From data challenges to collaborative gig science. Coopetitive research process and platform

Andrey Ustyuzhanin
National Research University Higher School of Economics
austyuzhanin@hse.ru

Mikhail Belous
National Research University Higher School of Economics

Leyla Khatbullina
National Research University Higher School of Economics

Giles Strong
LIP – Laboratório de Instrumentação e Física Experimental de Partículas

Abstract

Extended CIML workshop abstract on collaborative research process and platform.

1 Extended abstract

Data science-rich approaches have demonstrated the possibility to foster significant breakthroughs in a variety of scientific domains [1,2,3]. Most of those results have been produced as a collaboration project between domain experts and data scientists. Although some domain-specific researches can learn data science tools and techniques to quite advanced level, it is not foreseen it will happen widely, so the gap between bleeding edge data science tools and data science expertise in specific domains will remain to exist. So there is a demand for the guidelines/process or even platform that would help collaboration between experts in different fields going.

We propose a novel 'co-research' process and 'coopetition' platform to support it for collaborative experience on real scientific challenges, which poses a design challenge in itself. The process unites different roles: problem owners, participants, mediators in a similar way to the agile process for software engineering [8].

The real research challenges also require flexibility from the platform to adapt to variability in metrics and data analysis protocols.

From the perspective of data scientist that collaborate on the specific challenge, it should guide their education/skill growth. We design the platform to accommodate for different technical skills of the participants. The mainstream of participants comes from educational tracks, so the collaborative challenges have to be embedded into training curricula. All submissions are made via a version control system, so it is possible to decompose each solution into thematic building blocks like data imputation, rescaling, binary classification, CNN's etc. Quality of the solution concerning the given figures of merit can suggest which training material would be profitable to study for the participants to improve his solution.

Mentors are a special kind of role in the system, that can are akin to scrum masters in software agile design process. They chair regular meetings and help to resolve problems for stuck participants.

Preprint. Under review.

Thanks to the focus on process the quantitative estimation for each solution becomes a background process without much attention and attachment.

The participants accept the consent that data they submit can be used for analysis and mining by the algorithms and other members. So this initiative may gain the potential to collect codebase for study problem of automatic code generation to solve similar problems in the future given data and the definition of the problem.

Technically the 'coopetition' platform is built on top of the GitHub and codelab [4], comet.ml [5], Trello [6], discourse [7] taking the best of those platforms that help to democratise data science for the breadth of scientific domains. The platform is still at its infancy [9]; nevertheless, we'll give examples of challenges we've organised so far using the platform and possibilities of the further extension.

References

[1] Radovic, A., Williams, M., Rousseau, D., Kagan, M., Bonacorsi, D., Himmel, A., ... Wongjirad, T. (2018). Machine learning at the energy and intensity frontiers of particle physics. Nature, 560(7716), 41.

[2] Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M. Muller, K. R. (2013). Assessment and validation of machine learning methods for predicting molecular atomization energies. Journal of Chemical Theory and Computation, 9(8), 3404-3419.

[3] Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T. F. G., Qin, C., Petersen, S. et al (2018). De novo structure prediction with deeplearning based scoring. Annu Rev Biochem, 77, 363-382.

[4] https://github.com/codalab/codalab-competitions

[5] https://comet.ml

[6] https://trello.com

[7] https://www.discourse.org/

[8] Schwaber, K., Beedle, M. (2002). Agile software development with Scrum (Vol. 1). Upper Saddle River: Prentice Hall.

[9] https://coopetition.coresearch.club