Questions for the Workshop

Andreas Rauber, Carlo Zwölf, Dimitris Koureas, Koenraad de Smedt, Peter Wittenburg

The focus of the online workshop on 14.4. should be on the development of data-intensive science and not so much on "scholarly communication" although they are often closely related. Current discussions and documents are often about the needs to invest more in training and education, to improve visibility, correct citation and acknowledgements, and legal aspects. We collected a few questions as examples for the kind of issues which we would like to see being addressed. We see three major areas of future science:

- smart **knowledge extraction** from large and complex data sets – typically to be carried out by a higher degree of automation and the field of AI methods,
- smart methods of **knowledge representation and combination** – yet underdeveloped but crucial in times where many labs globally participate in extracting knowledge and where the results will be heterogeneous,
- the **infrastructure** needed to carry out the work efficiently and reliably.

In our preparation discussions, the following questions were raised which could be points of discussion:

1. How will data science develop in the coming decades anticipating the huge amounts of data being created and processed, and their inherent complexity given the many smart devices that are deployed everywhere?

2. In such a scenario data quality and trust will play an important role. Which mechanisms will help to establish the required trust.

3. Will data science give impulses to cross-disciplinary/cross-silo work? If so, which mechanisms would be needed to carry out such data science efficiently?

4. Science and reproducibility are twins until now. Will this remain and if so, how can reproducibility be ensured, and which infrastructural mechanisms are required?

5. What kind of common infrastructure and services are required to maintain a leading role for European data scientists? Which kinds of infrastructural support will be required to not load the data scientists which could better be done by data managers and stewards? What are the necessary timelines?

6. How will labs look like in about 10 years from now? What will be automated and what not? Which external services will be needed?

7. Are there already flagship projects in data science and/or data business that can indicate directions of developments? How will they evolve?

Some answers

Jean Claude Burgelman

These are all very relevant questions. I don't have another one to add, but i would sharpen some, by introducing the importance of machines led data science. The data volume will be sobbing, that only machines can make sure we make sense out of it. Data science will in my view in the first place, machine led science and that will imply a lot for where and how these data are gathered, processes etc …

Arturo H. Ariño

From the point of view of biodiversity data science (BDS), the proposed general questions are relevant to the field and should be readily addressed.

Two crucial points I think are of particular relevance is how to deal with biased, opportunistic data availability (which is the general case for BDS), as opposed to uniformly-sampled data, in order to make inferences. This may not be an additional question but just an aspect of Q2. Another aspect that might be covered as part of Q2, IMHO, is the issue of fitness-for-purpose, which is often confused with (but is actually just related to) Data Quality.

Q4 is particularly well crafted (not that the others aren't, but this one is even sharper), as it addresses a core issue within Data Science made up of opportunistic data which, by definition, are fluid and often irreproducible when not coming from experiments but from opportunistic collection--it's just the methods that could ensure reliability. I believe this is particularly relevant for BDS.

Finally, one issue I don't see mentioned in the questions (other than implicitly) is that of sensitive data, also related to privacy and ownership in some areas, which I believe will become increasingly relevant. Perhaps it could be referred to in Q6, although it is an issue by itself. OTOH, building a large list of questions is best avoided so perhaps this issue could be embedded somehow within the questions, rather than raising a new one.

Matthias Scheffler

He will indeed address the 3 dimensions mentioned above (knowledge extraction, knowledge representation and combination, infrastructure) based on their cutting-edge data science about materials. When addressing these dimensions, he will reflect on some of the questions which we raised.