sustainability

Article

Big Data and the Ethical Implications of Data Privacy in Higher Education Research

Diana Florea [1] **and Silvia Florea** [2,*]

[1] Faculty of Letters and Arts, Lucian Blaga University of Sibiu, 550024 Sibiu, Romania; diana.florea@ulbsibiu.ro
[2] UNESCO Chair in Quality Management of Higher Education and Lifelong Learning, Lucian Blaga University of Sibiu, 550024 Sibiu, Romania
* Correspondence: silvia.florea@ulbsibiu.ro

Received: 17 September 2020; Accepted: 19 October 2020; Published: 21 October 2020

Abstract: Despite the claimed worth and huge interest regarding the increasing volumes of complex data sets and the rewarding promise to improve research, there is, however, a growing concern regarding data privacy that affects both qualitative and quantitative higher education research. Within the contemporary debates on the impact of Big Data on the nature of higher education research and the effective ways to harmonize Big Data practice with privacy restrictions and regulations, this study sets out to qualitatively examine current issues regarding data privacy, anonymity, informed consent and confidentiality in data-centric higher education research, with a focus on the data collector, data subject and data user. We argue that within current regulations, data protection of research subjects concerns more data collection and disclosure and insufficiently describes use, having procedural implications for both the complex nature of higher education (HE) research and the type of research data being collected. We work our argument through an examination of several factors that call for a reconsideration of data privacy and access to private information in HE research. The conclusions indicate that Big Data-centric HE research is increasingly becoming a mainstream research paradigm which needs to address critical data privacy issues before being widely embraced.

Keywords: Big Data; privacy; higher education; research; ethics

1. Introduction

The 21st century society is driven by massive investments in data collection, storage and distribution and scientific knowledge is ever more deeply digitized into large datasets of information known as "Big Data". An entire new research field—that of data science has emerged and, along with it, the traditional paradigm of data collection has changed, making way for an abundance of readily usable data that not only creates a data-driven observational world for researchers but also allows them to conduct research and scientific experiments that are exclusively based on data collection, curation and storage. However, the Big Data potential comes with a price, as it poses serious challenges both to personal privacy and the instruments/tools and methods that can be used to protect it. Such challenges have been emerging because technologies are capable of collecting so much data and processing them so effectively (by means of data mining and a wide array of analytics) that it has recently become clear that not only privacy can hardly be protected by means of traditional technologies but also that the policies issued to protect privacy need to reconsider more closely what is (and hence is not) technologically feasible to protect. Within such currently intense debates [1,2], the related privacy and security concerns have been clarified [3]; while information *privacy* represents the benefit to exert control over how personal information is collected and used, information *security* describes the practice of protecting information through the use of technology. Various challenges concerning Big Data privacy have already been addressed, especially within new frameworks for privacy conformity in ETL areas (Extract, Transform and Load) [4–6], data generation (falsifying data and access restriction) and data storage (in scalable storage infrastructures) stages. Nonetheless, less uniform procedures have proven to be applicable in what concerns Big Data privacy in data processing, particularly in what concerns safeguarding information from unsolicited disclosure and extracting useful information without breaching subjects' privacy. These aspects will form the basis of our approach with regard to higher education research.

Without any doubt, the use of Big Data in higher education (HE) research contributes to the improvement of teaching and learning processes, to making progress assessments of individuals and education systems, and to optimizing decision-making and education governance. HE research has therefore changed, partly because researchers now rely on data that are "easily found" not "created" and partly because the pressure is increasingly now on weighing, interpreting and using data-evidence in order to contribute critical research to the body of progressive scientific knowledge of the field. This brings about a normative construction of skepticism among both researchers and subjects alike, and a growing tension emerging from ensuring privacy, confidentiality and anonymity in particularly Big Data-based HE research. While HE research is generally grounded on an acceptance of empirical and theory-based information as a public good, the use of such data to attain the potential for societal welfare must be nonetheless appropriately regulated so as to protect research participants from related individual and social harms. This calls for an examination of not only how technological advances as well as research in the Big Data era have complicated both data privacy and data access but also how human and social action in HE research can remain a pooled resource for improving scientific advancement. On the other hand, the consent of student research subjects to being objectified, tracked, measured, analyzed and indexed as well as profiled for educational and social purposes operates, we argue, within a still lax framework of current policies and regulations that serves the interested gaze of institutions and researchers rather than the privacy and interests of participating subjects [7]. In so doing, further threats are posed by potentially manipulative core elements of group and individual contexts retrievable from BigData analytics, by the increasingly stable positioning of data users as Big Brother observers, and not to a lesser extent, by the danger of commodifying observed individuals into clusters of predictable behavioral traits to be instrumentalized according to end-user purposes and interests.

This article is a position paper that proposes a qualitative study that uses a descriptive method based on content analysis to examine the ethical implications of data privacy and the related implications of anonymity, consent and confidentiality issues for data-centric HE research emerging within the framework of current policies and regulations. It is beside the scope of this study to provide a comprehensive account of the present-day controversy around data privacy—it would be downright unrealistic to do so—however, several core and relevant data privacy issues will be discussed as emerging from the perspective provided by the data collector-subject-user relationship in HE research explored within the framework of current European policies and regulations. The argument that we extend is that while data protection is mostly restricted in what concerns data collection and disclosure, it is nonetheless more lax in what regards data (re)use, having implications for data operability in HE research.

The study is divided into the following sections. The Section 1 provides an introductory overview of Big Data and Section 2 reviews Big Data in HE research relative to such notions as anonymity, confidentiality and informed consent. Section 3 sets data privacy against a more recently regulatory European framework of data protection discussing several factors that call for a reconsideration of data privacy and access to private information in HE research. The last part of the study presents the conclusions and the implications for Big Data-centric HE research.

2. Big Data in HE Research

Big Data is supporting data for new data-intensive research and involves a dynamic growth of data that spans a variety of computational and social areas from finance to medicine, from marketing to climate science, from urban planning to health and security. It goes well beyond mere data generation, increasingly being not only a mode of acquiring knowledge through data but also an equally strong belief in the informative value of fine-grained structures, features and patterns inductively retrieved from observed massive datasets [8,9]. Big Data is generally conceptualized along structural and functional dimensions [10] and is supported by science information technology and statistics which operate as critical enablers of this newly emerging research paradigm. While the structural dimension describes Big Data features such as volume, velocity and variety [11], value and verification [12], flexibility [13], data source generation by social media applications, mobile devices and networks [14,15], a more functional approach to Big Data includes not only the related techniques and specific working instruments but also the utilization of innovative technologies designed to retrieve, collect, store, disseminate, manage and parse heterogeneous and large datasets [16,17].

Due to Big Data, traditional function-based static systems such as transport, energy and education have now become more dynamic data-driven and data-assisted networks that require complex analysis techniques seeking to identify relational and semantic interpretations of the phenomena underlying the data. Such is the case of HE which has routinely collected, analyzed and used student data (for example, data analytics regarding learning environments, management processes, retrieved from social media and/or from teaching and learning management systems) for a large number of purposes, including marketing, instructional design, student support, enrolment planning, operational resource allocation and institutional reporting. Such an availability of huge amounts of data compiled in institutional databases makes HE research an increasing data-intensive field that allows for the use of tools, techniques and processes from data science [18,19]. With such tremendous educational data available, HE researchers can now better examine subgroups within a wide diversity of populations [20], teachers can use visualization techniques [21], students can benefit from better designed learning trajectories [22], while universities can address their own institutional challenges [23,24]. Big Data in HE education can be used to support teaching, learning and administration albeit digital artifacts can generate data that may effectively intertwine with other categories [25,26]. Teaching and learning activities are generally analyzed by means of data analytics collected from more flexible, technology-enabled online learning platforms (MOOCs, Coursera, FutureLearn, Knewton Platform or Dreambox Learning, etc.), Open Universities learning management systems (Moodle, Blackboard, etc.) or user-generated data from social media, social networks, crowd-sourcing (Edmodo, ResearchGate, Academia.edu, Viki, etc.) and labor analytics (Jobseekers, Island recruiting, etc.). In turn, academic and management processes are particularly informed by data analytics retrieved from various data-dashboards that make available comparable international data to researchers, institutional managers, education providers and other stakeholders (GESIS, Eurostat, OECD Education GPS18, UNESCO databases, student tracking surveys, to name only a few). All these educational data analytics have emerged as learning analytics [27–29], academic analytics [30–32], research analytics and labor market analytics which contribute to the improvement of teaching and learning [33,34], form the basis for the redesign of new and competitive models of HE education, enhance organizational management and serve as a foundation for systemic change.

Big Data-based HE research is also increasingly reflective of the researcher's concern about the "what" at the expense of the "how" and "why" research questions that lie at the core of traditional education and social science (ESS) research [28]. Identifying causes rather than merely describing issues can better inform the successful design of optimal strategies to achieve desirable educational outcomes, albeit arguments in favor of correlational analyses that are apt to generate effective interventions even in the absence of causality have been more recently advanced [35]. Methodology-wise, Big Data HE research continues all methodology traditions, encompassing quantitative research defined by positivist epistemology, qualitative methods characterized by interpretivism and the overarching epistemology of pragmatism that characterizes the mixed methods in ESS research [36]. It embraces new forms of empiricism that go beyond the quantitative and qualitative traditions being shaped by dynamic data-changing contexts [37], findings and interactions [16,38]. Such a complex nature and huge potential of Big Data is increasingly making Big Data a mainstream paradigm in HE research which is both promising and challenging alike. However, the old ways of addressing research data undermine contemporary researchers' ability to maximize the qualities that make Big Data so appealing, not least in relation to the critical data privacy issues that will be discussed within the next sections.

2.1. Ethical Issues of Privacy

The ethical issues related to privacy, confidentiality and anonymity in using Big Data in HE are similar to those that apply to all ESS research, having emerged along with the relatively swift change of the research paradigm from old traditional research to more recent Big Data research. Recent studies have analyzed this transition and have captured the technical bottlenecks for Big Data institutional integration, comparison and usage, leading to rising concerns regarding both individual and institutional protection through authentication and security protocols [39]. Arguably, open data and Big Data brokers in HE seek to maintain *neutral* data collection and curation practices while representations of objects and subjects of study are generally warranted to serve effectively as data from which information can be meaningfully retrieved. However, the often unsafe haven of personal data repositories containing students' individual data that are commonly cross-matched with other end-user attributes and/or data correlations have inevitably increased the risks and liabilities of the HE data ecosystems to an extent that the latter exceeds the golden promised HE research returns. At the same time, much data remains, more often than not, fragmented and silo-stored whereas an increasing number of data brokers and private companies are beginning to curate HE education specific data and aggregate it into more meaningful 'analytics tools' that can be provided, and may be sold back, to third-parties or other education stakeholders. This poses another danger caused by which type of end-user of data-built products and services (organization, stakeholder, institution, etc.) is morally and legally entitled to reap the benefits from products and services derived from the use of student personal data, all the more so as the fast pace of commercialization of personal data is increasingly undermining subjects' trust and confidence while multiplying end-users' benefits, interests and fortunes. The process, outside careful control, may trigger serious threats to student safety and security as well as jeopardize the large utility of Big Data as appropriate consent decisions from subjects are mostly, if not always, treated *collectively* and not *individually*. Consequently, two conflicting views need reconciling within the argument of this paper: On the one hand, there is a huge potential in the completely new generated classes of information of Big Data to benefit HE and contribute critically to the public good; on the other, such a potential must be harnessed and placed in agreement with mechanisms of data access, privacy and ownership. In other words, while regulatory standards and financial incentives should be provided to attract owners into sharing data, individuals' interests and the society at large are equally to be served. This calls for a reconsideration of data information flows within current regulations across the whole spectrum that includes data acquisition and modes of aggregation to data usage and transmission across multiple HE entities. Specialized literature abounds in recent studies that seek to determine not only the extent to which privacy and Big Data are incompatible but also whether reconfigured choices can adequately meet the contemporary need for additional enforced constraints.

2.2. Anonymity, Confidentiality and Privacy

The emerging worrisome aspects of the fast-paced development of Big Data applications call for a new understanding of anonymity, confidentiality and informed consent. In HE research "anonymization is a methodological axiom" (p. 547) whereas confidentiality is a norm, both being central to ethical ESS research practice [40]. In general, privacy is determined by confidentiality and anonymity and it refers to the capacity to share knowledge about an individual selectively, though not publicly. While confidentiality defines a more or less explicit researcher-participant agreement on the basis of which no traceable record of the participant's data is to be disclosed, maintaining authorized limitations on information access and disclosure—anonymity, on the other hand,"obliterates the link between data and a specific person not so much to protect privacy but, in a sense, to bypass it entirely" [10] (p. 49). The nature of the social contract between higher education institutions (HEIs) and students has, in many ways, thus irrevocably changed [20,41], however, it is still grounded on the assumption that the collection, assessment and application of student data are done in the best interest of the student. In a research context, subjects (be they students, research participants, individuals, teachers, administration personnel, etc.) are therefore assured by researchers that data collected from them remain untraceable, either by employing pseudonyms (tokenization) for individuals and locations or by altering reported features of participants (including gender, age group, occupation, etc.). However, in practice, the specifics of education and social science (ESS) qualitative research may slightly complicate this process, the most commonly identified issues being those regarding anonymization in variable-sized educational communities, isolated geographical locations, and anonymization relative to intended research duration, audience variety, etc. In addition, anonymization of individuals is increasingly dependent on the researcher's complex decision of treating differently those who desire to be identified, on the one hand, and those who prefer to remain anonymous, on the other. Along this line, (p. 417) argue that "anonymization of the location of the research may prove more or less practical or impractical—depending on its distinctiveness—and more or less desirable, depending on its importance in providing the social context of the analysis that is being developed" [42]. Anonymity is also problematic with regard to long-term engagement with groups of subjects who know each other [43], in small educational communities, and as a particular issue in ethnographic work in HE or short-term observational research [44,45]. Boosted by the upsurge of Big Data, anonymity and informed consent have emerged as panaceas designed to counteract the troubling issues posed by the wide range of emerging applications of Big Data; thus "anonymization seems to take data outside the scope of privacy, as it no longer maps onto identifiable subjects, while allowing information subjects to give or withhold consent maps onto the dominant conception of privacy as control over information about oneself" [10] (p. 45). On the other hand, informed consent requires an even more careful approach in what concerns data management and data disclosure matters. The literature swirls around demonstrations of how anonymity fails to be provided when individual records and data sets contain more information than needed (such as location, vanity information, etc.) and when malevolent interests are at stake. Linkage and differencing attacks as well as re-identification mechanisms have been counteracted by a wide variety of approaches (such as K-anonymity, synthetic data and differential privacy) that are designed by data controllers to remove the threats for subject identification and make them hence unreachable [46]. In many different, yet converging ways, such responses to the challenges addressed by confidentiality and anonymity bespeak the centrality of privacy whose instrumental value lies in the complex modes of information flows. Privacy is critical, in particular because it is inherent in the values of confidentiality and anonymity and because it addresses a set of complex ethical issues in Big Data-centric research. Nonetheless, the interests of individual privacy and data confidentiality should be balanced against the social benefits of HE research access and use. This brings about a two-fold perspective on privacy on the basis of which researchers can not only manage privacy on behalf of individuals but also should profess a 'radical honesty' by acknowledging difficulties with true de-identification of research participants [47]. According to Nissenbaum (2011) [48], informed consent should be thus inclusive of information concerning the nature of harm acknowledged across process and over time, an understanding of the risk of privacy breaches as well as an optimal appraisal of privacy within the Big Data contextual conditions. However, students may evaluate their own privacy in quite different ways, ranging from trusting incomplete or excessive information to using heuristics that serve a wide array of personal interests and may standardize deflections from sound decision-making. This makes informed consent a complex freedom of choice that operates within a clear understanding of what that choice entails. This is to say that, with so much interconnected data, it is possible not only for a subject to release information about another subject, without that person knowing that their own data are being retrieved but also for a small group of individuals to volunteer information about themselves that serves as data characterizing larger groups of students. Along these lines, the challenge to the HE research community remains to balance the autonomy in informed consent and research utility with the *risk* of providing access and to engage in collaborative reuse of data.

3. Privacy, HE Research and the New European Perspective

The development of cloud storage, high-speed data networks and fast development of data processing mechanisms has spurred the establishment of a harmonized legal instrument that would overrule all national legislations [49]. The European Union's General Data Protection Regulation (hereafter abbreviated as GDPR) entered into force on 25 May 2018, as a substitute of the old legal framework dating back to 1995. In contrast to the United States legal framework that pursues a more sectoral regulation (health privacy, children's privacy, etc.), the European GDPR approach applies an overarching perspective within a broad definition of processing personal data, public/private actors and operations, and introduces new compliance obligations and a set of high sanctions associated with them.

Under GDPR, personal data processing involves a wide range of operations, from data collection to data erasure, which include retrieval, storage, transmission and use. The laid-out principles that govern personal data analysis require that Data Controllers and Data Processors operate fairly, transparently and legally, using data exclusively for the purposes for which it was collected and deleting it when no longer necessary. The data subject is therefore entitled to access, modify, delete their personal data and provide discretionary consent to its use. Central to the new GDPR is the core view that individuals must authorize approval for their personal information to be processed. This involves a number of consequences for research, stemming, on the one hand, from the apparently conflicting objectives within Europe to facilitate the information flow across the European Research Area via cross-border access, and on the other, from making such flow of personal data comply with the legal stipulations concerning the fundamental rights to privacy and to personal data protection. The introduction of the derogation in Article 89 of GDPR, which allows for a general personal data use by exempting research from the consent obligation, raises justified concerns regarding its *real* applicability for research purposes [50,51]. While the GDPR admits the necessity to foster various research types (such as scientific, statistical, historical research), it does not, however, provide a full definition of what represents scientific research, being unclear when stating that "the processing of personal data for scientific research purposes should be interpreted *in a broad manner* including for example technological development and demonstration, fundamental research, applied research and privately funded research" (GDPR: Recital 33, our emphasis). Furthermore, associated concerns regarding this derogation facilitate a loop that permits personal data processing in such ways which, albeit lawful, could not only cause harm to subjects [49] but also benefit end-users whose purposes may not be exclusively research-oriented. A closer look at HE research and the ways in which Big Data and the new regulations impact the nature of HE research will reveal five core identifiable and problematic issues along the Big Data collector-subject-user cline which call, we hold, for a revision of privacy and its procedural technicalities in HE research projects.

1. Firstly, the lack of GDPR explicitness regarding data property neutralizes the relationship between data-owner and data user. In more specific terms, there is no clear indication about who owns the new research data, whether it is the information subject (the student, teacher), the data custodian (the data collecting HE institution), the researcher (the added-value contributor to research), the society at large, or simply the ultimate data buyer (a HE stakeholder or any third party purchasing an interest in the data). Such lack of clarity additionally stems from the fact that some regulations treat data as information whereas others treat it as property, which complicates matters and poses data management risks arising from the everlasting nature of digital data.

2. Secondly, there is a conflicting nature between the general provision of lawful data collection practices, scope limitation and personal data minimization on the one hand and the exemption for scientific research on the other. In other words, while other sectors of research (such as biomedical, health research) are more restricted in ensuring good data governance and rigorous data-centric security controls over how they obtain, store, process and analyze data—in HE research, the specific ethical issues are not tackled through the GDPR as the EU has no conferred competency to harmonize legislations in this field. Hence, the restrictive content of the controversial Article 89 that specifies rules that must be harmonized with the national or EU laws shows that the EU has merely a support competency leaving room for national legislations to be enforced. However, how far and how well can these safeguards capture all the mechanisms for personal data processing in HEIs and Big Data HE research? How is the tension between the public nature of HE research and the corporate stakeholders alleviated when dealing with sectorial Codes of Conduct, binding private stakeholder rule sand data protection seals? The implications of such safeguards and tools for the HE research communities that seek data protection standardization while maximizing data sharing advantages operate both within HEIs and between third parties (public or private) and HEIs. For example, in the case of a research making use of labor market analytics that may help universities identify jobs for their graduates and which may inform institutional/national/regional graduate tracking studies, the privacy protocols regarding the Big Data that are mined from millions of advertised jobs as well as the management of graduates' CVs must be harmonized across the multiple entities that use and process that data. Furthermore, the data subjects' interests must be weighed against those of third parties' Data Controllers that must ensure data protection and privacy. Thus, while consent may look, within a university, like an attractive legal research protocol due to its ease of application, with third parties' (public and private) or corporate stakeholders' consent, it may be considered to be a more optional legal obligation, which means that it can be removed at any time by the participating data subject. This mismatch of visions, protocols and unstandardized approaches are likely to stall the research process, complicate collaboration and hinder data use and data portability.
3. A third critical aspect concerns *the type of consent in HE research* albeit the GDPR defines and sets clear provisions for pseudonymization, encryption, informed consent and anonymized data. The broad nature of consent and the GDPR provision, according to which, under certain provisions, "data subjects should have the opportunity to give their consent only to *certain areas of research or parts of research projects*" (Recital 33 GDPR, our emphasis), are indicative of a problematic issue for the data collector-subject-user relationship that we explore within this study. A large-scale Big Data project, for example, examining a number of educational communities of students and professors alike in which anonymization of research subjects has been made, names of data subjects have been removed and informed consent has been obtained, is likely to contain pictures of buildings, streets and open spaces besides research data per se. The broad consent, may cover here, areas regarding both subjects' personal data and their use for different project sections/areas. However, there is an unclear area concerning the consent regarding pictures. The difficulty in handling the pictures lawfully would allow for identification of both research-included students and out-of-survey pooled individuals in identifiable settings. Additionally, in HE research informed by social media analytics, data processing and student personal data are manageable under the provisions of a consent that is exclusively restricted to researchers (data collector) and participating subjects and not to data users (any third party, HEI or stakeholder) and/or data portability. Extending this critical issue, since Big Data analysis may lead to all sorts of discriminations, stereotype perpetuation, life-choice limitation, judgments, harassment, etc., the consent should be descriptive of the individual's agreement to personal data processing, the scope of all activities agreed therein as well as, most importantly, the expectations that are likely to be violated if the subject agrees to provide personal data. That is, rather than describe what will be achieved via and during the actual research, the focus in a HE informed consent

should be unambiguously on the “what?”, “how?”, “for whom?” and “for what?” (research part) in the undertaken research.

4. Another important aspect is related to the fragmented nature of ethics in Big Data-based research and the uneven treatment of the data collector-subject-user relationship across the EU-28 HE systems. Sectorally, HE research is principled in ethical guidelines, codes of conduct, standards for ethical research practices, etc., and sets of criteria are defined for proper research conduct in order to maximize research quality and address research integrity. However, a single look at the EU28 National Codes of Ethics (NCE) shows that there is considerable variation, fragmentariness and diversity in approaches. While biomedical ethics is generally included in NCE, many countries have separate bioethics codes and activities carried out within separate national ethics councils for the life sciences. The European Code of Conduct for Research Integrity, published in its last 2017 version [52], has emerged as an attempt at unifying national approaches and visions, being relevant to both privately funded and public research (researchers, universities, funding institutions, academies, learned societies, publishers, etc.) despite its acknowledged limitations in use and applicability. Since not only social, economic political and technological factors but also changes altering the research environment are very likely to impact research regulating values and principles, the Code remains a living document that needs constant updating and harmonizing.
5. A fifth factor refers to data portability, use and data-sharing issues in HE Big Data research. The legitimate need and necessity (of stakeholders, HEIs, other entities, etc.) to ensure (and thereby gain from) research accessibility to personal data often clashes with national privacy laws and Eurostat policies [49], which is why cross-border data transfer brings about issues pertaining to effectiveness and lawfulness of use and reuse of data. As a result, more effective data management systems, such as FAIR—based on the four foundational principles of Findability, Accessibility, Interoperability, and Reusability—are in place and have been more recently instrumentalized for fields such as digital humanities [53]. Since the context of the ongoing debate and critical reflection regarding data-sharing and reuse of data is very complex and multifaceted, we will only refer to the more general data management plans for data generated in HE publicly funded experiments that aim at striking a balance between the public data openness and privacy protection. Two conflicting trends define the current data ecosystem in HE research: On the one hand, the emergence of a variety of large-scale data repositories that range from a university to global repositories such as Mendeley Data, DataHub, FigShare, EUDat and include multi-formatted available data types is apt to complicate privacy protection of research subjects due to lack of insufficient restrictions on the deposited data descriptors. On the other hand, data management in HE research, in line with Open Science and FAIR principles, fosters an increasingly swift transition from human-readable data to machine-readable data, which requires the “interpretive” judgment of HE researchers to be coupled with the scientific effort of data scientists. While the former is involved in data (re)use and the latter in data collection and curation, the participating subject’s position is somewhat unclear, being, at best, fragmented between a wide-scoped, multi-user and multi-staged research. For example, in projects dealing with large sets of Moodle- or Blackboard-generated data on student interactions, academic learning progression, etc., personal data protection falls with the learning management system (LMS) administrator. The researcher whose interest lies in assessing students’ choices, attitudes, learning environment or progress over a range of tasks and period of time may use the information for general analyses; however, if more in-depth examinations are targeted and data collected on stigmatized, or otherwise sensitive behaviors are of research interest, the responsibility for the informed consent and privacy of subjects involved will rest with the researcher as well. In this case, data privacy implicates protecting the individuals whereas confidentiality involves protecting the information, hence, a subject’s personal data while involving both remains arbitrary and largely project-determined. Additionally, the distinction between “the information a subject provides” (such as personal data or by creating a LMS

platform account) and "the information data users (researchers, third parties, administrators, etc.) get from their use of a subject's data" underscores a policy and practices that any Google map user is likely to understand. Complications may arise when research is large-scale and multi-phased, when data is shared across multiple entities and when research involves large datasets of de-identified information whose amount of stripping in qualitative datasets may prove problematic for research data validation.

4. Conclusions

Ethics in privacy data management has moved from being a marginal issue in traditional HE research to one of the core concerns in Big Data HE research. In this paper, we have not only mapped the different challenges and applications of protective and responsible data privacy procedures in research, but also argued that the ethical dimension of Big Data-based research should seek to achieve a balance between research scope, utility and subjects' privacy on the one hand and the benefits and risks of providing access and engage in collaborative use of data on the other. While such tension is reflective of the huge potential of Big Data in HE research that is both promising and challenging for researchers, it nonetheless entails a pressing necessity to understand the wide-ranging implications of Big Data use and operationalization for the future HE research agendas. Such a critical implication concerns data privacy that, we hold within the argument of this study, can be successfully ensured by optimal information stewardship, responsible data management and more protective data-centric security.

Author Contributions: Both authors contributed in equal shares to this article and bear full responsibility for its content. All authors have read and agreed to the published version of the manuscript.

Funding: Project financed by Lucian Blaga University of Sibiu and Hasso Plattner Foundation research grants LBUS-IRG-2020-06.

Acknowledgments: The authors would like to acknowledge the Lucian Blaga University of Sibiu and the Hasso Plattner Foundation for providing financial support through LBUS-IRG-2020-06 which made this study possible.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

EU	European Union
HE	Higher Education
NCE	National Codes of Ethics
GDPR	General Data Protection Regulation

References

1. Mehmood, A.; Natgunanathan, I.; Xiang, Y.; Hua, G.; Guo, S. Protection of Big Data Privacy. *IEEE Access* **2016**, *4*, 1821–1834. [CrossRef]
2. Sokolova, M.; Matwin, S. Personal Privacy Protection in Time of Big Data. In *Challenges in Computational Statistics and Data Mining*; Springer: Berlin, Germany, 2015; pp. 365–380.
3. Jain, P.; Gyanchandani, M.; Khare, N. Big data privacy: A technological perspective and review. *J. Big Data* **2016**, *3*, 25. [CrossRef]
4. Lane, J.; Stodden, V.; Bender, S.; Nissenbaum, H. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*; Cambridge University: Cambridge, UK, 2013.
5. Hoffman, S. Medical Big Data and Big Data Quality Problems. *SSRN Electron. J.* **2014**, *21*, 289. [CrossRef]
6. Mattioli, M. Disclosing Big Data. *Minn. Law Rev.* **2014**, *99*, 535.
7. Khan, A. Book review: Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for Human Future at the New Frontier of Power. *Soc. Chang.* **2019**, *49*, 735–738. [CrossRef]
8. Boyd, D.; Crawford, K. Critical Questions for Big Data. *Inf. Commun. Soc.* **2012**, *15*, 662–679. [CrossRef]
9. Housley, W.; Procter, R.; Edwards, A.; Burnap, P.; Williams, M.; Sloan, L.; Rana, O.F.; Morgan, J.; Voss, A.; Greenhill, A. Big and broad social data and the sociological imagination: A collaborative response. *Big Data Soc.* **2014**, *1*, 1–15. [CrossRef]

10. Barocas, S.; Nissenbaum, H. Big Data's End Run around Anonymity and Consent. In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*; Lane, J., Victoria, S., Bender, S., Nissenbaum, H., Eds.; Cambridge University Press: Cambridge, UK, 2014; pp. 44–75.
11. Douglas, L. 3D Data Management: Controlling Data Volume, Velocity and Variety. Gartner Report. 2001. Available online: https://gtnr.it/2VqBPPs (accessed on 24 April 2020).
12. Maneth, S.; Poulovassilis, A. Data Science. *Comput. J.* **2016**, *60*, 285–286. [CrossRef]
13. Kitchin, R. Big data and human geography. *Dialog. Hum. Geogr.* **2013**, *3*, 262–267. [CrossRef]
14. Snijders, C.; Matzat, U.; Reips, U.D. "Big Data": Big gaps of knowledge in the field of internet science. *Int. J. Internet Sci.* **2012**, *7*, 1–5.
15. Ward, J.S.; Barker, A.D. Undefined by Data: A Survey of Big Data Definitions. *arXiv* **2013**, arXiv:1309.5821.
16. Tolle, K.M.; Tansley, D.S.W.; Hey, T. The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]. *Proc. IEEE* **2011**, *99*, 1334–1337. [CrossRef]
17. Dede, C.; Ho, A.; Mitros, P. Big Data Analysis in Higher Education: Promises and Pitfalls. Educause [Review]. 2016. Available online: https://bit.ly/2HSWlk8 (accessed on 2 March 2020).
18. Waller, M.A.; Fawcett, S.E. Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *J. Bus. Logist.* **2013**, *34*, 77–84. [CrossRef]
19. Klašnja-Milićević, A.; Ivanović, M.; Budimac, Z. Data science in education: Big data and learning analytics. *Comput. Appl. Eng. Educ.* **2017**, *25*, 1066–1078. [CrossRef]
20. Mayer- Schonberger, V.; Cukier, K. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*; Houghton Mifflin Harcourt: Boston, MA, USA, 2013.
21. Greer, J.; Mark, M. Evaluation Methods for Intelligent Tutoring Systems Revisited. *Int. J. Artif. Intell. Educ.* **2015**, *26*, 387–392. [CrossRef]
22. McKenney, S.; Mor, Y. Supporting teachers in data-informed educational design. *Br. J. Educ. Technol.* **2015**, *46*, 265–279. [CrossRef]
23. Daniel, B.K.; Butson, R. Technology enhanced analytics (TEA) in higher education. In Proceedings of the International Conference on Educational Technologies (ICEduTech), Kuala Lumpur, Malaysia, 29 November–1 December 2013; Kommers, P., Issa, T., Sharef, N.M., Isaıas, P., Eds.; IADIS Press: Lisbon, Portugal, 2013; pp. 89–96.
24. Florea, S.; Cecile, H.M. Governance and Adaptation to Innovative Modes of Higher Education Provision. *Manag. Sustain. Dev.* **2014**, *6*, 35–38. [CrossRef]
25. Daniel, B. Big Data and analytics in higher education: Opportunities and challenges. *Br. J. Educ. Technol.* **2014**, *46*, 904–920. [CrossRef]
26. Beneito-Montagut, R. Big Data and Educational Research. In *The BERA/SAGE Handbook of Educational Research: Two Volume Set*; SAGE Publications: New York, NY, USA, 2017; pp. 913–934.
27. Clow, D. An overview of learning analytics. *Teach. High. Educ.* **2013**, *18*, 683–695. [CrossRef]
28. Daniel, B.K. (Ed.) *Big Data and Learning Analytics in Higher Education: Current Theory and Practice*; Springer: New York, NY, USA, 2017.
29. Prinsloo, P.; Slade, S. Mapping Responsible Learning Analytics. In *Responsible Analytics and Data Mining in Education*; Routledge: London, UK, 2018; pp. 63–79.
30. Campbell, J.P.; DeBlois, P.B.; Oblinger, D.G. Academic analytics: A new tool for a new era. *Educ. Rev.* **2007**, *42*, 40.
31. Shields, R. Following the leader? Network models of "world-class" universities on Twitter. *High. Educ.* **2015**, *71*, 253–268. [CrossRef]
32. Souto-Otero, M.; Beneito-Montagut, R. From governing through data to governmentality through data: Artefacts, strategies and the digital turn. *Eur. Educ. Res. J.* **2015**, *15*, 14–33. [CrossRef]
33. Van Harmelen, M. Analytics for Understanding Research: CETIS Analytics Series. Available online: http://publications.cetis.org.uk/wp-content/uploads/2012/12/Analytics-for-Understanding-Research-Vol1-No4.pdf (accessed on 24 April 2020).
34. Kobayashi, V.B.; Mol, S.; Kismihok, G. Labour Market Driven Learning Analytics. *J. Learn. Anal.* **2014**, *1*, 207–210. [CrossRef]
35. Mayer-Schönberger, V. Big Data for cardiology: Novel discovery? *Eur. Hear. J.* **2015**, *37*, 996–1001. [CrossRef]
36. Greene, J.C. Engaging Critical Issues in Social Inquiry by Mixing Methods. *Am. Behav. Sci.* **2012**, *56*, 755–773. [CrossRef]

37. Kitchin, R. Big Data, new epistemologies and paradigm shifts. *Big Data Soc.* **2014**, *1*, 205395171452848. [CrossRef]
38. Daniel, B.K. Big Data and data science: A critical review of issues for educational research. *Br. J. Educ. Technol.* **2017**, *50*, 101–113. [CrossRef]
39. Miyares, J.; Catalano, D. Institutional Analytics Is Hard Work: A Five-Year Journey. *Educ. Rev.* **2016**. Available online: https://er.educause.edu/~{}/media/files/articles/2016/8/erm1656.pdf (accessed on 2 March 2020).
40. Nespor, J. Anonymity and Place in Qualitative Inquiry. *Qual. Inq.* **2000**, *6*, 546–569. [CrossRef]
41. Bauman, Z.; Lyon, D. *Liquid Surveillance*; Polity Press: Cambridge, UK, 2013.
42. Wiles, R.; Crow, G.; Heath, S.; Charles, V. The Management of Confidentiality and Anonymity in Social Research. *Int. J. Soc. Res. Methodol.* **2008**, *11*, 417–428. [CrossRef]
43. Walford, G.; Massey, A. (Eds.) *Explorations in Methodology*; Studies in Educational Ethnography; Emerald Publishing Limited: Bingley, UK, 1999.
44. Moosa, D. Challenges to anonymity and representation in educational qualitative research in a small community: A reflection on my research journey. *Comp. J. Comp. Int. Educ.* **2013**, *43*, 483–495. [CrossRef]
45. Troman, G.; Jeffrey, B.; Walford, G. (Eds.) *Methodological Issues and Practices in Ethnography*; Studies in Educational Ethnography; Emerald Publishing Limited: Bingley, UK, 2005.
46. Dwork, C.; Lane, J.; Stodden, V.; Bender, S.; Nissenbaum, H. Differential Privacy: A Cryptographic Approach to Private Data Analysis. In *Privacy, Big Data, and the Public Good*; Cambridge University Press: Cambridge, UK, 2014; pp. 296–322.
47. Wilbanks, J.; Lane, J.; Stodden, V.; Bender, S.; Nissenbaum, H. Portable Approaches to Informed Consent and Open Data. In *Privacy, Big Data, and the Public Good*; Cambridge University Press: Cambridge, UK, 2014; pp. 234–252.
48. Nissenbaum, H. A Contextual Approach to Privacy Online. *Daedalus* **2011**, *140*, 32–48. [CrossRef]
49. Elias, P.; Lane, J.; Stodden, V.; Bender, S.; Nissenbaum, H. A European Perspective on Research and Big Data Analysis. In *Privacy, Big Data, and the Public Good*; Cambridge University Press: Cambridge, UK, 2014; pp. 173–191.
50. Chassang, G. The impact of the EU general data protection regulation on scientific research. *Ecancermedicalscience* **2017**, *11*, 709. [CrossRef]
51. Mondschein, C.F.; Monda, C. The EU's General Data Protection Regulation (GDPR) in a Research Context. In *Fundamentals of Clinical Data Science*; Kubben, P., Dumontier, M., Dekker, A., Eds.; Springer: Cham, Switzerland, 2019.
52. *The European Code of Conduct for Research Integrity*; All European Academies: Berlin, Germany, 2017. Available online: https://bit.ly/2VmdwlQ (accessed on 14 May 2020).
53. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.