Statistical Journal of the IAOS 39 (2023) 773–782
DOI 10.3233/SJI-230060
IOS Press

Data science skills for the next generation of statisticians

Laura Antonucci[a], Antonio Balzanella[a], Elvira Bruno[b], Crocetta Crocetta[c], Simone Di Zio[d], Lara Fontanella[d], Maurizio Sanarico[b], Bruno Scarpa[e], Rosanna Verde[a,*] and Giorgio Vittadini[f]
[a] *Università della Campania "Luigi Vanvitelli", Caserta, Italy*
[b] *SDG Group, Milano, Italy*
[c] *Università degli Studi di Bari "Aldo Moro", Bari, Italy*
[d] *Università degli Studi "Gabriele d'Annunzio", Chieti/Pescara, Italy*
[e] *Università degli Studi di Padova, Padova, Italy*
[f] *Università Milano Bicocca, Milano, Italy*

Abstract. This paper analyses the future prospects of statistics as a profession and how data science will change it. Indeed, according to Hadley Wickham, Chief Scientist at Rstudio, "a data scientist is a useful statistician", establishing a strong connection between data science and applied statistics.

In this direction, the aim is to look to the future by proposing a structural approach to future scenarios. Some possible definitions of data science are then discussed, considering the relationship with statistics as a scientific discipline. The focus then turns to an assessment of the skills required by the labor market for data scientists and the specific characteristics of this profession. Finally, the phases of a data science project are considered, outlining how these can be exploited by a statistician.

Keywords: Data science, skills, data scientist, data scientist job, data scientist salary

1. Introduction

A standardized definition or a universally accepted set of skills for a data scientist does not exist, because it is a new and constantly evolving professional figure. For example, Ho et al. [1] define data scientist as a person possessing "the abilities to collect, clean, extract, transform, and load the data. In addition, apply statistical, analytical, and machine learning techniques to draw insights from the data. And most importantly, a data scientist must be able to communicate the findings in both written and spoken form." However, this definition is probably starting to become obsolete because, for example, a data scientist can be theoretical or applied. The first tries to devise methods and strategies to improve the ability to deal with data in a well-founded way, while the second looks for the best practical approach to solve a business/applied problem and helps to make it available for enhancing the everyday operations.

Less problematic is the definition of statistician, having a much longer history. Therefore, it can certainly be said that a statistician is a professional who specializes in the field of statistics and collect, organize, analyze, and interpret data to help make informed decisions, draw conclusions, and solve real-world problems in various fields such as science, business, economics, social sciences, and more (for a discussion on the matter see Hayford [2]).

Starting from this deliberately generic definition, we can say that statistics is a branch of mathematics and a discipline that involves the collection, analysis, interpretation, presentation, and organization of data. It provides a systematic framework for dealing with data, summarizing information, and making informed decisions based on empirical evidence. So, in short, statis-

[*] Corresponding author: Rosanna Verde, Università della Campania "Luigi Vanvitelli", Caserta, Italy. E-mail: rosanna.verde@unicampania.it.

tics can be defined as "the science of learning from data" [3].

We could report many other definitions of both – data scientist and statistician – but it is clear that the two professions have a large overlap area and, above all, while statistics and the profession of statisticians are ancient and well-established, the professional role of data scientist is considerably more recent and continuously evolving.

In a more drastic way, a popular joke claims that "a data scientist is a statistician who lives in San Francisco" outlining the modern and fashion style of data scientists compared with a "elderly" statistician.

The relationship between data scientist and statistician is at the center of an ongoing and very fervent debate, and in this work, we intend to add some fundamental aspects to contribute to this debate.

To give an idea of the scope of this topic, in [4] authors analysed over 100 publications from statistics and data science, outlining an interesting picture of the overlaps and differences between these two disciplinary fields. From this research it clearly emerges that some scholars argue that statistics isn't necessary for data science, but the findings emphasize the complementary relationship between the two. According to this research, statistics provides a foundation for data science, ensuring reliability and validity, while data science extends statistics to Big Data. Data scientists should recognize the importance of statistics, and statisticians should embrace the capabilities of data science [4]. However, although it seems clear that the two disciplines complement and compensate for each other, it is equally evident that statisticians often consider data science a threat.

In this paper, we would like to analyse what are the future perspectives of statistics as a profession and how data science is going to modify it.

In the following, we will first give a look to the future by proposing a tentative normative scenario and then we discuss some possible definition of data science, by considering the relation with statistics as a scientific discipline. We will then look at the skills required by the job market for data scientists and the specific characteristics of this profession. Finally, we will consider the steps of a data science project by outlining how those can be leveraged by a statistician.

2. Futures studies and futures literacy: A tentative normative scenario

When we talk about future skills, we can think of outlining scenarios for the future of statisticians and data scientists. A Futures Studies approach can help us in asking ourselves the right questions and stimulate reflections on what the jobs of tomorrow may be in which statisticians will be called upon to play a key role. Over the last half-century, the study of the future moved from forecasting the future to the broader concept of foresight, which includes shaping multiple futures, drawing desirable futures and anticipatory decision-making. Even though future thinking has been crucial since the beginning of civilization, it is only in the middle of the nineteenth century that scholars start talking about Futures Studies (FS) as a new paradigm [5–7].

Futures Studies refers to a multidisciplinary scientific research field [8,9] regarded by many – not without criticism – as a science in its own right [10–12].

Closely related to the FS we find the concept of Futures Literacy (FL), a new form of literacy on which the attention of national governments, international organizations (like UNESCO, EU, OECD) and scientific research is more and more focused [13]. FL concerns the ability to imagine several futures: what the decision makers and even each single person can do today to set the course to the desired future depend on how people are future literate. The ability to become future literate becomes the capacity to make better today's decisions [14,15]. FL is a set of skills, that can be learned and taught, and we believe that statisticians must also become more *future literate*

The next generations must become more skilled at "using-the-future" because a) the future does not yet exist but must be imagined and shaped; b) humans can learn to imagine the future (therefore it is a skill that can be developed); and c) the future can/must be built, in a proactive approach and not passively endured (passive approach). Hence, also the next generation of statisticians should be more *future literate*.

Talking about the future of this contrast/overlap of professions, a question we all need to address is the following: will data scientist be in the future of statistics? In other words, we ask ourselves whether the two professional fields will move in the direction of greater overlap – to the point of merging – or will move away to the point of constituting two separate and different scientific fields. It is difficult to give an answer but at the same time, it is very useful to try to outline a normative scenario, i.e., a scenario that we statisticians would like to come true, for example, in the next 10 years. This scenario was not outlined according to the scenario method (which would require specific research) but it is simply an exercise proposed by the authors to stimulate a debate on this topic. We also remember that a scenario is never a prediction, but only one of the many plausible images of a distant (here desirable) future.

In this scenario, the data scientist will be a statistician with a solid theoretical and methodological background, who will know how to select quality data. He/she will know how to frame a complex problem (data), able to use the appropriate software and capable of interpreting the results (information). Also, data scientists will be able to produce results that are usable for decision-making and will have the ability to produce results that are understandable to non-professionals, with particular emphasis on reproducibility of research, communication, and visualization. Finally, we want the data scientist of the future to be able to "produce" data when these are not available (new problems and very complex problems). And, because of the last point, we can imagine the statistician of the future as a person capable of applying mixed methods, namely a mixing of qualitative and quantitative methodologies. In fact, in studying complex phenomena, many scholars claim that a mixed method (namely research approaches which combine in a single research strategy a mix of methods as well as involves collecting, analyzing, and exploiting different types of data) is desirable, given the need to analyze the problem from many perspectives [16,17].

If we agree that this is a desirable scenario, then the question that arises is: what must we do today to realize this scenario? Some answers to this question are contained in the previous sections of this article.

Another reflection about the skills necessary for the next generation of Statisticians is that according to the majority of futurologists, many of the jobs that will exist in 2030 haven't even been invented yet. As a consequence, many of the jobs that statisticians will do in the future do not exist today. To this end, there are many studies that try to outline what the jobs of the future will be. Among the many, we mention here the report on the 100 jobs of the future [18]. In this report, given the importance of data processing, there are several jobs under the category of "data jobs", so jobs that statisticians could do. We find the *Algorithm interpreter*, the *Behaviour prediction analyst*, the *Data commodities broker*, the *Data farmer*, the *Data privacy strategist*, the *Data storage solutions designer*, the *Data waste recycler*, the *Forensic data analyst*, the *Freelance virtual clutter organizer* and the *Predictive regulation analyst* [18].

So, not only do we have to deal with the issue of big data analysis and artificial intelligence, but it also becomes crucial to understand how statisticians must be prepared today to be called upon to carry out jobs of this type tomorrow.

After the reflections made so far, we pose one last question: "Where is the statistic headed?" Once again, it is hard to answer but we can conclude with a series of points to reflect on.

- Many jobs of the future will deal with DATA (an inevitable consequence of the so-called datafication of society).
- Statisticians are called upon to proactively anticipate these new jobs.
- There is a need for a correct dialogue with other disciplines (Mathematics, Computer science, Information science, etc.) since the future must be built together.
- To survive in a fast-changing world, we must abandon the Business-As-Usual Thinking (i.e., continue with the present course of action) and start making strategic decisions in view of the futures that emerges on the horizon.
- We must provide a university education which takes Future Literacy into account.
- Another interesting trend consists of the so-called Knowledge-Guided AI, which encompasses combining science with data-driven methods. Examples are physics- informed AI in meteorology, climatology and fluid dynamics, epidemiology-informed machine learning, distribution-informed AI, etc. Statisticians could have a core role if endowed with the appropriate mindset and toolkit.

We are aware that these points derive from a minimal tentative scenario, so free of disruptive events, and a future work might be the construction of 3–4 scenarios through, for example, a Delphi survey. However, in order to stimulate a discussion on the future of statistics and statisticians, we believe that they contain relevant and strategic elements, which probably constitute the key factors of future changes.

3. Data science

Although William Cleveland was the first that in 2001 coined the term data science, by claiming that "results in data science should be judged by the extent to which they enable the analyst to learn from data" [19], it is only in 2008 that this term starts to be vastly used thanks to DJ Patil and Jeff Hammerbacher, founder and Chief Scientist of Cloudera. With this term they intended to denote a new professional figure capable of carrying out analyses on large masses of data in order to extract relevant information that can bring value to the company in which they work. Today, it is difficult to identify a data scientist unambiguously, as the term encompasses a broad range of skills and expertise. They are expected to support public and private administrations in making decisions based on data.

Ben Baumer [20] defines data science as "an emerging interdisciplinary field that combines elements of mathematics, statistics, computer science, and knowledge in a particular application domain for the purpose of extracting meaningful information from the increasingly sophisticated array of data available in many settings", which includes the most relevant aspects widely accepted: it combines elements of mathematics, statistics, computer science and knowledge of an application domain.

The title "Data Scientist" is ambitious on the one hand, as it requires extensive training and knowledge of data management, analysis, forecasting, and communication tools, as well as domain-specific knowledge. On the other hand, it is restrictive because it encompasses many different professions, including data architect, data engineering, data modelling, data analyst, machine learning, and AI analyst.

Moreover, with the increasing presence of data scientists in different fields, it is difficult to refer to "a data scientist"; we should rather refer to different "data scientists" based on their specialized experience in various fields. Data scientists can be analysts of genetic data, physical data, social-economic data, life-science data, or have a role as analysts of business processes, digital marketing, official statistics, and more. This perspective reflects the counterpart of statistics, which related to its longer history begun in the modern sense in the 18th century, includes several branches. In this sense, we can consider a general distinction between applied statistics, theoretical statistics, and mathematical statistics. Moreover, depending on its specialization domain and application in several scientific and humanistic areas, statistics is also shaped as biostatistics, economic statistics, social statistics, demography, epidemiology, medical statistics, psychological statistics, statistical mechanics, and engineering statistics. In relation to these various domains of the statistics, the "statisticians" come from different backgrounds of study, which contribute to their specialization.

The focus should move from the general competencies of a data scientist to the specialized experience developed in various fields. Today, the figure of the data scientist pervades all professional and scientific domains, particularly in an era when data, which represent a wealth of knowledge, are widely accessible due to technological advancements. However, it is important to note that data, by their nature, are characterized by variability and uncertainty, requiring appropriate skills to provide reliable and effective answers.

4. Data scientists: Skills from the job market

Today's job market strongly seeks experienced data scientists to fill the need for data support at any level and function. There is no doubt, therefore, that a career in this field offers promising prospects for future generations, with good salaries and exponentially growing demand for such positions over the last decade.

However, it needs to be investigated again! It should ensure the fulfilment of expectations for a satisfying job that aligns with the skills of new generations, primarily acquired through bachelor's and master's programs. For workers, skills equate to employability and social mobility. Labor market operators often fail to recognize the actual abilities of data scientists and do not differentiate between various roles within the field of data science.

In this sense, it is interesting to report the main skills required by the job market for the different data scientist figures. According to the worldwide survey "State of Data Science 2022" performed by Anaconda Inc., the top five most important skills/areas of expertise in data science organizations are: engineering skills (38.12%), probability and statistics (33.26%), business knowledge (32.22%), communication skills (30.56%), and big data management (29.24%). The same survey also reports that most (62.51%) of organizations are at least moderately concerned about the potential impact of a talent shortage, generating a mismatch between the availability of new talents and the demand from companies in terms of both skills and numerical shortage.

In the view of understanding the potential mismatch between the demand and the offer in terms of data science skills, we also highlight which kinds of data are usually analyzed by data scientists and the methods used most regularly. The recent report "Data Science Salary Report 2023" by BigCloud (https://bigcloud.global) provides a detailed analysis of these issues based on a survey that includes 1300 responses from people involved in data science as workers (93%) or students (7%).

According to this report, data scientists mainly work on the following data typologies: 1) Relational; 2) Text; 3) Image; 4) Time series; 5) Video; 6) Audio; 7) Sensor. The main methods, according to this survey, are reported in Table 1.

The surveys conducted, however, reveal a data scientist figure that is still too oriented towards the use

Table 1
Main methods used in data science

Data science methods used most regularly	Percentage of respondents
Random forests	42%
Neural networks	41%
Logistic regression	40%
Decision trees	38%
Gradient boosted machines	36%
Ensemble methods	28%
Bayesian techniques	26%
Convolutional neural networks (CNNs)	25%
Support vector machines (SVMs)	20%
Recurrent neural networks (RNNs)	17%

of IT tools and machine learning software (Python), especially when considering the demand of companies that focus on business intelligence and digital marketing. On the other hand, there is a lack of surveys on the demand for data scientists in public administration and government institutions, as well as in research organizations. In more scientific contexts, there are only a few analyses on the demand and role of data scientists as an integrated figure in a multidisciplinary context with competences in the analysis of high-frequency, complex, structured, and unstructured data (e.g., analysts of genetic, physical, and environmental data).

5. Data scientists: Salaries and perspectives from the job market

As highlighted above, the relevance of the data scientist profession is emphasized by good salaries that are increasing in Europe, with the perspective of a rising career. A negative note is the still much lower percentage of female Data Scientists compared to their male colleagues. The report by BigCloud mentioned in the previous chapter also analyzes the salaries in Europe and the UK for Data Scientists. A meaningful comparison among countries is rather difficult, as the roles are often not comparable due to the diversity of positions and competences, as already pointed out. An attempt was made in this study to compare data on the average salary of 'Data Scientist' and 'Senior Data Scientist' median salary, average increase, and average bonus for the six countries considered: France, Germany, Italy, the Netherlands, Switzerland, and the UK. The data visualization was produced by us by homogenizing the currencies to Euros (Figs 1 and 2). The conversion rate for GBP £ and CHF was on the 20th of October 2022.

From this survey, it also emerges that 40% of workers are satisfied with their salary and that 62% had an increment of their salary in the past two years in the range 1–15%.

Further evidence to confirm the job market's interest in the data scientist role, comes from a comparison of salaries between statisticians and data scientists. Although these two roles partially overlap, salary data from the Economic Research Institute (https://www.erieri.com/), summarized in Fig. 3, reveal that data scientists have higher salaries than statisticians in all countries considered. These figures also suggest a potential increase from 6 percent to 14 percent over the next five years.

According to the 2020 U.S. Emerging Jobs Report from LinkedIn (https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/emerging-jobs-report/Emerging_Jobs_Report_U.S._FINAL.pdf), the data science field has topped the Emerging Jobs list for three years. It is a specialty that continues to grow significantly (35%) across all industries. The data in this report indicates that some of this growth can likely be attributed to the evolution of previously existing jobs, such as statisticians, and increased emphasis on data in academic research.

In Europe, the rising occupations have been evaluated by the LinkedIn reports "Jobs on the rise 2021." These look at the roles experiencing the highest growth between April and October 2020, compared to the previous year. The job categories are ranked by a combination of growth and size of demand. Focusing on the job category "Artificial Intelligence and Data Science" in France it is in 15th place with a 40% increase in recruitment in 2020 (with only 23% of women recruited). In Germany, the same job category is in 13th place, and Amazon recorded the newest hires in Germany. This is not surprising, as the company has benefited the most from the crisis this year and is now investing in the expansion of its e-commerce platform. In Spain, in December 2020, there was an announcement of a €600 million investment in artificial intelligence as part of its plan to transform the national economy. Therefore, it is no surprise that jobs in artificial intelligence and data science increased by 64% during 2020 and ranked

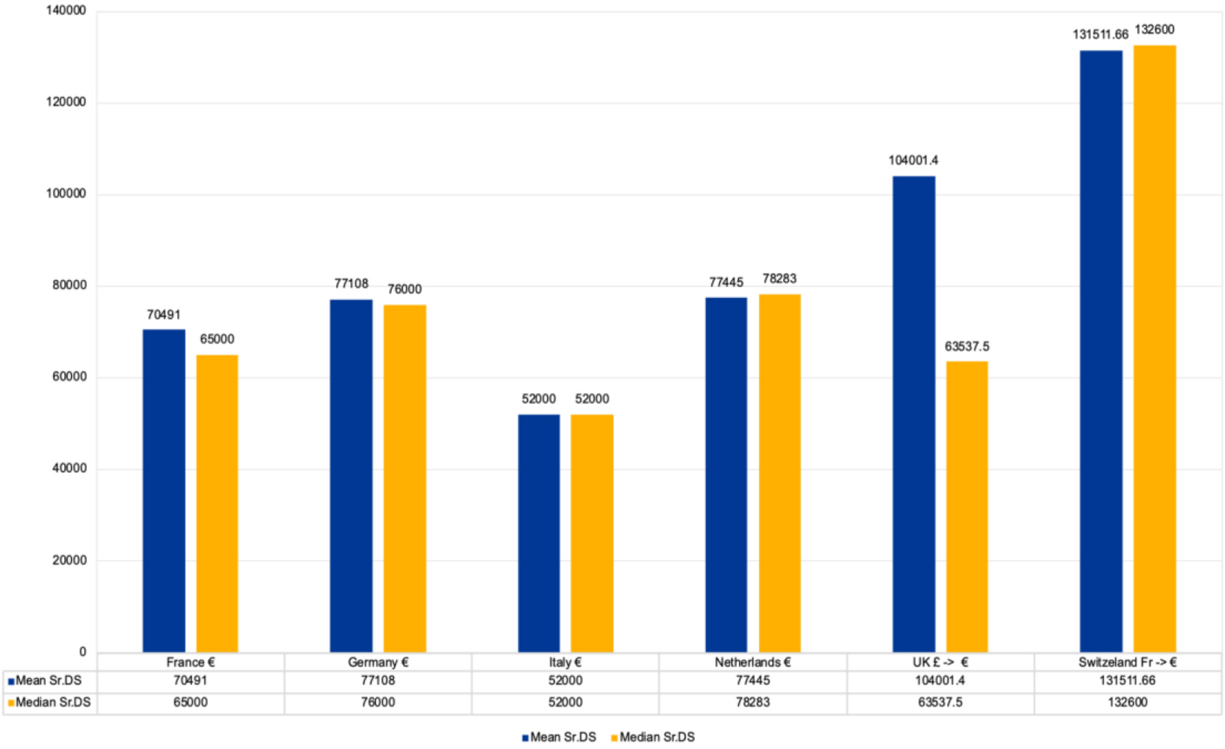


Fig. 1. Salary senior data scientist.

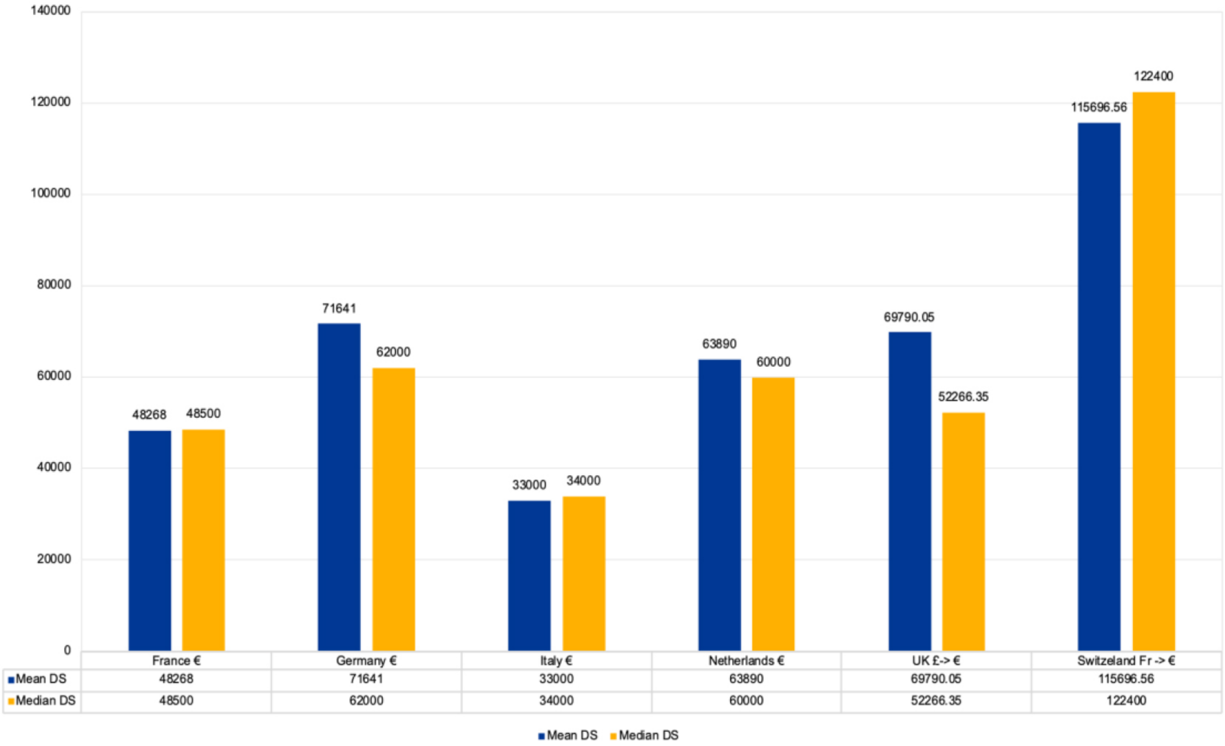
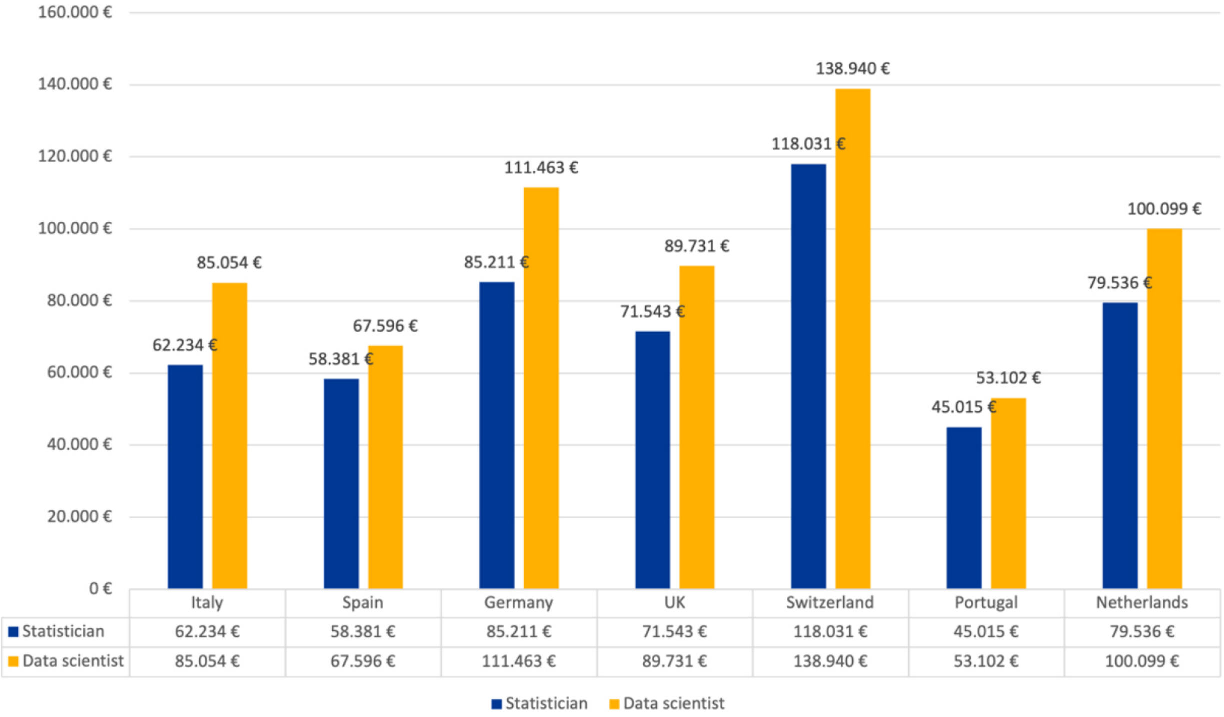


Fig. 2. Salary data scientist.



	Italy	Spain	Germany	UK	Switzerland	Portugal	Netherlands
■ Statistician	62.234 €	58.381 €	85.211 €	71.543 €	118.031 €	45.015 €	79.536 €
■ Data scientist	85.054 €	67.596 €	111.463 €	89.731 €	138.940 €	53.102 €	100.099 €

Fig. 3. Salary statisticians vs data scientists according to the economic research institute.

15th. In Sweden, the closely related category “Data & Analytics” is in 15th place with a 38% growth in 2020. This category saw more male hires (65%) than female hires. Finally, in Italy, this category is not among the top 15 positions.

6. The future of data science

Hadley Wickham [21] summarizes quite well the steps needed for a data science project: “I think there are three main steps in a data science project: you collect data (and questions), analyze it (using visualization and models), then communicate the results.”

In the following we will consider these three steps (data collection, data analysis, communication), by looking to the specific data science skills for a future statistician.

6.1. Collecting data

Often, data must be downloaded from the internet or from social media (Twitter, Instagram, Tripadvisor, ...). As mentioned before, job market requires the data scientist to be skilled on sophisticated and modern tools of data collection (use of API application programming interface, web scraping, etc.), to be prepared on using tools of data engineering related for example to storage huge amount of data, and to be able to integrate data coming from multiple disparate sources.

However, much more than the technologies and tools necessary to these tasks, we believe that the most important skill for a data scientist should be the ability to think about data, to evaluate the consequences and the aftermath of his possible choices. As an example, let’s consider the quote by Mike Loukides [22] claiming that using data is not really what people mean by data science: “a data application acquires its value from the data itself and creates more data as a result. It’s not just an application with data; it’s a data product. Data science enables the creation of data products”. This idea of making data as a result or a product is clearly connected with the idea that data are not coming from an experiment, but they are already available given the new technologies advancements, with all the risks connected with “swimming with data” and finds something that, in fact, is not present in the data, just because we keep looking into them. The quite famous quote by the Nobel prize 1991 in Economic Sciences Ronald H. Coase “If you torture the data long enough, Nature will always confess” remind us that we cannot be agnostic to the data collection process!

However, it is well known that a specificity of the Big Data era is the availability of many datasets covering large percentages of their respective populations, yet they were never intended to be probabilistic samples, but it clearly would be foolish to ignore such big datasets only because they are not probabilistic or representative. Xiao-Li-Meng [23] discusses these topics by claiming the big data paradox: "the more the data, the surer we fool ourselves" and proposing an identity linking data quantity, data quality and problem difficulty to measure the quality of whatever data science result.

6.2. Analysing data

The role of statistical models. Statistics found its way as the backbone to strengthen the scientific method in disciplines where it was already used and to extend it to other, most difficult disciplines. The statistical science on the practical ground develops methods to quantify uncertainty in mathematical models based.

Statistical science in the AI/big data world, somewhat dominated by software development and coding, can have a distinct and relevant role by working with the interdisciplinary teams involved in MLOps and data science, and at the same time present and apply methods rooted in the statistician mindset and way to solve problems.

Making causal inference is a related topic, the most common way to carry out causal inference is by using counterfactuals. To define a counterfactual in observational data with hundred or more variables is not straightforward, a review is proposed in [24]. Miller [25] makes a detailed description of the missing link between the current research on explanations from the fields of philosophy, psychology, and cognitive science. According to him there are three main aspects that an explainable AI system must have in order to achieve explainability: (1) people seek explanations of for why some event happened, instead of another, which suggests a need for counterfactual explanations; (2) recommendations can focus on a selective number of causes (not all of them), which suggests the need for causality in XAI (avoid the user to be overwhelmed by potential causes); and (3) explanations should consist in conversations and interactions with a user promoting an explanation process where the user engages in and learns the explanations.

Figure 3 proposes a taxonomy of explainable AI as put forward by Belle et al. in [26].

Some tools to support explainability are the Dalex package in R and python (https://dalex.drwhy.ai/) and the Captum package in the Pytorch ecosystem.

An interesting branch is that of causal transfer learning [27] based on a paradigm called Joint Causal Inference and rooted in Structural Causal Models. Another approach is that of the deconfounder proposed by David Blei and coll. [28]. Both methods are suitable for multiple cause problems.

Another interesting approach to causal analysis is the Targeted Minimum Loss Estimation framework. This latter takes the view that no model is absolutely the best when working with complex data sets, therefore, an ensemble method, termed SuperLearner is proposed as a first step and then a plug.in estimator is obtained for the causal parameter. The ensemble is not composed by many perturbed executions of the same algorithm, as occurs for instance in random forest, gradient boosting machine and XGboost, but by a set of algorithms selected just because of their heterogeneity. The key idea is that such diverse algorithms could best represent different aspects/part of the data and by combining them using weights obtained by a generalized cross-validation process, we could improve over the individual method. An example of model set might be, random forest, logistic regression, naïve Bayes, K-NN, Xgboost and CART.

When considering interpretability and explainability there are three different perspectives, namely predictive accuracy, descriptive accuracy and relevancy. **Predictive accuracy** is easy to measure using consolidated methods and metrics, classification matrices (also called confusion matrices), ROC curves and so on. **Descriptive accuracy** can be described as the degree to which an interpretation method objectively represents the relationships learned by the ML models. **Relevancy**: An interpretation is relevant if it provides insight for a particular type of users into a chosen domain problem. While descriptive accuracy is a hot topic, with many researchers working trying to devise methods to provide a better understanding of the learnt representation of the model, also with some new approaches based on advanced mathematics, relevancy is harder to define and formalize. The long and somewhat unique consulting practice that characterize the statistical procession might offer some interesting space to contribute to the endeavor towards a better understanding of complex models, especially when considering relevancy.

We can say that complex models require specific tools and methods to allow data scientists and analysts to understand or disentangle, also partially, how, and

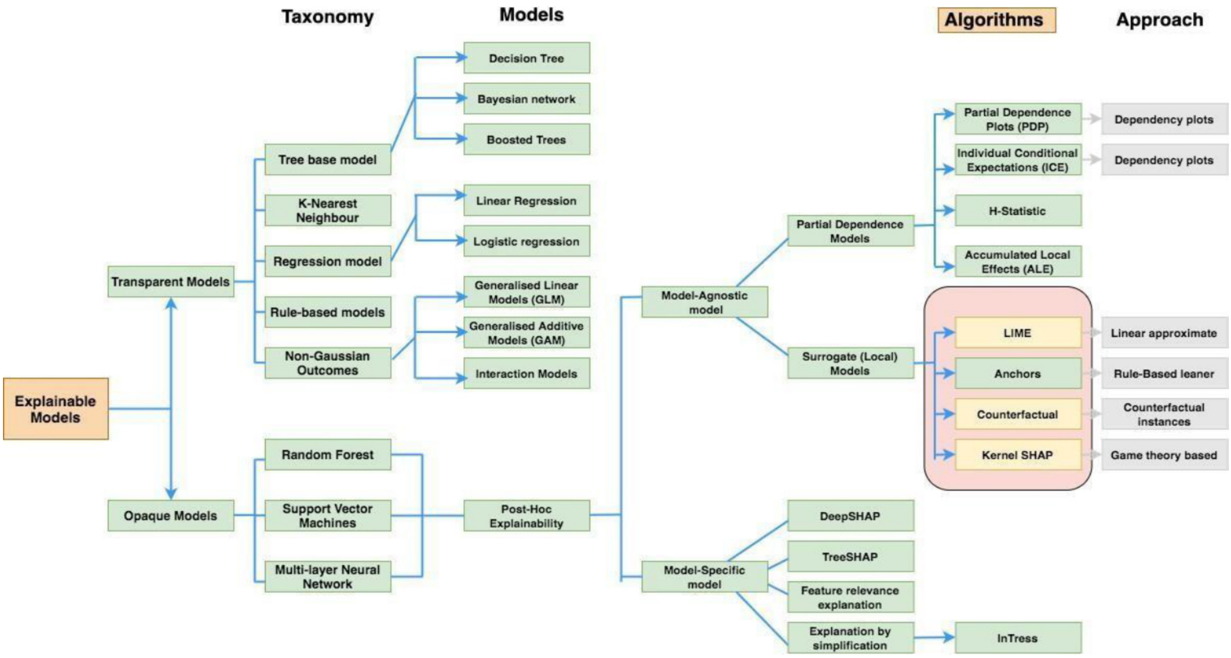


Fig. 4. Taxonomy of explainable AI based on the taxonomy proposed by Belle and Papantonis.

why they work well or not. Final users are, most of the times, not in a position to use the output of such tools, still being able to convey the message to them is essential for acceptance and real-world use of the model's results. The latter sentence highlights the necessity of developing special skills devoted to such a task.

6.3. *Communicate the results*

The last step of a data science project is the communication of what has been obtained by the analysis. Soft skills and abilities of data visualization are crucial as well as skills on storytelling when presenting results to clients. In addition, communication skills nowadays are also needed to show to the wide public the analytical results of a research: outreach skills as well as ability to manage social media are certainly relevant.

7. Conclusions

The advise of Leo Breiman [29] "Advising a young person today, I say: Take statistics, but remember that the great adventure of statistics is in gathering and using data to solve interesting and important real world problems." is probably even more current twenty years later, and it will also be valid in the future if statistics as a scientific discipline will be able to include all the skills provided also from different disciplines.

We have seen in previous sections that statisticians can significantly contribute to the improvement of data science projects by bringing their specific expertise. As shown in Section 5, according to Hadley Wickham [21], a data science project includes data collection, data analysis, and communication of results. Statisticians can impact each one of these steps by providing valuable insights and expertise. However, the main contribution of statisticians is to ensure the consistency of the whole chain from the definition of the goal to the communication of results. This includes planning the data collection to account for the representativeness of the population, choosing suitable data pre-processing methods whose impact on the analysis is correctly considered, selecting an analysis tool that considers both features and limitations, choosing appropriate visualization tools, and providing a critical and detailed interpretation of the results.

In this sense, statisticians adapt to the emerging and future challenges of data science assuming the role of a big brother who ensures the statistical soundness of the entire data processing chain. This includes opening black boxes and understanding their insides, as recently recalled by Brad Efron's opinion about Artificial Intelligence: "[. . .] and I sometimes think that the AI crowd is not critical enough – is a little too facile. The whole point of science is to open up black boxes, understand their insides, and build better boxes for the purposes of mankind" (Brad Efron, An Interview with Brad Efron of Stanford, www.b-eye-network.com/view/9947).

Acknowledgments

This article is the result of the discussion promoted by the Italian Statistical Society on the role that statisticians can play in the development of data science. The authors are grateful to colleagues who have contributed to enrich this work with their suggestions.

References

[1] Ho A, Nguyen A, Pafford JL, Slater R. A Data Science Approach to Defining a Data Scientist. SMU Data Science Review. 2019; 2(3): Article 4.

[2] Hayford FL. An Inquiry into the Nature and Causes of Statisticians. Journal of the American Statistical Association. 1941; 36(213): 1-10. doi: 10.1080/01621459.1941.10502065.

[3] Wild CJ, Utts JM, Horton NJ. What Is Statistics? In: Ben-Zvi D, Makar K, Garfield J, (eds) International Handbook of Research in Statistics Education. Springer International Handbooks of Education. Springer, Cham. 2018. doi: 10.1007/978-3-319-66195-7_1.

[4] Hassani H, Beneki C, Silva ES, Vandeput N, Madsen DO. The science of statistics versus data science: What is the future? Technological Forecasting and Social Change. 2021; 173: 121111.

[5] Masini E. Rethinking futures studies. Futures. 2006; 38: 1158-1168.

[6] Kuosa T. Evolution of futures studies. Futures. 2011; 43: 327-336.

[7] Son H. The history of Western futures studies: An exploration of the intellectual traditions and three-phase periodization. Futures. 2015; 66: 120-137.

[8] Bell W. Foundations of Futures Studies I: History, Purposes, Knowledge. New Brunswick, NJ: Transaction Publishers. 1997.

[9] Dator J. Futures studies. In Bainbridge WS, (Ed.). Leadership in science and technology. Thousand Oaks, California: Sage Reference Series (1), 2011, pp. 32-40.

[10] Gidley JM. The Future: A Very Short Introduction. Oxford University Press, (2017).

[11] Inayatullah S. Deconstructing and reconstructing the future: Predictive, cultural and critical epistemologies. Futures. 1990; 22(2): 115-141.

[12] Miller R. Transforming the Future: Anticipation in the 21st Century, London: Routledge (2018).

[13] Di Zio S, Tontodimamma A, Del Gobbo E, Fontanella L. Exploring the research dynamics of futures studies: An analysis of six top journals. Futures. 2023; 153: 103232. doi: 10.1016/j.futures.2023.103232.

[14] Miller R. Futures literacy: A hybrid strategic scenario method. Futures. 2007; 39: 341-362.

[15] Miller R, Poli R, Rossel P. The discipline of anticipation: Exploring key issues. IN: fumee.org. (2013).

[16] Johnson RB, Onwuegbuzie AJ. Mixed Methods Research: A Paradigm Whose Time Has Come, Educational Researcher. 2004; 33(7): 14-26.

[17] Sale JEM, Lohfeld LH, Brazil K. Revisiting the quantitative-qualitative debate: implications for mixed-methods research. Quality and Quantity. 2002; 36(1): 43-53.

[18] Tytler R, Bridgstock R, White P, Mather D, McCandless T, Grant-Iramu M. 100 jobs of the future. Published by Deakin University. (2019).

[19] Cleveland WS. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. International Statistical Review/Revue Internationale de Statistique. 2001; 69(1): 21-26.

[20] Baumer B. A Data Science Course for Undergraduates: Thinking With Data, The American Statistician. 2015; 69(4): 334-342. doi: 10.1080/00031305.2015.1081105.

[21] Wickham H. Tidy Data. Journal of Statistical Software. 2014; 59(10): 1-23. doi: 10.18637/jss.v059.i10.

[22] Loukides M. What Is Data Science? O'Reilly Media, Inc. ISBN: 9781449336097; (2011).

[23] Xiao Li M. Statistical Paradises and Paradoxes in Big Data (I). The Annals of Applied Statistics. 2018; 12(2): 685-726. doi: 10.1214/18-AOAS1161SF.

[24] Chou YL, Moreira C, Bruza P, Ouyang C, Jorge J. Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications arxivorg/abs/2103.04244. 2021.

[25] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. 2019; 267: 1-38.

[26] Belle V, Papantonis I. Principles and practice of explainable machine learning. Frontiers in big Data. 2021; 39.

[27] Mooij JM, Magliacane S, Claassen T. Joint Causal Inference from Multiple Contexts. Journal of Machine Learning Research. 2020; 21: 1-108.

[28] Wang Y, Blei DM. The Blessing of Multiple Causes. Journal of the American Statistical Association. 2020, pp. 1574-1596.

[29] Olshen RA. Conversation with Leo Breiman. Statistical Science. 2001; 16(2): 184-198.