Towards a data science collaboratory

Towards a Data Science Collaboratory

Joaquin Vanschoren[1], Bernd Bischl[2], Frank Hutter[3], Michele Sebag[4], Balazs Kegl[4], Matthias Schmid[5], Giulio Napolitano[5], Katy Wolstencroft[6], Alan R. Williams[7], and Neil Lawrence[8]

[1] Eindhoven University of Technology, Eindhoven, Netherlands,
j.vanschoren@tue.nl
[2] Ludwig-Maximilians-University Munich, Munich, Germany
bernd.bischl@stat.uni-muenchen.de
[3] Albert-Ludwigs-Universität Freiburg, Freiburg, Germany
fh@informatik.uni-freiburg.de
[4] Université Paris Sud, Paris, France
michele.sebag@lri.fr, balazs.kegl@gmail.com
[5] Universität Bonn, Bonn, Germany
{matthias.schmid,giulio}@imbie.meb.uni-bonn.de
[6] Universiteit Leiden, Leiden, The Netherlands
k.j.wolstencroft@liacs.leidenuniv.nl
[7] University of Manchester, Manchester, U.K.
alan.r.williams@manchester.ac.uk
[8] University of Sheffield, Sheffield, U.K.
n.lawrence@dcs.shef.ac.uk

Abstract. Data-driven research requires many people from different domains to collaborate efficiently. The domain scientist collects and analyzes scientific data, the data scientist develops new techniques, and the tool developer implements, optimizes and maintains existing techniques to be used throughout science and industry. Today, however, this data science expertise lies fragmented in loosely connected communities and scattered over many people, making it very hard to find the right expertise, data and tools at the right time. Collaborations are typically small and cross-domain knowledge transfer through the literature is slow. Although progress has been made, it is far from easy for one to build on the latest results of the other and collaborate effortlessly across domains. This slows down data-driven research and innovation, drives up costs and exacerbates the risks associated with the inappropriate use of data science techniques.

We propose to create an open, online collaboration platform, a 'collaboratory' for data-driven research, that brings together data scientists, domain scientists and tool developers on the same platform. It will enable data scientists to evaluate their latest techniques on many current scientific datasets, allow domain scientists to discover which techniques work best on their data, and engage tool developers to share in the latest developments. It will change the scale of collaborations from small to potentially massive, and from periodic to real-time. This will be an inclusive movement operating across academia, healthcare, and industry, and empower more students to engage in data science.

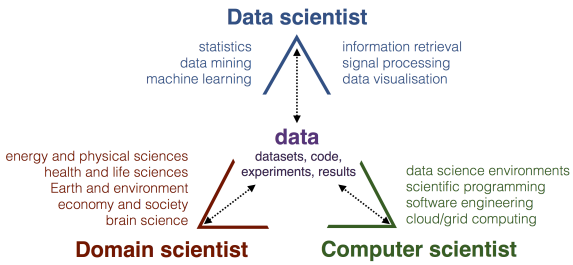

Fig. 1. Roles within the data science ecosystem and the gaps between them.

1 Motivation

Our increasing ability to collect, store, interconnect and analyse large sets of data is revolutionizing science, healthcare, and industry, transforming every-day life in front of our eyes. Today, however, data science expertise lies fragmented in loosely connected communities and scattered over many people, making it very hard to find the right expertise, data and tools at the right time. This creates deep scientific and societal challenges, slowing down data-driven research and innovation, driving up costs and exacerbating the risks associated with the inappropriate use of data science techniques.

1.1 Large amount of underpowered and irreproducible research

While scientific papers are being produced at a tremendous rate, reproducible and readily applicable major discoveries are far fewer. In biomedicine alone, an estimated 85% of research resources are wasted [17, 18] because of false or exaggerated findings (often the result of less than rigorous analysis), lack of transparency and reproducibility, and lack of large-scale, interdisciplinary collaboration. Irreproducibility and overly general conclusions have also been highlighted as among the most important challenges to preclinical research on drug targets [2] and data mining research [15].

1.2 The fragmented data science ecosystem

Data-driven research requires many different people to collaborate efficiently: the domain scientist collects and analyzes scientific data to study or discover phenomena; the data scientist designs and analyzes algorithms, to solve specific problems or benchmark on a domain of data sets; the tool developer implements, optimizes and maintains existing techniques to be used by many people in science and industry [20]. This is shown in Fig. 1. However, it is far from easy for these people to collaborate efficiently.

First, there is a gap between data scientists and domain scientists that slows down the rate of research tremendously. Because domain scientists have a more

limited view on the state of the art in data science, they are often unsure about the latest and most appropriate techniques, and either spend a lot of time on literature research and trial and error or make suboptimal choices. Data scientists, on the other hand, often don't speak the language of domain scientists, hence missing opportunities to work interactively with them and innovate in their respective fields. Knowledge transfer through the literature is slow: findings are spread over millions of papers, based on tacit domain-specific knowledge and community-specific jargon, and results are hard to reproduce or build upon. Small-scale collaboration, likewise, requires a significant time investment, and luck, to find the people that have the right skills, resources and the time to delve into a new problem. Moreover, experts tend to leverage their own specialized work which is not necessarily the best solution for the given problem. In all, scientists spend vast amounts of time on tasks that others could do in a fraction of that time, that could be done much better using novel/better techniques, or that could be automated altogether.

There is also a gap between academia and industry. Many innovative enterprises need advanced data science expertise in order to compete, but often cannot easily access the latest scientific methods and scientific data. Large companies are collecting and acquiring large amounts of data (e.g., Facebook's purchase of WhatsApp), and acquiring key data scientists (e.g., Facebook hiring NYU's Yann LeCun, Baidu hiring Stanford's Andrew Ng, Google buying DeepMind and DNNResearch). This is a clear sign of the success of the field, but it creates a major challenge for smaller firms that want to enter the market: they are debilitated by lack of access to data, and lack of access to expertise [21].

In healthcare, medical professionals are confronting diseases that result from a large range of causes (e.g., Alzheimer's, heart disease, cancer), or that coevolve with clinical advances (drug resistance). Scientists and clinicians, to better characterize the causes of a disease, and understand how diseases manage to subvert existing cures, now require large-scale, heavily interconnected data sources and novel techniques for analysis. These include data sources that characterize people's genotype (high-resolution sequencing) and phenotype (gene expression, clinical measurements, wearable sensors to track activity and body signals). Again, domain scientists and data scientists should work closely together to tackle these hard challenges, so that the best techniques can be leveraged to speed up scientific work. An additional major challenge is ensuring that privacy is protected whilst leveraging data for wider social benefit [21, 22].

1.3 Fragmented access to tools and data

As shown in Fig. 1, what connects these communities is data, in the form of datasets, computer programs and experimental results. There exist many wonderful open data repositories for scientists [12], however, for many data scientists it is far from obvious how to access, search and understand these data islands, and they often choose to benchmark their algorithms on a limited set of old datasets. As a result, the resulting empirical results are of little value to domain scientists and industry.

Vice versa, for domain scientists it is far from obvious how to access and understand the latest algorithms, or simply compare them to assess which are most useful. There are great software tools and libraries for data scientists, but they too are islands that focus on more standard techniques, and even then it is not obvious which components are most useful. Empirical evaluations of all these algorithms on large numbers of datasets are typically not organized online. Some are 'printed' in papers, using varying experimental setups, which makes them virtually impossible to build on.

In short, while it is theoretically possible for these communities to build directly on each other's work, in practice there is a lot of friction that inhibits efficient collaboration. When scientists need to understand new data formats, study source code, learn new programming languages, email the authors for code or data, let alone sign complex paperwork, they typically 'give up' and do things the old fashioned way, sticking to the simple tools and data they already know. We need to rethink the way we share data and code, and remove friction so that others can effortlessly access and build on them.

1.4 Manpower

The McKinsey Global Institute is expecting a shortage of 140,000 to 190,000 data scientists by 2018 in the US alone [23]. Moreover, the success of data science in industry means that large, multinational companies are attracting many data scientists from the academic sector. This makes it increasingly difficult to find competent data scientists to work in important public societal domains (e.g. health care, environment, transportation, energy) and scientific research.

However, it is unlikely that we can solve this problem by simply training armies of data scientists. The problem is that data science itself does not scale because of the fragmentation and inefficiencies discussed above. We need to facilitate access to the very best data, tools and training materials, enable frictionless collaboration to 'connect brains', and automate much of the drudge work that currently slows down data science. This will also allow more students to succeed.

2 An online collaboratory

To address these issues, we propose to scale up the practice of data science from small-scale local collaborations to frictionless, real-time, massively collaborative online collaborations. This goal can be broken down into smaller, specific objectives as outlined below.

2.1 Organized, easy access to data sources and software tools

To address the fragmentation of tools and data sources, we need to bring together domain scientists (including the teams that curate open data) and data scientists (including those that maintain data analysis toolboxes). The goal is to extract actionable datasets from large scientific databases and identify key

software components, and then annotate them with a practical base vocabulary to create a structured index, a 'search engine for data science', that allows users to easily search for datasets, algorithms, and workflows. This will help domain scientists to quickly find useful tools (e.g. scalable clustering algorithms), allow data scientists to test their algorithms on a large range of scientific datasets, and give students easy access to the state of the art.

2.2 Change the scale of collaboration

To help bridge the gap between data scientists and domain scientists, we need to bring them together on the same online collaboration platform, or collaboratory, for data-driven research. In this collaboratory, data scientists and domain scientists should be able to collaborate and build directly on each other's results. A domain scientist can share a dataset of interest and issue an open invitation to anyone in the world (or a circle of trusted people) to help analyse it. Data scientists respond by sharing experiments showing how well their particular solution works, and other people can again build on that. Crucially, all experiments should be linked to the underlying data sets and workflows, ensuring reproducibility and also creating a single, large, organized body of research. This enables large-scale collaborations across domains because everyone can build on the combined results of all other researchers. At the same time, it creates a reference where companies can discover interesting datasets, techniques, and people to collaborate with (or hire), as well as a learning environment for students to learn practical data science in an engaging way, actively interacting with the community, and possibly collaborating with top scientists. This will help them gain visibility and prepare them to fulfill key roles in industry and academic research. To protect preliminary work, the collaboratory should naturally support social networking and privacy settings, and scientists keep ownership of all shared resources and experiments.

2.3 Change the speed of collaboration

The collaboratory should be very easy to use so that researchers can focus on their work without distraction. Scientists should be able to share experiments automatically (in the background), from the tools they already use, and from the computational infrastructure they prefer. This can be achieved by extending these tools to seamlessly connect to the collaboratory, and download or upload results together with all the necessary metadata to ensure reproducibility. As a result, experiments can be shared in real time: other people (in your collaboration) will be able to comment on or build on your results the second they are uploaded. It would even be possible to further automate experimentation, e.g., to run algorithms on all newly uploaded datasets.

2.4 Automating data science

Expertise in data science is not easily expressed: it is often a 'gut feeling' about which techniques may help or not. However, having a large collection of data

science experiments on all kinds of data sets offers unprecedented opportunities for machine learning techniques to discover patterns in which approaches work best on certain types of data, and thus make informed decisions based on many prior observations. For instance, they could recommend specific algorithms (e.g. which outlier detection to run), or intelligently optimize large parameter spaces, thus saving a lot of time. This will allow both academic and industrial data scientists to design better data analysis pipelines much faster, and thus be more productive.

2.5 Cross-domain collaboration protocols

To address the problem of underpowered, irreproducible research, we can follow the example of other sciences, such as genetic epidemiology, which have addressed this issue by adopting large-scale collaborative research with a strong replication culture [7], promoting the open sharing of data, protocols, and software [34, 29], adopting more appropriate statistical methods [26, 19], and providing continued education of scientists in research methods and statistical literacy [8]. Likewise, we should bring together domain scientists, data scientists and tool developers to develop a common understanding of specific research problems (e.g., genome-wide association studies). Crucially, they need to establish which metadata is necessary to maximize reproducibility, how to handle data confidentiality, which are the best practices in solving these problems (such as proper statistical significance testing), and how the end result can be objectively evaluated (if possible at all). This should result in protocols defining how datasets are annotated and how experiments are stored, so that experiments can be reproduced and collaborations can scale.

2.6 Provide a new incentive structure for research

Finally, the collaboratory should make it clear how scientists should be credited for their contributions, and automatically track the impact of these contributions online. Beyond tracking citations, it could track the reuse of datasets, code and experiments within the collaboratory, as well as social interactions (e.g. follows, shares, downloads) in an open and transparent manner (so that misuse is also easy to spot). Hence, it will allow scientists to demonstrate the broader impact of their work, make them more visible, and make it much easier for others to build on their work and to collaborate with them.

3 Illustration: A user story

To sketch the current working situation of data scientists, and bring across the possible impact of the proposed collaboratory, we formulate a user story featuring Elisa, a young fictitious researcher working at one of Europe's flagship bio-informatics institutes. She is analyzing her data, but is unsure about which

data analysis techniques will work best. She spends weeks sifting through the literature, but as she tries out different prior approaches, she finds that many won't scale to her high-dimensional data, some are only available in unfamiliar tools (or not at all), and others looked promising but worked poorly. Next, she spends a lot of time optimizing the parameters of her workflow in a time-consuming grid search, painstakingly recording results in notebooks, spreadsheets and databases. When she wants to compare against the state of the art, she again spends weeks hunting down data and code and emailing authors for help. After she submits her paper, the reviewers ask her to rerun experiments because she used a statistical method incorrectly. Finally, when her paper is finally published, she still needs to manually annotate and upload her results into a public database. When a prominent colleague emails her a question a year later, she discovers that she unfortunately lost track of some crucial details and cant reproduce her earlier results.

Then, she discovers that her data analysis software was updated with an collaboratory plugin. She has recently created a new dataset and she wants to try it out. When she uploads her dataset to the collaboratory, it automatically analyses it and provides an easy-to-read overview of interesting statistics, as well as a list of very similar datasets. Results obtained by other scientists on these datasets enable a direct comparison of existing techniques. She immediately sees which techniques show most promise. In the comments, she finds advice on which parameters to tune and which implementations are most scalable. She runs a new set of experiments with her favourite data analysis tool and a collaboratory plugin: her results are now automatically uploaded, with all details necessary for reproducibility, and appear online, visible only to her, organized and linked to her dataset. While the results are streaming in, Elisa connects to all her trusted colleagues so she can share and discuss her results with them. They can now follow her research in real time, and quickly provide some useful suggestions. In the mean time, the collaboratory is also continuously learning from her experiments: on demand, it can recommend specific techniques (e.g. feature selection techniques) and intelligently optimize all the parameters of her workflow. In the end, she creates an online study out of her experiments, makes it public, and puts a backlink in her paper so that others can find the full details online. This further enlarges the collective body of structured knowledge, empowering the researcher after her. As her methods and results are easily accessible, applicable and modifiable, her work is quickly reused, cited more often, and the collaboratory tracks this to show this impact. People also contact her directly, leading to many new collaborations.

4 Related work

This collaboratory would not have to be created from scratch. There already exist several tools and infrastructures that form a solid foundation.

First, there are many open data sources, that vary considerably across domains. In biodiversity, there exists a community-agreed species occurrence reg-

istry (GBIF), with a single API and data format [10]. The environmental sciences use a federated, semantic registry for environmental-related data, DataONE [24]. Bioinformatics is driven by disparate tools and databases, but also has common data formats such as FASTA [28], and model formats such as SBML [16]. ELIXIR [9] offers a tools and data services registry leveraging other registries. Finally, in pharmacology, there exist data exploration portals that provide a common interface to disparate data sources, such as OpenPHACTS [37], BioGPS [40] and MyGene.info [39].

Moreover, initiatives like DataONE and CEDAR[1] promote cross-disciplinary standards to describe datasets with metadata, building on the ISA-tools community standard [31] for describing and aggregating studies, and BioSharing.org [32], a registry used by many journals for policies, standards and databases in the biological and biomedical sciences. Workflow sharing platforms, such as myExperiment [13], SHIWA [27] and Galaxy Tool Shed [5] allow scientists to share and find data analysis workflows. SEEK [38] is a platform designed to store, annotate, share and interlink heterogeneous data and models.

Other tools to share research in a reproducible and reusable way are Research Objects [1], publishable bundles of datasets, code and other materials, as well as IPython/Jupyter Notebooks [33], documents that mix programming code, text and other media, and that can be re-executed or altered at any time. Recomputation.org [11] is an initiative for making computational experiments available in reproducible and reusable form as virtual machines.

There are also several initiatives to improve the practice of scientific data analysis through community guidelines and training. The STRATOS initiative[2] aims to provide guidelines for the design of observational medical research, including the use of appropriate statistical methods and accurate result interpretation. The Data Carpentry workshops[3] teach researchers to analyse data more effectively and productively. Finally, data science challenges, such as those run on Kaggle [6], also help bridge the gap between data scientists and domain scientists. They focus data scientists attention on a specific problem, identify talented people, and encourage data scientists and domain scientists to collaborate in teams. However, the traditional approach to data challenges will not work when solving open problems, with open ended data and without a 'correct' solution; it also emphasizes competition and inhibits collaboration.

There has also been great progress in algorithm selection and configuration techniques. The Automatic Statistician[4] is a service that automatically analyzes a dataset and returns statistics and models in an easy-to-read report. Modern work on combined algorithm selection and hyperparameter optimization is also very effective [35] and can be substantially improved by learning across datasets.

What is still missing in the state-of-the-art are community-driven, open collaboration platforms that bring together domain scientists and data scientists on

[1] http://med.stanford.edu/cedar.html
[2] http://www.stratos-initiative.org/
[3] http://www.datacarpentry.org
[4] http://www.automaticstatistician.com/

the same platform. Doing so will allow data scientists to access disparate open data sources and download many datasets to evaluate their techniques. Domain researchers will be able to find implementations of techniques that perform specific tasks, ranked by how well they performed on prior datasets, and so decide which techniques to try. Ideally, this is realized by building on existing platforms, standards, data and workflow repositories. This will maximise sustainability and broad impact, as these components and the collaboration platform will co-evolve.

5 OpenML

One platform that already takes a step in this direction is OpenML, an online machine learning platform where researchers can *automatically* log and share data, code, and experiments in fine detail and organize them online to work and collaborate more effectively [36]. It organizes the following information:

Data sets Data sets can be shared publicly or within *circles* of researchers. They can be uploaded or simply linked from existing scientific data repositories. For known data formats, OpenML will automatically analyze and annotate the data sets with measurable characteristics, so that they can be searched and analyzed based on this meta-data. Data sets can be updated and are automatically versioned. Currently however, OpenML is not not linked to the open data sources discussed above, and it understands only tabular data formats.

Tasks Data sets typically serve as input for scientific *tasks*, defining which inputs are given, which outputs are expected to be returned, and what scientific protocols should be used. By creating a task, scientists can challenge the data science community to come up with the best solution to solve it. Hence, they are similar to data mining challenges on platforms such as Kaggle [6], except that they are collaborative and real-time: scientists are encouraged to post their solution as quickly as possible (as in scientific publishing), and other scientists can immediately build on that solution. OpenML builds human and machine-readable descriptions of such tasks. Currently however, the set of supported tasks is mostly limited towards core machine learning studies. Further discussions are needed between domain scientists to properly define a wider range of tasks.

Flows Flows are implementations of data analysis workflows. They can be single algorithm implementations, scripts (e.g., in R) or workflows (e.g., in tools such as RapidMiner [30] and KNIME [3]). They are again shared publicly or within *circles*, can be uploaded or linked from existing repositories (e.g. the workflow repositories discussed above), and updates are automatically versioned. Ideally, they are wrappers around existing software that take OpenML tasks as inputs. This allows automatic execution of algorithms on new data sets, but this is not required. The platform is designed so that flows can be run on any computational infrastructure, such as the researchers own computers, and the results uploaded. There is currently no support to run IPython, R, or other code on the servers.

Runs Runs are the results of executing flows on tasks, uploaded to the platform. They are fully reproducible, linked to the data set and flow versions, hyperparameter settings, and information on the authors and computational hardware. They are also reusable, containing non-aggregated results depending on the task. Where possible, runs are evaluated on the server to allow objective comparisons, using a broad range of evaluation measures. Because runs are automatically linked to the underlying tasks, flows, and authors, they can be easily searched and compared across different data sets and flows.

Integrations OpenML features an extensive REST API to find and upload data sets, download tasks, find and upload flows, and download or upload runs. Moreover, programming APIs are offered in Java, R and Python to allow easy integration into existing software tools. For instance, using the OpenML[5] package for R, one can authenticate, search and download datasets, and upload the results of machine learning experiments in just a few lines of code. Using these APIs, it is also directly integrated in machine learning toolboxes such as WEKA [14], MOA [4], HubMiner[6], and mlr[7]. While this is a good start, a lot of key data analysis toolboxes, platforms, and programming APIs are still missing.

Website OpenML.org is a website offering easy access to most OpenML functionality. It allows users to browse through all shared datasets, flows and runs. It compares all results obtained on specific tasks and flows, and has dedicated pages for each data set, task, flow and run with all known details. When logged in, you can also upload new data sets and flows, create new tasks, add comments, and organize information through tagging and wiki-like editing.

6 Summary

Many sciences have made significant breakthroughs by adopting online tools that help organize, structure and analyze detailed scientific data online [25]. One of the most important reasons for this is that large amounts of (open) data can now be analysed by algorithms. As we have shown, however, the required data science expertise is fragmented over many people and communities, slowing down data-driven research and innovation. To alleviate this fragmentation, we propose to fast-track the creation of an online collaboratory for data-driven science based on open data, open source tools, and open science. We hope to bring together data scientists, tool developers, and domain scientists from many different domains to see how current tools can be connected online, and how best practices can be shared. We expect that this networked approach to data science has great potential to speed up the rate of discovery in data-driven science. It will lead to more effective collaborations between researchers and higher efficiency and creativity in research. Indeed, if scientists and data scientists everywhere are

[5] `https://github.com/openml/r`
[6] `http://mloss.org/software/view/574/`
[7] `https://github.com/berndbischl/mlr`

part of the same 'experimentation system in the sky', they can all become much more productive, try out new ideas very quickly, and efficiently learn from each other. This has substantial long-term benefits:

We can automate many aspects of data science research. Today, a lot of the data scientist's time is devoted to activities that are clerical or mechanical: trying many techniques, finding related work, and comparing different approaches. Through the proposed platform, data scientists will be able to make informed decisions without needing to set up large sets of experiments every step along the way. Guesswork will be largely eliminated, replaced with a data-driven approach to data science. Moreover, it will allow companies to speed up data-driven analysis and drive down costs. Tasks that used to require a team of data scientists could be achieved by the domain scientist alone with the help of automated tools and a worldwide support network.

It will facilitate the transfer of data science innovations across domains. Indeed, expertise on data analysis in one field can be very useful to analyze data in another. For instance, expertise in analysing high-dimensional gene expression data could be used to analyse high-dimensional molecule descriptions.

Facilitating online scientific collaboration has great potential to speed up data-driven science in industry. Research in drug discovery, bio-informatics, health and many other industries can be sped up through frictionless, large-scale collaboration with data scientists, which in turn may speed up the development of new drugs, treatments, and generally improving our quality of life.

A whole generation of new data scientists can be trained much more effectively. Knowledge that is currently scattered over thousands of papers will be united in a single organized resource. Students will be able to actively interact with the community, learning important skills for their professional careers.

An easy to use search engine for data science will make it easier to find open datasets, as well as all results obtained from them (visualizations, models,...). This may inspire young people to become scientists, steep students in data-driven research, encourage entrepreneurs to leverage this data, and facilitate data journalism and citizen science initiatives. Moreover, it can be used to focus the data science community on important societal challenges. Data scientists are naturally drawn to places that offer interesting data, and love to use novel tools to find patterns in that data. Through the collaboratory, it becomes much easier for data scientists to get involved.

References

1. Bechhofer, S., De Roure, D., Gamble, M., Goble, C., Buchan, I.: Research objects: Towards exchange and reuse of digital knowledge. The Future of the Web for Collaborative Science (2010)

2. Begley CG, E.L.: Drug development: Raise standards for preclinical cancer research. Nature 483, 531–533 (2012)
3. Berthold, M., Cebron, N., Dill, F., Gabriel, T., Kotter, T., Meini, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: KNIME: The Konstanz information miner. Studies in Classification, Data Analysis, and Knowledge Organization 5, 319–326 (2008)
4. Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B.: MOA: Massive Online Analysis. Journal of Machine Learning Research 11, 1601–1604 (2010)
5. Blankenberg, D., Von Kuster, G., Bouvier, E., Baker, D., Afgan, E., Stoler, N., Taylor, J., Nekrutenko, A., et al.: Dissemination of scientific software with galaxy toolshed. Genome Biol 15(2), 403 (2014)
6. Carpenter, J.: May the best analyst win. Science 331(6018), 698–699 (2011)
7. Chanock, S., Manolio, T., Boehnke, M., Boerwinkle, E.e.a.: Replicating genotype-phenotype associations. Nature 447(7145), 655–660 (2007)
8. Collins, F., Tabak, L.: NIH plans to enhance reproducibility. Nature 505, 612–613 (2014)
9. Crosswell, L.C., Thornton, J.M.: ELIXIR: a distributed infrastructure for european biological data. Trends in biotechnology 30(5), 241–242 (2012)
10. Edwards, J.L., Lane, M.A., Nielsen, E.S.: Interoperability of biodiversity databases: biodiversity information on every desktop. Science 289(5488), 2312–2314 (2000)
11. Gent, I.P.: The recomputation manifesto. arXiv preprint arXiv:1304.3674 (2013)
12. Gibbs, S.: EU commits 14.4m to support open data (2014), `http://gu.com/p/432m4/sbl`
13. Goble, C.A., De Roure, D.C.: myExperiment: social networking for workflow-using e-scientists. In: Proceedings of the 2nd workshop on Workflows in support of large-scale science. pp. 1–2. ACM (2007)
14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: An update. SIGKDD Explorat. 11(1), 10–18 (2009)
15. Hirsh, H.: Data mining research: Current status and future opportunities. Statistical Analysis and Data Mining 1(2), 104–107 (2008)
16. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., et al.: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19(4), 524–531 (2003)
17. Ioannidis, J.: Why most published research findings are false. PLoS Med 2(8), e214 (2005)
18. Ioannidis, J.: Why most discovered true associations are inflated. Epidemiology 19, 640–648 (2008)
19. Johnson, V.: Scientific method: statistical errors. Proceedings of the National Acadademy of Sciences USA 110, 19313–19317 (2014)
20. Kegl, B.: The data science ecosystem (2015), `https://medium.com/@balazskegl/the-data-science-ecosystem-678459ba6013`
21. Lawrence, N.: Open data science (2014), `http://inverseprobability.com/2014/07/01/open-data-science/`
22. Lawrence, N.: Beware the rise of the digital oligarchy (2015), `http://gu.com/p/469qm/sbl`
23. McKinsey Global Institute: Big data: The next frontier for competition, `http://www.mckinsey.com/features/big_data`
24. Michener, W.K., Allard, S., Budden, A., Cook, R.B., Douglass, K., Frame, M., Kelling, S., Koskela, R., Tenopir, C., Vieglais, D.A.: Participatory design of dataoneenabling cyberinfrastructure for the biological and environmental sciences. Ecological Informatics 11, 5–15 (2012)

25. Nielsen, M.: Reinventing discovery: the new era of networked science. Princeton University Press (2012)
26. Nuzzo, R.: Scientific method: statistical errors. Nature 506, 150–152 (2014)
27. Olabarriaga, S., Pierantoni, G., Taffoni, G., Sciacca, E., Jaghoori, M., Korkhov, V., Castelli, G., Vuerli, C., Becciani, U., Carley, E., et al.: Scientific workflow management–for whom? In: e-Science, 2014 IEEE 10th International Conference on. vol. 1, pp. 298–305. IEEE (2014)
28. Pearson, W.R., Lipman, D.J.: Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences 85(8), 2444–2448 (1988)
29. Peng, R.: Reproducible research in computational science. Science 334, 1226–1227 (2011)
30. van Rijn, J.N., Umaashankar, V., Fischer, S., Bischl, B., Torgo, L., Gao, B., Winter, P., Wiswedel, B., Berthold, M.R., Vanschoren, J.: A RapidMiner extension for open machine learning. In: RCOMM 2013. pp. 59–70 (2013)
31. Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., et al.: ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics 26(18), 2354–2356 (2010)
32. Sansone, S.A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., et al.: Toward interoperable bioscience data. Nature genetics 44(2), 121–126 (2012)
33. Shen, H.: Interactive notebooks: Sharing the code. Nature 515(7525), 151–152 (2014)
34. Stodden, V., Guo, P., Ma, Z.: Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. PLoS ONE 8, e67111 (2013)
35. Thornton, C., Hutter, F., Hoos, H.H., Leyton-Brown, K.: Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 847–855. KDD '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2487575.2487629
36. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: OpenML: Networked science in machine learning. SIGKDD Explorations 15(2), 49–60 (2013)
37. Williams, A.J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E.L., Evelo, C.T., Blomberg, N., Ecker, G., Goble, C., et al.: Open PHACTS: semantic interoperability for drug discovery. Drug discovery today 17(21), 1188–1198 (2012)
38. Wolstencroft, K., Owen, S., du Preez, F., Krebs, O., Mueller, W., Goble, C., Snoep, J.L.: The SEEK: a platform for sharing data and models in systems biology. Methods Enzymol 500, 629–655 (2011)
39. Wu, C., MacLeod, I., Su, A.I.: BioGPS and MyGene.info: organizing online, gene-centric information. Nucleic acids research p. gks1114 (2012)
40. Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C.L., Haase, J., Janes, J., Huss, J.W., et al.: BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol 10(11), R130 (2009)