

Working as a Data Scientist with real-world clinical data

Berber Snoeijer, PHARMO Institute NV, Utrecht, The Netherlands

ABSTRACT

The profession of Data Scientist requires many different skills. Depending on the field of action more or less focus is needed on these different skills. However, the visualisation, creativity, high programming skills and hacker mindset are more than necessary to be able to find solutions for the day to day caretaker in Dutch Health care. This presentation will focus on the general skills associated with the function of data scientist and how they are used by the data scientists in our organisation.

INTRODUCTION

Working as a Data Scientist is a very interesting and challenging job where different skills are combined. Combining technical skills with communicative skills and combining insight in data structures with insight in customer needs are often not the things that are combined within a single professional. Therefore, if one is equally skilled in both then he/she could become a data scientist. However, there are many aspects in the job one has to like and has to get experienced in. Below are some of these aspects mentioned.

WORKING AS A DATA SCIENTIST

DATA INTEGRATION AND PRESENTATION

At the PHARMO Institute different data sources are available like a GP database, a laboratory assessments dataset and a large claims dataset comprising more than 1 fifth of all claims in The Netherlands.

The data of these different source have to be combined for research and reporting. This can be done by probabilistic coupling method or by an exact coupling method. The probabilistic coupling method has been used for more than 15 years now and is based on variables like gender, birth date and GP. The exact method requires more extensive privacy measures and better registration and is now being implemented. Combined different data sources is more difficult than it seems in first instance. For example, it is possible that a patient is registered more than once, data is entered at the wrong place or patient characteristics of two different patients are very similar. All these real life errors and options have to be taken care of.

As a result there will be a big relational database that can be used for clinical research or mirror reports to the caretakers. The presentation of data is important to make it available for others. Data structures and relations may be very complex. The data scientist has an overview of these structures and relations and can present the data as such that it becomes clear to others as well.

COLLABORATION

A data scientist is communicative both in speech and in writing. He/she must be able to make his ideas clear to others within the company. On the one hand the management who must approve the ideas of the data scientist and on the other hand the programmers and technicians who will take over new procedures and ideas of the data scientist.

HIGH-LEVEL PROGRAMMING

High-level programming is needed to get the data that is needed in a form that is usable. Different programming techniques like using macro's and macro variables, Base SAS, SQL, arrays, retain, statistics, creating graphs may be necessary and have to be combined to get the results that you need.

STATISTICS

Statistics are necessary to get insight into the data. This may vary from simple descriptive statics (in SAS like PROC MEANS and PROC UNIVARIATE) to different modeling methods (like PROC MIXED and PROC LOGISTIC).

SCIENCE

A background in science is needed to understand the clinical data. Results have to be interpreted in the right way and the data scientist has to know about scientific basics to be able to talk to caretakers and other customers.

PhUSE 2015

VISUALIZATIONS

Visualizing data and results is an important aspect of the data scientists job. Clear univocal graphs have to be made that can be used to explain complex results. On the other hand more complex explorative graphs can be used to get more insight in the data and data structures.

HACKER MINDSET AND PROBLEM SOLVER

Working with real life clinical data means that input data can be wrong at every level and detail. This can be everything what you can expect and what you cannot expect. Therefore, a data scientist must be aware of where the data comes from and what happens with the data afterwards. All entry points and entries might go wrong. For example, a patient number is in most cases representing a real number. But it might also be a container for all information that is not patient specific. Therefore one must be aware for birth dates like 01-01-1900, 11-11-1911 and 09-09-1999.

CONCLUSION

A brief summary of the tasks involved when working as a data scientist are mentioned above. yet, it is not limited to this. As can be seen from the above, the work of a data scientist is challenging, varying and requires an adaptive mind. Therefore, the data scientist needs to have the right interests, skills and he/she should like the big variations in his/her work.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Berber Snoeijer
PHARMO Institute NV
Van Deventerlaan 30-40
3528 AE Utrecht
Berber.snoeijer@pharmo.nl
www.pharmo.nl

Brand and product names are trademarks of their respective companies.