Geophysical Research Abstracts
Vol. 19, EGU2017-12179, 2017
EGU General Assembly 2017

The Swiss Data Science Center on a mission to empower reproducible, traceable and reusable science

Stanislaus Schymanski (1), Eric Bouillet (2), and Olivier Verscheure (2)

(1) Eidgenössische Technische Hochschule (ETH) Zurich, Department of Environmental Systems Science, Zurich, Switzerland (stanislaus.schymanski@env.ethz.ch), (2) Ecole Polytechnique Fédérale de Lausanne (EPFL), Swiss Data Science Center (SDSC), Lausanne, Switzerland

Our abilities to collect, store and analyse scientific data have sky-rocketed in the past decades, but at the same time, a disconnect between data scientists, domain experts and data providers has begun to emerge. Data scientists are developing more and more powerful algorithms for data mining and analysis, while data providers are making more and more data publicly available, and yet many, if not most, discoveries are based on specific data and/or algorithms that "are available from the authors upon request".

In the strong belief that scientific progress would be much faster if reproduction and re-use of such data and algorithms was made easier, the Swiss Data Science Center (SDSC) has committed to provide an open framework for the handling and tracking of scientific data and algorithms, from raw data and first principle equations to final data products and visualisations, modular simulation models and benchmark evaluation algorithms. Led jointly by EPFL and ETH Zurich, the SDSC is composed of a distributed multi-disciplinary team of data scientists and experts in select domains. The center aims to federate data providers, data and computer scientists, and subject-matter experts around a cutting-edge analytics platform offering user-friendly tooling and services to help with the adoption of Open Science, fostering research productivity and excellence.

In this presentation, we will discuss our vision of a high-scalable open but secure community-based platform for sharing, accessing, exploring, and analyzing scientific data in easily reproducible workflows, augmented by automated provenance and impact tracking, knowledge graphs, fine-grained access right and digital right management, and a variety of domain-specific software tools. For maximum interoperability, transparency and ease of use, we plan to utilize notebook interfaces wherever possible, such as Apache Zeppelin and Jupyter. Feedback and suggestions from the audience will be gratefully considered.