Workshop on HRMT – Modernizing statistics: how to get there? – Geneva, 15-17 October 2014

A new job for statisticians: the data scientist. Which skills, how to build them

(Antonio Ottaiano, ISTAT - Italian National Institute of Statistics)

1. Where data science begins

In the last few years, there has been an explosion in the amount of data available. People spend more and more time online, leaving behind tracks of the data they use. The web is full of *data-driven apps*: e-commerce applications, for instance; any web front end has a database behind it, collecting data from users. And we leave a data trail behind us whenever we surf the web, chat with our friends on Facebook, or buy something in a shop.

Mobile applications leave an even richer data exhaust: many of them are geolocated, and make available a great deal of data, all of which can be mined. Point-of-sale devices and frequent-shopper's cards make it possible to capture data also from retail transactions, not just from the online ones.

Data expands also because there is a lot of space to fill. The more storage is available, the more data you will find to put into it.

However, merely using data isn't really what we mean by "data science". As Mike Loukides explains in his *What is 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".

Significant worldwide companies (Google, Facebook, LinkedIn, Amazon) had their gains by using data creatively, turning it into something of value. By implementing its PageRank algorithm, for example, Google was among the first to use data outside of the page itself, in particular, the number of links pointing to a page. Tracking links made Google searches much more useful, and PageRank has been a key factor to the company's success. The PYMK (People You May Know) algorithm – that LinkedIn and Facebook use to suggest, starting from patterns of friendship relationships, people users should know - is another example of a creative and profitable use of data collected from users. Services like I Tunes analyses music users listen to and look for, and suggest them lists of songs they may want to buy.

All these are examples of "data products" that share the same feature: they are based on the fact that data collected from users (search terms, contacts, music) provides companies with added value for their business. This is where data science begins.

2. Using data effectively. The data scientist: role, activities, skills

Now, in such a huge sea of data, the question companies have to face is how to use data (not just their own data, but all the data available) effectively. In fact, how can they make that data useful?

Using data effectively requires something different from traditional statistics. What differentiates data science from statistics is that data science is characterized by what Loukides calls a "holistic approach". Such an approach involves gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.

This is a job for data scientists.

The title of “data scientist” was coined in 2008 by D.J. Patil, and Jeff Hammerbacher, who then led their teams of experts of data analysis at LinkedIn and Facebook: according to them, it was the title that seemed to best express people who used both data and science “to create something new”.

A data scientist can be seen as an evolution from the business or data analyst role.

Whereas a traditional data analyst looks at data from a single source, a data scientist will explore and examine data from multiple disparate sources, looking at it from many angles. The goal is different: discover previously hidden insight, to pick the right problems that have the most value to the organization.

Data scientists “make discoveries while swimming in data”, as Patil says. While doing that, they give structure to large quantities of formless data and make analysis possible. As they make discoveries, they communicate what they’ve learned and suggest its implications for new business directions.

The data scientist role has been described as “part analyst, part artist.” Anjul Bhambhri, vice president of big data products at IBM, says that “a data scientist is somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who really wants to learn and bring change to an organization."

In fact, often they are creative in displaying information visually and making the patterns they find clear.

Visualization is frequently the first step in analysis, and takes the data scientist through each step of his/her work. To make clear what the numbers mean, to explain which stories they are really telling, data scientists have to adopt a visualization approach and use visualization techniques and tools.

Summing up what we have said so far, we can now list a first set of activities a data scientist is required to carry out:

• looking for rich data sources

• working with large volumes of data

• cleaning the data

• crossing multiple datasets

• analyzing connections among data

• visualizing the data analysis.

What kind of person does all this? What abilities make a data scientist successful? What skills does he/she need?

Patil makes a list of what makes a good data scientist:

• *Technical expertise* in some scientific discipline.

• *Curiosity*, that’s the willing to go beyond the way data looks like, and find out what is hidden behind it.

• *Storytelling*: the ability to make a story based on the data available and to communicate it effectively.

• *Cleverness*: that is the ability to look at a problem in different ways and from many perspectives.

We are talking about interdisciplinary profiles, able to tackle issues from different points of view and to look at the problem they have to face as a whole, as part of a bigger one. Profiles that are required to be statisticians, mathematicians, programmers, even “artists”.

Are such profiles available on the job market? Who shall we look for?

The traditional backgrounds are not helpful anymore: a data management expert is not necessary so able to analyze data as he is in organizing data; vice versa, a quantitative analyst can be valuable in analyzing data, but not so able in managing and shaping a mass of unstructured data into a form that can be analysed.

So, if to find such competence profiles available on the job market is not so easy, could it be a better solution for an organization to create and grow, rather than hire, its own data scientists? Not necessarily a profile with all a data scientist's skills required, but an integrated team in which such skills are distributed among the members.

Building a group of data scientists was the road followed by D.J. Patil at Linkedin. That team - the team which turned out in developing products like PYMK - wasn't made only by statisticians, mathematicians and other "data people." It was a fully integrated group that included people working in design, web development, engineering, product marketing, all able in working with data. The aim was not to re-produce those silos that traditionally separated data people from engineering, from design, from marketing, and to make the data scientists' group a full product team responsible for designing, implementing, and maintaining data products.

At Linkedin, the strategy of building a cross-disciplinary group of experts of data science resulted successful.

But how can it work for a NSI? Can a NSI build integrated team of data scientists' competences also through training activities?

3. Working on data scientist's skills development

In Italy a reflection on data scientist's skills and training intervention started within the framework of the Italian Digital Agenda. Moving from the statement of the Scheveningen Memorandum on "Big Data and Official Statistics", adopted by the ESSC on 27 September 2013 ("Recognise that developing the necessary capabilities and skills to effectively explore Big Data is essential for their integration into the European Statistical System. This requires systematic efforts like appropriate training courses and establishing dedicated communities including academics for sharing experiences and best practice"), Istat gave its contribution both to the definition of the skill profile on data scientist and to the design of an introductory course on data science.

The skill profile (in Annex 1 an English translation from the Italian official document) was defined within the framework of the *Web professional profiles* list published by IWA (International Webmasters Association) Italy according to the CEN guidelines in the field of Generation 3 (G3) European ICT Profiles. Based on that, the data scientist profile was described in terms of mission, tasks, E-Competence Framework skills. Such a framework was the point from which the reflection at Istat on how to create and develop data scientist's skill started.

A project team was established, with the aim of integrating competences and experts coming from different areas: statistics, IT, organization, training.

The team started a debate about the target of the training activity: a choice was to be made between either to address both to statisticians and IT experts or to have separate training modules for different professional profiles. At the end, it was decided to design a training activity made of different steps: a first step being a seminar addressed to a wide audience, at an introductory level, with the aim of verifying within Istat the interest towards such topics and collect needs of deepening the subject, to be met at a later stage; in this second step, addressed to both statisticians and IT experts, traditional lectures will be integrated by "laboratory" activities, with the aim of promoting interaction among the different competences involved.

As for the contents, some macro-areas were identified: statistical methods, IT techniques and visualizations tools. A set of topics for each of them was defined as well, to be dealt with at a different level of detail according to the target: from the fundamental of database design and management to data mining; from big data platforms and applications (Hadoop, MapReduce, Cassandra, Hive) to programming languages (Python, Perl, Php), to data modeling and machine learning.

At the moment the design phase is still ongoing , so such contents haven't been organized yet in a training programme. The intention is to complete the step 1 (a 4 hours seminar) by the end of this year, and to deliver the second part of the programme (lectures and laboratoryfor approximately 80 – 100 hours) in the first quarter of 2015.

Conclusion

The recent explosion of digital data made available a large amount of data to be explored, analyzed, connected in order to build knowledge and create value. To take advantage of that, organisations more and more need "data scientists", people able not only to simply collect and report on data, but also able to look at those data from many perspectives, determine what they mean and suggest how to apply them in the business strategy.

Organisations can either hire such profiles or grow their own data scientists, investing in building and developing integrated team of data scientists' competences. It's an investment worth making, since – as Hal Varian, chief economist at Google, says – "the ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades".

References

1. M. Loukides, *What is Data Science*, O'Reilly Media, Inc.
2. D.J. Patil, *Building Data Science Teams*, O'Reilly Media, Inc.
3. T. H. Davenport, D.J. Patil, *Data Scientist: The Sexiest Job of the 21st Century*, Harvard Business Review

SkillProfiles.eu

PROFILE SHEET WSP-G3-024 "DATA SCIENTIST"

G3 Web Skills Profiles – version 2.0
Generation 3 European ICT Professional Profiles

Appendix to the official specification of June 30, 2014

Current version:	http://www.skillprofiles.eu/stable/g3/v2/profiles/WSP-G3-024.pdf
Previous version:	http://www.skillprofiles.eu/stable/g3/v1/profiles/WSP-G3-024.pdf
Last version:	http://www.skillprofiles.eu/stable/g3/profiles/WSP-G3-024.pdf
Editor:	▪ Pasquale Popolizio (Coordinatore Gruppo IWA Italy - Web Skills Profiles) ▪ Roberto Scano (Presidente IWA Italy) ▪ Alessandra de Seneen ▪ Concetta Ferruzzi (ISTAT) ▪ Silvia Losco (ISTAT) ▪ Antonio Ottaiano (ISTAT) ▪ Paolo Podda ▪ Emanuele Rizzardi ▪ Monica Scannapieco (ISTAT) ▪ Daniele Sghedoni ▪ Antonino Virgillito (ISTAT) ▪ Walter Vannini

Copyright

Table of contents

Profilo WSP-G3-024. Data Scientist

This section is normative.

The profile sheet, listed below and described in appendix B, is an integral part of the document, "G3 Web Skills Profiles - version 2.0 - Generation 3 European ICT Professional Profiles", official specification of 30 June 2014" [WSPG3-03].

Profile WSP-G3-024	**Data Scientist**
Summary definition	Professional profile encharged of activities related to collection, analysis, processing, explanation, dissemination, visualization of quantitative data for analytic, predicitive, strategic purposes.
Assignment	Data Scientist identifies, collects, processes, analyses, explains data related to organization in order to extract information, also through development of predicitive models, aiming at building advanced knowledge systems. Thanks to a deep knowledge of organisation's mission/business, data scientist identifies and accesses to data sources supporting organizational processes; data scientist chooses the most effective methods and models to address organisation's strategic decisions, and to develop operational plans; data scientist, starting from the data collected, produces development plans; he/she presents such programmes in the most appropriate way in order to support management decision, paying strong attention to issues related to effective visualization of information.
Documentation produced	**Accountable** ● Identifying and acquiring data. ● Analysing data ● Presentation of analysis either in textual or in graphic form ● Reporting **Responsible** ● Market profiling report. **Contributor** ● Sales plans. ● Marketing plans.
Primary duties	● To integrate professional profiles related to data analysis. ● Elicitation and needs analysis. ● Designing and preparing effective data analysis ● Identify relevant data and their sources. ● Collecting data. ● Data quality process ● Analysing data ● Building quantitative and qualitative models ● Building predictive models

	● Explaining analysis and models ● Presenting effectively analysis and models. ● Addresssing business needs ● co-operating with IT in order to define data collecting and management activities ● Co-operate with controller in order to develop analysis and reports supporting decisional processes
Assigned e-CF skills	● **A.6 Applications design:** Level e-3 ● **A.7 Monitoring IT trends**: Level e-4 ● **B.1 Applications development**: Level e-2 ● **B.3 Testing**: Level e-3 ● **B.5 Producing documentation**: Level e-3 ● **C.1 Assistance to users**: Level e-3 ● **C.3 Service delivery**: Level e-3 ● **C.4 Problem management**: Levels e-3, e-4
Abilities, knowledge	**Technical** ● Statistics. ● Analysis ● Quantitative and qualitative methods ● Clustering techniques (i.e. K-Mean, Fuzzy K-Mean). ● Multidimensional data modeling. ● Data visualisation (QlikView, Tableau, TIBCO Spotfire) ● Structured and unstructured data management ● Data quality management. ● Algorithms ● Inference ● Textual analysis ● Models and methods for decision making **Information Technology** ● SQL query language. ● ETL tools (Extract, Transform, Load). ● OLAP systems ● Statistical analysis systems (i.e.: R, SAS, SPSS). ● Scripting languages, like bash, PHP, PERL, Python. ● Data management platforms. **For development** ● PMML (Predictive Model Markup Language). ● Big Data platforms and applications (i.e.. Hadoop, MapReduce, Splunk, Cassandra). ● Machine Learning platforms (i.e.: apache Mahout). ● Business analytics.
Area of application of the KPI	● Number of projects appointed and completed.

Qualifications and certifications *(this section is for information purposes)*	● Masters/Training courses
Personal aptitudes *(this section is for information purposes)*	**Interpersonal and Organisational** ● Communication ● Leadership ● Teamworking and team management ● Creativity ● Flexibility ● Problem solving ● Value creation ● Business sense **Linguistic** ● Good knowledge of the national language or the language used by the working group - minimum level: B1 QCER. ● Good knowledge of the English language - minimum level: B2 QCER
Relationships and reporting lines *(this section is for information purposes)*	**Interacts with:** ● IT manager ● Master Data Manager ● Business lines managers ● Controller and business data analyst ● Top management (CIO, CFO, CEO, COO, …) ● Web Project Manager ● Digital Strategic Planner ● Knowledge Manager **Reports to:** ● Responsible for strategies/controller in big companies ● Responsabile for business/controller in medium size companiers ● Managers of SMEs

Appendices

Appendix A. Glossary

Informational

For the purposes of information and not required for compliance.

Note: The content required for compliance is referred to as "normative".

Normative

Required for obtaining compliance.

Note: Content listed as "informational" or "non-normative" is never necessary for compliance.

Appendix B. Profile Sheet Structure

The Web skills profiles are identified by an unambiguous code and are structured in reference to paragraph 4.2 of CWA CEN reference document, “European ICT Professional Profiles” [CWA-01], updated with respect to the “European e-Competence Framework 3.0” [CWA-02].

- Profile Title. Name - including the identification code - of the Web skill profile according to the unambiguous international catalogue from the IWA/HWG.
- Summary definition. Lists the primary purpose of the profile. The purpose is to give all stakeholders and users a brief, concise description of the specified Web skill profile, written in a form understandable by ICT professionals, managers, and Human Resources staff.
- Assignment. Describes the basic assignment of the profile. The purpose is to specify the working role defined in the Web Skill Profile.
- Documentation produced. Describes the documents produced by the job description as manager (guarantee), representative (support), and employee (contribution).
- Primary duties. Provides a list of typical tasks carried out by the profile. A task is an action undertaken to achieve a result in a broadly defined context and contributing to the definition of the profile.
- Assigned e-CF skills. Provides a list of the skills necessary (taken from the e-CF references) to carry out the assignment. A skill is the outcome of the previous definition of the Profile and helps to differentiate profiles.
- Abilities, knowledge. A list of abilities and knowledge necessary for the definition of the profile, subdivided into technical, IT, and improving abilities (strengthening the profile).
- Area of application of the KPI. Based on KPI (Key Performance Indicators), the area of application of the KPI is a more generic indicator, consistent with the grade level of the overall profile. It applies for adding depth to the assignment.
- Qualifications and certifications. These are the recommended, but not essential, qualifications and certifications for carrying out the activities in the profile. However, these

qualifications and certifications may be used for developing knowledge of specific skills within the profile.

- Personal aptitudes. A list of aptitudes supporting the abilities and knowledge, subdivided into interpersonal/organisational and linguistic. This section reports references to the QCER [CE-01], which promotes the understanding of specific language certifications, purely for informational purposes.

- Relationships and reporting lines. A list of Web skills profiles and not with whom the profile discusses (relationships) or reports (reporting lines). This section is for informational purposes.

Appendix C. References

[CC-01]	**Creative Commons** *Attribution – Share in the same way International 4.0 (CC BY-SA 4.0)* http://creativecommons.org/licenses/by-sa/4.0/deed
[CE-01]	**Council of Europe** *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* (gennaio 2002) http://www.coe.int/t/dg4/linguistic/CADRE_EN.asp
[CWA-01]	**CEN (European Committee for Standardization)** *CWA 16458:2012 European ICT Professional Profiles updated by e -CF version 3.0 competences* (marzo 2014) http://relaunch.ecompetences.eu/wp-content/uploads/2013/12/EU_ICT_Professional_Profiles_CWA_updated_by_e_CF_3.0.pdf
[CWA-02]	**CEN (European Committee for Standardization)** *CWA 16234:2014 Part 1. European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors* (marzo 2014) http://ecompetences.eu/wp-content/uploads/2014/02/European-e-Competence-Framework-3.0_CEN_CWA_16234-1_2014.pdf
[WSPG3-01]	**IWA (International Webmasters Association)** *G3 Web Skills Profiles - version 2.0 Generation 3 European ICT Professional Profiles Offcial specification of 30 June 2014* http://www.skillprofiles.eu/stable/g3/2013-06-30.pdf