NEW DIMENSIONS OF DATA SCIENCE PROFESSIONAL SKILLS AS EMERGED BY IDENTIFIED ETHICAL ISSUES: GDPR

S. Toleva-Stoimenova[1], K. Rasheva-Yordanova[1], D. Christozov[2]

[1]University of Library Studies and Information Technologies (BULGARIA)
[2]American University in Bulgaria (BULGARIA)

Abstract

Recent decades have witnessed an increased growth in data generated by humans and machines, giving birth to the Big Data paradigm. Analyzing Big Data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable. Moreover, Data Science has emerged as a new inter- and cross-disciplinary field, which includes statistics, informatics, computing, communication, management, and sociology. Data science focuses on analysis and uncovering hidden meaningful patterns, correlations, complex event processing and other insights.

This paper is related to Data Science and the Data Scientist's skills. The dimensions and challenges of Big Data and Data Science fundamental concept are briefly described. As the ways of capturing, collecting, aggregating and analyzing large and heterogeneous datasets needs powerful technologies and specific skills from Data Science a new profession emerges – a Data Scientist. The Data Scientist has the task of making sense out of the vast data and helping the organization in informed decision-making. It is therefore essential that the Data Scientist has to possess a lot of skills to face the serious data, process and management challenges. This study aims to discuss ethical concerns related to Big Data analytics which raises some topical issues about the Data Scientist's skills to reflect on.

The paper's primary focus is on some problems and constraints imposed to Big Data analytics according to the newly introduced GDPR. Organisations need to ensure that their data processing activities are carried out in accordance with the Data Protection Principles set out in the GDPR. As they are expected to be extremely challenging exploring the Data Scientist's ethical skills are timely to discuss. In this paper ethical skills are considered separately to emphasize the importance of meeting Data Protection requirements which will benefit both organisations and individuals in a Big Data context.

Keywords: Big Data, Data Science, Data Scientist's skills, ethical issues, GDPR.

1 INTRODUCTION

Recent decades have witnessed an increased growth in data generated by humans and machines, giving birth to the Big Data (BD) paradigm. The extraction of knowledge from the vast amounts of available digital information seems to be the next logical step in gradual transition from the „Information Age" to the „Knowledge Age". Analysing BD allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable. BD can deliver significant benefits for society and individuals in many areas such as health, scientific research, marketing, environment, etc. But there are serious concerns with the actual and potential impact of processing of huge amounts of data on the rights and freedoms of individuals, including their right to privacy.

This paper focuses on new challenges imposed to BD analytics according to General Data Protection Regulation (GDPR).

Access to personal information like buying preferences, call detail records and posts in social networks lead to increased privacy concerns [19,26]. Researchers have technical infrastructure to access the data from any data source including social networking sites, for future use whereas the users are unaware of the gains that can be generated from the information they posted [6]. These data were created in highly context-sensitive spaces, and it is entirely possible that some users would not give permission for their data to be used elsewhere. The BD and Data Science (DS) phenomenon include some implications underpinning to that effect. Should someone be included as a part of a large aggregate of data? What does it mean for someone to be spotlighted or to be analysed without

ISBN: 978-84-09-05948-5

knowing it? Who is responsible for making certain that individuals and communities are not hurt by the research process? What does informed consent look like? [6]

Data may be public (or semi-public) but this does not simplistically equate with full permission being given for all uses. BD researchers have to take into account that there is a considerable difference between being in public (i.e. sitting in a park) and being public (i.e. actively courting attention) [2].

This paper aims to bring awareness of the fundamental issues of privacy, security, governance and ethical aspects related to BD analytics. As the novel GDPR is expected to be extremely challenging for organizations exploring the Data Scientist's ethical skills are timely to discuss. In this paper ethical skills are considered separately to emphasize the importance of meeting data protection requirements which will benefit both organisations and individuals in a BD context.

The paper is organised as follows. The first section looks into BD nature and briefly explained DS fundamental concept. The second section gives an overview of GDPR and Data Protection Principles. The third section discusses Data Scientist's skills, specific to ethical conduct that he/she has to possess to meet the data protection requirements. In the conclusion, we argue that issues of DS professional ethics needs special attention in curriculum design to train DS specialists.

2 CHALLENGES FOR BD AND DS

The information generated and exchanged across networks has rapidly increased over the last two decades. This is due to advances in information and communication technologies, the digitalisation of production processes, the increasing use of electronic devices and networks, including the Internet of Things, cloud computing, etc.

The term BD refers to the large collection of heterogeneous unstructured data from different sources. They are not usually available in standard database formats we are used to operating with. BD has a complex nature because of the emergence of new forms of unstructured data generated by social media applications, web pages, blocks, log files, transactional applications, mobile and sensor networks, and other digital devices.

Researchers conceptualize BD along structural and functional dimensions. Many authors explicitly are based on its characteristics, such as the main definition of the 3 V's (Volume, Variety, Velocity) offered by Laney, 2001 [8, 12, 18, 23]. The volume relates to massive datasets, velocity relates to real-time data and variety relates to different sources of data. Other authors propose the 5 V's definition which adds Value and Veracity to the 3V's [1, 13, 14, 24]. Recently some researchers even use more V's to describe the BD [3, 20, 25].

As challenges of large volumes of data appeared and its structural diversity and complexity, the following relevant questions arise:

- The typical search helps us in discovering insights that have already been known, and it is useless in discovering things of which we are completely unaware.
- Query-based tools are time-consuming because search-based approaches require a virtually infinite number of queries.
- Statistical methods are largely limited to numerical data and inappropriate for unstructured data.

The functional dimension of BD includes powerful technologies and advanced algorithms for capturing, collecting, aggregating and analysing large and heterogeneous datasets [4]. First of all, tools and storage capabilities can handle BD. As BD is often noisy, unreliable, heterogeneous, dynamic in nature, it is captured, stored, mined, cleaned and integrated. After that comes the data analysis and modelling for BD. According to Christozov and Toleva the ability to address critical information, as well as verifying sources and considering constraints of applied technologies is a factor in generating useful knowledge from the acquired information [7]. The expectation from BD is that it may ultimately lead to better and more informed decisions.

Analytics refers to the methods used to analyse and acquire intelligence from BD. The literature contains a number of analytical processes and methods, such as text analytics, audio analytics, video analytics, social media analytics, predictive analysis of data [1], descriptive analytics, inquisitive analytics, prescriptive analytics and pre-emptive data analytics. Within these various BD analytics methods, there are a number of off the shelf software tools e.g. Hadoop, MapRecuce, Dyrad.

The broad challenges of BD can be grouped into three main categories, based on the data life cycle: data, process and management challenges [22]:

- *Data challenges* relate to the characteristics of the data itself (e.g. data volume, variety, velocity, veracity, volatility, quality, discovery and dogmatism).
- *Process challenges* are related to series of how techniques: how to capture data, how to integrate data, how to transform data, how to select the right model for analysis and how to provide the results.
- *Management challenges* cover for example privacy, security, governance and ethical aspects.

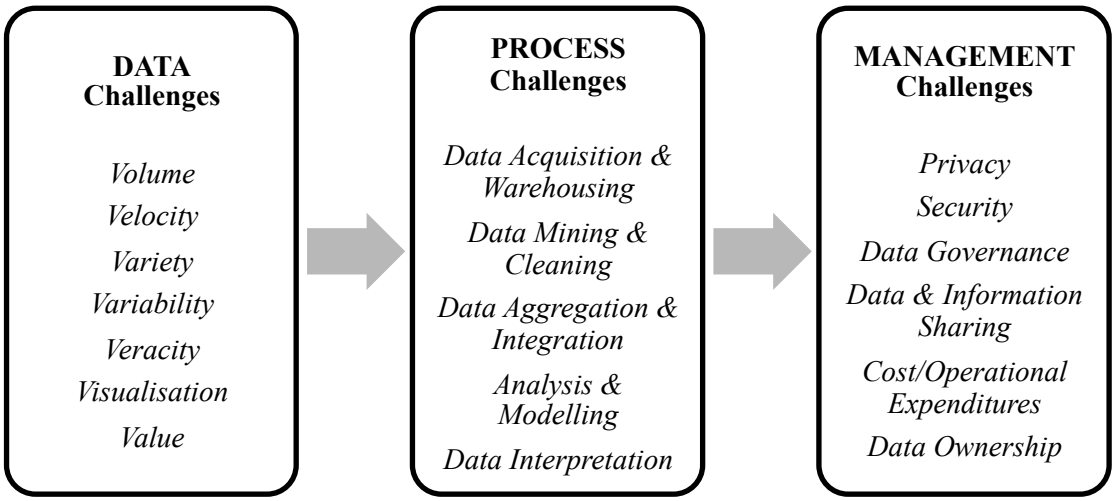

Figure 1. Data Lifecycle: Adapted from [22]

DS has emerged as a new inter- and cross-disciplinary field which includes statistics, informatics, computing, communication, management, and sociology. It is a generalized theory that provides algorithms for complex analysis of either structured or unstructured data, such as sophisticated statistical models, machine learning, neural networks, text analytics and other advanced data-mining techniques. DS focuses on analysis and uncovering hidden meaningful patterns, correlations, complex event processing and other insights. From DS a new profession emerges, the Data Scientist, who has the task of making sense out of the vast data and helping the organization in informed decision-making.

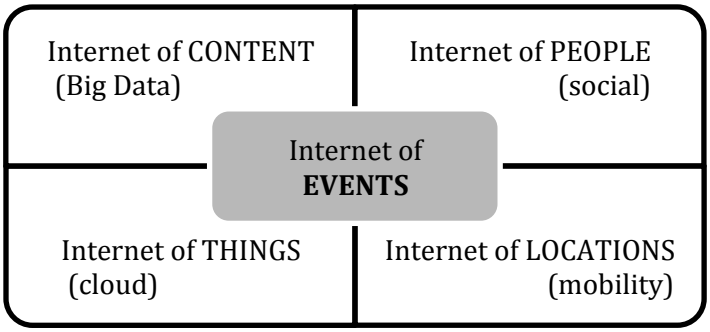

Figure 2. DS Data Sources: Adapted from [27]

DS aims to use the different data sources described in Figure 2 to answer questions grouped into the following four categories [27]:

- Reporting: What happened?
- Diagnosis: Why did it happen?
- Prediction: What will happen?
- Recommendation: What is the best that can happen?

Nowadays Data Scientist is faced with lots of serious data, process and management challenges, we have discussed earlier. According to [15, 16] the Data Scientist is expected to possess 3 generic skill categories - hard skills, soft skills and analytical skills. This paper is focused on another group of Data Scientist's skills related to some privacy aspects, these are what we call "ethical skills". As the novel GDPR is expected to be extremely challenging for organizations, it requires some ethical issues to be revised. In fact, ethical issues might arise in each segment if data processing, due to the multi-dimensionality of BD [5].

- The huge volume of data makes it possible for more pieces of valuable information to be identified or inferred than it was possible before.
- The high velocity of data makes feasible analysis in real time and thus a continuous refining of users' profiles.
- The variety of data sources make users traceable. In addition the diversity of data types allows data owners to build more complex and rich profiles of users.

The next two sections underline the importance of the GDPR challenge that Data Scientist has faced, exploring his/her specific ethical skills in this context.

3 GDPR

The GDPR have taken effect on May 25th, 2018, in order to keep pace with the modern digital information landscape. The regulation is applied to all organizations and businesses that process personal and marketing data from European residents. The GDPR conditions what and how personal data can be used commercially.

The predecessor to the GPDR is Data Protection Directive (DPD) of 1995 known as Directive 95/46/EC on the protection of individuals with regard to the processing of personal data. The DPD is focussed only on the principles of good governance without making no mention to a human right to data protection. By contrast, the GDPR „protects fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data" (Art. 1).

The GDPR principles are not so different from the principles set out in DPD. Some new requirements are introduced and the way these principles are applied. Moreover, the new regulation uses much broader definition of personal data. Personal data now means any information relating to an identified or identifiable natural person (data subject). A „natural person" is one who can be identified directly or indirectly; identifiers now include location, date, online identifiers or one or more factors specific to an individual's genetic, mental, cultural, economic, cultural, or social identity. (Art. 1)

There are six fundamental principles set out in GDPR to ensure the individual's rights and security of sensitive personal information, as follows [11,17].

Lawfulness, fairness and transparency. Art. 5(1)(a) Personal data must be processed lawfully, fairly and in a transparent manner in relation to the data subject. The references to lawfulness and fairness are the same as under the DPA. However, the reference to „transparent manner" is new and reflects the central importance of transparency to the GDPR. Transparency means explaining for which reasons organizations process which personal data. This principle does overlap with many of the elements of fairness.

Purpose limitations. Personal data of users must only be collected for „specified, explicit and legitimate purposes". Art. 5(1)(b) Data can only be used for a specific processing purpose that the subject has been made aware of and no other, without further consent. In summary, the purpose limitation principle states that personal data collected for one purpose should not be used for a new, incompatible, purpose. Compared to its predecessor the GDPR brings some limited changes to this principle. Further processing of personal data for archiving, scientific, historical or statistical purposes is still permitted, but is subject to the additional safeguards provided in **Art. 89**. This processing is not considered incompatible with the initial purposes.

Data minimization. Data collected on subject should be „adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed" Art. 5(1)(c)**.** The GDPR tightens the DPD restrictions further, stating that organisations should not collect data that is not necessary for a specified purpose that has been notified to data subjects. The GDPR introduces the „right to be forgotten" - to request erasure of personal data in several situations, for example, where the data is no

longer necessary for the purpose for which it was collected, or where it is processed on the basis of consent and the data subject withdraws that consent.

Accuracy. It is the responsibility of data collectors to ensure that the personal data is „accurate and kept up to date". Art. 5(1)(d)**.** Individuals have the right to request that inaccurate or incomplete data be erased or rectified within 30 days. This is very similar to the DPD. Accuracy also must be seen in the context of data hygiene, data management and data security in which accuracy mechanisms should be present, especially rectification mechanisms.

Data accuracy is important when large amounts of data are analysed, as with BD analytics. As incomplete (raw) data leads to incorrect analysis results plausibility algorithms can, at least partially, help to filter out and exclude incorrect data records. Moreover, the longer a data processing activity takes and the more extensive it is, the greater the risk of using incorrect data.

Storage limitations. It is expected personal data to be „kept in a form which permits identification of data subjects for no longer than necessary" Art. 5(1)(e)**,** i.e. data no longer required should be deleted. Neither the DPD nor the GDPR establish any minimum or maximum time periods for storing personal data. Clearly the longer personal data are retained, the less accurate the personal data are likely to be and the weaker the case for retaining the data will be. However, there are specific provisions on the processing of personal data for historical, statistical or scientific purposes.

Complying with storage limitations principle therefore ensures fair processing and facilitates compliance with the other principles stated above.

Integrity and confidentiality. This is the only principle that deals explicitly with security. The GDPR states that personal data must be processed „in a manner that ensures appropriate security of the personal data including protection against unlawful processing or accidental loss, destruction or damage, using appropriate technical or organisational measures.". Art. 5(1)(f). Organizations are responsible for ensuring that personal data are kept secure, both against external threats (e.g., malicious hackers) and internal threats (e.g., poorly trained employees). Currently, organisations use some coding techniques (e.g. pseudonymisation, cryptography or anonymisation technics), protected servers against external threats, closed-controlled system of data processing etc.

According to the GDPR the data subjects have eight rights, as follows:

Right to be informed. It is underpinned by the overarching concepts of accountability and transparency and is integral to complying with the Lawfulness, fairness and transparency principle.

Right to access. It gives individuals the right to obtain confirmation that their personal information are being processed and, if that is the case, to access and be provided with a copy of their personal data easily and at reasonable intervals in order to be aware of, and verify, the lawfulness of controller's processing activities.

Right to rectification. Data subjects are entitled to require a controller to rectify any errors in their personal data.

Right to be forgotten. Also known as Data Erasure, the right to be forgotten entitles the data subject to have the data controller erase his/her personal data, cease further dissemination of the data, and potentially have third parties halt processing of the data.

Right to restrict processing. In some circumstances, data subjects may not be entitled to require the controller to erase their personal data, but may be entitled to limit the purposes for which the controller can process those data (e.g., the exercise or defence of legal claims; protecting the rights of another person or entity; purposes that serve a substantial public interest; or such other purposes as the data subject may consent to).

Right to data portability. GDPR introduces the right for a data subject to receive the personal data concerning them, which they have previously provided in a „commonly use and machine readable format" and have the right to transmit that data to another controller.

Right to object. Data subjects have the right to object to the processing of their personal data for the purposes of direct marketing. In addition, personal data may be processed for scientific, historical or statistical purposes in the public interest, but individuals have a right to object to such processing.

Rights relating to automated processing. Data subjects have the right not to be subject to a decision based solely on automated processing which significantly affect them (including profiling).

4 DATA SCIENTIST'S ETHICAL SKILLS

The Data Scientist is considered to be able to write in programing languages like Python, R, Java, Ruby, Clojure, Matlab, Pig and SQL Besides, the Data Scientist should be familiar with the NLP, machine training, conceptual modelling, statistical analysis, predictive modelling and testing of hypotheses, working with databases. All of the above create **hard skills** group. The hard skills can be compared to what is expected from IT-professionals: subject matter expertise, data and technical skills, and maths and statistics knowledge [16].

The category **soft skills** comprises a great deal of non-technical communication skills, organizational business strategy and understanding of the architecture of the system [16]. Soft skills are psychological and emotional competences that helps people to deal effectively with challenges in personal or professional life. Unlike hard skills, soft skills are not job specific. The most essential soft skills are decision-making, problem solving, creative and critical thinking, effective communication, interpersonal relationship skills.

In [15,16] **analytical skills** are considered separately to emphasize the importance of analytical thinking of Data Scientist. The analytical skills are a subset of both hard and soft skills and are significant part of the professional profile of the Data Scientist.

Clearly, the Data Scientist is a professional with many qualifications. But we need to recognize that new job profiles and tasks have risen from the emergence of data-driven science. Generally, the knowledge produced by DS represents a new type of company asset. This data needs to be collected and protected according to the new GDPR that is why the Data Scientist have to possess some skills to meet the data protection requirements.

The GDPR itself requires every organization to appoint a data protection officer (DPO) whose main function it to promote and monitor compliance with the GDPR by a controller or processor and as a result, to protect the rights and freedoms of data subjects. We will refer to such skills as **ethical skills.**

The key components to the GDPR are not necessarily new and there have always been ethical norms in the digital information landscape. But today in the BD era the GDPR will have a major impact on DS activities, as far as Data Scientist operates with personal and sensitive information [10, 21]. In addition, BD analytics tends to involve collecting and analysing as much data as possible, and in many cases all the data points in a particular set, rather than a sample („n=all"). In this context, the GDPR principles of Purpose limitations, Data minimization and Storage Limitation should be applied throughout the entire life cycle of personal information, which goes against the principle of data warehousing and BD.

Generally, the DS process includes five steps: (1) defining the goal, (2) data collection, (3) data exploration and discovery of insights, (4) data modelling, (5) communicating and visualizing the result. It is evident that GDPR rules will introduce the biggest challenge in first, second and fifth step of the DS process.

The first step of the DS process is the identification of a purpose. It includes also defining what data will be collected and in what form e.g. consumer behaviour data processing activity.

The GDPR emphasizes that individuals have the right to be informed, prior the data collection, about what personal data will be collected, in which manner and for what purpose. The Data Scientist should then inform the data subject and ask for **consent**. Creating new applications for existing data will not be possible without consent. As per GDPR, consent must be informed, unambiguous, given with a clear affirmative act, and demonstrable.

Transparency also requires any information and communication with data subject to be easy to find and access and it happen in a clear and plain language.

The second step of the DS process is collecting the data which has to be considered with data minimization and storage limitation GDPR principles. In order to comply with minimization requirement, if a data controller is interested in information about a customer's preference regarding particular product, data such as marital status, salary, e-mail address, or telephone number are not necessary to find out a customer's perception of the product. As the Data Scientist is responsible for the protection of collected data (storage, all backup files, all versions and copies, etc.), he/she should be aware of the implications of the data storage principle. Data scientists should have insights about data storage and evaluate whether they are exposed to privacy violations. Both pseudonymisation and anonymization are encouraged in the GDPR and Data Scientists must be prepared to use proper

mechanisms that make them possible. Anonymization of data means that it irreversibly destroys any way of identifying a data subject**.** The resulting data should not be capable of singling any specific individual out, of being linked to other data about an individual, nor of being used to deduce an individual's identity. On the other hand, pseudonymisation is the process of replacing data that directly identifies an individual with data that indirectly identifies them (artificial identifiers or pseudonyms). Pseudonymisation does not remove all identifying information from the data but merely reduces the linkability of a dataset with the original identity of an individual (e.g., via an encryption scheme). Pseudonymous data are protected against identification, but they still are personal and allow re-identification, while anonymous data cannot be re-identified.

The fifth step of the DS process is communication and visualization, followed by implementation of the results. Transparency principle is clearly emphasized in the context of profiling, information duties and the demonstration of consent. The GDPR includes provisions dealing specifically with profiling, which is defined in Art. 4 as: „Any form of automated processing of personal data consisting of using those data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements." GDPR restrictions on the way popular profiling tools and user tracking technology, like cookies, could lead to many businesses requiring a data protection impact assessment for their profiling approach. Recital 71 of the GDPR also refers to examples of automated decision making "such as automatic refusal of an on-line credit application or e-recruiting practices without any human intervention". The wording here reflects the potentially intrusive nature of the types of automated profiling that are facilitated by BD analytics. The GDPR does not prevent automated decision making or profiling, but it does give individuals a qualified right not to be subject to purely automated decision making.

Data Scientists will likely be dealing with controllers and processors from different countries and therefore business cultures. Sometimes during the translation of cultural clichés and stereotypes into empirically verifiable datasets, subjectivity is introduced. In some circumstances even displaying different advertisements can mean that the users of that service are being profiled in a way that perpetuates discrimination, for example on the basis of race, sex, political opinion, religion, health status, etc. Related to the principle of transparency underpinning the GDPR is a recognition that the collection and processing of data should be carried out in a manner that prevents such discriminatory effects on persons. Data Scientists must have experience in dealing with different ways of thinking and doing business and have the flexibility to marshal these differences into a successful result.

According to the abovementioned steps of DS process we frame the most challenging tasks the Data Scientist is faced to achieve GDPR compliance (Table 1).

Table 1. Data Scientist's ethical skills required for collecting and processing of personal data

GDPR compliance tasks	**Data Scientist activities**	**Skills required**
Informed Consent	• Review the privacy notices, procedures, and contracts to address areas such as retention, security and data sharing. • Provide a model of declaration of consent. • Obtain explicit consent for every data collecting and processing activities. • Document the records of consent with all the personal data itself. • Collect personal data. • Periodic control of the validity of consent obtained and data retention periods in relation to the purposes for which data are collected. • Write in plain and clear language, free from professional and technical jargon.	• Databases; • Programming; • Linguistic proficiency; • Communication skills.
Pseudonymisation and Anonymization	• Process personal and sensitive data. • Use proper coding technics to personal data	• Programming;

	collected (masking, scrambling, encryption, etc.). • Protect pseudonymous and anonymous data against unauthorized access and re-identification.	• Databases; • Information security; • Networking; • Data processing; • Data ingestion; • Data mining; • Data preparation.
Profiling	• Provide data protection impact assessment. • Analyse data and discover structure, relationships and data rules. • Give information about profiling and profiling effects to the data subjects. • Give individuals a qualified right not to be subject to purely automated decision making. • Prevent discriminatory effects on persons.	• Advanced statistic; • Data processing; • Data ingestion; • Data mining; • Data preparation; • Machine learning; • Communication skills.

As shown in the table above the Data Scientist's ethical skills belong to both hard and soft skills sets. The hard skills are compulsory for Data Scientist to accomplish their activities for collecting and processing of personal data. These are technical skills, such as programming, databases, data handling, etc., as seen in the last column in the Table 1. Mastering different quantitative research questions needs Data Scientist to have some analytical skills (advanced statistics, modelling and machine learning, etc.). On the other hand, Data Scientist has to communicate with data subjects to comply with their right to the protection of personal data according to GDPR. These skills are from the set of their soft skills.

To sum up, the ability to handle data is a necessity for Data Scientists. However, handling BD nowadays is a challenging task and to succeed at this complex and highly non-linear discipline the data scientist's skill set needs to be adapted to the data protection requirements and the situation at hand. Ethical behaviour in this emerging field with so many possibilities, and where technology limits have shifted so dramatically is an obligatory part of DS professionals.

5 CONCLUSIONS

Some researchers have argued that Norbert Wiener was among the first who suggest the notion of "computer ethics". As DS matures as a field and increasingly affects the human life, it needs professional ethics code to establish the type of relationship between the professionals and the society including clients, research subjects, users of their services, etc. Although people have different privacy boundaries the privacy is a basic human need even for people that have nothing to hide. Clearly, Data Scientists cannot just take data and rely only on automated processing algorithms, they need to understand outcomes and should be aware of the ethical and legal implications.

Ethical conduct of a person possesses rights to access and capability to explore personal data of large groups of individuals raises new, previously unknown challenges. Complexity of setting the boundaries of what is acceptable and what is not is no more issue of common sense. GDPR regulates the use of personal data is only the legal side of professional ethics in the area of DS.

Drawing from this work we argue that issues of DS professional ethics needs special attention in curriculum design to train DS specialists. The curriculum has to build students' ethical imaginations and skills for collecting, storing, sharing and analyzing data derived from human subjects including data used in algorithms. The future Data Scientists need to be aware of the tenets of informed consent, discrimination, and privacy. Our work is aimed at developing a better understanding of these ethical issues related to Data Scientist's skills.

ACKNOWLEDGEMENTS

This work has been supported by National Science Fund at the Ministry of Education and Science, Republic of Bulgaria, within the Project DM 12/4 - 20/12/2017.

REFERENCES

[1] A. Gandomi, M. Haider, “Beyond the hype: BD concepts, methods, and analytics,” International Journal of Information Management, vol.35, no.2, pp.137-144, 2015.

[2] A. Marwick, D. Boyd, “To See and Be Seen: Celebrity Practice on Twitter,“ Convergence: The International Journal of Research into New Media Technologies, vol.17, pp.139-158, 2011.

[3] B. Ristevski, M. Chen, “Big Data Analytics in Medicine and Healthcare,” Journal of integrative bioinformatics, May 2018.

[4] C. Dede, A. Ho, P. Mitros. "BD Analysis in Higher Education: Promises and Pitfalls," EDUCAUSE Review 51, no. 5, September/October, 2016.

[5] C. K. Emani, N. Cullot, Ch. Nicolle, “Understandable Big Data: A survey, ” Computer Science Review, vol.17, pp. 70 – 81, 2015.

[6] D. Boyd, K. Crawford, “Critical Questions for BD: Provocations for a Cultural, Technological, and Scholarly Phenomenon,” Information, Communication, & Society, vol.15, pp.662-679, 2012.

[7] D. Christozov, S.Toleva-Stoimenova,”BD Literacy: A New Dimension of Digital Divide, Barriers in Learning via Exploring Big Data,” in Big Data: Concepts, Methodologies, Tools, and Applications (J. Girard., K. Berg, D. Klein eds.), IGI Global, pp.2300-2315, 2015.

[8] D. Laney, “3D data management: Controlling data volume, velocity and variety,” Tech. Rep. 949, META Group, 2001.

[9] G. Chen, K. Chen, D. Jiang, B. Ooi, L. Shi, H. Vo, S. Wu, “E3: an elastic execution engine for scalable data processing,” Journal of Information Processing, vol.20, no.1, pp.65–76, 2012.

[10] G. Vojkovic, “Will the GDPR slow down development of Smart Cities?,” 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018), Opatija, Croatia, pp. 1495-1497, 2018.

[11] G. Latchams, “A practical guide to the General Data Protection Regulation”, 2017. Retrieved from https://www.gregglatchams.com/wp-content/uploads/A-Practical-Guide-to-the-GDPR-Gregg-Latchams-v1-Sept-2017-1.pdf

[12] H. Chen, R. Chiang, V. Storey, "Business Intelligence and Analytics: From BD to Big Impact," MIS Quarterly, vol.36, no.4, pp.1165-1188, 2012.

[13] IBM and Said Business School, ‘‘Analytics: The real-world use of big data: How innovative enterprises extract value from uncertain data,’’ IBM Institute for Business Value and Said Business School Executive Report, Oct. 2012.

[14] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers, “BD: The next frontier for innovation, competition, and productivity,” McKinsey & Company, 2011. Retrieved from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

[15] K. Rasheva-Yordanova, E. Iliev, B. Nikolova, “Analytical Thinking As A Key Competence For Overcoming The Data Science Divide,” in Proceedings of EDULEARN18 Conference 2-4 July 2018, Palma, Mallorca, Spain, pp 7892-7898, 2018.

[16] K. Rasheva-Yordanova, V. Chantov, I. Kostadinova, E. Iliev, P. Petrova, B. Nikolova, “Forming of Data Science Competence for Bridging the Digital Divide,”

[17] M. Ch. Addis, M. Cutar, “The General Data Protection Regulation (GDPR), Emerging Technologies and UK Organisations: Awareness, Implementation and Readiness,” UKAIS 2018 Proceeding. Retrieved from https://www.ukais.org/resources/Documents/ukais%202018%20proceedings%20papers/paper_39.pdf

[18] P. Zikopoulos, C. Eaton, “Understanding BD: Analytics for Enterprise Class Hadoop and Streaming Data,” McGraw-Hill Education, 2011.

[19] S. Kaisler, F. Armour, J. Espinosa, W. Money, “BD: Issues and challenges moving forward,” in 46th Hawaii International onference on System Sciences (HICSS) , Hawaii. pp. 995-1004, 2013.

[20] S. S. Owais, N. S. Hussein, “Extract Five Categories CPIVW from the 9 V’s Characteristics of the Big Data,” International Journal of Advanced Computer Science and Applications, vol. 7, no.3, 2016.

[21] The United Kingdom Information Commissioner's Office, “BD, artificial intelligence, mashine learning and data protection,” Version 2.2., March 2017.

[22] U. Sivarajah, M. M. Kamal, Z. Irani, V. Weerakkody, “Critical analysis of Big Data challenges and analytical methods,” Journal of Business Research, vol. 70, pp. 263–286, 2017.

[23] X. Dong, D. Srivastava, “Big data integration,” in Proceeding of IEEE 29th International Conference on Data Engineering (ICDE2013), pp. 1245–1248, 2013.

[24] Y. Demchenko, P. Grosso, C. de Laat, P. Membrey, “Addressing big data issues in scientific data infrastructure,” in Proceedings of the IEEE International Conference on Collaboration Technologies and Systems (CTS '13), pp. 48–55, 2013.

[25] Y. Demchenko, C. de Laat, P. Membrey, “Defining Architecture Components of the Big Data Ecosystem”, In Proceeding of International Conference on Collaboration Technologies and Systems (CTS’2014), Retrieved from https://www.researchgate.net/publication/269272409_ Defining_architecture_components_of_the_Big_Data_Ecosystem

[26] V. Benjamins, “BD: From hype to reality?” In: Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14), pp. 2:1-2:2, ACM, New York, NY, USA, 2014.

[27] W.M.P. van der Aalst, “Data Scientist: The Engineer of the Future,” in Proceedings of the I-ESA Conference (K. Mertins, F. Benaben, R. Poler, and J. Bourrieres, eds.), vol. 7 of Enterprise Interoperability , pp. 13–28. Springer-Verlag, Berlin, 2014.