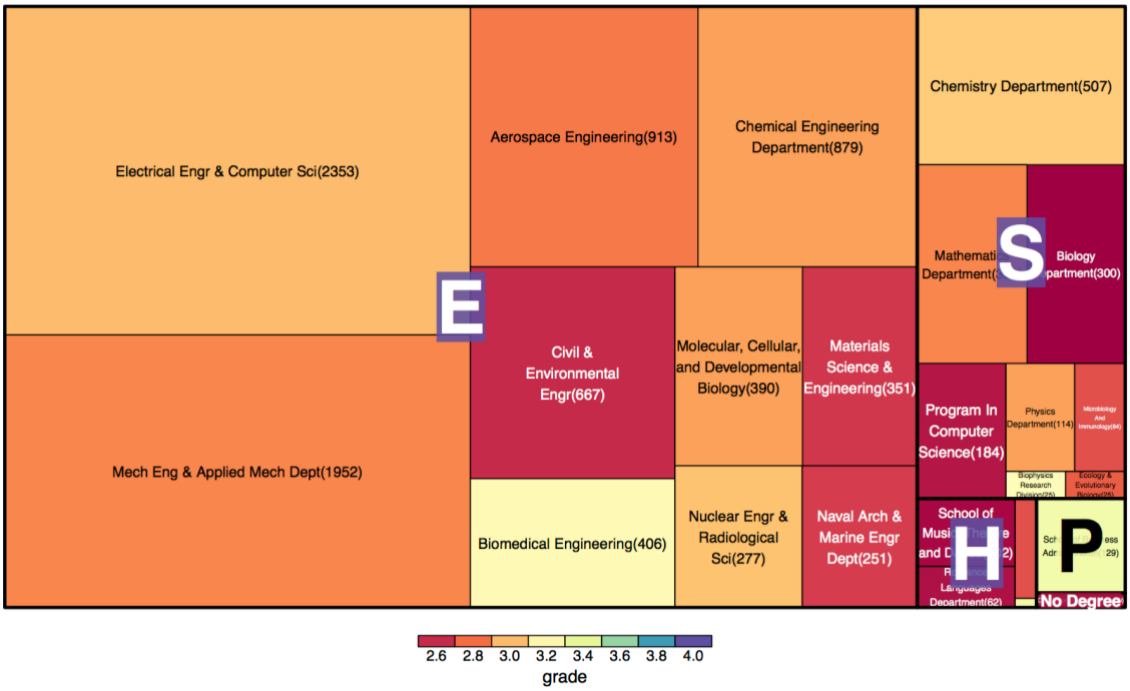

Figure 4: Tree map of student course grades in Physics 140 by major

ISSN 1929-7750 (online).

The LARC dataset contains the essential ingredients of this plot (grade in Physics 140 and eventual major of the student), which are quickly accessible and organized for the inner join (on student ID). These sorts of plots provide a good overview of the connection between a course and the academic pursuits of the students in a way that is interesting to both the academic analytics and the learning analytics communities.

Another example used the LARC dataset to describe students' co-enrolment social networks in selected majors and investigate what these networks revealed about the relatedness of the academic careers of the students in those majors (Figure 5). For large majors on campus, this visualization confirms much of what the researchers expected: many science majors cluster together, but math and physics don't cluster well with other majors. Certain majors are also notable for their betweenness among majors (e.g., psychology), acting as a sort of bridge or connection between science and social science. This result subsequently led the researchers to begin to quantify individual student social networks and to make new measures of student experience on campus, including the academic diversity of students' peers in their social network.

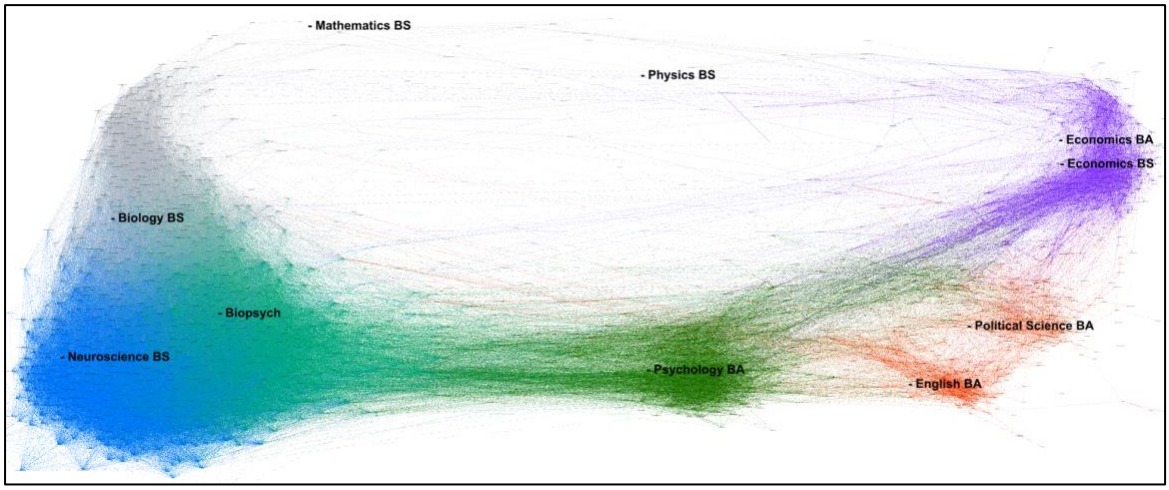

Figure 5: Network plot of all students in selected majors showing the relatedness of the students based on the courses they are concurrently enrolled in

A third example asks, "What are the typical term-by-term sequences of chemistry courses taken by chemistry majors?" This sunburst diagram (Figure 6) is a screen capture of an interactive tool in which the researchers consider up to 13 terms of the student's career (represent by the 13 rings, the innermost being the 1st term, the outermost being the 13th term).

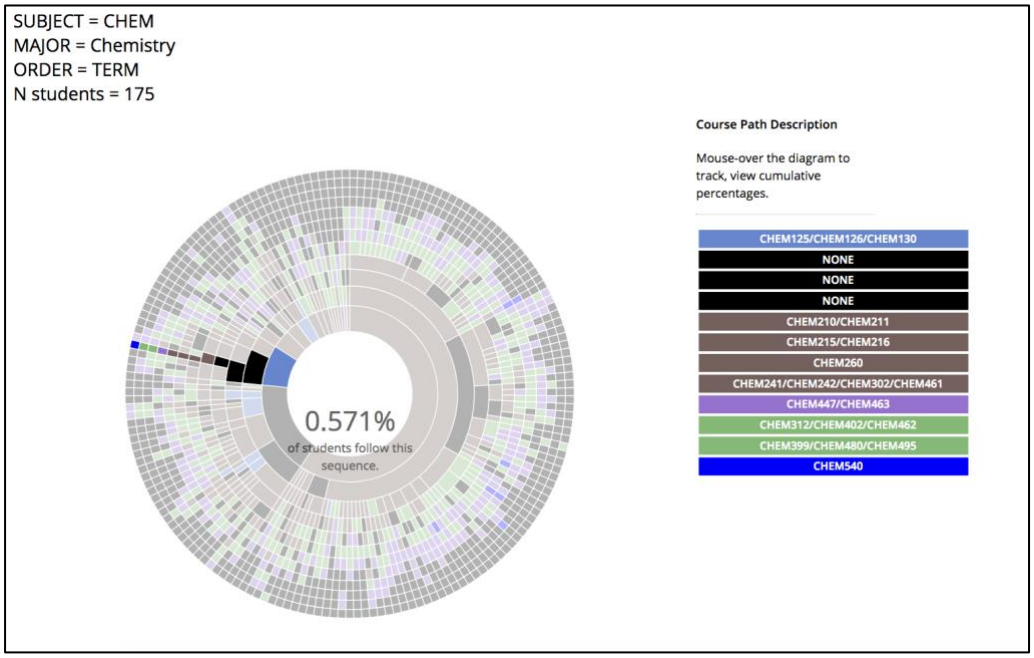

Figure 6: Sunburst diagram showing example sequence of chemistry courses taken (term by term) for a Bachelor of Science chemistry major student

ISSN 1929-7750 (online).

Hovering the mouse over the plot highlights a particular pathway that some fraction *f* of students followed in their pursuit of a Chemistry Bachelor of Science (BS) degree. The breadcrumb trail to the right shows, in order, the chemistry courses taken in each of those terms (or NONE if no chemistry course was taken in the selected term). The particular sequence highlighted here reveals that 0.57% (exactly 1 of the 176 chemistry majors) followed this precise path. This exercise revealed to the researchers the uniqueness of course-taking paths among many majors — even within a major, the required classes and the terms in the career in which they are taken are usually observed one time over the whole career. The same plot (not labelled) reveals that during the first two terms, chemistry students follow quite similar pathways — 53.7% take Chemistry 210 and Chemistry 215 their first and second terms.

A final example shows how the LARC dataset can be used to visualize the academic career of an individual student, how that pathway differs from that of a matched peer, and the average GPA for the courses attempted overall. In Figure 7, the selected student's career is shown on the blue line. The average GPA of peer students matched on test scores, sex, ethnicity, and other demographic factors is shown with the green line. The average course grades for the courses (for all students) attempted are shown with the red line. Here, we can see that the selected student attempted many math and science courses and performed well below the course average and below the average of matched peers. As the student adjusted their course-taking behaviour to focus on social science and humanities courses, specifically psychology, their GPA recovered and the student eventually completed a biopsychology, cognition, and neuroscience bachelor of arts degree. This is an example of how the LARC dataset can be used to better understand individual students' pathways and plan for future learning analytics–driven interventions that might prevent large dips in student performance and ultimately improve the students' learning outcomes.

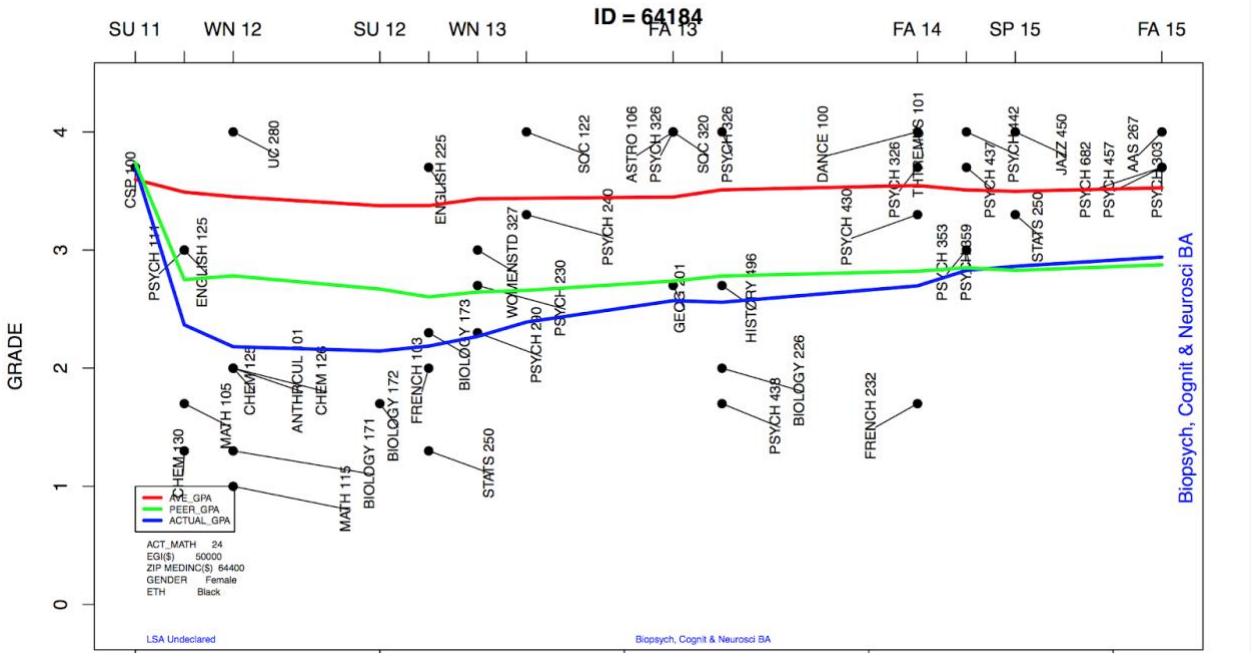

Figure 7: Diagram showing selected student, matched student, and average GPA for courses taken throughout the selected student's academic career

5. Considerations for Other Institutions Seeking to Design a Learning Analytics Dataset

While the LARC dataset was designed to meet the needs of learning analytics researchers at U-M, the project team has expressed at several academic conferences (Lonn, 2017; Lonn & Auerbach, 2018) that they would be very excited to see other institutions in higher education create similar datasets from their administrative data for the research community. In this section, the authors suggest six areas that other institutions might want to consider before they embark on such a project.

5.1. Facilitate Partnerships across the Institution

The LARC project would not have been successful without the commitment of the team members from a wide variety of departments, roles, levels, and institutional knowledge. Furthermore, a host of individuals behind the scenes supported this effort and justified the resources necessary to ensure its success. As they say, "it takes a village," not only to raise a child but also to accomplish any multifaceted task, particularly if the outcome is a common good that can benefit many individuals and units within the institution.

5.2. Learn the Language of Collaboration

Identifying partners is only the beginning. Within the context of the cross-disciplinary collaborations that are increasingly common within learning analytics and data science, collaborations often require quite a bit of level setting. Gray & Szalay (2004) illustrate how individuals with different areas of expertise learn to understand one another in their example of computer scientists working with astronomy domain experts and using a set of 20 questions that would require significant processing power to help frame the requirements for a set of technical solutions. This is somewhat analogous to the list of "common" learning analytics questions that the LARC team brainstormed to help define the scope of the dataset (see section 3.2). In Gray & Szalay's (2004) example, the computer scientists needed to learn how to effectively communicate with the astronomers, similar to how business intelligence and computer science experts in higher education might need to learn how to communicate with learning analytics researchers.

5.3. Obtain Buy-In from Relevant Data Stewards

Learning analytics, along with other new initiatives in higher education, requires leadership to navigate significant and strategic changes to organizational culture and behaviour (Baepler & Murdoch, 2010; Norris, Baer, & Offerman, 2009). Particularly if the dataset crosses multiple data stewards, it is important to get complete buy-in and an understanding that this effort has the potential to positively impact many areas of the institution.

5.4. Identify the Institution's 80% Rule Early On

As mentioned in section 3.3, it is nearly impossible to design a dataset to satisfy all members of a research community, regardless of its age or maturity. It is therefore important to take stock of where researchers are at the individual institution and design for their needs, as well as for the interests of administrators and institutional leaders. Importantly, the kinds of questions that are common to specific institutions will vary.

5.5. Design toward National and International Standards Whenever Possible

The Common Education Data Standards (CEDS) project (https://ceds.ed.gov) was just launching when LARC was being designed. Unfortunately, the standards for higher education were not yet mature enough to guide the development of the LARC dataset. However, as these standards and other multi-institutional efforts continue to grow, it is important that designers stay informed of their development to facilitate later potential joining of institutional data.

5.6. Acknowledge the Complexity, and Design to What Is Appropriate

Today, LARC is one of several learning analytics datasets at the University of Michigan. There are datasets for the learning management system, massive open online courses, and real-time event data (learning record store), as well as several unit-controlled student datasets. While LARC might directly tie in or include one or more of these datasets in the future, it was not practical to design such a massive dataset at the time the project team was designing the first iteration. Future needs of learning analytics researchers will help guide the priorities for what data elements are needed to best address the needs of this evolving domain.

6. Conclusion

While this paper presents the LARC dataset, it also offers a framework for other institutions to do likewise. A key element of the methodology herein is its implicit acknowledgement and handling of institutional idiosyncrasies. The process is generalizable to any institution with digitized student records, interested and motivated researchers, and an accommodating administration. However, the type of institution, the specific questions in which it is interested, the resources, and the data it has available all drive the realization of tailored local solutions. The work presented here exemplifies the process that other institutions might follow.

The LARC dataset was made available to researchers at U-M in September 2016. Since that time, over 50 investigators have requested and received access to the dataset and conducted analyses using the simplified data structures. Other institutions have also inquired about the process to construct a similar dataset at their own institutions. Locally, the LARC model has helped facilitate ongoing conversations between ITS, the institutional research and planning office, enrolment management staff, and researchers, providing a common touchpoint and example from which to approach future collaborations. For example, a new institutional dataset that will simplify, automate, and normalize enrolment data collected from the national student clearinghouse (https://studentclearinghouse.org) is largely following a development model first designed for the LARC dataset.

The LARC dataset was a collaborative project born out of a nascent community of scholars and staff members across an institution that were willing to devote their time and energy to a common goal that would serve as a collective resource for current and future researchers. The dataset today serves as a model for future collaborations that can extend beyond the scope of learning analytics and inform how faculty, students, staff, and administrators can best leverage the greatest resources any institution has at its disposal: the collective knowledge and expertise of the members of the institutional community.

Declaration of Conflicting Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work has been supported by the National Science Foundation (NSF) WIDER grant DUE-1347697 for the REBUILD project, by NSF IUSE grant DUE-1625397, and by the University of Michigan's Institutional Learning Analytics committee.

Acknowledgments

The authors would like to thank the original members of the LARC committee team for their input and guidance in the creation of the dataset as well as the current members of the University of Michigan's Institutional Learning Analytics committee. Many thanks also to the dedicated staff members in the business intelligence team within ITS Information Quest at Michigan. Finally, the creation and continued development of the LARC dataset would not have been possible without the vocal and financial support of members of the University of Michigan administration, particularly Kedra Ishop, Paul Robinson, Sean DeMonner, Cassandra Callaghan, and Vijay Thiruvengadam.

References

Arnold, K. E., Lonn, S., & Pistilli, M. D. (2014). An exercise in institutional reflection: The Learning Analytics Readiness Instrument (LARI). In A. Pardo & S. Teasley (Eds.), *Proceedings of the 4th International Conference on Learning Analytics and Knowledge (LAK '14),* 24–28 March 2014, Indianapolis, IN, USA (pp. 163–167). New York: ACM. https://dx.doi.org/10.1145/2567574.2567621

Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning, 4*(2), 1–9. https://dx.doi.org/10.20429/ijsotl.2010.040217

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSE Review, 42*(4), 40–57. https://er.educause.edu/articles/2007/7/academic-analytics-a-new-tool-for-a-new-era

Gray, J., & Szalay, A. (2004). *Where the Rubber Meets the Sky: Bridging the Gap Between Databases and Science.* Technical Report no. MSR-TR-2004-110. Redmond, WA, USA: Microsoft Research. https://arxiv.org/pdf/cs/0502011.pdf

Lonn, S. (2017). The LARC project: Normalizing student data for IR and learning analytics. Presentation at the Association for Institutional Research Forum, 30 May–1 June 2017, Washington, D.C. http://airforum2017.azurewebsites.net/SessionDetail.aspx?id=89100

Lonn, S., & Auerbach, G. (2018). The data is flat: Enabling learning analytics research using institutional student data. Presentation at the Higher Education Data Warehousing Forum, 8–10 April 2018, Corvallis, OR, USA. https://hedw.org/hedwpresentation/2018-the-data-is-flat-enabling-learning-analytics-research-using-institutional-student-data/

Lonn, S., McKay, T. A., & Teasley, S. D. (2017) Cultivating institutional capacities for learning analytics. In J. Zilvinskis & V. Borden (Eds.), *New Directions for Higher Education,* no. 179 (pp. 53–63). San Francisco, CA, USA: Jossey-Bass. https://dx.doi.org/10.1002/he.20243

Long, P., and Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, *46*(5), 31–40. https://er.educause.edu/~/media/files/article-downloads/erm1151.pdf

Lupton, D. (2016). *The Quantified Self: A Sociology of Self-Tracking*. Cambridge: Polity Press. https://dx.doi.org/10.1111/1467-9566.12495

Nam, S., Lonn, S., Brown, T., Davis, C. S., & Koch, D. (2014). Customized course advising: Investigating engineering student success with incoming profiles and patterns of concurrent course enrollment. In *Proceedings of the 4th International Conference on Learning Analytics and Knowledge (LAK '14),* 24–28 March 2014, Indianapolis, IN, USA (pp. 16–25). New York: ACM. https://dx.doi.org/10.1145/2567574.2567589

Norris, D., Baer, L., & Offerman, M. (2009). A national agenda for action analytics. Paper presented at the National

Symposium on Action Analytics, 21–23 September, 2009, Minneapolis, MN, USA. http://lindabaer.efoliomn.com/uploads/settinganationalagendaforactionanalytics101509.pdf

Pink, S., Ruckenstein, M., Willim, R., & Duque, M. (2018). Broken data: Conceptualising data in an emerging world. *Big Data & Society*, *5*(1), 1–13. https://dx.doi.org/10.1177/2053951717753228

Prinsloo, P., Slade, S., & Khalil, M. (2018). Stuck in the middle? Making sense of the impact of micro, meso and macro institutional, structural and organisational factors on implementing learning analytics. In *Proceedings of the European Distance and E-Learning Network Annual Conference*, 17–20 June 2018, Genova, Italy (pp. 326–334). Budapest, Hungary: European Distance and E-Learning Network. https://www.researchgate.net/publication/325870750

Shultz, G. V., Winschel, G. A., & Gottfried, A. (2015). Impact of general chemistry on student achievement and progression to subsequent chemistry courses: A regression discontinuity analysis. *Journal of Chemical Education*, *92*(9), 1449–1455. https://dx.doi.org/10.1021/acs.jchemed.5b00209

Stonebraker, M., Frew, J., Gardels, K., & Meredith, J. (1993). The Sequoia 2000 storage benchmark. *ACM SIGMOD Record*, *22*(2), 2–11. New York: ACM. https://dx.doi.org/10.1145/170036.170038

Szalay, A. S., Kunszt, P. Z., Thakar, A., Gray, J., Slutz, D., & Brunner, R. J. (2000). Designing and mining multi-terabyte astronomy archives: The Sloan Digital Sky Survey. *ACM SIGMOD Record*, *29*(2), 451–462. https://dx.doi.org/10.1145/335191.335439

York, D. G., Adelman, J., Anderson Jr., J. E., Anderson, S. F., Annis, J., Bahcall, N. A., ... & Boroski, W. N. (2000). The Sloan Digital Sky Survey: Technical summary. *The Astronomical Journal*, *120*(3), 1579–1587. https://dx.doi.org/10.1086/301513