Understanding Data Science Framework & studying its aspects

Pradnya Khot[1], Aniket Milkhe[2], Gayatree Sorte[3]

[1]Student, Department MCA, DY Patil Institute of master of Computer Application and Management.
[2]Student, Department of Computer Science & Engineering, Prof Ram Meghe College of Engineering & Management.
[3]Associate Software Engineer, GlobalLogic India

---***---

Abstract - *The present data upset isn't just about big data, it is about data, all things considered, and types. While the issues of volume and velocity introduced by the ingestion of gigantic measures of data stay predominant, it is the quickly creating difficulties being introduced by the third v, variety, which requires more consideration. The requirement for a far-reaching way to deal with find, access, repurpose, and genuinely coordinate every one of the assortments of data is the thing that has driven us to the improvement of a data science system that structures our establishment of doing data science. Remarkable highlights in this system incorporate issue distinguishing proof, data revelation, data administration and ingestion, and morals. A contextual investigation is utilized to show the structure in real life. We close with a conversation of the significant job for data keenness.*

Key Words: Data Science, Big Data, Data Acumen, Data discovery.

1. INTRODUCTION

Data science is that the translational research field par excellence that begins with translation: the particular problem that must be solved. It integrates many actors and fields of practice and is suitable for team science. The article recounts developing a framework to know what it means to do data science. Our data science research approach is predicated on handling real and widespread problems of public politics. It is a search model that starts with translation, working directly with communities or stakeholders and that centers the problems. The "research pull" versus a "research push" to get the research foundation for data science.

For data science, the processing of varied problems in several domains creates synergies and a general need for research and thus a search impulse. This data science framework justifies the refinement of scientific practices around data ethics and data insight (literacy). A brief discussion of those topics concludes the article.

2. DATA SCIENCE FRAMEWORK

Conceptual illustrations are proposed to seize the existence cycle of data science. An easy Google look for "Data Science" exhibits pages and pages with images. These numbers have overlapping attributes and may thoroughly summarize unique additives of the data science system. It is important to head past the conceptual framework and create a framework that may be operationalized for real data science practice. Our data science framework gives a complete method to fixing data science issues and builds basis for research. [1,2]

There are four features of our framework that range from different frameworks and are defined in element below. First, we specify the trouble to be addressed and constantly persists within the framework, wherein the data science studies is grounded on trouble. Second, we approach data discovery, as a pre-eminent undertaking and no longer as an afterthought. Third, governance and records ingestion play an important function in fostering trust and setting up data sharing protocols. Fourth, we actively join the ethics of data science to all elements of the framework. Subsequent, we describe the elements of the data science framework. Although the body is defined linearly, it is further away from being a linear system represented with the aid of using a circular arrow. [3]

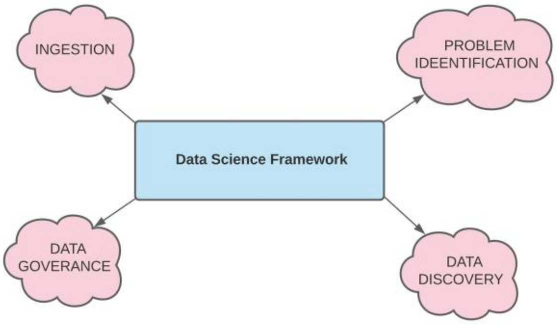

Fig 1: Data Science Framework

2.1 Problem Identification

Data Science brings disciplines and communities together to conduct interdisciplinary research that propounds comprehension into current and future societal challenges. Data becomes a common language for communication between the disciplines. The data science process begins with identifying the problem. It is achieved through reviews of traditional literature, including scrutiny (electronic reports from government, industry, and nonprofits) to find best practice. Knowledge (domain) also plays a role in translating information obtained from understanding the phenomena underlying the data. [4,5]

Knowing the domain provides the context to define, evaluate, and interpret the results at each stage of the investigation. They can take many forms resulting from an understanding of theory, modeling, or the underlying changes observed in the data.

2.2 Data Discovery

Data discovery is the recognition of undeveloped data sources that is probably associated with the precise subject of interest. Data pipelines and related equipment commonly begin on the factor of ingestion (Weber, 2018). The distinctive aspect of the framework is to commence the data pipeline with data discovery. The intention of the data discovery procedure is to think deeply and imaginatively about all the data, to seize all the possible types of data that would be beneficial to the issues at hand, and to collect a listing of these data sources. [6]

A crucial part of doing data science is to concentrate at the enormous reuse of existing data in conceptual improvement work. Data science methodologies provide possibilities to wrangle this data and observe it to investigate questions. In comparison to conventional studies approaches, data science expertise's studies permit researchers to study all sources of statistics. The purpose of this approach is that data collection can be subjected to ongoing differences information and knowledge.

Khan, Uddin and Gupta (2014) study the significance of range in data science sources. In the identical types of data too, for instance administrative facts, the issues (studies question) drive their use and applicability of the statistics content material to the subject beneath consideration. This range determines which field discoveries may be made ("Data Diversity", 2019). Researchers and difficulty depend on specialists decide "what the data are for any unique purpose, how the data is interpreted, and what proof is appropriate". An indistinguishable perspective is that data is "relational" and it's meaning is primarily based on its history, their attributes and the interpretation of the data when analyzed (Leonelli, 2019). [7,8,9]

The integration of data from distinctive sources implies the improvement of techniques primarily based totally on statistical concepts that verify the usability of the data (Economic Commission for Europe of the United Nations, 2014, p. 2015). [11,12] These integrated data sources offer the opportunity of assessing social repute to study and solution questions which have been tough to clear up within side the past. This demonstrates that the usefulness and applicability of the data vary depending on its use and field. The data are regularly incomplete, tough to access, unclean and unrepresentative. Due to the want for governance throughout a multiple groups and organizations, there can also be regulations on data access, facts linking and redistribution. Finally, reused facts can pose methodological problems in terms of inference or generate information from data, usually in the form of statistical, computational and theoretical models (Japec et al., 2015; Keller, Shipp and Schroeder, 2016) [17]

To this end, it is been taken into consideration it is beneficial to outline the data in four categories, designed, administrative, opportunity and procedural. The outcomes are a richer supply of data to useful resource trouble fixing and better inform the research plan. One caveat is the want to weigh the price of reusing present data in opposition to the brand-new data collection, wondering whether new experiments might produce quicker outcomes than locating and reusing data. In our experience, the advantages of reusing present data sources regularly outweigh those costs. Most importantly, it affords clues of facts gaps for the worthwhile improvement of recent facts collections.

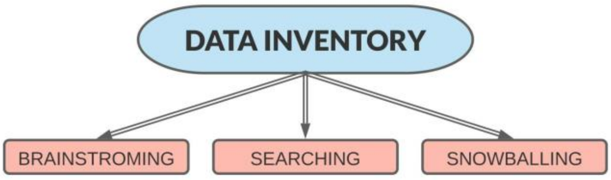

Fig 2- Data Inventory

2.3 Data Governance and Ingestion

Data governance is the establishment and enforcement of rules and procedures related to the access, dissemination and destruction of data. Data ingestion is the process of bringing data to the data management platform (s).

The combination of different data sources can lead to privacy and confidentiality issues, often due to conflicting interests between researchers and funders working together. For the sake of clarity, privacy refers to the amount of personal information that individuals allow others to access about themselves and confidentiality is procedure to keep the data secure. [15,16]

Information systems should be set up for case processing to ensure that only social workers have access to this private data and the granted access authorizations. Our focus is on policy analysis. Casework requires identification of individuals and families for the info to be useful, policy analysis doesn't. For casework, information systems must be found out to make sure that only social workers have access to those private data and approvals granted for access. Our focus is policy analysis.

Data governance necessitates tools to identify, manage, interpret and disseminate data (Leonelli, 2019). These tools are necessary to facilitate decision-making about the different ways of handling and evaluating data and to articulate conflicts between data sources; thereby changing research priorities to take into account not only publications but also data infrastructure and Data retention.

Governance and data ingestion best practices are part of the training of all members of the research team as well recorded in formal data management plans. The resulting modified read and write data or the code that can generate the modified data is generated from the original data sources, is stored on a secure server and is only accessible via secure remote access. The researchers do not have direct access to the data files. For these projects, access to data is mediated using various data analysis tools hosted on our own secure servers that connect to the data server via authenticated protocols (Keller, Shipp & Schroeder, 2016).[20,22]

2.4 Data Wrangling

These next stages of running the Data Science Framework activities quality, readiness, linkage, and exploration assessment activities can easily take up most of your project time and resources and help assess the quality of the data. Cleanse and enrich the available raw data in a more processed format. It is also known as data munging. This helps data scientists to accelerate the decision-making process and thus benefits the company. This method is used by a wide variety of high-profile companies, partly because of the advantages it has and partly because of the large amount of data that the company analyzes and processes. The technical properties of data prior to analysis have been found to be extremely useful in helping organizations quickly analyze large amounts of data. Data dispute is very important. Because this is the only way to use raw data.

In practical business environments, user information sometimes comes in different parts from different environments. Sometimes we store this information on multiple computers in different spreadsheets, and in different systems, which sometimes can lead to data redundancy, incorrect data, or missing data. In order to create a transparent and efficient system for data management, it is best to have all data available in one central place so that it can be used.

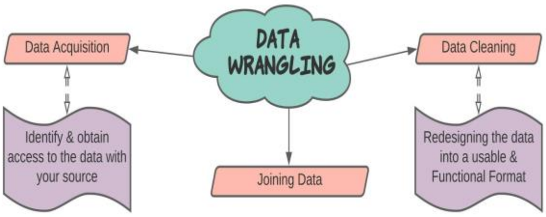
DATA WRANGLING
Data Acquisition
Identify & obtain access to the data with your source
Joining Data
Data Cleaning
Redesigning the data into a usable & Functional Format

Fig 3- Data Wrangling

2.5 Fitness-for-Use Assessment

Data adequacy was introduced in the 1990s from a management and an industry perspective (Wang and Stone, 1996) and later extended to official statistics by Brackstone (1999). Fitness-for-use begins off evolved with assessing the restrictions imposed at the statistics through the unique statistical techniques that will be used and if inferences are to be made whether or not or now no longer the statistics are consultant of the population to which the inferences extend. This evaluation ranges from simple descriptive tabulations and visualizations to complex analyzes. Finally, the usability must characterize the content of the information in the results.

2.6 Statistical Modelling and Analyses

Statistics and a statistical modeling are key to drawing strong conclusions from incomplete information (Adhikari & DeNero, 2019). Statistics offer regular and simple phrases and definitions for describing the connection among observations and conclusions. An appropriate statistical analysis depends on the research question, the intended use of the data in support of the research hypothesis, and the assumptions required for a particular statistical method (Leek and Peng, 2015). The ethical dimensions include ensuring accountability, transparency and the absence of algorithmic bias.

2.7 Communication and Dissemination

The communication includes sharing data, well-documented codes, working documents and dissemination through conference presentations, publications and social media. These steps are essential to ensure that the processes and results are transparent and reproducible (Berman et al., 2018). The step is to tell the story of the analysis and convey the context, the purpose and implications of the research and results (Berinato, 2019; Wing, 2019). Visuals, case studies, and other supporting evidence back up the results.

Communication and dissemination also are critical for constructing and preserving a network of practice, which may include dissemination through portals, databases and repositories, workshops and conferences, and the creation of new journals to protect privacy and the ethical dimension of investigation.

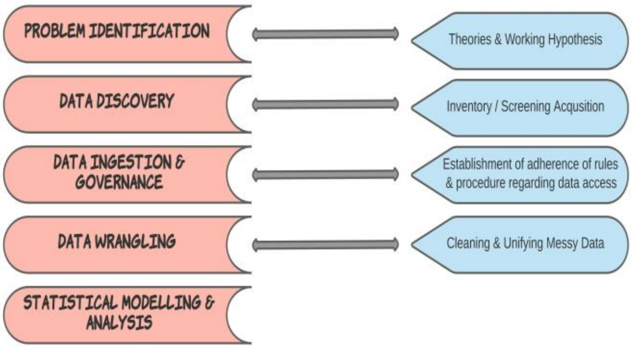
PROBLEM IDENTIFICATION
DATA DISCOVERY
DATA INGESTION & GOVERNANCE
DATA WRANGLING
STATISTICAL MODELLING & ANALYSIS
Theories & Working Hypothesis
Inventory / Screening Acqusition
Establishment of adherence of rules & procedure regarding data access
Cleaning & Unifying Messy Data

Fig 4: Overall Framework

3. Data Acumen

During the time spent doing data science, we have discovered that a considerable lot of the customers of this exploration do not have adequate data acumen and in this manner can be overpowered with how to utilize data-driven bits of knowledge. It is unreasonable to imagine that most of chiefs are data researchers. Indeed, even with space information, some proficiency in data science areas is valuable, including the underpinnings of likelihood and measurements to illuminate dynamic under vulnerability (Kleinberg, Ludwig, Mullainathan, and Obermeyer, 2015).

Data acumen, generally alluded to as data proficiency, seems, by all accounts, to be first presented during the 2000s as sociologies started to embrace and utilize freely open data (Prado and Marzal, 2013). We characterize data acumen as the capacity to make great decisions about the utilization of data to help issue arrangements. It isn't just the premise of factual and quantitative examination; it is a basic component to further develop society what's more, a vital initial step to measurable agreement. The requirement for strategy and other chiefs with data acumen is developing in corresponding with the gigantic repurposing of a wide range of data sources (Bughin, Seong, Manyika, Chui, and Joshi, 2018). [25,26,27]

Data acumen is both a standard and general idea. A data educated individual ought to reasonably comprehend the fundamentals of data science, (e.g., the data science structure depicted in Figure 1is a decent guide), and have the option to express inquiries that expect data to give proof:

What is the issue?

What are the examination inquiries to help the issue?

What data sources may educate the inquiries? Why?

How are these data conceived? What are the inclinations and moral contemplations?

What are the discoveries? Do they bode well? Do I trust them? How might I utilize them?

A data proficient individual comprehends the whole cycle, regardless of whether they don't have the right stuff to attempt the measurable examination. Data acumen requires a comprehension of how data are conceived, and why that matters for assessing the nature of the data for the examination question being tended to. As numerous sorts of data are found and repurposed to resolve scientific inquiries, this part of data education is progressively significant. Being data proficient is imperative to know why our instinct may not frequently be correct (Kahneman, 2011). We accept that building data limit and acumen of choice creators is a significant aspect of data science.[29]

4. CONCLUSIONS

We highlighted data discovery as a crucial however frequently unnoticed step in maximum data science frameworks. Without information discovery, we would flip to handy data sources. Data Discovery extends the power of data science with the aid of using taking into consideration many new information sources, now no longer simply technical sources. We additionally expand new behaviors with the aid of using taking a principles-primarily based totally method to moral issues as an essential feature at some stage in the information technology lifecycle. Each step of the data science framework consists of the documentation of the selections made, the strategies used and the effects that provide the opportunity of reusing and reusing the data. sharing and reproducibility.

Our data science framework gives a rigorous and repeatable, but flexible, basis for data science. The framework can function an evolving roadmap for data science as collectively to address the ever-converting data environment. Develop information insights among stakeholders, concern remember specialists and choice makers.

REFERENCES

[1] Adhikari, A., & DeNero, J. (2019). The foundations of data science. Retrieved December 1, 2019, from

https://www.inferentialthinking.com/chapters/intro#The-Foundations-of-Data-Science

[2] American Physical Society. (2019). Ethics and values. Retrieved from

https://www.aps.org/policy/statements/index.cfm

[3] Borgman, C. L. (2019). The Lives and After Lives of Data. Harvard Data Science Review, 1(1).

https://doi.org/10.1162/99608f92.9a36bdb6

[4] Box, G. E. P., Hunter, W. G., & Hunter J. S. (1978). Statistics for experimenters. Hoboken, NJ: Wiley.pp.563-571

[5] Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. Stamford, CT: McKinsey Global Institute.

[6] Berkeley School of Information. (2019). What is data science? Retrieved December 1, 2019, from

https://datascience.berkeley.edu/about/what-is-data-science/

[7] Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., Salary, A. (2018).

Realizing the potential of data science. Communications of the ACM, 61(4), 67–72.

https://dl.acm.org/citation.cfm?doid=3200906.3188721

[8] Brackstone, G. (1999). Managing data quality in a statistical agency. Survey

[9] Methodology, 25(2), 139–150. https://repositorio.cepal.org//handle/11362/16457

[10] Committee on Professional Ethics of the American Statistical Association. (2018). Ethical guidelines for statistical practice. Retrieved from https://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf

[11] Dasu, T., & Johnson, T. (2003). Exploratory data mining and data cleaning. Hoboken, NJ: Wiley.

[12] Data diversity. (2019, January 11). Nature Human Behaviour, 3, 1–2. https://doi.org/10.1038/s41562-018-0525-y

[13] Data for Democracy. (2018). A community-engineered ethical framework for practical application in your data work. Global Data Ethics Project. Retrieved December 1, 2019, from

https://www.datafordemocracy.org/documents/GDEP-Ethics-Framework-Principles-one-sheet.pdf

[14] De Veaux, R., Hoerl, R., & Snee, R., (2016). Big data and the missing links. Statistical Analysis and Data Mining, 9, 411–416. https://doi.org/10.1002/sam.11303

[15] Editorial: Nature research integrity is much more than misconduct [Editorial]. (2019, June 6). Nature 570, 5. https://doi.org/10.1038/d41586-019-01727-0

[16] Garber, Allan. (2019). Data science: What the educated citizen needs to know. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.88ba42cb

[17] Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., . . .Usher, A. (2015). Big data in survey research: AAPOR task force report. Public Opinion Quarterly, 79, 839–880.

https://doi.org/10.1093/poq/nfv039

[18] Kahneman, D. (2011). Thinking, fast and slow. New York, NY: Farrar, Straus and Giroux.

[19] Keller-McNulty, S. (2007). From data to policy: Scientific excellence is our future. Journal of the

American Statistical Association, 102, 395–399. https://doi.org/10.1198/016214507000000275

[20] Keller, S. A., Shipp, S., & Schroeder, A. (2016). Does big data change the privacy landscape? A review of the issues. Annual Review of Statistics and Its Application, 3, 161–180. https://doi.org/10.1146/annurevstatistics- 041715-033453

[21] Keller, S., Korkmaz, G., Orr, M., Schroeder, A., & Shipp, S. (2017). The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches. Annual Review of Statistics and Its Application, 4, 85–108. https://doi.org/10.1146/annurev-statistics-060116-054114

[22] Keller, S., Korkmaz, G., Robbins, C., Shipp, S. (2018) Opportunities to observe and measure intangible inputs to innovation: Definitions, operationalization, and examples. Proceedings of the National

Academy of Sciences (PNAS), 115, 12638–12645. https://doi.org/10.1073/pnas.1800467115

[23] Keller, S., Lancaster, V., & Shipp, S. (2017). Building capacity for data driven governance: Creating a new foundation for democracy. Statistics and Public Policy, 4, 1–11.

https://doi.org/10.1080/2330443X.2017.1374897

[24] Leonelli, S. (2019). Data governance is key to interpretation: Reconceptualizing data in data science. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.17405bb6

[25] National Research Council. (2007). Engaging privacy and information technology in a digital age. Washington, DC: National Academies Press.

[26] United Nations Economic Commission for Europe (UNECE). (2014). A suggested framework for the quality of big data. Retrieved December 1, 2019, from https://statswiki.unece.org/download/attachments/108102944/Big%20Data%20Quality%20Framework

%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2

[27] Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers.

Journal of Management Information Systems, 12(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

[28] Weber, B. (2018, May 17). Data science for startups: Data pipelines (Pt. 3). Towards Data Science.

Retrieved from https://towardsdatascience.com/data-science-for-startups-data-pipelines-786f6746a59a

[29] Wickham, H. (2014). Tidy data. Journal of Statistical Software, 59(10), 1–23.

https://doi.org/10.18637/jss.v059.i10

[30] Wing, J. M. (2019). The data life cycle. Harvard Data Science Review, 1(1).

https://doi.org/10.1162/99608f92.e26845b4