

Blagoj Ristevski¹ / Ming Chen²

Big Data Analytics in Medicine and Healthcare

¹ “St. Kliment Ohridski” University – Bitola, Faculty of Information and Communication Technologies, ul. Partizanska bb, 7000 Bitola, Republic of Macedonia, E-mail: blagoj.ristevski@fikt.edu.mk. <http://orcid.org/0000-0002-8356-1203>.

² Department of Bioinformatics, College of Life Sciences, Zhejiang University Zijingang Campus, Hangzhou, P.R. China

Abstract:

This paper surveys big data with highlighting the big data analytics in medicine and healthcare. Big data characteristics: value, volume, velocity, variety, veracity and variability are described. Big data analytics in medicine and healthcare covers integration and analysis of large amount of complex heterogeneous data such as various – omics data (genomics, epigenomics, transcriptomics, proteomics, metabolomics, interactomics, pharmacogenomics, diseasomics), biomedical data and electronic health records data. We underline the challenging issues about big data privacy and security. Regarding big data characteristics, some directions of using suitable and promising open-source distributed data processing software platform are given.

Keywords: Big Data Analytics, Data Mining, Health Informatics, Healthcare Information Systems

DOI: 10.1515/jib-2017-0030

Received: April 7, 2017; **Revised:** January 16, 2018; **Accepted:** March 20, 2018

1 Introduction

To obtain the best services and care for the patients, healthcare organizations in many countries have proposed various models of healthcare information systems. These models for personalized, predictive, participatory and preventive medicine are based on using of electronic health records (EHRs) and huge amounts of complex biomedical data and high-quality – omics data [1].

Contemporarily genomics and postgenomics technologies produce huge amounts of raw data about complex biochemical and regulatory processes in the living organisms [2]. These -omics data are heterogeneous, and very often they are stored in different data formats. Similar to these - omics data, the EHRs data are also in heterogeneous formats. The EHRs data can be structured, semi-structured or unstructured; discrete or continuous.

Big data in healthcare and medicine refers to these various large and complex data, which they are difficult to analyse and manage with traditional software or hardware [3], [4]. Big data analytics covers integration of heterogeneous data, data quality control, analysis, modeling, interpretation and validation [5]. Application of big data analytics provides comprehensive knowledge discovering from the available huge amount of data.

Particularly, big data analytics in medicine and healthcare enables analysis of the large datasets from thousands of patients, identifying clusters and correlation between datasets, as well as developing predictive models using data mining techniques [2]. Big data analytics in medicine and healthcare integrates analysis of several scientific areas such as bioinformatics, medical imaging, sensor informatics, medical informatics and health informatics. A survey of big data cases in medical and healthcare institutions/organizations is given in [6].


The new knowledge discovered by big data analytics techniques should provide comprehensive benefits to the patients, clinicians and health policy makers [7].

The remainder of the paper is organized as follows. Related work is described in the second section. Section 3 describes characteristics of big data, while big data analytics is depicted in the subsequent section. The next section explains some challenging issues about big data analytics techniques, while big data privacy and security are described in Section 6. Last section concludes this paper with discussion and further works.

2 Related Work

The rapid development of the emerging information technologies, experimental technologies and methods, cloud computing, the Internet of Things, social networks supplies the amounts of generated data that is growing tremendously in numerous research fields [8].

Blagoj Ristevski is the corresponding author.

 ©2018, Blagoj Ristevski and Ming Chen, published by De Gruyter.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.

On this point, contemporarily genomics and postgenomics technologies produce huge amounts of raw data about complex biochemical and regulatory processes in the living organisms [2]. These high throughput – omics data provide comprehensive insight towards different kinds of molecular profiles, changes and interactions, such as knowledge allied to the genome, epigenome, transcriptome, proteome, metabolome, interactome, pharmacogenome, diseasome, etc. [9]. These – omics data are heterogeneous and very often stored in different data formats. The main aims and characteristics of the different – omics disciplines are tabled in Table 1.

Table 1: The main aims of the variety of – omics disciplines.

– omics	The aim of study
Genomics	Study of the set of all genes in an organism, in a broader context non-coding parts of DNA are subject of study
Epigenomics	Study of all epigenomic modifications on the genetic material within a cell
Transcriptomics	Study of the expression level of all RNAs in particular cell, or cell population
Proteomics	Study of all possible interactions that a protein can present, complete set of proteins expressed by a genome in a given cell type or organism, under defined conditions, at a given time
Metalobomics	Study of the whole set of the metabolites (small-molecule compounds) within a cell, an organelle, a tissue, an organ or an organism
Interactomics	Study of the entire set of interactions (both: physical and indirect interactions) between and among proteins and other molecules within a particular cell and consequences of those interactions. These interactions are displayed as graphs and called biological networks
Pharmacogenomics	Study which combines pharmacology and genomics in order to analyse the role of the genome in individual's drug response
Diseasomics	Study of all diseases and disorders of an organism, often focusing on those diseases and disorders caused by genetic modifications

Similar to these – omics data, the EHRs data are also stored in heterogeneous formats. The EHRs data, which can be structured, semi-structured or unstructured; discrete or continuous, contain personal patients' data, clinical notes, diagnoses, administrative data, charts, tables, prescriptions, procedures, lab tests, medical images, magnetic resonance imaging (MRI), ultrasound, computer tomography (CT) data. Some of these data are acquired from wearable sensors or capture from medical monitoring devices, with different collection frequency [5] that makes these data to have complex features and high dimensions [10]. Dealing with noisiness and incompleteness of EHRs are still challenging task and these shortcomings should be consider while applying data mining techniques [11].

These growing amounts of various – omics data need to be collect, clean, store, transform, transfer, visualize and deliver in a suitable manner to be represented to the clinicians [12]. The processing of these big data in medicine and healthcare can be accelerating by using cloud computing and powerful multicore central processing units (CPUs), graphics processing units (GPU) and field-programmable gate arrays (FPGAs) with parallel processing methods.

3 Big Data Characteristics

The term big data is described by the following characteristics: *value*, *volume*, *velocity*, *variety*, *veracity* and *variability*, denoted as 6 “Vs” [13], [14], shown in Figure 1. Besides these 6 “Vs”, some authors has defined more than these 6 properties to describe big data characteristics [15].

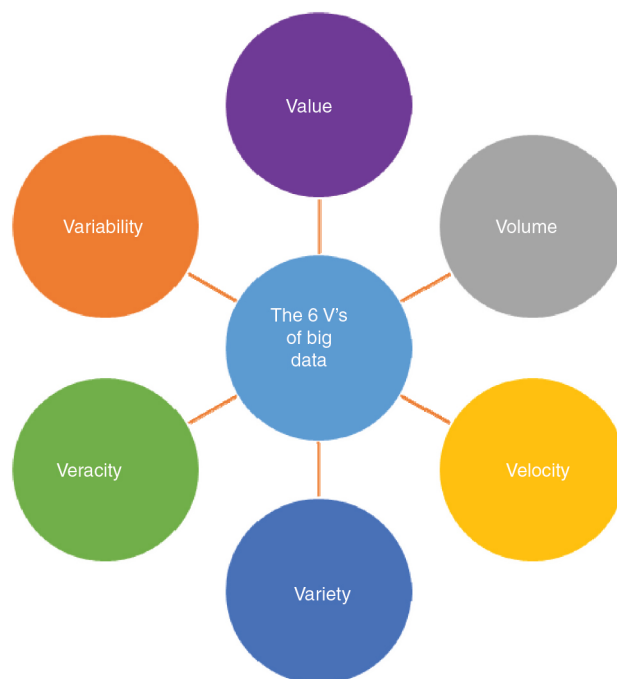


Figure 1: The 6 V's of big data.

The *volume* of health and medical data is expected to raise intensely in the years ahead, usually measured in terabytes, petabytes even yottabytes [14], [16]. Volume refers to the amount of data, while *velocity* refers to data in motion as well as and to the speed and frequency of data creation, processing and analysis. Complexity and heterogeneity of multiple datasets, which can be structured, semi-structured and unstructured, refer to the *variety*. *Veracity* refers to the data quality, relevance, uncertainty, reliability and predictive value [14], while *variability* regards about consistency of the data over time. The *value* of the big data refers to their coherent analysis, which should be valuable to the patients and clinicians.

Considering the big data characteristics, data searching, storage and analysis, a very appropriate and promising software platform for development of applications that can handle big data in medicine and healthcare is the open-source distributed data processing platform Apache Hadoop MapReduce [1], [17] that is based on data-intensive computing and NoSQL data modeling techniques [18].

4 Big Data Analytics

Applications of big data analytics can improve the patient-based service, to detect spreading diseases earlier, generate new insights into disease mechanisms, monitor the quality of the medical and healthcare institutions as well as provide better treatment methods [19], [20], [21].

Data mining techniques employed on EHRs, web and social media data enable identifying the optimal practical guidelines in the hospitals, identifying the association rules in the EHRs [22] and revealing the disease monitoring and health-based trends. Moreover, integration and analysis of the data with different nature, such as social and scientific, can lead to new knowledge and intelligence, exploring new hypothesis, identifying hidden patterns [14].

Nowadays, smart phones are excellent platforms to deliver personal messages to patients to involve them in behavioral changes to improve their wellbeing and health conditions. The mobile phone messages can substitute delivering of medical and motivational advices to the patients [14].

5 Challenges in Big Data Analytics

Regarding collection of large amount data, some challenging issues should be considered. Obtaining high-throughput – omics data is tied to the cost of experimental measurements. Concerning heterogeneity of the data sources, the noise of the experimental – omics data and the variety of the experimental techniques, environmental conditions, biological nature should be considered, before integration of these heterogeneous data

and before employing of the data mining methods. Different data mining techniques can be applied on these heterogeneous biomedical data sets, such as: anomaly detection, clustering, classification, association rules as well as summarization and visualization of those big data sets.

These shortcomings might lead to the unreliability of some of the data points, such as missing values or outliers. Despite of these drawbacks of the – omics data, EHRs data are very influenced by the staff who entered the patient's data, which can lead to entering missing values, incorrect data as a result of mistakes, misunderstanding or wrong interpretation of the original data [5]. Integration of data from various databases and standardization for laboratory protocols and values still remain challenging issues [10].

High dimensionality of the – omics data means, that there have many more dimensions or features than the number of samples, and on the other side the EHRs data which regard to the individuals/patients, makes data mining techniques to be more challenging task.

The subsequent stage is the pre-processing of the data, which usually envelop handling noisy data, outliers, missing values, data transformation and normalization. This data pre-processing enables to be applied statistical techniques and data mining methods and thus the big data analytics quality and outcomes can improve and can result with discovering of novel knowledge. This novel knowledge obtained by integration of the – omics and EHRs data should results with improving of the implemented healthcare to the patients as well to advanced decision making by the healthcare decision policy makers.

6 Big Data Privacy and Security

Two important issues towards big data in healthcare and medicine are security and privacy of the individuals/patients [14], [23]. All medical data are very sensitive and different countries consider these data as legally possessed by the patients [2]. To address these security and privacy challenges, the big data analytics software solutions should use advanced encryption algorithms and pseudo-anonymization of the personal data. These software solutions should provide security on the network level and authentication for all involved users, guarantee privacy and security, as well as set up good governance standards and practices.

7 Discussion and Future Work

Big data analytics in medicine and healthcare is very promising process of integrating, exploring and analysing of large amount complex heterogeneous data with different nature: biomedical data, experimental data, electronic health records data and social media data. Integration of such diverse data makes big data analytics to intertwine several fields, such as bioinformatics, medical imaging, sensor informatics, medical informatics, health informatics and computational biomedicine. As a further work, the big data characteristics provide very appropriate basis to use promising software platforms for development of applications that can handle big data in medicine and healthcare. One such platform is the open-source distributed data processing platform Apache Hadoop MapReduce that use massive parallel processing (MPP) [20], [24]. These applications should enable applying data mining techniques to these heterogeneous and complex data to reveal hidden patterns and novel knowledge from the data.

Recent hardware innovations in processor technology, newer kinds of memories/network architecture will minimize the time spent in moving the data from storage to the processor in a distributed setting [25].

Acknowledgements

This paper was supported by the Ministry of Education and Science of the Republic of Macedonia and the Ministry of Science and Technology (MOST) of the Government of the People's Republic of China.

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

- [1] Yang C, Li C, Wang Q, Chung D, Zhao H. Implications of pleiotropy: challenges and opportunities for mining big data in biomedicine. *Front Genet* 2015;6:229.
- [2] Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. *IEEE J Biomed Health Inform* 2015;19:1209–15.
- [3] Kankanhalli A, Hahn J, Tan S, Gao G. Big data and analytics in healthcare: introduction to the special section. *Inform Syst Front* 2016;18:233–5.
- [4] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inform Sci Syst* 2014;2:3.
- [5] Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. –Omic and Electronic Health Record Big Data Analytics for Precision Medicine. *IEEE Trans Biomed Eng* 2017;64:263–73.
- [6] Wang Y, Kung LA, Wang WY, Cegielski CG. An integrated big data analytics-enabled transformation model: application to health care. *Inf Manag* 2017;55:64–79.
- [7] El-Gayar O, Timsina P. Opportunities for business intelligence and big data analytics in evidence based medicine. In: *System Sciences (HICSS), 2014 47th Hawaii international conference on* 2014:749–57.
- [8] Gu D, Li J, Li X, Liang C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. *Int J Med Inform* 2017;98:22–32.
- [9] Gligorijević V, Malod-Dognin N, Pržulj N. Integrative methods for analyzing big data in precision medicine. *Proteomics* 2016;16:741–58.
- [10] Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. *Biomed Inform Insights* 2016;8:1.
- [11] Gaitanou P, Garoufallou E, Balatsoukas P. The effectiveness of big data in health care: a systematic review. In: *Metadata and semantics research*. 2014:141–53.
- [12] Lillo-Castellano JM, Mora-Jimenez I, Santiago-Mozos R, Chavarria-Asso F, Cano-González A, García-Alberola A, et al. Symmetrical compression distance for arrhythmia discrimination in cloud-based big-data services. *IEEE J Biomed Health Inform* 2015;19:1253–63.
- [13] Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang GZ. Big data for health. *IEEE J Biomed Health Inform* 2015;19:1193–1208.
- [14] Archenaa J, Anita EM. A survey of big data analytics in healthcare and government. *Procedia Comput Sci* 2015;50:408–13.
- [15] Borne K. Top 10 big data challenges – a serious look at 10 big data V’s. *MAPR*, 2014:NO4, 80.
- [16] Hermon R, Williams PA. Big data in healthcare: what is it used for? In: *Australian Ehealth Informatics and Security Conference*. 2014:40–9.
- [17] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM* 2008;51:107–13.
- [18] Trifonova OP, Il’in VA, Kolker EV, Lisitsa AV. Big data in biology and medicine. *Acta Naturae* 2013;5:13–6.
- [19] Agarwal M, Adhil M, Talukder AK. Multi-omics multi-scale big data analytics for cancer genomics. In: *International Conference on Big Data Analytics*. Cham, Switzerland: Springer International Publishing; 2015:228–43.
- [20] He KY, Ge D, He MM. Big data analytics for genomic medicine. *Int J Mol Sci* 2017;18:412.
- [21] Tan SL, Gao G, Koch S. Big data and analytics in healthcare. *Methods Inf Med* 2015;54:546–7.
- [22] Dinov ID, Heavner B, Tang M, Glusman G, Chard K, Darcy M, et al. Predictive big data analytics: a study of Parkinson’s disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PLoS One* 2016;11:e0157077.
- [23] Costa FF. Big data in biomedicine. *Drug Discov Today* 2014;19:433–40.
- [24] Yao Q, Tian Y, Li PF, Tian LL, Qian YM, Li JS. Design and development of a medical big data processing system based on Hadoop. *J Med Syst* 2015;39:23.
- [25] Kambatla K, Kollias G, Kumar V, Grama A. Trends in big data analytics. *J Parallel Distrib Comput* 2014;74:2561–73.