MORF: A Post-Mortem

MORF: A Post-Mortem

Ryan S. Baker

University of Pennsylvania, ryanshaunbaker@gmail.com

Stephen Hutt

University of Denver, Stephen.Hutt@du.edu

There has been increasing interest in data enclaves in recent years, both in education and other fields. Data enclaves make it possible to conduct analysis on large-scale and higher-risk data sets, while protecting the privacy of the individuals whose data is included in the data sets, thus mitigating risks around data disclosure. In this article, we provide a post-mortem on the MORF (MOoc Replication Framework) 2.1 infrastructure, a data enclave expected to sunset and be replaced in the upcoming years, reviewing the core factors that reduced its usefulness for the community. We discuss challenges to researchers in terms of usability, including challenges involving learning to use core technologies, working with data that cannot be directly viewed, debugging, and working with restricted outputs. Our post-mortem discusses possibilities for ways that future infrastructures could get past these challenges.

CCS CONCEPTS • Human-centered computing; • Social and professional topics; • Applied computing → Education; Interactive learning environments

Additional Keywords and Phrases: Data enclave, MORF, Privacy

ACM Reference Format:
First Author's Name, Initials, and Last Name, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

1 INTRODUCTION

Learning analytics and educational data mining began with widely available data sets. Accounts of the early history of the field (e.g., [6, 26, 5]) credit the availability of data repositories like the PSLC DataShop [18] and landmark data sets such as OULAD [19] as being one of the key enabling conditions for the emergence of the field. No longer was participation in the field limited to researchers at the small number of institutions that had their own data; the broader community of scientists could conduct analyses, build models, and debate the meaning of specific findings in the context of the data itself.

However, the specific data sets that became available shaped the research questions studied, in both positive and negative fashions. Many of the large data sets available at scale involve responses to straightforward items, where answers can be graded as a binary correct or incorrect, items are tagged with skills, and timing data is available. This has led to a focus on a small number of research topics, such as algorithms that attempt to predict future student correctness and prediction of MOOC stopout. In the specific case of MOOCs, other data -- involving textual data from MOOC forums -- has been less available and, therefore, less used. In general, concerns around data privacy have limited the availability of many types of educational data where it could be possible to re-identify students, such as discussion forums, collaborative chat, and videos of classrooms. Some data of these types has been shared, but to a much more limited degree than correctness and stopout data.

In other fields, this challenge of sharing sensitive data has led to the development of data enclaves: secure, controlled environments where sensitive or confidential data can be accessed and analyzed by researchers [20]. Data enclaves often provide tools for analysis within the environment but restrict the ability to download or remove raw data to ensure confidentiality [23]. Data are used in a range of fields, including healthcare [24, 21], economics [22], and geospatial data [17].

Broadly, over the last few years, public concern about data privacy has grown [28], legislation worldwide around data privacy has expanded [16], and the desire for privacy-protecting solutions such as data enclaves has expanded. In education, the interest in data enclaves has grown sufficiently that large-scale projects have now commenced to create data enclaves, such as the NSF-funded SafeInsights project.

In this paper, we present a retrospective on the MORF data enclave [15], one of the more prominent data enclave projects in the educational research community. MORF began in 2017, and is anticipated to merge into the SafeInsights project upon completion of that infrastructure. As such, though MORF lives on today, its days are likely numbered, and -- with development about to begin on SafeInsights as of this writing -- it seems like a relevant time to consider MORF's successes and failures. In a sense, this article can be seen as a post-mortem on an infrastructure that is not quite fully dead, but headed in that direction.

At its height, MORF contained data from millions of students taking hundreds of MOOCs at three universities, and contributed to a range of scientific projects that would have been highly difficult to conduct without its specialized data. And yet, despite the uniqueness of its data and the opportunity it offered, it was barely used at all (and only briefly so) outside of its home universities. It was an infrastructure and enclave that routinely generated enthusiasm at conferences, and seldom generated any use. Despite the rapid expansion of privacy-protection legislation worldwide, MORF's open and publicly available source code never seems to have been adopted beyond its initial university base (unlike other data repositories such as the LearnSphere of distinct repositories that emerged from the PSLC DataShop -- [18]).

As such, in this article, we come not to praise MORF, but to bury it -- to discuss the lessons learned from this lovely but unloved infrastructure, to understand its failures, and to report the challenges that we ultimately failed to solve. We do so in the hopes that by being transparent and candid about this work, future data enclaves in education, including SafeInsights, will learn from our mistakes, fix the problems we failed to fix, and solve the challenges that must be solved. Only by doing so can data enclaves achieve their full potential to become part of the future practice of learning analytics research.

2 HISTORY OF MORF

MORF's beginnings were very different from its eventual development as a data enclave. Several key elements were already in place, though. MORF's name -- the MOoc Replication Framework -- reflects those key elements. First, it involved MOOCs. And second, it supported replication. The initial vision of MORF was a framework that would support analysis on MOOC data without allowing direct access to or viewing of the data. But the initial vision supported a much narrower range of analysis. In its first version (internally referred to as 1.0), MORF supported researchers in specifying if-then production rules [9], where the “if” clause was some aspect (or set of aspects, with an AND between them) of the student’s behavior or a MOOC, and the “then” clause was some outcome. The “then” clause was typically whether the student completed the course with a certificate or not. The researcher could then test how many MOOC courses the finding held across, and overall statistical significance of the finding. For example, [1] examined whether 15 aspects of student behavior predicted whether a student earned a certificate, across 29 instances of 17 courses. An example of such an if-then rule was:

```
IF Participant indicated they are taking the course for Credit
THEN Participant is likely to earn a certificate.
```

The second version of the framework (2.0) extended the framework’s functionality considerably, enabling researchers to conduct arbitrary analyses involving predicting an outcome, but still without the ability to directly access or view data. This framework, first described in [10] and revised and described in greater detail in [15] (version 2.1), required users to input their code into a Docker container, and to use pre-approved output functions.

At its peak of number of data providers, MORF included data from MOOCs at three universities: the University of Pennsylvania, the University of Michigan, and the University of Edinburgh. At that time MORF had data from over 1.5 million distinct students taking hundreds of courses. However, since then, both Michigan and Edinburgh have discontinued participation in MORF, and only the University of Pennsylvania remains actively involved. However, the amount of data available has grown to over 10 million distinct students.

Although several research groups started projects involving MORF, the only projects that have reached the point of publication directly involved at least one researcher with some level of affiliation at University of Pennsylvania or the University of Michigan. However, this research included analyses of the factors leading to failure to complete a certificate [1], the differences in learning benefits between on-demand hints and automated scaffolding [29], attempting to replicate published difference in algorithm performance [11, 13], studying new algorithmic bias evaluation techniques [12], studying demographic differences in participation in MOOC discussion forums [3], quantitative ethnography analyses of the impact of COVID-19 on student participation in MOOCs [25], and investigating the generalizability of predictive models across countries [2].

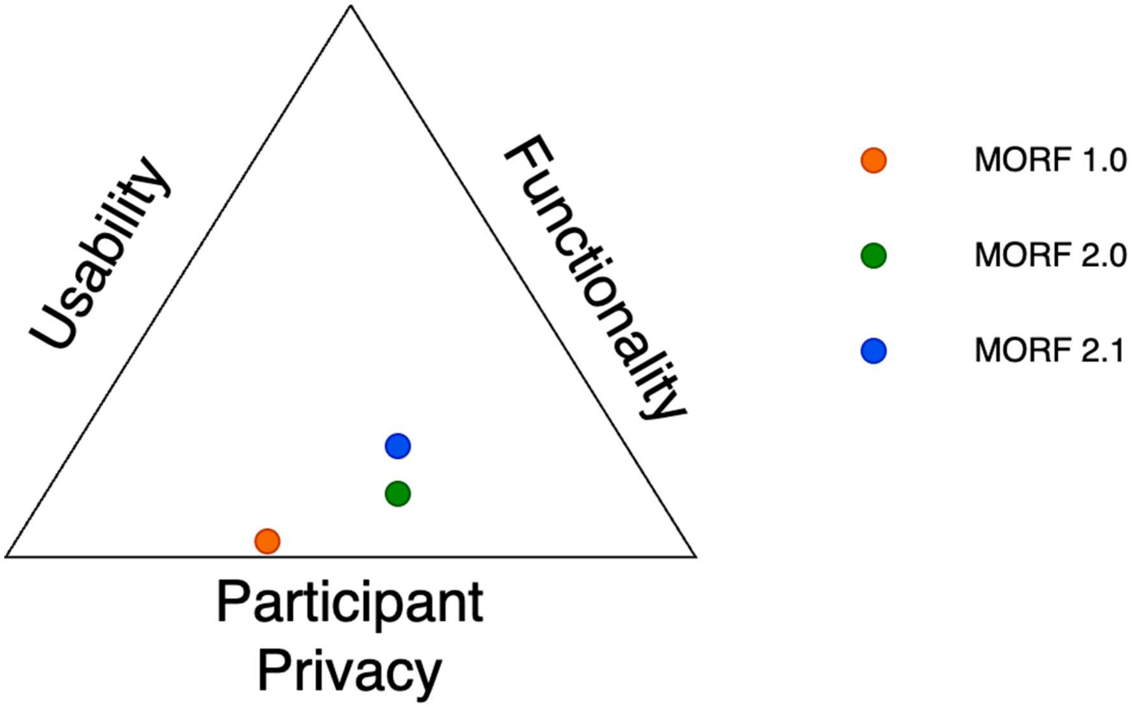

Figure 1. The Trade-off Between Factors Critical to the use of a Data Enclave.

3 TRADE-OFFS AND CHALLENGES

The initial design of MORF (both 1.0 and 2.0) prioritized student and instructor privacy above all other considerations. Given the nature of a privacy-preserving enclave, this focus was essential to ensuring the confidentiality and security of user data and critical to university approval of MORF data sharing. Our goal has always been to increase accessibility of data, maximizing potential analysis without the need to release personally identifying information (PII).

However, over time, MORF has encountered a trade-off between three critical factors: participant privacy, functionality, and usability (See Figure 1). While our design successfully maximized privacy, the constraints introduced by these measures have led to limitations in the system's functionality and usability. Although revisions to MORF have aimed to improve usability and functionality without compromising participant privacy, these adjustments have not yielded sufficient improvements to lead to substantial researcher adoption. In this section, we discuss the rationale behind these design decisions, the challenges they created (including what we learned from informal feedback obtained from users throughout the life of the project thus far), and the strategies employed to address them.

3.1 Language agnostic design

MORF was intentionally designed to be programming language agnostic, meaning that we, as its creators, did not want to impose the use of any particular programming language or analysis method. Recognizing the diversity of analytical approaches, we sought to accommodate as many languages as possible. While R and Python may have been the most commonly used in the field, we did not wish to limit someone who wished to use Matlab or Stata for example. To achieve this, MORF was designed to accept Docker containers, which allowed users to specify the environment for their analyses, including the programming language, version, and required libraries. This design decision was rooted in our desire to maximize accessibility. By not restricting users to a particular language or analysis method, we aimed to expand the system's potential use cases, enabling researchers to freely configure their projects within the MORF framework.

However, early usage revealed that the requirement to submit Docker containers created an unintended barrier for many users. A significant number of users lacked the technical expertise to create the necessary Docker containers, which hindered their ability to run analyses within the framework. As [14] notes, Docker introduces a number of significant technical challenges for new developers. To better support less experienced developers, subsequent versions of MORF embedded the process of creating Docker containers into the pipeline itself. Users were only required to specify the container setup, and we provided several common use-case examples as templates. These templates included a container for Scikit-learn analysis in Python and another for R, preloaded with common libraries.

Despite these improvements, we continued to receive feedback that the Docker container process remained confusing, with users unsure about what configurations were necessary. Although we offered additional resources to guide users, the process still presented a critical usability challenge.

Reflecting on this, we find ourselves attempting to balance functionality with usability. While language-agnostic design (facilitated by Docker) broadens MORF's functionality by accommodating a wide variety of analyses, it simultaneously reduces usability for users unfamiliar with Docker technology. This presents a paradox: improving functionality can lead to decreased usability, ultimately preventing some analyses from being conducted. A key consideration for the future is whether specifying a single language and providing more detailed resources and support would be more effective, even if it limits the system's flexibility.

3.2 Data use vs. data visibility

At the core of MORF, and privacy-preserving enclaves more broadly, is the protection of participant privacy. While MORF allows users full access to data for the purpose of analysis, the data itself is not directly visible to users. This means that manual exploration of the data is not possible. This design decision was driven by the need to maintain strict privacy standards, ensuring that raw data never leaves the enclave in compliance with data agreements.

However, we have received feedback from users indicating the need for some form of data visibility during the analysis design phase. Users expressed that without being able to view the data, it is difficult to fully account for its nature when developing an analysis plan. To address this need while still preserving privacy, we opted not to allow users to manipulate the data directly in its original form, which could have added unnecessary complexity. Instead, we provided detailed data schemas for all datasets within MORF, along with additional resources from the original data providers that offered comprehensive descriptions of the datasets. However, some users still reported that this was unacceptable.

To further support users in their development process, we created synthetic versions of the data. These synthetic datasets could be shared openly, allowing users to design and test their analysis code offline. Once their code was ready, it could be submitted to MORF and run on the full, authentic dataset. The synthetic data was intended to facilitate offline development while maintaining the privacy protections for real participant data.

Despite these efforts, some users found the sheer volume of data in MORF overwhelming. The challenge of narrowing down the vast datasets to something relevant to their research question(s)—both conceptually and in terms of writing effective SQL queries—proved to be a significant hurdle. Users reported this hurdle was exacerbated by not being able to view the data or inspect individual tables, to see if data were relevant to their research.

One potential solution to this challenge is to reduce the degrees of freedom in the system. Instead of allowing users to query the data in any way they wish, we could introduce a layer that assists in query construction, guiding users within a set of predefined parameters. While this approach could reduce flexibility for some users, it has the potential to improve usability, particularly for those who struggle with crafting complex data queries or navigating large datasets.

3.3 Debugging

A key aspect of any programming task is debugging. It is rare for code to function perfectly on the first attempt, and MORF jobs are no exception to this. However, we could not guarantee that providing detailed debug output would preserve participant privacy due to the nature of error messages—which often include snippets of the data used. To address this concern, we initially suppressed detailed error messages and replaced them with more generic feedback. Users would receive high-level error types (e.g., "Syntax Error" or "Runtime Error"), but without specific information about what caused the error.

This approach posed significant challenges for users trying to debug their code within the MORF environment. Without access to detailed error messages, it was difficult for them to identify and resolve the root causes of issues in their code. One attempt to mitigate this problem was through the introduction of synthetic data. Users were encouraged to test their code offline with full access to debug output before submitting their jobs to MORF. While this solution addressed some debugging needs, it did not resolve SQL errors or broader issues related to Docker, as those were not part of the offline testing environment.

To provide further support, we introduced a human-in-the-loop mechanism, where a member of the MORF team would review error messages and either release them or redact them as necessary to provide additional information to the user. This allowed for more targeted feedback while still maintaining privacy protections.

Despite this effort, the support mechanism did not yield the desired improvement in the debugging experience. One possible explanation relates to the delay introduced by this process. Programming tasks often involve a cycle of quickly identifying and fixing issues, but the need to email the MORF team, wait for support, and then re-submit jobs interrupted this flow. For users who encountered multiple sequential errors—a common occurrence in programming—this fragmented the debugging process into multiple, shorter bursts, reducing the efficiency and fluidity of troubleshooting.

This experience highlights the tension between privacy and usability. Users have a need for real-time debugging support, and this raises the question of how to streamline error reporting while maintaining necessary privacy safeguards.

3.4 Restricting outputs

While MORF provided considerable freedom in analysis (in theory, at least), we had to impose stricter controls on the format of output data to ensure the privacy of participants and instructors. Users were not allowed to self-define their outputs, as this would have posed significant risks. If allowed unrestricted outputs, a user could, for example, output the entire database or a subsection of it, circumventing privacy protections. Additionally, more sophisticated methods, such as encoding the database and outputting it, could bypass simple detectors intended to prevent unauthorized access to raw data.

To mitigate these risks, we limited the format of output data to pre-defined structures, which could then be processed and securely delivered to users. Given MORF's focus on predictive modeling, the output was primarily restricted to traditional predictive modeling metrics, such as accuracy, precision, recall, and other standard performance measures. While this approach preserved privacy, it also introduced an unintended limitation on the kinds of analyses that could be conducted within MORF and highlighted a tension between privacy and functionality.

To address these constraints, we offered users the opportunity to design their own output functions for the framework, which could then be submitted for review. Once reviewed and approved, these functions were integrated into the platform, broadening the range of analytical possibilities. Additionally, we invited users to suggest output types or methods of analysis for cases where they did not wish to develop their own functions, such as Epistemic Network Analysis (ENA). We then built functionality for these types of output into MORF, which became used in research (e.g., [25]).

Despite these efforts, we observed that the amount of work required to customize analysis outputs remained a barrier for many users. Those who preferred not to use the existing output functions often found the process of developing new ones too burdensome, leading some to abandon the platform altogether.

As we look toward the future of privacy-preserving enclaves like MORF, the need to control output data remains essential. However, it is clear that a more streamlined and flexible approach is required—one that balances the need for privacy with the freedom for users to define outputs that suit their research objectives. Developing such a system will be key to enhancing the usability of MORF or platforms like it without compromising the integrity of its privacy protections.

3.5 Reflection on challenges

In reflecting on the challenges encountered with MORF, it is clear that while we succeeded in prioritizing participant privacy and ensuring a flexible, language-agnostic platform, these strengths often came at the cost of usability and functionality. The trade-offs we faced—between privacy and user experience, between flexibility and ease of use—highlight the complex nature of designing a privacy-preserving system that meets the diverse needs of researchers. Our efforts to mitigate challenges were hopefully steps in the right direction but did not fully address the barriers users faced in practice, and thus did not solve the problems. Despite these shortcomings, it is our hope that the lessons learned will be valuable. Hopefully, future data enclaves in education can refine and simplify these systems, making them more accessible without compromising on the fundamental principles of privacy and security. It is our hope that the lessons learned from MORF can lead to a next generation of effective and usable tools for the research community.

4 WHAT’S NEXT?

The need for privacy-preserving enclaves, especially in education and learning settings, does not seem likely to go away. As more funders require data-sharing protocols and policies and we observe a greater push to transparency in science, the need for data-sharing solutions that protect student and instructor privacy is only growing. The very nature of the work we do as a community, often with learner populations under 18, often involving PII, means that sharing data is a complicated problem [4]. Increased legislation around these topics also places additional requirements on researchers [16], so the question then becomes, how do we do better next time?

Ultimately, future enclaves in education will need to solve many of the problems that we were unable to fix in the MORF infrastructure. Despite considerable initial interest by external partners, initial enthusiasm often turned to frustration and disengagement given the various challenges involved in using the platform. Users struggled with Docker, struggled with debugging, and struggled with working around limited outputs. These human-computer interaction challenges proved sufficiently large that no project was able to succeed in MORF that did not have members of Penn or

Michigan involved in some substantial fashion (even in the few cases where these involvements were informal). As such, the development of data enclaves that are more successful than MORF will ultimately involve a more successful partnership with users than we were able to achieve. Co-design and usability testing were part of our approach to developing MORF throughout, and we continually sought user feedback (if informally). That user feedback informed our design and also informs the discussion within this article. However, there were limitations to these efforts. For one thing, though we conducted usability testing internally (including with individuals who were not technical experts), we did not go beyond our internal team and close colleagues to a wider set of potential users. This turned out to be an error, as our internal users turned out to be more willing to put in the effort to learn the new technologies required than eventual external users. As such, we recommend that future such efforts bring in a broader set of external users for usability and co-design research.

Beyond that, it still remains a question as to how the challenges encountered by MORF can be surmounted. Is it as simple as pouring more resources into usability engineering and co-design partnerships? Or are there more fundamental challenges to designing data enclaves for use by educational researchers? It seems like an obvious recommendation to suggest that members of our community actively research and deploy the solutions used in other communities, but this is what we did, through the use of tools such as Docker and active participation at National Science Foundation meetings whether other communities presented their tools and their lessons learned. Somehow, it was not enough.

The other possibility is that the fundamental issue for MORF was not the quality of our solution but the demand for it. The learning analytics and educational data mining communities have been unusually successful at creating a range of open data sets that can be downloaded on demand or with minimal effort. With all this data available, the cost of learning to work with MORF may simply have been hard to justify for external researchers who could simply investigate different research questions than the limited set that could only be done with the type of data found in MORF. The unique studies possible using MORF involving analysis of learners across large numbers of MOOCs, the same learner across multiple MOOCs, or unrestricted discussion forum data, may simply have been insufficiently compelling to justify the considerable additional work required to use MORF. If this is the case, then future data enclaves may want to increase initial demand (and therefore build a community of scholars able to use their functionality) by offering incentives to users, such as research funding. This initial seed community may in turn bring in a wider community, both by supporting collaborators in getting started, and by publishing their code within the enclave, encouraging reproduction of analysis and progressive science (building on past analyses).

Another possibility for the future is that alternate technologies will become sufficiently useful and powerful to eliminate the need for data enclaves in educational research. In this vision, data enclaves such as MORF will become the Allotheria of data protection techniques -- a direction that seemed promising for a time, but ultimately lost out. For example, recent work in automated deidentification suggests that methods that use some user info (such as a list of names) can be highly effective [8] and that even methods with no external information can capture a large proportion of PII, even catching mistakes in deidentification made by human researchers and not caught by other human researchers [27]. If the next generation of large language models and approaches for using them can reach closer to 100% recall of PII (including not just names but other information that can identify a learner), there may no longer be a need for data enclaves. Similarly, approaches like obfuscation [7] may reach a point where we are confident that demographic data can be shared without risking a specific individual being identified by it.

Ultimately, the field will need to find a way to satisfice all elements of the satisfaction triangle we have presented above: offering sufficient ability to analyze learner data for a range of purposes (functionality), in a way that is feasible for a range of researchers (usability), and that properly protects the privacy of the participants who contributed the data (privacy). We believe that data enclaves remain an important part of the solution, and hope that the lessons learned from the MORF project will contribute to these developments, even if in the end the MORF infrastructure itself joins the Allotheria, the Multics operating system, and the Zeppelin in history's list of initially exciting approaches that didn't quite work out as well as might have initially been hoped for.

ACKNOWLEDGMENTS

We would like to thank NSF grants #DUE-2153481 and #DRL-1931419. We would also like to thank the MORF development team, including but not limited to Christopher Brooks, Josh Gardner, Miggy Andres, Michael Mogessie, and Maciej Pankiewicz, and all of the MORF platform's users.

REFERENCES

[1] Juan Miguel L. Andres, Ryan S. Baker, Dragan Gašević, George Siemens, Scott A. Crossley, and Srećko Joksimović. 2018. Studying MOOC completion at scale using the MOOC replication framework. In Proceedings of the 8th international conference on learning analytics and knowledge, 71-78.

[2] Juan-Miguel Andres-Bray, Stephen Hutt, and Ryan S. Baker. 2023 Exploring Cross-Country Prediction Model Generalizability in MOOCs. In Proceedings of the Tenth ACM Conference on Learning@ Scale, 183-194.

[3] Juan-Miguel L. Andres-Bray, Jaclyn L. Ocumpaugh, and Ryan S. Baker. 2019. Hello? Who is posting, who is answering, and who is succeeding in Massive Open Online Courses. Poster paper. In Proceedings of the 12th International Conference on Educational Data Mining (EDM), 492-495.

[4] Ryan S. Baker, Stephen Hutt, Christopher A. Brooks, Namrata Srivastava, and Caitlin Mills. 2024. Open Science and Educational Data Mining: Which Practices Matter Most?. In Proceedings of the 17th International Conference on Educational Data Mining (EDM), Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society, 279–287.

[5] Ryan S. Baker and George Siemens. 2014. Learning analytics and educational data mining. In Cambridge handbook of the leaning sciences, Third Edition.

[6] Ryan S. Baker and Kalina Yacef. 2009 The state of educational data mining in 2009: A review and future visions. Journal of educational data mining 1, 1, 3-17.

[7] David E. Bakken, R. Rarameswaran, Douglas M. Blough, Andy A. Franz, and Ty J. Palmer. 2004 Data obfuscation: Anonymity and desensitization of usable data sets. IEEE Security & Privacy 2, 6, 34-41.

[8] Nigel Bosch, R. Crues, Najmuddin Shaik, and Luc Paquette. 2020 "Hello, [REDACTED]": Protecting Student Privacy in Analyses of Online Discussion Forums. In Proceedings of the International Conference on Educational Data Mining.

[9] Ernest Friedman-Hill. 2003. Jess in action: rule-based systems in Java. Simon and Schuster.

[10] Josh Gardner, Christopher Brooks, Juan Miguel Andres, and Ryan S. Baker. 2018a. MORF: A framework for predictive modeling and replication at scale with privacy-restricted MOOC data. In 2018 IEEE International Conference on Big Data (Big Data), 3235-3244. IEEE.

[11] Josh Gardner, Christopher Brooks, Juan Miguel Andres, and Ryan Baker. 2018b. Replicating MOOC predictive models at scale. In Proceedings of the Fifth Annual ACM Conference on Learning at Scale, 1-10.

[12] Josh Gardner, Christopher Brooks, and Ryan Baker. 2019b. Evaluating the fairness of predictive student models through slicing analysis. In Proceedings of the 9th international conference on learning analytics & knowledge, 225-234.

[13] Josh Gardner, Yuming Yang, Ryan S. Baker, and Christopher Brooks. 2019a. Modeling and Experimental Design for MOOC Dropout Prediction: A Replication Perspective. International Educational Data Mining Society.

[14] Mubin UI Haque, Leonardo Horn Iwaya, and M. Ali Babar. 2020. Challenges in docker development: A large-scale study using stack overflow. In Proceedings of the 14th ACM/IEEE international symposium on empirical software engineering and measurement (ESEM), 1-11. https://dl.acm.org/doi/abs/10.1145/3382494.3410693

[15] Stephen Hutt, Ryan S. Baker, Michael Mogessie Ashenafi, Juan Miguel Andres-Bray, and Christopher Brooks. 2022. Controlled outputs, full data: A privacy-protecting infrastructure for MOOC data. British Journal of Educational Technology 53, 4, 756-775.

[16] Stephen Hutt, Sanchari Das, and Ryan S. Baker. 2023. The Right to Be Forgotten and Educational Data Mining: Challenges and Paths Forward. In Proceedings of the 16th International Conference on Educational Data Mining (EDM 2023). International Educational Data Mining Society. https://eric.ed.gov/?id=ED630886

[17] David Johnson, Mohammad Mushtaq, Nishaad Rao, and Noura Insolera. 2022. Updating the Geospatial Data in the PSID Restricted Data Enclave for 2005-2017. https://psidonline.isr.umich.edu/publications/Papers/tsp/2022-01_Geocode_Update.pdf

[18] Kenneth R. Koedinger, Ryan SJd Baker, Kyle Cunningham, Alida Skogsholm, Brett Leber, and John Stamper. 2010. A data repository for the EDM community: The PSLC DataShop. Handbook of educational data mining 43, 43-56.

[19] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. 2017. Open university learning analytics dataset. Scientific data 4, 1, 1-8.

[20] Julia Lane and Claudia Schur. 2010. Balancing access to health data and privacy: a review of the issues and approaches for the future. Health services research 45, 5p2, 1456-1467.

[21] Daniel Meza, Basil Khuder, Joseph I. Bailey, Sharon R. Rosenberg, Ravi Kalhan, and Paul A. Reyfman. 2021. Mortality from COVID-19 in Patients with COPD: A US Study in the N3C Data Enclave. International Journal of Chronic Obstructive Pulmonary Disease, 2323-2326.

[22] Abhishek Nagaraj and Matteo Tranchero. 2023. How does data access shape science? Evidence from the impact of US Census's research data centers on economics research. No. w31372. National Bureau of Economic Research.

[23] Beth A. Plale, Eleanor Dickson, Inna Kouper, Samitha Harshani Liyanage, Yu Ma, Robert H. McDonald, John A. Walsh, and Sachith Withana. 2019. Safe open science for restricted data. Data and Information Management 3, 1, 50-60.

[24] Richard Platt and Tracy Lieu. 2018. Data enclaves for sharing information derived from clinical and administrative data. JAMA 320, 8, 753-754.

[25] Jade Pratt and Marie Gordon. In press. The Impacts of COVID-19 on Student Interaction in a Massive Open Online Course. In Sixth International Conference on Quantitative Ethnography

[26] George Siemens. 2013. Learning analytics: The emergence of a discipline. American Behavioral Scientist 57, 10, 1380-1400.

[27] Shreya Singhal, Andres Felipe Zambrano, Maciej Pankiewicz, Xiner Liu, Chelsea Porter, and Ryan S. Baker. 2024. De-Identifying Student Personally Identifying Information with GPT-4. In Proceedings of the 17th International Conference on Educational Data Mining, 559-565.

[28] Marshall W. Van Alstyne and Alisa Lenart. 2020. Using data and respecting users. Communications of the ACM 63, 11, 28-30. https://doi.org/10.1145/3423998

[29] Yiqiu Zhou, Juan Miguel Andres-Bray, Stephen Hutt, Korinn Ostrow, and Ryan S. Baker. 2021. A comparison of hints vs. scaffolding in a MOOC with adult learners. In Proceedings of the International Conference on Artificial Intelligence in Education, 427-432. Cham: Springer International Publishing.