THE FUTURE OF DATA: UTOPIA OR DYSTOPIA?

Assof Prof. Aleksandrina Murdzheva[1],

Nikolay Vlaychev[2]

e-mail: amurdjeva@unwe.bg[1],

e-mail: nikolay_vlaychev@scalefocus.com[2]

Abstract

The journey into the future of data is not merely a technical discussion; it is a profound exploration of conceptual, philosophical, business, and ethical aspects that are rapidly reshaping our world. Is the future of data a utopia or a dystopia, and how do we ensure the choice is ours? The speed of technological development often outpaces its societal adoption, posing significant challenges that demand immediate, critical attention.

Ключови думи: Data, AI, data usage, data ethics, data storage

JEL: O03

The Unstoppable Snowball of Data Generation

Creation vs. Meaningfulness

The modern world is characterized by an incessant, accelerating generation of data. Everything—from our daily interactions to complex global systems—is contributing to this digital deluge. With the adoption of technologies like the Internet of Things (IoT), the concept of blockchain, and the rise of generative AI creating synthetic data, nowadays the world is living in interesting days. The consensus is clear: humanity does not have the luxury of choosing to stop generating data. It is an unstoppable snowball driven by the systems and media we have established.

The volume of data worldwide is growing exponentially, with growth driven primarily by a few key sources in recent years [1] [2] [3].

While precise figures in zettabytes (ZB) for the growth of each individual source (such as Social Media vs. IoT) are not readily available in public reports, the statistics provide a clear picture of the key drivers and the breakdown by data format.

- *Current Volume*: Projections for 2024 place the global volume of data (created, captured, copied, and consumed) at around 149 zettabytes (ZB).
- *Projected Growth*: The volume is expected to grow to 181 ZB by the end of 2025, representing an annual growth rate of over 20%
- *Historical Context*: Approximately 80% of all data in the world has been generated in the last two to four years alone.

Volume growth is driven by the following key categories related to social and corporate behavior:

[1] Assoc Professor, Information technologies and communications, University of National and World Economy, Sofia, Bulgaria

[2] Software Engineer at ScaleFocus, Bulgaria, https://www.linkedin.com/in/nvlaychev/

Source	Description of the Contribution	Key Statistics
User-Generated Data	Data created by end users through personal devices and online activity (photos, videos, posts).	**70% of all global data** is generated by users.
AI and ML	AI systems are **"data vampires"** that continuously generate, process, and manipulate vast data sets during training and operation.	AI is a key driver of data search, including synthetic data.
Social Media & Streaming	Platforms like TikTok, YouTube, and Instagram generate billions of daily uploads, especially **high-resolution videos** ($4K$ and $8K$), which account for the largest share of internet traffic.	Videos account for more than half of internet traffic.
Internet of Things	Sensors and connected devices in industry and households generate data streams in real-time.	Their growing use has been cited as a major driver of growth.
Cloud Solutions	Moving to the cloud removes storage limitations, allowing organizations to store **more data for longer periods** and support real-time analytics.	The cloud helps create and retain larger volumes of data.

One of the clearest indicators of the source of data is its structure. The majority of the data generated today comes from non-traditional sources (such as social media and video), leading to the predominance of the unstructured format:

Data Format	Share of Total Volume	Growth Rate
Unstructured Data	**80% to 80% of the total volume**	They grow approximately **three times faster** than structured data.
Structured Data	About 20% of the total volume	Include traditional data from databases and transactional systems.

Unstructured data (emails, text messages, blog posts, videos, audio, etc.) is the largest and fastest-growing segment, directly reflecting the dominant contribution of consumer and media sources.

However, the core dilemma is not the volume, but the utility of this torrent. Simply having more data does not automatically translate into wisdom or efficiency. More data does not mean more knowledge. More data does not mean that the business and the sociaty understand our business better. More data does not mean that people are better in our everyday lives.

While the creative segment of humanity needs more data to develop, invent, and understand the world, a crucial question arises for the rest: are these vast quantities of data useful, or do they produce more chaos in our business and everyday life, rather than produce something useful? The immediate challenge is not to stop generation, but how to handle this one that we have now.

The Data Storage Paradox: Capacity vs. Processing

In terms of raw data storage, the industry has achieved remarkable success. A huge progress has been made in the field of storage of ever-increasing volume-the powerful big data technologies and even the advent of quantum computing for data processing. The technology world is successful - powerful technologies how to store a large volume of data are here. Today, what was once a scary amount years ago (terabytes) is now commonplace.

Yet, this success in storage has created a new, deeper challenge in processing and insight extraction**.** Storing the data is only half the battle. The true difficulty lies in making the data meaningful and valuable. This is captured perfectly with the metaphor of a library: It's like having a library without

bookshelves. So you have just an empty room that physically you can store your books, so your data on top of each other. But if you need to find something, it costs you a lot of efforts and a lot of time to get it.

The storage problem vs. usage problem

The future challenge, therefore, shifts from being a storage problem to a usage problem. A new, powerful ideas on how to process and learn from the data is needed. This need has spurred innovative ideas in data science and computer engineering, such as exploring DNA storage [4] and hologram storage [7], essentially stealing some ideas from the nature to solve social and business cases.

The most innovative research and trends in the field of Big Data analysis currently revolve around the deeper integration of Artificial Intelligence (AI), new data architectures, and a focus on real-time processing and ethical use.

Here are the most significant innovative research areas:

1. Advanced AI and Machine Learning Techniques

The use of AI and ML has moved beyond simple predictive models to highly sophisticated, autonomous systems:

- Augmented Analytics [8] [9]: This involves using AI and Machine Learning to automate and enhance the entire data analysis workflow. It helps non-data-scientists generate insights, find patterns, and explain findings in natural language, drastically improving the speed of decision-making.
- Deep Learning (DL): Research continues to push the boundaries of neural networks, particularly in:
    - Natural Language Processing (NLP): Advancements like Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) are transforming how unstructured text data (customer feedback, documents, etc.) is understood, summarized, and queried for analysis.[6]
    - Computer Vision: Using Deep Learning models like Convolutional Neural Networks (CNNs) to extract valuable insights from image and video data at scale (e.g., medical imaging diagnostics, monitoring in manufacturing).
- Agentic AI: This is an emerging and highly innovative area where AI systems can perform entire workflows autonomously—setting goals, planning tasks, executing actions, and adapting to feedback without continuous human intervention.

Evolving Data Architectures and Real-Time Processing

Innovative research is focused on new data processing models to handle data volume, velocity, and variety more effectively [10] [11]:

- Data Mesh: A shift from centralized data warehouses/lakes to a decentralized architecture. It treats data as a product, with domain-specific teams owning and serving their data assets. This architecture is designed for massive scale and speed, improving data accessibility and quality across the organization.
- Data Fabric: This is a unified, intelligent platform that uses AI and machine learning to integrate and manage data from multiple disparate sources (on-premises, cloud, edge) without having to physically move all the data. It's focused on making data readily available and governed.
- Edge Analytics: With the proliferation of Internet of Things (IoT) devices, this trend focuses on performing analysis directly on the data source (the "edge") instead of sending everything to a central cloud. This is critical for real-time applications like predictive maintenance and autonomous vehicles, where latency must be near zero.

Ethical and Secure Data Usage

As Big Data becomes more influential, research is heavily invested in its responsible application [5]:

- Responsible AI and Ethics: A major research focus is on mitigating algorithmic bias, ensuring fairness, and developing mechanisms for transparency (explainability, or XAI) and accountability in AI/ML models that make critical decisions.[12]
- Data Observability: This involves applying DevOps principles to data pipelines to ensure data quality, freshness, and reliability. It uses automated monitoring and alerting to detect anomalies in data (schema changes, sudden volume drops) before they corrupt analysis or model training.[13]
- Privacy-Enhancing Technologies (PETs): Innovative research into techniques like Homomorphic Encryption and Federated Learning allows data to be analyzed or models to be trained while the underlying data remains encrypted or localized, respectively. This addresses increasingly stringent data privacy regulations.

The Ethical Crossroads: Privacy and Ownership

The Privacy Challenge: Technology vs. Behavior

The conversation naturally transitioned to data privacy, revealing one of the most significant philosophical hurdles of the data age. The panel agreed that privacy is crucial for both businesses (which expend great effort to anonymize and secure data) and individuals, whose personal data is constantly exposed and interconnected with professional and daily lives.

However, the problem is not a lack of protective technology. The technological solutions for protection are numerous and "most of these ideas are successful". The true crisis lies in the fact that social behavior actively undermines legal protection.

The laws and all these compliance documents, they try to protect our data, but our social behavior actually exposes this data without any boundaries.

The regulation system is often too slow and unable to keep up with the fast-changing ways humans behave online. Users often expose sensitive data (like pictures or personal identifiers) on platforms without considering the consequences, sometimes becoming victims to be followed or to be targeted.

The implication is profound: the privacy problem is fundamentally a **non-technological problem**. It is a challenge related to ethics, behavior, and social consciousness. The solution, therefore, requires education and critical thinking, rather than merely passing more laws.

The Ownership Enigma

The privacy dilemma is inseparable from the question of data ownership. The panel raised the perplexing question: when a user interacts with a technology that creates data, who is the true owner—the individual or the company that created the technology?

The example of **IoT (Internet of Things)** devices illustrates this paradox vividly. While a knowledgeable user might be able to change this pattern and transfer the data to a personal server, this highlights a deeply confusing situation where one possesses something but may not be able to fully own or protect it.

In most cases, the user (you) is considered the "data subject" and retains the rights to control and access your personal data, but the company that collects, processes, and stores the data is generally considered the legal "data controller" or "data holder."

1. *Data from IoT Devices (Smart Homes, Wearables, etc.)*

   For data generated by an Internet of Things (IoT) device, ownership is usually determined by a combination of the device's Terms of Service (ToS), the type of data, and legal regulations like GDPR.

- Personal Data (Data Subject Rights): For data that is personally identifiable (like your name, location history, or health data from a fitness tracker), you (the individual user) have significant rights to access, correct, delete, and port that data, especially under comprehensive laws like the EU's General Data Protection Regulation (GDPR).[1] This is often described as "ownership as control" rather than traditional property ownership.
- Non-Personal/Aggregated Data (Company Control): The manufacturer, service provider, or data processing platform owner often claims ownership or an exclusive right to use the non-personal data (or personal data that has been anonymized or aggregated) that is collected. For example, a smart thermostat company may own the aggregated data on average temperature settings to improve its product or sell general market insights.
- Contractual Clauses: The fine print in the Terms of Service you agree to when setting up the device usually grants the company a broad license to use and monetize the collected data.

  New Legislation: Laws like the EU Data Act [6] are attempting to give users of connected products (like IoT devices) greater control and access over the data they generate, even for non-personal data.

2. Data from Your Mobile Phone

The data generated by your mobile phone is handled by several different entities, none of whom typically have exclusive ownership over all of it.

- You (The User/Data Subject): You own the content you create (photos, contacts, texts). More importantly, you have privacy rights over your personal data.[3] You are the one who grants or revokes permissions for apps to access things like your location, microphone, and camera.[4]
- App Developers (Third Parties):Apps (like social media, games, or utilities) collect data based on the permissions you grant and the privacy policies you agree to. They typically own the database they create from your usage patterns and may use or sell this data for targeted advertising.
- Mobile Carrier/Phone Provider (ISP):Your carrier (e.g., T-Mobile, Verizon) can see information about your data consumption, general location (via cell tower connection), the metadata of your calls and texts (who you called, when, and for how long), and sometimes the websites you visit (if you're using their cellular network without a VPN).[5] They use this for billing and may also use it for marketing purposes, but in many jurisdictions, they are restricted by law from sharing the content of your communications.
- Operating System Providers (Google/Apple):The company that provides your phone's operating system (Android or iOS) collects extensive data on your usage, app activity, and location to provide services, personalize your experience, and target ads. You typically agree to this when setting up the device.

  The rapid evolution of technology—specifically cloud solutions and blockchain initiatives—has radically altered the traditional concept of ownership. Because users are often "unable to store them, we are unable to protect them [data], and actually... Or to track what is happening with this data," the opportunity to own one's own things is lost. This is amplified by the fact that owning one's own data has become "very expensive".

Data ownership is often replaced by a structure of complex rights and controls:[6]

- Customers have rights over your personal data (especially in regions with strong privacy laws).[7]
- Companies have contractual rights to use and monetize the data they collect and process.[8]

The challenge to regulatory bodies is significant. Technology has created a world with a new reality, and lawyers and data engineers are not able to respond to this very high speed. Ownership and privacy are intertwined; the absence of one compromises the other. The resolution, according to the discussion,

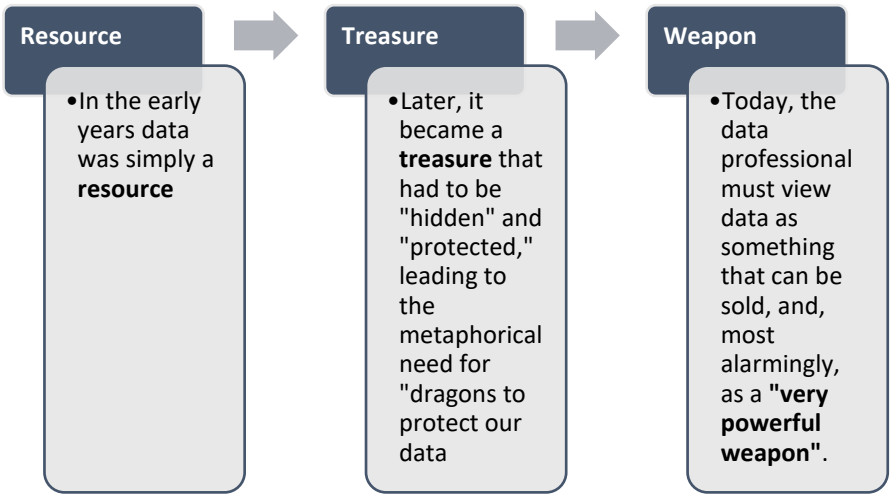
might be defined by one thing: the possibility to choose. Ownership is defined by the freedom to choose whether to share data or to keep it protected.

Data's Transformation: Resource, Treasure, Weapon

Monetization and Risk

The conceptual value of data has undergone a startling transformation, driven by its monetization. The phrase "data is the new oil" encapsulates its immense business importance. In 2006, Clive Humby, a British mathematician and data science entrepreneur, introduced the phrase “Data is the New Oil”. In simple words, what he said is that - “Data is the new oil. It is a valuable source. However unrefined data is not useful. It must be converted into valuable entities that drive profit. Companies now actively monetize their data, leading to the concern that they may be selling not just their own data, but *our* data.

And the data identity evolves from Resource to Weapon



This final identity—data as a weapon—necessitates the invention of more agile technology that can adapt to rapidly changing situations, as today's reality is often quite different from tomorrow's. This volatility and risk make data a gold mine for some, but a potential nightmare when its power is unleashed.

The Need for New Definitions

The current, traditional definition of data is woefully inadequate for the modern age. The complexity introduced by privacy, ownership, monetization, and weaponization demands a conceptual shift.

The current definition of data is not enough. The data is not only bytes, strings. It's something more and to the definition more things should be added so to be able to define the whole thing the whole technology environment creates and chalanges.

The signed professionals in computer science fields should convene to provide another, more conceptual definition of data. Data is no longer merely the raw zeros and ones of information theory [12][13]. It is intrinsically linked to behavior, business, and social dynamics. Redefining the term is crucial for the technology world to be able to adapt and create solutions that cover all these new aspects, particularly concerning storage, privacy, and ownership management.

The AI Game Changer

The Data Vampire and Synthetic Reality

Artificial Intelligence (AI) as the ultimate game changer and a central driver of the data apocalypse or utopia. AI was starkly defined as a data vampire. AI technologies consume data in such volumes that actually AI is one of the reasons that increased the need of data, AI is the reason this noble, this data generation to be unstoppable because AI will continue to develop when there are more data.

AI creates a dilemma regarding the data it consumes—is it real or synthetic? This question of authenticity mirrors dilemmas in other spheres of life (e.g., food or clothing) but holds unique power in the digital realm.

The need for AI to generate synthetic data is itself a paradox. According to Gartner the forecast is that 60% of data used for AI will be synthetic by 2024, and 75% of businesses will use generated customer data by 2026 [14]. It arises because either not enough data is produced by real-life systems, or because the existing real data is not accessible—it is hidden due to privacy concerns, or the owner refuses to provide it. Thus, the privacy barriers that society seeks to establish are simultaneously driving AI to create an artificial reality.

The dilemma of real versus synthetic data ultimately comes down to data reliability. Now, with the introduction of AI, the responsibility for quality is shared with a non-human entity, forcing the industry to adapt its entire framework for verifying information.

Hallucination and Critical Thinking

The output of AI is not guaranteed to be reliable - the phenomenon of AI hallucination, a scenario where a machine learning model, if "not learn[ed]... quite enough," becomes prone to error and bias. This risk is amplified by our own incomplete understanding of the parameters driving these systems.

In the new environment, the most vital human skill is **critical thinking** - It's not to know everything, but to ask yourself, is it a valid one, how I can verify that. Without critical thinking, AI will produce a fake reality, unreliable decisions. While technologies are needed to produce the knowledge required to distinguish between real and synthetic data, the human element—curiosity and the willingness to debug the source like code—is essential. The responsibility to detect and check for the synthetic, fake, or biased content—whether in the news, static data, or AI-generated responses—falls upon the users. The future of data with AI is bright only if customers remain critical thinking people. If society chooses merely to consume the results without questioning, the future is not so bright.

The Evolving Role of the Data Professional

Beyond Technology

The complexities of the new data reality have fundamentally altered the role and responsibilities of the data professional—the data engineer, data scientist, or database developer. They are no longer just custodians of technology; they must become strategists balancing ethics and efficiency. The new role of data professionals is defined from the observation that often we have to solve problems with technologies, but actually the problem is not a technology problem.

This requires professionals to adopt their cells because their responsibilities have changed. It is no longer enough to be a "technology person". The modern data professional must consider and find scalable solutions for non-technical aspects like privacy, ownership, and ever-changing social behavior. System architecture decisions, such as choosing between cloud and on-premise solutions, are now driven not just by functionality, but by the need to satisfy the ownership and privacy requirements of the data.

The professional development path has become very interesting because it demands curiosity about many things outside the purely technological domain. Designing systems requires considering the entire spectrum of human behavior and business realities, rather than falling back on traditional ways of using data.

Aligning Technology and Society

The complexity is further compounded by the fundamental shift in the responsibility for data quality. In the past, data quality was largely the responsibility of the data engineers—they were often guilty if the data was unreliable in BI or data warehouse projects.

Today, the game has changed with the introduction of AI as a new player. Now, a technology created by people—AI—will often be responsible for the source used in knowledge extraction systems. This

creates a disorienting feedback loop where it is difficult to distinguish where something ends and something begins in our play.

The new engagement for data professionals is to bridge the gap between technology and society, to align the two technologies and how people think about it and how they use it. A new responsabilty is here to teach them how to consume it correctly. Data professionals must act as ambassadors, explaining how systems work so that the public is aware of "possible traps or misleading". The task is not merely to implement, but to teach and guide consumption.

The Mandate for Adaptation and Redefinition

The future's outcome is not set in stone; it is contingent upon human effort and willingness to evolve.

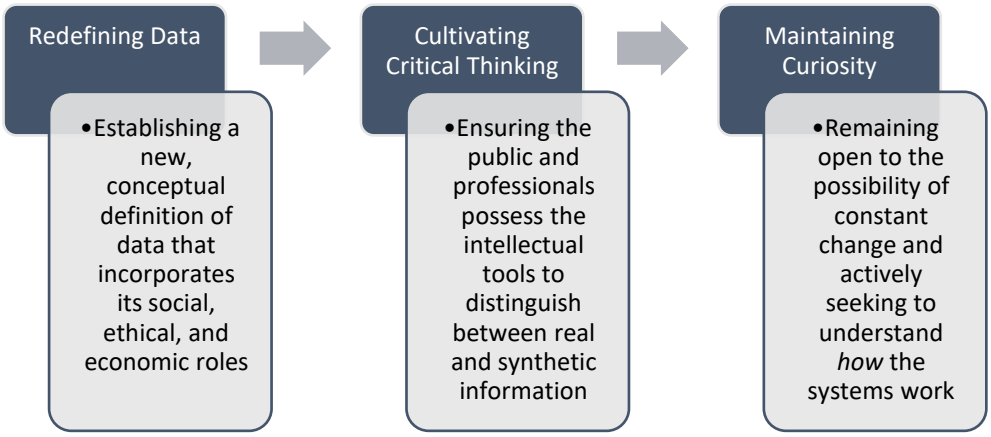
The Necessity of New Problem Solving

Once believed that all technologies in data processing and management solve problems, the current reality suggests the opposite. Nowadays all these technology actually... are creating problems and not always technology problems. These new problems are complex, non-traditional challenges that demand curiosity, critical thinking, and a willingness to adapt. The humorous observation was made that perhaps the industry is bored already and thus uses technology to produce new problems because there don't have enough problems" Yet, beneath the humor is the serious truth that the industry must accept these challenges as its new reality.

The Path to Control

To retain the power of choice—to steer away from dystopia—requires a fundamental commitment to growth and an open mindset.

The path involves three key mandates:



Conclusion: The Choice is Ours

The future of data is not predetermined. It is not an inevitable utopia, nor is it a guaranteed dystopia. The power to shape this future resides entirely with humanity. The journey ahead requires a radical acceptance of the complexity introduced by the digital age. Professionals must embrace a role that transcends pure technology, becoming ethical strategists and educators. Society must accept the necessity of developing its collective critical thinking to combat the risks of synthetic reality.

The ultimate message is one of responsibility: must stay open, remain curious, and never forget that technology often races ahead of our ability to adapt. By actively engaging with the problems our technology creates, by redefining our core concepts, and by choosing to think critically, humanity can ensure that the choice remains **ours**.

References

1. Taylor P. , Amount of data created, consumed, and stored 2010-2023, with forecasts to 2028, Nov 2025, Statista, https://www.statista.com/statistics/871513/worldwide-data-created/
2. Gkikas D., Prokopis T, AI in Consumer Behavior, Advances in Artificial Intelligence-based Technologies (pp.147-176), 2022
3. Bartley K., Big data statistics: How much data is there in the world?, 2025, https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-world/#:~:text=In%202024%2C%20the%20global%20volume,by%20the%20end%20of%202025.
4. Yiming, Dong & Sun, Fajia & Ping, Zhi & Ouyang, Qi & Qian, Long. (2020). DNA storage: Research landscape and future prospects. National Science Review. 7. 1092-1107. 10.1093/nsr/nwaa007. https://www.researchgate.net/publication/345359466_DNA_storage_Research_landscape_and_future_prospects#:~:text=Abstract%20and%20Figures,and%20sustainable%20data%20storage%20solution.
5. Opeyemi, Kehinde & Samuel, Adebis. (2023). Big Data Analytics: Emerging Trends, Technologies, and Innovations for Future., https://www.researchgate.net/publication/390630564_Big_Data_Analytics_Emerging_Trends_Technologies_and_Innovations_for_Future#:~:text=areas%20like%20drug%20discovery%2C%20financial,more%20accessible%20to%20non%2Dexperts.
6. Data Act explained [2023], https://eur-lex.europa.eu/eli/reg/2023/2854/oj, https://digital-strategy.ec.europa.eu/en/factpages/data-act-explained#:~:text=A%20data%20holder%20must%20have,connected%20product%20or%20related%20service.
7. Lin, Xiao & Liu, Jinpeng & Hao, Jianying & Wang, Kun & Yuanying, Zhang & Li, Hui & Horimai, Hideyoshi & Tan, Xiaodi. (2020). Collinear holographic data storage technologies. Opto-Electronic Advances. 3. 19000401-19000408. 10.29026/oea.2020.190004.
8. Minu, M S & Ahmad, Zoya. (2020). Augmented Analytics: The Future of Business Intelligence. 5. 7-13. 10.5281/zenodo.3757837.
9. Alghamdi, Noorah & Al-Baity, Heyam. (2022). Augmented Analytics Driven by AI: A Digital Transformation beyond Business Intelligence. Sensors. 22. 8071. 10.3390/s22208071.
10. Lu, Jiaheng & Holubova, Irena & Cautis, Bogdan. (2018). Multi-model Databases and Tightly Integrated Polystores: Current Practices, Comparisons, and Open Challenges. 2301-2302. 10.1145/3269206.3274269.
11. Yasuko Matsubara et al, MicroAdapt: Self-Evolutionary Dynamic Modeling Algorithms for Time-evolving Data Streams, Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (2025). DOI: 10.1145/3711896.3737048
12. Elmasri, R., & Navathe, S. B. (2016). Fundamentals of Database Systems (7th ed.). Pearson Education.
13. Floridi, L. (2010). Information: A Very Short Introduction. Oxford University Press.
14. Gartner Identifies Top Trends Shaping the Future of Data Science and Machine Learning [2023] https://www.gartner.com/en/newsroom/press-releases/2023-08-01-gartner-identifies-top-trends-shaping-future-of-data-science-and-machine-learning