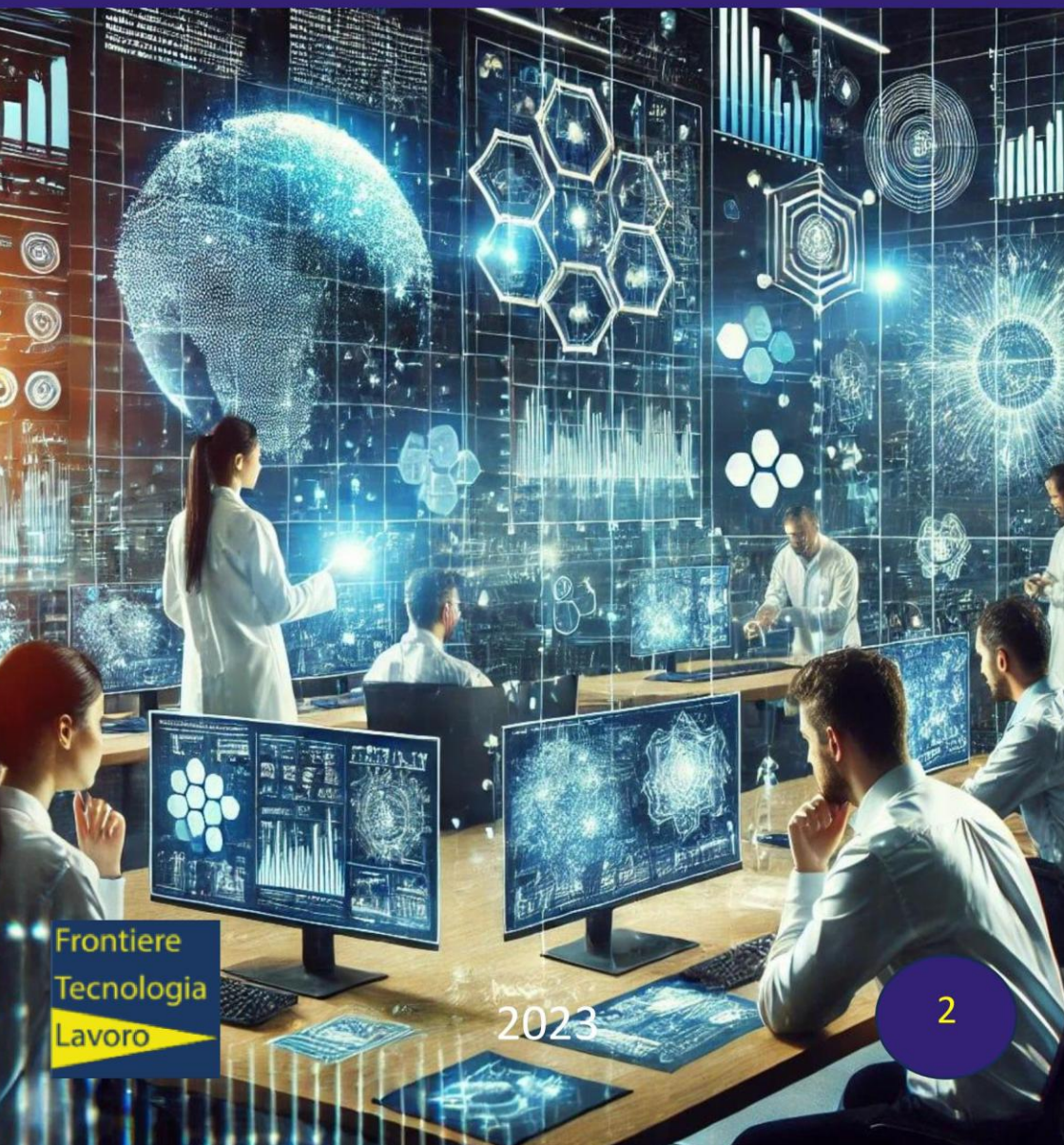


MASSIMO TAMIATTI

THE DATA SCIENTIST



Frontiere
Tecnologia
Lavoro

2023

2

What is it? This is a scientific dissemination website that studies the impact of new technologies on the world of work. Changes today are very fast and exponential. Professions are not static but very dynamic, so the ongoing evolutions can make them obsolete, transform them, or create entirely new ones. What skills will be necessary to face all this? Upskilling and reskilling, but not only.

Why? After 25 years of research, I have realized that to go beyond the boundaries, numbers are not enough. Traditional approaches are obsolete and tied to the old way of doing research. Today, we need to study the weak signals that are right before our eyes but that we see without really noticing. Stopping to consider these signals and imagining multiple possible futures will be the task of Frontiere. Traditional sources respond in real-time but only photograph the present. Today, however, there is a need to make predictions at least five years ahead, without neglecting the past and without forgetting that every topic addressed has a history to consider.

Who is it for? It's easy to make long-term predictions that cannot be verified. I am convinced that the research world needs an immediate confrontation with the world of education and orientation, and with young people who risk studying for professions that may not exist in a few years. They will be here to exchange opinions with the undecided, the disoriented, and those who intend to start over. Young people need to understand that there is a world above and a world below in the job market. A world of those who program algorithms (few and well-paid) and those who work for algorithms (many and poorly paid). We need to understand what to do to escape a job market that risks condemning young people to €1,000 per month salaries for life.

How will it be? Each topic will be addressed with light publications, with a strong emphasis on temporal depth. The publications will be accompanied by short videos (elivri) that get straight to the point. There will be an open chat for young people on the future scenarios of work, the profiles required, and the necessary skills.



Linkedin: <https://it.linkedin.com/in/massimo-tamiatti-77605168>

Facebook: <https://www.facebook.com/profile.php?id=100006516969397>

Telegram: <https://t.me/frontieretecnologialavoro>

DATA SCIENTIST

THE PROFESSIONS OF INNOVATION

The developments of Artificial Intelligence (AI) are certainly driven by Big Data. The exponential growth of devices and digital technologies (smartphones, computers, IoT sensors, etc.), online activities (including purchases, video viewing, music streaming), social media and sharing platforms, the evolution of storage technologies, and data processing (distributed databases and cloud storage systems) have contributed to the production of Big Data in recent years. The availability of large amounts of data is fundamental for training AI models.

But what exactly are Big Data?

The definition of Big Data refers to data that contain greater variety, arrive in increasing volumes, and with greater speed. This concept is also known as the three Vs. In simple terms, Big Data are larger and more complex datasets, mainly coming from new data sources. These datasets are so voluminous that traditional data processing software is unable to handle them. But these enormous volumes of data can be used to address business problems that could not have been addressed before. In 2001, Doug Laney, then Vice President and Service Director of Meta Group, described in a report on Big Data the "3V model": Volume, Velocity, and Variety. A simple and concise model to define the new data generated by the increase in information sources and, more generally, by the evolution of technologies. Laney's paradigm was later enriched by the variables of Veracity and Variability, and

for this reason, we speak of the "5Vs" of Big Data. Over the years, however, a sixth V has been added, focused on the value generated by the data.

The three Vs of Big Data

- Volume

The quantity of data is important. With Big Data, large volumes of unstructured and low-density data must be processed. It could be data of different value, from Twitter feeds to those of a web page, both on a mobile app and on sensor-enabled devices. For some organizations, it could be tens of terabytes of data; for others, even hundreds of petabytes(1).

- Velocity

This refers to the speed at which data is received and (perhaps) analyzed, for the consequent possible reaction. Normally, the highest speed of data flows directly into memory rather than being written to disk. Some Internet-enabled products work in real-time and require timely evaluations.

- Variety

This refers to the many types of data available. Traditional data types were structured and perfectly suited to a relational database. With the advent of Big Data, unstructured data(2) also arrives. Unstructured and semi-structured data types, such as text, audio, and video, require additional preliminary processing to derive meaning and support metadata(3).

The model has recently been enriched with two new Vs that define how these new data should be used:

- Veracity

Among industry professionals, some say, "Bad data is worse than no data." Data must be reliable and tell the truth. With Big Data, this challenge is even more difficult to address: data management technologies change, the speed at which data is collected increases, and sources multiply. However, the quality and integrity of information remain a fundamental pillar for useful and reliable analysis.

- Variability

As we have already mentioned, data is always much more and presents itself in different formats and from different contexts. The mutability of its meaning is an aspect to consider when data is interpreted, even more so if done by a user working in a business line and not just the Data Scientist.

A large amount of data is too much for our PC. This is why another new technology, Cloud Computing, becomes strategic: it allows loading an enormous amount of data into the Cloud and then retrieving only what is of interest. Cloud Computing is the distribution via the internet of essential services such as servers, databases, and software. A provider essentially makes storage, processing, and data transmission services available to its users on demand. In this way, companies and individuals no longer need to download software onto their computers (or other devices) to use certain programs or applications: they can do so in the Cloud, remotely, eliminating the need for physical servers (saving on related purchase, management, maintenance, and malfunction risk costs)(4). The combination of big data and cloud computing is revolutionizing the world of technology and business. Cloud computing provides a scalable and flexible infrastructure capable of handling large amounts of data, while big data provides the insights and analyses necessary to fully exploit such data.

From mines to refineries

Big Data has been defined in recent years as the new oil or the new gold, that is, an invaluable source of value. It is possible to compare them to immense mines from which information is extracted, processed, and, as in a refinery, value is extracted. However, merely collecting data, even using the best technologies available on the market, does not guarantee the quality of the information and, above all, does not ensure the extraction of knowledge. Talking about data, information, and knowledge means talking about different aspects related to each other. By definition, data is a coded representation of an entity, a phenomenon, a transaction, or an event. Information is the result of a data analysis process that often has meaning only for those who operate in the domain of data generation. Knowledge is obtained when a person uses information to make decisions and carry out conscious actions, that is, when information is used to be put "into practice." To implement this process and ensure that Big Data can be transformed into information to be used in business processes, building knowledge to improve performance, analytical tools are necessary. It is essential to consider their value. For this reason, Big Data Analytics(5) methodologies are fundamental. Through their use, a company can extract "value," that is, make more informed, timely, and conscious decisions from the vast world of Big Data. It is increasingly clear that the quality of data is as important as its quantity. The challenge is not only to collect massive data but also to ensure that it is aligned with business objectives and thus generates value(6). Due to their breadth and heterogeneity, data needs to be "extracted" from the various sources from which it comes (websites, social media, IoT devices present in stores, etc.) and then undergo various processes (management, integration, harmonization, analysis, etc.) capable of transforming it into value for the business.

The discipline that deals with data

The discipline that deals with data is "Data Science." We are surrounded by Big Data, Open Data, cloud, and databases produced by devices and IoT sensors, voice assistants, GPS, smartphones, etc. It is in this constantly evolving context, practically unstoppable, that Data Science (data science(7)) emerges, a complex discipline that embraces the entire information production chain, from collection to processing and critical analysis of results, considered strategic for the future. Data Science is an interdisciplinary field that uses scientific methods, algorithms, and systems to extract knowledge from data in various forms. Data Science combines concepts and techniques from different disciplines, including mathematics, statistics, computer science, and specific domains of the sector, to understand data, identify patterns, make predictions, and support informed decision-making. This field is widely used in various sectors, including business, science, healthcare, finance, marketing, and more. Data Science, unlike other scientific disciplines, does not use data to study a human or natural phenomenon; its object of study is the data itself. Studying it means knowing and understanding the world around us, understanding the causal links between events, predicting behaviors, anticipating needs, and reducing errors. With data, for example, we can truly understand the profiles of customers on an e-commerce platform or define market trends. For this reason, the areas of application of Data Science are numerous and constantly evolving—there is IoT, Machine Learning, AI, and perhaps even Quantum Computing in the future, and with them, the required skills are also growing and changing(8). Data Science is still a frontier discipline that we are building today. It is impossible to say what path this discipline will take in the near future or what skills or attitudes it will require. We are undoubtedly in the midst of change and immersed in an ocean of information and questions waiting to be processed and satisfied. What is certain, however, is that those who work with data never get bored(9). The year 2023 was a year of great

change in the world of Data Science and AI. We witnessed the boom of generative AI in 2023 thanks to language models like ChatGPT, Claude, and the still-announced Google Gemini. This has had an impact on the profile of the Data Scientist, which is rapidly changing. In this world of Big Data, many important figures gravitate, not just Data Scientists(10). Often working in synergy, especially with the Data Engineer(11) and the Data Analyst(12). But certainly, Data Scientists are a key figure. And it is important to clarify that it is not the evolution of the Data Analyst, as one might think, because while the Data Analyst primarily works on well-defined data to answer business questions, the Data Scientist tackles more complex problems, using advanced analysis techniques to generate innovative solutions. Therefore, the Data Scientist cannot be considered an evolution of the Data Analyst but rather a distinct professional figure with a different set of skills and responsibilities(13). When trying to understand the scenarios produced by new technologies and, in particular, the jobs that derive from them, we know for certain that many jobs will be eliminated, others will be transformed, and others will be completely new, that is, they did not exist previously. In this case, we are faced with a professional figure that is transforming. The role of Data Scientist began to emerge as a distinct profession toward the end of the 2000s and the beginning of the 2010s when the accumulation of large amounts of digital data made it necessary to have a new professional figure capable of extracting value from this data through statistical analysis, Machine Learning, and other techniques.

However, the term "Data Scientist" may have been coined only in 2008 by DJ Patil and Jeff Hammerbacher, both of whom worked at LinkedIn and Facebook at the time. Since then, more and more companies recognize the value of having data analysis experts on their teams. The Data Scientist is a figure that does not come from the future but has a crucial role in shaping it. This profession, like few others, is linked to the technological and scientific development and progress of today's world and the world of tomorrow(14).

It is the first time in human history that we are collecting more data than we are able to use, and this overproduction of information has created the need to have a specialized expert who can exploit the knowledge in favor of companies. This is the Data Scientist, considered a scientist: a professional profile that has appeared in company organizational charts only a few years ago but will have an increasingly important role(15). The knowledge and information acquired just three years ago make us understand that while the technical, mathematical, and scientific bases have remained substantially unchanged, the development of generative AI is rapidly making most professional profiles obsolete. This is due to the speed at which AI and Data Science are progressing in parallel, making it difficult to compare them to other fields. However, it is important to note that other sectors, such as the medical sector, with recent developments in vaccines, are advancing significantly because they are now intrinsically linked to the development of AI.

But what are the motivations that drive someone to pursue this career? Certainly, the awareness of being part of a profession linked to the technological and scientific progress of today's world, but not only that of the world that will be tomorrow: the feeling of being part of the future that awaits us and contributing to its development, because a person who works on predictions always has their head in the future. The professions that revolve around the Data Science sector will be increasingly linked to the development of AI. AIs represent one of the greatest technological changes of all time and will affect every aspect of our lives. Therefore, we must prepare to face the different domains of application in every aspect of our lives. The world of AI and Data Science is growing exponentially. Three years ago, generative AIs were not even considered in university master's programs; now, they are at the base of the development of the world and progress and already influence the lives of all of us.

When I grow up, I want to be a Miner...

Defined by the Harvard Business Review as the most fascinating profession of the 21st century, that of the Data Scientist is an increasingly sought-after position by companies in Italy and abroad(16). The Data Science sector can be considered a discipline within the broader field of AI; it particularly focuses on extracting information and knowledge through data analysis(17). Certainly, the data scientist is one of the professions of the near future. The future is made of data; we have seen this in the part related to Big Data, and specializing in some discipline that embraces them seems to be the smartest path, but a (simple) degree in mathematics or statistics is not enough to face the complexity of Big Data Analytics. Multidisciplinary preparation with diverse skills mixed wisely is required. It is a somewhat new figure born precisely from the need to manage data. But what does it do? It collects all the data available to the company to create useful information for making decisions and being more competitive in the market. The Data Scientist has a degree in mathematics, engineering, or statistics, in short, in STEM disciplines(18), but must also have a deep knowledge of programming languages(19).

But what is the path that allows you to become a Data Scientist?

It's a bit like running a marathon; it takes a lot of time and, above all, study and dedication. It is not something that allows you to find a job immediately after a six-month course. The path is not substitutive but integrative, and at the base, there must be at least a three-year degree in a STEM discipline. Solid foundations in programming and statistics must be built. As for programming(20), the best language is undoubtedly Python. It is a generational language with which you can do whatever you want, not just statistical analysis. As for Statistics, the focus should be on descriptive and

inferential statistics; additionally, it is necessary to have the basics to perform hypothesis tests and more than a smattering of probabilities and regression analyses. It is important to focus on EDA (Exploratory Data Analysis). This acronym refers to the set of critical processes at the beginning of a Data Science process, which aims to identify particular patterns(21) within the data, to identify anomalies, or simply to understand what the variables are and how they behave. It is very important to learn about the data you are working on before talking about Machine Learning and AI. Once the theory has been studied, practice follows, that is, the exploration of data using generative AI tools. A ChatGPT with a "Code Interpreter"(22) will suffice to load your CSV(23) and perform analyses supported by AI. This will not only speed up learning but also provide a kind of coach to help you in the initial phase of Data Science, that is, studying and understanding the data you are working on. In particular, staying updated on the latest trends and tools in the world of AI. Once the rules of statistics, programming, and data analysis have been learned, one can focus on learning Machine Learning(ML)(24). Here, one must start with the basics of mathematics. Two things must be learned primarily: linear algebra and calculus, and therefore derivatives. This is the mathematics that will be needed at least to master the basics of Machine Learning. It is then necessary to understand how a Machine Learning project works. It will be important to understand what the techniques are for splitting a "dataset into training and test sets" and a "validation set"(25). It is also necessary to delve into some classic ML problems: the issue of imbalanced classes; understanding the meaning of evaluation metrics such as accuracy, precision, and recall; going deeper to understand what a ROC curve is and what a confusion matrix(26) is. Then it will be necessary to understand what the main Machine Learning algorithms are and how they work, both supervised and unsupervised types. Then choose a topic of interest and dive into Kaggle(27). After mastering Machine Learning techniques, it will be necessary to enter the field of Deep

Learning(28). It will be necessary to understand the basics of how neural networks work and learn what the main Python frameworks (structures) are. Once the reconciliation techniques have been learned, familiarity with the concept of Transfer Learning(29) will be gained, which will certainly be very useful, to then deepen the main thematic areas such as computer vision, that is, how AI is applied to images and videos, the basics of natural language, and therefore AI applied to texts. Once the models have been learned, it will be possible to understand how to put them into production. That is, going from the code written in your notebook that only you can see to publicizing your creations to the world. The last thing will be the creation of a portfolio. In any case, Machine Learning, AI, knowledge of programming languages, and mastery in database management allow the Data Scientist to best perform their work and generate value for the business.

But what exactly does a Data Scientist do?

The Data Scientist is a professional who has entered the business landscape only in recent years. This role is closely linked to the massive spread of data, even in real-time, which can be analyzed to extract fundamental information for business development. The Data Scientist collects data from various sources, such as business applications, databases, web analytics, social media, and electronic documents. Subsequently, they structure this heterogeneous data into organized and accessible formats. They are a specialized figure who continuously updates and has a deep knowledge of programming languages such as Python and R(30). They then use software, algorithms, statistical methods, and Machine Learning tools for data analysis and can develop Machine Learning and Deep Learning algorithms.

They explore databases, perform specific queries(31), and extract useful information (Data Mining). This process allows identifying, within a large

amount of data, only those specific pieces of information that represent value for the company. The Data Scientist builds correlation models, demonstrates causality, develops scenarios, and creates predictive models. They can therefore use statistical models for analysis and predictions. Unlike a Data Analyst, the Data Scientist does not just seek the solution to a specific problem. They must process Big Data in various ways to discover hidden and unexpected trends, opportunities, and critical issues that can have a significant impact on the business. Therefore, it is essential that the Data Scientist has a deep understanding of business dynamics and the functioning of the company for which they work. In addition to extracting information from data, the Data Scientist must make it usable for managers. The creation of reports and data visualization are an integral part of their work. They understand market dynamics through quantitative analysis and produce clear visualizations of data. In conclusion, the Data Scientist can be considered a hybrid figure between a statistician and a computer scientist. They must also possess a perfect mix of technical skills and soft skills. In addition to technical skills, soft skills such as curiosity, effective communication, teamwork, and problem-solving are fundamental when working in heterogeneous teams that are constantly evolving and connected.

What will be the role of Data Scientists in the coming years?

It is important to always be prepared to face changes in trajectories to remain competitive within the sector. If until a few years ago, the task of Data Scientists was to analyze data, train Machine Learning models, and collaborate with Business Experts to understand which projects to develop, today we must acknowledge that models have become increasingly complex. Therefore, in 2023, it is common to use pre-trained models such as LLM(32) or various types of predictive models and compare them to determine which

is the best to use. In this context, the model-building phase has not become obsolete but has certainly undergone profound changes. Furthermore, the entire phase of data analysis and exploration is no longer conducted as in the past. Until a few years ago, Data Science was mainly entrusted to Expert Analysts, people capable of working with tables, numbers, graphs, and who had to be able to write code. Today, however, the work of the Data Scientist is confronted with the innovations brought by generative AI. But let's give a practical example: "The other day we had to analyze the data of a YouTube channel; it was about 180 videos published in the last 2 years, and after exporting the file to CSV, we did not insert the data into a Python notebook to analyze it from scratch, but we loaded it directly into the 'Code Interpreter' environment of ChatGPT(33). Then we started a dialogue with an AI to extract information, generate graphs, describe the data, and find strategies to maximize the growth of this channel. Therefore, in this specific case, we performed work that falls halfway between a Data Analyst and a Data Scientist using 'GPT for' as support. It seemed incredible to do all this without writing a single line of code, or rather, the Python code was generated directly by the language model on the AI guided only by a simple command. This allowed us to complete the analysis in half an hour; an analysis that would otherwise have taken two days if we had had to write the code from scratch, having to import the data, clean it, and generate the graphs. If a Data Scientist does not take into account the tools of generative AI, in the short term, they will no longer be competitive in their sector. A Data Scientist, a Data Analyst anchored to old working methods, that is, for example, building graphs from scratch or conducting exploratory analyses without the help of a language model, will be extremely uncompetitive and slower in the coming months and therefore less sought after by companies, especially innovative ones that keep up with the times. Therefore, what will become fundamental for Data Scientists from 2024 onward will be to learn the tools based on LLM for code writing and, at the same time, learn Auto Machine Learning tools

that allow selecting the best model, avoiding losing hours in project development, making the process faster and more performant because it starts from a pre-training perhaps done by a Big Tech, to verify the available data, which guarantees better performance. Another important observation is that the work of the Data Scientist will become increasingly strategic. If in the past they dedicated little time to code writing compared to searching for solutions for the business, in the future, it is imaginable that they will dedicate even less time to code writing and will be more involved in dialogue with the Software Engineer to engineer the solutions. This implies that the Data Scientist will be closer to a programmer who implements models and discusses the engineering of solutions rather than building code from scratch. The positive aspect of this evolution will be the greater value attributed to critical thinking, creativity, and intuition in decision-making compared to operational work, which is often the most mechanical and least human part. In short, we will free ourselves from the most repetitive activities. Instead of dedicating days to the exploration and writing of code for routine analyses when new data arrives, we will be able to focus on more complex challenges, such as the strategic interpretations of data and innovative decisions on how to use them to increase added value within an organization. At the same time, if you are a Data Scientist who loves to write "a lot," it might be more advantageous to dedicate yourself to putting Machine Learning models into production or Data Engineering, where there is still a lot of code to write. In summary, Data Scientists will increasingly collaborate with AI models to optimize work. The message is clear: we should not consider AI as a threat to our work but as a partner to work better and faster(34).

Where do Data Scientists find work? Data Scientists are employed in the finance, retail, e-commerce, and marketing sectors. However, this professional figure is becoming increasingly widespread in various sectors: there are job opportunities in manufacturing, logistics and transportation, IT

and telecommunications companies, banks and insurance companies, scientific fields, healthcare, and Public Administration. The work of the Data Scientist is mainly carried out in the office, although some companies offer the possibility of working remotely with flexible hours. Therefore, it is particularly important that the Data Scientist has a deep understanding of business dynamics and the functioning of the company for which they work. This allows them to identify the challenges and problems they must face, as well as the solutions they can find during data analysis. For example, a Data Scientist can help a company better define its market niche, segment its target audience, optimize marketing strategies, or rethink product development processes based on actual data (data-driven). It is also important that the Data Scientist makes the extracted information accessible to managers: another part of their work is related to data visualization (data visualization) and report creation. Using graphs, maps, tables, and diagrams, the Data Scientist makes their discoveries more understandable and usable, demonstrating the need to take certain actions to increase the company's competitiveness(35). In Italy, in 2022, 49% of large companies had at least one Data Scientist on staff, and their number grew by 28% of the sample(36). Despite a widespread willingness to experiment, few Italian companies today can be considered truly data-oriented, that is, capable of fully exploiting the available data to improve their performance. For Data Science to have a concrete impact, it is necessary to create a data culture that involves more and more workers at all levels in the daily use of insights (intuitions) and analysis results(37).

The possible quantum leap

Starting from the premise that the future is not predicted but created, let's try to imagine the evolution of the work of the Data Scientist even in such a dynamic and rapidly transforming world as that of Machine Learning and AI.

It is known that one of the main tasks assigned to Data Scientists is to "predict" the future. However, the future of Data Scientists as a profession is not at all predictable. New technologies are profoundly changing the responsibilities and activities carried out by Data Scientists, and further transformations could totally change the nature of this work. Some Data Scientists may have the opportunity to make a "quantum leap." This is because quantum computing will have a significant impact on data science professions. Quantum computers will be able to process large amounts of data efficiently and effectively. To use a quantum computer, classical algorithms cannot be used. It is necessary to invent new algorithms that exploit the properties of quantum mechanics and allow information to be extracted from data in this way. Quantum data scientists must therefore understand how to use quantum algorithms. In particular, quantum data scientists must be able to understand the basic principles of quantum mechanics, understand how quantum computers work, understand how to program a quantum computer, and, most importantly, understand how to use a quantum algorithm to solve a particular real problem(38). I would say that in this case, we are on unknown and unexplored terrain, certainly full of possibilities and technological development but also full of pitfalls.

Bibliography and Webography

1. The petabyte is the unit of measurement for large data centers (data processing centers). A petabyte is equivalent to one million gigabytes and can store an amount of data that is simply unimaginable for the average user.
2. Structured data and unstructured data represent two broad categories of data that can be collected. Structured data is data that fits easily into data tables and includes discrete data types such as numbers, short text, and dates. Unstructured data does not fit perfectly into a data table due to its size or nature: for example, audio and video files and large text documents.
3. Mirella Castigni, "Volume, variety, velocity: the 3 Vs of Big Data" in "zerounoweb.it" on September 20, 2022.<
<<https://www.zerounoweb.it/big-data/volume-varietà-velocità-le-3-v-dei-big-data/>>
4. "Cloud Computing and Big Data: a perfect combination" in "doxee.com."
<<https://www.doxee.com/it/blog/tecnologia/cloud-computing-e-big-data/>>
5. Analytics represent the process of discovering, interpreting, and communicating significant patterns in data. In short, analytics help us see meaningful data that would otherwise go unnoticed.
6. Alessandro Piva, "The 5Vs of Big Data: from Volume to Value" in "blog.osservatori.net" on December 21, 2023.
<https://blog.osservatori.net/it_it/le-5v-dei-big-data>
7. Data Science emerged as an interdisciplinary field in the 2000s, with the term "data scientist" attributed to DJ Patil and Jeff Hammerbacher in 2008.
8. Luca Malinverno, Data Scientist "The era of data, AI, and Analytics, PLM and Manufacturing: professionals and experiences" on "youtube.com" in 2021.
<<https://www.youtube.com/watch?v=CWYbWryyVKA>>
9. Franco Pigoli, Data Scientist and Education Manager in "dms.synergo.es"
<<https://dms.synergo.es/partner/franco-pigoli-di-porini-tra-i-protagonisti-del-data-management-summititaly/>>

10. Important figures in the world of Big Data include Data Engineer, Machine Learning Engineer, Data Analyst, Data Architect.
11. Data Engineer is responsible for the Data Pipeline and provides data to the Data Scientist in a format suitable for analysis.
12. Data Analyst manages descriptive activities and constitutes the link with business lines.
13. Irene di Deo, "Data Analyst, who they are, what they do, and how much they earn" in "blog.osservatori.net" on February 21, 2019.
<https://blog.osservatori.net/it_it/data-analyst-cosa-fa-stipendio>
14. Alexander Paolo Zeisberg, "Introduction to the world of data science" on "youtube.com."
15. Maria Grazia Tecchia, "One of the most sought-after profiles..." in "digitalic.it" on December 1, 2017.
<<https://www.digitalic.it/hardware-software/data-scientist-chi-e>>
16. "Data scientist: what they do and why they are important for companies" in "beta80group.it" on March 1, 2022.
<<https://news.beta80group.it/data-scientist-cosa-fa-e-perche-e-importante-per-le-aziende>>
17. "Artificial Intelligence and Data Science: what they are and how they work" in "talentgarden.it" on August 9, 2023.
<<https://blog.talentgarden.com/it/blog/data/intelligenza-artificiale-e-data-science-cosa-sono-e-comefunzionano>>
18. The term STEM is an acronym formed from the initials of four different disciplines in English: Science, Technology, Engineering, and Mathematics. It was first used in 2001 during a conference of the National Science Foundation to indicate the set of scientific subjects for which greater dissemination was hoped. Science, technology, engineering, and mathematics were the disciplines on which massive investment was necessary to meet the needs of the economic fabric.

19. "The 16 most used programming languages in 2024" in "shopify.com" on August 19, 2024.

<<https://www.shopify.com/it/blog/linguaggi-di-programmazione>>
20. Relevant programming languages include SQL, Python, and R.
21. A pattern is a synthetic representation of a set of data that expresses a recurring or exceptional model.
22. OpenAI's Code Interpreter represents a significant step forward compared to the traditional approach of ChatGPT. The latter is based on a large language model (LLM) to predict the next word in a sequence and guess an answer.
23. CSVs (Comma Separated Values, "comma-separated values") are simple text files that contain lists of data.
24. Machine learning is a data analysis method that automates the construction of analytical models. It is a branch of Artificial Intelligence and is based on the idea that systems can learn from data, identify patterns autonomously, and make decisions with minimal human intervention.
25. The division of a dataset into training, validation, and test sets is a crucial step in building a machine learning model because it allows the model to be trained on one set, fine-tuned on another, and then evaluated on a final set. Larger datasets benefit from a larger portion of training data, mixing and ensuring a normal distribution to avoid biases. During this process, it is necessary to split the entire dataset to succeed in our task.
26. Precision (precision) and recall (sensitivity) are two indicators used in machine learning to evaluate the quality of a decision-making or predictive model. Accuracy is the parameter of data accuracy and is related to the ROC curve, that is, a statistical technique that measures the accuracy of a diagnostic test over the entire range of possible values. A confusion matrix (or error matrix) is a method of visualization for the results of a classification algorithm.
27. Kaggle is an online community of Data Scientists and Machine Learning enthusiasts, founded in 2010.

28. Deep learning, a branch of machine learning, trains computers to perform tasks similar to those of humans, such as speech recognition, image identification, and prediction making. It improves the ability to classify, recognize, detect, and describe through data.
29. Transductive learning involves transferring knowledge from a specific source domain to a different but related target domain, with the primary goal of the target domain. It is particularly useful when labeled data from the target domain is scarce or absent.
30. Python is an extremely versatile and popular programming language used for various purposes, including web development, software, graphical interfaces, video games, data analysis, and automation of activities. R is a programming language specific to statistics and data processing and is based on concepts such as normal distribution, statistical tests, models, and regression.
31. In computing, a query, or interrogation, is a command written by the user to extract information from the content of a database.
32. A Large Language Model (LLM) is an advanced AI technology focused on understanding and analyzing text.
33. The Code Interpreter works differently from traditional ChatGPT: when a question is asked, it writes and executes a piece of computer code to find the answer. A process that allows for greater precision and reliability.
34. Alexander Paolo Zeisberg, "Introduction to the world of data science" on "youtube.com."
35. Patrizia Licata, "Data Scientist: the data expert who transforms information into value" in "peoplechange360.it" on October 20, 2023.

<<https://www.peoplechange360.it/futuro-del-lavoro/data-scientist-identikit-professionista-piu-ricercato-data-economy/>>
36. Patrizia Licata, "Data Scientist: the data expert who transforms information into value" in "peoplechange360.it" on October 20, 2023.

<<https://www.peoplechange360.it/futuro-del-lavoro/data-scientist-identikit-professionista-piu-ricercato-data-economy/it:»\:text=le\%20sfide\%20aziendali.,Anche\%20in\%20Italia\%20le\%20gra>>

ndi\%20aziende\%20assumono\%20gli\%20specialisti\%20dei,campione\%20(Fonte\%3A\%20PoliMi).>

37. "Big Data & Business Analytics Observatory 2022, investments grow by 20%" in "zerounoweb.it" on November 14, 2022.

<<https://www.zerounoweb.it/big-data/osservatorio-big-data-business-analytics-2022-gli-investimenti-crescono-del-20/>>

38. Fabio Moioli, "How the work of the data scientist evolves: predictions" in "agendadigitale.eu" on December 1, 2022.

<<https://www.agendadigitale.eu/cultura-digitale/come-si-evolve-il-lavoro-dello-scientista-dei-dati-leprevisioni/>>



Massimo Tamiatti

I have always been involved in the study and dissemination of topics related to new technologies, new jobs, and new skills. I do this in a completely new and different way compared to traditional research approaches because, in addition to consulting traditional sources, I consult sources from major strategic consulting firms, which do not have a static but dynamic vision of reality, as they focus on the future. I systematically use the web to search for the most up-to-date content, perhaps through webinars with privileged witnesses on the subject of the investigation or through some important influencer specialized in the topic being addressed. Obviously, today, I use Generative Artificial Intelligence tools. I do not neglect the present; I try to interpret the signals that we have before our eyes, which we are often used to looking at and not seeing, and which suggest what could happen. I do not neglect the past, that is, history, but in reality, many disciplines are mixed: economics, statistics, sociology, psychology, anthropology, and above all, I use imagination anchored to the facts that happen every day in the world. I consider numbers very important, but they are not everything; at the center, there must always be a particular attention to the privileged witnesses of the things that happen and their personal stories. Frontiere does not have the gift of foresight, but our team can make hypotheses, and among these hypotheses, I have realized that it is very likely to encounter the future. Spreading the culture of anticipation helps people understand what is happening in the world, especially in the world of work, helps them prepare, elaborating strategies, and organizing themselves to make very important conscious and correct choices for the effect they will have on their lives.

Institutional Profile

I was Chief Research Officer of the Agenzia Piemonte Lavoro, where I coordinated groups of researchers in the field of labor and vocational training policies until June 30, 2023. Since July 1, I have become part of the staff of the "Laboratorio Riccardo Revelli" as an honorary member. I am an expert in public employment services. Currently, I am a member of the group that coordinated the organization of the "Settimana del lavoro 2024" and related events on behalf of the Istituto per la memoria e cultura del lavoro, dell'impresa e dei diritti sociali (ISMEL), of which I am a member of the scientific committee. I am a member of the scientific committee of Polis policy, Accademia di alta formazione. I am the founder and director of the online scientific dissemination site <http://www.frontieretecnologialavoro.com> since 2022 and a member of the Associazione dei Futuristi italiani di Roberto Poli. I am an expert in the use of forecasting techniques applied to local labor markets, in collaboration with the start-up Skopia, specialized in future studies and anticipation, aimed at helping organizations develop anticipatory strategies to support complex decisions in areas of strong uncertainty. I have a degree in letters and philosophy, with a thesis in contemporary history and one in political science, both obtained from the University of Turin, and I am the author of numerous publications related to the labor market, and in particular to the new scenarios of the labor market, new professions, and new skills.

The following collaborated with me on this project: Gabriele Lamberti, Benjamin Dafku, Ivan and Federico Giacobino.