Big Data

King Hussein School of Computing Sciences

Computer Science Department

CS11449 – Computer and Society

Table of Contents

1. What is Big Data?

“Big data refers to extremely large and diverse collections of structured, unstructured, and semi-structured data that continue to grow exponentially over time. These datasets are so huge and complex in volume, velocity, and variety, that traditional data management systems cannot store, process, and analyze them.

The amount and availability of data is growing rapidly, spurred on by digital technology advancements, such as connectivity, mobility, the Internet of Things (IoT), and Artificial Intelligence (AI). As data continues to expand and proliferate, new big data tools are emerging to help companies collect, process, and analyze data at the speed needed to gain the most value from it.

Big data describes large and diverse datasets that are huge in volume and rapidly grow over time. Big data is used in machine learning, predictive modeling, and other advanced analytics to solve business problems and make informed decisions [1].

Three dimensions of big data are called the 3Vs: volume, velocity, and variety. These three basic data characteristics are essential for handling big datasets.

1. **Volume:** The overwhelming quantity of data created or gathered. Organizations today deal with huge volumes of data, often measured in terabytes, petabytes, or even exabytes, due to the development of digital devices and sensors.

2. **Velocity:** Refers to the speed at which data is generated, collected, and processed. An unprecedented amount of data is coming in from sensors, social media, and transactional systems, among other sources. Deriving insights and making timely decisions require real-time or near-real-time processing.

3. **Variety:** Discusses the range of sources and forms of data. Structured formats such as databases, semi-structured formats like XML or JSON, and unstructured formats like text documents, photos, videos, and social media posts are examples of different types of data. Keeping track of and evaluating this wide variety of data calls for.

Apart from the 3Vs, there are more Vs that are thought to offer a more thorough grasp of big data concerns. These extra Vs consist of:

4. **Veracity:** Discusses the data's dependability and credibility. Making sure that data is accurate and of high quality becomes increasingly important as data volume and variety increase. Inadequate data quality can result in inaccurate conclusions and choices.

5. **Value:** Discusses the importance or utility of the conclusions drawn from the data. The ultimate objective of big data analysis is to derive useful insights that can influence corporate choices, provide value, and enhance results.

6. **Variability:** The term describes the irregularity or turbulence in the data flow. Data processing and analysis can face difficulties due to the great variability of data streams in terms of volume, velocity, and diversity [2].

2. Usability of Big Data

1. **Big data in healthcare:** Big data in healthcare is essential for boosting medical research, maximizing resource allocation, and improving patient care. Healthcare professionals can find patterns and correlations in patient data through analysis, leading to more precise diagnoses and individualized treatment regimens. Big data analytics can also be used to better deliver healthcare, track disease outbreaks, and identify high-risk individuals.

2. **Big data in retail:** Big data offers vendors insightful information about the preferences and behavior of their customers. Retailers can target particular client categories, improve inventory management, and customize the shopping experience by evaluating customer data. Demand forecasting is made easier by big data analytics, which can also enable dynamic pricing strategies and increase supply chain efficiency.

3. **Big data in finance:** The finance industry heavily relies on big data analytics to assess risk, detect fraudulent transactions, and optimize investment strategies. By analyzing vast amounts of financial data, banks, and financial institutions can identify patterns that indicate potential fraud or money laundering activities. Big data in finance enables more accurate credit scoring, enhances fraud detection, and improves investment decision-making [3].

4. **Big data in AI and ML:** The technologies of artificial intelligence (AI) and machine learning (ML), together with the enormous field of big data, mark a significant advancement in computer power. The convergence of AI, ML, and big data signals a paradigm shift in how we approach problem-solving and decision-making across industries; it goes beyond simple technological improvement. By utilizing the enormous amounts of data that are produced every day, AI and ML algorithms can acquire previously unheard-of insights that empower companies to make more strategic and well-informed decisions. Furthermore, machine learning's iterative process, which is driven by constant data inflows, enables adaptive and self-improving systems that gradually improve accuracy and efficiency. As a result, the combination of big data, machine learning, and artificial intelligence transforms processes and spurs innovation, setting the stage for a time when intelligent systems will propel global progress and wealth.

3. Problem Statement

The widespread use of big data in the current digital era has brought both previously unheard-of opportunities and challenges to a wide range of industries. Organizations that use large datasets to spur innovation and obtain a competitive edge face a variety of legal and ethical challenges that require careful consideration and strategic management. Big data and legal and ethical issues combine to create complex quandaries that require careful examination to avoid potential pitfalls and promote responsible use.

Rapid advances in data collecting, storage, and analysis have overtaken the creation of strong legal frameworks, creating gaps and ambiguities that increase the likelihood of abuse and rights infringement. In addition, the sheer amount, speed, and diversity of data produced by many sources provide significant obstacles to maintaining adherence to current rules, such as privacy statutes and data protection laws, which were developed in a technologically distinct environment.

Concurrently, big data's ethical ramifications are a major worry, bringing up serious issues with responsibility, justice, and transparency in its application. The ingrained prejudices present in algorithms and datasets intensify societal inequalities and sustain discrimination, leading to significant consequences for both individuals and communities. Moreover, the lack of transparency in data collection methods and algorithmic decision-making procedures erodes confidence and intensifies concerns about privacy violations and loss of autonomy.

As a result, tackling the ethical and legal aspects of big data requires a multimodal strategy that incorporates stakeholder involvement, ethical frameworks, and legal compliance. To reduce possible harm and build stakeholder trust, organizations must proactively evaluate the risks connected to big data initiatives, put strong governance systems in place, and develop an ethically responsible culture. To create logical regulatory frameworks that combine promoting innovation with defending individual rights and societal values, legislators must also interact with industry stakeholders.

Given these difficulties, it is essential to conduct a thorough investigation of the ethical and legal concerns related to big data to support ethical best practices, guide decision-making processes, and preserve the values of accountability, openness, and fairness in the digital age. Stakeholders can navigate the rapidly changing big data world with integrity and vigilance by examining the complex interactions between technology, law, and ethics. This will help to ensure that the potential of big data to drive progress is used responsibly and fairly, benefiting society.

4. Advantages of Big Data

1. **Data-Driven Decision Making:** Big data enables organizations to make more informed decisions based on data analysis. An example is trend analysis, where companies analyze at what time of year their customers tend to purchase a certain product. This can prove very helpful when it comes to maximizing their profit.
2. **Improved Efficiency:** Big data can increase operational efficiency by identifying patterns and optimizing processes.
3. **Innovation and Research:** Big data can facilitate research and analysis through the analysis of large datasets.
4. **Fraud Detection:** Big data is essential for identifying fraud in a variety of sectors. Financial firms, for example, can instantly examine enormous amounts of transaction data to spot unusual or suspect trends that can point to fraud.
5. **Increases Productivity:** Big data solutions can help IT professionals work more productively. Big data solutions can automate the process of sorting through many sorts of data from different sources, freeing up personnel to concentrate on other important duties.

5. Disadvantages of Big Data

1. **Security Risks:** Adoption of new technology carries some risk, which makes sense as it becomes available in business. Businesses that use big data solutions are more vulnerable to cybersecurity risks since these solutions are popular targets for cybercriminals.
2. **Talent Gaps:** Since big data is a new field, there aren't enough IT professionals in the field to handle big data duties. A corporation can profit from having access to big data, but only if someone with a solid big data background works with it.
3. **Privacy Concerns:** Big data often involves the collection of vast amounts of personal information.
4. **Data Ownership and Control:** Determining who owns and controls the data can be ambiguous.

6. Principles of Big Data Ethics

Over the past ten years, a lot of studies have been conducted on big data ethics as academics and industry executives try to address public criticism of the use of big data. The definition of big data ethics is the study and advocacy of principles related to appropriate and inappropriate data use practices, with a focus on personal data. The goal of big data ethics is to establish a moral and ethical standard for using data.
The possibility for unethical use of data is outlined in the following major areas of concern for big data ethics:

1. **Informed Consent and Transparency:** Consent is the act of granting someone else your free will to do something to you. The most cautious, polite, and moral type of consent is informed consent. Giving participants a fair and accurate knowledge of how their data will be utilized necessitates a major effort on the part of the data collector. In the past, participation in a single study usually required informed consent for data collection. This kind of permission is rendered impossible by big data as the whole goal of big data mining, analytics, and studies is to find previously unthinkable patterns and trends among data points. Because neither the data collector nor the study participant may reasonably be expected to know or understand what will be gleaned from the data or how it will be utilized, permission cannot possibly be "informed." There have been changes made to the informed consent standard. The first permits secondary uses of data in advance and is referred to as "broad consent." The second method is called "tiered consent," which permits certain secondary uses of data, like cancer research, but not genetic research. Some contend that the idea of consent is being watered down by these more recent versions, leaving users vulnerable to unethical behavior. When information about possibly "unwilling" or ignorant data subjects is taken via social media platforms, more problems occur. Contracts for social media services frequently grant permission for the gathering, combining, and analysis of this kind of data. Ofcom did discover that, on average, 65% of internet users accept terms and conditions without reading them. Therefore, it makes sense to presume that a large number of end users could not be aware of the complete scope of data usage, which is increasingly being used for purposes other than digital advertising and social scientific research.

2. **Privacy Concerns:** Numerous notions, including liberty, autonomy, security, and, in a more contemporary meaning, data exposure and protection, are all part of the ethics of privacy. By dividing the concept of large data privacy into three areas, you can comprehend it:

    - The condition of privacy.
    - The right to privacy.
    - The loss of privacy and invasion.

    Because many existing privacy mechanisms are unable to protect sensitive data, the magnitude and velocity of big data present a severe risk. This has resulted in an exponential rise in cybercrime and data leaks.

    January 2021 was one instance of a big data leak that compromised the privacy of more than 200 million internet users. A growing Chinese social networking platform named Sociallarks experienced a breach because of several data security mistakes.

    Big data's increasing analytical capability raises additional concerns about how privacy may be impacted when personal data from several digital platforms is mined to build a complete picture of an individual without that person's express consent. For instance, a

person's digital footprint can be used to determine their political preferences, sexual orientation, social life, and other details when they apply for a job. Even though the applicant did not provide the material for review, any of this information could be used to deny their application for a job.

3. **Ownership:** Big data terminology avoids using the traditional or legal definition of ownership, which is the sole right to use, possess, and dispose of property. Instead, ownership in this sense relates to the capacity to profit from data advances, redistribute data, and modify data. Legislators have already decided that since data is neither property nor a commodity, it cannot be stolen. Despite this, internet users and customers who give useful information to businesses without profiting from it are not well-protected or compensated.

   Two categories can be used to separate data ownership:

   - The right to control data: edit, manage, share, and delete data.
   - The right to benefit from data: profit from the use or sale of data.

   Those who create data, such as Facebook users, do not inherently own the data, despite popular notions. Some even contend that using "free" internet services requires us to pay for the platform with the data we supply. But in today's environment, big data equals big money. When it comes to data ownership and the openness of the businesses that utilize and benefit from the data we share, a lot of internet users believe that the existing playing field is stacked against them.

4. **Bias and Discrimination:** Humans create algorithms, choose and prepare the data sets they analyze, and are biased beings. Thus far, a substantial body of research indicates that human biases are permeating technology and algorithms, adversely affecting people's lives and liberties. especially those who are members of our nation's minorities. The term "coded bias" has been linked to several well-known incidents, including the discovery of racial skin-type bias in commercial artificial intelligence systems developed by large US corporations by MIT lab researcher Joy Buolamwini. Buolamwini discovered that the software was trained on datasets including over 83% images of white people and 77% images of men. Because of these skewed datasets, the program's error rate in identifying white male faces is only 0.8%, while its mistake rate in identifying dark-skinned female faces is 20% in one case and 34% in the other two. These prejudices go across racial and gender boundaries and affect criminal profiling, housing, and poverty. Algorithm biases have been shown to affect our individual psyches and cognitive processes as they have permeated every aspect of daily life. When we believe that what we see online is a reflection of our world, a phenomenon happens. But what we see is frequently a customized reality that algorithms have generated based on our past viewing preferences. The algorithm filters out stuff that isn't likely to be enjoyable or agreeable for us and presents only that. Such filter bubbles

produce echo chambers and, in severe situations, can result in social exclusion, radicalization, and sectarianism.

5. **Data Quality and Accuracy:** In our real world, the data can be flawed. As the data can be incomplete and undermine our insights validity, produced by data analysis, this leads to flawed decisions with harmful consequences in most cases for our society. Ethical practices stress the necessity of multiple validation processes for the quality of the data gathered. Transparent sourcing is also a need when it comes to identifying limitations for the data. This is why the accuracy and reliability of the data play a huge role in the credibility of the data analytics and their future insights, furthermore, this affects Misinterpretation. Zoldan [6] underscores the issue that the utilization of Big Data for decision-making is often fraught with inaccuracies, and context gaps.

6. **Big Data Divide:** The term "big data divide" refers to the current situation in which access to and knowledge of large amounts of data are restricted to a small number of powerful organizations. These differences separate individuals who lack the financial, educational, and technological means to access and analyze large datasets, resulting in "haves" and "have nots" in the big data world. According to Tim Berners-Lee, people are cut off from data that could be extremely beneficial to their health because of the "big data divide." And while the market for apps that use data to improve our lives, health, finances, etc., is expanding, there is still no method for people to mine their data or link possible data silos that are overlooked by commercial software. Once more, the ethical conundrum of who owns the data we produce faces us; if it is not ours to alter, examine, and profit from according to our conditions, then we do not own it. When we consider algorithmic biases that classify people based on a compilation of data that they are unable to see, the data divide becomes even more problematic. For instance, profile software can identify someone as a high-risk candidate for criminal conduct, making it lawful for authorities to stop and search them or even to deny them accommodation in specific locations. Because of the "big data divide," people who are "data poor" are unable to comprehend the information or processes that were used to make choices about them and their lives [4].

7. **Security and Data Breaches:** The security of Big Data systems is paramount for safeguarding sensitive information and preventing data breaches. While all the steps to ensure one’s data privacy are taken, the issue of cyber threats still arises. Malicious actors are only imposing even more risks to the data we have in hand. The breach consequences vary as they may result in reputational damage, assets loss, or even more issues for the data owner. When implementing the concept of big data, various security measures should be taken to keep data integrity. For example: A company must ensure incident response plans to decrease the impacts of possible data breaches and keep their systems trustworthy.

8. **Social Implications and Power Dynamics:** Big data technology is being adopted worldwide and it also has been spreading widely in our modern-day technology world. This led to leaving a change by doing the following:

    1- Reaching social implications.

    2- Shaping power dynamics.

    3- Influencing decision-making.

    4- Redefining societal norms.

    To be able to gather such enormous datasets, analyze them, and interpret them in terms of algorithms and such leaves a huge amount of power in the hands of those who work on it. This raises the concern of:

    1. Data monopolies.
    2. Algorithmic governance.
    3. Erosion of privacy and individual autonomy.

    With these concerns in hand, there must be precautious protocols taken on the stakeholders of the data, there must be interdisciplinary collaboration, engagement, and ethical principles enforced and committed to. This leads to further prioritizing the public goods and values as well as privacy concerns.

    When we foster the decision-making process and monitor it to ensure its ethical compliance, we must promote data literacy, and control the concentration of power by decreasing the hands who own it. We therefore ensure big data potentials are harnessed yet also safeguarded against pitfalls.

7. General Data Protection Laws

- **CCPA and GDPR**

The emergence of big data has resulted in significant changes to how businesses gather, handle, and apply enormous volumes of data. However, this progress also raises several ethical and legal questions, especially concerning data protection and privacy. The General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) are two important laws that address these issues.

One of the most extensive data privacy regulations in the US is the California Consumer Privacy Act (CCPA), which was passed in 2018. Its main goal is to defend Californians' right to privacy and their rights as consumers. The California Consumer Privacy Act (CCPA) gives customers several rights, such as the ability to view and seek the deletion of their personal information, the ability to opt out of having their data sold, and the right to know what personal information is being collected. Moreover, it places responsibilities on companies concerning security protocols, consumer request fulfillment, and data practices that are transparent.

On the other hand, the European Union (EU) General Data Protection Regulation (GDPR) went into force in 2018. It was intended to give people more control over their data and to standardize data privacy rules throughout Europe. Any entity processing the personal data of EU persons is required to comply with the GDPR extraterritorially, regardless of its location. A few of the GDPR's most important requirements are getting express consent before processing data, putting data protection mechanisms in place, evaluating the effects of data protection, and designating data protection officers for specific companies.

The goals of the CCPA and GDPR are similar in that they both seek to protect people's right to privacy and make businesses liable for their data management practices. In the digital environment, they seek to advance accountability, transparency, and trust. Nonetheless, there are some significant distinctions between the two regulations in terms of their respective standards, enforcement methods, and scope.

The GDPR imposes stricter requirements for data protection throughout the EU and beyond, whereas the CCPA primarily focuses on improving consumer rights within the state of California. Depending on the seriousness of the infraction, non-compliance with these requirements may result in significant penalties, including fines of millions of dollars or euros.

Finally, it should be noted that the CCPA and GDPR are important turning points in the international endeavor to resolve the moral and legal issues raised by big data. In an increasingly data-driven world, corporations can show their dedication to upholding individuals' privacy rights while also reducing legal risks by adhering to these standards. However, navigating the intricate convergence of big data, privacy, and ethics still requires constant attention to detail and flexibility in response to changing legal environments.[5], [6].

- **Jordan PDPL (Personal Data Protection Law)**

The Jordan PDPL (Personal Data Protection Law), also known as Law No. 24 of 2023, is a major step toward protecting people's data in Jordan. This law, which went into effect on March 17, 2024, was enacted on September 17, 2023, to regulate the gathering, handling, and safeguarding of personal data while upholding the rights of individuals to privacy.

Law's Scope: Any data or information about a natural person who may be identified is covered by the Jordan PDPL, including sensitive personal data and criminal histories. Except for personal processing, it encompasses data processing operations carried out both before and after the law's passage.

Lawful Bases for Processing: Except in situations where special exceptions, such as those involving the public interest, medical treatment, or crime prevention, personal data processing requires the data subject's prior consent. Notably, neither the law nor third parties' legitimate interests are expressly permitted to be the basis for processing.

Principal attributes: The law places strict requirements on data processing, focusing on fairness, legality, transparency, and purpose limitation. Although it complies with international

standards, the concept of "data minimization" is not explicitly acknowledged. A variety of rights are granted to data subjects, such as the ability to access, correct, erase, and object to processing.

Protection Officer and Data Transfers: Unless there are specific circumstances, approval is required before data is sent to third parties. The law requires controllers to designate a data protection officer who will oversee adherence to data processing guidelines and efficiently manage complaints.

Penalties and Breach Notification: Controllers are required to take appropriate action to minimize harm and quickly notify the individuals who may have been impacted by a data breach. The consequences of breaking the law, which can include fines and license suspension, emphasize how crucial it is to follow the law.

Enforcement Mechanisms: A Personal Data Protection Board assists a specialized unit within the Ministry of Digital Economy and Entrepreneurship in monitoring compliance. This body ensures that the law is implemented effectively by approving standards, regulations, and licenses.

Future Implications: Companies doing business in Jordan will need to adjust to the terms of the PDPL by March 2025, which is the grace period. It is anticipated that additional laws would clarify licensing, consent processes, and disclosure requirements, highlighting the necessity of preventative compliance measures.

To sum up, the Jordan PDPL represents a significant advancement in the protection of people's privacy and data rights in Jordan, but its successful implementation will depend on strong compliance and enforcement procedures [7].

8. Conclusion

In the Big Data era, when information is the digital age's currency, companies that want to exploit data's potential while respecting people's rights and social values must navigate a complex legal and ethical landscape. The emergence of comprehensive data protection laws, like the California Consumer Privacy Act (CCPA), the General Data Protection Regulation (GDPR), and the Jordan Personal Data Protection Law (PDPL), as we have discussed throughout this presentation, highlights the global recognition of the need to regulate data practices in an increasingly interconnected world.

These legislative frameworks, which were put in place to protect personal data and lessen the dangers involved in processing it, offer businesses a vital platform on which to conduct themselves morally and responsibly. GDPR raises the bar for data protection internationally with its strict guidelines for gaining consent, notifying data breaches, and protecting data subjects' rights. In a similar vein, the CCPA represents a huge advancement in the field of digital privacy rights since it places a strong emphasis on giving customers control over their

personal information and enforcing penalties for non-compliance. The Jordan PDPL, which established extensive legislation to control the processing of personal data in Jordan, was recently enacted, demonstrating the growing global awareness of data privacy issues.

But while we weigh the effects of these legal frameworks, it's critical to recognize their limitations and the difficulties presented by the ever-changing technological landscape. By their very nature, laws are slow to adapt to new technological developments. They find it difficult to keep up with the rapidly developing powers of artificial intelligence, data analytics, and other emerging technologies. This disconnect between technology and legislation emphasizes how important it is for regulators to constantly innovate and adapt.

We believe that regulations surrounding big data ought to be more flexible, responsive, and future-focused, able to foresee the opportunities and problems that will arise from advances in technology. Regulatory organizations must maintain constant communication with academic institutions, industrial players, and civil society groups to guarantee that regulatory frameworks continue to be applicable and efficient and promote innovation.

Furthermore, although adhering to the law is necessary, moral action shouldn't be determined only by it. Regardless of legal requirements, businesses must act proactively to hold themselves responsible for moral data practices. Organizations can improve their long-term sustainability and competitiveness by cultivating a culture of transparency, accountability, and responsible data management. This can help them win over the trust and confidence of stakeholders such as investors and customers.

To sum up, the process of achieving moral and responsible data governance is a team effort that calls for cooperation from industry, government, and civil society. We can build a future where data-driven innovation coexists peacefully with privacy, security, and respect for human dignity by cooperating to close the gap between law and technology. It is up to us all to take advantage of this chance and open the door for a digital society that is more just, inclusive, and sustainable.

References

[1] "Big Data Defined: Examples and Benefits | Google Cloud." Accessed: Mar. 15, 2024. [Online]. Available: https://cloud.google.com/learn/what-is-big-data

[2] "6V's of Big Data - GeeksforGeeks." Accessed: Mar. 15, 2024. [Online]. Available: https://www.geeksforgeeks.org/5-vs-of-big-data/

[3] "Why Big Data is Important: Exploring Its Benefits and Uses | Institute of Data." Accessed: Mar. 15, 2024. [Online]. Available: https://www.institutedata.com/us/blog/why-big-data-is-important/

[4] B. D. Mittelstadt and L. Floridi, "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts," *Law, Governance and Technology Series*, vol. 29, pp. 445–480, 2016, doi: 10.1007/978-3-319-33525-4_19.

[5] "The general data protection regulation - Consilium." Accessed: Mar. 15, 2024. [Online]. Available: https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/

[6] "What is the CCPA (California Consumer Privacy Act)? | Cloudflare." Accessed: Mar. 15, 2024. [Online]. Available: https://www.cloudflare.com/learning/privacy/what-is-the-ccpa/

[7] "Jordan issues first personal data protection law: Clyde & Co." Accessed: Mar. 15, 2024. [Online]. Available: https://www.clydeco.com/en/insights/2023/10/jordan-issues-first-personal-data-protection-law