RE-DISCOVERING "BIG DATA" AND “DATA SCIENCE” IN GEODESY AND GEOMATICS

Ioannis D. Doukas[1]

[1] Professor of Geodesy and Geomatics, Dept. of Civil Engineering, Aristotle University of Thessaloniki, Greece, (jdoukas@civil.auth.gr)

Key words: *geodesy*; *geomatics*; (geo-) *big data*; *data science*; *data analytics*

ABSTRACT

In late years, a dramatic increase in data across the globe is happening. “Data science”, in parallel with the so-called "big data", monopolize the press, the internet, the public interest (via: text, applications, technology, public and private sector, offering-finding jobs...). Geodesy/Geomatics has traditionally been about data, dealing with a lot of data. Consequently, by considering data inference, algorithm development, related technology, and their blending multidisciplinarity which characterizes these sciences, it is not difficult to say that both geodesy and geomatics have also always been about the multidisciplinary field of “data science”. From the appearance of computers, on the one hand the data surrounding geodesy (and later, geomatics) and, on the other hand the geodetic problems-requirements of each era, were frequently resulting into heavy-duty tasks that tested (even surpassed) existing analytic methods and computers. In terms of modern methods of (geodetic) data-analysis, as well as computing power and speed requirements, the bar was always being set higher.

In the present paper, the "re-discovery" / "re-revelation" of the perpetual relationship among geodesy/geomatics (and their branches e.g. remote sensing, engineering geodesy, etc.) and “data science” - “big data”, is attempted. It is an old and fundamental relationship that we must always remember, revisit and refresh it, in the light of modern scientific and technological developments. After all, the new methodologies concerning “data science” - “big data”, their new tools and possibilities, their various solutions that circulate (directly or indirectly related to the branches of geodesy/geomatics), and their future trends, are very serious development-levers that have a permanent key-role in the evolvement of geodesy/geomatics in the 21st century. In this regard, some thoughts are paid about the necessary and permanent participation of “data science” in the curricula related to geodesy and geomatics.

I. INTRODUCTION

Nowadays, a wave of critical (scientific and technological, as well) events occur (on a daily basis) everywhere (around the globe mainly, but for the present….). Three of them, do play a dominant role related to the "area" of this paper:.

A. Products, strategies and processes, through the exploitation of new technologies, undergo a fundamental adjustment within an organization. All of this is contained in the general term "Digital transformation". In close relation to this, we have the following two:

B. The strong entry of connected manufacturing and the digital convergence between industry, business and other processes. All these events are meant by the cyber-physical transformation of manufacturing, called with the name "Industry 4.0" (i.e. the fourth industrial revolution). Cyber Physical Systems (CPS) (considered as the core of future production), are capable of substantial changing of both, the way the technical systems are constructed nowadays and, the way of interaction between people and their environment, in many domains. There are five key-areas that make up the backbone of Industry 4.0: (i). Data (collection, processing), (ii). Decentralization and service orientation, (iii). Networking – integration, (iv). Assistance systems and finally, (v). Self-organization and autonomy.

C. Since the term "Industry 4.0" is yet not very common outside Europe (mainly the German-speaking countries and Scandinavia are familiar with it), in many parts of the world the "Internet of Things (IoT)" and "Internet of Services (IoS)" (created and established from an open architecture, it is a global market of Web-services) are considered the equivalent alternative tools.

From a clearly technical aspect, the "Internet of Things (IoT) is a system of interrelated computing devices, mechanical and digital machines, objects, animals or people that are provided with unique identifiers and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction". From a simplified human aspect, the "Internet of Things means taking all the physical places and things in the world and connecting them to the internet".

These events have established the right conditions for the appearance of: "Data Science" and "Big Data". By taking into consideration the scientific fields and disciplines of Geodesy and Geomatics Engineering (GGE), they obviously have a direct (and strong) relationship with this new regime.

II. DATA SCIENCE-BRIEF HISTORY-TERMS

The field of Data Science (DS) gets its deep roots within Statistics (statistical models, etc.). The idea of DS spans many diverse fields, and slowly (but steadily) it has been making its way into the mainstream for over fifty (50) years (!). "Officially", the DS story starts in 1962 when the American mathematician John Tukey published the article: 'The Future of Data Analysis' (Tukey, J., 1962). In this article, Tucker calls for a reformation of statistics, as he refers to an unrecognized (yet) science that it will have as its subject the learning from data, or 'data analysis'. Twelve years later, the Danish computer science pioneer Peter Naur publishes the book: 'Concise Survey of Computer Methods'. The most decisive and important thing in this book, is the fact that the term "data science" is used many times (Naur, P., 1974).

After these two milestones, the story is still evolving (a huge number of sources can be found in the bibliography and the Internet, as well).

Data science (DS) is the (computing-based) science that deals with raw data, in order to extract meaningful information from it. From then on, it aims to communicate the information that has emerged in order to create value. It is a modern scientific field that, through the analysis of data relating to a particular "subject", aiming at the discovery of knowledge (Pierson, L., 2017). DS deals with data (unstructured, structured) and includes everything related to data cleaning, preparation and, finally analysis. On the other hand, it is a resultant-science with components: mathematics, statistics, software programming, "intelligent" data recording, understanding / realizing things in different ways. The difference between DS and "classical statistics" is based (to a small extent) on the use of "unusual" programming languages that are not used in statistics, but mainly, DS differs from statistics on the need for "expertise" in the specific subject-matter in which DS is used.

Statisticians usually have only a limited amount of expertise in fields outside of statistics, consequently they're almost always forced to consult with a subject matter expert to verify exactly what their findings mean and to decide the best direction in which to proceed. On the other hand, data scientists generally have (or are required to have) strong expertise in the field in which they work. By generating profound knowledge, they use their expertise to understand exactly what these indications mean in relation to the area in which they work. By quoting (Davenport and Patil, 2012) "A data scientist is a hybrid of data hacker, analyst, communicator, and trusted adviser". Indeed, an extremely powerful and rare combination is. Finally, when there are data-related problems that have to do with designing, building/implementing software solutions, then the solution is left to Data Engineering (DE) (Pierson, L., 2017).

Both, DS and DE, are dealing with the following data styles:

Structured: There is a "traditional" RDBMS (relational database management system) where data is stored, processed, and manipulated.

Unstructured: Data doesn't fit into a RDBMS, since usually is generated from human activities.

Semi-structured: Fact 1, data doesn't fit into a RDBMS. Fact 2, this data is structured by tags which help in creating a form of order (and hierarchy, as well) in this data.

II.1. Data Science-Elements: Big Data

Although the term 'Big Data' (BD) is "officially" introduced by the Oxford English Dictionary in 2013, Roger Mougalas (of O'Reilly Media) is the first person who uses this term in its 'modern context' (in 2005). He relates the term BD with a large dataset that is (almost) impossible to manage - process using 'traditional' tools. So, 2005 is the year in which the BD-revolution began and has been established since then. The companion of BD is Metadata which (in a simplified explanation) deals with "data about data" (Rouvroy, A.,2016)

BD can carry "big errors" such as: lack of consistency and reliability, "false" data, noise, lack of representativeness, incomplete information (of course there are also ethical issues, personal data issues, etc.) (Liu, J., et al. 2016). One of the major constraints to success with big data, is of course the human factor. Southekal gives a very detailed presentation of indicative BD-problems (Southekal, P., 2017).

Found in the Web, below is quoted some interesting information and facts about BD on Earth:

1900: The amount of human knowledge doubled every 100 years.

2013: IBM shared statistics showing 90% of the data in the world had been created within the last two years (!!) This means that now for university students, the newest knowledge that they receive during the first year of training already in the third year becomes obsolete.

Nowadays: Due to global "digitalization", the amount of human knowledge doubled every 2 years. At the same rate, the volume of new data produced by mankind is growing

2018-2025: The data volume generated worldwide is set to rise its current level of 33 zettabytes (2018) to around 175 zettabytes by 2025 !! (Reisnel, D., et al. 2018).

How big is 175 ZB?? :

-One zettabyte is equivalent to a trillion GBs

-In case someone is able to store the entire Global Datasphere on DVDs, the result will be a stack of DVDs that could get that person to the moon 23 times or circle Earth 222 times…….

According to the UN, these changes are no longer linear in time, they are exponential, that's why the new digital world is called 'exponential'. Consequently, the fact today is that BD is everywhere, so technically speaking, 'big data' now really means 'all data' — or just 'data'. It is interesting to recall that a century ago, the resource in question was oil. Nowadays, similar concerns are being raised by the giant-companies (p.e. Amazon, Apple, Facebook, Google, Microsoft….) that deal in data. In simple words, the oil of the digital era is data (Economist, 2017).

Big data 10Vs:

Traditional BD is defined (by a 2011-report from McKinsey Global Institute) as featuring one (or more) of the following 3 V's: Volume, Velocity, and Variety (Manyika et al. 2011). A rough internet search shows that, in less than 10 years, these Vs easily became ten (for the present, of course….), as follows:

#1: Volume: The best-known characteristic of BD.

#2: Velocity: The speed at which data is being generated, produced, created, or refreshed.

#3: Variety: There is structured data and mostly unstructured data (but also semi-structured) as well. Most big data seems to be unstructured, but let's not forget that besides "typical data" such as: audio, image, video files, and other text formats, there are also: log files, click data, machine - sensor data, etc.

#4: Variability: It refers to: (i). The number of inconsistencies in the data (ii). Multiple data dimensions, since there are many and different data types and sources (iii). The inconsistent speed at which BD is loaded into the database.

#5: Veracity: As any or all of the above properties increase, the veracity (confidence or trust in the data) drops. It refers more to the following: the provenance or reliability of the data source, its context, and how meaningful it is to the analysis based on it.

#6: Validity: Similar to veracity, it refers to how accurate and correct the data is for its intended use (about 60 % of a data scientist's time is spent cleansing data before being able to do any analysis).

#7: Vulnerability: Big data, undoubtedly brings new security concerns.

#8: Volatility: It is dealing with questions like: 'How old does data need to be before it is considered irrelevant, historic, or not useful any longer?', 'How long does data need to be kept for?'.

#9: Visualization: There are heavy-duty technical challenges (concerning: limitations of in-memory, poor scalability, poor technology, functionality, response time).

#10: Value: If there is no value derived from the data, then all the above characteristics of BD are meaningless !..

II.2. Data Science-Elements: Data Analytics

Data Analytics (DA): The science of examining raw data with the purpose of drawing conclusions about that information. It involves applying an algorithmic or mechanical process to derive insights. The focus of DA lies in inference, which is the process of deriving

conclusions that are solely depended on what the researcher already knows.

There have always been four types of DA:

Descriptive: It reports on the past. Using data aggregation and data mining, it responds to the questions, "What happened and Why?" and contributes to summarize results.

Diagnostic: It uses the data of the past to study the present. It responds to the questions "Why did this particular something happen?", "What went wrong?"

Predictive: It uses insights based on past and current data to predict the future (an event, a trend, etc.). Using statistical modeling and simulation, it responds to the question, "What will happen?" and contributes to informed decisions concerning the future.

Prescriptive: This is the most effective DA-tool to deal with data insights in order to generate value from it. It uses models to specify optimal behaviors, processes, structures, systems and actions. Using heuristics and optimization models, it responds to the question, "What should be done?" and contributes to making complex time-sensitive decisions.

The Evolution Of DA

1. Analytics 1.0 - The Era of Business Intelligence-BI (born in the mid-1950s). 2. Analytics 2.0-The Era of Big Data (emerged in the mid-2000s when internet and social media companies—Amazon, Google, eBay, etc.—began to amass and analyze new/different kinds of information). 3. Analytics 3.0-The Era of Data-Enriched Offerings. 4. Analytics 4.0-The Era of Automated Analytics: On the way to the next stage in analytic maturity. The combination of data-mining techniques and machine learning algorithms, along with the existing DA (descriptive, predictive, prescriptive), now comes to full fruition.

II.3. Data Science-Elements: Data Mining

Data Mining (or Knowledge Discovery in Databases-KDD): The process of sorting through large sets of data to identify patterns and create/establish relationships to solve problems through data. The number of standards and rules produced by data mining system is huge. In any case, measures of pattern interestingness (whether objective or subjective) can be utilized to direct the discovery process, and knowledge is represented by any identified interesting pattern. Representative data mining models indicatively are: Decision Trees, K-Means, Naive Bayes, Neural Networks, Support Vector Machines (SVM) (Jha, A. et al. 2016).

II.4. Data Science-Elements: Machine Learning

The superset is "Artificial Intelligence-AI", which refers to any kind of technique which enables computers to mimic the human intelligence, through the following means: decision trees, deep learning, 'if-then'-rules, logic, machine learning (ML).

Machine learning -ML-(or algorithmic learning): It is a subset of AI. It includes complex and impalpable statistical techniques, which enable machines to improve at tasks with experience. By applying (in an iterative manner) algorithmic models to data, hidden patterns or trends are discovered by computer, which are predictive-tools. Three are the main learning methods: (i) Supervised, (ii) Unsupervised, and (iii) Semi-supervised (Pierson, L., 2017).

Finally, a subset of ML is "Deep Learning" which contains algorithms permitting software to train itself, in order to carry out tasks. All these by using multilayered neural networks exposed to ….big data data.

II.4. Data Science-Elements: Methods, Software

The scientific field of DS is so wide, that it is impossible to be described effectively in a few pages. Some indicative "must-skills" are given here, if the target is one to become a:

A. Data Scientist:

Programming skills: R, Python (Python, Java, Perl, C/C++ are the most common coding languages that are used in DS).

Hadoop platform (a set of open source programs and procedures, which can be used as the "backbone" of big data operations, by everyone).

Apache Hive data warehouse (software that facilitates reading, writing, and managing large datasets residing in distributed storage using SQL)

Apache Pig (a platform for analyzing large data sets, which consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs).

SQL - NoSQL database coding

B. Big Data professional:

Analytical skills: To be able to make sense of the data.

Creativity: Capturing and applying new ideas and methods in order to: gather, interpret, and analyze a data strategy

Math-Stat skills: Knowledge and experience to old-fashioned "number crunching". This is extremely crucial, be it in DS, DA, or BD.

Computer science: Continuous pursuit (and need) to come up with algorithms to process data into insights.

Business skills: Understanding of both, the business objectives, as well as the underlying processes that drive the growth of the business.

C. Data Analyst:

Data Intuition: Ability to think like a data analyst

Programming skills: R, Python

Statistical-mathematical skills: Descriptive and inferential statistics, experimental designs.

ML-skills: Transforming/Mapping raw data - convert it into another format (Data wrangling or Data munging), with the intent of making it more appropriate and valuable.

Communication and Data Visualization skills.

III. RELATIONS WITH GEODESY & GEOMATICS

In the GeoBuiz Report (2018 Edition), five sections are found concerning the new digital era and the geodetic/geomatics arena: Global Geospatial Industry, GNSS & Positioning Technologies, GIS/Spatial Analytics Technologies, Earth Observation (EO) Technologies, and 3D Scanning Technologies. Of course, by breaking down the above sections, we could see sub-sections which are really "serious big data producers" like (in alphabetic order):

3D -AR (Augmented Reality), 3D-VR (Virtual Reality), Building Information Modelling (BIM, GeoBIM, Historic-BIM), Cadastral surveying, Cartography, Disaster Monitoring, Management, Earth observation (EO), Earth sciences, Engineering Surveying, Geocomputation, Geodesy, Geodetic Metrology (Industrial Metrology, 3D-Metrology, Large Volume Metrology, Technical Geodesy), Geography, Geomatics-Geoinformatics, Geosensors-Geosensorics (IoT, Edge devices), GeoWeb-SensorWeb, GIS and applications-Spatiotemporal analysis, GIS and Automated Mapping/Facilities Management, GNSS and applications (PNT, GPS-weather forecasting, tracking, etc.), High Definition Surveying, Hydrography-Hydrographic surveying, Laser Scanning/LIDAR, Location-based services, Photogrammetry, Remote Sensing, SLAM/SLAMMOT, Smart cities, Surveying-Topography, Unmanned Vehicles (air, ground, water)….

All the above do give BD and on top of that, it is obvious that DS is always and everywhere present. After all, people dealing with GGE are familiar with this new era of tools. Simply, because they have a good and long history in common. The whole "new thing" started (Davenport, T.H., 2014) as "Decision support" (1970-1985) where data analysis was performed to support decision making. Then (1980-1990), there was the new phase of "Executive support", i.e. analysis aiming to produce decisions by senior executives. In the decade (1990-2000) the development upgraded to OLAP-Online analytical processing, i.e. analysis of multidimensional data-tables with software. Between 1989-2005, the first choice was BI-Business intelligence (data-driven decisions, with emphasis on reporting). Between (2005-2010) the main role is played by Analytics, focused on math-stat analysis for decisions. Finally, from 2010 to present, BD-DS are everywhere, with a focus on very large, fast-moving, unstructured data.

This brief timeline of data-analysis is very well known to geodesists firstly (and geomatics-people, later) because:

(i). Their data was – depending on the time (each era) of the evolution of GGE - just …big data (ii). The methods they were using (use and will use) are parts or modifications or copies of the mentioned above techniques/methods.

The biggest part of the GGE-big data nowadays, is Spatial data i.e. discrete representations of continuous phenomena (Evans M.R. et. Al. 2019). Spatial data is represented with the following basic models: (a). Raster (grid): satellite images are good examples of raster data, (b). Vector: consist of points, lines, polygons, and their aggregate (or multi-) counterparts, and (c). Network: graphs consisting of spatial networks form another important data type used to represent road networks.

The instances of the above data types that exhibit at least one of the 3 V's (i.e. volume, velocity, and variety) are defined as "Spatial Big Data" (SBD) (Karimi, H.A. 2014). Even better, a fresh (modern) term already exists, "Geo Big Data" (Hayasi, Y. et al. 2016).

The major factors that have facilitated and strengthened the rise of IoT are: (i). Increased adoption of machine learning practices (ii). Increased deployment of sensor technologies, and (iii). Improved real-time data processing capabilities.

These three factors have a significant impact on Geospatial analysis: Sensor nodes (Doukas, I.D. 2016) create data which is timestamped and geo-location-stamped. The meaning of a "thing", in the context of the Internet of things (IoT), needed to be noted here: "Thing" is an entity or physical object that has a unique identifier, an embedded system and the ability to transfer data over a network. These "things" are surely expected to become active elements in business, information, research and social processes (Hassan, F. 2018). Many IoT applications consider both, the location of an edge device-"thing" (i.e. a network device that has an immediate or direct connection to the Internet or an external non-propriety network), as well as the proximity, with respect to other connected devices. Consequently, a heavy processing demand of multidimensional geospatial data, and DA capabilities as well, are both a must. Only a GIS application is the solution for such demanding tasks. The powerful alliance of GIS and IoT network and data technologies: (i). Makes attainable the real-time spatiotemporal DA, (ii). Enables geo-insights to be delivered at the right time and place, precisely when these insights are actionable.

IV. GGE-EDUCATION NEEDS

Because of the tight space limitations, the comments here are exclusively based on an excellent paper (Hong, T., et al. 2018). In most of their findings, by just replacing "power" (even "grid", "energy") with "GGE", the result is (almost) a perfect match, and it is easily applicable on GGE. ! Let's see the quotations of just three (3) of their key-remarks (but modified with GGE):

1. After gathering various perspectives from members of academia, industry, and government, we have to propose an interdisciplinary and entrepreneurial approach to revising the traditional GGE curriculum for training the next generation of "GGE data scientists"
2. Ideally, university GGE programs are supposed to provide fresh graduates with skills that meet the current and emerging needs of the GGE industry. However, the mainstream GGE curriculum in the past has rarely recognized DA as a crucial component.
3. After a decade-long GGE modernization effort, the GGE industry is now sitting on a gold mine of data. We now have the opportunity to dig into the data, gain valuable insights, and make the decisions needed to run the most complex man-made system on earth. Let's start with training the next generation of GGE data scientists!

V. LOOKING TO THE FUTURE

The Future of DS has two paths to expand:

Algorithms (p.e. Cognitive Machine Learning, Embedded Deep Learning, Hyperfast DA, Massive-scale graphs, Natural Language generation, Spatio-temporal Predictive DA, Structural database generation, etc.)

Applications (Blockchain and Cybersecurity, Disaster management Geosensorics, Healthcare, Internet of Things and …Internet of "Everything", Smart "Everything", etc.). We are in front of Tiny Stuff and Smart Dust: Computers smaller than a grain of sand, sprayed (to measure chemicals in the soil, for SHM-Structural Health Monitoring) or injected (yes, in the human body).

The general context is extra-turbulent. Of course, its subset of geo-context, is behaving (at least) the same. Well documented, accurate and on-time decisions needed to harmonize the GGE (science-profession) with these contemporary "intruders". On the foundations of that, the issue of harmonizing university education with DS is of crucial importance.

The circumstances impose that all (most probably) should agree with Albert Einstein: "Everything should be made as simple as possible, but not simpler".Because of the tight space limitations, the comments

REFERENCES

Davenport, T.H. (2014). Big Data at Work-Dispelling the Myths, Uncovering the Opportunities. ISBN 978-1-4221-6816-5, Harvard Business School Publishing.

Davenport, T.H. and Patil, D.J. (2012). Data Scientist: The Sexiest Job of the 21st Century, Harvard Business Review, October 2012.

Donoho, D. (2015). 50 Years of Data Science, Tukey Centennial workshop, Princeton NJ, Sept-18 -015.

Doukas, I.D. (2016). On the High-tech Onrush of Sensors, Geosensors, Sensor Fusion and their Networks. Their Influence on Geodesy and Geomatics. "Measuring and Mapping the Earth"-Honorary Volume dedicated to Professor Emeritus Christogeorgis Kaltsikis, pp. 148-164, Ziti Publications, Thessaloniki.

Economist (2017). Regulating the internet giants-The world's most valuable resource is no longer oil, but data, (economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data), (Accessed: 14-1-2019).

Evans M.R., Oliver D., Yang K., Zhou X., Ali R.Y., Shekhar S. (2019). Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities. In: Wang S., Goodchild M. (eds) CyberGIS for Geospatial Discovery and Innovation. GeoJournal Library, vol 118. Springer, Dordrech.

GeoBuiz Report (2018 Edition). Geospatial Industry Outlook & Readiness Index. Geospatial Media and Communications Pvt. Ltd.

Goodchild, M.F. (2018). Big Geodata. Comprehensive Geographic Information Systems, 19-25, DOI: 10.1016/b978-0-12-409548-9.09595-6.

Hajirahimova, M.S. and Aliyeva, A.S. (2017). About Big Data Measurement Methodologies and Indicators, International Journal of Modern Education and Computer Science (IJMECS), Vol.9, No.10, pp. 1-9, 2017 DOI: 10.5815/ijmecs.2017.10.01D.

Hasan, Q.F. (Ed.) (2018). Internet of Things A to Z: Technologies and Applications, First Edition. ISBN 978-1-111-945674-2, John Wiley & Sons, Inc.

Hayasi, Y., Suzuki, Y., Sato, S. and Tsukahara, K. (2016). Disaster Resilient Cities-Concepts and Practical Examples. ISBN: 978-0-12-809862-2, Elsevier Inc.

Hilbert, M. and López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information, Science 01 Apr 2011, Vol. 332, Issue 6025, pp. 60-65, DOI: 10.1126/science.1200970.

Hong, T., Gao, D. W., Laing, T., Kruchten, D., & Calzada, J. (2018). Training Energy Data Scientists : Universities and Industry Need to Work Together to Bridge the Talent Gap. IEEE Power and Energy Magazine, 16(3), 66–73. doi:10.1109/mpe.2018.2798759.

Jha, A., Dave, M. and Madan, S. (2016). A Review on the Study and Analysis of Big Data using Data Mining Techniques. International Journal of Latest Trends in Engineering and Technology (IJLTET). Vol. 6 (3), pp. 94-102.

Karimi, H.A. (Ed.) (2014). Big Data-Techniques and Technologies in Geoinformatics. ISBN 978-1-4665-8655-0, CRC Press.

Liu, J., Li, J., Li, W. and Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues, ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 115, pp. 134-142.

Manyika, J. et al. (2011). Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute.

Naur, P. (1974). Concise Survey of Computer Methods, Studentlitteratur, Lund, Sweden, ISBN 91-44-07881-1, pp. 1-397.

Pierson, L. (2017). Data Science for Dummies, 2nd Edition. ISBN 978-1-119-32763-9, John Wiley & Sons, Inc.

Reinsel, D., Gantz, J. and Rydning, J. (2018). Data Age 2025: The Digitization of the World- From Edge to Core. International Data Corporation (IDC).

Rouvroy, A. (2016). "Of Data And Men"-Fundamental Rights And Freedoms In A World Of Big Data. Bureau Of The Consultative Committee Of The Convention For The Protection Of Individuals With Regard To Automatic Processing Of Personal Data [ETS 108]).

Singh, K. (2018). Understanding Different Components & Roles in Data Science. (https://dimensionless.in/understanding-different-components-roles-in-data-science/) (Accessed: 31-3-2019).

Southekal, P. (2017). Data for Business Performance: The Goal-Question-Metric (GQM) Model to Transform Business Data into an Enterprise Asset, ISBN 978-1634621847, Technics Publications

Tukey, J. (1962). The Future of Data Analysis. The Annals of Mathematical Statistics, Vol. 33, No. 1 pp. 1-67.