Organizing crowd-sourced AI challenges in enterprise environments: opportunities and challenges

Mahtab Mirmomeni*, Isabell Kiral*, Subhrajit Roy*, Todd Mummert, Alan Braz, Jason Tsay, Jianbin Tang, Umar Asif, Thomas Schaffter, Eren Mehmet, Bruno De Assis Marques, Stefan Maetschke, Rania Khalaf†, Michal Rosen-Zvi†, John Cohn†, Gustavo Stolovitzky†, Stefan Harrer†

* These authors contributed equally to this work

† Corresponding authors: rkhalaf@us.ibm.com, rosen@il.ibm.com, johncohn@us.ibm.com, gustavo@us.ibm.com, sharrer@au.ibm.com

T. Schaffter is with Sage Bionetworks, USA. E. Mehmet is with the University of Illinois at Urbana-Champaign, USA. All other authors are with IBM Research in USA, Israel and Australia.

The need for crowd-sourced AI challenges to solve data science problems

Large-scale investments by enterprises operating in sectors such as for example pharma, healthcare, retail and finance often result in the generation of proprietary, unstructured datasets with substantial information content that can be extracted using AI technology and subsequently allows to make more intelligent, evidence based strategic decisions. Thus, as the data owners, enterprises have a strong interest in accessing such information. However, the abundance of data is often not matched by an equally strong supply of data science resources capable of developing and applying AI to drive insights from the data. Crowd-sourcing the analysis of proprietary data can solve this resourcing problem and at the same time accelerates the speed and innovation of AI solutions. For enterprises to be able to leverage AI crowd-sourcing, they need to find a way to allow external data scientists to build analysis software for their data, while at the same time not externalizing the data itself which often is rendered highly sensitive through its personal or economic value. Conventional 'Kaggle-style' AI crowd-sourcing ecosystems do not offer that feature but make the challenge data directly available to the solver community. Hence, such platforms are not suitable for hosting enterprise AI challenges. In an effort to circumvent the need to publicly share their data and still be able to use conventional crowd-sourcing platforms, some enterprises have resorted to using redacted data for enabling external crowd-sourced challenges [1] which generally compromises the quality of the model solutions. In other scenarios companies may use conventional crowd-sourcing platforms internally [2] but in these cases, they exclusively rely on internal data scientist resources which limits size and efficiency of the solver community substantially.

In order to overcome these limitations and truly harness the wisdom of the crowd, a novel type of challenge platform is needed that allows to build, test, evaluate and validate AI models on proprietary data while at the same time avoiding the need to grant the solver community access to the data itself. Such a platform can then be used to collaboratively employ data scientists from inside or outside an enterprise (or a combination thereof) to work together whilst keeping the enterprise data secure and protected. To ensure enterprise data security, we propose a model-to-data approach in which the challenge data is never directly accessed by the participants who will instead create models based on a small sample data provided by the enterprise. Participants then submit their sample models to a repository of models, residing within the enterprise. There, and shielded from participants, their submitted models will be evaluated on the actual data. Model performances will be determined based on an evaluation metric and the results of such evaluation runs will be handed back to the respective participants. Following this scheme, the enterprise challenge platform will keep the data shielded behind a firewall at all times while facilitating model ingestion into the firewalled model evaluator and extraction of model performances out of it.

An example: The Deep Learning Epilepsy Detection Challenge

Recently, we have introduced a first prototype of such a crowd-sourced AI challenge platform for enterprises using exclusively IBM technology. We successfully tested the platform by organising the IBM-internal Deep Learning Epilepsy Detection Challenge in collaboration with Temple University Hospital. While we describe the technical details of this IBM AI challenge platform in an accompanying paper [3], we provide a first-hand account of handling enterprise-specific features of organising and running a crowd-sourced AI challenge in the following section below.

The how-to of running crowdsourced AI competitions for enterprises

A comprehensive general description of the how-to of running crowd-sourced AI challenges is given by Stolovitzky et. Al. in Figure 2 of [4]. In the following section, we map the specific requirements and tasks related to running an enterprise crowdsourced challenge onto this flow chart.

a. **Define scientific use case and assemble challenge organising team:** The scientific use case needs to be designed in close collaboration with the enterprise, i.e., data owner, to ensure the challenge caters to the business need of the enterprise. The client needs to be represented on the challenge organizing team from the beginning onwards to guide the definition of the scientific use case in the initial stages of challenge planning.

b. **Solver community:** The solver community needs to be defined by the enterprise. The enterprise can decide to only allow its own data scientists to participate, to open the challenge to trusted 3rd parties, or to the public. The enterprise may also decide to engage with the platform owner and infrastructure provider as trusted party for recruiting data scientist to work on the challenge. As part of setting the challenge rules, the enterprise client should also determine which party owns models and code submitted during the challenge.

c. **Data access:** Data usage agreements for high-value proprietary datasets are generally complex legal arrangements and not trivial to put in place. Deep understanding of the business need of the enterprise is required to put these agreements together. Data integrity and safety are paramount features of the crowd-sourced AI enterprise challenge platform.

d. **Challenge platform:** The enterprise challenge platform needs to enable participants to participate in the challenge, while ensuring the challenge data to be kept secure and untouched by participants. Our platform supports a model-to-data approach in which the data of the enterprise is never directly accessed by the participants. The participants only get access to a small snapshot of the data allowing them to understand data format and other basic features necessary to build models for analysing it. They then containerize their models and submit them to the platform which will run their model against the firewalled challenge data. This approach is similar to the approach used in the DREAM mammography challenge which we co-organised [5]. Figure 1 below illustrates the data flow in our enterprise challenge ecosystem. A hybrid cloud framework is key to enabling data safekeeping. All data analytics processes, including model creation and submission as well as compute resource management and provisioning are facilitated by IBM Watson Studio and IBM Watson Machine Learning services.

e. **Incentives:** The enterprise client defines the incentives of the challenge, which could range from being paying participants, employing participants or offering them co-ownership of the IP and challenge results. More conventional participation incentives such as publication and presentation of outcomes in scientific journals and at technical conferences may also be relevant in enterprise challenge scenarios.

f. **Advertising:** Challenge advertisement is key to form the best possible solver community which in turn is part of the client agreement. Success of the challenge depends on the solver community which the enterprise has specified. Different scenarios for running challenges are the client engaging directly with the solver community or the client engaging with the platform provider who then engages the solver community.

g. **Analysis of the results:** The challenge organizing team needs to work closely with the enterprise to ensure that challenge results are interpreted and presented in a way that caters to the client's business need, i.e. answers the scientific use case which defined the conception of the challenge in the first place.

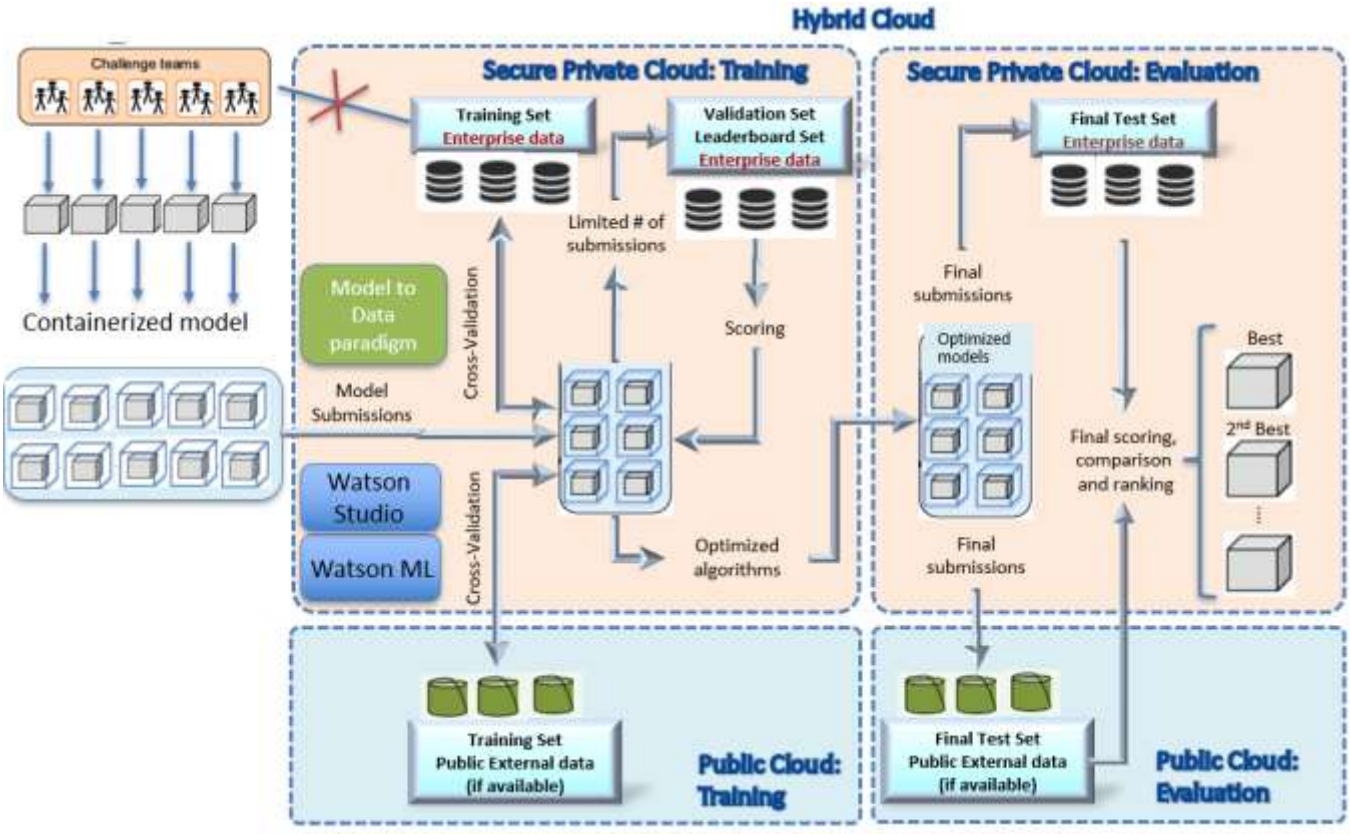

Figure 1: *Architecture of the IBM-built crowd-sourced AI challenge platform*

References

[1] https://www.australianmining.com.au/news/oz-minerals-unearthed-award-1m-prize-for-exploration-contest/ (2019).

[2] https://www.forbes.com/sites/ryanholmes/2018/09/28/why-the-new-open-data-initiative-by-microsoft-adobe-and-sap-could-revolutionize-customer-experience/#14b4095952e4 (2018).

[3] Kiral, I., Roy, S., Mummert, T., Braz, A. et. Al., 2019. The Deep Learning Epilepsy Detection Challenge: design, implementation and test of a new, crowd-sourced AI challenge ecosystem. submitted to *NeurIPS CiML 2019.*

[4] Suez-Rodriguez, S., Stolovitzky, G. et Al., 2016. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat. Rev. Genetics* 17, pp. 470-486.

[5] https://www.synapse.org/#!Synapse:syn4224222/wiki/401743 (2019).

Acknowledgements

The authors would like to thank Joseph Picone and Iyad Obeid from Temple University, USA for providing EEG data to IBM as part of the Deep Learning Epilepsy Detection Challenge.