Chapter 2

Data Used in Educational Data-Intensive Research

As early as the 1960s, computers began to fascinate educators. One of the first broadly implemented computer-based learning systems, PLATO (Programmed Logic for Automatic Teaching Operations), arrived 9 years before the first ARPANet transmission—the forerunner of the Internet—and 17 years before the Apple II popularized personal computing. As computers branched out beyond the realms of banking and scientific calculations and into personal applications, the idea of using computers to support teaching and learning gained widespread acceptance (Cuban, 1986). While interest was sparked early on, it took many years for technologies to become widely adopted and implemented with any depth in schools and universities (Collins & Halverson, 2009; Krumm, 2012). The story of technology integration in educational organizations intersects with data-intensive research in important ways: Some of the first technologies to be broadly adopted—learning management systems and intelligent tutoring systems—represent key touch points for the fields of learning analytics and educational data mining, respectively (Baker & Siemens, 2014).

Turning data into knowledge has until very recently been a manual activity (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Manual analysis has been the norm in schools and universities since the dawn of the Progressive era and the proliferation of Scientific Management practices (Tyack & Cuban, 1995). In more recent times, increased use of technology has led to the collection and storage of data that push on the capabilities of most manual approaches. In addition, as the volume of data has increased, so too has the need to combine data from across multiple technological platforms, like administrative data systems and digital learning environments, to better understand the processes and outcomes of teaching and learning. As combining data becomes more important, computer-based techniques are often required to merge, process, and analyze these data, all in an effort to unlock potential insights.

In this chapter, we introduce two of three foundational topics related to data-intensive research in education—*data* and *workflow*. In Chapter 3,

we discuss the third foundational topic—analytical *methods*. None of these topics is unique to data-intensive research *per se*. For example, a workflow is a job-specific set of processes that transform inputs into outputs. All research involves some type of workflow—collecting data and analyzing it using recognized methods (i.e., inputs) to generate new knowledge (i.e., outputs). Over the next two chapters, we emphasize what is distinctive about data-intensive research across these general topics.

In what follows, we begin by outlining three general types of data used in educational data-intensive research. While we describe each somewhat in isolation, in practice, and in many of the examples that we cite, researchers regularly find value in combining different types of data. Following our discussion of three types of educational data, we discuss unique opportunities and challenges associated with using educational data as part of a data-intensive project. Based on our descriptions of three data types, we then introduce a generic workflow that outlines the ways in which data from multiple sources can be analyzed, interpreted, and translated into change ideas that are taken up as part of a formal research study or local improvement project.

Types of Educational Data Used in Data-Intensive Research

In this section, we describe three broad types of data that are perhaps best characterized by the technologies in which they are captured and stored: (1) digital learning environments, (2) administrative data systems, and (3) sensors and recording devices. Data from *digital learning environments*, perhaps more than any other, have fueled data-intensive research in education (Roschelle & Krumm, 2015; Winne, 2017). Games, simulations, and tutoring systems as well as the increased amount of teaching and learning that is occurring through online courses and Massively Online Open Courses (MOOCs) are all creating more and more data on more and more students.

A second type of data fueling data-intensive research in education comes from *administrative data systems*. These systems are used in schools and districts as well as at the level of state and federal governments in the United States to collect and store information associated with delivering some type of service (Figlio, Karbownik, & Salvanes, 2017). For example, with investments from the U.S. Department of Education, states throughout the U.S. have created statewide longitudinal data systems that collect and store data on individual students over time. Data stored in these systems can include standardized test performances, attendance, and major behavioral infractions. Increasingly potent as tools for research, administrative data systems are creating opportunities for researchers and interested practitioners to jointly interpret data to both improve services

as well as answer questions that are useful to the broader research community (Connelly, Playford, Gayles, & Dibben, 2016).

Lastly, as data from digital learning environments have been increasingly collected and stored, so too are data being collected from *sensors and recording devices*, such as video and audio data. Sensor and recording device data have increased in availability through newly developed instruments that capture biometric data and the ability to parse audio and video recordings using machine learning and artificial intelligence techniques. In education, data from sensors and recording devices have been combined with data from digital learning environments, like intelligent tutoring systems (e.g., Bosch, Chen, Baker, Shute, & D'Mello, 2015). These multiple data streams have been blended together to advance researchers' understanding of student learning and factors affecting learning over time within these environments.

Digital Learning Environments

In the following sections, we highlight three digital learning environments based on the degree to which they are used in schools and universities and in their prominence in the research literature: intelligent tutoring systems, learning management systems, and MOOCs.

Intelligent Tutoring Systems

An intelligent tutoring system (ITS) is a type of digital learning environment that applies artificial intelligence to students' interactions with the system. ITSs often employ three *models* that drive the adaptations that a system makes based on a student's input: (1) an expert, or *domain model*, which organizes the skills and strategies in the domain, (2) a *student model* of what a student understands about the domain that is inferred from their performances on learning tasks and (3) an *instructional*, or *pedagogical model*, of common mistakes and misconceptions along with a corresponding feedback strategy (Anderson, Corbett, Koedinger, & Pelletier, 1995). ITSs collect information on students, their progress in the system, and interactions that they engage in during a learning task. ITSs provide feedback to students in the form of hints, strategies, and different ways to practice the skills on which they need help (Razzaq & Heffernan, 2006). The same data that the ITS uses to figure out how to respond to a student's actions can also be used by human analysts to gain a detailed picture of learning processes and the behaviors learners engage in (Baker, 2016). For example, work by Baker and colleagues (Baker, D'Mello, Rodrigo, & Graesser, 2010) illustrate how data from ITSs can be used to detect a variety of behaviors and affective states such as boredom and frustration. These studies help in building knowledge

related to how students learn as well as support potential improvements to the ITSs themselves (e.g., Roll et al., 2006).

Learning Management Systems

LMSs are "web-based systems that allow instructors and students to share instructional materials, make class announcements, submit and return course assignments, and communicate with each other online" (Lonn & Teasley, 2009, p. 686). As noted previously, LMSs, along with ITSs, helped give rise to the fields of learning analytics and educational data mining, respectively. LMSs typically collect information on learning resources (e.g., digital files posted by an instructor) that students accessed and when they accessed them as well as when students accessed an assessment and how well they did on the assessment. Currently, LMSs are more widely used in higher education than in K–12, and they tend to be adopted on a campus-wide basis with the intent that all online and blended courses offered by a college or university are supported by the same LMS. Using data collected and stored by a campus's LMS, Krumm (2012) examined approximately 20,000 courses taught at the University of Michigan. Major takeaways from these analyses revealed that most instructors use relatively few tools that are provided by the LMS but that factors such as the college one teaches in and the enrollment size of the course can affect the number of tools used. In general, instructors favor using tools that make their teaching more efficient as opposed to rethinking how they teach (Lonn & Teasley, 2009). Said differently, while LMSs can be considered widely adopted, they are often not central to teaching and learning. However, when these systems are more central to instruction, researchers have found ways to use data from these systems to drive early warning systems. One such tool that allows instructors to provide feedback to students based on their interactions with an LMS is Course Signals, which was originally developed and deployed at Purdue University (see Arnold & Pistilli, 2012).

Massive Open Online Courses

When MOOCs burst onto the higher education scene in 2010, course enrollments reached hundreds of thousands of students (Means, Bakia, & Murphy, 2014). Critics have been quick to point to the relatively small percentage of enrollees who actually completed these free online courses, and the hype around MOOCs has abated. Nevertheless, the MOOC learning platforms designed for very large enrollments, such as Coursera and edX, have endured, with large numbers of people taking courses on these platforms, including for academic credit. The data generated as thousands of learners use these platforms in a single course continues

to be a major source of data for researchers (e.g., Evans, Baker, & Dee, 2016; Gasevic, Kovanovic, Joksimovic, & Siemens, 2014; Ho et al., 2015; Zhu et al., 2016). As these systems evolve, they continue to develop new features and functionality that capture granular data closer to ITSs than LMSs (e.g., Aleven et al., 2017).

Administrative Data Systems

A second type of data fueling data-intensive research in education stems from *administrative data systems*. These systems are used at school and district levels as well as at the level of entire states. In this section, we describe two types of administrative data systems: student information systems and statewide-longitudinal data systems.

Student Information Systems

Student information systems (SISs) are digital systems used by schools and universities to store student-level information. They are, in many ways, the central data repositories for educational organizations as they collect and store multiple data elements on students, including demographics, attendance, and academic performances. SISs are different from learning management systems, but the two can be integrated. While LMSs are often used as student-facing repositories of digital resources and activities, SISs are teacher- and administrator-facing repositories of student demographic and learning-outcome data. SISs play a key role in data-intensive research because they offer a ready source of data on educational outcomes (e.g., grades) and demographic information, which can play a role in evaluations of technology-based interventions as well as early warning system research (e.g., Bowers, Sprott, & Taff, 2013).

Statewide Longitudinal Data Systems

In 2005, the U.S. Department of Education began giving grants to states to develop statewide longitudinal data systems (SLDSs). Among the requirements for SLDSs developed with these funds was the use of a unique statewide identifier for every student; storage of each student's demographic characteristics and enrollment history and scores on state accountability tests; and the ability to link the student's K–12 data with the state's higher education data system. According to the National Center for Education Statistics, by 2015, 84 percent of statewide longitudinal data systems contained unique student identifiers, 88 percent contained demographic and enrollment history data, and 57 percent could link to higher education data systems. For the first time, there was a data infrastructure in a majority of states that provided the potential to examine,

for example, educational outcomes at the scale of an entire state. State level and university-based researchers increasingly leveraged these data for both accountability and reporting purposes as well as district-and school-improvement purposes. Knowles (2016), for example, used data from Wisconsin's SLDS to develop an early warning system for students at risk of dropping out of high school.

Sensors and Recording Devices

As data from digital learning environments have been increasingly collected and stored, so too have data been collected from *sensors and recording devices*. Location, physical movement, and speech can all be tracked and analyzed using a variety of different sensors—small, often single data stream devices. Fitness sensors that measure, for example, steps taken or heart rate have been used in educational contexts to promote healthy behavior changes in youth (e.g., Schaefer, Ching, Breen, & German, 2016). Thus, sensor and recording device data have increased through instruments that capture biometric data, which are quantifications of an individual's physical activity. Moreover, the ability to parse audio and video recordings using machine learning and artificial intelligence techniques has opened up opportunities to analyze familiar forms of data, such as audio and video files, at larger and larger scales. Hand in hand with different algorithms have been multiple advances in collection and storage of these data in various digital formats (Baker & Siemens, 2014).

An important recent advance involves blending multiple data streams from sensors, recording devices, and digital learning environments (Blikstein, 2013; Liu, Davenport, & Stamper, 2010). Merging, or fusing, data from multiple systems can allow researchers to identify patterns across the different data streams that have been brought together. These *multimodal* investigations are providing new insights into basic factors affecting learning (e.g., Woolf et al., 2009). Understanding what learners do as they engage in learning tasks can drive digital learning environment adaptations. Recent work suggests that combining sensor data with data from digital learning environments can support accurately identifying multiple affective and engagement-related states (e.g., D'Mello, Dieterle, & Duckworth, 2017).

In the same way that data from sensors can be used to measure specific behaviors over time, audio and video recording data can be used by to detect facial expressions (e.g., Bosch, D'Mello, Ocumpaugh, Baker, & Shute, 2016) and body language (e.g., Grover et al., 2016). Audio data can be used for speech recognition, and even without analyzing the meaning of the recorded utterances, speech prosody (i.e., stress and intonation) can be used to make inferences about the emotional state of speakers.

For example, D'Angelo et al. (2015) are building speech-based learning analytics for collaboration that can support teachers to identify what is occurring in small groups, thereby enabling teachers to direct their attention to less well-functioning groups. Pilot data have shown that combining speech activity (i.e., who is talking when) with the actions of collaborators in digital learning environments can identify turn-sharing and frustration.

Characteristics of Educational Data

The three types of education data described previously are intended to be overarching categories with which to think about the rapidly expanding types of data used to understand and improve teaching and learning. As can be seen in the previous examples, many researchers and research groups combine data from across these categories, and much of what are considered *big data* in education fall into one or more of the categories described previously. But what exactly are big data? Many scientific disciplines work with large, complex datasets (Dede, 2015), and the term big data is a relative and regularly shifting assessment of "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Manyika et al., 2011). In a similar way, *data-intensive research* is a relative term that speaks to both the data and the research field in which the data are collected and analyzed. By current standards, datasets used in educational data-intensive projects have hundreds of thousands or millions of observations or hundreds or thousands of *features*. Depending upon the analytical method used, many of these datasets require software and hardware with specific capabilities to analyze, and the specific hardware and software will vary depending upon the ultimate purposes one has for an analysis.

In our own work, we regularly draw on data from LMSs and SISs. And over the years, we have developed a degree of familiarity with how to wrangle, explore, and model these types of data. Data from LMSs are often similar to one another but different from other types of educational data one could use in a data-intensive research project. Reasons for why these data are similar to one another but different from other types of data involve (1) the tasks that students are engaged in and (2) how data from those tasks are collected and stored by the technology. As noted previously, the types of data that are most often collected from LMSs include learning resources selected and when, as well as learning activities, such as assessments completed and when. These data can be substantively different from, for example, game-based learning environments because, at multiple levels, the types of activities that students are engaged in within a digital game are often dramatically different from an LMS (e.g., Owen, Ramirez, Salmon, & Halverson, 2014). Thus, working with and making sense of data require becoming familiar with the activities of the digital

learning environment as well as the ways in which data from those activities are captured and stored for later use.

As one explores data from different technologies, one is likely to experience both structured and unstructured data—as well as variations in between. Structured data does not have a precise definition. In general, it is any kind of data organized into a table with rows and columns. Therefore, structured data have an explicit organization, and more often than not, structured data are housed in well-defined relational databases. One of the benefits of structured data is that they can more easily be manipulated and analyzed than unstructured data, such as large segments of text, audio, and video. While similarly lacking a precise definition, what makes unstructured data *unstructured* is that it does not have an explicit, predefined organization. Thus, tabular organization must be provided after the fact, often requiring significant wrangling and pre-processing. For example, when assessing samples of student writing, each sample needs to be converted into a list of numeric features, many thousands of them, each of which captures a different characteristic (Rutstein & Neikrasz, 2016). These numeric features can then be modeled using supervised machine learning algorithms across training and testing data. Known outcomes from individuals who scored the same writing samples train an algorithm. After adequate training and testing, the algorithm can be put into production in order to score new, unseen texts. Figure 2.1 illustrates this workflow. The latter part of this workflow is made possible by providing tabular, numeric structure to the previously unstructured data.

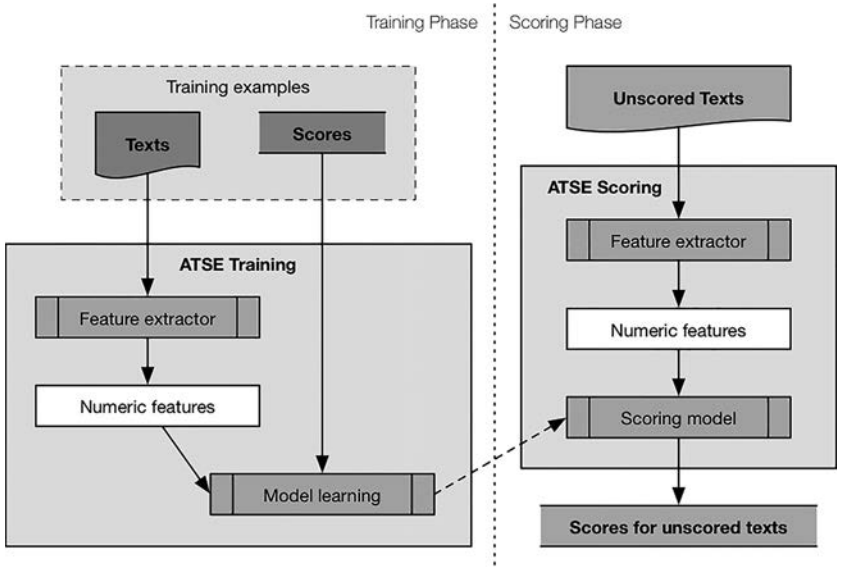

Figure 2.1 Training and Scoring Phases in an Automated Text Scoring Engine (ATSE)

Oftentimes, discussions of big data in education mention anywhere from 3 to 7 "Vs": volume, velocity, variety, veracity, variability, visualization, and value (e.g., van Rijmenam, 2013). The intent of these Vs is to help distinguish the types of data one is likely to use in data-intensive research as opposed to more traditional modes of inquiry. The four Vs that are most germane to data as opposed to how the data are used include volume, velocity, variety, and veracity. Volume is about the amount of data available, which is often affected by the number of observations, the number of features per observation, or both. Velocity addresses the rate at which data are generated; for example, every click that a student makes within a digital learning environment can lead to a rate of multiple clicks per minute, and over multiple minutes spread out over multiple days, volume and velocity can become closely related ideas. Variety describes the different types of observations, or events, that can be gleaned from a technology. For example, a digital game environment, from the same session, can continuously track a player's screen coordinates as well as specific interactions within the game environment, all of which can lead to highly variable data over time. Veracity captures the degree to which a user can trust data. There are no standard units of measure for veracity, but data can be untrustworthy for a variety of reasons. For some administrative data systems, individuals inputting the data can use the system differently than intended, which means end-users of the data, such as researchers, need to understand as deeply as possible how and why data are entered into and stored within a system (Figlio et al., 2017).

To ground the four Vs in an educational example, as a researcher, imagine working with several high schools in a large district to help them in identifying patterns in students' attendance over time. Instead of looking at whether a student was present or absent for the entire school day, the participating high schools are interested in the patterns that manifest by following individual students over time on a period-by-period basis. Teachers in participating high schools recorded whether or not a student was present across seven instructional periods, which yields seven measures per day, per student. For 6,000 students across the high schools, this would create a table with over 7.5 million cells for a 180-day school year. Following a single cohort of students from grades 9 through grade 12, then, would produce over 30 million unique observations that could be mined to look for patterns in which classes are missed most often and the relationships between missing class and overall school performance. Over a four-year period, the *volume* of a final dataset, depending upon when downloaded and analyzed, will reach the numbers identified previously. The *velocity* of these attendance data could best be thought of as hourly (i.e., at least during a typical school day). Importantly, these data are marginally low variety in the form of "present" or "not present" for a given hour of the day, such as "Period 1." If one is interested in the specific

courses a student was present or absent for, such as "Period 2 Geometry," this level of detail increases the variety. *Veracity* is about the degree to which one can trust a row, column, or cell of data. For example: what does it mean to be counted as "late" for a class? The rule that defines lateness may or may not match the expectations of end-users of the data.

The four Vs described previously offer useful ways of thinking about characteristics of big data, but these characteristics may not ultimately address what can be unique about educational data used in a data-intensive research project. What is unique about working with educational data, especially within the context of formal schooling environments, is the interaction between the technology and the environment in which the technology is used. Focusing on the technology, we have noted the importance of the *tasks that students engage in* within a digital learning environment as well as the ways that *data from those tasks are collected and stored* by the technology. A rich digital learning task that does not capture granular data on what students do within the task, by definition, will not be useful for data-intensive research as requisite data are not collected. Less rich tasks that capture data on what students do, on the other hand, may also not be useful for data-intensive research as these data often fall victim to the garbage in, garbage out principle (Mislevy, Behrens, DiCerbo, & Levy, 2012). Rich tasks that are specifically developed so that students can generate meaningful events represent the best initial set of technology-specific circumstances for analyzing educational data (Schwartz & Arena, 2013; Shute & Ventura, 2013). However, technologies are not used in a vacuum—when taken up in schools, a technology will be used by students and teachers who can have different goals from those of the technology's developer.

Two other characteristics of educational data based on the interaction between a technology and the environment in which it is used include *coverage* and *centrality*. Coverage as we use the term denotes the number of students within an educational organization who use a given technology from which data are collected. Centrality denotes the degree to which the technology is used as a core element, or facilitator, of instruction, i.e., how students interact with one another, the instructor, and content to be learned (Cohen, Raudenbush, & Ball, 2003). Coverage can be important because a technology from which data will later be used may in fact not be used by large numbers of instructors or students. Fewer students or more narrow groups of students (i.e., two dimensions of coverage) will dramatically affect the claims a researcher may seek to make based on the particular coverage of a technology he or she is analyzing. Given the diversity of content areas and instructional approaches in schools, the types of technologies with the most coverage by default tend to be those that facilitate more generic instructional interactions, such as accessing resources and submitting assessments. LMSs are a prototypical example

of a broad coverage, generic technology as they can be used in all content areas and nearly all grades. Other technologies that offer high degrees of coverage include administrative data systems, as they track similar data elements for nearly all students.

Coverage and centrality, much like the four Vs described previously, are not inherently positive or negative characteristics. Broad coverage and non-central data from student information systems largely fuel early warning system research and development. Highly central technologies that have broad coverage are rare. One challenge in working with broad coverage systems like LMSs is that there are often large amounts of variation across educational units, such as classrooms and courses, using the technology. Given these differences, data from broad coverage technologies often necessitate that special attention be paid to unit-to-unit differences.

Task richness and how data are collected from tasks; coverage and centrality; the 4 Vs; and traditional considerations of quality educational research, such as overall research design, all factor into using educational data for data-intensive research. Building on these general characteristics, we now turn our attention to practical concerns around accessing and sharing educational data.

Challenges and Opportunities in Working With Educational Data

There are multiple challenges as well as opportunities in working with educational data. Opportunities include building new knowledge as well as engaging in practical school improvement work—and new ideas yet to be developed. Data from digital learning environments as well as sensors and recording devices offer unique opportunities because they can be used to measure educational *processes* as they unfold over time. A core tenet of improvement is that changes in outcomes are dependent upon changes in processes (Langley et al., 2009). As many of the articles cited previously demonstrate, rigorous analyses of process data can also be used to build new understandings of how people learn. While there are a number of opportunities, privacy and security remain large and looming challenges in working with educational data. In Chapter 4, we detail many of these issues. For the purposes of understanding the types of data introduced previously, in this section, we describe several challenges and opportunities facing researchers in working with educational data as part of a data-intensive project.

A big challenge involves working with data from across multiple technologies, such as digital learning environments and administrative data systems. Issues of different identifiers used across technologies as well as duplicated entries can make merging datasets a labor-intensive and

sometimes error-prone activity. A related challenge to this is making sure that the right students are present in the right datasets, which is often most noticeable after different datasets have been merged together. The integrity of samples of students directly implicates the veracity of educational data used in data-intensive research, as we described previously. The data a researcher eventually analyzes depends upon the business rules of the database as well as the informal rules around how individuals input and make use of data within these systems. For example, what counts as an "enrolled student" in a college course that uses an LMS can be far from clear-cut using LMS data alone. Thus, for projects geared toward predicting students who are likely to drop out of a course, corroborating data from across multiple sources can become a critical activity. Ultimately, it is where intended and actual uses for a technology conflict that working with data from across multiple datasets can prove problematic because intended uses are easy to communicate through data dictionaries and other written materials—informal and non-standard uses, less so.

Solutions to some of these challenges have included services offered by for-profit companies and industry groups that support normalizing student rosters across technologies (e.g., Clever and OneRoster). Other approaches include growing numbers of educational organizations developing and housing more and more data in data warehouses, which often contain common identifiers across databases. Moreover, there are growing standards movements that are intended to help create more common data models for administrative data systems (e.g., Ed-Fi Alliance) and digital learning environments (e.g., Experience API, Caliper). In general, these efforts address *interoperability*; programs such as the Schools Interoperability Framework (SIF) and Common Education Data Standards (CEDS) have emerged from consensus among vendors on how data can be exchanged across systems.

Accessing Data

A challenge for both new and experienced researchers and educational data scientists is accessing data. State departments of education and individual school districts, starting in the early 2000s, began using administrative data systems and making their data available to researchers for well-defined research purposes. Once in the hands of researchers, these data were analyzed, reported on, and oftentimes destroyed in line with more or less well-defined data-use agreements. In many ways, data has been open to researchers with legitimate research purposes for a long time. Similarly, data collected by a digital learning environment could be accessed and analyzed by researchers with legitimate research purposes. In education and the physical and social sciences more generally, there is a growing movement where various datasets are being made publicly

available. These efforts are moving data from servers once only accessible by researchers into public repositories that are creating opportunities for researchers to explore new questions and individuals new to data-intensive research in education to develop skills using often highly structured and well-documented datasets.

Under the labels of "open data" and "reproducible science," a variety of data sources are being opened up to broader audiences. The basic idea behind the open data movement is that anyone can access or use a dataset, and key to this movement is not just accessibility but usability. Making open data usable means making it accessible in machine-readable, structured, granular, and well-documented formats. On a case-by-case basis, individual research projects have made data available to external audiences (e.g., the Study of Instructional Improvement at the University of Michigan). These efforts can support replication of results as well as new explorations. Researchers in other sciences have proposed principles for enhancing the reproducibility of those results that are based on computational methods. They argue that while sharing data is useful, unless the computational software and workflow are also made available, the "computational reproducibility" of the findings cannot be assured. "Access to the computational steps taken to process data and generate findings is as important as access to data themselves" (Stodden et al., 2016, p. 1240).

The rise of structured, machine-readable data permits researchers to combine information or search for new patterns and new insights. The National Center for Education Statistics (NCES) is another resource for a variety of accessible, well-documented datasets (e.g., the Common Core of Data). More recently, datasets from tutoring systems and large online courses (e.g., MOOCs) are also being used in this way. For example, Harvard and MIT, in 2014, released de-identified data from open online courses, containing the original learning data from the 16 HarvardX and MITx courses offered in 2012–13 (Ho et al., 2015). Researchers at the Pittsburgh Science of Learning Center have developed and maintained the "world's largest repository of learning interaction data" in DataShop (Koedinger et al., 2010). DataShop contains data from multiple online educational environments, is open access, and is designed to provide researchers with a place to share data as well as analytic tools.

Data-Intensive Research Workflow

The forerunner to data-intensive research, and therefore learning analytics and educational data mining, is a field of inquiry referred to as knowledge discovery in databases (KDD). The phrase was initially used in the late 1980s, and it was coined to emphasize that knowledge was the key outcome of any data-driven inquiry. From the outset, KDD referred to an overall workflow: "data preparation, data selection, data cleaning,

incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data" (Fayyad et al., 1996, p. 39). As we noted at the outset of the chapter, a workflow is a set of processes that transform inputs into outputs across multiple steps and decisions. A key input into this workflow consists of the types of data detailed previously. In this section, we introduce a generic workflow that is intended to support researchers, practitioners, and data scientists prepare for a data-intensive analysis and communicate one's findings. This workflow is based on workflows that have been documented by general data science practitioners (e.g., Guo, 2012; Wickham & Grolemund, 2017) as well as workflows that are based on practitioners' use of data in schools (e.g., Marsh, 2012).

A common workflow carried out using shared data analysis tools can make for efficient, reproducible data-intensive research (see Figure 2.2). In Chapters 6 and 7, we place this workflow within a broader set of phases that we use to help researchers and practitioners organize their collaboration around data-intensive analyses as well as co-developing and testing change ideas inspired by their analyses. The workflow described in the next sections comprises five steps: (1) prepare, (2) wrangle, (3) explore, (4) model, and (5) communicate. In Chapter 3, we go more in-depth into steps 2–4.

Prepare

First and foremost, data-intensive research involves defining and refining one or more research questions. Having a clear set of research questions helps a team identify what data to collect and formulate potential analytical strategies. Along with clear questions, it can be useful to identify what gets collected and stored by a technology—not all potentially useful data are collected by a technology and not all data collected by a technology are useful. In an education context, understanding the *activity system* in

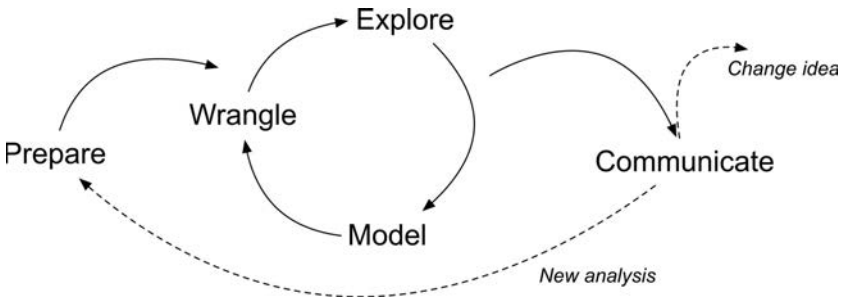

Figure 2.2 Steps of Data-Intensive Research Workflow

which a technology is used can be crucial for ultimately making sense of data, in particular from digital learning environments (Roschelle, Knudsen, & Hegedus, 2009). Some instructional activity systems can include, among many factors, the actions and intentions of teachers and the goals that they have for students—from serving as a reward to students for completing work early to providing students' primary interactions with a course's content. All of these uses for a technology can affect the conclusions one can draw from data stemming from the technology as these different uses influence which students interact with it in the first place as well as what they do within the technology (Murphy et al., 2014). Being prepared for a data-intensive analysis, therefore, involves refining research questions and developing an understanding of where the data come from.

Wrangle

Wrangling data, sometimes referred to as munging or pre-processing entails the work of *manipulating*, *cleaning*, *transforming*, and *merging* data. At a basic level, manipulating involves identifying, acquiring, and importing data into analysis software; cleaning data involves ensuring that each variable is in its own column, each observation is in its own row, and each value is in its own cell within a dataset (Wickham & Grolemund, 2017). Data cleaning also involves identifying and remediating missing data, extreme values, and ensuring consistent use of identifier, key, or linking variables. Data wrangling can also involve transforming variables, such as recoding categorical variables and rescaling continuous variables. These types of transformations are the initial building blocks for exploratory data analysis. Along with manipulation, cleaning, and transforming data, merging data is an important component of data wrangling. One of the earliest and biggest value-adds that a data scientist can bring to a formal research project or local improvement project is merging once disparate data sources. For example, merging data from a student information system that stores student grades with data from a digital learning environment that stores students' longitudinal interactions within a specific technology can be used to unlock the relationships between what students do or do not do on a day-to-day basis with how they performed on a longer-term outcome, such as a course grade. Merging data on what students do, i.e., process data, with how well they do, i.e., outcome data, are the building blocks of multiple types of *models*, described later.

Explore

Exploratory data analysis is a widely covered topic that captures some combination of *data visualization* and *feature engineering*. Data visualization involves graphically representing one or more variables, whereby

the goal of data visualization, according to Behrens (1997), "is to discover patterns in data that allow researchers to build rich mental models of the phenomenon being examined" (p. 154). Discovering patterns in data entails generating questions about one's data, visualizing relationships between and among variables, and creating as well as selecting features for subsequent data modeling. Feature engineering is the process of creating new variables within a dataset, which goes above and beyond the work of recoding and rescaling variables. For example, using data from an ITS, Baker, Gowda, and Corbett (2011) created new features, such as *the length of time a student paused after reading a hint*. Veeramachaneni, O'Reilly, and Taylor (2014) used brainstorming and crowd-sourcing techniques to develop features—such as *the difference in grade between current lab grade and average of student's past lab grade*—that were used to predict when students would stop actively participating in a MOOC course. Feature engineering draws on substantive knowledge from theory or practice, experience with a particular data system, and general experience in data-intensive research.

Model

Modeling involves developing a mathematical summary of a dataset. There are two general types of modeling approaches: unsupervised and supervised learning. Unsupervised learning algorithms can be used to understand the structure of one's dataset. Supervised models, on the other hand, help to quantify relationships between features and a known outcome. Known outcomes are also commonly referred to as labels or dependent variables. A known outcome can include longer-term results of complex processes, such as dropping out of high school (Knowles, 2016), or shorter-term results like being off task (Hershkovitz, Baker, Gobert, Wixon, & Sao Pedro, 2013). Features used in a supervised learning model can also be referred to as predictors or regressors. Other names for features include attributes, independent variables, or simply—variables.

Unsupervised learning algorithms are often characterized as exploratory because unlike supervised learning models, they cannot be easily evaluated against a ground truth, or known outcome. When using supervised learning models, on the other hand, one can test a model's predictions against known outcomes. Supervised learning, or predictive modeling, involves two broad approaches: classification and regression. Classification algorithms model categorical outcomes (e.g., yes or no outcomes); regression algorithms characterize continuous outcomes (e.g., test scores). A model, the result of model-*ing*, can refer to either a general algorithm or a particular algorithm that has been applied to a particular dataset. When used to refer to a general algorithm, a model is a set of mathematical rules; in specific form, a model mathematically summarizes relationships within particular datasets (James, Witten, Hastie, & Tibshirani, 2013).

The process of modeling involves both *building* and *evaluation*. Building a model entails selecting features from a dataset and applying one or more algorithms to the dataset. Those who build a model are evaluating its performance using a variety of techniques, such as bootstrapping or cross-validation. Formally evaluating a model involves assessing its performance (i.e., how well it classifies categorical outcomes or predicts continuous values) on data that were not used to build the model. The steps involved in modeling, much like exploratory data analysis, are iterative and build on one another over time.

Communicate

Communicating what one has learned involves *selecting* among those analyses that are most important and most useful to an intended audience. In addition, one must choose a form for displaying that information, such as a graph or table in static or interactive form. After creating initial versions of data products, research teams often spend time refining or *polishing* them, by adding or editing titles, labels, and notations and by working with colors and shapes to highlight key points. In addition, writing a *narrative* to accompany the data products is important and involves, at a minimum, pairing a data product with its related research question, describing how best to interpret the data product, and explaining the ways in which the data product helps answer the research question. These three steps—select, polish, and narrate—are intended to create a stand-alone data product that the intended audiences can use to inform their work.

The workflow cited previously lays out a series of steps for engaging in data-intensive research. Having a workflow creates multiple benefits and is intended to help both new and experienced educational data scientists create more reproducible data products, share analyses with internal and external audiences, and provide a structure for updating one's analyses over time. The workflow can help in achieving these goals by providing a key set of activities to address and an order in which to address them. While each step can and will be engaged in different ways across individuals and teams, each step represents an important one for almost any researcher or data scientist.

At the beginning of this section, we presented a somewhat linear movement across these five steps, from left to right in Figure 2.2. While there is often a great deal of iteration that occurs from wrangling to exploring to modeling, at any given time in a project one can be engaged in an activity that is difficult to put into any one step alone. Over time, we have come to see the workflow as overlapping activities as much as steps. Figure 2.3 is an alternative rendering of the workflow that captures the ways in which activities overlap and can be difficult to disentangle as

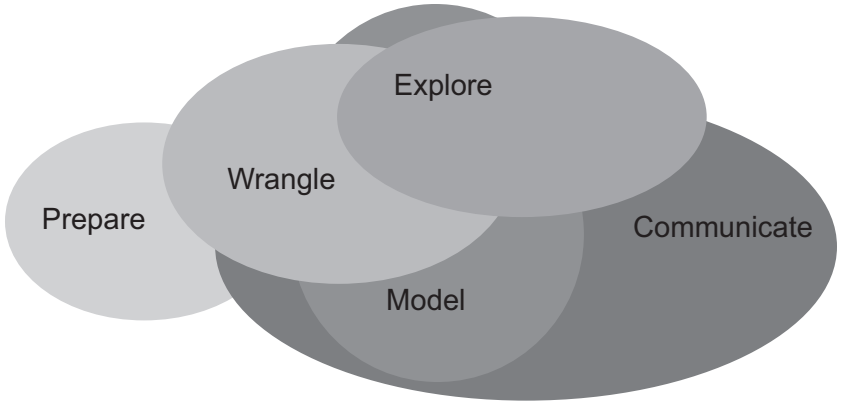

Figure 2.3 Overlapping Activities Within the Data-Intensive Research Workflow

distinct steps—especially while engaged in a project. For example, *communicate*, in practice, is not a single step that occurs at the end of a formal modeling process. On the contrary, communication is happening throughout a project, and it is often only a matter of degrees that separates how much selecting, polishing, and narrating is involved in preparing for a research group's lab meeting and a formal presentation to a client or partner. Regardless of whether one is engaged in a formal research study or local improvement effort, when working with multiple complex datasets it is often the case that preparing, wrangling, exploring, modeling, and communicating will need to take place in more or less structured ways.

Conclusion

The increasing use of technology in schools and universities is fueling the collection of ever more data on more and more students. Across learning environments of all kinds, there are three major sources of data that data-intensive researchers regularly draw upon: (1) digital learning environments, (2) administrative data systems, and (3) sensors and recording devices. In this chapter, we introduced a data-intensive research workflow that individuals and teams can draw on as they work with these types of data. This workflow is made up of five steps that address key elements of moving from identifying a dataset to producing a data product that answers an important question for researchers, practitioners, or both. This workflow will be used throughout this book. In the next chapter, we focus on three steps: wrangle, explore, and model and describe specific analytical techniques that researchers and data scientists can use in carrying out these steps.