Excerpt from Learning Analytics Goes to School: A Collaborative Approach to Improving Education

By Andrew Krumm, Barbara Means, and Marie Bienkowski

https://www.routledge.com/Learning-Analytics-Goes-to-School-A-Collaborative-Approach-to-Improving/Krumm-Means-Bienkowski/p/book/9781138121836

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

Figure 2.2 Steps of Data-Intensive Research Workflow

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 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 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.

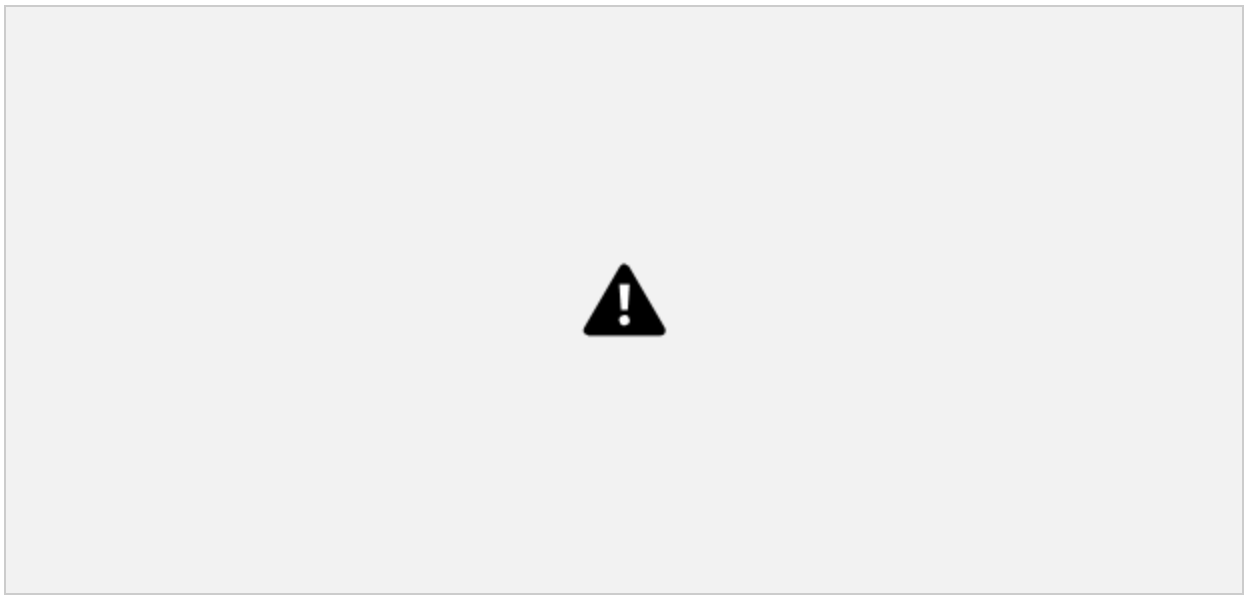

Figure 2.3 Overlapping Activities Within the Data-Intensive Research Workflow