Marianne Faro

DATA SCIENCE REQUIRES A LOT OF HARD AND UNREMITTING WORK

The CIO as data factory manager

Data science is hot. You only have to read the job offers and blogs. Every company deals with big data, is building large data lakes and is looking for exciting things to do with machine learning or deep learning. Every university delivers groups of data scientists that are expected to infiltrate the companies, wondering how they will make digital transformation possible with their algorithms.

Unfortunately, however, data science consists for more than 90% of hard and unremitting work. Not on algorithms and beautiful machine learning models, but on data. On methodological work and automation. On setting up a factory for your data processing.

Data-wrangling

The data you gather are rarely as beautiful as the Kaggle sets that are clean and drawn up with a clear definition of a problem. So the first steps in data science consist of scripting how to collect the data, to think of a way to clean them up, what to do with outliers and anomalies, how to handle missing values, and whether you have to anonymize data from the moment of ingestion onwards in the context of e.g. the GPDR. JADS (Jheronimus Academy of Data Science in Den Bosch) uses a recognisable term for this: 'data wrangling'. It is essential to take this step right from the start: think about your data pipeline.

Data storage

Parallel to this, you have to think about the job that was previously performed by database engineers or warehouse experts. Do you choose a relational database, NoSQL, or unstructured database to store your wrangled and analyzed data? What do you use to store and quickly process terabytes of images? Do you need to set up a more systematic system with multiple nodes or do you think about this later on? How do you protect the data from loss or misuse? Do you only have to store the results of your machine learning model or also the intermediate steps and inputs? Thinking about data storage is also part of your data pipeline and something that should reflect in your decision-making.

Analytics

Using the tidied and smartly saved dataset, you can then work with analytics to achieve your first insights. If you want to be smart, you need to know what exactly is happening, what data are significant, and what patterns you need to identify. There are plenty of data, but for now, only people can interpret it. For example, you need domain knowledge for correlations. Because what can be an 'aha' moment for you as a statistician because you find a strong correlation from the data, is a 'duh' truism moment for the domain expert: of course you will get wet when it rains.

The same applies to the algorithms and models applied by a data scientist. There are plenty of libraries and applicable models, but if you do not understand what they do with your data, which variables are important or not, when there is overfitting, you will not get very far. Then you have a cool model with a lousy outcome.

CIO as data factory manager

It is the task of the CIO or CDO to direct this. Direct this by using a good data pipeline. Direct this by understanding the data and domain knowledge. Direct this by asking many questions and contemplating. Data science is difficult and is certainly not a miracle cure. Seeing through the system requires years of deepening, and you must always continue to find the business case with common sense and domain knowledge.

Marianne Faro is the director of Itility.