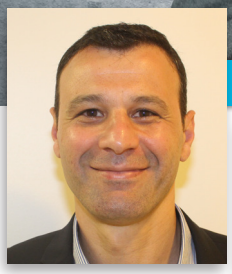
KEYRUS
insight into value

DATA SCIENTIST, AN ELITE PROFESSION IN THE PROCESS OF BEING DEMOCRATIZED

EXPERT OPINION



Bouzid Ait Amir | Manager in the Analytics Center – **Keyrus**

& **Nicolas Marivin** | Operations Manager in the Big Data & Analytics Dept – **Keyrus**

For 5 years now, the Data Scientist has been riding high at the top of the list of those employee profiles that are the most sought-after and, due to their scarcity, the most difficult to recruit. This situation could change in the short term: the arrival of enabling tools is going to speed up a sort of democratization of Data Science and force those who are seen as the current elite of the data professions to branch out into new skill areas.

It is widely agreed that Data Science is at the crossroads of various disciplines: applied mathematics, statistics, Machine Learning, Computer Science, Business Intelligence, Data Visualization... Faced with this list, it comes as no surprise that the expression "Swiss Army knife" is so often used to describe the Data Scientist! It is all the more justified when you consider that, as well as these numerous skills, which are already broadly multidisciplinary, you can add in those of a sound knowledge of business issues, coupled with a talent for communication, which is essential for establishing dialogue with the enterprise's various business functions. Such is the profile of the "true" Data Scientist – the one for whom all the world's major enterprises are looking to carry out a mission involving very high strategic stakes: to detect, in data, regardless of its nature, new levers for creating value for the enterprise.

IMPOSSIBLE TO FIND, OR IRREPLACEABLE?

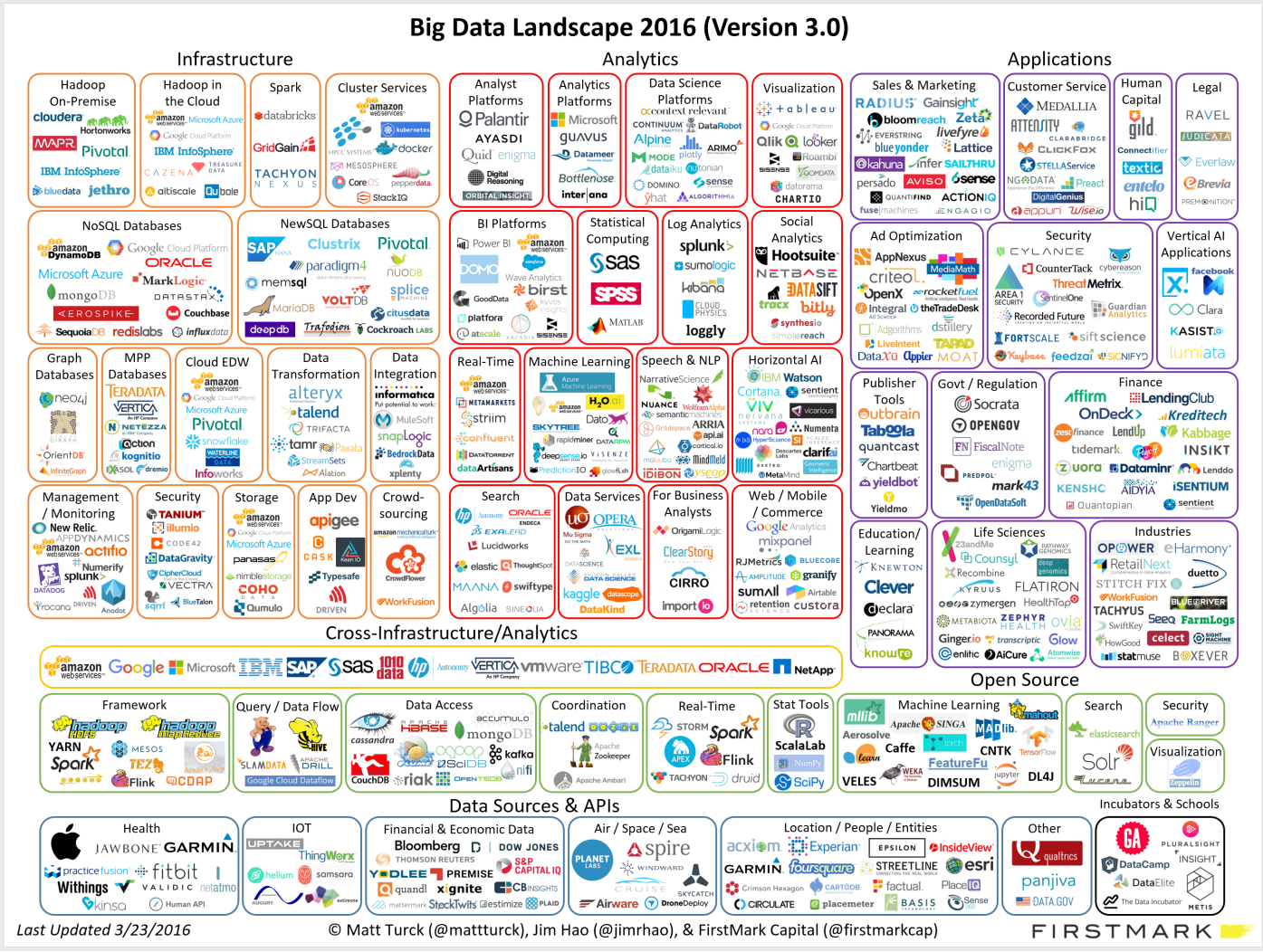
If we keep to this elitist definition of the Data Scientist, we have to accept the obvious: the population eligible to fill this function is very small and not enough to meet the demand. Not only are Data Scientists rare, but, on top of that, the best of them turn towards the GAFA[1] which, aside from offering them substantial amounts of money to take a job with them, present them with often particularly attractive career development prospects.

To understand this scarcity, it is interesting to consider what, according to many articles that have appeared in recent years, distinguishes Data Scientists from other data professionals. It is said that Data Scientists are more "relevant" than Data Analysts because they go more deeply into the analysis of data and apply more sophisticated methods for the purpose not of solving a problem, but of discovering new lines of thought. They also set themselves apart from pure statisticians, if this quotation, which has appeared all over the Web, is anything to go by: "Data Scientist - Person who is better at statistics than any software engineer and better at software engineering than any statistician"[2] . Specialists in Data Mining, a discipline often considered to be the first step of Data Science, are, in turn, less well up on computer science than Data Scientists. Finally, whilst mathematicians have a better theoretical knowledge of models, they often lack that data culture that characterizes Data Scientists.

AN ELITE THAT IS PURELY A PRODUCT OF CURRENT ECONOMIC CIRCUMSTANCES?

The profession of Data Scientist was born out of a dual "deluge": on the one hand, the deluge of data, frequently described by its three "Vs" for volume, velocity, and variety; on the other hand, the deluge of IT solutions, which has been very well demonstrated by the development of the "Big Data Landscape" between 2014 and 2016. In this rapidly changing landscape, Data Scientists are heaven-sent for the Big Data era: knowing the tools and mastering the methods as they do, they, more than any others, are capable of "making data talk" –

[1] GAFA is the acronym made up of the best known giants of the web (Google, Apple, Facebook, Amazon)
[2] It was seen for the first time in May 2012, in a tweet by the Data Scientist Josh Will. With no shortage of petty squabbling, one also comes across this more grating definition: "Data Scientist – Statistician living in the Silicon Valley"...



whether it is to find within them unexpected avenues for growth, or to uncover value-destroying phenomena.

With regard to the deluge of data, the volume is not a real problem: solutions, both for storage and analytical processing, manage this aspect fully. The velocity, and above all the variety, remain, however, real issues. Indeed, the creation of value comes from the combined analysis of three categories of data: internal data, de-siloed and structured in SQL Data Warehouses; semi-structured data such as Web, machine or XML logs; and the so-called unstructured textual, image and video etc. data. The skill of the Data Scientist lies in reconciling all this information by using the appropriate tool or tools, regardless of the language or technological environment. The Data Scientist makes insights emerge from the raw data by creating or choosing the most relevant features[3] and applying the right Machine Learning method or, increasingly, blends of models based on a sublayer made up of a large number of pre-existing models.

As things stand at present, it can be said that the demand for Data Scientists of a high level is directly linked to the complexity and variety of data sources. It is justified, above all, by the need to master the tools and programming languages specific to the Big Data ecosystem. However, due to the very changes in this ecosystem in the short term, these technical skills will be less and less essential.

RAPID EMERGENCE OF NEW ENABLING TOOLS

A few years ago, the processing of semi- and unstructured data required real expertise in Data Management and development. Software for Text Mining and, more broadly, automatic language processing were neither very relevant, nor sufficiently effective. Today, it is possible to find solutions on the market, offered under licence or especially as Open Source, that make it possible to process these unstructured data in a much more automated way. While you definitely still need to have numerous analytical skills, a very large part of the pre-processing is already integrated and very easy to use. For semi-structured data, the changes are even more radical. You can now structure this type of data by calling up just one line of code, using Python, for example. You

[3] *First level of transformation applied to raw data to structure them and/or aggregate them to provide data processing that is more summarized and can be more easily exploited.*

can even find on the market solutions that allow you to do Data Blending[4], regardless of the data's format (structured/unstructured) or the storage format (SQL and NoSQL database).

Everything points to the management of data variety ceasing very soon to be a problem. The need to master a multitude of programming languages to address the diversity of business problems will also diminish. Not so long ago, you really had to be an IT specialist to write code using Hadoop or Spark. Today, Hadoop is mainly used as a data warehousing solution. Spark, on the other hand, is making its mark by offering a data management structure and functionalities that are becoming increasingly simple to implement with each new version. Whilst there is not yet in existence an analytical solution making it possible to perform any type of processing, within an environment that is, or is not, distributed, and in a programming language that is stable and mastered by the greatest number of people, the trend towards this is very real.

RAPID EMERGENCE OF NEW ENABLING TOOLS

Data Science is in the process of equipping itself with tools that will make it more widely accessible. We have every cause to be happy about this for enterprises which, due to a lack of means and skills, were stuck on the sidelines of the Data Economy. This democratization through tools enlarges the pool of human resources from which those enterprises will be able to draw and enables them to empower employees with profiles which, while related, are less technical and multi-skilled than that of the Data Scientist as still currently defined. This will notably be the case for the Data Miner, who will be able to rely on Machine Learning models that are already integrated and configured, and on classic application solutions with interfaces that can be used by non-experts and a programming language accessible to all.

But what becomes of Data Scientists in this scenario? The integration of part of their technical skills into tools gives them the opportunity to develop two essential aspects of their activity: the ability to communicate about the results obtained, which involves in particular mastering Data Visualization tools, and working more closely with business functions, which is essential to give practical effect to the contributions of Data Science and embed the data culture at the very heart of the activity. For those who have a real appetite for these two career directions, there is no reason why the profession of Data Scientist should not continue to be "the sexiest job of the 21st century"[5] for a few years yet.

About the authors

Nicolas Marivin
Operations Manager in the Big Data & Analytics department, Nicolas Marivin possesses more than 15 years' experience in the field of data valorization. In addition to his activities managing Business Analytics, Data Science, and Big Data offerings, he is involved in information system innovation and modernization assignments.

Bouzid Ait Amir
A graduate in Econometrics from the Toulouse School of Economics and in Computer Science from Télécom Bretagne, Bouzid AÏT AMIR has 15 years' worth of expertise in the Data Science field, which has enabled him to assist France's largest enterprises with their transformation and the implementation of innovative analytical solutions.

[4] Concept created by a software company that reflects the fact to reconcile data types and formats in agile mode.
[5] Reference to the famous article by Thomas H. Davenport and D.J. Patil, "Data Scientist: The Sexiest Job of the 21st Century", Harvard Business Review, October 2012.

A propos de Keyrus

Keyrus, creator of value in the era of Data and Digital
An international player in consulting and technologies and a specialist in Data and Digital, Keyrus is dedicated to helping enterprises take advantage of the Data and Digital paradigm to enhance their performance, facilitate and accelerate their transformation, and generate new drivers of growth, competitiveness, and sustainability.
Placing innovation at the heart of its strategy, Keyrus is developing a value proposition that is unique in the market and centred around an innovative offering founded upon a combination of three major and convergent areas of expertise:

- **Data Intelligence**
Data Science - Big Data Analytics – Business Intelligence – EIM – CPM/EPM

- **Digital Experience**
Innovation & Digital Strategy – Digital Marketing & CRM – Digital Commerce – Digital Performance – User Experience

- **Management & Transformation Consulting**
Strategy & Innovation – Digital Transformation – Performance Management – Project Support

Present in 15 countries on 4 continents, the **Keyrus** Group has 2500 employees.
Keyrus is quoted in compartment C of the Eurolist of Euronext Paris (Compartment C/Small caps – ISIN Code: FR0004029411 – Reuters : KEYR.PA – Bloomberg : KEY : FP)

Further information at : www.keyrus.com

KEYRUS Marketing and Communication Department
155 rue Anatole France, 92593 Levallois-Perret Cedex – Tel. : +33 (0)1 41 34 10 00 - marketing@keyrus.com