The Technology Behind Big Data

Geethanjali Jujjavarapu & Udbhav Tiwari

Centre for Internet & Society, India

Introduction

Defining big data is a disputed area in the field of computer science[1], there is some consensus on a basic structure to its definition[2]. Big data is data that is collected in the form of datasets that has three main criteria: size, variety & velocity, all of which operate at an immense scale[3]. It is 'big' in size, often running into petabytes of information, has vast variety within its components, and is created, captured and analysed at an incredibly rapid velocity. All of this also makes big data difficult to handle using traditional technological tools and techniques.

This paper will attempt to perform a high-level literature review of the most commonly used technological tools and processes in the big data life cycle. The big data life cycle is a conceptual construct that can be used to study the various stages that typically occur in collecting, storing and analysing big data, along with the principles that can govern these processes. The big data life cycle consists of four components, which will also be the key structural points of the paper, namely: Data Acquisition, Data Awareness, Data Analytics & Data Governance.[4] The paper will focus on the aspects that the author believes are relevant for analysing the technological impact of big data on both technology itself and society at large.

Scope: The scope of the paper is to study the technology used in big data using the "Life Cycle of Big Data" as model structure to categorise & study the vast range of technologies that are involved in big data. However, the paper will be limited to the study of technology related directly to the big data life cycle. It shall specifically exclude the use/utilisation of big data from its scope since big data is most often being fed into other, unrelated technologies for consumption leading to rather limitless possibilities.

---

[1] EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012)

[2] Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)

[3] Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research 52(1), 11–19 (2010) [4] Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013)

Goal: Goal of the paper is twofold: a.) to use the available literature on the technological aspects of big data, to perform a brief overview of the technology in the field and b.) to frame the relevant research questions for studying the technology of big data and its possible impact on society.

Data Acquisition

Acquiring big data has two main sub components to it, the first being sensing the existence of the data' itself and the second, the stage of collecting and storing this data. Both of these subcomponents are incredibly diverse fields, with lots of rapid change occurring in the technology utilised to carry out these tasks. The section will provide a brief overview of the subcomponents and then discuss the technology used to fulfil the tasks.

Data Sensing

Data does not exist in a vacuum and is always created as a part of a larger process, especially in the aspect of modern technology. Therefore, the source of the data itself plays a vital role in determining how it can be captured and analysed in the larger scheme of things. Entities constantly emit information into the environment that can be utilised for the purposes of big data, leading to two main kinds of data: data that is "born digital" or "born analogue."[4]

Born Digital Data

Information that is "born digital," is created, by a user or by a digital system, specifically for use by a computer or data- processing system. This is a vast range of information and newer fields are being added to this category on a daily basis. It includes, as a short, indicative list: email and text messaging, any form of digital input, including keyboards, mouse interactions and touch screens, GPS location data, data from daily home appliances (Internet of Things), etc. All of this data can be tracked and tagged to users as well as be aggregated to form a larger picture, massively increasing the scope of what may constitute the 'data' in big data.

Some indicative uses of how such born digital data is catalogued by technological solutions on the user side, prior to being sent for collection/storage are:

[4] Big Data and Privacy: A Technological Perspective - President's Council of Advisors on Science and Technology (May 2014)

a.) Cookies - There are small, often just text, files that are left on user devices by websites in order to that visit, task or action (for example, logging into an email account) with a subsequent event.[5] (for example, revisiting the website)

b.) Website Analytics[6] - Various services, such as Google Analytics, Piwik, etc., can use JavaScript and other web development languages to record a very detailed, intimate track of a user's actions on a website, including how long a user hovers above a link, the time spent on the website/application and in some cases, even the time spent specific aspects of the page.

c.) GPS[7] - With the almost pervasive usage of smartphones with basic location capabilities, GPS sensors on these devices are used to provide regular, minute driven updates to applications, operating systems and even third parties about the user's location. Modern variations such as A-GPS can be used to provide basic positioning information even without satellite coverage, vastly expanding the indoor capabilities of location collection.

All of these instances of sensing born digital data are common terms, used in daily parlance by billions of people from all over the world, which is a symbolic of just how deeply they have pervaded into our daily lifestyle. Apart from privacy & security concerns this in turn also leads to an exponential increase in the data available to collect for any interested party.

Sensor Data

Information is said to be "analogue" when it contains characteristics of the physical world, such as images, video, heartbeats, etc. Such information becomes electronic when processed by a "sensor," a device that can record physical phenomena and convert it into digital information. Some examples to better illustrate information that is born analogue but collected via digital means are:

---

[5] Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." MIS quarterly 36.4 (2012): 1165-1188.

[6] Chandramouli, Badrish, Jonathan Goldstein, and Songyun Duan. "Temporal analytics on big data for web advertising." 2012 IEEE 28th international conference on data engineering. IEEE, 2012.

[7] Laurila, Juha K., et al. "The mobile data challenge: Big data for mobile computing research." Pervasive Computing. No. EPFL-CONF-192489. 2012.

a.) Voice and/or video content on devices - Apart from phone calls and other forms communication, video and voice based interactions have started to regularly be captured to provide enhanced services. These include Google Now[8], Cortana[9] and other digital assistants as well as voice guided navigation systems in cars, etc.

b.) Personal health data such as heartbeats, blood pressure, respiration, velocity, etc. - This personal, potentially very powerful information is collected by dedicated sensors on devices such as Fitbit[10], Mi Band[11], etc. as well as by increasingly sophisticated smartphone applications such as Google Fit that can do so without any special device.

c.) Camera on Home Appliances - Cameras and sensors on devices such as video game consoles (Kinect[12] being a relevant example) can record detailed human interactions, which can be mined for vast amounts of information apart from carrying out the basic interactions with the devices itself.

While not as vast a category as born digital data, the increasingly lower costs of technology and ubiquitous usage of digital, networked devices is leading to information that was traditionally analogue in nature to be captured for use at a rapidly increasing rate.

Data Collection & Storage

Traditional data was normally processed using the Extract, Transform, Load (ETL) methodology, which was used to collect the data from outside sources, modify the data to fit needs, and then upload the data into the data storage system for future use.[13] Technology such

---

[8] Lazer, David, et al. "The parable of Google flu: traps in big data analysis." *Science* 343.6176 (2014): 12031205.

[9] *ibid*

[10] Banaee, Hadi, Mobyen Uddin Ahmed, and Amy Loutfi. "Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges." *Sensors* 13.12 (2013): 17472-17500.

[11] *ibid*

[12] Chung, Eric S., John D. Davis, and Jaewon Lee. "Linqits: Big data on little clients." *ACM SIGARCH Computer Architecture News*. Vol. 41. No. 3. ACM, 2013.

[13] Kornelson, Kevin Paul, et al. "Method and system for developing extract transform load systems for data warehouses." U.S. Patent No. 7,139,779. 21 Nov. 2006.

as spreadsheets, RDBMS databases, Structured Query Languages (SQL), etc. were all initially used to carry out these tasks, more often than not manually. [14]

However, for big data, the methodology traditionally followed is both inefficient and insufficient to meet the demands of modern use. Therefore, the Magnetic, Agile, Deep (MAD) process is used to collect and store data[1516]. The needs and benefits of such a system are: attracting all the data sources regardless of their quality (magnetic), logical and physical contents of storage systems adapting to the rapid data evolution in big data (agile) and complex algorithmic statistical analysis required of big data on a very short notice[17]. (deep)

The technology used to perform data storage using the MAD process requires vast amount of processing power, which is very difficult to create in a single, physical space/unit for nonstate or research entities, who cannot afford supercomputers. Therefore, most solutions used in big data rely on two major components to store data: distributed systems and Massive Parallel Processing [18] (MPP) that run on non-relational (in-memory) database systems. Database performance and reliability is traditionally gauged using pure performance metrics (FLOPS per second, etc.) as well as the Atomicity, consistency, isolation, durability (ACID) criteria.[19] The most commonly used database systems for big data applications are given below. The specific operational qualities and performance of each of these databases is beyond the scope of this review but the common criteria that makes them well suited for big data storage have been delineated below.

---

[14] Henry, Scott, et al. "Engineering trade study: extract, transform, load tools for data migration." *2005 IEEE Design Symposium, Systems and Information Engineering*. IEEE, 2005.

[15] Cohen, Jeffrey, et al. "MAD skills: new analysis practices for big data." *Proceedings of the VLDB Endowment*

[16] .2 (2009): 1481-1492.

[17] Elgendy, Nada, and Ahmed Elragal. "Big data analytics: a literature review paper." *Industrial Conference on Data Mining*. Springer International Publishing, 2014.

[18] Wu, Xindong, et al. "Data mining with big data." *IEEE transactions on knowledge and data engineering* 26.1 (2014): 97-107.

[19] Supra Note 17

Non-relational databases

Databases traditionally used to be structured entities that operated solely on the ability to correlate information stored in them using explicitly defined relationships. Even prior to the advent of big data, this outlook was turning out to be a limiting factor in how large amounts of stored information could be leveraged, this led to the evolution of non relational database systems. Before going into them in detail, a basic primer on their data transfer protocols will be helpful in understanding their operation.

A protocol is a model that structures instructions in a particular manner so that it can be reproduced from one system to another[2021]. The protocols which govern technology in the case of big data have gone through many stages of evolution, starting off with simple HTML based systems[22], which then evolved to XML driven SOAP systems[23], which led to JavaScript Object Notation, or JSON[24], the currently used form for in most big database systems. JSON is an open format used to transfer data objects, using human-readable text and is the basis for most of the commonly used non-relational database management systems. Examples of Non-relational databases also known as NoSQL databases, include MongoDB[25], Couchbase[26], etc. They were developed for both managing as well as storing unstructured data. They aim for scaling, flexibility, and simplified development. Such databases rather focus on the high-performance scalable data storage, and allow tasks to be written in the application layer instead of databases specific languages, allowing for greater interoperability.[27]

---

[20] Hu, Han, et al. "Toward scalable systems for big data analytics: A technology tutorial." *IEEE Access* 2 (2014):

[21] -687.

[22] Kurt Cagle, Understanding the Big Data Lifecycle - LinkedIn Pulse (2015)

[23] Coyle, Frank P. *XML, Web services, and the data revolution*. Addison-Wesley Longman Publishing Co., Inc., 2002.

[24] Pautasso, Cesare, Olaf Zimmermann, and Frank Leymann. "Restful web services vs. big'web services: making the right architectural decision." *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008.

[25] Banker, Kyle. *MongoDB in action*. Manning Publications Co., 2011

[26] McCreary, Dan, and Ann Kelly. "Making sense of NoSQL." *Shelter Island: Manning* (2014): 19-20.

[27] *ibid*

In-Memory Databases

In order to overcome performance limitation of traditional database systems, some modern databases now use in-memory databases. These systems manage the data in the RAM memory of the server, thus eliminating storage disk input/output. This allows for almost realtime responses from the database, in comparisons to minutes or hours required on traditional database systems. This improvement in the performance is so massive that, entirely new applications are being developed for using IMDB systems.[28] These IMDB systems are also being used for advanced analytics on big data, especially to increase the access speed to data and increase the scoring rate of analytic models for analysis.[29] Examples of IMDB include VoltDB[30], NuoDB[31], SolidDB[32] and Apache Spark[33].

Hybrid Systems

These are the two major systems used to store data prior to it being processed or analysed in a big data application. However, the divide between data storage and data management is a slim one and most database systems also contain various unique attributes that cater them to specific kinds of analysis. (as can be seen from the IMDB example above) One example of a very commonly used Hybrid system that deals with storage as well as awareness of the data is Apache Hadoop[33], which is detailed below.

Apache Hadoop

Hadoop consists of two main components: the HDFS for the big data storage, and MapReduce for big data analytics, each of which will be detailed in their respective section.

---

[28] Zhang, Hao, et al. "In-memory big data management and processing: A survey." *IEEE Transactions on Knowledge and Data Engineering* 27.7 (2015): 1920-1948.

[29] *ibid*

[30] *ibid*

[31] Supra Note 20

[32] Ballard, Chuck, et al. *IBM solidDB: Delivering Data with Extreme Speed*. IBM Redbooks, 2011.

[33] Shanahan, James G., and Laing Dai. "Large scale distributed data science using apache spark." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. [33] Shvachko, Konstantin, et al. "The hadoop distributed file system." *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*. IEEE, 2010.

1. The HDFS[34][35] storage function in Hadoop provides a reliable distributed file system, stored across multiple systems for processing & redundancy reasons. The file system is optimized for large files, as single files are split into blocks and spread across systems known as cluster nodes.[36] Additionally, the data is protected among the nodes by a replication mechanism, which ensures availability even if any node fails. Further, there are two types of nodes: Data Nodes and Name Nodes.[37] Data is stored in the form of file blocks across the multiple Data Nodes while the Name Node acts as an intermediary between the client and the Data Node, where it directs the requesting client to the particular Data Node which contains the requested data.

This operating structure for storing data also has various variations within Hadoop such as HBase for key/value pair type queries (a NoSQL based system), Hive for relational type queries, etc. Hadoop's redundancy, speed, ability to run on commodity hardware, industry support and rapid pace of development have led to it being almost co-equivalently associated with big data.[38]

Data Awareness

Data Awareness, in the context of big data, is the task of creating a scheme of relationships within a set of data, to allow different users of the data to determine a fluid yet valid context and utilise it for their desired tasks.[39] It is a relatively new field, in which most of the work is currently being done on semantic structures to allow data to gain context in an interoperable format, in contrast to the current system where data is given context using unique, model specific constructs.[40] (such as XML Schemes, etc.)

---

[34] Borthakur, Dhruba. "The hadoop distributed file system: Architecture and design." *Hadoop Project Website*

[35] .2007 (2007): 21.

[36] *ibid*

[37] *ibid*

[38] Zikopoulos, Paul, and Chris Eaton. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.

[39] Bizer, Christian, et al. "The meaningful use of big data: four perspectives--four challenges." *ACM SIGMOD Record* 40.4 (2012): 56-60.

[40] Kaisler, Stephen, et al. "Big data: issues and challenges moving forward." *System Sciences (HICSS), 2013 46th Hawaii International Conference on*. IEEE, 2013.

Some of the original work on this field was carried out in the form of utilising the Resource Description Framework (RDF), which was built primarily to allow describing of data in a portable manner, especially being agnostic towards platforms and systems for Semantic Web at the W3C. SPARQL is the language used to implement RDF based designs but both largely remain underutilised in both the public domain as well as big data. Authors such as Kurt Cagle[41] and Bob DuCharme[42] predict its explosion in the next couple of years. Companies have also started realising the value of interoperable context, with Oracle Spatial[43] and IBM's DB2[44] already including RDF and SPARQL support in the past 3 years.

While underutilised, the rapid developments taking place in the field will make the impact that data awareness may have on big data as big as Hadoop and maybe even SQL. Some aspects of it are already beginning to be used in Artificial Intelligence, Natural Language Processing, etc. with tremendous scope for development.[45]

Data Processing & Analytics

Data Processing largely has three primary goals: a. determines if the data collected is internally consistent; b. make the data meaningful to other systems or users using either metaphors or analogy they can understand; and (what many consider most importantly) provide predictions about future events and behaviours based upon past data and trends.[46]

Being a very vast field with rapidly changing technologies governing its operation, this section will largely concentrate on the most commonly used technologies in data analytics.

---

[41] Supra Note 21

[42] DuCharme, Bob. "What Do RDF and SPARQL bring to Big Data Projects?." *Big Data* 1.1 (2013): 38-41.

[43] Zhong, Yunqin, et al. "Towards parallel spatial query processing for big spatial data." *Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International*. IEEE, 2012.

[44] Ma, Li, et al. "Effective and efficient semantic web data management over DB2." *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008.

[45] Lohr, Steve. "The age of big data." *New York Times* 11 (2012).

[46] Pääkkönen, Pekka, and Daniel Pakkala. "Reference architecture and classification of technologies, products and services for big data systems." *Big Data Research* 2.4 (2015): 166-186.

Data analytics requires four primary conditions to be met in order to carry out effective processing: fast, data loading, fast query processing, efficient utilisation of storage and adaptivity to dynamic workload patterns. The analytical model most commonly associated with meeting this criteria and with big data in general is MapReduce, detailed below. There are other, more niche models and algorithms (such as Project Voldemort[47] used by LinkedIn), which are used in big data but they are beyond the scope of the review, and more information about them can be read at article linked in the previous citation. (Reference architecture and classification of technologies, products and services for big data system)

MapReduce

MapReduce is a generic parallel programming concept, derived from the "Map" and "Reduce" of functional programming languages, which makes it particularly suited for big data operations. It is at the core of Hadoop[48], and performs the data processing and analytics functions in other big data systems as well.[49] The fundamental premise of MapReduce is scaling out rather than scaling up, i.e., (adding more numerical resources, rather than increasing the power of a single system)[50]

MapReduce operates by breaking a task down into steps and executing the steps in parallel, across many systems. This comes with two advantages, a reduction in the time needed to finish the task and also a decrease in the amount of resources one has to expend to perform the task, in both power and energy. This model makes it ideally suited for the large data sets and quick response times required of big data operations generally.

The first step of a MapReduce job is to correlate the input values to a set of keys/value pairs as output. The "Map" function then partitions the processing tasks into smaller tasks, and assigns them to the appropriate key/value pairs.[51] This allows unstructured data, such as plain text, to

---

[47] Sumbaly, Roshan, et al. "Serving large-scale batch computed data with project voldemort." *Proceedings of the 10th USENIX conference on File and Storage Technologies*. USENIX Association, 2012.

[48] Bar-Sinai, Michael. "Big Data Technology Literature Review." *arXiv preprint arXiv:1506.08978* (2015).

[49] ibid

[50] Condie, Tyson, et al. "MapReduce Online." *Nsdi*. Vol. 10. No. 4. 2010.

[51] Supra Note 47

be mapped to a structured key/value pair. As an example, the key could be the punctuation in a sentence and the value of the pair could be the number of occurrences of the punctuation overall. This output of the Map function is then passed on "Reduce" function.[52] Reduce then collects and combines this output, using identical key/value pairs, to provide the final result of the task.[53] These steps are carried using the Job Tracker & Task Tracker in Hadoop but different systems have different methodologies to carry out similar tasks.

Data Governance

Data Governance is the act of managing raw big data as well as the processed information that arises from big data in order to meet legal, regulatory and business imposed requirements. While there is no standardized format for data governance, there have been increasing call with various sectors (especially healthcare) to create such a format to ensure reliable, secure and consistent big data utilisation across the board. The following tactics and techniques have been utilised or suggested for data governance, with varying degrees of success:

1. **Zero-knowledge systems**: This technological proposal maintains secrecy with respect to the low-level data while allowing encrypted data to be examined for certain higherlevel abstractions.[54] For the system to be zero-knowledge, the client's system will have to encrypt the data and send it to the storage provider. Due to this, the provider stores the data in the encrypted format and cannot decipher the same unless he/she is in possession of the key which will decrypt the data into plaintext. This allows the individual to store his data with a storage provider while also maintaining anonymity of the details contained in such information. However, these are currently just beginning to be used in simple situations. As of now, they are not expandable to unstructured and complex cases and have to be developed marginally before they can be used for research and data mining purposes.

[52] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: a flexible data processing tool." *Communications of the ACM* 53.1 (2010): 72-77.

[53] ibid

[54] Big Data and Privacy: A Technological Perspective, White House, https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy__may_2014

2. **Homomorphic encryption**: Homomorphic encryption is a privacy preserving technique which performs searches and other computations over data that is encrypted while also protecting the individual's privacy.[55] This technique has however been considered to be impractical and is deemed to be an unlikely policy alternative for near future purposes in the context of preserving privacy in the age of big data.[56]
3. **Multi-party computation**: In this technique, computation is done on encrypted distributed data stores.[57] This mechanism is closely related to homomorphic encryption where individual data is kept private using encryption algorithms called "collusion-robust" while the same is used to calculate statistics.[58] The parties involved are aware of some private data and each of them use a protocol which produces results based on the information they are aware of and the information they are not aware of, without revealing the data they are not already aware of.[59] Multi-party computations thus help in generating useful data for statistical and research purposes without compromising the privacy of the individuals.
4. **Differential Privacy**: Although this technological development is related to encryption, it follows a different technique. Differential privacy aims at maximizing the precision of computations and database queries while reducing the identifiability of the data owners who have records in the database, usually through obfuscation of query results.[60] This is widely applied today in the existence of big data in order to ensure preservation of privacy while trying to reap the benefits of large scale data collection.[61]

---

[55] Tene, Omer, and Jules Polonetsky. "Big data for all: Privacy and user control in the age of analytics." *Nw. J. Tech. & Intell. Prop.* 11 (2012): xxvii.

[56] Big Data and Privacy: A Technological Perspective, White House, https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy__may_2014

[57] Privacy by design in big data, ENISA

[58] Big Data and Privacy: A Technological Perspective, White House, https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy__may_2014

[59] Id

[60] Id

[61] Tene, Omer, and Jules Polonetsky. "Privacy in the age of big data: a time for big decisions." *Stanford Law Review Online* 64 (2012): 63.

5. **Searchable encryption**: Through this mechanism, the data subject can make certain data searchable while minimizing exposure and maximizing privacy.[62] The data owner can make his information available through search engines by providing the data in an encrypted format but by adding tags consisting of certain keywords which can be deciphered by the search engine. This encrypted data shows up in the search results when searched with these particular keywords but can only be read when the person is in possession of the key which is required for decrypting the information.

   This technique of encryption provides maximum security to the individual's data and preserves privacy to the greatest possible extent.
6. **K-anonymity**: The property of k-anonymity is being applied in the present day in order to preserve privacy and avoid re-identification.[63] A certain data set is said to possess the property of k-anonymity if individual specific data can be released and used for various purposes without re-identification. The analysis of the data should be carried out without attributing the data to the individual to whom it belongs and should give scientific guarantees for the same.
7. **Identity Management Systems**: These systems enable the individuals to establish and safeguard their identities, explain those identities with the help of attributes, follow the activity of their identities and also delete their identities if they wish to.[64] It uses cryptographic schemes and protocols to make anonymous or pseudonymous the identities and credentials of the individuals before analysing the data.
8. **Privacy Preserving Data Publishing**: This is a method in which the analysts are provided with the individual's personal information with the ability to decipher particular information from the database while preventing the inference of certain other information which might lead to a breach of privacy.[65] Data which is essential for the

---

[62] Lane, Julia, et al., eds. *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge University Press, 2014.

[63] Crawford, Kate, and Jason Schultz. "Big data and due process: Toward a framework to redress predictive privacy harms." *BCL Rev.* 55 (2014): 93.

[64] http://homes.esat.kuleuven.be/~sguerses/papers/DanezisGuersesSurveillancePets2010.pdf

[65] Seda Gurses and George Danezis, A critical review of 10 years of privacy technology, August 12th 2010, http://homes.esat.kuleuven.be/~sguerses/papers/DanezisGuersesSurveillancePets2010.pdf

analysis will be provided for processing while sensitive data will not be disclosed. This tool primarily focuses on microdata.

9. **Privacy Preserving Data Mining**: This mechanism uses perturbation methods and randomization along with cryptography in order to permit data mining on a filtered version of the data which does not contain any form of sensitive information. PPDM focuses on data mining results unlike PPDP.[66]

Conclusion

Studying the technology surrounding big data has led to two major observations: the rapid pace of development in the industry and the stark lack of industry standards or government regulations directed towards big data technologies. These observations have been the primary motivating factor for framing further research in the field. Understanding how to deal with big data technologically, rather than just the potential regulation of possible harms after the technological processes have been performed might be critical for the human rights dialogue as these processes become even more extensive, opaque and technologically complicated.

[66] Id