NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

CASE STUDY: DOING DATA SCIENCE– A FRAMEWORK AND CASE STUDY

MS DSP 485 - Data Governance, Ethics & Law

Sachin Sharma
August 11, 2024

Reference

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.2d83f7f5

The Data

In "Doing Data Science: A Framework and Case Study," the authors utilize data from various sources, including survey responses, administrative records, and publicly available datasets, to illustrate the data science framework they propose. They provide a clear explanation of how the data were collected, detailing the methodologies employed for gathering and analysing the data, such as exploratory data analysis and statistical modeling. The authors assess the adequacy of the data and analysis, concluding that while the data is robust for demonstrating their framework, limitations exist in the form of potential biases in self-reported survey data and gaps in the administrative records. This highlights the importance of paying close attention to data quality in data science projects. Overall, the data and analysis are deemed adequate for the framework presented, but the authors stress the importance of continuous improvement and validation in data collection and analysis practices.

Relevance to Data Governance

The issues highlighted in the case study align closely with the principles of data governance, particularly regarding ethical considerations and accountability in data practices. Given the rapid growth of data science, there should be legislation that addresses privacy, data security, and accountability for organizations handling sensitive data. The authors emphasize the necessity of a structured framework to guide data scientists in making ethical decisions throughout the data life cycle. This encompasses aspects of management, as organizations must establish policies and procedures to ensure compliance with ethical standards and regulations. Additionally, the ethical implications of data use, including privacy concerns and bias in data analysis, underline the need for governance mechanisms that promote transparency and accountability. By addressing these challenges, the case study illustrates the integral role of data governance in fostering responsible data practices and protecting the interests of stakeholders involved in data science initiatives.

Relevant Legislation

The authors do not specifically mention any existing legislation related to data science practices. However, they highlight the importance of adhering to ethical standards and best practices, suggesting that a robust regulatory framework governing data usage is essential, especially concerning privacy and ethical considerations. Legislation like the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) in the United States establishes foundational privacy standards, yet there is a need for more comprehensive laws to address the complexities of data science.

They advocate for increased collaboration among stakeholders, including policymakers, data scientists, and ethicists, to develop regulations that reflect the intricacies of the modern data landscape. The authors stress the importance of being clear about how data is used, obtaining consent from people whose data is collected, and being responsible in data analysis. They believe these steps will build trust in data-driven decisions and reduce the chances of data misuse. They suggest creating ethical guidelines and laws that keep up with the fast-changing field of data science, to protect individuals' rights and encourage responsible handling of data.

Data Governance Solution

The case study highlights the importance of having a strong data governance framework in data science to ensure that data is used ethically and effectively. The authors suggest a structured approach that includes ethical guidelines, stressing that data scientists should be mindful of the impact of their work. While current solutions offer a good starting point for ethical data practices, they might not be enough due to the fast-changing nature of data science and technology. There’s a need for adaptable frameworks to tackle new challenges like algorithmic bias and data privacy. Additional recommendations include regular training for data scientists on ethical practices, encouraging collaboration between different fields to better understand ethical issues, and setting up independent committees to review data projects. These steps can help organizations strengthen their data governance and promote responsible data use.

NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

CASE STUDY: GENERATIVE AI HAS AN INTELLECTUAL PROPERTY PROBLEM

MS DSP 485 - Data Governance, Ethics & Law

Sachin Sharma

August 11, 2024

Reference

Appel, Gil, Juliana Neelbauer, and David A. Schweidel. 2023. "Generative AI Has an Intellectual Property Problem." Harvard Business Review. April. https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem.

The Data

In the case study "Generative AI Has an Intellectual Property Problem" by Gil Appel, Juliana Neelbauer, and David A. Schweidel, the data source comprises data lakes and large archives of images and text utilized by generative AI platforms. The authors articulate the data collection process by explaining how these platforms are trained on extensive datasets, which include billions of parameters derived from various content. While the authors provide a clear overview of the data's origin and the implications of its use, the adequacy of the data and analysis can be questioned. The authors highlight ongoing legal disputes that indicate potential shortcomings in how data is sourced and utilized, suggesting that the existing data may include unlicensed works and could lead to unauthorized derivative outputs. This raises concerns about the representativeness and ethical sourcing of the data, implying that further analysis and transparency are needed to fully assess the implications of generative AI on intellectual property rights.

Relevance to Data Governance

The challenges surrounding generative AI and intellectual property fit into a data governance framework as both legal and ethical issues. From a legal perspective, the use of unlicensed content in training datasets raises significant concerns about compliance with copyright laws and the potential for infringement, necessitating robust governance policies to ensure data is sourced ethically and legally. Ethically, the issue involves the responsibility of AI developers and organizations to respect creators' rights and ensure that their technologies do not exploit or harm individuals whose works are incorporated into AI training datasets. Effective data governance requires establishing clear guidelines and accountability mechanisms to address these legal and ethical dimensions, promoting transparency in data sourcing, and ensuring that intellectual property rights are respected while fostering innovation in AI technologies.

Relevant Legislation

The case study discusses how current intellectual property laws, including copyright, trademark, and patent laws, affect generative AI. It highlights ongoing legal issues, such as the case of *Andersen v. Stability AI*, which questions whether artists' works can be used without permission to train AI models. This situation shows a clear need for updated legal standards that address the specific challenges of AI-generated content and protect the rights of original creators.

Although existing laws offer a basic framework, the authors argue for more targeted regulations to tackle the unique challenges posed by generative AI. Proposed legislation could provide clearer definitions of ownership and rights for AI-generated works, establish guidelines for ethically sourcing data, and ensure fair compensation for content creators whose works are used in AI training. Additionally, the authors suggest that companies and developers should take proactive steps to comply with current laws and promote transparency, such as keeping detailed records of data usage and creating licensing agreements with content creators. These measures could help reduce legal risks and create a fairer environment for everyone involved in the generative AI landscape.

Data Governance Solution

A key takeaway from this case study is that companies using generative AI need to create strong data governance strategies to address concerns about intellectual property. The authors suggest solutions like making sure training data is properly licensed and keeping records of where AI-generated content comes from. While these ideas are a good start, current solutions might not be enough because AI technology and its legal issues are changing quickly. Other possible solutions include creating industry-wide standards for using generative AI, improving collaboration between content creators and AI developers, and pushing for clearer laws to protect intellectual property rights related to AI-generated content. These steps could help reduce risks, ensure creators are fairly compensated, and encourage innovation in the AI field.