Data Privacy and the Student Data Warehouse

By Michael Brown
Institutions of higher education in the US have increasingly relied on digital technologies that help to collect, archive, and circulate data about students and their contexts, facilitating the use of a student data warehouse (SDW) to support institutional planning. The SDW refers to the people, policies, practices, and technologies that facilitate the movement of data about students and their behaviors throughout an education organization. These warehouses have become a fulcrum technology in the function of postsecondary organizations, but the construction of SDWs often occurs in the background of institutional work, opaque to students, faculty, and staff. As a consequence, we know very little about the SDW and the decision-making that guides its organization and implementation.

Prior research suggests that the lack of transparency in SDW operations disadvantages students’ ability to manage their individual rights to privacy. SDWs collect demographic and behavioral information about students, but students are often unaware of this, and education institutions provide few avenues to allow students to control their data.

Building on Linnet Taylor’s conception of data justice, this brief outlines the development of the SDW, the data privacy risks associated with it, and a decision-making framework featuring three actions that organizations should take to guide their decisions about the SDW:

- Establish equitable data collection and archiving systems that allow students to understand and control the data that represents them.
- Ensure trust and transparency in the use of data.
- Empower students, faculty, and staff to be informed agents in the governance of their data.

As institutional leaders make decisions about the SDW, what it contains, and who has access to it, a student-centered approach to privacy would actively engage students in governance activities. To develop that approach, institutions should also ask the following questions:

- If we incorporate a new technology, is it transparent to users, particularly to students?
- Do individuals have the opportunity to control data collection and storage?
- If not, what is the rationale for constraining individual autonomy?

Recent research suggests that institutions are not engaging in this work. Michael Brown observed that faculty became resistant to data-driven teaching and learning initiatives when they realized that, like their students, they would also be subject to the opaque algorithms that classified student behavior. Brown and Carrie Klein observed that institutions rarely updated their policies to reflect the kind of dynamic, real-time data collection that occurred on campus, constraining students’ access to their data and placing narrow limits on which information students could reasonably expect to keep private. This current approach, lacking reflection and proactive planning, has strong potential for harm and privacy violations. By actively engaging stakeholders in governance of their data, postsecondary institutions can serve as a model for how other institutions and stakeholders in our datafied society could govern the collection, circulation, and assetization of data.
In an ongoing research project on higher education stakeholders’ knowledge of institutional data practices, an administrator was asked to identify where and how her institution stored data about students on her campus. She pulled up a diagram illustrating several tools and technologies on campus that collected data about students: the learning management system, card swipe access points at various campus buildings, interactions with the registrar, financial aid office, and academic advisors. All of these tools, even when they spoke to each other (such as when registration records are sent to the learning management system to ensure that students are enrolled in a course site), ended their circuitous route at a small square in the center of the diagram labeled “SDW,” for student data warehouse.

Since 1995, institutions have invested substantial resources in the digital architecture that supports campus life. This infrastructure is often invisible, and yet it plays an increasingly central role in how institutions function, which Ben Williamson calls the invisible architecture of datafication. The architecture emerged to serve various needs; a central data repository made it easier to track who had access to student data and how people used the data. It also made it easier to act on student data. Through learning and academic analytic initiatives, the SDW functions as a fulcrum technology that allows data representations of students to circulate around campus, facilitating decision-making and action. This research brief outlines broad trends in higher education that have fostered the construction and maintenance of the SDW; identifies the potential challenges that students experience in attempting to manage their data in the SDW; and outlines a decision-making framework, previously co-authored with Dr. Carrie Klein, to guide institutions in the development and operation of SDWs.

What Is a Student Data Warehouse?
The student data warehouse refers to a range of people, policies, practices, and technologies that facilitate the movement of data about students and their behaviors throughout an education organization. SDWs now contain various data assets, including administrative and record-keeping data related to students’ personally identifiable information (age, gender, race, nationality, place of residence); trace data of their engagement with
digital technologies on campus (such as clicks in the learning management system or card swipes to enter a building); and academic records, including disciplinary actions. This range of collection marks a sea change from the kinds of education records that institutions previously stored. The SDW often exists in a liminal space in campus operations, as it often has no physical footprint in the campus environment. The key function of the student data warehouse is to aggregate, organize, extract, and archive student data to support organizational functions. Organizations achieve this by linking data collection and extraction technologies (which include a wide range of digital and face-to-face practices) with organizational and archival technologies. The warehouse, then, is not simply the servers on which data is stored but, rather, the technologies, people, and practices that facilitate data fulfillment. Warehouses do not exist simply for storage. They serve as a link in the supply chain between production and consumer.

Commentary on the SDW focuses on who has access to student data, how they gain access, and for what purpose. These are important questions that institutions should wrestle with as part of their ethic of care and concern for students. A parallel conversation, however, is needed about how the fulfillment of data resources to their objectives happens through the SDW. This conversation is necessary because the SDW has changed institutional record keeping such that institutions now have millions of data assets about individual students. Expecting an individual to keep track of and correct this data—as the Family Educational Rights and Privacy Act (FERPA) does—is unreasonable. To help students, faculty, and staff better understand and control their data, institutions need to develop education and transparency about the SDW and its functions.

Two recent research projects demonstrate how failing to account for these concerns creates environments rife with potential for privacy violations. In research with faculty using a new dashboard technology, instructors were surprised and frustrated to learn that the technology was collecting and reporting data about the assessments they used, attendance by their students, and their choices of instructional artifacts back to their departments. This violation of their classroom autonomy caused faculty to question other aspects of the technology and deter their use of the tool in the classroom. When faculty found themselves subject to opaque surveillance activities, they became increasingly concerned with the intrusion of surveillance technologies in all aspects of their teaching work. As a result, instructors often abandoned the dashboard, wary of its potential to collect data in ways that were not explained and over which they and their students had little control. This deterred the faculty’s adoption of the dashboard but did not lead to conversations within the department about data collection, extraction, or archiving activities. Moreover, faculty were not inclined to help their departments figure out how to create student-centered technology systems. Instead, they built classrooms increasingly organized around defending against the intrusion of data-extraction technologies.

In a related study of institutional data privacy policies, institutions often failed to update their internal policies to reflect the ways that technology works in contemporary institutions. Policies were outdated, treated data as static artifacts that had to be manually updated, and were often silent on institutions’ potential uses of data. Such policies’ lack of structure to govern data use means that students have little control of these representations and almost no recourse regarding who their data is shared with as long as the sharing meets the requirement of a legitimate educational purpose, which policies rarely define. Uncharted territories are not well suited to participatory governance, but the lack of relevant, student-centered, up-to-date policies means that local (and in many cases individual) decisions about data use and sharing reign.

If we extend the metaphor of a warehouse, it is rare (or perhaps unheard of) that the subjects from whom are extracted should participate in governance over operations, but students, instructors, and other stakeholders are fundamentally different from other subjects in the process of fulfillment. As Janja Komljenovic noted in her recent call for further research on the political economy of data assets in higher education, data representations are fundamentally different kinds of assets (or commodities) from books or shoes. Data assets are intimately connected to the individuals from whom they are extracted, and institutions often use the assets to guide organizational decision-making about the treatment of the individuals that the data assets represent. When the SDW extracts data from users, they are disadvantaged because they are no longer in control of their representation and, particularly for students, are no longer in control of
how institutions use the data to shape other aspects of the users’ experience. They are stripped of their autonomy, including their ability to control how they appear in data systems. Institutions need to adopt a different approach to the system of student data fulfillment that eschews the dehumanizing approach of the warehouse.

How Did SDWs Develop?

How Did SDWs Develop?

Starting with Civil War benefits in the late nineteenth century and continuing to the creation of social security cards in the early twentieth century, the US government developed technologies of documentary identity that allowed individuals to assert who they were and claim entitlements. Higher education institutions were increasingly expected to develop their own documentary bookkeeping to help individuals document their earned credentials and to keep track of entitlements. The institutional student data system involved an ad hoc network of physical records, stored in various locations, without standardized requirements. Throughout the twentieth century, the proliferation of information about students on campus and the potential for errors in this data limited students’ opportunities. In The Known Citizen, a study of how public notions of privacy evolved in the US, Sara Igo argues that FERPA emerged from parents’ concerns over “children being locked into a records prison via questionable and potentially damaging information in public school files.”

In response to FERPA’s implementation, the student information system has shifted from an input-focused archive to a process-focused network of practices and policies. Federal reporting requirements increased alongside policy mandates that make institutions responsible for protecting student data (and to ensure the data’s accuracy). The creation of the Integrated Post-Secondary Data System (IPEDS) survey, the passage of the Clery Act in 1990, and expanding data collection activities related to federal and state accountability and accreditation regimes increased the need for information technology that could store data about students and their behaviors.

Fortunately, from the mid 1990s through the early twenty-first century, education institutions’ ability to store information also increased substantially through the development of computational power and the construction of digital storage systems on campus. Institutions invested substantial resources in the movement from what stakeholders call the “miles of tape” system to digital and cloud-based storage. Institutions hired specialized staff to maintain the new data infrastructure and to manage new reporting and archiving requirements. As
federal policy, credentialing systems, and the wide array of digital instructional technologies have made student data warehouses essential infrastructural technologies, a core question remains regarding how to promote students' autonomy and privacy: how should the SDW be configured?

Problems of Delivery and Open Doors: A Framework for Decision-Making

As a starting point, the following decision-making framework comprises three action items that should guide every institutional decision about the SDW's organization:

› Establish equitable data collection and archiving systems that allow students to understand and control the data that represents them.
› Ensure trust and transparency in the use of data.
› Empower students, faculty, and staff to be informed agents in the governance of their data.

The action items relate to three principles drawn from Linnet Taylor’s 2017 article *What Is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally.* Taylor argues that data should be visible, individuals should be able to make decisions about their engagement with technology, and individuals should be free from discrimination.

With those action items in mind, institutions can ask guiding questions that inform the implementation of technology and the SDW's configuration. For example, if an institution incorporates a new technology, is it transparent to users (particularly to students)? Do individuals have the opportunity to control their engagement with technology? If not, what is the rationale for constraining individual autonomy? For each of the above action items, the framework includes an associated problem, the resulting technological and organizational obligations, and the individual abilities and rights impacted (see Table 1 for a summary).

<table>
<thead>
<tr>
<th>Action Items</th>
<th>Associated Problem</th>
<th>Technical Obligations</th>
<th>Organizational Obligations</th>
<th>Individual Abilities and Rights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Establishing equitable systems</td>
<td>Algorithmic and institutional bias and discrimination</td>
<td>Identifying and preventing discrimination and bias in algorithms</td>
<td>Co-creating inclusive structures, policies, and practices with stakeholders</td>
<td>Ability to report and challenge discrimination and bias</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Acknowledging, interrogating, and addressing structural inequities</td>
<td></td>
</tr>
<tr>
<td>Ensuring transparency and trust</td>
<td>Opaque algorithms and interventions and organizational policies</td>
<td>Ensuring transparent and accurate data and representation</td>
<td>Actively including student voices in analytics use and outcomes</td>
<td>Ability to rely on data accuracy and aligned data to meet co-determined needs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Providing clear data visualizations, interventions, and terms of use</td>
<td>Center individual and group contexts in decision- and policy-making</td>
<td></td>
</tr>
<tr>
<td>Empowering informed agency</td>
<td>Data power divides and dataveillance</td>
<td>Sharing benefits of data with stakeholders</td>
<td>Ensuring student rights and privacy protection through accountability</td>
<td>Ability to access data and choice in representation or (in)visibility</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Allowing students to access and manage their data</td>
<td>Educating students about their data and related rights</td>
<td>Right to be informed, educated, and empowered to action</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Active inclusion of stakeholders in tool development</td>
<td></td>
<td>Right to exercise self-governance over data and data use</td>
</tr>
</tbody>
</table>

Stakeholders in US higher education have called for institutional environments that promote equity; center students; and empower students, faculty, and staff to participate in institutional governance. In part, this means establishing equitable student data systems in which users have control over their representations.
ensures that institutions meet their ethic of care and acknowledges that institutions have been slow to address discriminatory practices in the past. Allowing users to control their representations also increases institutions’ ability to anticipate harm, either from biased algorithms or from privacy violations resulting from surveillance-level data collection. As a result of historically unequal systems, institutions need to work proactively to establish trust through transparent, collaborative governance. To participate in collaborative governance, students, faculty, and staff need to be empowered and informed of their rights within the institution.

Ensure Transparency and Trust

To ensure transparency and trust, institutions must confront the opaque algorithms, digital tools, interventions, and organizational policies that can harm students.

A prominent example of how opaque predictive algorithms can cause harm was the plan by former Mount St. Mary’s president Simon Newman, who suggested the university might intervene with at risk students by encouraging them to withdraw even before the semester started. In his parlance, the institution would use student data to “put a Glock to their heads,” in order to increase retention numbers by 4 percent.19

In terms of data systems and data-collection and -extraction technology, FERPA already requires institutions to ensure accurate data. To foster trust and transparency, institutions need to go beyond FERPA’s obligations, to ensure transparent data collection and provide clear data visualizations, interventions, and terms of use. To trust institutional data use, students need to understand when and why institutions act upon student data. To achieve this, institutions should actively include student voices in the development of analytics applications and center students’ needs and expectations in policymaking.

To address the potential for algorithmic and institutional bias (and the discrimination that it facilitates20), institutions must begin by clearly accounting for how data moves through the organization: how it enters, exits, and who encounters data as it circulates. Research indicates that few institutions have
undertaken this essential initial task. However, if institutions continue to approach the SDW from the perspective of a supply chain—failing to account for how doors to the warehouse open, for whom, and under what circumstances—they will leave the supply chain at risk of breakdown. Specifically, institutions inevitably violate students’ privacy rights when they open the SDW to third-party vendors that can access, collect, and use student data either to intervene with students or to enrich themselves by monetizing data assets. Recent controversy over the use of e-proctoring technologies highlights the ways in which the SDW can extend beyond the classroom into students’ homes. Students reported having to use a webcam to film their bedrooms before they could access e-proctoring software required to take an exam. E-proctors benefit from the violation of students’ privacy rights because companies use the data they collect during these activities to enhance the development of their commercial software. To avoid this, institutions need to engage users and stakeholders in the creation of policies and practices that guide data use. Individuals need to have recourse, including the ability to report and challenge discrimination and bias as they encounter it, and to challenge the technologies supporting this data collection when they infringe on students’ privacy rights.

Empower Individuals

Finally, this framework asks institutions to empower individuals to control how, when, where, and for what purpose they show up in institutional data sets. This means identifying data power divides (both among units within the institution and between the institution and the individual) and the potentially numerous forms of dataveillance in which the institution engages. Degli Esposti defines dataveillance as “the systematic monitoring of people or groups, by means of digital information management systems, in order to regulate or govern their behavior.” When Virginia Commonwealth University proposed a dataveillance-like program called RAM Attend that tracks students’ classroom attendance via mobile phones, nearly 60 percent of students opted out of the pilot study.

To address data power divides, institutions need to share with stakeholders their objectives and the potential benefits of data collection activities. Additionally, institutions need to allow students to access and manage their data, and should involve stakeholders in the development of teaching and learning technologies. To ensure that students can exercise their rights, institutions also need to educate students about their data and related rights. Similarly, institutional policies should reflect students’ right to access data, express choice in their representations, and opt out of the long term storage of their data via educational technologies (a process that many institutions already implement for plagiarism detectors).

Postsecondary institutions can serve as a model for how other aspects of the datafied society could govern the collection, circulation, and assetization of data. Specifically, these institutions can envision a future in which individuals have autonomy over their data representations and can determine their relationship to the institution. As this brief outlines, higher education institutions can achieve this by building SDWs that are student-centered, which involves initiating the three action items, asking the associated questions, and engaging students, faculty, and staff in their responses. Otherwise, students will find themselves and their data trapped in the black box of the student data warehouse.
References


Endnotes


2 Ibid.


9 M. Brown, *Seeing Students at Scale: How Faculty in Large Lecture Courses Act Upon Learning Analytics Dashboard Data*.

10 Ibid.


13 Ibid.


15 IPEDS requires institutions to report various data about students, staff, and institutional resources. The Clery Act mandates reporting about violent crime that occurs on or near campus.


