How to be responsible in all the steps of a data science pipeline: The case of the Italian public sector

Monica Scannapieco1,* and Antonino Virgillito2
1Istat (Italian National Institute of Statistics), Via Balbo 16, 00184 Rome, Italy
2Agenzia delle Entrate (Italian Revenue Agency), Via Giorgione 106, 00147 Rome, Italy
*Correspondence: scannapi@istat.it
https://doi.org/10.1016/j.patter.2021.100393

The paper highlights how each step of a data science pipeline can be performed in a “responsible” way, taking into account privacy, ethics, and quality issues. Several examples from the Italian public sector contribute to clarifying how data collections and data analyses can be carried out under a responsible view.

An extensive discussion is currently in progress at different levels of society (including with academics, technicians, and public decision makers), revolving around ethical aspects related to the extensive usage of data and the application of automated processes and artificial intelligence (AI) methods.

Data scientists are deeply involved in this debate, as they manage firsthand all the sensitive choices related to data treatment. This is particularly true for data scientists working in the public sector, as they are responsible for the realization of data-driven products that can have a significant impact on people’s lives. An example is given in the fiscal domain by data analyses necessary to detect tax frauds and prevent tax evasion. A further example is provided by the computation of crucial statistics that drive public decision making, as it was in the case of COVID-19 emergency responses.

Moreover, the role of public administrations is perceived as particularly invasive for the privacy of citizens. This puts data scientists in a position in which they have to raise their awareness about transparency and take accountability of their actions, so they can be recognized as trusted by the general public.

A call for “responsible” data science is then paramount in this setting. This often seems to translate mainly in controlling the “algorithmic” part of the analysis process. Although this is surely one of the aspects that must be taken into account, our idea is that responsibility instead affects all the phases of a data science pipeline, from the acquisition of data to the delivery and dissemination of information products.

We sketch our ideas of a responsible data science process by analyzing the choices that can be made at different steps of a common data science pipeline, taking, as examples, two radically different uses of data science in the public sector: on one hand, official statistics, where data science is used mainly to produce statistics from “big data” sources, and on the other hand, fiscal or social security agencies carrying out risk analysis activities based on their internal data.

The paper concludes with the presentation of a standardization effort on modeling a generic data science pipeline, where a “trust management” business function that takes into account all the issues described along the paper is explicitly introduced.

**Responsible data access**

It is necessary to put in place responsible measures that take into account privacy and/or ethical issues starting from the very first phases of a data science pipeline, namely data access and data acquisition. Let us consider the following two examples, namely (1) accessing mobile network operator (MNO) data in a privacy-preserving way and (2) acquiring Web data from public web sites in an ethical way.

**Example of responsible access to MNO data**

MNO data is a big data category that can serve many purposes in the public administrations sector; as an example, persons mobility analyses are very useful to support policies related to transports and tourism. They have also proven to be a relevant source for timely decision making during the COVID-19 emergency (see e.g., the statistics provided by the Spanish statistical institute1).

However, MNO data are personal data, hence the requirement to put in place privacy-preservation methods since the beginning of their treatment. To this aim, national statistical organizations (NSOs) are starting to invest in privacy-enhancing technologies (PETs)2 to access these data in a privacy-preserving way; PETs include techniques like secure multiparty computation, homomorphic encryption, trusted execution environments, etc. The main idea is that data are not moved from the data holder to the NSO’s premises but rather remain at the data holder’s premises only being accessed by the NSO, without any direct acquisition of raw and elementary data.

The current effort by NSOs is to engage in a partnership relationship with private data holders like MNO data providers, in order to be able to define shared access modalities that can be implemented by means of PETs. In this way, statistics based on MNO data can be provided by NSOs in a way that final users can trust both the quality of the provided results and the guarantees of their own privacy preservation.

**Example of a responsible acquisition of web data**

The automated acquisition of web-available data can be done either via dedicated application programming interfaces (APIs)
that the data holder purposefully set up or via web-scraping programs, i.e., dedicated software programs that access and download (portions of) websites.

An example of a project involving web-scraping activities in the official statistics domain is the “Enterprise Characteristics” project, performed within the European projects ESSnet Big Data Pilots I and II. In these projects, massive web-scraping activities were set up, accessing hundreds of thousands of enterprises’ websites with the purpose of analyzing the information they provide on their websites and estimating characterstics like presence of web order facilities, presence of online job advertisements, etc.

Within such projects, a proposal was made to make web-scraping programs respect a “netiquette,” i.e., a set of principles and operational practices ensuring that “web-scraped data are used in an appropriate and ethical manner that limits the burden on website owners and survey respondents to the greatest extent possible.” The proposed principles include (1) seek to minimize burden on the website owners, (2) honor requests made by website owners to refrain from scraping their websites, (3) protect all personal data in all statistics and research outputs and seek ethical advice when web-scraping data that may identify individuals, (4) apply scientific principles in the production of statistics and research based on web-scraped data and consider other sources of data, and (5) abide by all applicable legislation and monitor the evolving legal situation.

This is a further notable measure put in place to carry out data acquisition in a respectful and ethical way.

**Responsible data analysis**

We refer to “data analysis” as the set of core activities of data scientists, featuring the extensive manipulation of raw data with the objective to disclose its hidden characteristics through, for example, statistical indicators or predictions obtained by machine learning (ML) algorithms. A fundamental question is how these outputs can be used in the context of a public administration. The General Data Protection Regulation (GDPR) imposes a strict rule on this point for personal data, forbidding, in essence, the implementation of any completely automated process that produces a direct legislative effect on citizens. Hence, the use of data science in the public sector is basically targeted to support human operators in their decisions and activities rather than to replace them through the introduction of extensive automation. However, such a human-in-the-loop scenario is not by itself a guarantee that the results of the treatment as a whole are completely ethical.

One particularly relevant concept in this sense is data fairness. Fairness can be defined in terms of lack of discrimination, i.e., treating someone differently. It is possible to distinguish between two categories of fairness: individual fairness, for which similar predictions are given to similar individuals, and group fairness (also known as statistical fairness), for which different groups are treated equally independently of a particular race, gender, or sexual orientation. A very famous example of group unfairness is the COMPAS algorithm used by the Department of Corrections in Wisconsin, New York, and Florida that has led to harsher sentencing toward African Americans. As we said, such a situation is not possible in the current European legislative framework; however, unfair algorithmic decisions can still occur, possibly providing misleading indications to human operators.

As ML algorithms base their predictions on training data and improve them with the growth of such data, it is important to properly identify the possible sources of unfairness while preparing the training set. In a typical project, the creation and curation of training datasets consists of assigning a label to the training dataset on which the algorithm bases its prediction on unseen/unlabeled data. Labeling can be done in two ways: manually or through the use of historical data.

Manual labeling is largely a human-based activity involving several people: domain experts, data scientists, ML experts, etc. In this scenario, the judgement over a particular label can be different if done by different people, and this inconsistency in the labeling process can eventually lead to inaccurate algorithmic decisions.

On the other hand, labels can be directly derived by existing historical data. A typical example in risk analysis applications is the use of data that records the activity of human auditors in the past that identified frauds or any other illegitimate behavior. Using that as training data means having the algorithms somehow “reproduce” the choices made by auditors in the cases selection automatically. However, these data are inherently biased because manual case selection is naturally targeted toward situations that expose more evidently risky characteristics and less dangerous cases are under-represented. Again, the algorithm can make wrong decisions because it lacks sufficiently representative examples.

In substance, humans typically suffer from conscious and unconscious biases, and these can be incorporated in training sets even inadvertently, perpetuating and amplifying existing inequalities and unfair choices. Data scientists should be aware of this and take sufficient time and attention analyzing training data before feeding it to the algorithm, making sure that all situations are properly represented.

Overall, we can say that data-related human design decisions affect learning outcomes throughout the entire process pipeline, even if, at a certain point, these decisions seem to disappear in the black-box, “magic” approach of ML algorithms. It is a still-open problem to figure out concrete solutions on how to discover and eliminate unintended unfair biases from the training datasets and/or to create “by design” datasets that are natively “fair.” A suggestion is to never underestimate the extensive use of data exploration and statistical analysis in any stage of the data preparation. Domain knowledge is also fundamental, as domain experts can quickly recognize anomalies that are not detected by data scientists. Explainability of algorithms is also an important element in this sense, giving human operators indications about the factors that drove the choices of the algorithm.

**A standard data science pipeline for official statistics**

The systematic introduction of trust in the context of a data science pipeline can be facilitated by standardizing the phases that compose the pipeline itself. A contribution, in this sense, within the domain of official statistics has been realized within
a project financed by Eurostat, the statistical office of the European Union. The project, named ESSnet Big Data Pilots II, defined BREAL (Big data REference Architecture and Layers), a European reference architecture for using big data sources in official statistics. Even if the planned application domain of BREAL is official statistics, the architecture is indeed applicable to many other domains, and, in particular, it can be used by public and private organizations that would like to follow a defined and controlled way of using data science for their purposes.

The standard data science pipeline proposed by BREAL is shown in Figure 1 and encompasses three major business process areas:

- Development and information discovery—where the exploration of the data source, its integration with other data, and the discovery of information take place.
- Production—actually creating statistical products through the use of raw data sources.
- Continuous improvement—monitoring and assessing the data source usage with a focus on representativeness issues and the validity of the models used.

The main drivers to the first phase of BREAL’s life cycle are two business functions, namely specifying needs and new data sources exploration. The development and information discovery process is supported and served by the business functions of metadata and trust management. This latest function is crucial to the BREAL life cycle and is intended to take into account all of the issues described throughout this paper. In other words, trust management is a cross-cutting function that, if implemented, has an impact on all the phases of a responsible data science pipeline.

**Conclusion**

In this paper, we discussed some of the issues that should be tackled in order to achieve responsible treatment of data throughout all the steps of a data science pipeline. As also highlighted by Srivastava et al., several communities are called to contribute to this vision, including data-driven enterprises, market professionals, academic researchers, and public sector scientists and domain experts, the latter being appointed with the social responsibility connected with the mentioned ethical and privacy issues. We believe that standardization efforts such as BREAL can help clearly identify roles and responsibilities of each of these actors.

**REFERENCES**

1. INE (Istituto Nazionale di Statistica) Statistical information for the analysis of the impact of the COVID-19 crisis. https://www.ine.es/covid/covid_movilidad_en.htm.
2. UN Handbook on Privacy Preserving Computation Techniques (2018). https://unstats.un.org/bigdata/task-teams/privacy/UN%20Handbook%20for%20Privacy-Preserving%20Techniques.pdf.
3. ESSnet Big Data II From exploration to exploitation. Pilots 1 and 2. https://ec.europa.eu/eurostat/cros/essnet-big-data-2_en.
4. WPC ESSnet Proposal for a Web Scraping Policy. https://ec.europa.eu/eurostat/cros/content/WPC_ESSnet_Webscraping_policy_draft_en.
5. General Data Protection Regulation (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. https://eur-lex.europa.eu/eli/reg/2016/679/oj.
6. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through Awareness. ITCS ’12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference.
7. Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias (ProPublica).
8. European Commission. BREAL- Big Data Reference Architecture and Layers. https://ec.europa.eu/eurostat/cros/content/wpt-milestones-and-deliverables_en.
9. Srivastava, D., Scannapieco, M., and Redman, T.C. (2019). Ensuring High-Quality Private Data for Responsible Data Science: Vision and Challenges. J. Data and Information 11, 1–9.

**About the authors**

Monica Scannapieco is currently head of the “Information and Application Architecture” division at the Italian National Institute of Statistics, where she has worked since 2006 promoting and leading innovation projects on data management. She earned a university degree in computer engineering with honors and a PhD in computer engineering.
Dr. Antonino Virgillito is senior data scientist at Agenzia delle Entrate (Italian Revenue Agency), where he works on the analysis of large-scale datasets of fiscal data through machine learning and advanced analytics techniques. With a PhD in computer engineering, Dr. Virgillito is an expert in machine learning, big data technologies, data engineering, and business intelligence. He previously worked at Istat (Italian National Institute of Statistics), IRISA-INRIA (France), and Università di Roma “La Sapienza” and has collaborated with United Nations Economic Commission for Europe (UNECE).