Published in proceedings of the ACM Special Interest Group on Information Retrievel (SIGIR) Forum 2016, Pisa, Italy, 17-21 July 2016, which should be cited to refer to this work.

Report on the Cloud-Based Evaluation Approaches Workshop 2015

Henning Müller, Jayashree Kalpathy–Cramer, Allan Hanbury,
Keyvan Farahani, Rinat Sergeev, Jin H. Paik, Arno Klein, Antonio Criminisi,
Andrew Trister, Thea Norman, David Kennedy, Ganapati Srinivasa,
Artem Mamonov, Nina Preuss
henning.mueller@hevs.ch

Abstract

Data analysis requires new approaches in many domains for evaluating tools and techniques, particularly when the data sets grow large and more complex. Evaluation–as–a–service (EaaS) was coined as a term to represent evaluation approaches based on APIs, virtual machines or source code submission, different from the common paradigm of evaluating techniques on a distributed test collection, tasks and submitted results files. Such new approaches become necessary when data sets become extremely *large*, contain *confidential* information or might *change* quickly over time. The workshop on cloud–based evaluation (CBE) took place in Boston, MA, USA on November 5, 2015 and explored several approaches for data analysis evaluation and frameworks in this field. The objective was to include several stakeholders from academic partners, companies to funding agencies to cover various interests and viewpoints in the discussion of evaluation infrastructures. The workshop focused on the biomedical domain but the results are easily applicable to many domains of information analysis and retrieval.

1 Introduction

Scientific challenges such as TREC[1] (Text REtrieval Conference) and CLEF[2] (Conference and Labs of the Evaluation Forum) have helped many researchers to evaluate their tools and algorithms and compare them to strong baselines in a common standard setting. Based largely on the Cranfield paradigm of distributing a test collection, query topics and then ground truth for the evaluation, this type of evaluation has been run for many years despite criticism that interactive tools are often neglected. Interactive retrieval evaluation was attempted several times but often with low participation. With extremely large data sets the distribution of data becomes a problem — sending hard disks by post has its disadvantages. Costs for locally storing and analyzing extremely large data sets that are now

[1] http://trec.nist.gov/

[2] http://www.clef-campaign.org/

available are high, excluding researchers from lower resource countries and giving labs with larger computer infrastructures a strong advantage. More powerful servers allow the testing of an extended set of parameters or allow more training with techniques such as deep learning. Also quickly changing data sets or confidential data, for example patient data in the biomedical domain, are problematic. Data can only be released with privacy protection in place and also this leaves a risk of data abuse, as data sets can be linked with other available data resources and thus may allow a re–identification of subjects.

At a workshop in March 2015 in Switzerland, several new ideas and methods to overcome the limitations of current paradigms for evaluation were discussed [6]. These approaches were put under the common term Evaluation–as–a–Service (EaaS), meaning that there is a communication via services and executable software and not only on the basis of data. A white paper on EaaS was also made available in late 2015 [4]. One guiding principle is the idea to move algorithms to the data and not the data to the algorithms [5], with running software actually becoming easier to move than the data. Virtual machines or Docker[3] containers make moving executable software much easier than having to compile and configure complex source code that depends on many libraries and tools.

2 Approaches presented

The CBE workshop started with an overview of the approaches developed by the participants to address the problem of evaluating huge, private or real–time data. The program of the CBE workshop and the minutes including many of the discussions are available[4].

2.1 VISCERAL and EaaS

A first presentation described the general EaaS principles described in the EaaS white paper and the VISCERAL (Visual Concept Extraction Challenge in Radiology) project [7]. In VISCERAL, the evaluation of medical image analysis and retrieval algorithms was entirely organized in a cloud environment. A small training data set and a virtual machine (VM) were made available to participants, who installed and optimized their algorithms and then the organizers took over the VMs for running the algorithms on the test data. Figure 1 shows this principle. The experience shows that using VMs created an overhead for participants in the challenge and can reduce participation but it solves the majority of the technical problems. Unfortunately, different VMs (Amazon vs. Azure) are not compatible and thus a VM is not fully mobile, meaning it has to be reinstalled when used on a different cloud. Docker can be a solution for this. On the other hand operating system flexibility is high in virtual machines where also very old versions and libraries can be supported. Keeping data and software together means that the approaches are fully reproducible and code can be rerun when new data become available.

2.2 NCI and the Coding 4 Cancer lung challenge

Similar to other government institutions the ideas of open science and crowdsourcing have also started to be used by the National Cancer Institute (NCI). Challenges have become an

[3] http://www.docker.com/

[4] http://www.martinos.org/cloudWorkshop/

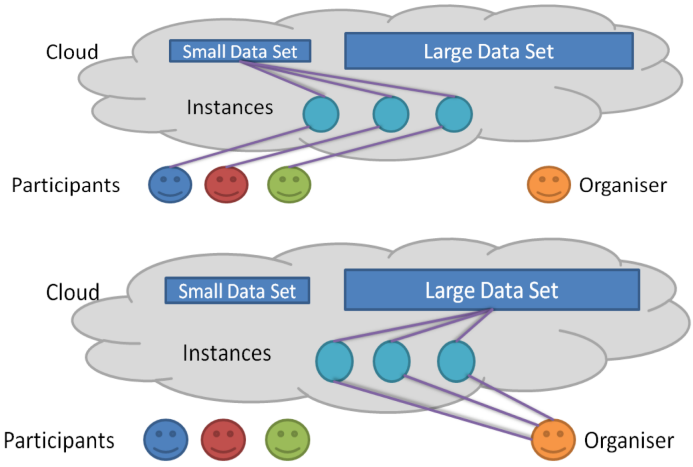

Figure 1: In VISCERAL, the participants of a challenge obtain their own computing instance in the cloud (a VM), linked to a small training dataset of the same structure as the large test set. Software for carrying out the competition is placed in the instances by the participants and the challenge organizer then executes the algorithms on the test data.

important aspect to work on open data and open science. The Coding4Cancer[5] challenges are planned to be run entirely inside a closed environment where the participants need to submit code and cannot view or access the test data, only getting access to a small training data set. This avoids cheating, which may be more of a problem when prize money is offered. Submitting algorithms for independent assessment and restricting data access also allows a potential reuse of the code on new data and avoids manual optimizations on the test data. The protection of the patient data is also essential in this case as only the training data need to be anonymized manually and are visible to the participants. Matching medical data from several sources creates a particular risk for patient identification and this can be limited if the data are only analyzed in a closed environment.

2.3 Microsoft Azure as a scientific platform

Microsoft Research Inc. highlighted its interest in scientific challenges, for example in machine learning and medical imaging. Concrete applications that resulted from research in these domains inside Microsoft Research were also presented. One important project is the CodaLab[6] platform for organizing data science challenges. The open source platform is strongly supported by Microsoft and it has so far mainly been used in artificial intelligence challenges and for medical imaging. A community has been built around the platform to ensure its sustainability. It allows participant registrations, result submission, a leaderboard and is open for extensions. It is integrated with the Microsoft Azure cloud computing environment, which allows for an efficient integration of code and data.

[5] http://www.coding4cancer.org/
[6] http://www.codalab.org/

2.4 Sage Bionetworks challenges

Since 2011, Sage Bionetworks, in partnership with the open science DREAM community (Dream Challenges[7]) has been organizing 6–8 crowdsourcing algorithmic challenges a year focused on solving problems in systems biology and translational medicine. DREAM Challenges run on Sages Synapse[8] platform : the Synapse infrastructure has evolved over time and moved from requiring challenge participants to submit their vector of prediction and source code file towards scoring challenges based on the submission of Docker containerized re–runnable models. Sages Docker registry will be used to support the upcoming Digital Mammography DREAM Challenge[9] , the first Coding4Cancer Challenge that will run, hosting 100TB of data in the cloud, and focus on reducing the recall rate in mammogram screening. The Docker approach is well–suited for Sages open challenges that each tend to attract 300–700 participants from diverse backgrounds: Docker Engine supports Linux, Windows and OSX as well as having many machine instances on clouds (i.e., AWS, Azure, IBM-SoftLAYER). Notably, Sage is also archiving these Docker containerized Challenge results as an open research resource that can be used as the starting point for others model development.

2.5 Intel Collaborative Cancer Cloud and beyond

Also Intel Corporation, as a hardware manufacturer, sees the medical information analysis domain as one of the main future drivers for computing demand and has worked with clinical and academic partners to develop the Collaborative Cancer Cloud (CCC[10]). The main idea is that medical data can reside inside the institutions where they are produced and they can be analyzed and treated there with only aggregated results being transferred between institutions in a safe way for multi–center studies. Code can be distributed via Docker containers between the compute nodes and secure computation assures that no confidential data can be revealed. A central node can aggregate data from the institutions and can organize the code distribution.

2.6 NITRC – Neuroimaging Informatics Tools and Resources Clearinghouse

NITRC[11] follows the general idea that code and data should belong together and should be accessible and shared in a community. Since 2006, NITRC has made data sets and tools available for download or in cloud environments (i.e. in VMs) and has standardized interfaces and formats to make an exchange easier or at least possible. Currently centered around the neuroimaging community, the tools have helped to improve the reproducibility of research results and have fostered reuse of tools and services. This can be used as an example for what is also possible in other scientific domains and research communities if sharing is more widespread.

[7] http://www.dreamchallenges.org/

[8] http://www.synapse.org/

[9] https://www.synapse.org/#!Synapse:syn4224222/wiki/231837

[10] http://www.intel.com/content/www/us/en/healthcare-it/collaborations/ohsu-intel-collaboration-video.html

[11] https://www.nitrc.org/

3 Motivations & incentives for challenges

As scientific challenges in data science have become a much larger field than 10–15 years ago, much research has been conducted on motivations and incentives for such challenges [2]. With public administrations and public agencies in the US using crowdsourcing challenges[12] as a way to use collective, distributed knowledge, understanding incentives and motivations becomes important to maximize crowdsourcing's potential. Depending on the groups of individuals working on crowdsourced data science challenges the motivations can strongly differ. Some individuals are clearly attracted by prize money and thus try to win challenges whereas for other groups, fame is more important. In terms of scientific careers, publications can be an important factor. People might also have an interest in symbolic prices (free t–shirts or similar), as the feeling to be part of a community is a very important factor. Prize money should also be in relation to the gains — it is not always better to have higher prize money as non–experts might be deterred by a large prize money and would not participate, feeling that they would have no chance of winning.

Challenges also need to take into account a reasonable timeline, leaving the participants the time to get familiar with the data and give feedback on the initial setup, data format, etc. This can lead to a challenge being composed of several parts from a trial run, then a full run with limited prize money and a final run with higher prize money. The objective is often to find an optimal solution and very often the best results are not obtained by people working directly in the field but rather by groups working in adjacent fields [1]. This fact may be linked to these groups thinking out of the box and not only following traditional paths for a solution. Using staged challenges is another possibility to create collaborations of groups with different skills. For example, in medical image analysis a first challenge can be on noise reduction in images, a second on feature extraction and a third on fusion and machine learning. Groups can participate in a specific stage or several and combined solutions can then credit people who participated in one or more stages of the best solutions. Having interim winners combine forces for a better solution can also create a very stimulating environment.

4 Sustainability of research infrastructures

A major challenge of such common research infrastructures is that the sustainability is not trivial [3]. Many projects have been funded for a limited amount of time, but beyond the project the infrastructure and data can only be sustained based on personal initiatives or with the help of temporary small amounts of funding, for example in collaboration with infrastructure providers. This is particularly the case when not only storage is needed for data download but a computing infrastructure. Both data and software need to be available long–term so that they do not depend directly on short–term funding cycles, as is currently the practice in most science disciplines. If data and code are available, then the entire system remains reproducible and can be a baseline that is valid for several years and can help produce a more significant and lasting impact.

If a majority of the scientific data sets used in publications could be made available for sharing in such a way there also need to be incentives for researchers to clean the data and make a clear evaluation scenario available together with the data to foster reuse. This can be

[12] `http://www.challenge.gov/`

via data publications, for example. *Nature Scientific Data* shows that the major publishers have realized how important data sets are in many scientific fields. It can be expected that good data sources can become highly cited.

As an outcome of the workshop it was proposed that very likely a public–private partnership is needed to sustain such a system and that researchers and also funding bodies need to work with infrastructure providers to define needs, costs and possibilities to maximize the outcomes against expended efforts. This requires particularly long–term support and has to be across national or regional borders. In the scientific domain, international collaborations are the norm and much research is not linked to only a national or regional funding body — in this case globally shared infrastructures are also useful.

5 Visibility and promotion

One important aspect discussed at the workshop was around making the various concepts and initiatives better known to the stakeholders from researchers to funding organizations and companies. This is important as only a combined effort can help to make Evaluation–as–a–Service a success and a large participation is necessary to see the main benefits. Isolated efforts can help to show the advantages in small–scale studies. If automatically funded research projects would have free storage for data over ten years or more and if infrastructure providers would give easy and unbureaucratic access to computing linked to the storage this could make many data science areas much more reproducible. It can allow running developed techniques on various data sets to objectively see advantages and disadvantages.

6 Conclusions

The workshop on cloud–based evaluation brought together over 15 researchers with a variety of roles and backgrounds largely in the biomedical domain. The topic of reproducible science and related infrastructures was discussed deeply, as this is often (and rightfully) criticized in the current big data research area. The initial topic of cloud–based evaluation was a starting point but the discussion went far beyond this topic. Besides reproducibility of data science and comparison of new approaches with strong baselines, it is seen as important to avoid producing many incompatible small data sets and tools that cannot be reused. It is also better to have efforts to reuse software wherever possible and share data and work on extensions of existing sets rather than creating something totally different. Cloud–based approaches can help to effectively provide sustainable, scalable, extensible, and interoperable environments for evaluations and benchmarking through open science.

References

[1] Lars Bo Jeppesen and Karim R. Lakhani. Marginality and problem-solving effectiveness in broadcast search. *Organization Science*, 21(5):1016–1033, 2010.

[2] Kevin J. Boudreau, Nicola Lacetera, and Karim R. Lakhani. Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science*, 25(5):843–863, 2011.

[3] Philip E Bourne, Jon R Lorsch, and Eric D. Green. Sustaining the big–data ecosystem. *Nature*, 527, 2015.

[4] Allan Hanbury, Henning Müller, Krisztian Balog, Torben Brodt, Gordon V. Cormack, Ivan Eggel, Tim Gollub, Frank Hopfgartner, Jayashree Kalpathy-Cramer, Noriko Kando, Anastasia Krithara, Jimmy Lin, Simon Mercer, and Martin Potthast. Evaluation–as–a–service: Overview and outlook. *ArXiv*, 1512.07454, 2015.

[5] Allan Hanbury, Henning Müller, Georg Langs, Marc André Weber, Bjoern H. Menze, and Tomas Salas Fernandez. Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis. In *CLEF conference*, Springer Lecture Notes in Computer Science, 2012.

[6] Frank Hopfgartner, Allan Hanbury, Henning Müller, Noriko Kando, Simon Mercer, Jayashree Kalpathy-Cramer, Martin Potthast, Tim Gollub, Anastasia Krithara, Jimmy Lin, Krisztian Balog, and Ivan Eggel. Report on the evaluation–as–a–service (eaas) expert workshop. *ACM SIGIR Forum*, 49(1):57–65, 2015.

[7] Georg Langs, Henning Müller, Bjoern H. Menze, and Allan Hanbury. Visceral: Towards large data in medical imaging — challenges and directions. *Lecture Notes in Computer Science*, 7723:92–98, 2013.