Open Innovation in Data Science through Evaluation-as-a-Service

Allan Hanbury and Henning Müller

Contributor names and short CVs

Dr. Allan Hanbury is Senior Researcher at the Vienna University of Technology. He has an M.Sc. in physics from the University of Cape Town, South Africa, a PhD in applied mathematics from Mines ParisTech, France, and the habilitation in practical informatics from the Vienna University of Technology, Austria. He was scientific coordinator of the EU-funded Khresmoi IP on medical information search and analysis, and is coordinator of the VISCERAL support action on benchmarking in big medical data and coordinator of the KConnect H2020 Innovation Action on setting up a data-value chain for medical text processing and analysis. He was leader of the Evaluation, Integration and Standards work package of the MUSCLE EU Network of Excellence, and leads a number of Austrian national projects. His research interests include data science, multimodal information retrieval, and the evaluation of information retrieval systems. He is author or co-author of over 100 publications in refereed journals and international conferences.

Prof. Dr. Henning Müller studied medical informatics at the University of Heidelberg, Germany, then worked at Daimler-Benz research in Portland, OR, USA. From 1998-2002 he worked on his PhD degree at the University of Geneva, Switzerland with a research stay at Monash University, Melbourne, Australia. Since 2002 Henning has been working for medical informatics at the University hospital of Geneva. Since 2007 he has been a full professor at the HES-SO Valais and since 2011 he is responsible for the eHealth unit of the school. Since 2014 he is also professor at the medical faculty of the University of Geneva. Henning is coordinator of the Khresmoi project, scientific coordinator of the VISCERAL project and initiator of the ImageCLEF benchmark. He has authored over 400 scientific papers and is in the editorial board of several journals.

Type of the presentation proposed

Impact contribution

Title of the presentation

Open Innovation in Data Science through Evaluation-as-a-Service

Summary of the presentation (<100 words)

Companies and organisations are increasingly adopting an Open Innovation approach to solving their data analytics problems – they make data and problem specifications publicly available, and offer prizes to the top solution submitted. The almost ubiquitous approach to running such public evaluation involves distributing the data to the participants for processing on their own computers. Recently, it is often found that this data distribution approach is not practical, as the data may be: huge, private or real-time. This talk will present a roadmap toward Evaluation-as-a-Service, a paradigm allowing Open Innovation in Data Science while avoiding the need to distribute data.

Extended abstract of the presentation

Many companies and organisations realise that the Data Scientists that potentially have the best solutions to the data analytics problems that they are facing do not necessarily work for them. For this reason, companies and organisations are adopting an Open Innovation approach to solving their data analytics problems – they make data and problem specifications publicly available, and often offer prizes to the top solution submitted. Examples of platforms on which such Data Science competitions are made available are Kaggle[1] and Dream Challenges.[2]

There is a long tradition within the academic community of organising public challenges, benchmarks, competition, and evaluation campaigns, in particular in the areas of data mining, machine learning, information retrieval and computer vision such as PASCAL, TREC[3] (Text Retrieval Conference) and CLEF[4] (Cross-Language Evaluation Forum). The techniques that have been developed over time to run these public evaluations are highly applicable to Open Innovation in Data Science.

The almost ubiquitous approach to running such public evaluation involves distributing the data to the participants so that they perform the evaluation locally on their own computing infrastructure. In particular in recent years, it is often found that this approach of distributing data is not practical, as the data may be:

- **Huge** - in order to obtain realistic evaluation results, the evaluation should be done on realistic amounts of data. For example, in the case of web search, this could be Petabytes of data. The current common approach of sending this data on hard disks through the postal service has major disadvantages.
- **Non-distributable** - In many cases, it is not permitted to distribute data due to privacy, terms of service or commercial sensitivity of the data. Privacy is the major concern for medical patient records. Even though one would be permitted by law to distribute anonymized medical records, automated anonymization procedures needed for large scale anonymisation are usually not trusted sufficiently by the owners of the data. The Twitter terms of service forbid redistribution of Tweets, while query logs are not made available for researchers after the AOL debacle in 2006. Distribution of company documents for the evaluation of enterprise search or analytics would not be permitted due to the commercial sensitivity of the data.
- **Real-time** - Companies working on real-time systems, such as recommender systems, are often not interested in evaluation results obtained on static historical data, in particular if these data have to be anonymised to allow distribution, as these results are too far removed from their operative requirements.

Some data fall into two or all three of these categories.

There are currently a number of initiatives to allow evaluation to be done without the necessity of distributing the data, in order to solve the challenges described above. The initiatives all basically use the idea of *Evaluation-as-a-Service* (EaaS). The data are not distributed, but are stored on a central infrastructure. Access to the data is made possible either by making available APIs to access the data in a controlled way, or by providing Virtual Machines (VMs) that can access the data on which solutions should be implemented. By choosing to centralise the data, the challenges of distributing it

[1] https://www.kaggle.com/competitions
[2] http://dreamchallenges.org
[3] http://trec.nist.gov
[4] http://www.clef-initiative.eu

are resolved. Strong privacy controls can also be implemented, such as only allowing the VMs to access confidential data during execution of the installed programs, once the VMs have been sandboxed to cut off all access. Also, the data can be updated as regularly as needed to allow programs to be run on actual data. With algorithms installed in virtual machines, the existing systems can always be reused as a strong baseline on new data.

On the 5-6 March 2015, twelve representatives of initiatives that organise evaluations or competitions using the Evaluation-as-a-Service paradigm met for a workshop in Sierre, Switzerland. Initiatives from Europe, the USA and Japan, were represented, providing a wide spectrum of approaches that have been adopted for EaaS. The initiatives included: TREC Microblog,[5] BioASQ,[6] Living Labs for Information Retrieval Evaluation,[7] PAN (uncovering plagiarism, authorship, and social software misuse),[8] and a National Cancer Institute (NCI) project on Informatics Tools for Optimized Imaging Biomarkers for Cancer Research & Discovery.[9] The workshop provided a forum in which to exchange information about the approaches adopted to EaaS and the experience gained. The result of the workshop is a roadmap on the development of EaaS presented in a White Paper, which will be available before the EDF takes place. The workshop was funded by the European Science Foundation (ESF) and the VISCERAL[10] FP7 project.

The presentation at the EDF will present the EaaS roadmap and experiences, as well as the challenges that have been identified. The roadmap will cover:

- Technical aspects, such as EaaS system variants, system usability and reproducibility of results;
- Emotional aspects, including fears and incentives for EaaS organisers, participants, data and task providers, infrastructure providers, and funding agencies; and
- Regulatory aspects, such as legal considerations and sustainability of EaaS infrastructure.

The presentation will also encourage further members to join the EaaS community that developed as an outcome of the workshop (http://eaas.cc/). Further development of this community will allow effective implementation of EaaS approaches and technologies, and hence further encourage companies to adopt Open Innovation approaches to solve their Data Science challenges.

[5] https://github.com/lintool/twitter-tools/wiki/TREC-2014-Track-Guidelines

[6] http://www.bioasq.org/

[7] http://living-labs.net/

[8] http://pan.webis.de/

[9] http://grantome.com/grant/NIH/U24-CA180927-01A1

[10] http://visceral.eu