Issues in Information Systems
Volume 20, Issue 1, pp. 178-185, 2019

PriVacy: THE MISSING "V" IN BIG DATA ANALYTICS

Christopher B. Davison, Ball State University, cbdavison@bsu.edu
Allen D. Truell, Ball State University, atruell@bsu.edu
Edward J. Lazaros, Ball State University, ejlazaros@bsu.edu
Jensen J. Zhao, Ball State University, jzhao@bsu.edu

ABSTRACT

The descriptive models of Big Data found in the research literature generally neglect privacy as a component of big data. Utilizing Cavoukian's (2011) Privacy by Design framework as a foundation, the authors propose an additional "V" to the Big Data descriptive model: priVacy. The purpose of this paper is to present an expanded theoretical model of Big Data with privacy as a necessary component. The implication is that privacy is an aspect of many data science domains including data analytics and cybersecurity with many concomitant legal and political ramifications.

Keywords: Privacy, Big Data, Cybersecurity, Data Model, Design, Data Analytics.

INTRODUCTION

Big data is a ubiquitous facet of modern life. Big data and data analytics provide competitive advantage and profitability to organizations. As such, the market-driven rush to obtain this data often results in overlooked privacy considerations. Consider the IBM Watson Health analytics that aggregates massive amounts of lifestyle and behavior data related to healthcare. This data is evaluated for socioeconomic factors and provided to insurance companies (Allen, 2018). In this paper, the authors draw upon current research on big data models and privacy research to create a privacy cognizant big data model.

To present privacy in the big data model context, the authors begin this paper by providing a literature review. The literature review will set the context and provide definitions associated with big data and information privacy. Following that, an expanded big data framework will be presented and discussed.

LITERATURE REVIEW

Privacy as a Right

Warren and Brandeis (1890) argue the right to privacy in their seminal work published in the Harvard Law Review. This work is widely regard as the first legal conception of privacy as a right (Glancy, 1979) and is currently cited on Google Scholar more that 22,000 times.

In addition to the conceptualization of privacy as a right in the Warren and Brandeis (1890) work, they argue that the law lacks protections for what they deem the private person. They cite malicious instances of photography and the publication of unauthorized photographs as evidence for the need to codify privacy as a right with legal redress. Furthermore, they recant the "evil of the invasion of privacy by newspapers" (p. 195) as a particularly egregious affront to individual privacy.

Solidifying the concept of privacy, Warren and Brandies (1890) cite a Judge Cooley's definition of privacy as the right to be let alone. Today, an expanded definition of information privacy is given as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" (Cavoukian, 2011, p. 6).

Interestingly, Warren and Brandeis (1890) discuss the intricacies and complexities of modern life making it necessary for individuals to be able to retreat to a private space and gain solitude Today, most people would agree that modern life is even more intricate and complex thus making the right to privacy even more necessary.

Privacy and Cybersecurity

It is often the case that technologists rush to create technologies and provide privacy and security controls after the fact. Whether a rush to market by businesses or to be the first to publish by scientists, this rush to create quite often leads to neglect. Privacy and security are sometimes at odds with the basic desire to just make something work . A prototype is often built with little to no privacy or security in its working structure. Adding privacy and security after the prototype further complicates the goal of keeping data private and secure (Cavoukian, 2011). This is known as the concept of shoehorning.

Data privacy and cybersecurity issues are often interrelated (Murrill, Liu, Thompson, 2012). Many sorts of data breaches or information theft can result in the loss of privacy and private information. Landwehr (2012) argues that privacy and cybersecurity are joined. To make a computational system secure is to control the release of data. This secured control over the data is what Landwehr considers privacy.

Cybersecurity regarding small, mobile sensors (e.g., fitbands) and the networks that carry the data is another large domain of research. Handel, Shcrieber, Rothermaler, and Ivanova (2018) point out that hardly any manufacturers encrypt their Bluetooth data. From a privacy perspective, this indicates medical and health related data are being broadcast from these devices. Many fitband devices transmit medical information such as heart rate, galvanic skin response and even sleep activity. Any adversary within Bluetooth range can easily intercept this traffic.

Big data presents special cybersecurity issues. The very nature of big data is large quantities of data at high speeds. According to Benjelloun and Lahcen (2019), big data security generally aims to ensure monitoring, detect vulnerabilities, threats, and abnormal behaviors. However, due to the voluminous nature of the data, adding security layers to an already burned system can create more overhead and less responsiveness. Thus, a trade-off between data security and systems performance exists in the big data domain.

Privacy in the Legal and Regulatory Environment

The Federal Privacy Act of 1974 provides general privacy protection laws and controls the information privacy practices of federal agencies. The Department of Justice stipulates that this law establishes a code of fair information practices that governs the collection, maintenance, use, and dissemination of information about individuals. This law applies only to federal agencies and, while often reflected in state law, does not apply beyond federal agencies.

The Electronic Communications Privacy Act of 1986 (ECPA) is an extension of the federal wire-tapping (telephone) to transmission of electronic data by computer. According to the Department of Justice, the ECPA has been amended by the Communications Assistance for Law Enforcement Act (CALEA) of 1994, the USA PATRIOT Act (2001), the USA PATRIOT reauthorization acts (2006), and the FISA Amendments Act (2008). Criticisms of the ECPA include that it is outdated by modern technologies, that the law does not provide adequate protections to all forms of electronic communications, and that the law is not comprehensive in protecting "in transit" data communications from intercept.

The Stored Communications Act was enacted as Title II of the ECPA. It is a law that addresses voluntary and compelled disclosure of stored wire and electronic communications and transactional records held by third-party internet service providers (ISPs). Th intent of the Act is to limit the ability of the government to force ISPs to relinquish data such as log files and content. Additionally, the Act provides criminal penalties for anyone who intentionally accesses (without authorization) a facility through which an electronic communication service is provided.

The Computer Fraud and Abuse Act of 1984 (CFAA) was enacted as an amendment to existing computer fraud law. The law prohibits accessing a computer without authorization, or in excess of authorization. Prior to computer-specific criminal laws, computer crimes were prosecuted as mail and wire fraud, but the applying law was often insufficient.

There are a number of other U.S. laws that provide legal remedy for consumers including the Identity Theft Enforcement and Restitution Act of 2008. However, most of these laws concern data theft, monetary loss, and fraudulent activity and provide remedy for those crimes. Privacy protection appears to be of lesser importance than fraud and theft.

In the European Union, the General Data Protection Regulation 2016/679 (GDPR) was created in May of 2016 and implemented in May of 2018. It is a regulation on data protection and privacy for all individuals within the European Union (EU) and the European Economic Area. The law also addresses the transfer of personal data outside of the EU. Compliance with GDPR resulted in a number of lawsuits (e.g., European Center for Digital Rights vs. Google) and

European users being blocked from websites (e.g., The Chicago Tribune). In their research on information privacy, Verteş-Olteanu and Racolta (2019) present a criticism of the GDPR and the European effort to legislate and enforce privacy rights.

Privacy Regulatory Insufficiency

Bartholomew's (2016) research on smart-grid data privacy reveals that lack of both regulatory control of big data and the legal remedies for breach of privacy. He addresses the push for smart-grid technology and modernization of the electrical infrastructure without the concomitant privacy protection. He recommends that Massachusetts institute privacy regulations that build upon the state's existing privacy laws, incorporate best practices of other states, support and conform to federal privacy laws, and adhere to the DOE and the Federal Smart Grid Task Force165 Voluntary Code of Conduct (VCC). One intent of the VCC is to encourage innovation while appropriately protecting the privacy and confidentiality of customer data. It should be noted that the VCC is not law but is intended to be voluntarily adopted by smart-grid providers.

Zarsky (2019) discusses an interesting and often over-looked aspect of data privacy: manipulation. Zarsky proposed that big data and the increasing sophisticated techniques available to analyze big data can lead to manipulation. He defines manipulation as influencing people to behave in a certain manner (tailored responses) and continually customizing these behaviors by virtue of the highly available big data. Furthermore, Zarsky argues that this manipulation is often non-transparent, and the manipulated individual is unaware of the situation. From this manipulation-based paradigm, Zarsky argues for the need of privacy preserving technologies.

Privacy by Design

To address the need for privacy in technologies, Cavoukian (2011) presents the concept of Privacy by Design (PbD). In her work, she discusses privacy controls that are embedded in every stage of development as opposed to being "bolted on" (p. 28) after the fact. She presents a holistic approach to privacy that encompasses technology design aspects, life cycle protections, legal aspects, and political issues.

In PbD, privacy is foundational design principle. Privacy comes before and during the build; not an after the fact add-on or adjustment. It goes beyond legal and regulatory compliance to provide a design framework that is flexible and encourages innovation while still delivering results. For a graphical depiction of Cavoukian's (2001) PbD model, see Figure 1.

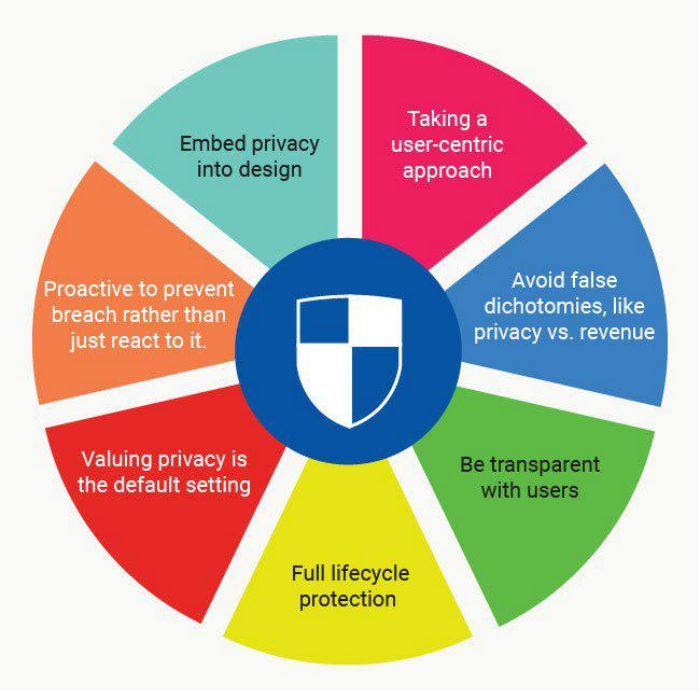

Figure 1. Privacy by Design. This figure illustrates the seven principles of Privacy by Design (Mackie, 2018).

Big Data Model

As it is a recent phenomenon, big data's definition is a moving target. The size and scope of what constitutes big data can vary. In the general sense, the size of big data is an *N* where traditional Relational Database Management Systems (RDBMS) cannot scale to process the enormity of the data (EMC Education Services, 2015). In today's sense, this can be measured in petabytes of information. However, as that number is a variable and data management software systems undergo continual updating and improving, it could be much larger tomorrow. According to Hashem, Yaqoob, Anuar, Mokhtar, Gani and Khan (2015), big data is more than size but it is encompasses the integration of techniques and technologies to uncover hidden values in complex, massive, and heterogeneous data sets.

Laney (2001), provides a more holistic and three dimensional definition for the growth and challenges associated with big data. His definition addresses not only the size but the scope of big data: The Three Vs (Volume, Velocity, and Variety). Laney's research is graphically depicted in Figure 2.

In Laney's (2001) work, volume refers to sheer size (quantity) of the data. Volume is perhaps the most discussed and understood characteristic of the definition as the term big data itself implies volume and is relative to size.

Velocity refers to the speed at which big data is generated and processed. An example of this is the streams of data generated in multi-modal sensing environments. Large sensor arrays generating massive streams of low-level data can potentially overwhelm networks and data management systems. As these sensor streams converge to their destination, the sheer velocity of the data presents interesting challenges.

Finally, variety refers to the heterogeneous nature of big data. The data could be structured in the typical RDBMS sense or quite unstructured such as combinations of video, text, and various file formats streaming into a big data system. Managing, cataloging, indexing, and providing retrievals and analytics on top of a variety of data pose unique challenges to system designers.

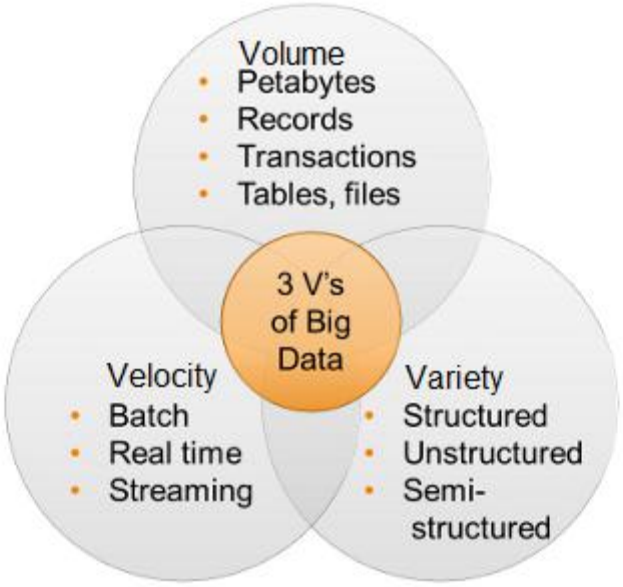

Figure 2. The three V's of big data. This figure illustrates the three original elements of big data (Big Data (Hadoop) Glossary, 2015).

In more recent research, the model depicting big data has expanded in the number of Vs that define the concept (Davison, 2015). The added Vs to the model include: veracity, variability, and value (see Figure 3). There are some models that expand the Vs further, however, these models are new and subject to both publication and debate.

The concept of veracity in the big data framework pertains to noise. With too much noise in the data, the data will not be accurate. As such, the data cannot be trusted to be truthful.

Variability in big data can refer to several phenomena. Often, variability refers to change and the rate of change in data. However, it can mean inconsistency in data that is addressed by outlier detection methodologies. It may also indicate data modeling discrepancies or data type mismatches.

The value of big data refers to its usefulness to business. In that context, the profitability of the data is examined. Often, large and expensive infrastructures are required for big data analytics. If the data produced does not justify the expensive of its procurement and analysis, then the value of the big data is suspect.

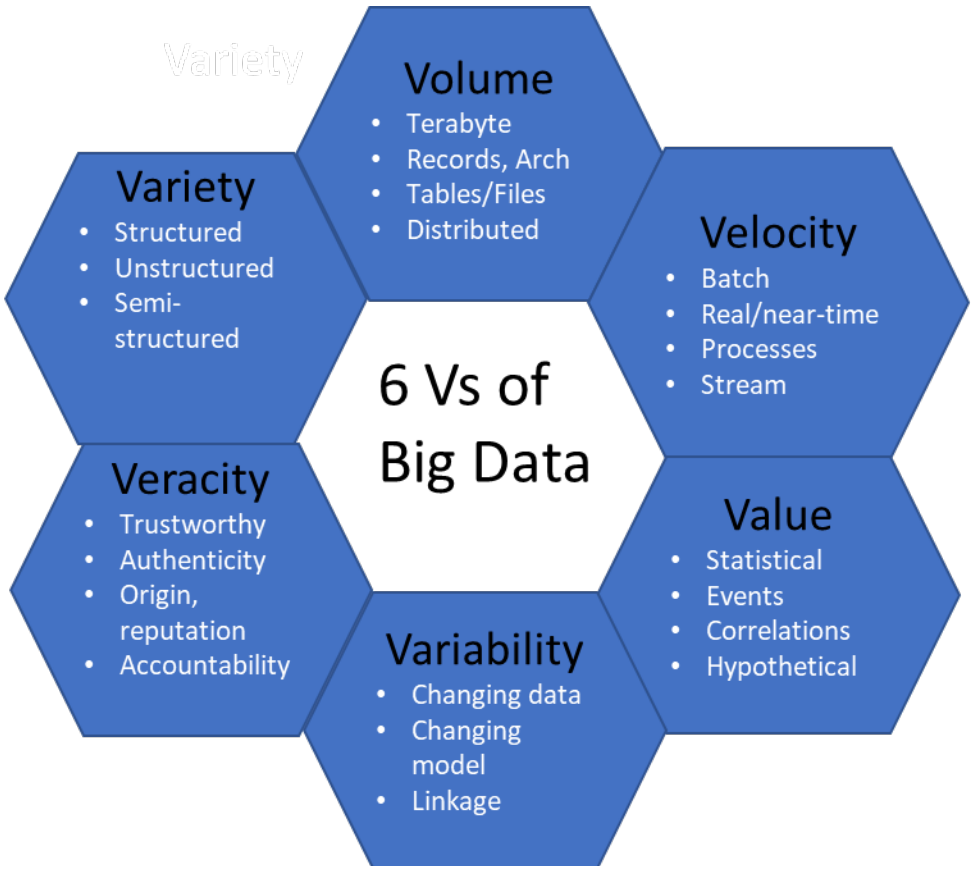

Figure 3. The six V's of big data. This figure illustrates an expanded framework of big data (Andreu-Perez, Poon, Merrifield, Wong, & Yang, 2015).

PRIVACY INCLUSIVE BIG DATA MODEL

Notably absent in the research and models on big data is the concept of privacy. As privacy and security in big data are often interrelated (Murrill, Liu, Thompson, 2012), the authors of this paper argue that privacy, and by extension: security, should be a necessary component of the big data descriptive model.

In this proposed big data model, privacy would encompass many aspects of security due to the interrelated nature of the two concepts as well as privacy policies and informed consent. PbD in big data, as Cavoukian, (2011) discusses, should be an integrated design component. Privacy of big data should be given equal weight as other components such as volume, variety, and velocity (see Figure 4).

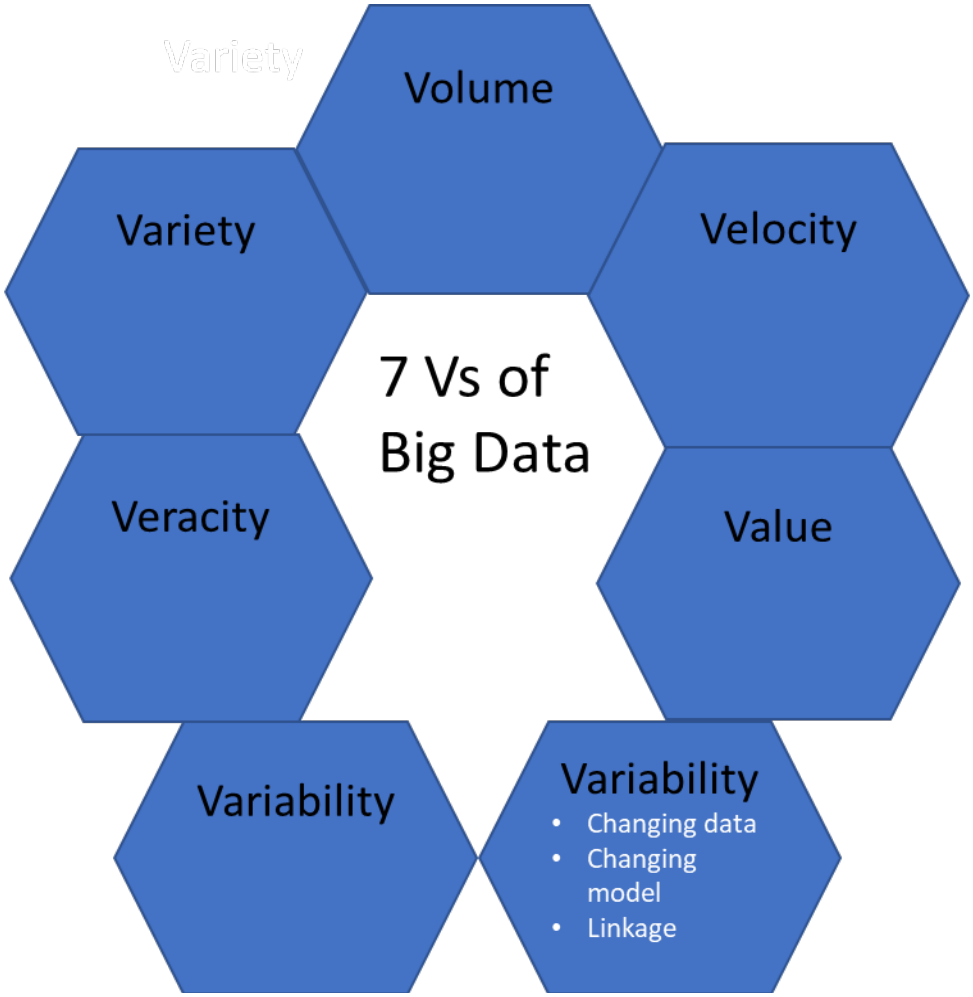

Figure 4. The seven V's of big data. This figure illustrates the model of big data proposed in this paper.

Aspects of Privacy in the Privacy Inclusive Big Data Model

Cybersecurity is an important component of big data (Kuhn, 2018). As cybersecurity and privacy are interrelated, the authors incorporated security constructs as a factor of privacy in this proposed big data model. The security issues discussed by Benjelloun and Lahcen (2019) such as monitoring, detection, threat assessment, identifying vulnerabilities and abnormal behaviors, and discovering threats should all be part of the privacy enhanced big data model. A rigorously developed (PbD) privacy/cybersecurity construct also provides regulatory compliance and added legal protection.

Strong, understandable and transparent privacy polices within the big data model is another important construct. Clearly elucidating the types and amounts of data (Rice & Sniffen, 2018) should be part of a well-crafted privacy policy. Additionally, the usage, processing, storage (retention), sharing, and legal requirements governing the data should be explained in the privacy policy. In his arguments for a well-conceived privacy policy, Brook (2011) makes the case for Notice (conspicuous disclosure of privacy polices), Choice (where customers can opt in or out), Access/Correction (allowing customers access to and correction of data), Security (steps taken to guard against unauthorized access) and Enforcement (steps taken to enforce the privacy policy).

Finally, informed consent is another component of the privacy aspect of big data. Brook (2011) terms informed consent as "choice" (para. 19) to participate in data collection or withdraw from data collection activities. After providing transparent data use policy, informed consent provides the user the ability to control the feed (or lack thereof) of their information. Users may opt-in or opt-out (and be afforded the opportunity at any time) of any data collection activities. Tavani (2007) argues that this would enable "informed choices (that is, they would have limited control)" (p. 17) with regard to how the data is used.

Limitations of the Study
The work on this version of the big data model is based on the 6V model presented by Andreu-Perez, Poon, Merrifield, Wong, and Yang (2015). While this is a popular big data model, the authors of the paper note that it is not the only big data model in existence.

Another limitation of this study is the ever-expanding number of models of big data. The authors chose to expand the 6V model as this model is present in a number of works and citations. There are some big data models that are larger with up to 10 Vs (Firican, 2017).

Suggestions for Future Directions
The big data model will undoubtedly expand to encompass more Vs. As this model expands, one suggested area for future research is to examine the fit of privacy within the expanded model. The fit of privacy should be re-evaluated as subsequent Vs are introduced into the model.

CONCLUSION

In this paper, the authors made the case for inclusion of priVacy as a component of the model that describes big data. Privacy is an often-overlooked aspect of big data. The engineering and scientific priorities of big data appear to be in processing, resource requirements, analytics, and to some lesser extent, usability. As such, the rush to provide, store and capitalize on big data has resulted in a negligence of privacy

Within the privacy aspect of big data, the authors of this paper made the case for the following components of data privacy: security, policy, and consent. Security/Cyber-security in big data guards against the unauthorized access of personal information. Policy, with the proposed privacy construct, entails providing the customer with clear, transparent, and detailed information on how the data-collecting organization will use the customer's personal information prior to disclosure of said information. Finally, informed consent is providing the customer (i.e., data producer) the facility to opt-in or opt-out of any data collection activities. The opt-in and opt-out options are not to be understood as a one-time selection, but as an always available privacy option for the customer.

REFERENCES

Allen, M. (2018). Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates. *Propublica*. Retrieved from https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates

Andreu-Perez, J., Poon, C. C., Merrifield, R. D., Wong, S. T., & Yang, G. Z. (2015). Big data for health. *IEEE journal of biomedical and health informatics*, *19*(4), 1193-1208.

Bartholomew, A. (2016). The Smart Grid in Massachusetts: A Proposal for a Consumer Data Privacy Policy. *Boston College Environmental Affairs Law Review*, *43*(1), 79–110.

Big Data (Hadoop) Glossary (2015). 3Vs of Big data. Retrieved from: https://maheshwaranm.blogspot.com/2015/12/glossary.html

Benjelloun, F. Z., & Lahcen, A. A. (2019). Big data security: Challenges, recommendations and solutions. In *Web Services: Concepts, Methodologies, Tools, and Applications* (pp. 25-38). IGI Global.

Brook, J.C. (2001). Components of a privacy policy. *CIPP Guide*. Retrieved from https://www.cippguide.org/2011/08/09/components-of-a-privacy-policy/

Cavoukian, A. (2011). Privacy by design: origins, meaning, and prospects for assuring privacy and trust in the information era. In *Privacy protection measures and technologies in business organizations: aspects and standards* (pp. 170-208). IGI Global.

Davison, C. B. (2015). Addressing the Challenges of Teaching Big Data in Technical Education. *CTE Journal*, *3*(1), 43-50.

EMC Education Services (2015). *Data Science and Big Data Analytics*. Indianapolis, IN: Wiley.

Firican, G. (2017). The 10 Vs of Big data. Retrieved from https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Glancy, D. J. (1979). The Invention of the Right to Privacy. *Arizona Law Review, 21*(1), 1-39.

Handel, T., Schreiber, M., Rothmaler, K., & Ivanova, G. (2018). Data Security and Raw Data Access of Contemporary Mobile Sensor Devices. In *World Congress on Medical Physics and Biomedical Engineering 2018* (pp. 397-400). Springer, Singapore.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues, *Information Systems* (47), 98-115.

Kuhn, M. L. (2018). 147 Million Social Security Numbers for Sale: Developing Data Protection Legislation After Mass Cybersecurity Breaches. *Iowa Law Review*, *104*(1), 417–445.

Laney, D. (2001). 3-d data management: Controlling data volume, velocity and variety, "META Group," Research Note, February 2001.

Mackie, J. (2018). Privacy by design. Retrieved from https://www.termsfeed.com/blog/privacy-design

Murrill, B. J., Liu, E. C., & Thompson, R. M. (2012, February). Smart meter data: Privacy and cybersecurity. Congressional Research Service, Library of Congress.

Tavani, H. T. (2007). Philosophical theories of privacy: Implications for an adequate online privacy policy. *Metaphilosophy*, *38*(1), 1-22.

Verteş-Olteanu, A., & Racolta, R. (2019). The Rise and Fall of Information Privacy. *Juridical Current*, *22*(1).
Warren, S. D., & Brandeis, L. D. (1890). The right to privacy. *Harvard law review*, 193-220.

Zarsky, T. (2019). Privacy and Manipulation in the Digital Age. *Theoretical Inquires in Law, 20*(1), 157.