Data Science - more than just Machine Learning: A summary of the Data Science Session at INFORMATIK 2020

Birgitta König-Ries[1], Klemens Böhm[2]

Abstract: In this short article, we briefly summarize the Data Science session at INFORMATIK 2020. With three invited talks, the session focused on data-science challenges beyond the development of new machine learning models.

Keywords: Data Science; Data Management

1 Introduction

Too often, data science is reduced to machine learning (ML) and the development of new ML models. While the three talks that we had invited were very heterogeneous with respect to background (one talk from academia, two from industry) and thematic focus, they all agreed on one point: While the development of new ML models is obviously an important part of data science, it is just one part out of several important ones. The talks highlighted these other aspects very well, as discussed below.

2 Invited Talks

Benno Stein from Bauhaus University in Weimar introduced the Internet Archive hosted at Bauhaus University and its partner universities in Halle, Leipzig and Paderborn as a representative sample. On the one hand, this copy of a large subset of all web data is the basis for numerous research activities by his group. They range from work on argumentation and celebrity profiling to vandalism detection. On the other hand, to successfully build and maintain a suitable copy, specific research efforts are needed. Examples include work on Webpage segmentation and on the quality analysis of web crawling.

In the second talk of the session, **Torsten Grabs** from Snowflake argued that today, machine learning models are successfully being developed, but that their use in production environments often fails. He quoted a customer stating that ” we are good at building models that we don’t use”. Two of the main hurdles towards moving from development

[1] Friedrich Schiller University Jena, Heinz Nixdorf Chair for Distributed Information Systems, Jena, Germany, birgitta.koenig-ries@uni-jena.de
[2] Karlsruhe Institute of Technology (KIT), IPD, Karlsruhe, Germany, klemens.boehm@kit.edu

doi:10.18420/inf2020_04

to production are the lack of integration between data and ML platforms and the large heterogeneity of existing ML platforms; each platform is best suited for a certain subset of use cases. Torsten Grabs encouraged research on environments that facilitate declarative integration of data and analysis, analogously to what SQL and the relational algebra did for business intelligence.

The third and final talk of the session focused on a popular modeling paradigm to represent knowledge, namely Knowledge Graphs. The talk was given by **Steffen Lamparter** from Corporate Technology, Siemens AG. He argued against the development of a unified, company-wide Knowledge Graph and in favor of the ad hoc and use-case specific development of small Knowledge Graphs. The talk showed examples of a wide variety of use cases of different complexity supported by these graphs that integrate structured data (e.g., from DBMS), semi-structured data (e.g., from documents) and human knowledge. Siemens aims for comprehensive tool support that enables engineers (i.e., domain experts) to build knowledge graphs without the help of data scientists.

3 Discussion and Conclusion

A recurrent theme in the discussions during and after the session was the strong need in both industry and academia for highly qualified data scientists with a strong understanding of the (mathematical) foundations of machine learning. This understanding is a prerequisite for solving the challenges identified in the session.