Andreas Geyer-Schulz (Karlsruhe Institute of Technology)

Data at the Core: Archives of Data Science, Series B: Organizing a Journal Around Data Sets.

Topic 1 – Bringing in information from where we can get it

Keywords: information services, e-publishing, scientific infrastructure

Introduction

Scientific journals are part of the scientific infrastructure. In the field of data science they should support generation, access, storing, distribution, analyzing, visualizing, and using data from advanced information technologies. Big data poses new challenges to scientists, for example the evaluation of machine learning results or the integration of data collecting services into a globally distributed infrastructure like the Internet. In addition, for many data analysis tasks, the distribution of labor between scientists becomes an additional challenge, especially, if the necessary collaboration spans several scientific fields.

The Archives of Data Science bundle of academic journals has the purpose to provide a low cost/high quality innovative publishing outlet to the German Classification Society (GfKl e.V.) and related data science and statistics societies which meets the goals of many of its members, namely timely journal publication and wide and international availability. The journal Archives of Data Science, Series B (Data Sets, Algorithms, Processes, and Services) addresses the problems of big data. In this contribution we concentrate on the concept and services of this journal which is organized around data sets.

The journal covers scientific articles which improve methods, algorithms, and processes over the whole data life cycle. In addition, papers on data analysis processes, services, and (scientific) infrastructures are welcome.

Methods / Problem statement

Concepts The Archives of Data Science, Series B is organized around data sets. The journal follows the traditional structure of volumes and numbers, however with a non-traditional interpretation of what constitutes a number and a volume.

A number is an open ended stream of articles which starts with a seminal article on a data set (head article) and continues with articles which propose innovative ways of 'handling' the data set (tail articles). A head article describes a data set and provides access to it. All data sets must be open. Such a seminal article describes at least the structure of the data set and the interfaces available to access the data set. We also define data sets in a wide sense: e.g. comma-separated files, relational data-bases, open linked data, data-harvesting processes, data generators, and interfaces to data streams. In addition, the measurement process for the data set must be explained in detail, restrictions and possible problems of the measurement process should be covered.

The authors of such a seminal article are also expected to provide problems and questions (ideally a challenge) that they would like to see solved by analyzing the data set. Tail articles describe innovative ways of generating, accessing, storing, distributing, analyzing, visualizing, and using data in the broadest sense. Each article must be complemented by a well-documented open source version of a software package which implements the metho

Results / Proposed solution

Services In this section we discuss two essential services of Archives of Data Science, Series B:

1. Data Store. Data sets will be linked with articles via DOIs. In principle, small to medium size data sets can be stored in the permanent storage area of Archives of Data Science, Series B. For big data sets, currently external datastores are needed.

2. Dynamic Binding of Volumes: Volumes provide two orthogonal dimensions of combining articles: The first dimension provides for a given challenge a set of comparable solutions by different methods and algorithms. The second dimension shows how different challenges can be addressed by variations of the same data analysis method. Over time, collection volumes can be produced by recombining articles to address a variety of scientific goals.

Conclusions

Conclusions The organization of Archives of Data Science, Series B addresses the following problems:

- The incentive problem for scientists who are owners of data sets. Publication of a data set becomes publishable and the data set can be quoted. If the data set is interesting enough, this article will often be quoted.

- Quality management issues: Since algorithms and methods work with the same open datasets, results become comparable and scientists can replicate data analytic studies. Open source allows source code level analysis of programs and algorithms.

- Division of labor: Division of labor between scientific groups is improved, because results of different methods become comparable. Since tail articles can be dedicated to solving small problems of a data set (e.g. missing data, outliers, improved data structures, data compression, ...), scientists can build incrementally on the work of others.

- Two feedback cycles: One feedback cycle aims at improving methods and algorithms, the second aims at improving data sets and measurement processes. (Invited to session: Access to statistical data and research by Karl Mosler.)