Collaborative Big Data Review for Educational Impact

Tamara Williams, Xiaoyue Cheng, Mahbubul Majumder, Matt Hastings, Hongwook Suh, Kunal Dash, and Jian Ju Yeo

Abstract

Big data is a unique field of study which requires specialized analytics. The field of education has a lot of data: individual student test scores, attendance, behavior, and demographic data are just some of the regularly collected information year after year. Individual student data across an entire state over several years quickly becomes big data. Collaboration between education experts and big data experts is needed in order to maximize the use and impact of educational big data. The goal of big data collaboration is to improve systems and schools in order to serve students most effectively. The purpose of this article is to offer a new conceptual framework titled Prepare, Do, Share as a protocol of collaborative big data review and to share the experience of one such collaboration as a replicable example.

Key Words: big data, collaboration, research practice partnership, higher education, community partner, Prepare Do Share, school attendance

Introduction

There is a lot of individual student data collected in K–12 education. Each year, state departments of education collect individual student data such as individual student statewide assessment scores, daily attendance, behavior infractions, and demographic data. This data set collected year after year provides

a rich source of information that can be used to inform state-level policy recommendations as well as promote school and district improvement. To do so requires both an understanding of big data analytics and a practitioner's understanding of the data itself. Both areas of expertise are needed in order to ensure accurate analysis and interpretation of results. This article introduces the Prepare, Do, Share Framework as a set of protocol steps recommended for collaborative teams in analyzing big educational data.

Background

Big data is a recent term used to describe data sets that are large, continuing to grow, and have variety in the data elements. The concept of 3V (Volume, Velocity, and Variety) is taken from industry analyst Doug Laney (2001), and it is now widely used to describe big data. Big data science is a field of study which requires a unique understanding of both the power and deception of big data. Expertise in statistical approaches, computer programs, and scientific ways of handling big data is a key to harness the valuable information for decision making purposes.

Annual educational student-level state reporting to educational agencies produces large data sets that grow each year. The volume of this data is large. Consider the state of Nebraska where there are over 300,000 students in public education (Nebraska Department of Education, 2017). Each year, school districts report demographic information for each student and other student data including school attendance, school enrollment, participation in specialized services such as limited English proficiency (LEP), special education services, and high ability services (HAL), as well as student outcome data like state test scores over multiple subjects. In any given year, one student can generate several rows of data. Multiplied by 300,000 and growing each year, educational data certainly has robust volume, velocity, and variety.

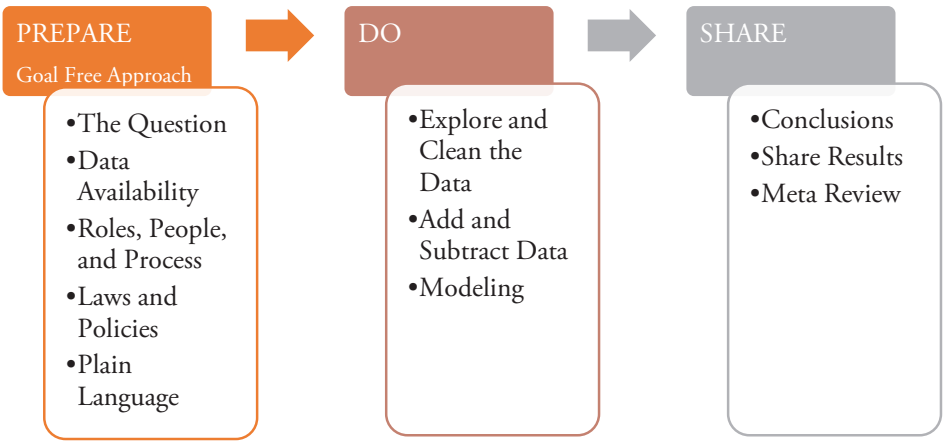
Educational data can be confusing to noneducators. Statistics and big data protocols can be confusing for nonstatisticians. Therefore, a collaborative approach between big data experts and education insiders can help leverage the power of big data and the insight of educational practitioners. These two areas of expertise combined are ideal to investigate problems of practice and uncover solutions that promote education improvement. As established by Biag (2017), in-depth explanations of research-practice partnerships (Coburn, Penuel, & Geil, 2013) are lacking in peer-reviewed academic publications. The framework presented in this article contributes to the literature as an example of research-practice partnership at the state-level.

Prepare, Do, Share Conceptual Framework

This article introduces the Prepare, Do, Share Framework as a set of protocol steps recommended for collaborative teams in analyzing big educational data. This model's structure is similar to Deming's Plan, Do, Check, Act model (1986), which is a process to make collaborative changes that lead to improvement in a manner of continuous quality improvement.

Unlike Deming, the Prepare, Do, Share Framework is rooted in data collaboration, not business. Also, Deming's Plan stage emphasizes establishing clear goals and objectives for the collaborative work. This is deemphasized in the Prepare, Do, Share Framework so as to focus the user on the neutralizing Goal Free approach. The following sections of this article explain the Prepare, Do, Share Framework and provide a detailed example of the framework in action.

Figure 1. Prepare, Do, Share Conceptual Framework



Prepare

The first phase of the Prepare, Do, Share Framework is Prepare (see Figure 1). In Prepare, the collaborative team is setting the stage for efficient communication and trust building. The Prepare phase is when boundaries and protocols are transparently reviewed. When collaborating with data, ethics, integrity, trust, and adherence to laws and protocol are essential.

Goal Free Approach

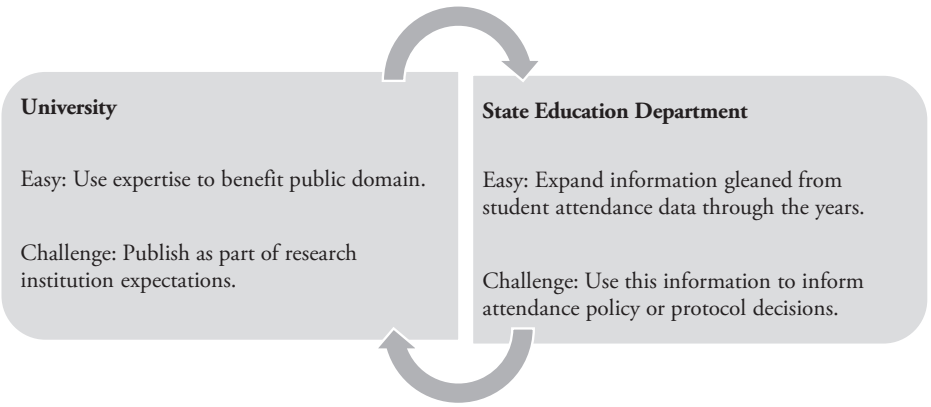
Goal free evaluation (GFE) is when the evaluation team remains intentionally unaware of the project's detailed goals and, instead, analyzes the data interactions neutrally (Mertens & Wilson, 2019). A goal free approach helps remind participants that the exploration process will be governed by neutral

data relationships instead of external policy or political agenda. In other words, all parties agree to set aside personal agendas and let the data relationships initiate further conversation and exploration. Goal free, however, is difficult. Goal free requires self-reflection and team trust. Realistically, all members of a collaborative team have some personal or professional agenda in mind before entering the collaboration (Ferman & Hill, 2004). With a shared understanding of removing external factors and being honest in desired end results, trust and efficiency become accelerators to the project.

PREPARE
Goal Free Approach
- The Question
- Data Availability
- Roles, People, and Process
- Laws and Policies
- Plain Language

Both collaborators in our project were driven by socially acceptable reasons (easy) and personally or politically motivated reasons (challenge). By acknowledging the easy and challenging factors of participation, the collaborative team was able to hold each other accountable if decision making trended toward only benefiting the personal or political motivation. For example, in the rushed desire to publish an interesting result, team members held each other accountable for continued data modeling to confirm consistency in results. It was the responsibility of the entire team to slow down the process and investigate several iterations of the analysis.

Figure 2. Collaborator's Motivation



Additionally, there were many professional details that helped boost the team's successful work towards a goal free approach. The university statisticians had no prior understanding of the educational data elements. This helped ensure the beginning analyses were driven by interesting numerical interactions and not preconceived, hoped-for interactions. The state department team was the bridge between data results and possible implications for policy or protocol decisions. They did not hint at possible actions that might come from the

results of analysis and did not share any expected or desired results. The university team educational leader had practitioner knowledge of the data elements but was not steeped in current state or federal accountability conversations. Team members were patient and allowed the data relationships to determine the direction of collaborative conversation and exploration.

The Question

The collaborative team must, first, determine an interesting and worthwhile question. Big data exploration uses unique resources, so the question needs to be significant enough that when it is answered, the result will help inform more than just local analyses of the same question. The question's answer should either inform state and district policy or have such an impact that it could alter the daily actions of students, teachers, and leaders at the building level.

In our collaborative project, the Nebraska Department of Education was interested in identifying the threshold number of absent days after which the impact on student outcomes necessitates prevention or intervention in Nebraska public schools. The answer to this question could impact policy decisions for federal accountability reporting as well as help determine attendance protocol recommendations for state accountability and local systems of support for students. For the purpose of this project, student outcomes were defined as state math and reading test scores. Future collaboration may also consider on-time graduation as a separate student outcome to be reviewed.

Data Availability

The larger the scope of a data set, the larger the implication of the analysis. As described earlier, annual educational data clearly meets the three V's of big data: volume, velocity, and variety. Existence of data is not an issue. Consistency of the data within the data set and ethical access to the data can limit what is available to the collaborative team.

Many questions can be answered by the data that is already collected, the on-the-shelf data. As researchers, it is easy to brainstorm many possible "what about…" paths of inquiry. These inquiries should be captured and saved for later. First, learn what the available data unearths. Only intrude on others' time and effort when the available data has been exhausted and there is no other possible avenue to answer an important question without collecting new or more data.

Lastly, the data used in the project needs to be available for all members of the data set. For example, educators already collect formative and summative information at the classroom, building, and district levels to inform learning and teaching. However, these local data are rarely captured statewide in a consistent manner. To determine the elements of the data set, consider the

question, "What is the largest geography that collects the same data?" This least common denominator will be the data set for the project.

In our collaborative project, the state education department was seeking a statewide answer. Therefore, the analysis needed to be limited to the data available consistently from all school districts across the state. Multiple years of state-reported data elements such as student-level state test scores, national test scores, attendance, enrollment, and demographics can be included in this analysis. However, classroom assessments, district-level Response to Intervention data, or other local data could not be included in the current analysis because it was not consistent between all school districts.

Roles, People, and Process

Data analysis involves a series of small decisions that aggregate together to limit or possibly sway a result, so having people committed to facts and neutral review is important. A combination of different types of personalities, roles, experiences, and expertise is important. Throughout the project, the team will make critical decisions including data interpretation, adding new data, and eliminating data. Each decision should be tracked, including the date, who was involved in the decision, and the resulting business rule from the decision. Designate a person to keep track of the Process for the data review project.

In our collaborative project, the data team from the state education department provided clarity on elements of the data, the university statisticians provided data expertise, and the third university partner was a former school administrator with a practitioner's understanding of the data elements. Having multiple roles of expertise and diverse professional experiences increased understanding of the decade of student-level data, the anomalies, and relationships of data elements.

Additionally, a secure shared folder between all parties allowed transparency of the process. Data scripts and resulting data images were maintained in an organized manner by the data experts so that all team members could read the data decisions and view the data results. Entire team phone calls were used periodically for formative progress checks and data clarification conversations.

Laws and Policies

Student educational data is regulated by the Family Educational Rights and Privacy Act (FERPA, 1974). State and local data protocols are designed to protect student data and uphold FERPA. Be strict about signed data agreements and data security. Do not take on the risk of highly sensitive data without the technology to protect the data.

In our collaborative project, data was being transferred from the state education department to the university. Administrators, attorneys, and research

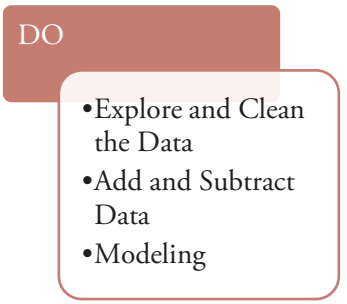
directors from both institutions were involved in creating the data agreement which outlined data transfer, data storage, and data destruction. The length of the agreement was for one year, ensuring that an end date for the project was determined and closure procedures outlined.

Plain Language

The diverse expertise of each professional is typically paired with the academic language of that area of expertise. The challenge is to find non-lingo terms to replace specialized vocabulary. Define acronyms and explain the context of each. Determine dependent and independent variables within the data.

In our collaborative project, many education terms needed to be defined such as individualized education program (IEP), alternate assessment (AA), English language learner levels (ELL level 1, 2, …), contracted services, open enrollment, option enrollment, and so on. In the process of explaining each term, the team created an informal data dictionary which described each data element and the type of data it represented. Similarly, the university statisticians explained modeling decisions in plain language so as to get collaborative feedback in the analysis process.

Do



The second phase of the Prepare, Do, Share Framework is Do (see Figure 1). In Do, the collaborative team is getting their hands on the data and doing math. The Do phase is engaging. This is a time to reference the previous "what about" inquiry list that was set aside earlier. It may be possible to explore some of these inquiries within the cleaned data set.

Explore and Clean the Data

This is, by far, the bulk of work for the team. All data is messy, especially large data sets. Data is messy for many reasons, such as changes in reporting protocols through the years, reporting errors, misunderstanding of a data element, and so on. Large data sets have multiple elements. Plot those elements over time. What general trends emerge? Do the anticipated positive and negative relationships between elements show up? How dispersed is the data? Does the data take on certain shapes?

In our collaborative project, cleaning the data was both straightforward and complex. For example, in the data set, KG and K both reference kindergarten, the year before first grade. In the decade of data, there were times that reporting used KG, and other times it was K. This is a straightforward naming change.

There are also more complex examples, such as the result of district boundary changes across the state through the years. The data also showed that some school districts disappeared, and new districts appeared through the years. What was going on? Nebraska, like other highly rural states, has experienced years of district consolidation and elimination. Thus, even the simple plot of population over time by location was important to review. Additionally, comparing the state data files to government census shape files also helped the team see the years of district reorganization throughout the state. With this understanding of the data, the team decided that data from all districts in all years could be used for the data analysis while using the most recent shape files for data visualization so that the results could be applied to the current reality.

Add and Subtract Data

New variables can be created based on the existing data. Additionally, if data is not pertinent to the questions being explored, it can be eliminated. Each time a data element is removed, review the data for impact using the strategies completed in exploring and cleaning the data.

In our collaborative project, the team needed to calculate new variables and also eliminate other variables from the original data set. Notes for each data decision were kept within the code of statistical programming software R (R Core Team, 2019) and were accessible to the state department team throughout the project. Overall for the project, there were approximately 20 business rules. Some examples of business rules for this project are listed below.

1. *Remove AA.* The team determined that state test scores implied general education state tests and excluded alternate assessments taken by students who have significant cognitive disabilities.
2. *50% FTE.* Some students in the raw data set had a full-time equivalency (FTE) less than half of the day. Only students who spent at least half of their day in school were included in the analyzed data.
3. *In/Out of Nebraska*. Students only in Nebraska for less than a few months in the school year were removed from the data. To be in the analyzed data, students needed to have attended any Nebraska school for at least 75% of a school year, regardless of transferring in and out of Nebraska schools.
4. *1st – 12th*. Due to data anomalies, student attendance data before first grade was eliminated.
5. *Math & Reading only*. Nebraska has consistently tested state math and state reading through the years. Additionally, there have been years of testing state science skills, state writing skills, and state speaking and listening. Due to the consistency of data, the team decided to limit the outcome test

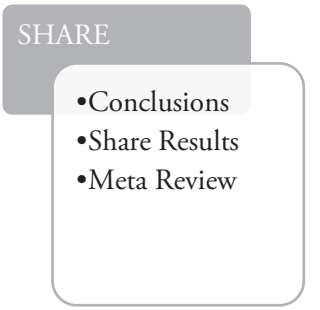
scores in this project to only reading and math and eliminated other test scores from the analyzed data set. The addition of other subject tests, which are not given at all grade levels, did not enhance or change the outcome analysis. Thus, limiting the data set to math and reading (the reading test was replaced by ELA since 2016–17) did not change the shape of the data, but did simplify the data cleaning.

6. *%Ab*. The percent absent (%Ab) is calculated per student per year. Given days present and days absent for each student, each year, the team was able to calculate percent absent from all possible days of enrollment per student per year. The complexity of this variable comes when a student attended multiple Nebraska schools in one year. To calculate this student's percent absent for that year, the team calculated all possible days the student should have been in session for the time enrolled in each district and then calculated all days present for that student while in each district. The ratio, thus, is the percent absent for that student that year.
7. *If ever homeless or highly mobile that year, then yes*. If a student transferred in and out of schools throughout the year, some instances in the data might show a student as homeless or highly mobile, but in another instance in the data that year, the student is not marked as homeless or highly mobile. The team determined that these two markers are risk factors that are important to study. The possible error in overidentifying these two characteristics was reviewed by seeing the impact of changing yes/no for these variables within the model.

Modeling

In our collaborative project, we focused on specific questions to answer based on the data. For this we considered various statistical and machine learning models to fit the data so that those questions could be answered. For example, one burning question is how many school days can a student miss without hampering learning goals? To answer these types of questions, it is important to control over various demographics of the students and other variables related to school characteristics. Statistical modeling is suitable to fit the data in such a way that allows answering the question. It also provides flexibility to incorporate individual random differences. On the other hand, machine learning models are useful to predict student performance and identify features that may be important for achieving a desired outcome. These could provide important information from the data to develop important policy related to student governance, such as allowable missing school days, starting point of early intervention for selective students to make sure they do not fail, and so on.

Share



The third phase of the Prepare, Do, Share Framework is Share (see Figure 1). In Share, the collaborative team shares their results within the team as well as with stakeholders. This is when feedback and outside perspective is important for academic review. Stakeholders, now in possession of the results, will use the information to inform decision making.

Conclusions

The collaborative team will discover answers to the original questions in the project. Additionally, there will likely be many other results that were discovered along the way. Be sure to collect and share all of the results. One benefit of big data collaborative review is the residual information that is developed as part of the process.

In our collaborative project, the Nebraska Department of Education was interested in identifying the threshold number of absent days after which the impact on student outcomes necessitates prevention or intervention in Nebraska public schools. The team answered this question as it relates to student math and reading state test scores. It was found that the absences starting at 4% of the school year (about 6 days) have significant impact on student state test scores. Also, absences totaling around 9% of the school year have the maximum impact on state test scores. While all absences are important, it is essential that schools begin to respond no later than when 4% of the school year is missed.

Additionally, several findings about student attendance already established in the literature were confirmed. For example, attendance is correlated with student achievement, and attendance impacts student populations differently (Epstein & Sheldon 2002; Ginsburg, Jordan, & Chang, 2014; Tanner-Smith & Wilson, 2013).

Share Results

Share the results on both a small and large stage. For each of the expert roles in the collaborative team, share the results with that professional peer group. Statisticians should share modeling and analysis results with other statisticians. Education experts can share with others practicing in the education field. Be open to the feedback. Once the results have survived small group feedback, take the results to a larger audience.

In our collaborative project, the result of decreasing impact of percent absent on state test results after 9% of the year was shared with the state department team. While the university team was satisfied with this result, the director of the state department team asked the important question, "What about when

it starts to tick up?" That is, when does the most dramatic impact on state test scores *begin* so that schools can be sure they are responding no later than that mark? This important question was not considered until the first firm results were shared on a small stage. Another example of a small stage was the university statisticians sharing their modeling and analysis results with other peer statisticians. In line with the data agreements at the beginning of the project, no data was shared with other peers, but the process and theoretical underpinnings were shared, debated, and defended with academic colleagues. Lastly, the state department invited the university team to present at several local education meetings and conferences. This allowed preliminary and final results to be shared with a variety of educators and educational leaders.

Meta Review

Meta review allows the team to reflect on improvements for the future and compare results with what is found in the current literature. In our collaborative project, there were many successes around transparency and trust as described earlier. In future collaborations, we would create a shared, online graphic organizer aligned with the Prepare, Do, Share Framework to keep track of the detailed processes. This graphic organizer could serve as a formative check agenda for collaborative team meetings.

Conclusion

Educational data is a robust data set that easily meets the 3 V's for big data: volume, velocity, and variety. Each year, individual student data reported to the state grows and can be used to inform state-level policy recommendations. Big data analysis for educational impact requires both an understanding of big data analytics and a practitioner's understanding of the educational data itself. Both areas of expertise are needed in order to ensure accurate analysis and interpretation of results. The goal of this collaboration is system improvement to benefit students. Collaborative teams should use the Prepare, Do, Share Framework (see Figure 1) to ensure efficient, accurate, and successful collaboration. The Prepare, Do, Share Framework was introduced in this article as well as detailed examples of the Framework used in action.

References

Biag, M. (2017). Building a village through data: A research-practice partnership to improve youth outcomes. *School Community Journal*, *27*(1), 9–27. http://www.adi.org/journal/2017ss/BiagSpring2017.pdf

Coburn, C. E., Penuel, W. R., & Geil, K. (2013). *Research–practice partnerships at the district level: A new strategy for leveraging research for educational improvement*. W. Grant Foundation.

Deming, W. E. (1986). *Out of the crisis*. Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Epstein, J. L., & Sheldon, S. B. (2002). Present and accounted for: Improving student attendance through family and community involvement. *The Journal of Educational Research*, *95*(5), 308–318.

Family Educational Rights and Privacy Act, 20 U.S.C. § 1232g; 34 CFR Part 99 (1974).

Ferman, B., & Hill, T. L. (2004). The challenges of agenda conflict in higher education–community research partnerships: Views from the community side. *Journal of Urban Affairs*, *26*(2), 241–257.

Ginsburg, A., Jordan, P., & Chang, H. (2014, August). *Absences add up: How school attendance influences student success*. https://www.attendanceworks.org/wp-content/uploads/2017/05/Absenses-Add-Up_September-3rd-2014.pdf

Laney, D., (2001). 3–D data management: Controlling data volume, velocity, and variety (File 949). *Application Delivery Strategies*. META Group.

Mertens, D., & Wilson, A. (2019). *Program evaluation theory and practice: A comprehensive guide* (2nd ed.). Guilford Press.

Nebraska Department of Education. (2017). *Nebraska public schools state snapshot*. https://nep.education.ne.gov/statedata.html

R Core Team. (2019). *R: A language and environment for statistical computing*. https://www.R-project.org/

Tanner-Smith, E., & Wilson, S. (2013). A meta-analysis of the effects of dropout prevention programs on school absenteeism. *Prevention Science: The Official Journal of the Society for Prevention Research, 14*(5), 468–478.

Tamara Williams is an assistant professor in the Educational Leadership Department at the University of Nebraska at Omaha. Research interests include culturally responsive leadership, data-informed leadership, school improvement, program evaluation, and school and community partnerships. Correspondence concerning this article may be addressed to Dr. Tamara Williams, 6001 Dodge Street, Roskens 312, Omaha, NE 68182, or email tamarawilliams@unomaha.edu

Xiaoyue Cheng is an assistant professor in the Department of Mathematics at the University of Nebraska at Omaha. Research interests include data visualization and interactive graphics, data exploratory analysis, data mining, and statistical classification.

Mahbubul Majumder is an associate professor in the Department of Mathematics at the University of Nebraska at Omaha. Research interests include exploratory data analysis, data visualization and visual inference, statistical modeling, and data science.

Matt Hastings is the senior administrator for the Office of Data, Research & Evaluation at the Nebraska Department of Education. Research interests include education workforce longitudinal data systems, school–community partnerships, and education policy.

Hongwook Suh is the director of research and evaluation and psychometrician lead for the Nebraska Department of Education. Research interests include test development, data management, and school outcomes.

Kunal Dash is a statistical research analyst for the Nebraska Department of Education. Research interests include improving student outcomes.

Jian Ju Yeo is also a statistical research analyst for the Nebraska Department of Education. Research interests include improving student outcomes.