NTTS Conference 2015

Bruxelles, 10 – 12 March 2015

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

Antonio Ottaiano, *Italian National Institute of Statistics*

Istat

Explosion of data available

- millions of people on the web
  - data driven apps
  - mobile applications
  - data trail

What do we 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" (Mike Loukides, What is Data Science?)

Where Data science begins

- PYMK - LinkedIn
- Page Rank - Google

How to use data effectively?

- Holistic approach
  - data gathering
  - data processing
  - turn data into a story
  - presenting

This is a job for *Data scientists*!

- people who use both data and science *"to create something new"* (D.J. Patil, J. Hammerbacher, 2008)

- explore data from multiple sources
- pick the right problem
- "make discoveries while swimming in data"

Activities

- 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 makes a good data scientist

- Technical expertise
- Curiosity
- Storytelling
- Cleverness

An interdisciplinary profile

- Are such profiles available?
- Who shall we look for?

Let's grow our own data scientists – definition of a skill profile

1) COMPETENCE: Analytical/ Computing Skills

S/he must be accomplished in analytical methods, and have an appreciation and understanding of information presented in mathematical terms; have the ability to extract the key messages or underlying trends present within data; and be able to present statistical results and concepts, both orally and in writing, in a confident and professional manner.

2) COMPETENCE: Delivering Quality and Ethical Analysis

A Data Scientist must maintain the highest levels of integrity when carrying out his/her work. S/he will continuously check, cross-check and quality assure work to ensure accuracy and to produce technically accurate figures and work of a high standard. An effective Data Scientist will anticipate problems and test data to ensure it makes sense.

3) COMPETENCE: Process Management Skills & Creative problem solving

An effective Data Scientist will be structured and methodical in his/her approach to work. S/he will have strong project management skills, with the ability to prioritise and make decisions in a timely manner. S/he takes responsibility for his/her own work and can come up with solutions to the problems encountered.

4) COMPETENCE: Teamwork

A Data Scientist must work collaboratively with others, building networks and communicating clearly with them. When required, s/he can convince others to supply information and will provide advice and assistance when the need arises. An effective Data Scientist will support his/her colleagues and will take the views of others on board.

5) COMPETENCE: Communication of Information

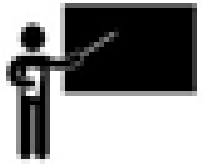
A Data Scientist must produce objective information in a fluent manner and following a logical structure. S/he will provide data in a manner that is appropriate for the target audience and can project a credible and confident image to internal and external parties. S/he takes steps to ensure that people do not misinterpret data and outputs, anticipating where this may occur.

6) COMPETENCE: Developing Analytical Expertise

A Data Scientist must develop expertise in his/her area of specialism but also be open to learning new skills to perform in a new environment. S/he will be flexible in his/her work and willing to work outside of his/her comfort zone. An effective Data Scientist will take on board feedback from managers and colleagues.

Let’s grow our own data scientists – design of training activities

- Methodologies:
  - seminar
  - lectures
  - laboratory



- Contents:
  - Data analysis methods and tools
  - Big Data management methods and tools
  - Using Big Data for statistical purposes