CHANGING DATA SOURCES IN THE AGE OF MACHINE LEARNING FOR OFFICIAL STATISTICS

CEDRIC DE BOOM & MICHAEL REUSENS

Statistics Flanders, Brussels, Belgium

ABSTRACT

Data science has become increasingly essential for the production of official statistics, as it enables the automated collection, processing, and analysis of large amounts of data. With such data science practices in place, it enables more timely, more insightful and more flexible reporting. However, the quality and integrity of data-science-driven statistics rely on the accuracy and reliability of the data sources and the machine learning techniques that support them. In particular, changes in data sources are inevitable to occur and pose significant risks that are crucial to address in the context of machine learning for official statistics.

This paper gives an overview of the main risks, liabilities, and uncertainties associated with changing data sources in the context of machine learning for official statistics. We provide a checklist of the most prevalent origins and causes of changing data sources; not only on a technical level but also regarding ownership, ethics, regulation, and public perception. Next, we highlight the repercussions of changing data sources on statistical reporting. These include technical effects such as concept drift, bias, availability, validity, accuracy and completeness, but also the neutrality and potential discontinuation of the statistical offering. We offer a few important precautionary measures, such as enhancing robustness in both data sourcing and statistical techniques, and thorough monitoring. In doing so, machine learning-based official statistics can maintain integrity, reliability, consistency, and relevance in policy-making, decision-making, and public discourse.

1 INTRODUCTION

The field of statistics has long played a critical role in informing policy decisions, driving innovation, and advancing scientific knowledge. Traditional statistical methods such as surveys and censuses have provided valuable insights into a wide range of topics, from population demographics to economic trends and public opinion. However, in recent years, the increasing availability of open and large data sources has opened up new opportunities for statistical analysis. In particular, the rise of machine learning has transformed the field of statistics, enabling the analysis of massive datasets, the identification of complex patterns and relationships, non-linear forecasting, etc. [1, 2]. Machine learning algorithms can be used to analyze data from a wide range of sources, providing insights that traditional survey methods may not capture.

The use of machine learning for official statistics has the potential to provide more timely, accurate and comprehensive insights into a wide range of societal topics

[3]. By leveraging the vast amounts of data that are generated by individuals and entities on a daily basis, statistical agencies can gain a more nuanced understanding of trends and patterns, and respond more quickly to emerging issues.

However, this shift towards machine learning also presents a number of challenges. In particular, there are concerns about data quality, privacy, and security, as well as the need for appropriate technical skills and infrastructure [4, 5], as well as challenges related to explainability, accuracy, reproducibility, timeliness, and cost effectiveness [6]. As statistical agencies grapple with these challenges, it is essential to ensure that the benefits of machine learning are balanced against the risks and that the resulting insights are both accurate and representative. In this paper, we explore the changing data sources in the age of machine learning for official statistics, as we believe that this pervasive issue largely remains underexposed, as we will explain in Section 2.2. In that respect, we highlight some of the key considerations for statistical agencies looking to incorporate machine learning into their workflows in Section 3, by zooming in on the causes and risks associated with using external data sources, the consequences on using such sources for statistical production, and, finally, a set of mitigations that should ideally be incorporated in any genuine deployment of machine learning for official statistics.

2 MACHINE LEARNING FOR OFFICIAL STATISTICS

The data abundance in governmental, corporate, social and personal contexts, both online and offline, becomes a tantalizing source and opportunity for the improvement and expansion of official statistics. For example, to inquire about the overall satisfaction with life of its citizens, a nation could organize periodic surveys. But when this nation has access to its citizens' social media posts, likes, reader's letters, media consumption, ticket sales, (online) shopping carts, etc. it could use all of these data as a proxy to extract novel and innovative statistical insights [7]. Typically, the end result is either a new statistic, a statistic that complements an existing one, or an *ersatz* statistic that aims to replace one or more existing statistics. We will briefly zoom in on the different components of which such a novel statistic is generally comprised.

2.1 Machine learning

To derive novel insights and innovative statistics from data, data scientists and statistical researchers often use a wide variety of powerful tools that are in the realm of machine learning[1]. In this paper we will not go into too much detail about machine learning. However, the use of data sources together with machine learning models can cause unwanted effects regarding statistical production, see further in Section 3. So, we deem it useful to provide a high-level overview of the typical process that involves machine learning for official statistics.

Machine learning is a subdiscipline of artificial intelligence that enables machines to learn from data and improve their performance over time without being explicitly programmed. This approach involves building algorithms that automatically learn from data to identify patterns, relationships, and structures that may be difficult or impossible for humans to discern. The typical process of designing a machine

1 Although originally (and technically) they imply different methods and techniques, the terms machine learning, data science, artificial intelligence, deep learning... are nowadays considered interchangeable. In this paper we consistently use the term machine learning to denote the scientific discipline concerned with learning the most optimal model parameters based on data. It is a subdiscipline of artificial intelligence, while deep learning is a subdiscipline of machine learning. Data science encompasses both machine learning as well as data preparation, analytics and visualization.

learning algorithm consists of two main phases: training and inference. During the training phase, the parameters of a machine learning model are tuned to solve a specific task. For this, a wide variety of data sources can be used, or even outputs from existing or pre-trained machine learning models. After the model is trained, its parameters are kept fixed so that the model can be used to predict outcomes or identify patterns in new or previously unseen data. This process is called inference. It is important to keep in mind the distinction between training and inference. After training, the model remains unchanged, and it remains unchanged until it is retrained again. Later, in Section 3, we will focus on the disparities that this can cause w.r.t. the inference phase and, in consequence, official statistics production.

Supervised learning is one of the most common types of machine learning used for official statistics. This method involves training a model on a labeled dataset, where each data point has a known outcome or target variable. The model learns to associate features in the data with the target variable, enabling it to make predictions on new data with similar features. In the context of official statistics, supervised learning can for example be used to predict the happiness of an individual based on their Twitter profile [8]. Unsupervised learning, on the other hand, is used when the target variable is unknown or the goal is to identify patterns or relationships within the data. In this approach, the machine learning model learns to recognize similarities and differences among input data without explicit guidance from labeled data. In the context of official statistics, unsupervised learning can for example be used to identify citizens, companies or events that are similar to each other on one or more aspects that could be hidden from plain sight [9].

Machine learning can be used to complement or even replace official statistics, and its ability to nowcast and forecast is an extremely valuable addition. Modern machine learning models, tools and hardware can analyze vast amounts of data in real-time or near-real-time, providing more up-to-date and precise estimates of e.g. economic and social trends. By incorporating machine learning into official statistical production, one can benefit from the strengths of both approaches and make more informed decisions based on the most current and accurate data [10].

2.2 External data sources

Let's focus on the data sources that will power such machine learning models. Their nature, size, structure, frequency... can be vastly different, they must typically be gathered 'in the wild' and should often be combined with each other to extract meaningful insights. Compared to more traditional data sources for official statistics, they may present unique and appealing characteristics such as:

Broad-spectrum – Covers a wide variety of topics.

Diversity – A large variety of sources to cover different perspectives.

Availability – Lots of data is freely and easily accessible.

Size – Some datasets can be enormous, sometimes even complete.

Structure – Not only tabular data, but also images, video, text, audio, etc.

Timeliness – (Near) up-to-date and real-time information.

Frequency – Raw data on various, even very fine-grained time scales.

Granularity – Raw data on various, even fine-grained levels of detail.

Coverage – Various locations and regions can be filtered and covered.

On the other hand, before all this data is ready to be exerted for machine learning and official statistical production, a few challenges need to be overcome, such as:

Data quality – Data may contain errors, biases, or missing values that need to be addressed to ensure accuracy and reliability.

Data interpretation – Understanding the context and meaning of data can be difficult, especially when dealing with unstructured data such as text or images.

Data integration – Combining data from different sources with varying structures and formats can be challenging and time-consuming.

Selection bias – Proper randomization or compiling representative population samples can be challenging, and it greatly depends on the underlying data origins.

Operationalization bias – Reproducibility can be difficult as it depends on many implicit, hidden, and/or production-specific design choices [11, 12].

Computational resources – Processing and analyzing large amounts of data may require significant computational resources.

Privacy and security – Sensitive data may need to be protected and anonymized to ensure privacy and security.

Data ethics – Data collection and use should adhere to ethical principles.

Fairness and justness – The end solution should ideally be as neutral as possible and should not discriminate [13].

Cost – All of the above requires resources, budgets and a talented workforce. In addition, the data source itself might need to be purchased. In 2016, McKinsey reported that many companies have started to specialize in acquiring and selling data [14].

With the right tools, workforce, technological advances, mindset, and legislative support, these challenges can and should be manageable. The most challenging piece of the puzzle, however – and one that is more than often ignored – is the *lack of control* you can exert over the data sources that are externally gathered. As a national statistics agency, traditionally, survey data and administrative records that power official statistics are completely under your own control. But once you start exploiting external data sources to power novel, innovative, complimentary or ersatz statistics, this lack of control of your data should never be ignored, and if possible, should be front and center on your agenda early on in the process.

As the popular saying goes: "With great power comes great responsibility" (from Spider-Man, 2002). Having control and power over your data is essential to fulfilling your responsibilities as a statistics agency. However, in the world of data, the opposite is often true: with great amounts of external data comes great powerlessness. Therefore, it is crucial to prioritize the issue of data control when incorporating external sources into official statistics. Taking the time to establish proper protocols and procedures for external data management can prevent a multitude of issues down the line and ensure that the data you rely on remain accurate and trustworthy.

This paper delves into the pervasive problem of powerlessness and lack of control, unraveling the multifaceted aspects, risks, and pitfalls that arise from utilizing external data sources for machine learning in official statistics. We will explore the concepts of 'change' and 'consequence' in their most expansive interpretations to comprehensively tackle this question.

3 THE CHALLENGE OF CHANGING DATA SOURCES

Relying heavily on external data sources for machine learning in official statistics comes with significant risks. Such a dependence can leave statistical agencies vulnerable since they have limited control over these sources. This situation is similar to how our global economy, mobility, and prosperity were once highly dependent on the availability of oil. Since the prices and availability of these precious resources are often beyond our control, countries can do nothing but endure price fluctuations and shortages. Clive Humby proclaimed in 2006 that "data is the new oil", given its powerful intrinsic value. However, his statement keeps holding true in terms of vulnerability, powerlessness, and lack of control over external providers.

In the following paragraphs, we will delve into the various types and causes of data changes. We will then discuss the ramifications of changing data sources for machine learning in official statistics. Finally, we will provide a list of best practices and tips, although it is important to remember that there is no free lunch: whenever we incorporate external data, we expose ourselves to the risk of future changes in these data sources.

3.1 Types and Causes of Changing Data Sources

3.1.1 Data types and schemas

A change in data types or schemas refers to modifications made to the data formats or the structure in which the data are stored and offered. These types of changes may arise due to a need to accommodate future use cases or business requirements, to eliminate technical debt, or to improve data storage and retrieval efficiency. Even the most innocent changes – e.g. integers becoming floats, data columns that are added or removed... – can break entire pipelines. In the most fortunate of cases the runtime environment will throw errors that reveal the cause of these data changes. In other cases, however, the data changes remain undetected and secretly wreak havoc in the pipeline. If the pipeline contains machine learning components, data type changes can e.g. induce feature mismatches – discrepancies between the feature distributions at train and inference time – that lead to unreliable predictions.

It is important to be vigilant about changes in data types or schemas, as even seemingly minor adjustments can have significant impacts further down the data pipeline. To mitigate these risks, it is advisable to stay informed about data change announcements from providers and implement robust data checks during data ingestion, ranging from simple data (type) validation to full-blown automated feature analysis, outlier detection, etc. Additionally, the deployment of effective monitoring systems can help catch machine learning failures quickly and prevent potentially costly errors.

3.1.2 Sharing and collection technology

Data can be shared and collected using many different technologies, such as APIs, queues, network drives, external drives, e-mail... but also web scraping, online analytics tools, sensor networks... Changes in these technologies inevitably occur from time to time. For example, API endpoints often need to be updated to improve functionality and performance. Changes may be made to the API's data structures or methods, to provide more efficient or comprehensive data access. In addition, changes may be made to the API to address security vulnerabilities or to ensure compliance with new regulations or standards. Furthermore, changes in business requirements or strategy may also lead to changes in API endpoints. For instance, a company may introduce new products or services, modify their existing offerings, or change their pricing.

A recent, telling example is the Twitter API. In 2021, Twitter launched version 2 of its popular API that introduced many changes in endpoints, data fields, pricing... compared to version 1.1. Twitter encouraged developers to migrate to this new API offering, but for many use cases such a migration would introduce breaking changes that, in their turn, would impact entire data processing pipelines, statistics production, etc. For the time being, Twitter offered both version 1.1 and version 2 of their API in parallel, which caused many to bury their heads in the sand. The situation got even worse when Elon Musk acquired Twitter in 2022 and decided to suspend all existing API offerings. Instead, in 2023 a new enterprise tier was introduced that put a price of more than 40 thousand USD per month on any reasonably effective use of the API. This caused great dissatisfaction in the development and research community and many initiatives were abandoned.

3.1.3 Concept drift

Concept drift is related to changes in the data distribution between train and test time, which can have multiple causes [15, 16]. Changes in business logic can induce information shifts, for example, when categorical variables are expanded with additional categories or when the meaning of certain data fields is altered. A particular pervasive issue is the calculation of derived data fields, especially when those calculations are not transparent or proprietary. In the age of machine learning, you should always assume that derived data fields can be the result of a model prediction; when this model is updated without your knowing, the derived data fields will have a (slightly) different data distribution, which will cause issues in dependent machine learning models. But even when data fields are not the result of a model's prediction, it is important to periodically reevaluate and retrain models, since many sociological and economic processes are naturally prone to concept drift themselves.

3.1.4 Frequency and interruptions

A change in data frequency refers to modifications made to the rate at which data is collected or updated, which can happen deliberately or randomly. Deliberate changes may arise due to shifts in business requirements or technological choices. Random shifts are most often attributed to noisy factors such as network issues, component failures, downtime... or (human) errors. Such changes in frequency can impact machine learning components dramatically. E.g. when periodical data is sampled every minute instead of every second, the data distribution changes on which the model was trained. To mitigate these risks, data pipelines should be designed to monitor changes in incoming data frequency.

3.1.5 Ownership and discontinuation

Worse than interruptions is downright discontinuation of the data source, which has immediate consequences on the future existence of the statistic. Also, a change of ownership of the data source – e.g. when acquired by another company – is not a fictional scenario, and it can trigger any of the risks that are discussed in this section. Building redundancy by diversifying data sources is a useful mitigating strategy to avoid single points of failures.

3.1.6 Legal properties

Legal changes refer to modifications made to the legal landscape that governs the collection, storage, and use of data. This type of change may arise due to new privacy laws, contractual obligations, or changes in the cost of data access or storage. One cause of this type of change is the adoption of new regulations, such as GDPR, which require companies to comply with stricter rules for collecting and

processing data. Additionally, changes in the cost of data access or storage may require companies to modify their data sources or methods to reduce costs, which can have contractual consequences. If possible, negotiate airtight SLAs with the data provider and make sure to attribute enough attention to future data changes.

3.1.7 Ethics and public perception

Ethical considerations and public perceptions can affect data collection methods and sources. If certain data sources or variables are considered controversial or intrusive, there may be a shift towards alternative sources, which may require a refresh of the used machine learning models. It can also impact the way machine learning models are designed and trained. If certain variables or factors are considered discriminatory or unethical, there may be a push towards eliminating or adjusting them to reduce algorithmic bias. Changes in ethics or public perceptions can also result in greater accountability and the need for transparency. Stakeholders may demand more openness and clarity around the use of algorithms, data sources, and decision-making processes. This can lead to greater scrutiny and oversight of machine learning models, which may impact their performance if not adequately addressed, especially when black-box models need to be replaced by more interpretative variants [17, 18, 19]. Finally, public trust can be significantly affected. If stakeholders perceive that machine learning is being used inappropriately or unethically, they may lose faith in the integrity and reliability of official statistics. This can have significant consequences for public policy and decision making.

3.2 Consequences of Changing Data Sources

When data sources change, there will be consequences for official statistics production, especially if there are machine learning components involved. We will broadly but briefly cover a variety of areas that can be impacted, some of which have already been mentioned above.

Concept drift – Concept drift means that the underlying patterns and relationships in the data may change over time, which can lead to model deterioration or loss of accuracy. This issue can be particularly relevant when dealing with long-term trends, as changes in societal norms, technology, or other external factors can influence data over time [16].

Model staleness – When a model becomes outdated, it no longer reflects current trends or patterns in the data. This can occur if the machine learning model is not updated frequently enough to keep pace with changing data sources. As a result, the model may not perform as well as it once did, leading to less accurate official statistics.

Bias and neutrality – Changing data sources can also introduce bias or incorrect data, which can impact the neutrality of the statistics produced, or which can lead to the phenomenon "garbage in, garbage out" [20, 21]. Since it is essential that official statistics remain neutral and objective, this will negatively impact the accuracy and validity of these statistics.

Availability – If data become unavailable (for a certain period in time or indefinitely) or are limited in scope, this may impact the ability to produce accurate and timely official statistics.

Integration – A change in data sources can cause a domino effect when multiple statistics or models rely on this data source. Especially be mindful when the output of machine learning models is used as input for other machine learning models, either directly or indirectly as part of a larger data pipeline. Since the predictions of

a machine learning model can become unreliable when the input data change, this prediction shift itself is a changing data source for other models.

Extra labor – The risk of changing data sources requires additional resources and labor to mitigate the effects of such changes, monitor the occurrence of changes, and ensure that the new data are properly integrated into existing machine learning models. This has tremendous impacts on the costs and timeline of the produced statistics, and it may also require a significant team expansion.

Breaking changes or discontinuation – In some cases, changing data sources may cause the impossibility of producing a statistic any further. If this is the case, it may be necessary to stop offering the statistic altogether or find alternative data sources that can produce accurate and reliable official statistics. When alternative data sources are found, there will almost always be a mismatch with the original data source that has an impact on the resulting statistic, resulting in a breaking change. In that case, it is important to overcome the mismatches as best as possible – e.g. in terms of statistical properties – and, certainly, to be transparent about the breaking change, e.g. by indicating on a graph when exactly the data source was changed.

Quality metrics – Finally, changing data sources imply changes in timeliness, validity, accuracy, completeness, consistency and other quality metrics w.r.t. the produced official statistics [6]. Ensuring that resulting statistic continues to meet these quality metrics remains critical.

3.3 Mitigating Changing Data Sources

As has been illustrated above, changes in data sources can significantly impact the performance of a machine learning model. The effects can be diverse, ranging from introducing biases in the data to producing incorrect results. This can have serious implications, especially when the model is being used for official statistics, where accuracy and reliability are of paramount importance. Therefore, it is essential to take measures to prevent and mitigate such changes. This is not an easy task, as the consequences can be diverse, and the required efforts to mitigate them are often time-consuming and not straightforward. We do not claim to have definite answers. However, we will propose several recommendations and best practices, including performing a risk analysis, monitoring, diversifying data sources, building technical robustness, using data normalization techniques, and incorporating data validation processes.

Risk analysis – Performing a risk analysis before incorporating a new data source is an essential step in mitigating the impact of changes in data sources. This analysis involves identifying the potential risks associated with the data source, which we have covered in Section 3.1. The analysis should be comprehensive, considering both technical and non-technical aspects of the data source, and should ideally include potential solutions for the identified risks. This will often force you to face the hard truth and will lead you to decide that the candidate data sources are not adequate or reliable enough. Trade-offs will nevertheless need to be considered, depending on the use case at hand.

Monitoring – Monitoring everything that is relevant is another crucial step in mitigating the impact of changing data sources. It involves tracking various aspects of the data sources, the machine learning models, and their outputs to detect and respond to changes promptly. Draft a list of variables and quantities that must be continuously tracked to ensure that the models remain reliable and accurate over time. For this, inspiration can be drawn from the discussed topics in Section 3.1,

but it will vary from use case to use case, as well as the nature of the models that have been used.

Supervised models, for example, can be tested against a reference test set or a historical reference model; if the accuracy, precision or recall starts to deviate significantly from this reference set, it should be flagged. On the other hand, monitoring the performance of unsupervised models can be more challenging, because there is no clear performance measure that can be directly computed. One approach is to monitor the model's ability to detect patterns and clusters in the data. It is possible to use a reference test set or reference model for this, but the informative metrics – e.g. cluster similarity, homogeneity, separation... – are more abstract and somewhat harder to interpret. Another approach is to visualize projections of certain interesting data points in the learned latent spaces or preferably a reduction thereof, which greatly benefits interpretability but makes it harder to convert it into hard numbers. As a suggestion, a good balance between interpretability and hard performance metrics is found when clusters are tested against pre-existing domain knowledge, e.g. by listing similar data points for given queries. Simply monitoring whether expected similarities emerge or not can provide powerful signals about model and data performance. Another effective approach is to create proxy supervised tasks that rely on the output of the unsupervised model. Monitoring the model's performance on such proxy tasks can provide insights into the quality and usefulness of the unsupervised model's output.

Diversification – Diversifying data sources is another important measure, but is easier said than done. One challenge of using multiple data sources is the potential for conflicts or inconsistencies between the sources. Different data sources may have different formats, schemas, and levels of quality, which can create discrepancies and inconsistencies that must be resolved before the data can be used in the model. Therefore, data normalization is key. Additionally, integrating multiple data sources can be a complex and time-consuming process. It can also create additional computational overhead, which may impact the model's scalability and portability. Finally, finding relevant and reliable data sources can be a challenging task, particularly for specialized or niche domains. It may require extensive research and communication with data providers to retrieve relevant data. Again, this story is about economical, technical and practical trade-offs, and is of course highly use-case-dependent.

Technical robustness – Building technical robustness is paramount and requires significant engineering efforts. Building an automated, data-driven statistic that is resistant to changing data sources such as errors, outliers, outages, time-dependent variability, etc. ensures consistency in the statistical offering. Using data normalization techniques and incorporating data validation into the pipeline are essential measures, but robust technical implementations also require thorough unit and integration testing, failover and deduplication, scalability solutions, security measures, etc. Of course, this is an entire field of study on its own.

Legal robustness – Finally, we believe that agreeing on clear legal guidelines is the best mitigation strategy to counter the risk of changing data sources, for example, by closing formal data sharing agreements or SLAa with data providers. Such agreements should specify the terms and conditions under which the data can be shared, as well as the legal responsibilities of each party. In particular, the agreements should specify the legal consequences of non-compliance.

4 CONCLUSION

In this paper we have investigated the risks and consequences of changing data sources when using machine learning for official statistics. The list is long and covers many different aspects, ranging from statistical issues and model inconsistencies to technical problems and ethical considerations. We have also looked at a few potential mitigation strategies. However, we admit that these strategies do not provide all the adequate answers and might leave the reader unsatisfied or, worse still, beguiled, as the solutions require many additional resources and efforts. As we have stressed a couple of times in this paper, this is a story of trade-offs. Depending on the use case at hand, some trade-offs might be easier to handle than other ones. However, in the context of official statistics, our advice is to not tread lightly on these matters and to minimize the risk of losing control over your data sources as much as possible. This takes time, effort and careful planning with a horizon of multiple years. To end on a positive note, despite the challenges associated with changing data sources, machine learning offers many opportunities for official statistics. By being aware of the risks and taking necessary precautions, statistical agencies can leverage these opportunities while maintaining the integrity and reliability of their data-driven products. We hope that our checklist of risks and mitigation strategies provides a useful starting point for statistical agencies and practitioners to ensure the robustness of their machine learning-based statistical reporting.

REFERENCES

[1] Stuart J. Russell and Peter Norvig. *Artificial Intelligence*. Pearson Education, 2009.

[2] Trevor Hastie, Jerome Friedman, and Robert Tisbshirani. *The elements of Statistical Learning*. Springer, 2017.

[3] UNECE. Machine learning for official statistics. Technical report, UNECE, 2022.

[4] Hossein Hassani, Gilbert Saporta, and Emmanuel Sirimal Silva. Data mining and official statistics: The past, the present and the future. *Big Data*, 2(1):34–43, March 2014.

[5] Marco Puts and Piet Daas. Machine learning from the perspective of official statistic. *The Survey Statistician*, 84:12–17, July 2021.

[6] Wesley Yung, Siu-Ming Tam, Bart Buelens, Hugh Chipman, Florian Dumpert, Gabriele Ascari, Fabiana Rocci, Joep Burger, and InKyung Choi. A quality framework for statistical algorithms. *Statistical Journal of the IAOS*, 38(1):291–308, March 2022.

[7] Wesley Yung. The Evolution of Official Statistics in a Changing World. *Harvard Data Science Review*, 3(4), oct 28 2021.

[8] Manon Reusens, Michael Reusens, Marc Callens, Bart Baesens, et al. Benchmark study for flemish twitter sentiment analysis. *Social Science Research Network*, 2022.

[9] Annelien Crijns, Victor Vanhullebusch, Manon Reusens, Michael Reusens, and Bart Baesens. Topic modelling applied on innovation studies of flemish companies. *Journal of Business Analytics*, pages 1–12, 2023.

[10] Sevgui Erman, Eric Rancourt, Yanick Beaucage, and Andre Loranger. The Use of Data Science in a National Statistical Office. *Harvard Data Science Review*, 4(4), oct 27 2022.

[11] Matthias Haucke, Rink Hoekstra, and Don van Ravenzwaaij. When numbers fail: do researchers agree on operationalization of published research? *Royal Society Open Science*, 8(9):191354, September 2021.

[12] Nagireddy Neelakanteswar Reddy. Operationalization bias: A suboptimal research practice in psychology. December 2022.

[13] Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. From fair predictions to just decisions? conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology*, 2022.

[14] Nicolaus Henke, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, and Guru Sethupathy. The age of analytics: competing in a data-driven world. Technical report, McKinsey & Company, 2016.

[15] Hanqing Hu, Mehmed Kantardzic, and Tegjyot S Sethi. No free lunch theorem for concept drift detection in streaming data classification: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(2):e1327, 2020.

[16] Firas Bayram, Bestoun S. Ahmed, and Andreas Kassler. From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems*, 245:108632, June 2022.

[17] Alicja Gosiewska, Anna Kozak, and Przemysław Biecek. Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems*, 150:113556, November 2021.

[18] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019.

[19] Alex John London. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, 49(1):15–21, January 2019.

[20] Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300, December 2020.

[21] R. Stuart Geiger, Dominique Cope, Jamie Ip, Marsha Lotosh, Aayush Shah, Jenny Weng, and Rebekah Tang. "garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies*, 2(3):795–827, 2021.