THE DARK SIDE OF BIG DATA

...the beginning of the end of privacy?

Ankita Kundra

Critical Writing- Memes and Internet (CW-002-0X)

Research Paper

June 5, 2018

Preceptor : Anunaya Rajhans

Ankita Kundra

Research Paper – The Dark Side of Big Data

5 June 2018

Let it sink in for a moment -- Google is aware of our browsing habits, Amazon monitors our shopping preferences, Linkedin knows our educational background and work experience, Facebook knows our social relationships while mobile operators know our precise locations and calling patterns. Whenever we use our smartphones, chat with friends and acquaintances on social networking sites or shop online – in each of these every day activities, we end up producing a trail of data called digital footprints. All this complex datasets coming from different sources, also known as "Big Data", is then analyzed by state as well as big corporations for various reasons ranging from providing customized online ads to manipulation of user behavior to something potentially much more sinister. As the data storing capabilities and data processing technologies become much more sophisticated, the security breaches and privacy violations in the digital world are likely to lead to consequences much beyond relative trifles like showing customized ads to consumers. One can already see the signs of imminent danger of big data in the form of Cambridge Analytica[1] fiasco in US and Aadhar leaks[2] in India which have bring to forth important questions like- Is the concept of privacy dead in the digital age? Is it like a currency that can be traded for innovation and security? Are the existing rules capable enough to protect us against the security breaches that might happen due to transfer of data to a third party? This paper shall explore some of these questions and make a claim **that the existing legal**

[1] Personal data of millions of Americans was shared by Faceboook to political consultancy firm Cambridge Analytica which was apparently used by it to influence US elections https://www.cnbc.com/2018/03/21/facebook-cambridge-analytica-scandal-everything-you-need-to-know.html

[2] A report in The Tribune has claimed that an "agent" available on WhatsApp facilitated access with a login ID and password to the particulars of any Aadhaar number. http://www.bbc.com/news/world-asia-india-42575443

framework may not be able to protect us against the potential tyranny of big data, thereby making a case for new set of rules and regulations. In doing so, the paper shall discuss the privacy threats posed by big data, predictive analytics enabled by it and the problems inherent with deification of data; and also how the existing laws do not take into consideration any of these threats. In the process, we shall also discuss the relevance of privacy in today's age and why the cause of privacy should be taken up by the state even if the public is seemingly indifferent to it. This paper's focus on the dark side of big data should in no way be mistaken as an oversight or indifference to the positive potential of big data as an enabler in confronting some of the world's most critical problems. While fully acknowledging the positive impacts of big data, the scope of the paper is limited to **making readers aware about the potential havoc that the big data can cause and the inability of existing laws to mitigate these risks**. It is to be noted that the paper shall not go into the realm of suggesting the possible rules and legalities that can help to mitigate the negative impacts of big data.

Big data refers to voluminous and complex datasets that are difficult to analyze, store, transfer or visualize using traditional data-processing software. The concept of big data gained prominence in the early 2000s when Gartner's Research Vice President Doug Laney articulated the now-mainstream definition of big data as the three Vs- Volume, Velocity and Variety (What Is Big Data n.d.). In this, volume refers to the amount of data, variety refers to the number of types of data and velocity refers to the speed of data processing. Thus, the term Big Data is defined as "a new generation of technologies and architectures, designed to economically separate value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery and analysis" (Manyika J 2011). In simpler terms-- with the penetration of internet and digital technologies, the volume of information produced grew so large that old

data-processing system could no longer store this information in its memory, leave alone process it. This led to innovation in data processing technologies like Google's MapReduce and Yahoo's Hadoop which made analyzing and interpreting such vast amounts of data possible. The internet companies which collected this huge amount of data started using these data processing technologies to analyze and interpret the data primarily for financial incentives. It is to be noted that in early 2000s, only a quarter of world's stored information was in digital format while today, after two decades, over ninety eight percent of the stored information is in digital format (Viktor Mayer Schonberger 2012). As the data storage costs continue to plummet and analytical tools become even more powerful, the size and scale of data collection, storage, monitoring and analysis would increase even more. The insights one can gain from this big data can potentially change markets, organizations, relationship between citizens and governments and even citizens among themselves.

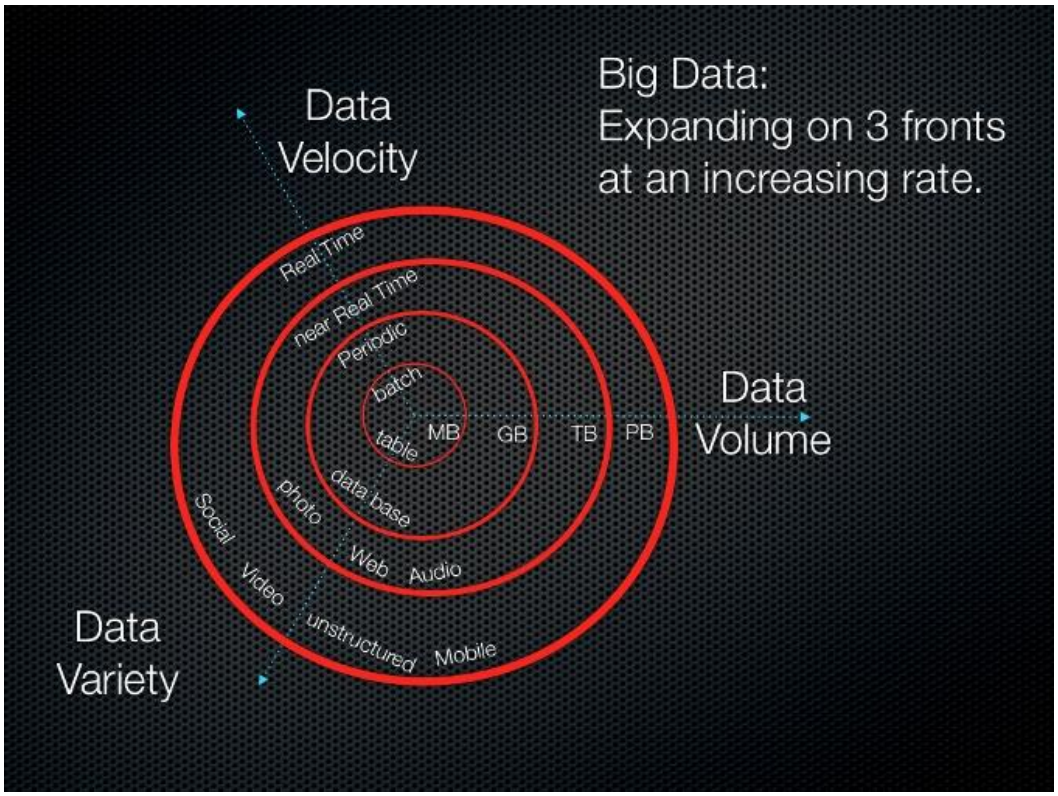

Figure 1 : Three Vs of Big Data
Source : *https://whatis.techtarget.com/definition/3Vs*

Why Big Data should worry us?

"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it"

-Dan Ariely, Behavorial Economics (2013)

The above assertion by Ariely, presumably, has little relevance today in 2018. It is arguable that Big Data is an overused term often used by people having little understanding of it. It may also be true that many people claim to have had a "big data" experience without actually having it. But the constant growth of big data companies, Data Science seemingly maintaining its position as the "sexiest job of 21st century"[3] (Davenport 2012) and constant improvement in data processing technologies somewhat contradicts with the assertion that big data is mostly a hype like teenage sex.

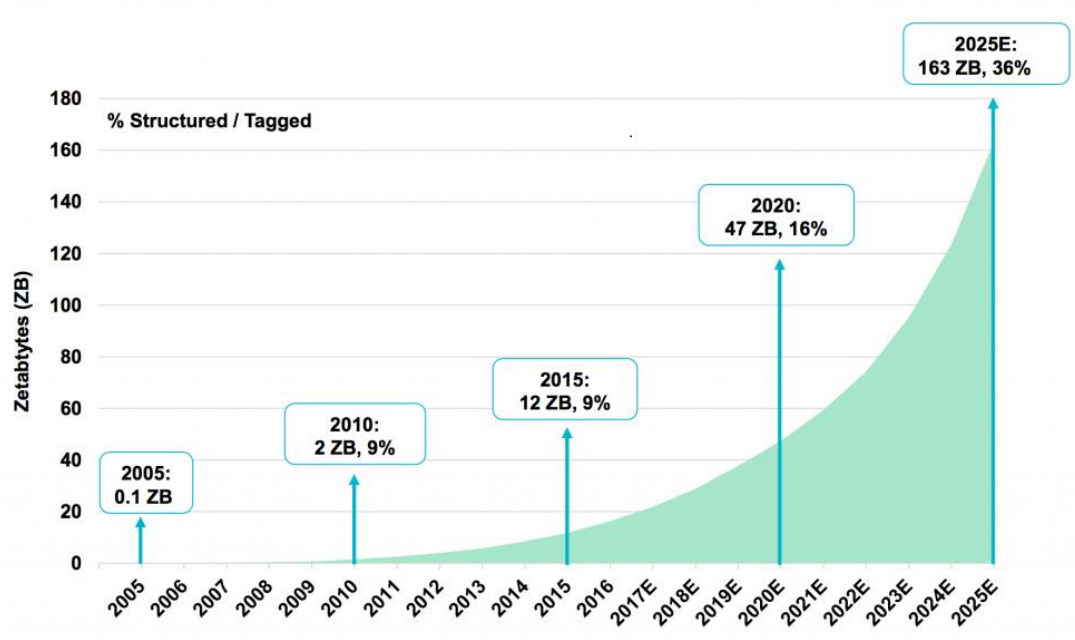

Figure 2 : Data Volume Growth (KP Internet Trends 2017)
Source : https://www.fourquadrant.com/go-to-market-big-data-analytics-research/

[3] An article in Harvard Business Review declared Data Science as the sexiest job of the 21st century https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Today, the data extracted from internet is being used by organizations across all sectors including health, education, FMCGs etc for reasons varying from better pricing strategy, more accurate demand estimation, targeted online advertisement, better product strategy to much more. It is not only the private sector but also the state which is using the big data for various reasons like targeted social welfare schemes, bolstering security of citizens, deciding on parole of prisoners etc. It is important to note that an organization is often required to store all data in one location in order to facilitate this analysis. A higher data volume increases the probability of the datasets containing valuable and sensitive information, making it a more appealing target for hackers. Various cases of data breaches like Aadhar leaks and Facebook leaks prove that the hacking of big data or its unauthorized use is not so uncommon today. The claims that the fear around big data is exaggerated and similar to our initial anxieties about electricity or television are questionable. The dangers of not regulating big data is not limited to relatively trivial issues like getting customized online ads or protecting some puritan notion of privacy. The stakes involved are much higher and one just needs to go through some of the previous recorded cases of devious use of data to understand that.

History is a replete with instances of data-misuse to aid and abet some of the most heinous crimes on mankind. In 1943, US Census Bureau handed over block addresses of Japanese Americans to US government to facilitate their internment after 1941 Pearl Habor attack by Japan on US (Minkel 2007). The Netherlands' comprehensive civil records were used by the invading Nazis to round up Jews during World War 2[4]. Closer home in Gujarat, during the Godhra riots of 2002, there was systematic misuse of voter cards and licenses to identify and target Muslim localities, businesses and vehicles (Misuse of voters list in Gujarat riots alleged

[4] https://www.facinghistory.org/resource-library/text/anti-jewish-measures-netherlands-and-belgium-between-1940-and-1944

2002). If data had the potential to create such havoc in the days of small data, it is terrifying to imagine what it can cause in today's times of big data when it is possible to know the real time location of a person. And while targeted killings are indeed heinous, they are also more visible to public eye which inherently gives a space for a system of checks and balances. Big data can also be misused for other "sophisticated" purposes in a way that small data cannot - like manipulation of human behavior-- that may not even grab public eye balls but have significant repercussions on society.

The potential dangers of big data are very real. The next sections of the paper would discuss the specific threats posed by big data by enabling privacy breaches, predictive analytics and deification of data itself.

Big Data and Privacy

"People have really gotten comfortable not only sharing more information & different kinds but more openly and with more people. That social norm is just something that's evolved over time"

-Mark Zuckerberg, Facebook founder (2010)

"You have zero privacy anyway, get over it"

- Scott McNealy, Sun Microsystems founder

Many commentators especially the owners of the internet companies including Zuckerberg and McNealy argue that the concept of privacy is dead in the digital age. For them, it is a foregone conclusion that if one is on internet, their data would be collected, stored, analyzed and distributed. Privacy is seen as a currency which could be traded in favor of other social values like security, efficiency, technological innovation and free speech etc. The argument goes that in today's world when people are willingly sharing their intimate pictures, banal opinions/ideas,

details of everyday mundane activities, as well as employment and relationship statuses on different social netwoking sites, then what is the fuss about privacy? What is the big deal when the same data is stored and analyzed to enhance business capabilities, organizational efficiencies and even national security? The fact that the recent Cambridge Analytica scandal could not make a dent in Facebook's quaterly profits is often used to bolster the argument that privacy is not desirable to people anymore (Facebook posts record growth in first quarter despite privacy scandal 2018). "But I have nothing to hide" is another argument one commonly gets to hear during discussions on big data's invasion of privacy. This part of the paper shall focus on some of the problems with the above line of argument and assert that the big data does not only increase the risk to privacy but also changes the character of this risk. But before we proceed any further, it is pertinent to discuss what exactly entails privacy.

Privacy is a difficult term to define with its conception evolving over time, cultures and contexts. In older times, privacy was defined as seclusion, safety and security primarily in physical terms. But with the coming of digital technology and availability of most intimate of information outside the realm of physical space, this conception of privacy was found to be too narrow and exclusive. Today, sensitive information like medical records or criminal history is not stored with a person, in his/her property or a physical institutional structure but in a central data base kept much away from the person. To encompass this changing reality, the concept of privacy has been broadened in contemporary discourse to include "freedom of thought, control over one's body, solitude in one's home, control over personal information, freedom from surveillance, protection of one's reputation, and protection from searches and interrogations among other things" (Pozen, 2016). Privacy is important to ensure autonomy of an individual. Most of our learning and individuality comes from our interactions with others in a private

sphere when we assume that no one is observing us. With the knowledge that we are being constantly observed, the process of complete self development may not take place. Julie E. Cohen, in her article *What Privacy is for* in the Harvard Law Review rightly describes privacy as something that "protects our subjectivity from the pervasive efforts of commercial and government actors to render individual and communities fixed, transparent and predictable" (Cohen 2012). Privacy is something that is necessary for us to form an identity that is not dictated by the social conditions that largely influence our thinking, decisions and actions. Another reason why privacy matters is that it helps one to evade from unnecessary explanations and justifications. We do a lot of things in private which when judged by someone afar might seem odd, embarrassing or even worse. More importantly, privacy acts as a limit on the power of state and private corporations. The more someone knows about us, the more power they can have over us. Personal data can be used to affect our reputations, influence our decisions, shape our behaviors and at its worst, even kill us. Thus, privacy is just as relevant in big data age as it was in any other time.

The rampant collection, storage and distribution of personal data enabled by "big data" can have an adverse impact on the cherished social value of privacy. Many commentators insist that the positive externalities of big data are so huge that an "irrelevant social value like privacy" should be unthinkingly traded off in exchange, as has already been discussed in first part of this section. Edward Snowden's famous statement- "Arguing that you don't care about the right to privacy because you have nothing to hide is no different than saying you don't care about free speech because you have nothing to say" –is extremely relevant here. People from specific ethnicities-- identifiable by the government records of their personal data-- targeted during World War 2 also had "nothing to hide". When Internet users are aware of large-scale data collection, data storage

and surveillance, they may self-censor their behavior due to the fear of unexpected consequences, which potentially has a chilling effect on their freedom of speech and expression. While Facebook's net revenue may not have been affected much by the multiple cases of data breaches against it over the years, it has definitely caused a sense of paranoia among users with some believing that the site uses the phone's microphone to listen to their conversations.

Figure 3 : Twitter Screenshots showing misgivings against Facebook
Source : *https://twitter.com/_Jaan__/status/991913131201908736*, *https://twitter.com/DubbleJump/status/989676567130529798*

The perception that facebook is spying via the phone's microphone has become so widespread over the years that the the company had to write a blog post back in 2016 strongly denying such rumours (https://newsroom.fb.com/news/h/facebook-does-not-use-your-phones-microphone-for-ads-or-news-feed-stories/) and again in April 2018, the company founder Mark Zuckerberg had to deny this claim in his testimony in US Congress. This puts a question mark on the claim that people do not care about their privacy any more . More importantly, the very argument that people do not really care about privacy because "despite of all the privacy breaches, they continue to be on internet" or "they share most intimate of details online" is based on flawed premises. First, it is the agency of an individual to share his/her personal details with their intimate group of friends or even to a wider community. This does not imply that the individual has given an informed consent for storing this data, sharing it with a third part user, or for

aggregating the data from multiple sources -- credit card transactions, Facebook activities, Google searches, online purchases etc -- to map his/her complete profile. Secondly, it's absurd to expect people to not carry a credit card or not have an email address or cut themselves out from social networking sites in today's digital age. With the state itself backing digital transactions across the world, it is unfair to then tell people that they might as well opt out of the system of digital payments if they want to avoid having their purchases monitored. "Buyer Beware" ends up putting too much onus on an individual. People do not have to test their food or airlines or automobiles for safety; it is the government that regulates and ensures that these options are safe. The government should similarly protect its citizens from the perils of big data by bringing forth effective regulations and legal framework. Importantly, it needs to ensure that it itself does not breach the citizens' privacy in name of security or targeted social welfare schemes. We shall now see why we need to change the concepts and rules around privacy to meet the demands of big data.

The rules and laws that have been protecting privacy till now may no longer work in the age of big data. Right to be let alone, control of one's personal information, anonymization and right to opt out were some of the concepts used to hitherto protect privacy but big data presents a challenge to each of these concepts of privacy protection. The traditional physical and decisional conception of privacy involving "right to be let alone" made a fundamental distinction between "being observed", which can accompany any act made in public, versus "being identified", a "separate and more intrusive act" (Geer 2017). We consent to be observed constantly; we rarely consent to be identified. Today this distinction has eroded courtesy the rapid advancement of digital technologies and the accompanying rise of the field broadly called data science. Now, we are being continuously sampled and monitored. Thus, in today's world of big data, if we are

being observed, it is a given that we are being identified as well. Thus, the conventional difference between "being identified" and "being observed" no longer exists and this distinction cannot be translated into law. Another important principle that has been used by privacy laws across the world has been to allow the individuals to control whether, how and by whom their personal information may be processed (Viktor Mayer Schonberger 2012). This principle is the reason behind the formulaic system of "notice and consent", followed by several internet companies wherein users are told at the time of collection as to which information is being gathered and for what purpose and whether they are fine with the data being sold to a third party user. But the problem with this approach is that in the world of big data, most of the innovative secondary use of data is not thought about at the time when it is collected. How can citizens give informed consent about something if they have incomplete information about it? The system of "notice and consent" also puts a lot of onus on the individual by putting the liability on him/her to go through the privacy terms and conditions. This system protects data users as well as third party users from any accountability for malicious use of data as long as consent has been taken before the collection of data. Another technical approach to protecting privacy, anonymization, may also not work effectively in many circumstances today. Data anonymization is a type of information sanitization in which personally identifiable information - names, address, credit card number, social security number etc- are removed or encrypted from the datasets so that the people whom the data describe remain anonymous. Many internet companies that maintain huge database of users claim to follow data anonymization before selling the information to a third party in order to ensure privacy of the users. Unfortunately, re-identification science disrupts the privacy policy by enabling in de-anonymization (Ohm 2010). Given enough data, perfect anonymization is impossible to achieve no matter how much one tries. Researchers have shown

that not only personal identifiable data but also people's connections with each other are vulnerable to de-anonymization. The "opt out" principle of protecting privacy wherein a website has the right to collect, store and distribute data unless the consumer chooses to opt out also doesn't work since in a huge database, even opting out leaves a trace. Moreover, many consumers may not understand the company's privacy policy or not go on the privacy tab to opt out of the data collection feature, even if they do not want their data to be collected and stored. The principle of "opt in" where companies can collect, store and share data only if users choose to opt in, on the other hand, would adversely impact innovation, efficiency and positive use of data and thus, is not a good alternative to opt out.

The methods that are usually proposed for protecting privacy of a consumer – inform and consent, anonymization and opt out option – clearly fail in case of big data, thereby making a need for new concepts and rules around privacy.

Predictive analytics enabled by Big Data

Predictive analysis is the use of historical patterns and time-tested organizational strategies to make predictions about specific outcomes in the future. Decisions involved in predictive analytics are often automated and use advanced algorithm techniques like regression analysis[5] to come up with predictions about the future (Hamilton 2017). Humans are programmed to recognize patterns even if they aren't any and machine learning tools make that process of pattern recognition faster. Machine-learning algorithms recognize and analyze patterns in an initial training data set and then look for similar patterns in the new data to make predictions about the future. Thus, if they learn the wrong pattern from the data, the analysis would be flawed as well.

[5] Regression analysis is a set of statistical processes for estimating the relationships among variables. It helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. https://en.wikipedia.org/wiki/Regression_analysis

The old adage "a computer is a device that allows you to make mistakes far faster than paper and pencil" is extremely relevant here. Today, predictive analysis is being used by companies for differential pricing wherein price of an online item shown to a potential consumer is based on the prediction about his/her maximum willingness to pay (Kshetri 2014). The exploitation of personal information of people leads to different people paying differently for the same product.

Predictive analytics is also being used today for predictive policing. The quantum of sentence and the approval of parole for a criminal offence is increasingly being determined by predictive analytics in many states across US[6]. The historical data on the recidivism rates of samples of known criminals is analyzed to determine the factors that are statistically related to recidivism which is then incorporated into a mathematical model that predicts whether a criminal is likely to commit a crime again. The predicted likelihood of committing a crime is then used to decide the quantum of sentence given to him/her. There are increasing talks about using predictive analytics to preempt commission of a criminal offence by predicting which set of people are likely to commit violent crimes and then scrutinizing them and giving them special care etc to preclude a crime from being committed. The problem with this system is that since big data has the ability to predict human actions increasingly accurately, we might get tempted to judge people not only by what they did but also by what we predicted they would do. Propensities begin to matter more than actions. By relying on data- driven solutions to reduce risk to society, we discount the value of individual responsibility and holding people accountable for their actions. Thus, there is an urgent need for expanding our understanding of justice so as to include safeguards for human agency. By guaranteeing human agency, it can be ensured that we are judged on the basis of our actions and not simply on the basis of predicted analysis of big data. Policy on big data should

[6] https://theconversation.com/we-use-big-data-to-sentence-criminals-but-can-the-algorithms-really-tell-us-what-we-need-to-know-77931

clearly spell out that though we might be held responsible for our past actions but definitely not on the basis of statistical predictions of our future actions.

Deification of Data

"In God we trust – all others bring data", this sentiment that we used to commonly see in tech companies is becoming more and more pervasive in everyday life. As more and more aspects of our life becomes datafied, the solutions proposed by policy makers involve collecting more and more data. One must be aware that the quality of underlying data can be poor, biased or miss-analyzed knowingly or unknowingly. Data at times can fail to capture what it wants to quantify or we might mistake correlations for causations. Data set needs to be seen with skepticism and it needs to be realized that data is a tool, not a course of action. It should not be seen as a substitute to judgment. A policy memo on big data should be aware of this shortcoming of data and not treat it as a final say on anything. Perhaps, with increasing use of big data, we might also need data auditors.

Conclusion

The paper began with discussion about what exactly entails big data and if the fear around it is merely a manifestation of human anxieties about the "unknown". We saw how the big data has enabled collection, storage and distribution of data at a scale never seen before in human history. This huge volume of data can indeed help in efficient and rational decision making but it poses a huge threat to citizens' privacy and at its worst, can be used to aid and abet ugly crimes on humanity. This was followed by a discussion on privacy and how it would continue to be relevant irrespective of changing times and contexts. The paper then discussed how big data increases the risk to privacy and also changes the nature of this risk. None of the old concepts of

protecting privacy including "right to be let alone" and "control of personal information" seems to work in the big data age, increasing the urgency for some brainstorming on privacy concepts and laws. Big data has also enabled predictive analytics which is being used today for predictive policing. The biggest problem with predictive policing is that it leads to a system of punishment on the basis of propensities and not actions. Thus, there is a need for expanding our understanding of justice to include safeguards for human agency. Any policy on big data should guarantee that people shall continue to be judged on the basis of personal responsibilities and actual behavior and not some statistical based predictions of their behavior. Big Data can also lead to wrong analysis and predictions and one need to be mindful of the fallibility of data. In this overreliance and deification of data, we must not forget the adage "Lies, damned lies and statistics". While discussing the potential dangers of big data and the inadequacy of present legal framework to mitigate this danger, the paper does not propose any solution. The objective of the paper was to start a discourse on the dark side of big data so that readers realize the need for new policy framework for big data. The privacy and protection rights have evolved from industrial era to digital era; big data calls for further evolution of these rights, both conceptually and legally.

Works Cited

Bhatia, Gautam. "The Supreme Court's Right to Privacy Judgment – I..." *Live Law*, August 2017.

Cohen, Julie E. "What Privacy Is For." *Harvard Law Review*, Novembwr 23, 2012.

Cohn, Cindy. "In Quest of Privacy in the Digital Age." *The New York Times*, October 18, 2017.

Davenport, Thomas H. "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review*, 2012.

Geer, Andrew Burt and Dan. "The End of Privacy." *The New York Times*, Oct 5, 2017.

Loon, Ronald van. "What is Big Data And How Does It Work?" *Data Science Central.* December 12, 2017. https://www.datasciencecentral.com/profiles/blogs/what-is-big-data-and-how-does-it-work.

Manyika J, Chui M. *Big data: the next frontier for innovation, competition, and productivity.* New York : Mickensy Global Institute, 2011.

Minkel, JR. "Confirmed: The U.S. Census Bureau Gave Up Names of Japanese-Americans in WW II." *Scientific American*, March 30, 2007.

"Misuse of voters list in Gujarat riots alleged." *Times of India .* March 12, 2002. https://timesofindia.indiatimes.com/india/Misuse-of-voters-list-in-Gujarat-riots-alleged/articleshow/3541858.cms.

Morgan, Jacob. "Privacy Is Completely And Utterly Dead, And We Killed It." August 2014.

Ohm, Paul. "Broken Promises of Privacy : Responding to the suprising failure of Anonymization." *UCLA Law Review*, 2010.

Ross, Ron. "Why Security and Privacy Matter in a Digital World." *NIST*, September 2017.

Sadowski, Jathan. "Why Does Privacy Matter? One Scholar's Answer." *The Atlantic*, February 26, 2013.

Schneier, Bruce, interview by Gazette. *On internet privacy, be very afraid* (August 25, 2017).

Singh, Parminder Jeet. "Privacy in the Digital Age." *The Hindu*, August 8, 2017.

The Guardian. "Facebook posts record growth in first quarter despite privacy scandal." April 2018.

Viktor Mayer Schonberger, Kenneth Cukier. *Big Data : The Essential Guide to Work, Life and Learning iIn the Age of Insight.* 2012."What Is Big Data." *SAS: The Power to know.* https://www.sas.com/en_in/insights/big-data/what-is-big-data.html.