Data Science 1 (2017) 1–5
DOI 10.3233/DS-170013
IOS Press

Data Science – Methods, infrastructure, and applications

Michel Dumontier [a] and Tobias Kuhn [b]

[a] *Maastricht University, Maastricht, The Netherlands*
E-mail: michel.dumontier@maastrichtuniversity.nl; ORCID: https://orcid.org/0000-0003-4727-9435
[b] *Department of Computer Science, Vrije Universiteit Amsterdam, The Netherlands*
E-mail: t.kuhn@vu.nl; ORCID: https://orcid.org/0000-0002-1267-0234

1. About Data Science

Science has always been about data. Observational data points have served as the evidence that has allowed us assess, accept, and discard scientific theories. But in the last decades, scientific data has grown dramatically in both size and importance. Data Science is therefore not a new science discipline, but rather a new pair of glasses – a new paradigm – to look at problems and questions in the existing disciplines with the new possibilities of data analytics in mind. It also stands for the development that data, when properly linked, transcend disciplines and can enable new sorts of interdisciplinary research fields and even breed entirely new areas. The focus on data also immediately highlights other important and urgent issues in contemporary science, namely the reproducibility of results, the responsible treatment of potentially sensitive data, the transparency and openness of scientific data and processes, the attribution and recognition of data gathering and curation efforts, and the now widely accepted requirement that scientific data should be Findable, Accessible, Interoperable, and Reusable (FAIR). With this journal, called *Data Science*, we intend to give this type of research the focus and attention we think it deserves.

2. The journal

Data Science is an interdisciplinary journal that covers aspects around scientific data over the whole range from data creation, mining, discovery, curation, modeling, processing, and management to analysis, prediction, visualization, user interaction, communication, sharing, and re-use. The ultimate goal is to unleash the power of scientific data to deepen our understanding of physical, biological, and digital systems, gain insight into human social and economic behaviour, and design new solutions for the future. We are interested in general methods and concepts, as well as specific tools, infrastructures, and applications. The rising importance of scientific data, both big and small, brings with it a wealth of challenges to combine structured, but often siloed data with messy, incomplete, and unstructured data from text, audio, visual content such as sensor and weblog data. New methods to extract, transport, pool, refine, store, analyze, and visualize data are needed to unleash their power while simultaneously making tools and workflows easier to use by the public at large. The journal invites contributions ranging from theoretical and foundational research to platforms, methods, applications, and tools in all areas. We welcome papers which add a social, geographical, and temporal dimension to Data Science research, as well as application-oriented papers that prepare and use data in discovery research.

Our journal has a number of features to maximize the transparency, speed, and quality with which results are published and made available for current and future reuse and interpretation. First of all, Data Science is an open access journal, which increases the visibility and enables simple access and use of the reported results. Article processing charges will be waived for the first year and charges thereafter will be reasonable and competitive. We are furthermore committed to minimizing the time it takes to obtain a decision on a submitted manuscript. For that reason, Data Science gives reviewers only 10 days to respond and aims for sending out first decisions on submissions within weeks rather than months.

In order to increase the visibility and recognition of reviewers and to promote accountability in the reviewing process, Data Science has opted for reviews to be open and attributable. All reviews are made freely available under CC-BY licenses after a decision has been made for the submission (independent of whether the decision was to accept or reject). In addition to solicited reviews, everybody is welcome to submit additional reviews and comments for papers that are under review. Reviews are non-anonymous by default, although reviewers can request to stay anonymous. The journal attributes the work of editors and non-anonymous reviewers in all published articles.

All Data Science submissions will moreover be publicly available as preprints right away. Publishing preprints has the advantage of establishing a precedent for the work while it undergoes peer review. Manuscripts submitted to Data Science are made available as preprints prior to reviewing so that reviewers and others are free to not only read, but also share submitted papers. Preprints will remain available after reviewing, independent of whether the paper was accepted or rejected for publication.

Enabling access to content in a manner that conforms to community standards is a key part of the FAIR principles, ensuring that these results can be easily reused in other contexts. Data Science requires authors to represent and provide any data used or produced in their studies with community-based data formats and metadata standards. These data should furthermore be made openly available free of charge, unless privacy or other well founded concerns apply.

Finally, in order to experiment with better communication methods for the future of scientific communication, we encourage authors to write their papers in HTML and to provide (meta)data with formal semantics, as a step towards the vision of semantic publishing, which will allow us – to a certain extent – to automatically integrate, combine, organize, and reuse scientific knowledge.

3. This issue

This inaugural issue features position papers on various aspects around Data Science. Specifically, these aspects cover new types of insights we can gain from data, new types of data that have become available and require new methods and tools, urgent social issues that stem from these new types of data-driven research, and new approaches on the role of data in the scientific publishing process.

3.1. New kinds of insights from data

New data with increased coverage and size might allow us to arrive at new types of insights. For example, could such data let us predict the outbreak of conflicts and wars? In his paper “conflict forecasting and its limits”, Thomas Chadefaux explores this question, which is both very important for humanity and very complicated to answer. He argues that the degree to which we can find answers depends on whether conflicts behave like clocks (predictable), clouds (difficult to predict), or black swans (unpredictable). Before we can make real progress on prediction of conflicts, we need to understand the fundamental nature of such events.

3.2. New kinds of data

In the last decades, we have also witnessed the emergence of entirely new types of data. Specifically, three papers of the inaugural issue look into the new kinds of data that are rooted in symbolic logic with languages of precisely defined syntax and semantics, and how such kind of data can be combined with the prevalent statistical and machine learning approaches for data analytics. A fourth paper focuses on the fact that such semantic data increasingly come in the form of continuous and dynamic data streams, rather than discrete releases of static datasets.

Lawrence E. Hunter argues in his paper entitled “Knowledge-based biomedical Data Science” that Data Science in general, and in the domain of biomedicine in particular, can benefit a great deal from the existing body of research on computational knowledge representation and reasoning. Applications can include logical inference on ontology annotations of domain entities, such as genes and biological processes, and on the biomedical literature where these entities are further described, as well as the automated generation and evaluation of hypotheses.

Robert Hoehndorf and Núria Queralt-Rosinach make a similar argument in their paper “Data science and symbolic AI: Synergies, challenges and opportunities” proposing a symbiotic combination of statistical and symbolic Data Science, thereby combining symbolic AI approaches like ontologies and reasoning with statistical approaches like machine learning and probabilistic models. Such a symbiotic system can consume data and – importantly – existing knowledge to reliably produce new knowledge.

Xander Wilcke, Peter Bloem, and Victor de Boer explain in their paper “The knowledge graph as the default data model for learning on heterogeneous knowledge” how we can use techniques and models that originate from symbolic AI research, specifically knowledge graphs and Linked Data, as underpinning unifying model for all kinds of machine learning approaches. With the rise of deep learning, raw data has become the preferred input instead of manually constructed features, but this doesn’t work well for heterogeneous data, for example data that includes images, sound, and text. A formal knowledge graph is – maybe surprisingly – in a sense “rawer” than heterogeneous raw data, because all information is preserved in a general and uniform manner and can potentially be capitalized on by a machine learning algorithm.

Daniele by Dell’Aglio, Emanuele Della Valle, Frank van Harmelen, and Abraham Bernstein then focus in their paper “Stream reasoning: A survey and outlook” on dynamic streams of such formally represented data. They explain how the field has evolved in its ten years of existence and sketch the open challenges they see on the road ahead.

3.3. Social aspects around Data Science

Data Science is by nature cross-disciplinary, with the potential to bridge virtually any scientific endeavors within or between sectors. A series of five articles in this journal address the social aspects and connecting abilities of Data Science, when optimally engaged in research, education, and business. In “Maintaining intellectual diversity in Data Science”, Richard P. Mann and Olivia Woolley-Meza argue that intellectual diversity leads to the combined approaches that are needed to address complex phenomena. In order to enrich the research community, appropriate incentives should be implemented to support diverse thought and approaches, which may be initially perceived as unfashionable.

Furthermore, multi-disciplinary teams must include a data scientist to rapidly advance biomedical research, as Manisha Desai posits in her paper, “The integration of the data scientist into the team: Implications and challenges”. Team science, interdisciplinary collaboration across disciplines to address important scientific questions, is key to successful biomedical research. Yet, academic medical centers do not yet have the professional advancement structures in place to incentivize the integration and promotion of data scientists into the clinical and translational research team.

To promote cross-disciplinary Data Science, Evangelos Pournaras describes in his paper,“Cross-disciplinary higher education of Data Science – beyond the computer science student”, the experience of designing and implementing a postgraduate Data Science course for students of all disciplines. While the experience of developing the course was challenging, the course was deemed to be highly rewarding for the variety of students engaged. Innovative curricular designs like this demonstrate the possibilities for disciplines to learn from one another based on connecting nature of Data Science.

In her paper, “Thoughtful artificial intelligence: Forging a new partnership for Data Science and scientific discovery”, Yolanda Gil describes a vision for thoughtful artificial intelligence systems, which would partner with the scientist in driving research and discovery beyond what individual researchers can accomplish individually, and perhaps with less error and bias than can be manually achieved. Human creativity and insight could overcome the shortcomings of AI, just as AI can overcome human shortcomings in performing science.

Christine Chichester, finally, argues in her paper “Valorizing omics visualization for discovery” that we should start appreciate that visualizations are more than just eye candy for presentation purposes, but rather that they form a crucial step in data-driven knowledge discovery. Visualization efforts need to be given proper credit and need to be encouraged and supported to advance our knowledge discovery abilities.

3.4. Scientific publishing

Finally, the inaugural issue contains two papers on semantic publishing, which is the field of research on how we can use semantic technologies to improve the communication and dissemination of scientific findings and metadata. This includes our own paper entitled “Genuine semantic publishing”, in which we argue that we should return to the original and literal meaning of *semantic publishing* by letting the authors themselves represent their results in a semantic notation in the first place, and release these representations as prime publication elements.

While we argue for requiring authors to do some extra work, Silvio Peroni takes the opposite stance in his paper “Automating semantic publishing”, declaring that with appropriate methods a large amount of formal and semantic structure can automatically extracted from already written papers, thereby minimizing the additional effort needed on the side of the authors. He then presents a specific approach and framework to gradually building semantic information starting from low-level syntactic elements in a markup language like HTML.

4. Invitation for contributions

For the upcoming issues we are looking for different kinds of papers on the topics outlined above. Most importantly, we are inviting the submission of research papers that report on original research. But we are also interested in position papers that present relevant and novel discussion points around the journal topics in a thorough manner, like the papers of this inaugural issue. We moreover also accept survey papers on the state of the art of relevant topics to serve as introductory and overview texts for interested readers.

We furthermore welcome suggestions for special issues. The first special issue call on "Special Issue. Distributed Ledgers: Making Data Science More Open, Transparent, and Accountable" is already open and accepting submissions. We aim for a continuous stream of such special issues on a variety of Data Science topics. More information about the journal and its open calls can be found on our website https://datasciencehub.net.