Global Data Consortium: Artificial Intelligence's Essential Role in the Future of Universities

George Siemens, Srecko Joksimovic, Damien Coyle, Walter Reilly, Shane Dawson, and Rose Luckin

American Council on Education®

Working Draft as of May 1, 2024

ACE and the American Council on Education are registered marks of the American Council on Education and may not be used or reproduced without the express written permission of ACE.

American Council on Education
One Dupont Circle NW
Washington, DC 20036

© 2024. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

EXECUTIVE SUMMARY

Objective: The Global Data Consortium (GDC) seeks to place academic institutions at the forefront of artificial intelligence (AI) innovation by promoting collaboration among researchers, colleges, industry leaders, and governments. The GDC aims to democratize access to educational data and leverage the potential of AI to address disadvantages and enhance learning and teaching outcomes for all.

Benefits: The consortium promises accelerated AI research, development of AI models tailored to education, access to diverse datasets, and improved AI capabilities within institutions. It also aims to increase equity and graduation rates in education through AI support.

Access for Institutions: Participating institutions will gain access to a secure data infrastructure, a community of experts, AI models and algorithms, and opportunities for collaborative research and professional development. Additionally, they will be at the forefront of cutting-edge practices, leveraging the latest AI advancements in practical, real-world educational applications to drive innovation.

Responsible AI Practices: A key focus is on using synthetic data to protect privacy and minimize algorithmic biases. The consortium will establish best practices in policy work to guide ethical and responsible AI use in education.

Technical Framework: This document describes a decentralized data architecture based on the principles of data mesh, an innovative approach that emphasizes distributed data ownership and scalable, agile data management. It outlines the infrastructure's scalability, performance, and security features, emphasizing the importance of self-serve data infrastructure and federated governance.

Ethical, Responsible, and Representative: The technical design of the data consortium ensures that student data remains secure, private, and under the direct control of the member institution. The GDC focuses on membership from all institutions, including Historically Black Colleges and Universities (HBCUs), Hispanic-Serving Institutions (HSIs), and small regional institutions.

INTRODUCTION

Postsecondary education faces persistent and increasing challenges reflected in stagnant completion rates, concerning equity gaps, soaring costs, gaps between skills required by graduates for successful employment and skills developed in the education system, and the need to respond to new technologies. At the same time, there are pressing concerns related to the transformative impact of AI on teaching and learning practice. While a few institutions are grappling with this emerging technology, the majority of colleges and universities are currently peripheral to the AI conversation.

The impetus for change and innovation has been driven by big tech (Google, Microsoft, Amazon, Nvidia, Tencent, Open AI), leaving the postsecondary sector underprepared and absent from central initiatives. However, the massive educational data that exists in universities can provide insights and breakthroughs in understanding teaching practices, the learning process as it unfolds, and how to support *all* learners holistically through AI models. Data quantity and quality determines the quality of AI models. Development of foundation models—the basis of tools like ChatGPT and Gemini—requires gigabytes or even petabytes of high-quality data, reflective of data agreements and web scraping with sites like X, Reddit, and YouTube.

However, much of the education data is unstructured and scattered, resulting in missed opportunities to create AI products that are tailored to the specific needs of education. The American Council on Education is establishing the GDC to directly respond to this challenge. The GDC enables institutions to innovate and position universities globally as central agents for advancing AI technologies in education. By fostering a collaborative ecosystem, this consortium will harness the collective expertise and resources to drive breakthroughs in AI research and its application in postsecondary education.

Universities currently have a limited and peripheral role in the evolution and deployment of AI. A data consortium will enable sharing of data, capabilities, models, and expertise and can be transformative in postsecondary education. This consortium will push postsecondary education to the forefront of the conversation and accelerate AI-supported educational technology. This collaborative approach advances the AI development process and ensures that the products are highly tailored to enhance educational outcomes.

GLOBAL CONSORTIUM: VISION AND LOGIC

The GDC is envisioned as a cooperative network of colleges and universities, their leaders, and experts dedicated to the responsible collection, sharing, and utilization of data for developing AI products that benefit students. It is a group of organizations that pool data in order to achieve outputs that are not possible for them as single entities. This is a proven approach, as demonstrated in sectors like finance for fraud detection (e.g., FICO's Falcon Intelligence Network) and health care, where shared data enables the system as a whole to function more impactfully for end users (e.g., Cancer Research Data Commons, Chaimeleon, European Health Data and Evidence Network).

Our vision for the GDC is to create a sustainable model and platform that accelerates AI innovation and addresses systemic challenges through advanced data analytics and machine learning. The logic underpinning this vision is grounded in the belief that shared data, knowledge, and resources can catalyze significant advancements in AI. Leveraging the wellspring of university data will result in responsibly and ethically developed models and products that can be deployed in universities.

PREVIOUS EDUCATION CONSORTIUM ATTEMPTS

Data consortia are not new to the educational sector. Numerous existing projects have attempted to engage in sharing but have generally been limited in scope and impact. These projects fall into roughly five categories: surveys, massive open online courses (MOOCs), LearnSphere, data infrastructure "localized" to individual institutions, and aggregate initiatives.

It is important to acknowledge the historical precedent of postsecondary education data collected through national surveys. In 1962, the Inter-university Consortium for Political and Social Research (ICPSR) was one of the first initiatives to grant open access to a centralized store of critical research datasets. Around the same time, the need for a means to measure academic proficiency nationwide gained support, and the National Assessment of Educational Progress (NAEP) was established. The oldest and most comprehensive survey of college student information was established at ACE in 1966: the Cooperative Institutional Research Program (CIRP). CIRP is now the longest running and largest such survey with data from over 15 million students. Finally, the College and Beyond II project hosted at the ICPSR collected student data over 20 years, linking college experiences to long-term outcomes. These surveys are critical infrastructure in shaping educational policy and continue to empower educators and researchers to understand educational outcomes.

MOOCdb was born out of the excitement that MOOCs brought to education when EdX and Coursera launched in 2012 (Veeramachaneni et al. 2013). The research opportunities provided by digital educational platforms serving millions of global learners seemed boundless. However, a lack of a standardized data schema created significant hurdles for researchers, limiting their capacity to share and aggregate data. According to Veeramachaneni et al. (2013), researchers were spending more than half their time cleaning datasets rather than creating models and conducting analyses. To address this problem, MOOCdb provided a standardized data schema for data extraction from EdX and Coursera courses. In this way, MOOCdb facilitated data sharing between institutions without the data access and storage required of a data consortium.

The MOOC Replication Framework (MORF) built on this earlier work to address research challenges and replication failures by providing a platform-as-a-service consisting of data as well as experimental and inferential infrastructure (Gardner et al. 2018). The platform provides data, a computing environment, and a means to reproduce any analysis or experiment conducted on the platform, all while protecting student data privacy. As such, the MORF is based on the premise of sending "code to data" rather than downloading data and analyzing locally. At the outset, three institutions provided hundreds of MOOC courses to the platform. Currently, only one institution is continuing to upload new data.

LearnSphere is a project based out of Carnegie Mellon University that represents the current best practice for the storage and sharing of student interaction data from digital learning experiences. LearnSphere is a landing page for several data repositories and for Tigris, an analysis, computation, and workflow sharing platform. The repositories, known as DataShops (e.g., Koedinger et al. 2010), provide secure storage for curated data and an interface to browse, search, and download data. There is a mix of open, by request, and private data. The DataShops require data imports to adhere to the same data standard, facilitating reproducibility of analyses between datasets. Tigris is an analysis and computing platform that allows users to create modeling and analysis pipelines and modules and to share them for reuse by other users. Tigris is integrated with DataShop to supply analysis data.

Another category of data platform is growing in prominence within existing universities. These are designed from the ground up for integrating data across student experiences at an institution, but so far, none of these platforms have been deployed across multiple institutions. Instead, they generally serve institutional data access needs for research and administrative purposes. Arizona State University's (ASU) Learning at Scale is an example. It combines data from digital learning experiences, student profiles and trajectories, demographic data, and written assignments. These data are available internally and externally by request.

Some emerging initiatives seek to aggregate institutional data. The Unizin Data Platform (UDP) is a learning data ecosystem that integrates and normalizes data from Student Information System (SIS), Learning Management System (LMS), and digital learning tools within institutions as a platform-as-a-service. UDP is a cloud-native solution that is designed to support analytics and dashboards within institutions in lieu of their current infrastructure. UDP emphasizes communities of practice for research and analytics to share models and tools between institutions.

A final initiative that bears mentioning, SEERNet, seeks to advance education research and practice by connecting digital learning platform developers with researchers and educators. In doing so, the Institute for Educational Science–funded initiative will leverage existing digital platform infrastructure to advance education research and practice.

Table 1. Data Initiatives

Data Initiative	Description	Examples
Survey	Nationwide surveys on enrollment, tuition, aggregate proficiencies, etc. Centrally collected and distributed to inform policy and research	ACE's CIRP Department of Education's NAEP
MOOC	Data schemas and repositories designed to structure MOOC courses so that data can be easily extracted and analyzed	MOOCdb, MORF
LearnSphere	Data repositories and shareable analysis tools–datasets typically include interaction data as part of a learning platform	DataShop and Tigris, partnered with Carnegie Mellon University, The University of Memphis, Massachusetts Institute of Technology (MIT), Stanford University
University-Localized	Comprehensive student data sharing available upon request and approval: demographic, student profile, course profile, discussion boards, written assignments	ASU's Learning at Scale–all data generated within ASU
Aggregate	A learning ecosystem that integrates and normalizes data from SIS, LMS, and digital platforms and provides learning analytics	Unizin, a membership-based organization of postsecondary education institutions focused on technology solutions but with focus on a unified data platform

These data consortia have enabled researchers and universities to begin sharing data for research and—in some cases (such as Unizin)—to support the digital transformation of institutions. Missing from existing initiatives are the key GDC points of focus: institutional capacity building with AI, collection of datasets to drive AI tool and model development, networked support in initiating and deploying new data-centric technologies, a modern mesh-based data architecture, use of cutting-edge privacy practices (such as the use of synthetic data), and international collaboration to ensure data reflects diverse populations.

TECHNICAL

The technical environment for the GDC is planned to enable local (domain) control of data while enabling multi-institutional data sharing. Key features of this infrastructure include advanced security protocols, scalable data storage solutions, intuitive data sharing platforms, and comprehensive support for data interoperability. These components are designed to facilitate seamless collaboration among partners, ensuring data is both protected and readily accessible for collaborative research and innovation.

The foundation of the data consortium is the data mesh architecture, where each partner organization is responsible for managing their data products. This approach represents a transformative shift in managing and sharing data, addressing limitations and bottlenecks experienced by organizations using centralized architectures such as data lakes and centralized repositories. *Data mesh* as a concept emerged in 2019, promoting a decentralized approach to data management and commonly structured around four main principles: i) domain ownership, ii) data as a product, iii) the self-serve data infrastructure, and iv) the federated data governance.

In general, the domain ownership principle shifts the responsibility for data to domain teams, mitigating the bottlenecks introduced by centralized data teams and ensuring that those closest to the context from which data were collected remain the custodians of the data. To enable members to develop their capabilities with AI and developing AI products, the GDC will offer training, share best practices, and make technical resources and expertise available to ensure that all partners can successfully implement initiatives within their institution. In particular, the GDC will offer a *core technical and innovation team* to assist members with planning and implementing institutional projects. This core service will ensure that all institutional members—regardless of resources—can participate and develop internal capabilities.

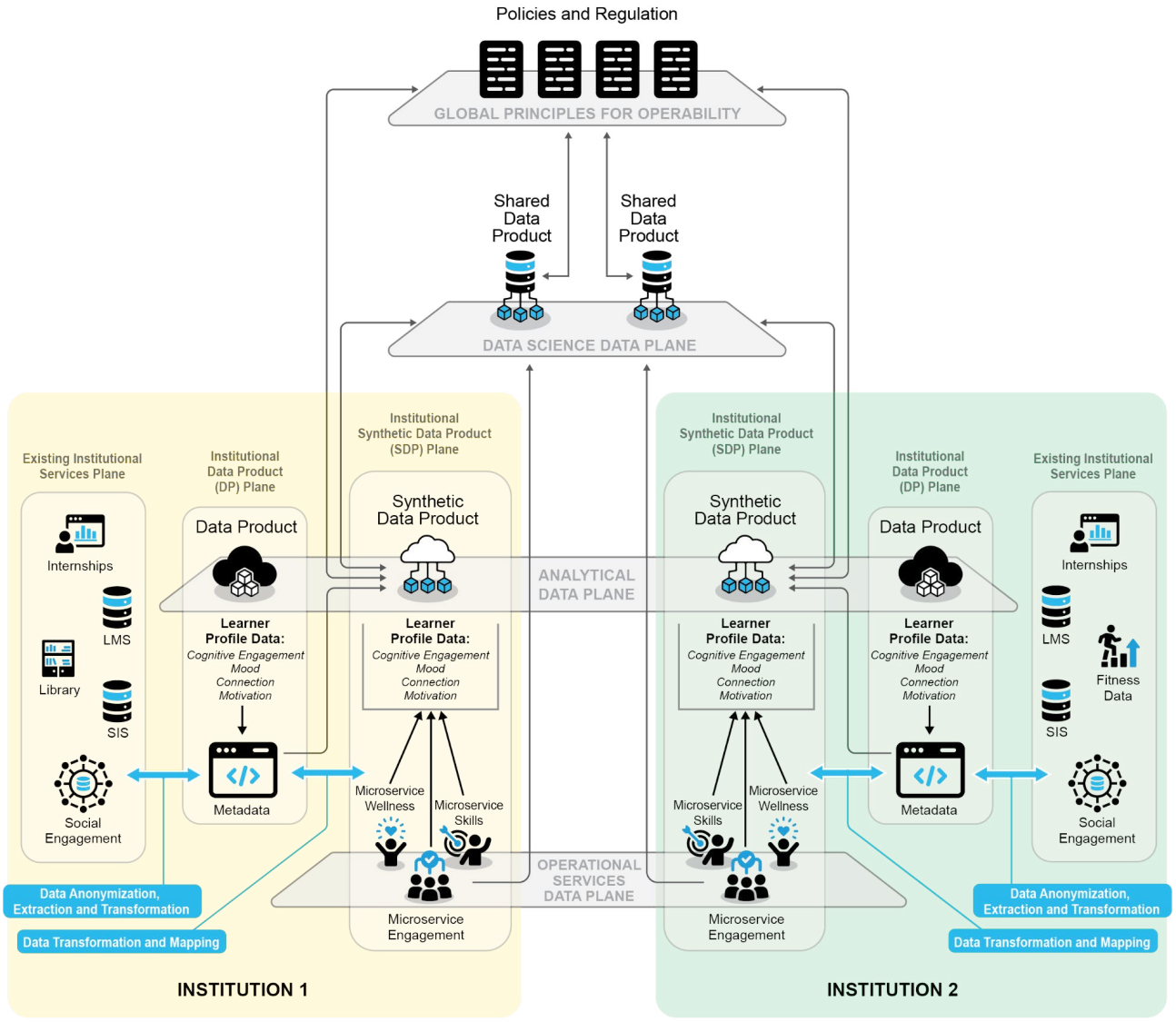
The principle of treating data as a product mandates that data teams approach their datasets with the same rigor and care as tangible goods, guaranteeing that the data adheres to stringent quality and usability standards for use beyond their specific domain and individual institutional settings. The GDC encourages partnering institutions to focus on the entire lifecycle of data—from creation and storage to documentation and maintenance—with an emphasis on making data easily accessible, understandable, and usable for others. By implementing the data as a product principle within the data consortium, it will enable a more dynamic and efficient data sharing environment, providing a rich and interoperable data infrastructure.

The self-serve data infrastructure platform introduces platform thinking to data infrastructure, where a dedicated team provides the tools and systems for partnering teams to build, execute, and maintain interoperable data products.

The federated governance principle ensures that decentralization does not lead to fragmentation. By standardizing interoperability and promoting organizational and industry regulations, federated governance fosters a cohesive data ecosystem that respects the consortium's collective goals and compliance requirements.

Figure 1 shows the interaction between the four principles and provides a high-level overview of the proposed architecture.

Figure 1. GDC Architecture: A Conceptual Overview



GDC infrastructure is a nuanced implementation of self-serve platform capabilities, segmented into distinct but interconnected planes. These planes are conceptualized as areas of operation, each integrated with but distinct from the others. This approach is similar to the coexistence of physical and consciousness planes, as explained by Dehghani. Each plane serves a specific function, contributing to the overall ecosystem while retaining its distinct identity.

In this structure, both horizontal and vertical planes articulate the complexity and scope of data management strategies. Horizontally, the data is organized into tiers that reflect the nature of data ownership and its intended use: original institutional data that remains within the jurisdiction of its creators; extracted data earmarked for collaborative analysis within the consortium, balancing institutional responsibility with collective benefit; and synthetically generated data, designed to be openly shared across the consortium to stimulate innovation while safeguarding privacy and ensuring data fidelity. The interaction between the horizontal planes is driven by the institutional requirements, ensuring student data privacy.

Vertically, these planes are functional domains that facilitate operations across the horizontal tiers—from data collection, processing, and storage to analysis, dissemination, and application. This architecture allows for a dynamic and flexible approach to data management, supporting the GDC's commitment to responsible AI principles (as outlined below), including ethical usage, privacy preservation, and governance. These planes establish a framework that respects the complexities of data in educational settings. It is important to distinguish that vertical planes are conceptualized in the same way as argued by Dehghani. However, the introduction of horizontal planes serves to ensure that institutional requirements can be satisfied without necessarily engaging with the rest of the GDC ecosystem.

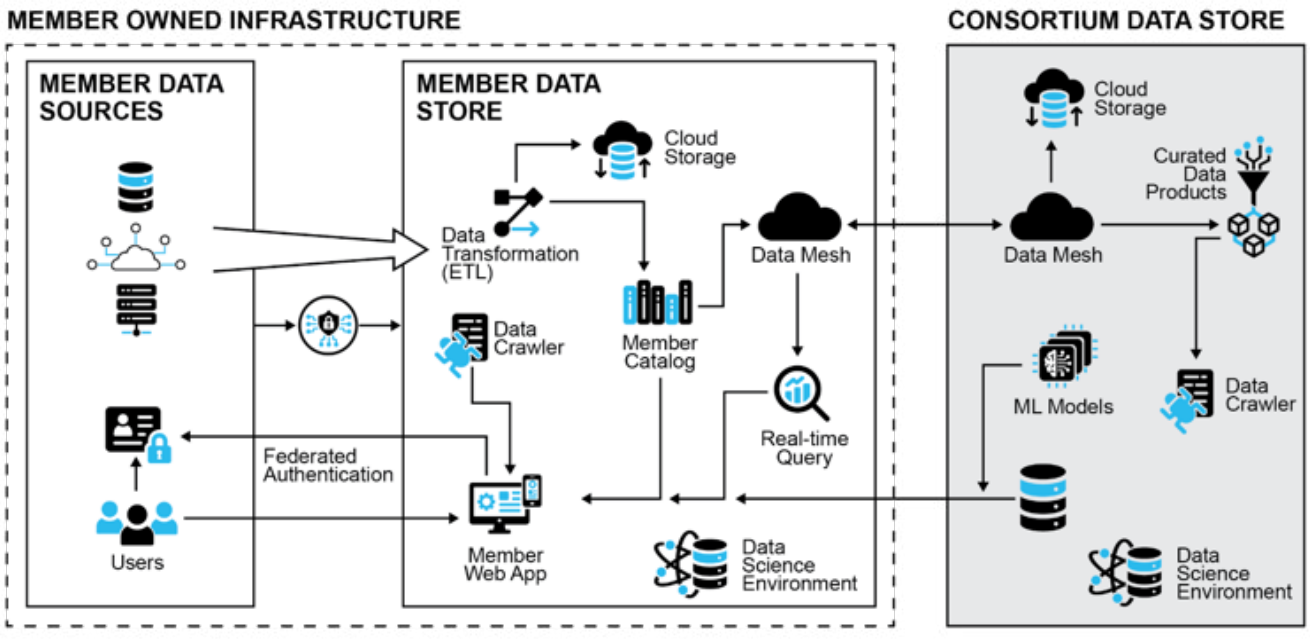
A MODERN DATA INFRASTRUCTURE

Modern cloud applications and data platforms share common architectural practices and principles. Solutions should:

- Be scalable and elastic
- Be highly available and performant
- Utilize serverless managed cloud services
- Use decoupled architecture
- Utilize Infrastructure as Code (IaC) and Continuous Integration/Continuous Deployment (CI/CD) to orchestrate the software development lifecycle via DevOps practices
- Include security, compliance, monitoring, auditing, and notification services

Figure 2 details how member (individual university) data flows from source data to member store data to the consortium data store. There is some overlap between functionality that exists at a member data store level and what exists at the consortium level. For example, the data science environments (cloud-based toolkits for working with data) are present at both member and consortium levels.

Figure 2. Data Flow



DATA ENVIRONMENT AND FLOW

A key principle of the data infrastructure platform is to provide member institutions with easily deployable infrastructure to connect with the consortium while they maintain complete autonomy of their own data. The provisioned infrastructure will allow member institutions full data governance controls to manage any source system connected to the member data store and provide tools to assist in any data preparation required prior to sharing (data curation) and controls over what data is approved for sharing to the consortium data store.

The infrastructure provides member institutions ETL (Extract, Transform and Load) and ELT (Extract, Load, Transform) tooling and expertise to ingest data into the Member Data Store and stage and curate data ready for sharing. This tooling supports a wide range of traditional and modern data sources for data ingestion, including relational databases, data warehouses, data lakes, and streaming data.

Members have full control over what datasets are shared with the consortium. Once data is curated and ready for sharing, the member institution can publish the dataset via an approval process within a supplied GDC Member Web Application. This notifies the consortium that the new dataset exists and allows the consortium to query the dataset remotely via the data mesh infrastructure.

Member Institutions will have access to a secure planned GDC Member Web Application that provides users with an interface to navigate a detailed data catalog, including:

- Draft datasets present in the member's local data store
- Published datasets present in the member's local data store shared with the consortium
- Published datasets shared by the consortium

The data catalog will allow search of available datasets along with detailed descriptions, field names, types, and associated metadata to assist users in finding, inspecting, and interrogating data. Users from member institutions can also subscribe to receive email notifications of any significant updates to consortium datasets of particular interest.

Data stored either in the Members Data Store or from the Consortium Data Store (over the Data Mesh) can be queried in a centralized analytics service via natural language query. Member institutions can use this service to provide connectivity directly to their existing analytics platforms or leverage the native tools in the cloud provider's environment to provide rich dashboarding and graphing capability or data science notebooks environments within the Member Data Store account. In addition, the web application will facilitate the sharing of algorithms, machine learning models, data science projects in the form of Jupyter Notebooks, and associated resource materials between member institutions via the consortium data store.

The GDC Member Web Application (see figure 2) will be secured using federated identity, allowing member institutions to utilize existing Security Assertion Markup Language–based identity providers to have complete control in how users access data and authenticate.

This infrastructure minimizes the overall heavy lifting of provisioning the Member Data Store by utilizing the cloud-based marketplace and private subscriptions. Using this provisioning tool, key services will be automatically deployed/configured to support member's access to/from the remote consortium infrastructure.

WHAT DOES THIS LOOK LIKE PRACTICALLY?

Institutions will have control over their data and determine what gets shared with consortia members. Data will be shared on a solution-oriented basis. This means that data is shared as it relates to a consortia project or to a particular capability that members want to realize. Only data that addresses a clear need is made available—rather than sharing for the sake of sharing.

Consider the following project. Consortia members are interested in reducing student dropouts. To initiate the project, network members meet to discuss the project scope and need, available institutional data, models to work with data, use cases for adoption of results, mechanisms to confirm effectiveness of the models developed, and technical needs to run the project. By defining data needed, the members can locally prepare required data, deliver to the mesh as a data product, and coordinate with the GDC core technical team on models. Most universities have a range of data on student help-seeking behavior as well as institutional responses. This pairing of data with evaluations later in the student's academic journey is a type of data that will provide rapid value for the institution to begin planning targeted help resources specific to individual learner needs.

While this is not a complex data project, since it relies on existing learning analytics and machine learning approaches, it outlines the process for engagement between members: central support, network learning pods, data access and sharing, and shared model training and development. More involved AI product projects, such as developing an AI-enhanced tutor, would require more data, more computation capabilities, and more cross-network engagement.

RESPONSIBLE AI

The GDC is committed to the principles of responsible AI, including the use of synthetic data to ensure privacy and mitigate biases. This approach will enable the safe and ethical development of AI applications, prioritizing human values and ethical considerations. Drawing on the comprehensive insights from contemporary literature (e.g., Baker and Xiang 2023, Dignum 2023), the foundation of responsible AI rests on the pillars of the thoughtful application of synthetic data, privacy, ethics, and governance.

Synthetic data is emerging as an important practice within responsible AI, as it offers a pathway to address privacy concerns (Savage 2023). Synthetic datasets are artificial constructs that are engineered to mimic the statistical properties of real-world data without compromising individual privacy. This allows us to simulate diverse educational scenarios and outcomes, safeguarding the privacy of students while providing rich datasets for AI development and research. Recent evidence further shows that synthetic data can provide better utility than real data, given that the right balance is achieved between accuracy and privacy (Veeramachaneni et al. 2013). In that sense, we envision that only synthetic data would be shared within the GDC ecosystem, while partnering institutions would remain exclusive users of their learner data.

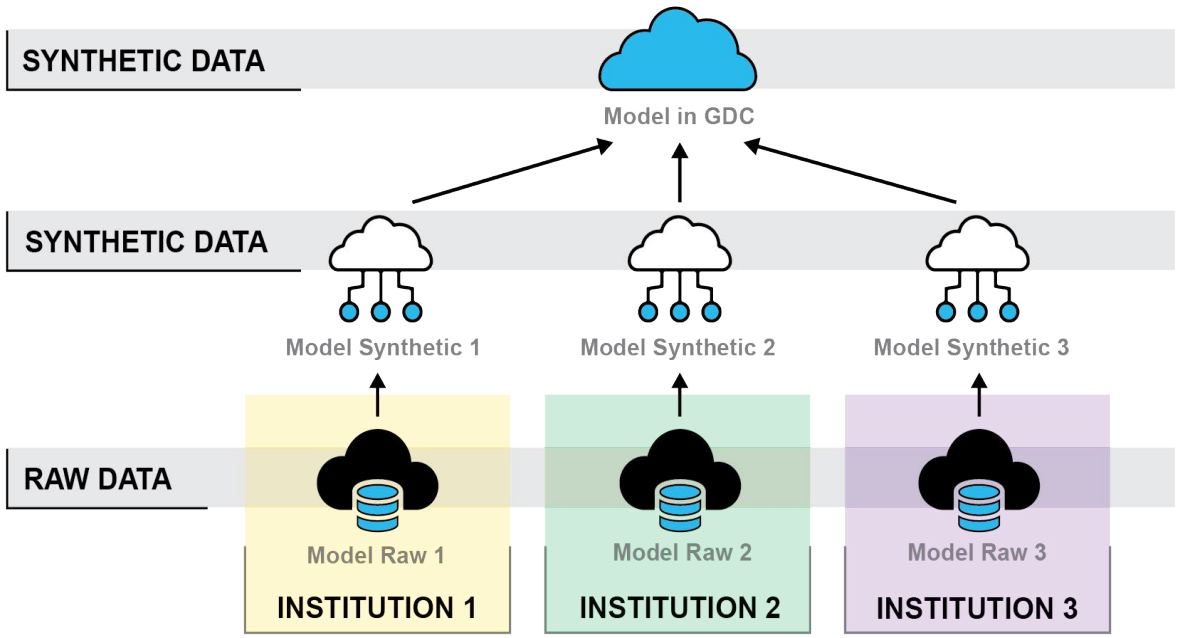
In parallel with the use of synthetic data, our dedication to **privacy** encompasses stringent protocols for data handling. These protocols must align with both domestic (institutional) and international data protection regulations, ensuring the confidentiality and integrity of student data. This commitment extends to developing systems that enable us to harness the benefits of AI without encroaching upon the individual's right to privacy. This entails adhering to principles laid out in significant regulatory frameworks such as the European Union (EU) Artificial Intelligence Act, which categorizes our system as high-risk. Such classification mandates a rigorous adherence to data protection principles throughout the AI system's life cycle, incorporating data minimization and protection by design and default. This is one of the drivers for deploying a multi-plane platform (figure 1), where our horizontal planes enable privacy by design and ensure the adherence to existing and developing regulatory frameworks. Furthermore, it involves the deployment of sophisticated technologies enabling algorithms to operate on data without necessitating its transfer or duplication, thereby upholding strict data governance measures.

Building on our foundation of responsible AI and the strategic use of synthetic data, our approach to **ethics** takes center stage in ensuring the development and deployment of AI systems under the GDC that are technologically advanced as well as fundamentally aligned with human values. Drawing from the European Commission's Ethics Guidelines for Trustworthy AI, the following four ethical requirements are particularly important: i) human agency and oversight, ii) diversity, nondiscrimination, and fairness, iii) societal and environmental wellbeing, and iv) accountability. In doing so, we integrate mechanisms such as human-in-the-loop, human-on-the-loop, and human-in-command to empower individuals with control over AI systems. This allows for informed decision-making while

fostering a symbiotic relationship between AI systems and their human users, ensuring that our consortium's AI applications enhance rather than supplant human capabilities. One of the primary motivators for the GDC's emphasis on responsible AI is the deliberate effort to eliminate unfair biases within AI systems. AI technologies should be accessible to everyone as a foundational mission. This approach is crucial in mitigating risks of marginalization and discrimination, reinforcing commitment to fostering an equitable AI landscape. Proactive measures in this direction reflect a deep-seated belief that the true potential of AI is realized when it uplifts and benefits all learners. Finally, the proposed operationalization establishes clear accountability mechanisms. Through mechanisms such as auditability and accessible redress, responsibility for AI outcomes is clearly defined and actionable, fostering a culture of accountability within the consortium.

Finally, the GDC is committed to upholding excellence in its **governance framework**. This commitment is reinforced by adherence to the current and developing regulatory frameworks, such as the EU Artificial Intelligence Act and the European Commission's Ethics Guidelines for Trustworthy AI. Governance practices will ensure rigorous validation and testing of AI systems, guided by principles that prioritize ethical design choices, meticulous data collection, and comprehensive data preparation processes. Such measures are crucial to creating AI systems that are resilient, secure, and free of biases, complying with the rigorous requirements set for high-risk AI systems under the EU Artificial Intelligence Act. Moreover, the consortium will establish robust oversight and accountability mechanisms within its governance model. Competent human oversight monitors the deployment of AI systems, ensuring their operation aligns with data protection targets and standards.

Figure 3. Raw-Synthetic Data Model



Drawing on the principles outlined above, figure 3 illustrates a layered approach to developing synthetic data within the GDC ecosystem, harmonizing with our established principles of responsible AI. The proposed structure indicates a flow of information from raw data to synthetic data models, aligning with our commitment to privacy and adherence to data protection principles as mandated by the existing regulatory frameworks. As such, we facilitate the responsible sharing of synthetic data within the consortium while allowing partnering institutions to retain exclusive control over their raw data. This approach embodies our operationalization strategy, which encompasses strict governance, robust oversight, and accountability, thereby ensuring that our AI applications adhere to technical and ethical standards.

INVITATION

The American Council on Education extends an open invitation to all interested parties to join this global initiative. By contributing knowledge, resources, or expertise, you can help shape the future of AI and its positive impact on society. Until June 1, 2024, we are soliciting commentary on any aspect of this paper, including technical design, concerns about deployment, ethics and privacy, and configuration of the consortium. Please email gdc@acenet.edu to offer your comments.

CONCLUSION

The data consortium will be a game changer—catalyzing innovation, research, and student success gains for member institutions. By participating in the consortium, member organizations will secure access and collaborative opportunities far beyond what any single institution could achieve alone. Additionally, the GDC will supercharge each member's capabilities and potential.

The GDC can develop groundbreaking education AI models to support increased student wellbeing and retention. Across member institutions, learner completion rates will rise; equity will increase and expand; and learning will be transformed. The consortium's policy work will cement responsible data and AI best practice into the foundations of educational innovation in postsecondary education. This collaborative data consortium approach will take postsecondary education to a new level through shared data resources, expertise, and ambition. We will unleash unprecedented beneficial innovation at scale. This is an opportunity for postsecondary education to lead the way in shaping the future of education, with the consortium model becoming a field-advancing template for collaboration within the postsecondary education sector.

REFERENCES

Baker, Stephanie, and Wei Xiang. 2023. "Explainable AI is Responsible AI: How Explainability Creates Trustworthy and Socially Responsible Artificial Intelligence." *arXiv* preprint, 2312.01555.

Dignum, Virginia. 2023. "Responsible Artificial Intelligence: Recommendations and Lessons Learned." In *Responsible AI in Africa: Challenges and Opportunities*, edited by Damian Okaibedi Eke, Kutoma Wakunuma, and Simisola Akintoye, 195–214. Cham: Palgrave Macmillan.

Gardner, Josh, Christopher Brooks, Juan Miguel Andres, and Ryan S. Baker. 2018. "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) Proceedings*, edited by Naoki Abe, Huan Liu, Calton Pu, Xiaohua Hu, Nesreen Ahmed, Mu Qiao, Yang Song, Donald Kossmann, Bing Liu, Kisung Lee, Jiliang Tang, Jingrui He, and Jeffrey Saltz, 3235–3244. Piscataway, NJ: IEEE.

Koedinger, Kenneth R., Ryan S.J.d. Baker, Kyle Cunningham, Alida Skogsholm, Brett Leber, and John Stamper. 2010. "A Data Repository for the EDM Community: The PSLC DataShop." In *Handbook of Educational Data Mining*, edited by Cristóbal Romero, Sebastian Ventura, Mykola Pechenizkiy, and Ryan S.J.d. Baker, 43–56. Boca Raton, FL: CRC Press.

Savage, Neil. 2023. "Synthetic Data Could Be Better Than Real Data." *Nature*, April 27, 2023. https://www.nature.com/articles/d41586-023-01445-8.

Veeramachaneni, Kalyan, Franck Dernoncourt, Colin Taylor, Zachary Pardos, and Una-May O'Reilly. 2013. "MOOCdb: Developing Data Standards for MOOC Data Science." In *AIED 2013 Workshops Proceedings, Volume 1*, co-chairs Zachary A. Pardos and Emily Schneider, 17–24. Leeds, UK: International AI in Education Society.

ACE
American
Council on
Education®