A galaxy of data challenges

By organizing Kaggle competitions, astrophysicist Thomas Kitching can focus on asking the right questions.

In science we always have more questions than answers. Finding the right tools to tackle these questions is an essential part of research. This challenge has recently taken on a new shape as many fields transition from data-scarce to data-rich paradigms and individual researchers realize they do not possess the tools to analyse large heterogeneous datasets. One strategy is to pursue interdisciplinary research and collaborations. Another is to seek the wisdom of the crowd.

This is the approach that we took by engaging with Kaggle, the data science platform. Kaggle runs competitions in which organizers upload a dataset and pose a problem or question based on that data. Organizers define a metric for scoring submissions, and provide training data for competitors to play around with and train their models. A prize is also offered, which can be anything from cash to a job interview.

From the organizer's perspective, Kaggle is a source for myriads of data scientists worldwide who can try to solve your data science problem. The key is to ask the right question, and provide the right prize. From the competitor's perspective, you can get access to new datasets, interact with other data scientists, and potentially get rich or land a new job.

We have run three Kaggle competitions since 2010 that have posed the problems of measuring galaxy shapes[1], determining the position of dark matter in noisy data[2] and classifying galaxy types[3], and there is currently a competition to help classify supernovae observations[4]. Running public competitions to spur scientific advances is not a new idea, and has a long heritage dating back to at least the Longitude Prize — a challenge set in 1713 to accurately determine the longitudinal position of a ship, with a maximum prize of up to £20,000 (equivalent to over £2 million in 2018). Fast-forward to the twenty-first century and such competitions thrive by the Internet, through which a nearly unlimited number of potential competitors can be instantly reached. Furthermore, Kaggle competitions are software rather than hardware based,

Credit: ESA/Hubble and NASA

and so solutions can be tested, improved and combined, and evolve at a rapid rate, with no specialist equipment required.

It can be difficult for organizers to offer good prizes. However, we identified a new mode of impact by partnering with a company — Winton Capital Management — who sponsored monetary prizes for the competitions. In return we allowed them to offer job interviews for data science positions to winners of the competitions. This is a win-win-win scenario: astronomy wins because we get good competitors solving our problems, the sponsoring company wins because they can reduce recruitment costs by getting access to people who have proven their data science skills, and the competitors win because they get a chance to solve astronomy mysteries and earn a position on the leader board while (hopefully) having fun.

During our competitions, new ideas were revealed for successfully solving the problems, leading to scientific papers describing the results. While the exact implementation of these ideas needed refinement before applying to real data, the kernel of these new directions for analysis began on the Kaggle leader board. However, it was pointed out in one competition — in a not entirely tongue in cheek manner — that the best way to win would be to model not the data, but the person who created the competition, in order to determine what assumptions were made. Indeed, from my perspective as a competition setter this is an astute observation, and one that future challenges should carefully consider.

But is running competitions the best way to generate new ideas and advance science? Organizing a competition is stressful, and while the majority of competitors are collaborative some competitors get, well… competitive — which is not always a pleasant experience. More broadly, the approach of crowdsourcing scientific advances to a competition could be seen as a step towards a more market-driven approach to science, particularly when prizes are monetary in nature. In a survival of the fittest, almost everyone ends up a loser on the leader board. So perhaps it would be better to collaborate rather than compete, and Kaggle is now moving towards that model by encouraging people to upload and explore datasets together.

A broader question is whether crowdsourced competitions are a step towards a mode of science in which scientists only ever ask the questions, but rely on others to find the answers. Taking this further: we use human competitors now, but perhaps in the future AI competitors will answer the questions we set. This reminds me of Isaac Asimov's Multivac stories, where scientists are those who ask the right questions to an AI, which provides the answers.

But for the moment at least, by admitting our limited knowledge and reaching out for help, these competitions provide a data playground in which problems can be solved and new friendships formed. ❐

Thomas Kitching
Mullard Space Science Laboratory, UCL, London, UK.
e-mail: t.kitching@ucl.ac.uk

Published online: 11 February 2019
https://doi.org/10.1038/s42256-018-0016-x

References

1. Mapping dark matter. *Kaggle* https://www.kaggle.com/c/mdm (2012).
2. Observing dark worlds. *Kaggle* https://www.kaggle.com/c/DarkWorlds (2013).
3. Galaxy zoo. *Kaggle* https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge (2014).
4. PLAsTiCC astronomical classification. *Kaggle* https://www.kaggle.com/c/PLAsTiCC-2018 (2018).